SYLLABUS OF BASIC EDUCATION

Complete Online Text References

Fall 2022

Exam 8

V05 2022_06_10

ACTUARIAL STANDARDS BOARD

Actuarial Standard
of Practice
No. 12

Risk Classification (for All Practice Areas)

Revised Edition

Developed by the
Task Force to Revise ASOP No. 12 of the
General Committee of the
Actuarial Standards Board

Adopted by the
Actuarial Standards Board
December 2005
Updated for Deviation Language Effective May 1, 2011

(Doc. No. 132)

ASOP No. 12—December 2005

T A B L E O F C O N T E N T S

APPENDIXES

ASOP No. 12—December 2005

December 2005

TO: Members of the American Academy of Actuaries and Other Persons Interested in Risk Classification (for All Practice Areas)

FROM: Actuarial Standards Board (ASB)

SUBJ: Actuarial Standard of Practice (ASOP) No. 12

This booklet contains the final version of a revision of ASOP No. 12, now titled *Risk Classification (for All Practice Areas).*

Background

In 1989, the Actuarial Standards Board adopted the original ASOP No. 12, then titled *Concerning Risk Classification*. The original ASOP No. 12 was developed as the need for more formal guidance on risk classification increased as the selection process became more complex and more subject to public scrutiny. In light of the evolution in practice since then, as well as the adoption of a new format for standards, the ASB believed it was appropriate to revise this standard in order to reflect current generally accepted actuarial practice.

Exposure Draft

The exposure draft of this ASOP was approved for exposure in September 2004 with a comment deadline of March 15, 2005. Twenty-two comment letters were received and considered in developing the final standard. A summary of the substantive issues contained in the exposure draft comment letters and the responses are provided in appendix 2.

The most significant changes from the exposure draft were as follows:

1. The task force clarified language relating to the interaction of applicable law and this standard.

2. The task force revised the definition of "adverse selection."

3. The task force reworded the definition of "financial or personal security system" and included examples.

4. The words "equitable" and "fair" were added in section 3.2.1 but defined in a very limited context that is applicable only to rates.

5. With respect to the operation of the standard, the task force added language that clarifies that this standard in all respects applies only to professional services with respect to designing, reviewing, or changing risk classification systems.

6. Sections 4.1 and 4.2 were combined into a new section 4.1, Communications and Disclosures, which was revised for clarity. The placement of communication requirements throughout the proposed standard was examined, and a sentence regarding disclosure was removed from section 3.3.3 and incorporated into section 4.1. A similar change was made by adding a new sentence in section 4.1 to correspond to the guidance in section 3.4.1.

   In addition, the disclosure requirement in section 4 for the actuary to consider providing quantitative analyses was removed and replaced by a new section 3.4.4, which guides the actuary to consider performing such analyses, depending on the purpose, nature, and scope of the assignment.

The task force thanks everyone who took the time to contribute comments on the exposure draft.

The ASB voted in December 2005 to adopt this standard.

Task Force to Revise ASOP No. 12

Mark E. Litow, Chairperson

David J. Christianson
Arnold A. Dicke
Paul R. Fleischacker
Joan E. Herman
Barbara J. Lautzenheiser
Charles L. McClenahan
Donna C. Novak
Ronnie Susan Thierman
Kevin B. Thompson

General Committee of the ASB

W.H. Odell, Chairperson

Charles A. Bryan
Thomas K. Custis
Burton D. Jay
Mark E. Litow
Chester J. Szczepanski
Ronnie Susan Thierman

ASOP No. 12—December 2005

Actuarial Standards Board

Michael A. LaMonica, Chairperson

Cecil D. Bykerk	William A. Reimert
William C. Cutlip	Lawrence J. Sher
Lew H. Nathan	Karen F. Terry
Godfrey Perrott	William C. Weller

ACTUARIAL STANDARD OF PRACTICE NO. 12

RISK CLASSIFICATION (FOR ALL PRACTICE AREAS)

STANDARD OF PRACTICE

Section 1. Purpose, Scope, Cross References, and Effective Date

1.1 Purpose—This actuarial standard of practice (ASOP) provides guidance to actuaries when performing professional services with respect to designing, reviewing, or changing risk classification systems.

1.2 Scope—This standard applies to all actuaries when performing professional services with respect to designing, reviewing, or changing risk classification systems used in connection with financial or personal security systems, as defined in section 2.4, regarding the classification of individuals or entities into groups intended to reflect the relative likelihood of expected outcomes. Such professional services may include expert testimony, regulatory activities, legislative activities, or statements concerning public policy, to the extent these activities involve designing, reviewing, or changing a risk classification system used in connection with a specific financial or personal security system.

Throughout this standard, any reference to performing professional services with respect to designing, reviewing, or changing a risk classification system also includes giving advice with respect to that risk classification system.

Risk classification can affect and be affected by many actuarial activities, such as the setting of rates, contributions, reserves, benefits, dividends, or experience refunds; the analysis or projection of quantitative or qualitative experience or results; underwriting actions; and developing assumptions, for example, for pension valuations or optional forms of benefits. This standard applies to actuaries when performing such activities to the extent such activities directly or indirectly involve designing, reviewing, or changing a risk classification system. This standard also applies to actuaries when performing such activities to the extent that such activities directly or indirectly are likely to have a material effect, in the actuary's professional judgment, on the intended purpose or expected outcome of the risk classification system.

If the actuary departs from the guidance set forth in this standard in order to comply with applicable law (statutes, regulations, and other legally binding authority), or for any other reason the actuary deems appropriate, the actuary should refer to section 4.

1.3 Cross References—When this standard refers to the provisions of other documents, the reference includes the referenced documents as they may be amended or restated in the

future, and any successor to them, by whatever name called. If any amended or restated document differs materially from the originally referenced document, the actuary should consider the guidance in this standard to the extent it is applicable and appropriate.

1.4 Effective Date—This standard will be effective for any professional service commenced on or after May 1, 2006.

Section 2. Definitions

The terms below are defined for use in this actuarial standard of practice.

2.1 Advice—An actuary's communication or other work product in oral, written, or electronic form setting forth the actuary's professional opinion or recommendations concerning work that falls within the scope of this standard.

2.2 Adverse Selection—Actions taken by one party using risk characteristics or other information known to or suspected by that party that cause a financial disadvantage to the financial or personal security system *(sometimes referred to as antiselection).*

2.3 Credibility—A measure of the predictive value in a given application that the actuary attaches to a particular body of data (predictive is used here in the statistical sense and not in the sense of predicting the future).

2.4 Financial or Personal Security System—A private or governmental entity or program that is intended to mitigate the impact of unfavorable outcomes of contingent events. Examples of financial or personal security systems include auto insurance, homeowners insurance, life insurance, and pension plans, where the mitigation primarily takes the form of financial payments; prepaid health plans and continuing care retirement communities, where the mitigation primarily takes the form of direct service to the individual; and other systems, where the mitigation may be a combination of financial payments and direct services.

2.5 Homogeneity—The degree to which the expected outcomes within a risk class have comparable value.

2.6 Practical—Realistic in approach, given the purpose, nature, and scope of the assignment and any constraints, including cost and time considerations.

2.7 Risk(s)—Individuals or entities covered by financial or personal security systems.

2.8 Risk Characteristics—Measurable or observable factors or characteristics that are used to assign each risk to one of the risk classes of a risk classification system.

2.9 Risk Class—A set of risks grouped together under a risk classification system.

2.10 Risk Classification System—A system used to assign risks to groups based upon the expected cost or benefit of the coverage or services provided.

Section 3. Analysis of Issues and Recommended Practices

3.1 Introduction—This section provides guidance for actuaries when performing professional services with respect to designing, reviewing, or changing a risk classification system. Approaches to risk classification can vary significantly and it is appropriate for the actuary to exercise considerable professional judgment when providing such services, including making appropriate use of statistical tools. Sections 3 and 4 are intended to provide guidance to assist the actuary in exercising professional judgment when applying various acceptable approaches.

3.2 Considerations in the Selection of Risk Characteristics—Risk characteristics are important structural components of a risk classification system. When selecting which risk characteristics to use in a risk classification system, the actuary should consider the following:

3.2.1 Relationship of Risk Characteristics and Expected Outcomes—The actuary should select risk characteristics that are related to expected outcomes. A relationship between a risk characteristic and an expected outcome, such as cost, is demonstrated if it can be shown that the variation in actual or reasonably anticipated experience correlates to the risk characteristic. In demonstrating a relationship, the actuary may use relevant information from any reliable source, including statistical or other mathematical analysis of available data. The actuary may also use clinical experience and expert opinion.

Rates within a risk classification system would be considered equitable if differences in rates reflect material differences in expected cost for risk characteristics. In the context of rates, the word *fair* is often used in place of the word *equitable*.

The actuary should consider the interdependence of risk characteristics. To the extent the actuary expects the interdependence to have a material impact on the operation of the risk classification system, the actuary should make appropriate adjustments.

Sometimes it is appropriate for the actuary to make inferences without specific demonstration. For example, it might not be necessary to demonstrate that persons with seriously impaired, uncorrected vision would represent higher risks as operators of motor vehicles.

3.2.2 Causality—While the actuary should select risk characteristics that are related to expected outcomes, it is not necessary for the actuary to establish a cause and

effect relationship between the risk characteristic and expected outcome in order to use a specific risk characteristic.

3.2.3 Objectivity—The actuary should select risk characteristics that are capable of being objectively determined. A risk characteristic is objectively determinable if it is based on readily verifiable observable facts that cannot be easily manipulated. For example, a risk classification of "blindness" is not objective, whereas a risk classification of "vision corrected to no better than 20/100" is objective.

3.2.4 Practicality—The actuary's selection of a risk characteristic should reflect the tradeoffs between practical and other relevant considerations. Practical considerations that may be relevant include, but are not limited to, the cost, time, and effort needed to evaluate the risk characteristic, the ongoing cost of administration, the acceptability of the usage of the characteristic, and the potential usage of different characteristics that would produce equivalent results.

3.2.5 Applicable Law—The actuary should consider whether compliance with applicable law creates significant limitations on the choice of risk characteristics.

3.2.6 Industry Practices—When selecting risk characteristics, the actuary should consider usual and customary risk classification practices for the type of financial or personal security system under consideration.

3.2.7 Business Practices—When selecting risk characteristics, the actuary should consider limitations created by business practices related to the financial or personal security system as known to the actuary and consider whether such limitations are likely to have a significant impact on the risk classification system.

3.3 Considerations in Establishing Risk Classes—A risk classification system assigns each risk to a risk class based on the results of measuring or observing its risk characteristics. When establishing risk classes for a financial or personal security system, the actuary should consider and document any known significant choices or judgments made, whether by the actuary or by others, with respect to the following:

3.3.1 Intended Use—The actuary should select a risk classification system that is appropriate for the intended use. Different sets of risk classes may be appropriate for different purposes. For example, when setting reserves for an insurance coverage, the actuary may choose to subdivide or combine some of the risk classes used as a basis for rates.

3.3.2 Actuarial Considerations—When establishing risk classes, the actuary should consider the following, which are often interrelated:

a. Adverse Selection—If the variation in expected outcomes within a risk class is too great, adverse selection is likely to occur. To the extent practical, the actuary should establish risk classes such that each has sufficient homogeneity with respect to expected outcomes to satisfy the purpose for which the risk classification system is intended.

b. Credibility—It is desirable that risk classes in a risk classification system be large enough to allow credible statistical inferences regarding expected outcomes. When the available data are not sufficient for this purpose, the actuary should balance considerations of predictability with considerations of homogeneity. The actuary should use professional judgment in achieving this balance.

c. Practicality—The actuary should use professional judgment in balancing the potentially conflicting objectives of accuracy and efficiency, as well as in minimizing the potential effects of adverse selection. The cost, time, and effort needed to assign risks to appropriate risk classes will increase with the number of risk classes.

3.3.3 Other Considerations—When establishing risk classes, the actuary should (a) comply with applicable law; (b) consider industry practices for that type of financial or personal security system as known to the actuary; and (c) consider limitations created by business practices of the financial or personal security system as known to the actuary.

3.3.4 Reasonableness of Results—When establishing risk classes, the actuary should consider the reasonableness of the results that proceed from the intended use of the risk classes (for example, the consistency of the patterns of rates, values, or factors among risk classes).

3.4 Testing the Risk Classification System—Upon the establishment of the risk classification system and upon subsequent review, the actuary should, if appropriate, test the long-term viability of the financial or personal security system. When performing such tests subsequent to the establishment of the risk classification system, the actuary should evaluate emerging experience and determine whether there is any significant need for change.

3.4.1 Effect of Adverse Selection—Adverse selection can potentially threaten the long-term viability of a financial or personal security system. The actuary should assess the potential effects of adverse selection that may result or have resulted from the design or implementation of the risk classification system. Whenever the effects of adverse selection are expected to be material, the actuary should, when

practical, estimate the potential impact and recommend appropriate measures to mitigate the impact.

3.4.2 Risk Classes Used for Testing—The actuary should consider using a different set of risk classes for testing long-term viability than was used as the basis for determining the assigned values if this is likely to improve the meaningfulness of the tests. For example, if a risk classification system is gender-neutral, the actuary might separate the classes based on gender when performing a test of long-term viability.

3.4.3 Effect of Changes—If the risk classification system has changed, or if business or industry practices have changed, the actuary should consider testing the effects of such changes in accordance with the guidance of this standard.

3.4.4 Quantitative Analyses—Depending on the purpose, nature, and scope of the assignment, the actuary should consider performing quantitative analyses of the impact of the following to the extent they are generally known and reasonably available to the actuary:

a. significant limitations due to compliance with applicable law;

b. significant departures from industry practices;

c. significant limitations created by business practices of the financial or personal security system;

d. any changes in the risk classes or the assigned values based upon the actuary's determination that experience indicates a significant need for a change; and

e. any expected material effects of adverse selection.

3.5 Reliance on Data or Other Information Supplied by Others—When relying on data or other information supplied by others, the actuary should refer to ASOP No. 23, *Data Quality*, for guidance.

3.6 Documentation—The actuary should document the assumptions and methodologies used in designing, reviewing, or changing a risk classification system in compliance with the requirements of ASOP No. 41, *Actuarial Communications*. The actuary should also prepare and retain documentation to demonstrate compliance with the disclosure requirements of section 4.1.

Section 4. Communications and Disclosures

4.1 Communications and Disclosures—When issuing actuarial communications under this standard, the actuary should comply with ASOP Nos. 23 and 41. In addition, the actuarial communications should disclose any known significant impact resulting from the following to the extent they are generally known and reasonably available to the actuary:

a. significant limitations due to compliance with applicable law;

b. significant departures from industry practices;

c. significant limitations created by business practices related to the financial or personal security system;

d. a determination by the actuary that experience indicates a significant need for change, such as changes in the risk classes or the assigned values; and

e. expected material effects of adverse selection;

f. the disclosure in ASOP No. 41, section 4.2, if any material assumption or method was prescribed by applicable law (statutes, regulations, and other legally binding authority);

g. the disclosure in ASOP No. 41, section 4.3, if the actuary states reliance on other sources and thereby disclaims responsibility for any material assumption or method selected by a party other than the actuary; and

h the disclosure in ASOP No. 41, section 4.4, if, in the actuary's professional judgment, the actuary has otherwise deviated materially from the guidance of this ASOP.

The actuarial communications should also disclose any recommendations developed by the actuary to mitigate the potential impact of adverse selection.

Appendix 1

Background and Current Practices

Note: The following appendix is provided for informational purposes but is not part of the standard of practice.

Background

Risk classification has been a fundamental part of actuarial practice since the beginning of the profession. The financial distress and inequity that can result from ignoring the impact of differences in risk characteristics was dramatically illustrated by the failure of the nineteenth-century assessment societies, where life insurance was provided at rates that disregarded age. Failure to adhere to actuarial principles regarding risk classification for voluntary coverages can result in underutilization of the financial or personal security system by, and thus lack of coverage for, lower risk individuals, and can result in coverage at insufficient rates for higher risk individuals, which threatens the viability of the entire system.

Adverse selection may result from the design of the classification system, or may be the result of externally mandated constraints on risk classification. Classes that are overly broad may produce unexpected changes in the distribution of risk characteristics. For example, if an insurer chooses not to screen for a specific risk characteristic, or a jurisdiction precludes screening for that characteristic, this may result in individuals with the characteristic applying for coverage in greater numbers and/or amounts, leading to increased overall costs.

Risk classification is generally used to treat participants with similar risk characteristics in a consistent manner, to permit economic incentives to operate and thereby encourage widespread availability of coverage, and to protect the soundness of the system.

The following actuarial literature provides additional background and context with respect to risk classification:

1. In 1957, the Society of Actuaries published *Selection of Risks* by Pearce Shepherd and Andrew Webster, which educated several generations of actuaries and is still a useful reference.

2. In 1980, the American Academy of Actuaries published the *Risk Classification Statement of Principles*, which has enjoyed widespread acceptance in the actuarial profession. At the time of this revision of ASOP No. 12, the American Academy of Actuaries was developing a white paper regarding risk classification principles.

3. In 1992, the Committee on Actuarial Principles of the Society of Actuaries published "Principles of Actuarial Science," which discusses risk classification in the context of the principles on which actuarial science is based.

Current Practices

Over the years, a multitude of risk classification systems have been designed, put into use, and modified as a result of experience. Advances in medical science, economics, and other disciplines, as well as in actuarial science itself, are likely to result in continued evolution of these systems. While future developments cannot be foreseen with accuracy, practicing actuaries can take reasonable steps to keep abreast of emerging and current practices. These practices may vary significantly by area of practice. For example, the risk classes for voluntary life insurance may be subdivided to reflect the applicant's state of health, smoking habits, and occupation, while these factors are usually not considered in pension systems.

Innovations in risk classification systems may engender considerable controversy. The potential use of genetic tests to classify risks for life and health insurance is a current example. In some cases, such controversy results in legislation or regulation. The use of postal codes, for example, has been outlawed for some types of coverage. For the most part, however, the legal test for risk classification has remained unchanged for several decades; risk classification is allowed so long as it is "based on sound actuarial principles" and "related to actual or reasonably anticipated experience."

Risk classification issues in some instances may pose a dilemma for an actuary working in the public policy arena when political considerations support a system that contradicts to some degree practices called for in this ASOP. Also, when designing, reviewing, or changing a risk classification system, actuaries may perform professional services related to a designated set of specific assumptions that place certain restraints on the risk classification system.

In such situations, it is important for those requesting such professional services to have the benefit of professional actuarial advice.

This ASOP is not intended to prevent the actuary from performing professional services in the situations described above. In such situations, the communication and disclosure guidance in section 4.1 will be particularly pertinent, and current section 4.1(e), which requires disclosure of any known significant impact resulting from expected material effects of adverse deviation, may well apply. Section 4.1(a), which relates to applicable law, and section 4.1(b), which relates to industry practices, may also be pertinent.

Appendix 2

Comments on the Exposure Draft and Responses

The exposure draft of this revision of ASOP No. 12, *Risk Classification for All Practice Areas*, was issued in September 2004 with a comment deadline of March 15, 2005. Twenty-two comment letters were received, some of which were submitted on behalf of multiple commentators, such as by firms or committees. For purposes of this appendix, the term "commentator" may refer to more than one person associated with a particular comment letter. The task force carefully considered all comments received. Summarized below are the significant issues and questions contained in the comment letters and the responses, which may have resulted from ASB, General Committee, or task force discussion. Unless otherwise noted, the section numbers and titles used below refer to those in the exposure draft.

GENERAL COMMENTS	
Comment	Several commentators suggested various editorial changes in addition to those addressed specifically below.
Response	The task force implemented such suggestions if they enhanced clarity and did not alter the intent of the section.
Comment	One commentator noted that the ASOP should deal with the ability of an insured to misrepresent or manipulate its classification.
Response	The task force believed that the considerations raised by the commentator are adequately addressed by sections 3.2.3 and 3.2.4.
Comment	One commentator thought that a section on public and social policy considerations should be added to the standard.
Response	The task force believed that social and public policy considerations, while essential aspects of the way the public views the profession, did not belong in an ASOP dealing with the actuarial aspects of risk classification.
Comment	One commentator questioned whether the ASOP would apply to company selection criteria (tiering criteria) and schedule-rating criteria that may be part of a rating scheme.
Response	The task force believes that the ASOP applies to the extent the selection or schedule rating criteria, used by a company as part of the risk classification system, creates the potential for adverse selection.
Comment	One commentator believed that the ASOP could conflict with proposed state legislation to ban credit as a rating variable and suggested adding an additional consideration in section 3 that the actuary should select risk characteristics in order to avoid controversy or lawsuits.
Response	The task force believes it has addressed issues regarding applicable law, industry practices, business practices, and testing the risk classification system under various scenarios.
Comment	In the transmittal memorandum of the exposure draft, the task force asked whether the key changes from the previous standard were appropriate.
Response	Several commentators responded that the changes were appropriate and some suggested additional changes that are discussed in this appendix.

Comment	One commentator expressed concern regarding the expansion of scope and the implications in actuarial work that would be otherwise unrelated to risk classification and the expansion of scope to the public policy arena in general.
Response	The task force has added modified wording in the standard to clarify that in all cases the standard applies only in respect to design, reviewing, or changing risk classification systems related to financial or personal security systems.
Comment	Two commentators believed that the revised standard should discuss the purposes of risk classification similar to the discussion in the previous standard. One commentator noted the discussion about encouraging "widespread availability of coverage" in particular.
Response	The task force retained a brief discussion of the purposes of risk classification in appendix 1 but did not believe it was appropriate for the ASOP to provide additional education about the purposes of risk classification. The task force noted that a white paper on risk classification that could contain such material is being developed.
Comment	Several commentators noted that the previous ASOP No. 12 had been very useful in court proceedings and recommended that the task force retain some of the wording in section 5 of the previous ASOP. One commentator suggested strengthening the revised standard so that actuarial testimony would be given greater weight by the courts in interpreting rate standards. Another commentator suggested strengthening the ASOP by adding an explicit statement that one objective during the development and use of risk classification systems is to minimize adverse selection.
Response	The task force reviewed the revised standard with these concerns in mind but concluded that the revised standard represents current generally accepted practice and provides an appropriate level of guidance. The task force considered the specific suggestions with respect to additional wording and incorporated some of the wording regarding adverse selection from the old section 5.5 into appendix 1.
Comment	In the transmittal memorandum of the exposure draft, the task force asked whether it was appropriate for the ASOP not to use the terms "equitable" and "fair." Two commentators believed that the ASOP should use or define these concepts because they have been used in court proceedings, but the majority of commentators believed that it was appropriate not to define them and that the standard adequately addressed these concepts.
Response	The task force agreed that the ASOP should not define subjective qualities such as "equitable" and "fair." As the result of ASB deliberation on this issue, language was added to section 3.2.1 to discuss what was meant by the terms "equitable" and "fair." These terms are intended to apply to a risk classification system only to the extent the risk classification system applies to rates. As such, a formal definition was not added. Court decisions notwithstanding, there is no general agreement as to what characterizes "equitable" classification systems or "fair" discrimination. The task force also considered the possibility that further discussions about such issues might become part of the proposed white paper on risk classification that the American Academy of Actuaries is developing.
Comment	One commentator questioned why the standard offered separate guidance for "risk characteristics" (section 3.2) and "risk classes" (section 3.3). Another commentator believed there should be greater differentiation between the concepts of "risk characteristic" and "risk classification."
Response	The task force believed that the ASOP uses these terms appropriately and made no change.
Comment	One commentator thought that section 3.3.2 should include guidance on appropriately matching the risk with the outcome when establishing a risk class.
Response	The task force believed that section 3.2.1 addressed this comment and noted that section 3.3.2(a) requires sufficient homogeneity with respect to outcomes.

Section 1.2, Scope	
Comment	In the transmittal memorandum of the exposure draft, the task force asked whether it was appropriate to include the actuary's advice within the scope of the standard. Several commentators agreed that including guidance on actuarial advice was appropriate. One commentator believed that the disclosure requirements in section 4 could be burdensome to an actuary who has provided brief oral advice.
Response	The task force kept actuarial advice within the scope of the standard and intended that the disclosure requirements in section 4 should apply to any actuarial advice that falls within the scope of the standard.
Comment	One commentator questioned what was meant by "legislative activities" as an example of a professional service.
Response	The task force intended that "legislative activities" could include drafting legislation, for example.
Comment	Several commentators questioned the meaning of "personal security system." One commentator questioned whether the definition of "financial or personal security system" would exclude share-based payment systems from the scope of the standard. The commentator recommended that the standard be revised to include such systems.
Response	The task force intended that the ASOP should apply if share-based payment systems or stock options were part of a financial or personal security system, as defined in the section 2.5. If such plans were not part of a financial or personal security system, the ASOP would not apply. The task force chose not to expand the scope to include such plans in all situations but did clarify the definition of "financial or personal security system."
SECTION 2. DEFINITIONS	
Comment	One commentator suggested that a definition of experience be included, citing the definition of "experience" in the previous ASOP (old section 2.5), which includes the wording, "Experience may include estimates where data are incomplete or insufficient."
Response	The task force agreed that experience may include estimates where data are incomplete or insufficient but did not believe that the old definition was necessary in the revised ASOP.
Comment	One commentator suggested that a definition of "reasonable" be included.
Response	The task force disagreed and did not add a definition of "reasonable."
Section 2.1, Advice	
Comment	One commentator suggested that "other work product" was not needed, since the standard already listed "an actuary's oral, written, or electronic communication."
Response	The task force revised the language to clarify that "communication or other work product" was intended.
Comment	One commentator believed that a definition for "advice" is not needed.
Response	The task force disagreed and retained the definition of advice.
Section 2.2, Adverse Selection	
Comment	In the transmittal memorandum of the exposure draft, the task force asked if the definition of "adverse selection" was appropriate or whether an alternative definition (included in the transmittal letter) would be preferable. Many commentators responded, some agreeing with the original, some with the alternative, and some suggested other wording. The other wording was most often to change the phrase, "take financial advantage of."
Response	The task force believed that some of the reasoning on the part of the commentators who preferred the current version did not accurately describe adverse selection. The task force ultimately decided to use the alternative definition in the standard and believed that it better addressed some commentators' concerns that the other definition could have a negative connotation with respect to motivation.

Comment	One commentator suggested that "antiselection" is synonymous with adverse selection and that should be made clear in the definition.
Response	The task force agreed and added that reference.
Section 2.4, Credibility (now 2.3)	
Comment	Two commentators believed that within the definition of "credibility" the language concerning "predictive" was confusing.
Response	The task force retained the definition as it is used in several other ASOPs.
Section 2.5, Financial or Personal Security System (now 2.4)	
Comment	Several commentators questioned the meaning of "personal security system."
Response	The task force clarified the definition.
Comment	One commentator suggested that "impact" be modified to read "financial impact."
Response	The task force disagreed and revised the definition of "financial and security systems" to delineate the impacts.
Section 2.6, Homogeneity (now 2.5)	
Comment	One commentator believed the definition of "homogeneity" needed revisions to include the concept of grouping similar risks. Another commentator found the definition unclear.
Response	The task force believes that the current definition is appropriate for this ASOP.
Section 2.7, Practical (now 2.6)	
Comment	One commentator believed the definition of "practical" was much too broad and needed to be more actuarial in nature. Alternatively, the commentator suggested dropping it and relying on section 3.2.4.
Response	The task force believed the definition was appropriate and made no change. Section 3.2.4 addresses actuarial practice with respect to practicality. While "practical" is used there and in other places, it is always modified by its context.
Section 2.8, Risk(s) (now 2.7)	
Comment	One commentator suggested that the definition of risks as individuals or entities seemed too limiting and noted that covered risks can also include pieces of property or events.
Response	The task force disagreed, believing that "entity" could encompass property and events.
Comment	One commentator suggested that a unit of risk be defined at the basic unit of risk.
Response	The task force disagreed and made no change.
Section 2.9, Risk Characteristics (now 2.8)	
Comment	One commentator suggested defining risk characteristics as "measurable or observable factors or characteristics, each of which is measured by grouping similar risks into risk classes."
Response	The task force disagreed and made no change.
Section 2.11, Risk Classification System (now 2.10)	
Comment	One commentator believes the definition of "risk classification system" is circular since "classify" is used in the definition.
Response	The task force agreed and revised the wording.
Comment	One commentator recommended that the term "risks" be changed to "similar risks" in this definition just as in the old definition of risk classification that used the phrase "grouping risks with similar risk characteristics."
Response	The task force disagreed and made no change.
Comment	One commentator suggested replacing "groups" with "classes."
Response	The task force disagreed and made no change.

SECTION 3. ANALYSIS OF ISSUES AND RECOMMENDED PRACTICES	
Section 3.2.1, Relationship of Risk Characteristics and Expected Outcomes	
Comment	One commentator expressed concern with the standard's differentiation between the section's quantitative and subjective factors.
Response	The task force did not intend to be prescriptive as to how to quantify the ratings scheme and believed that the ASOP was sufficiently specific. The ASOP does not address rate adequacy. Selection is the focus, not quantification.
Comment	One commentator believed that "clinical" was not an appropriate adjective to describe the experience an actuary is allowed to use.
Response	The task force intentionally used the term "clinical."
Comment	One commentator believed that if the classification cannot be measured by actual insurance data, then it is not really a risk classification system.
Response	The task force disagreed and made no change.
Comment	One commentator suggested that the three points addressing why risk classification is generally used be moved to background information.
Response	The task force agreed that such educational language was more appropriate in an appendix than in the body of the ASOP and has moved it.
Comment	One commentator believed that it may be difficult to deal with the process and procedures involved with considering the interdependence of risk characteristics and their potential impact on the operation of the risk classification system.
Response	The task force did not change the language to address this comment but notes that section 3.2.4 addresses considerations regarding practicality.
Section 3.2.2, Causality	
Comment	A number of commentators expressed concern with establishing a cause-and-effect relationship while others thought the standard did not go far enough in this regard.
Response	The task force agreed that, where there is a demonstrable cause-and-effect relationship between a risk characteristic and the expected outcome, it is appropriate for the actuary to include such a demonstration. However, the task force recognized that there can be significant relationships between risk characteristics and expected outcomes where a cause-and-effect relationship cannot be demonstrated.
Section 3.2.4, Practicality	
Comment	Two commentators suggested the use of examples of practical considerations.
Response	The task force revised the section to indicate that the language shows examples of practical considerations.
Comment	One commentator suggested that "theoretical," as used in section 3.2.4, be defined.
Response	The task force replaced "theoretical" with "other relevant."
Section 3.2.5, Applicable Law	
Comment	One commentator thought that the proposed language in this section was much too broad.
Response	The task force disagreed with the comment and made no change.

Section 3.3, Considerations in Establishing Risk Classes	
Comment	One commentator expressed concern that the documentation requirements for these considerations represented an increase from the previous version.
Response	The task force thought the documentation requirements were appropriate and necessary and made no change.
Section 3.3.1, Intended Use	
Comment	One commentator noted that stratifying data sets in loss reserving is different from risk classification, which is done to price risks, and believed that loss reserving permits more flexibility. The commentator stated that the definition of a risk classification system does not apply to loss reserving.
Response	The task force agreed with the first concepts but disagreed with the final sentence and therefore made no change.
Section 3.3.2, Actuarial Considerations	
Comment	With respect to section 3.3.2(a), one commentator suggested replacing the word "for" in the first line with "within" for clarification.
Response	The task force agreed and made the suggested change.
Comment	With respect to section 3.3.2(b), two commentators questioned what was intended by the use of the term "large enough."
Response	The task force believed the language was sufficiently clear and made no change.
Comment	One commentator pointed out that there are often classes that, individually, have associated experience with low statistical credibility and believed that alternatives to credibility should be included in section 3.3.2(b).
Response	While the task force agreed that there are situations in which actuarially sound classification plans will have individual classes where the experience has low statistical credibility, the task force believed that credibility is a desirable characteristic of risk classes within a risk classification system and that no expansion to include alternatives was necessary.
Comment	One commentator suggested replacing "statistical predictions" with "predictions" in section 3.3.2(b) to avoid the implication that underlying statistics were required. Another commentator suggested that the term "predictions" needed explanation.
Response	The task force agreed with these comments and replaced "predictions" with "inferences" and edited the language to improve its clarity.
Comment	One commentator suggested that the last sentence of section 3.3.2(b), while accurate, was irrelevant.
Response	The task force agreed and eliminated the sentence.
Comment	With respect to section 3.3.2(c), one commentator suggested the need for definitions of "accuracy" and "efficiency."
Response	The task force believed that the existing language regarding the actuary's professional judgment was sufficient in determining the meaning of "accuracy" and "efficiency" and did not add a definition of either word.

Comment	Several commentators suggested that section 3.3.2(d) be eliminated. A number of those commentators also pointed out that the language was both inconsistent with current actuarial practice and inappropriate as an implied requirement.
Response	The task force agreed and deleted the section.
Section 3.3.3, Other Considerations	
Comment	Several commentators pointed out that the last sentence of the section was unclear and might inadvertently require a degree of testing and determination that was not intended.
Response	The task force deleted the last sentence of the section. In addition, section 4.1, Communications and Disclosures, was clarified as to what disclosures are appropriate.
Section 3.3.4, Reasonableness of Results	
Comment	One commentator found the parenthetical wording confusing.
Response	The task force believed the examples were appropriate and made no change.
Comment	One commentator found this section ambiguous in the context of establishing risk classes. Another commentator suggested that a cost-based definition of reasonable be added or that the section be deleted entirely.
Response	The task force retained the section but clarified the wording by mentioning the intended use of the risk classes. The task force did not believe additional clarification of "reasonableness" was necessary because reasonableness is a subjective concept that may depend on the actuary's professional judgment. The task force also notes that the *Introduction to the Actuarial Standards of Practice* discusses this concept in further detail.
Section 3.4, Testing the Risk Classification System	
Comment	One commentator indicated that it may be preferable to substitute the word "or" for "and" on the second line so that the sentence reads, "Upon establishment of the risk classification system or upon subsequent review. … "
Response	The task force did not agree and believed the word "and" was appropriate because testing should be carried out both upon establishment and upon subsequent review.
Comment	One commentator wanted to substitute "continuing" for "long-term" viability in the second line. The commentator believed that the usual issue is the current and near-future viability of a system, not its long-term prognosis. Also, another commentator said that the requirement to "test long-term viability" is new and questioned its meaning.
Response	The task force considered alternative wording but ultimately decided that the existing wording best reflected that the actuary should check the risk classification system for viability both in the short-term and in the long-term.

Comment	One commentator believed that testing the system is set out as something the actuary should do, if appropriate, rather than as something the actuary should consider. The commentator believed that the paragraph implied a duty to test in some situations, without describing explicitly what those situations would be (i.e., when testing would be "appropriate"). The commentator suspected that the situations described in sections 3.4.1–3.4.3 were the kind of situations that the task force had in mind as situations where long-term testing would be "appropriate." However, as currently written, the commentator thought that a stronger duty could be implied. The commentator suggested that section 3.4 itself should read, "…the actuary should consider testing the long-term viability of the risk classification system. …"
Response	The task force believed that the existing wording conveyed the concept that the actuary considers whether testing is appropriate and made no change.
Section 3.5, Reliance on Data Supplied by Others (now Reliance on Data or Other Information Supplied by Others)	
Comment	One commentator believed that the provision for reliance on data supplied by others was not needed in this ASOP because ASOP No. 23, *Data Quality*, addresses this.
Response	This task force agreed and revised the section to refer to ASOP No. 23, using wording consistent with other recently adopted ASOPs and exposure drafts.
SECTION 4. COMMUNICATIONS AND DISLOSURES	
Section 4.1, Communications (now Communications and Disclosures)	
Comment	One commentator suggested changing the phrase "when issuing actuarial communications under this standard" to "when issuing actuarial communications that include elements of actuarial work within the scope of this standard."
Response	The task force retained the original language to be consistent with other ASOPs.
Section 4.2, Disclosures (now 4.1, Communications and Disclosures)	
Comment	One commentator stated that some of the disclosures, notably section 4.2(a) and 4.2(c) (now 4.1(a) and 4.1(c)), are impractical, since they might require the actuary to begin with the universe and then disclose everything that is not utilized. The commentator suggested replacing these disclosure requirements with a communication that defends the choice of risk classification system and notes in that defense how compliance with applicable law and business practices affected the selection, rather than describing all the alternatives that would have been available in the absence of such constraints.
Response	The task force did not agree that the requirement to disclose significant limitations required a discussion of all alternatives that would have been available in the absence of legal or business constraints. The task force noted that the listed disclosures proceed from considerations required in section 3 and modified the wording of the disclosure requirements to be more consistent with that section, including revising the lead-in sentence to require disclosure of the significant impact of such considerations.
Comment	One commentator stated that the disclosure issue is heightened by the expansion of scope into the public policy arena and stated that excessive disclosure requirements may weaken the actuary's ability to influence the discussion of public policy.
Response	The task force disagreed with the comment and noted that, while the scope of the standard now includes regulatory activities, legislative activities, and statements regarding public policy, the scope does so only in the context of the performance of professional services.

Comment	One commentator suggested deleting section 4.2(a) (now 4.1(a)), which requires disclosure of significant limitations due to compliance with applicable law, noting that other ASOPs have tended not to include this requirement except where the limitations seriously distort the work product.
Response	The task force disagreed with this comment, noting that significant limitations on the choice of risk characteristics are likely to distort the risk classification system and therefore should be disclosed.
Comment	Several commentators expressed opinions regarding the requirement that the actuary should disclose whether quantitative analyses were performed relative to items being disclosed. One commentator expressed strong objection to this requirement, asserting that the requirement would be counter-productive and would reduce the number of quantitative analyses being done. Another commentator agreed and noted that the disclosure issue was heightened by the expansion of scope to the public policy arena, where an advocacy position may be taken. A third commentator objected to the requirement to disclose that quantitative analyses were *not* done but suggested requiring that any analyses that were done be summarized. A fourth commentator suggested exempting certain of the required disclosures from the requirement to consider quantification. A fifth commentator pointed out that, while the actuary was required to disclose whether quantitative analyses were performed, the actuary was only required to consider providing the results of those analyses in the disclosure.
Response	The disclosure requirement for the actuary to consider providing quantitative analyses of the impact of the items being disclosed was removed, and instead similar wording was added as a new section 3.4.4, Quantitative Analyses, which guides the actuary to consider performing such analyses, depending on the purpose, nature, and scope of the assignment.
Comment	In the transmittal letter for the exposure draft in request for comment #6, the task force asked whether there were any situations in which the requirement in section 4.2(c) (now 4.1(c)) to disclose any significant limitations created by business practices of the financial or personal security system would not be appropriate. Two comments were received, both agreeing with the appropriateness of the requirement.
Response	The task force retained the requirement.
Comment	Two commentators suggested substituting "indicates" for "creates" in section 4.2(d) (now 4.1(d)).
Response	The task force agreed, changed the wording as suggested, and made other revisions for clarity.
Comment	In the transmittal letter for the exposure draft in request for comment #7, the task force asked whether the requirement in 4.2(e) (now 4.1(e)) to disclose the effects of adverse selection was appropriate. Three commentators addressed this request for comment, and all agreed the requirement was appropriate. However, one commentator suggested that there be no requirement to quantify the impact.
Response	The task force retained the requirement in what is now 4.1(e) and also removed the requirement to consider providing quantitative analyses. Additionally, the task force deleted section 4.2(f) after determining that it was already covered by ASOP No. 41, Actuarial Communications, to which section 4.1 refers.
APPENDIX (now Appendix 1)	
Comment	One commentator expressed concern with the citing of the textbook *Selection of Risks* by Shepherd and Webster.
Response	The task force believed that citing the Shepherd and Webster book was appropriate but added a new lead-in sentence to the citation to indicate that the references cited provide additional background and context with respect to risk classification.

CAS MONOGRAPH SERIES
NUMBER 2

DISTRIBUTIONS FOR ACTUARIES

David Bahnemann

CASUALTY ACTUARIAL SOCIETY

This monograph contains a brief exposition of the standard probability distributions and their applications by property/casualty actuaries. The focus is on the use of parametric distributions fitted to empirical claim data to solve standard actuarial problems, such as creation of increased limit factors, pricing of deductibles, and evaluating the effect of aggregate limits.

A native of Minnesota, David Bahnemann studied mathematics and statistics at the University of Minnesota and at Stanford University. After teaching mathematics at Northwest Missouri State University for 18 years, he joined the actuarial department at the St. Paul Companies in St. Paul, Minnesota. For the next 25 years he provided actuarial support to several excess and surplus lines underwriting departments. While at the St. Paul (later known as St. Paul Travelers, and then Travelers) he was involved in large-account and program pricing. During this time he also created several computer-based pricing tools for use by both underwriters and actuaries. During retirement he divides his time between White Bear Lake and Burntside Lake in Minnesota.

DISTRIBUTIONS FOR ACTUARIES

David Bahnemann

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, Virginia 22203
www.casact.org
(703) 276-3100

Distributions for Actuaries
By David Bahnemann

Library of Congress Cataloging-in-Publication Data
Bahnemann, David
Distributions for Actuaries / David Bahnemann
978-0-9624762-8-0 (print edition)
978-0-9624762-9-7 (electronic edition)
1. Actuarial science. 2. Distribution (Probability theory) 3. Insurance—Mathematical models.
I. Bahnemann, David.

To Abbie,
Lisa & Greta

2015 CAS Monograph Editorial Board

Contents

Foreword

This is the second monograph in the recently introduced CAS Monograph Series. A CAS monograph is an authoritative, peer reviewed, in-depth work on an important topic within the property and casualty actuarial practice.

In this monograph David Bahnemann brings together two perennially important elements of actuarial practice: a solid academic presentation of parametric distributions coupled with the application of these distributions in the actuarial paradigm.

Bahnemann taught mathematics at the university level for nineteen years, thus developing an excellent appreciation for what works and what does not work in presenting and conveying technical subject matter. Following that, he worked for more than two decades in applying this knowledge to all types of real actuarial problems that actuaries face every day. Hence, we have this rare presentation of mathematics that actuaries use whenever distributions are involved.

This monograph is useful for those wishing to learn the subject matter for the first time as well as for practicing actuaries who wish to have in their bookcase a "desk reference manual" for use whenever faced with a problem involving parametric distributions.

This work clearly is a labor of love in which Bahnemann has brought together in a single volume his entire professional life experience in this field. The CAS is grateful for his effort in producing this monograph as well as the gift it represents to the CAS and its members.

C. K. "Stan" Khury
Chairperson
Monograph Editorial Board

Preface

This monograph contains a brief exposition of the standard probability distributions—and their fundamental applications—commonly encountered by property/casualty actuaries. Specifically, it includes the basic distributional topics that I had occasion to use during the 25 years I provided actuarial support to the excess and surplus lines underwriting departments at the St. Paul Companies (now Travelers). The emphasis is on a clear, informal presentation of the basic concepts, and there has been no attempt to provide an exhaustive (and possibly, exhausting) compendium of every possible topic and technique. Moreover, the focus is clearly on the use of parametric distributions fitted to empirical claim data to solve standard actuarial problems—creation of increased limit factors, pricing of deductibles, evaluating the effect of aggregate limits, and so on.

A prerequisite for understanding this material is an upper-level undergraduate course in mathematical—that is, calculus-based—probability and statistics, and the mathematical level of this monograph is similar to that in such a course.

I envision two possible uses of this monograph—first, as a study aid when the reader is first learning the material, and later as a handy on-the-shelf reference and source of ideas when faced with a distributional problem. The work contains more than six dozen worked-out illustrative examples and more than 170 problems that can serve as a help in mastering the fundamentals, as well as extending the basic ideas and providing applications beyond those presented in the text.

Chapter 1 contains a brief review of basic concepts from probability and mathematical statistics. Moreover, this chapter also provides an introduction to the notational conventions used throughout the text. Chapters 2 and 3, respectively, introduce the most commonly used probability distributions for claim size and claim counts. Many of the examples in this pair of chapters illustrate methods of fitting a probability distribution from a given parametric distribution family to a set of claim data. Chapter 4 is devoted to the properties of aggregate loss distributions and to some of the standard techniques for approximating values of such distributions. Chapter 5 takes up the concepts of excess claims and layers of insurance, ideas which find application in Chapter 6 to the modeling of such common policy provisions as deductibles and limits.

Projects like this never see the light of day without the assistance of many individuals. I am indebted to former St. Paul/Travelers colleagues David Warren and Nancy Braithwaite, who helped bring the manuscript to the attention of the Casualty Actuarial Society and recruit persons to check the problem solutions. I am grateful to the anonymous reviewers who made valuable suggestions for improvement and to the team of volunteers who

verified the problem answers and identified errors: Kendall McDonald, Ira Robbin, Heidi Holtti, Su Fei Ang, Mikalai Filon, George Schuler, Patrick Filmore, Andrew Scott, Kevin Hanson, and Rachel Larson. At the CAS Donna Royston provided excellent and thoughtful editorial support. I am particularly indebted to Stan Khury, whose enthusiasm for the project was essential. All these generous contributors deserve my heartfelt thanks. Finally, above all, I owe an enormous debt of gratitude to my wife, Abbie, whose encouragement and support never faltered, and without which the manuscript would not have been completed.

David Bahnemann

1. Introduction

Property/casualty insurance policies are written to cover policyholder losses that arise from certain unpredictable events. These events, which occur more or less randomly over time, must happen during the time period the policy is in effect in order to qualify as insured events. To cite just a few possibilities, an insured event could be property damage due to fire or storm, medical treatment due to illness, or personal injury due to accident or professional malpractice. The occurrence of such an event can trigger a claim against the policy.

In order to determine a reasonable premium charge for a policy, actuaries must be able to quantify the random aspects of the underlying claim process. In particular, they must be able to construct appropriate probability models for the incidence and size of claims, topics which are the subjects of Chapters 3 and 2, respectively. We begin here in Chapter 1 with a brief summary of basic probability concepts.

A Note on Notation. In addition to providing a review of the probability prerequisites, this chapter establishes most of the notational conventions used throughout the subsequent chapters. In general, the notation is consistent with standard usage employed by expositors of probability and mathematical statistics. Probability spaces are denoted by upper-case Greek letters and probability events are denoted by upper-case Roman letters. The probability of a general random-variable-related event is usually denoted by $\Pr\{\cdot\}$. As usual, cumulative probability functions are denoted by $F(\cdot)$ and probability density functions by the associated lower-case Roman letter: $f(\cdot)$. For most parametric distributional families, parameters are denoted by lower-case Greek letters. Random variables are denoted by upper-case Roman letters, with X or Y denoting a claim-size variable, N a claim-count variable, and S an aggregate-loss variable. In every case, the introduction of a concept is accompanied by sufficient mathematical display to establish the applicable notational conventions.

1.1. Probability Spaces

Consider an experiment of chance for which the outcome cannot be predicted in advance. For example, tossing a coin and observing whether it lands Heads (H) or Tails (T) is an experiment with a set of two possible, but unpredictable outcomes: $\{H, T\}$. The roll of a single die or pair of dice, the blind selection of objects from a well-mixed collection such as cards from a shuffled deck, the time to failure of an electronic or mechanical component, or the occurrence of an insurance claim—each

can be interpreted as an experiment of chance with outcomes that cannot be predicted in advance.

A set Ω of all possible distinct outcomes of an experiment of chance is called a ***sample space*** for the experiment. Each element ω of Ω is referred to as an ***elementary outcome***. A performance of the experiment, obtaining one of the elementary outcomes as a result, is a ***trial*** of the experiment.

Note that different sets of elementary outcomes may be defined for any given experiment, depending on what attributes of the outcomes are of particular interest. For example, if an experiment consists of tossing a coin twice in succession, then one set of elementary outcomes could consist of all ordered pairs of Heads and Tails:

$$\Omega_1 = \{HH,\ HT,\ TH,\ TT\}.$$

If the order is unimportant, then the elementary outcomes could be the unordered pairs of Heads and Tails: $\Omega_2 = \{\{H,\ H\},\ \{H,\ T\},\ \{T,\ T\}\}$. Alternatively, if only the number of Heads obtained is material, then $\Omega_3 = \{0,\ 1,\ 2\}$ would suffice as a sample space. However, sometimes selecting a sample space for which the elementary outcomes can be assigned equal probabilities makes all subsequent probability calculations easier—see Example 1.2(*a*).

An ***event*** E for an experiment of chance is a subset of the sample space: $E \subseteq \Omega$. If, at a trial of the experiment, outcome $\omega \in \Omega$ is obtained and it also happens that $\omega \in E$, then one says that ***event E has occurred***.

Example 1.1. (***a***) An experiment consists of tossing a coin three times in succession and observing the resulting sequence of Heads and Tails. A sample space consists of eight elementary outcomes, each an ordered triple of *H*s and *T*s:

$$\Omega = \{HHH,\ THH,\ HTH,\ HHT,\ HTT,\ THT,\ TTH,\ TTT\}. \tag{1.1}$$

Thus, if on the first and second tosses the coin falls Heads and on the third Tails, then the outcome of the trial is *HHT*. The event E of obtaining at most one Heads among the three tosses is defined by the set $E = \{HTT,\ THT,\ TTH,\ TTT\}$.

(***b***) Another experiment consists of rolling a pair of dice and observing the number of spots on each die in turn (a *die* being a cube whose six faces are marked with one through six spots). There are 36 elementary outcomes in the sample space:

$$\begin{aligned}\Omega = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.\end{aligned} \tag{1.2}$$

For example, if one observes five spots on the first die and two spots on the second, then the outcome of the trial is (5,2).

The set $E = \{(1,6),\ (2,5),\ (3,4),\ (4,3),\ (5,2),\ (6,1)\}$ defines the event that the sum of the spots is seven. The event

$$F = \{(1,4),(2,4),(3,4),(4,4),(5,4),(6,4),(4,1),(4,2),(4,3),(4,5),(4,6)\}$$

occurs when four spots are obtained on at least one die. Having obtained the outcome (5,2) on a trial, we observe that event E has occurred because $(5,2) \in E$ and that event F has not occurred because $(5,2) \notin F$.

(***c***) An insurance policy pays at most $200,000 for an incurred claim. The issuing of such a policy can be interpreted as a trial of an experiment of chance for which the uncertain outcome is the occurrence (or non-occurrence) of one or more claims. A reasonable set of elementary outcomes would be the number of incurred claims: $\Omega_1 = \{0, 1, 2, 3, \ldots\}$. In addition, the occurrence of a claim can be interpreted as another experiment of chance for which the size of the claim is the unpredictable outcome. For this experiment the sample space can be expressed as an interval of real numbers: $\Omega_2 = [0; 200{,}000]$.[1] ■

Generally, most—but not necessarily all—subsets of Ω can be considered events for an experiment with sample space Ω. In order to define probabilities for the events of an experiment of chance in a reasonable way, the set S of events must have certain properties. In particular, (*i*) S must contain Ω and (*ii*) S must contain the complement $E^c = \{\omega \in \Omega : \omega \notin E\}$ whenever $E \in S$. Moreover, (*iii*) S must contain the union of every countable collection of events in S.[2] A collection of sets with these properties is called a σ-*algebra* or *Borel field*.[3] When the sample space Ω is finite or countably infinite, it is customary to assume that S is just the set of all subsets of Ω. The alternate case for the sample space that is uncountably infinite will be discussed briefly in Section 2.

Consider now an experiment of chance with a sample space Ω and a set of events S. To construct a ***probability space*** (Ω, S, P) for the experiment, one must assign a real number $P(E)$ to each event E—the ***probability*** of the event—that serves as a measure of the likelihood of the event will occur in a trial of the experiment. An event that is certain to occur—that is, the event Ω—is assigned the maximum probability of 1, and all other events have a probability measure between 0 and 1. A real-valued function P defined on the set of events S is called a ***probability set function*** if it satisfies the following three axioms:

$$\mathbf{A_1} \quad P(\Omega) = 1,$$

$$\mathbf{A_2} \quad 0 \le P(E) \le 1 \text{ for } E \in S,$$

$$\mathbf{A_3} \quad \text{If } \{E_1, E_2, E_3, \ldots\} \text{ is a countable collection of disjoint events, that is, } E_i \cap E_j = \varnothing \text{ for } i \ne j, \text{ then } P\left(\bigcup_i E_i\right) = \sum_i P(E_i). \tag{1.3}$$

Other properties of function P can be derived from axioms $\mathbf{A_1}$, $\mathbf{A_2}$, $\mathbf{A_3}$ and the properties of S. Verification of the following set of statements is requested in Problem 1.2.

---

[1] In reality the value of an insurance claim is expressed in whole monetary units (cents or dollars, for example), but it is convenient to assume that all values on the continuous interval are possible outcomes.

[2] We observe the usual convention that a ***countable*** set A contains either a finite or countably infinite number of elements. Set A is ***countably infinite*** if it can be put into one-to-one correspondence with the set of positive integers.

[3] After the French mathematician, Emile Borel (1871–1956). Borel was a pioneer in the development of modern measure theory and the theory of functions. Throughout the period 1905–1950, he published more than 50 papers and several longer works in probability theory.

Properties of P(x)

Assume that E, F ∈ S are events for a probability space (Ω,*S*,*P*). *Then*

$$(\boldsymbol{a})\quad P(E^c) = 1 - P(E). \tag{1.4}$$

$$(\boldsymbol{b})\quad P(\varnothing) = 0. \tag{1.5}$$

$$(\boldsymbol{c})\quad P(E) + P(F) = P(E \cup F) + P(E \cap F). \tag{1.6}$$

$$(\boldsymbol{d})\quad P(E) = P(E \cap F) + P(E \cap F^c). \tag{1.7}$$

$$(\boldsymbol{e})\quad \text{If } E \subseteq F \text{ then } P(E) \le P(F). \tag{1.8}$$

$$(\boldsymbol{f})\quad \text{If } E = \{\omega_1, \omega_2, \omega_3, \ldots\} \text{ is a countable subset of } \Omega, \text{ then } P(E) = \textstyle\sum_i P(\{\omega_i\}). \tag{1.9}$$

There are many ways to assign the probability function P for a probability space (Ω,*S*,*P*). Methods range from those founded on *a priori* assumptions about the underlying experiment to methods based on analyses of sample data.

In the special case in which Ω is a finite set of n elementary outcomes, there are often situations in which the outcomes can be assumed, by *a priori* reasoning based on symmetry arguments, to have equal probabilities: $P(\{\omega\}) = 1/n$ for each $\omega \in \Omega$. For example, in the toss of single fair coin (that is, a coin of uniform composition with a symmetrical shape), it is reasonable to assume that outcomes Heads and Tails are equally probable: $P(H) = P(T) = 1/2$. Similarly, single objects selected blindly ("at random") from a collection of n similar objects can also be assumed to be equally probable. Thus, a specified card drawn from a well-shuffled bridge deck would have probability 1/52.[4]

In the finite case for which the n elementary outcomes are assigned equal probability, the probability $P(E)$ of an event E containing m elementary outcomes can be calculated by the following formula based on property (1.9) above, where $\#(E)$ denotes the number of elements in the set E:

$$P(E) = \sum_{\omega \in E} P(\{\omega\}) = \frac{\#(E)}{\#(\Omega)} = \frac{m}{n}. \tag{1.10}$$

Example 1.2. (***a***) As in Example 1.1(*a*), an experiment involves tossing a coin three times and observing the sequence of Heads and Tails. The sample space Ω is displayed in equation (1.1). If the coin is assumed to be fair, then it makes sense to assign equal probabilities to the eight elementary outcomes in Ω: $P(\{\omega\}) = 1/8$ for each $\omega \in \Omega$. Applying formula (1.10) to the event

$$E = \{HTT, THT, TTH, TTT\}$$

yields the probability of obtaining at most one Head: $P(E) = 4/8 = 0.5000$.

[4] A bridge deck contains 52 distinct playing cards, divided into four suits of 13 cards each: Hearts, Diamonds, Spades, Clubs. Each suit contains 10 numbered cards—Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10—and three Face cards: Jack, Queen, King. Hearts and Diamonds are red cards; Spades and Clubs are black cards.

(***b***) An experiment consists of drawing a single card at random from a bridge deck, so that each of the 52 elementary outcomes is assigned probability 1/52. We define events E and F by

$$E: \textit{card drawn is a Face card} \quad \text{and} \quad F: \textit{card drawn is a Heart.}$$

The probabilities of events E, E^c, F, $E \cap F$, and $E \cup F$ are calculated from formula (1.10):

$$P(E) = \frac{\#(E)}{52} = \frac{12}{52} = 0.2308,$$

$$P\left(E^c\right) = \frac{\#\left(E^c\right)}{52} = \frac{52-12}{52} = 0.7692,$$

$$P(F) = \frac{\#(F)}{52} = \frac{13}{52} = 0.2500,$$

$$P(E \cap F) = \frac{\#(E \cap F)}{52} = \frac{3}{52} = 0.0577,$$

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = \frac{12}{52} + \frac{13}{52} - \frac{3}{52} = \frac{22}{52} = 0.4231.$$

(***c***) An experiment consists of dealing a hand of five cards at random from a standard deck of 52. Since the order of the cards is immaterial, the number of elementary outcomes is given by the combinatoric formula ${}_nC_k = n!/[k!\ (n-k)!]$ for the number of distinct selections (or combinations) of k objects from a collection of n distinguishable objects:

$${}_{52}C_5 = \frac{52!}{5!47!} = 2{,}598{,}960.$$

Let E be the event of obtaining five cards from the same suit, and let F be the event of obtaining no face cards. Thus,

$$P(E) = \frac{(4)({}_{13}C_5)}{{}_{52}C_5} = \frac{5{,}148}{2{,}598{,}960} = 0.0020,$$

$$P(F) = \frac{{}_{40}C_5}{{}_{52}C_5} = \frac{658{,}008}{2{,}598{,}960} = 0.2532,$$

$$P(E \cap F) = \frac{(4)({}_{10}C_5)}{{}_{52}C_5} = \frac{1{,}008}{2{,}598{,}960} = 0.0004. \blacksquare$$

One of the most useful probability concepts is that of ***conditional probability***. Often one has partial information about the result of an experiment of chance, information which can alter the likelihood that a particular event could occur. For instance, consider the experiment of Example 1.1(*b*) involving the roll of a pair of

dice. Assuming that each die is fair, we assign the probability 1/36 to each elementary outcome in (1.2). As a result, the probability of obtaining a total of seven spots (event E in that example) is $P(E) = 6/36 = 0.1667$. However, this probability changes if we know that event F has already occurred, namely, that at least one die shows four spots. In this case, the number of possible elementary outcomes has been reduced from 36 in Ω to only 11 in event F. In addition, there are only two outcomes in E that are also in F—that is, $E \cap F = \{(3,4), (4,3)\}$—and these remain equally probable. Thus, the conditional probability of E given that F has occurred, denoted by $P(E|F)$, is

$$P(E|F) = \frac{\#(E \cap F)}{\#(F)} = \frac{2}{11} = 0.1818.$$

However, the first quotient in this equation could also be expressed as

$$\frac{\#(E \cap F)}{\#(F)} = \frac{\#(E \cap F)/\#(\Omega)}{\#(F)/\#(\Omega)} = \frac{P(E \cap F)}{P(F)},$$

which can be generalized to provide a definition for conditional probability. If $P(F) > 0$, then $P(E|F)$, the probability of event E, given that event F has occurred, is defined by

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{1.11}$$

In addition, one can express (1.11) in the following multiplicative form, which is satisfied even when $P(F) = 0$:

$$P(E \cap F) = P(F) \cdot P(E|F). \tag{1.12}$$

Equation (1.12) is occasionally useful in calculating $P(E \cap F)$, as in Example 1.3(*b*).

Example 1.3. (***a***) An experiment consists of tossing a fair coin two times in succession and observing the resulting sequence of Heads and Tails. The sample space contains four equally probable outcomes: $\Omega = \{HH, HT, TH, TT\}$.

The probability of obtaining two Tails (event E), given that at least one of the coins lands Tails (event F), is therefore

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(\{TT\})}{P(\{HT, TH, TT\})} = \frac{1/4}{3/4} = \frac{1}{3}.$$

(***b***) An urn contains eight white chips and five black chips. Two chips are drawn at random without replacing the first chip before drawing the second—at each draw the chips in the urn are equally likely to be drawn.

Let E_1 denote the event that the first chip is white, and let E_2 denote the event that the second chip is white. Clearly,

$$P(E_1) = \frac{8}{13} \quad \text{and} \quad P(E_2|E_1) = \frac{8-1}{13-1} = \frac{7}{12}.$$

Thus, (1.12) implies that the probability $P(E_1 \cap E_2)$ that both chips are white is

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2|E_1) = \frac{8}{13} \cdot \frac{7}{12} = \frac{56}{156} = 0.3590. \blacksquare$$

It is possible, however, that the occurrence of event F does not alter the probability of E, that is, $P(E \mid F) = P(E)$. In this situation, we have

$$P(E \cap F) = P(E) \cdot P(F). \tag{1.13}$$

Events E and F for which equation (1.13) holds are said to be ***stochastically independent*** (or merely ***independent***) ***events***; otherwise, they are said to be ***dependent events***.

Example 1.4. Consider again the experiment of Example 1.1(*b*), involving the roll of two fair dice. The 36 equally probable elementary outcomes are displayed in (1.2). Let E_7 denote the event of obtaining a total of seven spots, and F_2 denote the event that the first die shows two spots. Thus,

$$P(E_7) = \frac{6}{36} = \frac{1}{6} \quad \text{and} \quad P(F_2) = \frac{6}{36} = \frac{1}{6}.$$

Since

$$P(E_7 \cap F_2) = P(\{(2,5)\}) = \frac{1}{36} = P(E_7) \cdot P(F_2),$$

events E_7 and F_2 are independent, by definition.

On the other hand, let E_5 be the event of obtaining a total of five spots, so that $P(E_5) = 4/36 = 1/9$. In this case,

$$P(E_5 \cap F_2) = P(\{(2,3)\}) = \frac{1}{36}.$$

Therefore, E_5 and F_2 are dependent events:

$$P(E_5) \cdot P(F_2) = \frac{1}{9} \cdot \frac{1}{6} \neq \frac{1}{36} = P(E_5 \cap F_2). \blacksquare$$

1.2. Random Variables and Probability Distributions

When working with a random phenomenon modeled by a probability space, one is often more concerned with some numerical function of the outcomes in the sample space than in the actual set of outcomes. For example, interpreting the occurrence of an insurance claim as the outcome of a random experiment, actuaries usually focus on the monetary *amount* of the claim. From another perspective, they may be primarily interested in the *number* of claims occurring during the policy term.

Assume that (Ω,S,P) is a probability space for an experiment of chance. Consider now a function X defined on the sample space Ω that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$. The function X is called a ***random variable*** on Ω provided that for every real number x the set $\{\omega \in \Omega : X(\omega) \le x\}$ is an event in S (such a function X is a ***measurable function*** with respect to S). The ***range space*** or ***value space*** of X is the range of the function X:

$$R_X = \{x \in \Re : x = X(\omega) \text{ for some } \omega \in \Omega\}.$$

We denote random variables by upper-case letters—*X, Y, N*. Specific values that a random variable assumes are represented by lower-case letters—*x, y, n*.

Most commonly encountered random variables can be classified as one of two major types—the ***discrete type*** or the ***continuous type***—although actuaries also meet the ***mixed type*** of variable that combines features of both the discrete and the continuous. A discrete random variable X has a countable range space, $R_X = \{x_1, x_2, x_3, \ldots\}$, whereas for continuous random variables the range space consists of one or more intervals of real numbers—finite or infinite in length.

Assume now that (Ω,S,P) is a probability space and that X is a random variable defined on Ω. The set function P_X defined on subsets of the real numbers $\Re$ is called a ***probability distribution*** (or merely ***distribution***) for X provided it assigns to a set A of real numbers the probability that X takes on a value in A:[5]

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\}). \tag{1.14}$$

In particular, the probability that X lies in the semi-infinite interval $(-\infty, x]$ is

$$P_X((-\infty, x]) = P(\{\omega \in \Omega : X(\omega) \le x\}). \tag{1.15}$$

Example 1.5. An experiment consists of tossing three fair coins. As discussed in Examples 1.1(*a*) and 1.2(*a*), the sample space (1.1) consists of eight elementary outcomes, each with probability 1/8. Let the discrete random variable X denote the number of Heads obtained. There are clearly four possible values for X: $R_X = \{0, 1, 2, 3\}$. Thus, variable X is defined on the sample space by

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = TTT \\ 1 & \text{if } \omega \in \{HTT, THT, TTH\} \\ 2 & \text{if } \omega \in \{HHT, HTH, THH\} \\ 3 & \text{if } \omega = HHH. \end{cases}$$

[5] Technically speaking, set A must belong to the σ–algebra $\mathcal{B}$ generated by all the semi-infinite intervals $(-\infty, x]$. The resulting induced probability space $(\Re, \mathcal{B}, P_X)$ is defined on all of $\Re$. Virtually every set of real numbers encountered in practice belongs to $\mathcal{B}$. Details of this formal approach to random variables and probability distributions can be found in an advanced textbook of probability.

Each value of X defines an event in the underlying sample space, with an associated probability. Thus, the probability set function P defined on the sample space induces in a natural way a probability distribution P_X for the random variable X:

$$P_X(\{0\}) = \Pr\{X = 0\} = P(\{TTT\}) = \frac{1}{8},$$

$$P_X(\{1\}) = \Pr\{X = 1\} = P(\{HTT, THT, TTH\}) = \frac{3}{8},$$

$$P_X(\{2\}) = \Pr\{X = 2\} = P(\{HHT, HTH, THH\}) = \frac{3}{8},$$

$$P_X(\{3\}) = \Pr\{X = 3\} = P(\{HHH\}) = \frac{1}{8}. \blacksquare \qquad (1.16)$$

In the final set of equations (1.16) of this example we introduced a somewhat simplified notation. If $H(X)$ is a statement about the values of X that can be true or false, then $\Pr\{H(X)\}$ represents the more precise expression

$$\Pr\{H(X)\} = P_X(\{x: H(x) \text{ is true}\}).$$

Probability Distribution Functions

In practice, however, the probability distribution for a random variable X is usually expressed by a function defined directly on the real-number values of X. Specifically, one often generates a probability distribution for X by means of a ***probability density function f*** (abbreviated ***p.d.f.***) defined on all of the real numbers $\Re$. (The density function for X may be denoted by f_X when it is important to distinguish the random variable from other variables in a given context.) Function f has a distinctive form, depending on whether X is of the discrete type or continuous type. Beginning with the discrete case, we shall in the following discussion take up these two types, as well as the mixed type, in turn.

Often f depends on a set of one or more numbers $\Theta = \langle \theta_1, \theta_2, \ldots, \theta_r \rangle$, which can vary over a range of values, each value-set of numbers determining a specific density function. Such numbers are called ***parameters***. The resulting distributions are then said to belong to a ***parametric distribution family***.

Whenever X has a countable range space $R_X = \{x_1, x_2, x_3, \ldots\}$, X is said to be a discrete random variable. To serve as a probability density function in the discrete case, function f must have properties (*i*), (*ii*), (*iii*) listed below. In the discrete case f is also called a ***probability mass function***.

$$\begin{aligned} &(i) && f(x_i) \geq 0 && \text{for } x_i \in R_X, \\ &(ii) && f(x) = 0 && \text{for } x \notin R_X, \\ &(iii) && \textstyle\sum_i f(x_i) = 1, && i = 1, 2, 3, \ldots. \end{aligned} \qquad (1.17)$$

Function P_X is defined by setting $P_X(\{x_i\}) = f(x_i)$ for $i = 1, 2, 3, \ldots$. Moreover, for a set A, $A \subseteq \Re$, it follows that

$$P_X(A) = \sum_{x_i \in A} f(x_i). \tag{1.18}$$

The next example describes three common families of discrete distributions.

Example 1.6. (***a***) The simplest non-trivial random variable takes on only two distinct values: {0,1}. Such a variable X can be defined on any probability space relative to a fixed event E with $P(E) = p$, where $0 < p < 1$:

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E. \end{cases}$$

A trial resulting in the occurrence of E is often termed a "success," whereas the occurrence of the complement E^c is called a "failure." Therefore, the probability of obtaining a success is $\Pr\{X = 1\} = p$, and the probability mass function is

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{if } x \notin \{0,1\}. \end{cases}$$

The probability distribution with this function is called a ***Bernoulli distribution*** with parameter p, and X is accordingly known as a ***Bernoulli random variable***.[6]

(***b***) Another family of discrete distributions, related to the Bernoulli, comprises the ***binomial distributions*** with parameters n and p. X is a binomial random variable if X equals the number x of successes, each with probability p, obtained in n independent Bernoulli trials ($n = 1, 2, 3, \ldots$).[7] The probability of x successes in n trials is $\Pr\{X = x\} = {}_nC_x\, p^x\, (1 - p)^{n-x}$, and the probability mass function is

$$f(x) = \begin{cases} {}_nC_x\, p^x (1-p)^{n-x} & \text{if } x \in \{0, 1, 2, \ldots, n\} \\ 0 & \text{if } x \notin \{0, 1, 2, \ldots, n\}. \end{cases} \tag{1.19}$$

---

[6] The Bernoulli variable is named for Jacob [James] Bernoulli (1654–1705), a prominent member of the Bernoulli family of Swiss mathematicians. His most significant work in probability, the *Ars conjectandi,* was published posthumously in Basel in 1713.

[7] Informally, we say that successive trials of a single experiment or trials of separate experiments are said to be independent whenever the probabilities of the outcomes in one trial do not depend on those of another. In particular, if event E is associated with a certain trial and event F with another trial in the sequence, then $\Pr\{E \text{ and } F\} = P(E) \cdot P(F)$. A more formal treatment of this topic can be found in a standard probability theory text.

The factors ${}_nC_x$ in (1.19) are the ordinary binomial coefficients. The Binomial Theorem is used at step (2) in the following verification that the probabilities in (1.19) all sum to 1:

$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} {}_nC_x p^x (1-p)^{n-x} \overset{(2)}{=} (p+1-p)^n = 1.$$

(***c***) Consider a Bernoulli experiment with two elementary outcomes: success or failure. Independent trials with Pr{*success*} = p are performed until the first success is obtained. For this experiment the elementary outcomes in Ω form a countably infinite set of sequences beginning with a number (0, 1, 2, . . .) of failures (F) and ending with a single success (S):

$$\Omega = \{S, FS, FFS, FFFS, FFFFS, FFFFFS, \ldots\}.$$

Let N denote a random variable with value equal to the number n of trials required to obtain the first success. Independence of the component Bernoulli trials implies that for $n = 1, 2, 3, \ldots$ the probability mass function is

$$f(n) = \Pr\left\{ \textit{first } S \textit{ obtained on the } n^{th} \textit{ trial} \right\} = (1-p)^{n-1} p. \tag{1.20}$$

Note that the sum of an infinite geometric series is used at step (2) in the following verification of the sum of all nonzero probabilities:

$$\sum_{n=1}^{\infty} f(n) = p \sum_{n=0}^{\infty} (1-p)^n \overset{(2)}{=} p \cdot \frac{1}{1-(1-p)} = 1.$$

A distribution with probability mass function (1.20) is accordingly called a ***geometric distribution*** with parameter p. Refer also to Problem 3.21. ■

A probability distribution for a random variable X can also be characterized by a function related to the probability density function, the ***cumulative distribution function*** (sometimes shortened to ***distribution function*** or abbreviated ***c.d.f.***). This function is denoted by F—or by F_X whenever the dependence on X must be emphasized—and is defined for all real numbers x by

$$F(x) = \Pr\{X \le x\}. \tag{1.21}$$

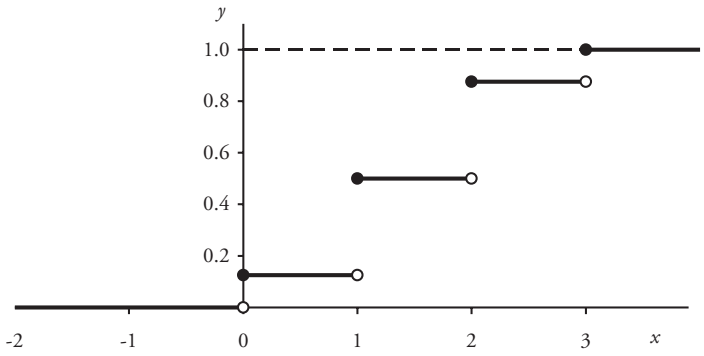
Therefore, if X is a discrete variable with range space $R_X = \{x_1, x_2, x_3, \ldots\}$ with a probability mass function f satisfying (1.17), then function F is given by

$$F(x) = \sum_{x_i \le x} f(x_i). \tag{1.22}$$

Example 1.7. Let X be the random variable of Example 1.5. The probability density function can be expressed by the table

# Heads x	0	1	2	3
$f(x)$	0.125	0.375	0.375	0.125

Figure 1.1. Cumulative Distribution Function $y = F(x)$ [Example 1.7]



As usual, we assume that $f(x) = 0$ for $x \notin R_X = \{0, 1, 2, 3\}$. A graph of the cumulative distribution function, below, is shown in Figure 1.1:

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 0.125 & \text{if } 0 \le x < 1 \\ 0.500 & \text{if } 1 \le x < 2 \\ 0.875 & \text{if } 2 \le x < 3 \\ 1.000 & \text{if } 3 \le x < \infty. \blacksquare \end{cases}$$

Example 1.7 illustrates the fact that for a discrete random variable X, $F(x)$ has a jump discontinuity at each value $x_i \in R_X$ for which $f(x_i) > 0$. Moreover, the height of the jump at x_i is just $f(x_i)$. Elsewhere the function is constant:

$$F(x) = \begin{cases} 0 & \text{if } x < x_1 \\ \sum_{j=1}^{i-1} f(x_j) & \text{if } x_{i-1} \le x < x_i, i = 2, 3, \ldots \\ 1 & \text{if } \max R_X \text{ exists and } x \ge \max R_X. \end{cases}$$

Thus, for every probability distribution defined on a discrete random variable the cumulative distribution function $F(x)$ is a step function.

Suppose now that random variable X is a non-discrete variable. This means the range space R_X is a uncountable set, and we shall further assume that R_X consists of one or more intervals (of finite or infinite length) of real numbers. To serve as a probability density function in this case f must be defined on all of $\Re$ and be Riemann integrable there ("Riemann integrable" generally means that the function has at most a countable

set of points of discontinuity).[8] Function f must also have properties analogous to those of the discrete case (1.17):

$$\begin{aligned} &(i)\ \ f(x) \geq 0 \qquad \text{for } x \in R_X, \\ &(ii)\ \ f(x) = 0 \qquad \text{for } x \notin R_X, \\ &(iii)\ \ \int_{-\infty}^{\infty} f(x)\,dx = 1. \end{aligned} \tag{1.23}$$

If such a density function exists, the probability function P_X is defined for a set A of real numbers by the integral

$$P_X(A) = \int_A f(x)\,dx. \tag{1.24}$$

Thus, for example,

$$P_X([a,b]) = \int_a^b f(x)\,dx, \quad [a,b] \subseteq \Re. \tag{1.25}$$

In particular, the cumulative distribution function is given by

$$F(x) = \int_{-\infty}^{x} f(u)\,du. \tag{1.26}$$

A basic theorem of calculus guarantees that for such an integrand $f(x)$, the function $F(x)$ is a continuous function of x on all of $\Re$. Moreover, when the density function $f(x)$ is continuous at x, then $F(x)$ is also differentiable at x, with $F'(x) = f(x)$. As a result, the random variable X and its associated probability distribution P_X are said to be ***continuous***.

The next example illustrates a trio of important continuous distributions.

Example 1.8. (***a***) Random variable X takes on values throughout an interval $[\alpha, \beta]$ of real numbers ($\alpha < \beta$). Variable X is said to have a ***uniform distribution*** on $[\alpha, \beta]$ if the probability density function is given by

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } x \in [\alpha, \beta] \\ 0 & \text{if } x < \alpha \text{ or } x > \beta. \end{cases}$$

[8] Named for the German professor of mathematics, Bernhard Riemann (1826–1866), who gave the first rigorous definition, the Riemann integral is the ordinary integral of elementary calculus. Riemann's approach to integration was later extended by other mathematicians, notably Henri Lebesgue (1875–1941). Although today the most general and rigorous treatments of probability are founded on the Lebesgue theory of measure and integration, the Riemann approach (and its generalization by Stieltjes, discussed in the next section) is adequate for the present work.

Thus, the probability that X lies in a subinterval $[c, d]$, where $\alpha \le c < d \le \beta$, is proportional to the length of the subinterval:

$$P_X([c,d]) = \int_c^d f(x)\,dx = \frac{1}{\beta-\alpha}\int_c^d dx = \frac{d-c}{\beta-\alpha}.$$

(***b***) Let X denote the size of an insurance claim that is unrestricted by any policy limit. Then it is reasonable to consider the nonnegative real numbers as the range space of X: $R_X = [0, \infty)$. When X has the probability density function

$$f(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ (1/\beta)e^{-x/\beta} & \text{if } 0 \le x < \infty \quad (\beta > 0) \end{cases}$$

X is said to have an ***exponential distribution***. The cumulative distribution function is

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - e^{-x/\beta} & \text{if } 0 \le x < \infty. \end{cases}$$

In the case $\beta = 200$ the probability that X falls in the interval [300,400] is

$$\Pr\{300 \le X \le 400\} = \frac{1}{200}\int_{300}^{400} e^{-x/200}\,dx = F(400) - F(300) = 0.0878.$$

(***c***) The random variable Z with the important ***standard normal distribution***—known also as the ***Gaussian distribution***[9]—has a continuous nonzero density function defined on all of $\Re$:

$$f(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}z^2\right), \quad -\infty < z < \infty. \tag{1.27}$$

Because f is a function of z^2, the distribution is symmetric about $z = 0$.

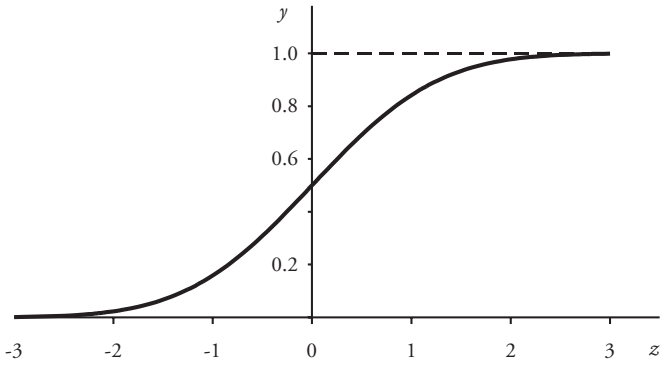
The cumulative distribution function, denoted in this case by the special notation $\Phi(z)$, is therefore

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left(-\tfrac{1}{2}u^2\right)du. \tag{1.28}$$

Because the integral in (1.28) cannot be evaluated by elementary methods of calculus, values of Φ must be obtained by some approximation method—refer to Appendix A.1 for details. A graph of $y = \Phi(z)$ is shown in Figure 1.2. ■

[9] The German mathematician Karl Friedrich Gauss (1777–1855) is widely acknowledged as the greatest mathematician of the nineteenth century. Working at the University of Göttingen, he made significant contributions to a broad range of fields in mathematics and physics. He used the normal distribution to model the distribution of measurement errors.

Figure 1.2. Cumulative Distribution Function $y = \Phi(z)$ [Example 1.8(*c*)]



In addition to the special properties for cumulative distribution functions for discrete and continuous random variables already mentioned, the function $F(x)$ has a number of general properties, listed below.

Properties of F(x)

Assume that c is an arbitrary real constant. Then

$$(\boldsymbol{a})\ \ 0 \le F(x) \le 1 \quad \textit{for all } x \in \Re. \tag{1.29}$$

$$(\boldsymbol{b})\ \ F(x_1) \le F(x_2) \quad \textit{for } x_1 < x_2. \tag{1.30}$$

$$(\boldsymbol{c})\ \ \lim_{x\to\infty} F(x) = 1 \quad \textit{and} \quad \lim_{x\to-\infty} F(x) = 0. \tag{1.31}$$

$$(\boldsymbol{d})\ \ \Pr\{X = c\} = \lim_{x\to c+} F(x) - \lim_{x\to c-} F(x) = F(c+) - F(c-). \tag{1.32}$$

$$(\boldsymbol{e})\ \ F(x) \textit{ is continuous from the right, that is, } \lim_{x\to c+} F(x) = F(c). \tag{1.33}$$

Proof:

(***a***) The inequality follows from the definition (1.21) of F as a probability.

(***b***) The inequality follows from

$$F(x_2) - F(x_1) = \Pr\{x_1 < X \le x_2\} \ge 0.$$

(***c***) Let $\langle x_n \rangle$ be an increasing sequence of reals with $\lim_{n\to\infty} x_n = \infty$. Then $\langle I_n \rangle = \langle (-\infty, x_n] \rangle$ is an ascending sequence of intervals, with $\bigcup_n I_n = (-\infty, \infty)$ and $P_X(I_n) = F(x_n)$. Applying the result of Problem 1.3(*a*):

$$\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} P_X(I_n) = P_X(\bigcup\nolimits_n I_n) = P_X((-\infty, \infty)) = 1.$$

(***d***) The sequence $\langle I_n \rangle = \langle (c - \frac{1}{n}, c + \frac{1}{n}] \rangle$ is a descending sequence of intervals, with $\{c\} = \bigcap_n I_n$. The result of Problem 1.3(*b*) yields

$$\begin{aligned} \Pr\{X = c\} = P_X\left(\bigcap_n I_n\right) &= \lim_{n\to\infty} P_X(I_n) \\ &= \lim_{n\to\infty}\left(F\left(c + \frac{1}{n}\right) - F\left(c - \frac{1}{n}\right)\right) \\ &= F(c+) - F(c-). \end{aligned}$$

(***e***) Let $\langle x_n \rangle$ be a decreasing sequence of reals with $\lim_{n\to\infty} x_n = c$. Then $\langle I_n \rangle = \langle (-\infty, x_n] \rangle$ is a descending sequence of intervals, with $\bigcap_n I_n = (-\infty, c]$ and $P_X(I_n) = F(x_n)$. Again applying Problem 1.3(*b*):

$$\lim_{n\to c+} F(x) = \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} P_X(I_n) = P_X\left(\bigcap_n I_n\right) = P_X\left((-\infty, c]\right) = F(c). \blacksquare$$

Note that property (*d*) above implies that if X is a continuous random variable then $\Pr\{X = x\} = 0$ for every real x.

Occasionally one encounters random variables that are neither entirely discrete nor entirely continuous but whose distribution is a hybrid of these two main types. Such a random variable is said to have a ***mixed distribution***, with a cumulative distribution function of the form described below.

A distribution function F is of the mixed type if the function can be expressed as

$$F(x) = \omega_1 F_1(x) + \omega_2 F_2(x), \tag{1.34}$$

where $F_1(x)$ is the distribution function of a continuous variable X_1, $F_2(x)$ is the distribution function of a discrete random variable X_2 with $R_{X_2} = \{x_i\}$, and the nonnegative numbers ω_1 and ω_2 satisfy $\omega_1 + \omega_2 = 1$. Here the numbers ω_1 and ω_2 can be interpreted as the respective probabilities of being (1) in the continuous state or (2) in the discrete state. A graph of $y = F(x)$ is continuous except at the points $\{x_i\}$, where it has a jump discontinuity of height $\omega_2 f_{X_2}(x_i)$.

The next example illustrates some types of mixed distributions often encountered in the modeling of property/casualty claim processes.

Example 1.9. The distribution of an unlimited claim-size random variable Y has the exponential c.d.f.

$$F_Y(x) = \begin{cases} 0 & \text{if } -\infty \le x < 0 \\ 1 - e^{-0.01x} & \text{if } 0 \le x < \infty. \end{cases}$$

(***a***) The variable X is distributed like Y for positive x, but takes on the value 0 with probability 0.25. Thus, the distribution of X is a mixed distribution with a discrete lump of probability at $x = 0$. Since

$$\omega_1 = \textit{Probability of being in the continuous state} = 1 - \Pr\{X = 0\} = 0.75,$$

the cumulative distribution function of X has the form

$$F(x) = (0.75)\begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - e^{-0.01x} & \text{if } 0 \le x < \infty \end{cases} + (0.25)\begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 & \text{if } 0 \le x < \infty \end{cases}$$

$$= \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - 0.75e^{-0.01x} & \text{if } 0 \le x < \infty. \end{cases}$$

(***b***) Alternatively, suppose that X is distributed like Y, but is limited above by the value 200—that is, claims less than or equal to 200 are paid at full value, but claims greater than 200 are paid at the maximum value of 200. In this case, however,

$$\omega_1 = \textit{Probability of being in the continuous state} = F_Y(200) = 1 - e^{-2}.$$

Therefore,

$$F(x) = \left(1 - e^{-2}\right)\begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{1 - e^{-0.01x}}{1 - e^{-2}} & \text{if } 0 \le x < 200 \\ 1 & \text{if } 200 \le x < \infty \end{cases} + e^{-2}\begin{cases} 0 & \text{if } -\infty < x < 200 \\ 1 & \text{if } 200 \le x < \infty \end{cases}$$

$$= \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - e^{-0.01x} & \text{if } 0 \le x < 200 \\ 1 & \text{if } 200 \le x < \infty. \end{cases}$$

(***c***) Finally, again assume that X is distributed like Y, but simultaneously has both modifications described in parts (*a*) and (*b*): it takes on the value 0 with probability $\Pr\{X = 0\} = 0.25$ and is limited above by the value 200. Thus, the modified variable X_c has a mixed distribution with two discrete lumps of probability mass, one at $x = 0$ and another at $x = 200$. Observe that

$$\omega_1 = \textit{Probability of being in the continuous state} = 0.75F_Y(200) = 0.75\left(1 - e^{-2}\right).$$

Variable X then has the cumulative distribution function

$$F(x) = 0.75\left(1 - e^{-2}\right)\begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{1 - e^{-0.01x}}{1 - e^{-2}} & \text{if } 0 \le x < 200 \\ 1 & \text{if } 200 \le x < \infty \end{cases}$$

$$+\left(0.25+0.75e^{-2}\right)\begin{cases}0 & \text{if } -\infty < x < 0\\ \dfrac{0.25}{0.25+0.75e^{-2}} & \text{if } 0 \le x < 200\\ 1 & \text{if } 200 \le x < \infty\end{cases}$$

$$=\begin{cases}0 & \text{if } -\infty < x < 0\\ 1-0.75e^{-0.01x} & \text{if } 0 \le x < 200\\ 1 & \text{if } 200 \le x < \infty.\end{cases}$$

Graphs of $y = F(x)$ for parts (a), (b), and (c) are shown in Figures 1.3, 1.4, and 1.5, respectively. ■

Figure 1.3. Mixed Distribution Function $y = F(x)$ [Example 1.9(*a*)]

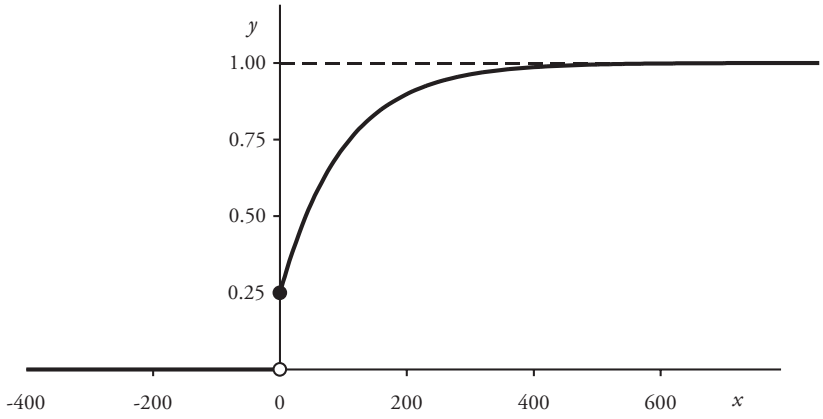


Figure 1.4. Mixed Distribution Function $y = F(x)$ [Example 1.9(*b*)]

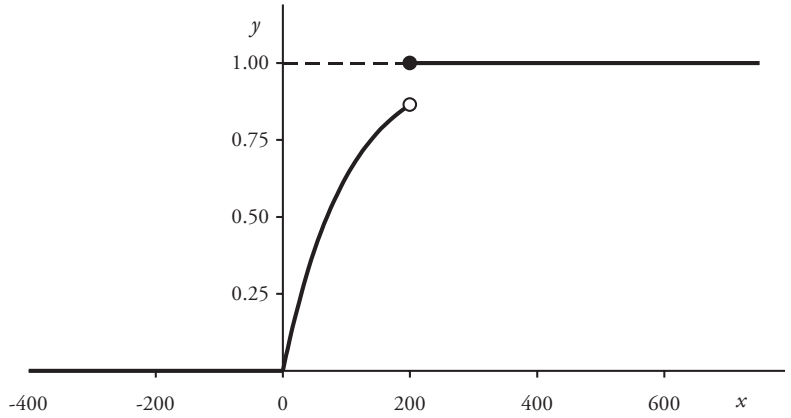
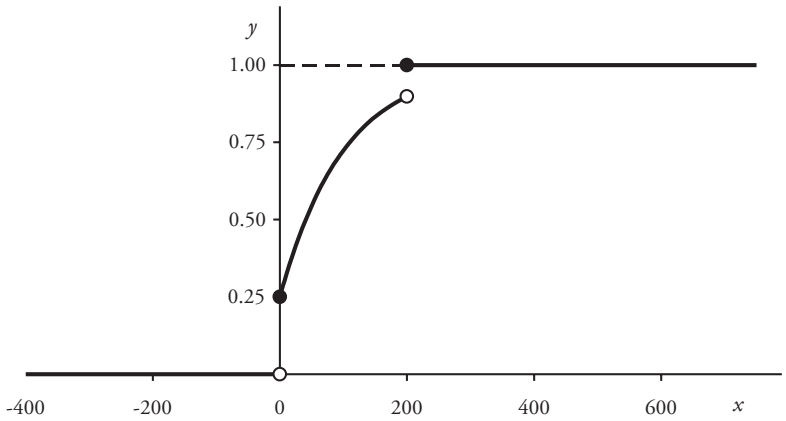


Figure 1.5. Mixed Distribution Function $y = F(x)$ [Example 1.9(*c*)]



Joint Distributions

It is frequently necessary to work with two or more random variables at the same time, recognizing that the values of one variable may influence the values of another. Accordingly, one must consider the probability distribution of the variables jointly. For example, suppose that X and Y are random variables with respective density functions $f_X(x)$ and $f_Y(y)$. We define $F(x, y)$, the ***joint cumulative distribution function*** of X and Y, by

$$F(x, y) = \Pr\{X \le x \text{ and } Y \le y\}, \quad -\infty < x, y < \infty. \tag{1.35}$$

The ***joint probability density function*** $f(x, y)$ is a function that, in the case that X and Y are both discrete variables, satisfies

$$f(x_i, y_j) = \Pr\{X = x_i \text{ and } Y = y_j\}, \quad x_i \in R_X, y_j \in R_Y, \tag{1.36}$$

as well as

$$f_X(x_i) = \sum_{y_j \in R_Y} f(x_i, y_j), \quad x_i \in R_X,$$

$$f_Y(y_j) = \sum_{x_i \in R_X} f(x_i, y_j), \quad y_j \in R_Y. \tag{1.37}$$

In this context, functions $f_X(x)$ and $f_Y(y)$ are called the ***marginal probability density functions*** of X and Y, respectively.

In the case that X and Y are both continuous variables, the density function $f(x, y)$ must satisfy

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, du\, dv, \quad -\infty < x, y < \infty, \tag{1.38}$$

with the marginal density functions given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx. \tag{1.39}$$

When it is true that the density functions satisfy the relation

$$f(x, y) = f_X(x) \cdot f_Y(y), \quad -\infty < x, y < \infty, \tag{1.40}$$

we say that random variables X and Y are ***independent***. For independent random variables, the probability distribution of one variable is not affected by the values of the other. In particular, the probability $\Pr\{X \le x\}$ is mathematically independent of the value of Y, and vice versa. Consequently, we also have

$$F(x, y) = F_X(x) \cdot F_Y(y), \quad -\infty < x, y < \infty. \tag{1.41}$$

1.3. Mathematical Expectation

One of the most useful random-variable concepts is that of mathematical expectation, which we define in the following way. Assume that X is a random variable with p.d.f. f—and range space $R_X = \{x_i\}$ if X is discrete. Let g be a function such that

$$\sum_i g(x_i) f(x_i) \quad \text{or} \quad \int_{-\infty}^{\infty} g(x) f(x)\,dx,$$

depending on whether X is discrete or continuous, respectively, exists as a finite number. Thus, whenever the above expression is an infinite series or improper Riemann integral, it must be convergent. The ***expectation*** or ***expected value*** of $g(X)$ is denoted by $E[g(X)]$, and it is defined by

$$E[g(X)] = \begin{cases} \sum_i g(x_i) f(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f(x)\,dx & \text{if } X \text{ is continuous.} \end{cases} \tag{1.42}$$

A Note on Integrals. Students of integration theory will recognize that the dual expressions in (1.42) can be represented by a single formula in which the integral is of the Riemann–Stieltjes type, as opposed to the ordinary Riemann integral of elementary calculus:[10]

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\,dF(x). \tag{1.43}$$

Without going into the theoretical details, unnecessary for the present discussion and which can be obtained from a text of real analysis, the Riemann–Stieltjes integral

[10] Thomas Jan Stieltjes (1856–1894) was a prominent Dutch mathematician who made contributions to continued fractions, number theory, and analysis. Appearing in his 1894 paper, "*Recherches sur les fractions continues,*" his was the first published generalization of the Riemann integral. Details concerning the Riemann–Stieltjes integral can be found in textbooks of probability theory or advanced calculus; for example, refer to McCord and Moroney [14], pp. 82–92, or Apostol [2], pp. 140–182.

has the following properties which support the use of the expression in (1.43). Whenever $F(x)$ is a nondecreasing differentiable function for which $F'(x) = f(x)$ and $f(x)$ is Riemann-integrable, the Stieltjes integral in (1.43) reduces to the Riemann integral

$$\int_{-\infty}^{\infty} g(x)\,dF(x) = \int_{-\infty}^{\infty} g(x)\,f(x)\,dx. \tag{1.44}$$

In the case that $F(x)$ is a nondecreasing step function with jumps at a countable set of values $\{x_i\}$ and with the height of the jump at x_i equal to $f(x_i)$, then

$$\int_{-\infty}^{\infty} g(x)\,dF(x) = \sum_i g(x_i)\big(F(x_i) - F(x_{i-1}-)\big) = \sum_i g(x_i)\,f(x_i). \tag{1.45}$$

As a result, we are justified in using the notation of (1.43) in place of (1.42), with the integral $\int_{-\infty}^{\infty} g(x)\,dF(x)$ interpreted as a Riemann–Stieltjes integral.

Properties (1.46) through (1.50) below are straightforward consequences of definition (1.42). Verification is requested in Problem 1.8.

Properties of E[g(X)]

Assume that c is a real constant and that h and g are functions for which $E[g(X)]$ and $E[h(X)]$ exist. Then

$$(\boldsymbol{a})\quad E[c] = c. \tag{1.46}$$

$$(\boldsymbol{b})\quad E\big[c\,g(X)\big] = c\,E\big[g(X)\big]. \tag{1.47}$$

$$(\boldsymbol{c})\quad E\big[g(X) + h(X)\big] = E\big[g(X)\big] + E\big[h(X)\big]. \tag{1.48}$$

$$(\boldsymbol{d})\quad E\big[g(X)\big] \le E\big[h(X)\big] \textit{ whenever } g(x) \le h(x) \textit{ for all } x. \tag{1.49}$$

$$(\boldsymbol{e})\quad \big|E\big[g(X)\big]\big| \le E\big[\big|g(X)\big|\big]. \tag{1.50}$$

One of the most important expected values for a random variable X is the ***mean*** $E[X]$, obtained from (1.43) when $g(X) = X$:

$$E[X] = \int_{-\infty}^{\infty} x\,dF(x). \tag{1.51}$$

In addition, the expectation of $g(X) = (X - E[X])^2$ defines the ***variance*** of X:

$$Var[X] = E\Big[\big(X - E[X]\big)^2\Big]. \tag{1.52}$$

The mean is a familiar measure of central tendency. For claim-size distributions, discussed in Chapter 2, the mean $E[X]$ is often called the ***severity***.[11] The variance is

[11] Actuaries sometimes use the term "severity" as a synonym for "claim size" when referring to claim-size distributions as "severity distributions." However, in this monograph we shall consistently use the term to denote *mean claim size.*

a standard measure of the dispersion of the distribution—the larger the variance, the more widely dispersed over the range space is the unit mass of probability. The square root of the variance is known as the ***standard deviation***: $SD[X] = \sqrt{Var[X]}$.

Properties of Var[X]
If c is a real constant, then

$$(\boldsymbol{a})\ \ Var[c] = 0. \tag{1.53}$$

$$(\boldsymbol{b})\ \ Var[c\,X] = c^2\,Var[X]. \tag{1.54}$$

$$(\boldsymbol{c})\ \ Var[X] = E[X^2] - (E[X])^2. \tag{1.55}$$

Proofs of these variance properties are requested in Problem 1.9.

Generalizing the expected values involved in definitions (1.51) and (1.52), we define the expectation of $g(X) = X^m$ for $m = 1, 2, 3, \ldots$:

$$E[X^m] = \int_{-\infty}^{\infty} x^m\, dF(x). \tag{1.56}$$

When the expression in (1.56) exists, the expected value $E[X^m]$ is called the ***m^{th} moment about 0*** (or more simply, the ***m^{th} moment***) of X. In addition, the expected value $E[(X - E[X])^m]$ is called the ***m^{th} central moment*** of X. Accordingly, $Var[X]$ is the second central moment of X.

Although we have defined the moments, as well as the variance and other moment-based entities, as characteristics of a random variable, it is customary to refer to them interchangeably as properties of the random variable and of its associated probability distribution. In a formal treatment of probability these concepts can be defined separately, but we shall not do so here.

Example 1.10. (***a***) Random variable X has a Bernoulli distribution with parameter p. Then

$$E[X] = (0)(1-p) + (1)\,p = p,$$
$$E[X^2] = (0^2)(1-p) + (1^2)\,p = p,$$

so that equation (1.55) yields $Var[X] = p(1 - p)$.

(***b***) Variable X is uniformly distributed on $[\alpha, \beta]$. Integrating over the interval, we obtain the mean and variance as functions of parameters α and β:

$$E[X] = \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} x\,dx = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{\alpha+\beta}{2},$$

$$Var[X] = \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} x^2\,dx - (E[X])^2 = \frac{\beta^2+\alpha\beta+\alpha^2}{3} - \frac{(\alpha+\beta)^2}{4} = \frac{(\beta-\alpha)^2}{12}.$$

(***c***) Z has the standard normal distribution with p.d.f. (1.27). Then

$$E[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \exp\left(-\tfrac{1}{2} z^2\right) dz = 0,$$

$$E[Z^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \exp\left(-\tfrac{1}{2} z^2\right) dz = \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = 1,$$

and so $Var[Z] = 1 - 0^2 = 1$. ■

Finally, we examine another special expected value for a random variable X, namely that of the function $g(X) = e^{tX}$. If there exists a positive number K such that the expectation $E[g(X)] = E[e^{tX}]$ exists for all $|t| < K$, then the resulting function of t, $M(t) = E[e^{tX}]$, is called the ***moment-generating function*** of X. Thus,

$$M(t) = E\left[e^{tX}\right] = \int_{-\infty}^{\infty} \exp(tx)\, dF(x).^{12} \tag{1.57}$$

Moment-generating functions play an important role in probability. When it exists, the moment-generating function of a random variable is unique and, moreover, completely characterizes the probability distribution of the variable. That is, two random variables with the same moment-generating function have the same distribution. In addition, when it exists, the m^{th} derivative of $M(t)$ evaluated at $t = 0$ is just the m^{th} moment:

$$\frac{d^m}{dt^m} M(t)\bigg|_{t=0} = E\left[X^m\right], \quad m = 1, 2, 3, \ldots. \tag{1.58}$$

Proofs of both these assertions about the moment-generating function can be found in a standard text of probability theory.

Example 1.11. (***a***) Assume that random variable X has a binomial distribution with parameters n and p ($n = 1, 2, 3, \ldots$ and $0 < p < 1$). Function $M(t)$ exists for all real t, and

$$M(t) = \sum_{x=0}^{n} e^{tx} f(x) = \sum_{x=0}^{n} {}_nC_x \left(pe^t\right)^x (1-p)^{n-x} = \left(1 - p + pe^t\right)^n.$$

The first two derivatives are

$$\frac{d}{dt} M(t) = n\left(1 - p + pe^t\right)^{n-1} pe^t,$$

$$\frac{d^2}{dt^2} M(t) = \begin{cases} n(n-1)\left(1 - p + pe^t\right)^{n-2} p^2 e^{2t} + n\left(1 - p + pe^t\right)^{n-1} pe^t & \text{if } n \geq 2 \\ pe^t & \text{if } n = 1. \end{cases}$$

[12] For a continuous random variable the moment-generating function is a type of Laplace transform of its probability density function. Although useful when it exists, the moment-generating function fails to exist for a number of important distributions, notably the lognormal family discussed in Chapter 2. However, the *characteristic function* of a random variable X, defined as the complex-valued function $E[\exp(itX)]$, always exists and has similar moment-generating properties. The uniqueness of the moment-generating function is usually obtained from a corresponding uniqueness theorem about the characteristic function. For example, see Parzen [18], pp. 400–404. Refer also to Section 4.5.

Evaluating these derivatives at $t = 0$, we obtain

$$E[X] = M'(0) = np,$$

$$E[X^2] = M''(0) = n(n-1)p^2 + np,$$

and hence $Var[X] = n(n-1)p^2 + np - (np)^2 = np(1-p)$.

(***b***) Assume that X is exponentially distributed with parameter β. Then

$$M(t) = \frac{1}{\beta}\int_0^\infty e^{tx}e^{-x/\beta}\,dx = \frac{1}{1-\beta t}, \quad -\infty < t < 1/\beta,$$

and the first two derivatives are

$$\frac{d}{dt}M(t) = \frac{\beta}{(1-\beta t)^2} \quad \text{and} \quad \frac{d^2}{dt^2}M(t) = \frac{2\beta^2}{(1-\beta t)^3}.$$

Therefore,

$$E[X] = M'(0) = \beta \quad \text{and} \quad E[X^2] = M''(0) = 2\beta^2,$$

so that $Var[X] = 2\beta^2 - \beta^2 = \beta^2$. ■

For two or more random variables considered jointly, there are several useful and important results about expectations.

First, consider $g(X, Y) = X + Y$. The expectation $E[X + Y]$ is just the sum $E[X] + E[Y]$. If, for example, X and Y are continuous variables, then

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}x\left[\int_{-\infty}^{\infty}f(x,y)\,dy\right]dx + \int_{-\infty}^{\infty}y\left[\int_{-\infty}^{\infty}f(x,y)\,dx\right]dy \\ &= \int_{-\infty}^{\infty}x\,f_X(x)\,dx + \int_{-\infty}^{\infty}y\,f_Y(y)\,dy \\ &= E[X] + E[Y]. \end{aligned} \tag{1.59}$$

Secondly, assume that X and Y are *independent* random variables and that g and h are functions for which $E[g(X)]$ and $E[h(Y)]$ each exist. Then the expectation of the product XY is the product of the expected values:

$$E[g(X)\cdot h(Y)] = E[g(X)]\cdot E[h(Y)]. \tag{1.60}$$

In the continuous case, we have

$$\begin{aligned} E[g(X)\cdot h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x)h(y)f_X(x)f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}g(x)f_X(x)\,dx \cdot \int_{-\infty}^{\infty}h(y)f_Y(y)\,dy \\ &= E[g(X)]\cdot E[h(Y)], \end{aligned}$$

and a similar argument applies in the discrete case.

Again, suppose that X and Y are independent random variables. Then

$$Var[X+Y] = Var[X] + Var[Y]. \tag{1.61}$$

To verify (1.61), use (1.59) and (1.60) to obtain

$$E\left[(X+Y)^2\right] = E\left[X^2 + 2XY + Y^2\right] = E\left[X^2\right] + 2E[X]E[Y] + E\left[Y^2\right],$$

and so

$$\begin{aligned} Var[X+Y] &= E\left[X^2\right] + 2E[X]E[Y] + E\left[Y^2\right] - (E[X] + E[Y])^2 \\ &= E\left[X^2\right] - (E[X])^2 + E\left[Y^2\right] - (E[Y])^2 \\ &= Var[X] + Var[Y]. \end{aligned}$$

When random variable Y is the sum of n independent, identically distributed random variables, $Y = X_1 + X_2 + \cdots + X_n$, there is a useful result involving the moment-generating functions. If each variable X_i has the same distribution as some random variable X, the generating function of Y can be expressed in terms of $M_X(t)$. Observe that the independence of the variables $\{X_i\}$ is used at step (3):

$$\begin{aligned} M_Y(t) &= E\left[\exp\left(tX_1 + tX_2 + \cdots + tX_n\right)\right] = E\left[\exp(tX_1)\exp(tX_2)\ldots\exp(tX_n)\right] \\ &\overset{(3)}{=} \prod_{i=1}^{n} E\left[\exp(tX_i)\right] \\ &= \left(M_X(t)\right)^n. \end{aligned} \tag{1.62}$$

Example 1.12. Assume that random variable Y is the sum of n independent random variables $\{X_i\}$, each having the distribution of X, a Bernoulli random variable with parameter p:

$$Y = X_1 + X_2 + \cdots + X_n.$$

The moment-generating function of each variable X_i exists for all real t, and

$$M_{X_i}(t) = M_X(t) = E\left[\exp(tX)\right] = e^{0t}(1-p) + e^{1t}p = 1 - p + pe^t.$$

Thus, by (1.62) the generating function for the sum Y is

$$M_Y(t) = \left(1 - p + pe^t\right)^n.$$

Because this is the moment-generating function of a binomial distribution with parameters n and p, the uniqueness of the generating function implies that Y is binomially distributed with parameters n and p. ■

1.4. Random Samples

The problem of fitting a parametric distribution to a set of claim data, discussed briefly in the next section, relies heavily on the theory of sampling from a population,

a collection of objects with identical distributional characteristics. For example, an actuary is often interested in inferring the distribution of sizes or numbers of claims from a set of data obtained from a portfolio of similar policies. We begin with the definition of random sample, which is fundamental to the discussion that follows.

An ordered set $\langle X_1, X_2, \ldots, X_n \rangle$ of independent, identically-distributed random variables is a ***random sample*** from a population random variable X if each X_i has the distribution of X. Thus, the distribution of X_i does not depend on the value of any other random variable X_j $(i \neq j)$ in the sample. In practice, a random sample of size n may be generated by performing n successive independent trials of a single experiment—for example, tossing a coin n times—or perhaps by making n selections from a collection of similar objects, each time replacing the selected object before making the next selection, a method called selection with replacement. The set of particular values of the random variables $\langle X_i \rangle$, denoted by $\langle x_1, x_2, \ldots, x_n \rangle$, is referred to as a set of ***sample observations***. A ***statistic*** is a function of the sample variables $\langle X_1, X_2, \ldots, X_n \rangle$.

The sample moments, analogs of formula (1.56) for moments of the population distribution, are useful statistics in the analysis of sample data:

$$M_m = \frac{1}{n}\sum_{i=1}^{n} X_i^m, \quad m = 1, 2, 3, \ldots. \tag{1.63}$$

The first moment is sometimes denoted by $\bar{X}$:

$$\bar{X} = M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.64}$$

and the sample variance by S^2:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = M_2 - M_1^2. \tag{1.65}$$

Because they are functions of random variables, statistics are also random variables, with probability distributions induced by that of the population random variable. For instance, if the distribution of the population variable X has mean $E[X] = \mu$ and variance $Var[X] = \sigma^2$, then

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \mu,$$

$$Var\left[\bar{X}\right] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}. \tag{1.66}$$

In addition,

$$E\left[S^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu + \mu - \bar{X}\right)^2\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \left(\bar{X} - \mu\right)^2\right]$$

$$= \sigma^2 - Var\left[\bar{X}\right]$$

$$= \frac{n-1}{n}\sigma^2. \tag{1.67}$$

Assume that $\langle x_1, x_2, \ldots, x_n \rangle$ is a set of observations from the random sample $\langle X_1, X_2, \ldots, X_n \rangle$ of size n. The ***sample distribution function*** or ***empirical distribution function*** $F_n(x)$ is defined for all real x by

$$F_n(x) = \sum_{x_i \le x} \frac{1}{n} = \frac{\#\ observations \le x}{n}. \tag{1.68}$$

Although it is not itself a probability distribution function, $F_n(x)$ has the same form as a cumulative distribution function for a discrete random variable defined on a finite set of equally probable outcomes. $F_n(x)$ therefore has the properties of such a function—specifically, (1.29) through (1.33). In particular, the first moment evaluated at the observations $\langle x_1, x_2, \ldots, x_n \rangle$ is the mean of the sample distribution:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.69}$$

The variance of the sample distribution is defined similarly:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.70}$$

Often, for ease of analysis, data in a set of observations from a random sample are grouped into a collection of disjoint intervals or cells: $\{(c_{k-1}, c_k]\}$ $(k = 1, 2, \ldots, m)$. In this case, (1.69) has the form

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{m} n_k a_k, \tag{1.71}$$

where

$$n_k = \#\ observations\ in\ (c_{k-1}, c_k] \quad \text{and} \quad n = \sum_{k=1}^{m} n_k,$$

$$a_k = \sum_{x_i \in (c_{k-1}, c_k]} \frac{x_i}{n_k} = average\ observation\ in\ (c_{k-1}, c_k].$$

If the sum of the observations in the k^{th} cell is unavailable, so that a_k is not known exactly, one could approximate the average a_k in formula (1.71) by the interval midpoint:

$$a_k \approx \frac{1}{2}(c_{k-1} + c_k). \tag{1.72}$$

1.5. Fitting Distributions

Actuaries frequently find it desirable to fit a parametric distribution model to a set of claim data, both for the purpose of smoothing the empirical distribution but also for interpolating among or extrapolating beyond the existing data. The problem of extrapolation is particularly important in describing the behavior of the very large claims in a claim-size distribution—the probability of such claims is usually so small that in any given sample of claim-size data the number of large claims is insufficient to characterize adequately the right-hand tail of the underlying population distribution. In this section we review the rudimentary details of several methods used to fit probability models to data.

We begin with a finite set of claim data $\langle x_1, x_2, \ldots, x_n \rangle$, which can be interpreted as a set of particular values for a random sample $\langle X_1, X_2, \ldots, X_n \rangle$ from a population random variable X with an unknown distribution. The variables $\langle X_i \rangle$ are independent, and each has the same distribution as X, representing the results of a single random selection from the population variable X. Here, X could be either a claim-size or claim-count variable, as discussed in Chapters 2 and 3. The aim is to find a probability model for the distribution of X, consistent with the sample observations $\langle x_1, x_2, \ldots, x_n \rangle$.

In practice, in order to justify the interpretation as a random sample from a single population, claim data must often be adjusted in order that all are on the same basis. For example, claim-size data obtained from multiple policy or accident years may very well require the application of trend factors to remove the effects of monetary inflation over time.

Methods of fitting models to sample data usually depend on first selecting a distribution family, that is, a collection of distribution functions $\{F_\Theta(x)\}$ indexed by a finite set of numeric parameters $\Theta = \langle \theta_1, \theta_2, \ldots, \theta_r \rangle$. The choice of such a family can be arbitrary, but should take into consideration any known or desired properties of the distribution under investigation.

Having chosen such a family, one must next identify the particular member of the selected distribution family that, according to some selection criterion, best describes the data. This is usually done by finding an appropriate ***point estimate*** for each distribution parameter—usually in the form of a statistic, that is, a function of the sample random variables:

$$\hat{\theta}_i = g_i(X_1, \ldots, X_n), \quad i = 1, 2, \ldots, r. \tag{1.73}$$

For example, the sample-moment statistics (1.63) are useful in this regard.

In a given situation there may exist several possible parameter estimators, statistics which could differ in their ease of computation or in the general properties of estimators deemed desirable by statisticians. The latter include the three estimator properties described below—*bias, consistency, efficiency.*

Assume that X is a random variable with a distribution that depends on the unknown parameter θ. Let $\langle X_1, X_2, \ldots, X_n \rangle$ be a random sample of X of size n and assume that

$\hat{\theta}_n$ is a function of the sample random variables, a statistic whose distribution depends on the parameter θ.

- $\hat{\theta}_n$ is said to be an ***unbiased estimate*** of θ whenever the mean of $\hat{\theta}_n$ is just θ: $E[\hat{\theta}_n] = \theta$. For example, the expected value of the sample mean $\bar{X}$ is $E[\bar{X}] = E[X]$. Thus, if $E[X] = \theta$, then $\hat{\theta}_n = \bar{X}$ is an unbiased estimate of θ. However, because $E[S^2] = (1 - \frac{1}{n})\, Var[X]$, the sample variance statistic is a *biased* estimate of $Var[X]$, although the bias $\frac{1}{n} Var[X]$ is insignificant for large samples.
- $\hat{\theta}_n$ is said to be a ***consistent estimate*** of θ if $\hat{\theta}_n$ converges in probability to θ, that is,

$$\lim_{n\to\infty} \Pr\left\{\left|\hat{\theta}_n - \theta\right| < \varepsilon\right\} = 1 \quad \text{for all } \varepsilon > 0.$$

  It can be shown that when $Var[X]$ is finite $\bar{X}$ converges in probability to $E[X]$, and so, as in the above example, $\hat{\theta}_n = \bar{X}$ is a consistent estimate of $\theta = E[X]$. In addition, if $\hat{\theta}_n$ is an unbiased estimate of θ and if $\lim_{n\to\infty} Var[\hat{\theta}_n] = 0$, then $\hat{\theta}_n$ is also a consistent estimate of θ.
- Suppose that $\hat{\theta}_n$ is an unbiased estimate of θ and that for all estimates $\hat{\theta}_n^*$ for which $E[\hat{\theta}_n^*] = \theta$, we have $Var[\hat{\theta}_n] \le Var[\hat{\theta}_n^*]$ for all θ. In this case $\hat{\theta}_n$ is said to be an ***unbiased, minimum variance estimate*** of θ. Statistic $\hat{\theta}_n$ is also called the ***most efficient estimator*** of θ.

To find an optimal fitted distribution it is often advisable to try more than one method of calculating a set of parameter estimates $\hat{\Theta}$ and sometimes work with more than one distribution family. Then, after deciding on a particular parameter estimate, one should in the final step evaluate how well the distribution $F_{\hat{\Theta}}(x)$ fits the sample data. One could do this with an informal comparison of the fitted and empirical distribution functions or more rigorously by employing a standard goodness-of-fit test, such as that based on the chi-square statistic.

Briefly described below are four useful techniques of parameter estimation: the *method of moments,* the *maximum-likelihood method,* the *minimum chi-square method,* and *minimum-distance methods.*[13]

Method-of-Moments Estimation

First proposed by the English statistician Karl Pearson, this is the oldest technique of estimating parameters and perhaps the easiest to apply in practice. The method-of-moments method is based on the usually reasonable assumption that the sample moments are good estimates of the corresponding population moments.[14]

Accordingly, one computes successive sample moments $M_m = \frac{1}{n}\sum_{i=1}^{n} x_i^m$ evaluated at the sample data points $\langle x_1, x_2, \ldots, x_n \rangle$ and then equates them to the corresponding

[13] More complete discussions of estimation techniques can be found in standard mathematical statistics texts. For example, see Hogg and Craig [7], chapter 6, or Lindgren [13], chapter 5.

[14] Karl Pearson (1857–1936), founder of the field of mathematical statistics, established the world's first college department of statistics at University College London. His contributions include the foundations of statistical hypothesis testing and decision theory, and he is the eponymous inventor of the chi-square goodness-of-fit test.

moments of the assumed distributional model, which depend on the unknown parameters Θ:

$$M_m = E_\Theta\left[X^m\right], \quad m = 1, 2, 3, \ldots. \tag{1.74}$$

One must use as many of these equations as is necessary to determine the parameters uniquely—in general, when there are r parameters to estimate use (1.74) for $m = 1, 2, \ldots, r$. The resulting system of equations could then be solved to obtain $\hat{\Theta} = \langle \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r \rangle$ in terms of the observed data values $\langle x_i \rangle$.

Method-of-moments estimates have the advantage of usually being very easy to calculate, but they do not always have the desirable properties indicated above—they are often consistent, but are sometimes biased.

Example 1.13. A possibly unbalanced die is rolled and the number of spots on the upper surface is observed. We define a Bernoulli random variable:

$$X = \begin{cases} 1 & \text{if \# spots} = 6 \\ 0 & \text{if \# spots} \neq 6. \end{cases}$$

The probability distribution for X has a single unknown parameter

$$p = \Pr\{X = 1\} = \Pr\{\text{\# spots} = 6\}$$

with the probability mass function

$$f(x) = \begin{cases} p^x (1-p)^{1-x} & \text{if } x \in \{0, 1\} \\ 0 & \text{if } x \notin \{0, 1\}. \end{cases}$$

It is evident that $E[X] = p$.

In order to estimate parameter p, the die is rolled 50 times, creating a random sample of size $n = 50$. Twelve sixes are observed, so that $\sum_{i=1}^{50} x_i = 12$. The method-of-moments estimate of parameter p is obtained from the unbiased, consistent estimator $\hat{p} = \bar{x}$:

$$\hat{p} = \frac{12}{50} = 0.2400. \blacksquare$$

Maximum-Likelihood Estimation

Simply stated, a maximum-likelihood estimator $\hat{\Theta}$ of distribution parameters Θ specifies the member $F_{\hat{\Theta}}(x)$ of a distribution family which maximizes the probability of obtaining the values $\langle x_1, x_2, \ldots, x_n \rangle$ actually observed in a random sample $\langle X_1, X_2, \ldots, X_n \rangle$.

To begin, let $\langle X_1, X_2, \ldots, X_n \rangle$ be a random sample from a probability distribution with density function $f_\Theta(x)$. The joint probability density function for the sample is then $\prod_{i=1}^{n} f_\Theta(x_i)$. Evaluated at the observed sample values $\langle x_1, x_2, \ldots, x_n \rangle$, this product

can be regarded as a function of the parameters Θ. As such, it is called the ***likelihood function*** of the random sample:

$$L(\Theta) = \prod_{i=1}^{n} f_{\Theta}(x_i). \tag{1.75}$$

Therefore, $\hat{\Theta} = \langle \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r \rangle$ is a maximum-likelihood estimator if it yields a maximum value for the likelihood function:

$$L\left(\hat{\Theta}\right) \geq L(\Theta) \quad \text{for all } \Theta.$$

Because they are located at points at which the likelihood function attains an extreme value, maximum-likelihood estimators are usually, but not always, unique.

An analytic solution of the maximum-value problem can be sought by setting the r partial derivatives of the likelihood function equal to 0 and solving the resulting system of equations. However, in most situations it is easier to solve the equivalent problem of maximizing the ***log-likelihood function*** $\log L(\Theta)$, using the same technique:[15]

$$\begin{cases} \dfrac{\partial}{\partial \theta_1} \log L(\theta_1, \ldots, \theta_r) = 0 \\ \quad \vdots \qquad\qquad\qquad \vdots \\ \dfrac{\partial}{\partial \theta_r} \log L(\theta_1, \ldots, \theta_r) = 0. \end{cases}$$

(Recall that $\log x$ is an increasing function of x, so that whenever $L(\hat{\Theta})$ is a maximum value of the likelihood function, $\log L(\hat{\Theta})$ is a maximum value of $\log L(\hat{\Theta})$, and conversely.) If, as it often does, the analytic approach proves intractable, one could employ an iterative solving algorithm, available in many computer software packages.

Maximum-likelihood estimators are usually consistent and efficient, but not always unbiased, estimators of the distribution parameters.

Example 1.14. Returning to the problem of Example 1.13, we now determine the maximum-likelihood estimator of $p = \Pr\{X = 1\}$ for a sample of size n. The likelihood function is

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

and so the log-likelihood function is

$$\log L(p) = (\log p)\sum_{i=1}^{n} x_i + \log(1-p)\left(n - \sum_{i=1}^{n} x_i\right).$$

[15] Throughout this monograph, $\log x$ denotes the natural (base e) logarithm function of x.

Therefore, the equation

$$\frac{\partial}{\partial p}\log L(p)=\frac{1}{p}\sum_{i=1}^{n}x_i-\frac{1}{1-p}\left(n-\sum_{i=1}^{n}x_i\right)=0$$

has the solution $\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$, a maximum-likelihood estimate of p. ■

Minimum Chi-Square Estimation

The minimum chi-square estimator $\hat{\Theta}$ of distribution parameters Θ specifies the member $F_{\hat{\Theta}}(x)$ of a selected distribution family that minimizes an associated chi-square statistic. This statistic is identical in form to that used in the classic Pearson chi-square goodness-of-fit test.

To construct the chi-square statistic one must first group the data from a random sample of size n into a smaller number m of classes or cells. If the population distribution is of the continuous type, like that of most size-of-loss random variables, then the cells may take the form of intervals of real numbers. Otherwise, if the distribution is discrete and the random variable is integer-valued—like that of a claim-count random variable—then the cells must be subsets of the nonnegative integers.

We then calculate the cell frequencies:

$$n_k=\#\ observations\ in\ the\ k^{th}\ cell\ (k=1,2,\ldots,m)\quad\text{and}\quad n=\sum_{k=1}^{m}n_k.$$

The statistic $\chi^2(\Theta)$ is given by

$$\chi^2(\Theta)=\sum_{k=1}^{m}\frac{|n_k-\phi_k(\Theta)|^2}{\phi_k(\Theta)},\tag{1.76}$$

where $\phi_k(\Theta)$ is the expected number of sample observations in the k^{th} cell, based on the population distribution with parameters Θ. For example, if the random variable X has a continuous distribution, with cells of the form $(c_{k-1}, c_k]$, then

$$\phi_k(\Theta)=n\cdot(F_\Theta(c_k)-F_\Theta(c_{k-1})).$$

Since each expected value $\phi_k(\Theta)$ is a function of Θ, so is $\chi^2(\Theta)$, and a minimum chi-square estimate of Θ is a value $\hat{\Theta}$ at which the statistic achieves a minimum value:

$$\chi^2(\hat{\Theta})\le\chi^2(\Theta)\ \text{for all}\ \Theta.$$

Calculation of $\hat{\Theta}$ is complicated by the fact that both numerator and denominator of $\chi^2(\Theta)$ depend on Θ. However, use of a computer-implemented iterative solving algorithm is a practical way of overcoming such computational complexities.

One advantage of using minimum chi-square estimation of the distribution parameters, of course, is the fact that at the end of the procedure one has a built-in

goodness-of-fit test available. The value of the chi-square statistic under the assumption of the fitted distribution—the null hypothesis—has already been computed. For illustrations of this method, refer to Examples 2.8 and 2.11.

Minimum-Distance Estimation

As with the minimum chi-square method discussed previously, minimum-distance methods are applied to grouped sample data. In particular, the method is most useful in estimating parameters for a random variable X with a continuous distribution. Suppose, for example, the n sample values have been assigned to m cell intervals of the form $(c_{k-1}, c_k]$, where

$$n_k = \#\, observations\ in\ (c_{k-1}, c_k] \quad (k = 1, 2, \ldots, m) \quad \text{and} \quad n = \sum_{k=1}^{m} n_k.$$

The empirical sample distribution function at the cell boundary point c_k is

$$F_n(c_k) = \frac{1}{n}\sum_{i=1}^{k} n_i. \tag{1.77}$$

One minimum-distance estimator of parameter Θ is the value $\hat{\Theta}$ that minimizes the "distance" $D(\Theta)$ between the sample and parametric distribution functions, $F_n(x)$ and $F_\Theta(x)$, evaluated at the cell boundary points:

$$D(\Theta) = \sqrt{\sum_{k=1}^{m} \left| F_n(c_k) - F_\Theta(c_k) \right|^2}. \tag{1.78}$$

Clearly, $D(\Theta)$ is a function of Θ, and $\hat{\Theta}$ must satisfy $D(\hat{\Theta}) \le D(\Theta)$ for all Θ.

An analytic solution of a minimum-distance problem is unlikely to be straightforward, but as in the case of minimum chi-square estimation, $\hat{\Theta}$ can usually be obtained by applying a computer utility or software that implements an iterative solving algorithm. Minimum-distance methods in an actuarial setting are more fully discussed in a paper by Klugman and Parsa [12]. Examples of minimum-distance fitting can be found in Examples 2.9 and 2.10.

1.6. Problems

1.1 Let Ω be the sample space for an experiment of chance.

(***a***) Show that the set $\{\varnothing, E, E^c, \Omega\}$ is a σ-algebra.

(***b***) Assume that S is a σ-algebra, with $E, F \in S$. Show that $E \cap F \in S$.

(***c***) Assume that $\Omega = \{a, b, c, d\}$ and that $\{a\}$ and $\{b,c\}$ are events. Find the smallest σ-algebra S containing this pair of events—this is the σ-algebra generated by $\{a\}$ and $\{b,c\}$.

1.2 Assume that (Ω, S, P) is a probability space. Verify the following properties of P.

(***a***) Equation (1.4). (***b***) Equation (1.5).

(***c***) Equation (1.6). (***d***) Property (1.7).

(***e***) Equation (1.8). (***f***) Property (1.9).

1.3 Assume that (Ω, S, P) is a probability space. Verify the following properties of the probability function P.

(***a***) If $E_1 \subseteq E_2 \subseteq E_3 \subseteq \cdots$ is an ascending sequence of sets in S, then $P(\bigcup_n E_n) = \lim_{n\to\infty} P(E_n)$.

(***b***) If $E_1 \supseteq E_2 \supseteq E_3 \supseteq \cdots$ is a descending sequence of sets in S, then $P(\bigcap_n E_n) = \lim_{n\to\infty} P(E_n)$.

1.4 An urn contains three red and four black chips. Two chips are drawn at random without replacement. Calculate:

(***a***) the probability that both chips are red.

(***b***) the probability that both chips are black.

(***c***) the expected number of red chips.

1.5 For a probability space (Ω, S, P) with events E and F show that

(***a***) $P(E) = P(F) \cdot P(E|F) + P(F^c) \cdot P(E|F^c)$.

(***b***) If E and F are independent, then $P(E \cup F) = 1 - P(E^c)P(F^c)$.

1.6 Consider the following generalization of Example 1.4. Two fair dice are rolled. Let E_m denote the event of obtaining a total of m spots ($m = 2, 3, \ldots, 12$) and F_n the event that the first die shows n ($n = 1, 2, \ldots, 6$) spots. For what values of (m, n) are events E_m and F_n independent? Explain.

1.7 A random variable X takes on five values with nonzero probability: $R_X = \{1, 2, 3, 4, 5\}$. The probability mass function is tabulated below, where k is a constant.

x	1	2	3	4	5
$f(x)$	k^2	$0.50k$	k	$0.25k$	0.50

Calculate:

(***a***) k.

(***b***) $F(2)$.

(***c***) $\Pr\{X \text{ is odd}\}$.

(***d***) $E[X]$.

(***e***) $Var[X]$.

1.8 Verify the following properties of the expected value $E[g(X)]$.

(***a***) Equation (1.46).

(***b***) Equation (1.47).

(***c***) Equation (1.48).

(***d***) Property (1.49).

(***e***) Property (1.50).

1.9 Verify the following properties of the variance $Var[X]$.

(***a***) Equation (1.53).

(***b***) Equation (1.54).

(***c***) Equation (1.55).

1.10 A discrete random variable N has the countably infinite range space $R_N = \{1, 2, 3, \ldots\}$.

(***a***) Can the outcomes in R_N be assigned equal probabilities?

(***b***) Find the constant k such that $f(n) = k\,p^n$ $(0 < p < 1, n \in R_N)$ is a probability mass function for N.

(***c***) Using the function of part (b), calculate $E[N]$ and $Var[N]$.

1.11 A random variable X has the cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - 0.25e^{-0.50x} & \text{if } 0 \le x < \infty. \end{cases}$$

Calculate:

(***a***) $\Pr\{X = 0\}$. (***b***) $\Pr\{X = 1\}$. (***c***) $\Pr\{X < 1\}$.
(***d***) $\Pr\{1 < X < 2\}$. (***e***) $E[X]$. (***f***) $Var[X]$.

1.12 Assume random variable X has a continuous distribution, with c.d.f. $F(x)$.

(***a***) Show that $\Pr\{X = c\} = 0$ for all real c.
(***b***) Show that $\Pr\{a < X < b\} = \Pr\{a \le X \le b\} = F(b) - F(a)$ for all a and b.

1.13 Evaluate these Riemann–Stieltjes integrals.

(***a***) $\int_0^1 x\, d(x^2)$. (***b***) $\int_1^3 x^2\, d(\log x)$.
(***c***) $\int_0^\infty d(1 - e^{-x})$. (***d***) $\int_0^5 x\, d(\llbracket x \rrbracket)$.[16]

1.14 Evaluate these Riemann–Stieltjes integrals, in which F denotes the discrete cumulative distribution function of Example 1.7.

(***a***) $\int_0^\infty x\, dF(x)$. (***b***) $\int_0^\infty x^2\, dF(x)$. (***c***) $\int_0^\infty \exp(tx)\, dF(x)$.

1.15 Evaluate these Riemann–Stieltjes integrals, in which F denotes the continuous c.d.f. of the uniform distribution of Example 1.8(*a*).

(***a***) $\int_0^\infty x\, dF(x)$. (***b***) $\int_0^\infty x^2\, dF(x)$. (***c***) $\int_0^\infty \exp(tx)\, dF(x)$.

1.16 A random variable X has the distribution function

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - (0.20)e^{-x/100} - (0.80)e^{-x/200} & \text{if } 0 \le x < \infty. \end{cases}$$

(***a***) Show that the distribution of X can be interpreted as the mixture of two exponential distributions.
(***b***) Determine $E[X]$ and $Var[X]$.

1.17 Calculate $E[X]$ and $Var[X]$ for the random variable X of Example 1.8(*a*).

1.18 Calculate $E[X]$ and $Var[X]$ for the random variable X_a of Example 1.9(*a*).

1.19 Assume that X is a random variable with $E[X] = 0$ and $Var[X] = 1$. Calculate $E[Y]$ and $Var[Y]$ for $Y = \sigma X + \mu$.

1.20 Find the moment-generating functions for:

(***a***) the standard normal distribution.
(***b***) the normal distribution of $Y = \sigma Z + \mu$, where Z is the standard normal random variable.
(***c***) the distribution of X, uniformly distributed on the interval $[\alpha, \beta]$.

---

[16] $\llbracket x \rrbracket$ denotes the *greatest integer function,* defined for every real x as the unique integer m satisfying $m \le x < m + 1$. For example, $\llbracket 5 \rrbracket = 5$, $\llbracket \pi \rrbracket = 3$, $\llbracket -1.5 \rrbracket = -2$.

1.21 Random variable N has the geometric distribution defined by (1.20). Determine:
(***a***) $M(t)$. (***b***) $E[N]$. (***c***) $Var[N]$.
(***d***) the probability that an odd number of trials is required to obtain the first success.

1.22 Justify the algebraic rearrangement in the second step of (1.67).

1.23 Find the maximum-likelihood estimator of the parameter β for the exponential distribution of Example 1.8(*b*).

1.24 Random variable X has a mixed probability density function f defined by

$$f(x) = \sum_{k=1}^{n} \omega_k f_k(x), \text{ where } 0 < \omega_k < 1 \text{ and } \sum_{k=1}^{n} \omega_k = 1.$$

Show that $E[X]$ and $Var[X]$ are given by

$$E[X] = \sum_{k=1}^{n} \omega_k \mu_k \quad \text{and} \quad Var[X] = \sum_{k=1}^{n} \omega_k \sigma_k^2 + \sum_{k=1}^{n} \omega_k (\mu_k - E[X])^2,$$

where μ_k and σ_k^2 are the respective mean and variance of the k^{th} distribution.

2. Claim Size

Every property/casualty claim process involves two independent random variables: the ***claim-size random variable*** and the ***claim-count random variable***. These two variables combine to create a third fundamental claim variable, the ***aggregate-loss random variable***, values of which represent the total claim amount generated by the underlying claim process. We shall investigate each of these variables and their related distributions in turn. Distributions of claim-size variables are studied in this chapter, the claim-count variable is the subject of Chapter 3, and then aggregate-loss distributions are taken up in Chapter 4.

A claim-size variable has an associated probability distribution called a ***size-of-loss distribution***, often shortened to ***loss distribution***. A set of empirical claim data, being finite, always has a discrete distribution, but as we shall see, a set of claim data can be usefully interpreted as a sample drawn from an underlying claim-size population assumed to have a continuous loss distribution.

To model the size of property/casualty insurance claims, actuaries employ a variety of parametric families of continuous distributions. The most popular probability distributions used for this purpose, including the lognormal and Pareto families, are studied in this chapter.

2.1. Claim-Size Random Variables

A claim-size random variable, if based on a finite population of claims or on a finite sample of claims from a larger population, always has a discrete distribution. However, for many actuarial calculations it is useful to assume that the sizes of the underlying claim population are modeled by a continuous distribution, usually one of the standard parametric distributions discussed later in this chapter. Thus the task of the actuary often is to fit a continuous parametric claim-size distribution to a discrete sample of claim data. In addition, as we shall see, distributions of various derived random variables are neither wholly discrete nor continuous, but of the mixed discrete/continuous type.

Claim-size variables, by their very nature, take on only nonnegative values. Thus for all such variables X, $\Pr\{X < 0\} = 0$. That is, $F_X(x) = 0$ for all $x < 0$. The probability density function $f(x)$ for a continuous size-of-loss distribution for which claim size is unbounded (or unlimited) from above takes on positive values over a semi-infinite interval of the form $0 \le \xi < x < \infty$. For positive b in this interval, the portion of the distribution defined on the subinterval (b, ∞) is called the ***long tail*** of the distribution.

Alternately, the part of the distribution defined on the finite subinterval (ξ, b), extending to the left and bounded below by 0, is called the ***short tail***. Clearly, such distributions cannot be symmetric.

The ***skewness*** of the distribution, defined as the normalized third central moment, is a measure of distribution symmetry:

$$Sk[X] = \frac{E\left[(X - E[X])^3\right]}{(Var[X])^{3/2}} = \frac{E\left[X^3\right] - 3E[X]E\left[X^2\right] + 2(E[X])^3}{\left(E\left[X^2\right] - (E[X])^2\right)^{3/2}}.$$

The larger the absolute value $|Sk[X]|$ the more asymmetric is the distribution. Symmetric random variables X, on the other hand, always have zero skewness—it is easy to verify that $X = -X$ implies $Sk[X] = Sk[-X] = -Sk[X]$, that is, $Sk[X] = 0$. The standard normal variable Z, for example, has $Sk[Z] = 0$. However, for a continuous, unlimited loss distribution, with its infinite long tail, skewness is usually positive—corresponding to greater probability density toward the left end of the distribution. Such a distribution is said to be ***positively skewed***.

The following three examples illustrate these fundamental properties of claim-size distributions.

Example 2.1. A discrete claim-size random variable X has the finite set of values $R_X = \{0, 50, 100, 200\}$, with probability mass function

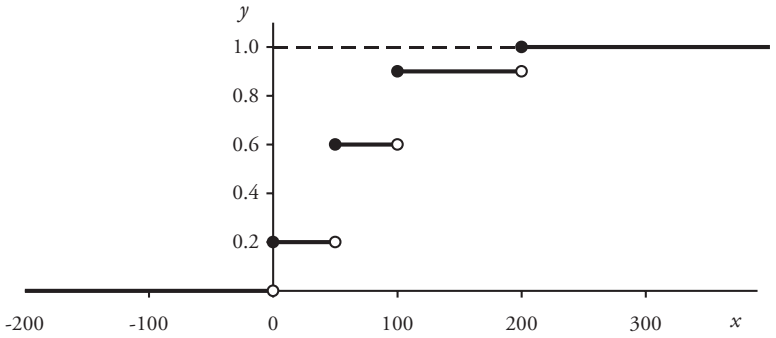
Claim Size x	0	50	100	200
$f(x)$	0.20	0.40	0.30	0.10

A graph of the cumulative distribution function $F(x)$ is shown in Figure 2.1. The severity (mean) and variance of variable X are

$$E[X] = \frac{2}{10}(0) + \frac{4}{10}(50) + \frac{3}{10}(100) + \frac{1}{10}(200) = 70,$$

$$Var[X] = \frac{2}{10}(0 - 70)^2 + \frac{4}{10}(50 - 70)^2 + \frac{3}{10}(100 - 70)^2 + \frac{1}{10}(200 - 70)^2 = 3{,}100.$$

Figure 2.1. Discrete Cumulative Distribution Function [Example 2.1]



In addition, the third central moment is

$$E\left[(X-E[X])^3\right]$$

$$=\frac{2}{10}(0-70)^3+\frac{4}{10}(50-70)^3+\frac{3}{10}(100-70)^3+\frac{1}{10}(200-70)^3=156{,}000,$$

so that $Sk[X] = (156{,}000)/(3{,}100)^{3/2} = 0.9038$. ■

Example 2.2. A continuous claim-size variable X has the exponential cumulative distribution function

$$F(x)=\begin{cases}0 & \text{if } -\infty < x < 0\\ 1-e^{-x/250} & \text{if } 0 \le x < \infty,\end{cases}$$

a graph of which is shown in Figure 2.2. A probability density function for X is therefore given by

$$f(x)=\begin{cases}0 & \text{if } -\infty < x \le 0\\ (1/250)\,e^{-x/250} & \text{if } 0 < x < \infty.\end{cases}$$

Consequently,

$$E[X]=\frac{1}{250}\int_0^\infty x\,e^{-x/250}\,dx=250,$$

$$Var[X]=\frac{1}{250}\int_0^\infty (x-250)^2\,e^{-x/250}\,dx=62{,}500,$$

and the skewness is given by

$$Sk[X]=\frac{(1/250)\int_0^\infty (x-250)^3\,e^{-x/250}\,dx}{(62{,}500)^{3/2}}=\frac{31{,}250{,}000}{15{,}625{,}000}=2.0000\,.$$ ■

Figure 2.2. Continuous Cumulative Distribution Function [Example 2.2]

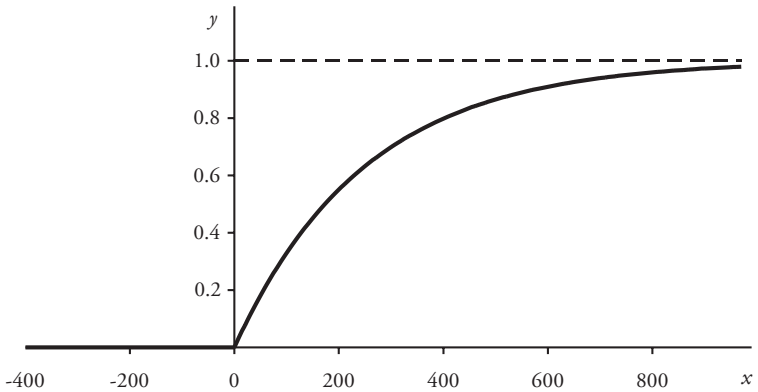
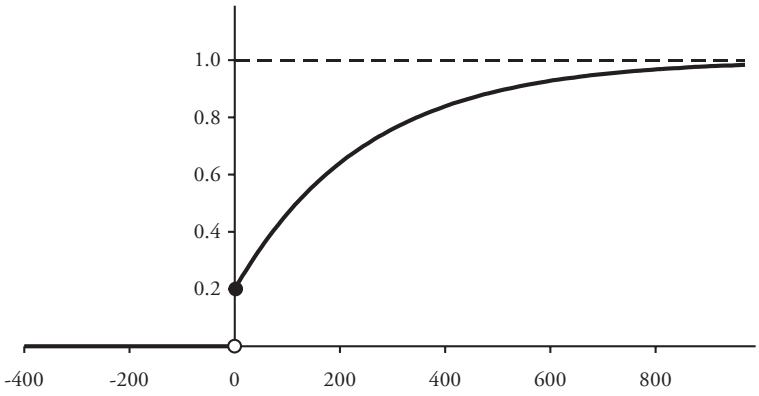


Figure 2.3. Mixed Cumulative Distribution Function [Example 2.3]



Example 2.3. For claim-size random variable Y the probability of a claim of size zero is $\Pr\{Y=0\}=0.20$. But for positive values Y is distributed conditionally as variable X in Example 2.2—that is,

$$\Pr\{Y \le y \mid Y>0\} = \frac{\Pr\{0<Y\le y\}}{0.80} = 1-e^{-y/250}, \quad 0<y<\infty.$$

Therefore, the cumulative distribution function of Y (see Figure 2.3) is given by

$$F_Y(y) = \begin{cases} 0 & \text{if } -\infty < y < 0 \\ 1-0.80\,e^{-y/250} & \text{if } 0 \le y < \infty. \end{cases}$$

Obviously, Y has a mixed discrete/continuous distribution—$F_Y(y)$ is continuous for all $y \neq 0$, with a jump discontinuity at $y=0$.

The mean, variance, and skewness of Y are, respectively,

$$E[Y] = (0.20)(0)+(0.80)(250) = 200,$$

$$Var[Y] = (0.20)(0-200)^2+(0.80)(62{,}500) = 58{,}000,$$

$$Sk[Y] = \frac{(0.20)(0-200)^3+(0.80)(31{,}250{,}000)}{(58{,}000)^{3/2}} = 1.9043.$$

A comparison with Example 2.2 reveals the effect of transferring 20% of the total probability to the value $y=0$—the mean, variance, and skewness of Y are all smaller than those of X. ■

2.2. Limited Moments

Actuaries seldom use continuous parametric size-of-loss distributions in their pure form, that is, without restrictions placed on the size of claims. The reason for this, of course, is that property/casualty insurance policies almost always specify some type

of limitation on the claim amount payable under the policy. Consequently, every unlimited probability distribution used to model insurance claim sizes must be modified appropriately to reflect whatever policy conditions are in place.

The most widely encountered condition of this type is a ***policy occurrence limit*** that caps each claim amount at a specified maximum value. Assume that X denotes an unlimited claim-size random variable, for which $\Pr\{X < 0\} = 0$, and that individual claim amounts are then restricted by a policy limit l. The effective claim size *from the viewpoint of the insurer* is the limited random variable Y, defined by

$$Y = \min\{X, l\} = \begin{cases} X & \text{if } 0 \le X < l \\ l & \text{if } l \le X < \infty. \end{cases} \tag{2.1}$$

The insurer pays in full those claims less than l and for all other claims pays the maximum amount l. Because the policy limit serves to conceal the actual size of each claim larger than l, variable X modified in this way is said to be ***censored at l***. In terms of the function F_X, the cumulative distribution function of variable Y is given by the formula

$$F_Y(y) = \Pr\{Y \le y\} = \begin{cases} F_X(y) & \text{if } -\infty < y < l \\ 1 & \text{if } l \le y < \infty. \end{cases}$$

Accordingly, the distribution of Y can have a discrete lump of nonzero probability at $y = l$, of size

$$\Pr\{Y = l\} = F_Y(l) - F_Y(l-) = 1 - F_X(l-).$$

In particular, if F_X is everywhere continuous with $0 < F_X(l) < 1$, then variable Y has a mixed discrete/continuous distribution—F_Y is continuous for all $y \ne l$, and it has a single jump discontinuity at $y = l$, with $\Pr\{Y = l\} = 1 - F_X(l) > 0$.

With respect to X, the mean of the censored variable Y is referred to as the ***limited expected value*** or ***limited severity of X***. It is denoted by $E[X; l]$ and represented by the Riemann–Stieltjes integral formula (refer to Section 1.3):

$$E[X; l] = E[Y] = \int_0^\infty y\, dF_Y(y) = \int_0^l x\, dF_X(x) + l \cdot (1 - F_X(l)). \tag{2.2}$$

If X is continuous, then there exists a function f_X such that $dF_X(x) = f_X(x)dx$, and

$$E[X; l] = \int_0^l x f_X(x)\, dx + l \cdot (1 - F_X(l)). \tag{2.3}$$

In the case that variable X has a discrete set of values $\{x_i\}$, (2.2) has the form

$$E[X; l] = \sum_{x_i \le l} x_i\, f_X(x_i) + l \cdot \sum_{x_i > l} f_X(x_i). \tag{2.4}$$

Equation (2.2) is easily generalized to a formula for limited moments of all orders m, where $m = 1, 2, 3, \ldots$:

$$E\left[X^m; l\right] = \int_0^l x^m \, dF_X(x) + l^m \cdot \left(1 - F_X(l)\right). \tag{2.5}$$

In the discrete case for which the values $\langle x_i \rangle$ constitute n observations for a random sample $\langle X_1, X_2, \ldots, X_n \rangle$ from a population claim-size random variable X, we denote by $\hat{X}$ the variable with the sample distribution $f_n(x_i) = 1/n$, $(1 \le i \le n)$. The ***sample limited expected value*** $E_n[\hat{X}; l]$ is a special case of (2.4):

$$E_n\left[\hat{X}; l\right] = \frac{1}{n}\sum_{i=1}^{n} \min\{x_i, l\} = \frac{1}{n}\sum_{x_i \le l} x_i + \frac{1}{n}\sum_{x_i > l} l. \tag{2.6}$$

Sometimes sample observations are grouped by size into a finite number m of non-overlapping intervals of the form $(c_{k-1}, c_k]$, where $k = 1, 2, \ldots, m$ and for which it is possible that $c_m = \infty$. Often only the claim count—and occasionally the total claim amount—in each interval is known. Whenever this is the case, probabilities for the discrete sample distribution can only be calculated accurately at the finite interval endpoints $\{c_k\}$. This is also true for the sample limited severity $E_n[\hat{X}; l]$, which is exactly computable only when $l = c_j$, $j = 1, 2, \ldots, m$. At such a point, formula (2.6) becomes

$$E_n\left[\hat{X}; c_j\right] = \frac{1}{n}\sum_{k=1}^{j} n_k a_k + \frac{c_j}{n}\sum_{k=j+1}^{m} n_k, \tag{2.7}$$

where n_k and a_k are, respectively, the number of claims and the average claim size in the k^{th} group interval $(c_{k-1}, c_k]$ and where $n = \sum_{k=1}^{m} n_k$. If the total claim amount ξ_k in $(c_{k-1}, c_k]$ is known, then it is evident that $\xi_k = n_k a_k$, or $a_k = \xi_k / n_k$. Otherwise, for group intervals of finite width, a_k can be approximated by the interval midpoint: $a_k \approx \frac{1}{2}(c_{k-1} + c_k)$. This approximation, of course, is consistent with the assumption that claim sizes are distributed uniformly throughout each group interval $(c_{k-1}, c_k]$—refer to Problem 2.5.

It is useful to consider the limited expected value $E[X; x]$ as a *function* of the variable limit x, defined on the semi-infinite interval $0 \le x < \infty$ by

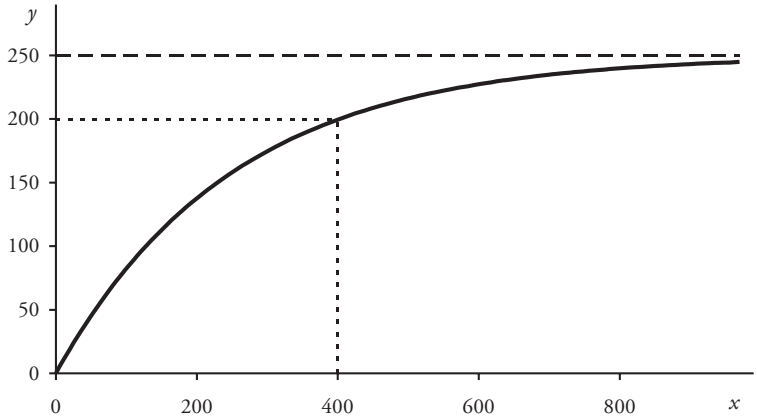
$$E\left[X; x\right] = \int_0^x u \, dF(u) + x \cdot \left(1 - F(x)\right). \tag{2.8}$$

$E[X; x]$ exists as a finite number for all $0 \le x < \infty$, even when $E[X]$ does not exist. Proof of this fact is requested in Problem 2.7.

The next examples illustrate the limited expected function $E[X; x]$ in three important cases—the first with a continuous variable X, the second with a discrete variable, and the third with a grouped claim sample.

Example 2.4. Assume that the continuous random variable X is distributed as in Example 2.2. Then the first three moments of X limited at 400 are

Figure 2.4. Limited Severity Function $y = E[X; x]$ [Example 2.4]



$$E[X;400]=\frac{1}{250}\int_0^{400} x\,e^{-x/250}\,dx+400\,e^{-1.6}=(250)\left(1-e^{-1.6}\right)=199.53,$$

$$E[X^2;400]=\frac{1}{250}\int_0^{400} x^2e^{-x/250}\,dx+(400)^2e^{-1.6}=(125{,}000)\left(1-2.6e^{-1.6}\right)=59{,}384,$$

$$E[X^3;400]=\frac{1}{250}\int_0^{400} x^3e^{-x/250}\,dx+(400)^3e^{-1.6}=20{,}310{,}141.$$

Therefore, the mean and variance of the censored variable $Y = \min\{X, 400\}$ are $E[Y] = 199.53$ and $Var[Y] = 59{,}384 - (199.53)^2 = 19{,}572$. The skewness is

$$Sk[Y]=\frac{20{,}310{,}141-(3)(199.53)(59{,}384)+(2)(199.53)^3}{(19{,}572)^{3/2}}=0.2377.$$

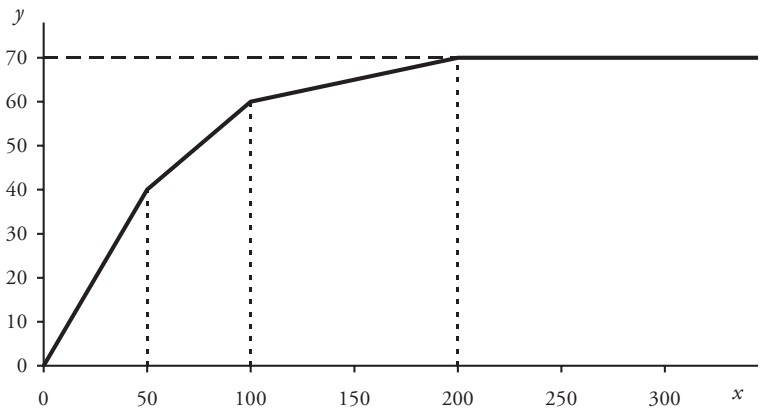
Censoring reduces not only the mean but also the variance and skewness of a random variable. The limited severity function for variable X, a graph of which is shown in Figure 2.4, is $E[X; x] = 250(1 - e^{-x/250})$. ■

Example 2.5. Consider the discrete claim-size variable X of Example 2.1. The limited severity function for X is the continuous, piecewise-linear function

$$E[X;x]=\begin{cases}0.8x & \text{if } 0\le x<50\\ 0.4x+20 & \text{if } 50\le x<100\\ 0.1x+50 & \text{if } 100\le x<200\\ 70 & \text{if } 200\le x<\infty.\end{cases}$$

Figure 2.5 displays a graph of $y = E[X; x]$. ■

Figure 2.5. Limited Severity Function $y = E[X; x]$ [Example 2.5]



Example 2.6. A random sample of 200 claims is drawn from a population with an unknown claim-size distribution.[17] These observations are grouped by size into nine group intervals of the form $(c_{k-1}, c_k]$, where $c_0 = 0$ and $c_k = 500(k + 1)$ for $k = 1, 2, \ldots, 9$. The results are displayed in the table.

Size Group	# Claims
0–1,000	42
1,001–1,500	61
1,501–2,000	47
2,001–2,500	26
2,501–3,000	14
3,001–3,500	7
3,501–4,000	2
4,001–4,500	1
4,501–5,000	0
Total	200

To calculate limited expected values at the endpoints of each group interval, we use formula (2.7) with the average claim size in the i^{th} group approximated by the interval midpoint: $a_k \approx \frac{1}{2}(c_{k-1} + c_k)$. For example, the approximate sample limited severity at $c_3 = 2{,}000$ is

$$E_{200}\left[\hat{X}; 2{,}000\right] \approx \frac{(42)(500) + (61)(1{,}250) + (47)(1{,}750)}{200}$$

$$+ \frac{(26 + 14 + 7 + 2 + 1)(2{,}000)}{200} = 1{,}398.$$

The complete set of limited expected values at the group interval endpoints is displayed in Table 2.1, along with values of the sample cumulative distribution function F_{200} at the same points. A graph of $y = E_{200}[\hat{X}; x]$ for $x = c_k$, $0 \le k \le 9$, is shown in Figure 2.6. ■

The limited severity functions of the previous examples exhibit some mathematical attributes that are shared by all such functions, notably those properties listed below.

[17] In the time-honored tradition of the textbook example, claim data used in the examples and problems throughout this monograph have been selected to illustrate clearly the concepts under study rather than obtained strictly from potentially messier real-life insurance data.

Table 2.1. Sample Limited Severities [Example 2.6]

Size x	$F_{200}(x)$	$E_{200}[\hat{X}; x]$
0	0.0000	0
1,000	0.2100	895
1,500	0.5150	1,214
2,000	0.7500	1,398
2,500	0.8800	1,490
3,000	0.9500	1,533
3,500	0.9850	1,549
4,000	0.9950	1,554
4,500	1.0000	1,555
5,000	1.0000	1,555

Properties of $E[X; x]$

Assume that X is a claim-size random variable, for which $\Pr\{X < 0\} = 0$. *Then*

$$(\boldsymbol{a}) \quad E[X; x_1] \le E[X; x_2] \text{ for all } 0 \le x_1 < x_2 < \infty. \tag{2.9}$$

$$(\boldsymbol{b}) \quad E[X; x] \text{ is continuous for all } 0 \le x < \infty. \tag{2.10}$$

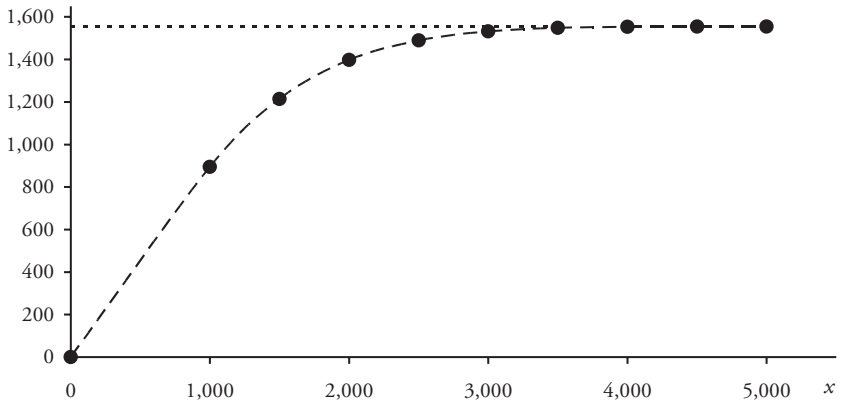
$$(\boldsymbol{c}) \quad \text{If } E[X] \text{ exists, then } E[X; x] \le E[X] \text{ for all } 0 \le x < \infty. \tag{2.11}$$

$$(\boldsymbol{d}) \quad \text{If } E[X] \text{ exists, then } \lim_{x\to\infty} E[X; x] = E[X]. \tag{2.12}$$

$$(\boldsymbol{e}) \quad E[X; x] \text{ is a concave function on } 0 \le x < \infty. \tag{2.13}$$

$$(\boldsymbol{f}) \quad E[aX + b; x] = a\,E[X; (x-b)/a] + b \text{ for constants } a > 0 \text{ and } b. \tag{2.14}$$

Figure 2.6. Sample Limited Severity Function [Example 2.6]



Proof:

(***a***) Assume that numbers (x_1, x_2) satisfy the inequality $0 \le x_1 < x_2 < \infty$. Then for random variables $Y_1 = \min\{X, x_1\}$ and $Y_2 = \min\{X, x_2\}$ we have $Y_1 \le Y_2$, so that $E[X; x_1] = E[Y_1] \le E[Y_2] = E[X; x_2]$.

(***b***) Again, assume that $0 \le x_1 < x_2 < \infty$. Then

$$0 \le \min\{X, x_2\} - \min\{X, x_1\} \le x_2 - x_1,$$

which implies that $0 \le E[X; x_2] - E[X; x_1] \le x_2 - x_1$. This means that $E[X; x]$ is uniformly continuous—and hence continuous—on $0 \le x < \infty$.

(***c***) Since $Y = \min\{X, x\} \le X$, we have $E[X; x] = E[Y] \le E[X]$.

(***d***) Existence of $E[X]$ implies that

$$\lim_{x\to\infty} \int_0^x u\, dF_X(u) = E[X] \quad \text{and} \quad \lim_{x\to\infty} \int_x^\infty u\, dF_X(u) = 0.$$

But

$$0 \le x(1 - F(x)) = \int_x^\infty x\, dF_X(u) \le \int_x^\infty u\, dF_X(u) \to 0 \quad \text{as } x \to \infty,$$

so that

$$\lim_{x\to\infty} E[X; x] = \lim_{x\to\infty} \int_0^x u\, dF_X(u) + \lim_{x\to\infty} x(1 - F(x)) = E[X] + 0.$$

(***e***) Let random variable $Y(x) = \min\{X, x\}$ be a function of x on $0 \le x < \infty$.

Then, for $0 \le x_1 < x_2 < \infty$ and $0 \le t \le 1$,

$$tY(x_1) + (1-t)Y(x_2) \le Y(t x_1 + (1-t) x_2).$$

This implies that

$$t E[X; x_1] + (1-t) E[X; x_2] \le E[X; t x_1 + (1-t) x_2] \quad \text{for } 0 \le t \le 1,$$

which means that $E[X; x]$ is a concave function.

(***f***) Let $Y = \min\{aX + b, x\}$. Then

$$Y = \min\{aX, x - b\} + b = a \min\{X, (x-b)/a\} + b.$$

Therefore, $E[aX + b; x] = E[Y] = aE[X; (x - b)/a] + b$, as required. ■

2.3. Gamma Distributions

Gamma distributions comprise a versatile family of probability distributions, with many applications in statistics and probability. Property/casualty actuaries have found them useful in constructing a variety of insurance models—parameter uncertainty for claim-count distributions, approximation of aggregate-loss distributions, and occasionally as claim-size distributions.

The ***gamma distribution*** with positive parameters (α, β) is defined by the probability density function

$$f(x)=\begin{cases}0 & \text{if } -\infty < x \le 0\\ \dfrac{1}{\beta^{\alpha}\,\Gamma(\alpha)}\,x^{\alpha-1}e^{-x/\beta} & \text{if } 0 < x < \infty \qquad (\alpha>0,\beta>0).\end{cases} \tag{2.15}$$

Except for a discontinuity at $x = 0$ when $0 < \alpha < 1$, f is an everywhere-continuous function of x. Thus, gamma distributions are of the continuous type.

The symbol Γ in (2.15) denotes the ***gamma function***,[18] defined for positive x by the convergent improper integral

$$\Gamma(x)=\int_0^{\infty} u^{x-1}e^{-u}\,du, \quad 0 < x < \infty. \tag{2.16}$$

The integral formula of Problem 2.10(*a*) implies that $\int_0^{\infty} x^{\alpha-1}\,e^{-x/\beta}\,dx = \beta^{\alpha}\,\Gamma(\alpha)$. So $\int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} f(x)dx = 1$, thus confirming that f is indeed a density function.

The gamma function is continuous on its domain and has derivatives of all orders there.[19] The following properties of the function are the most useful for our purposes. Verifications are requested in Problem 2.11.

Properties of $\Gamma(x)$

$$(\boldsymbol{a})\quad \Gamma(1)=1. \tag{2.17}$$

$$(\boldsymbol{b})\quad \Gamma(x+1)=x\,\Gamma(x), \quad 0 < x < \infty. \tag{2.18}$$

$$(\boldsymbol{c})\quad \Gamma(n+1)=n!, \quad n=1,2,3,\ldots. \tag{2.19}$$

$$(\boldsymbol{d})\quad \Gamma(x+n)/\Gamma(x)=\prod_{i=0}^{n-1}(x+i), \quad 0 < x < \infty, \quad n=1,2,3,\ldots. \tag{2.20}$$

Property (2.19) shows that $\Gamma(x)$ is an extension, to all positive real numbers, of the factorial function $n! = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n$ (where n is a positive integer), providing continuous interpolation between successive integer factorials.

The ***incomplete gamma function*** $\Gamma(x, \alpha)$ is handy in representing gamma-related distribution functions. It is defined for positive real x by the integral

$$\Gamma(x,\alpha)=\int_0^{x} u^{\alpha-1}e^{-u}\,du \quad (\alpha>0). \tag{2.21}$$

This integral is an ordinary proper integral whenever $\alpha \ge 1$ and is a convergent improper integral when $0 < \alpha < 1$. It is obvious that $\lim_{x\to\infty} \Gamma(x, \alpha) = \Gamma(\alpha)$.

---

[18] The gamma function was introduced in 1730 by Swiss mathematician Leonhard Euler (1707–1783) as a generalization of the factorial function $x!$ to nonintegral values of x. Euler proposed the integral formula $\Gamma(x) = \int_0^1 [\log(1/u)]^{x-1}\,du$, which is equivalent to (2.16). The traditional Γ notation is due to French mathematician Adrien-Marie Legendre (1752–1833).

[19] Proofs of continuity and differentiability can be found in a standard advanced calculus text.

Suppose now that random variable X has a gamma distribution with density function f as defined in (2.15). The cumulative distribution function for X can be conveniently expressed in terms of an incomplete gamma function. To observe this, start with the integral $I(x) = \int_0^x f(u)du$ $(0 < x < \infty)$ and apply the change-of-variable substitution $u = \beta v$:

$$I(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^x u^{\alpha-1} e^{-u/\beta}\, du = \frac{1}{\Gamma(\alpha)} \int_0^{x/\beta} v^{\alpha-1} e^{-v}\, dv = \frac{\Gamma(x/\beta, \alpha)}{\Gamma(\alpha)}.$$

The gamma (α, β) cumulative distribution function is therefore

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{\Gamma(x/\beta, \alpha)}{\Gamma(\alpha)} & \text{if } 0 \le x < \infty. \end{cases} \tag{2.22}$$

To derive now general formulas for the m^{th} moments of X, both unlimited and limited, begin this time with the integral $I_m(x) = \int_0^x u^m f(u)du$ and apply the same substitution $u = \beta v$ as before. For $m = 1, 2, 3, \ldots$

$$I_m(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^x u^m u^{\alpha-1} e^{-u/\beta}\, du = \frac{\beta^m}{\Gamma(\alpha)} \int_0^{x/\beta} v^{\alpha+m-1} e^{-v}\, dv = \frac{\beta^m}{\Gamma(\alpha)} \Gamma(x/\beta, \alpha + m).$$

Hence,

$$E[X^m] = \lim_{x\to\infty} I_m(x) = \frac{\beta^m}{\Gamma(\alpha)} \Gamma(\alpha + m) = \alpha(\alpha+1)\cdots(\alpha+m-1)\beta^m, \tag{2.23}$$

$$\begin{aligned} E[X^m; x] &= I_m(x) + x^m (1 - F(x)) \\ &= E[X^m] \cdot \frac{\Gamma(x/\beta, \alpha + m)}{\Gamma(\alpha + m)} + x^m \left(1 - \frac{\Gamma(x/\beta, \alpha)}{\Gamma(\alpha)}\right). \end{aligned} \tag{2.24}$$

From (2.23) it follows that

$$E[X] = \alpha\beta, \quad Var[X] = \alpha\beta^2, \quad Sk[X] = \frac{2}{\sqrt{\alpha}}. \tag{2.25}$$

Random variables with the gamma (α, β) distribution also have an important reproductive property: *the sum of independent gamma variables having the same* β *parameter is also gamma-distributed.* This result is readily obtained from an argument based on the moment-generating function for a gamma (α, β) random variable X:

$$\begin{aligned} M_X(t) = E[e^{tX}] &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x(1/\beta - t)}\, dx \\ &\overset{(2)}{=} \frac{1}{\beta^\alpha \Gamma(\alpha)} \cdot \Gamma(\alpha)(1/\beta - t)^{-\alpha} \\ &= (1 - \beta t)^{-\alpha}, \quad -\infty < t < 1/\beta, \end{aligned} \tag{2.26}$$

where the integral formula of Problem 2.10(*a*) was used at step (2). The restriction $-\infty < t < 1/\beta$ guarantees that, as a function of t, the improper integral in the first step is convergent.

Assume now that $\{X_i\}$ is a finite collection of independent gamma-distributed random variables, with identical β parameters but possibly different α parameters: $\{(\alpha_i, \beta)\}$. Independence among the X_i implies that the generating function for the sum $Y = \Sigma_i X_i$ is the product of the component generating functions. Hence,

$$M_Y(t) = \prod_i M_{X_i}(t) = \prod_i (1-\beta t)^{-\alpha_i} = (1-\beta t)^{-\Sigma_i \alpha_i}, \quad -\infty < t < 1/\beta.$$

This is the generating function of a gamma random variable. The uniqueness of the generating function thus implies that Y has a gamma $(\Sigma_i \alpha_i, \beta)$ distribution.

There are two important special cases of the gamma distribution worth noting.

(***a***) The first instance, for which the exponent parameter α is fixed at $\alpha = 1$, is the familiar exponential distribution,[20] with cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - e^{-x/\beta} & \text{if } 0 \le x < \infty \quad (\beta > 0), \end{cases} \tag{2.27}$$

and probability density function

$$f(x) = \begin{cases} 0 & \text{if } -\infty < x \le 0 \\ (1/\beta) e^{-x/\beta} & \text{if } 0 < x < \infty. \end{cases} \tag{2.28}$$

The mean, variance, and skewness for a random variable X with an exponential distribution are likewise special cases of the general formulas (2.25):

$$E[X] = \beta, \quad Var[X] = \beta^2, \quad Sk[X] = 2. \tag{2.29}$$

Furthermore,

$$E[X; x] = \beta\left(1 - e^{-x/\beta}\right) = E[X]\left(1 - e^{-x/\beta}\right). \tag{2.30}$$

We have already encountered an example of this distribution type—the claim-size random variable of Example 2.2 is exponentially distributed with $\beta = 250$. Exponential distributions have a number of actuarial applications, but they have limited practical value as size-of-loss distributions. With only a single parameter available, the exponential family is usually not flexible enough to provide a good fit to an empirical set of sample claim data. However, it is possible to adopt the Insurance Services Office (ISO) approach

[20] The exponential distribution is sometimes known as the *Laplace distribution,* in honor of French mathematician and physicist Pierre-Simon Laplace (1749–1847). Laplace made important contributions to analysis and celestial mechanics, as well as to probability. His 1812 treatise, *Théorie Analytique des Probabilités,* provided an early mathematical basis for the subject. In it he wrote "the theory of probabilities is at bottom nothing but common sense reduced to calculus. . . ."

and work with a *mixture* of exponential distributions. For example, the mixture of two such distributions has a three-parameter distribution function of the form

$$F(x)=\begin{cases}0 & \text{if } -\infty<x<0\\ 1-\omega e^{-x/\beta_1}-(1-\omega)e^{-x/\beta_2} & \text{if } 0\le x<\infty \quad (\beta_1>0,\beta_2>0,0<\omega<1).\end{cases} \tag{2.31}$$

Bureau actuaries at ISO use mixtures of up to twelve exponential distributions, involving as many as 23 parameters, to model claim size in the current ISO increased limits factor methodology.[21]

(***b***) Another important special case of the gamma distribution occurs when $\alpha=\frac{1}{2}n$ (n a positive integer) and $\beta=2$. This distribution is known as the ***chi-square distribution with n degrees of freedom*** and is denoted by $\chi^2(n)$. The probability density function is, accordingly,

$$f(x)=\begin{cases}0 & \text{if } -\infty<x\le 0\\ \dfrac{1}{2^{n/2}\,\Gamma\left(\frac{1}{2}n\right)}x^{n/2-1}e^{-x/2} & \text{if } 0<x<\infty.\end{cases} \tag{2.32}$$

A random variable X with the $\chi^2(n)$ distribution therefore has mean, variance, and skewness

$$E[X]=n,\quad Var[X]=2n,\quad Sk[X]=2\sqrt{\frac{2}{n}}. \tag{2.33}$$

The chi-square arises naturally as the distribution of the sum of squares of independent standard normal random variables. It figures prominently in the classic goodness-of-fit test—the so-called chi-square test, first introduced by British statistician Karl Pearson in 1900. In such a test the calculated test statistic is distributed under the null hypothesis according to a chi-square model.

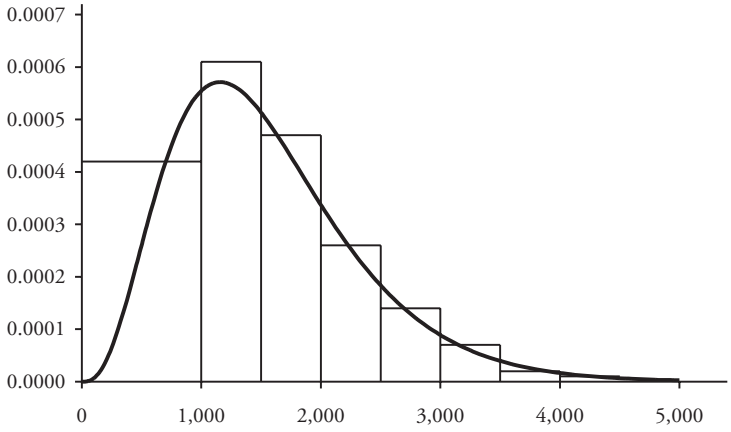
Because the gamma function and the associated distribution functions are defined by integrals with integrands having no elementary antiderivatives, evaluation of these functions necessarily involves some type of approximation. Some standard approximations are discussed in Appendix A.1.

Example 2.7. Return now to the grouped sample of 200 claims of Example 2.6. We shall attempt to fit a gamma distribution model to these data.

Begin by assuming that these data represent a random sample of claims drawn from a population having a gamma distribution with unknown parameters (α, β). To use the method-of-moments technique for estimating (α, β), we compute the first and second sample moments M_1 and M_2, based on the midpoint approximation to the average

[21] For a description of this approach, refer to the "Explanatory Memorandum" section of a current ISO Actuarial Service Circular for increased limits data and analysis (Jersey City, NJ: Insurance Services Office, Inc.); refer also to Keatinge [11]. For a discussion of mixed probability distributions like that in formula (2.31) as probability-weighted sums of conditional distributions, refer to Section 3.3.

Figure 2.7. Histogram with Gamma Density Function [Example 2.7]



claim size in each group: $M_1 = 1{,}555$ and $M_2 = 3{,}036{,}875$ (refer to Problem 2.6). Substituting these numbers into formulas (2.25) for the gamma mean and variance and then solving for α and β yields the joint method-of-moments estimators:

$$\hat{\alpha} = \frac{M_1^2}{M_2 - M_1^2} = \frac{(1{,}555)^2}{3{,}036{,}875 - (1{,}555)^2} = 3.907288,$$

$$\hat{\beta} = \frac{M_2 - M_1^2}{M_1} = \frac{618{,}850}{1{,}555} = 397.931.$$

The skewness of the resulting distribution is $2/\sqrt{3.907288} = 1.0118$.

The implied gamma probability density function is graphed in Figure 2.7 with a histogram of the sample distribution. Table 2.2 displays the limited expected values

Table 2.2. Tail Probabilities and Limited Severities [Example 2.7]

	Pr {$X > x$}		$E[X; x]$	
Size x	Sample	Gamma	Sample	Gamma
1,000	0.7900	0.7382	895	924
1,500	0.4850	0.4604	1,214	1,223
2,000	0.2500	0.2465	1,398	1,396
2,500	0.1200	0.1186	1,490	1,484
3,000	0.0500	0.0528	1,533	1,525
3,500	0.0150	0.0222	1,549	1,543
4,000	0.0050	0.0089	1,554	1,550
4,500	0.0000	0.0035	1,555	1,553
5,000	0.0000	0.0013	1,555	1,554

and tail probabilities at the group endpoints and compares the sample statistics to the corresponding gamma distribution values.

We test the goodness of fit of this gamma distribution by using the Pearson chi-square test. To implement this test, define six cells by taking the first five groups in the table of Example 2.6 and, to avoid low-frequency cells, combine the remaining four groups into a single cell. The resulting seven cell boundaries are $\{c_k\} = \{0, 1000, 1500, 2000, 2500, 3000, \infty\}$, where $k = 0, 1, 2, \ldots, 6$. The observed k^{th} cell frequency is just the tabulated sample frequency n_k for the k^{th} cell $(c_{k-1}, c_k]$. The expected frequency in the k^{th} cell is implied by the selected gamma distribution: $\phi_k(\hat{\alpha}, \hat{\beta}) = (200)(F_{\hat{\alpha},\hat{\beta}}\}(c_k) - F_{\hat{\alpha},\hat{\beta}}(c_{k-1}))$, where $F_{\hat{\alpha},\hat{\beta}}(c_6) = 1$. The chi-square statistic then has the value

$$\chi^2 = \sum_{k=1}^{6} \frac{\left(n_k - \phi_k\left(\hat{\alpha},\hat{\beta}\right)\right)^2}{\phi_k\left(\hat{\alpha},\hat{\beta}\right)}$$

$$= \frac{(42-52.37)^2}{52.37} + \frac{(61-55.56)^2}{55.56} + \frac{(47-42.77)^2}{42.77}$$

$$+ \frac{(26-25.58)^2}{25.58} + \frac{(14-13.16)^2}{13.16} + \frac{(10-10.57)^2}{10.57}$$

$$= 3.096.$$

When testing the fit of a distribution based on parameters estimated from a sample, χ^2 has $q - r - 1$ degrees of freedom, where $q = \#$ cells and $r = \#$ estimated parameters. In this example $d.f. = 6 - 2 - 1 = 3$, and so the rejection limit at the 5% significance level is $\chi^2_{0.95}(3) = 7.815$. Because $\chi^2 < 7.815$, we do not reject the null hypothesis that the fitted gamma distribution provides a reasonable description of the population claim size. ■

2.4. Lognormal Distributions

Applications of lognormal distributions are commonly found in a variety of fields—physics, reliability theory, biology, economics, to name a few. Moreover, they are widely used in property/casualty insurance to model claim size. Like their gamma-distributed counterparts, lognormal random variables take on only nonnegative values, and the distribution is positively skewed. The shape of the lognormal probability density curve $y = f(x)$ is typical of many continuous claim-size distributions—the curve rises to a maximum value in the short tail of the distribution (that is, the mode occurs at a relatively small positive value of x) and then declines asymptotically to $y = 0$ as $x \to \infty$.

Random variable X has a ***lognormal distribution*** with parameters (μ, σ) if, and only if, $\log X$ is normally distributed with mean μ and variance σ^2. Therefore, the lognormal variable X can be expressed as $X = e^{\sigma Z + \mu}$, where Z is the standard normal random variable. As a consequence, the lognormal cumulative distribution function is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x \le 0 \\ \Phi\left(\dfrac{\log x - \mu}{\sigma}\right) & \text{if } 0 < x < \infty \quad (-\infty < \mu < \infty, \sigma > 0). \end{cases} \tag{2.34}$$

Again, log x denotes the natural (base e) logarithm function of x, and Φ denotes the standard normal distribution function:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du, \quad -\infty < z < \infty.$$

The continuous lognormal variable X has probability density function

$$f_X(x) = \begin{cases} 0 & \text{if } -\infty < x \le 0 \\ \dfrac{1}{\sigma\sqrt{2\pi}\, x} \exp\left(-\tfrac{1}{2}(\log x - \mu)^2/\sigma^2\right) & \text{if } 0 < x < \infty. \end{cases} \tag{2.35}$$

Keep in mind that parameters (μ, σ) represent the mean and standard deviation not of X, but that of the normally-distributed variable log X. Although the lognormal variable X has finite moments of all orders, it turns out that $E[e^{tX}]$ is infinite for all $t > 0$, so the moment-generating function $M_X(t)$ does not exist. Nevertheless, the m^{th} moments of X are obtainable from the generating function of the standard normal variable Z, that is, from $M_Z(t) = \exp(\frac{1}{2}t^2)$. In fact, for $m = 1, 2, 3, \ldots$

$$E[X^m] = E\left[\left(e^{\sigma Z+\mu}\right)^m\right] = e^{m\mu} M_Z(m\sigma) = \exp\left(m\mu + \tfrac{1}{2}m^2\sigma^2\right). \tag{2.36}$$

The mean, variance, and skewness follow directly:

$$E[X] = e^{\mu+\sigma^2/2},$$

$$Var[X] = \left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2},$$

$$Sk[X] = \left(e^{\sigma^2} + 2\right)\sqrt{e^{\sigma^2} - 1}. \tag{2.37}$$

To derive a formula for the limited m^{th} moments, begin by evaluating the integral $I_m(x) = \int_0^x t^m f(t)dt$. The change-of-variable substitution $v = (\log t - \mu)/\sigma$ at step (2) does the trick:

$$\begin{aligned} I_m(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_0^x t^{m-1} \exp\left(-\tfrac{1}{2}(\log t - \mu)^2/\sigma^2\right) dt \\ &\overset{(2)}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\log x-\mu)/\sigma} \exp(m\sigma v + m\mu) \exp\left(-\tfrac{1}{2}v^2\right) dv \\ &= \frac{\exp\left(m\mu + \frac{1}{2}m^2\sigma^2\right)}{\sqrt{2\pi}} \int_{-\infty}^{(\log x-\mu)/\sigma} \exp\left(-\tfrac{1}{2}(v - m\sigma)^2\right) dv \\ &= \exp\left(m\mu + \tfrac{1}{2}m^2\sigma^2\right) \cdot \Phi\left(\frac{\log x - \mu - m\sigma^2}{\sigma}\right). \end{aligned} \tag{2.38}$$

Consequently,

$$E\left[X^m;x\right] = I_m(x) + x^m\left(1-F(x)\right)$$

$$= E\left[X^m\right]\cdot\Phi\left(\frac{\log x-\mu-m\sigma^2}{\sigma}\right)+x^m\,\Phi\left(\frac{-\log x+\mu}{\sigma}\right). \qquad (2.39)$$

As in the case of gamma-related distributions, evaluation of the normal and lognormal functions requires some sort of approximation. Microsoft Excel users find the worksheet functions LOGNORM.DIST and LOGNORM.INV useful—refer to Appendix A.1.

Example 2.8. Returning again to the grouped claim-size data of Example 2.6, we now attempt to fit a lognormal distribution model. This time, however, we shall use the minimum chi-square technique to estimate the distribution parameters—that is, we set the lognormal parameters μ and σ equal to the joint values for which the chi-square statistic $\chi^2(\mu, \sigma)$, as a function of the variable parameters μ and σ, achieves a minimum value.

Set up six cells as in Example 2.7, defined by the seven cell boundaries: $\{c_k\} = \{0, 1000, 1500, 2000, 2500, 3000, \infty\}$, where $k = 0, 1, 2, \ldots, 6$. As usual, the observed cell frequency is just the tabulated sample frequency n_k for the cell $(c_{k-1}, c_k]$. Expected frequencies $\phi_k(\mu, \sigma)$ are those derived from the lognormal distribution: $\phi_k(\mu, \sigma) = (200)(F_{\mu,\sigma}(c_k) - F_{\mu,\sigma}(c_{k-1}))$, in which $F_{\mu,\sigma}(x)$ is the lognormal cumulative distribution function (note that $F_{\mu,\sigma}(c_6) = 1$). The chi-square statistic then, as a function of μ and σ, is

$$\chi^2(\mu,\sigma) = \sum_{k=1}^{6}\frac{\left(n_k-\phi_k(\mu,\sigma)\right)^2}{\phi_k(\mu,\sigma)} -$$

$$= \frac{\left(42-\phi_1(\mu,\sigma)\right)^2}{\phi_1(\mu,\sigma)} + \frac{\left(61-\phi_2(\mu,\sigma)\right)^2}{\phi_2(\mu,\sigma)} + \frac{\left(47-\phi_3(\mu,\sigma)\right)^2}{\phi_3(\mu,\sigma)}$$

$$+ \frac{\left(26-\phi_4(\mu,\sigma)\right)^2}{\phi_4(\mu,\sigma)} + \frac{\left(14-\phi_5(\mu,\sigma)\right)^2}{\phi_5(\mu,\sigma)} + \frac{\left(10-\phi_6(\mu,\sigma)\right)^2}{\phi_6(\mu,\sigma)}.$$

To find values that minimize $\chi^2(\mu, \sigma)$ by analytic methods would be a daunting task, but computer software applications that use iterative algorithms often handle such problems with ease. In this example the Microsoft Excel Solver returns $(\hat{\mu}, \hat{\sigma}) = (7.274670, 0.442525)$, corresponding to a minimum value of $\chi^2(\hat{\mu}, \hat{\sigma}) = 1.828$. The minimum chi-square estimates have a built-in goodness-of-fit test—because $\chi^2(\hat{\mu}, \hat{\sigma})$ is less than the 5% rejection limit $\chi^2_{0.95}(3) = 7.815$, the fitted distribution, as in Example 2.7, is a reasonable model of the data.

The graph of the fitted probability density function is shown in Figure 2.8 along with the histogram of the observed empirical distribution. The sample and fitted lognormal limited expected values and tail probabilities at the group interval endpoints are displayed in Table 2.3.

It is instructive to compare the present lognormal model with the gamma model of Example 2.7. The two distributions have similar severities—1,555 for the gamma model

Figure 2.8. Histogram with Lognormal Density Function [Example 2.8]

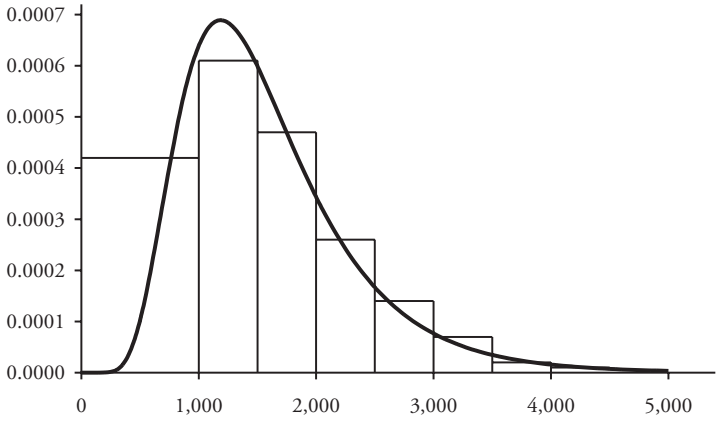


Table 2.3. Tail Probabilities and Limited Severities [Example 2.8]

	Pr {$X > x$}		$E[X; x]$	
Size x	Sample	Lognormal	Sample	Lognormal
1,000	0.7900	0.7965	895	958
1,500	0.4850	0.4653	1,214	1,273
2,000	0.2500	0.2305	1,398	1,441
2,500	0.1200	0.1072	1,490	1,522
3,000	0.0500	0.0491	1,533	1,559
3,500	0.0150	0.0227	1,549	1,576
4,000	0.0050	0.0106	1,554	1,584
4,500	0.0000	0.0051	1,555	1,588
5,000	0.0000	0.0025	1,555	1,590

and 1,592 for the lognormal—but the lognormal has the larger standard deviation and skewness: $SD = 2{,}369.675$ and $Sk = 1.4959$, compared to 786.670 and 1.0118, respectively, for the gamma. Appropriately, on the interval $0 < x < 3{,}500$ the two distributions are similar, but beyond $x = 3{,}500$ the lognormal model consistently has the larger tail probability. ■

2.5. Pareto Distributions

Pareto distributions bear the name of the eponymous Italian sociologist and economist Vilfredo Pareto (1843–1923), who first proposed using them in an 1896 textbook.[22] The distribution has long been attractive to property/casualty actuaries.

[22] In his *Cours d'Économie Politique* (Paris, 1896–97), based on lectures in economics given at Switzerland's University of Lausanne, Pareto introduced what has become known as Pareto's Law of Income Distribution. The law asserts that within a given population the proportion of individuals with incomes larger than x is modeled by a function of the general form C/x^{α}.

The computationally simple form of the distribution function—requiring only algebraic calculations and no limit processes—and the typically heavy long tail have made the Pareto family the distributional family of choice to model claim size in a variety of actuarial applications.

The classical Pareto distribution applies only to random variables with values larger than a fixed positive number γ. Such variables have a continuous cumulative distribution function of the form

$$F(x)=\begin{cases}0 & \text{if } -\infty < x < \gamma \\ 1-\left(\dfrac{\gamma}{x}\right)^{\alpha} & \text{if } \gamma \le x < \infty \quad (\alpha>0, \gamma>0)\end{cases} \tag{2.40}$$

and a corresponding density function

$$f(x)=\begin{cases}0 & \text{if } -\infty < x < \gamma \\ \dfrac{\alpha\gamma^{\alpha}}{x^{\alpha+1}} & \text{if } \gamma \le x < \infty.\end{cases} \tag{2.41}$$

Unlike the gamma- and lognormally-distributed random variables, which have moments of all orders, there exist for a Pareto random variable X only a finite number of moments. The existence of the m^{th} moment depends on the size of parameter α. In particular, the following improper integral converges—and $E[X^m]$ exists—whenever $m < \alpha$:

$$E\left[X^m\right]=\alpha\gamma^{\alpha}\int_{\gamma}^{\infty} x^{m-\alpha-1}\,dx=\frac{\alpha\gamma^{m}}{\alpha-m} \quad (m=1,2,3,\ldots,m<\alpha). \tag{2.42}$$

For example, the mean $E[X]$ exists if $\alpha > 1$, but $\alpha > 2$ is required for the variance also to exist:

$$E[X]=\frac{\alpha\gamma}{\alpha-1} \quad (\alpha>1),$$

$$Var[X]=\frac{\alpha\gamma^{2}}{(\alpha-1)^{2}(\alpha-2)} \quad (\alpha>2). \tag{2.43}$$

On the other hand, limited moments exist for all values of parameter α. For example, the limited severity function is

$$E[X;x]=\begin{cases}\gamma+\gamma\log\left(\dfrac{x}{\gamma}\right) & \text{if } \alpha=1 \\ \dfrac{\alpha\gamma}{\alpha-1}\left[1-\dfrac{1}{\alpha}\left(\dfrac{\gamma}{x}\right)^{\alpha-1}\right] & \text{if } \alpha\neq 1.\end{cases} \tag{2.44}$$

For computational convenience the classical Pareto distribution is sometimes transformed into the so-called "single-parameter" Pareto distribution. If random variable X has the distribution function (2.40), then dividing each claim by γ yields the rescaled variable $Y = X/\gamma$. This transform standardizes the minimum claim size at 1 and reduces the set of parameters from $\{\alpha, \gamma\}$ to $\{\alpha\}$.[23] The distribution function of the transformed variable is then

$$F_Y(y) = \Pr\{X/\gamma \le y\} = \Pr\{X \le \gamma y\} = F_X(\gamma y) = \begin{cases} 0 & \text{if } -\infty < y < 1 \\ 1 - \dfrac{1}{y^\alpha} & \text{if } 1 \le y < \infty. \end{cases} \tag{2.45}$$

Not surprisingly, the distribution function of Y is a special case of (2.40), for which parameter $\gamma = 1$.

As a claim-size distribution the classical Pareto distribution models only those claims in excess of a specified positive amount—a disadvantage in some applications. The most widely used form of the Pareto distribution gets around this restriction by shifting the minimum claim size to 0, as described below.

Suppose that random variable Y has the single-parameter distribution function (2.45). Applied to Y, the linear transformation $L(Y) = \beta(Y - 1) = X$, for which $\beta > 0$, first shifts the lower limit to 0 and then scales the claim size by the constant multiplier β. Consequently, $F_X(x)$ satisfies for all x

$$F_X(x) = \Pr\{L(Y) \le x\} = \Pr\{\beta Y - \beta \le x\} = \Pr\left\{Y \le \frac{x+\beta}{\beta}\right\} = F_Y\left(\frac{x+\beta}{\beta}\right).$$

The resulting random variable X is said to have the ***shifted Pareto distribution***,[24] with distribution functions

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - \left(\dfrac{\beta}{x+\beta}\right)^\alpha & \text{if } 0 \le x < \infty \quad (\alpha > 0, \beta > 0), \end{cases} \tag{2.46}$$

$$f_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}} & \text{if } 0 \le x < \infty. \end{cases} \tag{2.47}$$

[23] The distribution of Y is a single-parameter distribution only in the sense that function F_Y in (2.45) formally depends on just the single parameter α. Parameter γ is still present, however, in the preliminary scaling of random variable X.

[24] When it is unnecessary to maintain the distinction between the classical Pareto distribution function (2.40) and its shifted counterpart, most actuaries refer to (2.46) simply as the *Pareto distribution function.*

To obtain the m^{th} moments (m = 1, 2, 3, . . .) we first evaluate the integral $I_m(x) = \int_0^x u^m f(u)\, du$ by applying the change-of-variable substitution $u = \beta/v - \beta$:

$$I_m(x) = \alpha\beta^\alpha \int_0^x \frac{u^m}{(u+\beta)^{\alpha+1}} du = \alpha\beta^m \int_{\beta/(x+\beta)}^1 v^{\alpha-m-1}(1-v)^m\, dv. \qquad (2.48)$$

Then $E[X^m]$ is obtained as the limit

$$E[X^m] = \lim_{x\to\infty} I_m(x) = \alpha\beta^m \int_0^1 v^{\alpha-m-1}(1-v)^m\, dv. \qquad (2.49)$$

The integral in the right member of equation (2.49) is the special beta function $B(\alpha - m, m + 1)$. After applying a well known relation linking the beta and gamma functions,

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\, dx = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \quad (p>0, q>0),$$

we obtain

$$E[X^m] = \alpha\beta^m \frac{\Gamma(\alpha-m)\Gamma(m+1)}{\Gamma(\alpha+1)} = \frac{m!\beta^m}{(\alpha-1)(\alpha-2)\cdots(\alpha-m)} \quad (m<\alpha). \qquad (2.50)$$

Formula (2.50) yields

$$E[X] = \frac{\beta}{\alpha-1} \quad (\alpha>1),$$

$$Var[X] = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} \quad (\alpha>2). \qquad (2.51)$$

To develop now a formula for the limited m^{th} moments, start again with the integral $I_m(x)$ in equation (2.48). The binomial formula is applied at the second step in the following sequence, and the integration in the final step requires that $\alpha \neq 1, 2, \ldots, m$:

$$\begin{aligned} I_m(x) &= E[X^m] - \alpha\beta^m \int_0^{\beta/(x+\beta)} v^{\alpha-m-1}(1-v)^m\, dv \\ &= E[X^m] - \alpha\beta^m \int_0^{\beta/(x+\beta)} \left(\sum_{k=0}^m {}_mC_k (-1)^k v^{\alpha-m-1+k} \right) dv \\ &= E[X^m] - \alpha\beta^m \sum_{k=0}^m {}_mC_k (-1)^k \frac{v^{\alpha-m+k}}{\alpha-m+k}\Bigg|_0^{\beta/(x+\beta)} \\ &= E[X^m] - \alpha\left(\frac{\beta}{x+\beta}\right)^\alpha \sum_{k=0}^m {}_mC_k \frac{(-1)^k \beta^k (x+\beta)^{m-k}}{\alpha-m+k}. \end{aligned} \qquad (2.52)$$

Therefore, for $\alpha \neq 1, 2, \ldots, m$

$$E\left[X^m\right] = I_m(x) + x^m\left(\frac{\beta}{x+\beta}\right)^{\alpha}$$

$$= E\left[X^m\right] - \alpha\left(\frac{\beta}{x+\beta}\right)^{\alpha}\left(\sum_{k=0}^{m} {}_mC_k \frac{(-1)^k \beta^k (x+\beta)^{m-k}}{\alpha - m + k} - \frac{x^m}{\alpha}\right). \qquad (2.53)$$

In particular,

$$E[X; x] = \frac{\beta}{\alpha - 1}\left[1 - \left(\frac{\beta}{x+\beta}\right)^{\alpha-1}\right] \quad (\alpha \neq 1),$$

$$E\left[X^2; x\right] = E\left[X^2\right] - \alpha\left(\frac{\beta}{x+\beta}\right)^{\alpha}\left(\frac{(x+\beta)^2}{\alpha - 2} - \frac{2\beta(x+\beta)}{\alpha - 1} + \frac{\beta^2 - x^2}{\alpha}\right) \quad (\alpha \neq 1, 2).$$

(2.54)

The limited severity in the case that $\alpha = 1$ is requested in Problem 2.21.

Example 2.9. The table displays the observations from a random sample of 200 claims drawn from a population with an unknown claim-size distribution, grouped by size into ten groups. Note that in this case the right-most group interval is a semi-infinite interval: $10{,}000 < x < \infty$.

Size Group	# Claims
0–2,000	56
2,001–3,000	22
3,001–4,000	18
4,001–5,000	16
5,001–6,000	14
6,001–7,000	12
7,001–8,000	10
8,001–9,000	8
9,001–10,000	7
>10,000	37
Total	200

In this example we fit a shifted Pareto (α, β) distribution to these data by minimizing, as a function of α and β, the "distance" between the sample limited expected values and the corresponding Pareto statistics at the finite endpoints of the sample groups: $c_k = 1000\ k$, where $k = 0, 2, 3, \ldots, 10$. That is, the desired least-squares estimators $(\hat{\alpha}, \hat{\beta})$ are parameter values that minimize the quasi-distance function

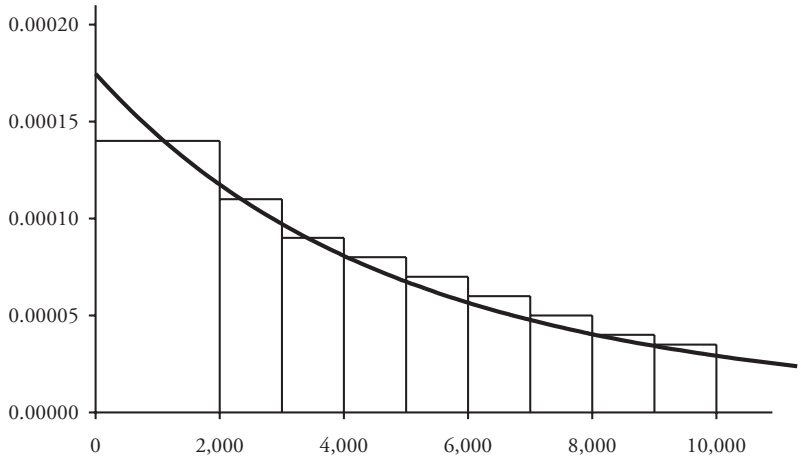
$$D(\alpha,\beta) = \sqrt{\sum_{i=2}^{10}\left|E_{\alpha,\beta}[X; c_i] - E_n\left[\hat{X}; c_i\right]\right|^2}.$$

In this equation variable X has a shifted Pareto (α, β) distribution and $\hat{X}$ has the discrete sample distribution. The components of $D(\alpha, \beta)$ are thus defined by

$$E_{\alpha,\beta}[X; c_i] = \frac{\beta}{\alpha - 1}\left[1 - \left(\frac{\beta}{c_i + \beta}\right)^{\alpha-1}\right] \quad \text{and}$$

$$E_n\left[\hat{X}; c_i\right] = \frac{1}{2n}\sum_{k=2}^{i} n_k\left(c_{k-1} + c_k\right) + \frac{c_i}{n}\sum_{k=i+1}^{11} n_k, \quad i = 2, 3, \ldots, 10,$$

Figure 2.9. Histogram with Pareto Density Function [Example 2.9]



where n_k = # claims in the k^{th} group $(c_{k-1}, c_k]$, $n_{11} = 37$, and $n = 200$. The average claim size for each finite group interval has been set to the interval midpoint in the formula for $E_n[\hat{X}; c_i]$. Solving iteratively yields $\hat{\alpha} = 6.000000$ and $\hat{\beta} = 34{,}355.3719$.[25] Figure 2.9 compares the graph of the implied Pareto density function to the histogram of the observed sample distribution.

Using the ten distribution groups as cells, we apply the chi-square test and obtain $\chi^2 = 1.793$. Since *d.f.* = 10 − 2 − 1 = 7, the rejection limit is $\chi^2_{0.95}(7) = 14.067$. The fitted Pareto distribution is therefore, at the 5% level of significance, a reasonable fit to these claim data. Table 2.4 compares the sample and Pareto limited expected values and tail probabilities. ■

2.6. Estimation with Modified Data

The fact that most available insurance claim-size data are modified by such common policy conditions as limits and deductibles presents additional challenges to the problem of fitting a distributional model to the unknown underlying *unmodified, unlimited* claim-size distribution for a portfolio of policies. In this section we consider some possible techniques for parameter estimation under such conditions. Generally speaking, one must either adjust the data to remove the effects of the policy modifications or modify the parametric distribution formulas to model the data modifications—or use a combination of both approaches. We begin in Example 2.10 with a set of policy data censored by a policy limit and then in Example 2.11 take up the problem of data both censored by a policy limit and truncated by a deductible.

Example 2.10. A sample contains n = 1,500 claims from a large portfolio of policies, each with a policy limit of $300,000. These claim data are summarized in Table 2.5. For a sequence of selected claim sizes x the number and total amount for

[25] The indicated solution was obtained by using the Solver utility in Microsoft Excel. To facilitate the iterative process, parameter α was arbitrarily fixed at $\hat{\alpha} = 6$ and the corresponding $\hat{\beta}$ obtained iteratively.

Table 2.4. Tail Probabilities and Limited Severities [Example 2.9]

	$\Pr\{X > x\}$		$E[X; x]$	
Size x	Sample	Pareto	Sample	Pareto
2,000	0.7200	0.7121	1,720	1,693
3,000	0.6100	0.6051	2,385	2,350
4,000	0.5200	0.5164	2,950	2,910
5,000	0.4400	0.4425	3,430	3,388
6,000	0.3700	0.3807	3,835	3,799
7,000	0.3100	0.3287	4,175	4,153
8,000	0.2600	0.2848	4,460	4,459
9,000	0.2200	0.2476	4,700	4,724
10,000	0.1850	0.2159	4,903	4,956

claims less than or equal x have been tabulated. Moreover, there are 23 claims with the policy-limit value of 300,000. We wish to use these censored data to find an unlimited lognormal distribution for the unmodified claim population underlying this portfolio. Such a distribution will be useful in creating a set of increased limit factors for pricing policy limits greater than 300,000 (this topic is discussed in detail in Chapter 6). As in Example 2.9, we shall use the method of minimum-distance estimation to obtain the desired parameters.

Note that for all claim sizes x, $x \le 300{,}000$, sample limited expected values can be calculated accurately from the summarized sample data—for example, at $x = 5{,}000$ we have

$$E_n\left[\hat{X};5{,}000\right] = \frac{1{,}102{,}272 - (5{,}000)(1{,}500 - 1{,}096)}{1{,}500} = 2{,}082.$$

Table 2.5. Censored Data and Limited Severities [Example 2.10]

Size x	# Claims $\le x$	Σ Claims $\le x$	$E_n[\hat{X}; x]$	$E_{\mu,\sigma}[X; x]$
1,000	729	225,138	664	648
5,000	1,096	1,102,272	2,082	2,091
10,000	1,208	1,918,947	3,226	3,239
25,000	1,326	3,752,091	5,401	5,410
50,000	1,391	6,007,543	7,638	7,580
100,000	1,440	9,234,739	10,156	10,168
200,000	1,468	13,100,561	13,000	13,069
300,000	1,500	22,343,455	14,896	14,850

(However, the same cannot be said for any $x > 300{,}000$.) The lognormal limited mean $E_{\mu,\sigma}[X; x]$ is obtained from formula (2.39) with $m = 1$:

$$E_{\mu,\sigma}[X; x] = \exp\left(\mu + \tfrac{1}{2}\sigma^2\right) \cdot \Phi\left(\frac{\log x - \mu - \sigma^2}{\sigma}\right) + x \cdot \Phi\left(\frac{-\log x + \mu}{\sigma}\right).$$

Minimizing the quasi-distance function

$$D(\mu,\sigma) = \sqrt{\sum_x \left| E_{\mu,\sigma}[X; x] - E_n\left[\hat{X}; x\right]\right|^2}$$

over all parameter values yields $(\hat{\mu}, \hat{\sigma}) = (6.9852, 2.5850)$, and the resulting limited expected values are displayed in the fifth column of Table 2.5.

As usual with grouped claim data, we can easily apply the chi-square statistic to the nine cells defined by the sequence of claim sizes in Table 2.5:

$$\{0, 1K, 5K, 10K, 25K, 50K, 100K, 200K, 300K, \infty\}.$$

We set the observed frequency of the last cell $(300K, \infty)$ to be 23, the number of limit claims. The chi-square statistic $\chi^2 = 2.763$ is less than the 5% rejection limit for *d.f.* = 6, $\chi^2_{0.95}(6) = 12.6$, so we conclude that the lognormal with fitted parameters $(\hat{\mu}, \hat{\sigma})$ is an acceptable distribution for the underlying claim population for this portfolio. ■

The policy condition known as a ***straight deductible*** eliminates all claims less than or equal to the deductible amount *d,* where $d > 0$, and it reduces the size of larger claims by *d.* (Section 6.5 contains a more extended discussion of deductible concepts.) Thus, a claim sample generated by a portfolio of policies with a straight deductible would be missing all original claims of size *d* or less and sizes of the remaining claims would be reduced by the amount *d.* A sample or random variable with this property is said to be ***truncated below by d and shifted by d***. The next example illustrates how an unlimited parametric distribution model could be fitted to an underlying claim population, given only a sample of truncated and shifted claim-size data.

Example 2.11. A sample contains 770 claims from a portfolio of identical policies, each with a policy limit of $200,000 and a straight deductible of $1,000. Thus, the sample data have been censored above at 200,000 and then truncated below by 1,000 and shifted by the same amount. Adding back the 1,000 deductible amount to each claim in the sample removes the shift effect of the deductible, resulting in an adjusted sample of claim sizes censored above by 200,000 and truncated below by 1,000. These adjusted claim sizes have been sorted into 12 groups, with the observed group frequencies displayed in Table 2.6. Thirty-one claims valued at the policy limit were placed in the last group $(200K, \infty)$.

To fit a lognormal distribution to the underlying unmodified claim population, we shall use the minimum chi-square method of parameter estimation.

Table 2.6. Truncated and Censored Data [Example 2.11]

Size Group	Obs # Claims	Exp # Claims
0–1*K*	0	0
1*K*–5*K*	367	360
5*K*–10*K*	112	122
10*K*–25*K*	118	121
25*K*–50*K*	65	63
50*K*–75*K*	36	27
75*K*–100*K*	13	16
100*K*–125*K*	10	10
125*K*–150*K*	8	7
150*K*–175*K*	6	5
175*K*–200*K*	4	4
>200*K*	31	34
Total	770	770

Note that if X is a non-truncated random variable, then X truncated below by 1,000 is defined only for $1{,}000 < X < \infty$ by

$$X_{1000} = X.$$

Thus, the expected frequency of cell $(c_{k-1}, c_k]$ $(k = 1, 2, \ldots 12)$ in terms of the cumulative distribution function $F_{\mu,\sigma}(x)$ of an unmodified, unlimited lognormal distribution is

$$\phi_k(\mu,\sigma) = (770)\frac{\left(F_{\mu,\sigma}(c_k) - F_{\mu,\sigma}(c_{k\text{-}1})\right)}{1 - F_{\mu,\sigma}(1{,}000)}.$$

Minimizing the chi-square statistic

$$\chi^2(\mu,\sigma) = \sum_{k=1}^{12} \frac{(n_k - \phi_k(\mu,\sigma))^2}{\phi_k(\mu,\sigma)}$$

over all μ and σ yields the estimated parameters $(\hat{\mu},\hat{\sigma}) = (6.6916, 2.6965)$. Corresponding expected cell frequencies are shown in the third column of Table 2.6.

The minimum chi-square statistic $\chi^2(\hat{\mu}, \hat{\sigma}) = 4.691$ is less than the 5% rejection limit $\chi^2_{0.95}(8) = 15.5$, so we can conclude that the fitted lognormal distribution is an acceptable description of the underlying unlimited claim-size distribution. ■

Other examples of parameter estimation based on modified data can be found in Problems 2.39 and 2.43. In addition, we shall return to the important concept of truncated

random variables and data in Sections 5.1 and 6.5, as well as in Problems 2.41 and 2.42. For a slightly different approach to the estimation problem addressed in Example 2.11, refer to Example 5.3 and Problem 5.23.

2.7. Transformations

With claim-size random variables, as with random variables in general, one can create new variables by transforming existing ones. This is often done in order to create claim-size models with predetermined properties or with properties somewhat different from, but related to, those of a known variable. In this section we shall focus on those functions that transform one continuous claim-size random variable into another.

Assume that T is a strictly increasing—and hence invertible—continuous and differentiable function that maps a set of nonnegative real numbers into itself. If X is a continuous claim-size variable, then $Y = T(X)$ is also a continuous random variable with nonnegative values. As such, Y is also a possible random variable for the size of insurance claims. Because T is an increasing function, distribution functions for X and Y are related by

$$F_Y(y) = \Pr\{T(X) \le y\} = \Pr\{X \le T^{-1}(y)\} = F_X\left(T^{-1}(y)\right), \quad 0 < y < \infty. \tag{2.55}$$

Moreover, if F_X is differentiable at $x = T^{-1}(y)$, then

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(T^{-1}(y)\right) = f_X\left(T^{-1}(y)\right) \frac{d}{dy} T^{-1}(y). \tag{2.56}$$

The simplest such function is the linear transformation $L(X) = aX + b$, where a and b are real constants and $a > 0$. For a continuous variable X and $Y = L(X) = aX + b$, distribution functions (2.55) and (2.56) become, respectively,

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right) \quad \text{and} \quad f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right), \quad 0 < y < \infty. \tag{2.57}$$

Linear transformations appear in a variety of probability settings, where they are used to translate the values of a random variable, up or down by a fixed amount, or to rescale the values by applying a constant multiplier. For example, we previously observed in Section 2.5 that the linear transformation $L(Y) = \beta(Y - 1) = X$ creates a shifted Pareto random variable X from the classical single-parameter Pareto variable Y. In addition, we encountered in Section 2.4 the transformation $T(Z) = e^{\sigma Z + \mu} = X$, transforming the standard normal variable Z first by the linear function $\sigma Z + \mu$ and then by the exponential function, to define X as a lognormal claim-size random variable.

With regard to distribution characteristics, it is well known that $L(X) = aX + b$ transforms the mean of the random variable when either $a \neq 1$ or $b > 0$—specifically, $E[L(X)] = L(E[X]) = aE[X] + b$. It also transforms the variance whenever $a \neq 1$: $Var[L(X)] = a^2\, Var[X]$. However, when $a > 0$ the skewness of a distribution remains unchanged under such a linear transform: $Sk[L(X)] = Sk[X]$. Proof of this invariance property is requested in Problem 2.26.

Table 2.7. Effect of $L_c(X)$ on Distribution Parameters

Distribution Family	X Parameters	$L_c(X)$ Parameters
Normal	μ, σ	$c\mu, c\sigma$
Gamma	α, β	$\alpha, c\beta$
Exponential	β	$c\beta$
Lognormal	μ, σ	$\mu + \log c, \sigma$
Shifted Pareto	α, β	$\alpha, c\beta$
Weibull	β, δ	$c\beta, \delta$
Burr	α, β, δ	$\alpha, c^{\delta}\beta, \delta$

In some cases the linear transformation $L(X)$ does not change the parametric family of the initial distribution of variable X, but only alters the parameters within the family. For example, consider

$$L_c(X) = cX, \quad c > 0. \tag{2.58}$$

If random variable X has a gamma (α, β) distribution, then the transformed variable $Y = L_c(X)$ has the cumulative distribution function

$$F_Y(y) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{\Gamma(y/(c\beta), \alpha)}{\Gamma(\alpha)} & \text{if } 0 \le x < \infty. \end{cases} \tag{2.59}$$

Consequently, Y is also gamma-distributed, but with parameters $(\alpha, c\beta)$. A similar outcome is obtained when transformation (2.58) is applied to a random variable with one of several other common parametric distributions. The results of applying L_c to X with the normal, exponential, lognormal, and Pareto distributions—as well as with the Weibull and Burr distributions defined in Examples 2.12 and 2.14—are shown in Table 2.7.

Another class of transformations important to the study of claim-size random variables are those functions having the form

$$T(X) = cX^{1/\delta} \quad (c > 0, \delta > 0) \tag{2.60}$$

Such a transformation can be employed to produce a variable $T(X)$ with distributional tail characteristics differing from those of X, as illustrated in Examples 2.12 and 2.13. Parameter δ serves to alter the thickness of the long tail of the distribution. In general, the distribution of $T(X)$ has a heavier long tail than that of X whenever $0 < \delta < 1$. On the other hand, if $\delta > 1$, then X has the heavier-tailed distribution.

Example 2.12. Consider the random variable X, defined by

$$X = T(Y) = \beta^{(\delta-1)/\delta} Y^{1/\delta} \quad (\beta > 0, \delta > 0)$$

in which transformation T is a special case of (2.60). If Y has an exponential (β) distribution, then formula (2.55) yields for the transformed variable X the cumulative distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - \exp\left(-(x/\beta)^{\delta}\right) & \text{if } 0 \le x < \infty \quad (\beta > 0, \delta > 0). \end{cases} \tag{2.61}$$

Here X is said to have a ***Weibull distribution***, after Swedish engineer E.H.W. Weibull (1887–1979). In a 1939 paper Weibull proposed the distribution as a model for the random failure time of various parts of mechanical systems, for which $F(t) = \Pr\{\textit{failure time} \le t\}$. A later paper [22], published by Weibull in 1951, served to promote the distribution in the U.S.

The Weibull distribution is known for its exceptional ability to fit a wide variety of data, and it is widely employed in reliability engineering and failure analysis. Of course, when $\delta = 1$ the distribution reduces to the special case of an exponential distribution. Otherwise—especially when $0 < \delta < 1$—it is useful in modeling size-of-loss distributions.

Formulas for the moments of the Weibull distribution are obtained from the integral $I_m(x) = \int_0^x u^m f(u)\, du$, where the p.d.f. is $f(x) = (\delta/\beta^{\delta})\, x^{\delta-1}\, e^{-(x/\beta)^{\delta}}$:

$$I_m(x) = \frac{\delta}{\beta^{\delta}} \int_0^x u^{m+\delta-1} \exp\left(-(u/\beta)^{\delta}\right) du \overset{(2)}{=} \beta^m \int_0^{(x/\beta)^{\delta}} v^{m/\delta} e^{-v}\, dv = \beta^m \Gamma\left((x/\beta)^{\delta}, 1 + m/\delta\right),$$

for $m = 1, 2, 3, \ldots$. Note that the change-of-variable substitution $v = (u/\beta)^{\delta}$ was used at step (2). Therefore, the Weibull m^{th} moments are

$$E\left[X^m\right] = \lim_{x\to\infty} I_m(x) = \beta^m \int_0^{\infty} v^{m/\delta} e^{-v}\, dv = \beta^m\, \Gamma(1 + m/\delta), \tag{2.62}$$

$$\begin{aligned} E\left[X^m; x\right] &= I_m(x) + x^m\left(1 - F(x)\right) \\ &= E\left[X^m\right] \cdot \frac{\Gamma\left((x/\beta)^{\delta}, 1 + m/\delta\right)}{\Gamma(1 + m/\delta)} + x^m e^{-(x/\beta)^{\delta}}. \end{aligned} \tag{2.63}$$

Again, because the Weibull distribution functions and moments are expressed in terms of the gamma function, evaluation necessarily involves the use of approximation techniques. ■

The next example, using a set of related Weibull variables, illustrates how the size of parameter δ in (2.61) affects the distribution of probability in the long tail of a claim-size distribution.

Example 2.13. Consider three related Weibull random variables: X_1 has parameters $(\beta_1, \delta_1) = (220.653, 0.80)$, X_2 has parameters $(\beta_2, \delta_2) = (250, 1.00)$, and $(\beta_3, \delta_3) = (265.774, 1.20)$ are parameters for X_3. The means of X_1, X_2, and X_3 are therefore identical:

$$E[X_1] = (220.653)\Gamma(1 + 1/0.80) = 250,$$

$$E[X_2] = (250)\Gamma(2) = 250,$$

$$E[X_3] = (265.774)\Gamma(1 + 1/1.20) = 250.$$

Tail probabilities for these variables are compared in Table 2.8, clearly indicating the effect of parameter δ on the thickness of the long tail. ■

Example 2.14. As another application of transformation (2.60) consider now the random variable X defined by $X = T(Y) = Y^{1/\delta}$, where Y has the shifted Pareto (α, β) distribution:

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1 - \left(\dfrac{\beta}{x^{\delta} + \beta}\right)^{\alpha} & \text{if } 0 \le x < \infty \quad (\alpha > 0, \beta > 0, \delta > 0), \end{cases} \tag{2.64}$$

$$f_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{\alpha\beta^{\alpha}\delta x^{\delta-1}}{(x^{\delta} + \beta)^{\alpha+1}} & \text{if } 0 \le x < \infty. \end{cases} \tag{2.65}$$

Such a transformed distribution is called the ***Burr distribution***, after the Purdue University statistician Irving Wingate Burr, who first proposed its use. Clearly, the Burr distribution is a generalization of the shifted Pareto, to which it reduces when $\delta = 1$. Burr made numerous contributions to reliability theory, statistical quality control, and distribution theory. He introduced the distribution in 1942 as one suitable for modeling failure times in reliability engineering. ■

Table 2.8. Weibull Tail Probabilities [Example 2.13]

Size x	$\Pr\{X_1 > x\}$ $\delta_1 = 0.80$ $\beta_1 = 220.653$	$\Pr\{X_2 > x\}$ $\delta_2 = 1.00$ $\beta_2 = 250.000$	$\Pr\{X_3 > x\}$ $\delta_3 = 1.20$ $\beta_3 = 265.774$
200	0.3968	0.4493	0.4912
300	0.2784	0.3012	0.3146
400	0.2000	0.2019	0.1953
500	0.1460	0.1353	0.1183
600	0.1079	0.0907	0.0702
700	0.0806	0.0608	0.0409
800	0.0607	0.0408	0.0235
900	0.0460	0.0273	0.0133
1,000	0.0351	0.0183	0.0074

2.8. Inflation Effects

When the claim process to be modeled is subject to some type of inflationary pressure applied over time, one must account for this in a probability model for the size of claims. Such time-dependent forces can arise from a variety of sources. *Monetary inflation* results from the changing, usually declining, value of the underlying currency. *Social* or *judicial inflation* occurs when changes take place in the societal or legal environment—changes that often affect the size of insurance claims. In contrast to monetary inflation, which usually gives rise to increasing claim size, social and judicial inflation could possibly result in a decrease as well as an increase in the size of claims. Of course, these types of inflationary pressure can also affect the *frequency* of claims, the subject addressed in Chapter 3.

Often a claim-size distribution, whether an empirical distribution based on a population of actual claims or a continuous parametric model as discussed in this chapter, must be adjusted for inflationary trend to account for past changes or to model change projected for the future. The simplest approach is to assume that all claims in the population are impacted in the same way by inflation, as is clearly the case with monetary inflation. Accordingly, we shall first study the concept of *uniform trend* and then take up one approach to the concept of *variable trend.*

Suppose that claim-size random variable X is subject to a uniform inflationary trend over a period of time. This means that every claim, large and small, changes by the same percentage during the time period. That is, X is transformed into a new random variable $Y = T(X) = \tau X$, where $\tau = 1 + r$ is the trend factor and r is the inflation rate for the period.

For example, assume that claim size is increasing at the uniform rate of 5% *per annum.* Then the constant trend factor for a single year is $\tau_1 = 1.05$, whereas for a three-year period the factor is $\tau_3 = (1.05)^3 = 1.1576$.

Transformation $T(X) = \tau X$ is a linear transformation of the form (2.58), and so the cumulative distribution function of the transformed variable Y is a special case of formula (2.57):

$$F_Y(y) = F_X(y/\tau), \quad 0 < y < \infty. \tag{2.66}$$

For example, if X has a lognormal (μ, σ) distribution, then

$$F_Y(y) = \begin{cases} 0 & \text{if } -\infty < y \le 0 \\ \Phi\left(\dfrac{\log y - \log\tau - \mu}{\sigma}\right) & \text{if } 0 < y < \infty. \end{cases}$$

As shown previously in Table 2.7, Y also has a lognormal distribution, but with parameters $(\mu + \log\tau, \sigma)$.

When a uniform trend factor is applied to a censored random variable the nonlinearity of the limited expected value $E[X; x]$ with respect to the random variable X serves to modify the effect of inflation on the average censored claim size. For example,

consider a claim-size random variable X subject to the fixed positive upper limit l. The average claim size before trending is $E[X; l]$, and the severity after applying $\tau = 1 + r$ is given by equation (2.14):

$$E[\tau X; l] = \tau\, E[X; l/\tau]. \tag{2.67}$$

If $\tilde{r}$ denotes the effective rate of change on the censored variable, then

$$1 + \tilde{r} = \frac{E[\tau X; l]}{E[X; l]} = (1 + r)\frac{E[X; l/(1+r)]}{E[X; l]}. \tag{2.68}$$

$E[X; x]$ is a nondecreasing function of x, so it follows that

$$|\tilde{r}| \le |r|. \tag{2.69}$$

This overall reduction in the effect of inflation on claim size is due to the fact that all censored claims—those larger than l—are unchanged by the force of inflation. For example, if $x > l$ and $r > 0$ then x and $(1 + r)x$ are each replaced by l in the calculation of $E[X; l]$ and $E[\tau X; l]$.

However, if limit l is subjected to the same trend factor as the claim size—so that after trending the severity is $E[\tau X; \tau l]$—then this leveraging effect of the upper limit disappears. Proof of this assertion is requested in Problem 2.32.

Example 2.15. Random variable X has a shifted Pareto distribution with $(\alpha, \beta) = (2; 3{,}000)$. The average claim size subject to a policy limit of \$8,000 is

$$E[X; 8{,}000] = (3{,}000)\left(1 - \frac{3{,}000}{8{,}000 + 3{,}000}\right) = 2{,}182.$$

Application of a uniform 10% trend to X yields the limited severity

$$E[1.1X; 8{,}000] = (1.1)E[X; 8{,}000/1.1] = (3{,}300)\left(1 - \frac{3{,}000}{8{,}000/1.1 + 3{,}000}\right) = 2{,}336.$$

Consequently, the effective inflation rate for the limited variable is less than the nominal 10% rate: $\tilde{r} = 2{,}336/2{,}182 - 1 = 7.1\%$.

If, on the other hand, X is subjected to a negative annual trend of −5% so that $\tau = 0.95$, then

$$E[0.95X; 8{,}000] = (0.95)E[X; 8{,}000/0.95] = 2{,}101.$$

In this case,

$$\tilde{r} = \frac{2{,}101}{2{,}182} - 1 = -3.7\%. \blacksquare$$

The assumption of a uniform trend—claims of all sizes are subject to the same rate of change—is not always satisfied in practice. Empirical evidence sometimes suggests that the trend factor should in some way be an increasing function of the claim-size variable X. In a study of non-uniform trend models Sheldon Rosenberg and Aaron Halpert proposed an annual trend factor of the form

$$\tau(x) = a\,x^{b} \quad (a>0, b>0).^{26} \tag{2.70}$$

Note that factor (2.70) reduces to the uniform case when $b = 0$; otherwise $\tau(x)$ is an increasing function of x.

The trended random variable Y is therefore

$$Y = \tau(X)\cdot X = a\,X^{b+1}. \tag{2.71}$$

When X has a lognormal (μ, σ) distribution the distribution function of the trended variable Y is

$$F_Y(y) = \Phi\left(\frac{\log\left((y/a)^{1/(b+1)}\right)-\mu}{\sigma}\right) = \Phi\left(\frac{\log y-(b+1)\mu-\log a}{(b+1)\sigma}\right), \quad 0<y<\infty. \tag{2.72}$$

This implies that Y is also lognormally distributed, with parameters

$$(\tilde{\mu},\tilde{\sigma}) = \big((b+1)\mu+\log a,(b+1)\sigma\big).$$

However, X and Y in (2.71) do not always belong to the same distribution family. If X has a shifted Pareto (α, β) distribution, for example, then Y has distribution function

$$F_Y(y) = 1-\left(\frac{\beta}{(y/a)^{1/(b+1)}+\beta}\right)^{\alpha} = 1-\left(\frac{a^{1/(b+1)}\beta}{y^{1/(b+1)}+a^{1/(b+1)}\beta}\right)^{\alpha}, \quad 0\le y<\infty, \tag{2.73}$$

which defines a Burr distribution with parameters

$$\big(\tilde{\alpha},\tilde{\beta},\tilde{\delta}\big) = \big(\alpha, a^{1/(b+1)}\beta, 1/(b+1)\big).$$

Example 2.16. Random variable X has a lognormal distribution with parameters $(\mu, \sigma) = (7.2, 0.476)$ and thus has mean

$$E[X] = \exp\left(7.2+\tfrac{1}{2}(0.476)^2\right) = 1{,}500.$$

Applying the variable trend factor $\tau(x) = 0.96x^{0.0183}$ yields a new lognormal variable $X_\tau = \tau(X)\cdot X = 0.96X^{1.0183}$ with mean

$$E[X_\tau] = \exp\left((1.0183)(7.2)+\log 0.96+\tfrac{1}{2}(1.0183)^2(0.476)^2\right) = 1{,}650.$$

[26] Rosenberg and Halpert [20], p. 466.

Table 2.9. Variable Trend Factor [Example 2.16]

Claim Size x	$\tau(x) = 0.96x^{0.0183}$
100	1.0444
500	1.0756
750	1.0836
1,000	1.0894
1,500	1.0975
1,650	1.1000
2,000	1.1033
3,000	1.1115
4,000	1.1173
5,000	1.1219

Trending has increased the overall unlimited mean of the distribution by 10%: $E[X_\tau]/E[X] = 1{,}650/1{,}500 = 1.10$. Table 2.9 displays values of the variable trend factor τ for several claim sizes. ■

2.9. Problems

2.1 A continuous claim-size random variable X takes on values larger than or equal to 1,000 and has the Pareto cumulative distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 1{,}000 \\ 1 - (1{,}000/x)^3 & \text{if } 1{,}000 \le x < \infty. \end{cases}$$

Evaluate:

(***a***) $E[X]$. (***b***) $Var[X]$.

(***c***) $\Pr\{X > 2{,}000\}$. (***d***) $E[X; 2{,}000]$.

2.2 Claim-size random variable Y has the cumulative distribution function

$$F_Y(y) = \begin{cases} 0 & \text{if } -\infty < y < 500 \\ 1 - (0.75)(500/y)^2 & \text{if } 500 \le y < \infty. \end{cases}$$

Evaluate:

(***a***) $E[Y]$. (***b***) $Var[Y]$.

(***c***) $\Pr\{Y = 500\}$. (***d***) $E[Y; 1{,}000]$.

2.3 Derive the limited severity function for random variable Y with the mixed distribution of Example 2.3.

2.4 The table displays the grouped claim sample data of Example 2.6, but with the total claim amount in each group now included. Calculate the mean of the sample distribution, as well as the limited severities at the endpoints of each group interval. Compare these results to those obtained in Example 2.6 and explain the observed differences.

Size Group	# Claims	Total Claim Amount
0–1,000	42	20,370
1,001–1,500	61	74,725
1,501–2,000	47	82,250
2,001–2,500	26	57,200
2,501–3,000	14	37,800
3,001–3,500	7	22,400
3,501–4,000	2	7,200
4,001–4,500	1	4,400
4,501–5,000	0	0
Total	200	306,345

2.5 Demonstrate that the midpoint approximation to a_k in formula (2.7) is consistent with the assumption, but does not necessarily imply, that claims are distributed uniformly on each group interval of finite width.

2.6 Using the notation of formula (2.7) for grouped sample data and the midpoint approximation $a_k \approx \frac{1}{2}(c_{k-1} + c_k)$, develop formulas for approximating the sample moments M_1 and M_2.

2.7 Verify that for every claim-size random variable X, $E[X; x]$ exists as a finite number. Cite an example for which $E[X; x] < E[X] = \infty$.

2.8 Show that for a discrete claim-size variable X, $E[X; x]$ is a piecewise linear function on the interval $0 \le x < \infty$.

2.9 Assume that X is a continuous claim-size random variable with a density function $f(x) = F'(x)$ continuous on the interval $0 < x < \infty$.

(***a***) Show that function $E[X; x]$ is differentiable on $0 < x < \infty$.

(***b***) Prove that the limited severity function for X can be expressed as $E[X; x] = \int_0^x (1 - F(u))\, du$.

(***c***) Use the second derivative test from elementary calculus to verify that function $E[X; x]$ is concave.

2.10 Show that the gamma function $\Gamma(x)$ defined by equation (2.16) can also be expressed by each of these integral formulas.

(***a***) $\Gamma(x) = c^x \int_0^\infty u^{x-1}\, e^{-cu}\, du,\ c > 0.$

(***b***) $\Gamma(x) = 2\int_0^\infty u^{2x-1}\, e^{-u^2}\, du.$

(***c***) $\Gamma(x) = \int_0^1 (\log(1/u))^{x-1}\, du.$

2.11 Prove these properties of the gamma function $\Gamma(x)$.
(***a***) Equation (2.17). (***b***) Equation (2.18).
(***c***) Equation (2.19). (***d***) Equation (2.20).

2.12 Derive these values of $\Gamma(x)$.
(***a***) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.
(***b***) $\Gamma\left(n+\frac{1}{2}\right) = \dfrac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n}\sqrt{\pi}, \quad n = 1, 2, 3, \ldots$.

2.13 Random variable X has a gamma (α, β) distribution for which $E[X] = \sqrt{Var[X]}$. What can be said about α and β?

2.14 Assume that X has an exponential distribution. For $a > 0$ and $b > 0$ calculate $\Pr\{X > a + b \mid X > a\}$. Interpret the result.

2.15 Assume that X has the mixed exponential distribution with cumulative distribution function (2.31). Calculate:
(***a***) $E[X]$. (***b***) $Var[X]$. (***c***) $E[X; x]$.

2.16 Use the minimum chi-square method to estimate the gamma parameters of Example 2.7. Compare the mean and variance of the resulting gamma distribution with the sample statistics. Which of the two gamma distributions, that obtained by the method-of-moments or that obtained by the minimum chi-square method, provides a better fit to the data?

2.17 In a certain claim population the claim-size random variable X is distributed lognormally with $(\mu, \sigma) = (6.3210, 1.6000)$. Calculate:
(***a***) $E[X]$. (***b***) $Median[X]$.[27] (***c***) $Var[X]$.
(***d***) $\Pr\{X > 3{,}000\}$. (***e***) $\Pr\{1{,}000 < X < 3{,}000\}$.
(***f***) $E[X; 3{,}000]$. (***g***) $E[X \mid X > 3{,}000]$.

2.18 (***a***) Calculate the mean and variance of the minimum chi-square fitted distribution of Example 2.8 and compare with the sample statistics.
(***b***) Calculate the method-of-moments estimators of parameters μ and σ for fitting a lognormal distribution model to the data of Example 2.6. Compare the result with that obtained in Example 2.8.
(***c***) Which of the parameter estimates—the method-of-moments or the minimum chi-square—is likely to provide the better fit in Example 2.8?

2.19 A claim-size variable X has a shifted Pareto distribution with parameters $(\alpha, \beta) = (3; 4{,}000)$. Calculate:
(***a***) $E[X]$. (***b***) $Median[X]$. (***c***) $Var[X]$.
(***d***) $\Pr\{X > 3{,}000\}$. (***e***) $\Pr\{1{,}000 < X < 3{,}000\}$.
(***f***) $E[X; 3{,}000]$. (***g***) $E[X \mid X > 3{,}000]$.

[27] Recall that m is the ***median*** of a continuous distribution for random variable X provided that $F_X(m) = 0.50$.

2.20 Show that if $Var[X]$ exists for a shifted Pareto (α, β) random variable X, then $Var[X] > (E[X])^2$.

2.21 Derive a formula for $E[X; x]$ when X has the shifted Pareto (α, β) distribution for which $\alpha = 1$.

2.22 Claim-size variable X is defined on a population from which a random sample $\langle X_1, X_2, \ldots, X_n \rangle$ is drawn. Let $\langle x_i \rangle$ be a set of observations for such a sample. Verify the following method-of-moments estimators for the indicated distribution parameters, where M_1 and M_2 are the first two sample moments.

(***a***) Estimator of the gamma parameter β when α is known: $\hat{\beta} = M_1/\alpha$.

(***b***) Joint estimators of the lognormal (μ, σ) parameters:

$$\hat{\mu} = \log\left(M_1^2/\sqrt{M_2}\right) \quad \text{and} \quad \hat{\sigma} = \sqrt{\log\left(M_2/M_1^2\right)}.$$

(***c***) Estimator of the lognormal parameter μ when parameter σ is known: $\hat{\mu} = \log(M_1) - \frac{1}{2}\sigma^2$.

(***d***) Joint estimators of the shifted Pareto (α, β) parameters:

$$\hat{\alpha} = \frac{2\left(M_2 - M_1^2\right)}{M_2 - 2M_1^2} \quad \text{and} \quad \hat{\beta} = \frac{M_1 M_2}{M_2 - 2M_1^2}.$$

2.23 Verify the following maximum-likelihood estimators for the indicated distribution parameters.

(***a***) Estimator of the gamma parameter β when α is known: $\hat{\beta} = M_1/\alpha$.

(***b***) Joint estimators of the lognormal (μ, σ) parameters:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \log x_i \quad \text{and} \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log x_i - \hat{\mu}\right)^2}.$$

(***c***) Estimator of the lognormal parameter σ when μ is known:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log x_i - \mu\right)^2}.$$

(***d***) Estimator of the shifted Pareto parameter α when β is known:

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n}\log(x_i + \beta) - \log\beta}.$$

2.24 Obtain formulas for the median of each continuous distribution.

(***a***) exponential (β). (***b***) lognormal (μ, σ).

(***c***) shifted Pareto (α, β). (***d***) Weibull (β, δ).

2.25 For an unlimited population of 5,000 claims the mean claim size is 1,000 with a standard deviation of 2,000. Estimate the number of claims that are larger than 1,000, assuming that the size-of-loss distribution is:

(***a***) gamma. (***b***) lognormal. (***c***) shifted Pareto.

2.26 Prove: if $Sk[X]$ exists for random variable X, then $Sk[L(X)] = Sk[X]$ for all linear transformations $L(X) = aX + b$ for which $a > 0$.

2.27 Assume that X is distributed according to the classical Pareto distribution function (2.40) with parameters (α, γ). Find a linear transformation L so that $Y = L(X)$ has the shifted Pareto (α, β) distribution (2.46).

2.28 Assume that random variable U is uniformly distributed on the interval $0 < u < 1$ and that parameters (α, β, δ) are all positive. In each case determine the distribution of the transformed variable X.

(***a***) $X = -2 \log U$.
(***b***) $X = \beta(U^{-1/\alpha} - 1)$.
(***c***) $X = (\beta(U^{-1/\alpha} - 1))^{1/\delta}$.
(***d***) $X = \beta(-\log U)^{1/\delta}$.
(***e***) $X = \log(1 + Y/\beta)$, where Y has a shifted Pareto (α, β) distribution.

2.29 Random variable X has a Burr (α, β, δ) distribution with cumulative distribution function (2.64).

(***a***) Derive a formula for $E[X^m]$, $m = 1, 2, 3, \ldots$.
(***b***) Derive a formula for $E[X; x]$.

2.30 Random variable Y is defined by $Y = T(X) = e^X$, where X is gamma (α, β) distributed. Y is said to have the ***loggamma distribution***.

(***a***) Derive the cumulative distribution function for Y.
(***b***) Derive a formula for $E[Y^m]$, $m = 1, 2, 3, \ldots$.

2.31 The ***coefficient of variation CV[X]*** of a random variable X is defined as the ratio of the standard deviation to the mean: $CV[X] = SD[X]/E[X]$. Show that an application of the uniform trend transformation $T(X) = \tau X$ $(\tau > 0)$ leaves both the coefficient of variation and the skewness invariant.

2.32 Prove that the damping effect of a positive upper limit l on a uniform trend rate disappears when the limit l is subjected to the same trend factor as the claim size.

2.33 Claim-size random variable Y is obtained by applying to X the variable trend factor $\tau(x) = ax^b$. Determine the distribution of Y when the distribution of X is:

(***a***) exponential (β).
(***b***) shifted Pareto (α, β).
(***c***) Weibull (β, δ).
(***d***) Burr (α, β, δ).

2.34 For claim-size random variable X let $p = \Pr\{X \le E[X]\}$. Determine p when X has the following distributions. How does p compare to 0.50?

(***a***) exponential (β).
(***b***) gamma (α, β).
(***c***) lognormal (μ, σ).
(***d***) shifted Pareto (α, β), $\alpha > 1$.

2.35 Assume that n random variables $\{X_i\}$ are independent and identically distributed with an exponential (β) distribution. Show that the distribution of $Y = \sum_{i=1}^{n} X_i$ is gamma (n, β).

2.36 Assume that claim-size variable X has a lognormal (μ, σ) distribution. Verify that the conditional mean $E[X \mid X > a]$ is given by

$$E[X \mid X > a] = E[X] \frac{\Phi\left(\left(-\log a + \mu + \sigma^2\right)/\sigma\right)}{\Phi\left(\left(-\log a + \mu\right)/\sigma\right)} \quad (a > 0).$$

2.37 A parametric family of continuous probability distributions has a ***scale parameter*** θ whenever the probability density function $f_\theta(x)$ depending on parameter θ can be written in the form

$$f_\theta(x) = \frac{1}{\theta} f_1(x/\theta).$$

For each family of distributions identify the scale parameter, if any.

(***a***) normal (μ, σ).
(***b***) gamma (α, β).
(***c***) lognormal (μ, σ).
(***d***) shifted Pareto (α, β).
(***e***) Weibull (β, δ).
(***f***) Burr (α, β, δ).

2.38 Assume that X is a continuous random variable whose distribution has a scale parameter θ. Show that $c\theta$ is a scale parameter for the distribution of variable cX.

2.39 The grouped data displayed in the table represent the sizes of a random sample of claims drawn from an unlimited population and then censored at the value 100,000.

Size Group	# Claims
0–10,000	78
10,001–20,000	27
20,001–30,000	21
30,001–40,000	15
40,001–50,000	12
50,001–60,000	10
60,001–80,000	8
80,001–99,999	7
100,000	22
Total	200

(***a***) Obtain estimates of parameters $(\hat{\mu}, \hat{\sigma})$ for a lognormal model fit to the underlying unlimited population distribution by minimizing the distance between the sample limited severities at the eight group endpoints 10,000 through 100,000 and those implied by the lognormal model at the same points, as in Example 2.10.
(***b***) Compare the sample mean M_1 to the limited severity at 100,000 implied by the lognormal $(\hat{\mu}, \hat{\sigma})$ distribution.
(***c***) Use the chi-square test, with eight cells, to test the goodness-of-fit of the lognormal $(\hat{\mu}, \hat{\sigma})$ distribution.

2.40 To the sample data of Problem 2.39, fit a lognormal model to the underlying claim-size distribution by using the minimum chi-square method applied to the nine cells with the boundary points

$$\{0;\ 10{,}000;\ 20{,}000;\ 30{,}000;\ 40{,}000;\ 50{,}000;\ 60{,}000;\ 80{,}000;\ 100{,}000;\ \infty\}.$$

Note that the observed frequency in the ninth cell $(10{,}000;\ \infty)$ is 22.

2.41 For the unlimited claim-size random variable X and positive limit l, the random variable Y defined only on the interval $0 \le X \le l$ by

$$Y = X, \quad 0 \le X \le l$$

represents variable X ***truncated from above at*** l.

(***a***) Derive the cumulative distribution and density functions for Y.
(***b***) Obtain a formula for $E[Y]$.

2.42 For the unlimited claim-size random variable X and positive limit a, the random variable Y defined only for $X > a$ by

$$Y = X, \quad a < X < \infty$$

represents variable X ***truncated from below at a***.

(***a***) Derive the cumulative distribution and density functions for Y.

(***b***) Obtain a formula for $E[Y]$.

2.43 The table displays the result of a random sample of claims drawn from an unlimited population, truncated above at size 50,000.

(***a***) Using the minimum chi-square method with ten cells for estimating parameters, fit a lognormal model to the underlying non-truncated population distribution.

(***b***) Compare the sample mean to the severity of the fitted lognormal distribution truncated above at 50,000.

Size Group	# Claims
0–5,000	104
5,001–10,000	120
10,001–15,000	81
15,001–20,000	54
20,001–25,000	45
25,001–30,000	32
30,001–35,000	26
35,001–40,000	18
40,001–45,000	12
45,001–50,000	8
Total	500

3. Claim Counts

This chapter is devoted to probability models associated with the number of claims generated either by a single policy or by a portfolio of policies in property/casualty insurance. We study first some aspects of the basic claim process by means of a simplified example and then turn to the standard claim-count models. Our main emphasis is on the two most important parametric families, the Poisson and negative binomial distributions, with special attention paid to the modeling of both parameter uncertainty and claim contagion.

3.1. An Elementary Claim Process

The incidence of insurance claims is most usefully modeled as a random process, continuous throughout a fixed time interval. For a single policy this period is the length of time the policy remains in force—the policy term, typically one year. The basic random process must be endowed with a probability structure rich enough to support the essential random variables. The most important such time-dependent random variable—the claim count—is the principal focus of this chapter. Values of the claim-count variable, which we denote by N, are just the numbers of insured events occurring during the policy term that give rise to claims against the policy.

Begin by considering a simple discrete model of a claim process, based on the following pair of assumptions about claims arising from a single policy:

- **$\mathbf{B_1}$** During a short time interval the probability of a single claim occurring is a fixed number p $(0 \le p \le 1)$ and the probability of two or more claims occurring is zero.
- **$\mathbf{B_2}$** The numbers of claims occurring in disjoint short time intervals, each with the probability structure described in $\mathbf{B_1}$, are independent random variables.

In other words, the number of claims occurring during a single short interval is a Bernoulli random variable—it takes on the value 1 with probability p and the value 0 with probability $1 - p$.

Now let N_m denote the total number of claims occurring in m adjacent, but non-overlapping intervals for each of which assumptions $\mathbf{B_1}$ and $\mathbf{B_2}$ both hold. It is evident that N_m is the sum of m independent Bernoulli random variables, and so it has a binomial distribution with parameters (m, p) and probability function

$$\Pr\{N_m = n\} = {}_mC_n p^n (1-p)^{m-n}, \quad n = 0, 1, 2, \ldots, m. \tag{3.1}$$

The mean and variance of N_m are mp and $mp(1 - p)$, respectively.

Example 3.1. The probability that an individual policyholder makes a claim in any single day is 0.003. Assuming that at most one claim per day is possible and that claims on successive days are independent, the binomial distribution (3.1) with $p = 0.003$ applies to any time period comprised of m successive days.

For example, during a 30-day period the respective probabilities of no claims and a single claim are

$$\Pr\{N_{30} = 0\} = {}_{30}C_0 (0.003)^0 (0.997)^{30} = 0.9138,$$

$$\Pr\{N_{30} = 1\} = {}_{30}C_1 (0.003)^1 (0.997)^{29} = 0.0825.$$

As a result, the probability of two or more claims is $1 - 0.9138 - 0.0825 = 0.0037$. The expected number of claims for the 30-day period is $mp = (30)(0.003) = 0.0900$.

Probabilities of claims occurring during a full year can be computed in a similar way. For example, the probability of two claims in a 365-day year is

$$\Pr\{N_{365} = 2\} = {}_{365}C_2 (0.003)^2 (0.997)^{363} = 0.2009,$$

and the expected number of claims for the year is $(365)(0.003) = 1.0950$. ■

Example 3.1 indicates how to apply the binomial model to a policy term of reasonable length—one year, for example. In that example, this was accomplished by partitioning the policy term into disjoint short time intervals, during each of which at most one claim is possible, thereby dividing the period into discrete units with separate but identical probability structures. However, such a discrete-time approach is conceptually at variance with the intuitive view that a claim process should be a *continuous* one. It seems desirable, therefore, to find a continuous-time model for the process.

Passage from a discrete model to a continuous one always requires some type of limit procedure. To accomplish this in the case of the binomial model, first partition the policy period into m short subintervals of equal length, each with a Bernoulli probability structure specified by $\mathbf{B}_1$, for which p is the probability of a single claim. The total number of claims in all these subintervals then has the binomial probability function (3.1) with parameters (m, p).

Next, we allow the number of subintervals to become infinite in such a way that *the expected number of claims for the total policy period remains unchanged.* That is, as $m \to \infty$ parameters m and p must always satisfy $mp = \lambda$ for some positive constant λ. This implies that the probability p of a claim in each subinterval approaches zero as the number of subintervals becomes arbitrarily large—or equivalently, as the subinterval length becomes arbitrarily small. Then, for each nonnegative integer n, the probability of obtaining n claims is

$$\Pr\{n \text{ claims}\} = \lim_{\substack{m\to\infty \\ mp=\lambda}} {}_mC_n \, p^n (1-p)^{m-n}$$

$$= \lim_{m\to\infty} \frac{m(m-1)\ldots(m-n+1)}{n!} \left(\frac{\lambda}{m}\right)^n \left(1-\frac{\lambda}{m}\right)^{m-n}$$

$$= \lim_{m\to\infty} \frac{m}{m}\cdot\frac{m-1}{m}\cdots\frac{m-n+1}{m}\cdot\frac{\lambda^n}{n!}\cdot\left(1-\frac{\lambda}{m}\right)^{-n}\left(1-\frac{\lambda}{m}\right)^{m}$$

$$= \frac{\lambda^n e^{-\lambda}}{n!}. \tag{3.2}$$

The final step is a consequence of a familiar limit theorem from elementary calculus: $\lim_{m\to\infty}(1+x/m)^m = e^x$.

Formula (3.2) expressing $\lambda^n e^{-\lambda}/n!$ as the limit of binomial probabilities was first derived by Siméon-Denis Poisson (1781–1840), French mathematician and mathematical physicist extraordinaire.[28] The resulting probability distribution with probabilities given by (3.2) subsequently came to be known as a ***Poisson distribution***.

Poisson distributions have been applied to a diverse range of random events occurring throughout some time interval (or, alternatively, some type of spatial configuration). The number of alpha particles emitted from a radioactive source during a fixed time period, the number of defective products in a lot of manufactured items, the number of calls arriving at a telephone switchboard during an hour, the number of cells visible under a microscope in a certain region, the number of hurricanes striking the North American Atlantic coast in a single year—all have been successfully modeled as Poisson processes.

Because of results like (3.2) the Poisson distribution has also come to play a prominent role in modeling claim processes in property/casualty insurance. In the next section, we derive this probability distribution directly from a set of general assumptions.

3.2. Poisson Claim Processes

Experience has shown that the claim process in property/casualty insurance is often a ***Poisson process***. This means that claims occur over time in accordance with the following set of assumptions, sometimes referred to as the ***Poisson postulates***. In statements $\{\mathbf{A_1}, \mathbf{A_2}, \mathbf{A_3}, \mathbf{A_4}, \mathbf{A_5}\}$, $P_n(t)$ is the probability that n claims occur during a time interval of length t, $0 \le t < \infty$.

$\mathbf{A_1}$ The numbers of claims[29] occurring in disjoint time intervals are independent random variables.

$\mathbf{A_2}$ The probability structure is time-invariant—that is, for all $a \ge 0$ the probability of n claims occurring in the interval between a and $a + t$ equals $P_n(t)$. Thus, the distribution of the number of claims occurring during an interval depends on the length of the interval but not on the endpoints.

---

[28] Poisson's derivation appeared in his 1837 treatise on probability, *Recherchés sur la probabilité des jugements en matière criminelle et en matière civile* [*Research on the probability of criminal and civil verdicts*]. Poisson published more than 300 papers on mathematics, including the fields of analysis and probability, and on a wide range of topics in physics. His memorable adage, "Life is good for only two things, discovering mathematics and teaching mathematics," is undoubtedly best appreciated by other mathematicians.

[29] In this special formulation of the Poisson postulates the underlying random event is the occurrence of an insurance claim during a specified time interval. However, as indicated above, the general Poisson process can be applied to a variety of random events occurring in time or space.

A_3 The probability of a single claim occurring in a short interval of length h, $h > 0$, is approximately proportional to h:

$$P_1(h) = \lambda h + o(h) \text{ for some positive constant } \lambda.^{30}$$

Parameter λ is the **time density** of the incidence of claims—the average number of claims per unit of time.

A_4 The probability of more than one claim occurring in a short interval of length h is approximately zero: $\sum_{n=2}^{\infty} P_n(h) = o(h)$.

A_5 In an interval of length $t = 0$, $P_0(0) = 1$ and $P_n(0) = 0$ for $n > 0$.

Although these assumptions are often satisfied in practice, there are situations involving the incidence of insurance claims in which one or more of them fails to hold in a significant way. For example, $\mathbf{A_2}$ and $\mathbf{A_3}$ imply that the density λ of claims per unit time remains constant over time, an assumption usually valid in the short run but which might fail in the long run. The assumption of independence in $\mathbf{A_1}$ fails whenever the occurrence of a claim alters the probability of later claims. This phenomenon of claim contagion will be explored later, in Section 3.4. Finally, $\mathbf{A_4}$ is incompatible with the occurrence of multiple simultaneous claims, as is the case when two or more individuals are injured in the same accident. Such a violation of postulate $\mathbf{A_4}$ can be avoided by always defining "claim" to refer to a single insured event, without regard to the number of claimants involved.

A general formula for the Poisson probability function $P_n(t)$ can be derived directly from postulates $\{\mathbf{A_1}, \mathbf{A_2}, \mathbf{A_3}, \mathbf{A_4}, \mathbf{A_5}\}$. First, observe that $\mathbf{A_3}$ and $\mathbf{A_4}$ together imply that the probability $P_0(h)$ of zero claims in a short interval of positive length h is given by

$$P_0(h) = 1 - \sum_{n=1}^{\infty} P_n(h) = 1 - \lambda h + o(h).$$

Moreover, the independence and time-invariance properties $\mathbf{A_1}$ and $\mathbf{A_2}$ imply that the probability of zero claims in the interval $(0, t + h)$ can be expressed as

$$P_0(t + h) = P_0(t)\,P_0(h).$$

Extending this last equation to the case of n claims, where $n \geq 1$, we obtain

$$P_n(t + h) = P_n(t)\,P_0(h) + P_{n-1}(t)\,P_1(h) + \cdots + P_0(t)\,P_n(h), \tag{3.3}$$

verification of which is requested in Problem 3.5. Finally, combining the last three equations with postulate $\mathbf{A_4}$ produces

$$P_0(t + h) = P_0(t)(1 - \lambda h + o(h)),$$

$$P_n(t + h) = P_n(t)(1 - \lambda h + o(h)) + P_{n-1}(t)\,(\lambda h + o(h)) + \textstyle\sum_{i=2}^{n} P_{n-i}(t)\,o(h), \quad n \geq 1.$$

[30] The expression $o(h)$, pronounced "little oh of h," denotes a function of h that approaches 0 faster than h, so that $A = o(h)$ means $\lim_{h \to 0} A/h = 0$. If $A = o(h)$ and $B = o(h)$, then $A + B = o(h)$ also.

These equations can now be used to obtain appropriate expressions for the derivative, with respect to t, of probability function $P_n(t)$.

The case $n = 0$ yields the differential equation

$$\frac{d}{dt}P_0(t) = \lim_{h\to 0}\frac{P_0(t+h)-P_0(t)}{h} = \lim_{h\to 0}\frac{P_0(t)(-\lambda h + o(h))}{h} = -\lambda P_0(t).$$

The initial condition $P_0(0) = 1$ supplied by $\mathbf{A_5}$ gives rise to the unique solution $P_0(t) = e^{-\lambda t}$. Similarly, the derivative in the case $n \geq 1$ is

$$\begin{aligned}\frac{d}{dt}P_n(t) &= \lim_{h\to 0}\frac{P_n(t+h)-P_n(t)}{h}\\ &= \lim_{h\to 0}\frac{P_n(t)(-\lambda h + o(h)) + P_{n-1}(t)(\lambda h + o(h)) + o(h)}{h}\\ &= -\lambda P_n(t) + \lambda P_{n-1}(t).\end{aligned}$$

When $n = 1$ the solution of this differential equation is $P_1(t) = \lambda t e^{-\lambda t}$. Continuing inductively for $n = 2, 3, 4, \ldots$ yields the general Poisson probability function

$$P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \quad n = 0, 1, 2, \ldots. \tag{3.4}$$

Example 3.2. The claim process for a certain liability policy is Poisson, and claims occur at a constant rate of 0.04 per year. If the policy term is one year, then formula (3.4) with $t = 1$ and $\lambda = 0.04$ applies. The probabilities for zero, one, and two claims against the policy are, respectively,

$$P_0(1) = e^{-0.04} = 0.9608,$$

$$P_1(1) = (0.04)e^{-0.04} = 0.0384,$$

$$P_2(1) = \frac{(0.04)^2 e^{-0.04}}{2!} = 0.0008.$$

On the other hand, to obtain probabilities for claims arising during an 18-month period put $t = 1.5$ and $\lambda = 0.04$ into the same formula. Thus

$$P_0(1.5) = e^{-0.06} = 0.9418,$$

$$P_1(1.5) = (0.06)e^{-0.06} = 0.0565,$$

$$P_2(1.5) = \frac{(0.06)^2 e^{-0.06}}{2!} = 0.0017. \blacksquare$$

In property/casualty insurance applications the claim-count random variable of greatest interest is the number of claims occurring during a fixed time period, usually

a single policy term. It is appropriate then to take the basic time unit in the Poisson process to be the length of this fixed period and adjust the parameter λ to represent the claim density during the selected period. We denote the resulting random variable by N and use the customary $f(n)$ to denote the associated probability mass function, which has the simplified form

$$f(n) = \frac{\lambda^n e^{-\lambda}}{n!}, \quad n = 0, 1, 2, \ldots. \tag{3.5}$$

The moment-generating function $M(t)$ for claim-count N with distribution (3.5) exists for all real t:

$$M(t) = E\left[e^{tN}\right] = \sum_{n=0}^{\infty} e^{tn} \frac{\lambda^n e^{-\lambda}}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\left(\lambda e^t\right)^n}{n!} = \exp\left(\lambda e^t - \lambda\right). \tag{3.6}$$

This function has derivatives of all orders, and so all moments of N can be obtained from the successive derivatives of $M(t)$ evaluated at $t = 0$:

$$E[N] = M'(0) = \lambda e^t M(t)\big|_{t=0} = \lambda,$$

$$E\left[N^2\right] = M''(0) = \left(\lambda e^t + \lambda^2 e^{2t}\right) M(t)\big|_{t=0} = \lambda + \lambda^2,$$

$$E\left[N^3\right] = M'''(0) = \left(\lambda e^t + 3\lambda^2 e^{2t} + \lambda^3 e^{3t}\right) M(t)\big|_{t=0} = \lambda + 3\lambda^2 + \lambda^3.$$

It is not surprising, given the role that λ plays in the Poisson postulates, that $E[N] = \lambda$. In addition, the variance and skewness are

$$Var[N] = E\left[N^2\right] - (E[N])^2 = \lambda + \lambda^2 - \lambda^2 = \lambda, \tag{3.7}$$

$$Sk[N] = \frac{E\left[(N - E[N])^3\right]}{(Var[N])^{3/2}} = \frac{\lambda}{\lambda^{3/2}} = \frac{1}{\sqrt{\lambda}}. \tag{3.8}$$

Poisson random variables have a distinctive property that serves to characterize this distributional family—the mean and variance are equal, each identical to the distribution parameter λ.

The next example illustrates one way to fit a Poisson distribution to a set of claim data.

# Claims	# Policies
0	868
1	118
2	11
3	2
4	1
≥5	0
Total	1,000

Example 3.3. During a single policy period of one year a certain portfolio of 1,000 identical insurance policies generated 150 claims. These data have been summarized in the table by the number of claims per policy per year. We wish to find a Poisson distribution for the claim-count variable N for an individual policy selected from the portfolio.

Table 3.1. Claim-Count Distributions [Example 3.3]

	Pr{$N = n$}	
# Claims n	Sample	Poisson ($\lambda = 0.15$)
0	0.8680	0.8607
1	0.1180	0.1291
2	0.0110	0.0097
3	0.0020	0.0005
4	0.0010	0.0000
≥5	0.0000	0.0000
Mean	0.1500	0.1500
Variance	0.1735	0.1500

To do so, one can interpret these data as observations for a sample of size 1,000 drawn from a population of policies with identical Poisson claim-count distributions and unknown Poisson parameter λ. The method-of-moments estimate of parameter λ is just the sample average: $\hat{\lambda} = 150/1{,}000 = 0.15$ claims per policy per year. It is also the case that $\hat{\lambda}$ is a maximum-likelihood estimator of the parameter—see Problem 3.7.

In addition, the distribution based on $\hat{\lambda}$ can be interpreted as a parametric distribution fit to the empirical distribution of the portfolio data. Table 3.1 compares the sample distribution to that implied by the Poisson formula $f(n) = (0.15)^n e^{-0.15}/n!$.

Visual inspection of the tabulated values shows that the Poisson probabilities are close to the sample values. However, one can test the goodness of fit in a more formal way, as with the Pearson chi-square test. First compute the Pearson statistic relative to the three cells containing policies with 0, 1, and 2 or more claims, respectively—grouping in this way avoids creating cells with frequencies that are too small. Here n_k denotes the observed policy frequency in the k^{th} cell. The expected cell frequency $\phi_k(\hat{\lambda})$ predicted by the Poisson $(\hat{\lambda})$ distribution is $\phi_k(\hat{\lambda}) = 1{,}000\hat{p}_k$, where $\hat{p}_k$ is the Poisson $(\hat{\lambda})$ probability of being in the k^{th} cell. Then

$$\chi^2 = \sum_{k=1}^{3} \frac{\left(n_k - \phi_k(\hat{\lambda})\right)^2}{\phi_k(\hat{\lambda})} = \frac{(868-860.7)^2}{860.7} + \frac{(118-129.1)^2}{129.1} + \frac{(14-10.2)^2}{10.2} = 2.43.$$

The χ^2 statistic is approximately chi-square distributed, with degrees of freedom

$$d.f. = \#\ cells - \#\ estimated\ parameters - 1 = 3 - 1 - 1 = 1.$$

The 95^{th} percentile of the chi-square distribution with $d.f. = 1$ is $\chi^2_{0.95}(1) = 3.84$. Because $\chi^2 = 2.43 < \chi^2_{0.95}(1)$, we conclude at the 5% significance level that the Poisson distribution is an acceptable model for these data. ■

3.3. Parameter Uncertainty

It is sometimes the case that an insurance claim process is not strictly Poisson because the density parameter λ fails to be uniform throughout a population or is itself subject to some type of random fluctuation. For example, in a portfolio of insurance policies for which the random variable N is Poisson-distributed, the expected number of claims—the Poisson parameter λ—might vary from one insured to another. What is the distribution of N for a policy selected at random from such a population? Or consider the situation in which the distribution of the number of wind-damage claims depends on a parameter λ that varies with random changes in some key weather variables. How should one model the distribution of N in such a case?

Answers to questions like these can be obtained by means of a mixture of Poisson distributions, in which the parameter λ is itself taken to be a random variable. The resulting variable N having such a mixed distribution is called a model of ***parameter uncertainty***.

To model parameter uncertainty in the Poisson case, begin by assuming that a given population of policies has a finite number m of parameter states $\{S_i\}$, where $1 \le i \le m$. In each state S_i the claim process is Poisson with claim density α_i and $\Pr\{\textit{being in state } S_i\} = p_i$, where $p_1 + p_2 + \cdots + p_m = 1$. Now let λ denote a random variable with the set of values $\{\alpha_i\}$ and the associated discrete probability distribution, for which $E[\lambda] = \sum_{i=1}^{m} \alpha_i p_i$. In this context, variable λ is called the ***mixing parameter*** for the distribution of N.

For example, consider a portfolio of policies comprised of m disjoint subgroups, where the claim-count distribution in the i^{th} subgroup is Poisson with mean α_i. The probability p_i of obtaining a single policy from the i^{th} subgroup in a random selection from this mixed portfolio is just the fraction of the total number of portfolio policies that belong to the i^{th} subgroup. The probability function of the claim-count variable N for such a randomly selected policy is then specified for each $n = 0, 1, 2, \ldots$ by the conditional probability formula

$$f_N(n) = \sum_{i=1}^{m} \Pr\{N = n | \lambda = \alpha_i\} \cdot \Pr\{\lambda = \alpha_i\} = \frac{1}{n!} \sum_{i=1}^{m} \alpha_i^n e^{-\alpha_i} p_i. \tag{3.9}$$

The expected value of this mixed distribution turns out to be, quite reasonably, the probability-weighted average of the $\{\alpha_i\}$, that is, $E[\lambda]$:

$$E[N] = \sum_{n=0}^{\infty} n \sum_{i=1}^{m} \Pr\{n | \lambda = \alpha_i\} p_i = \sum_{i=1}^{m} p_i \sum_{n=0}^{\infty} n \Pr\{n | \lambda = \alpha_i\} = \sum_{i=1}^{m} p_i \alpha_i = E[\lambda]. \tag{3.10}$$

The second moment is obtained in a similar way:

$$E[N^2] = \sum_{n=0}^{\infty} n^2 \sum_{i=1}^{m} \Pr\{n | \lambda = \alpha_i\} p_i = \sum_{i=1}^{m} p_i \sum_{n=0}^{\infty} n^2 \Pr\{n | \lambda = \alpha_i\} = \sum_{i=1}^{m} (\alpha_i + \alpha_i^2) p_i.$$

As a result, one can express $Var[N]$ in terms of the mean and variance of λ:

$$Var[N] = E[\lambda] + \sum_{i=1}^{m} \alpha_i^2 p_i - (E[\lambda])^2 = E[\lambda] + Var[\lambda]. \tag{3.11}$$

It is evident from (3.10) and (3.11) that when N has a mixed Poisson distribution, $Var[N] = E[N]$ if, and only if, $Var[\lambda] = 0$ (in which case the variable λ is constant). Therefore, a mixture of distinct Poisson distributions—for which $Var[\lambda] > 0$—cannot itself be a Poisson distribution.

Example 3.4. A portfolio of 100 insurance policies for which the claim counts are Poisson-distributed produces an overall average of 0.51 claims per policy per year. However, this portfolio consists of four policy subgroups, representing four parameter states, with expected claim counts ranging from 0.10 to 1.40, as shown in the table. Also tabulated are the numbers of policies in each subgroup. Consequently, the distribution of N for a policy selected at random from this portfolio has a mixed Poisson distribution with probabilities

State	Density α_i	# Policies
S_1	0.10	20
S_2	0.35	40
S_3	0.70	30
S_4	1.40	10
Total	0.51	100

$$f_N(n) = \frac{(0.10)^n e^{-0.10}(0.2) + (0.35)^n e^{-0.35}(0.4) + (0.70)^n e^{-0.70}(0.3) + (1.40)^n e^{-1.40}(0.1)}{n!}$$

As expected, $E[N] = 0.51$, and the variance exceeds the mean: $Var[N] = 0.6439 > E[N]$. Probabilities for N are shown in Table 3.2, where they are compared with those of the single Poisson distribution for which $\lambda = 0.51$.

Table 3.2. Claim-Count Distributions [Example 3.4]

	Pr{$N = n$}	
# Claims n	Mixed Poisson	Poisson ($\lambda = 0.51$)
0	0.6365	0.6005
1	0.2556	0.3063
2	0.0788	0.0781
3	0.0218	0.0133
4	0.0056	0.0017
5	0.0013	0.0002
6	0.0003	0.0000
≥7	0.0001	0.0000
Mean	0.5100	0.5100
Variance	0.6439	0.5100

This example effectively illustrates how mixing several distinct Poisson distributions serves to increase the dispersion of the claim-count distribution over what would be expected in the case of a single Poisson distribution. ■

The discrete conditional probability formula (3.9) is readily generalized to the case in which the variable λ has a *continuous* density function $f_\lambda(u)$ on the interval $0 < u < \infty$ for which $\int_0^\infty f_\lambda(u)du = 1$. Thus, the continuous analog of (3.9) for claim counts $n = 0, 1, 2, \ldots$ is given by the integral formula

$$f_N(n) = \int_0^\infty \Pr\{N = n|\lambda = u\} f_\lambda(u)du = \frac{1}{n!}\int_0^\infty u^n e^{-u} f_\lambda(u)du.^{31} \tag{3.12}$$

Formulas (3.10) and (3.11) for the mean and variance of N also hold in the continuous case. For example, in the following continuous analog of (3.10) integration over the semi-infinite interval $0 < u < \infty$ replaces summation over the finite set of parameter values $\{\alpha_i\}$:

$$E[N] = \sum_{n=0}^{\infty} n\int_0^\infty \frac{u^n e^{-u}}{n!} f_\lambda(u)du = \int_0^\infty \left(\sum_{n=0}^{\infty} n\frac{u^n e^{-u}}{n!}\right) f_\lambda(u)du = \int_0^\infty u\, f_\lambda(u)du = E[\lambda]. \tag{3.13}$$

As before, the second moment is $E[N^2] = E[\lambda] + E[\lambda^2]$, so that

$$Var[N] = E[\lambda] + Var[\lambda] = E[N] + Var[\lambda]. \tag{3.14}$$

Moreover,

$$Sk[N] = \frac{E[\lambda] + 3Var[\lambda] + E\left[(\lambda - E[\lambda])^3\right]}{(E[\lambda] + Var[\lambda])^{3/2}}. \tag{3.15}$$

Example 3.5. Parameter λ for a mixed Poisson distribution has an exponential density function: $f_\lambda(u) = 4e^{-4u}$, $0 < u < \infty$. Consequently, $E[\lambda] = 0.25$ and $Var[\lambda] = 0.0625$. The claim-count probabilities for the mixed distribution follow from (3.12):

$$f_N(n) = \frac{4}{n!}\int_0^\infty u^n e^{-5u}du = \frac{4}{n!}\Gamma(n+1)5^{-(n+1)} = (0.8)(0.2)^n, \quad n = 0, 1, 2, \ldots.$$

This distribution is an instance of the ***geometric distribution***—see Problem 3.21. Table 3.3 displays probabilities for the distribution, again compared to those of the related, but less-dispersed single Poisson distribution. ■

---

[31] By using the Riemann–Stieltjes integral, formulas (3.9) and (3.12) can both be expressed by the single integral formula $f_N(n) = (1/n!)\int_0^\infty u^n e^{-u} dF_{\lambda(u)}$, where the function $F_\lambda(u)$ is the cumulative distribution function for the variable mixing parameter λ.

Table 3.3. Claim-Count Distributions [Example 3.5]

	Pr{$N = n$}	
# Claims n	Mixed Poisson	Poisson ($\lambda = 0.25$)
0	0.8000	0.7788
1	0.1600	0.1947
2	0.0320	0.0243
3	0.0064	0.0020
4	0.0013	0.0001
5	0.0003	0.0000
≥6	0.0001	0.0000
Mean	0.2500	0.2500
Variance	0.3125	0.2500

The mixed Poisson distribution provides a powerful alternative to the single Poisson distribution in modeling insurance claim data. However, the mixed distribution approach does require that one must have some a priori knowledge of the distribution of the variable parameter λ. In cases where claim data can reasonably be partitioned into a finite number of homogenous subgroups, as in Example 3.4, there is usually no difficulty in constructing the mixed distribution. On the other hand, if one is presented with a sample of claim data for which the mean and variance are quite different, then finding an appropriate parametric distribution in the absence of additional information about the population is more challenging. A common solution to this problem, which involves a special family of distributions for the variable λ, is the topic of the next section.

3.4. Negative Binomial Distributions

It is often the case that claim-count data yield a sample distribution with markedly different mean and variance and that very little is known about the actual population distribution. Nevertheless, the actuary can be faced with the problem of fitting a parametric distribution to such data in order to model the claim-count probabilities of a policy selected at random from the underlying population.

In situations like this it has proven useful to suppose that the population has a mixed Poisson distribution and, in the absence of any other information, to *assume* a particular distribution for the variable parameter λ. Gamma distributions are almost always used for this purpose because of the useful analytic form of the resulting mixed distribution.

We start by assuming that the mixing parameter λ has a gamma distribution with positive parameters α and $\beta = \nu/\alpha$ and the resulting density function

$$f_\lambda(u) = \frac{(\alpha/\nu)^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-(\alpha/\nu)u}, \quad 0 < u < \infty. \tag{3.16}$$

This special choice of α and β is contrived to yield convenient forms for the means and variances of λ and N. In particular, $E[\lambda] = \nu$ and $Var[\lambda] = \nu^2/\alpha$. The probability function for N is then obtained by substituting (3.16) into formula (3.12) for each $n = 0, 1, 2, \ldots$:

$$\begin{aligned} f_N(n) &= \int_0^\infty \frac{u^n e^{-u}}{n!} \cdot \frac{(\alpha/\nu)^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-(\alpha/\nu)u} du \\ &= \frac{(\alpha/\nu)^\alpha}{n!\Gamma(\alpha)} \int_0^\infty u^{\alpha+n-1} e^{-\left(\frac{\nu+\alpha}{\nu}\right)u} du \\ &= \frac{(\alpha/\nu)^\alpha}{n!\Gamma(\alpha)} \Gamma(\alpha+n) \left(\frac{\nu+\alpha}{\nu}\right)^{-(\alpha+n)} \\ &= \frac{\Gamma(\alpha+n)}{n!\Gamma(\alpha)} \left(\frac{\alpha}{\nu+\alpha}\right)^\alpha \left(\frac{\nu}{\nu+\alpha}\right)^n. \end{aligned} \tag{3.17}$$

The mean, variance, and skewness of N are thus given by formulas (3.13), (3.14), and (3.15):

$$\begin{aligned} E[N] &= \nu, \\ Var[N] &= \nu + \frac{\nu^2}{\alpha}, \\ Sk[N] &= \frac{\alpha^3\nu + 3\alpha^2\nu^2 + 2\alpha\nu^3}{(\alpha^2\nu + \alpha\nu^2)^{3/2}}. \end{aligned} \tag{3.18}$$

The mixed probability distribution defined by (3.17) is a member of the ***negative binomial*** distribution family. Such a distribution has a probability function, with parameters r and q, of the general form

$$f(n) = \binom{r+n-1}{n} q^r (1-q)^n \quad (r > 0, 0 < q < 1), \quad n = 0, 1, 2, \ldots. \tag{3.19}$$

The leading coefficient in this function is an instance of the ***general binomial coefficient***, defined for all real x and $n = 0, 1, 2, 3, \ldots$ by

$$\binom{x}{n} = \frac{\Gamma(x+1)}{n!\Gamma(x-n+1)} = \begin{cases} 1 & \text{if } n = 0 \\ \dfrac{x(x-1)(x-2)\cdots(x-n+1)}{n!} & \text{if } n > 0. \end{cases} \tag{3.20}$$

Putting $r = \alpha$ and $q = \alpha/(\nu + \alpha)$ into (3.19) and then applying (3.20) shows that probability function (3.17) does indeed belong to the negative binomial family.

The general binomial coefficient first arose in connection with Isaac Newton's ***binomial series***, a convergent series expansion valid for all real s:

$$(1+x)^s = \sum_{n=0}^{\infty} \binom{s}{n} x^n, \quad -1 < x < 1. \tag{3.21}$$

When the exponent s is a positive integer m, $\binom{s}{n} = \binom{m}{n}$ is identical to the standard binomial coefficient ${}_mC_n$ of combinatorial analysis:

$$\binom{m}{n} = {}_mC_n = \begin{cases} 0 & \text{if } n > m \\ \dfrac{m!}{n!(m-n)!} & \text{if } 0 \le n \le m. \end{cases}$$

Moreover, whenever $s = m$ series (3.21) reduces to the familiar finite sum of the elementary Binomial Theorem.

The negative binomial distribution is so called because the probabilities in (3.19) are fixed multiples of terms from a convergent binomial series with negative exponent. This becomes evident after using the identity

$$\binom{r+n-1}{n} = (-1)^n \binom{-r}{n}, \quad r > 0 \tag{3.22}$$

to restate probability function (3.19) as

$$f(n) = \binom{-r}{n} q^r (q-1)^n. \tag{3.23}$$

Verification of $\Sigma_{n=0}^{\infty} f(n) = 1$ follows immediately from (3.21) with $x = q - 1$ and negative exponent $s = -r$:

$$\sum_{n=0}^{\infty} f(n) = q^r \sum_{n=0}^{\infty} \binom{-r}{n} (q-1)^n = q^r (1+q-1)^{-r} = 1.$$

The negative binomial moment-generating function is obtained in a similar way:

$$M(t) = \sum_{n=0}^{\infty} e^{tn} \binom{-r}{n} q^r (q-1)^n = q^r \left(1 + (q-1)e^t\right)^{-r}, \quad -\infty < t < -\log(1-q). \tag{3.24}$$

Returning to the mixed distribution (3.17), one can see that for a given mean ν the size of $Var[N]$ relative to $E[N]$ is determined by parameter α. The larger the value of α, the more nearly equal are $Var[N]$ and $E[N]$ and conversely. One can therefore interpret $1/\alpha$ as a measure of parameter uncertainty—of how significantly the mixed distribution deviates from a single Poisson.

Not surprisingly, the Poisson distribution is a limiting case—that is, for fixed mean ν the negative binomial distribution (3.17) tends to a Poisson distribution as $\alpha \to \infty$. This can be easily demonstrated using the moment-generating function. The generating function of distribution (3.17) is

$$M(t) = \left(1 - \frac{\nu}{\alpha}\left(e^t - 1\right)\right)^{-\alpha}, \quad -\infty < t < \log((\nu + \alpha)/\nu),$$

obtained by substituting $r = \alpha$ and $q = \alpha/(\nu + \alpha)$ into (3.24). Passing to the limit as $\alpha \to \infty$ while holding ν constant yields

$$\lim_{\alpha\to\infty} M(t) = \lim_{\alpha\to\infty}\left(1 - \frac{\nu}{\alpha}\left(e^t - 1\right)\right)^{-\alpha} = \exp\left(\nu\left(e^t - 1\right)\right), \tag{3.25}$$

the moment-generating function of a Poisson random variable with mean ν. The conclusion follows from the uniqueness property of the generating function.

One of the earliest uses of the negative binomial as a mixed Poisson distribution was in modeling the concept of *accident proneness.* The number of accidents incurred by individual members of a population group was assumed to be Poisson-distributed, but with different parameters—the more "accident-prone" members having larger Poisson parameters and those less so having smaller expected values. In the realm of property/casualty insurance, actuaries began to apply the negative binomial distribution to automobile liability in the 1950s and 1960s.[32] Since then, the distribution has enjoyed a wide range of applicability.

Example 3.6. Claim-count data from a sample of 5,000 automobile liability policies are displayed in the table. Here the mean 0.1238 and variance 0.130074 are unequal. This inequality suggests that the policies were possibly drawn not from a homogeneous population of Poisson-distributed policies but from a mix of policies with different Poisson distributions.

# Claims	# Policies
0	4,429
1	528
2	39
3	3
4	1
≥5	0
Total	5,000

To obtain the distribution of claim counts for a single policy selected at random from this population, we shall assume a negative binomial distribution of the form (3.17) and search for appropriate parameter estimates $(\hat{\alpha}, \hat{\nu})$. Derivation of maximum-likelihood parameter estimates for the negative binomial distribution involves some difficult, but not insurmountable, complexities,[33] whereas the method-of-moments estimates are easily calculated. Opting

[32] For an example of such early applications of the negative binomial distribution in auto insurance see Hewitt [6].

[33] For example, refer to Simon [21]. In most cases the maximum-likelihood equations are solvable only by iteration, but Simon observes that the method "will usually produce answers very similar to the method-of-moments" estimates.

Table 3.4. Negative Binomial Distribution [Example 3.6]

	Pr{*n* claims}	
# Claims *n*	Sample	Negative binomial
0	0.8858	0.8862
1	0.1056	0.1044
2	0.0078	0.0087
3	0.0006	0.0006
4	0.0002	0.0000
≥5	0.0000	0.0000

for the latter approach, we set $\hat{\nu} = 0.1238$ and then solve the variance formula $Var = \hat{\nu} + \hat{\nu}^2/\hat{\alpha} = 0.130074$ for $\hat{\alpha}$:

$$\hat{\alpha} = \frac{(0.1238)^2}{0.130074 - 0.1238} = 2.44285.$$

Probabilities calculated from the resulting mass function

$$f(n) = \frac{\Gamma(2.44285 + n)}{n!\Gamma(2.44285)}(0.951766)(0.048234)^n, \quad n = 0, 1, 2, \ldots,$$

are displayed in Table 3.4. The chi-square statistic based on the four cells corresponding to 0, 1, 2, and 3 or more claims is $\chi^2 = 0.7753$. The degrees-of-freedom parameter is *d.f.* = 4 − 2 − 1 = 1, so the rejection limit for a test at the 5% significance level is $\chi^2_{0.95}(1) = 3.84$. Because $\chi^2 < 3.84$, one can conclude that the negative binomial distribution is an acceptable model.

Recall that a negative binomial distribution can be interpreted as a mixture of Poisson distributions with a variable gamma-distributed mixing parameter λ. In this case an *implied* gamma density function for the variable mixing parameter λ can be obtained by putting $\hat{\alpha}$ and $\hat{\nu}$ into formula (3.16):

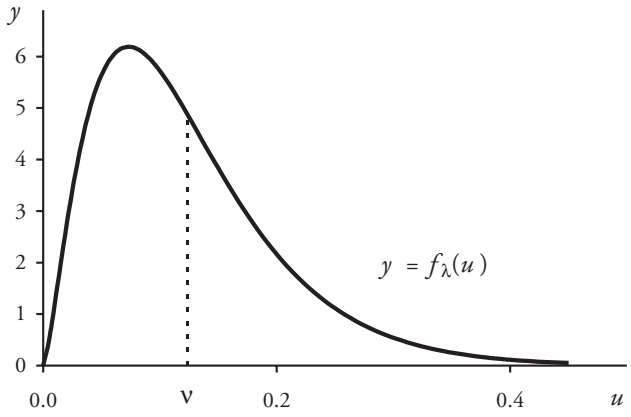
$$f_\lambda(u) = (1141.264)\, u^{1.44285} e^{-19.7322u}, \quad 0 < u < \infty.$$

Here $E[\lambda] = 0.1238 = \nu$ and $Var[\lambda] = 0.006274$. A graph of this density function $y = f_\lambda(u)$ is shown in Figure 3.1. ■

3.5. Claim Contagion

One of the assumptions of the Poisson claim-count process, that of independence of successive claims, is not always satisfied. This happens whenever the occurrence of a claim changes the probability of subsequent claims. For example, a successful products

Figure 3.1. Implied Gamma Density Function for Mixing Parameter λ [Example 3.6]



liability claim against a manufacturer often increases the likelihood that similar claims will be brought in the future—a classic example of ***claim contagion***. A standard approach to modeling such a contagion process is based on an urn model proposed by the Hungarian mathematician George Pólya (1887–1985). Pólya models have since been used to model a variety of contamination processes, including the spread of contagious diseases.[34]

In the Pólya model, an urn initially contains w white balls and b black balls. A trial consists of drawing one ball at random, noting its color, and then replacing it together with c additional balls of the same color. Obtaining a white ball on the first trial therefore increases the probability of selecting a white ball on the next trial. The probability function for the number W_m of white balls obtained in m trials, is derived by conventional combinatorial methods:

$$\Pr\{W_m = n\} = P_n(w, b, c; m) = {}_mC_n \frac{\prod_{i=0}^{n-1}(w + ic) \prod_{i=0}^{m-n-1}(b + ic)}{\prod_{i=0}^{m-1}(w + b + ic)}, \quad n = 0, 1, \ldots, m. \quad (3.26)$$

A distribution with probabilities $P_n(w, b, c; m)$ is known as a ***Pólya distribution***.[35] The ratio $\gamma = c/w$ is customarily called the ***degree of contagion***. When there is no contagion—that is, when $c = \gamma = 0$—the Pólya distribution is identical to the simpler binomial distribution for which the probability of drawing a white ball remains constant throughout successive trials.

---

[34] Pólya's original contamination model first appeared in a 1923 paper by Pólya and F. Eggenberger "Über die Statistik der vergetteter Vorgänge," *Zeitschrift für Angewandte Mathematik und Mechanik,* III, 279–289]. The present formulation is based on that presented by William Feller in his classic probability textbook: Feller [4], pp. 118–121, 142–143. Urn models have been used to model probability distributions ever since they were introduced by Swiss mathematician Jacob Bernoulli to describe the two-outcome experiment underlying the random variable now known as a Bernoulli variable.

[35] Although history and logic dictate that (3.26) should be called the Pólya distribution, some authors apply that name instead to the associated negative binomial distribution (3.27).

To derive the moments of the distribution for W_m, let p denote the probability of drawing a white ball in the first Pólya trial: $p = w/(w+b)$. Then

$$E[W_m] = \sum_{n=0}^{m} nP_n(w,b,c;m) = \frac{mw}{w+b}\sum_{n=1}^{m} P_{n-1}(w+c,b,c;m-1)$$

$$= mp\sum_{k=0}^{m-1} P_k(w+c,b,c;m-1)$$

$$= mp.$$

Similar reasoning yields

$$E[W_m^2] = mp + m(m-1)\frac{w+c}{w+b+c}p = mp + m(m-1)p^2\frac{1+\gamma}{1+p\gamma},$$

so that

$$Var[W_m] = mp\left(1+(m-1)\frac{1+\gamma}{1+p\gamma}p - mp\right).$$

It is instructive to compare these formulas to the respective mean mp and variance $mp(1-p)$ of the related binomial distribution. Clearly, the two distributions have identical means, each equal to mp. The Pólya distribution, as one would reasonably expect, has the larger variance: $Var[W_m] \geq mp(1-p)$.

Example 3.7. Pólya trials are conducted with an urn that initially contains $w = 10$ white balls and $b = 5$ black balls. Corresponding to $c = 2$, the degree of contagion is $\gamma = 2/10$. The initial probability is therefore $p = E[W_1] = 10/15$. Various probabilities for drawing white balls in the first three trials are given by

$$\Pr\{\textit{white on 1st trial}\} = \frac{10}{15} = 0.6667,$$

$$\Pr\{\textit{white on 2nd trial} \mid \textit{white on 1st trial}\} = \frac{12}{17} = 0.7059,$$

$$\Pr\{\textit{white on 2nd trial} \mid \textit{black on 1st trial}\} = \frac{10}{17} = 0.5882,$$

$$\Pr\{\textit{white on 2nd trial}\} = \frac{12}{17}\cdot\frac{10}{15} + \frac{10}{17}\cdot\frac{5}{15} = 0.6667,$$

$$\Pr\{\textit{white on 3rd trial} \mid \textit{white on 1st \& 2nd trials}\} = \frac{14}{19} = 0.7368,$$

$$\Pr\{\textit{white on 3rd trial} \mid \textit{black on 1st \& 2nd trials}\} = \frac{10}{19} = 0.5263.$$

Table 3.5. Pólya and Binomial Probabilities [Example 3.7]

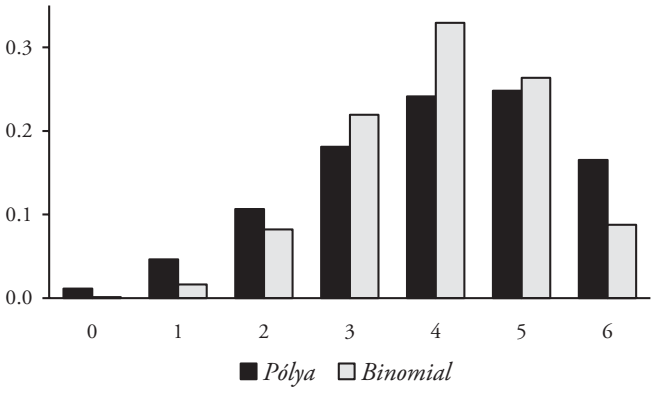
	Pr{n white balls in 6 trials}	
# White Balls n	Pólya ($c = 2$)	Binomial ($c = 0$)
0	0.0115	0.0014
1	0.0462	0.0165
2	0.1066	0.0823
3	0.1809	0.2195
4	0.2412	0.3292
5	0.2481	0.2634
6	0.1654	0.0878
Mean	4.0000	4.0000
Variance	2.1176	1.3333

The probabilities $P_n(10,5,2;6)$ for obtaining n white balls in six successive Pólya trials are shown in Table 3.5 and compared with those for the related binomial distribution, for which $c = 0$. The expected count for each distribution is $mp = 4$, but the Pólya distribution with positive contagion has the larger variance, a fact clearly evident in Figure 3.2. ■

To interpret the Pólya urn model as a claim process, identify the draw of a white ball in a Pólya trial with the occurrence of a claim. However, contagion in the urn model occurs at discrete times, after each draw from the urn. Modeling a time-continuous claim process, this time with contagion, again requires some type of limit process.

We proceed as in Section 3.1, where a Poisson distribution arose as the limit of a sequence of binomial distributions. Again, partition the basic time period—the policy

Figure 3.2. Pólya and Binomial Distributions [Example 3.7]



term—into m subintervals of equal length and perform one Pólya trial per subinterval. Then let $m \to \infty$ in such a way that the expected number of white balls (that is, claims) in the time period remains constant: $mp = \nu > 0$. Passage to the limit is carried out in such a way that the degree of contagion γ remains constant, as well. As before, the probability structure in each subinterval changes with m so that $p \to 0$ as $m \to \infty$. Whenever $\gamma > 0$ the limit of the Pólya probability function (3.26) is

$$\lim_{\substack{m\to\infty \\ mp=\nu}} P_n(w, b, c; m) = \binom{1/\gamma + n - 1}{n} \left(\frac{1}{1+\gamma\nu}\right)^{1/\gamma} \left(\frac{\gamma\nu}{1+\gamma\nu}\right)^n, \quad n = 0, 1, 2, \ldots. \tag{3.27}$$

Comparison to formula (3.19), after setting $r = 1/\gamma$ and $q = 1/(1 + \gamma\nu)$, reveals the limiting distribution to be negative binomial with mean ν and variance $\nu + \gamma\nu^2$.

Thus we have observed that two distinct situations—claim contagion in this section and parameter uncertainty with a gamma-distributed λ discussed in Section 3.4—give rise to the same distribution family for the claim-count variable N. In fact, the negative binomial distribution in (3.27), with mean ν and contagion parameter γ, is identical to the negative binomial distribution (3.17) with mean ν and uncertainty parameter $\alpha = 1/\gamma$. It is not surprising, then, that the negative binomial distribution remains the principal alternative to the Poisson for modeling the distribution of property/casualty claim counts.

We conclude this section with an outline of the proof of limit formula (3.27) in which the negative binomial is obtained as the limiting case of the Pólya distribution for a contagion model.

***Proof of* (3.27):** Start by expressing the Pólya probability function in terms of the general binomial coefficients, where $\gamma = c/w$ and $p = w/(w + b)$:

$$P_n(w, b, c; m) = \binom{\frac{1}{\gamma} + n - 1}{n} \binom{\frac{1-p}{\gamma p} + m - n - 1}{m - n} \Bigg/ \binom{\frac{1}{\gamma p} + m - 1}{m}.$$

An application of the identity in Problem 3.27(b) yields

$$P_n(w, b, c; m) = \binom{\frac{1}{\gamma} + n - 1}{n} \cdot \underbrace{\frac{\binom{\frac{1-p}{\gamma p} + m - n - 1}{m - n}}{\binom{\frac{1}{\gamma p} + m - n - 1}{m - n}}}_{A} \cdot \underbrace{\frac{\prod_{i=1}^{n}\left(1 - \frac{i-1}{m}\right)}{\prod_{i=1}^{n}\left(\frac{1}{\gamma p m} + \frac{m-i}{m}\right)}}_{B}. \tag{3.28}$$

To complete the proof, we evaluate the limits of quotients A and B in turn.

First, apply to quotient A in equation (3.28) formula (3.20) defining the general binomial coefficient as a ratio of gamma function expressions. This yields

$$A = \frac{\binom{\frac{1-p}{\gamma p} + m - n - 1}{m-n}}{\binom{\frac{1}{\gamma p} + m - n - 1}{m-n}} = \frac{\Gamma\left(\frac{1-p}{\gamma p} + m - n\right)}{\Gamma\left(\frac{1}{\gamma p} + m - n\right)} \cdot \frac{\Gamma\left(\frac{1}{\gamma p}\right)}{\Gamma\left(\frac{1-p}{\gamma p}\right)}.$$

The limit of A is based on the asymptotic relation $\Gamma(x) \sim \sqrt{2\pi}e^{-x}x^{x-1/2}$.[36] Applying this to the gamma functions in the last equation, substituting $Q = \gamma pm - \gamma pn$, and observing that factors involving $\sqrt{2\pi}e^{-x}$ cancel, one obtains

$$A \sim \frac{\left(\frac{1-p+\gamma pm-\gamma pn}{\gamma p}\right)^{\frac{1+\gamma pm-\gamma pn}{\gamma p}-\frac{1}{\gamma}-\frac{1}{2}}}{\left(\frac{1+\gamma pm-\gamma pn}{\gamma p}\right)^{\frac{1+\gamma pm-\gamma pn}{\gamma p}-\frac{1}{2}}} \cdot \frac{\left(\frac{1}{\gamma p}\right)^{\frac{1}{\gamma p}-\frac{1}{2}}}{\left(\frac{1-p}{\gamma p}\right)^{\frac{1}{\gamma p}-\frac{1}{\gamma}-\frac{1}{2}}}$$

$$= \left(\frac{1-p+Q}{1+Q}\right)^{\frac{1+Q}{\gamma p}} (1-p)^{-1/(\gamma p)} \left(\frac{1-p}{1-p+Q}\right)^{1/\gamma} \sqrt{\frac{(1-p)(1+Q)}{1-p+Q}}.$$

An application of the limit formulas

$$\lim_{\substack{m\to\infty \\ mp=\nu}} Q = \gamma\nu \quad \text{and} \quad \lim_{\delta\to 0}(1+\delta x)^{1/\delta} = e^x$$

yields the desired limit:

$$\lim_{\substack{m\to\infty \\ mp=\nu}} A = e^{-1/\gamma} \cdot e^{1/\gamma} \cdot \left(\frac{1}{\gamma\nu+1}\right)^{1/\gamma} \cdot 1 = \left(\frac{1}{\gamma\nu+1}\right)^{1/\gamma}.$$

For quotient B in equation (3.28) we have

$$\lim_{\substack{m\to\infty \\ mp=\nu}} B = \lim_{\substack{m\to\infty \\ mp=\nu}} \frac{\left(1-\frac{0}{m}\right)\left(1-\frac{1}{m}\right)\cdots\left(1-\frac{n-1}{m}\right)}{\left(\frac{1}{\gamma pm}+\frac{m-1}{m}\right)\left(\frac{1}{\gamma pm}+\frac{m-2}{m}\right)\cdots\left(\frac{1}{\gamma pm}+\frac{m-n}{m}\right)}$$

[36] This relation is a generalization of Stirling's approximation formula: $n! \sim \sqrt{2\pi n}(n/e)^n$, n a positive integer; see Feller [4], pp. 52–54, 66. $f(x) \sim g(x)$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$. Therefore, $f(x) \sim g(x)$ and $\lim_{x\to\infty} g(x) = L$ together imply that $\lim_{x\to\infty} f(x) = L$.

$$= \frac{1}{\left(\frac{1}{\gamma\nu}+1\right)^{n}}$$

$$= \left(\frac{\gamma\nu}{1+\gamma\nu}\right)^{n}.$$

This completes the proof. ■

3.6. Portfolio Claims

Up to now our focus has been on modeling the claim process for a single policy. However, it is also important to find probability models that describe the aggregate behavior of entire portfolios of similar policies. In a variety of situations it is possible to infer the distribution of the portfolio claim count from those of the individual component policies.

For example, if the claim process for each policy in a portfolio of policies is Poisson, what can be said about the distribution of N, the total number of *portfolio* claims that occur during a policy period? The answer lies in the reproductive property of Poisson variables—that is, the sum of mutually independent Poisson random variables is also Poisson-distributed. This fact follows from an argument based on the moment-generating function.

Let $N = N_1 + N_2 + \cdots + N_m$ be the sum of m independent Poisson random variables. If $E[N_i] = \lambda_i$, where $1 \le i \le m$, then

$$M_N(t) = E\left[\exp\left(t\sum_{i=1}^{m} N_i\right)\right] = \prod_{i=1}^{m} E\left[e^{tN_i}\right] = \prod_{i=1}^{m} \exp\left(\lambda_i e^t - \lambda_i\right)$$

$$= \exp\left(\left(\sum_{i=1}^{m}\lambda_i\right)\left(e^t - 1\right)\right),$$

the moment-generating function for a Poisson variable with parameter $\sum_{i=1}^{m}\lambda_i$. The uniqueness property of the generating function implies that N must be Poisson-distributed with mean $\sum_{i=1}^{m}\lambda_i$.

In the special case that each policy in a portfolio of m policies has the same Poisson distribution with expected value λ, it is evident that the portfolio claim-count variable N has a Poisson distribution with parameter $m\lambda$.

One can similarly show by means of the moment-generating function that the sum $N = N_1 + N_2 + \cdots + N_m$ of mutually independent negative binomial variables, identically distributed as in (3.17) with parameters (α, ν), has a negative binomial distribution with parameters $(m\alpha, m\nu)$:

$$M_N(t) = \prod_{i=1}^{m} E\left[e^{tN_i}\right] = \left(\left(1 - \frac{\nu}{\alpha}\left(e^t - 1\right)\right)^{-\alpha}\right)^{m}$$

$$= \left(1 - \frac{1}{m\alpha} m\nu\left(e^t - 1\right)\right)^{-m\alpha}.$$

On the other hand, the sum N does not necessarily have a negative binomial distribution when the $\{N_i\}$ have different α and ν parameters. Consequently, the sum of independent claim-count random variables, each with a contagion structure as described in Section 3.5, is not always itself a negative binomial contagion model. Nevertheless, it is often desirable to be able to treat such a distribution as if it *were* such a model. One can do this by defining the contagion parameter γ for an arbitrary claim-count variable N in a way that is consistent with the negative binomial case, namely,

$$\gamma = \frac{Var[N] - E[N]}{(E[N])^2}, \tag{3.29}$$

obtained by rearranging the negative binomial formula

$$Var[N] = E[N] + \gamma(E[N])^2.$$

Formula (3.29) implies that the contagion parameter for a Poisson random variable is $\gamma = 0$, as one would reasonably expect.

Example 3.8. Consider a group of 100 identical policies, each with a Poisson claim process and an expected annual claim count of 0.035 per policy. What is the probability that these policies in aggregate generate five or more claims during a single year?

The portfolio claim-count variable N is the sum of identically distributed Poisson variables. Under the reasonable assumption that the claim processes associated with these policies are independent, the reproductive property of the Poisson process implies that N also has a Poisson distribution, with parameter $\lambda = (100)(0.035) = 3.50$. Therefore,

$$\Pr\{N \geq 5\} = 1 - e^{-3.50}\left(1 + 3.50 + \frac{1}{2}(3.50)^2 + \frac{1}{6}(3.50)^3 + \frac{1}{24}(3.50)^4\right) = 0.2746. \blacksquare$$

Example 3.9. A portfolio contains 20 independent, identically-distributed policies subject to claim contagion. Each policy has an expected claim count of $\nu = 0.150$ per year and contagion parameter $\gamma = 0.400$. Thus the distribution of the portfolio claim count N is negative binomial, with

$$E[N] = (20)(0.150) = 3.000,$$

$$Var[N] = (20)(0.150) + (0.400)(20)(0.150)^2 = 3.180.$$

Formula (3.29) implies a portfolio contagion parameter of

$$\gamma = \frac{3.180 - 3.000}{(3.000)^2} = 0.020. \blacksquare$$

3.7. Problems

3.1 Random variable N has a binomial (m, p) distribution.
(***a***) Use $M_N(t)$ to derive the mean, variance, and skewness for N.
(***b***) Evaluate $\lim_{\substack{m\to\infty \\ mp=\lambda}} M_N(t)$, where $\lambda > 0$. What conclusion can be drawn?

3.2 Verify that the Poisson probability function (3.4) satisfies $\sum_{n=0}^{\infty} P_n(t) = 1$.

3.3 Assume that claim-count variable N has probability function $f(n) = \lambda^n e^{-\lambda}/n!$ What are the values of λ and $\Pr\{N \le 3\}$ in each case?
(***a***) $E[N] = 3.20$. (***b***) $Var[N] = 2.50$.
(***c***) $Sk[N] = 0.40$. (***d***) $f(1) = f(2)$.
(***e***) $E[e^{tN}] = e^{4e^t}/e^4$. (***f***) $f(0) = 0.80$.

3.4 The policy claim count for a liability line of insurance is Poisson-distributed with constant density of 0.10 claims per policy per year. Compute the probability that a single policy has exactly two claims when the policy term is:
(***a***) 6 months. (***b***) 15 months. (***c***) 24 months.

3.5 Use mathematical induction to verify equation (3.3) for the decomposition of $P_n(t + h)$ in the derivation of the Poisson probability function.

3.6 Prove that the Poisson probability function $f(n) = \lambda^n e^{-\lambda}/n!$ has a maximum value at $n = [\![\lambda]\!]$, where $[\![x]\!]$ denotes the greatest integer function. [*Hint*: show that f satisfies the recursion relation $f(n) = (\lambda/n) f(n-1)$ for $n = 1, 2, 3, \ldots$.]

3.7 Let $\langle n_i \rangle$ be a set of observations for a random sample of claim counts $\langle N_1, N_2, \ldots, N_m \rangle$ drawn from a Poisson-distributed population with unknown parameter λ. Prove that the sample mean $M_1 = \frac{1}{m}\sum_{i=1}^{m} n_i$ is a maximum-likelihood estimator for λ.

3.8 Show that the cumulative distribution function $F(n)$ for a Poisson (λ) random variable can be expressed at each nonnegative integer n by

$$F(n) = \sum_{k=0}^{n} \frac{\lambda^k e^{-\lambda}}{k!} = 1 - \frac{\Gamma(\lambda, n+1)}{n!}.$$

3.9 The distribution of policy-year claims in a portfolio of 6,000 identical policies is summarized in the table.

# Claims	# Policies
0	5,220
1	722
2	52
3	4
4	2
≥5	0
Total	6,000

(***a***) Fit a Poisson model to these data to obtain a probability function for the claim-count variable N for a single policy selected at random from this population.
(***b***) Check the goodness of fit of the resulting distribution with a chi-square test.

3.10 Derive these formulas for the moments of a random variable N with a mixed-Poisson distribution directly from probability function (3.12), thus verifying (3.14) and (3.15).
(***a***) $E[N^2] = E[\lambda] + E[\lambda^2]$. (***b***) $E[N^3] = E[\lambda] + 3E[\lambda^2] + E[\lambda^3]$.
(***c***) $E[(N - E[N])^3] = E[\lambda] + 3Var[\lambda] + E[(\lambda - E[\lambda])^3]$.

3.11 Assume that N has a mixed-Poisson distribution for which the mixing parameter λ has a gamma distribution with $(\alpha, \beta) = (2, 1)$. Compute:
(***a***) $E[N]$. (***b***) $Var[N]$. (***c***) $\Pr\{N \le 3\}$.

3.12 N is the claim-count variable for a policy selected at random from a population characterized by a mixture of Poisson distributions for which λ has a gamma distribution with $E[\lambda] = 0.100$ and $Var[\lambda] = 0.005$. Compute $f_N(n)$ for $n = 0, 1, 2, 3$.

3.13 Random variable N_1 is Poisson-distributed with $\lambda = 0.75$, variable N_2 has a mixed-Poisson distribution with $\Pr\{\lambda = 0.6\} = 0.75$ and $\Pr\{\lambda = 1.2\} = 0.25$, and variable N_3 has a mixed distribution for which $f_\lambda(u) = \frac{4}{3}e^{-(4/3)u}$.
(***a***) Show that $E[N_1] = E[N_2] = E[N_3]$.
(***b***) Compute $Var[N_1]$, $Var[N_2]$, and $Var[N_3]$.
(***c***) Compute $f_{N_i}(n)$ for $i = 1, 2, 3$ and $n = 0, 1, 2, 3, 4, 5$.

3.14 To the data of Example 3.6 fit a mixed Poisson distribution of the form

$$f(n) = \frac{1}{n!}\left(\omega_1\lambda_1^n e^{-\lambda_1} + \omega_2\lambda_2^n e^{-\lambda_2}\right), \quad \omega_1 + \omega_2 = 1.$$

(***a***) Compute method-of-moments parameter estimates $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\omega}_1$, $\hat{\omega}_2$.
(***b***) Compare the fit of the resulting distribution to that of the negative binomial distribution obtained in Example 3.6.

3.15 The table displays the incidence of claims from a portfolio of 10,000 annual policies. Fit a reasonable distribution model to these data. What assumptions must one make? Test the goodness of fit of the fitted distribution.

# Claims	# Policies
0	8,956
1	907
2	120
3	15
4	2
≥5	0
Total	10,000

3.16 Prove identity (3.22) for the general binomial coefficient.

3.17 Prove this identity: $\prod_{i=1}^{n-1}(x+i) = \Gamma(x+n)/\Gamma(x)$, $n = 2, 3, 4, \ldots$.

3.18 Derive these formulas for the mean and variance of a random variable N with the general negative binomial probability function (3.19).
(***a***) $E[N] = r(1-q)/q$. (***b***) $Var[N] = r(1-q)/q^2$.

3.19 Show that the negative binomial probability function (3.19) satisfies a recursion relation of the following form: there exist numbers $a > 0$ and $b > 0$ such that

$$f(n) = \frac{na + b}{n} f(n-1), \quad n = 1, 2, 3, \ldots.$$

3.20 Explain how the negative binomial probability function (3.19), whenever parameter r has a positive integer value, can be interpreted as $\Pr\{M = n\}$, where M is the number of failures occurring before the r^{th} success in a sequence of independent Bernoulli trials for which q is the probability of success in a single trial. The negative binomial distribution for which parameter r is a positive integer is sometimes called the ***Pascal distribution***, after French mathematician and philosopher Blaise Pascal (1623–1662).[37]

3.21 A random variable N with probability mass function

$$f(n) = p(1-p)^n, \quad 0 < p < 1, \quad n = 0, 1, 2, \ldots$$

has a ***geometric distribution*** with parameter p.

(***a***) Show that the geometric distribution is a special case of the Pascal distribution.

(***b***) Compute $E[N]$ and $Var[N]$ in terms of p.

3.22 A claim-count variable N is obtained as a mixture of geometric variables. Derive a formula for the probability function $f_N(n)$ under each of the following assumptions about the distribution of the variable parameter p.

(***a***) p is uniformly distributed on the interval $0 < u < 1$.

(***b***) p is distributed on the interval $0.10 < u < 1$ with $f_p(u) = 1/(u \log 10)$.

3.23 (***a***) Verify that the moment-generating function $M_\lambda(t)$ for the mixing parameter λ with the gamma probability density function (3.16) is

$$M_\lambda(t) = \left((\alpha - \nu t)/\alpha\right)^{-\alpha}.$$

(***b***) Show that the moment-generating function $M_N(t)$ for the mixed distribution (3.17) satisfies the equation $M_N(t) = M_\lambda(e^t - 1)$, where $M_\lambda(t)$ is the generating function of part (*a*).

(***c***) Prove that the relation $M_N(t) = M_\lambda(e^t - 1)$ holds for an arbitrary (not just gamma) distribution for the variable parameter λ.

---

[37] Pascal and fellow French mathematician Pierre de Fermat (1601–1665) are credited with establishing the mathematical foundations of probability. In a remarkable correspondence during the summer of 1654 they solved a celebrated problem from the realm of gambling: how should the stakes in a game of chance be divided between two equally skilled players when the game is interrupted? Pascal's easily generalized solution made use of the array of binomial coefficients that has since become known as Pascal's Triangle.

3.24 Let $\langle n_1, n_2, \ldots, n_m \rangle$ be observations of a random sample of size m drawn from a population with a negative binomial distribution (3.17) with unknown parameters (α, ν). Find formulas for the method-of-moments parameter estimates $(\hat{\alpha}, \hat{\nu})$.

3.25 For the Pólya distribution of Example 3.7, compute:
(***a***) Pr{*white on 3rd trial* | *white on 1st & black on 2nd trial*}.
(***b***) Pr{*white on 3rd trial* | *black on 1st & white on 2nd trial*}.
(***c***) Pr{*white on 3rd trial*}.

3.26 The number W_m of white balls drawn from an urn in m Pólya trials has probability function $P_n(100,25,5; m)$, $n = 0, 1, 2, \ldots, m$.
(***a***) What is the degree of contagion?
(***b***) Compute these probabilities:
Pr{*white on 1st trial*},
Pr{*white on 2nd trial* | *white on 1st trial*},
Pr{*white on 2nd trial* | *black on 1st trial*},
Pr{*white on 2nd trial*}.
(***c***) Compute the probabilities $P_n(100,25,5;4)$ for $n = 0, 1, 2, 3, 4$.
(***d***) What is the limiting distribution of W_m as $m \to \infty$ such that $mp = 3.2$ and the degree of contagion remain constant?

3.27 (***a***) Demonstrate that the Pólya probability function (3.26) can be expressed in terms of the general binomial coefficients as

$$P_n(w,b,c;m) = \binom{\frac{1}{\gamma}+n-1}{n}\binom{\frac{1-p}{\gamma p}+m-n-1}{m-n}\Bigg/\binom{\frac{1}{\gamma p}+m-1}{m},$$

where $p = w/(w+b)$ and $\gamma = c/w$.
(***b***) Show that the denominator in part (*a*) can be written as

$$\binom{\frac{1}{\gamma p}+m-1}{m} = \binom{\frac{1}{\gamma p}+m-n-1}{m-n}\prod_{i=1}^{n}\left(\frac{1}{\gamma p m}+\frac{m-i}{m}\right)\Bigg/\prod_{i=1}^{n}\left(1-\frac{i-1}{m}\right).$$

3.28 N has a negative binomial distribution with mean 1.00 and contagion parameter $\gamma = 0.20$. Compute the probabilities of $N = 0, 1, 2, 3, 4, 5$ claims.

3.29 (***a***) Three independent claim-count variables (N_1, N_2, N_3) have respective means (10, 25, 5) and contagion parameters (0.35, 0.20, 0). Compute the contagion parameter for $N = N_1 + N_2 + N_3$.
(***b***) Let N be the sum of m independent claim-count random variables with means $\{\nu_i\}$ and contagion parameters $\{\gamma_i\}$. Prove that the contagion parameter for N is

$$\gamma = \textstyle\sum_{i=1}^{m}\gamma_i\nu_i^2 \Big/ \left(\sum_{i=1}^{m}\nu_i\right)^2.$$

3.30 A policyholder owns a fleet of 20 insured automobiles. The claim process for each vehicle is Poisson-distributed with claim density of 0.30 per year. Assuming that the individual claim processes are independent, find the probability of incurring at least six auto claims in a single year.

3.31 The claim-count variable for a portfolio of 8,000 policies is Poisson-distributed with an expected value of 650 claims. Assuming also that the claim-count variable for each policy has the same Poisson distribution, compute the expected number of policyholders that produce at least one claim.

3.32 In a portfolio of m identical policies, the claim count for every policy has the same negative binomial distribution with contagion parameter γ. If γ_m is the contagion parameter of the portfolio distribution, find $\lim_{m\to\infty}\gamma_m$. What does this imply about the nature of the portfolio distribution for large m?

3.33 The random time of occurrence—or ***waiting time***—for successive claims in a claim process is occasionally of interest. In the case that the process is Poisson, the distribution of the random variable T_n, the occurrence time of the n^{th} claim, has a particularly simple form.

Note that $T_n \le t$ is identical to the event that at least n claims occur in the time interval $[0, t)$. Thus, when the claim process is Poisson with parameter λ = # *claims per unit time*, probability formula (3.4) implies that the distribution function for T_n is

$$F_n(t) = \Pr\{T_n \le t\} = \begin{cases} 0 & \text{if } -\infty < t < 0 \\ \displaystyle\sum_{k=n}^{\infty} \frac{(\lambda t)^k e^{-\lambda t}}{k!} & \text{if } 0 \le t < \infty. \end{cases}$$

(***a***) Show that T_n has the gamma probability density function

$$f_n(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t}.$$

(***b***) Obtain $E[T_n]$ and $Var[T_n]$ in terms of n and λ.

(***c***) Let $\hat{T}_n$ denote the time between successive claims:

$$\hat{T}_n = \begin{cases} T_1 & \text{if } n = 1 \\ T_n - T_{n-1} & \text{if } n \ge 2. \end{cases}$$

Show that variables $\hat{T}_n$ are independent and that each has an exponential distribution with parameter $\beta = 1/\lambda$.

3.34 The claim-count variable for a property policy with a two-year term has a Poisson distribution with 0.215 claims per year.

(***a***) What is the expected time until the occurrence of the first claim?

(***b***) What is the probability that the first claim will occur within the first year? . . . the second year?

(***c***) What is the probability that the second claim will occur within the first year? . . . the second year?

3.35 For a certain claim process the claim-count variable N has a Poisson distribution with parameter λ, and the probability that any given claim is fraudulent is p. Find the distribution of $N^* = \#\ \textit{fraudulent claims}$. [*Hint*: for $n = 1, 2, 3, \ldots$

$$f_{N^*}(n) = \sum_{k=n}^{\infty} \Pr\{n \textit{ fraudulent claims} \mid N = k\} \cdot f_N(k)$$

$$= \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} f_N(k).]$$

4. Aggregate Claims

The probability distribution of the total claim amount S for a claim process is called an ***aggregate loss*** (or ***aggregate claim***) ***distribution***. Because S depends on two independent random variables—the number of claims N and the claim size X—the distribution of S is a ***compound distribution***, that is, an appropriate combination of the claim-count and claim-size distributions. In this chapter we describe how the aggregate distribution and its properties are derived from the component distributions of N and X and then discuss some practical methods for evaluating and approximating the distribution.

4.1. A Discrete Example

Before providing a general definition of the aggregate distribution in the next section, we illustrate the basic ideas with a simple discrete model in Example 4.1.

Example 4.1. Assume first that $n = 0, 1, 2$ are the only possible numbers of claims and that there exist just three potential claim sizes: {100, 200, 300}. Associated probability functions for N and X are shown in the following tables.

Claim Count *N*

# Claims n	$f_N(n)$
0	0.60
1	0.30
2	0.10
$E[N] = 0.50$, $Var[N] = 0.45$	

Claim Size *X*

Size x	$f_X(x)$
100	0.40
200	0.50
300	0.10
$E[X] = 170$, $Var[X] = 4{,}100$	

Thus, there are seven distinct total loss amounts: {0, 100, 200, 300, 400, 500, 600}. Probabilities for these values of S are defined by

$$f_S(s) = \Pr\{S = s\} = \sum_{n=0}^{2} \Pr\{S = s | N = n\} \cdot f_N(n). \tag{4.1}$$

Conditional probabilities $\Pr\{S = s|N = n\}$ in this formula are displayed here for each n value and all possible s values.

	$n=0$	$n=1$	$n=2$		
Amount s	0	100	100 + 100	100 + 200	100 + 300
		200	200 + 100	200 + 200	200 + 300
		300	300 + 100	300 + 200	300 + 300
$\Pr\{S = s\|N = n\}$	1.00	0.40	0.16	0.20	0.04
		0.50	0.20	0.25	0.05
		0.10	0.04	0.05	0.01

Inserting these tabulated probabilities into formula (4.1) yields values of the aggregate probability function. For example,

$$\begin{aligned} f_S(300) &= \Pr\{S = 300|N = 0\} f_N(0) + \Pr\{S = 300|N = 1\} f_N(1) \\ &\quad + \Pr\{S = 300|N = 2\} f_N(2) \\ &= (0)(0.60) + (0.10)(0.30) + (0.20 + 0.20)(0.10) \\ &= 0.0700. \end{aligned}$$

Other probabilities are obtained in a similar way and then assembled to form the distribution of S, shown in the table. Figure 4.1 displays a histogram of the discrete probability mass function f_S.

The expected loss amount for such a policy is $E[S] = 85$. In the next section, we shall see that it is not merely coincidental that $E[S] = (0.50)(170) = E[N]\,E[X]$. The premium charge for such a policy would therefore be \$85 plus a loading for the expense of doing business and a provision for profit and risk. ■

Aggregate Loss S

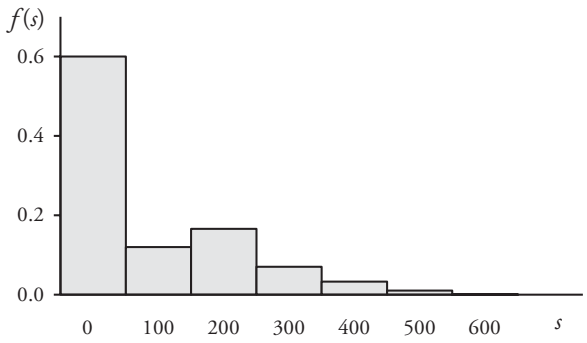
Amount s	$f_s(s)$	$F_s(s)$
0	0.6000	0.6000
100	0.1200	0.7200
200	0.1660	0.8860
300	0.0700	0.9560
400	0.0330	0.9890
500	0.0100	0.9990
600	0.0010	1.0000
$E[S] = 85$, $Var[S] = 15{,}055$		

4.2. Aggregate Distribution Properties

Example 4.1 shows how values for the aggregate random variable S can be generated in two steps: (*i*) select a number of claims $N = n$ and then (*ii*) choose n claim-size values for X. The sum of these n numbers is a single value for S. Assuming that the sizes of successive claims are mutually independent and also independent of the number of claims, one can define the aggregate random variable by

$$S = \begin{cases} 0 & \text{if } N = 0 \\ X_1 + X_2 + \cdots + X_N & \text{if } N > 0, \end{cases}$$

Figure 4.1. Aggregate Probability Mass Function [Example 4.1]



where $X_1, X_2, \ldots, X_N$ are independent random variables, all identical to X. This two-step generation of the aggregate variable S suggests how to construct the probability distribution for S from the component claim-count and claim-size distributions. In the discussion that follows, $f_N(n)$ denotes $\Pr\{N = n\}$, and $F(x)$ is the cumulative distribution function for X.

For every positive integer n define $Y_n = \sum_{k=1}^{n} X_k$ as the sum of n independent random variables, each identical to X. (For later convenience, define $Y_0 = 0$.) The distribution function of Y_n is the convolution of n replicates of $F(x)$:

$$F_n^*(y) = \Pr\{Y_n \le y\} = \underbrace{(F * F * \cdots * F)}_{n\text{-fold convolution}}(y), \quad n = 1, 2, 3, \ldots, -\infty < y < \infty.$$

The convolution of two functions is obtained by a standard integral formula, employed in the following recursive definition of the sequence $\langle F_n^*(y) \rangle$:[38]

$$F_0^*(y) = \begin{cases} 0 & \text{if } y < 0 \\ & \\ 1 & \text{if } y \ge 0 \end{cases} \quad \text{and}$$

$$F_n^*(y) = \left(F_{n-1}^* * F\right)(y) = \int_{-\infty}^{\infty} F_{n-1}^*(y-u)\, dF(u), \quad n = 1, 2, 3, \ldots. \tag{4.2}$$

Finally, the aggregate variable S has the compound distribution function

$$F_S(s) = \sum_{n=0}^{\infty} f_N(n) \Pr\{Y_n \le s | N = n\} = \sum_{n=0}^{\infty} f_N(n)\, F_n^*(s), \quad 0 \le s < \infty. \tag{4.3}$$

[38] Development of the convolution integral formula $(F_1 * F_2)(x) = \int_{-\infty}^{\infty} F_2(x-u)\, dF_1(u)$ for the distribution function of the sum of two independent random variables can be found in most textbooks of mathematical probability; see also Problem 4.2. In practice, it is usually easier to derive the distribution of the sum Y_n from the moment-generating or characteristic functions of the random variables involved than it is to perform the sequence of integrations indicated in (4.2).

The m^{th} moments of S ($m = 1, 2, 3, \ldots$) are related to the corresponding moments of the $\{Y_n\}$ variables by the equation

$$\begin{aligned} E\left[S^m\right] &= \int_0^\infty s^m \, dF_S(s) \\ &= \int_0^\infty s^m \sum_{n=0}^{\infty} f_N(n) \, dF_n^*(s) = \sum_{n=0}^{\infty} f_N(n) \int_0^\infty s^m \, dF_n^*(s) \\ &= \sum_{n=0}^{\infty} f_N(n) \, E\left[Y_n^m\right], \end{aligned} \tag{4.4}$$

provided the $E[Y_n^m]$ exist. Formulas for the first three moments of Y_n, displayed below, follow from the independence of the $\{X_k\}$ variables. Derivation of these formulas also depends on the fact that the second and third moments of a sum of independent random variables are the respective sums of the second and third moments of the summands.

$$\begin{aligned} E[Y_n] &= n E[X], \\ E\left[Y_n^2\right] &= E\left[(Y_n - E[Y_n])^2\right] + (E[Y_n])^2 \\ &= nE\left[(X - E[X])^2\right] + (n E[X])^2 \\ &= nVar[X] + n^2 (E[X])^2, \\ E\left[Y_n^3\right] &= E\left[(Y_n - E[Y_n])^3\right] + 3E[Y_n]E\left[Y_n^2\right] - 2(E[Y_n])^3 \\ &= nE\left[(X - E[X])^3\right] + 3n^2 E[X] Var[X] + n^3 (E[X])^3. \end{aligned}$$

Combining these results with equation (4.4) yields

$$E[S] = \sum_{n=1}^{\infty} f_N(n)(nE[X]) = \left(\sum_{n=0}^{\infty} n f_N(n) \right) E[X] = E[N]E[X], \tag{4.5}$$

$$\begin{aligned} E\left[S^2\right] &= \sum_{n=1}^{\infty} f_N(n) \left(nVar[X] + n^2 (E[X])^2 \right) \\ &= E[N]Var[X] + E\left[N^2\right](E[X])^2, \end{aligned} \tag{4.6}$$

$$\begin{aligned} E\left[S^3\right] &= \sum_{n=1}^{\infty} f_N(n) \left(nE\left[(X - E[X])^3\right] + 3n^2 E[X] Var[X] + n^3 (E[X])^3 \right) \\ &= E[N] E\left[(X - E[X])^3\right] + 3E\left[N^2\right] E[X] Var[X] \\ &\quad + E\left[N^3\right](E[X])^3. \end{aligned} \tag{4.7}$$

Therefore,

$$Var[S] = E[N]Var[X] + Var[N](E[X])^2,$$

$$Sk[S] = \frac{E[N]E\left[(X - E[X])^3\right] + 3Var[N]E[X]Var[X]}{(Var[S])^{3/2}}$$

$$+ \frac{E\left[(N - E[N])^3\right](E[X])^3}{(Var[S])^{3/2}}. \tag{4.8}$$

If N is distributed with mean $E[N] = \lambda$ and contagion parameter γ so that $Var[N] = \lambda + \gamma\lambda^2$, then formulas (4.5) and (4.8) become

$$E[S] = \lambda E[X],$$

$$Var[S] = \lambda E[X^2] + \gamma\lambda^2(E[X])^2,$$

$$Sk[S] = \frac{\lambda E[X^3] + 3\gamma\lambda^2 E[X]E[X^2] + 2\gamma^2\lambda^3(E[X])^3}{\left(\lambda E[X^2] + \gamma\lambda^2(E[X])^2\right)^{3/2}}. \tag{4.9}$$

In the special case that N has a Poisson distribution, these formulas reduce to

$$E[S] = \lambda E[X],$$

$$Var[S] = \lambda E[X^2],$$

$$Sk[S] = \frac{E[X^3]}{\sqrt{\lambda}\left(E[X^2]\right)^{3/2}}. \tag{4.10}$$

Derivations of (4.5)–(4.7) above, based on the fundamental equation (4.4) relating the moments of S to those of the sequence $\langle Y_n \rangle$, are completely straightforward. However, as with any random variable, these formulas can also be derived from the moment-generating function of S whenever that function exists.

To construct $M_S(t)$, start with the moment-generating function of variable Y_n. For each fixed n, $Y_n = \sum_{k=1}^{n} X_k$ is the sum of independent identical random variables, and therefore

$$M_{Y_n}(t) = E\left[\exp\left(t\sum_{k=1}^{n} X_k\right)\right] = \prod_{k=1}^{n} E[\exp(tX_k)] = (M_X(t))^n,$$

where $M_X(t)$ is the generating function for the common claim-size variable X. Accordingly, $M_S(t)$ is given by the series

$$M_S(t) = E[\exp(tY_N)] = \sum_{n=0}^{\infty} f_N(n) M_{Y_n}(t) = \sum_{n=0}^{\infty} f_N(n)(M_X(t))^n.$$

But this last formula can be interpreted as an expected value with respect to the distribution of N:

$$M_S(t) = \sum_{n=0}^{\infty} f_N(n)\left(M_X(t)\right)^n = \sum_{n=0}^{\infty} f_N(n)\exp\left(n\log M_X(t)\right) = E_N\left[\exp\left(n\log M_X(t)\right)\right].$$

Thus, in terms of the generating function M_N for N:

$$M_S(t) = M_N\left(\log M_X(t)\right). \tag{4.11}$$

As usual, $E[S]$ is now obtainable from (4.11) by differentiation:

$$E[S] = M_S'(0) = M_N'\left(\log M_X(t)\right)\frac{M_X'(t)}{M_X(t)}\bigg|_{t=0} = M_N'(0)\frac{M_X'(0)}{M_X(0)} = E[N]E[X].$$

Similar derivations of formulas (4.6) and (4.7) are requested in Problem 4.5.

Example 4.2. Assume that the claim-count random variable N has a Poisson (λ) distribution:

$$f_N(n) = \frac{\lambda^n e^{-\lambda}}{n!}, \quad n = 0, 1, 2, 3, \ldots$$

Moreover, suppose that claim size X is gamma (α, β) distributed:

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{\Gamma(x/\beta, \alpha)}{\Gamma(\alpha)} & \text{if } 0 \le x < \infty. \end{cases}$$

Accordingly, the moment-generating function for X is $M_X(t) = (1 - \beta t)^{-\alpha}$, and the generating function for the sum of n independent such gamma variables is the n^{th} power of $M_X(t)$:

$$\left(M_X(t)\right)^n = \left((1-\beta t)^{-\alpha}\right)^n = (1-\beta t)^{-n\alpha}, \quad -\infty < t < 1/\beta.$$

However, this is the generating function of a gamma variable with parameters ($n\alpha$, β), so the n-fold convolution of identical gamma-distributed variables also has a gamma distribution:

$$F_n^*(y) = \begin{cases} 0 & \text{if } -\infty < y < 0 \\ \dfrac{\Gamma(y/\beta, n\alpha)}{\Gamma(n\alpha)} & \text{if } 0 \le y < \infty, \quad n = 1, 2, 3, \ldots. \end{cases}$$

Note that deriving this convolution formula from the moment-generating function is considerably less onerous than carrying out the successive integrations indicated in formula (4.2). The distribution function for this combination of N and X can therefore be expressed in closed analytic form:

$$F_S(s) = \begin{cases} 0 & \text{if } -\infty < s < 0 \\ \displaystyle\sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \cdot \frac{\Gamma(s/\beta, n\alpha)}{\Gamma(n\alpha)} & \text{if } 0 \le s < \infty. \end{cases} \tag{4.12}$$

In the particular instance that $\lambda = 2.5$ and $(\alpha, \beta) = (3, 400)$, the Poisson formulas (4.10) imply that

$$E[S] = \lambda\alpha\beta = (2.5)(3)(400) = 3{,}000,$$

$$Var[S] = \lambda\alpha(\alpha+1)\beta^2 = (2.5)(3)(4)(400)^2 = 4{,}800{,}000,$$

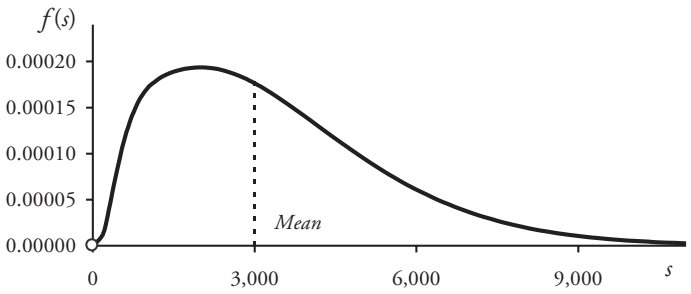
$$Sk[S] = \frac{\alpha+2}{\sqrt{\lambda\alpha(\alpha+1)}} = \frac{5}{\sqrt{30}} = 0.9129.$$

Values for the cumulative distribution function $F(s)$ in this special case are displayed in Table 4.1. The distribution has a discrete lump of probability of size $f_N(0) = e^{-2.5} = 0.0821$ at $s = 0$, but at all other s values $F(s)$ is continuous. A graph of the corresponding probability density function is shown in Figure 4.2. ■

Table 4.1. Aggregate Distribution [Example 4.2]

Amount s	$F(s)$
0	0.0821
500	0.1096
1,000	0.1867
2,000	0.3755
3,000	0.5613
4,000	0.7152
5,000	0.8273
6,000	0.9013
7,000	0.9465
8,000	0.9723
9,000	0.9863
10,000	0.9934

Figure 4.2. Aggregate Density Function [Example 4.2]



In Example 4.2 we observed that the convolution of a gamma cumulative distribution function with itself is also gamma, a fact which led to an easy-to-calculate exact formula for the aggregate distribution function of that example. However, this desirable *reproductive property*—the distribution of a sum of identical independent random variables having the same distribution type as the components—is shared by just a few families of distributions (notably the normal distributions, which are not generally useful as claim-size distributions). In fact, sums of the ubiquitous lognormal and Pareto distributions belong neither to their respective families nor to any other familiar parametric distribution family. As a consequence, actuaries have since the mid-1900s sought to develop various procedures for calculating values of an aggregate distribution. Among these are several approximations using easily calculable parametric distributions, algorithms featuring recursive formulas, Fourier-transform-based methods, and Monte Carlo simulation. The remainder of this chapter is devoted to some of the most important of these techniques.

4.3. Approximation by Matching Moments

In this section we discuss a traditional technique of approximation, the method of ***matching moments***, similar to the method-of-moments for fitting a distribution model to sample data. This approach is based on the not-unreasonable assumption that two distributions with identical moments of order m, where usually $1 \le m \le 3$, are sufficiently similar that one distribution can be used to approximate the other.[39]

The method consists of two steps:

(*i*) Calculate from the moments of the claim-count and claim-size distributions the required mean $\mu = E[S]$, variance $\sigma^2 = Var[S]$, and skewness $\kappa = Sk[S]$ according to formulas (4.5) and (4.8).

(*ii*) Select from a convenient parametric family of continuous distributions the particular member with matching respective first, second, and third moments.

[39] Although the method of matching moments usually gives reasonable results, the assumption on which it is based—that distributions with identical lowest moments are indeed comparable—could fail to hold. It is possible for distributions with identical first-, second-, and third-order moments to be significantly different. For a discussion of this "moment problem" see Pentikäinen [19] and the references cited there. However, Pentikäinen suggests that acceptable approximations are usually obtained by the method of matching moments when the variable X is restricted to a finite interval, as in the case that claim size is limited by policy conditions.

The cumulative distribution function of this selected parametric distribution could then serve as an approximation to F_S.

Normal Approximation

There are several reasons for considering a normal approximation to an aggregate-loss distribution. The variables Y_n defined in Section 4.2 are the sums of independent, identically distributed claim-size random variables. Thus, when n is large, the Central Limit Theorem implies that Y_n is approximately normally distributed. In addition, whenever the claim count is Poisson-distributed with mean λ (but not when N has a positive contagion parameter—see Problem 4.8), the Poisson formulas (4.10) imply that $Sk[S]$ is small for large values of λ. In fact, in the Poisson case, one can observe that $\lim_{\lambda\to\infty} Sk[S] = 0$, so that the distribution of S is asymptotically symmetric.

In such a case—when N is Poisson-distributed and λ is large—it is useful to try the approximation $S \approx Y = \sigma Z + \mu$, where Z is the standard normal variable. This is equivalent to asserting that the standardized variable

$$T(S) = \frac{S - \mu}{\sigma} \tag{4.13}$$

is approximately standard normal. The ***normal approximation*** $S \approx Y = \sigma Z + \mu$ (or equivalently, $T(S) \approx Z$), yields

$$\begin{aligned} F_S(s) &= \Pr\{S \le s\} = \Pr\{T(S) \le T(s)\} \\ &\approx \Pr\{Z \le T(s)\} = \Phi(T(s)) = \Phi((s-\mu)/\sigma), \quad 0 \le s < \infty. \end{aligned} \tag{4.14}$$

Because $E[Y] = E[S] = \mu$ and $Var[Y] = Var[S] = \sigma^2$, the normal approximation (4.14) certainly involves matching the first two, but not necessarily the third, moments of the distributions. Of course, the skewness of the symmetric variable $Y = \sigma Z + \mu$ is zero, whereas $Sk[S]$ is usually positive. Because of this, the normal approximation generally underestimates the probabilities of large claims. Moreover, it could assign significant probability to negative values of s and thereby fail to model acceptably the short tail of the aggregate distribution (for instance, refer to Example 4.3). Obviously, the normal approximation is useful only in those cases where S is nearly symmetric. In other situations one must look elsewhere for a satisfactory approximation.

Gamma Approximation

When S is notably skewed, one way to improve on the normal approximation is to match moments with a known skewed distribution. The versatile family of gamma distributions often provides a reasonable starting point.

For example, consider a gamma-distributed variable G with parameters (α, β). The required mean μ and variance σ^2 then determine α and β:

$$\begin{cases} \mu = \alpha\beta \\ \sigma^2 = \alpha\beta^2 \end{cases} \quad \text{implies} \quad \begin{cases} \alpha = \mu^2/\sigma^2 \\ \beta = \sigma^2/\mu. \end{cases}$$

With these parameters the distribution of G has the specified mean and variance, and it is also skewed, with $Sk[G] = 2\sigma/\mu > 0$. However, the utility of the gamma approximation $S \approx G$ depends on how close $2\sigma/\mu$ is to the desired skewness κ.

For a better approximation—one that matches all three parameters μ, σ, and κ—start again with a gamma variable G, except this time solve for the gamma parameters in terms of σ and κ:

$$\begin{cases} \sigma^2 = \alpha\beta^2 \\ \kappa = 2/\sqrt{\alpha} \end{cases} \quad \text{implies} \quad \begin{cases} \alpha = 4/\kappa^2 \\ \beta = \dfrac{1}{2}\sigma\kappa. \end{cases} \tag{4.15}$$

As before, the distribution of G is now completely determined, but with $E[G] = \alpha\beta = 2\sigma/\kappa$. In order to match the required aggregate mean μ we introduce the shifted variable $Y = G + \mu - 2\sigma/\kappa$, a random variable with all three specified properties:

$$E[Y] = E[G] + \mu - 2\sigma/\kappa = \mu,$$

$$Var[Y] = Var[G] = \sigma^2,$$

$$Sk[Y] = Sk[G] = \kappa.$$

The distribution function of the resulting ***shifted gamma approximation*** $S \approx Y$ is

$$\begin{aligned} F_S(s) \approx F_Y(s) &= F_G(s - \mu + 2\sigma/\kappa) \\ &= \Gamma((s-\mu)/\beta + \alpha; \alpha)/\Gamma(\alpha), \quad \mu - 2\sigma/\kappa \le s < \infty, \end{aligned} \tag{4.16}$$

where gamma parameters α and β are given by (4.15). Depending on the sign and magnitude of the quantity $\mu - 2\sigma/\kappa$, the shift of the origin sometimes adversely affects the modeling of the short tail of the distribution, as in the case of the normal approximation (again, refer to Example 4.3).

Normalizing Transformations

The normal approximation to S can also be improved by finding a refinement of the standardizing transformation (4.13) that allows for a better match of the third moments. Specifically, one could look for a transform $T(S)$ with not only the properties $E[T(S)] = 0$ and $Var[T(S)] = 1$ as in (4.13), but with the additional property that the transformed variable be symmetric, or nearly so: $Sk[T(S)] \approx 0$. If such a "symmetrizing" function T could be found, then the assumption $T(S) \approx Z$ is more likely to provide a satisfactory approximation to S. Such a transformation must necessarily be more complex than that of standardizing transformation (4.13). In particular, it cannot be *linear,* because the skewness property of a random variable is invariant under such a transformation (refer to Problem 2.26). Two such normalizing functions, described in this section, have been used extensively—the normal power and the Wilson–Hilferty transformations.

For a random variable S with mean μ, variance σ^2, and skewness κ the ***normal power transformation*** is defined by

$$T_{NP}(S) = \sqrt{\frac{6}{\kappa} \cdot \frac{S-\mu}{\sigma} + \frac{9}{\kappa^2} + 1} - \frac{3}{\kappa}. \tag{4.17}$$

It was proposed in 1969 by Finnish authors Lauri Kauppi and Pertti Ojantakanen, who were seeking an approximation to the Poisson case of an aggregate distribution.[40] Kauppi and Ojantakanen observed that for large values of S the standardized aggregate variable $(S-\mu)/\sigma$ could be approximated by a certain quadratic polynomial Q in the standard normal variable Z:

$$Q(Z) = \frac{\kappa}{6}\left(Z^2 - 1\right) + Z \approx \frac{S-\mu}{\sigma}. \tag{4.18}$$

This approximation formula is based on the so-called Edgeworth series expansion of a distribution function.[41]Solving the approximate equation (4.18) for Z yields formula (4.17) and the approximation

$$Z \approx T_{NP}(S) = Q^{-1}\left((S-\mu)/\sigma\right) = \sqrt{\frac{6}{\kappa} \cdot \frac{S-\mu}{\sigma} + \frac{9}{\kappa^2} + 1} - \frac{3}{\kappa}. \tag{4.19}$$

Thus, the ***normal power approximation*** to F_S is

$$F_S(s) \approx \Phi\left(T_{NP}(s)\right). \tag{4.20}$$

Formula (4.20) is generally applicable to the long tail of the distribution, the main region of interest in most applications. T_{NP} is somewhat less successful in modeling the short tail, but a refinement of $T_{NP}(s)$ for smaller values of s exists.[42] The Normal Power approach can generally be relied upon to give acceptable results whenever S is moderately skewed, say when $\kappa < 2$.

Another classic approach to this problem is based on the work of Harvard statisticians Edwin B. Wilson and Margaret M. Hilferty. In 1931 Wilson and Hilferty developed a transformation of the chi-square variable $X = \chi^2(m)$ with m degrees of freedom that yielded, approximately, the standard normal variable Z:

$$W(X) = \frac{\sqrt[3]{\frac{1}{m}X} - \left(1 - \frac{2}{9m}\right)}{\sqrt{\frac{2}{9m}}} \approx Z. \tag{4.21}$$

[40] Kauppi and Ojantakanen [10].

[41] A detailed derivation of the normal power approximation from the Edgeworth expansion can be found in Beard, Pentikäinen, and Pesonen [3], pp. 107–110, 355–356.

[42] *Ibid.,* pp. 116–117.

This transformation gave rise to a remarkably accurate approximation to the cumulative distribution function of the chi-square random variable:

$$F_{\chi^2(m)}(x) \approx \Phi(W(x)), \tag{4.22}$$

illustrated in Table 4.2. Since its initial appearance the Wilson–Hilferty transformation has been successfully generalized to other random variables—including, as we shall see, moderately skewed aggregate-loss variables.

To generalize (4.21) in this way, recall that $\chi^2(m)$ is gamma-distributed, with parameters $\alpha = (1/2)\,m$ and $\beta = 2$. Thus, $E[\chi^2(m)] = m$ and $Var[\chi^2(m)] = 2m$. Thus, for the scaled variable $Y = (1/m)\,\chi^2(m)$ we have $E[Y] = 1$ and $Var[Y] = 2/m$. Setting $v = \sqrt{Var[Y]}$, one can express transformation W in (4.21) as

$$W(Y) = \frac{3}{v}\left(Y^{1/3} - 1\right) + \frac{v}{3}. \tag{4.23}$$

It is a simple matter now to apply (4.23) to an arbitrary gamma random variable G with parameters (α, β). In this case, set $Y = G/(\alpha\beta)$, for which $v = 1/\sqrt{\alpha}$. Consequently,

$$W(Y) = 3\sqrt{\alpha}\left(Y^{1/3} - 1\right) + \frac{1}{3\sqrt{\alpha}},$$

or in terms of the variable G,

$$W(G) = 3\sqrt[3]{\alpha}\left(\frac{G - \alpha\beta}{\sqrt{\alpha}\beta} + \sqrt{\alpha}\right)^{1/3} - 3\sqrt{\alpha} + \frac{1}{3\sqrt{\alpha}}. \tag{4.24}$$

Table 4.2. Wilson-Hilferty Approximation to $\chi^2(10)$

x	$F_{\chi^2(10)}(x)$	$\Phi(W(x))$	Relative Error
3	0.0186	0.0193	+3.76%
6	0.1847	0.1837	–0.54%
9	0.4679	0.4672	–0.15%
12	0.7149	0.7155	+0.08%
15	0.8679	0.8686	+0.08%
18	0.9450	0.9453	+0.03%
21	0.9789	0.9789	+0.00%
24	0.9924	0.9923	–0.01%
27	0.9974	0.9973	–0.01%
30	0.9991	0.9991	0.00%

Replacing G with S and substituting $\alpha\beta = E[G] = \mu$, $\alpha\beta^2 = Var[G] = \sigma^2$, $2/\sqrt{\alpha} = Sk[G] = \kappa$ into (4.24) leads to a transformation of the aggregate variable S:

$$T_{WH}(S) = 3\left(\frac{2}{\kappa}\right)^{2/3}\left(\frac{S-\mu}{\sigma}+\frac{2}{\kappa}\right)^{1/3}-\frac{6}{\kappa}+\frac{\kappa}{6}. \tag{4.25}$$

As in (4.22), we obtain the ***Wilson–Hilferty approximation*** to distribution function $F_S(s)$:

$$F_S(s) \approx \Phi\left(T_{WH}(s)\right), \quad 0 \le s < \infty. \tag{4.26}$$

Because this approximation has been so successfully applied to gamma random variables, which in turn have provided acceptable approximations to a wide range of aggregate distributions, it is not surprising that the Wilson–Hilferty formula has proved to be useful in approximating aggregate distributions, as well.

In fact, all three approaches that take into consideration the skewness of S—the shifted gamma, the normal power, and the Wilson–Hilferty schemes—provide acceptable approximations to the aggregate-loss variable S whenever the skewness is not too large.

Example 4.3. The result of applying the normal (4.14), shifted gamma (4.16), normal power (4.20), and Wilson–Hilferty (4.26) approximations to the moderately skewed distribution of Example 4.2 are displayed in Table 4.3. The normal approximation clearly fails to yield a reasonable result, whereas the other three methods generate quite acceptable approximations to the long tail of the distribution.

Application of these same approximations to the Poisson/gamma distribution (4.12) for which $\lambda = 10$, $\alpha = 0.05$, and $\beta = 6{,}000$ yields the outcomes shown in Table 4.4.

Table 4.3. Approximations to $F_S(s)$: $\mu = 3{,}000$, $\sigma = 2{,}191$, $\kappa = 0.9129$ [Example 4.3]

Amount s	$F(s)$	Normal	Relative Error	Normal Power	Relative Error	Shifted Gamma	Relative Error	Wilson-Hilferty	Relative Error
0	0.0821	0.0855	+4.14%	0.0534	–34.96%	0.0459	–44.09%	0.0464	–43.48%
1,000	0.1867	0.1807	–3.21%	0.1900	+1.77%	0.1775	–4.93%	0.1765	–5.46%
2,000	0.3755	0.3240	–13.72%	0.3745	–0.27%	0.3680	–2.00%	0.3668	–2.32%
3,000	0.5613	0.5000	–10.92%	0.5591	–0.39%	0.5607	–0.11%	0.5605	–0.14%
4,000	0.7152	0.6760	–5.48%	0.7125	–0.38%	0.7185	+0.46%	0.7191	+0.55%
5,000	0.8273	0.8193	–0.97%	0.8245	–0.34%	0.8310	+0.45%	0.8318	+0.54%
6,000	0.9013	0.9145	+1.46%	0.8987	–0.29%	0.9038	+0.28%	0.9044	+0.34%
7,000	0.9465	0.9661	+2.07%	0.9443	–0.23%	0.9475	+0.11%	0.9478	+0.14%
8,000	0.9723	0.9888	+1.70%	0.9707	–0.16%	0.9724	+0.01%	0.9724	+0.01%
10,000	0.9934	0.9993	+0.59%	0.9927	–0.07%	0.9930	–0.04%	0.9929	–0.05%

Table 4.4. Approximations to $F_S(s)$: $\mu = 3{,}000$, $\sigma = 4{,}347$, $\kappa = 2.8293$ [Example 4.3]

Amount s	$F(s)$	Normal	Relative Error	Normal Power	Relative Error	Shifted Gamma	Relative Error	Wilson-Hilferty	Relative Error
0	0.00005	0.2451	—	0.4023	—	0.1228	—	0.1494	—
2,000	0.5922	0.4090	–30.94%	0.5866	–0.95%	0.5886	–0.61%	0.5835	–1.47%
4,000	0.7513	0.5910	–21.34%	0.7108	–5.39%	0.7504	–0.12%	0.7519	+0.08%
6,000	0.8401	0.7549	–10.14%	0.7978	–5.04%	0.8402	–0.01%	0.8443	+0.50%
8,000	0.8946	0.8750	–2.19%	0.8590	–3.98%	0.8949	+0.03%	0.8992	+0.51%
10,000	0.9294	0.9463	+1.82%	0.9020	–2.95%	0.9298	+0.04%	0.9333	+0.42%
12,000	0.9522	0.9808	+3.00%	0.9322	–2.10%	0.9525	+0.03%	0.9552	+0.32%
14,000	0.9674	0.9943	+2.78%	0.9532	–1.47%	0.9676	+0.02%	0.9694	+0.21%
16,000	0.9777	0.9986	+2.14%	0.9678	–1.01%	0.9778	+0.01%	0.9789	+0.12%
18,000	0.9846	0.9997	+1.53%	0.9779	–0.68%	0.9847	+0.01%	0.9853	+0.07%

This second distribution is considerably more skewed than that in Table 4.3, with $\mu = 3{,}000$, $\sigma = 4{,}347$, and $\kappa = 2.8293$. Again, as expected, the normal approximation is unsuitable. The shifted gamma and Wilson–Hilferty methods, however, produce generally satisfactory results, at least to the long tail, while the normal power approximation is less accurate. ■

4.4. Recursion

In contrast to the method of matching moments, in which the approximating distribution for the aggregate-loss random variable is selected from a family of continuous distributions, the next technique under consideration involves a discrete approximating distribution. Values of this distribution are calculated by means of a recursion formula for the aggregate probability function.

The recursion approach has been studied since the mid-1960s, when the Poisson case was first described by R. M. Adelson. It was later extended to other cases by such authors as H. H. Panjer.[43] We present in this section a basic formulation of the recursion method, which rests on a pair of assumptions, one for each of the variables N and X.

Suppose first that the claim count N has a distribution for which the probability function $f_N(n)$ satisfies, for some constants a and b, a recursion relation on n:

$$f_N(n) = \frac{na + b}{n} f_N(n-1), \quad n = 1, 2, 3, \ldots. \tag{4.27}$$

[43] Panjer [17].

Whenever N is Poisson-distributed with $E[N] = \lambda$, it is easy to show that probabilities $f_N(n)$ satisfy (4.27) with $a = 0$ and $b = \lambda$. This relation also holds in the negative binomial case—refer to Problem 3.19.

In addition, assume that claim-size variable X has a *discrete* structure, with only a finite number of equally spaced values x_k:

$$\{x_k\} = \{kh : k = 0, 1, 2, \ldots, \hat{m}\}, \quad \text{where } h > 0 \text{ is the constant } \textit{step length}. \tag{4.28}$$

We denote the probability mass function for X by

$$g(k) = \Pr\{X = x_k\} = f_X(x_k),$$

for which, as usual, $g(k) \geq 0$ and $\sum_{k=0}^{\infty} g(k) = \sum_{k=0}^{\hat{m}} g(k) = 1$. It is convenient to select $\hat{m}$ so that $\hat{m} = \max\{k : g(k) > 0\}$.

Again, let $Y_n = \sum_{i=1}^{n} X_i$ be the sum of n ($n \geq 1$) independent random variables, each identical to X. Because the component variables $\{X_i\}$ can have only values that are multiples of h, this is true for each Y_n and for the aggregate loss variable S, as well. Probabilities for Y_n are denoted by

$$g_n(m) = \Pr\{Y_n = mh\}, \quad m = 0, 1, 2, 3, \ldots,$$

where, by convention, $g_0(0) = 1$ and $g_0(m) = 0$ when $m \geq 1$. Thus, the probability function $f_S(m)$ for S has the form

$$f_S(m) = \Pr\{S = mh\} = \sum_{n=0}^{\infty} f_N(n)\, g_n(m), \quad m = 0, 1, 2, 3, \ldots. \tag{4.29}$$

Because of the independence of the $\{X_i\}$ it is easy to verify that the convolution probabilities $g_n(m)$ can be expressed recursively for positive n by

$$g_n(m) = \sum_{k=0}^{m} g(k)\, g_{n-1}(m-k), \quad m = 0, 1, 2, 3, \ldots. \tag{4.30}$$

In addition, observe that $g_n(0) = g^n(0)$ for each positive n, so that

$$f_S(0) = \sum_{n=0}^{\infty} f_N(n)\, g^n(0) = \begin{cases} f_N(0) & \text{if } g(0) = 0 \\ M_N(\log g(0)) & \text{if } g(0) > 0. \end{cases} \tag{4.31}$$

Finally, applying (4.27) to formula (4.29), we obtain

$$f_S(m) = \sum_{n=1}^{\infty} \left(a + \frac{b}{n}\right) f_N(n-1)\, g_n(m), \quad m = 1, 2, 3, \ldots. \tag{4.32}$$

Having established these preliminary results, we can now state and prove the main theorem about $f_S(m)$:

The probability function for the aggregate-loss variable with a claim-count distribution satisfying (4.27) *and a claim-size variable having the discrete structure* (4.28) *satisfies a recursion relation on the integer variable m:*

$$f_S(m) = \frac{1}{1 - ag(0)} \sum_{k=1}^{m} \left(a + \frac{b}{m} k \right) g(k) f_S(m-k), \quad m = 1, 2, 3, \ldots, \tag{4.33}$$

and $f_S(0)$ *is given by* (4.31).

Proof: Verification of formula (4.33) rests on an ingenious argument about conditional probabilities and expectations for the random variables involved in the sums $\{Y_n\}$ to create an alternative expression for $g_n(m)$.

Begin by considering the following conditional probability formula for X_n. Variables X_n and $X_1 + X_2 + \cdots + X_{n-1}$ are independent, so for each k and for each positive m for which $g_n(m) \neq 0$

$$\Pr\{X_n = kh | Y_n = mh\} = \frac{g(k) g_{n-1}(m-k)}{g_n(m)}.$$

Subject to the condition $Y_n = mh$, the expected value of X_n is therefore

$$E[X_n | Y_n = mh] = h \sum_{k=1}^{m} \frac{k g(k) g_{n-1}(m-k)}{g_n(m)}. \tag{4.34}$$

It is obvious that

$$\sum_{i=1}^{n} E[X_i | Y_n = mh] = E[Y_n | Y_n = mh] = mh. \tag{4.35}$$

However, the random variables $\{X_i\}$ are identical and independent, and they play symmetric roles in the definition of Y_n. This means that the expected values $E[X_i | Y_n = mh]$ must be identical, so that the sum in equation (4.35) must also equal $nE[X_n | Y_n = mh]$. Consequently, $E[X_n | Y_n = mh] = (m/n)h$. Substituting this value into equation (4.34) yields the alternate formula

$$g_n(m) = \frac{n}{m} \sum_{k=1}^{m} k g(k) g_{n-1}(m-k). \tag{4.36}$$

But $g(k)g_{n-1}(m-k) = 0$ whenever $g_n(m) = 0$, so (4.36) is valid for *all* values of m.

To conclude, apply (4.30) and (4.36) to formula (4.32) and obtain

$$f_S(m) = \sum_{n=1}^{\infty} a f_N(n-1) g_n(m) + \sum_{n=1}^{\infty} \frac{b}{n} f_N(n-1) g_n(m)$$

$$= \sum_{n=1}^{\infty} a f_N(n-1) \sum_{k=0}^{m} g(k) g_{n-1}(m-k) + \sum_{n=1}^{\infty} \frac{b}{n} f_N(n-1) \cdot \frac{n}{m} \sum_{k=1}^{m} k g(k) g_{n-1}(m-k)$$

$$= ag(0)\sum_{n=1}^{\infty} f_N(n-1)\,g_{n-1}(m) + \sum_{k=1}^{m}\left(a+\frac{b}{m}k\right)g(k)\sum_{n=1}^{\infty} f_N(n-1)\,g_{n-1}(m-k)$$

$$= ag(0)\,f_S(m) + \sum_{k=1}^{m}\left(a+\frac{b}{m}k\right)g(k)\,f_S(m-k).$$

Solving this equation for $f_S(m)$ yields (4.33), as required. ■

Example 4.4. Claim-count random variable N is Poisson-distributed with mean $\lambda = 1.75$. Variable X has a discrete structure of the form (4.28), with $h = 1{,}000$, $\hat{m} = 5$, and the tabulated probabilities $g(k)$.

k	x_k	$g(k)$
1	1,000	0.20
2	2,000	0.40
3	3,000	0.20
4	4,000	0.15
5	5,000	0.05
≥6	1,000 k	0.00

Applying formula (4.33) in succession yields the probability function for the aggregate variable S:

$$f_S(0) = e^{-1.75} = 0.1738,$$

$$f_S(1) = \frac{1.75}{1}(1)(0.20)(0.1738) = 0.0608,$$

$$f_S(2) = \frac{1.75}{2}[(1)(0.20)(0.0608) + (2)(0.40)(0.1738)] = 0.1323,$$

$$f_S(3) = \frac{1.75}{3}[(1)(0.20)(0.1323) + (2)(0.40)(0.0608) + (3)(0.20)(0.1738)]$$

$$= 0.1046,$$

$$f_S(4) = \frac{1.75}{4}[(1)(0.20)(0.1046) + (2)(0.40)(0.1323) + (3)(0.20)(0.0608)$$

$$+ (4)(0.15)(0.1738)] = 0.1170,$$

$$\ldots.$$

The cumulative probability function F is a step function:

$$F_S(s) = \sum_{k=0}^{[\![s/h]\!]} f_S(k).$$

Values of the derived discrete distribution functions for S are displayed in Table 4.5. ■

In order to use formula (4.33) to approximate the distribution of an aggregate-loss variable S for which the claim-size variable X is continuous, or continuous almost everywhere, one must first approximate X with a discrete variable of the form (4.28) by selecting parameters h and $\hat{m}$ and defining probabilities $g(k)$.

Table 4.5. Aggregate Distribution [Example 4.4]

Amount s	$f_S(s)$	$F_S(s)$
0	0.1738	0.1738
1,000	0.0608	0.2346
2,000	0.1323	0.3669
3,000	0.1046	0.4715
4,000	0.1170	0.5886
5,000	0.0932	0.6818
6,000	0.0786	0.7604
7,000	0.0641	0.8245
8,000	0.0499	0.8744
9,000	0.0377	0.9121
10,000	0.0274	0.9395
12,000	0.0138	0.9729
14,000	0.0063	0.9886
16,000	0.0027	0.9955

In general, greater accuracy is achieved by choosing h small and $\hat{m}$ large. However, there does exist a tradeoff. The recursive nature of the method requires the calculation of all preceding values $\{f_S(1), f_S(2), \ldots, f_S(m-1)\}$ before $f_S(m)$ can be evaluated, necessitating a large number of arithmetic operations in most applications. Calculation time can be adversely affected if $\hat{m}$ becomes too large.

Whenever X is censored—say, by a policy limit l—one should select h and $\hat{m}$ so that $\hat{m}h = l$. On the other hand, if X is unlimited, then $\hat{m}h$ must be large enough to guarantee that $1 - F_X(\hat{m}h)$ is small, as in Example 4.5.

Probabilities $g(k)$ can be defined in variety of ways. In general, one is faced with the problem of starting with a continuous probability distribution defined by F_X for *intervals* of X values and redistributing the total probability mass to a finite set of *point* values. One simple technique, often referred to as the ***midpoint method***, is to treat the lattice points $\{kh\}$ as the midpoints of certain intervals and then assign probabilities as follows:

$$
\begin{aligned}
g(0) &= F_X\left(\tfrac{1}{2}h\right), \\
g(k) &= F_X\left(\left(k+\tfrac{1}{2}\right)h\right) - F_X\left(\left(k-\tfrac{1}{2}\right)h\right), \quad k = 1, 2, \ldots, \hat{m}-1, \\
g(\hat{m}) &= 1 - F_X\left(\left(\hat{m}-\tfrac{1}{2}\right)h\right).
\end{aligned}
\tag{4.37}
$$

One difficulty with the midpoint method is that when h is large and $\hat{m}$ is small the discrete distribution may fail to have the same moments as the continuous distribution for X. This can often be improved by a careful selection of h and $\hat{m}$.

Example 4.5. Return again to the aggregate-loss variable of Example 4.2, in which N is Poisson with $E[N] = 2.5$ and X is gamma-distributed with $(\alpha, \beta) = (3, 400)$, so that $E[X] = 1{,}200$ and $Var[X] = 480{,}000$.

Now approximate the distribution function using recursion model (4.33), with the midpoint method for assigning claim-size probabilities and two choices for parameters h and $\hat{m}$: (A) $(h, \hat{m}) = (100, 60)$ and (B) $(h, \hat{m}) = (20, 300)$. Note that $\hat{m}h = 6{,}000$ in both cases, and that $F_X(6{,}000) = 0.99996$. Both sets of parameters return good approximations to $E[X]$ and $Var[X]$: (1,199.98; 480,642) for option (A) and (1,199.88; 479,846) for (B). Nevertheless, the two options do yield materially different approximations to $F_S(s)$, as shown in Table 4.6. ■

4.5. Fourier Approximation

We have already observed that the moment-generating function of a random variable is a Laplace transform of its probability density function f. In an analogous way, the ***characteristic function*** φ of a random variable is defined as a Fourier transform of the density function:

$$\varphi(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx \qquad \text{for all real } t\left(i = \sqrt{-1}\right). \tag{4.38}$$

Table 4.6. Aggregate Distribution, Discrete Approximations [Example 4.5]

Amount s	$F_S(s)$	*Approx* (A) $h = 100$	Relative Error	*Approx* (B) $h = 20$	Relative Error
0	0.0821	0.0821	0.00%	0.0821	0.00%
500	0.1096	0.1158	+5.66%	0.1108	+1.09%
1,000	0.1867	0.1956	+4.77%	0.1885	+0.96%
2,000	0.3755	0.3852	+2.58%	0.3775	+0.53%
3,000	0.5613	0.5699	+1.53%	0.5630	+0.30%
4,000	0.7152	0.7218	+0.92%	0.7165	+0.18%
5,000	0.8273	0.8318	+0.54%	0.8282	+0.11%
6,000	0.9013	0.9042	+0.32%	0.9019	+0.07%
7,000	0.9465	0.9482	+0.18%	0.9469	+0.04%
8,000	0.9723	0.9733	+0.10%	0.9725	+0.02%
9,000	0.9863	0.9868	+0.05%	0.9864	+0.01%
10,000	0.9934	0.9937	+0.03%	0.9935	+0.01%

Whereas the moment-generating function of a random variable could fail to exist, the expected value in (4.38) exists for every random variable. Moreover, to every characteristic function there corresponds a unique probability distribution, thus establishing a one-to-one correspondence between the set of probability distributions and the set of characteristic functions.

There exists a variety of formulas that invert formula (4.38) and allow recovery of the density function f—or equivalently, the distribution function F—from $\varphi(t)$. Particularly useful in this section is the inversion formula

$$\frac{F(x+)+F(x-)}{2}=\frac{1}{2}-\frac{1}{\pi}\int_0^\infty \frac{\Im\left(e^{-itx}\varphi(t)\right)}{t}\,dt, \tag{4.39}$$

where $F(x+)$ and $F(x-)$ are the respective right- and left-hand limits of F at x.[44]

The correspondence between distributions and characteristic functions has been exploited by several authors—notably in the early 1980s by Philip Heckman and Glenn Meyers[45]—to develop methods for approximating an aggregate distribution function. These methods generally involve setting up the approximating function in a such a way that an appropriate inversion formula becomes easy to evaluate. The general approach to using characteristic functions as the basis of an approximation is outlined in this section, with particular attention paid to the Heckman–Meyers approach.

The characteristic function of an aggregate variable S is defined in a manner analogous to that of the moment-generating function $M_S(t)$. For each positive integer n the characteristic function of $Y_n=\sum_{k=1}^n X_k$ is given by the product

$$\varphi_{Y_n}(t)=E\left[\exp\left(it\sum_{k=1}^n X_k\right)\right]=\prod_{k=1}^n E\left[\exp(it\,X_k)\right]=\left(\varphi_X(t)\right)^n,$$

where $\varphi_X(t)$ is that of the common claim-size distribution. Thus, $\varphi_S(t)$ is given by the series

$$\varphi_S(t)=E\left[\exp(itY_N)\right]=\sum_{n=0}^\infty f_N(n)\varphi_{Y_n}(t)=\sum_{n=0}^\infty f_N(n)\left(\varphi_X(t)\right)^n. \tag{4.40}$$

Finally, $\varphi_S(t)$ can be expressed in terms of $M_N(t)$, as in (4.11):

$$\varphi_S(t)=M_N\left(\log\varphi_X(t)\right). \tag{4.41}$$

In the case that N has a Poisson (λ) distribution, this equation becomes

$$\varphi_S(t)=\exp\left(\lambda\varphi_X(t)-\lambda\right). \tag{4.42}$$

---

[44] $\Im(a+ib)$ denotes the imaginary part of the complex number $a+ib$: $\Im(a+ib)=b$ and $|a+ib|=\sqrt{a^2+b^2}$. Also, $e^{i\theta}$ can be expressed as a complex number of the form $a+ib$ by means of Euler's Formula: $e^{i\theta}=\cos\theta+i\sin\theta$. Note that if F is continuous at x, then (1/2) $(F(x+)+F(x-))=F(x)$. For a derivation of inversion formula (4.39) consult, for example, Parzen [18], pp. 400–413.

[45] Heckman and Meyers [5].

On the other hand, if N has a negative binomial distribution with mean λ and contagion parameter γ ($\gamma \neq 0$), then

$$\varphi_S(t) = \left(1 + \lambda\gamma - \lambda\gamma\,\varphi_X(t)\right)^{-1/\gamma}. \tag{4.43}$$

Example 4.6. The characteristic function for the gamma (α, β) random variable is $\varphi(t) = (1 - i\beta t)^{-\alpha}$. Therefore, the aggregate variable S with a Poisson-distributed claim count with mean λ and a gamma (α, β) claim-size distribution has characteristic function $\varphi_S(t) = \exp(\lambda(1 - i\beta t)^{-\alpha} - \lambda)$. ■

The Heckman–Meyers algorithm begins with the definition of a piecewise-linear approximation to the cumulative distribution function $F_X(x)$ for claim-size variable X. As we shall soon discover, this approach gives rise to a computationally tractable characteristic function for S. We start by assuming that $F_X(x)$ is continuous on an interval $0 < x < l$. The Heckman-Meyers algorithm accordingly assumes that X is censored at l. If one must use an uncensored variable, choose l large enough so that $1 - F_X(l)$ is negligibly small. After partitioning the closed interval $[0, l]$ into m subintervals

$$0 = c_0 < c_1 < \cdots < c_{m-1} < c_m = l,$$

we approximate $F_X(x)$ by a piecewise-linear function $\hat{F}_X(x)$ with nodes at the points[46]

$$\left(c_k, F_X(c_k)\right), \quad k = 0, 1, 2, \ldots, m.$$

That is, the graph of $y = \hat{F}_X(x)$ on $[0, l]$ is a continuous polygonal curve connecting the endpoints $(0, F_X(0))$ and $(l, F_X(l))$. The associated probability density function $\hat{f}_X(x)$ is a step function—that is, $\hat{f}_X(x)$ is piecewise constant on $[0, l]$, with the sequence of constants defined by

$$\delta_k = \frac{F_X(c_k) - F_X(c_{k-1})}{c_k - c_{k-1}}, \quad k = 1, 2, \ldots, m.$$

Consequently, the characteristic function associated with the approximating distribution function $\hat{F}_X(x)$ is

$$\hat{\varphi}_X(t) = \sum_{k=1}^{m} \int_{c_{k-1}}^{c_k} \delta_k e^{itx}\,dx + \left(1 - F_X(c_m)\right) e^{ic_m t}$$

$$= \sum_{k=1}^{m} \int_{c_{k-1}}^{c_k} \delta_k (\cos tx + i \sin tx)\,dx + \left(1 - F_X(c_m)\right) e^{ic_m t}$$

[46] To improve the approximation to $F_X(x)$ it is advantageous to use a nonregular partition, with partition points closer together nearer $x = 0$, where the graph of $y = F(x)$ is steeper, and farther apart nearer $x = l$, where the graph is flatter. For example, the formula $c_k = \exp(\log(l)k/m)$ for $1 \leq k \leq m$ often works well—see Example 4.7.

$$= \frac{1}{t}\sum_{k=1}^{m}\delta_k\left(\sin c_k t - \sin c_{k-1}t\right) + \left(1 - F_X(c_m)\right)\cos c_m t$$

$$+ \frac{i}{t}\sum_{k=1}^{m}\delta_k\left(\cos c_{k-1}t - \cos\ c_k t\right) + i\left(1 - F_X(c_m)\right)\sin c_m t$$

$$= A(t) + iB(t),$$

where $A(t)$ and $B(t)$ denote the real and imaginary parts of $\hat{\varphi}_X(t)$, respectively. Note that a discrete lump of probability of size $1 - F_X(l)$ has been incorporated at the upper limit $l = c_m$.

We can now use formulas (4.42) and (4.43) to develop the characteristic function for the approximating aggregate distribution. In the Poisson case

$$\hat{\varphi}_S(t) = \exp\left(\lambda A(t) + i\lambda B(t) - \lambda\right) = R(t)e^{i\theta(t)},$$

where $R(t) = e^{\lambda A(t) - \lambda}$ and $\theta(t) = \lambda B(t)$. In the negative binomial case the function is

$$\hat{\varphi}_S(t) = \left(1 + \lambda\gamma - \lambda\gamma\left(A(t) + iB(t)\right)\right)^{-1/\gamma} = R(t)e^{i\theta(t)},$$

with

$$R(t) = \sqrt{\left(\left|1 + \lambda\gamma - \lambda\gamma\ A(t)\right|^2 + \left|\lambda\gamma\ B(t)\right|^2\right)^{-1/\gamma}} \text{ and}$$

$$\theta(t) = -\frac{1}{\gamma}\arctan\left(\frac{\lambda\gamma B(t)}{1 + \lambda\gamma - \lambda\gamma\ A(t)}\right).$$

The cumulative distribution function of S is recoverable from $\hat{\varphi}_S(t)$ by means of inversion formula (4.39):

$$F_S(s) \approx \hat{F}_S(s) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\Im\left(e^{-its}\,\hat{\varphi}_S(t)\right)}{t}\,dt$$

$$= \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\Im\left(e^{-its}R(t)e^{i\theta(t)}\right)}{t}\,dt$$

$$= \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{R(t)}{t}\sin\left(st - \theta(t)\right)dt, \qquad (4.44)$$

whenever $F_S(s)$ is continuous at s. In their paper [5], Heckman and Meyers use numerical integration to evaluate formula (4.44) at the required values of s.

Example 4.7. An application of the Heckman–Meyers algorithm to the aggregate distribution of Example 4.2 yields the results shown in Table 4.7.[47] Here the

[47] These results were obtained from an implementation of the Heckman–Meyers algorithm in a Microsoft Excel workbook created by the author.

Table 4.7. Heckman–Meyers Approximation [Example 4.7]

Amount s	$F(s)$	$\hat{F}(s)$	Relative Error
0	0.0821	0.0412	–49.82%
500	0.1096	0.1094	–0.15%
1,000	0.1867	0.1869	+0.10%
2,000	0.3755	0.3757	+0.06%
3,000	0.5613	0.5614	+0.01%
4,000	0.7152	0.7151	–0.01%
5,000	0.8273	0.8271	–0.02%
6,000	0.9013	0.9011	–0.02%
7,000	0.9465	0.9463	–0.02%
8,000	0.9723	0.9722	–0.01%
9,000	0.9863	0.9862	–0.01%
10,000	0.9934	0.9934	0.00%

basic claim-size interval of definition is taken as [0; 20,000], with partition points $c_k = \exp(\log(20{,}000)k/256)$, $k = 1, 2, \ldots, 256$. This choice of partition improves the accuracy of the approximation by placing the points closer together at the left end of the interval and farther apart at the right end, where the distribution is flatter. The approximation is highly accurate, except at the single point $s = 0$. At this exceptional point there is a discrete lump of probability, the probability of $N = 0$ claims. Such points of discrete probability give rise to jump discontinuities in the function $F_S(s)$, as discussed in the next section. ■

4.6. Discontinuities

When a generally continuous claim-size distribution has a nonzero probability amount at a positive singular point ξ, the corresponding aggregate distribution function has a jump discontinuity at all multiples of ξ. This phenomenon is always present when the continuous claim-size variable X is censored at a limit value l. The distribution of the modified variable has a discrete lump of probability in the amount of $1 - F_X(l)$ at $x = l$ and an aggregate distribution function based on the modified distribution would then have jump discontinuities at all positive integer multiples of l.

The size of the jump discontinuity in the aggregate distribution at $s = kl$, the k^{th} multiple of the claim limit l, is given by

$$f_N(k) \cdot (1 - F_X(l))^k, \quad k = 1, 2, 3, \ldots. \tag{4.45}$$

In situations where $E[N]$ is fairly large, the probability $f_N(k)$—and therefore the size of the discontinuity at k—is comfortably small. When this occurs the error of

approximation by a continuous function is negligible. On the other hand, when the expected number of claims is small, then the techniques discussed in Sections 4.3 and 4.5 can fail to provide a reasonable approximation at or near such a point of discontinuity. The next example illustrates this situation.

Example 4.8. Consider a gamma-distributed claim-size variable X with $(\alpha, \beta) = (2.5, 500)$. The unlimited mean is 1,250, but the distribution limited at $l = 2{,}000$ has a mean of 1,147. The limited distribution has a single discrete amount of probability of size 0.1562 at $l = 2{,}000$.

Compounding this claim-size variable with a Poisson claim-count variable with mean $\lambda = 1.308$ yields an aggregate random variable S with mean 1,500 = (1.308)(1,147). The aggregate distribution function F_S will then have a jump discontinuity at $s = 2{,}000$, the size of which is given by (4.45) with $k = 1$:

$$(1.308)\left(e^{-1.308}\right)(0.1562) = 0.0552.$$

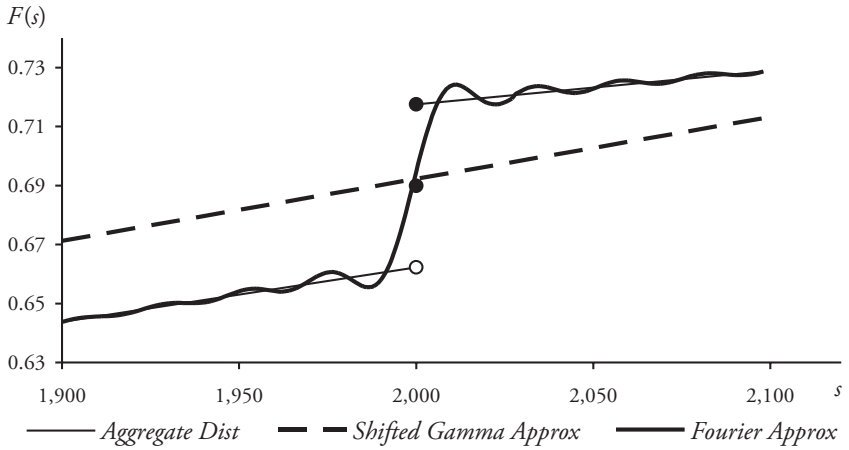
Approximating the aggregate distribution function by the shifted gamma approximation (4.16) yields the continuous function shown in the graph of Figure 4.3 as the dashed curve. This approximation has considerable error throughout a fairly broad interval about the discontinuity at $s = 2{,}000$.

By way of contrast, the Fourier approximation (4.44) based on the Heckman-Meyers algorithm does a better job of approximating the function near the singular point, but as a continuous function it also has difficulty at the point itself. This approximation is shown in Figure 4.3 as the solid curve. Note that the graph of the Heckman–Meyers function passes through the midpoint of the jump at $s = 2{,}000$. ■

4.7. Simulation

Methods for calculating or approximating aggregate loss distributions discussed in the previous sections all involve deterministic models—that is, numbers associated with the approximating distribution are all calculable from definite algorithms and functional

Figure 4.3. Approximations to *F*(*s*) at a Point of Discontinuity [Example 4.8]



formulas. In this section we turn to another classic approach to the problem—the method of distribution ***simulation***, often called ***Monte Carlo simulation*** in reference to its stochastic basis.

The simulation technique is conceptually simple and straightforward: first (*i*) generate a large random sample of selections from the parametric distribution in question, and then (*ii*) create the discrete distribution for this sample, a distribution which can be a useful approximation to the original parametric distribution. In the case of stochastic simulation, however, there is no empirical population of data from which to select a random sample. The sample points must be generated, either from a table of random numbers or by means of a computer random number generator. Such computer software packages—more accurately characterized as pseudo random number generators—typically generate numbers uniformly distributed between 0 and 1.

At the heart of the simulation method lies the following theorem, used to transform a number u randomly selected from a uniform distribution on the interval $0 < u < 1$ to a random value of a variable with a specified distribution. Thus, if F is the distribution function of random variable Y and u is a number randomly generated from the interval $0 < u < 1$, then $\tilde{F}^{-1}(u)$—where $\tilde{F}^{-1}$ is the generalized inverse function defined at (4.46)—is a randomly generated value of variable Y.

Assume that $F(y) = \Pr\{Y \le y\}$ is the cumulative distribution function for random variable Y. The generalized inverse function $\tilde{F}^{-1}$ is defined for each u in the open interval $(0, 1)$ by

$$\tilde{F}^{-1}(u) = \min\{\xi : u \le F(\xi)\}. \tag{4.46}$$

If random variable U is uniformly distributed on the interval $(0, 1)$, then random variable $\tilde{F}^{-1}(U)$ is identical to Y: $\tilde{F}^{-1}(U) = Y$.

Proof: Observe first that $\tilde{F}^{-1}$ has the following properties:

$$u \le F\left(\tilde{F}^{-1}(u)\right) \text{ for all } u \text{ in} (0,1), \tag{4.47}$$

$$\tilde{F}^{-1}(F(y)) \le y \text{ for all real } y, \tag{4.48}$$

$$\tilde{F}^{-1}(u) \text{ is a nondecreasing function of } u, \tag{4.49}$$

(refer to Problem 4.19). The theorem will be established if we can show that for all real y

$$\{u : \tilde{F}^{-1}(u) \le y\} = \{u : u \le F(y)\}. \tag{4.50}$$

Then, if random variable U is uniformly distributed on (0, 1), equation (4.50) implies that

$$\Pr\{F^{-1}(U) \le y\} = \Pr\{U \le F(y)\} = F(y), \quad -\infty < y < \infty.$$

As a result, random variables $\tilde{F}^{-1}(U)$ and Y have the same cumulative distribution function $F(y)$, and so they must be identical: $\tilde{F}^{-1}(U) = Y$.

To prove equation (4.50), assume that y is a fixed real number. First select u in (0, 1) so that $\tilde{F}^{-1}(u) \le y$. Because of (4.47) and the fact that F is a nondecreasing function

$$u \le F\left(\tilde{F}^{-1}(u)\right) \le F(y). \tag{4.51}$$

Conversely, suppose that $u \le F(y)$. Properties (4.48) and 4.49) imply that

$$\tilde{F}^{-1}(u) \le \tilde{F}^{-1}(F(y)) \le y. \tag{4.52}$$

Combining (4.51) and (4.52) yields the desired result. ■

Example 4.9. Suppose that X has the claim-size distribution of Example 4.1, with cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 100 \\ 0.40 & \text{if } 100 \le x < 200 \\ 0.90 & \text{if } 200 \le x < 300 \\ 1.00 & \text{if } 300 \le x < \infty. \end{cases}$$

(***a***) The inverse (4.46) is therefore given by

$$\tilde{F}^{-1}(u) = \begin{cases} 100 & \text{if } 0 < u \le 0.40 \\ 200 & \text{if } 0.40 < u \le 0.90 \\ 300 & \text{if } 0.90 < u < 1.00. \end{cases} \tag{4.53}$$

If U is uniformly distributed on the interval (0, 1), then $\tilde{F}^{-1}(U)$ takes on three possible values—100, 200, 300—with probabilities

$$\Pr\left\{\tilde{F}^{-1}(U) = 100\right\} = 0.40 - 0 = 0.40,$$

$$\Pr\left\{\tilde{F}^{-1}(U) = 200\right\} = 0.90 - 0.40 = 0.50,$$

$$\Pr\left\{\tilde{F}^{-1}(U) = 300\right\} = 1.00 - 0.90 = 0.10.$$

This verifies, of course, that random variables $\tilde{F}^{-1}(U)$ and X are identical.

(***b***) Suppose now that three trials of the random generation process are performed, generating random numbers $u_1 = 0.4547$, $u_2 = 0.9236$, and $u_3 = 0.2573$. The corresponding random values of X are obtained from formula (4.53): $x_1 = \tilde{F}^{-1}(u_1) = 200$, $x_2 = \tilde{F}^{-1}(u_2) = 300$, and $x_3 = \tilde{F}^{-1}(u_3) = 200$. ■

The next three examples illustrate methods for generating random values of commonly encountered claim-size and claim-count random variables.

Example 4.10. (***a***) Variable X_1 is exponentially distributed, with $F_1(x) = 1 - e^{-x/\beta}$ for $0 < x < \infty$. Because F_1 is strictly increasing for positive x, the function is invertible

Table 4.8. Random Values for Claim-Size Distributions [Example. 4.10(*d*)]

Trial	Uniform u	Exponential x_1	Pareto x_2	Std Normal z	Lognormal x_3
(1)	0.1854	410	216	–0.8950	18
(2)	0.3038	724	397	–0.5135	44
(3)	0.5498	1,596	981	0.1252	189
(4)	0.7953	3,172	2,420	0.8249	947
(5)	0.9774	7,580	11,304	2.0028	14,221

there in the ordinary sense, and the inverse defined in (4.46) is identical to the conventional inverse function:

$$x = F_1^{-1}(u) = -\beta \log(1-u), \quad 0 < u < 1.$$

Alternatively, because $1 - U$ is also distributed uniformly on the interval (0, 1), one can generate a random value for x by the equation $x = -\beta \log u$, $0 < u < 1$.

(***b***) Similarly, when random variable X_2 has a shifted Pareto distribution with distribution function $F_2(x) = 1 - (\beta/(x+\beta))^{\alpha}$ for $0 < x < \infty$, the inverse function is given by

$$x = F_2^{-1}(u) = \beta\left((1-u)^{-1/\alpha} - 1\right), \quad 0 < u < 1.$$

(***c***) Random variable X_3 has a lognormal distribution with parameters (μ, σ). Thus, if z is a randomly generated value of the standard normal distribution,[48] then $x = \exp(\mu + \sigma z)$ is a random value for X_3.

(***d***) Five values of u were randomly generated from the uniform distribution on the interval (0, 1). Corresponding random values for X_1 when $\beta = 2{,}000$, for X_2 when $(\alpha, \beta) = (2; 2{,}000)$, for standard normal Z, and for X_3 when $(\mu, \sigma) = (4.956, 2.3)$ are displayed in Table 4.8. ■

Example 4.11. (***a***) Variable X has a gamma distribution with probability density function

$$f(x) = \frac{1}{\beta^n \Gamma(n)} x^{n-1} e^{-x/\beta} \qquad (n = 1, 2, 3, \ldots, \beta > 0), \quad 0 < x < \infty.$$

The reproductive property of the gamma distribution implies that X has the same distribution as the sum of n independent random variables X_i, each with the exponential distribution with parameter β (refer to Section 2.3). Thus, to generate a random value

[48] Users of Microsoft Excel find the composite of two worksheet functions NORM.S.INV(RAND) convenient for generating random values of the standard normal variable Z. Refer to Appendix A.1 and also to Problem 4.26.

for X, generate values for each of the identical random variables X_i—the sum of these values is a random value for X:

$$x = x_1 + x_1 + \cdots + x_n.$$

(***b***) Variable X has a gamma distribution with $(\alpha, \beta) = (3, 400)$. Three random values are generated from the uniform distribution on (0, 1): 0.5349, 0.8762, 0.2009 (three different random values of the uniform distribution are required because the X_i must be independent). Thus, we have a random value for X:

$$\begin{aligned} x &= -(400)\log(1-0.5349) - (400)\log(1-0.8762) - (400)\log(1-0.2009) \\ &= 1{,}232. \end{aligned}$$

Example 4.12. (***a***) Variable N has a Poisson distribution with mean $\lambda = \#$ *claims per unit time*:

$$f_N(n) = \frac{\lambda^n e^{-\lambda}}{n!}, \quad n = 0, 1, 2, \ldots.$$

The random variable $\hat{T}_n$, where $\hat{T}_1 =$ *occurrence time of the first claim* and $\hat{T}_n =$ *time between the occurrence of the* $(n-1)^{st}$ *and the* n^{th} *claim* $(n > 1)$, has an exponential distribution with parameter $\beta = 1/\lambda$ [refer to Problem 3.33(*c*)]. Thus, the event that $N = n$ in a unit of time is equivalent to

$$\sum_{i=1}^{n} \hat{T}_i \le 1 < \sum_{i=1}^{n+1} \hat{T}_i. \tag{4.54}$$

If $\{u_i\}$ are values of the random variable U, uniformly distributed on the interval (0, 1), then by the result of Example 4.10(*a*) above, we have corresponding values of $\hat{T}_i$: $t_i = -(1/\lambda)\log u_i$. As a result, inequality (4.54) is satisfied whenever

$$\sum_{i=1}^{n} -(1/\lambda)\log u_i \le 1 < \sum_{i=1}^{n+1} -(1/\lambda)\log u_i.$$

After multiplying by -1 and applying the exponential function, we obtain the equivalent inequality

$$\prod_{i=1}^{n} u_i \ge e^{-\lambda} > \prod_{i=1}^{n+1} u_i. \tag{4.55}$$

Inequality (4.55) can now be used to generate a random value for N in the following way, a method easily programmed for computer implementation:

(*i*) Assume that $\langle u_i \rangle$, $i = 1, 2, 3, \ldots$, is a sequence of random values generated from the uniform distribution on the interval (0, 1).
(*ii*) If $u_1 < e^{-\lambda}$, then stop and set $n = 0$.

(*iii*) Otherwise, if $u_1u_2 < e^{-\lambda}$, then stop and set $n = 1$.
(*iv*) Otherwise, if $u_1u_2u_3 < e^{-\lambda}$, then stop and set $n = 2$.
(*v*) Otherwise, if $u_1u_2u_3u_4 < e^{-\lambda}$, then stop and set $n = 3$.
...

Continue in this way until (4.55) is satisfied by $i = m$:

$$\prod_{i=1}^{m+1} u_i < e^{-\lambda} \leq \prod_{i=1}^{m} u_i;$$

then stop and set $n = m$.

(***b***) Variable N has a Poisson distribution with mean $E[N] = \lambda = 1.500$. Successive products of numbers randomly generated from the uniform distribution on the interval (0, 1) are compared to $e^{-1.500} = 0.2231$ according to the procedure developed in part (*a*) above. Corresponding random values of N are then calculated, and the results of four such trials are displayed in Table 4.9. ■

The next two examples illustrate how Monte Carlo simulation methods can be used to generate random values of a compound aggregate loss random variable. To generate a single such value, one must first generate a random value for the claim-count variable N, say $N = n$, and then generate n values for the claim-size variable X. The sum of these claim-size amounts is a random value for the aggregate loss variable S.

Example 4.13. (***a***) For the aggregate variable S the claim-count N is Poisson-distributed with mean $\lambda = 1.500$. Claim-size X has a lognormal distribution with parameters $(\mu, \sigma) = (4.956, 2.300)$. Therefore, S has mean

$$E[S] = E[N]E[X] = (1.500)\exp\left(4.956 + \tfrac{1}{2}(2.300)^2\right) = 3{,}000.$$

For each random value n obtained for N we generate n random values for X, the sum of which is a random value for S. Table 4.10 displays the results of this procedure based on the four values for n generated in Example 4.12(*b*).

(***b***) One distinct advantage that Monte Carlo simulation has over other methods of approximating an aggregate loss distribution is the fact that it is easy to model various

Table 4.9. Random Values for Poisson Distribution [Example 4.12(*b*)]

Trial	u_1	u_2	u_3	u_4	Πu_i	$e^{-1.500}$	n
(1)	0.6791	0.7543	0.2391		0.1225	0.2231	2
(2)	0.1047				0.1047	0.2231	0
(3)	0.7591	0.4746	0.7205	0.3256	0.0845	0.2231	3
(4)	0.5029	0.2874			0.1445	0.2231	1

Table 4.10. Random Values for Aggregate-Loss Distribution [Example 4.13]

		Example 4.13(*a*)				Example 4.13(*b*)	
Trial	*n*	*u*	*z*	*x*	*s*	$\tilde{x}$	$\tilde{s}$
(1)	2	0.2871	–0.5619	39	—	39	—
		0.8945	1.2508	2,522	2,561	1,000	1,039
(2)	0	—	—	—	0	—	0
(3)	3	0.7387	0.6393	618	—	618	—
		0.3766	–0.3144	69	—	69	—
		0.9411	1.5641	5,185	5,872	1,000	1,687
(4)	1	0.6982	0.5192	469	469	469	469

policy conditions imposed on the size of claims—including complex deductible and/or limit restrictions. As an example, consider the imposition of a $1,000 policy limit on the size of claims in the claim process described in part (*a*) above. The results are shown in Table 4.10, where the limit has been imposed on each random claim size as it is generated, yielding the modified random values $\tilde{x}$ and the associated modified aggregate loss amounts $\tilde{s}$. ■

Example 4.14. Return now to the aggregate distribution of Example 4.2, for which N has a Poisson distribution with $\lambda = 2.5$ and X is gamma-distributed with parameters $(\alpha, \beta) = (3, 400)$. We generate 10,000 values of N—and for each of these a corresponding aggregate loss amount, thus creating a randomly generated sample of size 10,000. The resulting sample cumulative distribution function is an approximation to the aggregate distribution function $F(s)$. For example, 5,599 sample points have aggregate loss amounts less than or equal 3,000, so that

$$F_{10K}(3{,}000) = \frac{5{,}599}{10{,}000} = 0.5599 \approx F_S(3{,}000).$$

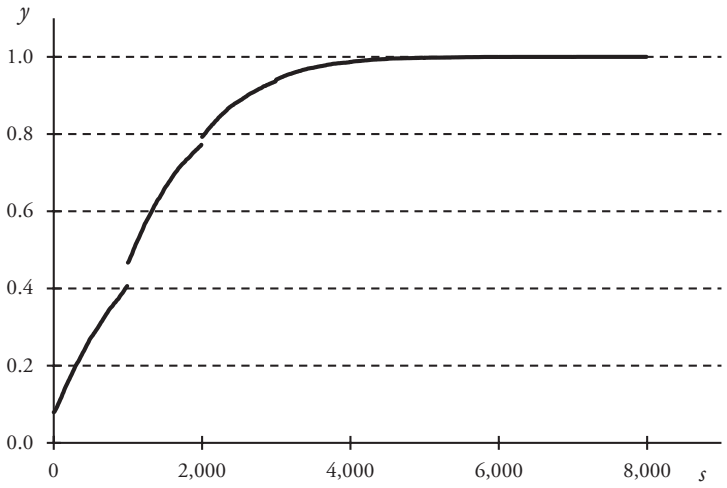
This compares favorably with the actual value of $F(3{,}000) = 0.5613$. Values of the cumulative distribution $F_{10K}(s)$ based on the generated sample are shown in Table 4.11 along with the exact values of $F(s)$. ■

Example 4.15. Consider the aggregate random variable S for which the claim count N is Poisson-distributed with $\lambda = 3$ and claim size X has a lognormal distribution with $(\mu, \sigma) = (6, 1.5)$. Moreover, claim size is limited by a policy limit of 1,000. As before, a sample of 10,000 random trials is generated, and the resulting aggregate distribution function created. A graph of $y = F_{10K}(s)$ is displayed in Figure 4.4, in which the discontinuity at multiples of the 1,000 limit is clearly evident. ■

Table 4.11. Approximation by Monte Carlo Simulation [Example 4.14]

Amount s	$F(s)$	$F_{10K}(s)$	Relative Error
0	0.0821	0.0832	+1.34%
500	0.1096	0.1084	–1.09%
1,000	0.1867	0.1874	+0.37%
2,000	0.3755	0.3796	+1.09%
3,000	0.5613	0.5599	–0.25%
4,000	0.7152	0.7130	–0.31%
5,000	0.8273	0.8342	+0.83%
6,000	0.9013	0.9042	+0.32%
7,000	0.9465	0.9468	+0.03%
8,000	0.9723	0.9719	–0.04%
9,000	0.9863	0.9865	+0.02%
10,000	0.9934	0.9935	+0.01%

Figure 4.4. Cumulative Distribution Function $y = F_{10K}(s)$ [Example 4.15]



4.8. Problems

4.1 Construct the discrete aggregate loss distribution based on these distributions for N and X.

Claim Count *N*

Count n	$f_N(n)$
0	0.20
1	0.40
2	0.25
3	0.15

Claim Size *X*

Size x	$f_x(x)$
500	0.10
1,000	0.40
1,500	0.30
2,000	0.20

4.2 Assume that $Y = X_1 + X_2$, where X_1 and X_2 are continuous, independent (not necessarily claim-size) random variables.

(***a***) X_1 and X_2 have respective probability density functions f_1 and f_2. Prove that

$$f_Y(y) = \int_{-\infty}^{\infty} f_1(y-x)\, f_2(x)\, dx.$$

(***b***) Assume now that X_1 and X_2 are identically distributed claim-size random variables, with common distribution function F and $F(x) = 0$ for $x < 0$. Show that F_Y can be expressed as

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \int_0^y F(y-x)\, dF(x) & \text{if } y \geq 0. \end{cases}$$

4.3 Verify that the recursion formula (4.2) yields $F_1^*(y) = F(y)$ for all y.

4.4 In each of the following cases construct a formula for $F_S(s)$ in terms of $f_N(n) = \Pr\{N = n\}$ and $F_X(x)$, the c.d.f. for X.

(***a***) $f_N(n) = 0$ for $n > 1$. (***b***) $f_N(n) = 0$ for $n > 2$.

4.5 Derive formulas (4.6) and (4.7) for $E[S^2]$ and $E[S^3]$ from the compound moment-generating function (4.11).

4.6 N is Poisson-distributed with mean λ, and X has an exponential (β) distribution. Derive explicit formulas for the aggregate distribution functions $f_S(s)$ and $F_S(s)$.

4.7 N is Poisson-distributed with mean $\lambda = 8$, and X has a gamma distribution with $\alpha = 0.2000$ and $\beta = 3{,}750$. Calculate the indicated values for $F_S(s)$.

Amount s	$F_S(s)$
0	_____
3,000	_____
6,000	_____
9,000	_____
12,000	_____
15,000	_____
18,000	_____
21,000	_____
24,000	_____
27,000	_____

4.8 λ and γ are the claim-count mean and contagion parameter, respectively, for an aggregate loss variable S. Prove that for fixed γ:

(***a***) $\lim_{\lambda\to\infty} CV[S] = \sqrt{\gamma}$. (***b***) $\lim_{\lambda\to\infty} Sk[S] = 2\sqrt{\gamma}$.

4.9 Verify the normal power inversion formula (4.19):

$$Q^{-1}\left((S-\mu)/\sigma\right) = T_{NP}(S),$$

where $T_{NP}(S)$ is given by (4.17).

4.10 Provide detailed derivations of the Wilson–Hilferty transformation formulas (4.24) and (4.25).

4.11 Derive from the Wilson–Hilferty chi-square approximation (4.22) a formula for $\chi^2_{0.95}(m)$, the 95^{th} percentile of the chi-square distribution with m degrees of freedom.

4.12 Use the formula obtained in Problem 4.11 to estimate the chi-square percentiles in the following table.

d.f. m	$\chi^2_{0.95}(m)$	Wilson–Hilferty	Relative Error
5	11.070	_____	_____%
10	18.307	_____	_____%
15	24.996	_____	_____%
20	31.410	_____	_____%
25	37.652	_____	_____%
30	43.773	_____	_____%

4.13 Tabulate the following approximations to the Poisson/gamma distribution function of Problem 4.7.

Amount s	$F_S(s)$	Normal	Relative Error	Normal Power	Relative Error	Shifted Gamma	Relative Error	Wilson-Hilferty	Relative Error
0	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
3,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
6,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
9,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
12,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
15,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
18,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
21,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
24,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%
27,000	_____	_____	_____%	_____	_____%	_____	_____%	_____	_____%

4.14 λ and γ are, respectively, the claim-count mean and contagion parameter for an aggregate loss variable S. Verify the following special cases of formula (4.31) for $f(0)$.

(***a***) If N is Poisson-distributed, then $f_S(0) = e^{\lambda g(0)-\lambda}$.

(***b***) If N has a negative binomial distribution, for which $\gamma \neq 0$, then

$$f_S(0) = (1+\gamma\lambda-\gamma\lambda\, g(0))^{-1/\gamma}.$$

4.15 Show that recursion formula (4.33) can be expressed in the following form, where $\hat{m}$ is defined by (4.28):

$$f_S(m) = \frac{1}{1-ag(0)} \sum_{k=1}^{\min\{m,\hat{m}\}} \left(a+\frac{b}{m}k\right) g(k)\, f_S(m-k), \qquad m = 1, 2, 3, \ldots.$$

4.16 Use the recursion method of Section 4.4 to calculate the cumulative distribution function of the aggregate random variable for which claim size X has the discrete distribution of Example 4.1 and N has a Poisson distribution with $\lambda = 1.35$.

4.17 Verify that the midpoint formulas of (4.37) actually define a discrete probability function.

4.18 Random variable U is uniformly distributed on the interval (0, 1).

(***a***) Show that the characteristic function of U is $\varphi_U(t) = (e^{it} - 1)/(it)$.

(***b***) Use formula (4.39) to recover $F_U(x)$ from $\varphi_U(t)$.

4.19 Let F be the cumulative distribution function for random variable Y. Prove these statements about the inverse function $\tilde{F}^{-1}$ defined by (4.46).

(***a***) $\tilde{F}^{-1}$ exists for all variables Y.

(***b***) $\tilde{F}^{-1}(u)$ is a nondecreasing function of u.

(***c***) $u \le F(\tilde{F}^{-1}(u))$ for all u in (0, 1).

(***d***) $\tilde{F}^{-1}(F(y)) \le y$ for all real y.

(***e***) Show by example that it is possible for the inequalities in (*c*) and (*d*) to be *strictly less than.*

(***f***) If F is strictly increasing, then $\tilde{F}^{-1}$ is the usual inverse of function F.

4.20 F is a strictly increasing cumulative distribution function for continuous random variable X. Prove: random variable $F(X)$ is uniformly distributed on the unit interval (0, 1).

4.21 Calculate the inverse function $\tilde{F}^{-1}(u)$ in the case that $F(x)$ is a Weibull distribution function (2.61).

4.22 Consider the following sequence of random selections from the uniform distribution on the interval (0, 1):

$$\langle 0.4695, 0.2871, 0.7527, 0.9106, 0.5538, 0.1189, 0.8853 \rangle.$$

Calculate the random value for N with a Poisson distribution ($\lambda = 3$) that is implied by the sequence.

4.23 Five random values of U, uniformly distributed on the interval (1,0), are shown in the table. Calculate corresponding random values for X_1 (exponential with $\beta = 2{,}000$), for X_2 (Pareto with $(\alpha, \beta) = (2.5; 3{,}000)$), for X_3 (lognormal with $(\mu, \sigma) = (5.181, 2.2)$), and for X_4 (Weibull with $(\beta, \delta) = (1{,}000; 0.5)$).

Trial	Uniform u	Exponential x_1	Pareto x_2	Lognormal x_3	Weibull x_4
(1)	0.2097	_____	_____	_____	_____
(2)	0.3562	_____	_____	_____	_____
(3)	0.6970	_____	_____	_____	_____
(4)	0.8245	_____	_____	_____	_____
(5)	0.9882	_____	_____	_____	_____

4.24 (***a***) Random variable N has a geometric distribution, with

$$f(n) = p(1-p)^n \quad (0 < p < 1), \quad n = 0, 1, 2, \ldots.$$

Show that random values of N can be generated by the formula

$$n = [\![(\log(1-u))/(\log(1-p)) - 1]\!],$$

where $[\![x]\!]$ denotes the greatest integer function. [*Hint*: the cumulative distribution function at positive integer n is

$$F_N(n) = 1 - (1-p)^{n+1}.$$

(***b***) Use moment-generating functions to show that the sum of m identical independent random variables, each distributed with a geometric distribution with parameter p has the special negative binomial distribution with probability density function

$$f(n) = \binom{m+n-1}{n} p^m (1-p)^n \quad (m = 1, 2, 3, \ldots, 0 < p < 1), \quad n = 0, 1, 2, \ldots.$$

(***c***) Describe a method for generating random values of a random variable with the negative binomial distribution defined in part (*b*).

4.25 (***a***) Random variables $\langle U_n \rangle$ (n = 1, 2, . . . , 12) are independent and uniformly distributed on the interval (0, 1). Show that the distribution of $X = \sum_{n=1}^{12} U_n - 6$ is approximately standard normal.

(***b***) Use the result of part (*a*) to devise a method of generating random values from a normal distribution with parameters (μ, σ).

(***c***) Use the result of part (*b*) to devise a method of generating random values from a lognormal distribution with parameters (μ, σ).

5. Excess Claims

We investigate in this chapter claim processes in which all claims are restricted to those larger in size than some fixed positive amount—that is, to claims that penetrate an excess layer of insurance. Distributions of such excess losses are critical to the quantification of such common policy provisions as deductibles and to the pricing of successive layers of coverage lying above a first-dollar, or primary, layer of insurance.

5.1. Excess Claim Size

Consider first an unlimited claim-size random variable X and a nonnegative constant a. The random variable Y defined by

$$Y = \begin{cases} 0 & \text{if } 0 \le X \le a \\ X - a & \text{if a} < X < \infty \end{cases}$$

represents the size of claims modified by a policy condition that imposes an underlying limit amount a. Here the insurer pays nothing if the claim size is a or less, and the sizes of all other claims are reduced by a. In this situation a could represent an amount retained by the insured, as in the case of a policy with a deductible, or for an umbrella or excess policy it might be the limit of an underlying primary policy.

The distribution function of variable Y is readily obtained from that of X:

$$F_Y(y) = \Pr\{Y \le y\} = \begin{cases} 0 & \text{if } -\infty < y < 0 \\ F_X(y + a) & \text{if } 0 \le y < \infty. \end{cases}$$

If $E[X]$ exists, then so does $E[Y]$. Moreover,

$$\begin{aligned} E[Y] &= \int_0^\infty y dF_X(y + a) \\ &= \int_a^\infty (u - a) dF_X(u) \\ &= \int_0^\infty u dF_X(u) - \int_0^a u dF_X(u) - a \int_a^\infty dF_X(u) \\ &= E[X] - E[X; a]. \end{aligned} \tag{5.1}$$

Clearly, $E[Y] \le E[X]$ whenever both expected values exist.

For random variable Y the probability that the insurer pays nothing,

$$F_Y(0) = \Pr\{Y = 0\} = \Pr\{X \le a\} = F_X(a),$$

is usually a positive number. However, insurers do not always see, nor are they usually interested in, claims for which $Y = 0$. It is therefore more useful, from an insurer's standpoint, to work with a related variable X_a, defined only for $X > a$:

$$X_a = X - a, \quad a < X < \infty. \tag{5.2}$$

X_a represents the ***excess of X over the limit a***, for which claims of size a or smaller are ignored and all others are reduced by the amount a. Thus modified, variable X is said to be ***truncated from below and shifted by a***. Variable X_a has a distribution function obtained conditionally from that of X—and in this case $F_{X_a}(0) = 0$:

$$F_{X_a}(x) = \Pr\{X - a \le x | X > a\} = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{F_X(x+a) - F_X(a)}{1 - F_X(\,)} & \text{if } 0 \le x < \infty. \end{cases} \tag{5.3}$$

Whenever $E[X]$ exists the expected value of X_a is

$$E[X_a] = \frac{\int_0^\infty x dF_X(x+a)}{1 - F_X(a)} = \frac{E[X] - E[X; a]}{1 - F_X(a)}. \tag{5.4}$$

[Compare this formula with that of (5.1).] Moreover, if all three moments $E[X]$, $E[X^2]$, and $E[X^3]$ exist, then the second and third moments of X_a are, respectively,

$$E[X_a^2] = \frac{E[X^2] - E[X^2; a] - 2a(E[X] - E[X; a])}{1 - F_X(a)}, \tag{5.5}$$

$$E[X_a^3] = \frac{E[X^3] - E[X^3; a] - 3a(E[X^2] - E[X^2; a])}{1 - F_X(a)} + \frac{3a^2(E[X] - E[X; a])}{1 - F_X(a)}. \tag{5.6}$$

The limited expected value of the excess random variable X_a is an obvious combination of limited severities of the unlimited claim-size variable X:

$$E[X_a; l] = \frac{\int_0^l x dF_X(x+a)}{1 - F_X(a)} + l\left(1 - \frac{F_X(l+a) - F_X(a)}{1 - F_X(a)}\right)$$

$$= \frac{\int_a^{a+l} (u-a) dF_X(u)}{1 - F_X(a)} + l\left(\frac{1 - F_X(a+l)}{1 - F_X(a)}\right)$$

$$= \frac{E[X; a+l] - E[X; a] - (a+l)(1-F_X(a+l)) + a(1-F_X(a))}{1-F_X(a)}$$

$$+ \frac{l(1-F_X(a+l)) - a(F_X(a+l) - F_X(a))}{1-F_X(a)}$$

$$= \frac{E[X; a+l] - E[X; a]}{1-F_X(a)}. \tag{5.7}$$

Example 5.1. Claim-size random variable X has an exponential distribution with mean β:

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1-e^{-x/\beta} & \text{if } 0 \le x < \infty. \end{cases}$$

For the exponential distribution family it is evident that the excess c.d.f. is independent of the size of limit a:

$$F_{X_a}(x) = \frac{\left(1-e^{-(x+a)/\beta}\right) - \left(1-e^{-a/\beta}\right)}{e^{-a/\beta}} = 1-e^{-x/\beta}, \quad 0 \le x < \infty.$$

As a consequence, the excess claim size X_a and unlimited claim size X have the *same* distribution. This means that the existence of a deductible or underlying coverage does not affect the distribution of claim size. In particular, $E[X_a] = E[X] = \beta$ for every limit a. ■

Example 5.2. Claim-size variable X has a Pareto distribution with probability density function

$$f_X(x) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}, \quad 0 < x < \infty.$$

Accordingly, the density function for X_d is

$$f_{X_d}(x) = \frac{f_X(x+d)}{1-F_X(d)} = \frac{\alpha\beta^\alpha}{(x+d+\beta)^{\alpha+1}} \Big/ \left(\frac{\beta}{d+\beta}\right)^\alpha = \frac{\alpha(d+\beta)^\alpha}{(x+d+\beta)^{\alpha+1}}, \quad 0 < x < \infty.$$

Hence, X_d is also Pareto-distributed, with parameters $(\alpha, d+\beta)$. The mean $E[X_d]$ exists whenever $\alpha > 1$, and it is an increasing linear function of the lower limit d:

$$E[X_d] = \frac{d+\beta}{\alpha-1}. \blacksquare$$

Example 5.3. The table below displays grouped claim-size data derived from a sample of 300 claims from an unlimited population with an unknown distribution.

Size Group	# Claims
0–5,000	139
5,001–10,000	68
10,001–15,000	32
15,001–20,000	15
20,001–25,000	11
25,001–30,000	8
30,001–35,000	5
35,001–40,000	4
40,001–45,000	4
45,001–48,500	14
Total	300

Before the data were tabulated these claims were censored by a $50,000 policy limit and then subjected to a $1,500 straight deductible. Using the minimum chi-square approach, we wish to find a lognormal distribution function $F_{\mu,\sigma}(x)$ for the population of the unlimited—non-truncated and non-censored—claims.

We begin by defining ten cells with boundaries $c_k = 5{,}000k$ ($k = 0, 1, \ldots, 9$) and $c_{10} = \infty$. The observed cell frequencies are just the tabulated group claim frequencies n_k. In particular, note that $n_{10} = 14$.

The expected cell frequencies $\phi_k(\mu, \sigma)$ are expressed in terms of the (as yet unknown) unlimited and unmodified population lognormal c.d.f. $F_{\mu,\sigma}(x)$. The probability $P_k(\mu, \sigma)$ of a claim being less than or equal c_k is

$$P_k(\mu,\sigma) = \begin{cases} \dfrac{F_{\mu,\sigma}(c_k + 1{,}500) - F_{\mu,\sigma}(1{,}500)}{1 - F_{\mu,\sigma}(1{,}500)} & \text{if } k = 1, 2, \ldots, 9 \\ 1 & \text{if } k = 10. \end{cases}$$

Therefore

$$\phi_k(\mu,\sigma) = (300)\left(P_k(\mu,\sigma) - P_{k-1}(\mu,\sigma)\right).$$

Minimizing the chi-square statistic

$$\chi^2(\mu,\sigma) = \sum_{k=1}^{10} \frac{\left(n_k - \phi_k(\mu,\sigma)\right)^2}{\phi_k(\mu,\sigma)}$$

as a function of μ and σ yields a minimum value of $\chi^2(\hat{\mu}, \hat{\sigma}) = 1.6610$ corresponding to the parameter estimates $(\hat{\mu}, \hat{\sigma}) = (8.67593, 1.18109)$.

Because sample data were truncated by the 1,500 deductible, the number of claims entirely eliminated by the deductible is unknown. However, one can estimate this number by means of $F_{\hat{\mu},\hat{\sigma}}(1{,}500) = 0.1243$:

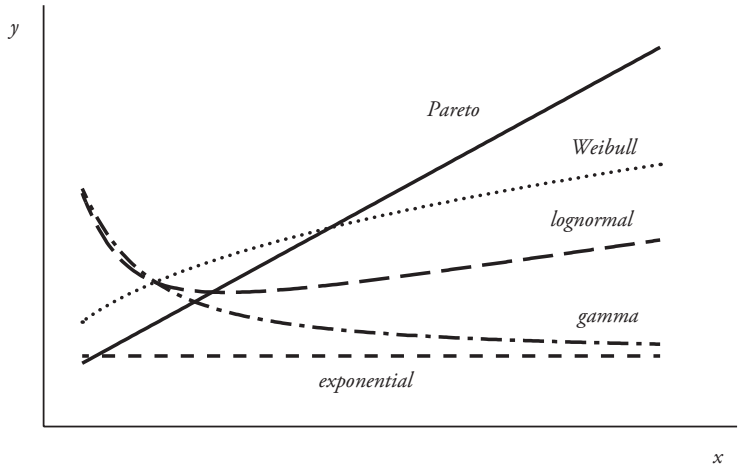
$$\#\,\textit{population claims} \le 1{,}500 \approx \frac{300}{1 - 0.1243}(0.1243) = (343)(0.1243) = 43. \blacksquare$$

5.2. Excess Severity

The expectation $E[X_a]$ obtained in (5.4), with respect to the unlimited random variable X, is called the ***mean excess claim size at a*** or ***excess severity at a***.[49] As with the limited expected value, we can express the mean excess claim size, when it exists,

[49] Illogically in the context of loss distributions, $E[X_a]$ is also known as the ***mean residual life at a***. The term, however, makes sense when the random variable X is a failure-time variable encountered in reliability theory. The expression apparently found its way into actuarial usage because many distributions used by actuaries have also played prominent roles in reliability theory.

Figure 5.1. Characteristic Excess Severity Function Graphs[50]



as a function of the associated limit x. In this context $E[X_x]$ is commonly denoted by $e(x)$—or by $e_X(x)$ when dependence on the random variable X must be indicated:

$$e_X(x) = \frac{E[X] - E[X; x]}{1 - F_X(x)}, \quad 0 < x < \infty. \tag{5.8}$$

The behavior of $e(x)$ for large values of x is characteristic for all distributions in a given parametric family and tends to differ from one such family to another. For example, when X is exponentially distributed, $e(x)$ is a constant function of x, as shown in Example 5.1. Example 5.2 indicates that for Pareto-distributed X with $\alpha > 1$, $e(x)$ is an increasing linear function of x. In the case of the lognormal family, $e(x)$ increases without bound as $x \to \infty$, whereas for gamma-distributed X the function decreases toward a horizontal asymptote as $x \to \infty$. The Weibull $e(x)$ function behaves like a/x^b for some a and b and large values of x. Typical shapes for the graph of $y = e(x)$ are shown in Figure 5.1. Refer to Problem 5.26 for hints on verifying these results.

The asymptotic behavior of $y = e(x)$ is occasionally useful when it comes to fitting a parametric distribution to a set of sample data. The shape of the graph of the sample excess severity function $e_n(x)$ may suggest an appropriate family of distributions. If this graph is approximately linear with positive slope, then a Pareto distribution could be used. If it is nearly constant for large x, a gamma or exponential model would be indicated. Otherwise, if the graph lies between these extremes, then a lognormal or Weibull distribution could be used.

[50] Figure 5.1 is suggested by a similar display in Hogg and Klugman [8], p. 109.

There is, however, a practical restriction in the use of this asymptotic test. The characteristic behavior of $y = e(x)$ becomes apparent only for large x, the region for which sample data is typically the most sparse. It is therefore essential that the claim data contain enough large claims so that $e_n(x)$ can be reliably calculated for sufficiently large values of x.

Example 5.4. The table displays grouped sample claim data for $n = 1{,}000$ policies.

Size Group	# Claims	Total Loss	Severity
0–100	100	6,000	60
101–500	300	95,000	317
501–1,000	240	145,000	604
1,001–2,000	185	260,000	1,405
2,001–4,000	140	450,000	3,214
4,001–5,000	15	66,000	4,400
5,001–10,000	20	150,000	7,500
Total	1,000	1,172,000	1,172

To investigate the behavior of the sample excess severity function for large x, begin by calculating values for the relevant sample statistics at the right-hand endpoints of the group intervals. For example, values of $F_n(2{,}000)$, $E_n[\hat{X};\, 2{,}000]$, and $e_n(2{,}000)$ for the discrete sample variable $\hat{X}$ are, respectively,

$$F_{1000}(2{,}000) = \frac{100 + 300 + 240 + 185}{1{,}000} = 0.8250,$$

$$E_{1000}\left[\hat{X};\, 2{,}000\right] = \frac{6{,}000 + 95{,}000 + 145{,}000 + 260{,}000}{1{,}000} + \frac{(140 + 15 + 20)(2{,}000)}{1{,}000} = 856,$$

$$e_{1000}(2{,}000) = \frac{1{,}172 - 856}{1 - 0.8250} = 1{,}806.$$

The complete set of end-point values is shown in Table 5.1.

The tabulated values of $e_n(x)$ along with a least-squares regression line are displayed graphically in Figure 5.2. It is evident that the sample values are very nearly aligned—the coefficient of determination for the linear regression is $R^2 = 0.9823$. A Pareto model is obviously indicated. To fit such a distribution, observe that the regression function $0.257335x + 1{,}208.50$ can be equated with the Pareto $e(x)$ and the resulting equation solved for parameters α and β:

$$0.257335x + 1{,}208.50 = (x + \beta)/(\alpha - 1).$$

Table 5.1. Sample and Pareto Excess Severity Functions [Example 5.4]

	Sample Distribution (n = 1,000)			Pareto Distribution		
Size x	$F_n(x)$	$E_n[\hat{X}; x]$	$e_n(x)$	$F(x)$	$E[X; x]$	$e(x)$
0	0.0000	0	1,172	0.0000	0	1,208
100	0.1000	96	1,196	0.0978	95	1,234
500	0.4000	401	1,285	0.3900	393	1,337
1,000	0.6400	606	1,572	0.6106	638	1,466
2,000	0.8250	856	1,806	0.8233	904	1,723
4,000	0.9650	1,096	2,171	0.9507	1,098	2,238
5,000	0.9800	1,122	2,500	0.9711	1,136	2,495
10,000	1.0000	1,172	—	0.9962	1,194	3,782

Thus, $(\alpha, \beta) = (4.88599; 4{,}696.22)$. Corresponding end-point values for this Pareto distribution are shown in Table 5.1 for comparison.

The technique of estimating Pareto parameters from the slope and intercept of the regression line seems to work well in this example, but it should be used with some caution. The slope of the regression line is sensitive to the size of the largest claims, and the calculated distribution parameters could be significantly affected by changes in just a few of these numbers. ■

Example 5.5. Figure 5.3 shows the graph of the sample excess severities for the data of Examples 2.6 and 2.7. Superimposed on this graph are the corresponding graphs of $y = e(x)$ for the fitted gamma and lognormal distributions obtained in those examples. The graph of the gamma model reasonably approximates that of the sample function, but the lognormal function diverges significantly from the sample values for $x > 1{,}500$. This suggests that of the two probability distributions obtained in Chapter 2 the gamma might provide the better fit. ■

Figure 5.2. Sample Excess Severities with Regression Line [Example 5.4]

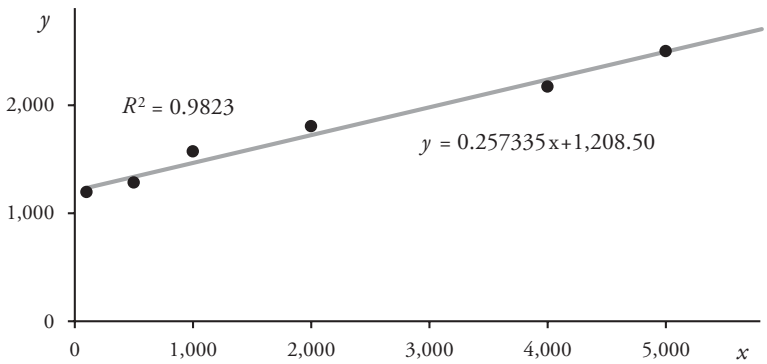
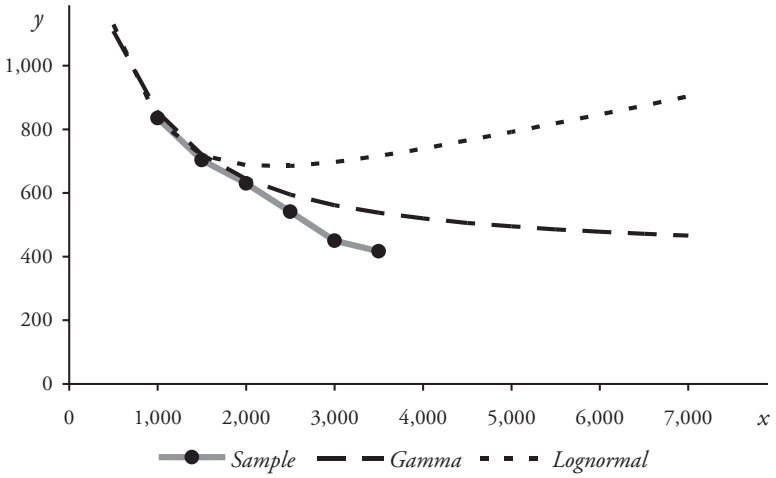


Figure 5.3. Excess Severity Functions [Example 5.5]



5.3. Layers of Coverage

In many situations an insurance policy may impose both an upper limit and a lower limit on the claims subject to the policy. How these are applied depends on the specific policy conditions—for example, on whether the lower limit represents a deductible or whether it is the limit of underlying coverage, as in the case of an umbrella or excess liability policy. We shall be primarily concerned with the latter case in this section and leave the main discussion of the deductible case to the next chapter.

If the excess variable X_a is subject to an upper limit l (as in the case of an excess policy written over underlying coverage), then the claim amount paid by the insurer is the unrestricted amount x first decreased by a and then limited by l. Such claims are said to belong to the ***layer of coverage*** defined by a and l. Limit a is called an ***underlying limit*** or ***attachment point***, whereas l is the ***layer limit*** or the ***width*** of the layer. An unlimited claim of size x is said to ***penetrate*** the layer whenever $x > a$.

If $a > 0$ the layer is called an ***excess layer***, whereas in the trivial case $a = 0$ claims in the layer are referred to as ***first-dollar*** or ***ground-up*** claims. The layer defined by a and l is sometimes denoted by the "interval" notation $(a, a + l]$—although a *layer* of coverage is conceptually different from an *interval* of claims. This distinction is explored in Problems 5.14 and 5.15.

In the case of a straight deductible, however, the deductible limit is generally applied *after* the policy limit. In this situation, the layer width is the policy limit reduced by the deductible size, $l - a$, so that the insured layer is $(a, l]$. Deductibles are explored in detail in Section 6.5.

Example 5.6 illustrates, in the context of a policy limit and deductible, how upper and lower policy limits serve to partition claims into a sequence of layers.

Example 5.6. An insurance policy with a \$3,000 limit and \$100 straight deductible defines a three-layer structure: (*i*) the deductible layer between 0 and 100, (*ii*) the insured layer of width 2,900 between 100 and 3,000, and (*iii*) an uninsured layer excess of 3,000. Note that the deductible effectively reduces the policy limit from 3,000 to 2,900.

Suppose that the occurrence of insured events during the policy period gives rise to four claims—of sizes 50, 600, 1,800, and 4,000—for a total of 6,450. Three claims penetrate the insured layer, and one of these is limited by the policy limit. The table shows how they are distributed among the three layers.

Layer	Claim 1	Claim 2	Claim 3	Claim 4	Total
[0; 100]	50	100	100	100	350
(100; 3,000]	**0**	**500**	**1,700**	**2,900**	**5,100**
(3,000; ∞)	0	0	0	1,000	1,000
Total	50	600	1,800	4,000	6,450

Here the insurer pays 5,100 in the insured layer, whereas the policyholder retains 1,350 of the total claim amount—350 within the deductible layer plus 1,000 in the uninsured layer above 3,000. ■

The random variable for claim size $X_{a,l}$ in the layer $(a,\ a+l]$ is defined on the interval $a < X < \infty$ by the equation

$$X_{a,l} = \begin{cases} X - a & \text{if } a < X \le a + l \\ l & \text{if } a + l < X < \infty. \end{cases} \tag{5.9}$$

Accordingly, the cumulative distribution function of variable $X_{a,l}$ is

$$F_{X_{a,l}}(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{F_X(x+a) - F_X(a)}{1 - F_X(a)} & \text{if } 0 \le x < l \\ 1 & \text{if } l \le x < \infty. \end{cases} \tag{5.10}$$

It is easy to verify that the moments of the layer distribution are just the limited moments of the excess variable X_a:

$$E\left[X_{a,l}\right] = E\left[X_a; l\right] = \frac{E\left[X; a+l\right] - E\left[X; a\right]}{1 - F_X(a)}, \tag{5.11}$$

$$E\left[X_{a,l}^2\right] = E\left[X_a^2; l\right] = \frac{E\left[X^2; a+l\right] - E\left[X^2; a\right] - 2a\left(E\left[X; a+l\right] - E\left[X; a\right]\right)}{1 - F_X(a)}, \tag{5.12}$$

$$E\left[X_{a,l}^3\right] = E\left[X_a^3; l\right] = \frac{E\left[X^3; a+l\right] - E\left[X^3; a\right] - 3a\left(E\left[X^2; a+l\right] - E\left[X^2; a\right]\right)}{1 - F_X(a)}$$

$$+ \frac{3a^2\left(E\left[X; a+l\right] - E\left[X; a\right]\right)}{1 - F_X(a)}. \tag{5.13}$$

Example 5.7. Random variable X has a Pareto distribution with parameters $(\alpha, \beta) = (2; 3{,}000)$. What is the average claim size in the layer 4,000 excess of the limit 5,000?

We first calculate the limited severities at the attachment point 5,000 and at $a + l = 9{,}000$. At $a = 5{,}000$

$$E[X; 5{,}000] = \frac{\beta}{\alpha - 1}\left(1 - \left(\frac{\beta}{5{,}000 + \beta}\right)^{\alpha - 1}\right) = (3{,}000)\left(1 - \frac{3{,}000}{8{,}000}\right) = 1{,}875.$$

A similar calculation yields $E[X; 9{,}000] = 2{,}250$. Therefore, the layer mean is

$$\frac{E[X; 9{,}000] - E[X; 5{,}000]}{1 - F_X(5{,}000)} = \frac{2{,}250 - 1{,}875}{1 - 0.8594} = 2{,}667. \blacksquare$$

Limits imposed on the size of claims serve to decrease the variability of a claim process. To compare the dispersion of different distributions in a meaningful way, one can use the ***coefficient of variation***. For variable X the coefficient of variation $CV[X]$ is defined as the ratio of the standard deviation to the mean:

$$CV[X] = \frac{\sqrt{Var[X]}}{E[X]} = \frac{SD[X]}{E[X]}. \tag{5.14}$$

Because the coefficient of variation is a dimension-less ratio, calculating CVs for random variables with different means can provide a basis for an apt comparison. In addition, $CV[X]$ has the useful property of remaining invariant whenever X is subjected to the linear transformation $L_c(X) = cX$, where $c > 0$ (refer to Problem 2.31): $CV[cX] = CV[X]$.

Example 5.8. A claim-size variable X has a lognormal distribution with parameters $(\mu, \sigma) = (5.9809, 1.800)$. Probabilities and first and second limited moments at limits 3,000 and 8,000 for this distribution are displayed in the table.

Limit l	$F_X(l)$	$E[X; l]$	$E[X^2; l]$
3,000	0.869761	891	1,853,050
8,000	0.952557	1,276	5,774,970
∞	1.000000	2,000	102,134,385

The coefficient of variation of the unlimited variable X is

$$CV[X] = \frac{\sqrt{102{,}134{,}385 - (2{,}000)^2}}{2{,}000} = 4.9531.$$

Not surprisingly, the distribution of X_{3K} has a smaller CV:

$$E[X_{3K}] = \frac{2{,}000 - 891}{1 - 0.869761} = 8{,}515,$$

$$CV[X_{3K}] = \frac{\sqrt{\dfrac{102{,}134{,}385 - 1{,}853{,}050 - 2(3{,}000)(2{,}000 - 891)}{1 - 0.869761} - (8{,}515)^2}}{8{,}515}$$

$$= 2.9858.$$

Restricting claims to the layer between 3,000 and 8,000 by imposing on X_{3K} an upper limit of 5,000 further reduces the coefficient of variation of the claim-size variable: $CV[X_{3K}; 5{,}000] = 0.6452$. ■

5.4. Excess Claim Counts

We now investigate the distribution characteristics of the random variable N_a, the number of claims excess of an underlying limit a. Because the very definition of an excess claim depends upon the size of the claim, distributions of excess claim counts involve not only the distribution of the ground-up claim count N, but also that of the unlimited claim size X.

If the distribution of X remains unchanged over time, then the probability of an excess claim also remains constant. Whenever this is true, the distribution of the excess claim count N_a is related in a simple way to the distribution of the number N of unrestricted, ground-up claims.

Let $F_X(x)$ be the cumulative distribution function for the claim-size variable X. The probability that a claim exceeds a is $p = 1 - F_X(a)$, and the probability of obtaining n such claims is given by the conditional probability formula (5.15) below. This distribution function for N_a is derived from the fact that the number n of excess claims, given the occurrence of k ground-up claims ($n \le k$), has a binomial distribution with parameters (k, p). The resulting formula is valid for every distribution of the ground-up claim-count variable N:

$$f_{N_a}(n) = \sum_{k=n}^{\infty} \Pr\{n \text{ excess claims} \mid N = k\} \cdot \Pr\{N = k\}$$

$$= \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} f_N(k), \quad n = 0, 1, 2, \ldots. \tag{5.15}$$

It is easy to show that $E[N_a] = pE[N]$:

$$E[N_a] = \sum_{n=0}^{\infty} n \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} f_N(k)$$

$$= \sum_{n=0}^{\infty} f_N(k) \sum_{n=0}^{k} n \binom{k}{n} p^n (1-p)^{k-n}$$

$$= \sum_{k=0}^{\infty} f_N(k)(kp)$$

$$= pE[N]. \tag{5.16}$$

In a similar way one can obtain a formula for the second moment:

$$E\left[N_a^2\right] = p^2 E\left[N^2\right] + p(1-p)E[N], \tag{5.17}$$

so that

$$Var[N_a] = p^2 Var[N] + p(1-p)E[N]. \tag{5.18}$$

If N is known to have a specific parametric distribution, one can often determine the exact distribution of N_a. For example, if N has a Poisson (λ) distribution, then (5.15) becomes

$$f_{N_a}(n) = \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} \frac{\lambda^k e^{-\lambda}}{k!}$$

$$= \frac{p^n \lambda^n e^{-\lambda}}{n!} \sum_{k=n}^{\infty} \frac{\lambda^{k-n}}{(k-n)!} (1-p)^{k-n}$$

$$= \frac{(p\lambda)^n e^{-\lambda}}{n!} \sum_{i=0}^{\infty} \frac{(\lambda - p\lambda)^i}{i!}$$

$$= \frac{(p\lambda)^n e^{-p\lambda}}{n!}.$$

This means that N_a is also Poisson-distributed, with parameter $\lambda_a = p\lambda$.

It is likewise true that if N has a negative binomial distribution of the form (3.17) with parameters (α, ν), then N_a has a negative binomial distribution as well, but with parameters (α, $p\nu$). A proof is requested in Problem 5.17.

Example 5.9. The number of claims for a ground-up claim process is Poisson-distributed with $\lambda = 15$. Moreover, the unlimited claim-size variable X has the lognormal distribution of Example 5.8.

Consequently, the number of claims that penetrate a policy layer with attachment point 3,000 also has a Poisson distribution. The expected layer claim count is

$$E\left[N_{3K}\right] = \lambda(1 - F_X(3{,}000)) = (15)(0.130239) = 1.9536. \blacksquare$$

5.5. Inflation Effects

In Chapter 2 we saw that the effect of a uniform inflationary trend factor applied to an unlimited claim-size variable is moderated by the presence of a policy limit. In particular, claims subjected to a positive rate of inflation r and limited by an upper

limit increase at a rate less than r. In this section we continue that previous discussion and explore the effects of uniform inflationary pressure on claims excess of a fixed lower limit.

Suppose that the inflation factor $\tau = 1 + r$ is applied to the ground-up claim size X with c.d.f. $F_X(x)$. Then the average claim sizes excess of the limit a before and after trending are, respectively,

$$E[X_a] = \frac{E[X] - E[X; a]}{1 - F_X(a)} \quad \text{and} \quad E[\tau X_a] = \frac{\tau E[X] - \tau E[X; a/\tau]}{1 - F_X(a/\tau)}.$$

Consequently, the effective trend factor $\tilde{\tau}$ for the excess claim size X_a is

$$\tilde{\tau} = \frac{E[\tau X_a]}{E[X_a]} = \tau \cdot \frac{E[X] - E[X; a/\tau]}{E[X] - E[X; a]} \cdot \frac{1 - F_X(a)}{1 - F_X(a/\tau)}. \tag{5.19}$$

Formula (5.19) for X_a can be easily generalized to the layer claim-size variable $X_{a,l}$, as requested in Problem 5.19.

Example 5.10. Claim-size random variable X is Pareto-distributed with parameters $(\alpha, \beta) = (2; 3{,}000)$ and is subject to a uniform annual inflation rate of $r = 10\%$. What is the annual trend rate for claims excess of 5,000?

The average excess claim size before trending is

$$e_X(5{,}000) = \frac{5{,}000 + 3{,}000}{2 - 1} = 8{,}000,$$

whereas the average trended claim size is

$$e_{1.10X}(5{,}000) = (1.10)\frac{5{,}000/1.10 + 3{,}000}{2 - 1} = 8{,}300.$$

Therefore, the effective excess trend rate is $\tilde{r} = 8{,}300/8{,}000 - 1 = 3.75\%$.

Similarly, the average trended claim size in the layer (5,000; 9,000] is

$$\frac{(1.10)E[X; 9{,}000/1.10] - (1.10)E[X; 5{,}000/1.10]}{1 - F_X(5{,}000/1.10)} = \frac{2{,}415 - 1{,}988}{1 - 0.841922} = 2{,}701.$$

The non-trended severity in this layer was found in Example 5.7 to be 2,667, so the rate of change for the layer claims is $\tilde{r} = 2{,}701/2{,}667 - 1 = 1.27\%$, yet another illustration of the damping effect of an upper limit. ■

Having just examined what happens to the *size* of excess claims when the unrestricted claim size is subject to inflation, we turn now to a related question: How does such an inflationary trend affect the *number* of excess claims? One would reasonably expect that, all other things being equal, a positive rate of inflation applied to the claim size

should increase the number of claims excess of a fixed limit a—after trending, all claims are larger, so there ought to be more of them that exceed the limit.

In fact, if $E[N]$ is the expected number of ground-up claims and τ_X is the claim-size trend factor, then the expected numbers of excess claims before and after trending are, respectively, $(1 - F_X(a))E[N]$ and $(1 - F_X(a/\tau_X))E[N]$. The effective trend factor $\tilde{\tau}_N$ for the excess claim count *due solely to the effect of inflation on the claim size X* is therefore given by

$$\tilde{\tau}_N = \frac{1 - F_X(a/\tau_X)}{1 - F_X(a)}. \tag{5.20}$$

To verify that a positive inflation rate applied to the unlimited claim size generally increases the excess claim count, observe that $\tau_X > 1$ implies that $a/\tau_X < a$, and so $F_X(a/\tau_X) \le F_X(a)$. Application of this last inequality to (5.20) yields $\tilde{\tau}_N \ge 1$, as expected. A similar argument shows that $\tilde{\tau}_N \le 1$ whenever $\tau_X < 1$.

Example 5.11. As in the previous example, claim-size variable X has a Pareto distribution with $(\alpha, \beta) = (2; 3{,}000)$. The effective annual trend factor for the number of claims excess of 5,000 due to 10% inflation in the claim size X is

$$\tilde{\tau}_N = \frac{1 - F_X(5{,}000/1.10)}{1 - F_X(5{,}000)} = \frac{1 - 0.841922}{1 - 0.859375} = 1.1241. \blacksquare$$

The 12.41% increase in the number of excess claims in the last example turned out to be larger than the basic claim-size inflation rate. But this is not always the case. Problem 5.21 shows that the rate of change in the number of excess claims can be either larger or smaller than the claim-size inflation rate.

Nevertheless, it *is* possible to generalize about the change in the total aggregate excess loss due to an inflationary trend applied to the unrestricted size of loss. The expected aggregate loss amount S for claims excess of limit a is

$$\begin{aligned} E[S] &= E[N_a]E[X_a] \\ &= (1 - F_X(a))E[N] \cdot \frac{E[X] - E[X; a]}{1 - F_X(a)} \\ &= E[N](E[X] - E[X; a]). \end{aligned}$$

Combining equations (5.19) and (5.20) yields the effective trend factor for the aggregate variable S:

$$\tilde{\tau}_S = \tau_X \frac{E[X] - E[X; a/\tau_X]}{E[X] - E[X; a]}. \tag{5.21}$$

As before, $\tau_X > 1$ implies that $a/\tau_X < a$ and $E[X; a/\tau_X] \le E[X; a]$. Consequently, the quotient expression in (5.21) cannot be less than 1, and so $\tilde{\tau}_S \ge \tau_X$. Similarly, $\tilde{\tau}_S \le \tau_X$ whenever

$\tau_X < 1$. As we have just demonstrated, the existence of a fixed underlying limit magnifies, or leverages, the effect of the basic uniform claim-size trend on the aggregate excess loss.

Example 5.12. As before in Examples 5.10 and 5.11, claim-size X has a Pareto distribution with $(\alpha, \beta) = (2; 3{,}000)$ and is subject to a uniform annual inflation rate of 10%. In addition, the ground-up claim count is increasing at an annual rate of 5%. What is the annual change in the total aggregate loss generated by claims excess of 5,000? How much of this change is due solely to claim-size inflation?

Example 5.11 showed that the claim count increases at a rate of 12.41% due to the increase in X, so the total increase in the claim count is

$$r_N = (1.05)(1.1241) - 1 = 18.03\%.$$

Since the excess claim size increases at a rate of 3.75%, as shown in Example 5.10, the total aggregate loss increases at the annual rate of

$$r_S = (1.1803)(1.0375) - 1 = 22.46\%.$$

Thus, 1.2246/1.05 − 1 = 16.6% is the annual rate of increase due only to the claim-size inflation. This result, of course, can also be obtained directly from equation (5.21):

$$\tilde{r}_S = (1.10)\frac{3{,}000 - 1{,}807}{3{,}000 - 1{,}875} - 1 = 16.6\%. \blacksquare$$

5.6. Aggregate Layer Claims

The aggregate-loss random variable S for claims in the excess layer $(a, a + l]$ is defined just as in Section 4.2, but with the modified variables N_a and $X_{a,l}$ as components. Formulas for the mean, variance, and skewness of S, in terms of the ground-up claim count N and unlimited claim size X, are obtained by applying equations (5.11), (5.12), (5.13), (5.16), and (5.18) to the formulas of (4.9).

For example, if the distribution of N has mean $E[N] = \lambda$ and contagion parameter γ so that $Var[N] = \lambda + \gamma\lambda^2$, then the layer mean, variance, and skewness can be obtained from the formulas

$$E[S] = \lambda\left(E[X; a + l] - E[X; a]\right), \tag{5.22}$$

$$Var[S] = \lambda\left(E[X^2; a + l] - E[X^2; a]\right) - 2aE[S] + \gamma\left(E[S]\right)^2, \tag{5.23}$$

$$\begin{aligned} Sk[S] \cdot \left(Var[S]\right)^{3/2} &= \lambda\left(E[X^3; a + l] - E[X^3; a]\right) - 3aVar[S] \\ &\quad - 3a^2E[S] + 3\gamma E[S]Var[S] - \gamma^2\left(E[S]\right)^3. \end{aligned} \tag{5.24}$$

Example 5.13. The components of a ground-up claim process are as described in Examples 5.8 and 5.9—that is, N has a Poisson distribution with mean $\lambda = 15$, and the claim-size variable X is lognormally distributed with parameters $(\mu, \sigma) = (5.9809,$

1.8000). What are the distribution characteristics for random variable S for claims in the layer 5,000 excess of 3,000?

Formula (5.22) yields the mean

$$E[S] = (15)(1,276 - 891) = 5,775,$$

and the variance and skewness are calculated from (5.23) and (5.24):

$$Var[S] = (15)(5,774,970 - 1,853,050) - (6,000)(5,775) = 24,178,800,$$

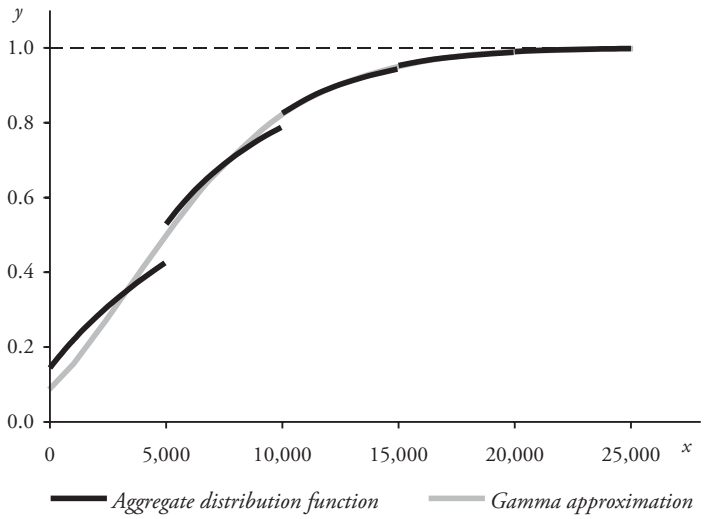
$$Sk[S] = \frac{(15)(37,049,701,689 - 4,790,705,259)}{(24,178,800)^{3/2}}$$

$$- \frac{(9,000)(24,178,800) - (3)(3,000)^2(5,775)}{(24,178,800)^{3/2}} = 0.92816.$$

Because the expected layer claim count $\lambda_{3K} = 1.9536$ is small, one should expect the cumulative distribution function for S to have significant discontinuities at the smaller multiples of the layer limit 5,000. This is clearly evident in Figure 5.4, which displays the graph of $y = F_S(x)$ as well as that of the continuous shifted gamma approximation to the function. ■

The distribution of Example 5.13 exhibits some properties typical of the distributions of aggregate loss in an excess layer. It is often the case, especially for small portfolios of policies or even for large single policies, that the expected layer claim count is small. As we have seen, this leads to jump discontinuities of substantial size at the lower end of the distribution, thus complicating the task of approximating the distribution with one of the continuous approximation models. Nevertheless, these methods can still return reasonable results for the long tail of the distribution, usually the most important region for applications of the aggregate distribution.

Figure 5.4. Layer Aggregate Loss Distribution Function [Example 5.13]



5.7. Problems

5.1 The claim-size random variable X for a claim process has an exponential distribution with mean 1,000. The expected number of claims for the ground-up claim process is 20. However, policy conditions limit claims to the layer between 1,000 and 3,000.

(***a***) Compute the mean and variance of the layer claim size.

(***b***) Compute the expected number of layer claims.

(***c***) How do the policy conditions alter the coefficient of variation of the claim-size variable?

(***d***) If a uniform inflation rate of 10% *per annum* is applied to X, what is the annual percentage increase in the layer claim size? . . . the layer claim count? . . . the total layer aggregate loss?

5.2 Compute $e_X(3{,}000)$ for the following distributions of X. Note that the unlimited severity for each distribution is the same: $E[X] = 2{,}000$.

(***a***) uniform on [0; 4,000].

(***b***) gamma, $(\alpha, \beta) = (2; 1{,}000)$.

(***c***) exponential, $\beta = 2{,}000$.

(***d***) shifted Pareto, $(\alpha, \beta) = (3; 4{,}000)$.

(***e***) lognormal, $(\mu, \sigma) = (5.9809, 1.8000)$.

5.3 Verify formulas (5.5) and (5.6) for the second and third moments of X_a.

5.4 Verify formulas (5.11), (5.12), and (5.13) for the moments of $X_{a,l}$.

5.5 Prove: $E[X_{a,l}] = e_X(a) - e_X(a+l)\dfrac{1-F_X(a+l)}{1-F_X(a)}$.

5.6 Claim-size variable X has the mixed cumulative distribution function $F(x) = \sum_{k=1}^{m}\omega_k F_k(x)$, where $\{F_k\}$ are the component distribution functions and the weights $\{\omega_k\}$ satisfy $\omega_k > 0$ and $\sum_{k=1}^{m}\omega_k = 1$. Show that

$$e_X(x) = \frac{\sum_{k=1}^{m}\omega_k e_k(x)(1-F_k(x))}{1-\sum_{k=1}^{m}\omega_k F_k(x)}, \quad 0 < x < \infty.$$

5.7 Compute $\Pr\{X_d > x\}$, where $0 < d < x$, and the distribution of X is:

(***a***) exponential (β).

(***b***) shifted Pareto (α, β).

5.8 Prove: If $E[X]$ exists, then $E[X] = E[X; x] + e(x)\,(1 - F(x))$ for all $x > 0$.

5.9 (***a***) Show that the excess severity function $e_X(x)$ can be expressed by the integral formula

$$e_X(x) = \int_x^\infty (u-x)\,dF_X(u) \Big/ \int_x^\infty dF_X(u), \quad 0 < x < \infty.$$

(***b***) The unlimited claim-size observations $\langle x_1, x_2, \ldots, x_n\rangle$ from a random sample of size n are grouped into a sequence of intervals of the form $(c_{k-1}, c_k]$, where

n_k = # *claims in the* k^{th} *interval,* $\Sigma_k n_k = n$, and $\bar{x}_k$ = *mean claim size in the* k^{th} *interval.* Show that the sample excess severity function $e_n(c_k)$ is

$$e_n(c_k) = \sum_{i>k} n_i(\bar{x}_i - c_k) \Big/ \sum_{i>k} n_i.$$

5.10 For the lognormal claim-size random variable X of Example 5.8 calculate:
(***a***) $CV[X; 3{,}000]$ and $CV[X; 8{,}000]$. Compare these numbers to $CV[X]$.
(***b***) $Sk[X]$, $Sk[X; 3{,}000]$, and $Sk[X; 8{,}000]$.

5.11 Calculate $CV[X]$ in terms of the distribution parameters when the distribution of X is:
(***a***) exponential (β).
(***b***) gamma (α, β).
(***c***) lognormal (μ, σ).
(***d***) shifted Pareto (α, β) with $\alpha > 2$.
(***e***) uniform on the interval $[0, a]$, $a > 0$.

5.12 Assume that the policy of Example 5.6 has a ground-up claim process with $E[N] = 5$ and that the claim-size variable X is Pareto-distributed with (α, β) = (3; 5,000). For each layer L defined in that example compute:
(***a***) probability P_L that a claim penetrates the layer L.
(***b***) expected number of layer claims $E[N_L]$.
(***c***) expected layer claim size $E[X_L]$.
(***d***) expected aggregate layer loss $E[S_L]$.

Layer L	P_L	$E[N_L]$	$E[X_L]$	$E[S_L]$
[0; 100]	_____	_____	_____	_____
(100; 3,000]	_____	_____	_____	_____
(3,000; ∞)	_____	_____	_____	_____
[0; ∞)	1.0000	5.0000	2,500.00	12,500

5.13 Assume that $E[X]$ exists and that the partition

$$0 = b_0 < b_1 < b_2 < \cdots < b_{m-1} < b_m = \infty$$

defines a sequence of m contiguous layers. Prove: if μ_k is the mean claim size for the k^{th} layer $(b_{k-1}, b_k]$ and $p_k = \Pr\{X > b_{k-1}\}$, then $E[X] = \Sigma_{k=1}^{m} p_k \mu_k$.

5.14 Let X denote an unlimited claim-size variable with distribution function F, and assume that $0 \le a < b$. The claim *interval* between a and b is just the set of claims of size X such that $a < X \le b$.
(***a***) Explain how the claim *interval* between a and b differs from the *layer* defined by a and b.
(***b***) If λ is the mean number of ground-up claims, what is the expected number of claims in the interval $a < X \le b$?
(***c***) Prove that the average claim size in the interval $a < X \le b$ is

$$E[X \mid a < X \le b] = \frac{E[X; b] - E[X; a] - b(1 - F(b)) + a(1 - F(a))}{F(b) - F(a)}.$$

5.15 For the grouped data of Example 5.4 the indicated groups can be used to define either a sequence of claim *intervals* or a sequence of *layers* of coverage. Calculate the average claim size for each interval and each layer.

Interval/Layer	Interval Mean	Layer Mean
(0; 100]	_____	_____
(100; 500]	_____	_____
(500; 1,000]	_____	_____
(1,000; 2,000]	_____	_____
(2,000; 4,000]	_____	_____
(4,000; 5,000]	_____	_____
(5,000; 10,000]	_____	_____

5.16 Verify formula (5.17) for the second moment of the excess claim-count random variable N_a.

5.17 Prove that whenever the ground-up claim count N has a negative binomial distribution of the form (3.17) with parameters (α, ν), then the distribution of the claim count N_a excess of an underlying limit a is also negative binomial, with parameters $(\alpha, p\nu)$, where X is the claim-size variable and $p = 1 - F_X(a)$.

5.18 The ground-up claim count N has mean λ and contagion parameter γ. Prove that for the excess claim count N_a, the contagion parameter is unchanged: $\gamma_a = \gamma$.

5.19 Derive a generalization of formula (5.19) for the effective trend factor $\tilde{\tau}$ associated with the layer claim size $X_{a,l}$.

5.20 Show that the leveraging effect on the aggregate excess loss disappears whenever the underlying limit a is also trended at the same rate as the claim-size variable X.

5.21 Claim-size random variable X is lognormally distributed with $(\mu, \sigma) = (5.9809, 1.8000)$ and is subject to an inflation rate of 10% *per annum.* Calculate the corresponding effective inflation rate on the excess claim count for each of the following underlying limits, thus demonstrating that the induced claim-count rate of change can be either more or less than the basic claim-size inflation rate.
(***a***) $a = 3{,}000$. (***b***) $a = 8{,}000$.

5.22 Variable X has a lognormal distribution as in Problem 5.21 and is also subject to a 10% inflation rate. Calculate the effective inflation rate on the excess aggregate loss for each of the following excess layers. What can be said about the effective layer inflation rate as compared to the basic rate of inflation?
(***a***) $(a, a + l] = (3{,}000; 5{,}000]$. (***b***) $(a, a + l] = (3{,}000; 8{,}000]$.

5.23 As in Example 5.3, the table summarizes grouped claim-size data from a sample of 1,000 claims. These claims are excess of a 500 straight deductible and have been censored by a policy limit of 100,000.

(***a***) Use the minimum chi-square method to obtain estimates of lognormal parameters for the ground-up population claim-size distribution.

(***b***) Estimate the number of claims eliminated by the policy deductible.

Size Group	# Claims
0–1,000	236
1,001–2,000	161
2,001–3,000	107
3,001–5,000	135
5,001–10,000	159
10,001–15,000	71
15,001–25,000	62
25,001–50,000	43
50,001–75,000	11
75,001–100,000	15
Total	1,000

5.24 Assume that X is a continuous claim-size random variable for which $E[X]$ exists as a finite number. Derive this integral formula for $e_X(x)$:

$$e_X(x) = \frac{\int_x^\infty (1 - F_X(u))\, du}{1 - F_X(x)}.$$

5.25 Establish the following asymptotic properties of the mean excess claim size function $e_X(x)$. In each case it is useful to express $e_X(x)$ by the integral formula of Problem 5.24.

(***a***) If X has the gamma density function

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta},$$

then $e_X(x) \approx x/(x/\beta + \alpha - 1)$ for large x so that $\lim_{x\to\infty} e_X(x) = \beta$. [*Hint*: Apply l'Hôpital's Rule.]

(***b***) If X has a Weibull density function

$$f(x) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} \exp\left(-(x/\beta)^\alpha\right), \quad 0 < x < \infty,$$

then $e_X(x) \approx \beta^\alpha/(\alpha x^{\alpha-1})$ for large x. [*Hint*: use l'Hôpital's Rule to show that

$$\lim_{x\to\infty} \frac{e_X(x)}{\beta^\alpha / \left(\alpha x^{\alpha-1}\right)} = 1.]$$

6. Limits and Deductibles

In this chapter we explore some common applications of the claim-count, claim-size, and aggregate-loss distributions in property/casualty insurance. In particular, we investigate the pricing of policies with various coverage limitations such as deductible options, and per-claim and aggregate limits with a variety of properties. We begin by reviewing some basic premium concepts and how they relate to distributional theory.

6.1. Premium Concepts

Every insurance policy has an associated loss process described jointly by a claim-count random variable N and a claim-size variable Y. In the following discussion Y represents the entire claim amount: the indemnity payment plus loss adjustment expense allocated to the claim, as limited by policy conditions. Allocated loss adjustment expenses (ALAE) are those incurred during the settlement process for an individual claim: attorneys' fees, investigation expense, expert witness fees, and the like. (Unallocated loss adjustment expenses, such as claim department overhead, are usually treated as general expenses and are not included in the policy aggregate loss.) The expected loss for the policy is then $E[N]E[Y]$, the mean of the policy aggregate loss distribution. Premium charged for such a policy is based on this expected loss, loaded for general expenses, underwriting profit, and a charge for risk.

The mean $E[N]$ of the claim-count variable represents the expected number of claims per policy. In most situations, the expected claim count is seen to depend on an ***exposure unit*** associated with the policy coverage. The exposure unit is usually chosen to have certain desirable characteristics: (*i*) it should be a meaningful indicator of the policy's expected number of claims—the more exposure units covered by the policy the greater the expected number of claims, and (*ii*) one should be able to determine an expected number of claims—constant over at least a moderate period of time—associated with a single exposure unit.

For example, a single auto is the customary exposure unit for an auto liability policy with a term of one year. For such a policy the expected number of policy claims $E[N]$ is obtained by multiplying the number of autos covered by the policy for a year, referred to as the number of ***vehicle years***, and the expected number of claims per auto per year—that is, the number of claims per vehicle year. Other common measures of exposure include dollars of annual payroll for workers' compensation policies, the number of objects manufactured in a year or dollars of annual sales for product liability coverages, and building area measured in square feet for premises liability coverages.

The expected number of claims per unit exposure is called the ***claim frequency***. If m denotes the number of exposure units and φ the claim frequency, then obviously

$$m\varphi = E[N]. \tag{6.1}$$

In addition, the ***claim severity*** is the average claim size $E[Y]$ for the policy. The product of frequency and severity, denoted by p, is called the ***pure premium***:

$$p = pure\ premium = (frequency)(severity) = \varphi E[Y]. \tag{6.2}$$

It is clear that the pure premium represents the *expected aggregate claim amount per unit of exposure.* Accordingly, exposure times pure premium yields the policy expected loss:

$$mp = m(\varphi E[Y]) = E[N]E[Y]. \tag{6.3}$$

In the case that a policy involves more than a single line of business—each with its own exposure, frequency, and severity—then the policy expected loss is obtained by summing over all component coverages the corresponding products of exposure, frequency, and severity.

Example 6.1. For a certain general liability coverage the exposure unit is \$1,000 of annual sales, the claim frequency is 0.000825 claims per \$1,000 sales per year, and the claim severity is \$5,200.

An insured has \$650,000 of sales revenue per year. Consequently, the number exposure units for an annual policy is

$$m = \frac{\$650{,}000}{\$1{,}000} = 650,$$

the pure premium is $p = (0.000825)(5{,}200) = 4.29$, and the expected loss for the policy is $mp = (650)(4.29) = 2{,}789$. ■

To calculate the policy premium one must first load the pure premium amount with a provision for general expenses, underwriting profit, and risk. ***General expenses*** include acquisition expense—commission paid to agents and brokers—salaries and overhead, taxes and fees, and other costs of doing business. ***Underwriting profit*** is the expected excess of premium over paid losses and expenses. (In some lines of business the underwriting profit could be zero, or even negative, in anticipation of an offset from investment income.) The ***risk charge*** is extra premium collected by the insurer to cover such contingencies as (*i*) random fluctuations of losses about the expected values and (*ii*) uncertainty inherent in the selection of critical parameters used in modeling the underlying loss process. Insurer risk from the first source is called ***process risk*** and that from the second, ***parameter risk***.

Provisions for expense, profit, and risk can be treated either as variable—loaded as a percent of the final premium amount—or as fixed—added as a dollar amount per unit of exposure to the pure premium. Agent and broker commission is generally a variable expense, whereas the overhead cost of issuing a policy could be loaded as a fixed expense.

If variable expenses, plus the load for profit and risk, constitute the fraction v of the total policy premium,[51] and fixed expenses are f dollars per unit exposure, then the modified pure premium

$$R = \frac{p+f}{1-v} \tag{6.4}$$

is the ***rate per unit exposure***. The number of exposure units m times the rate R yields the final policy premium P:

$$P = mR = \frac{m(p+f)}{1-v} = \frac{E[N]E[Y]+mf}{1-v}. \tag{6.5}$$

In the case that $f = 0$—that is, all expense amounts are assumed to be variable and expressed by the expense ratio v—the factor

$$\psi = \frac{1}{1-v} \tag{6.6}$$

is called a ***loss-cost multiplier***. Rate formula (6.4) then reduces to the simpler form $R = \psi p$, and the premium formula becomes

$$P = mR = m(\psi p) = \psi E[N]E[Y]. \tag{6.7}$$

In subsequent sections we shall generally assume that expenses are loaded by means of a loss-cost multiplier ψ, as in (6.7).

Example 6.2. A business owner wishes to buy annual insurance coverage for general liability and auto liability for a business operation that involves premises of 20,000 square feet and four automobiles. General and auto liability premiums are rated separately, as indicated below.

For the general liability coverage the insurer has determined a claim frequency of 0.004 per 1,000 square feet per year and a claim severity of 6,500. The general liability pure premium is therefore $p = (0.004)(6{,}500) = 26.00$. Variable expenses plus profit load amount to 30% of the premium; fixed expenses are 4.10 per exposure unit. Therefore, the general liability annual rate is

$$R_{GL} = \frac{26.00 + 4.10}{1 - 0.30} = 43.00 \text{ per } 1{,}000 \text{ square feet.}$$

For the auto coverage the claim frequency is 0.052 per vehicle year, with a claim severity of 2,800 and fixed expense of 9.80 per vehicle year, and so the auto liability rate is

$$R_{AL} = \frac{(0.052)(2{,}800) + 9.80}{1 - 0.30} = 222.00 \text{ per vehicle year.}$$

[51] As we shall see in Section 6.3, the risk load is often calculated as an amount that varies with the policy limit, as well as one that varies with premium.

Annual liability premium P is obtained by multiplying the number of exposure units and the rate for each coverage and summing the results:

$$P = \frac{20{,}000}{1{,}000} R_{GL} + 4R_{AL} = (20)(43) + (4)(222) = \$1{,}748. \blacksquare$$

6.2. Increased Limit Factors

The premium for many property/casualty policies is calculated first for a basic per-claim policy limit, and then this basic-limit premium is multiplied by an appropriate ***increased limit factor*** (ILF) to determine the full policy premium. A set of increased limit factors—one for each of the available policy limit options—can be obtained from an empirical loss distribution based on loss data organized around the required policy per-claim limits, or it can be derived from an appropriate parametric size-of-loss distribution. Such an analytic distribution fit to empirical sample data is often useful for obtaining factors for those higher limits for which data are either sparse or nonexistent.

Let P_b denote the policy premium at the basic per-claim limit b and P_l the premium at a policy per-claim limit l. Then the increased limit factor $I(l)$ is defined by

$$I(l) = \frac{P_l}{P_b}, \tag{6.8}$$

so that $P_l = P_b \cdot I(l)$. Note also that if the policy premium is based on formula (6.7), then $p_l = p_b \cdot I(l)$, where p_l is the pure premium associated with the limit l.

In the discussion that follows, $E_l[Y]$ is the policy severity, including both indemnity payment and allocated loss adjustment expense, appropriately modified by the policy limit l. When expenses are loaded by means of a loss-cost multiplier ψ, ILF formula (6.8) becomes

$$I(l) = \frac{P_l}{P_b} = \frac{\psi E[N] E_l[Y]}{\psi E[N] E_b[Y]} = \frac{E_l[Y]}{E_b[Y]}. \tag{6.9}$$

The specific form of $E_l[Y]$ depends on whether policy conditions stipulate that limit l applies to the full claim amount, including both indemnity and allocated loss adjustment expense portions of a claim, or whether it applies only to the indemnity payment.

Consider first the case that policy limit l applies to the total claim amount: indemnity loss plus loss adjustment expense. If X_t denotes the ground-up, unlimited total claim-size random variable, then the policy severity is $E_l[Y] = E[X_t; l]$. In these circumstances ILF formula (6.9) can be expressed as

$$I(l) = \frac{E[X_t; l]}{E[X_t; b]}. \tag{6.10}$$

On the other hand, suppose that the limit l applies only to the indemnity portion of the claim, as is usually the case. If random variable X denotes just the indemnity component of the claim, then one could write

$$E_l[Y] = E[X; l] + \varepsilon, \tag{6.11}$$

where ε is the average per-claim allocated loss adjustment expense, independent of the policy limit. In this case (6.9) has the form

$$I(l) = \frac{E[X; l] + \varepsilon}{E[X; b] + \varepsilon}. \tag{6.12}$$

Provision for loss adjustment expense in formula (6.11) is an overall average amount ε added to every claim, regardless of size. Amount ε can thus be interpreted as the mean of a loss adjustment expense random variable, but in (6.11) it is unnecessary to know exactly how that variable is distributed.

As an alternative to this approach, it is sometimes useful to assume that loss adjustment expense bears some functional relationship to the size of the indemnity payment. One simple scheme is to assume that loss adjustment expense is a fixed multiple u of the indemnity amount. This assumption can be approximately true provided that the indemnity payment is not too large. (An alternative, hybrid method of expense loading is described in Problem 6.6.) Again, assuming that the policy limit applies only to the indemnity portion of the claim, one can write

$$E_l[Y] = E[X; l] + u\,E[X; l] = E[X; l](1 + u). \tag{6.13}$$

Then ILF formula (6.9) becomes

$$I(l) = \frac{E[X; l](1 + u)}{E[X; b](1 + u)} = \frac{E[X; l]}{E[X; b]}. \tag{6.14}$$

The three approaches to loss adjustment expense incorporated into formulas (6.12), (6.13), and (6.14) can be combined into a single general formula for the policy severity:

$$E_l[Y] = (E[X; l] + \varepsilon)(1 + u). \tag{6.15}$$

In case that limit l applies to indemnity loss plus loss adjustment expense, set $X = X_t$ and $\varepsilon = u = 0$ in (6.15). Otherwise, when the limit applies only to the indemnity payment, let variable X represent the indemnity-only portion of a claim and set either $\varepsilon = 0$ or $u = 0$, as desired.

The Insurance Services Office (ISO) increased limits methodology treats allocated loss adjustment expense additively like the constant ε in formula (6.15) and loads unallocated adjustment expense multiplicatively like the factor $1 + u$ in that formula.[52]

[52] For an extended discussion of the ISO method, refer to a current ISO Actuarial Service Circular for increased limits data and analysis for General and/or Commercial Auto Liability (Jersey City, NJ: Insurance Services Office, Inc.).

Table 6.1. Increased Limit Factors [Example 6.3]

Limit l ($000)	$E[X; l]$	$I(l)$ ALAE = 2,200	$I(l)$ ALAE = 20%
100	8,896	1.0000	1.0000
500	13,626	1.4263	1.5317
750	14,668	1.5202	1.6488
1,000	15,345	1.5812	1.7249
2,000	16,738	1.7067	1.8815
3,000	17,390	1.7655	1.9548
4,000	17,782	1.8008	1.9989
5,000	18,048	1.8248	2.0288

Example 6.3. Indemnity losses for a portfolio of insurance policies have a lognormal claim-size distribution with parameters $(\mu, \sigma) = (7.000, 2.400)$. The policy per-claim limit applies only to the indemnity portion of a claim, and the average per-claim loss adjustment expense is 2,200. Claim frequency for these policies is $\varphi = 0.0005$ per exposure unit, and variable expenses equal 35% of premium.

A set of increased limits factors based on (6.15) with $b = 100{,}000$, $\varepsilon = 2{,}200$, and $u = 0$ is shown in the third column of Table 6.1. For example,

$$I(1{,}000K) = \frac{15{,}345 + 2{,}200}{8{,}896 + 2{,}200} = 1.5812.$$

For a policy with 400 exposure units, so that $E[N] = (400)(0.0005) = 0.2000$, the basic-limit premium is

$$P_{100K} = \frac{(0.2000)(8{,}896 + 2{,}200)}{1 - 0.35} = \$3{,}414.$$

The corresponding premium for a policy limit of 1,000,000 is therefore

$$P_{1{,}000K} = P_{100K} \cdot I(1{,}000K) = (3{,}414)(1.5812) = \$5{,}398.$$

Alternatively, if the loss adjustment expense is treated as 20% of the indemnity portion of the claim, then the resulting increased limit factors are displayed in the fourth column of Table 6.1. For example, in this case

$$I(1{,}000K) = \frac{(15{,}345)(1.20)}{(8{,}896)(1.20)} = 1.7249.$$

For the policy with 400 exposure units the basic-limit premium is

$$P_{100K} = \frac{(0.2000)(8{,}896)(1.20)}{1 - 0.35} = \$3{,}285,$$

and with a 1,000,000 limit,

$$P_{1{,}000K} = (3{,}285)(1.7249) = \$5{,}666. \blacksquare$$

Excess Layer Pricing

Increased limit factors can also be applied to price an excess layer of coverage, defined by a policy limit l and attachment point a ($l > 0$, $a > 0$), as discussed in Section 5.3. If φ and $E_{a,l}[Y]$ denote, respectively, the ground-up claim frequency and the severity for the policy layer $(a,\ a + l]$, then the layer pure premium is

$$p_{a,l} = \varphi\left(1 - F_{X_t}(a)\right) E_{a,l}[Y],$$

where X_t is the total ground-up claim amount. In the case that X_t is subject to the layer limits we rearrange the pure premium formula as follows:

$$\begin{aligned} p_{a,l} &= \varphi\left(1 - F_{X_t}(a)\right) \frac{E[X_t; a+l] - E[X_t; a]}{1 - F_{X_t}(a)} \\ &= \varphi\left(E[X_t; a+l] - E[X_t; a]\right) \\ &= \varphi E[X_t; b]\left(\frac{E[X_t; a+l]}{E[X_t; b]} - \frac{E[X_t; a]}{E[X_t; b]}\right). \end{aligned} \tag{6.16}$$

In this special case, a layer factor, applied to the basic-limit pure premium to calculate the pure premium for the excess layer, is just the difference of two ground-up increased limit factors of the form (6.9), namely,

$$p_{a,l} = p_b \cdot \left(I(a+l) - I(a)\right).$$

Since $P_{a,l} = m(\psi p_{a,l}) = m(\psi p_b)\ (I(a+l) - I(a))$, premium for the excess layer $(a,\ a + l]$ can be calculated by using the ***layer formula*** for policy premium P:

$$P = P_b \cdot \left(I(a+l) - I(a)\right). \tag{6.17}$$

The simplicity of this basic formula makes it very easy to apply. Because of this, it is widely used in increased limits pricing, even in situations where it is not strictly appropriate. For example, suppose that the layer limits l and a apply only to the indemnity portion of a claim and ALAE is added as in formula (6.11). Then the excess-layer premium based on that model would be

$$\begin{aligned} P_{a,l} &= \psi E[N]\left(1 - F_X(a)\right)\left(\frac{E[X; a+l] - E[X; a]}{1 - F_X(a)} + \varepsilon\right) \\ &= \psi E[N]\left(E[X; a+l] - E[X; a] + \left(1 - F_X(a)\right)\varepsilon\right). \end{aligned} \tag{6.18}$$

On the other hand, the layer formula for P yields

$$P = \psi E[N](E[X; b] + \varepsilon)\left(\frac{E[X; a+l] + \varepsilon}{E[X; b] + \varepsilon} - \frac{E[X; a] + \varepsilon}{E[X; b] + \varepsilon}\right)$$
$$= \psi E[N](E[X; a+l] - E[X; a]). \tag{6.19}$$

Notice that $P = P_{a,l}$ and that the load for loss adjustment expense has dropped out of the premium calculation in (6.19) entirely. In the situation where such an excess policy is written over a primary policy providing first-dollar coverage for the primary layer $[0, a]$, this state of affairs is consistent with the assumption that allocated loss adjustment expense is paid in its entirety by the primary insurer.

On the other hand, if loss adjustment expense is loaded by means of the factor $1 + u$, then

$$P_{a,l} = \psi E[N](E[X; a+l] - E[X; a])(1+u),$$

where the provision for ALAE in the excess premium is

$$ALAE = \psi E[N](E[X; a+l] - E[X; a])\, u.$$

The layer formula for P yields the premium amount

$$P = \psi E[N] E[X; b](1+u)\left(\frac{E[X; a+l]}{E[X; b]} - \frac{E[X; a]}{E[X; b]}\right)$$
$$= \psi E[N]\big(E[X; a+l] - E[X; a]\big)(1+u). \tag{6.20}$$

In this case, for which the ALAE multiplier u is the same for both primary and excess policies, the basic layer formula preserves the loss adjustment expense loading exactly and $P = P_{a,l}$.

Example 6.4. We return to the portfolio of policies described in Example 6.3 and calculate the premium for successive excess layers of insurance for a policy with $m = 400$. We use the ILFs constructed in that example under the assumption that the average per-claim ALAE payment is $\varepsilon = \$2{,}200$.

The basic limit premium was calculated in the previous example to be \$3,414. Thus, premium P for the layer (1,000,000; 2,000,000] given by the layer formula (6.17) is

$$P = (3{,}414)(1.7067 - 1.5812) = (3{,}414)(0.1255) = \$428.$$

Similarly, for the layer (2,000,000; 3,000,000], we obtain

$$P = (3{,}414)(1.7655 - 1.7067) = (3{,}414)(0.0588) = \$201.$$

Premium amounts for the successive million-dollar layers obtained from these layer factors applied to the basic-limit premium are displayed in Table 6.2. ■

Table 6.2. Layer Premium [Example 6.4]

Layer ($000)	Layer Factor	Premium
[0; 100]	1.0000	3,414
[0; 1,000]	1.5812	5,398
(1,000; 2,000]	0.1255	428
(2,000; 3,000]	0.0588	201
(3,000; 4,000]	0.0353	121
(4,000; 5,000]	0.0240	82

Consistency

The premiums calculated in Example 6.4 illustrate an important and desirable characteristic of excess layer pricing—premium amounts for successively higher layers of constant width decrease as the attachment point becomes larger. In that example, premium for the ground-up million-dollar layer is $5,398, and for successively higher layers of one-million-dollar width the calculated premium steadily declines with increasing attachment point: $428, $201, $121, $82. As we shall see, this is a property common to all pricing methods based on expected losses and reasonable distributions for the claim-size random variables.

Consider a set of increased limit factors based on the general severity formula (6.15). If the claim-size variable X has a continuous probability density function $f_X(x) = F'_X(x)$, then the ILF function $I(x)$ is twice differentiable with respect to the limit x (refer to Problem 2.9). Specifically, for all $x > 0$

$$I'(x) = \frac{(1+u)(1-F_X(x))}{E[X;b]+\varepsilon} \quad \text{and} \quad I''(x) = \frac{-(1+u)f_X(x)}{E[X;b]+\varepsilon} < 0.$$

A set of increased limit factors for which $I''(x) < 0$ for all limits x is said to be ***consistent***. Thus, every set of increased limit factors based on severity formula (6.15) for which the claim-size density function is continuous is always consistent.

Consistent sets of increased limit factors share a common property: the premium P calculated from the layer formula (6.17) applied to successive excess layers of *constant width* is a decreasing function of the attachment point limit. It is easy to verify this assertion in the case that the claim-size variable has a continuous probability density function. Assume that in the formula

$$P = P_b \cdot (I(x+l) - I(x))$$

the attachment point x is variable, whereas the layer width l is a fixed constant. Then the rate of change of premium P with respect to x is

$$\frac{dP}{dx} = P_b\left(\frac{d}{dx}I(x+l) - \frac{d}{dx}I(x)\right).$$

Consistency means that $I'(x)$ is a decreasing function of x, so that

$$\frac{d}{dx} I(x+l) < \frac{d}{dx} I(x) \quad \text{and hence} \quad \frac{dP}{dx} < 0.$$

Therefore, for each fixed l, premium P for the layer $(x, x+l]$ decreases as the attachment point x increases.

6.3. Risk Load

Increased limit factors based on expected-value concepts have generally been thought to be inadequate for pricing insurance policies with high limits or attachment points unless they were loaded with a charge for insurer risk. With lower claim probabilities for such policies, loss behavior associated with excess policies is more volatile and less predictable than that of primary policies with lower limits. Insurer risk due to such variability is process risk, in contrast to the parameter risk derived from estimation errors in selecting the claim-count and claim-size distributions. Process risk has long been understood by actuaries as a function of the variance of the basic stochastic claim process for a portfolio of policies or line of property/casualty insurance business.

In most approaches to risk-loaded increased limit factors, the risk load $\rho(l)$ is usually defined as an increasing function of the policy limit l, which is added to the expected total policy severity for the policy. The severity formula (6.15) is thus modified

$$E_l[Y] + \rho(l) = (E[X; l] + \varepsilon)(1+u) + \rho(l), \tag{6.21}$$

and the resulting risk-loaded increased limit factors are

$$I(l) = \frac{(E[X; l] + \varepsilon)(1+u) + \rho(l)}{(E[X; b] + \varepsilon)(1+u) + \rho(b)}. \tag{6.22}$$

The merits of different methods of quantifying process risk for increased limit factors have been debated since the mid-1970s. Robert Miccolis[53] suggested in 1977 that process risk load be added to the policy expected aggregate loss as a constant multiple of the variance of the policy aggregate indemnity-loss random variable S:

$$E[N]E[X; l] + k\,Var[S]$$
$$= E[N]\left(E[X; l] + k\frac{E[N]Var[X; l] + Var[N](E[X; l])^2}{E[N]} \right).$$

The multiplier k in this formula is selected arbitrarily to produce the desired level of risk loading. The risk load function $\rho(l)$ in formula (6.21) is thus given by

$$\rho(l) = k\frac{Var[S]}{E[N]} = k\left(E[X^2; l] + \delta(E[X; l])^2\right), \tag{6.23}$$

[53] Miccolis [16].

where $\delta = Var[N]/E[N] - 1$. Setting $\delta = 0$, of course, is consistent with the assumption that N has a Poisson distribution. Note also that when $\delta = 0$ the risk load $\rho(l)$ is independent of the claim-count random variable and dependent only upon the claim-size variable. Such a variance-based approach to process risk load was adopted by ISO in the early 1980s.[54]

By mid-decade, however, ISO changed to a method based on the standard deviation of the policy aggregate indemnity-loss distribution:

$$\begin{aligned}\rho(l) &= k\frac{\sqrt{Var[S]}}{E[N]} \\ &= \frac{k}{\sqrt{E[N]}}\sqrt{E[X^2; l] + \delta(E[X; l])^2} \\ &= k'\sqrt{E[X^2; l] + \delta(E[X; l])^2}, \end{aligned} \tag{6.24}$$

where δ is defined as in the variance formula (6.23). ISO actuaries cited several reasons for this change. Risk-loaded factors based on (6.23) and (6.24) with thick-tailed Pareto distributions for X were sometimes inconsistent (for example, refer to Problem 6.10). Moreover, it seemed preferable to express the risk load as a dollar amount rather than in terms of (dollars)^2. In 1991 ISO introduced a new risk-loading method that includes a measure of parameter risk as well as process risk. But this new method returned to the earlier variance approach to process risk.[55]

Example 6.5. We turn again to the portfolio of policies described in Example 6.3. Indemnity losses are distributed lognormally with parameters $(\mu, \sigma) = (7.000, 2.400)$, and allocated loss adjustment expense is 20% of the indemnity payment. We generate a set of risk-loaded increased limit factors by using formula (6.22) with $\varepsilon = 0$ and $u = 20\%$. The risk load $\rho(l)$ is obtained by the standard deviation method (6.24) with $k' = 0.0277$ and $\delta = 0$. Thus,

$$\rho(100{,}000) = (0.0277)\sqrt{512{,}509{,}058} = 627,$$

$$\rho(1{,}000{,}000) = (0.0277)\sqrt{5{,}283{,}276{,}848} = 2{,}013,$$

so that

$$I(1{,}000{,}000) = \frac{(15{,}345)(1.20) + 2{,}013}{(8{,}896)(1.20) + 627} = 1.8074.$$

Two sets of increased limit factors—risk-loaded and non-risk-loaded—are displayed in Table 6.3. The average increased limit factor in each column is obtained by using the indicated portfolio weights for the given set of limits. The ratio of these two averages

[54] Insurance Services Office [9].

[55] This approach is based on the paper by Glenn Meyers [15].

Table 6.3. Risk-Loaded Increased Limit Factors [Example 6.5]

Limit l ($000)	$E[X; l] \times 1.20$	Risk Load $\rho(l)$	$I(l)$ w/o Risk Load	$I(l)$ w/ Risk Load	Limit Weight
100	10,675	627	1.0000	1.0000	15%
500	16,351	1,473	1.5317	1.5770	10%
1,000	18,414	2,013	1.7249	1.8074	30%
2,000	20,086	2,663	1.8815	2.0128	20%
3,000	20,868	3,090	1.9548	2.1197	10%
4,000	21,338	3,410	1.9989	2.1897	10%
5,000	21,658	3,668	2.0288	2.2407	5%
Average			1.6938	1.7955	

indicates an increase of $1.7955/1.6938 - 1 = 6.0\%$. This means that using the risk-loaded factors on a portfolio with this distribution of policy limits would generate 6% more premium than would be obtained by using the unloaded factors. ■

Unfortunately, both the variance and standard deviation approaches to risk load are incompatible with the layer formula (6.17) for pricing an excess layer. For example, if the risk load $\rho_{a,l}$ for the excess layer $(a, a + l]$ is defined as a multiple of the variance of the aggregate indemnity-loss variable S for the layer, then $\rho_{a,l} \neq \rho(a + l) - \rho(a)$, where $\rho(l)$ is defined by (6.23).

To show this in a special case, we first calculate the layer risk load. For simplicity, assume that N is the ground-up claim-count variable with $\delta = 0$ and that X is the ground-up claim-size variable. Then

$$\rho_{a,l} = \frac{k\,Var[S]}{(1 - F_X(a))E[N]}$$
$$= k\left(E\left[X^2; a + l\right] - E\left[X^2; a\right] - 2a\left(E\left[X; a + l\right] - E\left[X; a\right]\right)\right).$$

But this means that

$$\rho_{a,l} = \rho(a + l) - \rho(a) - 2ka\left(E\left[X; a + l\right] - E\left[X; a\right]\right) < \rho(a + l) - \rho(a).$$

That is, the risk load ascribed to the layer $(a, a + l]$ by the basic layer formula, namely $\rho(a + l) - \rho(a)$, is larger than the risk load based on the actual variance of the layer aggregate-loss random variable.

Overstatement of the risk load remains a technical problem when one uses the basic layer formula with risk-loaded ILFs to price excess layers of insurance. Ideally, one should first determine the layer premium by applying the basic formula with non-risk-loaded factors and then add on the risk load for the layer. Such an approach,

probably too cumbersome to be widely adopted, is more frequently used by reinsurers providing excess-of-loss coverage.

6.4. Aggregate Limits

It is often the case with liability lines of insurance that policies are written not only with a per-claim limit but also with an ***aggregate limit*** as well. Whereas the per-claim limit is the maximum the insurance company would pay on a single claim, an aggregate limit is the maximum amount that would be paid during the policy term for all claims combined.

The policy expected loss under the restrictions imposed by a per-claim limit l and an aggregate limit L (where $L > l$) is just the expected value $E[S_l; L]$, where S_l is the aggregate-loss random variable based on a claim-size variable N and a claim-size variable X limited at l. Thus, the unlimited aggregate mean is

$$E[S_l] = E[N]E[X; l]. \tag{6.25}$$

It is often more efficient and accurate to calculate the expected aggregate loss eliminated by the limit L—namely $E[S_l] - E[S_l; L]$—and subtract this amount from the unlimited mean (6.25) than it is to calculate $E[S_l; L]$ directly. For example, if the (unlimited) aggregate distribution function $F_S(s)$ has been approximated by one of the deterministic models discussed in Chapter 4, then expected excess loss $E[S_l] - E[S_l; L]$ can be obtained from the integral formula

$$\begin{aligned} E[S_l] - E[S_l; L] &= \int_0^\infty s\, dF_S(s) - \int_0^L s\, dF_S(s) - L(1 - F_S(L)) \\ &= \int_L^\infty (s - L)\, dF_S(s). \end{aligned} \tag{6.26}$$

In practice, the improper integral in (6.26) is most easily evaluated by numerical integration techniques. When a deterministic approximation is not practicable, then an approximation to $E[S_l; L]$ could be obtained by stochastic simulation.

Thus, when N is the claim-count variable, X is the unlimited indemnity-only claim-size variable, and allocated loss adjustment expense is loaded multiplicatively by the factor $1 + u$, the increased limit factor from the basic limit b with no aggregate limit to a per-claim limit l combined with an aggregate limit L is given by

$$I(l, L) = \frac{\psi E[S_l; L](1 + u)}{\psi E[N]E[X; b](1 + u)} = \frac{E[S_l; L]}{E[N]E[X; b]}. \tag{6.27}$$

The next example illustrates the use of formula (6.27).

Example 6.6. For a portfolio of liability policies the unlimited indemnity claim size distributed lognormally with $(\mu, \sigma) = (7.000, 2.400)$. Claim count N is distributed so that $E[N] = 1.20$ with contagion parameter $\gamma = 0.100$. Allocated loss adjustment expense is assumed to be 20% of the indemnity payment.

At the basic limit of 500,000 with no aggregate limit the expected policy indemnity loss is

$$E[S_{0.5M}] = E[N]E[X; 500{,}000] = (1.20)(21{,}743) = 26{,}092.$$

With a per-claim limit of 2,000,000 the expected loss is

$$E[S_{2M}] = E[N]E[X; 2{,}000{,}000] = (1.20)(28{,}338) = 34{,}006.$$

Consider now the case for which the per-claim limit of 2,000,000 is accompanied by an aggregate limit of 3,000,0000. In addition to the mean 34,006, the aggregate-loss variable S_{2M} has $SD[S_{2M}] = 151{,}311$ and $Sk[S_{2M}] = 9.4728$. A numerical integration of $\int_{3M}^{\infty}(s - 3{,}000{,}000)\, d\tilde{F}(s)$, where $\tilde{F}(s)$ is the shifted gamma approximation to $F_{S_{2M}}(s)$, yielded

$$E[S_{2M}] - E[S_{2M}; 3{,}000{,}000] = 91.$$

Thus,

$$E[S_{2M}; 3{,}000{,}000] = 34{,}006 - 91 = 33{,}915.$$

The ILF for the 2,000,000/3,000,000 limit combination is therefore

$$I(2M, 3M) = \frac{33{,}915}{26{,}092} = 1.2998.$$

Compare this result with the factor for the 2,000,000 per-claim limit with no aggregate limit:

$$I(2M) = \frac{34{,}006}{26{,}092} = 1.3033.$$

The expected policy aggregate losses for several combinations of per-claim and aggregate limits for this portfolio, as well as the increased limit factors calculated from them, are shown in Table 6.4. ■

6.5. Deductibles

The deductible is a coverage modification often used to decrease the policy claim count by eliminating small claims less than the deductible amount. It also serves possibly to encourage the policyholder to take steps to prevent or limit the occurrence of claims. We discuss in this section the standard straight deductible, as well as the less common franchise and diminishing deductible options.

By reducing the amount paid by the insurer for some or all claims, deductible provisions also serve to lower the premium charged for a policy. Deductible premium credits are easily calculated with the help of the claim-size limited pure premium and the loss elimination ratio concepts. In many cases, where there are sufficient data available, deductible credits can be calculated empirically. In other cases, analytic models involving parametric distributions are useful.

Table 6.4. Expected Aggregate Loss and ILFs with Aggregate Limits [Example 6.6]

Per-Claim Limit	Aggregate Limit ($000)					
	1,000	2,000	3,000	4,000	5,000	Unlimited
500,000	26,050	26,092	26,092	26,092	26,092	26,092
1,000,000	29,702	30,306	30,333	30,335	30,335	30,335
2,000,000	—	33,524	33,915	33,988	34,002	34,006
3,000,000	—	—	35,421	35,696	35,781	35,821
4,000,000	—	—	—	36,604	36,808	36,949
5,000,000	—	—	—	—	37,428	37,733
500,000	0.9984	1.0000	1.0000	1.0000	1.0000	1.0000
1,000,000	1.1384	1.1615	1.1625	1.1626	1.1626	1.1626
2,000,000	—	1.2848	1.2998	1.3026	1.3032	1.3033
3,000,000	—	—	1.3575	1.3681	1.3713	1.3729
4,000,000	—	—	—	1.4029	1.4107	1.4161
5,000,000	—	—	—	—	1.4345	1.4462

Straight Deductible

The ***straight deductible*** is the most common deductible coverage modification. It eliminates, from the standpoint of the insurer, all claims less than or equal to the deductible amount d, and it reduces the size of larger claims by d. If X represents the unmodified, ground-up claim-size variable, excluding allocated loss adjustment expense, then application of a straight deductible of size d yields a modified random variable, truncated from below and shifted by d:

$$X_d = X - d, \quad d < X < \infty. \tag{6.28}$$

Clearly, claims net of such a deductible are excess over an underlying limit d, as discussed in Section 5.1.

For all deductible options discussed in this section we shall assume that allocated loss adjustment expense is not included in the deductible or policy limit and that the policy severity is modeled by the general formula (6.15) with ALAE parameters ε and u. Thus, the basic-limit pure premium before application of the deductible is

$$p_b = \varphi\left(E[X; b] + \varepsilon\right)(1 + u), \tag{6.29}$$

where b is the basic limit and φ the claim frequency.

In practice, the basic-limit pure premium (6.29) is modified to reflect the presence of a deductible by applying a ***deductible credit factor*** $C(d)$:

$$p_{d,b} = p_b \cdot \left(1 - C(d)\right), \quad 0 < d < b. \tag{6.30}$$

Of course, this implies that the deductible-modified policy premium is obtained in the same way:

$$P_{d,b} = P_b \cdot (1 - C(d)). \tag{6.31}$$

The deductible premium credit amount is therefore $P_b\, C(d)$. Note also that the existence of the deductible reduces the basic limit policy layer width to $b - d$.

A formula for the deductible credit factor $C(d)$ is easily derived by starting with the modified basic-limit pure premium, calculated from first principles as the product of the policy frequency and severity:

$$\begin{aligned} p_{d,b} &= \varphi\,(1 - F_X(d))\left(\frac{E[X; b] - E[X; d]}{1 - F_X(d)} + \varepsilon\right)(1 + u) \\ &= \varphi\,(E[X; b] - E[X; d] + (1 - F_X(d))\,\varepsilon)(1 + u). \end{aligned} \tag{6.32}$$

Equating the two expressions for $p_{d,b}$ in (6.30) and (6.32), we solve for $C(d)$:

$$C(d) = \frac{\varphi\,(E[X; d] + F_X(d)\varepsilon)(1 + u)}{p_b} = \frac{E[X; d] + F_X(d)\varepsilon}{E[X; b] + \varepsilon}. \tag{6.33}$$

The expression in (6.33) is merely the ratio of the pure premium eliminated by the deductible to that of the unmodified policy layer [0, b]. Accordingly, the ratio is referred to as a ***loss elimination ratio***. The loss elimination ratio concept is useful in quantifying the effects of a variety of coverage modifications mandated by policy conditions.

Formula (6.31) yields the basic-limit premium modified by the straight deductible d, but how should the premium for a higher limit l be so adjusted? Recall that $P_l = P_b \cdot I(l)$, where $I(l)$ is the increased limit factor for limit l with respect to the basic limit b, and that the premium credit amount for the existence of the deductible is $P_b \cdot C(d)$. Therefore,

$$P_{d,l} = P_l - P_b\, C(d) = P_b \cdot (I(l) - C(d)). \tag{6.34}$$

Example 6.7 As in Example 6.3, consider a portfolio of policies for which the ground-up indemnity claim size X has a lognormal distribution with parameters $(\mu, \sigma) = (7.000, 2.400)$ and allocated loss adjustment expense is $u = 20\%$ of the indemnity amount. The basic limit is $b = 100{,}000$. We calculate the credit factors, as well as the resulting frequency and severity, for five straight deductible options: {1,000; 2,000; 3,000; 4,000; 5,000; 10,000}. Results are tabulated in Table 6.5. For example, equation (6.33) with $u = 20\%$ and $\varepsilon = 0$ implies that

$$C(2{,}000) = \frac{E[X; d]}{E[X; b]} = \frac{1{,}111}{8{,}896} = 0.1249.$$

Table 6.5. Straight Deductible Credit Factors [Example 6.7]

Ded d	$E[X; d]$	$F_X(d)$	$C(d)$	Frequency	Severity	Pure Prem
0	0	0.0000	0.0000	0.000500	10,675	5.338
1,000	659	0.4847	0.0741	0.000258	19,182	4.942
2,000	1,111	0.5989	0.1249	0.000201	23,291	4.671
3,000	1,478	0.6625	0.1661	0.000169	26,375	4.451
4,000	1,793	0.7051	0.2016	0.000147	28,903	4.262
5,000	2,071	0.7364	0.2328	0.000132	31,070	4.095
10,000	3,144	0.8215	0.3534	0.000089	38,669	3.451

The ground-up claim frequency for this portfolio is $\varphi = 0.000500$. Consequently, for a policy with deductible $d = 2{,}000$ we have an deductible-adjusted frequency

$$\varphi(1 - F_X(d)) = (0.000500)(1 - 0.5989) = 0.000201.$$

In addition, the modified severity is

$$\frac{E[X; b] - E[X; d]}{1 - F_X(d)}(1 + u) = \frac{8{,}896 - 1{,}111}{1 - 0.5989}(1.20) = 23{,}291,$$

yielding the pure premium $p = (0.000201)(23{,}291) = 4.671$.

In Example 6.3 we calculated the basic-limit premium for a policy with 400 exposure units to be \$3,285. Accordingly, premium for this policy with a limit of 1,000,000 and a 2,000 deductible is $P = (3{,}285)(1.8074 - 0.1249) = \$5{,}527$. ■

Franchise Deductible

The ***franchise deductible*** was one of the first coverage modifications to arise. Marine underwriters from the earliest times used it with policies insuring cargo shipments. It is now utilized in some types of workers' compensation coverages. The franchise deductible eliminates all claims less than or equal to the deductible or "franchise" amount d, and claims in excess of d are paid in full. Consequently, application of a franchise deductible d to the unlimited, ground-up random variable X results in the truncated, but non-shifted variable

$$X_d = X, \quad d < X < \infty.$$

In this case, the deductible-modified basic-limit pure premium is

$$p_{b,d} = \varphi(E[X; b] - E[X; d] + (1 - F_X(d))(d + \varepsilon))(1 + u). \tag{6.35}$$

It is easy to show that the deductible credit factor for a franchise deductible of size d and basic limit pure premium given by (6.29) is

$$C(d) = \frac{E[X; d] - d(1 - F_X(d)) + F_X(d)\varepsilon}{E[X; b] + \varepsilon}. \tag{6.36}$$

Example 6.8. We recalculate the deductible credit factors of Example 6.7 in the case that d is a franchise deductible. For instance,

$$C(2{,}000) = \frac{E[X; d] - d(1 - F_X(d))}{E[X; b]} = \frac{1{,}111 - (2{,}000)(1 - 0.5989)}{8{,}896} = 0.0347.$$

Moreover, the resulting claim frequency and severity for a deductible of size 2,000 are, respectively,

$$\varphi = 0.000201 \quad \text{and} \quad \left(\frac{8{,}896 - 1{,}111}{1 - 0.5989} + 2{,}000\right)(1.20) = 25{,}691.$$

The full set of results is displayed in Table 6.6. As one would expect, the premium credit for a franchise deductible is less than that for a straight deductible of equal size—the straight deductible eliminates a larger fraction of the pure premium than is eliminated by the corresponding franchise deductible. ■

Diminishing Deductible

The ***diminishing*** (or ***disappearing***) ***deductible*** is an alternative that incorporates features of both the straight and franchise deductibles. Such a policy modification eliminates all claims less than a positive deductible amount d and pays in full all claims in excess of a larger amount D, $D > d$. Claims between d and D in size are paid net of a deductible amount that declines linearly from size d at $X = d$ to 0 at $X = D$—that

Table 6.6. Franchise Deductible Credit Factors [Example 6.8]

Ded d	$E[X; d]$	$F_X(d)$	$C(d)$	Frequency	Severity	Pure Prem
0	0	0.0000	0.0000	0.000500	10,675	5.338
1,000	659	0.4847	0.0162	0.000258	20,382	5.251
2,000	1,111	0.5989	0.0347	0.000201	25,691	5.152
3,000	1,478	0.6625	0.0523	0.000169	29,975	5.058
4,000	1,793	0.7051	0.0690	0.000147	33,703	4.970
5,000	2,071	0.7364	0.0846	0.000132	37,070	4.886
10,000	3,144	0.8215	0.1528	0.000089	50,669	4.522

is, the deductible "disappears" at D. Thus, the deductible amount, as a function of the unrestricted claim value x, is

$$Ded(x) = \begin{cases} \dfrac{d}{D-d}(D-x) & \text{if } d \le x \le D \\ 0 & \text{if } D < x < \infty. \end{cases} \tag{6.37}$$

Like the franchise deductible, the diminishing deductible has the advantage of eliminating, from the standpoint of the insurer, numerous small claims while at the same time paying larger claims in full. However, the diminishing deductible can be more difficult to administer.

It is a straight-forward exercise to show that the deductible-modified random variable $X_{d,D}$ is defined for $X > d$ by

$$X_{d,D} = \begin{cases} \dfrac{D}{D-d}(X-d) & \text{if } d < X \le D \\ X & \text{if } D < X < \infty. \end{cases} \tag{6.38}$$

The distribution function for variable $X_{d,D}$ is therefore

$$F_{X_{d,D}}(x) = \begin{cases} 0 & \text{if } x \le 0 \\ \dfrac{F_X\left(\dfrac{D-d}{D}x + d\right) - F_X(d)}{1 - F_X(d)} & \text{if } 0 \le x \le D \\ \dfrac{F_X(x) - F_X(d)}{1 - F_X(d)} & \text{if } D < x < \infty. \end{cases} \tag{6.39}$$

In the case that allocated loss adjustment expense is loaded multiplicatively (with $\varepsilon = 0$ in (6.29)), the credit factor $C(d, D)$ for the disappearing deductible defined by d and D is

$$\begin{aligned} C(d,D) = \frac{1}{E[X;b]}\Bigg[& E[X;D] - D(1-F_X(D)) \\ & - \frac{D}{D-d}\left(E[X;D] - E[X;d] - D(1-F_X(D)) + d(1-F_X(d))\right) \\ & + \frac{dD}{D-d}\left(F_X(D) - F_X(d)\right)\Bigg]. \end{aligned} \tag{6.40}$$

Example 6.9. We now calculate the deductible credit factors for the policies of Example 6.7 with a diminishing deductible for which $D = d + 1{,}000$. The factors are displayed in Table 6.7, compared with those obtained in Examples 6.7 and 6.8.

Table 6.7. Deductible Credit Factors [Example 6.9]

Ded d	Straight $C(d)$	Diminishing $C(d,D)$	Franchise $C(d)$
1,000	0.0741	0.0233	0.0162
2,000	0.1249	0.0424	0.0347
3,000	0.1661	0.0599	0.0523
4,000	0.2016	0.0766	0.0690
5,000	0.2328	0.0917	0.0864

Note that for a given d the straight deductible eliminates the largest percent of the total policy loss, the franchise deductible the least percentage, and the diminishing deductible eliminating an amount between the two extremes. ■

Deductibles and Inflation

Because the characteristics of claims net of a straight deductible are similar to those of excess claims, the deductible exerts a comparable leveraging effect on an inflationary trend, as discussed in Section 5.5.

Suppose, for example, that the pure premium for a policy with a fixed straight deductible of size d is subjected to a uniform trend factor $\tau = 1 + r$. Assuming first that claim-size variable X is unlimited from above by the policy conditions, we calculate the trended pure premium:

$$p_r = \varphi(1+r)\big(E[X] - E[X; d/\tau] + (1 - F_X(d/\tau))\varepsilon\big).$$

The effective trend factor is therefore

$$\tilde{\tau} = 1 + \tilde{r} = (1+r)\frac{E[X] - E[X; d/\tau] + (1 - F_X(d/\tau))\varepsilon}{E[X] - E[X; d] + (1 - F_X(d))\varepsilon}. \tag{6.41}$$

Both $E[X; x]$ and $F_X(x)$ are nondecreasing functions of x, so that

$$0 < r \le \tilde{r} \quad \text{or} \quad \tilde{r} \le r < 0. \tag{6.42}$$

Thus, in the absence of other policy limits, the straight deductible magnifies the effect of a uniform trend.

However, if policy claims are limited by an upper limit b, then

$$\tilde{\tau} = (1+r)\frac{E[X; b/\tau] - E[X; d/\tau] + (1 - F_X(d/\tau))\varepsilon}{E[X; b] - E[X; d] + (1 - F_X(d))\varepsilon}. \tag{6.43}$$

The damping effect of the upper limit in this case sometimes prevents inequalities (6.42) from holding for certain combinations of b, d, and r. This phenomenon is illustrated in the next example.

Example 6.10. A policy has a straight deductible of $d = 500$. The ground-up claim-size variable X has a shifted Pareto distribution with $(\alpha, \beta) = (2; 8{,}000)$. Moreover, $\varphi = 0.25$ and $\varepsilon = 50$. If claims are unlimited by policy conditions, then the policy pure premium is

$$\begin{aligned} p_0 &= \varphi\left(E[X] - E[X; d] + (1 - F_X(d))\varepsilon\right) \\ &= (0.25)\big(8{,}000 - 471 + (1 - 0.1142)(50)\big) \\ &= 1{,}893. \end{aligned}$$

However, if claims are subjected to a 5% uniform trend, then

$$\begin{aligned} p_{5\%} &= \varphi\left((1.05)E[X] - (1.05)E[X; d/1.05] + (1 - F_X(d/1.05))(1.05)\varepsilon\right) \\ &= (0.25)\big(8{,}400 - 472 + (1 - 0.1092)(52.5)\big) \\ &= 1{,}994. \end{aligned}$$

Thus, the effective trend rate is

$$\tilde{r} = \frac{p_{5\%}}{p_0} - 1 = \frac{1{,}994}{1{,}893} - 1 = 5.3\%,$$

greater than the nominal 5%.

On the other hand, if policy conditions limit claims to $l = 5{,}000$, replace $E[X]$ in the above calculations with the limited severity $E[X; 5{,}000] = 3{,}007$. Then

$$\tilde{r} = \frac{p_{5\%}}{p_0} - 1 = \frac{677.3}{662.7} - 1 = 2.2\%. \tag{6.44}$$

Here the natural increase in the effective trend rate has been dampened by the presence of the policy limit. ■

6.6. Problems

6.1 For a certain liability coverage the claim frequency is $\varphi = 0.00075$ and the severity is $E[Y] = 6{,}000$. For these policies expenses are all variable with $v = 25\%$. For a policy for an insured with 2,500 exposure units, calculate:

(***a***) pure premium p.
(***b***) expected # policy claims $E[N]$.
(***c***) policy expected loss.
(***d***) loss-cost multiplier γ.
(***e***) rate R.
(***f***) policy premium P.

6.2 A product liability policy is issued for a premium of \$13,000. The insured's exposure amount is \$2,100,000 of product sales, and the exposure unit is \$1,000 of sales. For this line of business the severity at the policy limit is 9,950, and the policy has a loss-cost multiplier $\psi = 1.60$. Calculate:

(***a***) rate R.
(***b***) pure premium p.
(***c***) claim frequency φ.
(***d***) expected # policy claims $E[N]$.
(***e***) policy expected loss.
(***f***) variable expense ratio v.

6.3 For the policy of Problem 6.2 instead of treating all expenses as variable assume that there is fixed expense of $0.50 per exposure unit and variable expense is 30% of premium. Calculate:

(***a***) rate R.
(***b***) pure premium p.
(***c***) claim frequency φ.
(***d***) expected # policy claims $E[N]$.
(***e***) policy expected loss.

6.4 The average claim frequency for a portfolio of policies is $\varphi = 0.000800$ per policy year. If the claim-count distribution is Poisson, compute the probability that an individual annual policy selected from this portfolio will give rise to more than a single claim when the number of exposure units is

(***a***) 1,000.
(***b***) 2,000.

6.5 Assume that the distribution of the unlimited indemnity claim size X for a portfolio of policies is lognormally distributed with $(\mu, \sigma) = (6.800, 2.600)$.

(***a***) Complete the following table of increased limit factors based on formula (6.12) with an average per-claim ALAE = 2,500.

Limit l	$E[X; l]$	ALAE	$I(l)$
100,000	9,178	2,500	1.000
250,000	_____	_____	_____
500,000	_____	_____	_____
1,000,000	_____	_____	_____
2,000,000	_____	_____	_____
5,000,000	_____	_____	_____

(***b***) Alternatively, assume that ALAE is 25% of the indemnity payment. Complete the following table of increased limit factors based on formula (6.14).

Limit l	$E[X; l]$	ALAE = 25%	$I(l)$
100,000	9,178	2,295	1.000
250,000	_____	_____	_____
500,000	_____	_____	_____
1,000,000	_____	_____	_____
2,000,000	_____	_____	_____
5,000,000	_____	_____	_____

6.6 Consider the following alternative to the two methods of loading allocated loss adjustment expense in an ILF formula—the per-claim average amount ε of formula (6.12) and the fixed multiple u of the indemnity payment in formula (6.14). Here the ALAE for smaller claims is loaded as a percent of the indemnity payment, and for larger claims ALAE is fixed at a constant per-claim average.

Let X be the unlimited claim-size (indemnity only) random variable. Assume that the allocated loss adjustment expense is a fixed multiple r ($r > 0$) of claim size X whenever X is not larger than the claim size c and that ALAE has the constant value rc when $X > c$. Thus the policy claim amount is

$$Y = \begin{cases} X + rX & \text{if } X \le c \\ X + rc & \text{if } X > c. \end{cases}$$

Show that in this case the policy severity at limit l is

$$E_l[Y] = \begin{cases} E[X; l](1 + r) & \text{if } l \le c \\ E[X; l] + r\,E[X; c] & \text{if } l > c. \end{cases}$$

6.7 For the portfolio of policies of Example 6.3 construct an ILF table using the method of loading allocated loss adjustment expense described in Problem 6.6. Assume that $r = 20\%$ and $c = 750{,}000$. Compare the results to those obtained in Example 6.3.

Limit l ($000)	$I(l)$ $\varepsilon = 2{,}200$	$I(l)$ $u = 20\%$	$I(l)$ limited
100	1.0000	1.0000	1.0000
500	1.4263	1.5317	_____
750	1.5202	1.6488	_____
1,000	1.5812	1.7249	_____
2,000	1.7067	1.8815	_____
3,000	1.7655	1.9548	_____
4,000	1.8008	1.9989	_____
5,000	1.8248	2.0288	_____

6.8 (***a***) Construct the following ILF table using the risk-loaded formula (6.22). Assume that the unlimited indemnity claim size X has a shifted Pareto distribution with $(\alpha, \beta) = (3; 6{,}000)$ and that $\varepsilon = 0$, $u = 20\%$. Use the standard deviation method of risk loading (6.24) with $k' = 0.5000$ and $\delta = 0.1000$.

Limit l	$E[X; l]$	ALAE	$\rho(l)$	$I(l)$ w/o RL	$I(l)$ w/ RL	Weight
1,000	796	159	447	1.0000	1.0000	10%
2,000	_____	_____	_____	_____	_____	5%
3,000	_____	_____	_____	_____	_____	15%
4,000	_____	_____	_____	_____	_____	15%
5,000	_____	_____	_____	_____	_____	25%
7,500	_____	_____	_____	_____	_____	10%
10,000	_____	_____	_____	_____	_____	20%

(***b***) Calculate the overall premium effect of using the risk-loaded factors in place of the unloaded factors.

6.9 Show that the risk-load parameter δ of formulas (6.23) and (6.24) can be expressed as $\delta = \gamma E[N]$, where γ denotes the contagion parameter for the claim-count variable N.

6.10 Construct the following table of increased limit factors using the risk-loaded formula (6.22). Assume that the unlimited indemnity claim size X has a shifted Pareto distribution with $(\alpha, \beta) = (0.780, 100)$ and that $\varepsilon = u = 0$. Use the variance method (6.23) for calculating the risk-load function $\rho(l)$ with $k = 0.0000005$ and $\delta = 0$. Calculate the layer factors for successive layers of 1,000,000 width and thereby demonstrate the inconsistency of this set of ILFs.

Limit l	$E[X; l]$	$\rho(l)$	$I(l)$	Layer Factor
1,000,000	2,994	622	1.0000	—
2,000,000	3,562	1,448	1.3858	0.3858
3,000,000	3,936	2,375	1.7458	0.3600
4,000,000	_____	_____	_____	_____
5,000,000	_____	_____	_____	_____
6,000,000	_____	_____	_____	_____
7,000,000	_____	_____	_____	_____
8,000,000	_____	_____	_____	_____
9,000,000	_____	_____	_____	_____
10,000,000	_____	_____	_____	_____

6.11 Consider a policy selected from the portfolio of Example 6.6 with per-claim limit l. Calculate the loss eliminated by the addition of an aggregate limit of size l and obtain the resulting loss elimination ratio when l equals:
(***a***) 1,000,000. (***b***) 2,000,000. (***c***) 3,000,000.
(***d***) 4,000,000. (***e***) 5,000,000.

6.12 (***a***) Assuming that variable expenses and profit load are 25% of premium, calculate the premium for a policy selected from the portfolio of Example 6.6 with a per-claim limit of 1,000,000 and no aggregate limit.
(***b***) What is the premium credit if the policy in part (*a*) is written with an aggregate limit of 1,000,000?

6.13 The following set of twelve (unadjusted) losses are incurred on a policy with a per-claim limit of $l = 20{,}000$ and a deductible of size $d = 2{,}000$:

$$\{1{,}000;\ 1{,}550;\ 1{,}700;\ 2{,}200;\ 2{,}500;\ 3{,}000;$$

$$5{,}200;\ 9{,}000;\ 11{,}000;\ 12{,}500;\ 15{,}000;\ 19{,}800\}.$$

Compute the total amount paid by the insurer (exclusive of loss adjustment expense) and the empirical loss elimination ratio for this set of claims if the deductible is

(***a***) a straight deductible. (***b***) a franchise deductible.

6.14 Repeat the calculations requested in Problem 6.13 if the policy in that problem has a per-claim limit of $l = 12{,}000$.

6.15 Derive formula (6.33) for the franchise deductible credit factor.

6.16 (***a***) Calculate the straight deductible factors and corresponding pure premiums for the portfolio of policies of Example 6.7, this time with ALAE parameters $\varepsilon = 500$ and $u = 0$.

(***b***) Calculate the franchise deductible factors and corresponding pure premiums for the portfolio of policies of Example 6.8 with ALAE parameters of part (*a*).

6.17 A policy has a basic limit $b = 5{,}000$ and a deductible of size d. Assume that the underlying claim size variable X has a shifted Pareto distribution with $(\alpha, \beta)=(3.00;10{,}000)$. Assume also that the unlimited claim frequency is $\varphi = 0.005$ and $\varepsilon = 250$. Compute $C(d)$, the policy frequency and severity, and the pure premium for deductibles of sizes {0; 250; 500; 750; 1,000} in the case of

(***a***) a straight deductible. (***b***) a franchise deductible.

6.18 Prove that on the interval $0 < d < b$ the deductible credit factors (6.33) and (6.36)

(***a***) are increasing functions of d.

(***b***) satisfy the inequality $0 < C(d) < 1$.

6.19 Verify formulas (6.38), (6.39), and (6.40) for the diminishing deductible.

6.20 Verify inequalities (6.42) for a deductible-modified uniform trend rate.

6.21 A portfolio of policies described in Example 6.7 has a total expected ground-up claim count of 500.

(***a***) For each of the indicated straight deductibles calculate (*i*) the total number of claims eliminated by the deductible and (*ii*) the percent of the total ground-up basic-limit loss eliminated by the deductible.

Deductible	# Claims Eliminated	% BL Premium Eliminated
1,000	_____	_____
2,000	_____	_____
3,000	_____	_____
4,000	_____	_____
5,000	_____	_____

(***b***) Assuming that the claim-size X is subjected to a 10% uniform trend, perform the same calculations requested in part (*a*).

Deductible	# Claims Eliminated	% BL Premium Eliminated
1,000	_____	_____
2,000	_____	_____
3,000	_____	_____
4,000	_____	_____
5,000	_____	_____

6.22 Verify the calculation of the pure premium amounts used in equation (6.44).

Appendix

A.1. Distribution Approximation

Normal Distributions

The cumulative distribution function of the standard normal random variable Z is defined by the integral

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\tfrac{1}{2}u^2\right) du, \quad -\infty < z < \infty.$$

This integral cannot be evaluated by the elementary method involving an antiderivative of the integrand. Consequently, mathematicians have developed approximation formulas involving easily calculated expressions. One such formula, based on a rational function, is cited by Abramowitz and Stegun:[56]

$$\Phi(z) \approx \begin{cases} Q(-z) & \text{if } -\infty < z < 0 \\ 1 - Q(z) & \text{if } 0 \le z < \infty, \end{cases} \tag{A.1}$$

where

$$Q(z) = \frac{1}{2}\left(1 + \sum_{k=1}^{6} a_k z^k\right)^{-16}$$

and

$$\begin{cases} a_1 = 0.0498673470 & a_2 = 0.0211410061 & a_3 = 0.0032776263 \\ a_4 = 0.0000380036 & a_5 = 0.0000488906 & a_6 = 0.0000053830. \end{cases}$$

The error in approximation (A.1) is bounded by 1.50×10^{-7}.

For users of Microsoft Excel, the built-in worksheet function NORM.S.DIST provides an approximation with precision similar to that of (A.1):

[56] Abramowitz and Stegun [1], p. 932. Formula (A.1) is one of several approximations to Φ included in this standard reference work.

$$\text{NORM.S.DIST}(z, \text{TRUE}) \approx \Phi(z), \quad -\infty < z < \infty$$

$$\text{NORM.S.DIST}(z, \text{FALSE}) \approx \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right). \tag{A.2}$$

In addition, Excel provides the related worksheet function NORM.DIST, which returns values of the distribution functions for the normal random variable $Y = \sigma Z + \mu$ with parameters μ and σ ($\mu > 0$, $\sigma > 0$):

$$\text{NORM.S.DIST}(y, \mu, \sigma, \text{TRUE}) \approx F_Y(y) = \Phi((y-\mu)/\sigma), \quad -\infty < z < \infty,$$

$$\text{NORM.S.DIST}(y, \mu, \sigma, \text{FALSE}) \approx f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\tfrac{1}{2}(y-\mu)^2/\sigma^2\right). \tag{A.3}$$

For the purpose of Monte Carlo simulation it is also useful to have available an approximation to the inverse function $\Phi^{-1}(u)$. Abramowitz and Stegun offer the following rational function (as usual, $\log x$ denotes the natural logarithm function):[57]

$$\Phi^{-1}(u) \approx x - \frac{b_0 x + b_1 x + b_2 x^2}{1 + c_1 x + c_2 x^2 + c_3 x^3}, \tag{A.4}$$

where

$$x = \begin{cases} \sqrt{-2\log u} & \text{if } 0 < u \le 0.5 \\ \sqrt{-2\log(1-u)} & \text{if } 0.5 < u < 1 \end{cases}$$

and

$$\begin{cases} b_0 = 2.515517 & b_1 = 0.802853 & b_2 = 0.010328 \\ c_1 = 1.432788 & c_2 = 0.189269 & c_3 = 0.001308. \end{cases}$$

The error in (A.4) is bounded by 4.50×10^{-4}. Excel also provides the useful worksheet function

$$\text{NORM.S.INV}(u) \approx \Phi^{-1}(u), \quad 0 < u < 1. \tag{A.5}$$

Gamma Distributions

The gamma function, defined by the convergent improper integral

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du, \quad 0 < x < \infty,$$

[57] *Ibid.*, p. 933.

can be approximated on the interval $1 \le x \le 2$ by the polynomial

$$\Gamma(x) \approx 1 + \sum_{k=1}^{8} d_k (x-1)^k, \quad 1 \le x \le 2, \tag{A.6}$$

where

$$\begin{cases} d_1 = -0.577191652 & d_2 = 0.988205891 & d_3 = -0.897056937 \\ d_4 = 0.918206857 & d_5 = -0.756704078 & d_6 = 0.482199394 \\ d_7 = -0.193527818 & d_8 = 0.035868343. \end{cases}$$

This approximation, of course, can be extended to all positive x by use of the recursive formula $\Gamma(x) = (x-1)\Gamma(x-1)$. Error in (A.6) is bounded by 3×10^{-7}.[58] In Microsoft Excel, $\Gamma(x)$ can be calculated by using the composition of two worksheet functions: EXP(GAMMALN(x)) $\approx \Gamma(x)$ for $x > 0$.

In Section 2.3 we showed that the gamma cumulative distribution function (2.16) can be expressed by the formula

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \dfrac{\Gamma(x/\beta, \alpha)}{\Gamma(\alpha)} & \text{if } 0 \le x < \infty \quad (\alpha > 0, \beta > 0), \end{cases} \tag{A.7}$$

where $\Gamma(x, \alpha)$ is the incomplete gamma function:

$$\Gamma(x, \alpha) = \int_0^x u^{\alpha-1} e^{-u}\, du \quad (\alpha > 0), \quad 0 \le x < \infty. \tag{A.8}$$

The incomplete gamma function (A.8) has a power series expansion:

$$\Gamma(x,\alpha) = x^\alpha e^{-x}\, \Gamma(\alpha) \sum_{k=0}^{\infty} \frac{x^k}{\Gamma(\alpha+k+1)} = x^\alpha e^{-x} \frac{\Gamma(\alpha)}{\Gamma(\alpha+1)} \left(1 + \sum_{k=1}^{\infty} \frac{x^k}{(\alpha+1)\ldots(\alpha+k)} \right),$$

so that the gamma distribution function (A.7) has a corresponding power series representation

$$F(x) = \frac{(x/\beta)^\alpha\, e^{-x/\beta}}{\Gamma(\alpha+1)} \left(1 + \sum_{k=1}^{\infty} \frac{(x/\beta)^k}{(\alpha+1)\ldots(\alpha+k)} \right). \tag{A.9}$$

An approximation to the gamma distribution function (A.7) can thus be obtained by using an appropriate partial sum of the series (A.9).

[58] *Ibid.*, p. 257.

Again, for users of Microsoft Excel, the worksheet function GAMMA.DIST provides an approximation to both the probability density and the cumulative distribution functions:

$$\text{GAMMA.DIST}(x, \alpha, \beta, \text{FALSE}) \approx f(x) = \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad 0 \le x < \infty,$$

$$\text{GAMMA.DIST}(x, \alpha, \beta, \text{TRUE}) \approx F(x) = \int_0^x f(x)\,dx. \tag{A.10}$$

In addition, the worksheet function GAMMA.INV returns an approximation to the inverse cumulative distribution function:

$$\text{GAMMA.INV}(u, \alpha, \beta) \approx F^{-1}(u), \quad 0 \le u < 1. \tag{A.11}$$

Lognormal Distributions

For the lognormal distribution function Microsoft Excel provides two worksheet functions. LOGNORM.DIST approximates the lognormal density and cumulative distribution functions with parameters μ and σ ($\mu > 0$, $\sigma > 0$):

$$\text{LOGNORM.DIST}(x, \mu, \sigma, \text{TRUE}) \approx F(x) = \Phi\big((\log x - \mu)/\sigma\big), \quad 0 < x < \infty,$$

$$\text{LOGNORM.DIST}(x, \mu, \sigma, \text{FALSE}) \approx \frac{1}{\sigma\sqrt{2\pi}\,x} \exp\Big(-\tfrac{1}{2}(\log x - \mu)^2 / \sigma^2\Big). \tag{A.12}$$

LOGNORM.INV provides values of the inverse c.d.f.:

$$\text{LOGNORM.INV}(u, \mu, \sigma) \approx F^{-1}(u), \quad 0 \le u < 1. \tag{A.13}$$

Weibull Distributions

As in the case of the previously mentioned distributions, the single Excel worksheet function WEIBULL.DIST provides an approximation to both the probability density and the cumulative distribution functions of the Weibull distribution with parameters β and δ ($\beta > 0$, $\delta > 0$):

$$\text{WEIBULL.DIST}(x, \delta, \beta, \text{TRUE}) \approx F(x) = 1 - \exp\Big(-(x/\beta)^{\delta}\Big), \quad 0 \le x < \infty,$$

$$\text{WEIBULL.DIST}(x, \delta, \beta, \text{FALSE}) \approx f(x) = \frac{\delta}{\beta^{\delta}} x^{\delta-1} \exp\Big(-(x/\beta)^{\delta}\Big). \tag{A.14}$$

A.2. Answers to Selected Problems

1.1 (***b***) *Hint:* $(E^c \cup F^c)^c = E \cap F$.
(***c***) $S = \{\varnothing, \{a\}, \{d\}, \{a, d\}, \{b, c\}, \{a, b, c\}, \{b, c, d\}, \Omega\}$.

1.2 (*a*) $P(E) + P(E^c) = P(E \cup E^c) = P(\Omega) = 1$.
(*e*) $P(F) = P(F \cap E) + P(F \cap E^c) = P(E) + P(F \cap E^c) \geq P(E)$.

1.3 (*a*) $\lim_{n\to\infty} P(E_n) = \lim_{n\to\infty} P(\bigcup_{k=1}^{n} E_k) = P(\bigcup_n E_n)$.

1.4 (*a*) 0.1429. (*b*) 0.2857. (*c*) 0.8571.

1.5 (*b*) $P(E \cup F) = P(E) + P(F) - P(E)P(F) = P(E)P(F^c) + P(F) = P(E)P(F^c) + 1 - P(F^c) = 1 - P(E^c)P(F^c)$.

1.7 (*a*) 0.2500. (*b*) 0.1875. (*c*) 0.8125. (*d*) 3.8125. (*e*) 1.7773.

1.10 (*b*) $(1-p)/p$. (*c*) $E[N] = 1/(1-p)$, $Var[N] = p/(1-p)^2$.

1.11 (*a*) 0.7500. (*b*) 0. (*c*) 0.8484. (*d*) 0.0597.
(*e*) 0.5000. (*f*) 1.7500.

1.12 (*b*) $\Pr\{X = x\} = 0$ implies $\Pr\{X < x\} = \Pr\{X \leq x\} = F(x)$ for all x.

1.13 (*a*) 2/3. (*b*) 4. (*c*) 1. (*d*) 15.

1.14 (*a*) 1.5. (*b*) 3. (*c*) $0.125 + 0.375e^t + 0.375e^{2t} + 0.125e^{3t}$.

1.15 (*a*) $\frac{1}{2}(\alpha + \beta)$. (*b*) $\frac{1}{3}(\alpha^2 + \alpha\beta + \beta^2)$. (*c*) $(e^{\beta t} - e^{\alpha t})/((\beta - \alpha)t)$.

1.16 (*b*) $E[X] = 180$, $Var[X] = 35{,}600$.

1.17 $E[X] = \frac{1}{2}(\alpha + \beta)$, $Var[X] = \frac{1}{12}(\beta - \alpha)^2$.

1.18 $E[X_a] = 75.00$, $Var[X_a] = 9{,}375$.

1.19 $E[Y] = \mu$, $Var[Y] = \sigma^2$.

1.20 (*b*) $\exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.

1.21 (*a*) $pe^t/[1 - (1-p)e^t]$, $t < \log(1-p)$. (*b*) $1/p$. (*c*) $(1-p)/p^2$.
(*d*) $p/[1 - (1-p)^2]$.

1.23 $\hat{\beta} = M_1$.

2.1 (*a*) 1,500. (*b*) 750,000. (*c*) 0.1250. (*d*) 1,375.

2.2 (*a*) 875. (*b*) does not exist. (*c*) 0.2500. (*d*) 687.50.

2.3 $(200)(1 - e^{-y/250})$.

2.4 $E[\hat{X}] = 1{,}532$; $E[\hat{X}; 1{,}000] = 892$, $E[\hat{X}; 1{,}500] = 1{,}203$.

2.6 $M_1 \approx \frac{1}{n}\sum_{k=1}^{m} n_k a_k$.

2.7 $E[X; x] \le \int_0^x u\, dF(u) + x < \infty$.

2.9 (***a***) $\frac{d}{dx}[\int_0^x u\, f(u)du + x(1 - F(x))] = 1 - F(x)$. (***b***) Use $E[X; 0] = 0$ with (*a*).

2.10 (***b***) Set $v = u^2$. (***c***) Set $v = \log(1/u)$.

2.11 (***b***) Integration by parts. (***d***) Apply (2.18) inductively to $\Gamma(x+1)/\Gamma(x) = x$.

2.13 $\alpha = 1$, $\beta = E[X]$.

2.14 $\Pr\{X > a + b \mid X > a\} = \Pr\{X > b\}$.

2.15 (***a***) $\omega\beta_1 + (1 - \omega)\beta_2$. (***b***) $\omega\beta_1^2 + (1 - \omega)\beta_2^2 + \omega(1 - \omega)(\beta_1 - \beta_2)^2$.

2.16 Excel Solver yields $(\hat{\alpha}, \hat{\beta}) = (4.7432, 337.31)$.[59]

2.17 (***b***) 556. (***d***) 0.1461. (***e***) 0.2108. (***f***) 1,023. (***g***) 9,689.

2.18 (***b***) $(\hat{\mu}, \hat{\sigma}) = (7.235292, 0.477366)$.

2.19 (***b***) 1,040. (***d***) 0.1866. (***e***) 0.3254. (***f***) 1,347. (***g***) 6,500.

2.20 $Var[X] = \frac{\alpha}{\alpha - 2}(E[X])^2 > (E[X])^2$.

2.21 $\beta(\log(x + \beta) - \log\beta)$.

2.24 (***a***) $\beta \log 2$. (***b***) e^{μ}. (***c***) $\beta(2^{1/\alpha} - 1)$.

2.25 (***a***) 1,282. (***b***) 1,315. (***c***) 1,428.

2.26 *Hint:* $E[(L(X) - E[L(X)])^3] = a^3 E[(X - E[X])^3]$, $Var[L(X)] = a^2 Var[X]$.

2.27 $L(X) = \beta(X/\gamma - 1)$.

2.28 (***a***) exponential (2). (***b***) shifted Pareto (α, β). (***c***) Burr (α, β, δ).
(***d***) Weibull (β, δ). (***e***) exponential $(1/\alpha)$.

2.29 (***a***) $\beta^{m/\delta}\,\Gamma(\alpha - m/\delta)\,\Gamma(1 + m/\delta)/\Gamma(\alpha)$, $\alpha\delta > m$.

2.30 (***a***) $F_Y(y) = \begin{cases} 0 & \text{if } -\infty < y \le 1 \\ \dfrac{\Gamma((\log y)/\beta, \alpha)}{\Gamma(\alpha)} & \text{if } 1 < y < \infty. \end{cases}$ (***b***) $E[Y^m] = (1 - m\beta)^{-\alpha}$.

2.31 $CV[\tau X] = \sqrt{Var[\tau X]}\,/\,E[\tau X] = (|\tau|/\tau)CV[X] = CV[X]$.

---

[59] Results obtained by an iterative algorithm applied by the Excel Solver may vary slightly, depending on how the problem is set up in the worksheet and the process is initiated.

2.32 $\tilde{\tau} = \tau E[X; \tau l/\tau]/E[X; l] = \tau$.

2.33 (***a***) Weibull $(a\beta^{b+1}, 1/(b+1))$. (***b***) Burr $(\alpha, \alpha\beta^{b+1}, 1/(b+1))$.
(***c***) Weibull $(a\beta^{b+1}, \delta/(b+1))$. (***d***) Burr $(\alpha, \alpha\beta^{b+1}, \delta/(b+1))$.

2.34 (***a***) 0.6321. (***b***) $\Gamma(\alpha, \alpha)/\Gamma(\alpha)$. (***c***) $\Phi(\sigma/2)$. (***d***) $1 - (1 - 1/\alpha)^{\alpha}$.

2.37 (***a***) none. (***b***) β. (***c***) none. (***d***) β. (***e***) β. (***f***) none.

2.38 $f_{c\theta}(x) = (1/c)\, f_{\theta}(x/c) = (1/c\theta)\, f_1(x/c\theta)$.

2.39 (***a***) Excel Solver yields $(\hat{\mu}, \hat{\sigma}) = (9.778102, 1.444776)$.
(***c***) $\chi^2 = 5.89 < \chi^2_{0.95}(5) = 11.1$.

2.40 Excel Solver yields $(\hat{\mu}, \hat{\sigma}) = (9.701968, 1.535797)$.

2.41 (***a***) $F_Y(y) = \begin{cases} F_X(y)/F_X(l) & \text{if } -\infty < y < l \\ 1 & \text{if } l \le y < \infty. \end{cases}$

2.42 (***a***) $F_Y(y) = \begin{cases} 0 & \text{if } -\infty < y \le a \\ \dfrac{F_X(y) - F_X(a)}{1 - F_X(a)} & \text{if } a \le y < \infty. \end{cases}$

2.43 (***a***) Excel Solver yields $(\hat{\mu}, \hat{\sigma}) = (9.495111, 1.084180)$.
(***b***) $M_1 = 14{,}840$, $E[Y] = 14{,}930$.

3.1 (***a***) $M'_N(0) = mp$, $M''_N(0) = mp + m(m-1)p^2$.
(***b***) $\lim\limits_{\substack{m\to\infty \\ mp=\lambda}} M_N(t) = \lim\limits_{m\to\infty} \left(1 + \dfrac{\lambda}{m}\left(e^t - 1\right)\right)^m = \exp\left(\lambda\left(e^t - 1\right)\right)$.

3.3 (***b***) 2.50, 0.7576. (***c***) 6.25, 0.1303. (***d***) 2.00, 0.8571.
(***e***) 4.00, 0.4335.

3.4 (***a***) 0.0012. (***b***) 0.0069. (***c***) 0.0164.

3.7 $\dfrac{d}{d\lambda} \log L(\lambda) = -m + \Sigma_{i=1}^{m}\, n_i/\lambda$.

3.8 Integrate $\int_{\lambda}^{\infty} t^n e^{-t}\, dt/n!$ by parts n times.

3.10 (***a***) $E\left[N^2\right] = \displaystyle\sum_{n=0}^{\infty} n^2 \int_0^{\infty} \frac{u^n e^{-u}}{n!} f_{\lambda}(u)\, du = \int_0^{\infty} \left(\sum_{n=0}^{\infty} n^2 \frac{u^n e^{-u}}{n!} \right) f_{\lambda}(u)\, du$

$$= \int_0^{\infty} \left(u + u^2\right) f_{\lambda}(u)\, du = E[\lambda] + E\left[\lambda^2\right].$$

3.11 (***a***) 2. (***b***) 4. (***c***) 0.8125.

3.12 0.9070, 0.0864, 0.0062, 0.0004.

3.13 (***b***) 0.7500, 0.8175, 1.3125.
(***c***)

*n**i*	(1)	(2)	(3)
0	0.4724	0.4869	0.5714
1	0.3543	0.3373	0.2449
2	0.1329	0.1283	0.1050
3	0.0332	0.0365	0.0450
4	0.0062	0.0087	0.0193
5	0.0009	0.0018	0.0083

3.14 (***a***) Excel Solver yields $\hat{\lambda}_1 = 0.114493$, $\hat{\lambda}_2 = 0.797855$, $\hat{p}_1 = 0.986380$.

3.15 Method-of-moments estimates are $(\hat{\alpha}, \hat{\nu}) = (0.685714, 0.120000)$. Using 4 cells {0, 1, 2, ≥3 claims}, $\chi^2 = 0.3288 < 3.8415 = \chi^2_{0.95}(1)$.

3.16 $$\binom{-r}{n} = \frac{-r(-r-1)(-r-2)\cdots(-r-n+1)}{n!} = (-1)^n \frac{(r+n-1)\cdots(r+2)(r+1)r}{n!}.$$

3.19 $a = 1 - q$, $b = (r-1)(1-q)$.

3.21 (***a***) Set $r = 1$, $q = p$ in (3.19). (***b***) $E[N] = (1-p)/p$, $Var[N] = (1-p)/p^2$.

3.22 (***a***) $f_N(n) = 1/((n+1)(n+2))$. (***b***) $f_N(n) = (0.9)^{n+1}/((n+1)\log 10)$.

3.23 (***a***) Substitute $\beta = \nu/\alpha$ into (2.26).

3.24 $\hat{\alpha} = M_1^2/(M_2 - M_1^2 - M_1)$, $\hat{\nu} = M_1$.

3.25 (***a***) 0.6316. (***b***) 0.6316. (***c***) 0.6667.

3.26 (***a***) $\gamma = 0.05$. (***b***) 0.8000, 0.8077, 0.7692, 0.8000.
(***c***) 0.0034, 0.0342, 0.1538, 0.3761, 0.4325.

3.27 *Hint:* Divide numerator and denominator of (3.26) by c^m. Observe that

$$(w+ic)/c = \frac{1}{\gamma} + i, (r+ic)/c = \frac{1-p}{\gamma p} + i, (w+r+ic)/c = \frac{1}{\gamma p} + i.$$

3.28 0.4019, 0.3349, 0.1674, 0.0651, 0.0217, 0.0065.

3.29 (***a***) $\gamma = 0.1000$.

3.30 0.5543.

3.31 624.

3.32 $\lim_{m\to\infty} \gamma_m = 0$.

3.33 (***a***) $$\frac{d}{dt}F_n(t) = \sum_{k=n}^{\infty}\left(\frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{(k-1)!} - \frac{\lambda(\lambda t)^k e^{-\lambda t}}{k!}\right) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}.$$
(***b***) $E[T_n] = n/\lambda$, $Var[T_n] = n/\lambda^2$.

3.34 (***a***) 4.651 years. (***b***) 0.1935, 0.1560. (***c***) 0.0201, 0.0497.

3.35 N^* is Poisson-distributed with parameter $p\lambda$.

4.1

Amount s	$F_S(s)$	Amount s	$F_S(s)$
0	0.2000		
500	0.2400	3,500	0.9047
1,000	0.4025	4,000	0.9507
1,500	0.5427	4,500	0.9781
2,000	0.6795	5,000	0.9934
2,500	0.7580	5,500	0.9988
3,000	0.8418	6,000	1.0000

4.4 (***a***) $F_S(s) = f_N(0) + f_N(1)F_X(s)$.
(***b***) $F_S(s) = f_N(0) + f_N(1)F_X(s) + f_N(2)(F_X * F_X)(s)$.

4.6 $f_S(s) = e^{-\lambda - s/\beta}\sum_{n=1}^{\infty}\frac{\lambda^n s^n}{\beta^n n!(n-1)!}$, $F_S(s) = e^{-\lambda - s/\beta}\sum_{n=0}^{\infty}\frac{\lambda^n}{n!}\cdot\sum_{k=n}^{\infty}\frac{s^k}{\beta^k k!}$, $s > 0$.

4.7 See Problem 4.13 solution.

4.11 $\chi^2_{0.95}(m) \approx m\left(\sqrt{2/(9m)}\left(\Phi^{-1}(0.95) - \sqrt{2/(9m)}\right) + 1\right)^3$.

4.12

d.f. m	$\chi^2_{0.95}(m)$	W-H	Rel Error
5	11.070	11.044	–0.24%
10	18.307	18.292	–0.08%
15	24.996	24.985	–0.04%
20	31.410	31.402	–0.03%
25	37.652	37.645	–0.02%
30	43.773	43.767	–0.01%

4.13

Amount s	$F_S(s)$	Normal	Relative Error	Normal Power	Relative Error	Shifted Gamma	Relative Error	Wilson-Hilferty	Relative Error
0	0.0003	0.1241	—	0.0756	—	0.0263	—	0.0312	—
3,000	0.3420	0.2819	–17.57%	0.3654	+6.84%	0.3362	–1.70%	0.3322	–2.87%
6,000	0.6070	0.5000	–17.63%	0.5981	–1.47%	0.6054	–0.26%	0.6043	–0.44%
9,000	0.7774	0.7181	–7.63%	0.7608	–2.14%	0.7782	+0.10%	0.7797	+0.30%
12,000	0.8782	0.8759	–0.26%	0.8642	–1.59%	0.8793	+0.13%	0.8810	+0.32%
15,000	0.9349	0.9584	+2.51%	0.9257	–0.98%	0.9356	+0.07%	0.9367	+0.19%
18,000	0.9658	0.9895	+2.45%	0.9605	–0.55%	0.9661	+0.03%	0.9666	+0.08%
21,000	0.9823	0.9981	+1.61%	0.9796	–0.27%	0.9824	+0.01%	0.9824	+0.01%
24,000	0.9910	0.9997	+0.88%	0.9896	–0.14%	0.9909	–0.01%	0.9908	–0.02%
27,000	0.9954	1.0000	+0.46%	0.9948	–0.06%	0.9953	–0.01%	0.9952	–0.02%

4.15 All terms in the sum (4.33) for which $k > \hat{m}$ are zero.

4.16

Amount s	$F_S(s)$
0	0.2592
500	0.2942
1,000	0.4366
1,500	0.5606
2,000	0.6838
3,000	0.8306
4,000	0.9223
5,000	0.9670
6,000	0.9872

4.17 $g(0) + \sum_{k=1}^{\hat{m}-1} g(k) + g(\hat{m}) =$

$$F_X\left(\tfrac{1}{2}h\right) - \left[F_X\left(\tfrac{1}{2}h\right) + F_X\left(\left(\hat{m} - \tfrac{1}{2}\right)h\right)\right] + 1 - F_X\left(\left(\hat{m} - \tfrac{1}{2}\right)h\right) = 1.$$

4.20 If $Y = F_X(X)$, then $F_Y(y) = \Pr\{F_X(X) \le y\} = \Pr\{X \le F_X^{-1}(y)\} = y$ for $0 < y < 1$.

4.21 $\widetilde{F}^{-1}(u) = \beta(-\log(1-u))^{1/\delta}$, $0 < u < 1$.

4.22 $n = 5$.

4.23

Trial	Uniform u	Exponential x_1	Pareto x_2	Lognormal x_3	Weibull x_4
(1)	0.2097	471	296	30	55
(2)	0.3562	881	578	79	194
(3)	0.6970	2,388	1,837	553	1,426
(4)	0.8245	3,480	3,017	1,384	3,028
(5)	0.9882	8,879	14,716	25,871	19,711

4.25 (***a***) Because $E\{U_n\} = \dfrac{1}{2}$ and $Var[U_n] = \dfrac{1}{12}$, $E[X] = 0$ and $Var[X] = 1$. The Central Limit Theorem implies that X is approximately normal.

5.1 (***a***) 865; 440,343. (***b***) 7.4. (***d***) 6.6%, 9.5%, 16.7%.

5.2 (***a***) 500. (***b***) 1,250. (***c***) 2,000. (***d***) 3,500. (***e***) 8,518.

5.7 (***a***) $e^{-x/\beta}$. (***b***) $((d+\beta)/(x+d+\beta))^{\alpha}$.

5.9 (***a***)
$$e_X(x) = \frac{\int_0^\infty u\,dF(u) - \left(\int_0^x u\,dF(u) + x\int_x^\infty dF(u)\right)}{1 - \int_0^x dF(u)} = \frac{\int_x^\infty (u-x)\,dF(u)}{\int_x^\infty dF(u)}.$$

5.10 (***a***) 1.1551, 1.5959. (***b***) 136.38, 1.1487, 2.2616.

5.11 (*a*) 1. (*b*) $\frac{1}{\alpha}\sqrt{\alpha}$. (*c*) $\sqrt{\exp(\sigma^2)-1}$.
(*d*) $\sqrt{\alpha/(\alpha-2)}$. (*e*) $\frac{1}{3}\sqrt{3}$.

5.12

Layer L	P_L	$E[N_L]$	$E[X_L]$	$E[S_L]$
[0, 100]	1.0000	5.0000	97.08	485
(100, 3000]	0.9423	4.7116	1,513.66	7,132
(3000, ∞)	0.2441	1.2207	4,000.00	4,883

5.13 *Hint:* $\mu_1 = E[X; b_1]$, $\mu_k = (E[X; b_k] - E[X; b_{k-1}])/p_k$, $1 < k < m$.

5.14 (*b*) $\lambda(F(b) - F(a))$. (*c*) $E[X \mid a < X \le b] = \int_a^b x\, dF(x) / \int_a^b dF(x)$.

5.15 Interval means: 60; 317; 604; 1,405; 3,214; 4,400; 7,500.
Layer means: 96; 339; 342; 694; 1,371; 743; 2,500.

5.16
$$E\left[N_a^2\right] = \sum_{n=0}^{\infty} n^2 \sum_{k=n}^{\infty} \binom{k}{n} p^n (1-p)^{k-n} \Pr\{N=k\}$$
$$= \sum_{k=0}^{\infty} \Pr\{N=k\} \sum_{n=0}^{k} n^2 \binom{k}{n} p^n (1-p)^{k-n}$$
$$= \sum_{k=0}^{\infty} \Pr\{N=k\}\left(k^2p^2 + kp - kp^2\right) = E\left[N^2\right]p^2 + E[N]p - E[N]p^2.$$

5.18
$$\gamma_a = \frac{Var[N_a] - E[N_a]}{(E[N_a])^2} = \frac{p^2 Var[N] + p(1-p)\lambda - p\lambda}{p^2\lambda^2} = \frac{Var[N]-\lambda}{\lambda^2} = \gamma.$$

5.20
$$\tilde{\tau}_S = \tau_X \frac{E[X] - E[X; (\tau_X a)/\tau_X]}{E[X] - E[X; a]} = \tau_X.$$

5.21 (*a*) 8.9%. (*b*) 11.5%.

5.22 (*a*) 9.5%. (*b*) 10.2%.

5.23 (*a*) $(\hat{\mu}, \hat{\sigma}) = (7.960294, 1.428801)$. (*b*) 125.

5.24 $E[X] - E[X; x] = \int_0^{\infty}(1 - F(u))du - \int_0^{x}(1 - F(u))du$.

6.1 (*a*) 4.50. (*b*) 1.875. (*c*) 11,250. (*d*) 1.3333. (*e*) 6.00. (*f*) 15,000.

6.2 (*a*) 6.1905. (*b*) 3.8690. (*c*) 0.0003888. (*d*) 0.8166. (*e*) 8,125. (*f*) 0.3750.

6.3 (*a*) 6.1905. (*b*) 3.8333. (*c*) 0.0003853. (*d*) 0.8090. (*e*) 8,050.

6.4 (*a*) 0.1912. (*b*) 0.4751.

6.5

Limit l	$E[X; l]$	ALAE [*a*]	$l(l)$ [*a*]	ALAE [*b*]	$l(l)$ [*b*]
100,000	9,178	2,500	1.0000	2,295	1.0000
250,000	12,548	2,500	1.2885	3,137	1.3671
500,000	15,180	2,500	1.5139	3,795	1.6539
1,000,000	17,702	2,500	1.7299	4,426	1.9288
2,000,000	19,968	2,500	1.9240	4,992	2.1756
5,000,000	22,404	2,500	2.1325	5,601	2.4410

6.7 ILFs: 1.0000, 1.5317, 1.6488, 1.7122, 1.8427, 1.9038, 1.9405, 1.9655.

6.8 (*a*)

Limit l	$E[X; l]$	ALAE	$\rho(l)$	$l(l)$ w/o RL	$l(l)$ w/ RL
1,000	796	159	447	1.0000	1.0000
2,000	1,313	263	778	1.6490	1.6787
3,000	1,667	333	1,034	2.0940	2.1645
4,000	1,920	384	1,238	2.4123	2.5267
5,000	2,107	421	1,404	2.6478	2.8055
7,500	2,407	481	1,710	3.0247	3.2805
10,000	2,578	516	1,919	3.2392	3.5759

6.9 $1 + \delta = Var[N]/E[N] = (E[N] + \gamma(E[N])^2)/E[N] = 1 + \gamma E[N]$.

6.10

Limit l	$E[X; l]$	$\rho(l)$	$l(l)$	Layer Factor
4,000K	4,223	3,374	2.1014	0.3556
5,000K	4,459	4,429	2.4585	0.3571
6,000K	4,660	5,533	2.8194	0.3609
7,000K	4,836	6,678	3.1849	0.3655
8,000K	4,994	7,859	3.5553	0.3705
9,000K	5,137	9,074	3.9309	0.3755
10,000K	5,268	10,319	3.4114	0.3806

6.11 (*a*) 0.0209. (*b*) 0.0142. (*c*) 0.0112. (*d*) 0.0093. (*e*) 0.0081.

6.12 (*a*) $40,447. (*b*) $844.

6.13 (*a*) 62,200; 0.2635. (*b*) 80,200; 0.0503.

6.14 (*a*) 50,900; 0.3973. (*b*) 68,900; 0.1841.

6.16 (*a*)

Ded d	$C(d)$	Pure Prem
1,000	0.0959	4.247
2,000	0.1501	3.993
3,000	0.1926	3.973
4,000	0.2283	3.625
5,000	0.2596	3.478
10,000	0.3783	2.921

(*b*)

Ded d	$C(d)$	Pure Prem
1,000	0.0411	4.505
2,000	0.0647	4.394
3,000	0.0848	4.300
4,000	0.1028	4.215
5,000	0.1193	4.137
10,000	0.1884	3.813

6.17 (*a*)

Ded d	$C(d)$	Frequency	Severity	Pure Prem
0	0.0000	0.005000	3,028	15.139
250	0.0855	0.004643	2,982	13.845
500	0.1648	0.004319	2,928	12.644
750	0.2385	0.004025	2,864	11.528
1,000	0.3071	0.003757	2,792	10.489

(*b*)

Ded d	$C(d)$	Frequency	Severity	Pure Prem
250	0.0088	0.004643	3,232	15.006
500	0.0221	0.004319	3,428	14.804
750	0.0391	0.004025	3,614	14.547
1,000	0.0590	0.003757	3,792	14.246

6.21 (*a*)

Ded	# Claims	% BL Prem
1,000	242	7.4%
2,000	299	12.5%
3,000	331	16.6%
4,000	353	20.2%
5,000	368	23.3%

(*b*)

Ded	# Claims	% BL Prem
1,000	234	7.1%
2,000	292	12.0%
3,000	324	16.1%
4,000	346	19.5%
5,000	362	22.6%

A.3. References

[1] Abramowitz, Milton, and Irene Stegun, eds. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover Publications, 1972 (reprint of a work originally published by the National Bureau of Standards, United States Department of Commerce, 1964).

[2] Apostol, Tom M. *Mathematical Analysis,* 2nd ed. Reading MA: Addison-Wesley, 1974.

[3] Beard, R. E., T. Pentikäinen, and E. Pesonen. *Risk Theory: The Stochastic Basis of Insurance,* 3rd ed. London: Chapman and Hall, 1984.

[4] Feller, William. *An Introduction to Probability Theory and Its Applications,* Vol. I, 3rd ed. New York: John Wiley & Sons, Inc., 1968.

[5] Heckman, Philip E., and Glenn G. Meyers. "The calculation of aggregate loss distributions from claim severity and claim count distributions," *Proceedings of the Casualty Actuarial Society,* LXX (1983), 22–61. Additional exhibits and FORTRAN code are in *PCAS,* LXXI (1984), 49–66.

[6] Hewitt, Charles C., Jr. "The negative binomial applied to the Canadian merit rating plan for individual automobile risks," *Proceedings of the Casualty Actuarial Society,* XLVII (1960), 55–65.

[7] Hogg, Robert V., and Allen T. Craig. *Introduction to Mathematical Statistics,* 4th ed. New York: Macmillan Publishing Co., Inc., 1978.

[8] Hogg, Robert V., and Stuart A. Klugman. *Loss Distributions.* New York: John Wiley & Sons, Inc., 1984.

[9] Insurance Services Office. *Report of the Increased Limits Subcommittee: A Review of Increased Limits Ratemaking.* New York: Insurance Services Office, Inc., 1980.

[10] Kauppi, Laurie, and Pertti Ojantakanen. "Approximations of the generalized Poisson function," *ASTIN Bulletin,* V (1969), 213–226.

[11] Keatinge, Clive. "Modeling losses with the mixed exponential distribution," *Proceedings of the Casualty Actuarial Society,* LXXXVI (1999), 654–698.

[12] Klugman, Stuart A., and A. Rahulji Parsa. "Minimum distance estimation of loss distributions," *Proceedings of the Casualty Actuarial Society,* LXXX (1993), 250–270.

[13] Lindgren, B. W. *Statistical Theory,* 2nd ed. New York: The Macmillan Company, 1968.

[14] McCord, James R., III, and Richard M. Moroney, Jr. *Introduction to Probability Theory.* New York: The Macmillan Company, 1964.

[15] Meyers, Glenn. *The competitive market equilibrium risk load formula for increased limits ratemaking.* New York: Insurance Services Office, Inc., 1991.

[16] Miccolis, Robert S. "On the theory of increased limits and excess of loss pricing," *Proceedings of the Casualty Actuarial Society,* LXIV (1977), 27–59.

[17] Panjer, H. H. "Recursive evaluation of a family of compound distributions," *ASTIN Bulletin,* XII (1981), 21–26.

[18] Parzen, Emanuel. *Modern Probability Theory and Its Applications.* New York: John Wiley & Sons, Inc., 1960.

[19] Pentikäinen, T. "Approximative evaluation of the distribution function of aggregate claims," *ASTIN Bulletin,* XVII (1987), 15–39.

[20] Rosenberg, Sheldon, and Aaron Halpert. "Adjusting size of loss distributions for trend," *Inflation Implications for Property-Casualty Insurance: Casualty Actuarial Society 1981 Discussion Paper Program,* 458–494.

[21] Simon, Leroy J. "Fitting negative binomial distributions by the method of maximum likelihood," *Proceedings of the Casualty Actuarial Society,* XLVIII (1961), 45–53.

[22] Weibull, E. H. W. "A statistical distribution function of wide applicability," *Journal of Applied Mechanics,* XVIII (1951), 293–297.

About the Author

A native of Minnesota, David Bahnemann studied mathematics and statistics at the University of Minnesota and at Stanford University. After teaching mathematics at Northwest Missouri State University for 18 years, he joined the actuarial department at the St. Paul Companies in St. Paul, Minnesota. For the next 25 years he provided actuarial support to several excess and surplus lines underwriting departments. While at the St. Paul (later known as St. Paul Travelers, and then Travelers) he was involved in large-account and program pricing. During this time he also created several computer-based pricing tools for use by both underwriters and actuaries. During retirement he divides his time between White Bear Lake and Burntside Lake in Minnesota.

ABOUT THE SERIES:

CAS monographs are authoritative, peer-reviewed, in-depth works focusing on important topics within property and casualty actuarial practice. For more information on the CAS Monograph Series, visit the CAS website at www.casact.org.

Expertise. Insight.
Solutions.

www.casact.org

Distributions for Actuaries
Errata

Page 145, Table, line 10: Replace 4 [claims] with 3 [claims]

Table, line 11: Replace 14 [claims] with 15 [claims]

line 8: Replace $n_{10} = 14$ with $n_{10} = 15$

line 13: Replace $k = 1, 2, \ldots, 9$ with $k = 0, 1, 2, \ldots, 9$

Page 170, line 12: Replace $f_X(x) = F'_X(x)$ with $f_X(x) = F'_X(x) > 0$

line 14: Delete factor $(1 + u)$ from both equations

line 17: Replace continuous with positive and continuous

Page 173, line 11 from bottom: Delete factor $(1 - F_X(a))$

Page 174, line 9: Replace claim-size variable N with claim-count variable N

Example 6.6, line 3 from bottom: Replace $(\mu, \sigma) = (7.000, 2.400)$ with $(\mu, \sigma) = (7.600, 2.400)$

Page 175, line 7: Replace $3{,}000{,}0000$ with $3{,}000{,}000$

Page 177, line 4 from bottom: Replace five straight with six straight

Page 182, line 15: Replace $3{,}007$ with $3{,}077$

Problem 6.1(d): Replace γ with ψ

Page 189, equation (A.3), both lines: Replace NORM.S.DIST with NORM.DIST

Page 199, Problem 6.10, line 8 of table:: Replace 3.4114 with 4.3114

AN ACTUARIAL NOTE ON THE CREDIBILITY OF EXPERIENCE OF A SINGLE PRIVATE PASSENGER CAR

BY
ROBERT A. BAILEY AND LEROY J. SIMON

The experience of the Canadian merit rating plan[1] for private passenger cars provides a means of evaluating the experience rating credibility of the experience of one car. The Canadian experience includes the experience of virtually every insurance company operating in Canada and is collated by the Statistical Agency (Canadian Underwriters' Association—Statistical Department) acting under instructions from the Superintendent of Insurance.

Merit ratings in Canada depend on the number of full years since the insured's most recent accident or since the insured became licensed. The ratings of A, X, Y and B correspond to three or more, two, one, and no years since the most recent accident or since licensing.[2] A + X would be the experience for two or more accident-free years and A + X + Y would be the experience for one or more accident-free years. Table 1 presents the data upon which this study is based. Earned premiums are converted to a common rate basis by use of the relationship in the rate structure that A:X:Y:B = 65:80:90:100. Other calculations in the table are self-explanatory. The authors have chosen to calculate Relative Claim Frequency on the basis of premium rather than car years. This avoids the maldistribution created by having higher claim frequency territories produce more X, Y, and B risks and also produce higher territorial premiums.

The experience rating formula commonly used may be expressed in the form:

$$\text{Modification} = ZR + (1 - Z) \quad \text{where}$$

$Z =$ credibility and

$R =$ the ratio of the actual losses to the expected losses.

If the modification is made equal to the subsequent experience of experience-rated risks relative to the average experience of all risks, and if R is made equal to the past experience on which the experience rating is based relative to the average of all risks, then the formula can be solved for the credibility. Where R = 0 as it is for accident-free risks, the credibility equals 1 — Modification. Referring to Table 1 and setting the Modification equal to the "Relative Claim Frequency", the credibilities obtained for a private passenger car for experience pe-

[1] See also "The Canadian Merit Rating Plan for Individual Automobile Risks," Herbert E. Wittick, P. C. A. S. XLV, pg. 214.

[2] Class 1A Select was introduced effective September 1, 1959 and uses a five-year period, but such risks are still a part of Class 1A in data used in the paper.

riods of one, two, or three years are shown in Table 2. For example, in Class 1A the Modification = .920 which gives Credibility = .080 as shown in Table 2 for a three-year period. As another example, in Class 5, A + X + Y, the Modification = .962 which gives Credibility = .038 as shown in Table 2 for a one-year period.

Table 2 also shows the average claim frequency of each class and the ratio of the three-year credibility to the annual claim frequency. If the variation of individual insureds' chances for an accident were the same within each class, the credibility (for experience rating) would be expected to vary approximately in proportion to the average claim frequency.[3] Classes 2, 3, 4 and 5 are more narrowly defined than Class 1, and the fact that the ratios in the last column of Table 2 for these classes are less than the ratio for Class 1 confirms the expectation that there is less variation of individual hazards in those classes. This also illustrates that credibility for experience rating depends not only on the volume of data in the experience period but also on the amount of variation of individual hazards within the class.

Table 3 shows the credibility of a two or three-year period in relation to the credibility for one year. If an individual insured's chance for an accident remained constant from one year to the next and if there were no risks leaving the class or no new risks entering the class, the credibilities for experience periods of one, two and three years would be expected to vary approximately in proportion to the number of years.[4] It should be remembered that experience rating is a procedure to find the deviation of an individual risk from the average risk and is different from class rate-making, which is a procedure to find the average and where an increase in the volume of the experience increases the reliability of the indication only in proportion to the square root of the volume. The fact that the relative credibilities in Table 3 for two and three years are much less than 2.00 and 3.00 is partially caused by risks entering and leaving the class. But it can be fully accounted for only if an individual insured's chance for an accident changes from time to time within a year and from one year to the next, or if the risk distribution of individual insureds has a marked skewness reflecting varying degrees of accident proneness.

If Class 1B risks have an average of 1.044 accidents in the year prior to the rating[5] the credibility for 1B risks for a one-year experience period is found to be:

$$\text{Modification} = ZR + (1 - Z)$$

$$1.476 = Z\,\frac{1.044}{.087} + 1 - Z$$

$$Z = .043$$

[3] See Appendix I.

[4] See Appendix I.

[5] See Appendix II.

This gives an interesting confirmation to the credibility of .046 produced by considering the combined A + X + Y group.

Tables 1, 2 and 3 are based on accident frequency in order to reduce chance fluctuations caused by variations in the size of claims. However, we noticed that B risks had an average claim cost consistently higher than average and A risks consistently lower. This tends to increase the credibility. Table 4 shows for Class 1, which has enough volume to make the average claim cost reliable, the same data as is presented in Tables 1, 2 and 3 except that losses are used instead of number of claims.

In summary, we feel that the Canadian merit rating data for private passenger cars leads to the following conclusions:

(1) The experience for one car for one year has significant and measurable credibility for experience rating.

(2) In a highly refined private passenger rating classification system which reflects inherent hazard, there would not be much accuracy in an individual risk merit rating plan, but where a wide range of hazard is encompassed within a classification, credibility is much larger.

(3) If we are given one year's experience and add a second year we increase the credibility roughly two-fifths. Given two years' experience, a third year will increase the credibility by one-sixth of its two-year value.

TABLE 1

Canada excluding Saskatchewan

Policy Years 1957 & 1958 as of June 30, 1959

Private Passenger Automobile Liability—Non-Farmers

Merit Rating	*Earned Car Years*	*Earned Prem. at Present B Rates*	*No. of Claims Incurred*	*Claim Freq. per $1000 of Prem.*	*Relative Claim Freq.*
Class 1 — Pleasure — no male operator under 25					
A	2,757,520	159,108,000	217,151	1.365	.920
X	130,706	7,910,000	13,792	1.744	1.175
Y	163,544	9,862,000	19,346	1.962	1.322
B	273,944	17,226,000	37,730	2.190	1.476
Total	3,325,714	194,106,000	288,019	1.484	1.000
A + X	2,888,226	167,018,000	230,943	1,383	.932
A + X + Y	3,051,770	176,880,000	250,289	1.415	.954
Class 2 — Pleasure — Non-principal male operator under 25					
A	130,535	11,840,000	14,506	1.225	.932
X	7,233	712,000	1,001	1.406	1.070
Y	9,726	944,000	1,430	1.515	1.153
B	21,504	1,992,000	3,421	1.717	1.307
Total	168,998	15,488,000	20,358	1.314	1.000
A + X	137,768	12,552,000	15,507	1.235	.940
A + X + Y	147,494	13,496,000	16,937	1.255	.955
Class 3 — Business use					
A	247,424	25,846,000	31,964	1.237	.920
X	15,868	1,783,000	2,695	1.511	1.123
Y	20,369	2,281,000	3,546	1.555	1.156
B	37,666	4,129,000	7,565	1.832	1.362
Total	321,327	34,039,000	45,770	1.345	1.000
A + X	263,292	27,629,000	34,659	1.254	.932
A + X + Y	283,661	29,910,000	38,205	1.277	.949
Class 4 — Unmarried owner or principal operator under 25					
A	156,871	18,450,000	22,884	1.240	.901
X	17,707	2,130,000	3,054	1.434	1.041
Y	21,089	2,523,000	3,618	1.434	1.041
B	56,730	6,608,000	11,345	1.717	1.247
Total	252,397	29,711,000	40,901	1.377	1.000
A + X	174,578	20,580,000	25,938	1.260	.915
A + X + Y	195,667	23,103,000	29,556	1.279	.929
Class 5 — Married owner or principal operator under 25					
A	64,130	5,349,000	6,560	1.226	.941
X	4,039	345,000	487	1.412	1.084
Y	4,869	413,000	613	1.484	1.139
B	8,601	761,000	1,291	1.696	1.302
Total	81,639	6,868,000	8,951	1.303	1.000
A + X	68,169	5,694,000	7,047	1.238	.950
A + X + Y	73,038	6,107,000	7,660	1.254	.962

TABLE 2

Class	Credibility 1 year	2 years	3 years	Claim Frequency per car year	Ratio 3 year cred. to annual claim frequency
1	.046	.068	.080	.087	.920
2	.045	.060	.068	.120	.567
3	.051	.068	.080	.142	.563
4	.071	.085	.099	.162	.611
5	.038	.050	.059	.110	.536

TABLE 3

	RELATIVE CREDIBILITY		
Class	1 year	2 years	3 years
1	1.00	1.48	1.74
2	1.00	1.33	1.51
3	1.00	1.33	1.57
4	1.00	1.20	1.39
5	1.00	1.32	1.55

TABLE 4

Canada excluding Saskatchewan

Policy Years 1957 & 1958 as of June 30, 1959

Private Passenger Automobile Liability—Non-Farmers

Merit Rating	Earned Premiums at Present B Rates	Incurred Losses	Loss Ratio	Relative Loss Ratio
Class 1—Pleasure—no male operator under 25				
A	159,108,000	63,191,000	.397	.911
X	7,910,000	4,055,000	.513	1.177
Y	9,862,000	5,552,000	.563	1.291
B	17,226,000	11,809,000	.686	1.573
Total	194,106,000	84,607,000	.436	1.000
A + X	167,018,000	67,246,000	.403	.924
A + X + Y	176,880,000	72,798,000	.412	.945

Credibility

Class	1 year	2 years	3 years
1	.055	.076	.089

Relative Credibility

Class	1 year	2 years	3 years
1	1.000	1.38	1.62

APPENDIX I

To illustrate that the credibilities would vary approximately in proportion to the number of years* for the first few years and for typical frequencies, consider a model in which 100,000 risks have an inherent hazard, as measured by their true claim frequency, of .05, 100,000 risks have a claim frequency of .10 and 50,000 risks have a frequency of .20. The number of persons claim-free for the past t years assuming a Poisson approximation to the distribution is as follows:

Frequency	$t=0$	$t=1$	$t=2$	$t=3$
.05	100,000	95,123	90,484	86,071
.10	100,000	90,484	81,873	74,082
.20	50,000	40,937	33,516	27,441
Total	250,000	226,544	205,873	187,594

The number of claims in the subsequent year will be:

Frequency	$t=0$	$t=1$	$t=2$	$t=3$
.05	5,000	4,756	4,524	4,304
.10	10,000	9,048	8,187	7,408
.20	10,000	8,187	6,703	5,488
Total	25,000	21,991	19,414	17,200
Claim frequency of total group	.10000	.09707	.09430	.09169
Relative to $t=0$	1.0000	.9707	.9430	.9169
Credibility		.0293	.0570	.0831
Relative credibility		1.000	1.945	2.836

APPENDIX II

Class 1B risks are known to have had one or more claims in the past year. Using the Poisson distribution as an approximation to the risk distribution (another curve which we have used in practice fits more exactly, but for theoretical considerations such as these, the Poisson is a good approximation), we observe that the number of persons having no claim last year is Ne^{-m}, where m is the claim frequency of the class and N is the radix or total number of persons in the population under consideration. Therefore, $N(1-e^{-m})$ persons produce the one or more claims with which we are concerned. The number of claims produced by the entire group is Nm. Hence the average number of claims produced by those risks which have one or more claims is $Nm/N(1-e^{-m})$ or $m/(1-e^{-m})$.

In our specific problem, the Class 1 claim frequency is .087 per car which means that risks that had one or more claims last year (and are Class 1B this year) had an average of $.087/(1-e^{-.087}) = 1.044$ claims.

* This illustration may be used equally as well to demonstrate that the credibilities vary approximately in proportion to the average annual frequency because in the Poisson distribution an increase in the annual frequency has the same effect as an increase in the length of time.

not immoral. Mr. Tarbell's paper indicates that we can learn from the N.A.U.A. He has clearly demonstrated that the N.A.U.A.'s ratemaking procedures are not crude. The N.A.U.A has done an excellent job—one worthy of actuarial approbation.

Once papers such as Mr. Tarbell's are printed in the P.C.A.S., another end is accomplished. We then have something available for all to discuss and to improve upon. This is a most desirable end. Our business is not static and our ratemaking procedures cannot be allowed to become staid or sterile. We must be alert to the requirements of the insuring public—probably the largest public of any American industry. What better way to lay the groundwork for this activity than by a general airing of the facts in the form of papers on ratemaking?

Papers on the fundamental ratemaking procedures of the various casualty, property and fire and accident and health lines have been sorely needed. Is not ratemaking basic to our industry? Is it not the actuary's main stock in trade? Regardless of where we work—for ourselves or for another; a private concern, an insurance department, a rating bureau, or an insurance company; an independent company or a bureau company; a stock company or a mutual company—regardless of our primary concern in our own particular job, do not all of our activities eventully devolve to ratemaking?

A start has been made, but additional papers on ratemaking are still needed. We should have a paper on General Liability ratemaking—an enormous task. The areas of burglary, fidelity and surety also require coverage. An important ratemaking area, almost completely devoid of papers in our *Proceedings,* is the Accident and Health field. We should have ratemaking papers on both Group and Individual Accident and Health. Accident and Health, incidentally, is a most timely and important topic.

These are the thoughts Mr. Tarbell's excellent paper has evoked from me.

DISCUSSIONS OF PAPERS READ AT THE NOVEMBER 1959 MEETING

AN ACTUARIAL NOTE ON THE CREDIBILITY OF EXPERIENCE OF A SINGLE PRIVATE PASSENGER CAR

BY

ROBERT A. BAILEY AND LEROY J. SIMON

Volume XLVI, Page 159

DISCUSSION BY W. J. HAZAM

The authors are to be congratulated for their very valuable contribution to our knowledge of credibility. Presented, as it was, at a time when a large segment of the industry is embarking on merit rating programs for individual private passenger risks, it provides a basis for the actuarial evaluation of plans now available and perhaps many we have yet to see.

While the data underlying the paper are exclusively the results under the

Canadian Merit Rating Plan,[a] the conclusions are not so geographically restricted. The most provocative of these conclusions is that the experience for one car-year has significant and measurable credibility. In the years prior to the current flurry of merit rating plans, this demonstrable fact had been all but lost, if at all recognized, in the generally prevailing opinion that merit rating was unfeasible. Our current plans may yet prove to be unfeasible. However, this paper demonstrates a means or concept by which to measure the actuarial justification for experience credits (credibilities) for one, two, three, etc., claim-free years.

In developing their credibilities, the authors have placed heavy reliance on frequencies in terms of premiums to correct for the maldistribution deriving from the use of an exposure base. I would be remiss as a reviewer to fail to point out that of which the authors are no doubt aware: that a premium base eliminates maldistribution only if (1) high frequency territories are also high premium territories and (2) if territorial differentials are proper. However, premium, although not perfect, is an improvement over exposure as a base for this type of study. The fact that either or both of these inherent assumptions may not always exist does not detract from the qualitative nature of the conclusions but may alter somewhat the basic relative frequencies of Table 1 and the consequent values in Tables 2 and 3.

The authors make the statement, ". . . the credibilities for experience periods of one, two, and three years would be expected to vary approximately in proportion to the number of years." This holds largely true only for low credibilities; large credibilities would render such a statement inaccurate. However, even in a low credibility area such as the authors are working with in the Canadian results, the theoretical relative credibilities would be less than 1.00, 2.00, and 3.00 for one, two, and three years claim free. For example, using the actuarially accepted $\frac{P}{P+K}$ formula for credibility in experience rating, the theoretical relativities to .046 (1 year credibility of class 1—see Table 2) would be as follows (Note: the k value of 2074 used below was derived on the assumption of 100 claims per year producing a one-year credibility of .046):

Credibility	Relative Credibility	Observed Result (Table 3)
$100 \quad \frac{100}{+\ 2074} = .046$	1.00	1.00
$200 \quad \frac{200}{+\ 2074} = .088$	1.91	1.48
$300 \quad \frac{300}{+\ 2074} = .126$	2.74	1.74

[a] See also "The Canadian Merit Rating Plan for Individual Automobile Risks" Herbert E. Wittick, CAS XLV, p. 214.

This observation should be added to the other reasons why the observed relative credibilities in Table 3 are not 1.00, 2.00, and 3.00.

It may be surmised from this approach to the Canadian results that, in a balanced merit rating plan, there is not enough credibility by class to warrant the magnitude of credits now being offered by many U. S. plans. We must remember, however, that these results are based strictly on claim frequencies, not claim frequencies plus convictions frequencies. Adding convictions no doubt helps substantiate larger credits but it is dubious that it will support current merit rating differentials, if the Canadian experience is at all indicative of what we might expect in this country.

This paper with its original concepts sets forth a basis for analysis of current U. S. plans when the data by class becomes available.

SOME CONSIDERATIONS ON AUTOMOBILE RATING SYSTEMS UTILIZING INDIVIDUAL DRIVING RECORDS

BY

LESTER B. DROPKIN

VOLUME XLVI, PAGE 165

Discussion by R. A. Bailey

As Mr. R. E. Beard, secretary and editor of Astin, said,[1]

> "The literature in the English language relating to analytical expressions of the risks involved in general insurance is scanty and largely limited to papers presented to International Congresses of Actuaries and the *Proceedings* of the Casualty Actuarial Society. There are, however, a number of contributions to the subject in various other languages, scattered over various journals, mainly, insurance publications of European countries, e.g. *Skandinavisk Aktuarietidskrift* and a few books."

The C.A.S. can rightfully be proud of its contributions in this field which have been ably enhanced by Mr. Dropkin's treatment of the negative binomial distribution.

The analytical expression of risk distributions provides a valuable insight into many practical problems. One of the important results of Mr. Dropkin's paper is a realization of the large amount of variation among individual risks. Automobile risks even within a single class or merit rating group are far from being all alike. In order to help visualize this variation there are shown in Figure 1 the graphs of the distribution of risks which Mr. Dropkin shows to be inherent in the negative binomial distribution. Four graphs are shown, all for an average accident frequency $\frac{r}{a} = .100$, and with variances of the accident frequency (not the variances of m, the inherent hazard) of $.120(r=\frac{1}{2})$, $.110(r=1)$, $.105(r=2)$ and $.101(r=10)$.

[1] Transactions of the XVth International Congress of Actuaries, Volume II, 1957, p. 230.

THE SWISS RE EXPOSURE CURVES AND THE MBBEFD[1] DISTRIBUTION CLASS

STEFAN BERNEGGER

ABSTRACT

A new two-parameter family of analytical functions will be introduced for the modelling of loss distributions and exposure curves. The curve family contains the Maxwell-Boltzmann, the Bose-Einstein and the Fermi-Dirac distributions, which are well known in statistical mechanics. The functions can be used for the modelling of loss distributions on the finite interval [0, 1] as well as on the interval [0, ∞]. The functions defined on the interval [0, 1] are discussed in detail and related to several Swiss Re exposure curves used in practice. The curves can be fitted to the first two moments μ and σ of a loss distribution or to the first moment μ and the total loss probability p.

1. INTRODUCTION

Whenever possible, the rating of non proportional (NP) reinsurance treaties should not only rely on the loss experience of the past, but also on actual exposure. For the case of per risk covers, exposure rating is based on risk profiles. All risks of similar size (SI, MPL or EML) belonging to the same risk category are summarized in a risk band. For the purpose of rating, all the risks belonging to one specific band are assumed to be homogeneous. They can thus be modelled with the help of one single loss distribution function.

The problem of exposure rating is how to divide the total premiums of one band between the ceding company and the reinsurer. The problem is solved in two steps. First, the overall risk premiums (per band) are estimated by applying an appropriate loss ratio to the gross premiums. In a second step, these risk premiums are divided into risk premiums for the retention and risk premiums for the cession. Due to the nature of NP reinsurance, this is possible only with the help of the loss distribution function.

However, the correct loss distribution function for an individual band of a risk profile is hardly known in practice. This lack of information is overcome with the help of distribution functions derived from large portfolios of similar risks. Such distribution functions are available in the form of so-called exposure curves. These curves directly permit the extraction of the risk premium ratio required by the reinsurer as a function of the deductible.

[1] Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac distribution

ASTIN BULLETIN, Vol. 27, No. 1, 1997, pp. 99-111

Often, underwriters have only a finite number of discrete exposure curves at their disposal. These curves are available in graphical or tabulated form, and are also implemented in computerized underwriting tools. One of the curves must be selected for each risk band, but it is not always clear which curve should be used. In such cases, the underwriter might also want to use a virtual curve lying between two of the discrete curves available to him.

This can be achieved by replacing the discrete curves with analytical exposure curves. Each set of parameters then defines another curve. If a continuous set of parameters is available, the exposure curves can be varied smoothly within the whole range of available curves. However, the curves must fulfill certain conditions which restrict the range of the parameters. In addition, practical problems can arise if a curve family with many (more than two) parameters is used. It might then become very difficult to find a set of parameters which can be associated with the information available for a class of risks. This problem can be overcome if a curve family is restricted to a one- or two-parameter subclass and if new parameters are introduced which can easily be interpreted by the underwriters.

In the following, the MBBEFD class of analytical exposure curves will be introduced. As will be seen, this class is very well suited for the modelling of exposure curves used in practice. Before analysing the MBBEFD curves in detail, some general relations between a distribution function and its related exposure curve will be discussed in section 2. These relations permit the derivation of the conditions to be fulfilled by exposure curves. The new, two-parameter class of distribution functions will then be introduced in section 3. Finally, several practical aspects, and the link to the well known Swiss Re property exposure curves Y_i, will be discussed in section 4.

Conventions

Following the notation used by Daykin et al in [1], we will denote stochastic variables by bold letters, e.g. $\mathbf{X}$ or $\mathbf{x}$. Monetary variables are denoted by capital letters, for instance, X or M, while ratio variables are denoted by small letters, for instance, $x = X/M$.

2. DISTRIBUTION FUNCTION AND EXPOSURE CURVE

2.1. Definition of the exposure curve

In the following, the relation between the distribution function F(x) defined on the interval [0, 1] and its limited expected value function $L(d) = E[\min(d, \mathbf{x})]$ will be discussed. Here, $d = D/M$ and $\mathbf{x} = \mathbf{X}/M$ represent the normalized deductible and the normalized loss, respectively. M is the maximum possible loss (MPL) and $\mathbf{X} \leq M$ the gross loss. The deductible D is the cedent's maximum retention under a non proportional reinsurance treaty. $M \cdot L(d)$ is the expected value of the losses retained by the cedent while $M \cdot (L(1) - L(d))$ is the expected value of the losses paid by the reinsurer. Thus, the ratio of the pure risk premiums retained by the cedent is given by the relative

limited expected value function G(d) = L(d)/L(1) [1]. The curve representing this function is also called the **exposure curve**:

$$G(d) = \frac{L(d)}{L(1)} = \frac{\int_0^d (1 - F(y))dy}{\int_0^1 (1 - F(y))dy} = \frac{\int_0^d (1 - F(y))dy}{E[x]} \tag{2.1}$$

Because of 1 – F(x) ≥ 0 and F'(x) = f(x) ≥ 0, G(d) is an increasing and concave function on the interval [0, 1]. In addition, G(0) = 0 and G(1) = 1 by definition.

2.2. Deriving the distribution function from the exposure curve

If the exposure curve G(x) is given, the corresponding distribution function F(x) can be derived from:

$$G'(d) = \frac{1 - F(d)}{E[x]} \tag{2.2}$$

With F(0) = 0 and G'(0) = 1/E[**x**] one obtains:

$$F(x) = \begin{cases} 1 & x = 1 \\ 1 - \dfrac{G'(x)}{G'(0)} & 0 \le x < 1 \end{cases} \tag{2.3}$$

Thus, F(x) and G(x) are equivalent representations of the loss distribution.

2.3. Total loss probability and expected value

The probability p for a total loss equals 1 – F(1⁻) and the expected (or average) loss μ equals E[**x**]. These two functionals of the distribution function F(x) can be derived directly from the derivatives of G(x) at x = 0 and x = 1:

$$\begin{aligned} \mu &= E[x] = \frac{1}{G'(0)} \\ p &= 1 - F(1^-) = \frac{G'(1)}{G'(0)} \end{aligned} \tag{2.4}$$

The fact that G(x) is a concave and increasing function on the interval [0 ,1] with G(0) = 0 and G(1) = 1 implies:

$$G'(0) \ge 1 \ge G'(1) \ge 0 \tag{2.5}$$

This is also reflected in the relation:

$$0 \le p \le \mu \le 1 \tag{2.6}$$

2.4. Unlimited distributions

If the distribution function F(X) is defined on the interval [0, ∞], the above relations have to be slightly modified. In this case there is no finite maximum loss M. However, the deductible D and the losses **X** can be normalized with respect to an arbitrary reference loss X_0, i.e. $\mathbf{x} = \mathbf{X}/X_0$ and $d = D/X_0$. G(d) is still a concave and increasing function with G(0) = 0 and G(∞) = 1. The expected value $\mu = E[\mathbf{x}]$ is also given by 1/G'(0), but there are no total losses, i.e. G'(∞) = 0.

3. THE MBBEFD CLASS OF TWO-PARAMETER EXPOSURE CURVES

3.1. Definition of the curve

In this section we will investigate the exposure curves and the related distribution functions defined by:

$$G(x) = \frac{\ln(a + b^x) - \ln(a+1)}{\ln(a+b) - \ln(a+1)} \tag{3.1 a}$$

The distribution function belonging to this exposure curve is given by:

$$F(x) = \begin{cases} 1 & x = 1 \\ 1 - \dfrac{(a+1)b^x}{a+b^x} & 0 \le x < 1 \end{cases} \tag{3.1 b}$$

The denominator and the term -ln(a + 1) in the nominator of (3.1 a) ensure that the boundary conditions G(0) = 0 and G(1) = 1 are fulfilled. As will be seen below, the cases a = {-1, 0, ∞} or b = {0, 1, ∞} have to be treated separately.

*Distribution functions of the type (3.1), defined on the interval [0, ∞] or [-∞, ∞], are very well known in statistical mechanics (**Maxwell-Boltzmann**, **Bose-Einstein**, **Fermi-Dirac** and **Planck** distribution). The implementation of these functions in risk theory does not mean that the distribution of insured losses can be derived from the theory of statistical mechanics. However, the MBBEFD distribution class defined in (3.1) shows itself to be very appropriate for the modelling of empirical loss distributions on the interval [0, 1].*

3.2. New parametrisation

The parameters {a, b} are restricted to those values, for which $G_{a,b}(x)$ is a real, increasing and concave function on the interval [0, 1]. It is easier to fulfill this condition by using the inverse g = 1/p of the total loss probability p as a curve parameter and to replace the parameter a in (3.1):

$$g = \frac{a+b}{(a+1)b}; \quad a = \frac{(g-1)b}{1-gb} \tag{3.2}$$

On the one hand, the condition $0 \le p \le 1$ is fulfilled only for $g \ge 1$. On the other hand, G(x) is a real function only for $b \ge 0$. It can be shown that no other restrictions regarding the set of parameters are necessary.

However, cases b = 1 (i.e. a = -1), b = 0 or g = 1 (i.e. a = 0) and $b \cdot g = 1$ (i.e. $a = \infty$) must be treated as special cases. The cases $b \cdot g = 1$ (i.e. $a = \infty$), $b \cdot g > 1$ (i.e. $a < 0$) and $b \cdot g < 1$ (i.e. $a > 0$) correspond to the MB, the BE and the FD distribution, respectively (cf. figure 4.1). By considering special cases b = 1, g = 1 and $b \cdot g = 1$ separately, all real, increasing and concave functions G(x) on the interval [0, 1] with G(0) = 0 and G(1) = 1 belonging to the MBBEFD class (3.1) can be represented as follows:

$$G_{b,g}(x) = \begin{cases} x & g = 1 \vee b = 0 \\ \dfrac{\ln\left(1+(g-1)x\right)}{\ln(g)} & b = 1 \wedge g > 1 \\ \dfrac{1-b^x}{1-b} & bg = 1 \wedge g > 1 \\ \dfrac{\ln\left(\dfrac{(g-1)b+(1-gb)b^x}{1-b}\right)}{\ln(gb)} & b > 0 \wedge b \neq 1 \wedge bg \neq 1 \wedge g > 1 \end{cases} \tag{3.3}$$

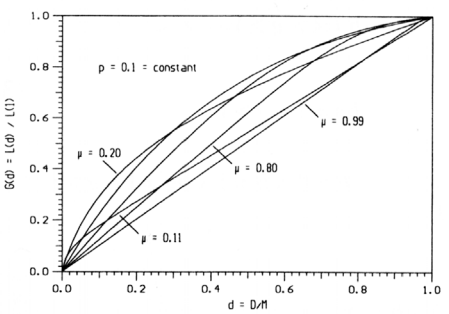


FIGURE 3.1: a) Set of MBBEFD exposure curves with constant parameter g = 1/p = 10 and μ = E[x] = 0.11, 0.2, 0.4, 0.6, 0.8, 0.99.

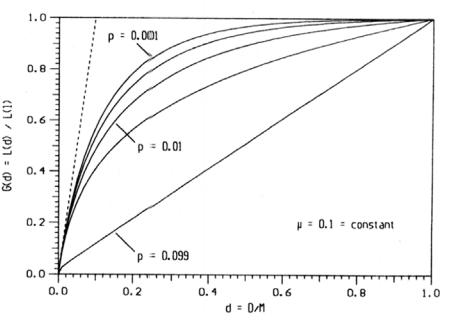


FIGURE 3.1: b) Set of MBBEFD exposure curves with constant μ = E[x] = 0.1 and p = 1/ g = 0.099, 0.031, 0.01, 0.0031, 0.001. The dashed line with slope $1/\mu$ represents the tangent at d = 0.

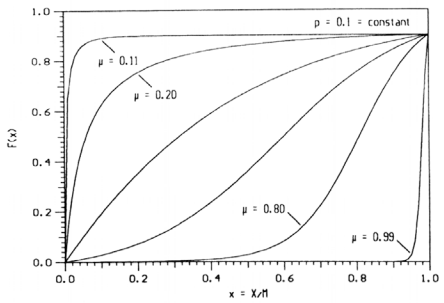


FIGURE 3.2: a) Distribution functions belonging to exposure curves of figure 3.1 a).

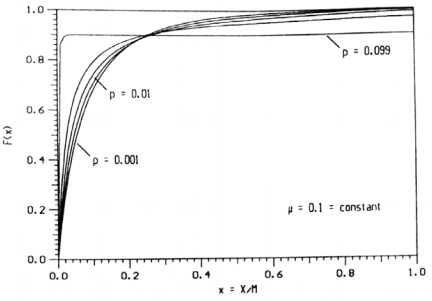


FIGURE 3.2: b) Distribution functions belonging to exposure curves of figure 3.1 b).

Examples of MBBEFD exposure curves are shown in figure 3.1. A set of curves with constant total loss probability p = 0.1 (i.e. g = 10) is represented in figure 3.1 a). Figure 3.1 b) contains a set of curves with constant expected value $\mu = 0.1$. The corresponding distribution functions are shown in figures 3.2 a) and b).

3.3. Derivatives

The derivatives of the exposure curves are given by:

$$G'(x) = \begin{cases} 1 & g = 1 \vee b = 0 \\ \dfrac{g-1}{\ln(g)\left(1+(g-1)x\right)} & b = 1 \wedge g > 1 \\ \dfrac{\ln(b)b^x}{b-1} & bg = 1 \wedge g > 1 \\ \dfrac{\ln(b)(1-gb)}{\ln(gb)\left((g-1)b^{1-x}+(1-gb)\right)} & b > 0 \wedge b \neq 1 \wedge bg \neq 1 \wedge g > 1 \end{cases} \tag{3.4}$$

with

$$G'(0) = \begin{cases} 1 & g = 1 \vee b = 0 \\ \dfrac{g-1}{\ln(g)} & b = 1 \wedge g > 1 \\ \dfrac{\ln(b)}{b-1} = \dfrac{\ln(g)g}{g-1} & bg = 1 \wedge g > 1 \\ \dfrac{\ln(b)(1-gb)}{\ln(gb)(1-b)} & b > 0 \wedge b \neq 1 \wedge bg \neq 1 \wedge g > 1 \end{cases} \tag{3.4 a}$$

and

$$G'(1) = \begin{cases} 1 & g = 1 \vee b = 0 \\ \dfrac{g-1}{\ln(g)g} & b = 1 \wedge g > 1 \\ \dfrac{\ln(b)b}{b-1} = \dfrac{\ln(g)}{g-1} & bg = 1 \wedge g > 1 \\ \dfrac{\ln(b)(1-gb)}{\ln(gb)g(1-b)} & b > 0 \wedge b \neq 1 \wedge bg \neq 1 \wedge g > 1 \end{cases} \tag{3.4 b}$$

The relation p = G'(1)/G'(0) = 1/g is obtained immediately from (3.4 a) and (3.4 b).

3.4. Expected value

According to (2.4) the expected value μ is given by:

$$\mu = E[x] = \frac{1}{G'(0)} = \begin{cases} 1 & g = 1 \vee b = 0 \\ \dfrac{\ln(g)}{g-1} & b = 1 \wedge g > 1 \\ \dfrac{b-1}{\ln(b)} = \dfrac{g-1}{\ln(g)g} & bg = 1 \wedge g > 1 \\ \dfrac{\ln(gb)(1-b)}{\ln(b)(1-gb)} & b > 0 \wedge b \neq 1 \wedge bg \neq 1 \wedge g > 1 \end{cases} \tag{3.5}$$

The expected value μ is represented as a function of the parameters b and g in figure 3.3 and discussed below in section 3.7.

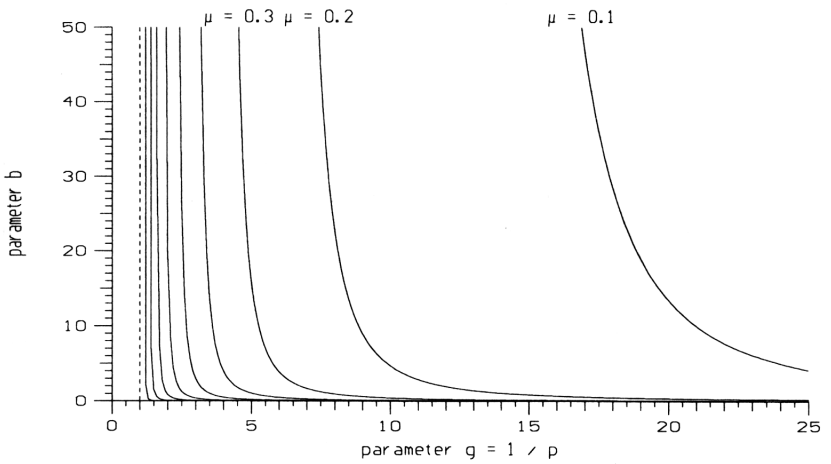


FIGURE 3.3: Parameter b as a function of g = 1/p for $\mu = E[\mathbf{x}] = 0.1, 0.2, \ldots 0.9$. The dashed line at g = 1 and the horizontal line at b = 0 represent the parameter sets {b, g} with $\mu = 1$.

3.5. Distribution function

According to (2.3), the distribution function belonging to the exposure curve $G_{b,g}(x)$ is given by:

$$F(x)=\begin{cases}1 & x=1\\ 0 & x<1\wedge(g=1\vee b=0)\\ 1-\dfrac{1}{1+(g-1)x} & x<1\wedge b=1\wedge g>1\\ 1-b^{x} & x<1\wedge bg=1\wedge g>1\\ 1-\dfrac{1-b}{(g-1)b^{1-x}+(1-gb)} & x<1\wedge b>0\wedge b\neq 1\wedge bg\neq 1\wedge g>1\end{cases} \tag{3.6}$$

The distribution functions belonging to the exposure curves of figure 3.1 are represented in figure 3.2. The set of distribution functions with constant total loss probability p = 0.1 (g = 10) is shown in figure 3.2 a). Figure 3.2 b) contains the set of distribution functions with constant expected value $\mu = 0.1$.

3.6. Density function

Because of the finite probability p = 1/g for a total loss, the density function f(x) = F'(x) is defined only on the interval [0, 1):

$$f(x)=\begin{cases}0 & g=1\vee b=0\\ \dfrac{g-1}{\left(1+(g-1)x\right)^{2}} & b=1\wedge g>1\\ -\ln(b)b^{x} & bg=1\wedge g>1\\ \dfrac{(b-1)(g-1)\ln(b)b^{1-x}}{\left((g-1)b^{1-x}+(1-gb)\right)^{2}} & b>0\wedge b\neq 1\wedge bg\neq 1\wedge g>1\end{cases} \tag{3.7}$$

3.7. Discussion

It is instructive to analyse the expected value $\mu = \mu(b, g)$ as a function of the parameters b and g (3.5). Figure 3.3. shows the range of permitted parameters in the {b, g} plane and the curves with constant expected value μ. One can see in figure 3.3 that $\mu_g(b)$ is a decreasing function of b (for g > 1 constant) and that $\mu_b(g)$ is a decreasing function of g (for b > 0 constant):

$$\begin{aligned}&\frac{\partial}{\partial b}\mu_g(b)\le 0\\ &\frac{\partial}{\partial g}\mu_b(g)\le 0\end{aligned}\qquad g>1\wedge b>0 \tag{3.8}$$

The expected value μ is related as follows to the extreme values of the parameters b and g:

$$\begin{aligned} &\lim_{b\to 0}\mu_g(b)=1; && \lim_{b\to\infty}\mu_g(b)=1/g=p \\ &\lim_{g\to 1}\mu_b(g)=1; && \lim_{g\to\infty}\mu_b(g)=0 \end{aligned} \tag{3.9}$$

3.8. Unlimited distributions

So far, only distributions defined on the interval [0, 1] have been discussed. However, as the MB, the BE and the FD distributions are defined on the interval [-∞, ∞] or [0, ∞], the MBBEFD distribution class can also be used for the modelling of loss distributions on the interval [0, ∞]. If the losses **X** and the deductible D are normalized with respect to an arbitrary reference loss X_0, then $\mathbf{x} = \mathbf{X}/X_0$ and $d = D/X_0$. The above formula can now be modified as follows:

$$G_{b,g}(x)=\begin{cases} 1-b^x & bg=1\wedge g>1 \\ \dfrac{\ln\left(\dfrac{(g-1)b+(1-gb)b^x}{1-b}\right)}{\ln\left(\dfrac{(g-1)b}{1-b}\right)} & 0<b<1\wedge bg\neq 1\wedge g>1 \end{cases} \tag{3.10}$$

$$G'(x)=\begin{cases} -\ln(b)b^x & bg=1\wedge g>1 \\ \dfrac{\ln(b)(1-gb)}{\ln\left(\dfrac{(g-1)b}{1-b}\right)\left((g-1)b^{1-x}+(1-gb)\right)} & 0<b<1\wedge bg\neq 1\wedge g>1 \end{cases} \tag{3.11}$$

$$G'(0)=\begin{cases} -\ln(b) & bg=1\wedge g>1 \\ \dfrac{\ln(b)(1-gb)}{\ln\left(\dfrac{(g-1)b}{1-b}\right)(1-b)} & 0<b<1\wedge bg\neq 1\wedge g>1 \end{cases} \tag{3.11 a}$$

$$G'(1)=\begin{cases} -\ln(b)b & bg=1\wedge g>1 \\ \dfrac{\ln(b)(1-gb)}{\ln\left(\dfrac{(g-1)b}{1-b}\right)g(1-b)} & 0<b<1\wedge bg\neq 1\wedge g>1 \end{cases} \tag{3.11 b}$$

$$G'(\infty)=0 \tag{3.11 c}$$

$$F(x)=\begin{cases} 1-b^x & bg=1\wedge g>1 \\ 1-\dfrac{1-b}{(g-1)b^{1-x}+(1-gb)} & 0<b<1\wedge bg\neq 1\wedge g>1 \end{cases} \tag{3.12}$$

The restriction $0 < b < 1$ is obtained immediately from (3.12) and the condition $F(\infty) = 1$, while the restriction $g > 1$ is obtained from (3.10), where the argument of the logarithm in the denominator must be greater than 0. The same restriction is also obtained from the relation $p = G'(1)/G'(0) = 1/g$, which is still valid. The parameter g is thus the inverse of the probability p of having a loss **X** exceeding the reference loss X_0.

4. CURVE FITTING

4.1. Expected value μ and total loss probability p

Because of (3.8) and (3.9), there exists exactly one distribution function belonging to the MBBEFD class for each given pair of functionals p and μ (cf. figure 3.3), provided that p and μ fulfill the conditions (2.6). The curve parameter $g = 1/p$ is obtained directly. The second curve parameter b can be calculated with the help of (3.5). Here, the following cases must be distinguished:

$$\begin{array}{lll} a) & \mu = 1 & \Rightarrow b = 0 \\ b) & \mu = \dfrac{g-1}{\ln(g)g} & \Rightarrow b = 1/g \\ c) & \mu = \dfrac{\ln(g)}{g-1} & \Rightarrow b = 1 \\ d) & \mu = 1/g & \Rightarrow b = \infty \\ e) & else & \Rightarrow 0 < b < \infty \wedge b \neq 1/g \wedge b \neq 1 \end{array} \tag{4.1}$$

In the general case e), the parameter b has to be calculated iteratively by solving the equation:

$$\mu = \frac{\ln(gb)(1-b)}{\ln(b)(1-gb)} \tag{4.2}$$

Because $\mu_g(b)$ is a decreasing function of b (3.8), the iteration causes no problems. An upper and a lower limit for b can be derived directly from (4.1).

4.2. Expected value μ and standard deviation σ

It is also possible to find a MBBEFD distribution assuming the first two moments (e.g. μ and σ) are known, provided the moments fulfill certain conditions. The first two moments of a distribution function with total loss probability p are given by:

$$\begin{aligned} \mu &= E[x] = p + \int_0^{1^-} x f(x) dx \\ \mu^2 + \sigma^2 &= E[x^2] = p + \int_0^{1^-} x^2 f(x) dx \leq \mu \end{aligned} \tag{4.3}$$

According to (4.3) the first two moments of F(x) and p must fulfill the following conditions:

$$\begin{aligned} \mu^2 &\le E[x^2] \le \mu \\ p &\le E[x^2] \end{aligned} \tag{4.4}$$

Calculation of g and b

Basic idea:
1. Start with $p^* = E[\mathbf{x}^2] \ge p$ as a first estimate (upper limit) for p, and calculate b^* and g^* for the given functionals μ and p^* with the method described in 4.1 above.
2. Compare the second moment $E^*[\mathbf{x}^2]$ with the given moment $E[\mathbf{x}^2]$ and find a new estimate for p^*.
3. Repeat until $E^*[\mathbf{x}^2]$ is close enough to $E[\mathbf{x}^2]$.

If the first moment μ is kept constant, then the second moment $E^*[\mathbf{x}^2]$ will be an increasing function of p^*. Thus the parameters g and b can be calculated without complications.

Remark: The second moment of the MBBEFD distribution has to be calculated numerically. This is best done by replacing F(x) with a discrete distribution function which has the same upper tail area $L(x_{i+1}) - L(x_i)$ as F(x) on each discretized interval $[x_i, x_{i+1}]$.

4.3. The MBBEFD distribution class and the Swiss Re Y_i property exposure curves

The Swiss Re Y_i exposure curves (i = 1 … 4) are very well known and widely used by non proportional property underwriters. As will be shown in this section, all these curves can be approximated very well with the help of a subclass of the MBBEFD exposure curves. In a first step, the parameters b_i and g_i have been evaluated for each curve i. By plotting the points belonging to these pairs of parameters in the {b, g} plane, we found that the points were lying on a smooth curve in the plane. In a next step, this curve was modelled as a function of a single curve parameter c. Finally, the parameters c_i representing the curves Y_i were evaluated.

The subclass of the one-parameter MBBEFD exposure curves is defined as follows:

$$G_c(x) = G_{b_c, g_c}(x) \tag{4.5}$$

with:

$$\begin{aligned} b_c &= b(c) = e^{3.1 - 0.15(1+c)c} \\ g_c &= g(c) = e^{(0.78 + 0.12c)c} \end{aligned} \tag{4.6}$$

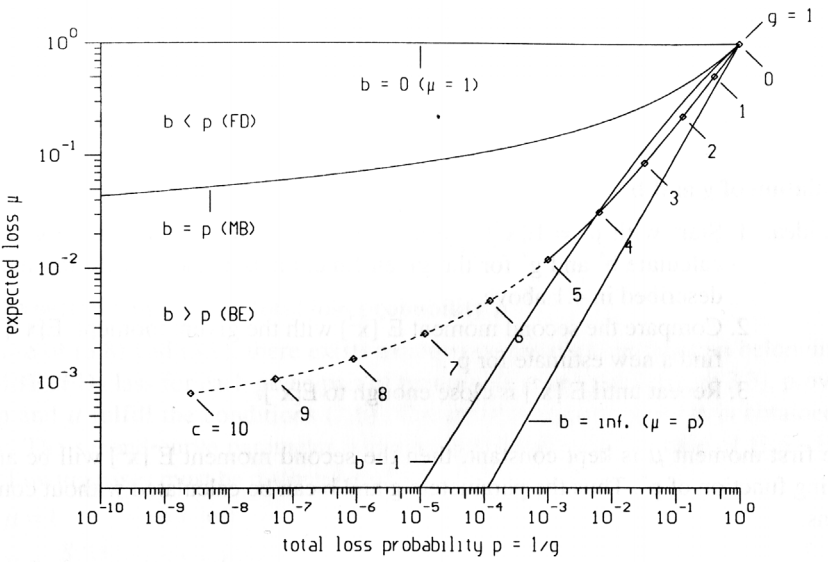


FIGURE 4.1: Range of parameters of the exposure curves $G_{b,g}(x)$. The expected value μ is shown as a function of p = 1/g for special cases b = 0, b = p, b = 1 and b = ∞. In addition, p and μ are shown as a function of the curve parameter c for c = 0 ... 10. The dashed part of this curve has no empirical counterparts.

The position of the curves c = 0 ... 10 in the {p, μ} plane is shown in figure 4.1 Here, the special cases b = 0, p, 1, ∞ and g = 1 are also shown.

The curves defined by c = 0.0, ..., 5.0, which are shown in figure 4.2, are related as follows to several exposure curves used in practice:

- The curve c = 0 represents a distribution of total losses only because of g(0) = 1.
- The four curves defined by c = {1.5, 2.0, 3.0 and 4.0} coincide very well with the Swiss Re curves $\{Y_1, Y_2, Y_3, Y_4\}$.
- The curve defined by c = 5.0 coincides very well with a Lloyd's curve used for the rating of industrial risks.

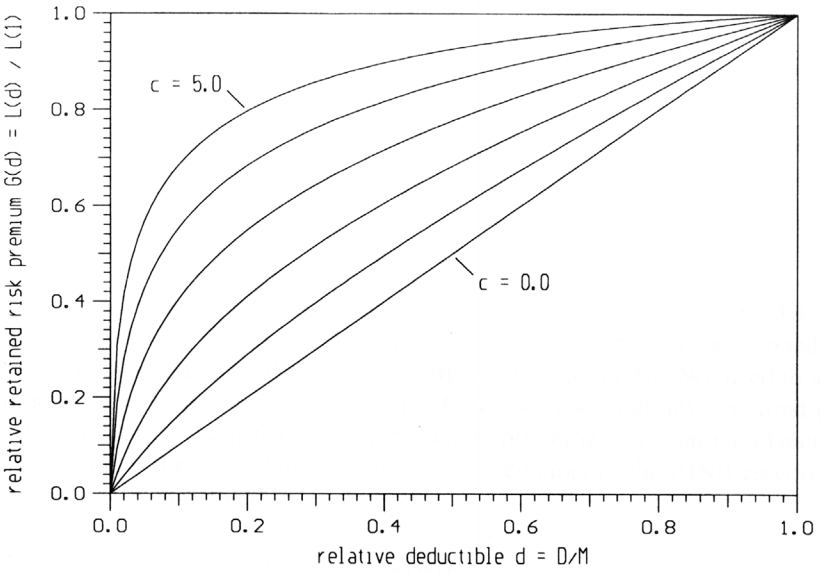


FIGURE 4.2: One-parameter subclass of the MBBEFD exposure curves, shown for c = 0.0, 1.0, 2.0, 3.0, 4.0 and 5.0.

Thus, the exposure curves defined in (4.6) are very well suited for practical purposes. The underwriter can use curve parameters which are very familiar to him. In addition, the class of exposure curves defined by (4.6) is continuous and the underwriter has at his disposal all curves lying between the individual curves Y_i, too.

REFERENCES

C.D. DAYKIN, T. PENTIKÄINEN AND M. PESONEN (1994) "Practical Risk Theory for Actuaries". *Chapman & Hall*, London.

Basics of Reinsurance Pricing

Actuarial Study Note

David R. Clark, FCAS

First Version 1996

Revised 2014

Basics of Reinsurance Pricing

Introduction

Like primary insurance, reinsurance is a mechanism for spreading risk. A reinsurer takes some portion of the risk assumed by the primary insurer (or other reinsurer) for premium charged. Most of the basic concepts for pricing this assumption of risk are the same as those underlying ratemaking for other types of insurance. This study note will assume a knowledge of basic ratemaking concepts on the part of the reader.

A major difference between reinsurance and primary insurance is that a reinsurance program is generally tailored more closely to the buyer; there is no such thing as the "average" reinsured or the "average" reinsurance price. Each contract must be individually priced to meet the particular needs and risk level of the reinsured. This leads to what might be called the “pricing paradox”:

> If you can precisely price a given contract, the ceding company will not want to buy it.

That is to say, if the historical experience is stable enough to provide data to make a precise expected loss estimate, then the reinsured would be willing to retain that risk. As such, the "basic" pricing tools are usually only a starting point in determining an adequate premium. The actuary proves his or her worth by knowing when the assumptions in these tools are not met and how to supplement the results with additional adjustments and judgment.

For the different types of reinsurance outlined in this study note, the basic pricing tools will be introduced in Section A, and criticisms and advanced topics will be introduced in Section B. Section A will include the methods generally accepted and standard throughout the industry. Section B will include areas which require the actuary's expertise but have not been solved to universal agreement.

This study note will focus on domestic treaty covers. Pricing for facultative covers or international (non-U.S.) treaties will not be addressed explicitly, but may be viewed as variations on the same themes. Differences exist in accounting, loss sensitive features and the amount of judgment needed, but the underlying theory does not change.

Finally, this study note will give numerical examples where needed. The numbers used are meant to illustrate the pricing techniques with realistic amounts, but in no way should be taken as recommendations for actual factors.

1. Proportional Treaties

Section 1A. Basic Tools

A proportional treaty is an agreement between a reinsurer and a ceding company (the reinsured) in which the reinsurer assumes a given percent of losses and premium. The simplest example of a proportional treaty is called "Quota Share". In a quota share treaty, the reinsurer receives a flat percent, say 50%, of the premium for the book of business reinsured. In exchange, the reinsurer pays 50% of losses, including allocated loss adjustment expenses, on the book. The reinsurer also pays the ceding company a ceding commission which is designed to reflect the differences in underwriting expenses incurred.

Another, somewhat more complicated, proportional treaty is known as "Surplus Share"; these are common on property business. A surplus share treaty allows the reinsured to limit its exposure on any one risk to a given amount (the "retained line"). The reinsurer assumes a part of the risk in proportion to the amount that the insured value exceeds the retained line, up to a given limit (expressed as a multiple of the retained line, or "number" of lines). An example should make this clear:

Retained Line: $100,000
1st Surplus: 4 lines ($400,000)

Risk	Insured Value	Retained Portion	1st Surplus Reinsured Portion	1st Surplus Percent
1	50,000	50,000	0	0%
2	100,000	100,000	0	0%
3	250,000	100,000	150,000	60%
4	500,000	100,000	400,000	80%
5	1,000,000	100,000	400,000	40%
6	10,000,000	100,000	400,000	4%

It is important to remember that this is not excess insurance. The retained line is only being used to establish the percent of the risk reinsured. Once the ceded percent is calculated, the reinsurer is responsible for that percent of any loss on the risk.

Other types of proportional treaties include fixed and variable quota share arrangements on excess business (e.g., commercial umbrella policies). For these contracts, the underlying business is excess of loss, but the reinsurer takes a proportional share of the ceding company's book. Umbrella treaties will be addressed in the section on casualty excess contracts.

The present section will focus primarily on a proportional property treaty. Most of the techniques described follow standard ratemaking procedures.

The following steps should be included in the pricing analysis for proportional treaties:

Step 1: Compile the historical experience on the treaty.

Assemble the historical premium and incurred losses on the treaty for five or more years. If this is not available, the gross experience (i.e., prior to the reinsurance treaty) should be adjusted "as if" the surplus share terms had been in place, to produce the hypothetical treaty experience. Because a surplus share treaty focuses on large risks, its experience may be different than the gross experience.

The treaty may be on a "losses occurring" basis for which earned premium and accident year losses should be used. Alternatively, the treaty may be on a "risks attaching" basis, which covers losses on policies written during the treaty period. For risks attaching treaties, written premium and the losses covered by those policies are used.

Step 2: Exclude catastrophe and shock losses.

Catastrophe losses are due to a single event, such as a hurricane or earthquake, which may affect a large number of risks. Shock losses are any other losses, usually affecting a single policy, which may distort the overall results. For property contracts, catastrophes are generally defined on a per-occurrence (multiple risk) basis, whereas shock losses are large losses due to a single risk. For casualty contracts, catastrophes may include certain types of claims impacting many insureds (e.g., environmental liability), whereas shock losses would represent a large settlement on a single policy.

Step 3: Adjust experience to ultimate level and project to future period.

The historical losses need to be developed to an ultimate basis. If the treaty experience is insufficient to estimate loss development factors, data from other sources may need to be used. Depending on the source of these factors, adjustments for the reporting lag to the reinsurer or the accident year / policy year differences may need to be made.

The next step is to adjust historical premiums to the future level. The starting point is historical changes in rates and average pricing factors (e.g., changes in schedule rating credits). Rate level adjustment factors can be calculated using the parallelogram method for "losses occurring" treaties. The impact of rate

changes anticipated during the treaty period must also be included. This is an area requiring some judgment, as these percents may not actually have been filed or approved at the time the treaty is being evaluated.

If the premium base is insured value (for property), or some other inflation sensitive base, then an exposure inflation factor should also be included in the adjustment of historical premium.

Finally, the losses need to be trended to the future period. Various sources are available for this adjustment, including the amounts used in the ceding company's own rate filings.

Step 4: Select the expected non-catastrophe loss ratio for the treaty.

If the data used in Step 3 is reliable, the expected loss ratio is simply equal to the average of the historical loss ratios adjusted to the future level. It is worthwhile comparing this amount to the ceding company's gross calendar year experience, available in its Annual Statement, and to industry averages.

Step 5: Load the expected non-catastrophe loss ratio for catastrophes.

Typically, there will be insufficient credibility in the historical loss experience to price a loading for catastrophe potential. However, this amount is critical to the evaluation of property treaties.

In the past, reinsurers had priced catastrophe loads based on "spreading" large losses over expected payback periods. A 1-in-20-year event would be included as a loading of 5% of the loss amount. The payback approach may still be used for casualty events but is only referenced as a reasonability check for property.

The most common procedure is now for a company to select a property catastrophe load based on an engineering-based model that incorporates the risk profile of the ceding company. These models will be discussed in Section 5A below.

Step 6: Estimate the combined ratio given ceding commission and other expenses.

After the total expected loss ratio is estimated, the other features of the treaty must be evaluated. These include:

1. Ceding Commission - often on a "sliding scale" basis (see Section 1B)
2. Reinsurer's general expenses and overhead
3. Brokerage fees (where applicable)

If the reinsurer's business is produced through a broker, there is typically a fee paid by the reinsurer as a percent of treaty premium. If the reinsurer markets the business directly to the ceding company, there is no brokerage fee, but the general expense loading may be higher.

Finally, the reinsurer must evaluate whether or not the projected combined ratio on the treaty is acceptable. The evaluation of treaty terms should take into account potential investment income and the risk level of the exposures to determine if they meet the target return of the reinsurer.

The remainder of this section will be devoted to an example of the pricing for a proportional treaty.

The ceding company has requested a property quota share treaty effective 1/1/97, to be written on a "losses occurring" basis. The submission includes six years of historical experience, rate changes, and a loss development triangle.

The first step involves compiling the historical experience, which in this case is six years with a partial period for 1996. The incurred losses shown are on an accident year basis and include case reserves and allocated loss adjustment expenses but do not include IBNR.

Accident Year Experience evaluated 9/30/96:

Accident Year	Earned Premium	Incurred Losses	Loss Ratio to date
1991	1,640,767	925,021	56.4%
1992	1,709,371	2,597,041 *	151.9%
1993	1,854,529	1,141,468	61.6%
1994	1,998,751	1,028,236	51.4%
1995	2,015,522	999,208	49.6%
1996	1,550,393	625,830	40.4%
Total	10,769,333	7,316,804	67.9%

*Includes 1,582,758 due to Hurricane Andrew

The catastrophe loss for Hurricane Andrew is identified in the 1992 period.

The losses, excluding the Andrew loss, are trended at 4% a year and developed to an ultimate basis. The development factor on the 1996 year is selected so as to project losses for the full year.

Accident Year	Incurred Losses (excl. cats)	LDF	Trend Factor at 4%	Trended Ultimate Incurred Losses
1991	925,021	1.000	1.265	1,170,152
1992	1,014,283	1.000	1.217	1,234,382
1993	1,141,468	1.000	1.170	1,335,518
1994	1,028,236	1.000	1.125	1,156,766
1995	999,208	1.075	1.082	1,162,229
1996	625,830	1.600	1.040	1,041,381
Total	5,734,046			7,100,428

In addition, the rate change information shown below is provided. It should be noted that the +10% rate increase to be effective 4/1/97 is an estimate based on the rate filing that the ceding company expects to make in the coming year. The rate level adjustment assumes that this amount will be approved.

Effective Date	Average Rate Change	
1/1/1991	2.00%	
1/1/1993	10.00%	
7/1/1994	-4.00%	
4/1/1997	10.00%	(pending)

The earned premium amounts above are then adjusted to the average 1997 rate level using factors based on a standard parallelogram method. The other adjustments are that the 1996 premium has been adjusted from a 9 month basis to a full year basis, and all premiums are trended based on average property value inflation of 3%.

Accident Year	Unadjusted Earned Premium	On Level Factor	Trend Factor at 3%	Earned Premium at 1997 Level
1991	1,640,767	1.096	1.194	2,147,147
1992	1,709,371	1.086	1.159	2,151,541
1993	1,854,529	1.034	1.126	2,159,198
1994	1,998,751	0.992	1.093	2,167,158
1995	2,015,522	1.023	1.061	2,187,654
1996	2,067,191	1.028	1.030	2,188,825
Total	11,286,131			13,001,523

The non-catastrophe loss ratio is estimated to be 54.6% based on the projections of loss and premium to the 1997 level.

Accident Year	Earned Premium at 1997 Level	Trended Ultimate Incurred Losses	Projected Loss Ratio
1991	2,147,147	1,170,152	54.5%
1992	2,151,541	1,234,382	57.4%
1993	2,159,198	1,335,518	61.9%
1994	2,167,158	1,156,766	53.4%
1995	2,187,654	1,162,229	53.1%
1996	2,188,825	1,041,381	47.6%
Total	13,001,523	7,100,428	54.6%

The loading for catastrophe losses now needs to be made. For the historical period, the catastrophe loss associated with Hurricane Andrew would have added about 15% to the loss ratio if it had not been excluded. A loading from a catastrophe model might add in a smaller amount. For our example, we will assume that we have selected a 10% loading for catastrophe losses, making our final expected loss ratio approximately 65%.

The final step in the evaluation is the determination of the reinsurer's combined ratio. A ceding commission of 30% has been suggested by the reinsured. The other expenses are listed below:

Expected Loss Ratio	65.0%
Ceding Commission	30.0%
Brokerage fees	5.0%
Administrative expenses	1.0%
Unallocated expenses	1.0%
Indicated Combined Ratio	102.0%

The reinsurance actuary must then evaluate the profitability of these proposed terms. A 102% combined ratio is unlikely to produce an acceptable return for the reinsurer so a reduction in the ceding commission may be the actuary's recommendation. Other provisions, such as a loss occurrence limit or adjustable features (discussed in the next section) may also be considered.

Section 1B. Special Features of Proportional Treaties

After the expected loss ratio is estimated for a proportional treaty, the actuary's work is not yet done. There will often remain disagreement between the ceding company and reinsurer about the loss ratio and the appropriate ceding commission. In theory, a reinsurer should "follow the fortunes" of the ceding company, but in practice their results may be quite different. Reinsuring a profitable insurer is no guarantee of profits for the reinsurer. In the negotiations to resolve these differences, adjustable features are often built into the treaty.

a) Sliding Scale Commission

A common adjustable feature is the "sliding scale" commission. A sliding scale commission is a percent of premium paid by the reinsurer to the ceding company which "slides" with the actual loss experience, subject to set minimum and maximum amounts.

For example:

Given the following commission terms:

Provisional Commission:	30%
Minimum Commission:	25% at a 65% loss ratio
Sliding 1:1 to	35% at a 55% loss ratio
Sliding .5:1 to a Maximum	45% at a 35% loss ratio

Then the results may follow, for different loss scenarios,

Actual Loss Ratio	Adjusted Commission
30% or below	45.0%
35%	45.0%
40%	42.5%
45%	40.0%
50%	37.5%
55%	35.0%
60%	30.0%
65% or above	25.0%

In a "balanced" plan, it is fair to simply calculate the ultimate commission for the expected loss ratio. However, this may not be appropriate if the expected loss ratio is towards one end of the slide. For example, if the expected loss ratio is 65%, the commission from a simple calculation is 25%, producing a 90% technical ratio (i.e., the sum of the loss and commission ratios). If the actual loss ratio is worse than 65%, the reinsurer suffers the full amount, but if the actual loss ratio is better than 65%, the reinsurer must pay additional commission.

It is more correct to view the loss ratio as a random variable and the expected loss ratio as the probability-weighted average of all possible outcomes. The expected ultimate commission ratio is then the average of all possible outcomes based on the loss ratio.

The simplest approach is to estimate the expected commission based on the historical loss ratios, adjusted to future level as above but including the catastrophe and shock losses. This is a good calculation to make as a reasonability check but may be distorted by historical catastrophes or years with low premium volume. It also leaves out many possible outcomes.

A better approach is the use of an aggregate loss distribution model. Several models are available and these are described in Section 4. The results of any of these models may be put into the following format:

Range of Loss Ratios	Average Loss Ratio in Range	Probability of being in Range	Sliding Scale Commission
0% - 35%	31.5%	0.025	45.0%
35% - 55%	46.9%	0.311	39.0%
55% - 65%	59.9%	0.222	30.1%
65% or above	82.2%	0.442	25.0%
0% or above	65.0%	1.000	31.0%

Note that in this example, the expected technical ratio is 96% (=65%+31%) rather than the 90% (=65%+25%) naively estimated above.

A further complication is the introduction of a carryforward provision in the commission. A carryforward provision allows that if the past loss ratios have been above the loss ratio corresponding to the minimum commission, then the excess loss amount can be included with the current year's loss in the estimate of the current year's commission. In the long run, this should help smooth the results.

Two approaches may be taken to pricing the impact of carryforward provisions. The first is to include any carryforward from past years and estimate the impact on the current year only. This amounts to shifting the slide by the amount of the carryforward. For example, if the carryforward from prior years amounts to a 5% addition to the loss ratio, the terms above would become:

Minimum Commission:	25% at a 60% current year loss ratio
Sliding 1:1 to	35% at a 50% current year loss ratio
Sliding .5:1 to a Maximum	45% at a 30% current year loss ratio

The analysis above would then be restated as follows:

Range of Loss Ratios	Average Loss Ratio in Range	Probability of being in Range	Sliding Scale Commission
0% - 30%	27.4%	0.006	45.0%
30% - 50%	43.0%	0.221	38.5%
50% - 60%	55.1%	0.222	29.9%
60% or above	78.3%	0.551	25.0%
0% or above	65.0%	1.000	29.2%

The problem with this approach is that it ignores the potential for carryforward beyond the current year. For example, in the first year of the program we would calculate the expected commission for the current year as though the program would be cancelled at the end of the year. The same price would result with or without the carryforward provision - which does not seem right because the benefit of the carryforward is ignored.

A second approach is to look at the "long run" of the contract. The sliding scale is modeled as applying to a longer block of years rather than just the single current year. The variance of the aggregate distribution would be reduced on the assumption that individual bad years would be smoothed by good experience on other years. The variance of the average loss ratio for a block of years should be significantly less than the variance of the loss ratio for a single year (roughly equal to dividing by the number of years in the block). As an example:

Range of Loss Ratios	Average Loss Ratio in Range	Probability of being in Range	Sliding Scale Commission
0% - 35%	34.1%	0.000	45.0%
35% - 55%	51.6%	0.118	36.7%
55% - 65%	60.4%	0.408	29.6%
65% or above	72.3%	0.474	25.0%
0% or above	65.0%	1.000	28.3%

This example reduces the aggregate variance, putting greater probability in the ranges closer to the expected loss ratio of 65%. The first problem with this approach is that the method for reducing the variance is not obvious; the example above reduces the standard deviation of the aggregate distribution by the square root of 5, assuming that the commission applies to a five-year block. A second problem is that it ignores the fact that the contract may not renew the following year, potentially leaving the reinsured with no carryforward benefit.

This issue is further complicated when a commission deficit can be carried forward but not a credit. There is no standard method for handling these questions so far as this author is aware.

b) Profit Commission

A profit commission subtracts the actual loss ratio, ceding commission and a "margin" for expenses from the treaty premium and returns a percent of this as additional commission. For example:

Actual Loss Ratio	55%	
Ceding Commission	25%	
Margin	10%	
Reinsurer's Profit	10%	(100%-55%-25%-10%)
Percent Returned	50%	(as a percent of Reinsurer's Profit)
Profit Commission	5%	(10% profit times 50%)

Like the sliding scale commission, this should be evaluated using an aggregate distribution on the loss ratio. Also like the sliding scale commission, there is some ambiguity concerning the handling of carryforward provisions.

c) Loss Corridors

A loss corridor provides that the ceding company will reassume a portion of the reinsurer's liability if the loss ratio exceeds a certain amount. For example, the corridor may be 75% of the layer from an 80% to a 90% loss ratio. If the reinsurer's loss ratio is 100% before the application of the loss corridor, then it will have a net ratio of 92.5% after its application, calculated as:

	Before Corridor	After Corridor	
Below corridor	80.0%	80.0%	100% capped at 80%
Within corridor	10.0%	2.5%	10% minus 75% of 90%-80%
Above corridor	10.0%	10.0%	100% minus 90%
Total Loss Ratio	100.0%	92.5%	

As above, the proper estimate of the impact of the loss corridor should be made using an aggregate distribution. The probability and expected values for the ranges below, within and above the corridor can be evaluated.

Range of Loss Ratios	Average Loss Ratio in Range	Probability of being in Range	Loss Ratio Net of Loss Corridor
0% - 80%	64.1%	0.650	64.1%
80% - 90%	84.7%	0.156	81.2%
90% or above	103.9%	0.194	96.4%
0% or above	75.0%	1.000	73.0%

For this example, the expected loss ratio is 75.0% before the application of the loss corridor. Even though this is less than the 80% attachment point for the corridor, the corridor still has the effect of lowering the reinsurer's expected loss ratio.

Many variations on these features can be used with a proportional treaty. Bear and Nemlick [1] provide further background on handling loss sensitive features. This should serve to illustrate that the actuary's job is not finished after the expected loss ratio is calculated.

2. Property Per Risk Excess Treaties

Section 2A. Experience and Exposure Rating Models

Property per risk excess treaties provide a limit of coverage in excess of the ceding company's retention. The layer applies on a "per risk" basis, which typically refers to a single property location. This is narrower than a "per occurrence" property excess treaty which applies to multiple risks to provide catastrophe protection.

The treaty premium is set as a percent of a subject premium base. The subject premium goes by the oxymoronic title "gross net earned premium income" (GNEPI) for losses occurring policies or "gross net written premium income" (GNWPI) for risks attaching policies. This premium is net of any other reinsurance inuring to the benefit of the per risk treaty, such as a surplus share treaty, but gross of the per risk treaty being priced.

The main tools available for pricing per risk treaties are experience and exposure rating.

a) Experience Rating

Experience rating is sometimes referred to as a "burn cost" model though that phrase more commonly denotes just the unadjusted experience and not the projected cost. The basic idea of experience rating is that the historical experience, adjusted properly, is the best predictor of future expectations. The analysis proceeds as follows:

Step 1:

> Gather the subject premium and historical losses for as many recent years as possible. Ten years should be sufficient, though the number of years relied upon in the final analysis should be a balance between credibility and responsiveness.

The historical losses should include all losses that would pierce the layer being priced after the application of trend factors.

Step 2:

Adjust the subject premium to the future level using rate, price and exposure inflation factors as outlined in the section on proportional treaties.

Step 3:

Apply loss inflation factors to the historical large losses and determine the amount included in the layer being analyzed. Sum up the amounts which fall in the layer for each historical period. If allocated expense (ALAE) applies pro-rata with losses, it should be added in individually for each loss.

Step 4:

Apply excess development factors to the summed losses for each period. As in any experience rating model, the loss development factors should be derived from the same ceding company data if possible. Along with the LDF, frequency trend, if determined to be needed, should be applied at this step.

Step 5:

Dividing the trended and developed layer losses by the adjusted subject premium produces loss costs by year. These may be averaged to project the expected loss cost.

The projected loss costs from this analysis should be randomly distributed about the average. If the loss costs are increasing or decreasing from the earliest to latest years in the experience period, then the assumptions of the model may need to be reexamined. The trend or development factors may be too high or low. Alternatively, there may have been shifts in the types of business or sizes of risks written by the ceding company.

As an example of experience rating for a property excess of loss treaty, assume the following terms are requested:

Effective Date:	1/1/97
Treaty Limit:	$400,000
Attachment Point:	$100,000

The losses shown below have been recorded for the treaty. For each loss, a 4% annual trend rate is applied to project the loss from its accident date to the average date in the prospective period. For each trended loss, we then calculate the portion that penetrates into the treaty layer being priced.

Accident Date	Untrended Total Loss	Trend Factor at 4%	Trended Total Loss	Loss in Treaty Layer
9/20/1988	240,946	1.411	339,975	239,975
10/11/1988	821,499	1.408	1,156,671	400,000
3/15/1989	158,129	1.385	219,009	119,009
6/21/1990	114,051	1.317	150,205	50,205
10/24/1990	78,043	1.300	101,456	1,456
1/10/1991	162,533	1.289	209,505	109,505
2/23/1992	324,298	1.234	400,184	300,184
4/30/1992	100,549	1.225	123,173	23,173
9/22/1992	75,476	1.206	91,024	0
1/1/1993	171,885	1.193	205,059	105,059
5/18/1993	94,218	1.175	110,706	10,706
8/1/1993	170,297	1.166	198,566	98,566
8/15/1994	87,133	1.119	97,502	0
7/12/1995	771,249	1.080	832,949	400,000

The losses that trend into the proposed layer are then summed for each historical accident year. The subject premium for each year is listed after adjustment for rate level changes and inflation trend of the insured values. The application of the loss development factor projects the ultimate trended loss cost for the treaty.

Accident Year	On Level Subject Premium	Trended Losses in Layer	LDF	Trended Ultimate in Layer	Loss Cost
1988	1,422,554	639,975	1.000	639,975	45.0%
1989	1,823,103	119,009	1.000	119,009	6.5%
1990	2,054,034	51,661	1.000	51,661	2.5%
1991	2,147,147	109,505	1.000	109,505	5.1%
1992	2,151,541	323,357	1.010	326,591	15.2%
1993	2,159,198	214,331	1.050	225,048	10.4%
1994	2,167,158	0	1.150	0	0.0%
1995	2,187,654	400,000	1.300	520,000	23.8%
Total	16,112,389	1,857,838		1,991,789	12.4%

b) Exposure Rating

The second pricing tool for property per risk treaties is exposure rating. The advantage of this approach over experience rating is that the current risk profile is modeled, not what was written years earlier. The exposure rating model is fairly simple, but may at first appear strange since nothing similar is found on the primary insurance side.

The approach was first developed by Ruth Salzmann in 1963 for Homeowners business and eventually adapted for commercial property as well. The method centers on an exposure curve (P). This represents the amount of loss capped at a given percent (*p*) of the insured value (*IV*) relative to the total value of the loss. This may be represented mathematically as:

$$P(p) = \frac{\int_0^{p \cdot IV} x \cdot f(x)\,dx + \int_{p \cdot IV}^{\infty} p \cdot IV \cdot f(x)\,dx}{\int_0^{\infty} x \cdot f(x)\,dx} = \frac{\int_0^{p \cdot IV} [1 - F(x)]\,dx}{E[X]}$$

where *f(x)* = distribution of individual loss dollar amount

For a property of a given insured value, we calculate the retention and limit as percents of that insured value. The portion of the expected loss on the risk which falls in the treaty layer is then given by:

P((Retention+Limit)/Insured Value) - P(Retention/Insured Value)

As an example, suppose the proposed treaty is intended to cover a per-risk layer of $400,000 excess of $100,000. For a single risk with an insured value of $500,000, we would calculate the difference between the exposure factors for 20% (from $100,000 / $500,000) and 100% (from $400,000+$100,000 / $500,000). From the table below, this results in an exposure factor of 44% (= 93%-49%).

Percent of I.V.	Exposure Factor
0%	0%
10%	37%
20%	49%
30%	57%
40%	64%
50%	70%
60%	76%
70%	81%
80%	85%
90%	89%
100%	93%
110%	97%
120%	100%

The exposure curve provided above is for illustration purposes only. The curve does allow for exposure above the insured value; this is due to the fact that often the limits profile provided does not include business interruption coverage for commercial policies or living expenses for homeowners policies.

For a portfolio of risks, this same calculation is performed on a distribution of premium by different ranges of insured values, known as the "limits profile". The limits profile must also be questioned to verify that the size of risk ranges are on a per location basis. If it is assembled using total values for policies covering multiple locations, distortions will result.

For the example below, it is assumed that all locations within the range are exactly equal to the midpoint of the range.

Treaty Limit:	$400,000
Treaty Retention:	$100,000

Range of Insured Values ($000s)	Midpoint	Retention as % of I.V.	Ret+Limit % of I.V.	Exposure Factor
20 - 100	60	167%	833%	0%
100 - 250	175	57%	286%	26%
250 - 1,000	625	16%	80%	41%
1,000 - 2,000	1,500	7%	33%	33%

Range of Insured Values ($000s)	Subject Premium	Expected Loss Ratio	Expected Losses	Reinsurer's Losses
20 - 100	682,000	65%	443,300	0
100 - 250	161,000	65%	104,650	27,209
250 - 1,000	285,000	65%	185,250	75,953
1,000 - 2,000	1,156,000	65%	751,400	247,962
Grand Total	2,284,000	65%	1,484,600	351,124

The reinsurer's loss cost is 15.37% (Reinsurer's Losses 351,124 over Subject Premium 2,284,000). This loss cost is then loaded for expenses and profit.

The expected loss ratio is of critical importance as the final rate will move proportionally with this amount. A rigorous projection of the expected loss ratio, following the procedures for proportional treaties, should be made.

An implicit assumption in the exposure rating approach outlined above is that the same exposure curve applies regardless of the size of the insured value. For example, the likelihood of a $10,000 loss on a $100,000 risk is equal to the likelihood of a $100,000 loss on a $1,000,000 risk. This assumption of scale independence may be appropriate for homeowners business, for which this technique was first developed, but may be a serious problem when applied to large commercial risks. The Lloyds scales, previously an industry standard, did not recognize this shortcoming.

Ludwig [4] gives an excellent, more detailed description of this topic.

Section 2B. Other Issues on Property Per Risk Treaties

After loss costs are estimated using the experience and exposure rating models, the actuary's task is to reconcile the results and select a final expected loss cost.

a) Free Cover

One difficulty in this reconciliation is the issue of "free cover". This refers to an experience rating in which no losses trend into the highest portion of the layer being priced. For example, if you are comparing prices for a layer $750,000 excess of $250,000 with a layer $250,000 excess of $250,000, and your largest trended loss is $500,000 from ground up, then you will produce the same loss cost for either option. The top $500,000 excess of $500,000 layer would be implicitly a "free cover". One approach to this problem is to use the experience rating as a basis for the lowest portion of the layer and then use the relativities in the exposure rating to project the higher layer.

The table below gives an example of this approach:

Layer to be Priced	Experience Rating Loss Cost	Exposure Rating Loss Cost	Selected Loss Cost
$250k xs $250k	16%	20%	16%
$500k xs $500k	0%	10%	8% *
$750k xs $250k	16%	30%	24%

* 8% = 16% x (10%/20%)

b) Credibility

A first measure of credibility is the number of claims expected during the historical period. Note that this is not the same as the actual number observed during the period. If credibility is set based solely on the historical number, then more credibility will be assigned to experience rating projections that are fortuitously worse than average.

Because the expected number of claims may not be easily calculable, the dollars of expected loss, based on the exposure rating, may be used. For example, if the exposure rating indicates that $2,000,000 in losses was expected during the historical period, but only $1,000,000 was actually observed, then the credibility given to the experience rating should still be based on the $2,000,000 expected.

As a second measure of credibility, it is appropriate to look at the year-to-year variation in the projected loss cost from each of the historical periods. Stability in this rate should add credibility even if the number of claims is relatively small.

Assigning credibility is, in part, a subjective exercise. Often significant credibility is given to experience rating simply because there are too many limitations to the exposure

rating alternative. Discussions on reconciling the experience and exposure rating results are given in Mashitz and Patrik [5] and Clark [9].

c) Inuring Reinsurance

An additional problem which may be encountered in both methods is that the excess treaty may apply to the ceding company's retention after a surplus share treaty is applied. The $750k xs $250k layer may apply to a $1,000,000 loss which is actually a 10% share of a $10,000,000 loss. For experience rating, the only accurate way to reflect this underlying reinsurance is to restate the historical loss experience on a basis net of the inuring reinsurance.

The exposure rating can be applied directly to a risk profile adjusted to reflect the terms of the inuring surplus share treaty. However, if the actuary has exposure curves varying by size of insured value, the curve should be selected based on the insured value before the surplus share is applied, but the exposure factor should apply to the subject premium after the surplus share is applied.

For example, suppose the ceding company from Section 2A decides to purchase a surplus share treaty in which it retains a maximum of $200,000 on any one risk. On its net retention, it then wishes to purchase a per-risk excess cover of $100,000 excess of $100,000. Its risk profile and the single exposure rating curve are the same as used in the earlier example.

Range of Insured Values ($000s)	Midpoint	Ins. Value after S/S	Gross Premium	GNEPI
20 - 100	60	60	682,000	682,000
100 - 250	175	175	161,000	161,000
250 - 1,000	625	200	285,000	91,200
1,000 - 2,000	1,500	200	1,156,000	154,133
Grand Total			2,284,000	1,088,333

Range of Insured Values ($000s)	Net Ins. Value	Retention % of I.V.	Ret+Limit % of I.V.	Exposure Factor
20 - 100	60	167%	333%	0%
100 - 250	175	57%	114%	24%
250 - 1,000	200	50%	100%	23%
1,000 - 2,000	200	50%	100%	23%

Range of Insured Values ($000s)	Subject Premium	Expected Loss Ratio	Expected Losses	Reinsurer's Losses
20 - 100	682,000	65%	443,300	0
100 - 250	161,000	65%	104,650	25,116
250 - 1,000	91,200	65%	59,280	13,634
1,000 - 2,000	154,133	65%	100,186	23,043
Grand Total	1,088,333	65%	707,416	61,793

The loss cost for the $100,000 excess of $100,000 layer is 5.68% (Reinsurer's Losses 61,793 over Subject Premium 1,088,333) for the per-risk excess treaty net of the surplus share. The exposure factors for the two highest ranges are the same because a single exposure curve is used.

3. Casualty Per Occurrence Excess Treaties

Section 3A. Experience and Exposure Rating Models

Like property excess, casualty lines use experience and exposure rating models. This discussion of casualty will refer to general liability (including products), auto liability and workers compensation. The same techniques described can be adapted for other casualty lines, such as professional liability, with some modifications.

Casualty per occurrence excess treaties are often separated into three categories:

Working Layer:

Low layer attachment which is expected to be penetrated, often multiple times in each annual period.

Exposed Excess:

Excess layer which attaches below some of the policy limits on the underlying business - that is, there are policies for which a full limit loss would cause a loss to the treaty. Typically, these losses will be less frequent and there will be some years in which the treaty layer is not penetrated.

Clash Covers:

High layer attachment excess - typically a loss on a single policy will not penetrate the treaty layer. A clash cover will be penetrated due to multiple policies involved in a single occurrence, or when extra-contractual obligations (ECO) or rulings awarding damages in excess of policy limits

(XPL) are determined in a settlement. The method for including allocated loss adjustment expenses in the treaty may also expose the clash layer.

The distinctions between these categories are generally soft in the pricing process. A perfect working layer would produce stable enough results to be retained by the ceding company. Experience rating techniques are still used even when the experience approaches the "exposed excess" category. On the other hand, for large ceding carriers, "clash" losses may be common enough that the experience rating procedure provides guidance for the price.

a) Experience Rating

The steps in the experience rating procedure follow those of property experience rating, but some additional complications arise.

Step 1:

Gather the subject premium and historical losses for as many years as possible. Along with the historical losses, it is very important that allocated loss adjustment expenses (ALAE) be captured separately from losses. For general liability and auto liability losses, the underlying policy limit should also be listed. For auto losses on a split limits rather than a combined single limit (CSL) basis, other modifications may be needed in order to separately cap losses for bodily injury and property damage.

Workers compensation (WC) losses will not have an explicit limit associated with them. However, because large workers compensation losses are often shown on a discounted case reserve basis, a request should be made for these losses on a full undiscounted basis. Further discussion of handling WC losses will be given in the next section.

Step 2:

Adjust the subject premium to the future level using rate, price and exposure inflation factors. These adjustment factors will vary for each line of business included.

Step 3:

Apply loss inflation factors to the individual historical losses. Inflation factors should also vary by line of business.

The selection for a source of loss inflation is difficult. The Insurance Services Office (ISO) estimates basic and total limits trend factors for general and auto liability for use in

ratemaking. Theoretically, what should be used is an unlimited trend factor derived from large losses only. Using losses capped at the underlying policy limit as a source may understate the final results. There is also an implicit assumption that the same trend factor applies to all losses regardless of amount. In the final analysis, the actuary must make a selection of loss inflation rates by year.

The trended losses must then be capped at applicable policy limits. This represents another problem for which there is no generally accepted solution. Theoretically, we want to cap losses at the limit applicable if the same policy were written in the future treaty period. One possible approach is to apply the historical policy limit to each trended loss; this leaves out the fact that the insured will generally increase its policy limits over time. A second approach is to apply the trend factor to the historical loss without applying a policy limit cap; this assumes that policy limits "drift" upwards to precisely match inflation. If this second approach is used, then the subject premium must also be adjusted to the level that would have been charged had the higher limits been in effect; otherwise an overstatement of the expected loss cost will result.

The discussion by Mata and Verheyen [10] gives some more advanced concepts on making use of exposure rating techniques to adjust for changes in the policy limit profile.

After the loss and ALAE amounts are trended, the portion of each in the treaty layer is calculated. Allocated expenses are usually included in one of two ways:

Pro-rata with loss:
ALAE in the layer allocated in proportion to losses.

ALAE as Part-of-Loss (aka "on top" or "add-on"):
ALAE added to loss and the treaty limit applies to the sum.

Example 1:

Trended Loss:	$640,000
Trended ALAE:	$320,000
Treaty Attachment:	$400,000
Treaty Limit:	$600,000

	Pro-Rata Treatment of ALAE			ALAE as Part-of-Loss
	Loss	ALAE	Loss+ALAE	Loss+ALAE
Retained	400,000	200,000	600,000	400,000
In Treaty	240,000	120,000	360,000	560,000
Above Treaty	0	0	0	0
Total	640,000	320,000	960,000	960,000

Example 2:

Trended Loss:	$920,000
Trended ALAE:	$460,000
Treaty Attachment:	$400,000
Treaty Limit:	$600,000

	Pro-Rata Treatment of ALAE			ALAE as Part-of-Loss
	Loss	ALAE	Loss+ALAE	Loss+ALAE
Retained	400,000	200,000	600,000	400,000
In Treaty	520,000	260,000	780,000	600,000
Above Treaty	0	0	0	380,000
Total	920,000	460,000	1,380,000	1,380,000

These two examples should serve to illustrate the two methods of including ALAE in a treaty. It should also be noted that the amount in the treaty layer is not necessarily higher or lower for either method, but depends on the actual experience.

Step 4:

Apply excess development factors to the summed losses for each period. For casualty lines, this step is critical due to the very large factors needed to reflect future development. If possible, historical patterns should be derived for the excess layer using ceding company data. Where this is not available, other benchmarks are needed.

The Reinsurance Association of America (RAA) publishes a loss development study on a biennial basis, which is considered an industry benchmark. The historical data in that study includes more than thirty years of development, broken out by line of business. Its statistics show a significant lag between reported losses for a primary company and a

reinsurer. The graphs included in the 2012 edition of the RAA Study, attached as a supplement to this study note, illustrate this lag.

The use of compiled industry data gives a level of stability to the estimate of excess development patterns that is often superior to that for individual ceding companies.

While the RAA statistics may be considered a benchmark, the user should remember that the data is simply what is reported by its members. Some cautions:

1. The reporting lag from the occurrence of an event to the establishment of a reinsurer's case reserve may vary by company. Included in the data is retrocessional business which may include several levels of reporting lag.

2. The mix of attachment points and limits is not cleanly broken out. In recent studies, the RAA has begun publishing statistics by attachment point ranges, but this data is considerably less stable than the total triangle. Loss development varies significantly for different attachment points so every effort should be made to adjust the selected factors to the layer of the treaty being priced.

3. The RAA requests data exclusive of Asbestos and Environmental claims which could distort the patterns. It cannot be known if all member companies have done this consistently. Other long term exposure claims, such as medical products, mold, or tobacco, are not excluded.

4. For workers compensation, the members may not handle the tabular discount on large claims in a consistent manner. If a ceding company reports a loss on a discounted basis, and the reinsurer establishes a case reserve as the amount of the discounted value that falls into the reinsured layer, a very high development factor may result due to the unwinding of the discount.

As a practical matter, having a very slow development pattern will often produce results showing either zero or very high projected ultimate layer losses by year. The actuary will often need to use smoothing techniques, such as a Bornhuetter-Ferguson approach or Cape Cod (aka Stanard-Bühlmann), to produce a final experience rate.

Step 5:

Dividing the trended and developed layer losses by the adjusted subject premium produces loss costs by year. These amounts are averaged and a final expected loss cost selected. The loss cost may be adjusted for the time value of money, expenses and risk load; these adjustments are dealt with in the last section of this study note.

b) Exposure Rating

The second pricing method is exposure rating. As was the case for property, this method estimates a loss cost based on the premium and limits expected to be exposed during the treaty period. The exposure rating approach uses a severity distribution, based on industry statistics, to estimate layer losses. The severity distribution is used to calculate increased limits factors (ILF) for general liability and auto liability, and excess loss factors (ELF) for workers compensation. The theory is the same for these different lines, but the practical calculation is different.

For all of these approaches, we begin with a Cumulative Distribution Function (CDF) representing the probability that a loss is a given size or smaller.

x =	random variable for size of loss
F(x) =	probability a loss is x or smaller, the CDF
f(x) =	density function, first derivative of F(x)
E[x] =	expected value or average unlimited loss
E[x;L] =	expected value of losses capped at L

The severity distribution is used to calculate expected losses in any given layer.

We define:

$$E[x;L] = E[min(x,L)] = \int_0^L x\, f(x)\, dx + \int_L^\infty L\, f(x)\, dx$$

$$ILF_{L,U} = \frac{E[x;U]}{E[x;L]} \qquad ELF_L = \frac{E[x]-E[x;L]}{E[x]}$$

For general liability and auto liability, one option is to use the truncated Pareto distribution for loss severity. The form of E[x;L] is given by

$$E[x;L] = P \cdot S + \left(\frac{1-P}{Q-1}\right) \cdot \left[(B+Q\cdot T) - (B+L)\cdot\left(\frac{B+T}{B+L}\right)^Q\right]$$

for $Q<>1$, $L>T$.

The five parameters for this distribution follow some intuitive meanings:

T = Truncation point, "small" losses are below this point, "large" losses follow a Pareto distribution
P = probability of a "small" loss
S = average small loss severity
B = scale parameter for Pareto distribution
Q = shape parameter for Pareto distribution

The scale of the distribution is easily adjusted for inflation by multiplying the parameters T, S and B by the same amount. Two limitations of this formula should be noted:

1. The formula shown above only applies for losses above the truncation point T. As a practical matter, this is not a problem as that parameter is set at an amount well below any treaty attachment point.
2. The excess factors for higher layers become very dependent on the Q parameter. This parameter must be watched very carefully when the curves are updated.

A curious note on the truncated Pareto distribution is that when B=0 and Q=1, the distribution becomes a log-logistic distribution of the form below.

$$E[x;L] = P \cdot S + (1-P) \cdot T \cdot \left[1 - ln\left(\frac{T}{L}\right)\right]$$

This has the property that expected losses in layers are equal if the limit and attachment point are in the same ratio.

$$E[x;U] - E[x;L] = E[x;kU] - E[x;kL] \quad \text{for any constant k}$$

This property may be approximated when the B parameter is small and the Q parameter is close to 1. It should be remembered, however, that this relationship holds for severities for individual claims, but not necessarily for treaty loss costs, which will decrease for higher layers due to fewer policies being exposed.

The “BQPST” form of the Pareto is not the only choice available. There is great flexibility possible with discrete mixture models. A discrete mixture is a weighted average of relatively simple curve forms that approximates a more complex but realistic shape.

A popular example of a mixture model is the Mixed Exponential, which is a weighted average of several exponential distributions.

$$E[x;L] = \sum_{j=1}^{N} w_j \cdot \mu_j \cdot \left(1 - exp\left(-\frac{L}{\mu_j}\right)\right)$$

$$\text{where} \qquad 1 = \sum_{j=1}^{N} w_j$$

Once a severity distribution is selected, an exposure factor can be calculated. This factor is analogous to the factor used for excess property and should likewise be applied to ground-up expected losses to estimate the loss cost to the treaty layer.

$$\text{Exposure Factor} = \frac{E[x; min(PL, AP + Lim)] - E[x; min(PL, AP)]}{E[x; PL]}$$

Where *PL* = Ceding Company Policy Limit
AP = Treaty Attachment Point
Lim = Treaty Limit

If the treaty includes ALAE in proportion to losses, this exposure factor can be applied to subject premium times an expected loss and ALAE ratio. If the ALAE is included with losses, the following exposure factor formula can be used:

$$\text{Exposure Factor} = \frac{E\left[x; min\left(PL, \frac{AP + Lim}{(1+e)}\right)\right] - E\left[x; min\left(PL, \frac{AP}{(1+e)}\right)\right]}{E[x; PL]}$$

Where:

PL = underlying Policy Limit applying to loss only
AP = Treaty Attachment Point applying to ALAE plus loss capped at PL
Lim = Treaty Limit applying to ALAE plus loss capped at PL
e = ALAE as a percent of loss capped at PL

The key assumption in both cases is that ALAE varies directly with capped indemnity loss. This is not an accurate model in that ALAE is not a constant percent of any given loss. For example, losses which close without an indemnity payment may still incur a large expense. In general, as the size of a loss increases, the ALAE as a percent of the loss will tend to decrease. The assumption that loss and ALAE are perfectly correlated will tend to result in an overstatement of expected amounts in the higher layers.

Another limitation of the formula for the latter case is that an exposure factor of zero will be applied to high layers which are indeed exposed. For example, if the underlying policy limit is $1,000,000 and the ALAE loading is 1.500, then a treaty attaching at $1,500,000 will not be considered exposed by this formula.

A more refined analysis of the effect of ALAE would require modeling of how ALAE varies with loss size.

Another use for the severity curves is for proportional treaties on excess business. These proportional treaties may be on a quota share basis, where the reinsurer takes a set percent of each contract the ceding company writes, or on a "cessions" basis for which the percent depends on the attachment point and limit written on each policy. A cessions basis treaty will typically require the ceding company to use increased limits factors to price the portion of its policies exposing the treaty. The exposure factors calculated above can be compared with the factors used by the ceding company in pricing its business.

For workers compensation, the severity distributions used most commonly come from the National Council on Compensation Insurance (NCCI), which publishes excess factors for retrospective rating plans in many states. Its curves vary by state and hazard group. The underlying data incorporates different injury types as well. It is not always possible to calculate the underlying severity distribution directly.

The NCCI curves, or other excess factors, can easily be approximated by an inverse power curve of the form:

$$ELF_L = \frac{E[x] - E[x;L]}{E[x]} = a \cdot L^{-b}$$

The parameters "*a*" and "*b*" are estimable from selected excess factors. The fitted factors behave in the higher layers much like the Pareto distribution described above.

Workers compensation does not have policy limits corresponding to those on liability policies. The WC limits refer instead to limitations on annual benefits specific to individual states. The exposure factor is therefore calculated using only the treaty attachment point (AP) and limit.

$$\text{Exposure Factor} = ELF_{AP} - ELF_{AP+Limit}$$

The exposure factor is estimated separately for each state and hazard group for which the ceding company projects premium for the treaty period. Expected loss ratios are also needed for each of these divisions.

An example of exposure rating would look as follows:

Treaty Limit:	750,000
Treaty Attachment Point:	250,000

State	H.G.	Standard Premium	ELR	ELF at 250,000	ELF at 1,000,000	Exposure Factor	Treaty Losses
AL	B	100,000	70%	0.030	0.006	0.024	1,680
AL	C	100,000	70%	0.040	0.008	0.032	2,240
NJ	B	100,000	85%	0.070	0.020	0.050	4,250
NJ	D	100,000	85%	0.100	0.035	0.065	5,525
		400,000					13,695

The loss cost for the treaty will be 3.42% (Treaty Losses 13,695 over Standard Premium 400,000).

Section 3B. Special Problems on Casualty Excess Treaties

This section will deal with a number of problems which commonly arise with casualty excess treaties. Issues about credibility or "free cover" have been addressed in the section on property per risk excess treaties, but should equally be considered for casualty treaties. The methods described are the author's suggestions and should not be viewed as the consensus opinion. However these issues are addressed, they cannot be ignored in the pricing process.

a) Including Umbrella Policies

The ceding company may include umbrella policies in the business subject to the treaty. These policies are excess of an underlying retention and "drop down" if an underlying aggregate is exhausted.

If the umbrella policies are above primary policies written by the ceding company, then it is best to consider the combination of the primary and excess as a single policy with a higher limit. For experience rating the primary and excess pieces are simply added together. When the umbrella policies are above primary policies from other carriers, the procedures are more difficult.

For experience rating, the main difficulty is in selecting the appropriate trend factor. The limit on the underlying policy should be added to losses on the umbrella policy before the application of trend, then subtracted after it:

Trended Loss = (Loss + Underlying Limit)·(Trend Factor) - Underlying Limit

This procedure will still leave out losses from the underlying policy which historically did not exhaust the underlying limit, but which would have after the application of a trend factor.

For exposure rating, the exposure factor on an excess policy is calculated as:

$$\text{Exposure Factor} = \frac{E[x; min(UL + PL, UL + AP + Lim)] - E[x; min(UL + PL, UL + AP)]}{E[x; UL + PL] - E[x; UL]}$$

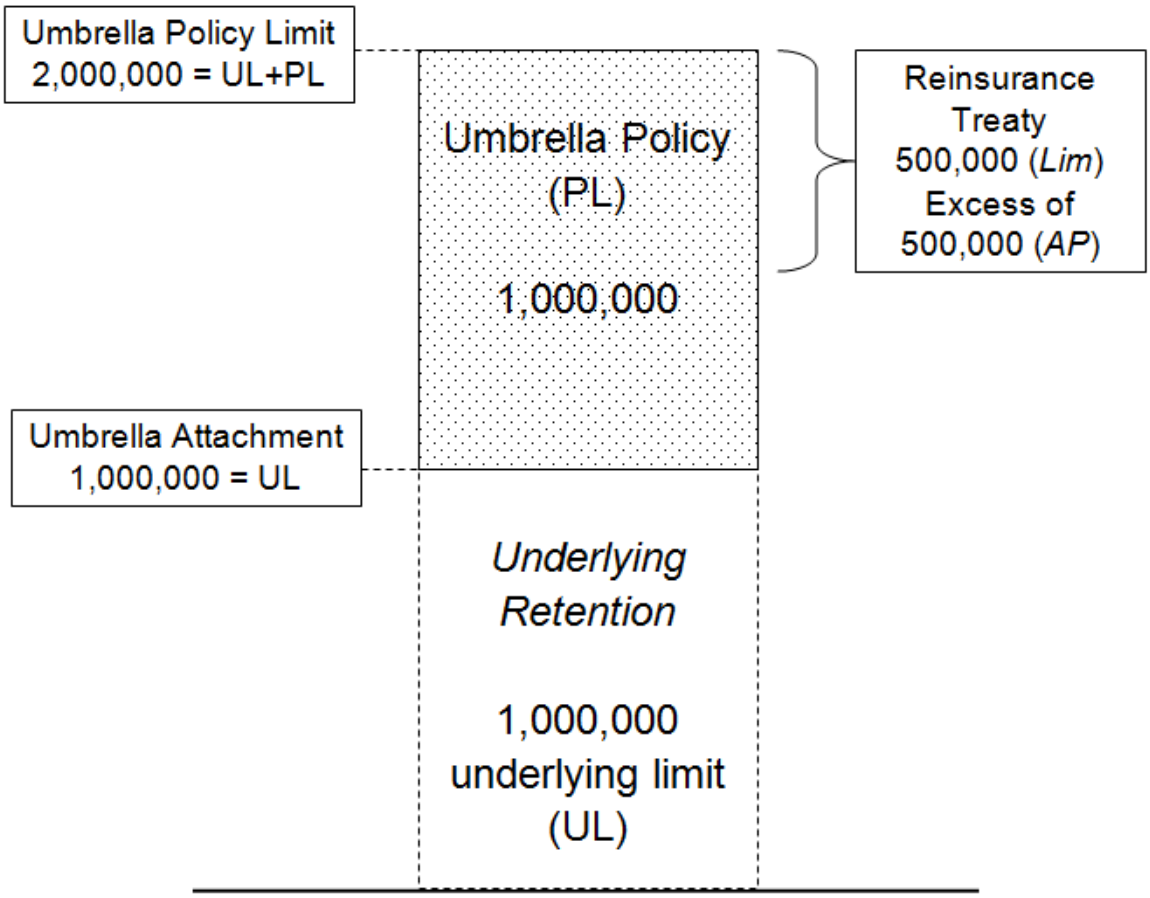
Where:

UL = Limit of Underlying Policies (attachment point of the umbrella)
PL = Policy Limit on Umbrella
AP = Treaty Attachment Point
Lim = Treaty Limit

For example, if the ceding company sells an umbrella policy for $1,000,000 excess of $1,000,000 and the treaty covers losses for the layer $500,000 excess of $500,000, then the exposure factor would be:

$$\text{Exposure Factor} = \frac{E[x; 2] - E[x; 1.5]}{E[x; 2] - E[x; 1]}$$

The graphic below illustrates how this treaty would apply.



This formula leaves out the possibility of the "drop down" feature of the umbrella policy. An approximation to include this additional exposure would be:

$$\text{Exposure Factor} = \frac{\{E[x;2] - E[x;1.5]\} \cdot (1-\phi) + \{E[x;1] - E[x;0.5]\} \cdot \phi}{\{E[x;2] - E[x;1]\} \cdot (1-\phi) + \{E[x;1] - 0\} \cdot \phi}$$

The ϕ in the formula represents the aggregate excess factor on the underlying policy. This is analogous to a Table M charge factor, and will be given a more explicit definition in the section on Aggregate Distributions.

b) Loss Sensitive Features

For working layer excess, the ceding company is often willing to retain more of the losses. In these cases, an annual aggregate deductible (AAD) may be used. The AAD allows the ceding company to retain the first losses in the layer, but maintain protection in case there are more losses than anticipated. The treaty then becomes an excess of aggregate cover, where the aggregate losses are per occurrence excess losses in the layer.

The savings due to aggregate deductibles can be estimated directly from the experience rating if they are set at a sufficiently low level (say, half of the expected value). A better approach is the use of an aggregate distribution model. An excess charge factor for a given AAD is defined as:

$$\phi_{AAD} = \frac{\int_{AAD}^{\infty}(y - AAD)\ g(y)\ dy}{E[y]}$$

where *g(y)* is the distribution of aggregate losses in the layer

The form of this expression may be recognized as that underlying Table M; it is also analogous to the ELF calculation used for per occurrence excess. This charge may be estimated from a number of different methods. These methods are outlined in a separate section on aggregate distributions.

The charge factor ϕ_{AAD} is multiplied by the loss cost for the layer gross of the AAD to estimate the net loss cost.

A second type of loss sensitive program is the "swing plan" which is a type of retrospective rating program. Actual losses to the layer are loaded for expenses and the result is charged back to the ceding company, subject to maximum and minimum constraints.

Swing plans likewise require aggregate distribution models to be evaluated correctly. A swing plan formula may work as follows:

Retro Premium = (Actual Layer Losses) x 100/80
Provisional Rate = 15%
Maximum Premium = 30% x Subject Premium
Minimum Premium = 10% x Subject Premium

For example, if actual losses in the layer are $100,000, then the ultimate premium paid to the reinsurer will be $125,000, subject to the maximum and minimum.

This formula may apply to a block of years instead of to a single year. Following the example of sliding scale commissions, the calculation of expected premium is as follows:

Range of Loss Cost	Probability	Average in Range	Loaded Loss Cost	Capped Premium
0% < LC < 8%	0.120	6.0%	7.5%	10.0%
8% < LC < 24%	0.630	18.0%	22.5%	22.5%
24% < LC	0.250	40.0%	50.0%	30.0%
Total	1.000	22.1%	27.6%	22.9%

In this example, the expected loss ratio is 96.5% (= 22.1%/22.9%), not 80% (from the 100/80 loading) because the maximum and minimum amounts are not in "balance". The loading, maximum or minimum rates can be adjusted to produce an acceptable loss ratio. A second issue on the swing plan is that the provisional rate of 15% is well below the expected ultimate swing plan premium rate of 22.9%; this difference is an added cash flow advantage for the ceding company which must be included in the final pricing evaluation.

c) Workers Compensation Experience Rating

As described above, experience rating for workers compensation may be distorted depending on how tabular discounts are taken into account. A way to avoid this distortion is to collect sufficient information for individual claimants to project their expected costs into the treaty layer. The information needed is:

1. Claimant's current age
2. Claimant's sex (M/F)
3. Estimate of annual indemnity cost including escalation, if any
4. Estimate of annual medical cost
5. Amounts paid to date

For claims with the potential for penetrating the layer, all future payments (both indemnity and medical costs adjusted for escalation) should be determined. For those potential payments which fall within the excess layer, an appropriate mortality factor should be applied to determine the expected amount in the treaty layer. It is important to note that some claims, for which the incurred amount reported by the ceding company falls below the treaty retention, will show an expected amount in the layer. A smaller development factor would then be needed to include only "true IBNR" claims.

4. Aggregate Distribution Models

Throughout the pricing discussions above, aggregate distribution models have been used for pricing a variety of treaty features. This section will outline a number of tools which can be used for these calculations. As a general rule, aggregate models produce results which are very sensitive to the input assumptions; wherever possible, sensitivity analysis on the parameters, or even several approaches, should be used.

All of the approaches in this section may be considered "advanced", but this is not to say that they are optional. Improper evaluation of features which vary with loss experience could lead to significant under- or over-pricing.

a) Empirical Distribution

For most of the adjustable features outlined in this paper, the historical experience can be used to estimate the impact of the adjustable feature. For example, if the actuary has five or more years of loss ratios on a surplus share treaty, then a sliding scale commission can be priced by calculating the commission as if the current terms had been in effect over the historical period (adjusted to current rate level).

The empirical approach is generally very easy to calculate and should be examined at least as a check on other methods. However, some caveats should be recognized:

1) The experience does not take into account all possible outcomes, and may miss the possibility of events outside of what has been observed.
2) If the volume or mix of business has been changing, then the volatility of the future period may be very different than the historical period.
3) If loss development has been performed using a Bornhuetter-Ferguson or Cape Cod method, then the historical periods may present an artificially smooth sequence of loss ratios that does not reflect future volatility.

b) Single Distribution Model

The single distribution approach assumes that the aggregate of all losses to the treaty follows a known CDF form. This is in contrast to a "collective risk" model for which there is explicit modeling of frequency and severity distributions.

A commonly used model is the lognormal distribution. The lognormal has been shown to be a reasonable approximation to empirical distributions, and most spreadsheet software applications allow it to be programmed directly.

The lognormal cumulative distribution function (CDF) has the form:

$$CDF = G(y) = \Phi\left(\frac{ln(y) - \mu}{\sigma}\right) = \int_0^y \frac{exp\left(-\frac{(ln(t) - \mu)^2}{2 \cdot \sigma^2}\right)}{t \cdot \sigma \cdot \sqrt{2\,\pi}}\, dt$$

The parameters can be easily set based on a method of moments, given an expected value and coefficient of variation (CV = standard deviation over the mean).

$$\sigma^2 = ln(CV^2 + 1) \qquad \mu = ln(mean) - \frac{\sigma^2}{2}$$

The limited expected loss function is given by:

$$E[y; L] = exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \Phi\left(\frac{ln(L) - \mu - \sigma^2}{\sigma}\right) + L \cdot \left[1 - \Phi\left(\frac{ln(L) - \mu}{\sigma}\right)\right]$$

Related to that is the excess charge function:

$$\text{Excess Charge Function} = \phi_L = \frac{E[y] - E[y; L]}{E[y]}$$

Finally, the expression for the conditional expected value within a given range is given by the formula below. The first term in the numerator and denominator is replaced by 1 if U is equal to infinity. The second term in the numerator and denominator is replaced by 0 if L=0.

$$E[y \mid L < y < U] = exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \frac{\Phi\left(\frac{ln(U) - \mu - \sigma^2}{\sigma}\right) - \Phi\left(\frac{ln(L) - \mu - \sigma^2}{\sigma}\right)}{\Phi\left(\frac{ln(U) - \mu}{\sigma}\right) - \Phi\left(\frac{ln(L) - \mu}{\sigma}\right)}$$

Where: *L* = lower end of range
U = upper end of range

The formula for the expected value in a given range is very useful because most adjustable features can be broken down into piecewise linear functions, and only the expected value is needed within each linear range.

For example, in the swing plan program illustrated above, the ultimate premium is at the minimum when the loss cost is below 8%, and it increases at a rate of 100/80 with loss until hitting the maximum of 30% at a 24% loss cost. The premium needs only to be estimated for the expected value in each of the three ranges. The ultimate premium estimates are then weighted together using the probabilities of the loss being in each range.

The procedure may be generalized as follows:

Aggregate Loss Range	Expected In Range	Probability In Range
0 to P1	E[y | 0<y<P1]	G(P1)
P1 to P2	E[y | P1<y<P2]	G(P2)-G(P1)
P2 to P3	E[y | P2<y<P3]	G(P3)-G(P2)
…	…	…
Pn & Above	E[y | Pn<y]	1-G(Pn)

This table can be set up for sliding scale commissions, profit commissions, swing plans, loss corridors, or many other common features. The formulae above make use of the lognormal distribution, which is often used in the actuarial literature and can be included in a spreadsheet program. Other curve forms, such as transformed gamma (see Venter [8]) or inverse Gaussian, have been recommended as also providing good fits to aggregate loss data.

The single distribution model has the advantage of being relatively simple to use, even when the source data is limited. A reasonable fit is provided even when frequency and severity distributions are not known. There are two main disadvantages: First, there is no allowance for the loss free scenario; in fact the lognormal is not defined for y=0. Second, there is no easy way to reflect the impact of changing per occurrence limits on the aggregate losses. Bear and Nemlick [1] offer several useful suggestions for modifying the single distribution model to overcome these disadvantages.

c) Recursive Calculation of Aggregate Distribution

The recursive formula, introduced into the actuarial literature by Panjer (see Panjer and Willmot [6]), is a very convenient tool for calculating an aggregate distribution for low frequency scenarios. The frequency distribution is assumed to be Poisson, negative binomial or binomial, and the severity distribution is defined in discrete steps.

For an example, assume that the frequency distribution is Poisson with a mean of λ.

This has the well-known form:

$$Pr(n) = \frac{\lambda^n \cdot e^{-\lambda}}{n!}$$

This can also be given the recursive form:

$$Pr(0) = e^{-\lambda} \qquad Pr(n) = \left(\frac{\lambda}{n}\right) \cdot Pr(n-1)$$

Next, a severity distribution must be defined. For the recursive formula, each possible severity must be equally spaced from the preceding amount. The largest severity may be set equal to the per occurrence limit on an excess treaty, or to the limit times a loading for ALAE. In this example, we will define:

Notation	Severity	Probability
S_1	250	.400
S_2	500	.150
S_3	750	.100
S_4	1,000	.350

This example uses four points, but the formula can be expanded to handle any finite number. The severity distribution must sum to one ($1=S_1+S_2+S_3+S_4$).

For the aggregate distribution, the probability of zero losses is simply equal to the Poisson probability of zero, Pr(0)=.050 for E[n]=λ=3. The probability of the aggregate losses totaling 250 is the probability of one loss, Pr(1)=.150, times the probability that that one loss is equal to 250, S_1=.400. This may be restated in terms of A_0:

$A_0 = Pr(0)$	.050
$A_1 = Pr(1) \cdot .400 = (\lambda/1) \cdot S_1 \cdot A_0$	.060

The probability that the aggregate distribution is 500 is the addition of two pieces: the probability of one 500 loss, plus the probability of two 250 losses. Again this can be restated recursively:

$$A_2 = \Pr(1) \bullet .150 + \Pr(2) \bullet .400 \bullet .400$$
$$= (\lambda/2) \bullet (S_1 \bullet A_1 + 2 \bullet S_2 \bullet A_0) \qquad .059$$

Likewise, the probabilities for higher amounts are easily calculable:

$$A_3 = (\lambda/3) \bullet (1 \bullet S_1 \bullet A_2 + 2 \bullet S_2 \bullet A_1 + 3 \bullet S_3 \bullet A_0) \qquad .057$$

$$A_4 = (\lambda/4) \bullet (1 \bullet S_1 \bullet A_3 + 2 \bullet S_2 \bullet A_2 + 3 \bullet S_3 \bullet A_1 + 4 \bullet S_4 \bullet A_0) \qquad .096$$

$$A_5 = (\lambda/5) \bullet (1 \bullet S_1 \bullet A_4 + 2 \bullet S_2 \bullet A_3 + 3 \bullet S_3 \bullet A_2 + 4 \bullet S_4 \bullet A_1) \qquad .094$$

Notice that for aggregate amounts above the largest possible individual severity, the number of terms does not increase. A simple table can be set up to illustrate this calculation:

Amount	Sev. Prob.	Agg. Prob.	Calculation
0	.000	.050	
250	.400	.060	(3/1)(.400 •.050)
500	.150	.059	(3/2)(.400 •.060+2 •.150 •.050)
750	.100	.057	(3/3)(.400 •.059+2 •.150 •.060+3 •.100 •.050)
1,000	.350	.096	(3/4)(.400 •.057+2 •.150 •.059+3 •.100 •.060+4 •.350 •.050)
1,250	.000	.094	(3/5)(.400 •.096+2 •.150 •.057+3 •.100 •.059+4 •.350 •.060)
1,500	.000	.083	(3/6)(.400 •.094+2 •.150 •.096+3 •.100 •.057+4 •.350 •.059)

This calculation continues indefinitely using the following formula:

$$A_k = \sum_{i=1}^{k} \frac{\lambda}{k} \cdot i \cdot S_i \cdot A_{k-i}$$

When the Poisson frequency distribution is used, the mean and variance of the aggregate distribution are easily estimated as:

$$\text{Mean} = \lambda (250 \bullet S_1 + 500 \bullet S_2 + 750 \bullet S_3 + 1000 \bullet S_4)$$

$$\text{Variance} = \lambda (250^2 \bullet S_1 + 500^2 \bullet S_2 + 750^2 \bullet S_3 + 1000^2 \bullet S_4)$$

The recursive formula can be generalized for frequency distributions other than Poisson:

$$A_k = \sum_{i=1}^{k} \left(a + \frac{b}{k} \cdot i \right) \cdot S_i \cdot A_{k-i}$$

The "a" and "b" parameters are defined as follows:

Poisson:

$a = 0$ $\quad b = \lambda$ $\quad Pr(n) = \frac{\lambda^n \cdot e^{-\lambda}}{n!}$

Negative Binomial:

$a = (1-p)$ $\quad b = (\alpha - 1) \cdot (1-p)$ $\quad Pr(n) = \binom{\alpha + n - 1}{n} \cdot p^{\alpha} \cdot (1-p)^n$

Binomial:

$a = \frac{p}{p-1}$ $\quad b = \frac{(M+1) \cdot p}{1-p}$ $\quad Pr(n) = \binom{M}{n} \cdot p^n \cdot (1-p)^{M-n}$

The use of a negative binomial or binomial frequency distribution allows for greater flexibility in the aggregate distribution.

The recursive formula has the major advantage of being simple to work with and providing an accurate handling of low frequency scenarios. The number of points evaluated on the severity distribution can be expanded to closely approximate continuous curves. The disadvantages are: 1) For higher expected frequencies, the calculation is inconvenient because all the probabilities up to the desired level must be calculated and 2) only a single severity distribution can be used in the analysis.

d) Other Collective Risk Models

In general, collective risk models are distributions for which frequency and severity are explicitly recognized. The recursive method outlined above is a straightforward example of a collective risk model. For handling continuous functions and higher expected frequencies, more advanced techniques may be needed.

A collective risk model assumes the severity of loss, represented by the random variable "x", has a given distribution. The aggregate loss is the sum of "n" of these severities, where "n" is also a random variable. Most aggregate loss models allow for more than a single severity distribution to be used.

The aggregate distribution may be evaluated using simulation or numerical methods. Numerical methods have been developed which can provide very close approximations to the theoretical distribution, with efficient computer time. The underlying calculations are well beyond the scope of this paper (see Heckman and Meyers [3], Robertson [7] or Wang [12] for detailed, readable accounts).

The inputs needed are the severity distribution(s) and parameters for the frequency distribution. Most models then will produce the cumulative distribution function G(y) and excess charge factor ϕ(y) at requested points.

The expected aggregate loss in a given range can be estimated as:

$$E[y \mid L < y < U] = \frac{E[y] \cdot \{\phi_L - \phi_U\} + L \cdot \{1 - G(L)\} - U \cdot \{1 - G(U)\}}{G(U) - G(L)}$$

The results of the aggregate model are very useful in pricing the adjustable features described in this study note. They become even more important on "pure" excess of aggregate covers such as Stop Loss treaties, which cover losses in excess of a set loss amount or loss ratio. The collective risk model is generally the best way to price these treaties but some words of caution are in order:

1. The complexity of the calculations can lead to a "black box" mentality - assuming the numbers must be right because of the accuracy of the computer. Whenever possible, more than one set of results should be produced, as a check on the sensitivity of the answer to the starting assumptions. Some basic statistics, such as the coefficient of variation (standard deviation over mean) and percentiles, should be compared to the empirical data for reasonability.
2. Most models assume that each occurrence is independent of the others and that the frequency and severity distributions are independent of each other. This may be a reasonable assumption in many cases, but could be false in others.
3. Some collective risk models use numerical methods with a large error term for low frequency scenarios. Check the output of the model; the expected error term should be given.
4. The aggregate distribution reflects the process variance of losses but does not reflect the full parameter variance. "Process variance" refers to the random fluctuation of actual results about the expected value. "Parameter variance" in the narrow sense refers to uncertainty about the parameters and may be calculable from outside sources. Some models allow for a prior distribution to apply to the selected parameters. "Parameter variance" in the broader sense of not being sure if you are even using the right model is harder to estimate and is

best reflected by repeated sensitivity analysis. This broader sense could perhaps be called "model risk".

5. Property Catastrophe Covers

Section 5A. Traditional Products and Methods

A property catastrophe cover provides protection for a catastrophic event, such as a hurricane or earthquake. The occurrence may often affect multiple risks and multiple policies. Typically, the catastrophe cover applies to the ceding company's retained exposure net of surplus share, per risk excess treaties and facultative certificates. That is, other reinsurance inures to the benefit of the catastrophe cover.

The limit is defined in excess of a total loss amount. A cover may be $10,000,000 in excess of $30,000,000 per occurrence. Because the limit is often a substantial dollar amount, the contract provides a limited number of reinstatements. Without reinstatements, the catastrophe cover would provide $10,000,000 of limit, but after the full layer is exhausted, there is no more protection. Additional reinstatements are available “pro-rata as to amount” and less often "pro-rata as to time".

Pro-rata as to amount means that if half the limit is exhausted, it can be reinstated for premium proportional to the amount reinstated:

Occurrence Limit:	$10,000,000
Annual Premium:	$2,000,000
Reinstatement Provision:	110% pro-rata as to amount
Actual Loss Amount:	$4,500,000
Reinstatement Premium:	$990,000 (= $2,000,000 X 1.10 X 4.5/10)

The treaty effectively has an aggregate limit equal to one plus the number of reinstatements, times the occurrence limit. For a cover with one reinstatement, the same results will be produced for four losses halfway through the layer as for two full limit losses.

Less frequently, the reinstatement premium is pro-rata as to time, meaning that the premium would be further reduced to reflect only the amount of time left in the policy period. Given the seasonal nature of some types of catastrophes (e.g., hurricanes), relatively few contracts include reinstatements pro-rata as to time.

Before the widespread development of catastrophe models there had been few tools available to systematically price catastrophe covers. The most common method was known as the payback approach, in which premium was set so the offered limit was paid back over a given period of time. For the example above, the payback period is five years, meaning that the $2,000,000 of annual premium would cover a single total loss of $10,000,000 every five years.

Catastrophe models are now the generally accepted approach for pricing of natural and some man-made events. There are four main components of typical catastrophe models:

- Event sets that simulate the covered hazards (e.g., hurricanes, earthquakes, terrorist events). These events cover the full range of possible sizes of a hazard at all relevant locations and are simulated based on estimates of frequency and intensity at specific locations.
- Calculation of local event intensity for each property within a portfolio.
- Estimation of damage for each property within a portfolio impacted by a given event.
- Insured loss estimates based on policies written by the ceding company.

The event sets are generally created and stored within the model prior to pricing of a catastrophe cover. The damage and insured loss estimates are specific to the portfolio written by the ceding company and therefore require additional information.

A catastrophe model will require several types of information:

1. Measure of exposure:
   This should be insured values, construction types, occupancies, along with attachment points for excess contracts.

2. Geographical information:
   Property address information is converted into latitude and longitude coordinates by a geocoding engine that is provided with the model. In some cases insured value information may be less precisely aggregated by zip code or state, resulting in less precise model results.

3. Terms of the insurance policies:
   Include deductible and coinsurance provisions of the original policies.

4. Details of inuring reinsurance:
   If a surplus share treaty inures to the benefit of the catastrophe treaty, any features such as occurrence caps or loss corridors will affect the catastrophe exposure.

The output of a catastrophe model is a distribution of possible losses on the subject business. The expected amount in the treaty layer, usually referred to as the average annual loss (AAL), can be calculated, along with its standard deviation. This can be used as a starting point for a loss cost on the cover.

The model output is also used for management of accumulations, so typically an Occurrence Exceedance Probability (OEP) curve is calculated. The OEP represents the probability that at least one event during the year will exceed a given loss amount. There may also be an Annual Exceedance Probability (AEP) given, which represents the probability that the total of all modeled events in a single year exceeds a given loss amount.

Catastrophe models are a major advance in the ability of insurers and reinsurers to assess their risks. There are additional items that may or may not be included in the results. If not explicitly modeled, these may need to be included more subjectively:

1. Workers compensation losses may be included within the cover. If there is an earthquake during standard working hours, this exposure could be substantial.

2. The inuring reinsurance terms may not be calculable by the model.

3. Even if earthquake coverage is not sold by the ceding company, there may still be exposure due to a "fire following" the earthquake.

4. Other coverage terms, such as the portion of policyholders purchasing replacement cost coverage instead of actual cash value, may be critical. After a major catastrophe event, there may be increased demand for materials and labor which raises the total cost borne by the insurer.

One last complication that should be addressed is due to the basis of coverage for the catastrophe cover: whether it is "losses occurring" during the period or "losses occurring on risks attaching" during the period. As before, "risks attaching" contracts cover losses on policies written during the treaty period. For risks attaching contracts, there is the potential for the reinsurer to pay twice on the same loss event.

Consider a treaty renewing on 1/1/95 for a layer of $10,000,000. A loss event takes place on 3/15/95. The ceding company has policies that are affected, some effective 7/1/94 and some effective 1/1/95. The catastrophe reinsurance treaty effective 1/1/94 covers the losses on the 7/1/94 policies and the treaty effective 1/1/95 covers the losses on the 1/1/95 policies. The reinsurer may end up paying $20,000,000 for the single event. To address this difficulty, many treaties include an “interlocking clause”, designed to equitably apportion losses that may be covered under more than one contract.

Section 5B. Alternative Risk Products

A great variety of products are grouped under the titles "financial reinsurance" or "finite reinsurance". For this study note, the term "finite risk" will refer to property catastrophe covers for which the maximum loss amount is reduced relative to traditional covers. This distinction is very soft because traditional covers are already "finite" in the sense that there is a definite limit that can be paid. Further, the relationship between the ceding company and the reinsurer on traditional covers is often viewed as a partnership; there is an unspoken understanding that the ceding company is expected to pay its own losses over the long term.

Two characteristics are common to most finite risk covers:

1. Multiple year features.
2. Loss sensitive features such as profit commissions and additional premium formulas.

For example, there may be a provision that the contract applies to a three-year period and is cancelable after the first or second year only if premium to date exceeds the loss payments. On the other side, there may be a profit commission which returns, say, 75% of premium if the contract is loss free for three years. In exchange for the profit commission, a relatively high annual premium is charged up front.

These types of features may greatly reduce the downside risk on the contract but it is rarely eliminated. The ceding company can only consider this insurance if two conditions are met:

1. The reinsurer assumes significant insurance risk under the reinsured portions of the underlying insurance agreements.
2. It is reasonably possible that the reinsurer may realize a significant loss from the transaction.

The reinsurance actuary is likely to be called upon to help quantify the risk on the contract to verify that these criteria are met. Timing risk as well as underwriting risk should be evaluated. The actuary charged with this task should refer to the American Academy of Actuaries' "Risk Transfer Testing Practice Note" [11] for guidance.

It is also important to remember that features like profit commissions may substantially reduce the "upside" of the contract from the reinsurer's perspective. It is all well and good to limit the loss to $500,000 in the event of a hurricane, but if this is in exchange for a maximum profit of $10,000 on a loss free year, then more attention needs to be given. In a limited sense, there is an equivalent traditional risk cover corresponding to the possible results from a finite risk cover.

Assume that the following terms are provided on a finite basis:

Annual Premium:	$2,500,000 (25% nominal rate on line)
Occurrence Limit:	$10,000,000
Profit commission:	80% after 10% margin on Annual Premium
Additional Premium:	50% of (Loss + Margin - Annual Premium)

	Loss Free Scenario	One Full Loss Scenario
Premium	$2,500,000	$2,500,000
Loss	$0	$10,000,000
Profit Commission	$1,800,000	$0
Add'l Premium	$0	$3,875,000
U/W Result*	$700,000	($3,625,000)

Now consider the following terms on a traditional basis:

	Loss Free Scenario	One Full Loss Scenario
Premium	$700,000	$700,000
Loss	$0	$4,325,000
U/W Result*	$700,000	($3,625,000)

* U/W result here excludes expenses

The rate on line for the traditional risk program is 16%, and produces an underwriting result equivalent to that of the more complex finite risk program. In this case, the

question becomes: would the reinsurer be willing to offer this cover on a traditional basis at a 16% rate on line? If not, the pricing for the alternative cover is also inadequate.

This type of analysis becomes more complicated when reinstatement provisions, expenses, and carryforward provisions from earlier years are taken into account, but those features can be reflected in an expanded analysis. More difficult are provisions in which the additional premium and profit commission percents change each year of the program or depend on whether the ceding company or the reinsurer cancels the cover.

The best approach to these programs is to estimate the different possible outcomes for a one-year time horizon. Using a simplifying assumption that any penetrations into the layer will exhaust the full limit, probabilities can be assigned to each scenario using a Poisson or other distribution.

Using a frequency distribution is convenient because the mean of the distribution is related to the "payback period" for traditional risk covers. The payback which produces an acceptable expected result can be compared to the results of catastrophe models or other pricing analysis.

A final consideration on finite reinsurance relates to the credit risk of the ceding company. In the example above, the reinsurer depends upon the contingent "additional premium" to minimize the downside risk on the contract. However, there is a new risk introduced that the ceding company will be financially unable to make the payment, especially after experiencing the loss that makes it necessary. A careful review of the ceding company's annual statement needs to be made.

6. Calculating the Final Price

Up to this point, this study note has focused on estimating the reinsurer's expected losses. The final program must be structured to cover this amount but also to cover the reinsurer's expenses and the risk that is borne by the stockholder. The timing of the payment of these amounts is also considered because investment income will contribute to profitability.

Turning first to expenses, it should be noted that the reinsurer's expenses are not the same as those of the ceding company. For instance, reinsurers are not subject to premium tax. The reinsurer's expenses can be broken into three types:

1. Expenses varying with premium
   - ceding commission paid to the reinsured
   - brokerage fees (where applicable)
   - federal excise tax (where applicable)

2. Fixed expenses
   - general overhead costs (salaries, real estate)
   - underwriting and claims audit expenses

   While these expenses may vary somewhat with the size of the account, it is clear that they would not increase simply by taking a larger share of a given treaty. These company expenses should be set independently of variable expenses such as ceding commissions.

   Similarly, an excess of loss reinsurance treaty may be quoted with and without an Annual Aggregate Deductible (AAD). The expected loss to the reinsurer net of the AAD is less than for the treaty without an AAD, but the expenses incurred may not be different. The reinsurer needs an expense structure that covers its costs regardless of whether an AAD is selected.

3. Expenses varying with losses
   - reinsurer's unallocated loss adjustment expenses

   This percent should also vary with the type of reinsurance contract. A working layer excess treaty may require extensive work; a quota share contract may require a review of a loss bordereau, but less claim file review. These amounts can be estimated after discussion with the claims department.

If it is desired simply to load the losses for these expense categories, the final premium could be estimated as:

$$\text{Premium} = \frac{\text{Loss Cost} \cdot (1 + \text{ULAE}) + \text{Fixed Expense}}{(1 - \text{Variable Expense \%})}$$

The "traditional" loading of 100/80, often applied on excess treaties, is an example of this formula. In that case, all expenses are considered variable with premium and assumed to total 20%.

However, consideration must also be given to the timing and risk elements of the contract. The cash flows on the treaty need to be estimated for the treaty, including premium and loss payments and any adjustable features (e.g., swing plan premium).

The considerations for profit or risk load are more complex, and beyond the scope of this study note. Feldblum [2] has given a description of various approaches to accounting for risk.

Bibliography

[1] Bear, Robert A. and Kenneth J. Nemlick "Pricing the Impact of Adjustable Features and Loss Sharing Provisions of Reinsurance Treaties" 1990 PCAS Vol. LXXVII

[2] Feldblum, Sholom "Risk Loads for Insurers" 1990 PCAS Vol. LXXVII

[3] Heckman, Philip E. and Glenn G. Meyers "The Calculation of Aggregate Loss Distributions from Claim Severity and Claim Count Distributions" 1983 PCAS Vol. LXX

[4] Ludwig, Stephen J. "An Exposure Rating Approach to Pricing Property Excess-of-Loss Reinsurance" 1991 PCAS Vol. LXXVIII

[5] Mashitz, Isaac and Gary Patrik "Credibility for Treaty Reinsurance Excess Pricing" 1990 CAS Discussion Paper Program

[6] Panjer, Harry H. and Gordon E. Willmot "Insurance Risk Models" Published by the Society of Actuaries 1992

[7] Robertson, John P. "The Computation of Aggregate Loss Distributions" 1992 PCAS Vol. LXXIX

[8] Venter, Gary "Transformed Beta and Gamma Distributions and Aggregate Losses" 1983 PCAS Vol. LXX

Additional references in Revised Study Note

[9] Clark, David R. "Credibility for a Tower of Excess Layers" CAS Variance Journal, 2011 Volume 5 Issue 1

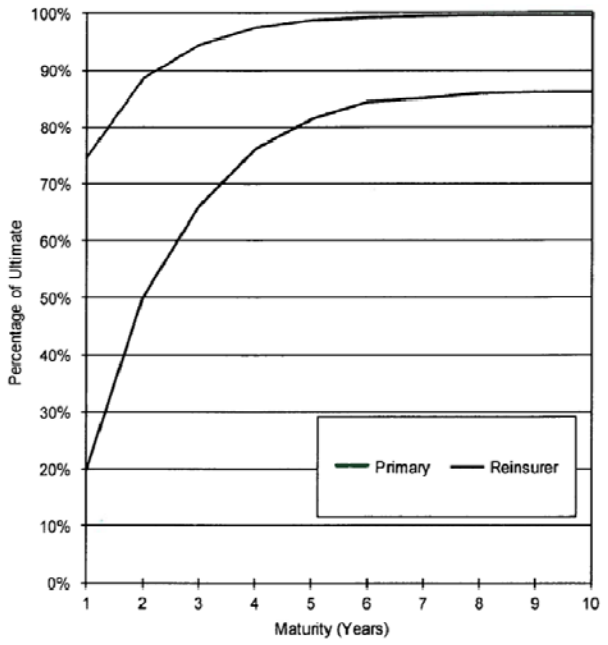
[10] Mata, Ana J. and Mark A. Verheyen "An Improved Method for Experience Rating Reinsurance Treaties using Exposure Rating Techniques" CAS Forum 2005 Spring

[11] "Risk Transfer Testing Practice Note" American Academy of Actuaries Committee on Property and Liability Financial Reporting (COPLFR), January 2007

[12] Wang, Shaun S. "Aggregation of Correlated Risk Portfolios: Models and Algorithms" 1998 PCAS Vol. LXXXV

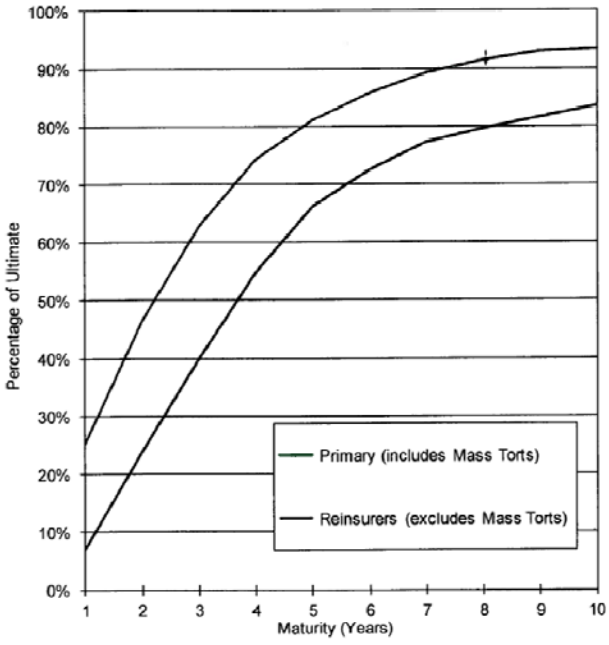
Supplement
Comparison of Primary and Reinsurance Development

Primary vs. Reinsurer
Historical Loss Development
Automobile Liability
Case Incurred Losses



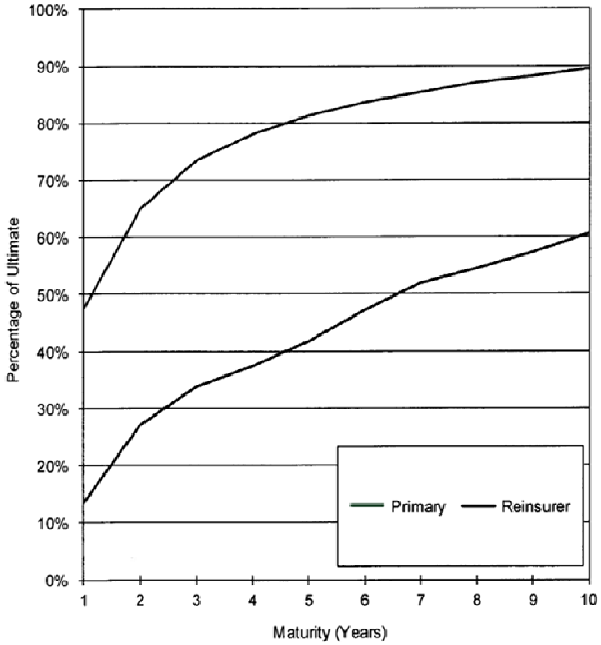
Primary Companies Data Source: A.M. Best Company (2010)
Reinsurance Companies Data Source: Exhibit A - 1.1

Primary vs. Reinsurer
Historical Loss Development
General Liability
Case Incurred Losses



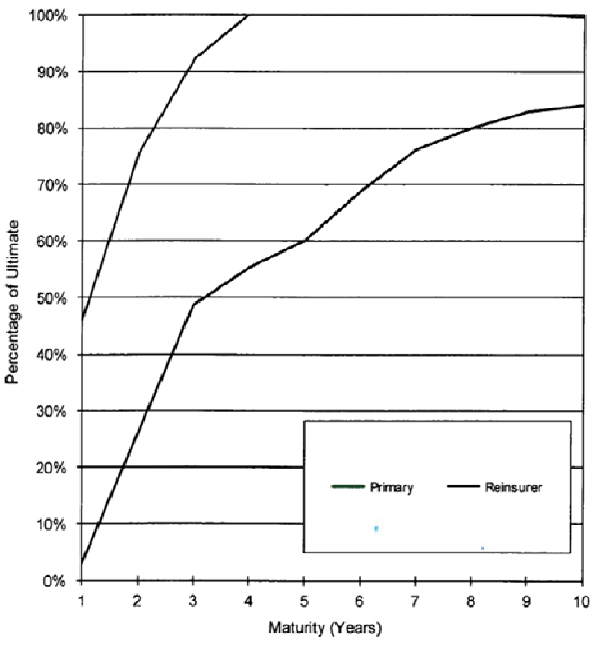
Primary Companies Data Source: A.M. Best Company (2010)
Reinsurance Companies Data Source: Exhibit A - 1.1

Primary vs. Reinsurer
Historical Loss Development
Workers Compensation
Case Incurred Losses



Primary Companies Data Source: A.M. Best Company (2010)
Reinsurance Companies Data Source: Exhibit A - 1.1

Primary vs. Reinsurer
Historical Loss Development
Medical Malpractice
Case Incurred Losses



Primary Companies Data Source: A.M. Best Company (2010)
Reinsurance Companies Data Source: Exhibit A - 1.1

Source: RAA Historical Loss Development Study 2012 Edition

Due to restrictions, please follow the link below:

Couret, J. and Venter, G., "Using Multi-Dimensional Credibility to Estimate Class Frequency Vectors in Workers Compensation," ASTIN Bulletin, Vol. 38, No. 1, May 2008, pp. 72-85.

INDIVIDUAL RISK RATING

Study Note, October 31, 2019

Ginda Kaplan Fisher
Lawrence McTaggart
Jill Petker
Rebecca Pettingell

Expertise. Insight.
Solutions.

Individual Risk Rating Study Note

Ginda Kaplan Fisher, Lawrence McTaggart, Jill Petker, and Rebecca Pettingell

Contents

Foreword

By Lawrence McTaggart

This study note introduces concepts and methods employed when supporting Excess, Deductible and Individual Risk pricing. The authors intend to provide a better experience for candidates by consolidating insights from several foundational papers, providing examples beyond U.S.-based workers' compensation practice, and introducing a few fresh insights.

Chapter 1 provides a summary of experience rating. An experience rating plan prospectively adjusts manual premium based on a policyholder's past experience. The more an individual risk's past experience differs from what is expected of risks in the rating manual classification, the greater the experience modification to the individual risk's manual premium.

Chapter 2 provides an overview of various loss sensitive rating plans. As insureds grow in size their appetite for risk grows. Loss sensitive rating plans allow insureds to retain a portion of their actual loss experience, fulfilling their desire to share in the risk, or reward, of their actual loss experience. Insureds and insurers negotiate the terms of loss sensitive policies, and an actuary will be asked to provide pricing for many different combinations of per-occurrence and aggregate retentions.

Chapter 3 introduces aggregate excess loss estimation. Estimates of aggregate excess loss contemplate both the severity of claims and the number of claims. The expected number of claims for a policy is, in part, a function of the size of the risk. Thus aggregate excess loss estimation considers risk size. Claim severity is a function, in part, of retentions and limits. Visualizing how a loss sensitive plan's insured retentions and insurer limits for both per-occurrence and aggregate boundaries is an important first step when pricing individual loss sensitive rating plans.

Chapter 4 concludes the study note with cautions associated with pricing excess and aggregate loss. Understanding a few of the ways bias can creep into an estimate begins to build the ability to discern estimates that may be biased, and defend estimates that are perceived by others to be biased.

The Excel-based Case Study applies the methods from the readings to a single set of fictional claims data. The Case Study is intended to provide greater clarity and understanding. In practice, or on the exam, the combinations of loss sensitive contract retentions, limits and aggregates is practically unlimited.

The authors hope this study note will help casualty actuaries world-wide understand and master these concepts.

Chapter 1: Experience Rating

By Rebecca Pettingell

1. Introduction/Definition

Experience Rating is the use of an insured's past loss experience to determine rates for a future exposure period. The compilation of rules, definitions, formulas, etc., needed to calculate such a rate is referred to as the experience rating plan.
It is generally in the form of a multiplicative factor applied to manual rates, such that:
Standard Premium = E-mod * Manual Premium.
An experience rating modification factor, or e-mod, greater than 1.0 implies the insured's experience is worse than average for its class. This is called a debit mod. A factor of less than 1.0 implies the insured's experience is better than average for its class. This is called a credit mod.
Note: a risk with a debit mod should not be viewed as a "bad" risk, nor should a risk with a credit mod be viewed as a "good" or "better" risk. The mod merely indicates the risk's expected loss ***relative to other risks in its class.***
There is a misconception that experience rating is an attempt to "charge back" or make up for the past loss experience of an insured. This is incorrect! What experience rating does is determine how much an insured's past loss experience is predictive of its future loss potential and incorporate that prediction into a prospective rate which is better tailored to that risk's loss potential.

2. Advantages of Experience Rating

There are several advantages to using experience rating. It allows us to account for differences between risks within a class. It also allows us to account for differences due to variables that are difficult, impractical, or impossible to quantify via rating variables. Experience rating is a further refinement of classification rating since an individual risk's rate can be further tailored to its loss potential beyond the use of the class and rating variables that make up an insurer's manual rates.

3. Differences within Class

Experience rating is particularly useful when insureds don't fit neatly into a rating class. This could happen if a risk has unique operations, or if the classification system is not sophisticated. The fewer number of rating classes or the broader the range of rating classes in a classification system, the more useful experience rating will be because it allows us to pick up differences within a rating class. Another way to think about this is experience rating allows us to account for the variance of the hypothetical means of the risks within a rating class.
Let's consider two different companies that are very similar in size and operations. Both would be in the same rating class because their operations are so similar. In one company, management is very safety conscious. They require all employees to complete regular safety courses and frequently inspect their premises for safety hazards. If there are any accidents or safety incidents, they conduct a thorough review to determine what went wrong and how another incident could be avoided in the future. In the second company, management thinks that safety is just common sense and spending time discussing safety is a waste of time and money.
If we used strictly manual rating criteria, these two companies would likely be rated the same or very similarly. However, it is pretty clear that the first company will likely have fewer claims and better loss experience. The application of an experience rating plan would likely pick up the

differences between these two companies and allow an insurance company to charge each of these insureds a rate that is more closely tailored to their loss potential.

4. Objectives/Goals

Experience rating accomplishes several objectives. First and foremost, it leads to greater risk equity. By charging a rate that is more commensurate with an insured risk's expected losses, we have increased fairness.

Second, experience rating creates an increased incentive for safety. By attaching a financial consequence to loss experience, there is an additional incentive to prevent or minimize losses on top of the incentives that already exist.

Also, experience rating enhances market competition. The same arguments that are made about classification rating systems enhancing market competition can also be made about experience rating. Since experience rating allows an insurer to charge a rate that is more in line with a risk's loss potential, the insurer will view more risks as being desirable to write. For example, imagine a risk that consistently has higher loss experience than other risks in its class. If an insurer has no mechanism to charge this risk a higher rate, it will not want to write this risk. However, by using experience rating the insurer can charge a premium that is more reflective of the risk's future loss potential.

5. Equity

When thinking about experience rating, it is natural to ask "Is it really equitable to base an insured's future rates on its past experience? Isn't this just a way to charge back an insured for poor loss experience?" Gary Venter answers this question very elegantly. In his article "Experience Rating—Equity and Predictive Accuracy," he states "to the extent that the loss experience is indicative of true differences from the classification average, it appears equitable to charge for it." The experience mod is intended to be a prospective measure of loss potential for the future exposure period. It is not intended to be a penalty or reward for past experience or to recoup past losses.

6. Credibility

Arguably the most important consideration in designing an experience rating plan is credibility—specifically how much credibility should be given to the individual insured's experience in the determination of the premium adjustment.

You can think of experience rating as a way of treating each risk as its own rating class. Just as an insurer might credibility weight the experience of a small rating class with the experience of the larger group it is a part of (e.g., the general liability experience of a small state might be credibility weighted with the country-wide indication), the experience of a single insured can be credibility weighted with that of other risks in its rating class.

First, let's review some basic credibility concepts. Then, we will explore those concepts in an individual risk rating and experience rating context, discussing specific credibility provisions and how they are incorporated into the determination of a risk's premium.

6.1 Credibility Review[1]

In determining a rate or premium (or, what is ultimately equivalent, the modification factor applied to the old rate or the manual rate), the amount of weight given to the insured's own experience represents the level of credibility ascribed to that experience. The complement of credibility is applied to the expected loss experience represented by the manual rate.

Over the years, a number of mathematical approaches to determining credibility have been explored—in particular:

- Classical credibility—also known as "limited fluctuation" credibility, since the volume of expected losses (or expected number of claims, or number of exposures) necessary for a risk's loss experience to be given full credibility is based upon the potential fluctuation of results from expected levels.
- Bühlmann credibility—also known as "greatest accuracy" or "least squares" credibility, as it involves the analysis of the variance associated with the stochastic situation being evaluated.
- Bayesian credibility, which updates prior hypotheses in light of emerging experience. Under certain circumstances, the Bühlmann and Bayesian credibility approaches give the same result.

Regardless of the particular approach used, there are certain characteristics that a credibility factor *Z* (which represents the level of credibility associated with a risk's observed loss experience) is expected to have:

- Z is a value between 0 and 1: $0 \leq Z \leq 1$.
- Z does not decrease as the size of the risk (the level of expected losses, or E) increases: $dZ/dE \geq 0$.
- As the size of a risk increases (i.e., as E increases), the ratio of Z to E decreases: $\frac{d}{dE}\left(\frac{Z}{E}\right) < 0$. This amounts to the charge for a loss of any given size decreasing as the size of the risk increases.

For purposes of experience rating, a Bühlmann credibility framework is used. The basic formula for credibility in this context is:

$$Z = \frac{E}{E + K}$$

where *K* is a constant for a particular situation. More specifically, in accordance with Bühlmann credibility,

$$K = \frac{Expected\ Value\ of\ the\ Process\ Variance}{Variance\ of\ the\ Hypothetical\ Means}.$$

Basically, the difference between a risk's actual loss experience and its expected loss experience can be divided into two categories. First, there is the variation that is purely random and results from the loss process being inherently stochastic, i.e., the process variance. Second, there is variation from the expected experience that is due to a risk being innately different from other risks within its class, i.e., the variance of the hypothetical means. We do not want to penalize or

[1] There are some excellent sources of information and explanation about credibility in the actuarial literature—e.g., Philbrick, 1981, "An examination of Credibility Concepts," *Proceedings of the Casualty Actuarial Society (PCAS)*, 68:195-219, and Mahler and Dean, 2001, "Chapter 8: Credibility," in *Foundations of Casualty Actuarial Science*, fourth edition, pp. 485-659.

reward a risk for experience that is truly random, but we do want the risk to take ownership of experience that is due to the risk's inherent differences. The weighting factor given to a risk's experience, Z, represents the portion of experience that is due to a risk's inherent differences. From this basic credibility framework emerges a set of formulas by which rates and premiums can be determined, either directly or as an adjustment to current rate levels, in light of the risk's own recent loss experience. While there are a number of specific formulas tailored to particular experience rating approaches, the general idea is reflected in a basic version of a rate modification factor. Letting M be the experience modification factor (or "mod"):

$$M = \frac{ZA + (1 - Z)E}{E} = \cdots = \frac{A + K}{E + K}$$

where the last expression can be derived algebraically from the previous one.

7. Credibility Issues in Experience Rating

Two common elements of experience rating plans that relate to the issue of credibility as applied in experience rating are:

1) MSL—The Maximum Single Loss is the amount at which individual large losses are capped when they are included in the calculation of a risk's experience, A. This prevents a single random event from exerting too much influence on the calculation of the mod.
2) Min and Max Adjustment—The calculated modification factor is often subject to a minimum and maximum value. These function as a final measure to ensure that the experience rating adjustment is not too extreme.

These two elements, along with the basic credibility framework and the credibility factor itself, Z, will vary with the size of the risk. The loss experience of larger risks will receive greater credibility than the loss experience of smaller risks. In practice, the size of a risk could be measured using manual premium, expected loss, expected number of claims, or an exposure base (such as sales receipts for GL or power units for commercial auto).

8. Split Loss Plans

Another potential feature of an experience rating plan is to separate the individual claims of a risk's loss experience into different layers. This plan is known as a split loss plan.
For example, let's suppose we have an experience rating plan which uses a single loss split at $5,000 and a risk has the following loss experience:

Claim #	Incurred Loss Amount
001	$1,150
002	$5,000
003	$3,000
004	$500
005	$50,000
006	$2,000
007	$10,000
008	$6,000
009	$350
010	$12,025
011	$4,500

Now let's look at the losses after we split them into layers of $0 – $5,000 and $5,000+.

Claim #	Incurred Loss Amount	Primary Loss Amount	Excess Loss Amount
001	$1,150	$1,150	$0
002	$5,000	$5,000	$0
003	$3,000	$3,000	$0
004	$500	$500	$0
005	$50,000	$5,000	$45,000
006	$2,000	$2,000	$0
007	$10,000	$5,000	$5,000
008	$6,000	$5,000	$1,000
009	$350	$350	$0
010	$12,025	$5,000	$7,025
011	$4,500	$4,500	$0
Total	$94,525	$36,500	$58,025

Now instead of comparing just the total loss experience of this risk to an expected amount, we will compare the primary and excess components independently.

Split rating plans—which ordinarily give much more weight to the small portion of losses than to the large portion—can also be thought of as a linear approximation to assigning credibility to the log of the loss amount. Many real loss distributions are skewed with extremely heavy tails. The most predictive estimate might be obtained by normalizing the distribution in some way, such as taking a logarithm of it. But that could lead to messy and complex rating algorithms. The split rating plan is a compromise between simplicity and precision.

One can view these primary and excess components of loss as representing the frequency and severity of the experience, respectively. If the limit for the primary portion of the loss is relatively low, when the actual primary losses exceed the expected primary losses it must mean there have been a higher number of losses than expected. Since the primary losses are truncated from above, a higher than expected outcome cannot be due to a single or small number of very large losses.

The excess component of the loss experience represents severity. It would be difficult for a large number of smaller losses to cause the excess portion of the loss experience to greatly exceed expectation unless the severity of the losses was higher than expected.

The NCCI WC Experience Rating Plan is an example of a split loss plan. The split plan has been found to produce empirically better results for the WC plan than a total or limited loss plan, perhaps because WC loss distributions are very heavy tailed[2]. This can also be thought of as separating the claim count uncertainty (the parameter risk, mostly driven by lots of small Med-only and TT claims) and the severity uncertainty (the process risk, driven by relatively few but influential Major PP, PT, and Fatal claims).

[2] See Gillam, W.R., "Parameterizing the Workers' Compensation Experience Rating Plan," PCAS LXXIX, 1992, pp21-56 and Meyers, G., "An Analysis of Experience Rating", PCAS LXXII 1985, pp278-317 for this result and discussions for and against using a split rating plan, and some practical and statistical considerations in choosing an experience rating plan structure.

With respect to credibility, a split plan necessitates a credibility-weighted modification factor (M) formula that is adjusted to reflect the two tiers of losses, primary and excess. Using the same underlying framework as described in the prior section, a split plan formula for the mod factor is:

$$M = \left[Z_p A_p + \left(1 - Z_p\right)E_p + Z_e A_e + (1 - Z_e)E_e\right] \div E$$

where A, E, and Z are defined as before, and the subscripts p and e refer to "primary" and "excess," respectively. This formula can then be algebraically manipulated to yield:

$$M = 1 + Z_p \frac{(A_p - E_p)}{E} + Z_e \frac{(A_e - E_e)}{E}.$$

This is the framework for determining the modification factor in the context of credibility factors. The NCCI uses a mathematically equivalent formula, but describes the components in terms of "weighting" and "ballast":

$$M = \frac{A_p + (1 - w)E_e + B + wA_e}{E_p + (1 - w)E_e + B + wE_e} = \frac{A_p + (1 - w)E_e + B + wA_e}{E + B}$$

where w is the excess loss weighting factor, B is the ballast value, and other terms are as defined earlier.

9. Schedule Rating

Schedule rating is a series of credits and debits that can be used to modify a risk's rates to reflect the risk's individual characteristics. Rates can be modified either upward (increased) or downward (decreased), depending on the expected impact on the risk's loss experience.

A schedule rating plan for commercial general liability coverage might look something like this:

		Range of Modifications:		
Risk Characteristic	Description	Credit		Debit
Location	Exposure inside the premises	5%	to	5%
	Exposure outside the premises	5%	to	5%
Premises	Condition and care of premises	10%	to	10%
Equipment	Type, condition, and care of equipment	10%	to	10%
Classification	Peculiarities of classification	10%	to	10%
Employees	Selection, training, supervision, experience	6%	to	6%
Cooperation	Medical facilities	2%	to	2%
	Safety Program	2%	to	2%

Under this plan, a risk that had a particularly good safety program might receive up to a 2% credit to its manual rates. However, a risk that has a much more inexperienced than average workforce could receive a debit of up to 6% to reflect the fact that inexperienced employees are correlated with worse than average loss experience.

One must be careful to prevent overlap when both schedule rating and experience rating will be applied to a policy. If a risk has made a recent change that will likely impact its loss experience, then it is appropriate to use a schedule credit (or debit). For example, suppose a risk has recently hired a full time safety manager who will oversee operations and be responsible for enforcing appropriate safety measures. This would be expected to have a favorable effect on loss experience and it would be appropriate to apply a schedule credit to reflect this expectation of improved loss experience. Contrast this to another risk who has always had a full-time safety manager on its staff. The effect of the safety manager on this second risk's experience will already be reflected in its loss experience since the safety manager was there during the experience period. If one applied a schedule credit for having a safety manager to this second risk, the effect of the safety manager would be double-counted—first by the schedule credit and

second by the experience mod. However, if the risk is too small to have fully credible experience, it might be appropriate to give some schedule credit for the safety manager—but less than for the first risk, since the impact of the safety manager is partially credited by the experience mod for the second risk, but not at all for the first.

10. Evaluating and Comparing Plans

The following definitions will be helpful for this section:

- Manual Premium—the manual premium refers to the premium calculated based on the criteria in the rating manual. In its simplest form, this is the exposure multiplied by the rates found in the rating manual. This is effectively the premium for a risk before the application of experience rating.
- Standard Premium—this is the premium after the application of the experience rating mod. This is sometimes also referred to as the modified premium, in reference to the fact that it includes the impact of the experience rating modification factor.

In the following discussions of Standard Premium and Standard Loss Ratios in this chapter, we ignore the schedule mod.

An effective experience rating plan should do two things. It should identify risk differences among otherwise similar risks, and it should adjust for them. There is a simple qualitative test that can be used to evaluate a plan based on these criteria, sometimes referred to as the Quintile Test, because it relies on observing the impact of the e-mod among quintiles of the set of risks subject to the plan[3].

The procedure is as follows:

- Rank order risks by the size of their mod and then collapse into five groups.
- Calculate the manual loss ratio and the standard (modified) loss ratio for each group.
- Observe any trends in the manual or standard loss ratios across the groups.

Consider the following sample of insurance risks which have been experience rated. They have already been ordered from lowest to highest mod.

Risk (1)	Manual Premium (2)	Loss (3)	Mod (4)
A	950	475	0.52
B	1,075	645	0.59
C	1,225	858	0.70
D	1,100	880	0.81
E	1,175	999	0.83
F	1,050	945	0.91
G	1,000	950	0.96
H	925	925	0.99
I	1,025	1,040	1.05

[3] Paul Dorweiler discussed a slightly more general version of this test as applied to New York Workmen's Compensation in his presidential address to the Casualty Actuarial Society at its twentieth anniversary, "A Survey of Risk Credibility in Experience Rating," *PCAS* XXI, 1934.

J	995	1,055	1.06
K	1,150	1,254	1.08
L	1,200	1,300	1.11
M	900	1,040	1.14
N	875	1,030	1.18
O	1,125	1,450	1.22

These fifteen risks would collapse into the following five groups:

Risk Group (5)	Manual Premium (6)	Loss (7)	Avg. Mod (8)	Manual Loss Ratio (9)=(7)/(6)	Standard Loss Ratio (10)=(7)/[(6)*(8)]
A-B-C	3,250	1,978	0.61	0.61	1.00
D-E-F	3,325	2,824	0.85	0.85	1.00
G-H-I	2,950	2,915	1.00	0.99	0.99
J-K-L	3,345	3,608	1.08	1.08	1.00
M-N-O	2,900	3,520	1.18	1.21	1.03

Column (6) is equal to the sum of the manual premium in column (2) for all the risks in the group; (7) is the sum of the loss in column (3) for all risks in the group; the average mod (8) is equal to the premium weighted average of the mods for each risk in the group.

The first thing we want to check is whether this experience rating plan correctly identifies differences in risks. To do this, we compare the manual loss ratios for the groups. In this plan, there is a distinct and upward trend in the manual loss ratio as the average modification factor increases. Risks with the lowest manual loss ratio received the lowest mods (i.e., they received the most credit) and the risks with the highest manual loss ratio received the highest mods. One would reasonably conclude that this plan does indeed identify differences in risks.

The second thing to check for is whether the plan reasonably adjusts for differences in the risks. To do this, we compare the standard loss ratios for the groups. Notice that the standard loss ratios are much less dispersed than the manual loss ratios and that there is no discernable trend in the standard loss ratios. It would be reasonable to conclude that this plan does indeed account for the differences in the risks.

Now let's look at this same set of risks, but use a different experience rating plan. In this example, the loss and premium experience for each risk is the same as in the previous example, but the mod for each risk is different.

Risk (1)	Manual Premium (2)	Loss (3)	Mod (4)
A	950	475	0.49
B	1,075	645	0.57
C	1,225	858	0.67
D	1,100	880	0.78
E	1,175	999	0.83
F	1,050	945	0.89
G	1,000	950	0.95
H	925	925	1.01

I	1,025	1,040	1.04
J	995	1,055	1.09
K	1,150	1,254	1.10
L	1,200	1,300	1.14
M	900	1,040	1.23
N	875	1,030	1.24
O	1,125	1,450	1.34

Grouping the risks and calculating the manual and standard loss ratios by group as we did above will give us:

Risk Group (5)	Manual Premium (6)	Loss (7)	Avg. Mod (8)	Manual Loss Ratio (9)=(7)/(6)	Standard Loss Ratio (10)=(7)/[(6)*(8)]
A-B-C	3,250	1,978	0.59	0.61	1.04
D-E-F	3,325	2,824	0.83	0.85	1.02
G-H-I	2,950	2,915	1.00	0.99	0.99
J-K-L	3,345	3,608	1.11	1.08	0.97
M-N-O	2,900	3,520	1.28	1.21	0.95

Notice that there is a downward trend in the standard loss ratios by group as the average mods increase. The risks with the best loss experience in the past now actually have higher loss ratios than the risks with the worst past loss experience in the group. This indicates that the experience rating plan is giving too much credibility to the risks' actual experience. The risks with the lowest mods are getting credit for better than average experience more than the experience is predictive of their future loss experience. The result is that their premium is reduced so much that their loss ratios are now higher than average. Likewise, the risks with the highest past loss experience are getting penalized too much under this plan. The result is that they now have lower loss ratios than the rest of the risks in the group.

This scenario is undesirable. Recall from earlier that the objectives of an experience rating plan include increasing equity and enhancing market competition. This second rating plan does not enhance equity because the risks with the highest mods are paying more premium than is equitable. This plan also does not enhance market competition. This plan generates a scenario where risks with higher past loss experience will generate a lower loss ratio and therefore higher profits. These risks will be more desirable for the insurer to write than risks with better past loss experience. This is contrary to the desire to enhance market competition by making ALL risks equal in terms of profit potential and therefore equally desirable to write.

Analogous to the previous example would be a scenario where there was an upward trend (higher standard loss ratios for groups with higher mods). This would indicate that the plan does not give enough credibility to actual experience. Risks with better than average past loss experience would not get enough credit and would continue to have lower than average loss ratios. Meanwhile, risks with higher than average past loss experience would continue to produce higher loss ratios even after the experience rating plan is applied. A good experience rating plan will have no discernable trend in the modified loss ratios.

We can also quantify the efficiency of an experience rating plan by comparing its results across the quintiles. Similar to above, we rank risks by their mod and combine them into five groups. We then calculate the manual and standard loss ratios for each group. For each plan, calculate the efficiency test statistic as the ratio of the variance of the standard loss ratio to the variance of the manual loss ratio[4]. The plan with the lower variance ratio is "better"—it does a better job at adjusting premium; i.e., it makes "risks of differing experience more equally desirable".

[4] This test statistic was first described by Robbin Gillam as a "quintiles test" in "Worker' Compensation Experience Rating: What Every Actuary should know", *PCAS* LXXIX, 1992, pp. 215-239.

Let's use the Efficiency Test to compare the experience rating plans in the last two examples. For clarity, we'll refer to the experience rating plan in the first example as Plan A and the second as Plan B.

Risk Group	Manual Loss Ratio	Plan A Standard Loss Ratio	Plan B Standard Loss Ratio
(1)	(2)	(3)	(4)
A-B-C	0.61	1.00	1.04
D-E-F	0.85	1.00	1.02
G-H-I	0.99	0.99	0.99
J-K-L	1.08	1.00	0.97
M-N-O	1.21	1.03	0.95
Sample Variance	0.0536	0.0002	0.0013
Efficiency Test Statistic		0.0039	0.0237

The test statistic for Plan A is lower than for Plan B. As expected, the result of the Efficiency Test is that Plan A is a better experience rating plan.

Questions

1. What are the objectives of experience rating?

2. Explain how experience rating increases equity.

3. Consider two insurance companies writing an identical line of business. Company A has developed a very sophisticated classification plan for rating risks which incorporates many different risk characteristics to assign a risk into one of several dozen classes. Company B has a much simpler rating plan which considers fewer risk characteristics and has only half a dozen rating classes. Which company would benefit more from using experience rating?

4. Explain how the use of experience rating can help an insurance company avoid adverse selection.

5. Discuss the concepts of process variance and the variance of the hypothetical means (VHM) and how they relate to experience rating.

6. Suppose you are pricing a risk (which will be experience rated). This particular risk has a significantly better safety program than most other risks. Would it be appropriate to apply a schedule credit to reflect lower than average loss potential due to this superior safety program?

Acknowledgments

This chapter would not have been possible without significant contributions by Rick Gorvett. I would also like to thank Ginda Fisher, Lawrence McTaggart, and Jill Petker for their support, advice, and assistance.

Chapter 2: Risk Sharing Through Retrospective Rating and Other Loss-Sensitive Rating Plans

By Jill Petker

1. Risk Sharing: Risk Retention and Risk Transfer

Retrospective rating and other loss-sensitive rating plans allow risk sharing between the insured and the insurer. This chapter will look at how risk sharing is achieved through retrospective rating, large deductibles, self-insurance arrangements, and other loss-sensitive rating plans. This risk-sharing contrasts with guaranteed-cost policies, where the insured's premium is fixed up front and the insured does not share in their own risk, except perhaps through a small deductible. We start with retrospective rating because it is a direct contrast to the experience rating that you already learned about through the Basic Ratemaking syllabus and in Chapter 1 of this study note. In current practice, however, large deductible plans are more common than retrospective rating. When sharing risk, the insured's risk tolerance and financial capacity are typically better suited to their retaining the risk associated with the more predictable primary losses while transferring the risk associated with the more volatile and uncertain per-occurrence excess losses to the insurer. However, even the primary layer of loss can be volatile (driven by frequency or even severity within the primary layer). Therefore, the primary risk that the insured retains is often limited in aggregate to a specified amount. The risk of having primary losses in excess of the insured's aggregate retention is transferred to the insurer.

The advantages to the insured of risk sharing through loss-sensitive rating plans include:

- An incentive for loss control, which affects their direct costs as well as indirect costs, such as lost productivity for Workers Compensation
- The immediate reflection of good loss experience, without the lag and credibility-weighting that come with experience rating
- Cash flow benefits from paid loss retrospective rating plans, large deductible plans, and self-insurance, all of which are described further below
- A possible reduction in premium-based taxes and assessments (under large deductible plans in particular)

The disadvantages to the insured include:

- Uncertain costs, compared to a fixed premium under guaranteed-cost plans
- The loss of the immediate tax deductibility of full guaranteed-cost premium
- The immediate reflection of bad loss experience
- Impact on future financial statements
- Ongoing administrative costs, e.g., paying bills long into the future
- The need to post security as collateral against credit risk (discussed further below)
- Added complexity, compared to guaranteed-cost plans

The advantages to the insurer include:

- The insured's incentive for loss control, which is stronger than the incentive provided by experience rating alone

- Ability to write some risks which the insurer would not find acceptable to write on a guaranteed-cost basis
- Less capital required to write policies under which the insured shares in their risk. (See section on capital and profit provisions below.)

The disadvantages to the insurer include:

- Higher administrative costs
- Credit risk (discussed below)
- A reduction in cash flow to the extent that insureds pay for their retained losses over time, as opposed to paying premium to cover all of their expected losses during the policy period, as under guaranteed-cost plans
- Insureds' tendency to second-guess claims handling and ALAE costs
- Insureds' tendency to question the size of profit provisions since they are taking on a share of the risk. (Again, see section on capital and profit provisions below.)

2. What is Retrospective Rating?

You have learned about how experience rating uses an insured's loss experience from historical policy periods to adjust their premium for the upcoming policy period. In contrast, retrospective rating uses an insured's loss experience from a policy period to adjust the premium for that same policy period. Adjustments to the policy premium are made "retrospectively" upon review of actual loss experience.

Risk sharing under a retrospective rating plan follows the format outlined in the section above, in that it is generally a primary layer of loss that is used to retrospectively adjust the policy premium. However, to protect the insured from volatility, the primary losses that influence the retrospectively rated premium will generally be subject to a maximum "ratable" loss amount. That maximum ratable loss amount may either be established directly, or it may correspond to a maximum premium amount. In addition, the primary losses that influence the retrospectively rated premium may be subject to a minimum ratable loss amount, which again may either be established directly or may correspond to a minimum premium amount.

3. The Retrospective Rating Formula

Premium under a retrospective rating plan is calculated as **Premium = (B + cL) × T**, where **B** is the basic premium amount, **c** is the loss conversion factor, **L** is the loss amount that will be used in the calculation, and **T** is the tax multiplier. Each of these components is discussed further below.[5]

B is called the **basic premium amount**. It reflects fixed charges (i.e., those that won't vary with actual losses), such as:

[5] If you are practicing in the US, you may find it helpful to review both NCCI's and ISO's retrospective rating plan manuals for detailed requirements and options specific to their lines of business. Examples of specifics related to those retrospective rating plans are included in the footnotes below. This chapter is intended to be a general discussion not specific to either the NCCI's or ISO's specific retrospective rating plans.

- Expenses for which the charge will be a fixed amount. Typical fixed expenses include underwriting expenses and commission (if commission is a fixed amount or a percentage of guaranteed-cost premium).
- Expected per-occurrence excess losses, if losses influencing the premium are subject to a per-occurrence loss limit. Estimating an appropriate charge for per-occurrence excess losses is often done by applying an expected ratio of excess/total losses to the total loss estimate for the policy. The estimation of expected ratios of excess/total losses is discussed in the CAS monograph *Distributions for Actuaries*. Sometimes you will see the per-occurrence excess loss component as a separate provision.[6] In that case, it may be called an excess loss premium. This excess loss premium generally includes a provision for the loss adjustment expenses associated with the per-occurrence excess losses (see more on this below).
- Expected aggregate excess losses, if losses influencing the premium are subject to a maximum ratable loss amount or if the retrospectively rated premium is subject to a maximum premium amount. This is often referred to as the **insurance charge**. Estimating an appropriate charge for aggregate excess losses will be discussed in Chapter 3. This component of the basic premium also generally includes a provision for the loss adjustment expenses associated with the aggregate excess losses.
- A credit if losses influencing the premium are subject to a minimum ratable loss amount or if the retrospectively rated premium is subject to a minimum premium amount. This amount is often referred to as the (insurance) **savings**. The combination of the savings and the insurance charge described above is often referred to as the **net insurance charge**. Estimating the savings will be discussed in Chapter 3.
- The underwriting profit provision, which will be discussed further below.

c is called the **loss conversion factor**. It covers expenses for which the charge is going to vary with actual losses. Typically the loss conversion factor would include loss adjustment expenses, and may also include loss-based assessments. If desired, expenses can be shifted back and forth between the basic premium and the loss conversion factor. However, it may not be prudent to charge for expenses that don't vary with losses through the loss conversion factor. If losses are lower than expected, those expenses will not be fully recouped.

To reflect the ability to shift expenses back and forth between the basic premium and the loss conversion factor, the following formula is often used to calculate the expense portion of the basic premium (as a percentage of the guaranteed-cost premium): $\mathbf{e - (c-1) \times E}$, where:

- **e** is the expense ratio underlying the guaranteed-cost premium. This expense ratio reflects the premium discount (recognizing that expenses are a lower percentage of premium for large accounts) but excludes premium-based taxes and assessments. It includes loss adjustment expenses.
- **c** is the selected loss conversion factor.

[6] Under NCCI's and ISO's retrospective rating plans, the charge for per-occurrence excess losses is a separate component in the formula. So the formula for retrospective premium becomes: Premium = (Basic Premium + Excess Loss Premium + c × Ratable Loss) × T

- **E** is the expected loss ratio underlying the guaranteed-cost premium.

You can see that as **c** increases, the expense portion of the basic decreases, and vice versa. **L** represents the losses that will be used to calculate the retrospective premium. These losses are often called **"ratable" losses** because they are used to calculate the retrospectively rated premium amount. Options related to these losses include:

- They may or may not include Allocated Loss Adjustment Expense (ALAE).[7] When the ratable losses include ALAE:
    - The loss adjustment expenses that would be typically covered through the loss conversion factor **c** would be just Unallocated Loss Adjustment Expense (ULAE).
    - Similarly, the expense ratio **e** would include ULAE but not ALAE.
    - The expected loss ratio **E** mentioned above would be an expected loss-and-ALAE ratio.

  When the ratable losses exclude ALAE:
    - The loss adjustment expenses that would be typically covered through the loss conversion factor **c** would be both ALAE and ULAE.
    - Similarly, the expense ratio **e** would include both ALAE and ULAE.
    - The expected loss ratio **E** mentioned above would be an expected loss-only ratio.
- They may or may not be subject to a per-occurrence loss limit. If they are, as mentioned above, the charge for the expected losses above the per-occurrence loss limit may be included in the basic premium amount or may be kept separate from the basic premium. [8,9]
    - Note that **E** in the formula above (for the expense amount to be included as fixed in the basic premium) represents total losses – limited to the policy limit if there is one, but not limited otherwise. As such, the loss conversion factor **c** is intended to be applied to total losses. Therefore, the charge for expected losses above the per-occurrence loss limit needs to have the loss conversion factor **c** applied to it.
- They may or may not be subject to an aggregate loss limit. If they are, as mentioned above, the charge for the expected (limited per occurrence, if applicable) losses above the aggregate loss limit is generally included in the basic premium amount.[10] This charge needs to have the loss conversion factor **c** applied to it for the same reason described above. Aggregate loss limits are often set in one of these two ways:
    - As a multiple of the expected losses that will be subject to the aggregate limit (i.e., either full losses or losses limited by the per-occurrence loss limit mentioned above). For example, if the per-occurrence loss limit is $250,000 and the expected limited

[7] Under ISO's retrospective rating plan, losses for commercial auto liability, general liability, and hospital professional liability must include ALAE. See their retrospective rating plan manual for details.

[8] A per-occurrence loss limit is required under ISO's retrospective rating plan but is optional under NCCI's plan. See their retrospective rating plan manuals for details.

[9] Under ISO's retrospective rating plan, ALAE is included on an unlimited basis for commercial auto liability, general liability, and hospital professional liability. However, there is an optional cross-lines accident limitation that does limit ALAE. See their retrospective rating plan manual for details.

[10] Under both NCCI's and ISO's retrospective rating plan, a maximum premium amount is required. See their retrospective rating plan manuals for details.

losses are $300,000, then the aggregate loss limit might be set at twice the expected limited losses, or $600,000. If the selected multiple is 2.5, then the aggregate loss limit would be $750,000.

- So that the maximum premium under the retrospective rating plan will be a multiple of the guaranteed-cost premium. For example, if the guaranteed-cost premium is $1,000,000 and the selected multiple is 1.25, then the aggregate loss limit would be set such that the maximum retrospectively rated premium would be $1,250,000. This method requires an iterative pricing approach, since backing into the implied maximum ratable loss increases the basic premium amount by adding a charge for the expected losses above that maximum ratable loss amount, in turn reducing the implied maximum ratable loss that will produce the selected maximum premium. Reducing the maximum ratable loss amount will then increase the insurance charge, thereby further reducing the implied maximum ratable loss. A few rounds of iteration should stabilize the maximum ratable loss amount.

Notice that under the second approach, the aggregate loss limit will automatically increase or decrease if exposures increase or decrease, since the exposure change will be reflected in the guaranteed-cost premium after the premium (exposure) audit. Under the first approach, the aggregate loss limit can also be made to increase or decrease with exposures if the limit is first calculated based on expected limited losses but then translated to a rate per exposure.

- They may or may not be subject to a minimum ratable loss amount. If they are, as mentioned above, the basic premium amount generally includes a credit. Note that even if there is no minimum ratable loss amount, there is still a minimum premium amount that will be charged – you can see from the retrospective rating formula that the minimum premium will be equal to the basic premium times the tax multiplier.
- They may be paid or incurred. If the premium is calculated based on paid losses, the plan is called a paid loss retrospective rating plan. If the premium is calculated based on incurred losses, the plan is called an incurred loss retrospective rating plan. Paid loss retrospective rating plans are often converted to an incurred loss basis at a pre-determined point in time (e.g., after five years).

There must be either a per-occurrence loss limit or an aggregate loss limit (or both) for there to be risk transfer to the insurer.

- If there is a per-occurrence loss limit but not an aggregate loss limit, and the coverage is Workers Compensation or Auto Liability (coverages with no aggregate policy limit), the insured is technically retaining an unlimited amount of loss exposure.
- If there is an aggregate loss limit but no per-occurrence loss limit, and if the aggregate loss limit is relatively low, then the aggregate loss limit can be used up by a few (and sometimes just one) large losses. This could eliminate the insured's loss control incentive before the policy is expired. If there is no per-occurrence loss limit and the aggregate loss limit is relatively high, the retrospective premium can be very unstable (driven by the volatility of large losses).

Note that total expected loss amount equals the sum of these components: the expected per-occurrence excess losses, the expected aggregate excess losses (net of any savings related to minimum ratable losses), and the expected ratable losses. Were it not for the increased incentive for loss control under loss-sensitive rating plans, the total expected loss amount under a retrospective rating plan would equal the total expected loss amount under a guaranteed-cost plan. It is a requirement under some retrospective rating plans filed in the US that the expected premium under a retrospective rating plan also equal the expected premium under a guaranteed-cost plan. However, this requirement (often called "the balance principle") does not make sense given the difference in risk transfer and the resulting difference in capital needed to support a retrospective rating plan vs a guaranteed-cost plan. See the section on Capital and Profit Provisions below.

For incurred loss retrospective rating plans, losses are typically first evaluated as of 6 months after the policy expiration (i.e. 18 months of development), and then annually thereafter. Since losses at 18 months are typically much lower than their ultimate level, the insured would typically receive a partial return of premium at that first evaluation. Then, as losses develop upwards at later evaluations, the insured would typically pay additional premium amounts. These additional premium amounts create credit risk for the insurer. See the section on Credit Risk below for a further discussion of credit risk and the ways in which insurers can mitigate that risk.

For paid loss retrospective rating plans, losses are typically evaluated monthly, beginning with the first month of the policy period. Therefore, the retrospectively rated premium amount increases as paid losses develop upwards. This creates credit risk for the insurer, which again, is discussed below.

T is called the **tax multiplier**. It is calculated as 1/(1 – tax rate), where the tax rate may include residual market and other premium-based assessments. If commission is a percentage of the net premium (i.e., net of retrospective rating adjustments), then commission would be included with the tax rate, and not in the basic premium amount.

4. Regulatory Approval and the Large Risk Alternative Rating Option (LRARO)

In the US, the pricing methodology and parameters for retrospective rating plans generally must be filed and approved by state regulators.[11] Those plan parameters include the expected loss ratio to be applied to guaranteed-cost premium in order to estimate the total expected losses, the expense ratio, per-occurrence excess losses as a ratio to total losses, the table of insurance charges that will be discussed further in Chapter 3, and the tax multiplier.

However, there is a Large Risk Alternative Rating Option under both ISO's and NCCI's retrospective rating plans in most states that allows large insureds to be retrospectively rated "as mutually agreed upon by carrier with insured." "Large" is generally defined in terms of standard premium individually or in any combination with WC, GL, Auto, Crime, and a few other lines of business. A key assumption underlying LRARO is that large risks are knowledgeable and sophisticated enough to negotiate with insurers their retrospective rating parameters. Although LRARO allows for pricing flexibility, pricing still must comply with regulatory principles and not be inadequate, excessive, or unfairly discriminatory.

In addition to allowing flexibility in pricing, LRARO also allows flexibility in structure. Examples include:

[11] Outside of the US, there is much less rate regulation for commercial insurance. Wherever you are practicing, be sure to understand and follow the rate regulation in place for that jurisdiction.

- NCCI's standard retrospective rating plan for WC only includes incurred loss retrospective rating plans. A paid loss basis requires the use of LRARO. Here, LRARO's pricing flexibility is important so that the insurer can reflect in its pricing the loss of investment income under a paid loss basis relative to an incurred loss basis.
- Maximum and minimum ratable loss amounts can be set directly, rather than indirectly through maximum and minimum premium amounts.
- The basic premium factor and/or the maximum and minimum ratable loss amounts can be based on exposures instead of standard premium, if that is deemed to be more appropriate or convenient.

5. Other Loss-Sensitive Plans

Other loss-sensitive plan types include the following:

- Large Deductibles: In the US, large deductibles for casualty lines of business are generally considered to be those at or above $100,000 per occurrence.
    - Because insurers wish to direct the handling of casualty claims from the start (and insureds are not typically set up to adjust and pay claims) insurers pay all claims up front and bill the insured for deductible reimbursements up to the per-occurrence deductible amount.
    - Like a retrospective rating plan, the losses that are subject to deductible reimbursement may or may not include ALAE.
    - Unlike a retrospective rating plan, however, the losses that are subject to deductible reimbursement must be subject to a per-occurrence loss limit (i.e., the deductible).
    - Like a retrospective rating plan, the insured's deductible reimbursements may or may not be capped at an aggregate deductible limit.
    - Unlike a retrospective rating plan, however, there is no analog to the minimum ratable loss amount. That is, there is no minimum deductible reimbursement.
    - Notice that the risk transfer under a large deductible is the same as the risk transfer under a retrospective rating plan that has a per-occurrence loss limit and a maximum ratable loss amount but does not have a minimum ratable loss amount. Per-occurrence and aggregate excess loss risk are transferred to the insurer.
    - Under a large deductible plan, the premium is generally fixed. However, the insured's cost (premium plus loss reimbursements under the deductible) is not fixed. A large deductible plan is considered to be a loss-sensitive plan because the insured's total cost for the policy period varies based on actual loss experience.
    - The premium for a large deductible must cover the same components that a retrospective rating plan's premium covers, with one exception: The premium does not cover the expected cost of losses below both the per-occurrence deductible and aggregate deductible limit.
    - Net premium (i.e., premium net of the deductible credit) still must cover the expected per-occurrence excess losses, expected aggregate excess losses (if

applicable), expenses, and an underwriting profit provision. Relative to the premium for a retrospectively rated policy:

- The charge for expected excess losses is the same.
- The provisions for most expenses are the same, except:
  - The provisions for premium tax and some premium-based assessments (if based on net-of-deductible premium) are lower because the deductible reimbursements are not premium and therefore are not subject to those taxes and assessments.
  - The provision for commission may be lower if commission is a percentage of net premium. (Alternatively, commission may be a percentage of guaranteed-cost premium, a flat allowance, or zero if a fee for service is paid directly by the insured to the agent or broker. Note that these same alternatives are available for retrospective rating plans as well.)
  - Because of these exceptions, the insured's expected cost is generally lower under a large deductible plan.

See question 9, the appendix, and the companion case study for examples of how to calculate the net premium under a large deductible plan.

- Note that as the deductible becomes large, the expected (excess) loss component of the premium can become very small relative to the expense and underwriting profit provisions. Thus the premium for a high deductible can appear to be surprisingly large. However, the expenses do not go away and the risk load applicable to the excess losses can be quite large due to the significant amount of both parameter risk and process variance.

- Self-Insured Retentions: Self-insured retentions (SIRs) are similar to large deductibles, but differ in these important ways:
  - In the US, self-insurance for Workers Compensation and Auto Liability requires regulatory approval, because they are both legally required coverages.
  - The insured is responsible for adjusting and paying claims or making arrangements for someone else to do those tasks. They may self-administer the claims or hire a third-party claims administrator. Many insurers have affiliated third-party claims administrators, so the insured may find it convenient to purchase claims-handling services from the same insurer from whom they buy excess loss coverage (per-occurrence and/or aggregate). However, some insureds value the control that choosing the party responsible for claims handling gives them. Insurers reimburse the insured for loss amounts in excess of the self-insured retention.
  - Because the insurer is not handling claims up front, it is not incurring ALAE for claims that stay within the self-insured retention. Therefore, the retention generally applies to pure loss only, and ALAE is generally shared pro-rata for claims that exceed the retention.

- In addition, because the insurer is not handling claims up front, only a minimal amount of ULAE is included in the premium for excess-over-SIR coverage. Therefore there is an even greater expense savings due to having a lower base for premium taxes and some premium-based assessments.
- Because the insurer is not responsible for claims until after they have been paid by the insured, the insurer does not take on credit risk for the loss-sensitive feature. See below for more on credit risk.
- The policy limit for excess-over-SIR coverage is generally not eroded by the self-insured retention. This contrasts with the limit for large deductible coverage, which generally is eroded by losses within the deductible. For example, excess-over-SIR coverage with a $1m limit over a $250k per-occurrence retention would cover the layer of losses between $250k and $1,250k. However, a large deductible policy with a $1m limit and a $250k deductible transfers the layer of losses between $250k and $1m, essentially only providing $750k of coverage. Note that when excess-over-SIR coverage is provided for Workers Compensation, a policy limit can be applied (as opposed to the usual statutory limits, which is essentially unlimited).

- Dividend Plans: Some dividend plans have loss-sensitive features that act similar to incurred loss retrospective rating plans, but with two important distinctions:
  - If the insured's losses are lower than expected, the money that is returned to the insured is not considered a premium credit for accounting purposes. Instead, it is considered to be an expense paid by the insurer. As such, there is no savings in premium-based taxes and assessments.
  - If the insured's losses are higher than expected, no additional money is collected from the insured. In this way, loss-sensitive dividend plans are not balanced in terms of the expected "ratable" losses.

  Like incurred loss retrospective rating plans, losses under loss-sensitive dividend plans are often evaluated six months after the policy expires and then annually thereafter. If reported losses develop upwards (the most likely scenario), the indicated dividend will decrease. Dividend amounts already paid at earlier evaluations may need to be partially recouped from the customer. Thus loss-sensitive dividend plans also create credit risk for the insurer, which is discussed below.
  Dividend payments generally are not contractually guaranteed and generally require approval from the insurer's board of directors.

6. Other Variations on Loss-Sensitive Plans

- Clash Coverage: When an insured has exposures covered by more than one loss-sensitive plan, they may wish to limit their exposure to a single occurrence that impacts their retentions across multiple lines of business. Often referred to as a Clash Deductible or Clash Aggregate, the coverage defines a single dollar amount for the sum of retained loss payments from an occurrence that impacts multiple lines of business. For example, an insured may have large deductible policies for Workers Compensation and Auto Liability

with deductibles of $250k and $100k, respectively. They may purchase clash coverage so that if an at-fault auto accident injures both their employee and a third party, their total retention will be only $300k instead of $350k. This coverage is difficult to price and may require the use of simulations with assumptions around frequencies, severities, and correlations between lines of business.

- Basket Aggregate Coverage: When an insured has exposures covered by more than one loss-sensitive plan, a Basket Aggregate (sometimes called Account Aggregate) policy can provide a total aggregate limit on all reimbursable or ratable losses from the underlying plans. Typically, the underlying plans are written with no aggregate deductible limits or maximum ratable loss amounts. A separate GL policy reimburses the insured for losses in excess of a specified maximum aggregate retention for the insured, up to a specified policy limit.

- Multi-Year Plans: Retrospective rating plans, large deductible plans, and basket aggregates are sometimes written as multi-year plans. Three years is a typical term. One goal is to stabilize costs by lengthening the experience period. The thought here is that good and bad years offset each other and reduce the insurance charge. However, loss trends for the longer policy period must be built into the charges for both per-occurrence and aggregate excess exposure. In addition, contract wording should allow for rate adjustment when exposures change significantly during the policy period. Also, credit risk increases as the insurer must evaluate the potential for the financial condition of the insured to deteriorate over a longer time horizon. Multi-year plans tend to become popular during soft markets as insureds attempt to lock in favorable rates.

- Captives: Captives are insurance companies formed to serve the insurance needs of their parent companies. They offer another avenue for risk sharing, although the risk sharing mechanism here is often reinsurance. That is, insurers will often write policies to provide coverage and then cede losses (usually primary, and usually limited to an aggregate amount) to the captive.

7. Credit Risk

Retrospective rating, large deductible, and loss-sensitive dividend plans subject insurers to credit risk. Insurers are depending on the customer to be willing and able to pay additional premium amounts, loss reimbursements, or returns of dividend amounts in the future. This is particularly true for paid loss retrospective rating plans and large deductible plans, but it is also true of incurred loss retrospective rating plans and loss-sensitive dividend plans at early maturities. Note that credit risk increases for long-tailed lines and for higher loss limits or deductibles. This is because the timeframe for collectible amounts grows longer, so the insurer is at greater risk of the insured becoming unable or unwilling to continue to pay those amounts during that timeframe.

There are several approaches available to insurers to protect themselves from this credit risk:

1. Security: The insurer can hold collateral against the amounts that are expected to be paid by the insured in the future. This approach can be used for either retrospective rating plans or

large deductible plans. For insureds with a weaker financial position, insurers may want to hold collateral to an amount higher in the range of loss outcomes.

2. Loss Development Factors: The insurer can apply loss development factors to the losses used in the retrospective premium or dividend formula. This is typically not done for paid loss retrospective rating plans, as those plans are typically intended to mimic the cash flows of a large deductible plan (see the appendix for examples of expected cash flows under an incurred retrospective rating plan vs. a large deductible plan). The loss development factors are generally established up front when the retrospective rating plan is written.[12] This option is not available for large deductible plans.
3. Holdbacks: The insurer and the insured can agree up front to defer all or a portion of retrospective premium adjustments and/or dividend payments until a specified maturity.[13] Again, this is typically not done for paid loss retrospective rating plans.

8. Setting Retention Levels

There are several considerations that should be taken into account when setting retentions levels for an insured:

- Per-occurrence retentions should generally be set so that the insured keeps the more predictable "working layer" of losses, which is the layer in which there is a relatively high rate of frequency. The insurer should take on the more volatile loss exposure above that level, where there is less frequency but where the claims can become quite large.
- The retentions should be within the insured's risk tolerance. Insureds who are more risk averse or who want more stability in their insurance-related costs will not feel comfortable taking on high retentions.
- The retentions should reflect the insured's financial capacity. When credit risk is an issue, the insurer may wish to set lower retentions in order to reduce credit risk.
- The retentions should increase with loss trend. If they do not, over time the effectiveness of the retentions will erode. This is particularly an issue for per-occurrence retentions, which are established as fixed dollar amounts. With inflation, more and more claims will exceed a fixed dollar amount. This is less of an issue for aggregate retentions, if they are set as a multiple of the expected primary losses or to produce a multiple of the guaranteed-cost premium.

9. Capital and Profit Provisions

With the exception of dividend plans, the risk transferred from the insured to the insurer under a loss-sensitive plan is lower than the risk transferred under a guaranteed-cost plan. This is

[12] Under both NCCI's and ISO's retrospective rating plans, the provision for IBNR is accomplished through factors that get applied to standard premium and then get multiplied by the loss conversion factor and the tax multiplier, for the first three adjustments (NCCI) or the first four adjustments (ISO). See their retrospective rating plan manuals for details.

[13] Holdbacks are not part of the NCCI's or ISO's filed retrospective rating plan manuals. They require the use of LRARO where those filed plans apply.

because the insured is retaining the risk for their own primary losses, up to an aggregate limit. Therefore, the capital required to support these plans is lower than the capital required to support a guaranteed-cost plan. Note, though, that the capital is not reduced in proportion to the loss sharing. As mentioned above, the customer is sharing in their less risky primary losses, and their risk is often capped. The insurer takes on the riskier per-occurrence and aggregate excess losses. Therefore the capital reduction is significantly less than the reduction in the expected loss dollars transferred to the insured. As a result, the profit provision (in dollars) is reduced, but is increased as a percentage of insured loss.

10. The Dissolution of Loss-Sensitive Rating Plans for Long-Tailed Lines

Retrospective rating adjustments and large deductible reimbursements typically continue until both parties agree to close the plan. (Sometimes there is a predetermined limit on the number of annual premium adjustments for an incurred retrospective rating plan.) An insured might want to close the plan in order to free up their balance sheet from the liabilities under the plan. This is often a desire if the insured is putting itself up for sale. Or they may want to eliminate the need to post collateral, thereby freeing up credit lines and/or saving costs associated with posting the collateral. An insurer might want to close the plan in order to eliminate the administrative costs associated with billing additional premium or loss reimbursement amounts. Or, if the insured is going through a bankruptcy or reorganization, it may be in the interests of both parties to close the plan. However, unless the insured and insurer are at least somewhat in agreement about the amount of future development on the losses under the plan, or unless the terms of closing the plan are predetermined at the time of sale, it is unlikely that an agreement on the cost of closing the plan will be reached.

Retrospective rating plans are generally closed through what is called a retrospective rating plan **closeout**. This closeout is generally achieved by applying final loss development factors to the losses in order to determine the final premium amount. As mentioned above, sometimes the terms of a future closeout are predetermined when the plan is initially written.

Large deductible plans may be closed through either a large deductible **buyout** or a **loss portfolio transfer**. A buyout is an agreement between the insurer and insured where, for a fee, the insurer assumes the liabilities related to the deductible layer of loss. These liabilities may include loss-based assessments associated with those losses. A loss portfolio transfer is a separate policy under which the insured's remaining loss obligations are ceded to an insurer or reinsurer.

Self-insured retentions are closed through loss portfolio transfers.

Questions

1. Why is there no credit risk related to self-insured retentions?
2. Given a tax rate of 5%, calculate the tax multiplier.
3. Given a tax multiplier of 1.05, calculate the tax rate.
4. Why is the tax multiplier minus 1 higher than the tax rate?
5. Given the following, calculate the amount of expenses (as a percentage of guaranteed-cost premium) that will be collected through the basic premium, as a percentage of the guaranteed-cost premium:
    - The loss conversion factor is 1.10.
    - The expected loss ratio is 0.70.
    - The expense ratio (excluding premium-based taxes and assessments) is 0.20.
6. Given the following, calculate the loss conversion factor:
    - The expense ratio (excluding premium-based taxes and assessments) is 0.25.
    - The expected loss ratio is 0.65.
    - The amount of expenses to be collected through the basic premium, as a percentage of the guaranteed-cost premium, is 0.15.
7. Given the following, calculate the retrospectively rated premium amount:
    1. The basic premium amount is $150,000.
    2. The loss conversion factor is 1.10.
    3. The tax multiplier is 1.031.
    4. The per-occurrence loss limit is $100,000.
    5. The maximum ratable loss amount is $500,000.
    6. There are 15 claims on the policy. Ten of those claims are under $10,000 and total $25,000. The other 5 claims have values of:
        - $15,000
        - $25,000
        - $50,000
        - $100,000
        - $1,000,000
8. How does the basic premium as a percentage of guaranteed-cost premium change as:
    - The loss conversion factor increases?
    - The loss limit increases?
    - The maximum premium or maximum ratable loss increases?
    - The minimum premium or minimum ratable loss increases?
    - The account size increases?

9. Given the following cost components, calculate the premium for a large deductible plan.
    - Fixed expenses are $35,000. This includes a flat dollar commission for the broker.
    - The underwriting profit provision is $5,000.
    - Loss-based expenses are 10% of losses.
    - The premium tax rate is 3%.
    - Expected losses are $300,000.
    - Expected losses limited to $250,000 per-occurrence are $270,000.
    - Expected losses limited to $250,000 per-occurrence and to $500,000 in aggregate are $260,000.
10. In what way is a loss-sensitive dividend plan unbalanced?
11. Under what conditions is the risk transfer the same for a retrospective rating plan and a large deductible plan?

Acknowledgments
I would like to thank the following people for their review and suggestions for improvement: Ginda Fisher, Howard Mahler, Kalev Maricq, Lawrence McTaggart, Fran Sarrel, Phillip Schiavone, Josh Taub, Matt Veibell, Amy Waldhauer, and Wade Warriner.
I would also like to thank the following people for answering questions related to content: Matt Hayden, Sandra Kipust, Diana O'Brien, Nancy Treitel-Moore, Jean Ruggieri, Diana Trent, Shane Vadbunker, and Chris Wallace.

Appendix: Examples of Expected Cash Flow

Examples ignore processing lags and assume no aggregate excess loss exposure.

Pricing Assumptions			
1	Initial Premium	1,100,000	
2	Expected Primary Loss & ALAE	600,000	
3	Expected Excess Loss & ALAE	300,000	
4	Commission	55,000	
5	General Expense	15,000	
6	Underwriting Profit Provision	5,000	
7	ULAE	10.0%	
8	Tax Rate	3.0%	
	Incurred Retrospective Rating Plan		
9	Basic Premium	405,000	= (3) x (10) + (4) + (5) + (6)
10	Loss Conversion Factor	1.100	= 1 + (7)
11	Tax Multiplier	1.031	= 1.0 / (1.0 - (8))
	Large Deductible Plan		
12	Premium	479,381	= {(3) + (4) + (5) + (6) + (7) * [(2) + (3)]} x (11)

Payment Patterns								
Time	Initial Premium	Primary Incurred Loss & ALAE	Primary Paid Loss & ALAE	Excess Paid Loss & ALAE	Total Paid Loss & ALAE	Commission	General Expense	ULAE
0.00	1.000					1.000	0.250	
0.25		0.107	0.021	0.001	0.014		0.438	0.073
0.50		0.263	0.072	0.005	0.050		0.625	0.162
0.75		0.454	0.145	0.020	0.103		0.813	0.265
1.00		0.655	0.234	0.050	0.173		1.000	0.380
1.50		0.773	0.409	0.150	0.323			0.492
2.50		0.879	0.635	0.350	0.540			0.655
3.50		0.939	0.798	0.600	0.732			0.799
4.50		0.974	0.904	0.800	0.869			0.902
5.50		0.989	0.956	0.900	0.937			0.953
6.50		0.997	0.977	0.950	0.968			0.976
7.50		1.000	1.000	1.000	1.000			1.000

Policyholder Cash Flows				
Incurred Retrospective Rating Plan				
Time	Primary Incurred Loss & ALAE	Premium[1]	Cumulative Cash Flow	Incremental Cash Flow
0.00	-	1,100,000	(1,100,000)	(1,100,000)
0.25	64,200	1,100,000	(1,100,000)	-
0.50	157,800	1,100,000	(1,100,000)	-
0.75	272,400	1,100,000	(1,100,000)	-
1.00	393,000	1,100,000	(1,100,000)	-
1.50	463,800	943,485	(943,485)	156,515
2.50	527,400	1,015,608	(1,015,608)	(72,124)
3.50	563,400	1,056,433	(1,056,433)	(40,825)
4.50	584,400	1,080,247	(1,080,247)	(23,814)
5.50	593,400	1,090,454	(1,090,454)	(10,206)
6.50	598,200	1,095,897	(1,095,897)	(5,443)
7.50	600,000	1,097,938	(1,097,938)	(2,041)
Large Deductible Plan				
Time	Premium	Deductible Loss Reimburse-ments	Cumulative Cash Flow[2]	Incremental Cash Flow
0.00	479,381	-	(479,381)	(479,381)
0.25	479,381	12,600	(491,981)	(12,600)
0.50	479,381	43,200	(522,581)	(30,600)
0.75	479,381	87,000	(566,381)	(43,800)
1.00	479,381	140,400	(619,781)	(53,400)
1.50	479,381	245,400	(724,781)	(105,000)
2.50	479,381	381,000	(860,381)	(135,600)
3.50	479,381	478,800	(958,181)	(97,800)
4.50	479,381	542,400	(1,021,781)	(63,600)
5.50	479,381	573,600	(1,052,981)	(31,200)
6.50	479,381	586,200	(1,065,581)	(12,600)
7.50	479,381	600,000	(1,079,381)	(13,800)

[1] Premium under the Incurred Retrospective Rating Plan begins as the Initial premium of $1,100,000. Starting at 18 months (time 1.5), the retrospective rating formula applies. Here, the policyholder gets a partial return of premium at 18 months, but then pays additional premium amounts at 30, 42, 54, 66, 78, and 90 months. These additional premium amounts create credit risk for the insurer.

[2] Cash flow for the policyholder includes both premium payments and deductible loss reimbursements.

Insurer Cash flows

Incurred Retrospective Rating Plan

Time	Premium		Total Paid Loss & ALAE	Commission	Premium Tax	General Expense	ULAE	Cumulative Cash Flow[3]	Incremental Cash Flow
0.00	1,100,000		-	55,000	33,000	3,750	-	1,008,250	1,008,250
0.25	1,100,000		12,900	55,000	33,000	6,570	6,570	985,960	(22,290)
0.50	1,100,000		44,700	55,000	33,000	9,375	14,580	943,345	(42,615)
0.75	1,100,000		93,000	55,000	33,000	12,195	23,850	882,955	(60,390)
1.00	1,100,000		155,400	55,000	33,000	15,000	34,200	807,400	(75,555)
1.50	943,485		290,400	55,000	28,305	15,000	44,280	510,500	(296,900)
2.50	1,015,608		486,000	55,000	30,468	15,000	58,950	370,190	(140,310)
3.50	1,056,433		658,800	55,000	31,693	15,000	71,910	224,030	(146,160)
4.50	1,080,247		782,400	55,000	32,407	15,000	81,180	114,260	(109,770)
5.50	1,090,454		843,600	55,000	32,714	15,000	85,770	58,370	(55,890)
6.50	1,095,897		871,200	55,000	32,877	15,000	87,840	33,980	(24,390)
7.50	1,097,938		900,000	55,000	32,938	15,000	90,000	5,000	(28,980)

Large Deductible Plan

Time	Premium	Deductible Loss Reimburse-ments	Total Paid Loss & ALAE	Commission	Premium Tax	General Expense	ULAE	Cumulative Cash Flow[4]	Incremental Cash Flow
0.00	479,381	-	-	55,000	14,381	3,750	-	406,250	406,250
0.25	479,381	12,600	12,900	55,000	14,381	6,570	6,570	396,560	(9,690)
0.50	479,381	43,200	44,700	55,000	14,381	9,375	14,580	384,545	(12,015)
0.75	479,381	87,000	93,000	55,000	14,381	12,195	23,850	367,955	(16,590)
1.00	479,381	140,400	155,400	55,000	14,381	15,000	34,200	345,800	(22,155)
1.50	479,381	245,400	290,400	55,000	14,381	15,000	44,280	305,720	(40,080)
2.50	479,381	381,000	486,000	55,000	14,381	15,000	58,950	231,050	(74,670)
3.50	479,381	478,800	658,800	55,000	14,381	15,000	71,910	143,090	(87,960)
4.50	479,381	542,400	782,400	55,000	14,381	15,000	81,180	73,820	(69,270)
5.50	479,381	573,600	843,600	55,000	14,381	15,000	85,770	39,230	(34,590)
6.50	479,381	586,200	871,200	55,000	14,381	15,000	87,840	22,160	(17,070)
7.50	479,381	600,000	900,000	55,000	14,381	15,000	90,000	5,000	(17,160)

[3] Insurer cash flows under the Incurred Retrospective Rating Plan equals the premium collected less losses and expenses paid.

[4] Insurer cash flows under the Large Deductible Plan equals the premium and deductible loss reimbursements collected less losses and expenses paid.

Chapter 3: Aggregate Excess Loss Cost Estimation

By Ginda Kaplan Fisher

1. Overview

1.1. Who Pays, and How Much?

A critical part of modeling the cost of an insurance contract is determining who pays, and how much. When an insurance policy includes risk sharing at an aggregate level, it can be quite challenging to model these aggregate losses and to determine the coverage responsibilities among the parties to an insurance contract—e.g., the policyholder and/or the insurer. How to do so will depend upon the specific nature and parameters of the contract. Estimating the cost of various slices of the aggregate losses is important in estimating insurance costs when:

- A retrospectively rated policy (or "retro") is considered, as retrospectively rated polices have a maximum ratable loss (max). The impact of aggregate losses on the policy premium are limited by the max.
- A retrospectively rated policy has a minimum ratable loss (min).
- A deductible policy has an aggregate limit.
- A policy is written over a self-insured retention, limiting the customer's aggregate losses.
- A (re)insurance policy has an aggregate limit on the total it will pay out, but the data (or mathematical functions used to estimate the data) used to price the policy is not subject to that limit.

Americans are probably most familiar with aggregate loss costs in health insurance. It is common for US health insurance policies to have a deductible and/or co-payment, but an annual limit on "out of pocket" costs. That is, there is an aggregate limit on the deductible plus co-payment (where a co-payment is really just another type of deductible—so the two combined are the total deductible for the policy).

For example, a policy might pay 80% of medical costs incurred after you pay a $2000 annual deductible. (That is, a 20% co-payment.) But your out-of-pocket medical costs will be capped at $10,000. So if you get very ill, the costs you are charged and the insurance payments might look like this:

Exhibit 3.1. Illustrative medical costs

Date	Gross Medical Cost Incurred	Payment toward Annual Deductible	Insured's Co-Payment	Insurance Payment	Insured's Cost for this month	Insured's cost so far this year
Jan	$1,000	$1,000	0	0	$1,000	$1,000
Feb	$5,000	$1,000	$800	$3,200	$1,800	$2,800
Mar	$20,000	0	$4,000	$16,000	$4,000	$6,800
Apr	$20,000	0	$3,200	$16,800	$3,200	$10,000
May	$10,000	0	0	$10,000	0	$10,000
Jun	$4,000	0	0	$4,000	0	$10,000

Here, you finished paying the $2000 flat deductible partway through February, and then paid 20% of the medical expenses incurred until you paid the out-of-pocket cap of $10,000 partway through April. In this example, you recovered in June and stopped incurring medical payments that year.

A commercial liability policy might have a per-claim deductible of $100K and an aggregate limit on the deductible of $500K. In the insurance industry, this type of policy is often referred to as a "large deductible policy" or a "large dollar deductible policy," in order to distinguish it from, for example, a Homeowners' policy with a $500 deductible. For simplicity, hereafter it will just be referred to as a deductible policy. A similar example for this policy might look like this:

Exhibit 3.2. Illustrative general liability costs

Date	Dollars of loss on claims that are each less than $100K	Number of claims over $100K	Dollars of loss on claims over $100K	Deductible	Insurance payment	Insured's cost so far this year
Q1	$132,500	0	0	$132,500	0	$132,500
Q2	$93,000	2	$350,000	$293,000	$150,000	$425,500
Q3	$105,000	0	0	$74,500	$30,500	$500,000
Q4	$122,500	1	$150,000	0	$272,500	$500,000

In this case, the insured pays all the losses on claims less than $100K, and pays the first $100K of each large claim, until the aggregate limit of the deductible is reached in Q3. After that, the insurance company pays the rest of the losses incurred under the policy. Of course, in typical years, the insured would not incur enough large claims to exhaust the aggregate limit.

The out-of-pocket maximum or aggregate limit on the deductible is a benefit to the insured (and a cost to the insurer). In general, the same mathematical tools can be used to estimate any "slice" of aggregate loss, whether a cost or a savings to the insurer. It is important to pay attention to which party benefits from any particular aggregate limit. When confused, it is often helpful to imagine a specific situation and ask, "how much does the insured pay before the insurer is responsible? How

much does the insurer pay before hitting its policy limits? How much is the insured responsible for above the policy limits?

This chapter focuses on retrospectively rated and deductible plans. It is clearer to develop the math of aggregate loss cost limitations in the context of a simple retrospective policy that has no per-occurrence loss limits. This allows us to delay introducing the complications of also needing to consider the impact of any per-occurrence limitations, so much of the chapter will be written from that perspective. Then this chapter will go on to explain how to incorporate per-occurrence limitations. Keep in mind that the tools described in this chapter can work for all the situations above.

This approach is consistent with the historical development of the math around aggregate insurance losses. Many of the early papers on aggregate excess loss costs were written from the perspective of US workers' compensation policies. Retrospective rating was introduced for workers' compensation a couple decades after the coverage was invented, as a way to more fairly charge premium to safer and less safe employers.[14] Workers' compensation policies have no policy limit on the insurer's liability (except for the limitations imposed by the human lifespan) and some retrospectively-rated polices have no per-claim loss limitation on ratable losses used in calculating the retrospective premium.

But the reader should be aware that deductible policies are far more important and widespread than retrospective policies today. For that reason, deductibles will be discussed alongside retros when the topic is relevant to policies with per-occurrence loss limitations.

For simplicity of language, this chapter will often refer to unlimited losses. This is a natural way to describe workers' compensation losses, because there is no limit to the insurer's liability under the policy. And most of this material was originally developed in the context of US worker's compensation, so this language is consistent with most of the literature. However, all the math works the same if you substitute "losses to policy limit" for "unlimited losses" when you are working with other coverages. In real life, the actuary must be careful to keep track of whether the word "limit" refers to a policy limit, a deductible limit, or some other limit.

From the point of view of the policyholder, a deductible with an aggregate limit looks the same as a retro with a loss limit (with respect to ultimate losses retained). For example, the insured who buys a large deductible policy with a deductible of $250,000 and an aggregate deductible limit of $500,000 is in essentially the same position as an insured who purchases a retro with a maximum that translates to $500,000 of loss, and a per loss limit of $250,000 (ignoring the fact that there might be some differences in the treatment of expenses). The language is a little different—what we call the per-claim (or per-occurrence) deductible on a large deductible policy corresponds to the loss limitation

[14] The first retrospective rating plan for Workmen's Compensation, as it was then called, was approved by Massachusetts in 1936, as described by Sydney Pinney in "Retrospective Rating Plan for Workmen's Compensation Risks," *PCAS* XXIV.

on a retro; what we call an aggregate limit on a deductible corresponds to the maximum on a retro—but the general structures are the same. In particular, the risk transfer is the same.[15]

The reader should be aware that the timing and accounting for the monies that flow between insurer and insured are different for different types of policies, even if the risk transfer is essentially the same. For example, in a retro plan, risk-sensitive future cash flows are typically premium, and typically they only happen once a year. Those cash flows are losses in a deductible plan, and the deductible losses may be billed and paid monthly. There are also plans with loss-sensitive dividends, which are an expense to the insurer (not premium or loss) and are usually calculated annually. But while the timing of cash flows and other aspects of the plans might differ, the expected ultimate loss, which is the subject of this chapter, is the same. Loss sensitive dividend plans and self-insured retention plans can have similar loss provisions, as discussed in chapter 2.

This chapter focuses mostly on aggregate limits of primary losses, because insurers typically have more information about those losses, and thus more methods of estimating them. But similar methods can be used to price policy limits when the actuary lacks a history of relevant data but has a reasonable idea of the underlying frequency and severity distributions.

1.2. Some definitions and notation to describe aggregate losses

It is important to remember that losses are random processes, and a particular outcome (for example, the losses that a risk incurs during a policy year) is unlikely to match the expected value.

First, consider a retrospectively rated policy with no per-claim limit. This is common on smaller policies, where the maximum ratable loss might easily be breached by one large claim.

The following notation and definitions are used throughout this chapter:

N: the random variable representing the number of claims that a risk incurs during the relevant period (usually the life of a policy).

The expected claim frequency is the expected number of claims divided by the exposures or premium of the risk. We might also consider the frequency of large or small claims.[16]

---

[15] Or nearly the same. There might be some differences due to the timing of the payments, and what sort of security is required.

[16] A policy might be written per-claim, or per-occurrence, but for simplicity, this chapter will refer to the insured event as a claim, and use the terms "claim" and "occurrence" interchangeably. In real life, there might be sub-limits per claim, as well as limits per-occurrence, or other differences between a claim and an occurrence. The same general methods can be used to estimate expected losses under such policies, but working out the details is beyond the scope of this chapter. Similarly, a loss sensitive rating plan might contemplate loss or loss + ALAE. The two would have different expected loss distributions. But investigating the expected difference is beyond the scope of this study note.

X: random variable representing a claim incurring to a risk.

The expected severity is E{**X**}, the expected value of a single claim, should it occur.

A: random variable representing the actual total aggregate loss incurring to a risk.[17]

E= E{**A**}: expected loss.

Note that E{**A**} = E{**N**}*E{**X**}

Entry ratio: r = $\frac{A}{E}$: the ratio of actual to expected losses (or, equivalently, the ratio of the actual policy loss ratio to the expected loss ratio)

For example, a policy was written on a commercial auto fleet. The underwriter expected total losses on the policy to be $200,000. At the end of the year, actual losses on the policy were estimated to have been $189,000. In this case, the entry ratio r = 189K/200K = 0.945.

If the premium for that policy was $250,000, the expected loss ratio would have been 80.0% ($200K/$250K). The actual loss ratio would have been 75.6% ($189K/$250K). The entry ratio calculated from loss ratios is 75.6%/80.0% = 0.945. The two methods of determining the entry ratio are equivalent: the loss ratio calculation is simply the loss calculation with both numerator and denominator divided by the premium.

The entry ratio, **r,** is also a random variable. Although policies of different sizes tend to have different distributions of **r**, similarly sized policies of the same type of coverage (e.g., commercial auto policies in the Midwest covering fleets of private passenger vehicles, with expected losses of a few million dollars) will behave similarly, and it is customary to estimate expected aggregate excess losses in terms of their entry ratio. When aggregate charges were published in tabular form, in printed books, the charges were calculated separately for various expected loss groups (ELGs) that were similar enough to group together for analysis, and the actuary "entered the table" at the appropriate entry ratio. Empirical studies of aggregate charges are still done by grouping similar policies in this way.

The approximate size of a policy has a large influence on how its aggregate excess losses will behave. This is mostly because the variance of the loss distribution is very sensitive to the expected number of claims. Historically, policies were grouped by their expected losses (into expected loss groups) as a proxy for the expected number of claims. Policies can also be grouped directly by their expected number of claims (into expected claim count groups.) The rationale for grouping policies by size (either expected loss or expected number of claims) is discussed later in this chapter, in section 6.

[17] Note that some textbooks use S to designate this amount.

ϕ(r): Table M charge[18] = the ratio of a risk's average amount of loss in excess of **r** times its expected loss, divided by the total expected loss, or the expected percent of losses excess of rE.

> ϕ(r) is also known as the **Aggregate Excess Loss Factor**, **Aggregate Excess Ratio**, **Excess Pure Premium Ratio**, or **Insurance Charge**. (Note that this chapter will use "insurance charge" to refer to an amount, not a ratio, but the phrase is used both ways in the literature.)

Table M: a collection of related aggregate excess loss factors (and related savings, defined below). When there is a per-occurrence limit, "Table M" will refer to those factors calculated ignoring the impact of that limit. See **Table M_D** below.

Insurance Charge: ϕ(r) times the expected loss, **E**.

> This value, the expected aggregate excess loss, is often called the insurance charge because on a retrospectively rated policy, this is the portion of the retrospective premium that is fixed and pays for losses.[19] (The other premium components are variable or pay for expenses.)
>
> For example, consider an insurer with a book of 5 similar policies, each with an expected loss of $100K. In a typical year, the actual losses on those policies are $80K; $90K; $100K; $110K; and $120K. The average loss for the book is, as expected, $100K per policy.
>
> (This is not a realistic example, it was chosen to be symmetric and with small variation for illustrative purposes only.)
>
> At r=1, the aggregate excess ratio, ϕ(**1**), is the portion of each loss above 100K, divided by the expected loss:
>
> (0+0+0+10K+20K)/(100K+100K+100K+100K+100K) = 0.06.
>
> At r = 0.6, the aggregate excess ratio, ϕ(**0.6**), is the portion of each loss above 60K (60K = 0.6 * expected loss):
>
> (20K+30K+40K+50K+60K)/(500K) = 0.40.
>
> At r = 1.2, the aggregate excess ratio, ϕ(**1.2**), is the portion of each loss above 120K, or zero.

ψ(r): **Table M Savings =** the expected amount by which the risk's actual aggregate loss falls short of **r** times the expected loss, divided by the expected loss, (so, an expected percent, not an expected amount.) Or,

---

[18] Historically, the National Council of Compensation Insurers (NCCI) published aggregate excess loss factors and aggregate minimum loss factors for use with retrospectively rated US workers' compensation in a large table. Those factors were referred to collectively as "Table M." For example, see *The 1965 Table M*, by LeRoy Simon, *PCAS* LII, 1965. The terminology has passed into common usage and will be used throughout this paper.

[19] If a retro policy also has a per claim loss-limit, the charge for that is sometimes considered part of the insurance charge, and sometimes considered a separate charge. The terminology is not entirely consistent across the industry, and the actuary should be careful to understand what is being measured or estimated.

$$\boldsymbol{\psi}(\mathbf{r}) = \textit{the expected value of } max\ [r - \frac{A}{E}, 0],$$

An empirical estimate over several similar risks could be calculated as:

$$\boldsymbol{\psi}(\mathbf{r}) = \sum_{i=1}^{N} max\ [r - \frac{A_i}{E_i}, 0]\ /N$$

ψ(r) is also known as the insurance savings or aggregate minimum loss factor. (As with the phrase "insurance charge", note that this chapter will use "insurance savings" to refer to an amount, not a ratio, but the phrase is used both ways in the literature.)

Retrospectively rated policies often have a minimum ratable loss as well as a maximum ratable loss. This is the minimum aggregate loss that factors into the retrospective premium calculation. Just as the maximum aggregate loss that the insured will pay for generates an insurance charge, the minimum ratable loss the insured will pay for even if it incurs no claims over the policy period generates an insurance savings that offsets the insurance charge (or is subtracted from the charge to generate a net insurance charge. The ratio of the net insurance charge to the expected loss is called the Net Table M charge, or the net aggregate loss factor).

Continuing the simple example as above, a book of 5 similar policies, each with an expected loss of $100K, and a typical loss distribution of $80K, $90K, $100K, $110K, and $120K:

At r=1, the insurance savings, $\psi(1)$, is the portion that each loss falls short of 100K, divided by the expected loss:

(20K+10K+0+0+0)/(100K+100K+100K+100K+100K) = 0.06.

At r = 0.6, the insurance savings, $\psi(0.6)$, is the portion that each loss falls short of 60K (0.6 * expected loss):

(0+0+0+0+0)/(500K) = zero.

At r = 1.2, the insurance savings, $\psi(1.2)$, is the portion that each loss falls short of 120K:

(40K + 30K + 20K + 10K + 0)/(500K) = 0.20.

Charges and Savings: More precisely, let

Y = **A**/**E**, actual loss in units of expected loss (i.e., the entry ratio)

F(Y) = the cumulative distribution function of Y.

Then

$$\phi(\mathbf{r}) = \int_r^\infty (y - r) dF(y)$$

and

$$\psi(\mathbf{r}) = \int_0^r (r - y) dF(y)$$

By definition, both $\phi(\mathbf{r})$ and $\psi(\mathbf{r})$ are non-negative for every r.

While the expected loss to the risk is **E**, there is often a great deal of variance in the distribution of **A**, the actual loss. For example, if 100 similar risks each have the same expected loss, E, we would expect some of them to actually have more loss, and others less than expected. Thus, in general, we expect both $\phi(\mathbf{r})$ and $\psi(\mathbf{r})$ to be positive numbers for most non-negative values of **r**.[20]

$\mathbf{A_D}$: The actual policy loss, with each claim or occurrence limited to D[21].

[20] Note that in unusual cases where all risks always have losses close to what is expected, the charges and savings are zero for many values of r, such as in the overly simple example of five policies, above.

[21] It was pointed out to the author that Bahnemann uses the same notation to refer to excess rather than primary losses. This is unfortunate for candidates studying both for CAS exam 8, but actuaries should always be aware of what notation means in context.

Many policies have per-occurrence limits as well as aggregate limits. If a retrospective policy has a per-occurrence limit, the actuary might estimate the expected excess loss separately, and then look at the function of limited losses.

Expected Primary Losses: $E\{A_D\}$, the expected value of the losses limited by the per-occurrence limit.

k: the excess ratio for the per-occurrence limit. That is, $k = \frac{E - E\{A_D\}}{E}$

Table M_D: A table of related aggregate excess loss factors and related saving developed using data in which the individual losses have been limited by a per-occurrence limit prior to being aggregated into policy outcomes for use in developing those charges and savings.

For example, $M_{\$100,000}$ has had a per-occurrence limit of $100,000 applied.

Limited Table M factors are developed exactly the same as unlimited Table M factors, except we use the distribution of limited (primary) losses.

r = $A_D / E\{A_D\}$, the entry ratio of the limited distribution, the actual policy loss on the policy in units of expected primary loss; and

$\mathbf{F_D}(r)$ = the cumulative distribution function of r, the limited losses whose unlimited cumulative distribution function was given by F.

Note that the limited Table M charge (or savings) in this case will be the ratio of a risk's average amount of limited loss in excess of (entry ratio) **r** times its expected limited loss, divided by the total expected limited loss.

Table L: It is also possible to calculate the total amount of loss that will be covered by the policy (per-occurrence excess plus aggregate excess) directly, as a single factor to expected loss. That amount is known as the Table L charge, for the California Table L.[22] It will be described in more detail in section 5.1.

Note that when considering Table L calculations, the entry ratio (r) is defined as the actual limited aggregate losses divided by the expected unlimited aggregate losses**.**

$\boldsymbol{\phi_D^*(r)}$: the Table L charge at entry ratio r and per-occurrence limit D for aggregate and per-occurrence loss. This is defined as the average difference between a risk's actual unlimited loss and its actual limited loss, plus the risk's limited loss in excess of rE.
The Table L insurance charge at entry ratio $r \geq 0$ is defined as:

$$\phi_D^*(r) = \int_r^\infty (y - r) dF^*(y) + k$$

[22] Skurnick, D., "The California Table L," *PCAS* LXI, 1974, pp. 117-140.

ψ^*_D(r): the Table L savings at entry ratio r and per-occurrence limit D, ψ^*_D(r), is defined as the average amount by which the risk's actual limited loss falls short of r times the expected unlimited loss.

$$\psi^*_D(r) = \int_0^r (r - y) dF^*(y)$$

Note that the Table L charge and savings are both expressed as ratios to expected unlimited loss.

As mentioned above, the claims covered by a deductible policy with an aggregate deductible limit are the same as the claims covered by a retrospectively rated policy with the same per-claim limit and maximum ratable loss entering the retrospective rating formula.

The amount of premium will be different, however. For a retrospective policy, the insurer pays all of the losses and the insured pays premium. The premium will be comparable to that of a fully insured policy, although the actual amount of premium to be paid to the insurer is uncertain until all the claims have settled. In contrast, for a deductible policy the insurer is reimbursed for losses below the deductible (subject to limit of the aggregate deductible amount) and the insured's premium is a fixed amount much smaller than that for a fully insured policy. For a deductible policy, the pure premium is the sum of the expected per-occurrence excess loss and the expected aggregate excess loss, and the total premium is the pure premium grossed up for the risk charge and other expenses. In the case of a deductible policy, the uncertainty is in the amount and timing of the loss reimbursements that the insured will have to pay to the insurer.

In Section 2 of this chapter we will try to give a better intuitive understanding of these entities by drawing pictures of them. This will be accompanied by descriptions of some important relevant calculations.

Questions

1. A policy has a $10,000 per-occurrence deductible, a $25,000 aggregate deductible limit, and a per-occurrence policy limit of $1M. Over the course of the policy, the insured incurs the following losses, in chronological sequence:

$3,000 $8,000 $14,000 $12,000 $18,000

Determine (i) the total insurance policy coverage, and (ii) the amount for which the insured is responsible after the insurance coverage, for each of the following:
(a) After the first three claims have been incurred
(b) After the first four claims have been incurred
(c) After all five claims have been incurred

2. Medium Manufacturing Company (MMC) buys a General Liability policy with a large deductible. The policy has a $250K per-claim deductible, covers claims up to $1M per claim (from the first dollar, so the insured amount is actually $1M less $250K, or $750K xs $250K[23]) with an aggregate limit on the policy of $5M and an aggregate limit on the deductible of $1M. During the policy period, MMC has the following claims:

- 25 small claims that collectively cost $500K
- 1 claim for $100K
- 1 claim for $300K
- 1 claim for $2M

(a) What is the total loss sustained by MMC prior to any consideration of insurance?

(b) What is MMC's total loss responsibility under the per-claim deductible (but before consideration of the aggregate limit of the deductible)?

(c) How much of MMC's deductible losses are above the aggregate limit on the deductible?

(d) How much is over the per-claim policy limit?

(e) How much loss would be paid by the insurer prior to consideration of the policy's aggregate limit?

(f) How much loss is over the policy's aggregate limit?

(g) How much in total will the insurance company need to pay for MMC's liability?

3. Let A, the total aggregate loss random variable, have a continuous uniform distribution from 0 to 100. Let E, the expected aggregate losses, be the mean of the uniform distribution, or 50. Find the Table M Insurance Charge associated with

[23] This is often denoted just "$750K x $250K", or even "750x250" in practice.

(a) A = 40

(b) A = 50

(c) A = 60

4. Let aggregate loss random variable A be an exponential distribution with a mean of 10. Find the Table M (Insurance) Savings associated with

(a) A = 5

(b) A = 10

(c) A = 15

2. Visualizing Aggregate Excess Losses

The mathematics of per-occurrence excess and aggregate excess loss coverage can often be challenging, so it can be helpful to think about the questions graphically. Section 2 of this chapter is adapted from a paper by Yoong-Sin Lee.[24] This paper is so widely used in the casualty actuarial field that the graphs he described are often referred to as "Lee diagrams."

While formulas are good for calculations, graphs often provide insight into the structure of a problem, and help with developing intuition. Many problems are hard to understand until you draw a picture. Per-occurrence excess and aggregate excess loss calculations can be unintuitive, and it's often helpful to draw a picture specifying which layers will be paid by which party before commencing with the calculations. A good graphical presentation can not only provide insight into the abstract relations, it can also make the mathematical procedure much easier to follow compared with algebraic manipulations. For those who always prefer algebra, it will serve at least as a very useful supplement to the algebraic treatment.

Note that a key feature of Lee diagrams is that "size" (severity, or aggregate loss, or entry ratio) is on the vertical axis, and the horizontal axis represents the cumulative claim count or cumulative % of loss distribution. In that sense, Lee diagrams are slightly different from what many actuaries are used to seeing with respect to probability functions.

2.1. Lee Diagrams of Severity Distributions

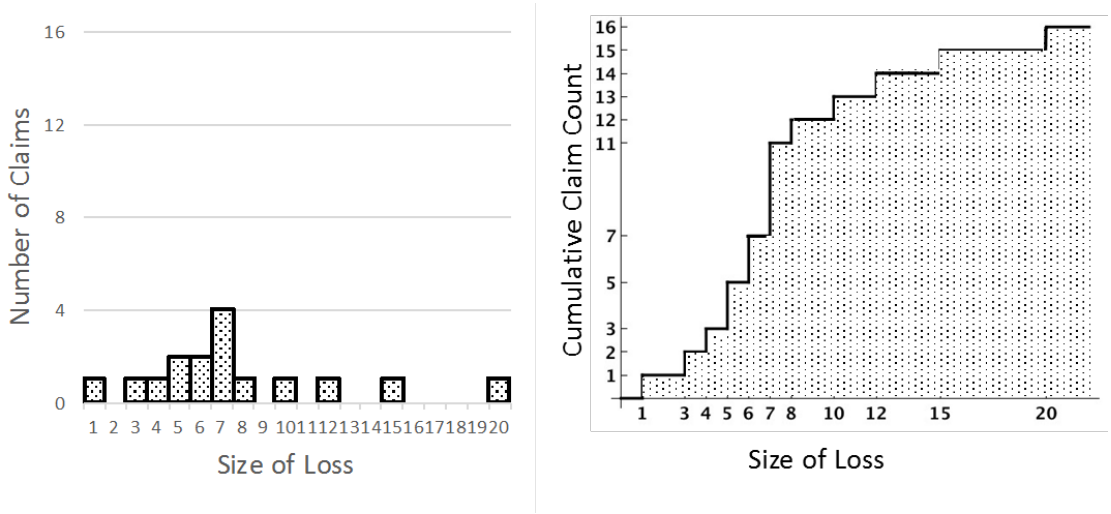
To develop the idea of what Lee diagrams look like, consider the case of per-occurrence deductibles and limits.

To start with, consider a large number of losses, of ordered sizes $x_1, x_2, \ldots, x_k$, occurring $n_1, n_2, \ldots, n_k$ times, respectively, with $n = n_1 + \ldots + n_k$.

We might, for example, look at the distribution of those losses. In Exhibit 3.3, we show the incremental and cumulative loss distribution of a set of losses, with size of loss on the X-axis.

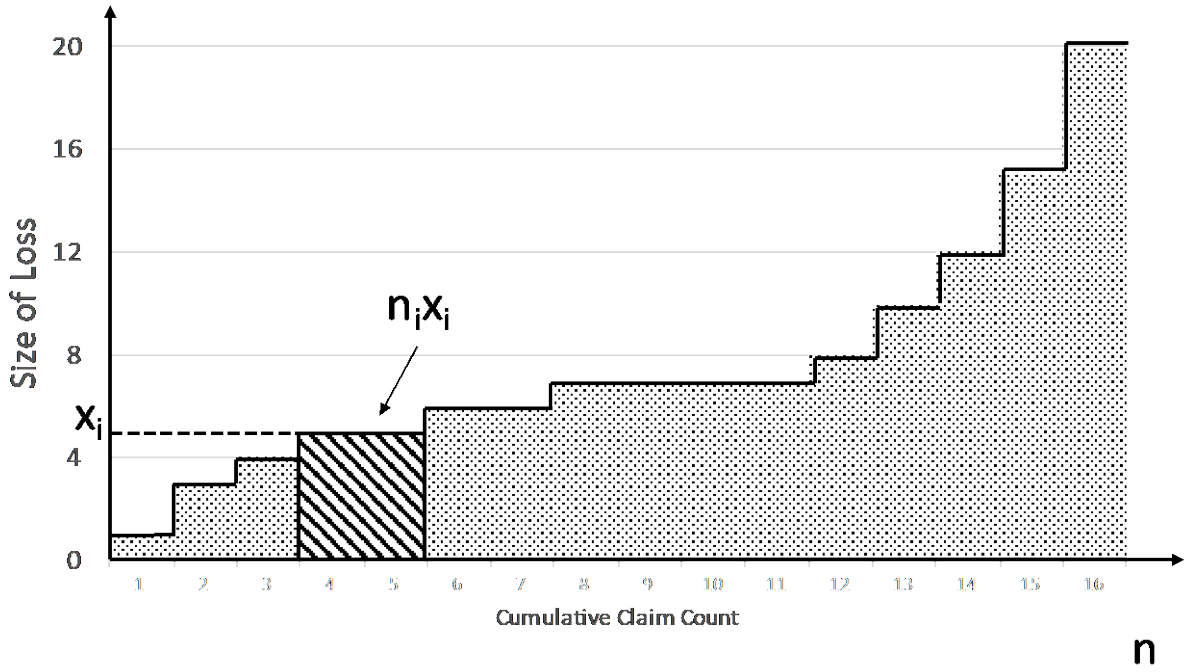
[24] Lee, Yoong-Sin, "The Mathematics of Excess of Loss Coverages and Retrospective Rating—A Graphical Approach," *PCAS* LXXV, 1988.

Exhibit 3.3. Size-of-loss on the X-axis
Incremental and Cumulative loss distributions



In Exhibit 3.4 we represent these same losses by means of a cumulative frequency curve, in which the y-axis represents the loss size, and the x-axis represents the cumulative number of losses, $c_i = n_l + \ldots + n_i$, $i \leq k$. This is how Lee diagrams are constructed.

Exhibit 3.4. A Cumulative Frequency Curve—Size-of-loss on the Y-axis



The curve is a step function (with argument along the vertical axis) which has a jump of n_i at the point x_i. Consider the shaded vertical strip in the graph. It has an area equal to $n_i x_i$. Summing all such vertical strips we have

Total amount of loss = $n_1x_1 + \ldots + n_kx_k$.

We may therefore interpret the area of the vertical strip corresponding to x_i as the amount of loss of size x_i, and the total enclosed area below the cumulative frequency curve as the total amount of loss.

In fact, we have a new way of viewing the cumulative frequency function curve. This curve can be constructed by arranging the losses in ascending order of magnitude, and laying them from left to right with each loss occupying a unit horizontal length.

Now let X be a random variable representing the amount of loss incurred by a risk. Define the cumulative distribution function (cdf) F(x) as

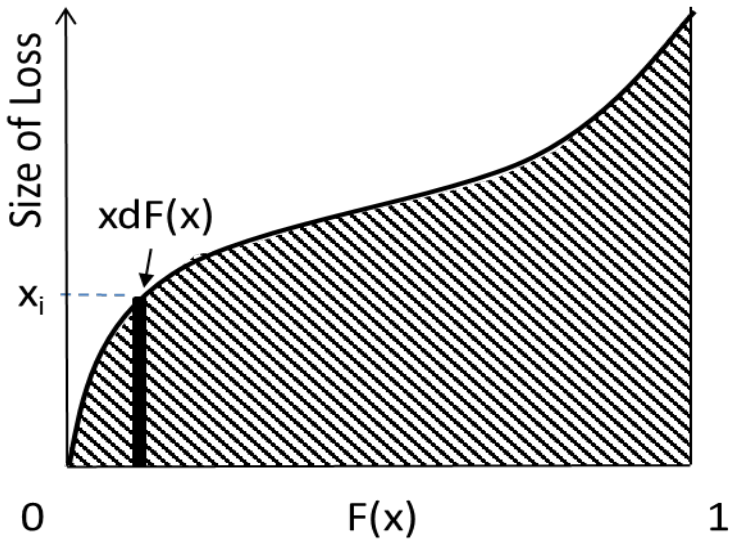
$$F(x) = \Pr(X \leq x).$$

Exhibit 3.5 shows the graph of a continuous cdf. Consider the vertical strip in the graph, with area xdF (x). If we sum up all these strips, we will obtain the expected value of X (where E{X} represents the expected value of a random variable X),

$$E\{X\} = \int_0^\infty xdF(x),$$

which is represented by the enclosed area below the cdf curve (the shaded area in the graph). We may interpret the expected loss as composed of losses of different sizes, and the strip xdF (x) as the contribution from losses of size between x and x+dx.

Exhibit 3.5. CDF Curve and Expectation



We can readily modify this diagram to visualize limits and deductibles:

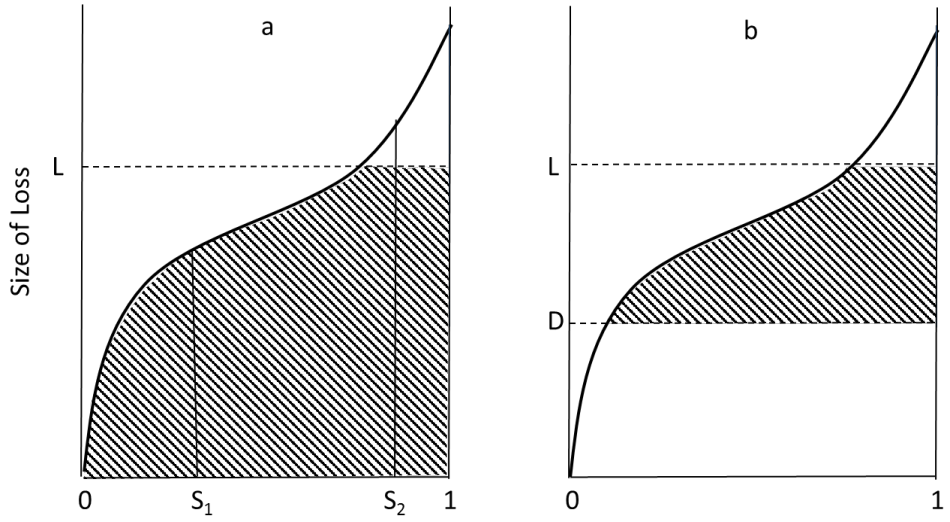
Limits:

Consider a coverage which pays for losses up to a limit L only. Exhibit 3.6(a) shows that a loss of size not more than L, such as S_1, is paid in full, while a loss of size S_2 which is greater than L, is paid only an amount L. By summing up vertical strips as before, except that strips with length greater than L are limited to length L, we obtain the expected payment per loss under such a coverage as the shaded area in Exhibit 3.6(a).

Deductibles:

Likewise, a coverage which pays for losses subject to a flat deductible D and up to limit L has expected payment per loss represented by the shaded area in Exhibit 3.6(b).

Exhibit 3.6. Expected Loss with (a) Limit and (b) Deductible



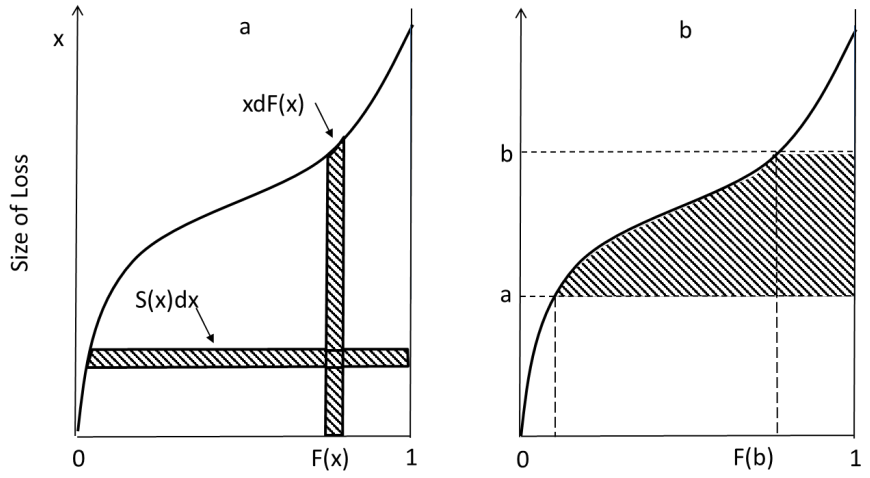
We have shown the integral along the x-axis, but as with any other measurement of area, one could just as well integrate in horizontal slices along the y-axis. One method is often easier than the other in actual practice, depending on what data is available and what curves are used to estimate the underlying process.

A vertical strip has area xdF(x), and we define

$$S(x) = 1 - F(x).$$

So a horizontal strip has area S(x) dx, as shown in Exhibit 3.7(a).

Exhibit 3.7. Size and Layer Views of Losses



Summing up the vertical strips and the horizontal strips separately must give us the same area, so we have

$$\int_0^\infty x dF(x) = \int_0^\infty S(x)dx = E\{X\}$$

This result can also be derived algebraically via integration by parts.

The two modes of summation correspond, in fact, to two views of the losses. The vertical strips group losses by size, whereas the horizontal strips group the loss amounts by layer. We may therefore call them the size method and the layer method. It is often more convenient to evaluate the expected loss in a layer by layer fashion, i.e., summing horizontal strips, than by the size method, i.e., summing vertical strips. For example, consider the layer of loss between a and b in Exhibit 3.7(b). The expected loss in this layer is represented by the shaded area. The layer method of summation gives simply

$$\int_a^b S(x)dx.$$

To express this integral by the size method is more difficult. However, some reflection, with the help of Exhibit 3.7(b), yields the following expression for the integral:

$$\int_a^b xdF(x) + bS(b) - aS(a).$$

Again, the equality of the two expressions can be established via integration by parts.

The more complicated expression derived from the size method is the form commonly found in the literature. Although the integral associated with the layer method is simple in form, S(x) is a function that is generally more difficult to integrate. This disadvantage vanishes, however, when the distribution is given numerically, as, for example, when actual experience is used. The retrospective rating Table M and Table L have been constructed by the layer method, as described in subsequent sections of this chapter; see also Simon[25] and Skurnick.[26]

2.2. Lee Diagrams of Aggregate Loss Distributions

Lee diagrams are a very effective way to visualize aggregate policy provisions. They can be used — looking at an individual policy — to keep track of who owes what when, and also to visualize the outcome of a large group of similar policies, to aid in calculating expected aggregate excess losses.

In practice, when pricing aggregate policy provisions, the actuary needs actual numbers, and those are commonly pre-calculated, or estimated with an accessible set of formulas. This study note will show how such factors can be calculated. It will start by considering the simple case, where there is no per-occurrence loss limitation. It will later consider the more general cases where a policy might have both per-occurrence and aggregate loss limitations, looking at limited Tables M and Table L.
The mathematical basis of Table M is a distribution of entry ratios and its underlying distribution of aggregate losses. Recall from Section 1.2 that r is the entry ratio (actual loss divided by expected loss).

$$Insurance\ Charge\ \ at\ r = X = \phi(r) = \int_r^\infty (y - r)f(y)dy$$

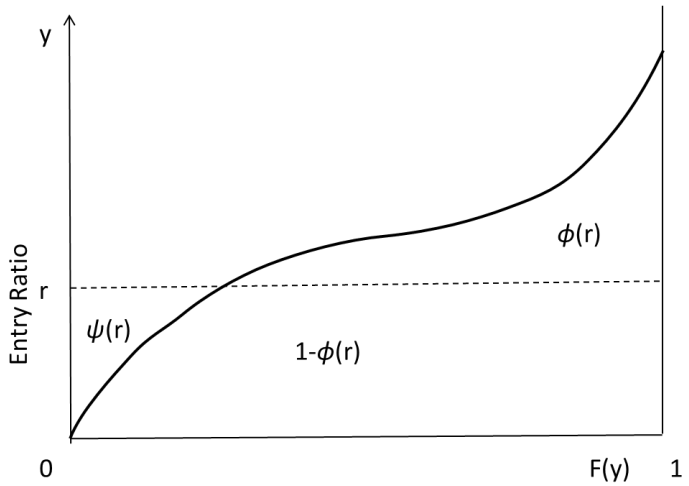
$$Insurance\ Savings\ at\ r = S = \psi(r) = \int_0^r (r - y)f(y)dy$$

$$S = X + r - 1\ or\ \psi(r) = \phi(r) + r - 1$$

(To be derived subsequently).
In estimating the expected aggregate excess loss, we need to consider the distribution of outcomes of the total claims on a policy. As with severity distributions, it can be helpful to visualize the data to gain an intuitive understanding of how the elements relate to each other. We can draw a picture, remembering that we are graphing the probability distribution of aggregate losses (which can be thought of as multiple simulations of a single policy).

[25] LeRoy J. Simon, "1965 Table M," *PCAS* LII, 1965, p. 1.
[26] David Skurnick, "The California Table L," *PCAS* LXI, 1974, p. 117.

Exhibit 3.8. Functions in Retrospective Rating



In Exhibit 3.8 the cdf F(y) is graphed against the entry ratio y. The functions $\phi(\mathbf{r})$ and $\psi(\mathbf{r})$ are represented by the areas indicated in the graph. A number of mathematical properties are now clearly demonstrated.

(1) By definition, the bounded area below the F(y) curve is equal to 1. Hence $\phi(0) = 1$.
(2) $\phi(r)$ is a decreasing function of r, and $\phi(r) \to 0$ as $r \to \infty$.
(3) $\psi(\mathbf{r})$ is an increasing function of r; its value is unbounded as $r \to \infty$.
(4) Consider a small strip at y= r in the graph. This shows that an increment d**r** from **r** will yield a decrease S(r)dr in $\phi(r)$. Hence

$$\phi'(r) = (d/dr)\, \phi(r) = -S(r).$$

Using the fact that S'(x) = -f(x), a second differentiation yields

$$\phi''(r) = f(r),$$

where f(r) is the density function of the entry ratio.[27] Similarly, we may deduce from Exhibit 3.8 that

$$\psi'(r) = (d/dr)\, \psi(r) = F(r)$$

and

$$\psi''(r) = f(r).$$

(5) Consider the area of the rectangle on the interval from 0 to r in Exhibit 3.8. This gives the relation

$$r = [1 - \phi(r)] + \psi(r)$$

or

$$\psi(r) = \phi(r) + r - 1; \qquad \text{(Formula 3.1)}$$

this is a fundamental relation connecting $\psi(r)$ and $\phi(r)$.
In general, consider a policy that has both a minimum ratable loss and a maximum ratable loss. Let L be the aggregate loss subject to a minimum of r_1E and a maximum of r_2E. So:

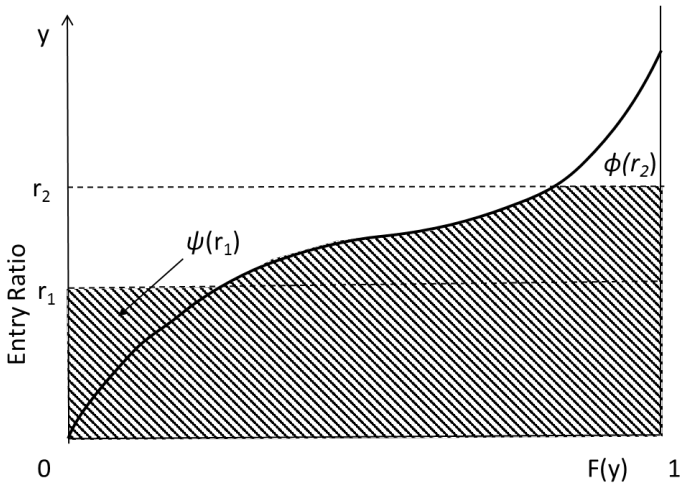
[27] Nels M. Valerius, "Risk Distributions Underlying Insurance Charges in the Retrospective Rating Plan," *PCAS* XXIX, 1942, p. 96.

$$L = \begin{cases} r_1E & \text{if } A <= r_1E \\ A & \text{if } r_1E <= A <= r_2E \\ r_2E & \text{if } r_2E <= A \end{cases}$$

If the actual loss is less than r_1E, L equals r_1E, the minimum loss. If the actual loss falls between r_1E and r_2E, L will be the actual loss. If the actual loss exceeds r_2E, the maximum loss, L will be r_2E.

Then a result more general than Formula 3.1 can also be obtained quite easily from examining Exhibit 3.9.

Exhibit 3.9. Expectation of Insured Loss (L) in Retrospective Rating



The shaded area in Figure 3.9 represents the quantity E{L}/E and we have

$$E\{L\}/E - \psi(r_1) + \phi(r_2) = 1,$$

or

$$E\{L\}/E = 1 + \psi(r_1) - \phi(r_2). \qquad \text{(Formula 3.2)}$$

See Skurnick.[28]

Lee diagrams can also be used to motivate the derivation of key formulas used in Retrospective Rating.
(Note that for ease of exposition, we ignore the tax factor in this chapter. Real premium calculations, would, of course, include a component for taxes.[29])
Recall from Chapter 2 that in a Retrospective Rating Plan, the retrospective premium R is given by

$$R = B + cA,$$

where B is the basic premium and c is the loss conversion factor (LCF), and where B is alternatively represented by

$$B = bP,$$

with P as the standard premium (before any applicable expense gradation) and b as the basic premium ratio.
For this section, we will assume the policy is subject to a maximum premium G and a minimum premium H.
Let L_G be actual loss that will produce the maximum premium:

$$G = B + cL_G$$

and let

[28] David Skurnick, "The California Table L," *PCAS* LXI, 1974, p. 117.
[29] The otherwise calculated retrospective premium would be multiplied by T, which is called the tax multiplier. Chapter 2 shows this more complete version of the formula.

$$r_G = L_G/E.$$

Similarly, define L_H to be

$$H = B + cL_H,$$
$$r_H = L_H/E.$$

Further, let

$$L = \begin{cases} L_H & \text{if } A <= L_H \\ A & \text{if } L_H <= A <= L_G \\ L_G & \text{if } L_G <= A \end{cases}$$

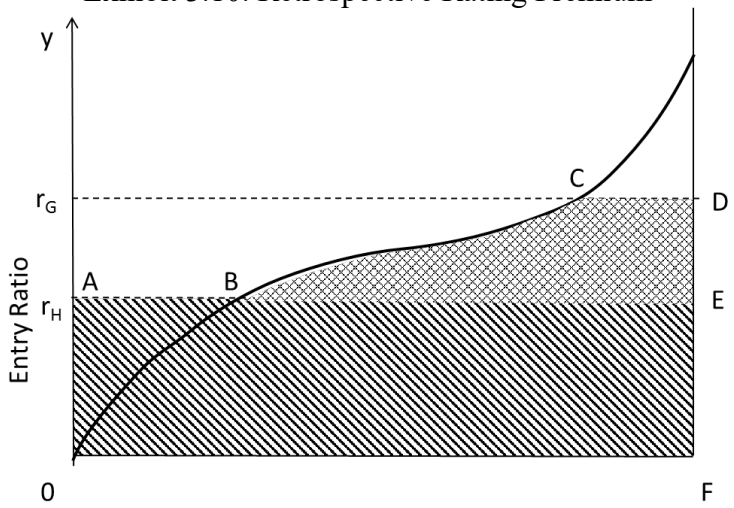
So if the actual loss is less than L_H, L equals L_H, the minimum ratable loss. If the actual loss falls between L_H and L_G, L will be the actual loss. If the actual loss exceeds L_G, the maximum ratable loss, L will be L_G.

Then the retrospective premium can be represented by

$$R = B + cL.$$

If we identify r_H and r_G with r_1 and r_2, respectively, then Exhibit 3.10 shows the quantity E{L}/E as the area of the shaded region OFDCBA.

Exhibit 3.10. Retrospective Rating Premium



It then follows that

$$E\{L\} = E - \phi(r_G)\, E + \psi(r_H)E$$

$$=E-I,$$

where

$$I = [\phi(r_G) - \psi(r_H)]E$$

is called the net insurance charge of Table M. If the plan is to cover the expected costs of the policy, the expected retrospective premium must be equal to the sum of the total expenses, e, and the expected loss, E:

$$E\{R\} = e + E.$$

On the other hand, it also follows from the above that

$$E\{R\} = B + c(E - I).$$

Equating these two quantities we obtain the basic premium in terms of the expense, expected loss, and the net insurance charge:

$$B + c(E - I) = e + E$$

or

$$B = e - (c - l)E + cI. \qquad \text{(Formula 3.3)}$$

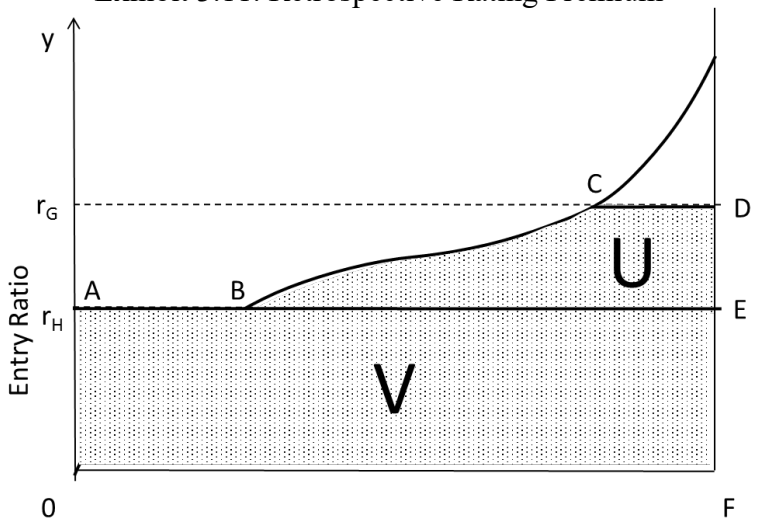
A formula relating the charge difference to the minimum premium, expected loss and expense provision has been used to facilitate the determination of retrospective rating values from specified maximum and minimum premiums. This formula can be derived with the help of Figure 3.11 below.

Consider the equation

$$R=B+cL$$

Taking the expectation of both sides, recalling that $E\{R\} = e + E$, and representing the expectation $E\{L\}/E$ by the shaded area of Exhibit 3.11 (areas U and V combined, equivalent to OFDCBA in Exhibit 3.10)

Exhibit 3.11. Retrospective Rating Premium



we have

$$e + E = B + cE[U+V].$$

On the other hand, we have for the minimum premium H:

$$H = B + cEr_H$$
$$= B + cE\,[V].$$

Taking the difference on both sides of the two equations above we have

$$(e+ E) - H = cE[U]$$
$$= cE\,[\phi(r_H) - \phi(r_G)].$$

Or

$$\phi(r_H) - \phi(r_G) = \frac{(e+E) - H}{cE}$$

(Formula 3.4)

We can also derive a formula relating the entry ratios themselves to the major plan parameters. The losses at the minimum premium are $r_H E$. So

$$H = cr_H E + B.$$

Similarly, the losses at the maximum premium are $r_G E$. So

$$G = cr_G E + B.$$

Subtracting these two equations you find

$$G - H = cE(r_G - r_H)$$

or

$$r_G - r_H = \frac{G - H}{cE}$$

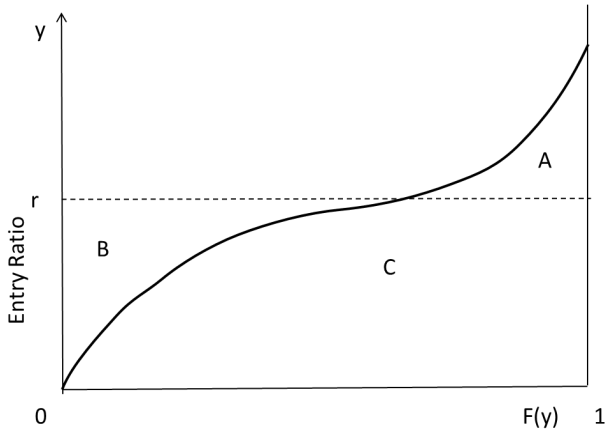
(Formula 3.5)

Formulas 3.4 and 3.5 can be used to determine the rating values given the maximum and minimum premiums. They are commonly referred to as the balance equations for aggregate losses.

One may interpret the difference in charge, $\phi(r_H) - \phi(r_G)$, as indicated by area U in Exhibit 3.11, to be the difference between the expected retrospective premium and the minimum premium, apart from conversion factor cE.

Questions

5. Label each of the three areas, A, B, and C in the Lee diagram below in terms of the Insurance Charge and the Insurance Savings.



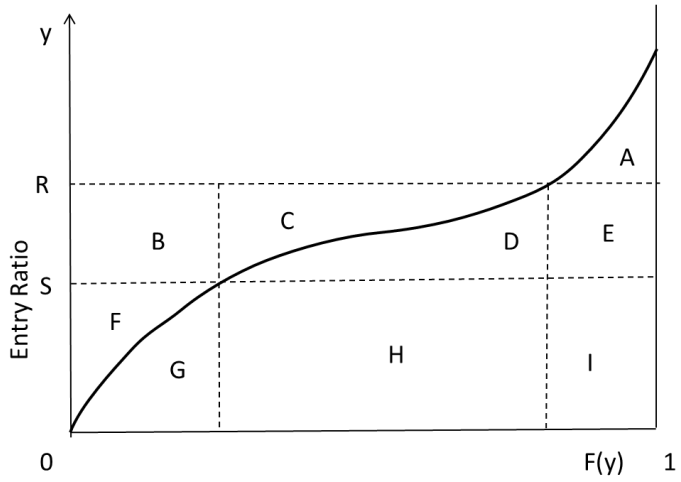
6. For the Lee diagram below, identify the areas associated with

(a) Insurance Charge at R

(b) Insurance Charge at S

(c) Insurance Savings at R

(d) Insurance Savings at S



3. Estimating Aggregate Loss Costs Using Table M

3.1. How Table M is Used

To summarize: Table M contains insurance charges and savings by entry ratio and by size of policy, possibly by limit, and other key considerations.[30] The entry ratio is defined as the aggregate losses divided by the expected aggregate losses (unlimited, or limited in the case of a limited Table M). Because different sized policies will have very different aggregate loss distributions, they must be grouped by approximate size, usually determined either by expected loss or expected number of claims[31] to estimate an appropriate insurance charge.

The mathematical basis of Table M is a distribution of entry ratios and its underlying distribution of aggregate losses.

$$Insurance\ Charge\ \ at\ r = X = \phi(r) = \int_r^\infty (y - r) f(y) dy$$

$$Insurance\ Savings\ at\ r = S = \psi(r) = \int_0^r (r - y) f(y) dy$$

$$S = X + r - 1\ or\ \psi(r) = \phi(r) + r - 1$$

For example, we might consider a workers' compensation insurance policy with expected aggregate losses of $200,000. It is a loss-sensitive policy, with a limit of $80,000 to the losses the insured is responsible for. That is, the maximum the insured will pay is the first $80,000 of aggregate losses, and the insurer will pay the rest. The entry ratio for the aggregate limit is 0.4 (=$80,000/$200,000). Assume that in this example the Table M for this sized insured has a corresponding insurance charge of 0.72 for an entry ratio of 0.4. Then the loss cost of the aggregate deductible policy is $144,000 (=$200,000*0.72), the expected losses to be owed by the insurer.

[30] Other key considerations include product being priced, types of losses covered, the jurisdiction where the policy is in force, etc. The size of the policy is highlighted because it has an enormous impact on the insurance charges and savings, as discussed later in this chapter.

[31] Grouping policies either by expected total loss (expected loss group) or expected number of claims (expected claim count group) serves the purpose of bucketing risks whose aggregate loss distributions have a similar variance component due to claim frequency. Grouping explicitly by expected number of claims has the advantage of getting at that aspect of the risk more directly, and is less subject to inflation. But the expected loss is a core element of any pricing exercise, and may be more readily available.

3.2. Empirical Construction of Table M[32]

While Table M can now be stored electronically, often as a function (or set of functions) of the plan parameters, rather than as a giant look-up table, the starting point for developing those functions are empirical methods. It is instructive to understand how this is done.

In order to construct Table M empirically, the first step is to obtain data on the annual aggregate losses for many insureds. The data needs to be split into groups of insureds which are expected to have similar distribution of aggregate losses. A separate analysis should be done for each of these groups.

Ideally, this means the insureds have a similar frequency-of-loss distribution and have similar patterns of claim severity. In practice, actuaries usually group insureds that are similar in size and are subject to similar risks. (E.g., workers' compensation risks engaged in moderately hazardous activities with between 35 and 40 expected claims.)

For each group, we need actual aggregate losses for the year, or the actual aggregate loss ratios, or some other measure that will allow us to compare a group's aggregate loss experience with that of the average risk of the group.[33] Typically, we use the average of the actual aggregate losses as the expected loss for the group. Note that if we use some other estimate of expected, the empirically calculated $\phi(0)$ will not equal 1.

The final published table of insurance charges is then organized by the groups examined (risk size and other key characteristics) as well as the entry ratio.

A representative sample of the data is shown in Exhibit 3.12, focusing on a single group.

[32] Section 3.2 is adapted from a study note by J. Eric Brosius, "Table M Construction," 2002, published by the Casualty Actuarial Society as part of the Syllabus of Exams.

[33] Developing those losses to ultimate without dampening the underlying variance is a complex problem which is beyond the scope of this study note, but which the practitioner should be aware of. A common solution is the use of a stochastic development procedure.

Exhibit 3.12.
Experience for a Group of Risks with Approximately 500 Expected Claims (N)

Risk	Actual Aggregate Loss
1	1,000,000
2	2,500,000
3	3,000,000
4	3,500,000
5	4,000,000
6	4,000,000
7	4,500,000
8	5,000,000
9	7,500,000
10	15,000,000

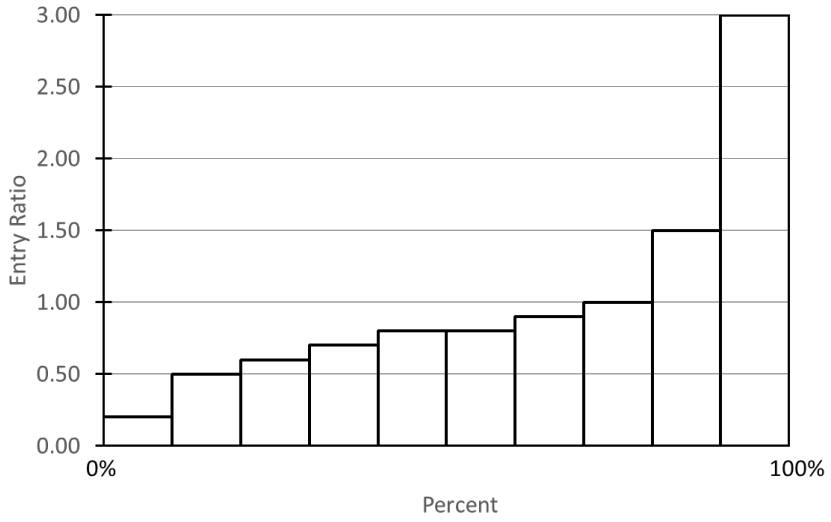
The second step is to compute entry ratios. As defined previously, the entry ratio is the ratio of the actual aggregate losses to the expected aggregate losses. In the example above, the empirical average aggregate loss per policy is $5M. So the entry ratios can be found as in Exhibit 3.13.

Exhibit 3.13.
Entry Ratios for a Group of Risks with N ~ 500,
and observed average aggregate loss of $5,000,000

Risk	Actual Aggregate Loss	Entry Ratio (r)
1	1,000,000	0.2
2	2,500,000	0.5
3	3,000,000	0.6
4	3,500,000	0.7
5	4,000,000	0.8
6	4,000,000	0.8
7	4,500,000	0.9
8	5,000,000	1.0
9	7,500,000	1.5
10	15,000,000	3.0

Then we will start to find the insurance charges in Table M. There are two methods for calculating the insurance charges: vertical slicing method and horizontal slicing method. The former is from the viewpoint of per risk, while the latter is from the viewpoint of per layer. It can be helpful to construct a Lee diagram of the data at this point.

Exhibit 3.14. Raw Data



(I) Vertical Slicing Method

We will explain the vertical slicing method first. In the example above, we will calculate the insurance charge of the entry ratio at 1.2, or $\phi(1.2)$.

Exhibit 3.15. Aggregate Excess at an Entry Ratio of 1.2

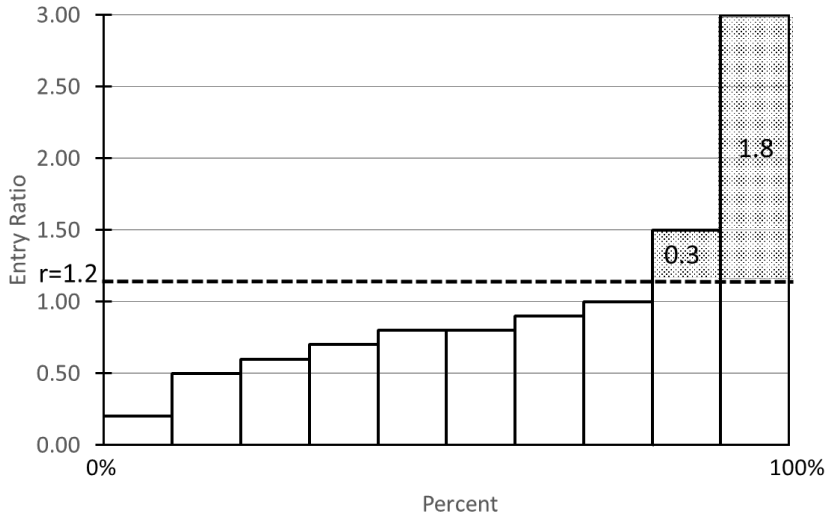


Exhibit 3.16. Calculation with vertical slices

Risk	Actual Aggregate Loss	Entry Ratio (r)	Excess of r=1.2
1	1,000,000	0.2	0
2	2,500,000	0.5	0
3	3,000,000	0.6	0
4	3,500,000	0.7	0
5	4,000,000	0.8	0

6	4,000,000	0.8	0
7	4,500,000	0.9	0
8	5,000,000	1	0
9	7,500,000	1.5	0.3
10	15,000,000	3	1.8

Then the average value of the excess column is the insurance charge of the entry ratio at 1.2. That is, $\phi(1.2)$=(0.3+1.8)/10=0.21. We can find the insurance charges for all the other entry ratios, using the same procedure. A Table M with an equal height of 0.1 can be constructed as below.

Exhibit 3.17. Table M: For N ~ 500

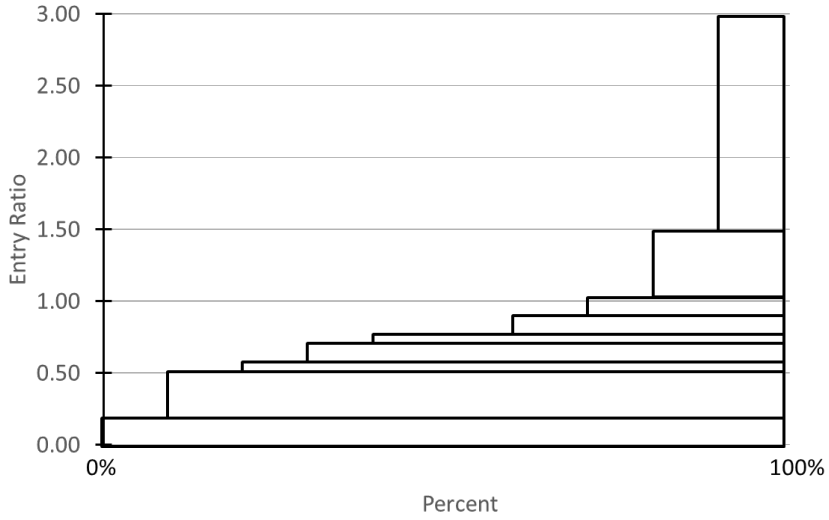
r	ϕ(r)
0	1
0.1	0.9
0.2	0.8
0.3	0.71
0.4	0.62
0.5	0.53
0.6	0.45
0.7	0.38
0.8	0.32
0.9	0.28
1.0	0.25
1.1	0.23
1.2	0.21
1.3	0.19
1.4	0.17
1.5	0.15
1.6	0.14
1.7	0.13
1.8	0.12
1.9	0.11
2.0	0.1
2.1	0.09
2.2	0.08
2.3	0.07
2.4	0.06
2.5	0.05
2.6	0.04
2.7	0.03
2.8	0.02
2.9	0.01
3.0	0

It can be seen that it takes lots of work to calculate the insurance charges for all the entry ratios if the vertical slicing method is used, although it is easy to understand and explain.

(II) Horizontal Slicing Method

Now, we will introduce the other method of constructing Table M, the horizontal slicing method. In comparison with the former method, the latter method is much easier to calculate insurance charges for multiple entry ratios although to some extent it is less intuitive and harder to explain. This is exactly comparable to slicing a distribution horizontally (instead of vertically) on a Lee diagram, as shown in Exhibit 3.18

Exhibit 3.18. Horizontal Slices



The procedure of the horizontal slicing method of calculating Table M charges is shown in Exhibit 3.19. Starting from the entry ratio (r) column, we find the number of risks in the group with the corresponding entry ratio as shown in the "# Risks" column. Then the "# Risks over r" shows the number of risks which have entry ratios exceeding a given entry ratio. The "% Risks over r" column converts the number in the "# Risks over r" into a percentage basis by dividing by the total number of risks (here it is 10). The "Difference in r" column shows the difference between the entry value in this row and the entry value in the next row. Finally, the last column in the table is the insurance charge. The last column begins from the bottom row which is 0 and then works up; the value in each row is equal to the value in the row beneath plus the product of the "% Risks over r" and the "Difference in r" in that row.

Exhibit 3.19. Calculation with horizontal slices

r	# Risks	# Risks over r	% Risks over r	Difference in r	ϕ(r)
0	0	10	100%	0.2	1
0.2	1	9	90%	0.3=0.5-0.2	0.8
0.5	1	8	80%	0.1=0.6-0.5	0.53
0.6	1	7	70%	0.1	0.45
0.7	1	6	60%	0.1	0.38
0.8	2	4	40%	0.1	0.32
0.9	1	3	30%	0.1	0.28
1.0	1	2	20%	0.5	0.25=0.15+20%*0.5
1.5	1	1	10%	1.5	0.15=0+10%*1.5
3.0	1	0	0%	0	0

Using the horizontal slicing method, we can construct a Table M with an equal height of 0.1, as shown previously in Exhibit 3.17. The results of the vertical and horizontal slicing methods are the same so long as we calculate horizontal slices at all the data points, because we use the same data. When a real Table M is constructed, the entry ratios are usually chosen so as to have intervals of 0.01 between rows.[34]

[34] If the entry ratios of the data points (the aggregate loss seen on a policy divided by the expected aggregate loss) falls between the "slices" chosen, we would not actually be adding the area of rectangles, and unless adjustments are made, the calculated charges will be slightly off. This is rarely a serious problem in real-life analyses, where many slices are used and there are enough observations that simple adjustments, such as adding the area of trapezoids instead of rectangles, and linearly interpolating between observed entry ratios, will yield adequate accuracy, but it can make a significant difference in the sort of simplified examples that might come up when studying this material. See question 7 for an example of this effect.

Finally, we can also calculate the insurance savings for each entry ratio by using the formula $\psi(r) = \phi(r) + r - 1$. If the observed data does not match all entry ratios we want in our table, we can interpolate. Exhibit 3.20 was developed by linearly interpolating the values in Exhibit 3.19. For example, the charge at 0.3 is gotten by linearly interpolating between the charges at 0.2 and 0.5 of 0.8 and 0.53 respectively:

$$(2/3)(0.8) + (1/3)(0.53) = 0.71.$$

Thus, the Table M with an equal height of 0.1 can be constructed as shown in Exhibit 3.20.

Exhibit 3.20. Table M: For N ~ 500

r	ϕ(r)	ψ(r)
0	1	0
0.1	0.9	0
0.2	0.8	0
0.3	0.71	0.01
0.4	0.62	0.02
0.5	0.53	0.03
0.6	0.45	0.05
0.7	0.38	0.08
0.8	0.32	0.12
0.9	0.28	0.18
1.0	0.25	0.25
1.1	0.23	0.33
1.2	0.21	0.41
1.3	0.19	0.49
1.4	0.17	0.57
1.5	0.15	0.65
1.6	0.14	0.74
1.7	0.13	0.83
1.8	0.12	0.92
1.9	0.11	1.01
2.0	0.1	1.1
2.1	0.09	1.19
2.2	0.08	1.28
2.3	0.07	1.37
2.4	0.06	1.46
2.5	0.05	1.55
2.6	0.04	1.64
2.7	0.03	1.73
2.8	0.02	1.82
2.9	0.01	1.91
3.0	0	2

3.3. Calculating Table M from a parameterized aggregate loss distribution

In many cases, the aggregate loss distribution can be modeled by parameterized functions which are amenable to manipulation. In a very simple example, one might assume that the number of claims can be modeled by a Poisson distribution and the severity of resulting claims can be modeled by a Pareto distribution. Or the aggregate loss distribution might be directly approximated using a lognormal distribution.

This might be done for a reinsurance contract, where there is not a statistically credible body of similar policies that can be used to construct an empirical aggregate loss density function. The actuary might, however, have evidence that similar policies tend to have loss frequency and severity distributions of certain general types, and might have nothing better on which to base their prices than the results of fitting the available claim data to those types of distributions. When data is thin, pricing actuaries should be careful to test the sensitivity of their loss cost estimates to a variety of assumptions. Even with an abundance of data, parameterized functions make it easy to develop a large number of consistent insurance charges.

Once the underlying frequency and severity distributions have been selected, the aggregate loss distribution can be simulated or, in many cases, calculated using a variety of closed-form methods. In either case, the resulting aggregate loss distribution can be used to generate Table M charges according to the horizontal slicing method described above. Chapter 4 of the CAS monograph "Distributions for Actuaries" by David Bahnemann discusses these methods, and those calculations are beyond the scope of this study note.[35]

In practice, a hybrid of empirical data and models is often used. For example, in order to accumulate enough data to have reasonably credible groups, we might not be able to split the data into buckets small enough to provide accurate charges across the whole range of the data. So the data might be split into a modest number of large groups, whose aggregate loss distribution can be fitted to parameterized distributions. In doing this, it is best to look at each policy's aggregate loss as a ratio to its expected loss, so as not to introduce extra variation due to the different expectations of loss. Once an empirical distribution is found, parameterized curves can be fit to it. Then the parameters can be interpolated to generate aggregate loss distributions for smaller, more homogeneous groups. The details of these procedures are beyond the scope of this study note, but the actuary should be aware that issues of loss development,[36] trend, and the heterogeneous nature of the underlying exposures all need to be considered.

Questions

7. Eight identical risks incur the following actual aggregate loss ratios, respectively:

20% 40% 40% 60% 80% 80% 120% 200%

Assume that the expected loss ratio for those risks is the observed average loss ratio.

(a) Construct a Table M showing the insurance charge for entry ratios from 0 to 3.0 in increments of 0.5.

---

[35] Two methods that have been used to create aggregate distributions from underlying frequency and severity distributions are the recursive method, described by Harry Panjer, "Recursive Evaluation of a Family of Compound Distributions," *Astin Bulletin*, Vol. 12, No. 1, 1981, pp. 22-26 and the Heckman-Meyers method, described by Philip E. Heckman and Glenn G. Meyers in "The Calculation of Aggregate Loss Distributions from Claim Severity and Claim Count Distributions," *PCAS* LXX, 1983.

See also D. Bahnemann, "Distributions for Actuaries," CAS Monograph # 2, Chapter 4.

[36] Some discussion of these topics can be found in H. C. Mahler, Discussion of "Retrospective Rating: 1997 Excess Loss Factors," *PCAS* LXXXV, 1998, pp. 316-344.

(b) Calculate the Insurance Charge at a 70% loss ratio.
(c) Calculate the Insurance Savings at a 70% loss ratio.
(d) Calculate the Insurance Charge at a 110% loss ratio.
(e) Calculate the Insurance Savings at a 110% loss ratio.

8. What are some advantages and disadvantages of using parameterized distributions to develop Tables M?

4. Estimating Limited Aggregate Excess Loss Costs[37]

4.1. Introduction of Limited Aggregate Deductible Policies

The original Table M was developed for retrospectively rated workers' compensation policies, and for historical reasons it was originally calculated based on aggregate losses with no per-claim limit.

But we often want to price aggregate insurance charges on limited losses. For example, a large dollar deductible workers' compensation policy might have a per-occurrence limit to ratable losses (or deductible) of $100,000 for each loss occurring to it. When the amount of a claim[38] is less than $100,000, the insured is responsible for the amount of the claim. If the amount of a claim exceeds the per-occurrence limit of $100,000, the insured will only be responsible for the first $100,000 of the loss.

However, if the insured also wanted to limit its total liability, it may also have negotiated an aggregate deductible limit of $250,000, so the insured would never have to pay more than $250,000 in losses occurring on this policy, regardless of actual experience. The policy could reach that limit if there are more than two claims larger than $100,000, or if there are lots of small claims, or some combination of the two. In this situation, the actuary needs to calculate the limited aggregate excess charge.

In pricing the loss portion of a deductible policy with an aggregate deductible limit, or a retrospectively rated policy with a per-occurrence limit, the actuary can either price for the excess losses and the aggregate deductible losses simultaneously (as is done in the California Table L, discussed in section 5) or can charge separately for losses in excess of the deductible and for the deductible losses in excess of the aggregate limit. We will first consider calculating the two charges separately—directly calculating a Table M appropriate to aggregate limited losses, which produces charges suitable to add to per-occurrence excess loss charges. We will then consider other methods of pricing such polices in section 5.

The following text will use the notation "limited Table M," or "Table M_D" where D is the limit or deductible amount to refer to a table of charges for the aggregate of limited losses.

4.2. Considerations for Table M_D

Often it is expedient to calculate the charges for the per-occurrence excess and the aggregate excess separately. The actuary might have enough data to update the estimate for the per-occurrence excess more frequently than the aggregate excess charge, or might have reasons to rely on different data sources for the two calculations. Both ISO and the NCCI take this approach for policies with a per-loss limitation on their retro plans, including an excess loss premium factor in addition to the Table M charge.

Throughout this section, we assume that the per-occurrence excess charge is known, and has been calculated based on losses not subject to an aggregate limit. So as not to double-count losses that might be subject to either the per-occurrence or the aggregate limit, the limited aggregate excess loss charge must be developed or estimated based on the distribution of limited losses, that is, losses to which the per-occurrence limit has already been applied.

---

[37] Section 4 is adapted from a study note by Ginda Kaplan Fisher, "Pricing Aggregates on Deductible Policies," 2002, published by the Casualty Actuarial Society as part of the Syllabus of Exams.

[38] This chapter assumes that a single insured occurrence will generate at most one claim, which would be subject to the per-occurrence limit. In real insured events, an occurrence can generate multiple claims which might apply to one or more insurance policies and interact with the limits of those policies in complicated ways. However, those details are beyond the scope of this study note.

For example, consider a policy which has a per-occurrence limit of $100,000 and an aggregate deductible limit of $250,000. Four claims occur:
$50,000
$50,000
$50,000
$300,000
After the first three claims, the insured is responsible for paying $150,000. Then the $300,000 claim occurs. The insured is only responsible for the first $100,000 of loss on that claim. But is the other $200,000 excess of the aggregate limit, or of the per-occurrence limit? This is an example of the overlap of the per-occurrence limit (deductible) and aggregate limit. It is customary to apply the per-occurrence limit first, so those $200,000 are considered excess of the per-occurrence limit, and should be contemplated in the per-occurrence excess charge.
If the actuary calculated Table M charges without limiting the aggregate losses for the effect of the $100,000 deductible, those $200,000 would increase the Table M charge, and there would be an overlap between the Table M charge and the per-occurrence excess charge, leading to inappropriate Table M charges. Simply limiting the aggregate losses for the effect of the $100,000 deductible before using any of the methods above to estimate Table M removes this problem.[39]
Actuaries can determine limited aggregate excess charges through the same methods used for any other aggregate excess loss: they can gather a large body of policy data which is expected to be similar to that for the policies being priced and build an empirical table or they can use information about the expected distribution of losses to model the charges.
The shape of the distribution of limited (or primary) losses is different from the shape of the distribution of the same losses when not subject to a limit[40], as the severity distribution can be quite different—nevertheless, it is just another loss distribution. In particular, all the same relationships used in constructing Table M charges apply to calculating limited loss insurance charges, as described above.
Because the size of the deductible has an impact on the shape of the aggregate loss distribution, a separate table M_D must be calculated for a wide range of deductibles, spanning the range of deductibles offered. Fortunately, this does not require masses of data at every deductible or loss limit. If the unlimited losses are known (as is often the case with both retrospectively rated and large deductible plans) the same losses can be used to calculate Table M charges at any limit simply by limiting each loss before adding it to the aggregate used. The actuary should be careful to limit individual occurrences prior to aggregating the losses of each policy.
In general, the lower the deductible (or the smaller the per-occurrence limit), the less variance there is in the severity distribution and thus the less variance there is in the resulting limited aggregate loss. This is because loss distributions tend to be positively skewed, with many small losses and few large losses. Therefore much of the variance of the severity distribution is driven by the extreme (high) losses, and after the application of the per-occurrence limit, the variance of severity is reduced. (Limiting the losses does not change the frequency distribution.) The reduction in variance of limited aggregate losses reduces the probability of unusually large

[39] Note that adjusting the excess charges to remove aggregate losses would be a much more complicated process, and would mean that excess charges would depend on the size of the policy, and not just the severity distribution of the losses.
[40] Or when subject to a much higher policy limit.

limited aggregate losses in a given year. Therefore, lower deductibles usually lead to lower insurance charges for entry ratios greater than 1.

4.3. Construction of Table M_D

When working with a limited Table M, it is important to remember to use limited losses consistently. The expected losses used in calculating the entry ratio must be the expected deductible (or limited) losses, and not the total expected ground-up losses on the policy.

For example, an insured has a per-occurrence deductible of $250,000 and its expected limited aggregate losses are $800,000. The aggregate deductible limit is $1,000,000. First, we need to compute the appropriate entry ratio, r:

$$1.25 = r = \$1{,}000{,}000/\$800{,}000$$

Assume that the insurance charge for an entry ratio of 1.25 in Table M_D with a per-occurrence limit of $250,000 is 0.18. Then the loss cost of the aggregate deductible limit is $144,000 (=$800,000*0.18).

The methods of constructing a Limited Table M or Table M_D are the same as those of constructing a standard Table M, except that the data of aggregate losses are required to be the limited aggregate losses rather than the unlimited aggregate losses. Therefore, in Table M_D the entry ratio (r) is defined as the actual limited aggregate losses divided by the expected limited aggregate losses.

For example, an insured risk has a per-occurrence limit (deductible) of 100,000 and it has five claims in a year. The five claims are shown in Exhibit 3.21.

Exhibit 3.21. Experience for a Group of Risks with a Per-Occurrence Limit of $100,000

Claim No.	Unlimited Amount	Limited Amount
1	60,000	60,000
2	70,000	70,000
3	90,000	90,000
4	110,000	100,000
5	120,000	100,000
Total	450,000	420,000

In this case, the unlimited aggregate losses of $450,000, the sum of all the five claims, can be used to construct a standard Table M. In order to construct a Limited Table M, however, we should use the sum of the limited aggregate losses, $420,000.

All the same methods that are used to construct unlimited Tables M (vertical or horizontal slicing of empirical data, manipulating parameterized loss distributions) can be used to construct a Table M_D. A Limited Table M has three dimensions: the expected size of the policy, the entry ratio, and the per-occurrence loss limit, D. Alternately, one can think of Table M_D as a set of tables, one for each deductible or per-occurrence loss limit.

Examples of using Tables M_D to price the insurance charge of an insurance policy with a deductible and an aggregate:

Expected total losses = $700,000
Deductible = $250,000
Expected Primary Losses[41] = $500,000
Expected number of claims = 60
Entry Ratio = 2.0 (which means the aggregate limit is 2.0 x $500,000 = $1,000,000)
Table M_D for policies around $500K in size is shown in Exhibit 3.22.[42]

Exhibit 3.22. Sample Table M_D

Insurance Charge Factor	Deductible		
Entry Ratio	100K	250K	500K
1.0	.240	.250	.260
1.5	.100	.110	.120
2.0	.030	.040	.050
2.5	.018	.022	.030

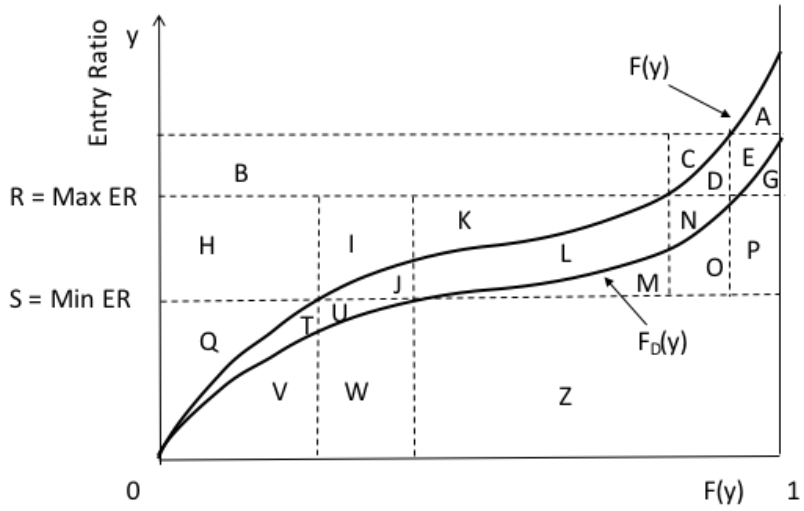
The factor at 250K for an entry ratio of 2.0 is 0.040, for an insurance charge of 0.040 x $500,000 = $20,000.
The total expected loss cost for this policy would be $220,000 ($20,000 plus the difference between $700,000 and $500,000).
Now consider the situation if the policy had been written with a deductible of $150,000.

Expected total losses = $700,000
Deductible = $150,000
Expected Primary Losses = $400,000
Entry Ratio = 2.5 (which means the aggregate limit is 2.5 x $400,000 = $1,000,000)
Note that the Expected Primary Losses are less because the deductible is now $150,000 rather than $250,000.
Using the table in Exhibit 3.22 again,
Interpolating[43] between the factor for an entry ratio of 2.5 at 100K (0.018) and at 250K (0.022) gives an insurance charge factor of .019, for an insurance charge of 0.019 x $400,000 = $7,733.
The total expected loss cost for this policy would be $307,733 ($7,733 plus the difference between $700,000 and $400,000).

Questions

9. For the Lee diagram below, identify the areas associated with
(a) Table M_D Charge at R
(b) Table M_D Savings at S
(c) Per-occurrence excess charge at D

[41] $E\{A_{250,000}\}$
[42] A real Table M_D would have many more entry ratios than this simplified example.
[43] Because the differences are small, any reasonable interpolation will do. I have used a linear interpolation for simplicity.



For questions 10 and 11,[44] please refer to chapter 2 for the retrospective premium formula, including the tax multiplier.

10. Let us assume a retrospectively rated insured had a basic premium of $30,000, an excess loss premium of $10,000, a loss conversion factor of 1.1, a tax multiplier of 1.05, an accident limit of $100,000, and a maximum premium of $250,000. Refer to chapter 2 for the retrospective premium formula including the tax factor.

a: If the insured has small losses totaling to $150,000 in year, what is the retro premium?
b: If the insured has small losses totaling to $200,000 in year, what is the retro premium?
c: If the insured has one large loss of $150,000 in year, what is the retro premium?
d: If the insured has one large loss of $150,000 in year plus $100,000 in small losses, what is the retro premium?

11. Let us assume a retrospectively rated insured had a basic premium of $300,000, an excess loss premium of $100,000, a loss conversion factor of 1.1, a tax multiplier of 1.05, an accident limit of $100,000, and a minimum premium of $650,000. Refer to chapter 2 for the retrospective rating formula, including the tax factor.
(a) If the insured has small losses totaling to $150,000 in year, what is the retro premium?
(b) If the insured has one large loss of $150,000 in year, what is the retro premium?

12. You price a retrospective policy with an expected loss of $150,000 and aggregate limit of $300,000, and find that the insurance charge is $15,000.
The customer requests that you also add a per-occurrence loss limitation of $100,000 to the losses subject to the retrospective calculation. You determine that if there were no aggregate limit, the cost of the per-occurrence limit would be $50,000
Would the combined charge for the per-occurrence limit and the aggregate limit be more, less, or the same as the sum of the two charges, $65,000? Why?

13. You are given the following table of insurance charges, by per-occurrence deductible:

[44] Questions 10 and 11 were adapted with permission from material written by Howard Mahler.

r	$100,000 deductible	$200,000 deductible
1.0	0.20	0.22
1.5	0.10	0.12
2.0	0.04	0.05
2.5	0.02	0.03

The expected unlimited losses are $40,000.

The expected primary losses at a per-occurrence limit of $100,000 are $20,000.

The expected primary losses at a per-occurrence limit of $200,000 are $30,000.

(a) A policy has a $100,000 per-occurrence deductible and a $40,000 aggregate deductible limit. Find the cost of the $40,000 aggregate deductible limit.
(b) Find the cost of the $40,000 aggregate deductible limit if the policy had a $200,000 per-occurrence deductible. (Use linear interpolation in the table, if necessary.)
(c) Which policy will the insurer charge more for? Why?

5. Other Methods of Combining Per-Occurrence and Aggregate Excess Loss Cost[45]

5.1. Estimating Per-Occurrence and Aggregate Combined Excess Loss Cost Using Table L

Consider the case of an insured with a per-occurrence limit of $50,000 for each loss occurring to it, and an aggregate limit of $250,000. For example, if the policy had the following claims:

$20,000

$30,000

$45,000

$55,000

$100,000

$120,000

The insured would be responsible for the first $50,000 of each claim, or

$$\$20K + \$30K + \$45K + 3 \times (\$50K) = \$245K\ .$$

If one more claim of $10,000 were incurred, the insured would only be responsible for an additional $5,000, because the aggregate limit on the limited loss would have been reached.

a. Table L and its Implication

Table L is a method to estimate a per-occurrence and aggregate combined excess policy simultaneously, in a single table. Like Table M_D, that table has three dimensions: the expected size of the policy, the entry ratio, and the per-occurrence loss limit. Alternately, one can think of Table L as set of tables, with one per each per-occurrence limit.

It is defined as follows: Assume that a formula for limiting or adjusting individual occurrences is given. The entry ratio (r) at any actual loss incurred by the risk is defined as the actual limited

[45] Section 5 is adapted from David Skurnick, "The California Table L," *PCAS* LXI, 1974; Yoong-Sin Lee, "The Mathematics of Excess of Loss Coverages and Retrospective Rating—A Graphical Approach," *PCAS* LXXV, 1988; and a study note by Ginda Kaplan Fisher, "Pricing Aggregates on Deductible Policies", 2002, published by the Casualty Actuarial Society as part of the Syllabus of Exams.

aggregate losses divided by the expected unlimited aggregate losses. The Table L charge at entry ratio r, $\phi_D^*(r)$, is defined as the average difference between a risk's actual unlimited loss and its actual loss limited to D, plus the risk's limited loss in excess of r times the risk's expected unlimited loss. The Table L savings at entry ratio r, $\psi_D^*(r)$, is defined as the average amount by which the risk's actual limited loss falls short of r times the expected unlimited loss. The Table L charge and savings are both expressed as ratios to expected unlimited loss[46].

This differs from Table M in that Table L looks at how much loss, on average, will be limited by the combination of the per-occurrence limit and the aggregate excess limit.

Recall that $F_D(Y)$ = the cumulative distribution function of Y, the limited losses whose unlimited cumulative distribution function was given by F. Then the Table L insurance charge at entry ratio $r \geq 0$ is defined as a formula:

$$\phi_D^*(r) = \int_r^\infty (y - r)\, dF_D(y) + k \qquad \text{(Formula 3.6)}$$

and

$$\psi_D^*(r) = \int_0^r (r - y) dF_D(y) \qquad \text{(Formula 3.7)}$$

where k is the excess ratio for the per-occurrence limit. That is,

$$k = \frac{E - E\{A_D\}}{E}, \qquad \text{(Formula 3.8)}$$

where E is the total expected loss, and $E\{A_D\}$ is the expected loss after application of the per-occurrence limit.

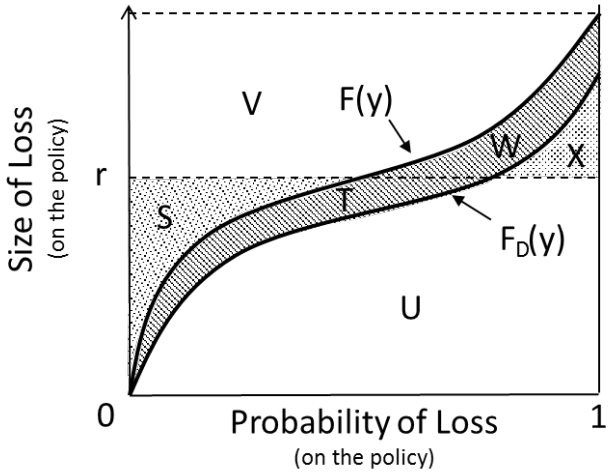
Note that if there is no loss limit, "k" will be zero, $F_D(y) = F(y)$, and the above formulas reduce to the Table M formulas.

Both the per-occurrence limit and the aggregate limit remove losses from the portion the insured owes (whether the ratable losses in a retro policy or the deductible losses in a deductible policy). If estimated separately, without considering both, the effects of the per-occurrence and aggregate limits overlap, as discussed in section 4. It is important not to double-count any losses excluded by these provisions. The formula for the Table L charge avoids this by using the limited distribution, F_D, in the integral.

We can see that the second part, k, of the $\phi_D^*(r)$ equation stands for the loss cost of the per-occurrence excess portion and the first part is the additional effect of the aggregate limit beyond that of the occurrence limit.

[46] Or policy limit loss, as mentioned in section 1.1, page 40.

Exhibit 3.23. Lee Diagram of Table L charge and savings



In Exhibit 3.23, the upper curve is F, the lower curve is F_D, and r corresponds to aggregate limit.
As with Table M, r = (aggregate limit) / (expected unlimited losses). The area under the upper curve (T+U+W+X) represents the unlimited loss distribution. It has area = 1, since all entities are defined in terms of the expected unlimited loss.
The area between the curves (T+W) represents the distribution of loss above the per-occurrence limit, and together they have area k.
The area under the lower curve (U+X) represents the distribution subject to the per-occurrence limit, or 1-k.
Area X represents the distribution of loss after application of the per-occurrence limit that is above the aggregate limit.
The Table M charge at entry ratio r (ignoring the per-occurrence limitation) is W+X.
The Table L charge at entry ratio r, $\phi_D^*(r)$, is T+W+X.
The Table M savings at entry ratio r (ignoring the per-occurrence limitation) is S.
The Table L savings at entry ratio r, $\psi_D^*(r)$, is S+T.
Also, r = S + T + U and 1 = the area under F(y) = T + U + W + X.
So $\phi_D^*(r) + r - 1 = (T + W + X) + (S + T + U) - (T + U + W + X) = T + S = \psi_D^*(r)$.
And (reading the above from right to left) the relationship between the insurance charge and the insurance saving in Table L is similar to that for Table M:

$$\psi_D^*(r) = \phi_D^*(r) + r - 1 . \qquad \text{(Formula 3.9)}$$

b. Construction of Table L

Here we will show an illustration of a Table L construction from empirical data.
To construct Table L, we need to obtain data on both the unlimited aggregate losses and the limited aggregate losses for each of the risks.

Exhibit 3.24. Experience for a Group of Risks with a Per-Occurrence Limit of $50,000

Risk	Actual Unlimited Aggregate Loss	Actual Limited Aggregate Loss
1	20,000	20,000
2	50,000	50,000
3	60,000	60,000
4	70,000	70,000

5	80,000	80,000
6	80,000	80,000
7	90,000	90,000
8	100,000	100,000
9	150,000	120,000
10	300,000	250,000
Average	100,000	92,000

First, we compute the excess ratio (k) for the per-occurrence limit:

$$k = 0.08 = (\$100,000-\$92,000) / \$100,000.$$

Next we compute the entry ratios. As stated previously, the entry ratio is the ratio of the actual limited aggregate losses to the expected unlimited aggregate losses. In this illustration, the expected unlimited aggregate losses of the group are $100,000. So the entry ratios, shown in Exhibit 3.25, are:

Exhibit 3.25. Entry Ratios for a Group of Risks with a Per Occurrence Limit of $50,000

Risk	Actual Unlimited Aggregate Loss	Actual Limited Aggregate Loss	Entry Ratio (r)
1	20,000	20,000	0.2
2	50,000	50,000	0.5
3	60,000	60,000	0.6
4	70,000	70,000	0.7
5	80,000	80,000	0.8
6	80,000	80,000	0.8
7	90,000	90,000	0.9
8	100,000	100,000	1
9	150,000	120,000	1.2
10	300,000	250,000	2.5

Then construct the Table L using the horizontal slicing method. The procedure is similar to the one we used to construct a Table M, and is shown in Exhibit 3.26. Note that the average r is 0.92 = 1-k.

Exhibit 3.26. Calculation of Table L

r	# Risks	# Risks over r	% Risks over r	Difference in r	ϕ_D^*(r)-k	ϕ_D^*(r)
0	0	10	100%	0.2	0.92	1
0.2	1	9	90%	0.3	0.72	0.8
0.5	1	8	80%	0.1	0.45	0.53
0.6	1	7	70%	0.1	0.37	0.45
0.7	1	6	60%	0.1	0.3	0.38
0.8	2	4	40%	0.1	0.24	0.32
0.9	1	3	30%	0.1	0.2	0.28
1	1	2	20%	0.2	0.17=0.13+20%*0.2	0.25
1.2	1	1	10%	1.3	0.13=0+10%*1.3	0.21
2.5	1	0	0%	0	0	0.08

Starting from the entry ratio (r) column, we find the number of risks in the group with the corresponding entry ratio as shown in the "# Risks" column. Then the "# Risks over r" shows the number of risks which have entry ratios exceeding a given entry ratio. The "% Risks over r" column converts the number in the "# Risks over r" into a percentage basis by dividing by the total number of risks (here it is 10). The "Difference in r" column shows the difference between the entry value in this row and the entry value in the next row.

Then, the "ϕ_D^*(r)-k" column is calculated similarly to ϕ(r) for an unlimited Table M. We begin from the bottom row with 0, as there are no expected losses greater than the largest entry ratio. Then we work up; the value in each row is equal to the value in the row beneath plus the product of the "% Risks over r" and the "Difference in r" in that row. Finally, the last column "ϕ_D^*(r)" is the "ϕ_D^*(r)-k" column plus the excess ratio (k) for the per-occurrence limit. Recall that k was calculated as 0.08 in the first step.

Note that if there is no loss limit, "k" will be zero, and this formula is the same as the formula for the Table M charge.
Finally, we can also calculate the insurance savings for each entry ratio by using the formula $\psi_D^*(r) = \phi_D^*(\mathrm{r}) + r - 1$. Thus, the Table L can be constructed as shown in Exhibit 3.27.

Exhibit 3.27. Table L

r	$\phi_D^*(r)$	$\psi_D^*(r)$
0	1	0
0.2	0.8	0
0.5	0.53	0.03
0.6	0.45	0.05
0.7	0.38	0.08
0.8	0.32	0.12
0.9	0.28	0.18
1	0.25	0.25
1.2	0.21	0.41
2.5	0.08	1.58

Note that as with calculating Table M_D, we do not need to have a large body of data at every per-occurrence loss limit in order to calculate Table L from empirical data. If the unlimited data is known at the claim level, we can create "as if" data at any per-occurrence loss limit. (When working with losses from coverages that might have varying policy limits, such as commercial auto insurance, it might be necessary to estimate "unlimited" claims/occurrences above lower per-occurrence policy limits if data from many policy limits is to be combined.)
As with Tables M, Tables L can also be calculated from simulated data (or other methods) if we have a parameterized loss distribution. But to calculate the Table L charge from simulated data, we need to separately simulate the number of claims and the severity of each claim, so that the per-claim loss limit can be appropriately applied. More detail is needed than the (unlimited) aggregate distribution, even if the excess ratio k is known.

5.2. The ICRLL Method

In some cases, it might be expedient to use an existing table of insurance charges, and apply reasonable modifications to it so it reflects the impact of limiting the losses. When tables were large printed documents, this was a very appealing option, even when there was adequate data or a robust enough model to explicitly calculate aggregate loss charges for a variety of deductible limits. Now, electronic or formulaic Tables M are generally available for both limited and unlimited policy losses, in which case, this sort of adjustment is not needed. However, until 2019, the NCCI published only unlimited Table M, and used an adjustment of this type in its workers' compensation rating manual: the Insurance Charge Reflecting Loss Limitation (ICRLL) procedure.[47] This procedure uses an unlimited Table M, and adjusts it to approximate Table M_D. The ICRLL method is presented here as an example of the sort of estimate an actuary can make when perfect data isn't available. Note that the 1998 NCCI Table M was published by Expected Loss Group (ELG) rather than by expected number of claims, and a State/Hazard Group adjustment was used to account for the different severities (and thus a different implied expected number of claims) within an ELG depending on the state and hazard group of the risk.

As mentioned above, Table M_D must be indexed by three variables: the expected size of the policy[48], the deductible, and the entry ratio. In effect, the ICRLL procedure can be used to map the three indices of M_D into the two used by the (unlimited) Table M, and can be thought of as a mapping of Table M_D onto Table M. Both the entry ratio and the size category (ELG) are modified to account for the deductible.

The Loss Group Adjustment Factor used in the ICRLL procedure is

$$\frac{1+0.8*k}{1-k} \qquad \text{(Formula 3.10)}$$

where k is the fraction of losses expected to be above the per-claim limit or deductible amount.[49]

For example, a workers' compensation insured has a per-occurrence limit of $250,000 and its expected limited aggregate losses are $490,000. In addition, its expected unlimited aggregate losses are $650,000. An aggregate deductible policy covers the insured and the aggregate deductible limit is $750,000.

We also know that this risk has a State/Hazard Group relativity of 0.9.

First, we compute the entry ratio: 1.53 = $750,000/$490,000.

Then we compute the ICRLL adjustment

$$\frac{1 + 0.8 * (\$650{,}000 - \$490{,}000)/\$650{,}000}{1 - (\$650{,}000 - \$490{,}000)/\$650{,}000} = 1.588$$

In an unlimited Table M, as excerpted below, the expected unlimited aggregate losses of $650,000 would correspond to Expected Loss Group 31. But in this case, we adjust the expected loss by the SHG and ICRLL adjustment to yield

$$\$650{,}000 \times 0.9 \times 1.588 = \$929{,}000.$$

So we will use an Expected Loss Group 29 to enter Table M.

---

[47] The ICRLL procedure was originally described by Ira Robbin in "Overlap Revisited—The 'Insurance Charge Reflecting Loss Limitation' Procedure," *Pricing*, Casualty Actuarial Society *Discussion Paper Program*, 1990, Volume 2.

[48] ICRLL used expected limited loss for this, consistent with grouping policies by expected unlimited loss for the unlimited Table M

[49] This excess ratio, k, was referred to as "ER" in the pre-2019 NCCI retrospective rating manual.

Exhibit 3.28. Table of Expected Loss Group[50]

Expected Loss Group	Range of Values
31	630,000-720,000
30	720,001-830,000
29	830,001-990,000
28	990,001-1,180,000
27	1,180,001-1,415,000
26	1,415,001-1,744,000

Looking this up in the excerpt of Table M below gives us a Table M charge of 0.1583, which indicates a dollar charge of 0.1583 x $490,000 or $77,567. This is the additional charge for the aggregate limit. The charge for the per-occurrence limit is $650,000 - $490,000 = $160,000. So the total expected loss cost for this policy is $160,000 + $77,567 = $237,567.

Exhibit 3.29. Table of Insurance Charges

	Expected Loss Group					
Entry Ratio	31	30	29	28	27	26
0.75	0.4150	0.4069	0.3989	0.3911	0.3833	0.3755
0.81	0.3864	0.3777	0.369	0.3605	0.3521	0.3436
1.07	0.2867	0.2764	0.2661	0.2557	0.2453	0.2349
1.15	0.2628	0.2522	0.2417	0.231	0.2203	0.2096
1.53	0.1797	0.169	**0.1583**	0.1476	0.1369	0.1261

Questions

14. Draw a Lee diagram illustrating a policy that has:

- A continuous uniform unlimited loss distribution from 0 to 500
- A continuous uniform limited loss distribution from 0 to 400
- An entry ratio of 1.5 times the expected unlimited loss

a) Label:
$\phi_D^*(1.5)$, the Table L charge at the entry ratio
$\psi_D^*(1.5)$, the Table L savings at the entry ratio

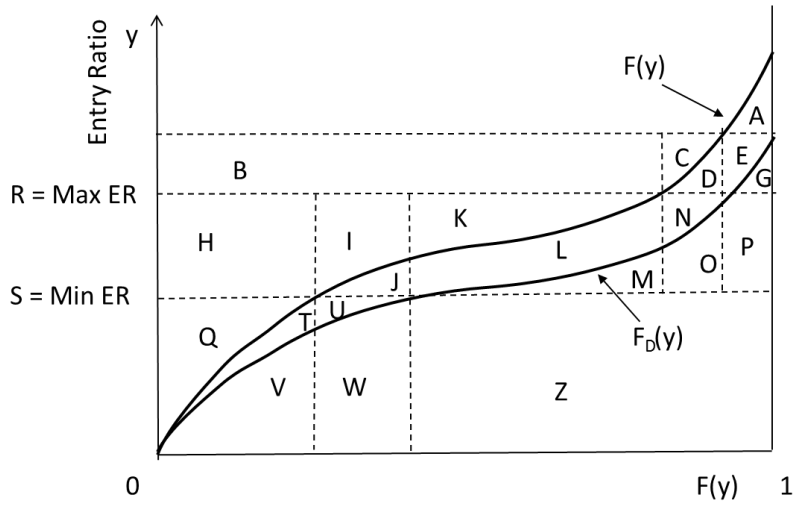
b) Calculate the value of
$\phi_D^*(1.5)$, the Table L charge at the entry ratio
$\psi_D^*(1.5)$, the Table L savings at the entry ratio

15. For the Lee diagram below, identify the areas associated with

[50] The Table of Expected Loss Groups changed over time, with inflation. This example is just illustrative.

(a) Table L Charge at R

(b) Table L Savings at S



16. What are some advantages to using ICRLL as compared to a limited Table M?

What are some disadvantages?

17.[51] A large dollar deductible workers' compensation policy requires the insured to reimburse the insurer for each occurrence up to $250,000, subject to an aggregate reimbursement of $1,200,000. The following attributes also apply to this policy:
Standard Premium: $1,000,000
Hazard Group Relativity: 0.900
Expected Unlimited Loss Ratio: 75%
K (Excess Ratio): 20%
Table M_D: Limited Insurance Charges with D = 250,000

Entry Ratio	1.0	1.5	2.0	2.5
Insurance Charge	0.250	0.110	0.040	0.022

Table of Expected Loss Ranges

Expected Loss Group	Expected Losses
31	630,000 − 720,000
30	720,001 − 830,000
29	830,001 − 990,000
28	990,001 − 1,180,000
27	1,180,001 − 1,415,000

Table M: Unlimited Insurance Charges

Entry Ratio	Expected Loss Group				
	31	30	29	28	27
0.5	0.415	0.407	0.399	0.391	0.383
1.0	0.386	0.378	0.369	0.361	0.352
1.5	0.287	0.276	0.266	0.256	0.245
2.0	0.263	0.252	0.242	0.231	0.220

a. Use a limited Table M approach to calculate the Insurance Charge.
b. Use the ICRLL procedure to calculate the total expected loss cost for this policy.

18. What was the purpose of the state/hazard group relativity? What implicit assumption is made when using a state/hazard group relativity?

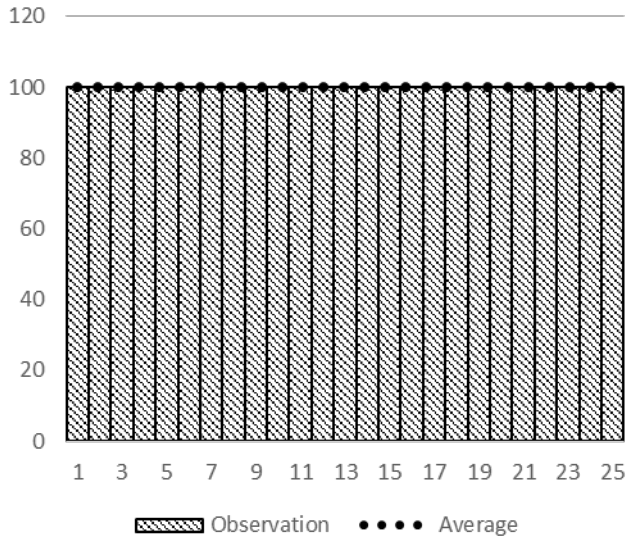
[51] Exercise 17 was adapted with permission from material written by Howard Mahler.

6. Understanding Aggregate Loss Distributions

To get an intuitive feel for how the distribution of deductible losses should behave, it is helpful to consider the extreme cases.
Consider first some extreme plan designs: A deductible policy with an infinite deductible but an aggregate limit on the deductible behaves like a retrospectively rated policy with a maximum, but no per-loss limitation and a minimum equal to basic times tax. Alternatively, a retrospectively rated policy with a per-loss loss limitation but an infinite maximum behaves exactly like a deductible policy with no aggregate limit.
Remember that a distribution of aggregate losses is information about the range of outcomes of many similar insurance policies. We will largely be concerned with the distribution of entry ratios, which are scaled with respect to expected aggregate losses. The shape of the distribution of entry ratios is largely driven by the variance of the underlying aggregate distribution. So it can be helpful to visualize some extreme outcomes, or extreme underlying severity distributions.

First, what would the aggregate loss distribution look like if every policy's losses were exactly equal to the expected losses? For example, every policy had exactly $100 of loss.

Exhibit 3.30. Twenty-five policies that all incur exactly their expected loss
(Only 25 shown so they can be seen)



The smallest outcome equals the expected loss of $100 equals the largest outcome.

At an entry ratio r = 0.8, the charge would be the part of the shaded area above the line y= (0.8 * E) = 80 divided by the total shaded area.

$$\phi(0.8) = (100\text{-}80)*25/(100*25) = 0.2.$$

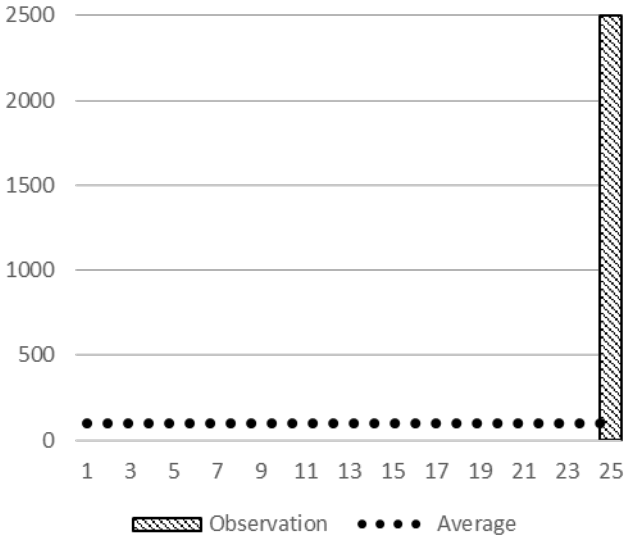
At an entry ratio r = 1.2, the charge would be the part of the shaded area above the line y= (1.2 * E) = 120 divided by the total shaded area. It can be seen there is no shaded area above 120, so

$$\phi(1.2) = 0$$

In fact, The Table M charge at any entry ratio r greater than or equal to 1 would be zero and the Table M charge for any entry ratio less than one would be 1-r.

Next consider the other extreme. What if a policy rarely had any losses, but if it had a loss, that loss would be enormous. For example, you might have a policy with a 1/10,000 chance of having any claims at all, but if it did have a claim, the claim was $1,000,000. This policy also has an expected loss of $100, but it has a very high variance. In this case, the Table M charge at an entry ratio of 1 would be nearly 1 (999,900/1,000,000 to be precise) because that one time in 10,000 when there is a loss, 999,900 of it will be in excess of the aggregate limit, and the other 9999 times when there is no loss the aggregate limit is irrelevant.

Exhibit 3.31.
Twenty-five policies, only one of which incurs any loss, with the same overall expected value as the policies pictured in Exhibit 3.30. (only 25 in this example, so they can be seen)



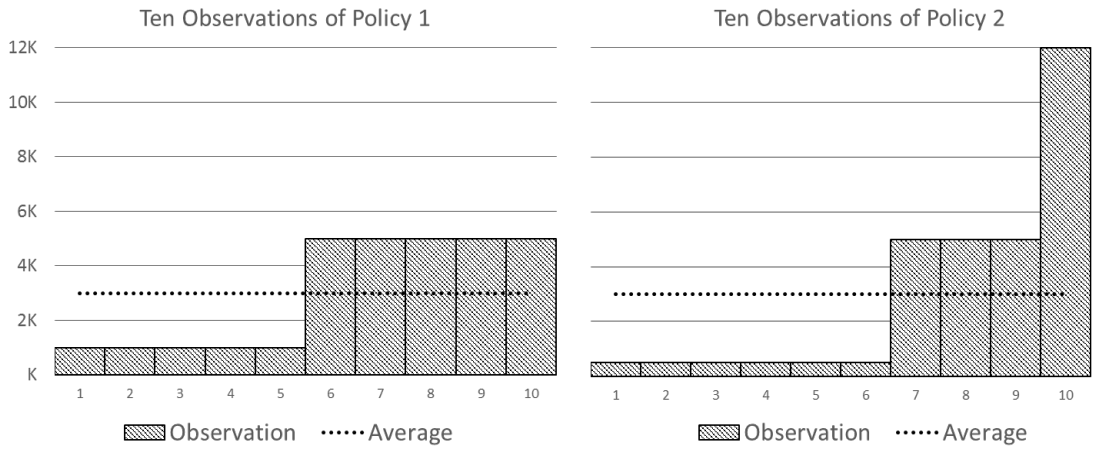
These extremely simple examples were chosen so the values are obvious and easy to calculate, but the same principles apply to real policies. Very small policies can be similar to the second extreme, because on very small policies the likelihood of even one claim is small, so most of the expected value of the aggregate loss is in the "tail,"[52] or unusually high outcomes (the rare cases when a loss occurs). In contrast, very large commercial policies are more like the first extreme.

[52] Mathematically, this is the right-hand tail of the distribution, but as most aggregate distributions that actuaries encounter are only defined between zero and infinity, the right hand tail is often referred to simply as "the tail."

They are expected to have a large number of claims, each of which is relatively small as compared to the total expected loss for the policy.
All other things being equal, the higher the expected number of claims, the lower the variance on the distribution of entry ratios, and the smaller the Table M charge is for entry ratios above 1. The possibility of extremely severe individual claims also increases the variance of the distribution of entry ratios.

For many types of insurance policies, the losses are driven by injuries to human beings. Some polices will tend to have more severe injuries than other policies (for example, a policy covering large trucks may have higher average liability severity than a policy covering private passenger vehicles, and a policy covering injuries to foundry workers will tend to have more severe claims than one covering injury to office workers.) But the difference in variance due to size of policy usually overwhelms those differences. That is, most of the difference in variance of aggregate loss among policies (and thus difference in Table M charge) is driven by the variance in the claim frequency.
But the variation of the severity distribution matters, too. For example, consider two policies, each of which has an expected frequency of 1 claim and no variance in the frequency—these imaginary polices will always have exactly 1 claim. Each claim on the first policy is equally likely to be $1000 or $5000. So the expected loss for the policy is $3000. Each claim on the second policy has a 60% change of costing $500, a 30% chance of costing $5000, and a 10% chance of costing $12,000. This policy also has an expected aggregate loss of $3000. Exhibit 3.32 compares the aggregate loss distributions of ten policies like the first, and ten like the second.

Exhibit 3.32. Comparing two policies with the same frequency and expected loss, but with different severity distributions



In this example, The charge at entry ratio of 1 is 0.33 = (2000*5)/(1000*5 + 5000*5) for policy 1 and 0.5 = (2000*3 + 9000)/(1000*6 + 5000*3 + 12,000) for policy 2.

For example, workers' compensation covers the same types of injuries to people all across the US. But some industries have a higher proportion of serious claims, and others have a higher proportion of minor claims. For instance, workers in metal foundries are subject to serious burns, whereas office workers are more likely to develop repetitive stress disorders. Because of this, the NCCI assigns workers' compensation job classes into seven Hazard Groups, A-G. Job classes in Hazard Group A have the smallest probability of serious injury leading to unusually high-severity claims, and those in Hazard Group G have the largest probability of a serious injury. Another driver of severity is location. The cost of the same injury may vary from place to place—medical care may be more expensive in one state than another. Also between the different states in the United States, workers' compensation laws vary significantly in the benefits they provide to injured workers for lost wages, and the courts may be more or less inclined to award very large liability verdicts.

In US workers' compensation, the NCCI reflects these differences by considering the state and hazard group of each risk when parametrizing the expected severity distribution used to generate aggregate loss factors. Similarly, when using the NCCI's new countrywide Table of Aggregate Loss Factors (Table M), the expected number of claims for a risk having a given expected loss will vary based on the average severity of its losses, which can be estimated by looking up the severity relativity of the appropriate state and hazard group mix of the policy. This can adjust the expected number of claims to reflect fewer (more) expected claims when a risk has a higher (lower) expected severity, so as to increase (decrease) the assumed variability of the risk. In general, for a given expected loss size, we treat a risk expected to have more severe individual claims as if it is smaller (and thus more variable) than a risk with the same expected loss due to a larger number of (on average) less severe claims.

To summarize, Table M shows different columns of aggregate loss factors for different expected claim count groups due to the impact of the size of a risk on the claim count distribution; all other things being equal, a larger insured has a tighter distribution of the ratio of observed claim

frequency over expected claim frequency than does a smaller insured. But severity distributions can vary as well, even within an insurance coverage.

If the severity distribution differs in scale, while having the same shape—in other words, the mean is different but the coefficient of variation and the skewness are approximately the same—simply adjusting the expected number of claims should yield reasonably accurate Table M charges (aggregate loss factors). However, if the difference is more extreme, we may need to also adjust the severity distribution, potentially needing to calculate a different Table M. Note that General Liability policies (especially products policies) and excess-of-loss reinsurance policies are more likely to differ significantly from other groups of policies due to their severity distribution.

Adjustments treating the differences as if they are driven mostly by scale have been used to adapt a table of expected aggregate charges developed from one coverage to be used for another coverage. For instance, some US insurers have used severity adjustment factors analogous to State/Hazard Group relativity factors in order to adapt a workers' compensation Table M to be used to estimate aggregate loss costs for Commercial Auto or General Liability. As always, care should be used when extrapolating that the resulting charges are reasonable. But sometimes there is not enough data to come up with a better estimate.

Questions

19. When all else is equal, if the variance of the loss distribution is larger, will the Table M charge be larger or smaller than with a smaller variance?

20. An actuary calculates the insurance charges on an aggregate deductible for a general liability policy for house painters. All the losses in the historical data used in the analysis resulted from inadequate and/or sloppy paint jobs, which were relatively inexpensive to fix. Later, it is discovered that some paint contained a toxic substance and those painters are liable for very expensive remediation of the painted properties.
The new claims are 10% as common as the historical claims. For every 10 claims that would have been expected before, there are now 11, one of which is cleaning up toxic paint.
Had this been known, the expected cost of a policy would have been twice the cost the actuary used.

(a) At an entry ratio of 2.00, with no per-occurrence loss limit, explain whether the insurance charge would increase, decrease, or stay the same.
(b) Explain how a per-occurrence limit would affect the change in the insurance charge for the aggregate deductible.

Acknowledgments

I would like to thank the many people who helped with everything from brainstorming the overall shape of this chapter to proofreading it. Jill Petker, Fran Sarrel, and Lawrence McTaggart helped with overall support and suggestions. X. Sherwin Li helped incorporate prior study notes and articles into this format and also helped review a draft. Dylan Lei and Matthew Iseler provided helpful and insightful feedback on a draft. Rick Gorvett provided some sample questions and pedagogic guidance. Teresa Lam, Jill Petker, Kirk Bitu, and Tom Daley helped me understand the changes to the NCCI retro rating plan. Nedzad Arnautovic, Elena Blagojevic, Brian Choi, Melanie_Dufour, Joe Harkman, Alex Leitheiser, Colin Rizzio, Josh Taub, and Matt Veibell, all helpfully submitted errata and suggestions to the earlier editions of this chapter. And I would especially like to thank Howard Mahler who reviewed multiple drafts of this chapter and found any number of typos and inelegant or confusing sections, suggested numerous clarifications and examples, and generously allowed me to use some of the exercises published in his study guide.

Index of equivalent terms

There are a great many names for many of the entities described in Chapter 3. As a convenience to the reader, I have tried to collect some of them here.

Table M Charge
Insurance Charge
Aggregate Excess Loss Factor
Aggregate Excess Ratio

Table M Saving
Insurance Saving
Aggregate Minimum Loss Factor

Net Insurance Charge
Net Aggregate Loss Factor
Net Table M Charge

Table M
Table of Insurance Charges
Table of Aggregate Loss Factors

Excess Loss Factor

Chapter 4: Concluding Remarks

By Ginda Kaplan Fisher

1. General Observations

In general, the premium for an insurance policy should pay for the expected costs, including the cost of capital supporting the policy. When retrospectively rated policies were developed, it was considered desirable that the expected premium to be paid by the insured would be the same, regardless of the retrospective policy provisions. (Obviously, the *actual* premium could vary, if actual losses were more or less than expected.) This was called a Balanced Plan.

Since then, large deductible policies and other policies that remove a significant fraction of the costs from "premium" have been developed. Also, as discussed in Chapter 2, the risk load and expected expenses to be paid may be significantly different with different policy provisions. There are still some highly regulated types of insurance where the expected premium must remain the same, but for most policy types, it is not necessary or desirable to balance the premium. It is still important that the pure premium or expected losses be balanced, however.

It is worth noting that there can be a great deal of uncertainty or risk in both the aggregate and per-occurrence excess loss. Especially when very high layers are insured, it is common for the risk charge to be greater than the loss cost for some portions of the coverage. Sometimes, the actual expected loss is so much less than the value of the risk that neither party cares much about the actual loss cost.

This study note focuses on loss cost, so it deals with those cases where the loss cost is significant *enough* to be worth estimating, and Chapter 3 provides tools for the actuary to estimate the cost of the aggregate excess loss. However, if the actuary uses these methods and comes up with some insignificant loss charge, it is usually appropriate to charge something for the coverage. The same is true for high layers of per-occurrence loss. Remember that if the customer wants to buy protection for some layer of risk, it is worth something to the insured. Maybe the insured knows something they aren't sharing. Even if the actuary is very comfortable with the total or primary loss pick,[53] the risk load and the expected expense of maintaining a loss-sensitive provision should be carefully evaluated.

Questions:

1. Would you expect a fair premium for a retrospective plan with a high aggregate loss limit to be more or less than the fair premium for a guaranteed cost plan covering the same risk?

2. Why should you generally recommend charging a non-negligible premium for high layers of aggregate or per-occurrence loss even though you estimated the loss cost for those layers to be negligible?

[53] A common name for the actuary's best estimate of E or $E(A_D)$. In some cases, rather than estimate the total expected loss, the actuary will use the more stable limited loss data to select a limited expected loss, the "primary loss pick" and estimate the other loss components from that.

2. Sensitivity of Table M charges to the Accuracy of the Loss Pick or Rate Adequacy

Also, whenever an actuary is pricing a loss sensitive plan (e.g., a deductible or retrospective policy) with an aggregate limit/maximum, the actuary should be aware of the leverage that the primary loss pick has on the insurance charge. This section has been adapted from a prior CAS study note[54]

It is tempting to think that this loss pick isn't very important, because the insured is responsible for those losses. This may be true if the entry ratio is very high and the deductible relatively low, as most of the insured losses will be in the per-occurrence excess portion, not the aggregate excess portion.[55] However, if the primary entry ratio is relatively low, or the deductible is very high, a significant portion of the expected insured losses will come from the aggregate. The loss pick might be inadequate on a large account because the underwriter has been optimistic, or on a small account because the state has demanded inadequate filed rates. In any case, as every actuary knows, it is hard to predict the level of future expected losses. An excessive loss pick will also lead to an inappropriate insurance charge.

The following example was priced using the Countrywide Table of Aggregate Loss Factors found in NCCI's Circular CIF-2017-32 "Countrywide–Announcement of Item R-1414–Revisions to Retrospective Rating Plan Manual Appendix B and All Related Rules and Endorsements"

Assume the actuary and underwriter expected there to be $500K total loss on a policy, and priced the policy accordingly. The underwriter sold an aggregate loss limit with an intended entry ratio of 2, or $1M. But in fact, the true expected loss is $550,000. Assume the error in estimation is due to frequency. What happens to the Table M charge?

The actuary determined this account would have about 42 claims, and so fit Expected Claim Group 41. The aggregate excess loss factor for an unlimited loss (k=0) in Expected Claim Group 41, at an entry ratio of 2 is 0.2108. So the Table M charge the actuary calculates is $500,000*0.2108 = $105,400.

But the true expected loss is $550,000. So the actual entry ratio at which aggregate losses will be capped is $1M/$550K = 1.82

The true expected number of claims is closer to 46, pushing the account into the slightly cheaper expected claim group 40. Even so, the aggregate excess loss factor in Expected Claim Group 40, at an entry ratio of 1.82 is 0.2228. And the true expected Table M charge is $550,000*0.2228 = $122,540. The aggregate limit has been underpriced by more than 16%.

[54] Ginda Kaplan Fisher, "Pricing Aggregates on Deductible Policies," 2002, published by the Casualty Actuarial Society as part of the Syllabus of Exams.

[55] Of course, if the excess portion is priced as a fraction of the primary loss pick, then the primary loss pick is important in pricing this component, too.

Exhibits 4.1 through 4.5 show an example of the impact on the insurance charge of an inadequate or excessive loss.[56]

In this example, the NCCI countrywide subtable 6 Table M charges were used, that is, this example represents a retrospective policy with a loss limitation of about 12%. However, the same effect would occur on any other insurance charge priced in a similar way (using Table M_D, ICRLL, etc.) Notice that the dollar error in insurance charges is greatest for large policies at low entry ratios, but the largest (absolute value) observed percent error in insurance charge is for a large policy at entry ratio 2. The percent error in the total expected losses for a deductible policy would also depend on the expected deductible losses. In any case, it is easy to see that adequate (primary) loss estimates are important to the profitability of a book of loss-sensitive policies.

Exhibit 4.1. If rates/loss picks are correct; Tables of % and $ Charge[57]

					At Entry Ratio				
true expected loss	Loss Pick	expected severity	true expected number of claims	Expected Claim Count group	1	1.2	1.7	2	3
3,000,000	3,000,000	12,000	250.0	29	0.25	0.18	0.07	0.04	0.01
1,000,000	1,000,000	12,000	83.3	36	0.32	0.25	0.13	0.09	0.02
500,000	500,000	12,000	41.7	41	0.37	0.31	0.18	0.13	0.04
100,000	100,000	12,000	8.3	57	0.54	0.49	0.39	0.34	0.23

true expected loss	Loss Pick	expected severity	true expected number of claims	Expected Claim Count group	1	1.2	1.7	2	3
3,000,000	3,000,000	12,000	250.0	29	762,600	543,900	220,500	124,500	16,500
1,000,000	1,000,000	12,000	83.3	36	319,400	247,900	128,400	85,300	20,700
500,000	500,000	12,000	41.7	41	186,300	152,750	91,600	66,550	22,400
100,000	100,000	12,000	8.3	57	54,440	49,310	39,170	34,440	23,240

[56] Using an inappropriate aggregate loss distribution can also produce significant pricing problems.
[57] $Charge based on true "expected loss."

If rates or loss picks are 10% inadequate, charges may be more than 20% inadequate:

Exhibit 4.2. If rates are 10% inadequate; Table of % and $Charge[58]

					Entry Ratio				
true expected loss	Loss Pick	expected severity	True expected number of claims	Expected Claim Count group	1	1.2	1.7	2	3
3,300,000	3,000,000	12,000	275.0	28	0.29	0.21	0.09	0.05	0.01
1,100,000	1,000,000	12,000	91.7	35	0.35	0.27	0.15	0.10	0.03
550,000	500,000	12,000	45.8	40	0.40	0.33	0.20	0.15	0.05
110,000	100,000	12,000	9.2	56	0.56	0.51	0.41	0.36	0.24

true expected loss	Loss Pick	expected severity	true expected number of claims	Expected Claim Count group	1	1.2	1.7	2	3
3,300,000	3,000,000	12,000	275.0	28	945,120	694,320	297,330	175,230	26,070
1,100,000	1,000,000	12,000	91.7	35	382,030	302,170	161,810	110,550	28,930
550,000	500,000	12,000	45.8	40	218,460	181,280	110,935	82,225	29,150
110,000	100,000	12,000	9.2	56	61,589	56,012	44,649	39,413	26,774

%error with 10% inadequate loss pick

% Error at Sold or intended Entry Ratio				
1	1.2	1.7	2	3
24%	28%	35%	41%	58%
20%	22%	26%	30%	40%
17%	19%	21%	24%	30%
13%	14%	14%	14%	15%

[58] $Charge based on true "expected loss."

If rates or loss picks are 10% excessive, charges may be more than 25% excessive:

Exhibit 4.3. If rates are 10% Excessive; Table of % and $Charge[59]

					Entry Ratio				
true expected loss	Loss Pick	expected severity	True expected number of claims	Expected Claim Count group	1	1.2	1.7	2	3
2,727,273	3,000,000	12,000	227.3	29	0.22	0.15	0.05	0.03	0.00
909,091	1,000,000	12,000	75.8	36	0.28	0.21	0.10	0.06	0.01
454,545	500,000	12,000	37.9	42	0.35	0.28	0.17	0.12	0.04
90,909	100,000	12,000	7.6	58	0.53	0.48	0.38	0.33	0.22

true expected loss	Loss Pick	expected severity	true expected number of claims	Expected Claim Count group	1	1.2	1.7	2	3
2,727,273	3,000,000	12,000	227.3	29	586,636	400,909	145,364	76,364	7,909
909,091	1,000,000	12,000	75.8	36	256,000	193,000	2,727	58,818	12,182
454,545	500,000	12,000	37.9	42	158,500	128,682	75,227	53,682	17,091
90,909	100,000	12,000	7.6	58	48,100	43,436	34,309	30,082	20,173

%error with 10% excessive loss pick

% Error at Sold or intended Entry Ratio				
1	1.2	1.7	2	3
-23%	-26%	-34%	-39%	-52%
-20%	-22%	-28%	-31%	-41%
-15%	-16%	-18%	-19%	-24%
-12%	-12%	-12%	-13%	-13%

[59] $Charge based on true "expected loss."

Questions:

3. Given a large deductible WC policy with the following features:

    - $2M expected total loss
    - Expected average severity of $10,000 per claim
    - The insured retains 86% of expected loss under the per-occurrence deductible (14% is expected to be excess of the deductible)
    - There is a limit on the aggregate deductible retained by the insured of $3M

    a.) What is the insurance charge for the aggregate limit?

If the account is larger than the pricing actuary realized, and the expected total losses should have been $2.5M, what should the insurance charge have been?

3. Consistency of Assumptions

The actuary should also be cautious of mismatched assumptions. Using different methods to calculate the per-occurrence excess charges and aggregate excess charges can sometimes lead to disjointed results. For instance, a company might have developed estimates of per-occurrence excess losses independently of the method used to develop its estimate of aggregate excess losses. Perhaps the company has estimated its own per-occurrence excess loss factors, but is relying on a rating bureau for aggregate excess loss factors. When this happens, a plan might come up with different pricing depending on how it is described. For instance, if the per-occurrence limit on a retrospectively rated plan is greater than or equal to the aggregate limit, the actuary's pricing model ought to recommend the same loss cost whether or not the per-occurrence limit is mentioned. But if the estimated charges were developed independently, that might not happen.

Mismatches in assumptions can creep into calculations in all sorts of places, including systematic errors. For instance, an actuary might look at a rating bureau's "pure premium" for a slice of the risk. But sometimes rating bureau pure-premium is "loaded" with various non-loss items, such as provisions for loss based assessments and LAE. If unadjusted rating bureau excess factors are multiplied by a loss estimate that doesn't include those components (and thus is smaller), excess losses can be underestimated, sometimes substantially so[60].

The actuary should be careful to monitor pricing assumptions for consistency and reasonability. When designing a pricing model, the actuary should compare the sum of the predicted primary, per-occurrence excess, and aggregate excess losses for various types of policies that might be written on various types of accounts, and ensure that the sums of the parts compare reasonably with the predicted total losses for those accounts. If not, an investigation of the assumptions used in estimating the per-occurrence excess and aggregate excess losses is in order.

[60] Whenever using factors from somewhere else, an actuary should ideally be familiar with the assumptions behind the calculation of those factors.

Acknowledgments

I would like to thank Jill Petker and Paul Ivanovskis for bringing many of these issues to my attention, and encouraging me to include them in the scope of a study note.

Solutions to Chapter Questions

Chapter 1 Answers

1. The objectives of experience rating are to:
    a. Increase equity
    b. Incentive for safety
    c. Enhance market competition

2. Experience rating adjusts a risk's rate to be more in line with that risk's expected loss experience. Risks whose expected loss experience is lower than average will pay less premium, and risks whose expected loss experience is higher will pay more premium.

3. Company B—since Company B has fewer rating classes, there will probably be more variation in risks within each rating class. The use of experience rating will allow the company to further tailor each risk's premium with its loss potential.

4. Without experience rating, a company would charge better than average and worse than average risks the same rate. Better than average risks might be able to find a lower rate with another company that recognizes the risk's lower loss potential. If enough of the better than average risks do this, the company will be left with only the worse than average risks.

5. In a group of risks, some of the difference in experience is due to underlying differences in the loss potential of the different risks. This is the variance of the hypothetical means. Some of the difference in experience among the risks is purely random, i.e., the process variance. Experience rating attempts to identify and adjust for the VHM, while at the same time not penalizing risks for differences in experience that are purely random.

6. Probably not. If the safety program in question does in fact reduce this risk's loss potential, this will be reflected in the risk's past experience and will be picked up by the experience rating. Using a schedule credit would double-count the expected benefit of the safety program. However, if the safety program is new (i.e., it was implemented during or after the experience period used by the experience rating program) then there is some expected benefit that would not be reflected in the past experience.

Chapter 2 Answers

1. There is no credit risk related to self-insured retentions because the insurer does not pay the retained losses up front, and therefore does not need to seek reimbursement from the self-insured.
2. 1.053 = 1 / (1 – 0.05)
3. 0.048 =1 - 1/1.05
4. The tax multiplier needs to account for the fact that premium tax is part of premium and therefore is itself taxed.
5. 13% = 0.20 – (1.10 – 1) x 0.70. That is, 13% of the guaranteed-cost premium will be collected as a fixed expense through the basic premium amount.
6. 1.15 = 1 + (0.25 - 0.15) / 0.65
7. Total losses, limited to the per-occurrence loss limit, are 315,000 = 25,000 + 15,000 + 25,000 + 50,000 + 2 x 100,000. This is below the maximum ratable loss amount.

   The retrospective premium is $511,892 = (150,000 + 1.1 x 315,000) x 1.031
8. As the loss conversion factor increases, expenses are shifted out of the basic premium, and the basic premium decreases.

   As the loss limit increases, the charge for per-occurrence excess exposure decreases, and the basic premium decreases.
   As the maximum premium or maximum ratable loss amount increases, the charge for aggregate excess exposure decreases, and the basic premium decreases.
   As the minimum premium or minimum ratable loss amount increases, the net charge for aggregate excess exposure (i.e., the net insurance charge) decreases, and the basic premium decreases.
   As the account size increases, there are two effects. The amount of premium discount increases, reducing the percentage expense provision in the basic premium. In addition, larger accounts have more stable loss experience, so the charge for aggregate excess exposure decreases. (The latter impact may become clearer after reading chapter 3.) For both of these reasons, the basic premium decreases.
9. The premium is calculated as:

35,000	fixed expense
30,000	loss-based expense = 300,000 x 10%
5,000	underwriting profit
30,000	per-occurrence excess = 300,000 – 270,000
10,000	aggregate excess = 270,000 – 260,000
110,000	subtotal
113,402	including premium tax = 110,000 x (1/.97)

10. A loss-sensitive dividend plan is unbalanced because if loss experience is better than expected, the insured receives a dividend, but if loss experience is worse than expected, the insured does not incur any additional costs.

11. The risk transfer is the same for a retrospective rating plan and a large deductible plan when:
    a. The loss limit for the retrospective rating plan equals the per-occurrence deductible.
    b. The maximum ratable loss amount for the retrospective rating plan equals the aggregate deductible limit.
    c. There is no minimum ratable loss amount for the retrospective rating plan.

Chapter 3 Answers

1.

Problem	Claim #	Claim $000	Occ.-Ltd. Agg. Sum	Occ. Excess	Agg. Excess	Total Insurance	Insured Payment
(a)	1	3	3	0	0	0	3
	2	8	11	0	0	0	8
	3	14	21	4	0	4	10
(b)	4	12	31	2	6	8	4
(c)	5	18	41	8	10	18	0

(a) Insurer = 0+0+4 = **4**; Insured = 3+8+10 = **21**
(b) Insurer = 4+8 = **12**; Insured = 21+4 = **25**
(c) Insurer = 12+18 = **30**; Insured = 25+0 = **25**

2.
(a) $500K + $100K + $300K + $2,000K = **$2,900K**
(b) $500K + $100K + $250K + $250K = **$1.1M**
(c) $1.1M - $1.0M = **$100K**
(d) Only the $2M claim breaches the per-claim policy limit, so **$1M**

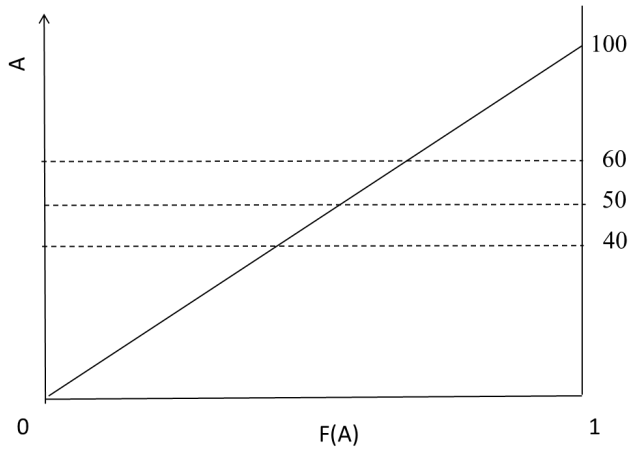
(e) Prior to application of the policy's aggregate limit, the policy would cover

- $0 on the small claims
- $0 on the $100K claim
- $50K on the $300K claim
- $750K on the $2M claim
- $100K for the aggregate on the deductible

For a total of **$900K**.
(f) **Zero** (900K < 5M)
(g) **$900K**

3.

The Table M insurance charge associated with a given outcome is the ratio of the area bounded by F(A) and that outcome to the total area under F(A). The total area under the curve F(A) = 100*1/2 = 50.



a) The area above the line A = 40 and below F(A) has area = 60 * 0.6 * ½ = 18

18/50 = 0.36

b) The area above the line A = 50 and below F(A) has area = 50 * 0.5 * ½ = 12.5

12.5/50 = 0.25

c) The area above the line A = 60 and below F(A) has area = 40 * 0.4 * ½ = 8
8/50 = 0.16

Calculus:
First we normalize the Lee diagram so that the area under the curve (the distribution of the probability of aggregate loss) adds to 1. Then

$$Let\ Y = {}^{A}/_{E} \sim Uniform(0,2)$$

$$Then\ f(y) = 0.50$$

(a) $r = {}^{40}/_{50} = 0.80$ $\phi(0.80) = \int_{0.8}^{2} 0.50(y - 0.80)dy = \mathbf{0.36}$

(b) $r = {}^{50}/_{50} = 1$ $\phi(1) = \int_{1}^{2} 0.50(y - 1)dy = \mathbf{0.25}$

(c) $r = {}^{60}/_{50} = 1.20$ $\phi(1.20) = \int_{1.2}^{2} 0.50(y - 1.20)dy = \mathbf{0.16}$

4.

$$E = 10 \quad Y = {}^{A}/_{E} \sim Expon(mean = 1) \quad f(y) = e^{-y}$$

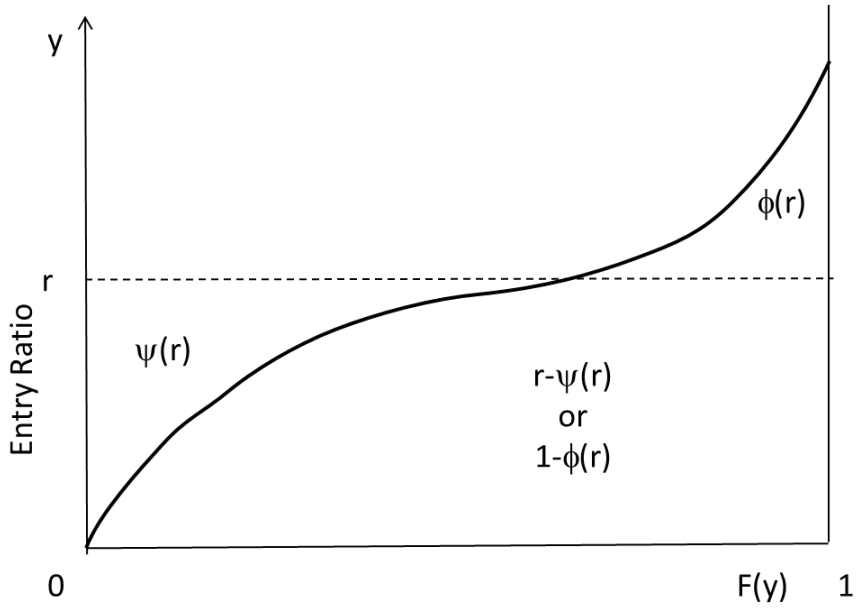
$$\psi(r) = \int_{0}^{r} (r - y)e^{-y}dy$$

(a) $r = {}^{5}/_{10} = 0.50$ $\psi(0.50) = \mathbf{0.1065}$

(b) $r = {}^{10}/_{10} = 1$ $\psi(1) = \mathbf{0.3679}$

(c) $r = {}^{15}/_{10} = 1.5$ $\psi(1.5) = \mathbf{0.7231}$

5.



6.

(a) $\Phi(R) = \mathbf{A}$

(b) $\Phi(S) = \mathbf{A + D + E}$

(c) $\psi(R) = \mathbf{B + C + F}$

(d) $\psi(S) = \mathbf{F}$

7.

(a) Step 1: Expected loss ratio = Average Loss Ratio = (20%+40%+40%+60%+80%+80%+120%+200%)/8 = 80%

So for each risk, r = loss ratio/0.8

To solve this precisely, we need to look at every slice at which there is at least one data point, plus all the other points we want factors for:

r	# Risks from prior to	# Risks Above	% Risks Above	Difference in r	$\Phi(r)$	$\psi(r) = \Phi(r)+r-1$
0	0	8	1.000	0.25	1.00	0
.25	1	7	0.875	0.25	.75	0
.50	2	5	0.625	0.25	.53125	.03125
.75	1	4	0.500	0.25	.25+.5*.25=.375	.375+.75-1=.125
1.0	2	2	0.250	0.50	.125+.25*.5=.25	0.25+1-1=.25
1.5	1	1	0.125	0.50	.125+.125*1=.125	0.125+1.5-1=0.625
2.0	0	1	0.125	0.50	0+.125*.5=.0625	.0625+2-1=1.0625
2.5	1	0	0.000	0.50	0	0+2.5-1=1.5
3.0	0	0	0.000	0	0	0+3-1=2

Total	8					

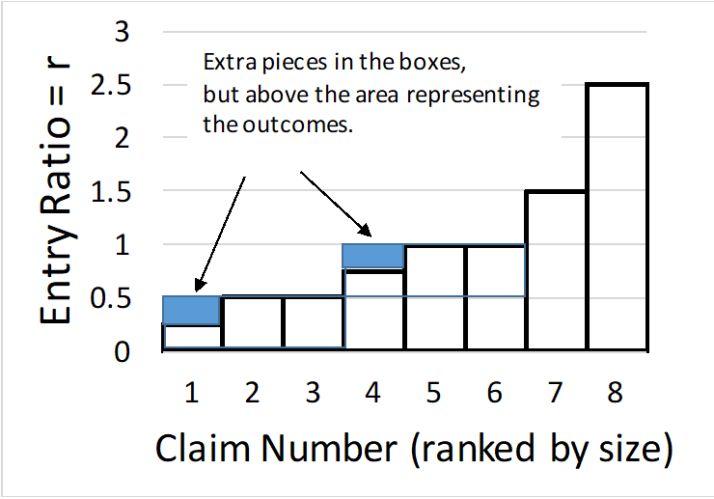
This can then be summarized as:

r	$\Phi(r)$	$\psi(r) = \Phi(r)+r-1$
0	1.00	0
.50	.53125	.03125
1.0	.25	.25
1.5	0625+.125*.5=.125	0.125+1.5-1=0.625
2.0	0+.125*.5=.0625	0.0625+2-1=1.0625
2.5	0	0+2.5-1=1.5
3.0	0	0+3-1=2
Total		

Note what happens if we just look at intervals of 0.5

r	Risks from prior to	# Risks Above	% Risks Above	fference in r	$\Phi(r)$	$\psi(r) = \Phi(r)+r-1$
0	0	8	1.000	0.5	1.0625	.0625
.50	3	5	0.625	0.5	.5625	.0625
1.0	3	2	0.250	0.5	.25	.25
1.5	1	1	0.125	0.5	0625+.125*.5=.125	0.125+1.5-1=0.625
2.0	0	1	0.125	0.5	0+.125*.5=.0625	0.0625+2-1=1.0625
2.5	1	0	0.000	0.5	0	0+2.5-1=1.5
3.0	0	0	0.000	0	0	0+3-1=2
Total	8					

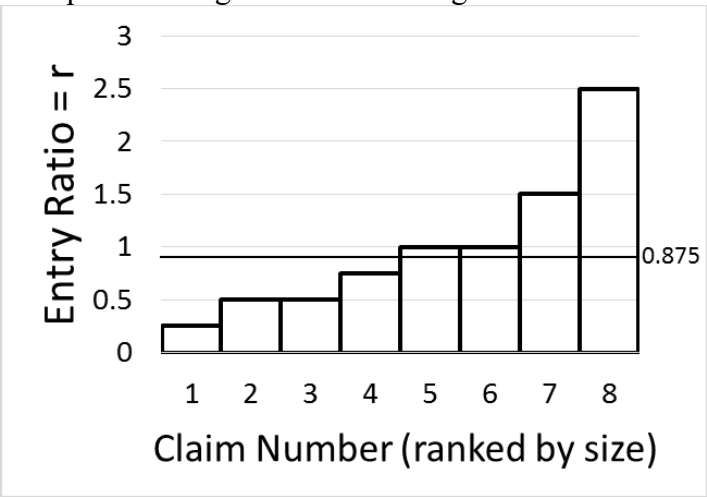
All looks well until we get to r = 0.5. Then we end up with 0.5625 for the charge, rather than 0.53125. The calculated charge is too large by 0.03125. The error grows at r=0 to 0.0625. This is because when we "integrate" under the curve, we are assuming each piece we are adding is a rectangle. But if we look at the Lee diagram of the outcomes:



We can see that there are two extra pieces in the boxes, but above the area representing the outcomes. Each has area = 0.25 x (1/8) = 0.03125. That is the source of the extra in the calculation. This is an example of the problem described in footnote 30 on page 66.

(b) 70% L/R ➔ r=0.875

It helps to look again at the Lee diagram to understand the situation



There are a few ways to approach this problem. First, we will solve it precisely, using the horizontal method:

The width of the distribution between 1.0 and 0.875 is 0.5, four of the 8 claims. So the additional insurance charge is 0.5 times the height of the additional band, or (1-0.875) * 0.5 = 0.0625. so

$\Phi(0.875) = \Phi(1.0) + 0.0625 = 0.25 + 0.625 =$ **0.3125**

We could also have interpolated. Had we estimated $\Phi(0.875)$ with a linear interpolation, we would have gotten:

$(\Phi(0.5) * (1\text{-}0.875) + \Phi(1.0) * (0.875 - 0.5))/(1.0 - 0.5) =$
$(0.53125 * 0.125 + 0.25 * 0.375)/(0.5) =$ **0.3203**

Note that the interpolation does not give the exact answer, but is reasonably close.

Since we are interested in the insurance charge at a single point, we might also use the vertical method.

Risk	Actual Agg. L/R	Entry Ratio	Excess of r = 70/80 = 0.875	Excess of r = 110/80 = 1.375
1	20%	0.25	0	0
2	40	0.50	0	0
3	40	0.50	0	0
4	60	0.75	0	0
5	80	1.00	0.125	0
6	80	1.00	0.125	0
7	120	1.50	0.625	0.125
8	200	2.50	1.625	1.125
Total:			**2.500**	**1.250**
Average			**0.3125**	**0.15625**

(c) $\psi(0.875) = 0.3125 + 0.875 - 1 =$ **0.1875**
(d) 110% L/R ➔ r=1.375. $\Phi(1.375) =$ **0.15625** (work done in section b)
(e) $\psi(1.375) = 0.15625 + 1.375 - 1 =$ **0.53125**

8. Using Parameterized distributions to develop Tables M
Advantages:

1. When you don't have a statistically credible group of policies to base your pricing on, but you have an idea of what shape the distribution of outcomes is likely to approximate, you can fit curves to what data you have.
2. When data is thin, and you have large gaps between empirical entry ratios, you don't have to rely on linear interpolation.
3. Even with large body of data, fitting distributions to frequency and severity can help develop charges that are consistent with the excess charges.

Disadvantages:

1. If the assumptions underlying the selected distribution aren't close enough to reality, you can generate plausible, internally consistent, precise, but misleading charges.
2. It might be more computationally complex to build a model than to use empirical data for the desired degree of precision.

9.

(a) M_D Charge at R = **G**
(b) M_D Savings at S = **Q+T+U**
(c) Per-Occ XS Charge at D = **A+D+E+J+L+N+T+U**

10.

(a) {40,000 + (1.1)(150,000)} (1.05) = 215,250.

Comment: The insured benefited from neither the maximum premium nor the accident limit.

(b) {40,000 + (1.1)(200,000)} (1.05) = 273,000. Limited to the maximum of $250,000.

Comment: The insured benefited from the maximum premium.

(c) {40,000 + (1.1)(100,000)} (1.05) = 157,500.

Comment: The insured benefited from the accident limit.

(d) {40,000 + (1.1)(200,000)} (1.05) = 273,000. Limited to the maximum of $250,000.

Comment: The accident limit decreased the losses entering the calculation, but the insured ended up paying the maximum premium anyway.

The last case is an example of the "overlap" between the effects of the maximum premium and the accident limit. In some years, even though there are large accidents, the accident limit will not provide any additional benefit to the insured beyond that provided by the maximum premium. In other words, for large accidents the accident limit and the maximum premium overlap

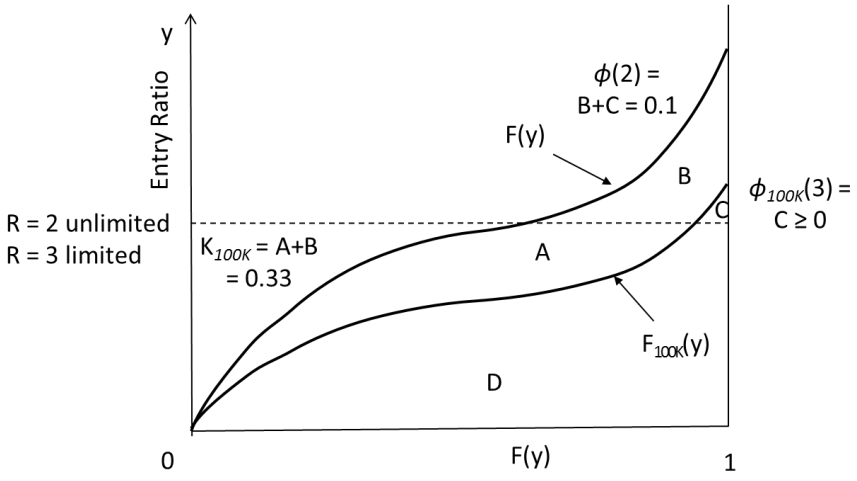
11.
(a) {400,000 + (1.1)(150,000)}(1.05) = 593,250. The insured pays the minimum premium, $650,000.
(b) {400,000 + (1.1)(100,000)}(1.05) = 535,500. The insured pays the minimum premium, $650,000.

The last case is an example of the "underlap" between the effects of the minimum premium and the accident limit. In some years, even though there are large accidents, the accident limit will not provide any benefit to the insured due to the minimum premium. This has a relatively small overall impact.

12.
E=150,000
Agg Limit = 300,000
R = 300 / 150 = 2.0
$\Phi(2.0)$ x 150,000 = 15,000
$\Phi(2.0) = 0.10$

With only Per-Occurrence Deductible:
k * E = 50,000 ➔ k = 50,000 / 150000 = 0.333

Note that the Aggregate Limit of 300,000 is three time the expected limited loss of 100,000, so the entry ratio, r, of the limited loss distribution is 300,000 / 100,000 = 3.



Together, the combined charge would be 0.333+0.10 = 0.433, or 0.433 x 150,000 = 65,000 However, the combined charge is very unlikely to be equal to 65,000. It will generally be less than 65,000 because there is overlap between the two charges, as shown by region B in the graph.

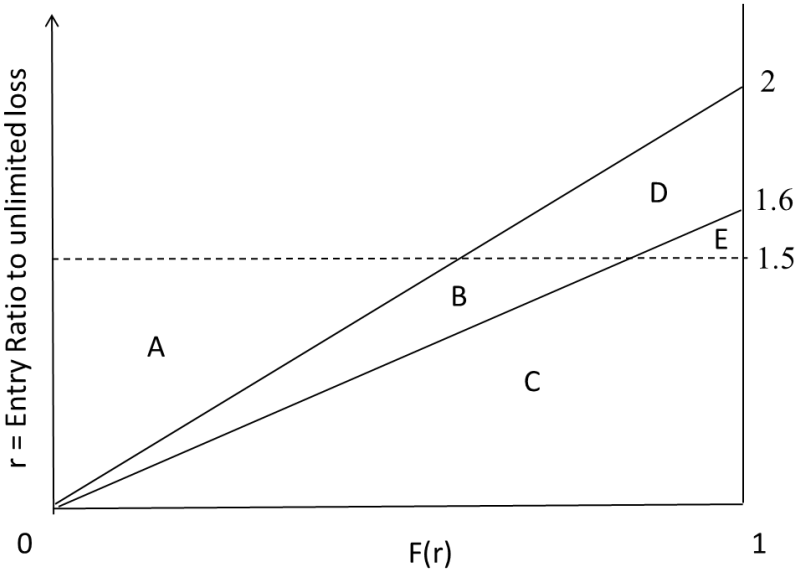
13.

a) The expected primary losses = 20,000.
Entry ratio = 40,000/20000=2.0
The aggregate excess loss factor is 0.04, so the insurance charge = 0.04 x 20,000=**800**.
(Which would be in addition to the $20,000 charge for the per-occurrence deductible, for a total expected loss cost within the policy of $20,800.)

b) The expected primary losses = 30,000.
Entry ratio = 40,000/30,000=1.333.
Interpolate the aggregate excess loss factors on the table to get
(1/3)*0.22 + (2/3)*.12 =0.1533 = the aggregate excess loss factor.
So the insurance charge is 0.1533 x 30000 = **4600**
(which would be in addition to the $10,000 charge for the per-occurrence deductible, for a total expected loss cost within the policy of $14,600.)

c) The insurer will charge more for (a) because even though the aggregate insurance charge is less than (b), the insured has a much smaller per-occurrence deductible which transfers more expected losses to the insurer.

14.



a) $\Phi^*_D(1.5) = B + D + E$
$\psi^*_D(1.5) = A + B$

b) The normalized area of the total loss (the area of the large triangle, B+C+D+E) is 1. The area of the limited loss is 400/500 times the total, or 0.8. It is also the area of the small triangle, C + E.
So the area of B + D is 0.2.

The area of E is the ½ the height times the length. The height is 1.6 – 1.5 = 0.1.
E is the same shape as C+E, which has height 1.6 and length 1. So the length of E must be (0.1/1.6) = 0.0625.
So the area of E is 0.1 * 0.0625 / 2 = 0.003125.

So $\Phi^*_D(1.5) = B + D + E = 0.2 + 0.003125 = 0.203125$
And
$\psi^*_D(1.5) = \Phi^*_D(1.5) + r - 1 = 0.203125 + 1.5 - 1 = 0.703125.$

15.

a) T+U+J+L+N+D+E+A+G
b) Q+U+T

16.

Advantages:

- When large tables were awkward (either as a vast pile or paper or a computer file that was difficult to store) ICRLL allowed a single unlimited Table M to be used to generate reasonable charges that would otherwise have required a large number of Tables M_D.
- The ICRLL procedure is an expedient way of approximating Table M_D when a suitable Table M_D is unavailable. For example, this method can be used to adjust a Table M developed based on one book of business to a similar book for which there isn't adequate data to develop its own Tables M.

Disadvantages:

- It is only an approximation to Table M_D and may introduce additional error in the estimate of the charge for the aggregate limit.

17.

a) Expected Unlimited Losses = 1,000,000 x 0.75 = 750,000
Expected Limited Losses = 750,000 x (1 – 0.2) = 600,000
r = 1,200,000 / 600,000 = 2.0 ➔ Ins Charge = 0.04 x 600,000 = 24,000

b) Loss Group Adjustment Ftr = $\frac{1+0.8(0.2)}{1-0.2} = 1.45$

Losses used in group selection = 750,000 x 0.90 x 1.45 = 978,750 ➔ ELG 29
r = 1,200,000 / 600,000 = 2.0
Insurance Charge Factor = 0.242
Total Expected Loss Cost = 0.2(750,000) + 0.242(600,000) = 295,200

18.

The State/Hazard Group Relativity makes an adjustment to reflect for differences in claim size by class of business and geographic location. For a given expected loss size, it treats a risk expected to have more severe individual claims as if it is smaller (and thus more variable) than a risk with the same expected loss resulting from a larger number of less severe claims.

The implicit assumption with using the State/Hazard Group Relativity is that the actual severity distribution has a similar shape as that which is used to determine the insurance charges, and differs mostly due to scale. If the difference is extreme, the severity distribution may need to be adjusted, potentially requiring a different Table M

19.

The Table M charge will be larger when the variance of the loss distribution is larger, all else being equal – see Exhibit 3.32.

20.

a) The insurance charge would increase. The toxic paint claims have a low frequency and a very high severity compared to the historical claims. This means they greatly increase the variance of the severity, which increases the variance of the aggregate losses. In particular, with the original assumptions, an insured would have had to experience twice the expected claims frequency to breach the aggregate, but with the revised understanding of the liability, a single large claim would be almost enough to do so. This will result in a much larger insurance charge at an entry ratio of 2.0.

b) A per-occurrence limit, assuming it is high enough to be above the historical losses, but substantially limits the toxic paint claims, would shift losses from the aggregate limit charge to the per-occurrence charge. Therefore, the charge for the aggregate deductible would decrease.

Chapter 4 Answers

1. The fair premium should be lower, as the insured is responsible for a significant fraction of the risk.

2. Because you need to price for the risk involved and the expenses of monitoring the claims experience. The risk includes the risk that the insured knows more about the liability than you do. The expenses may include annual or monthly reports to the insured whether or not any losses breach the insurable threshold.

3. Given a large deductible WC policy with the following features:

- $2M expected total loss
- Expected average severity of $10,000 per claim
- The insured retains 86% of expected loss under the per-occurrence deductible (14% is expected to be excess of the deductible)
- There is a limit on the aggregate deductible retained by the insured of $3M

b.) What is the insurance charge for the aggregate limit?

If the account is larger than the pricing actuary realized, and the expected total losses should have been $2.5M, what should the insurance charge have been?

References

Bahnemann, D. "Distributions for Actuaries," CAS Monograph # 2.

Brosius, J. Eric., "Table M Construction," CAS Study Note, 2002.

Couret, Jose and Gillam, William R., "Retrospective Rating: 1997 Excess Loss Factors," *PCAS* LXXXIV, 1997.

Fisher, Ginda Kaplan, "Pricing Aggregates on Deductible Policies," CAS Study Note, May 2002.

Gillam, W. R., "Workers' Compensation Experience Rating: What Every Actuary Should Know," *PCAS* LXXIX, 1992.

Gillam, W. R. and Snader, R.H., "Fundamentals of Individual Risk Rating," National Council on Compensation Insurance (Study Note), 1992

Heckman, Philip E. and Meyers, Glenn G., "The Calculation of Aggregate Loss Distributions from Claim Severity and Claim Count Distributions," *PCAS* LXX, 1983.

Insurance Services Office, Inc., *Commercial General Liability Experience and Schedule Rating Plan*, 2006. (or any other year)

Lee, Yoong-Sin, "The Mathematics of Excess of Loss Coverages and Retrospective Rating—A Graphical Approach," *PCAS* LXXV, 1988.

Mahler, Howard C., Discussion of "Retrospective Rating: 1997 Excess Loss Factors," *PCAS* LXXXV, 1998, Including Errata.

Miccolis, R. S., "On the Theory of Increased Limits and Excess of Loss Pricing," *PCAS* LXIV, 1977. Including discussion of paper: Rosenberg, S., *PCAS* LXIV, 1977, pp. 60-73.

National Council on Compensation Insurance, *Experience Rating Plan Manual for Workers Compensation and Employers Liability Insurance.*

National Council on Compensation Insurance, *Retrospective Rating Plan Manual for Workers Compensation and Employers Liability Insurance.*

National Council on Compensation Insurance, Circular CIF-2017-32 *Countrywide Announcement of Items R-1414 Revisions to Retrospecitve Rating Plan Manual Appendix B and all Related Rules and Endorsements,* 2017.

Panjer, Harry "Recursive Evaluation of a Family of Compound Distributions," *Astin Bulletin*, Vol. 12, No. 1, 1981, pp. 22-26

Robertson, J.P., "NCCI's 2007 Hazard Group Mapping," *Variance*, Vol. 3, Issue 2, 2009, Casualty Actuarial Society, pp. 194-213.

Robbin, Ira, "Overlap Revisited—The 'Insurance Charge Reflecting Loss Limitation' Procedure," *Pricing*, Casualty Actuarial Society *Discussion Paper Program*, 1990, Volume 2.

Skurnick, D., "The California Table L," *PCAS* LXI, 1974, pp. 117-140. Including discussion of this paper: Gillam, W.R., *PCAS* LXXX, 1993, pp. 353-365.

Teng. M. T. S., "Pricing Workers' Compensation Large Deductible and Excess Insurance," Casualty Actuarial Society Forum, Winter 1994, pp. 413-437.

Venter, G.G., "Experience Rating—Equity and Predictive Accuracy," *NCCI Digest*, April 1987, Volume II, Issue I, pp. 27-35.

CAS MONOGRAPH SERIES
NUMBER 5
Second Edition

GENERALIZED LINEAR MODELS FOR INSURANCE RATING
Second Edition

Mark Goldburd, FCAS, MAAA
Anand Khare, FCAS, FIA, CPCU
Dan Tevet, FCAS
Dmitriy Guller, FCAS

CASUALTY ACTUARIAL SOCIETY

This monograph is a comprehensive guide to creating an insurance rating plan using generalized linear models (GLMs), with an emphasis on application over theory. It is written for actuaries practicing in the property/casualty insurance industry and assumes the reader is familiar with actuarial terms and methods. The text includes a lengthy section on technical foundations that is presented using examples that are specific to the insurance industry. Other covered topics include the model-building process, data preparation, selection of model form, model refinement, and model validation. Extensions to the GLM are briefly discussed.

GENERALIZED LINEAR MODELS FOR INSURANCE RATING
Second Edition

Mark Goldburd, FCAS, MAAA
Anand Khare, FCAS, FIA, CPCU
Dan Tevet, FCAS
Dmitriy Guller, FCAS

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, Virginia 22203
www.casact.org
(703) 276-3100

Generalized Linear Models for Insurance Rating
By Mark Goldburd, Anand Khare, Dan Tevet, and Dmitriy Guller

Library of Congress Cataloging-in-Publication Data
Generalized Linear Models for Insurance Rating / Mark Goldburd, Anand Khare, Dan Tevet, and Dmitriy Guller
ISBN 978-1-7333294-3-9 (print edition)
ISBN 978-1-7333294-4-6 (electronic edition)
1. Actuarial science. 2. Classification ratemaking. 3. Insurance—mathematical models.
I. Goldburd, Mark. II. Khare, Anand. III. Tevet, Dan.

Contents

2019 CAS Monograph Editorial Board

Acknowledgments

The authors would like to thank the following people for their contributions:

Ali Ishaq, Leslie Marlo, Glenn Meyers, Stan Khury and the other past and present members of the Monograph Editorial Board, without whose efforts this text would not have been possible.

Jason Russ, Fran Sarrel and Delia Roberts, who coordinated with the authors on behalf of the Examination and Syllabus Committee.

The anonymous peer reviewers, whose thoughtful suggestions improved the quality of this text.

Eric Brosius, Paul Ivanovskis and Christopher Mascioli, who also served as reviewers and provided valuable feedback on earlier drafts of the text.

Josh Taub, who provided valuable feedback on the first edition of this text, which improved the quality of the second edition.

Margaret Tiller Sherwood, Howard Mahler, Jonathan Fox, Geoff Tims, Hernan Medina, and Eric Kitchens, whose thoughtful comments and suggestions for improvement further enhanced the section edition.

Donna Royston, who provided editorial support and coordinated production on behalf of the CAS.

1. Introduction

Generalized linear models have been in use for over thirty years, and there is no shortage of textbooks and scholarly articles on their underlying theory and application in solving any number of useful problems. Actuaries have for many years used GLMs to classify risks, but it is only relatively recently that levels of interest and rates of adoption have increased to the point where it now seems as though they are near-ubiquitous. GLMs are widely used in the personal lines insurance marketplace, especially in operations of meaningful scale. But as far as the authors are aware there is no single text written for the practicing actuary that serves as a definitive reference for the use of GLMs in classification ratemaking. This monograph aims to bridge that gap. Our ultimate goal is to give the knowledgeable reader all of the additional tools they need to build a market-ready classification plan from raw premium and loss data.

The target audience of this monograph is a credentialed or very nearly credentialed actuary working in the field of property/casualty or general insurance (for example, in the United States, a member or soon-to-be member of the Casualty Actuarial Society). It is assumed that the reader will be familiar with the material covered in the earlier exams of the CAS syllabus, including all of the Actuarial Standards of Practice and the ratemaking material covered in depth in Werner and Modlin's *Basic Ratemaking* (2010) (or their international equivalents, for readers outside the United States). Prior knowledge of the mathematics underlying GLMs will make for faster reading but is not absolutely necessary. Familiarity with a programming language is not required to read the text, but will be necessary to implement models.

If you should have a suggestion or discover any errors in this document, please contact the authors. Current contact information can be found in the CAS directory.

2. Overview of Technical Foundations

Generalized linear models (GLMs) are a means of modeling the relationship between a variable whose outcome we wish to predict and one or more explanatory variables.

The predicted variable is called the **target variable** and is denoted y. In property/casualty insurance ratemaking applications, the target variable is typically one of the following:

- Claim frequency (i.e., claims per exposure)
- Claim severity (i.e., dollars of loss per claim or occurrence)
- Pure premium (i.e., dollars of loss per exposure)
- Loss ratio (i.e., dollars of loss per dollar of premium)

For quantitative target variables such as those above, the GLM will produce an estimate of the *expected value* of the outcome.

For other applications, the target variable may be the occurrence or non-occurrence of a certain event. Examples include:

- Whether or not a policyholder will renew their policy.
- Whether a submitted claim contains fraud.

For such variables, a GLM can be applied to estimate the *probability* that the event will occur.

The explanatory variables, or **predictors**, are denoted $x_1 \ldots x_p$, where p is the number of predictors in the model. Potential predictors are typically any policy terms or policyholder characteristics that an insurer may wish to include in a rating plan. Some examples are:

- Type of vehicle, age, or marital status for personal auto insurance.
- Construction type, building age, or amount of insurance (AOI) for homeowners insurance.

2.1. The Components of the GLM

In a GLM, the outcome of the target variable is assumed to be driven by both a *systematic* component as well as a *random* component.

The **systematic component** refers to that portion of the variation in the outcomes that is related to the values of the predictors. For example, we may believe that driver age influences the expected claim frequency for a personal auto policy. If driver age

is included as a predictor in a frequency model, that effect is part of the systematic component.

The **random component** is the portion of the outcome driven by causes *other than* the predictors in our model. This includes the "pure randomness"—that is, the part driven by circumstances unpredictable even in theory—as well as that which may be predictable with additional variables that are not in our model. As an example of this last point, consider the effect of driver age, which we describe above as being part of the systematic component—*if* driver age is in the model. If driver age is *not* included in our model (either due to lack of data or for any other reason), then, from our perspective, its effect forms part of the random component.

In a general sense, our goal in modeling with GLMs is to "explain" as much of the variability in the outcome as we can using our predictors. In other words, we aim to shift as much of the variability as possible away from the random component and into the systematic component.

GLMs make explicit assumptions about both the random component and the systematic component. We will examine each in turn, beginning with the random component.

2.1.1. The Random Component: The Exponential Family

In a GLM, y—the target variable—is modeled as a random variable that follows a probability distribution. That distribution is assumed to be a member of the **exponential family** of distributions.

The exponential family is a class of distributions that have certain properties that are useful in fitting GLMs. It includes many well-known distributions, such as the normal, Poisson, gamma and binomial distributions. (It also includes a less widely known distribution—the Tweedie distribution—that is very useful in modeling insurance data; more on that later.) Selection and specification of the distribution is an important part of the model building process.

The randomness of the outcome of any particular risk (denoted y_i) may be formally expressed as follows:

$$y_i \sim Exponential\left(\mu_i, \phi\right) \tag{1}$$

Note that "*Exponential*" above does not refer to a specific distribution; rather, it is a placeholder for any member of the exponential family. The terms inside the parentheses refer to a common trait shared by all the distributions of the family: each member takes two parameters, μ and ϕ, where μ is the mean of the distribution. ϕ, the **dispersion** parameter, is related to the variance (but is not the variance!) and is discussed later in this chapter.

The parameter μ is of special interest: as the mean of the distribution, it represents the expected value of the outcome. The estimate of this parameter is said to be the "prediction" generated by the model—that is, the model's ultimate output.

If no information about each record other than the outcome were available, the best estimate of μ would be the same for each record—that is, the average of historical outcomes. However, GLMs allow us to use predictor variables to produce a better estimate, unique to each risk, based on the statistical relationships between the predictors and the target values in the historical data. Note the subscript i applied to μ in Equation 1 above, which denotes that the μ parameter in the distribution is record-specific. The subscript-less parameter ϕ, on the other hand, is assumed to be the same for all records.

2.1.2. The Systematic Component

GLMs model the relationship between μ_i (the model prediction) and the predictors as follows:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{2}$$

Equation 2 states that some specified *transformation* of μ_i (denoted $g(\mu_i)$) is equal to the **intercept** (denoted β_0) plus a linear combination of the predictors and the **coefficients**, which are denoted $\beta_1 \ldots \beta_p$. The values for the intercept (β_0) and the coefficients ($\beta_1 \ldots \beta_p$) are estimated by GLM software. The transformation of μ_i represented by the function $g(.)$ on the left-hand side of Equation 2 is called the **link function** and is specified by the user.

The right-hand side of Equation 2 is called the **linear predictor**; when calculated, it yields the value $g(\mu_i)$—that is, the model prediction transformed by our specified link function. Of course, the value $g(\mu_i)$ per se is of little interest; our primary interest lies in the value of μ_i itself. As such, after calculating the linear predictor, the model prediction is derived by applying the *inverse* of the function represented by $g(.)$ to the result.

The link function $g(.)$ serves to provide flexibility in relating the model prediction to the predictors: rather than requiring the mean of the target variable to be directly equal to the linear predictor, GLMs allow for a transformed value of the mean to be equal to it. However, the prediction must ultimately be driven by a linear combination of the predictors (hence the "linear" in "generalized linear model.")

In a general sense, the flexibility afforded by the ability to use a link function is a good thing because it gives us more options in specifying a model, thereby providing greater opportunity to construct a model that best reflects reality. However, when using GLMs to produce insurance rating plans, an added benefit is obtained when the link function is specified to be the natural log function (i.e., $g(x) = \ln(x)$): a GLM with that specification (called a **log link** GLM) has the property of producing a multiplicative rating structure.

Here's why: when a log link is specified, Equation 2 becomes

$$\ln \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

To derive μ_i, the inverse of the natural log function, or the natural exponential function, is applied to both sides of the equation:

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) = e^{\beta_0} \times e^{\beta_1 x_{i1}} \times \cdots \times e^{\beta_p x_{ip}}.$$

As demonstrated, the use of a log link results in the linear predictor—which begins as a series of additive terms—transforming into a series of multiplicative factors when deriving the model prediction.

Multiplicative models are the most common type of rating structure used for pricing insurance, due to a number of advantages they have over other structures. To name a few:

- They are simple and practical to implement.
- Having additive terms in a model can result in negative premiums, which doesn't make sense. With a multiplicative plan you guarantee positive premium without having to implement clunky patches like minimum premium rules.
- A multiplicative model has more intuitive appeal. It doesn't make much sense to say that having a violation should increase your auto premium by $500, regardless of whether your base premium is $1,000 or $10,000. Rather it makes more sense to say that the surcharge for having a violation is 10%.

For these and other reasons, log link models, which produce multiplicative structures, are usually the most natural model for insurance risk.

2.1.3. An Example

Suppose we construct a GLM to predict the severity of auto claims using driver age and marital status as predictors. The data we use contains 972 rows, with each row corresponding to a single claim. For each claim, the loss amount is recorded, along with several policyholder characteristics, among which are our predictors: driver age (in years, and denoted x_1 in our example) and marital status (coded as 0 = unmarried, 1 = married, and denoted x_2). We aim to produce a multiplicative rating algorithm, so a log link is used. We believe that the loss amount generated by a claim, after accounting for the effect of age and marital status, is random and follows a gamma distribution.

For this setup, our model inputs are:

- The data
- The model specifications:
  - *Target variable:* loss amount
  - *Predictors:* driver age (x_1) and marital status (x_2)
  - *Link function:* log
  - *Distribution:* gamma

The above are entered into the GLM fitting software. The outputs of the model fitting process are: estimates for the intercept, the two coefficients (for age and marital status), and the dispersion parameter (ϕ).

Suppose the software returns the following:

Parameter	*coefficient*
Intercept (β_0):	5.8
Coefficient for driver age (β_1):	0.1
Coefficient for marital status (β_2):	–0.15
Dispersion parameter (ϕ):	0.3

We then wish to use this information to predict average claim severity for a 25-year-old married driver. We use Equation 2, plugging in the following values: $\beta_0 = 5.8$, $\beta_1 = 0.1$, $\beta_2 = -0.15$, $x_1 = 25$, and $x_2 = 1$. We solve for μ_i, which represents average claim severity for this driver as indicated by the model. Per Equation 2,

$$g(\mu_i) = \ln \mu_i = 5.8 + (0.1)25 + (-0.15)1 = 8.15 \rightarrow \mu_i = \mathbf{3,463.38}$$

Thus, the model predicts the loss amount for a claim from this class of driver to follow a gamma distribution with parameters μ = 3,463.38 and ϕ = 0.3. The value 3,463.38 is the mean, or the expected severity for this driver; that figure may then be multiplied by an estimate of frequency to derive an expected pure premium which would underlie the rate charged for that class of driver.

Equivalently, the model prediction can be represented as a series of multiplicative rating factors by exponentiating both sides of the equation above:

$$\mu_i = \exp\left[5.8 + (0.1)25 + (-0.15)1\right] = e^{5.8} \times e^{0.1(25)} \times e^{-.15(1)}$$
$$= 330.30 \times 12.182 \times 0.861 = \mathbf{3,464.42}$$

which is similar to the result above. (The difference is due to rounding.)

The advantage of this last formulation is that it can be easily translated as a simple rating algorithm: begin with a "base" average severity of $330.30, and apply the factors applicable to driver age 25 and married drivers (12.182 and 0.861, respectively), to arrive at the expected severity for this particular class of driver: $3,464.

We might also use this model to predict mean severity for a 35-year-old unmarried driver; that prediction is exp[5.8 + (0.1)35 + (–0.15)0] = 10,938, meaning the loss amount follows a gamma distribution with parameters μ = 10,938 and ϕ = 0.3. Note that the ϕ parameter is the same as for the first driver in our example, since ϕ is constant for all risks in a GLM.

In this simple example, the specifications of the model—the distribution, the target variable and predictors to include—are given. In the real world, such decisions are often not straightforward. They are continually refined over many iterations of the model building process, and require a delicate balance of art and science.[1] The tools and concepts

[1] As for the link function, it is usually the case that the desirability of a multiplicative rating plan trumps all other considerations, so the log link is almost always used. One notable exception is where the target variable is binary (i.e., occurrence or non-occurrence of an event), for which a special link function must be used, as discussed later in this chapter.

Table 1. The Exponential Family Variance Functions

Distribution	Variance Function [$V(\mu)$]	Variance [$\phi V(\mu)$]
normal	1	ϕ
Poisson	μ	$\phi\mu$
gamma	μ^2	$\phi\mu^2$
inverse Gaussian	μ^3	$\phi\mu^3$
negative binomial[2]	$\mu(1+\kappa\mu)$	$\phi\mu(1+\kappa\mu)$
binomial	$\mu(1-\mu)$	$\phi\mu(1-\mu)$
Tweedie	μ^p	$\phi\mu^p$

that help guide proper model specification and selection for the purpose of building an optimal rating plan are the primary focus of this monograph.

2.2. Exponential Family Variance

The particulars of the exponential family of distributions are complex, and most are not important from the viewpoint of the practitioner and will not be covered in this monograph. [For a fuller treatment, see Clark and Thayer (2004).] However, it is necessary to understand the first two central moments of this family of distributions and how they relate to the parameters.

Mean. As noted above, the mean of every exponential family distribution is μ.

Variance. The variance is of the following form:

$$Var[y] = \phi V(\mu) \tag{3}$$

That is, the variance is equal to ϕ (the dispersion parameter) times some function of μ, denoted $V(\mu)$. The function $V(\mu)$ is called the **variance function**, and its actual definition depends on the specific distribution being used. Table 1 shows the variance functions for several of the exponential family distributions.

As shown in Table 1, for the normal distribution, the function $V(\mu)$ is a constant, and so the variance does not depend on μ. For all other distributions, however, $V(\mu)$ is a function of μ, and in most cases it is an increasing function. This is a desirable property in modeling insurance data, as we expect that higher-risk insureds (in GLM-speak, insureds with higher values of μ) would also have higher variance. Recall that a constraint of GLMs that we need to live with is that the ϕ parameter must be a

[2] Note that for the negative binomial distribution, the dispersion parameter ϕ is restricted to be 1. As such, although this table shows expressions for both the variance function and the variance (for the sake of completeness), they are in fact equivalent.

constant value for all risks. Thanks to the variance function of the exponential family, however, this doesn't mean the *variance* must be constant for all risks; our expectation of increasing variance with increasing risk can still be reflected in a GLM.

To illustrate this last point, recall our previous example, where we predicted the average severities for two drivers using the same model, with the predictions being $3,464 and $10,938. In both cases, the ϕ parameter was held constant at 0.3. Following Equation 3 and the gamma entry for $V(\mu)$ in Table 1, we can calculate the variance in loss amount for the first driver as $0.3 \times 3{,}464^2 = 3.6 \times 10^6$, while the second driver has a variance of $0.3 \times 10{,}938^2 = 35.9 \times 10^6$. Thus the higher-risk driver has a higher variance than the lower-risk driver (an intuitive assumption) despite the restriction of constant ϕ.

The third column in Table 1 reminds the reader that the variance function is *not* the variance. To get the actual variance, one must multiply the variance function by the estimated ϕ, which in effect serves to scale the variance for all risks by some constant amount.

2.3. Variable Significance

For each predictor specified in the model, the GLM software will return an estimate of its coefficient. However, it is important to recognize that those estimates are just that—estimates, and are themselves the result of a random process, since they were derived from data with random outcomes. If a different set of data were used, with all the same underlying characteristics but with different outcomes, the resulting estimated coefficients would be different.

An important question for each predictor then becomes: is the estimate of the coefficient reasonably close to the "true" coefficient? And, perhaps more importantly: does the predictor have *any* effect on the outcome at all? Or, is it the case that the predictor has no effect—that is, the "true" coefficient is zero, and the (non-zero) coefficient returned by the model-fitting procedure is merely the result of pure chance?

Standard GLM software provides several statistics for each coefficient to help answer those questions, among which are the *standard error, p-value,* and *confidence interval.*

2.3.1. Standard Error

As described above, the estimated coefficient is the result of a random process. The **standard error** is the estimated standard deviation of that random process. For example, a standard error of 0.15 assigned to a coefficient estimate may be thought of as follows: if this process—collecting a dataset of this size (with the same underlying characteristics but different outcomes) and putting it through the GLM software with the same specifications—were replicated many times, the standard deviation of the resulting estimates of the coefficient for this predictor would be approximately 0.15.

A small standard deviation indicates that the estimated coefficient is expected to be close to the "true" coefficient, giving us more confidence in the estimate. On the other hand, a large standard deviation tells us that a wide range of estimates could be achieved through randomness, making it less likely that the estimate we got is close to the true value.

Generally, larger datasets will produce estimates with smaller standard errors than smaller datasets. This is intuitive, as more data allows us to "see" patterns more clearly.

The standard error is also related to the estimated value of ϕ: the larger the estimate of ϕ, the larger the standard errors will be. This is because a larger ϕ implies more variance in the randomness of the outcomes, which creates more "noise" to obscure the "signal," resulting in larger standard errors.

2.3.2. p-value

A statistic closely related to the standard error (and indeed derived from the standard error) is the ***p*-value**. For a given coefficient estimate, the *p*-value is an estimate of the probability of a value of that magnitude (or higher) arising by pure chance.

For example, suppose a certain variable in our model yields a coefficient of 1.5 with a *p*-value of 0.0012. This indicates that, if this variable's true coefficient is zero, the probability of getting a coefficient of 1.5 or higher purely by chance is 0.0012.[3] In this case, it may be reasonable to conclude: since the odds of such a result arising by pure chance is small, it is therefore likely that the result reflects a real underlying effect—that is, the true coefficient is not zero. Such a variable is said to be **significant**.

On the other hand, if the *p*-value is, say, 0.52, it means that this variable—even if it has no effect—is much more likely to yield a coefficient of 1.5 or higher by chance; as such, we have no evidence from the model output that it has any effect at all. Note that this is not to say that we *have* evidence that it has *no* effect—it may be that the effect is actually there, but we would need a larger dataset to "see" it through our GLM.

Tests of significance are usually framed in terms of the **null hypothesis**—that is, the hypothesis that the true value of the variable in question is zero. For a *p*-value sufficiently small, we can reject the null hypothesis—that is, accept that the variable has a non-zero effect on the expected outcome. A common statistical rule of thumb is to reject the null hypothesis where the *p*-value is 0.05 or lower. However, while this value may seem small, note that it allows for a 1-in-20 chance of a variable being accepted as significant when it is not. Since in a typical insurance modeling project we are testing many variables, this threshold may be too high to protect against the possibility of spurious effects making it into the model.

2.3.3. Confidence Interval

As noted above, the *p*-value is used to guide our decision to accept or reject the hypothesis that the true coefficient is zero; if the *p*-value is sufficiently small, we reject it.

---

[3] It is perhaps worth clarifying here what is meant by "the probability of getting a coefficient of 1.5 or higher." Certainly, there is no randomness in the GLM fitting process; for any given set of data and model specifications, the GLM will produce the same result every time it is run, and so the probability of getting the coefficient of 1.5 with *this* data is 100%. However, recall that the estimates produced are random because they are derived from a dataset with random outcomes. Thus, the interpretation of the *p*-value may be stated as: *if* the true coefficient is zero—that is, the variable has no correlation with the outcome—there is a 0.0012 probability of the random outcomes in the data being realized in such a way that if the resultant dataset is entered into a GLM the estimated coefficient for this variable would be 1.5 or higher.

However, a hypothesis of zero is just one of many hypotheses that could conceivably be formulated and tested; we could just as easily hypothesize any other value and test against it, and the *p*-value would be inversely related to the degree to which the estimated coefficient differs from our hypothesized coefficient. It is then natural to ask: what *range* of values, if hypothesized, would *not* be rejected at our chosen *p*-value threshold? This range is called the **confidence interval**, and can be thought of as a reasonable range of estimates for the coefficient.

Confidence intervals are typically described by the complement of the *p*-value threshold used to compute them, expressed as a percentage. E.g., the confidence interval based on a *p*-value threshold of 0.05 is called the 95% confidence interval. SAS and other GLM software typically return the 95% confidence interval by default but provide the option to return a confidence interval for any chosen *p*-value threshold.

As an example: suppose, for a particular predictor, the GLM software returns a coefficient of 0.48, with a *p*-value of 0.00056 and a 95% confidence interval of [0.17, 0.79]. In this case, the low *p*-value indicates that the null hypothesis can be rejected. However, all values in the range 0.17 to 0.79 are sufficiently close to 0.48 such that, if set as initial hypotheses, the data would produce *p*-values of 0.05 or higher. Assuming that we are comfortable with a threshold of $p = 0.05$ for accept/reject decisions, hypotheses of values in that range would not be rejected, and so that range could be deemed to be a reasonable range of estimates.

2.4. Types of Predictor Variables

Predictor variables that go into a GLM are classified as being either *categorical* or *continuous,* and each of those types of variable is given a different treatment.

A **continuous variable** is a numeric variable that represents a measurement on a continuous scale. Examples include age, amount of insurance (in dollars), and population density.

A **categorical variable** is a variable that takes on one of two or more possible values, thereby assigning each risk to a "category." A categorical variable may be numeric or non-numeric. Examples are: vehicle primary use (one of either "commute" or "pleasure"); vehicle type (one of "sedan," "SUV," "truck," or "van"); or territory (a value from 1 to 8, representing the territory number). The distinct values that a categorical value may take on are called **levels**.

2.4.1. Treatment of Continuous Variables

The treatment of continuous variables in a GLM is straightforward: each continuous variable is input into the GLM as-is, and the GLM outputs a single coefficient for it. This results in the linear predictor holding a direct linear relationship with the value of the predictor: for each unit increase in the predictor, the linear predictor rises by the value of the coefficient (or declines, in the case of a negative coefficient). If a log link was used, this in turn results in the predicted value increasing or decreasing by some constant percentage for each unit increase in the predictor.

Logging Continuous Variables. When a log link is used, it is often appropriate to take the natural logs of continuous predictors before including them in the model, rather than placing them in the model in their original forms. This allows the scale of the predictors to match the scale of the entity they are linearly predicting, which in the case of a log link is the log of the mean of the outcome.

When a logged continuous predictor is placed in a log link model, the resulting coefficient becomes a *power transform* of the original variable. To see this mathematically, consider the simple case of a model with only an intercept term and a single continuous predictor x. Applying a log link, and logging predictor x, Equation 2 becomes:

$$\ln \mu = \beta_0 + \beta_1 \ln x$$

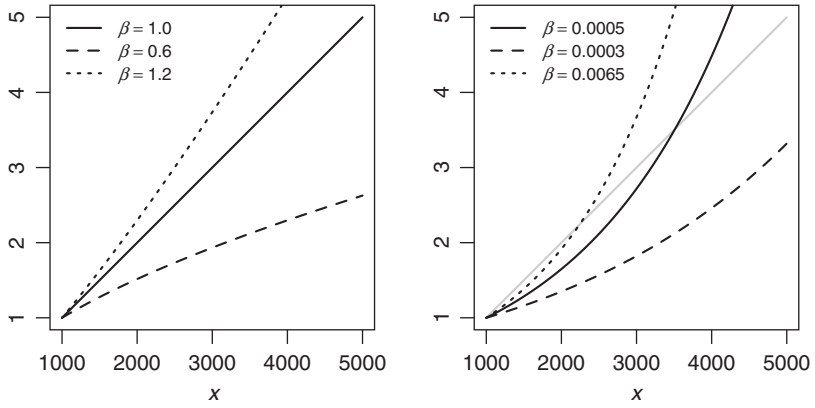
To derive μ, we exponentiate both sides:

$$\mu = e^{\beta_0} \times e^{\beta_1 \ln x} = e^{\beta_0} \times x^{\beta_1}$$

As demonstrated, when deriving the prediction, the coefficient β_1 becomes an exponent applied to the original variable x. To make this example more concrete, suppose x represents amount of insurance (AOI) in thousands of dollars; we log AOI and place it into a log link model, and the resulting coefficient is 0.62. We can use this information to derive a relativity factor for any AOI relative to a "base" AOI by raising the AOI to a power of 0.62 and dividing that by the base AOI raised to that same power. If our base AOI is \$100,000, the indicated relativity for \$200,000 of AOI is $200^{0.62}/100^{0.62} = 1.54$—in other words, a property with \$200,000 of AOI has an expected outcome 54% higher than that of a property with \$100,000 of AOI.

Including continuous predictors in their logged form allows a log link GLM flexibility in fitting the appropriate response curve. Some examples of the indicated response curves for different positive values of the coefficient are shown in the left panel of Figure 1. If the variable holds a direct linear relationship with the response, the estimated coefficient will

Figure 1. Indicated Response Curve for Logged Continuous Variable (*left*) and Unlogged Continuous Variable (*right*)



be near 1.0 (solid line). A coefficient between 0 and 1 (such as the 0.6 coefficient illustrated by the dashed line) would indicate that the mean response increases with the value of the predictor, but at a decreasing rate; this shape is often appropriate for predictors in insurance models. A coefficient greater than 1—such as 1.2, the dotted line—will yield a curve that increases at a mildly increasing rate. (Negative coefficients would yield response curves that are the "flipped" images of those illustrated here; a coefficient of −1.0 would indicate a direct inverse relationship, −0.6 would indicate a function that decreases at a decreasing rate, and the curve for −1.2 would be decreasing at an increasing rate.)

On the other hand, if the variable x is not logged, the response curve for any positive coefficient will always have the same basic shape: exponential growth, that is, increasing at an increasing rate. The right panel of Figure 1 illustrates the kinds of fits that might be produced for variables similar to those in the left panel if the variable x were not logged. As can be seen, a direct linear relationship (the gray line) is no longer an option. Only an exponential growth curve can be achieved; the magnitude of the growth varies with the coefficient. To be sure, there may be some instances where such a shape may be warranted; for example, if x is a temporal variable (such as year) meant to pick up trend effects, it may be desirable for x to yield an exponential growth relationship with the response, as trend is often modeled as an exponential function. In general, though, rather than viewing logging as a *transformation* of a continuous variable, it is often useful to consider the logged form of a variable the "natural" state of a predictor in a log link model, with the original (unlogged) variable viewed as a "transformation" that should only be used in certain specific cases.

Note that this suggestion is not due to any statistical law, but rather it is a rule of thumb specific to the context of insurance modeling, and is based on our *a priori* expectation as to the relationship between losses and the continuous predictors typically found in insurance models. For some variables, logging may not be feasible or practical. For example, variables that contain negative or zero values cannot be logged without a prior transformation. Also, for "artificial" continuous variables (such as credit scores) we may not have any *a priori* expectation as to whether the natural form or the logged form would better capture the loss response.

Also note that when including a logged continuous variable in a log link model, the underlying assumption is that the logged variable yields a linear relationship with the logged mean of the outcome. Certainly, there are many instances of predictors for which such will not be the case. An example is the effect of driver age on expected auto pure premium, which is typically at its highest for teen drivers and declines as drivers mature into their twenties and thirties, but rises again as the drivers enter their senior years. Regardless of whether the original variable has been logged or not, it is crucial to test the assumption of linearity and make adjustments where appropriate. Techniques for detecting and handling such non-linear effects will be discussed in Chapter 5.

2.4.2. Treatment of Categorical Variables

When a categorical variable is used in a GLM the treatment is a bit more involved. One of the levels is designated as the **base level**. Behind the scenes, the GLM software replaces the column in the input dataset containing the categorical variable with a

Table 2. Input Data to the GLM

freq	vtype	. . . *other predictors* . . .
0	SUV	. . .
0	truck	. . .
1	sedan	. . .
0	truck	. . .
0	van	. . .
. . .	. . .	. . .

series of indicator columns, one for each level of that variable *other than* the base level. Each of those columns takes on the values 0 or 1, with 1 indicating membership of that level. Those columns are treated as separate predictors, and each receives its own coefficient in the output. This resulting dataset is called the **design matrix**.

To illustrate: suppose, in an auto frequency model, we wish to include the categorical variable "vehicle type," which can be either "sedan," "SUV," "truck" or "van." We designate "sedan" as the base level.

Table 2 shows the target variable and vehicle type columns for the first five rows of our dataset. The vehicle type variable is named "vtype" in the data.

Prior to fitting the GLM, the software will transform the data to create indicator variables for each level of vtype other than "sedan," our base level. Table 3 shows the resulting design matrix.

Record 1, which is of vehicle type "SUV," has a 1 in the vtype:SUV column and zeros for all other columns relating to vehicle type. Similarly, record 2 has a 1 in the vtype:truck column and zeros in all the others. There is no column corresponding to vehicle type "sedan"; record 3's membership in that level is indicated by all three vehicle type columns being zero. Each of the newly-created vehicle type columns is treated as a separate predictor in Equation 2.

For a risk of any non-base level, when the values for the indicators columns are linearly combined with their respective coefficients in Equation 2, the coefficients

Table 3. Design Matrix

predictor:	freq	vtype:SUV	vtype:truck	vtype:van	. . . *other predictors* . . .
symbol:	y	x_1	x_2	x_3	$x_4 \ldots x_p$
	0	1	0	0	. . .
	0	0	1	0	. . .
	1	0	0	0	. . .
	0	0	1	0	. . .
	0	0	0	1	. . .
	. . .	. . .	. . .	. . .	. . .

relating to all other levels are multiplied by zero and drop out, while the coefficient relating to the level to which it belongs is multiplied by one and remains. For a risk of the base level, *all* the coefficients drop out. As such, the coefficient for each non-base level indicates the effect of being a member of that level *relative to* the base level.

Continuing with our example, suppose the GLM returns the estimates shown in Table 4 for the three non-base vehicle types.

To use this output to derive the linear predictor for an SUV, we plug the coefficients of Table 4 and the x predictors of Table 3 into Equation 2:

$$\begin{aligned} g(\mu) &= \beta_0 + 1.23 \times 1 + 0.57 \times 0 + (-0.30) \times 0 + \beta_4 x_4 + \cdots + \beta_p x_p \\ &= \beta_0 + 1.23 + \beta_4 x_4 + \cdots + \beta_p x_p \end{aligned}$$

As seen, all coefficients related to vehicle type for types other than "SUV" drop out of the equation, and only the coefficient for SUVs (1.23) remains. Since for a risk of vehicle type "sedan" *all* the vehicle type coefficients would drop out, the positive coefficient applied to SUVs indicates that their claim frequency is greater than that of sedans. Similarly, the negative coefficient attached to "van" indicates that claims are less frequent for vans than for sedans.

If a log link was used, a factor table for vehicle type can be constructed from this output by exponentiating each of the above coefficients. For the base level (sedan in this example) the factor is 1.000, since the effect of this vehicle type on the linear predictor is zero (and $e^0 = 1$). An SUV would get a rating factor of $e^{1.23} = 3.421$, indicating that the expected frequency for SUVs are 242% greater than that of sedans. The rating factor for a van would be $e^{-0.30} = 0.741$, indicating an expected frequency that is 25.9% lower than that of sedans.

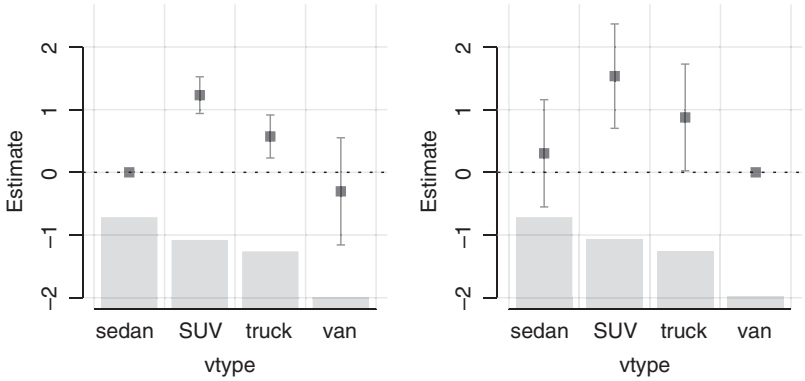
We now turn our attention to the significance statistics for this model (that is, the rightmost two columns of Table 4). These statistics help us assess our confidence in the values the parameters being non-zero. In the context of this model—where the parameters relate each level of vehicle type to the base level—a parameter of zero would mean that the level has the same mean frequency as the base level. It follows that the significance of the parameter measures the confidence that the level is significantly *different* from the base level.

The low p-value assigned to the vtype:SUV parameter indicates that the frequency for SUVs is significantly higher than that of sedans. For vans, on the other hand, the high p-value tells us that there is not enough evidence in the data to conclude that the frequency for vans is indeed lower than that of sedans.

Table 4. GLM Parameter Estimates for Vehicle Type

Parameter	Coefficient	Std. Error	p-Value
vtype:SUV (β_1)	1.23	0.149	<0.0001
vtype:truck (β_2)	0.57	0.175	0.0011
vtype:van (β_3)	–0.30	0.436	0.4871

Figure 2. Graphical representation of the parameter estimates for vehicle type, with "sedan" as the base level (*left panel*) and with "van" as the base level (*right panel*). The filled squares show the GLM estimates, and the error bars around them indicate the 95% confidence intervals around those estimates. The vertical gray bars at the bottom are proportional to the volume of data for each vehicle type.



A graphical representation of the estimates of Table 4 can be seen in the left panel of Figure 2. The filled squares show the GLM estimates, and the error bars around them indicate the 95% confidence intervals around those estimates. The vertical gray bars at the bottom are proportional to the volume of data for each vehicle type. We can see that "van," the level with the least amount of data, has the widest error bar. In general, for categorical variables, sparser levels tend to have wider standard errors, indicating less confidence in their parameter estimates, since those estimates are based on less data. The "van" error bar also crosses the zero line, indicating that this estimate is not significant at the 95% confidence level.

2.4.3. Choose Your Base Level Wisely!

In the above example, we've set the base level for vehicle type to be "sedan." Table 5 shows what the output would be had we used "van" as the base level instead.

This model is equivalent to that of Table 4 in that both would produce the same predictions. To be sure, the coefficients are different, but that is only because they are relating the levels to a different base. To see this, subtract the coefficient for "sedan"

Table 5. Parameter Estimates After Setting "van" as the Base Level

Parameter	Coefficient	Std. Error	*p*-Value
vtype:sedan (β_1)	0.30	0.436	0.4871
vtype:SUV (β_2)	1.53	0.425	0.0003
vtype:truck (β_3)	0.88	0.434	0.0440

from that of any of the other levels (using 0 for "van"), and compare the result to the corresponding coefficient on Table 4.

What *has* changed, though, are the significance statistics. Whereas for the prior model the "SUV" and "truck" estimates were highly significant, after running this model the *p*-values for both have increased, indicating less confidence in their estimates. The parameters are plotted in the right panel of Figure 2. We can see that the error bars have widened compared to the prior model.

To understand why, recall that the significance statistics for categorical variable parameters measure the confidence in any level being *different* from the base level. As such, to be confident about that relationship, we need confidence about both sides of it—the mean response of the parameter in question, as well as that of the base level. In this case, our base level has sparse data, which does not allow the model to get a good read on its mean frequency, and so we can't be certain about the relativity of any other level to it either.

As such, when using categorical variables, it is important to set the base level to be one with populous data—and not simply take the default base assigned by the software—so that our measures of significance will be most accurate.

2.5. Weights

Frequently, the dataset going into a GLM will include rows that represent the averages of the outcomes of groups of similar risks rather than the outcomes of individual risks. For example, in a claim severity dataset, one row might represent the average loss amount for several claims, all with the same values for all the predictor variables. Or, perhaps, a row in a pure premium dataset might represent the average pure premium for several exposures with the same characteristics (perhaps belonging to the same insured).

In such instances, it is intuitive that rows that represent a greater number of risks should carry more weight in the estimation of the model coefficients, as their outcome values are based on more data. GLMs accommodate that by allowing the user to include a **weight** variable, which specifies the weight given to each record in the estimation process.

The weight variable, usually denoted ω, formally works its way into the math of GLMs as a modification to the assumed variance. Recall that the exponential family variance is of the form $Var[y] = \phi V(\mu)$. When a weight variable is specified, the assumed variance for record i becomes

$$Var[y_i] = \frac{\phi V(\mu_i)}{\omega_i},$$

that is, the "regular" exponential family variance divided by the weight. The variance therefore holds an *inverse relation* to the weight.

When the weight variable is set to be the number of records that an aggregated row represents, this specification of variance neatly fits with our expectations of the variance for such aggregated records. Recall that a basic property of variances is that

for a random variable X, $Var\left[\left(\sum X_i\right)/n\right] = \frac{1}{n} Var[X]$; in other words, the variance of the average of n independent and identically distributed random variables is equal to $1/n$ times the variance of one such random variable. As such, a row representing the average loss amount of two claims would be expected to have half the variance of a single-claim row, and so on. Setting the weight to be the number of claims allows the GLM to reflect that expectation.

2.6. Offsets

When modeling for insurance rating plans, it is often the case that the scope of the project is not to update the entire plan at once; rather, some elements will be changed while others remain as-is. Some common examples:

- Rating algorithms typically begin with a base loss cost that varies by region or class, which is derived outside of the GLM-based analysis and may even be separately filed. The scope of the GLM project may be to update the rating factors only while the relative base loss costs remain static.
- When updating deductible factors, it is frequently desirable to calculate them using traditional loss elimination-based techniques, while the GLM is used for factors other than deductible. (Section 9.1 discusses this in more detail.)

In such instances, the "fixed" variable (base loss cost or deductible in the above examples) would not be assigned an estimated coefficient by the GLM. However, since it will be part of the rating plan the GLM is intended to produce, the GLM must be made aware of its existence so that the estimated coefficients for the other variables are optimal in its presence. GLMs provide the facility to do so through the use of an **offset**.

An offset is formally defined as a predictor whose coefficient is constrained to be 1. Mathematically, it is an added term to Equation 2:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \textit{offset}$$

When including an offset in a model, it is crucial that it be on the same "scale" as the linear predictor. In the case of a log link model, this requires the offset variable to be logged prior to inclusion in the model.

As an example, suppose the rating plan we intend to produce using a log-link GLM will include a factor for deductible, for which the base deductible level is $500, with the other options being $1,000 and $2,500. The deductible factors, having been separately estimated using non-GLM methods, are 0.95 for the $1,000 deductible and 0.90 for the $2,500 deductible. (The $500 deductible, being the base level, is assigned a factor of 1.00.) As we do not wish to alter these factors—but would like to ensure that the other factors estimated by the GLM are optimized for a rating plan that includes them—we include the deductible factors in the GLM as an offset.

To do so, we create a new variable, set to be ln(1.00) = 0 for those records with the base deductible of $500, ln(0.95) = –0.0513 for those records with $1,000 deductibles, and ln(0.90) = –0.1054 for records with $2,500 deductibles. That variable is set as the offset in the GLM specification.[4]

Multiple offsets can be included by simply adding them together (after first transforming them to the linear predictor scale). So, supposing we wish to offset a log-link model for both the territorial base loss cost and the deductible, a record for a risk in a territory with a base loss cost of $265 having a deductible factor of 0.90 would have its offset variable set to be ln(265) + ln(0.90) = 5.5797 + (–0.1054) = 5.4744.

Exposure Offsets. Offsets are also used when modeling a target variable that is expected to vary directly with time on risk or some other measure of exposure. An example would be where the target variable is the number of claims per policy for an auto book of business where the term lengths of the policies vary; all else equal, a policy covering two car years is expected to have twice the claims as a one-year policy. This expectation can be reflected in a log-link GLM by including the (logged) number of exposures—car years in this example—as an offset.

Note that this approach is distinct from modeling claims *frequency*, i.e., where the target variable is the number of claims divided by the number of exposures, which is the more common practice. In a frequency model, the number of exposures should be included as a weight, but not as an offset. In fact: a claim count model that includes exposure as an offset is *exactly equivalent* to a frequency model that includes exposure as a weight (but not as an offset)—that is, they will yield the same predictions, relativity factors and standard errors.[5]

To gain an intuition for this relationship, recall that an offset is an adjustment to the *mean*, while the weight is an adjustment to the *variance*. For a claim *count* model, additional exposures on a record carry the expectation of a greater number of claims, and so an offset is required. While the variance of the claim count is also expected to increase with increasing exposure—due to the exponential family's inherent expectation of greater mean implying greater variance—this is naturally handled by the GLM's assumed mean/variance relationship, and so no adjustment to variance (i.e., no weight variable) is necessary. For a claim *frequency* model, on the other hand, additional exposure carries the expectation of reduced variance (due to the larger volume of exposures yielding greater stability in the average frequency), but no change to the expected mean, and therefore a weight—but no offset—is needed.[6]

---

[4] This example is for a log-link GLM. For an example of the use of an offset in a logistic model, see CAS Exam 8, Fall 2018 Question 7. (Logistic regression is discussed later in this chapter.)

[5] Note that while this equivalence holds true for the Poisson (or overdispersed Poisson) distribution, it does not work for the negative binomial distribution since the two approaches may yield different estimates of the negative binomial parameter κ. (These distributions are discussed in the next section.)

[6] See Yan et al [2009] for a more detailed discussion of this equivalence and its derivation.

The following table summarizes this equivalence.

	Claim Count	Frequency
Target Variable	# of claims	$\frac{\#\ of\ claims}{\#\ of\ exposures}$
Distribution	Poisson	Poisson
Link	log	log
Weight	None	# of exposures
Offset	ln(# of exposures)	None

2.7. An Inventory of Distributions

The following sections describe several of the exponential family distributions available for use in a GLM, with a focus on the types of target variables typically modeled when developing rating plans: severity, frequency, pure premium and loss ratio.

2.7.1. Distributions for Severity

When modeling the severity of claims or occurrences, two commonly used distributions are the *gamma* and *inverse Gaussian* distributions.

Gamma. The gamma distribution is right-skewed, with a sharp peak and a long tail to the right, and it has a lower bound at zero. As these characteristics tend to be exhibited by empirical distributions of claim severity, the gamma is a natural fit (and indeed the most widely used distribution) for modeling severity in a GLM. The gamma variance function is $V(\mu) = \mu^2$, meaning that the assumed variance of the severity for any claim in a gamma model is proportional to an exponential function of its mean.

Figure 3 shows several examples of the gamma probability density function (pdf) curves for varying values of μ and ϕ. The two black lines illustrate gamma with $\phi = 1$, with means of 1 and 5. The two gray lines show gamma curves with those same means

Figure 3. The Gamma Distribution

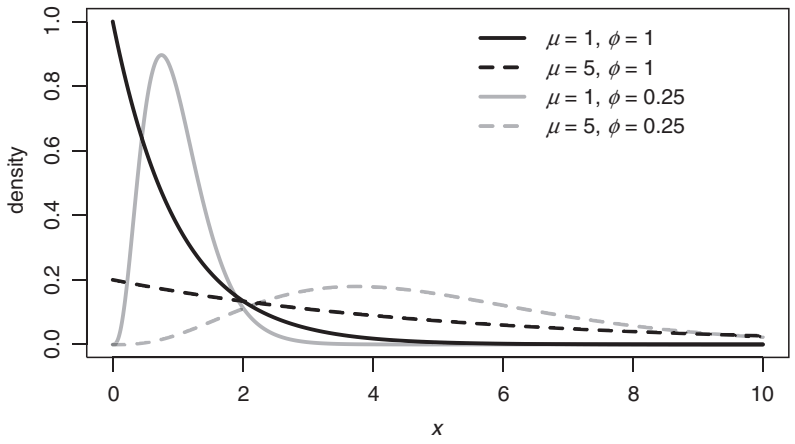
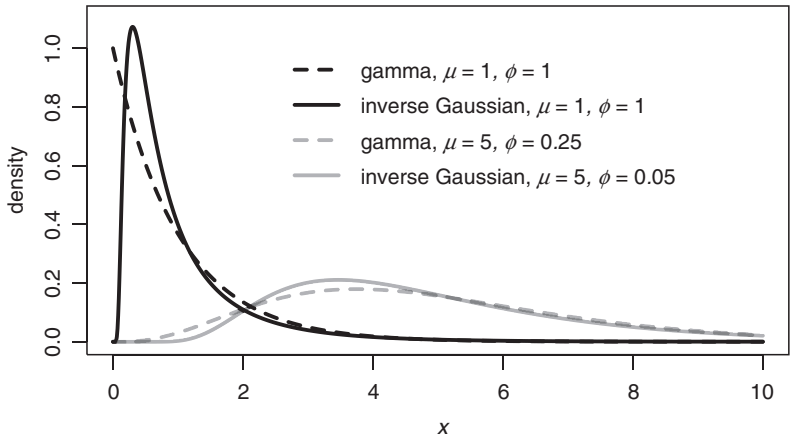


Figure 4. The Inverse Gaussian Distribution (as compared with the gamma distribution)



but with a lower value of ϕ. As you would expect, the gray lines indicate lower variance than their corresponding black lines. However, also note that the value of ϕ does not tell the full story of the variance. Comparing the two gray lines, it is clear that gamma with $\mu = 5$ (dashed line) has a much wider variance than gamma with $\mu = 1$ (solid line), despite their having the same value of ϕ. This, of course, is due to the variance function $V(\mu) = \mu^2$, which assigns higher variance to claims with higher expected means, and is a desirable characteristic when modeling severity in a GLM. We would expect claims with higher average severity to also exhibit higher variance in severity, and this property allows us to reflect that assumption in a GLM even though the dispersion parameter ϕ must be held constant for all claims.

Inverse Gaussian. The inverse Gaussian distribution, like the gamma distribution, is right-skewed with a lower bound at zero, which makes it another good choice for modeling severity. Compared to the gamma, it has a sharper peak and a wider tail, and is therefore appropriate for situations where the skewness of the severity curve is expected to be more extreme. (Later in this text we will discuss formal tests that can be applied to the data to assess the appropriateness of the various distributions.)

The variance function for the inverse Gaussian distribution is $V(\mu) = \mu^3$; like the gamma, the inverse Gaussian variance scales exponentially with the mean, but at a faster rate.

Figure 4 shows two examples of the inverse Gaussian distribution (the two solid lines) each compared to a gamma distribution with the same mean and variance (the dotted lines). As can be seen, the shapes of the inverse Gaussian distributions have sharper peaks and are more highly skewed than their gamma counterparts.[7]

[7] For the two $\mu = 5$ curves (the gray lines) in Figure 4, a gamma distribution with $\phi = 0.25$ is compared to an inverse Gaussian distribution with $\phi = 0.05$. This is so that the variance of the gamma curve ($\phi\mu^2 = 0.25 \times 5^2 = 6.25$) is equal to that of the inverse Gaussian curve ($\phi\mu^3 = 0.05 \times 5^3 = 6.25$). The intent is to demonstrate the difference in the curves that would be yielded by the two distributions *for the same data;* typically, the ϕ parameter under the inverse Gaussian distribution will be much lower than under the gamma distribution to compensate for the much larger values of $V(\mu)$ in keeping the overall assumed variance roughly constant.

2.7.2. Distributions for Frequency

When modeling claim frequency (e.g., expected claim count per unit of exposure or per dollar of premium), the most commonly used distribution is the *Poisson* distribution. Another available choice is the *negative binomial* distribution. Both are explained in the following sections.

Poisson. The Poisson distribution models the count of events occurring within a fixed time interval, and is widely used in actuarial science as a distribution for claim counts. Although the Poisson is typically a discrete distribution (defined only for integral values) its implementation in a GLM allows it to take on fractional values as well. This feature is useful when modeling claim frequency, where claim count is divided by a value such as exposure or premium, resulting in a non-integral target variable. (In such instances it is usually appropriate to set the GLM weight to be the denominator of frequency.)

The variance function for a Poisson distribution is $V(\mu) = \mu$, meaning that the variance increases *linearly* with the mean. In fact, for a "true" Poisson distribution, the variance *equals* the mean; stated in terms of the exponential family parameters, this would mean that $\phi = 1$ and so it drops out of the variance formula, leaving $Var[y] = \mu$. However, claim frequency is most often found to have variance that is greater than the mean, a phenomenon called **overdispersion**.

Overdispersion arises mainly because in addition to the natural variance arising from the Poisson process, there is another source of variance: the variation in risk level among the policyholders themselves. In statistical terms: in addition to the Poisson variance, there is variance in the Poisson mean (μ) among risks. To be sure, determining the appropriate mean, and thereby separating the good risks from bad risks, is precisely the purpose of our modeling exercise. However, our model will not be perfect; there will always be some variation in risk level among policyholders not explained by our model's predictors, and so the data will exhibit overdispersion.

One way to deal with this scenario is to use the **overdispersed Poisson** (ODP) distribution in place of the "true" Poisson. The overdispersed Poisson is similar to the Poisson distribution, with the main difference being the ϕ parameter: ODP allows it to take on any positive value rather than being restricted to 1 as is the case with the true Poisson.

When modeling claims frequency with the Poisson distribution, it is recommended that the overdispersed Poisson be used; otherwise, the variance will likely be understated, thereby distorting the model diagnostic measures such as standard error and p-value. (Note that the Poisson and ODP distributions will always produce the same estimates of coefficients, and therefore the same predictions; it is only the model diagnostics that will be affected.)

Negative Binomial. Another way of dealing with the overdispersion in the Poisson distribution resulting from random variation in the Poisson mean among risks is to treat the Poisson mean for any given risk as a random variable itself. Doing so, we would need another probability distribution to model the Poisson mean; a good

choice for that might be the gamma distribution. Such a setup would be stated mathematically as follows:

$$y \sim Poisson(\mu = \theta), \quad \theta \sim gamma(\ldots). \tag{4}$$

In words, the outcome (y) is Poisson-distributed with a mean of θ, where θ is itself random and gamma-distributed. This combination results in y following a **negative binomial** distribution.

For the negative binomial distribution, the standard exponential family dispersion parameter, ϕ, is restricted to be 1. However, this distribution includes a third parameter, κ, called the **overdispersion parameter**, which is related to the variance of the gamma distribution of Equation 4.

The negative binomial variance function is

$$V(\mu) = \mu(1 + \kappa\mu)$$

and so the κ parameter serves to "inflate" the variance over and above the mean, which would be the variance implied by the Poisson distribution. Indeed, as κ approaches zero, the negative binomial distribution approaches Poisson. (Note that for the negative binomial distribution, ϕ, restricted to be 1, drops out of the variance formula and thus the variance function $V(\mu)$ is the full expression of the variance.)

2.7.3. A Distribution for Pure Premium: the Tweedie Distribution

Modeling pure premium (or loss ratio) at the policy level has traditionally been challenging. To see why, consider the properties these measures exhibit, which would need to be approximated by the probability distribution used to describe them: they are most often zero, as most policies incur no loss; where they do incur a loss, the distribution of losses tends to be highly skewed. As such, the pdf would need to have most of its mass at zero, and the remaining mass skewed to the right. Fortunately, a rather remarkable distribution that can capture these properties does exist: the **Tweedie** distribution.

In addition to the standard exponential family parameters μ and ϕ, the Tweedie distribution introduces a third parameter, p, called the **power** parameter. p can take on any real number except those in the interval 0 to 1 (non-inclusive: 0 and 1 themselves are valid values). The variance function for Tweedie is $V(\mu) = \mu^p$.

One rather interesting characteristic of the Tweedie distribution is that several of the other exponential family distributions are in fact special cases of Tweedie, dependent on the value of p:

- A Tweedie with $p = 0$ is a normal distribution.
- A Tweedie with $p = 1$ is a Poisson distribution.
- A Tweedie with $p = 2$ is a gamma distribution.
- A Tweedie with $p = 3$ is an inverse Gaussian distribution.

Going further, thanks to the Tweedie distribution, our choices in modeling claim severity are not restricted to the moderately-skewed gamma distribution and

the extreme skewness of the inverse Gaussian. The Tweedie provides a *continuum* of distributions between those two by simply setting the value of p to be between 2 (gamma) and 3 (inverse Gaussian).

However, the area of the p parameter space we are most interested in is between 1 and 2. At the two ends of that range are Poisson, which is a good distribution for modeling frequency, and gamma, which is good for modeling severity. Between 1 and 2, Tweedie becomes a neat combination of Poisson and gamma, which is great for modeling pure premium or loss ratio—that is, the combined effects of frequency and severity. (For the remainder of this text, references to the Tweedie distribution refer to the specific case of a Tweedie where p is in the range [1,2].)

A Poisson Sum of Gammas. The Tweedie distribution models a "compound Poisson-gamma process." Where events (such as claims) occur following a Poisson process, and each event generates a random loss amount that follows a gamma distribution, the total loss amount for all events follows the Tweedie distribution. In this way the Tweedie distribution may be thought of as a "Poisson-distributed sum of gamma distributions."

In fact, the Tweedie parameters (μ, ϕ and p) bear a direct relationship to those of the underlying Poisson and gamma distributions; we will examine that more closely here.

Poisson has a single parameter, typically denoted λ, which is both the mean and the variance. (In prior sections we've referred to the Poisson mean by the symbol μ, following the Poisson's exponential family form. For this section, we'll use the "traditional" parameterizations of the underlying distributions, saving the symbol μ for the Tweedie mean.)

The gamma distribution takes two parameters: the shape and scale parameters, usually denoted α and θ, respectively. The mean is

$$E[y] = \alpha \cdot \theta, \tag{5}$$

and the coefficient of variation is

$$CV = 1/\sqrt{\alpha}\,. \tag{6}$$

The Tweedie mean can be related to those parameters as follows:

$$E[y] = \mu = \lambda \cdot \alpha \cdot \theta. \tag{7}$$

Notice that this is the product of the Poisson mean (λ) and the gamma mean ($\alpha \cdot \theta$), as we would expect—pure premium equals expected frequency times expected severity.

The power parameter (p) is

$$p = \frac{\alpha + 2}{\alpha + 1}. \tag{8}$$

As seen in Equation 8, the power parameter is purely a function of the gamma parameter α. Since α is itself a function of the gamma coefficient of variation (as can be seen by rearranging Equation 6 above), it follows that the p parameter is a function

of the gamma CV. Specifically, as the gamma CV approaches zero, p approaches 1; as the gamma CV gets arbitrarily large, p approaches 2. Values of p used in insurance modeling typically range between 1.5 and 1.8.

The left panel of Figure 5 shows an example of a Tweedie density function where $p = 1.02$. A value of p so close to 1 implies very little variance in the gamma (or severity) component, and so the randomness of the outcome is mainly driven by the random count of events (or, the frequency component). As such, the shape of the distribution resembles a Poisson distribution, with spikes at discrete points, but with a small amount of variation around each point. Also note that the distribution features a point mass at 0, which allows for the (likely) possibility of no claims.

The right panel of Figure 5 illustrates a Tweedie pdf for the more realistic case of $p = 1.67$. In this example, the gamma variation is considerably larger and therefore the discrete Poisson points are no longer visible. However, the distribution still assigns a significant probability to an outcome of 0.

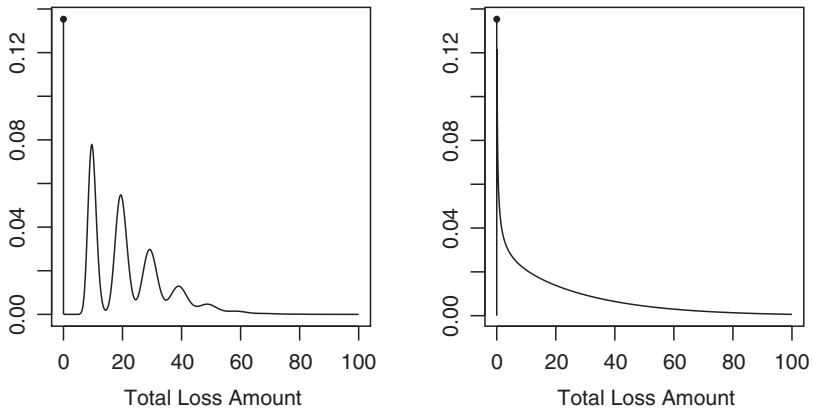
The formula for the Tweedie dispersion parameter (ϕ) is

$$\phi = \frac{\lambda^{1-p} \cdot (\alpha\theta)^{2-p}}{2-p}. \tag{9}$$

Through equations 7, 8, and 9, the Tweedie parameters can be derived from any combination of the Poisson parameter (λ) and gamma parameters (α and θ)—and vice versa, with some algebraic manipulation.

In a Tweedie GLM, the μ parameter varies by record, controlled by the linear predictor, while the ϕ and p parameters are set to be constant for all records. One important implication of this is that a Tweedie GLM contains the implicit assumption that frequency and severity "move in the same direction"—that is, where a predictor drives an increase in the target variable (pure premium or loss ratio), that increase is made up of an increase in both its frequency and severity components. (To see this, try the following exercise: begin with any set of μ, ϕ and p, and solve for λ, α and θ;

Figure 5. The Tweedie Distribution, with $p = 1.02$ (*left*) and $p = 1.67$ (*right*)



then, try increasing μ while holding ϕ and p constant. Both λ and the product $\alpha\theta$ will move upward.) This assumption doesn't always hold true, as often times variables in a model may have a positive effect on frequency while negatively affecting severity, or vice versa. However, Tweedie GLMs can be quite robust against such violations of its assumptions and still produce very strong models.

Determination of the *p* parameter. There are several ways the Tweedie p parameter may be determined:

- Some model-fitting software packages provide the functionality to estimate p as part of the model-fitting process. (Note that using this option may increase the computation time considerably, particularly for larger datasets.)
- Several candidate values of p can be considered and tested with the goal of optimizing a statistical measure such as log-likelihood (discussed in Chapter 6) or using cross-validation (discussed in Chapter 4).
- Alternatively, many modelers simply judgmentally select some value that makes sense (common choices being 1.6, 1.67 or 1.7). This may be the most practical in many scenarios, as the fine-tuning of p is unlikely to have a very material effect on the model estimates.

2.8. Logistic Regression

For some models, the target variable we wish to predict is not a numeric value, but rather the occurrence or non-occurrence of an event. Such variables are called *dichotomous* or *binary* variables. Examples are:

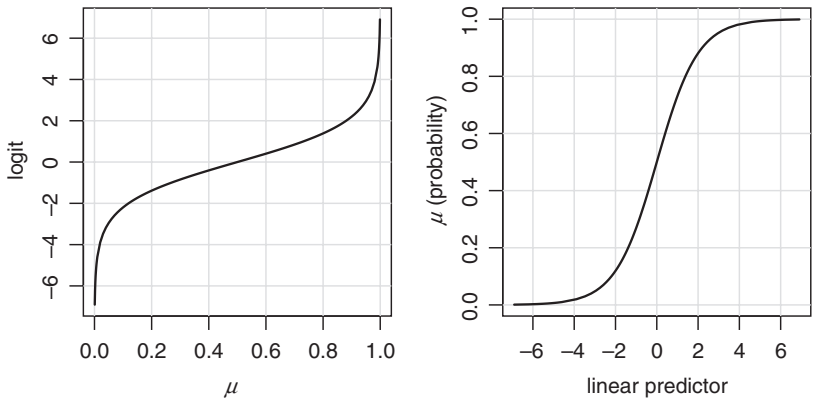
- Whether or not a policyholder will renew their policy.
- Whether a newly-opened claim will wind up exceeding some specified loss amount threshold.
- Whether a potential subrogation opportunity for a claim will be realized.

Such a model would be built based on a dataset of historical records of similar scenarios for which the outcome is currently known. The target variable, y_i, takes on the value of either 0 or 1, where 1 indicates that the event in question did occur, and 0 indicates that it did not.

Distribution. To model such a scenario in a GLM, the distribution of the target variables is set to be the binomial distribution. The mean of the binomial distribution—that is, the prediction generated by the model—is the *probability* that the event will occur.

Link Function. When modeling a dichotomous variable using the binomial distribution, a special type of link function must be used. Why not just use the log link? That's because a basic property of GLMs is that the linear predictor—that is, the right-hand side of Equation 2—is unbounded, and can take on any value in the range $[-\infty,+\infty]$. The mean of the binomial distribution, on the other hand, being a measure of probability, must be in the range [0,1]. As such, we will need a link function that can map a [0,1]-ranged value to be unbounded.

Figure 6. The Logit Function (*left*) and Its Inverse, the Logistic Function (*right*)



There are several link function that are available for this purpose, but the most common is the **logit** link function,[8] defined as follows:

$$g(\mu) = \ln \frac{\mu}{1-\mu}. \tag{10}$$

The left panel of Figure 6 shows a graph of the logit function. As can be seen, the logit approaches $-\infty$ as μ approaches zero, and becomes arbitrarily large as μ approaches 1.

The right-hand side of Figure 6 shows the inverse of the logit function, called the **logistic** function, defined as $1/(1 + e^{-x})$. In a GLM, this function translates the value of the linear predictor onto the prediction of probability. A large negative linear predictor would indicate a low probability of occurrence, and a large positive linear predictor would indicate a high probability; a linear predictor of zero would indicate that the probability is 50%.

The full specification of a logistic regression model can be summarized as follows:

$$y_i \sim binomial(\mu_i) \tag{11}$$

$$\ln \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{12}$$

Interpreting Results of a Logistic Model. The logit function of Equation 10 can be interpreted as the log of the *odds*, where the odds is defined as the ratio of the probability of occurrence to the probability of non-occurrence, or $\frac{\mu}{1-\mu}$. The odds is an alternate means of describing probability, which, unlike probability—which must lie in the region [0,1]—is unbounded in the positive direction. (Think of a near-certain event, which might be said to have "million-to-one" odds.)

[8] Others are the *probit* link and *complementary log-log* link, not covered in this text.

Exponentiating both sides of Equation 12, the logistic GLM equation becomes a multiplicative series of terms that produces the odds of occurrence. This leads to a natural interpretation of the coefficients of the GLM (after exponentiating) as describing the effect of the predictor variables on the odds. For example, a coefficient of 0.24 estimated for continuous predictor x indicates that a unit increase in x increases the odds by $e^{0.24} - 1 = 27\%$. A coefficient of 0.24 estimated for a given level of a categorical variable indicates that the odds for that level is 27% higher than that of the base level.

2.9. Correlation Among Predictors, Multicollinearity and Aliasing

Frequently, the predictors going into a GLM will exhibit correlation among them. Where such correlation is moderate, the GLM can handle that just fine. In fact, determining accurate estimates of relativities in the presence of correlated rating variables is a primary strength of GLMs versus univariate analyses; unlike univariate methods, the GLM will be able to sort out each variable's unique effect on the outcome, as distinct from the effect of any other variable that may correlate with it, thereby ensuring that no information is double-counted.

As such, before embarking on a GLM modeling project, it is important to understand the correlation structure among the predictors. This will aid in interpreting the GLM output—particularly in understanding significant deviations between the GLM indications versus what would be indicated by a series of univariate analyses of individual predictors.

Where the correlation between any two predictors is very large, however, the GLM may run into trouble. The high correlation means that much of the same information is entering the model twice. The GLM—forced not to double-count—will need to apportion the response effect between the two variables, and how precisely best to do so becomes a source of great uncertainty. As such, coefficients may behave erratically; it is not uncommon to see extremely high or low coefficients result in such scenarios. Furthermore, the standard errors associated with those coefficients will be large, and small perturbations in the data may swing the coefficient estimates wildly. Such a model is said to be *unstable.*

Such instability in a model should be avoided. As such it is important to look out for instances of high correlation prior to modeling, by examining two-way correlation tables. Where high correlation is detected, means of dealing with this include the following.

- For any group of correlated predictors, remove all but one from the model. While this is certainly the simplest approach, a potential downside is that there may be some unique information, distinct from the common information, contained in individual predictors that will not be considered in our modeling process.
- Pre-process the data using dimensionality-reduction techniques such as *principal components analysis* (PCA) or *factor analysis.* These methods create multiple new

variables from correlated groups of predictors. Those new variables exhibit little or no correlation between them—thereby making them much more useful in a GLM—and they may be representative of the different components of underlying information making up the original variables. The details of such techniques are beyond the scope of this paper.

Multicollinearity. Simple correlation between pairs of predictors are easy enough to detect using a correlation matrix. A more subtle potential problem may exist where two or more predictors in a model may be strongly predictive of a third, a situation known as **multicollinearity.** The same instability problems as above may result, since the information contained in the third variable is also present in the model in the form of the *combination* of the other two variables. However, the variable may not be highly correlated with either of the other two predictors *individually,* and so this effect will not show up in a correlation matrix, making it more difficult to detect.

A useful statistic for detecting multicollinearity is the **variance inflation factor** (VIF), which can be output by most statistical packages. The VIF for any predictor is a measure of how much the (squared) standard error for the predictor is increased due to the presence of collinearity with other predictors. It is determined for each predictor by running a linear model setting the predictor as the target and using all the *other* predictors as inputs, and measuring the predictive power of that model.

A common statistical rule of thumb is that a VIF greater than 10 is considered high. However, where large VIFs are indicated, it is important to look deeper into the collinearity structure in order to make an informed decision about how best to handle it in the model.

Aliasing. Where two predictors are *perfectly* correlated, they are said to be **aliased**, and the GLM will not have a unique solution. Most GLM fitting software will detect that and automatically drop one of those predictors from the model. Where they are *nearly* perfectly correlated, on the other hand, the software may not catch it and try to run the model anyway. Due to the extreme correlation, the model will be highly unstable; the fitting procedure may fail to converge, and even if the model run is successful the estimated coefficients will be nonsensical. Such problems can be avoided by looking out for and properly handling correlations among predictors, as discussed above.

2.10. Limitations of GLMs

This section discusses two important limitations inherent in GLMs that one should bear in mind when using them to construct rating plans.

1. GLMs Assign Full Credibility to the Data. The estimates produced by the GLM are fit under the assumption that the data are fully credible for every parameter. For any categorical variable in the model, the estimate of the coefficient for each level is the one which fits the training data best, with no consideration given to the thinness of the data on which it is based.

To gain an intuition for what this means in a practical sense, consider the following simple example. Suppose we run a GLM to estimate auto severity, and the GLM includes only one predictor: territory, a categorical variable with five levels, coded A through E. Volume of data varies greatly by territory, and the smallest territory, E, has only 8 claims.

After running this model, the prediction for each risk will simply be the overall average severity for its territory.

That's right. For a GLM with a single categorical variable as its only predictor, it actually makes no difference which distribution or link function is chosen, just so long as the GLM fitting process is able to converge. The answers will always be the same, and they will be the one-way averages of the target variable by levels of the categorical variable. (Of course, we would not need a GLM for this; we could get to the same place with a simple Excel worksheet.)

Now, continuing with our example, the indicated relativity for territory E, like the rest, will be based simply on the average severity for its 8 claims. As actuaries, if we had been using the one-way analysis to derive relativities, we would surely not select the raw indication for a territory with such little credibility with no modification; we would apply a credibility procedure, and, in absence of any additional information about the territory, probably select something closer to the statewide average. It stands to reason that for the GLM we should not just take the indicated relativity either.

To be sure, in such a scenario, the standard error for the territory E coefficient would be large, and its *p*-value high. In this way, the GLM *warns* you that the estimate is not fully credible—but does nothing about it.

Where multiple predictors or continuous variables are involved, the estimates are based on a more complicated procedure which could not be easily performed in Excel, and the answers would vary based on the chosen link function and distribution. However, the approach to deriving the estimates would similarly be one that gives full weight to the data at each level of each categorical variable.

Incorporating credibility into the GLM framework is generally beyond the scope of this monograph. However, Chapter 10 briefly discusses two extensions to the GLM that allow for credibility-like estimation methods: *generalized linear mixed models* (*GLMMs*) and *elastic net* GLMs.

2. GLMs Assume the Randomness of Outcomes is Uncorrelated. Another important assumption built into GLMs is that the random component of the outcome of the target variable is uncorrelated among the records in the training set. Note the qualification "random component" in that sentence—that's not the same thing as saying the outcomes are uncorrelated. If our auto severity model contains driver age and territory as predictors, we expect that drivers of similar ages or in the same territory would have similar outcomes, and thus be correlated in that way. After all, identifying and capturing such correlations is precisely the point of our modeling exercise. However, the assumption is that the *random* component of the outcome—which, from our vantage point, means the portion of the outcome driven by causes not in our model—are independent.

This assumption may be violated if there exist groups of records that are likely to have similar outcomes, perhaps due to some latent variable not captured by our model. The following are examples of where this may arise in insurance models:

- Frequently, the dataset going into an insurance GLM will comprise several years of policy data. Thus, there will be many instances where distinct records will actually be multiple renewals of the same policy. Those records are likely to have correlated outcomes; after all, a policyholder who is a bad driver in year 1 will likely still be a bad driver in years 2, 3 and 4.
- When modeling a line that includes a wind peril, policyholders in the same area will likely have similar outcomes, as the losses tend to be driven by storms that affect multiple insureds in the area at once.

Where the correlation is small, this is usually nothing to worry about; GLMs are quite robust against minor violations of their assumptions. However, it is important to be wary of instances of large correlation. Since the parameter estimates and significance statistics of a GLM are all derived "as if" all the random outcomes were independent, large instances of groups of correlated outcomes would cause the GLM to give undue weight to those events—essentially, picking up too much random noise—and produce sub-optimal predictions and over-optimistic measures of statistical significance.

There are several extensions to the GLM that allow one to account for such correlation in the data. One such method is the generalized linear mixed model (GLMM), briefly discussed in Section 10.1. Another is generalized estimating equations (GEE), not covered in this text.

3. The Model-Building Process

The prior chapter has covered the technical details of model construction. While this is a very important component of the model building process, it is important to understand all of the steps involved in the construction and evaluation of a predictive model. While each project has different objectives and considerations, any predictive modeling project should include the following components:

- Setting objectives and goals
- Communicating with key stakeholders
- Collecting and processing the necessary data for the analysis
- Conducting exploratory data analysis
- Specifying the form of the predictive model
- Evaluating the model output
- Validating the model
- Translating the model results into a product
- Maintaining the model
- Rebuilding the model

3.1. Setting Objectives and Goals

Before collecting any data or building any models, it is important to develop a clear understanding and to gain alignment on the scope and goals of the project. Important questions to ask include:

- What are the goals of the analysis? While the examples in this text focus on the construction of a rating plan, the goal of an analysis may be to develop a set of underwriting criteria or to determine the probability of a customer renewing a policy.
- Given the goals of the project, what is the appropriate data to collect? Is this data readily available, or will it be costly and time-consuming to obtain it?
- What is the time frame for completing the project?
- What are the key risks that may arise during the project, and how can these risks be mitigated?
- Who will work on the project, and do those analysts have the knowledge and expertise to complete the project in the desired timeframe?

3.2. Communicating with Key Stakeholders

One of the most common reasons for a project failing or falling significantly behind schedule is lack of alignment on the goals and outcomes of the project among its key stakeholders. Using the example of a rating plan, the modeler isn't just creating a predictive model, but rather constructing a new product that will (hopefully) enter the market. For this project, key stakeholders may include:

- Regulators: The goal of any predictive modeling project is to include all variables that are predictive and add lift to the model. However, many variables are considered off limits in pricing insurance risk, either due to legal and regulatory considerations or potential reputational risk. It is important to understand these limitations. These restrictions may also vary by state, as insurance is regulated at the state level.
- IT: The model results will likely need to be coded into a new rating system, and IT systems generally have limitations. Before and during model construction, it is important to communicate the desired rating structure to the programmers who will be coding the rating changes. Some components of the desired rating plan may not be feasible from an IT perspective, in which case it is important to be aware of those limitations early on and adjust the models accordingly. Furthermore, programming changes into IT systems has a cost, and so budget and availability of resources may further limit the rating plan that can be implemented.
- Agents/underwriters: Once the models are complete and turned into a product, someone will have to sell that product. If the new rating structure isn't understood by the policy producers, then it may be difficult to meet sales goals. By including agents in the discussion, the final product can better reflect their needs and concerns, which may in turn lead to a better business outcome.

3.3. Collecting and Processing Data

Collecting and processing data is often the most time-consuming component of a predictive modeling project, and modelers tend to underestimate the amount of time that will be required for this step. Most data is messy, so time must be spent figuring out how to clean the data, impute missing values, merge additional variables into the dataset, etc. Collecting and processing data are often iterative processes, as a modeler may discover later in the model-building process that a particular variable in the dataset is incorrect.

The data should also be split into at least two subsets, so that the model can be tested on data that was not used to build it. A strategy for validating the model should also be carefully formulated at this stage.

Chapter 4 discusses the process of collecting and preparing the data in greater detail.

3.4. Conducting Exploratory Data Analysis

Once the data has been collected, it is important to spend some time on exploratory data analysis (EDA) before beginning to construct models. EDA will help the modeler

better understand the nature of the data and the relationships between the target and explanatory variables. Helpful EDA plots include:

- Plotting each response variable versus the target variable to see what (if any) relationship exists. For continuous variables, such plots may help inform decisions on variable transformations.
- Plotting continuous response variables versus each other, to see the correlation between them.

3.5. Specifying Model Form

Key questions in specifying the model form include:

- What type of predictive model works best for this project and this data? While this text is focused on generalized linear models, other modeling frameworks (e.g., decision trees) may be more appropriate for some projects.
- What is the target variable, and which response variables should be included?
- Should transformations be applied to the target variable or to any of the response variables?
- Which link function should be used?

Chapter 5 further explores considerations related to the specification of the model form for GLMs.

3.6. Evaluating Model Output

Once there are preliminary results, the modeler should begin evaluating the output to determine next steps. Model evaluation involves:

- Assessing the overall fit of the model, and identifying areas in which the model fit can be improved.
- Analyzing the significance of each predictor variable, and removing or transforming variables accordingly.
- Comparing the lift of a newly constructed model over the existing model or rating structure.

These steps are detailed in Chapters 6 and 7.

3.7. Validating the Model

Model validation is a very important component of the modeling process, and should not be overlooked or rushed. The validation process is discussed in detail in Chapter 7.

3.8. Translating the Model into a Product

The ultimate goal of most modeling projects is to turn the final model into a product of some kind. In the insurance industry, this product is often a rating plan. Important considerations when turning the results of a modeling project into a final product include:

- Is the product clear and understandable? In particular, there should be no ambiguity in the risk classification, and a knowledgeable person should be able to clearly understand the structure of the product.

- Are there items included in the product that were not included in the model? Using the example of a rating plan, there are often rating factors included in the plan that are not part of the model because there is no data available on that variable. In such cases, it is important to understand the potential relationship between this variable and other variables that were included in the model. For example, if an insurer is offering a discount for safe driving behavior for the first time, this discount may overlap with other variables that were in the model. In such cases, it may be appropriate to apply judgmental adjustments to the variables in the rating plan.

3.9. Maintaining and Rebuilding the Model

The predictive accuracy of any model generally decreases over time, as the world changes and the data used to construct the model becomes less relevant. It is important to have a plan to maintain a model over time so that it does not become obsolete. Models should be periodically rebuilt in order to maximize their predictive accuracy, but in the interim it may be beneficial to refresh the existing model using newer data. This will allow model predictions to reflect the most recent experience.

4. Data Preparation and Considerations

Data preparation is one of the most important parts of the model-building process, and is usually the part of the process that takes the most time. Although every organization has different processes and systems for collecting, storing, and retrieving the data needed to build a classification plan, there are some common themes and situations with which all actuaries should be familiar.

It's important to remember that like the rest of the modeling process, the data preparation step is iterative. Correcting one error might help you discover another, and insights gleaned from the model-building process might prompt you to step back and revisit your approach to data preparation.

4.1. Combining Policy and Claim Data

In almost every case, the data most appropriate for use in building a classification plan is exposure-level premium (policy demographic) and loss (claim) data. Ideally, you would like to have a dataset with one record for each risk and each time period of interest. For some lines of business, it may suffice to attach claims to policy records and model at the policy level. For other lines, it may be beneficial to model at the level of individual risks within a policy. For example, when modeling for personal auto, claims should ideally be attached to the specific vehicles and drivers to which they pertain so that their characteristics can be included in the model as well.

The immediate difficulty with assembling such a dataset is that *premium and loss data tend not to be stored in the same place.* In many organizations, a policy-level premium database is housed within the underwriting area, and a claims database is housed within the claims area. In the normal course of business these two databases may never be matched against each other except at a very high level. So the first task of a modeling assignment is often to locate these two datasets and merge them.

If best practices have been followed and changes to these two datasets have tracked each other over time, merging them may not be time-consuming—it may even be trivial. But when dealing with legacy systems, or with policy and claims databases that have grown or evolved independently over time, problems may arise. The number of things that can go wrong is essentially unlimited. But here are some questions that the actuary may need to ask while in the process of doing a merge:

Are there timing considerations with respect to the way these databases are updated that might render some of the data unusable? If the policy database is updated at the end of every month and the claims database is updated daily, for example, the most recent claims data might not be usable because corresponding exposures are not available.

Is there a unique key that can be used to match the two databases to each other in such a way that each claim record has exactly one matching policy record? The answer to this question should always be "yes." If there are multiple policy records that match a single claim, merging may cause claims to be double counted. On the other hand, if the key does not match each claim to a policy record, some claim records may be orphaned.

What level of detail should the datasets be aggregated to before merging? This is a question whose answer is informed by both the goal of the model and practical considerations around resource limitations and run times. Data must often be aggregated across multiple dimensions. For the dimension of time, policy data is most often aggregated to the level of calendar year rather than any shorter period. Calendar-year data has several distinct advantages, among them that the calendar year is the usual policy period and that seasonality need not be addressed. When policy data is aggregated in this way, care must be taken to correctly count the exposures attributable to each record and store these exposure counts on the aggregated record. For example, a policy that is issued October 1 of a certain calendar year only contributes 25% of a full exposure to that year.

Claim data is usually also aggregated to policy and calendar year. If a particular policy has two $500 claims in a certain calendar year, the aggregated claim record would have only a claim count field with a value of 2 and a loss field with a value of $1000. Note that this treatment is not precise and that meaningful data is lost in the aggregation—in this example, the aggregated claim record could have also represented one claim of $900 and one claim of $100.

Depending on the goals of the model, further aggregation may be warranted. For example, in a book of small commercial property exposures, policies may be written at the level of the business entity, but demographic and loss data may be available by location. So a policy covering a business with two locations for one year may be aggregated to the business level (one exposure) or to the location level (two exposures). It usually makes sense to keep a finer level of detail in the model so that this information can be available to use for pricing, but if there are few enough businesses in the book with multiple locations, it may be more convenient to aggregate to the business level at the start of the project, retaining information on locations only in the form of a count.

Are there fields that can be safely discarded? There may be fields in either database which for whatever reason it would not make sense to consider in the model. Removal of these fields will speed up every other part of the model-building process.

But removal of fields is not something that should be done lightly, since costs to re-add them may be high if it's found that they're needed later in the process. A special case is when two fields contain identical or near-identical information, resulting in aliasing or near-aliasing. As discussed in Section 2.9, if you add both of these fields to your model, it will break; and, in any case, there is no reason to preserve a field that contains no new information.

Are there fields that should be in the database but aren't? There may be policyholder data that may be predictive of future loss that is collected at the underwriting step but not stored for later use. And there may be predictive data that is not collected at all. This goes beyond just the data preparation step of the process, but the actuary should be just as cognizant of what fields may be missing as they are of the fields that are currently available for use. The actuary's feedback to management on this issue may be critical to kickstarting the process of collecting new data and successfully evolving the classification plan over time.

4.2. Modifying the Data

Any dataset of sufficient size is likely to have errors. It's impossible to present a formulaic approach to error detection that will catch every possible error, and so human judgment is critical. But there are a few steps that should always be taken to attempt to catch and remedy some of the more common errors that can occur.

Check for duplicate records. If there are any records that are exactly identical, this likely represents an error of some sort. This check should be done prior to aggregation and combination of policy and claim data.

Cross-check categorical fields against available documentation. If database documentation indicates that a roof can be of type A, B, or C, but there are records where the roof type is coded as D, this must be investigated. Are these transcription errors, or is the documentation out of date?

Check numerical fields for unreasonable values. For every numerical field, there are ranges of values that can safely be dismissed as unreasonable, and ranges that might require further investigation. A record detailing an auto policy covering a truck with an original cost (new) of $30 can safely be called an error. But if that original cost is $5,000, investigation may be needed.

Decide how to handle each error or missing value that is discovered. The solution to duplicate records is easy—delete the duplicates. But fields with unreasonable or impossible values that cannot be corrected may be more difficult to handle. In a large enough dataset, deletion of every record that has an error might leave you with very few records from which to build a model. And, even worse, there might be something systematic about the presence of the error itself. For example, policies written out of a certain office may be consistently miscoded, while policies written out of other offices

aren't. In this case, deleting the offending records may leave you with no way to detect that this office also has less-skilled underwriters for a certain type of policy. A better solution is to replace erroneous or missing values with the mean or modal field value (to be used as the base level of your model), and add a new field for an error flag. The error flag can be included in the model and will proxy for the presence of the error.

Another means of handling missing or erroneous values in the data is to *impute* values for those predictors using information contained in the other predictors. This would involve building a second model, trained on the subset of data that is non-problematic, with the problem predictor as the target and all the *other* predictors as predictors.

Errors are not the only reason to modify your data. It may be appropriate to convert a continuous variable into a categorical variable (this is called "binning"), to reduce the number of levels in a categorical variable, to combine separate fields into new fields, or to separate a single field into multiple fields. But usually these sorts of modifications are made as a part of the model building process. Some of these modifications are covered in more detail in Chapter 5.

4.3. Splitting the Data

Before embarking on a modeling project, it is essential that the available data be split into at least two groups. One of those groups is called the **training set**. This is used to perform all the model-building steps—selecting the variables, determining the appropriate variable transformations, choosing the distribution, and so on. Another group of data, called the **test set** (or **holdout set**), will be used to assess the performance of the model and may also be used to choose among several candidate models.

Why do we do this? One reason is because attempting to test the performance of any model on the same set of data on which the model was built will produce over-optimistic results. After all, the model-fitting process optimizes the parameters to best fit the data used to train it, so we would expect it to perform better on this data than any other. Using the training data to compare our model to any model built on different data would give our model an unfair advantage.

Another reason is because, as we will see in later sections of this monograph, there are endless ways for us to make a GLM as complex as we wish. There may be many variables available to include. For any given variable, any number of polynomial terms or hinge functions can be created. We can also add interactions of any combination of variables (not to mention interactions of polynomial terms and hinge functions), and so on. As we increase the complexity, the fit to the training data will *always* get better. For data the model fitting process has *not* seen, on the other hand, additional complexity may not improve the performance of a model—in fact, it may actually make it worse.

For a GLM, model complexity is measured in terms of **degrees of freedom**, or the number of parameters estimated by the model-fitting procedure. Every continuous variable we include adds a degree of freedom. For a categorical variable, a degree of freedom is added for each non-base level. Furthermore, every polynomial term, every hinge function or interaction term—basically, anything for which the model will need

to estimate another parameter value—counts as a degree of freedom. As the name implies, each degree of freedom provides the model more freedom to fit the training data. Since the fitting procedure always optimizes the fit, additional flexibility to fit the data better means the model *will* fit the data better.

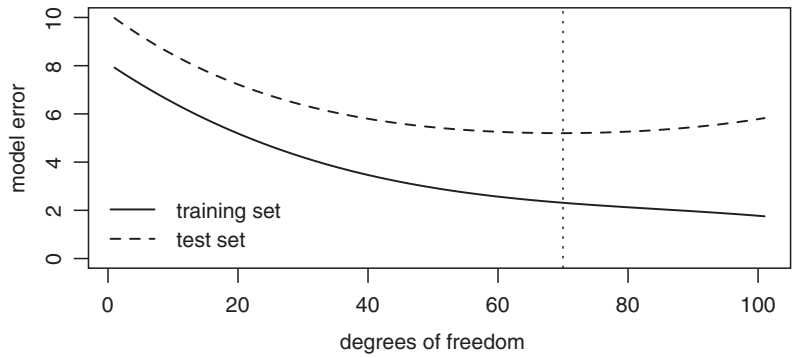
Figure 7 illustrates the relationship between the degrees of freedom and the performance of the model on the training set as well as on the test set (or any "unseen" data). Model performance is measured here by model error, or the degree to which the predictions "miss" the actual values, with lower error implying better model performance. As we can see, the performance on the training set is always better than on the test set. Increasing the complexity of the model improves the performance on both training set and the test set—up to a point. Beyond that point, the performance on the training set continues to improve—but on the test set, things get *worse.*

The reason for this deterioration of performance is because, with enough flexibility, the model is free to "explain" the randomness in the training set outcomes (called the **noise**) in addition to the part of the outcome driven by the systematic effects (called the **signal**). The noise in the training data would obviously not generalize to new data, so to the extent this information is in our model estimates this becomes a liability. A model that includes significant random noise in its parameter estimates is said to be **overfit**.

Our goal in modeling is to find the right balance where we pick up as much of the signal as possible with minimal noise, represented by the vertical dotted line in Figure 7. Thus, in addition to paying careful attention to the significance statistics and model fit diagnostics during the modeling process, it is critical to retain holdout data on which to test the resulting models. This out-of-sample testing allows for a truer assessment of the model's predictive power.

Since the divisions of data will remain intact throughout the entire modeling process, it is crucial to formulate a proper data splitting strategy before model building begins. The following sections discuss different approaches to splitting the data, as well as a possible alternative to splitting, called *cross validation.* In Chapter 7 we describe several tests that can be performed on the holdout set to choose among candidate models.

Figure 7. Illustration of the effect of model complexity (as measured by degrees of freedom, along the *x* axis), on the performance of the model (measured by model error, along the *y* axis) for both the training set and test set.



4.3.1. Train and Test

The simplest split to create is two subsets of the data, called the *training set* and the *test set.* The training set should be used for the entire model building process, beginning with the initial exploration of variables using univariate analyses, all the way through the model refinement. The test set is used when the model building is complete, to compare the resulting model against the existing rating plan and/or to assess the relative performance of several candidate models.

Typical proportions used for this split are 60% training/40% test or 70% training/30% test. Choice of split percentages involves a trade-off. More data available for the training set will allow for clearer views of patterns in the data. However, if too little data is left for the holdout, the final assessment of models will be have less certainty.

The split can be performed either by randomly allocating records between the two sets, or by splitting on the basis of a time variable such as calendar/accident year or month. The latter approach has the advantage in that the model validation is performed "out of time" as well as out of sample, giving us a more accurate view into how the model will perform on unseen years.

Out-of-time validation is especially important when modeling perils driven by common events that affect multiple policyholders at once. An example of this is the wind peril, for which a single storm will cause many incurred losses in the same area. If random sampling is used for the split, losses related to the same event will be present in both sets of data, and so the test set will not be true "unseen" data, since the model has already "seen" those events in the training set. This will result in over-optimistic validation results. Choosing a test set that covers different time periods than the training set will minimize such overlap and allow for better measures of how the model will perform on the completely unknown future.

4.3.2. Train, Validation and Test

If enough data is available, it may be useful to split the data three ways: in addition to the training and test sets, we create a *validation set.* The validation set is used to refine the model during the building process; the test set is held out until the end.

For example, a modeler may create an initial model using the training dataset, assess its performance on the validation dataset, and then make tweaks to the model based on the results. This is an iterative process. In this example, the validation dataset isn't really a holdout set, since the model is being adjusted based on its fit on the validation data.

Typical proportions used for this split are 40% for training, 30% for validation and 30% for test. Care should be taken that none of the subsets are too thin, otherwise their usefulness will be diminished.

4.3.3. Use Your Data Wisely!

A key caution regarding the use of a test set is that it be used *sparingly.* If too-frequent reference is made to the test set, or if too many choices of models are evaluated on it, it becomes less a test set and more of a training set; once a large part

of the modeling decision has been made based on how well it fits the test set, that fit becomes less indicative of how the model will behave on data that it has truly not seen.

Thus, the choice of how best to "spend" the available data throughout the refinement and validation of the model is an important part of the modeling strategy. Obviously, if a validation set is available (in addition to train and test), we have a bit more leeway, but the validation set will also diminish in usefulness if it is overused. As such, for a large part of the modeling process we will need to make use of the "in-sample" statistics—that is, the significant measures (such as *p*-values for parameter estimates and for the *F*-test, described in Section 6.2.1) derived using the training set.

As we may have many different ideas we wish to try in the course of refining and improving our model, the issue of precisely where reliance on the in-sample statistics will end and the validation or test metrics will begin should be carefully planned in advance.

An example strategy for this may be as follows. First, we might predefine a series of increasing levels of model complexity that we will evaluate as candidates for our final model. The simplest level of complexity might be to retain the current model and not change it all (yes, that should always be considered an option); as a second level, we may keep the structure of the current model intact, but change the numbers; for the third level, we may add some additional variables; the next level might add two-way interactions; subsequent levels may involve multiple-way interactions, subdivision of categorical variable groupings, and so on. Levels are ordered by the relative ease and cost of implementation. We build and refine a model at each level of complexity using the in-sample statistics (and validation set if available). When all the models are fully built, we evaluate them all on the test set, and their relative performance on this set is weighed together with all other business considerations in choosing which becomes the final model.

Once a final model is chosen, however, we would then go back and rebuild it using *all* of the data, so that the parameter estimates would be at their most credible.

4.3.4. Cross Validation

A common alternative to data splitting often used in predictive modeling is **cross validation**. Cross validation provides a means of assessing the performance of the model on unseen data through multiple splits of train and test.

There are several "flavors" of cross validation, but the most widely-used is called *k-fold* cross validation, for which the procedure is as follows:

1. Split the data into k groups, where k is a number we choose. (A common choice is 10.) Each group is called a *fold.* The split can either be done randomly or using a temporal variable such as calendar/accident year.
2. For the first fold:
   - *Train* the model using the *other* k–1 folds.
   - *Test* the model using the first fold.
3. Repeat step 2 for each of the remaining folds.

The output of this procedure is *k* estimates of model performance, each of which was assessed on data that its training procedure has not seen. Several models can be compared by running the procedure for each of them on the same set of folds and comparing their relative performances for each fold.

For most predictive modeling and machine learning applications, this is superior to a single train/test split, since *all* of the data is being used to test out-of-sample model performance as opposed to a single subset. However, it is often of limited usefulness for most insurance modeling applications, since cross validation has an important limitation: in order for it to be effective, the "training" phase of the procedure must encompass *all* the model-building steps. For a GLM, where the bulk of the model-building is the variable selection and transformation, that part would need to be included as well.

The reason for this is simple: if *all* the data was evaluated when deciding which variables to include, then even if the GLM fitting procedure was run on a subset of data, the remaining subset cannot be considered true "unseen" data. Some of the variables in our model may be there only because of outcomes "seen" in the test set.

Thus, using cross validation in place of a holdout set is only appropriate where a purely automated variable selection process is used. In such an instance, the same selection procedure can be run for each CV fold, and CV would then yield a good estimate (in fact, the best estimate) of out-of-sample performance. However, for most insurance applications, the variables are "hand-selected," with a great deal of care and judgment utilized along the way, and so proper cross validation is nearly impossible. Therefore, splitting the data at the outset and retaining that split throughout, as described in the prior sections, is the preferred approach.

Cross validation may still have some usefulness during the model building process. For example, when evaluating some of the model's "tuning parameters"—for example, how many polynomial terms to include, whether or not to use a certain variable as a weight, etc.—performing cross validation *within the training set* may yield valuable information on how a change to a model would affect its out-of-sample performance. However, the final model valuation should always be done using a distinct set of data held out until the end.

5. Selection of Model Form

Selecting the form of a predictive model is an iterative process, and is often more of an art than a science. As preliminary models are built and refined into final models, the model form is likely to evolve based on an analysis of the results.

In a generalized linear modeling framework, important decisions on the model form include:

- Choosing the target and predictor variables.
- Choosing a distribution for the target variable.
- Making decisions on the best form for the predictor variables, including whether to make them continuous or categorical, whether to apply transformations to the variables, and how best to group variables.

5.1. Choosing the Target Variable

Based on the scope of the modeling project, there may be several options for the target variable. When modeling a rating plan, for example, the target variable might be pure premium, claim frequency, or claim severity. If the goal of the project is instead to identify deficiencies in the existing rating plan, loss ratio may be a more appropriate target variable. Or when evaluating a set of underwriting restrictions, the probability of a large loss may be a good option.

The decision of which target variable to choose generally comes down to data availability and the preferences of the modeler. There is usually not one right answer, and it may be beneficial to try several options to see which one produces the best model.

5.1.1. Frequency/Severity versus Pure Premium

Where the ultimate goal of a model is to predict pure premium, there are two approaches we can use to get there.

1. Build two separate models: one with claims frequency—that is, count of claims per exposure—as the target, and another targeting claim severity, i.e., dollars of loss per claim. The individual models are then combined to form a pure premium model. Assuming log links were used for both, this combination of the two models is achieved by simply multiplying their corresponding relativity factors together.
2. Build a single model targeting pure premium, i.e., dollars of loss per exposure, using the Tweedie distribution.

This choice may be dictated by data constraints—for example, the data necessary to build separate models for claim frequency and severity may not be available. Furthermore, as the former approach requires building two models rather than one, time constraints may factor in as well, especially if a large number of pure premium models must be produced (e.g., when separately modeling multiple segments of the business or different perils).

However, where possible, the frequency/severity approach confers a number of advantages over pure premium modeling, some of which are as follows:

- Modeling frequency and severity separately often provides much more insight than a pure premium model, as it allows us to see the extent to which the various effects are frequency-driven versus severity-driven—information that may prove valuable in the model refinement process. Furthermore, some interesting effects may get "lost" when viewed on a pure premium basis due to *counteracting* effects on its components; for example, a variable that has a strong negative effect on frequency but an equally strong positive effect on severity would show up as a zero effect (and an insignificant variable!) in a pure premium model, and therefore go completely unnoticed. In such a case, while we may choose to deem the total effect of the variable a "wash" and not include it in our rating plan, that knowledge of the underlying effects may be useful in other business decisions.
- Each of frequency and severity is more stable—that is, it exhibits less random variance—than pure premium. Therefore, separating out those two sources of variance from the pure premium data effectively "cuts through the noise," enabling us to see effects in the data that we otherwise would not. For example, consider a variable that has a positive effect on frequency and no effect on severity, thereby having a positive total effect on pure premium. While this variable may show up as significant in a frequency model, when testing it in a pure premium model the high variance in severity may overwhelm the effect, rendering the variable insignificant. Thus, a predictive variable may be missed, leading to underfitting.
- Pure premium modeling can also lead to overfitting. Continuing with the above example of a variable that affects frequency only, if that variable *does* wind up included in our pure premium model, the model is forced to fit its coefficient to both the frequency and severity effects observed in the training data. To the extent the severity effect is spurious, that parameter is overfit.
- The distribution used to model pure premium—the Tweedie distribution—contains the implicit assumption that frequency and severity "move in the same direction"—that is, where a predictor drives an increase in the target variable (pure premium or loss ratio), that increase is made up of an increase in both its frequency and severity components. (See Section 2.7.3 for a detailed discussion on this.) Modeling frequency and severity separately frees us from this restriction.

5.1.2. Policies with Multiple Coverages and Perils

Where the line of business we are modeling includes several types of coverage, it is usually a good idea to separate out the data pertaining to each coverage and model them

separately. For example, when modeling for a Businessowners package policy that includes building, business personal property and liability coverage, each of those items should be separately modeled. We may also consider subdividing the data further and modeling each peril (or group of perils) individually; for example, for our Businessowners building model, we may wish to create separate models for fire and lightning, wind and hail, and all other.

Even if the final rating plan must be structured on an "all perils combined" basis, there may be benefit to modeling the perils separately, as that will allow us to tailor the models to the unique characteristics of each peril. We can always combine the models at the end. A simple method for combining separate by-peril models to form a combined all-peril model is as follows:

1. Use the by-peril models to generate predictions of expected loss due to each peril for some set of exposure data.
2. Add the peril predictions together to form an all-peril loss cost for each record.
3. Run a model on that data, using the all-peril loss cost calculated in Step 2 as the target, and the union of all the individual model predictors as the predictors.

The coefficients for the resulting model will yield the all-peril relativities implied by the underlying by-peril models for the mix of business in the data. Note that since the target data fed into this new model is extremely stable, this procedure doesn't require a whole lot of data. Rather, the focus should be on getting the mix of business right. The data used for this procedure should reflect the expected mix going forward, and so using only the most recent year may be ideal.

5.1.3. Transforming the Target Variable

In some modeling contexts, it may also be necessary or beneficial to transform the target variable in some way prior to modeling. Some considerations include:

- For pure premium, loss ratio or severity models, the presence of a few very large losses can have undue influence on the model results. In such cases, *capping* losses at a selected large loss threshold may yield a more robust and stable model. The cap point should be set high enough so that the target variable still captures the systematic variation in severity among risks, but not too high such that random large losses create excessive noise. (In Section 6.4 we discuss a formal statistical measure of a record's influence on the model results called *Cook's distance.* This statistic can also be used to alert the practitioner to instances where capping the target variable may be warranted.)
- In addition to the effect of individual large losses, it is also important to look out for catastrophic events that would cause a large number of losses at once, which can skew both frequency and severity effects. If possible, losses related to such events should be removed from the data entirely—thereby limiting the scope of the model to predicting *non-catastrophic* loss only—and a catastrophe model should separately be used to estimate the effect of catastrophes on the rating variables. If that is not an option, the effect of catastrophic losses should be tempered, either by adjusting

the value of the target variable downward or by decreasing the weight, so that these events should not unduly influence the parameter estimates.

- Where the data includes risks that are not at full loss maturity such that significant further loss development is to be expected (such recent accident year exposures for long-tailed lines), it may be necessary to *develop* the losses prior to modeling. Care should be taken so that the development factors applied match the type of entity being modeled. For example, for a severity model, the development factor should reflect only expected future development on *known claims*; for a pure premium or loss ratio model, the development factor should include the effect of pure IBNR claims as well.
- Where premium is used as the denominator of a ratio target variable (such as loss ratio), it may be necessary to on-level the premium.
- Where multi-year data is used, losses and/or exposures may need to be trended.

Note that for the latter three items on that list, as an alternative to applying those transformations, a temporal variable such as year can be included in the model. This variable would pick up any effects on the target related to time—such as trend, loss development and rate changes, for which the target has not been specifically adjusted—all at once. This is usually sufficient for most purposes, since the individual effects of development, trend, etc. are usually not of interest in models built for the purpose of rating. Rather, we wish to *control* for these effects so that they do not influence the parameter estimates of the rating variables, and the temporal variable does just that. Furthermore, the "control variable" approach also has the advantage in that the assumed temporal effects will be more "in tune" with the data the model is being estimated on.

On the other hand, there may be situations where adjusting the target using factors derived from other sources may be more appropriate. For example, where loss development factors are available that have been estimated from a wider, more credible body of data—perhaps incorporating industry data sources—those may provide a truer measure of development. Also, as there may already be established factors that have been assumed in other actuarial analyses of this same line of business (such as rate change analyses or reserve reviews) it may be preferable to use those in our rating factor model as well, so that all reviews of this line will be in sync. When doing so, however, it may be a good idea to try including the temporal variable even after the target has been adjusted; any significant temporal effects would then suggest a deficiency in the assumed factors, which can then be investigated.

5.2. Choosing the Distribution

Once the target variable is selected, the modeler must select a distribution for the target variable. This list of options is narrowed significantly based on the selected target variable. If modeling claim frequency, the distribution is likely to be either Poisson, negative binomial, or binomial (in the case of a logistic model). If modeling claim severity, common choices for the distribution are gamma and inverse Gaussian. The decision on which distribution to select may be based on an analysis of the deviance residuals, which is described in Section 6.3. It's important to realize, though, that the distribution

is very unlikely to fit the data perfectly. The goal is simply to find the distribution that fits the data most closely out of the set of possible options.

5.3. Variable Selection

For some modeling projects, the objective may be to simply update the relativity factors to be used in an *existing* rating plan. That is, the structure of the pricing formula will remain as-is, and only the numerical factors will change to reflect what is indicated by the most recent data. For such instances, **variable selection**—that is, choosing which variables to include in the model—is not an issue, as the choice of variables has been fixed at the outset.

Frequently, though, a rating plan update provides the company an opportunity to revisit the rating structure. Are there additional variables—not currently rated on, but available in the data—that may provide useful information about the target variable, thereby allowing us to more finely segment risks? Or, perhaps, a rating plan is being formulated for a line of business for the first time, and no prior model exists. In such cases, the choice of which variables to include becomes an important concern in the modeling process.

Certainly, a major criteria is variable significance—that is, we would like to be confident that the effect of the variable indicated by the GLM is the result of a true relationship between that predictor and the target, and not due to noise in the data. To that end, we are guided by the p-value, as described in Section 2.3.2. However, it is important to bear in mind a crucial limitation of the p-value: it says nothing about the probability of a coefficient being non-zero; it merely informs us of the probability of an estimated coefficient of that magnitude arising *if* the "true" coefficient *is* zero. In assessing our confidence in the indicated factor, the p-value should be viewed as one piece of information, which we combine with intuition and knowledge of the business to arrive at a decision on whether to include the variable. As such, there is no "magic number" p-value below which a variable should automatically be deemed significant.

In addition to statistical significance, other considerations for variable selection include:

- Will it be cost-effective to collect the value of this variable when writing new and renewal business?
- Does inclusion of this variable in a rating plan conform to actuarial standards of practice and regulatory requirements?
- Can the electronic quotation system be easily modified to handle the inclusion of this variable in the rating formula?

In practice, many different areas of the business may need to weigh in on the practicality and acceptability of any given variable in the final rating structure.

For more complex modeling projects—particularly where external data is attached to the insurer's own data to expand the predictive power—there may be hundreds or thousands of potential predictors to choose from, and variable selection becomes much more challenging. For such scenarios, a number of automated variable selection algorithms

exist that may aid in the process. (They may also add lots of spurious effects to the model if not used with appropriate care!) Those methods are beyond the scope of this paper.

5.4. Transformation of Variables

For any variable that is a potential predictor in our model, deciding whether or not to include it is not the end of the story. In many cases the variable will need to be transformed in some way such that the resulting model is a better fit to the data. Continuous and categorical variables each have considerations that would require transformation.

When including a continuous variable in a log-link model—logged, as discussed in Section 2.4.1—the model assumes a linear relationship between the log of the variable and the log of the mean of the target variable. However, this relationship doesn't always hold; some variables have a more complex relationship with the target variable that cannot be described by a straight line. For such instances, it is necessary to transform the variable in some way so that it can adequately model the effect.

To illustrate the ways a non-linear effect can be handled in a GLM, we will use the example of a multi-peril Businessowners building severity model that includes the building age (or age of construction) as one of its predictors. Building age is expressed in years, with a value of 1 signifying a new building.

Suppose, in this instance, the GLM returned a coefficient of −0.314 for log of building age. In log terms, this means that according to our model, each unit increase in the log of building age results in a 0.314 decline in the log of expected severity. We can also interpret this in "real terms": the expected severity for any building age relative to a new building is the age raised to −0.314. However, this is the best log-*linear* fit. But is a linear fit the best way to model the relationship?

The next section presents a useful graphical diagnostic that will allow us to find out.

5.4.1. Detecting Non-Linearity with Partial Residual Plots

The set of **partial residuals** for any predictor x_j in a model is defined as follows:

$$r_i = (y_i - \mu_i) g'(\mu_i) + \beta_j x_{ij}, \tag{13}$$

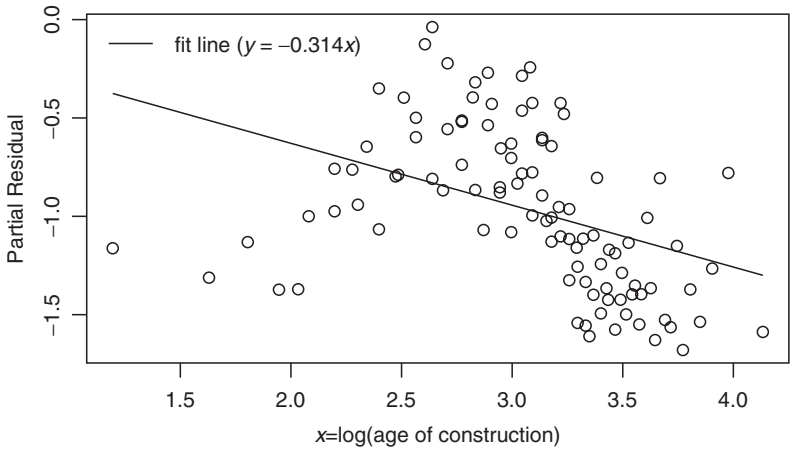
where $g'(\mu_i)$ is the first derivative of the link function. For a log link model, Equation 13 simplifies as follows:

$$r_i = \frac{y_i - \mu_i}{\mu_i} + \beta_j x_{ij}. \tag{14}$$

In Equation 14, the residual is calculated by subtracting the model prediction from the actual value, and then adjusted to bring it to a similar scale as the linear predictor (by dividing by μ_i)[9]. Then, $\beta_j x_{ij}$—that is, the part of the linear predictor that x_j is responsible for—is added back to the result. Thus, the partial residual may be thought

[9] Note that this is the "working residual" discussed in Section 6.3.2.

Figure 8. Partial Residual Plot of Age of Construction Variable



of as the actual value with all components of the model prediction *other than* the part driven by x_j subtracted out. (Hence the "partial" in "partial residual.") The variance in the partial residuals therefore contains the variance unexplained by our model in addition to the portion of the variance our model intends to explain with $\beta_j x_j$. We can then plot them against the model's estimate of $\beta_j x_j$ to see how well it did.

Figure 8 shows the partial residual plot for our example building age variable.[10] The model's linear estimate of the building age effect, or $-0.314x$, is superimposed over the plot. While the line may be the best *linear* fit to the points, it is certainly not the best fit, as the points are missing the line in a systematic way. The model is clearly over-predicting for risks where log building age is 2.5 (in real terms, building age 12) and lower. It under-predicts between 2.5 and 3.25, and once again over-predicts for older buildings. It is clear we will need something more flexible than a straight line to properly fit this data.

We present three ways such non-linearities can be accommodated within a GLM:

- binning the variable
- adding polynomial terms
- using piecewise linear functions.

Each of these approaches is discussed in the following sections.

5.4.2. Binning Continuous Predictors

One possible fix for non-linearity in a continuous variable is not to model it as continuous at all; rather, a new categorical variable is created where levels are defined as intervals over the range of the original variable. The model then treats it as it would any categorical variable; a coefficient is estimated for each interval, which applies to all risks falling within it.

[10] Note that despite this model having been built on around 50,000 records, the plot shows only 100 points. As 50,000 points would make for a very messy (and uninformative) scatterplot, the data has been *binned* prior to plotting. We discuss binning plotted residuals in Section 6.3.2. When binning partial residuals, the working weights, as described in that section, should be used.

Figure 9. Coefficient Estimates for the Bucketed Age of Construction Variable (*left*) and a Graphical Representation (*right*)

Variable	Estimate	Std. Err.
. . .	. . .	. . .
AoC: 11–14	0.622	0.117
AoC: 15–17	0.745	0.121
AoC: 18–20	0.561	0.124
AoC: 21–23	0.589	0.122
AoC: 24–26	0.344	0.128
AoC: 27–29	0.037	0.139
AoC: 30–33	–0.079	0.141
AoC: 34–39	–0.064	0.142
AoC: 39+	–0.139	0.147
. . .	. . .	. . .
. . .	. . .	. . .

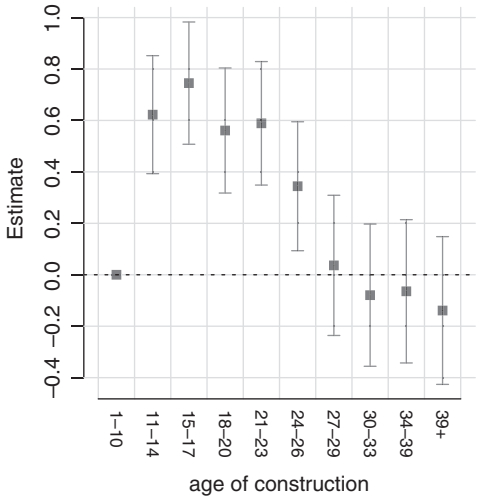


Figure 9 shows the results of running age of construction through our model as a categorical variable. For this example, ten bins were created. Interval boundaries were designed such that the bins contain roughly equal number of records, and building ages 1 through 10 was designated as the base level.

As the graphical plot of the coefficients shows, the model picked up a shape similar to that seen in the points of the partial residual plot. Average severity rises for buildings older than ten years, reaching a peak at the 15-to-17 year range, then gradually declining.

Binning a continuous variable frees the model from needing to constrain its assumed relationship with the target variable to any particular shape, as each level is allowed to float freely.

There are, however, some drawbacks to this approach.

In a general sense, binning a continuous variable, and therefore giving it a large number of parameters in our model, may violate the goal of parsimony, or keeping the model simple; as a rule, we shouldn't be giving the model more degrees of freedom than is necessary to adequately fit the data. The next paragraphs describe two more specific downsides to binning versus modeling a variable continuously.

Continuity in the Estimates is Not Guaranteed. Allowing each interval to move freely may not always be a good thing. The ordinal property of the levels of the binned variable have no meaning in the GLM; there is no way to force the GLM to have the estimates behave in any continuous fashion, and each estimate is derived independently of the others. Therefore, there is a risk that some estimates will be inconsistent with others due to random noise.

This pitfall is illustrated in the results shown in Figure 9. The building age effect on severity seems to be declining past 17 years. However, the 21–23 year factor is slightly higher than the 18–20 year factor. We have no reason to believe this break in the pattern is real, and it is most likely due to volatility in the data.

This issue may present an even bigger problem if the predictor variable is replacement cost of the building. The expectation is that, as the replacement cost increases, so does the expected loss cost for the policy. By including replacement cost in the model as a continuous variable (perhaps with some transformation applied), we can better ensure a monotonic relationship between replacement cost and predicted loss cost, which is a desirable outcome. If replacement cost is instead binned, there may be reversals in the variable coefficients due to volatility in the data. For example, buildings with a replacement cost of $300,000 may have a lower predicted loss cost than buildings with a $250,000 replacement cost, even though this result doesn't make sense.

In our building age example, note that the problem can be remedied somewhat by combining those two levels to a single level representing ages 18 through 23. Alternatively, we can manually smooth out the pattern when selecting factors.

Variation within Intervals is Ignored. Since each bin is assigned a single estimate, the model ignores any variation in severity that may exist *within* the bins. In our building age example, all buildings with ages between 1 and 10 years are assumed to have the same severity, which may not be the case. Of course, we could refine the interval boundaries to split that bin into two or more smaller ones. Doing so, however, would thin out the data, reducing the credibility of the estimates, thereby making them more susceptible to noise. Modeling building age as a continuous variable with a transformation (as discussed in the next sections) allows each building age to have a unique factor with no loss of credibility.

5.4.3. Adding Polynomial Terms

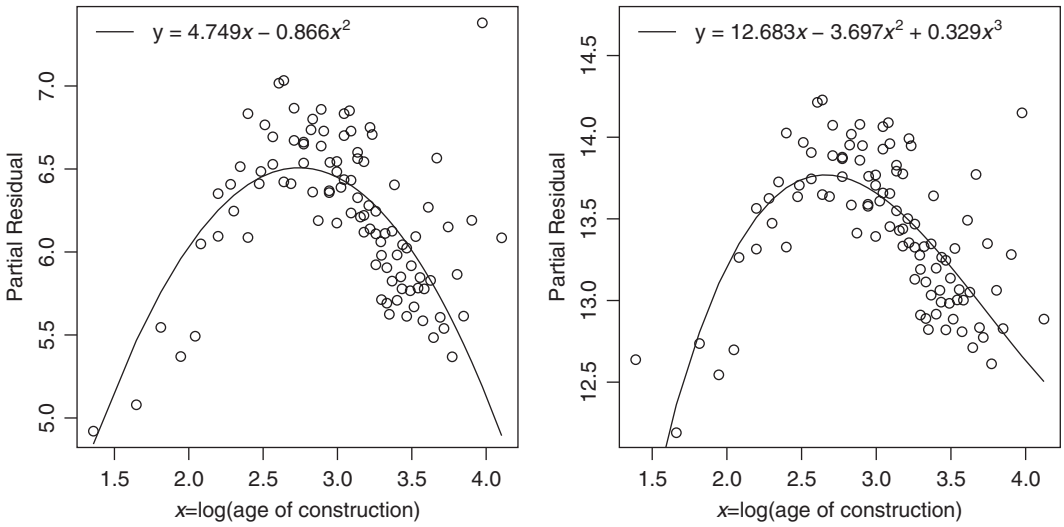
Another means of accommodating non-linearity in a linear model is to include the square, cube, or higher-order polynomials of the variable in the model in addition to the original variable. In such a model, the original variable and the polynomial terms are all treated as separate predictors, and a separate coefficient is estimated for each. This enables the model to fit curves to the data; the more polynomial terms that are provided, the more flexible the fit that can be achieved.

The left panel of Figure 10 shows the results of adding the square of the logged building age—in addition to log building age itself—to our model. In this example, the model estimated a coefficient of 4.749 for log building age (denoted here as x) and a coefficient of –0.866 for log building age squared (denoted as x^2). The graph shows the partial residuals with the curve formed by both building age terms superimposed.[11] Clearly this is a better fit to the data than the straight line shown in Figure 8.

In the right panel of Figure 10, a third term—the log building age cubed—was added. The additional freedom provided by this term allows the model to attenuate the downward slope on the right-hand side of the curve. This perhaps yields a better fit, as the points seem to indicate that the declining severity as building age increases does taper off toward the higher end of the scale.

[11] For this graph (as well as Figure 11) we extended the definition of partial residuals given in Equation 14 to include all terms related to the variable being evaluated (i.e., the $\beta_j x_{ij}$'s for all polynomial terms are added back to the working residual rather than the single $\beta_j x_{ij}$ term).

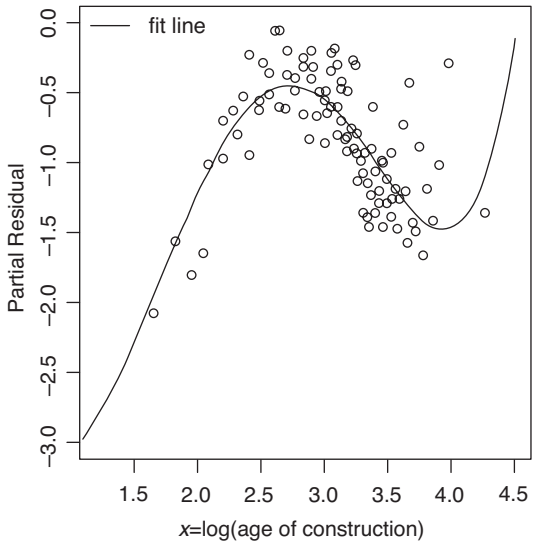
Figure 10. Partial Residual Plot of Age of Construction Variable using Two Polynomial Terms (*left*) and Three Polynomial Terms (*right*)



One potential downside to using polynomials is the loss of interpretability. From the coefficients alone it is often very difficult to discern the shape of the curve; to understand the model's indicated relationship of the predictor to the target variable it may be necessary to graph the polynomial function.

Another drawback is that polynomial functions have a tendency to behave erratically at the edges of the data, particularly for higher-order polynomials. For example, Figure 11 shows the partial residual plot that would result if we were to use *five* polynomial terms in our age of construction example. In this model, the fitted

Figure 11. Partial Residual Plot of Age of Construction Variable using Five Polynomial Terms



curve veers sharply upward near the upper bound of the data, and would most likely generate unreasonably high predictions for ages of construction higher than typical.

5.4.4. Using Piecewise Linear Functions

A third method for handling non-linear effects is to "break" the line at one or more points over the range of the variable, and allow the slope of the line to change at each break point.

Looking back at the partial residual plot in Figure 8, it is apparent that severity rises and reaches a peak at around age 2.75 (in log terms) and then declines. Thus, while a single straight line does not fit this pattern, a broken line—with a rising slope up to 2.75 and then declining—will likely do the job.

We can insert a break in the line at that point by defining a new variable as max(0,ln(*AoC*)–2.75), and adding it to the model. This new variable, called a *hinge function,* has a value of 0 for buildings with log age 2.75 or lower, and rises linearly thereafter, and so it will allow the GLM to capture the change in slope for older buildings versus newer ones.

Running the model with the addition of the hinge function breaking the line at 2.75 yields the partial output shown in Table 6.

For log(*AoC*) 2.75 and lower, the hinge function variable has a value of zero, and only the basic log(*AoC*) function varies; as such, the slope of the log-log response is 1.225. For log(*AoC*) above 2.75, on the other hand, both variables are in play. Thus, to calculate the log-log slope for older buildings, we must add the two coefficients together, yielding a slope of 1.225 + (–2.269) = –1.044. Thus, the log-log response is a positive slope for newer buildings and a negative slope for older buildings.

The left panel of Figure 12 shows the partial residual plot of log(*AoC*) under this model, with the broken line indicated by the model superimposed. This clearly does a much better job at fitting the points than the straight line of Figure 8.

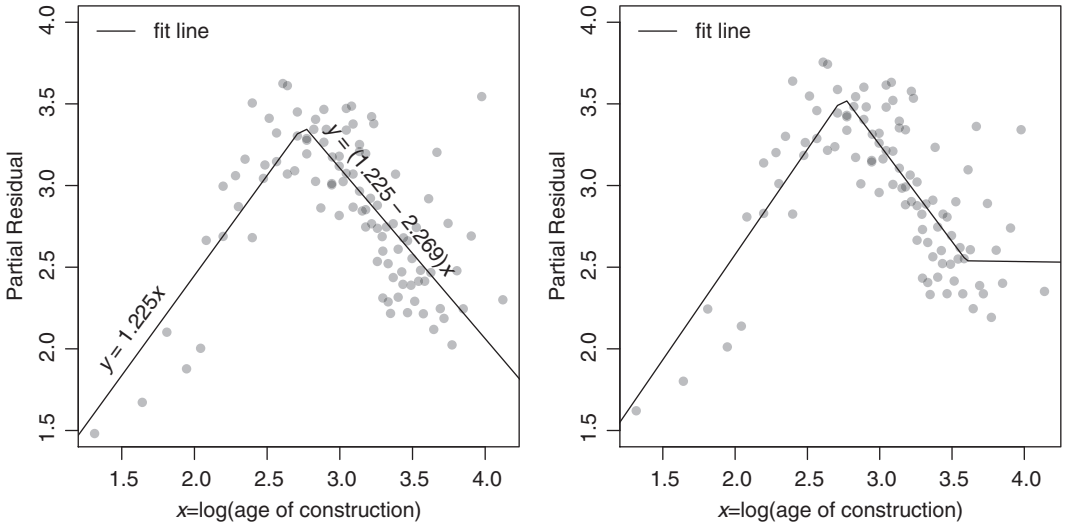
As the points seem to indicate that the downward slope tapers off at the far right of the graph, we may try to improve the fit further by adding another break at log(*AoC*) = 3.6. The resulting model output is shown in Table 7, and the right panel of Figure 12 graphs the partial residual plot.

The positive coefficient estimated for the second hinge function indicates that the slope of the line to the right of log(*AoC*) = 3.6 is higher than slope to the left of it. As the graphed fit line shows, this has the effect of nearly straightening out the steep

Table 6. Model Output After Adding a Hinge Function for a Break Point at log(*AoC*) = 2.75

Variable	Estimate	Std. Error	*p*-Value
. . .	. . .	. . .	. . .
log(*AoC*)	1.225	0.163	<0.0001
max(0,log(*AoC*)–2.75)	–2.269	0.201	<0.0001
. . .	. . .	. . .	. . .

Figure 12. Partial Residual Plot of Age of Construction Variable using a Break at 2.75 (*left*) and Breaks at Both 2.75 and 3.6 (*right*)



downward slope. The *p*-value of 0.0082 indicates that the evidence for a change in slope following the 3.6 point is strong, but not as strong as for a change following the 2.75 point. However, this may simply be due to that estimate being based on a relatively small subset of the data. As this leveling-off effect comports with our intuition, we may decide to keep the third hinge function term in the model.

The use of hinge functions allows one to fit a very wide range of non-linear patterns. Furthermore, the coefficients provided by the model can be easily interpreted as describing the change in slope at the break points; and, as we have seen, the significance statistics (such as *p*-value) indicate the degree of evidence of said change in slope.

One major drawback of this approach is that the break points must be selected by the user. Generally, break points are initially "guesstimated" by visual inspection of the partial residual plot, and they may be further refined by adjusting them to improve some measure of model fit (such as deviance, which is discussed in the next section). However, the GLM provides no mechanism for estimating them automatically. (In Chapter 10 we briefly discuss a useful model called *MARS,* a variant of the GLM, which, among other things, fits non-linear curves using hinge functions—and does it in an automated fashion with no need for tweaking by the user.)

Table 7. Adding an Additional Break Point at log(*AoC*) = 3.6

Variable	Estimate	Std. Error	*p*-Value
. . .	. . .	. . .	. . .
log(*AoC*)	1.289	0.159	<0.0001
max(0,log(*AoC*)–2.75)	–2.472	0.217	<0.0001
max(0,log(*AoC*)–3.60)	1.170	0.443	0.0082
. . .	. . .	. . .	. . .

Another potential downside is that while the fitted response curve is continuous, its first derivative is not—in other words, the fit line does not exhibit the "smooth" quality we would expect, but rather abruptly changes direction at our selected breakpoints.

5.4.5. Natural Cubic Splines

A more advanced method for handling non-linear effects—one that combines the concepts of polynomial functions and piecewise functions with breakpoints as discussed in the prior two sections—is the **natural cubic spline**. This is more mathematically complex than the prior two approaches, and we will not delve into the details here; the interested reader can refer to Hastie, Tibshirani & Friedman (Section 5.2.1 of 2nd Ed.) or Harrel (Section 2.4.4). We describe here some of its characteristics:

- The first and second derivatives of the fitted curve function are continuous—which in a practical sense means that the curve will appear fully "smooth" with no visible breaks in the pattern.
- The fits at the edges of the data (i.e., before the first selected breakpoint and after the last) are restricted to be linear, which curtails the potential for the kind of erratic edge behavior exhibited by regular polynomial functions.
- The use of breakpoints makes it more suitable than regular polynomial functions for modeling more complex effect responses, such as those with multiple rises and falls.

As with polynomial functions, natural cubic splines do not lend themselves to easy interpretation based on the model coefficients alone, but rather require graphical plotting to understand the modeled effect.

5.5. Grouping Categorical Variables

Some categorical predictor variables are binary or can only take on a small number of values. Others, though, can take on a large number of possible values, and for these variables it is generally necessary to group them prior to inclusion in the model. Consider, for example, driver age. If ungrouped, this variable is likely to consume too many degrees of freedom, which can lead to nonsensical results and the inability of the model to converge. In deciding how to group such variables, one strategy is to start with many levels and then begin grouping based on model coefficients and significance. For example, we may start with 30 buckets, then compare the coefficients for neighboring buckets. If one bucket is, say, drivers between the ages of 26 and 27, and another is drivers between 28 and 29, and the coefficients for these two levels are similar, we can create a new bucket for drivers between 26 and 29. This is generally an iterative process and requires balancing the competing priorities of predictive power, parsimony, and avoiding overfitting to the data.

5.6. Interactions

Thus far, we have focused on optimizing the selection and transformation of variables for our model under the assumption that each variable has an individual effect on the target variable. However, we may also wish to consider the hypothesis that two or more variables may have a *combined* effect on the target over and above their

individual effects. Stated differently, the effect of one predictor may depend on the level of another predictor, and vice-versa. Such a relationship is called an **interaction**.

An example of an interaction is illustrated in Figure 13. In this example we have two categorical variables: variable 1 has two levels, A and B, with A being the base level; variable 2 has levels X (the base) and Y.

The table in the left panel shows the mean response for each combination of levels with no interaction. For variable 1, the multiplicative factor for level B (relative to base level A) is 2.0, regardless of the level of variable 2. Similarly, the variable 2 relativity of level Y is 1.5, regardless of the level of variable 1. Simple enough.

The right panel shows an example of where an interaction is present. Where variable 2 is X, the relativity for level B is 2.0, as before; however, where variable 2 is Y, the level B relativity is 2.2. Or, we can state this effect in terms of variable 2: the relativity for level Y is either 1.50 or 1.65, depending on the level of variable 1.

Another way of describing the situation in the right panel of Figure 13 is as follows: for each of the two variables, there are **main effects**, where the relativity of level B is 2.0 and the relativity of level Y is 1.5; plus, there is an additional **interaction effect** of being both of level Y and level B—with a multiplicative factor of 1.1. This is the setup typically used in GLMs, and it allows us to use the GLM significance statistics to test the interaction effects distinctly from the main effects in order to determine whether the inclusion of an interaction significantly improves the model.

The above example illustrates the interaction of two categorical variables. It is also possible to interact two continuous variables, or a continuous variable with a categorical variable. In the following sections, we further explore the categorical/categorical interaction in a GLM, as well as the other two interaction types.

5.6.1. Interacting Two Categorical Variables

We present here a more concrete example to illustrate the handling of a categorical/categorical interaction in a GLM.

Suppose, for a commercial building claims frequency model, which uses a Poisson distribution and a log link, we include two categorical predictors: occupancy class, coded 1 through 4, with 1 being the base class; and sprinklered status, which can be either "no" or "yes," the latter indicating the presence of a sprinkler system in the building, with no sprinkler being the base case.

Figure 13. An Example of a Mean Response for Each Level of Two Categorical Variables Without an Interaction (*left panel*) and With an Interaction (*right panel*)

Without Interaction				With Interaction			
		Variable 1				Variable 1	
		A	**B**			**A**	**B**
Variable 2	**X**	10	20	Variable 2	**X**	10	20
	Y	15	30		**Y**	15	33

Table 8. Model with the Main Effects of Occupancy Class and Sprinklered Status

	Estimate	Std. Error	*p*-Value
(Intercept)	–10.8679	0.0184	<0.0001
occupancy:2	0.2117	0.0264	<0.0001
occupancy:3	0.4581	0.0262	<0.0001
occupancy:4	0.0910	0.0245	0.0005
sprinklered:Yes	–0.3046	0.0372	<0.0001

Running the model with the main effects only yields the output shown in Table 8. The coefficient of –0.3046 indicated for "sprinklered:yes" indicates a sprinklered discount of 26.3% (as $e^{-0.3046}-1 = -0.263$).

We then wish to test whether the sprinklered discount should vary by occupancy class. To do this, we add the interaction of those two variables in the model, in addition to the main effects. Behind the scenes, the model fitting software adds additional columns to the design matrix: a column for each combination of non-base levels for the two variables. Each of those columns is valued 1 where a risk has that combination of levels, and is 0 otherwise. These new columns are treated as distinct predictors in Equation 2, and so the coefficient estimated for each of those new predictors will indicate the added effect—above the main effects—of a risk having that combination of levels. In our example, this results in three additional predictors being added to our model: the combination of "sprinklered:yes" with each of occupancies 2, 3, and 4.

Running this model results in the output shown in Table 9. In this context, the coefficient of –0.2895 for the main effect "sprinklered:yes" indicates a discount of 25.1% *for occupancy class 1.* The three interaction effects yield the effect of having a sprinkler for the remaining three occupancy groups *relative* to the sprinklered effect for group 1.

Table 9. The Model with the Addition of the Interaction Term

	Estimate	Std. Error	*p*-Value
(Intercept)	–10.8690	0.0189	<0.0001
occupancy:2	0.2303	0.0253	<0.0001
occupancy:3	0.4588	0.0271	<0.0001
occupancy:4	0.0701	0.0273	0.0102
sprinklered:Yes	–0.2895	0.0729	0.0001
occupancy:2, sprinklered:Yes	–0.2847	0.1014	0.0050
occupancy:3, sprinklered:Yes	–0.0244	0.1255	0.8455
occupancy:4, sprinklered:Yes	0.2622	0.0981	0.0076

Looking at the row for the first interaction term, the negative coefficient indicates that occupancy class 2 should receive a steeper sprinklered discount than class 1; specifically, its indicated sprinklered factor is $e^{-0.2895-0.2847}$=0.563, or a 43.7% discount. The low *p*-value of 0.005 associated with that estimate indicates that the sprinklered factor for this class is indeed significantly different from that of class 1.

The interaction of occupancy class 3 with sprinklered shows a negative coefficient as well. However, it has a high *p*-value, indicating that the difference in sprinklered factors is not significant. Based on this, we may wish to simplify our model by combining class 3 with the base class for the purpose of this interaction.

The interaction term for occupancy class 4 has a significant positive coefficient of nearly equal magnitude to the negative coefficient of the main sprinklered effect. This result suggests that perhaps occupancy class 4 should not receive a sprinklered discount at all.

5.6.2. Interacting a Categorical Variable with a Continuous Variable

We extend the above example to add a continuous variable—amount of insurance (AOI)—to our frequency model. Following the practice discussed in Section 2.4.1, we will log AOI prior to inclusion in the model.

The main-effects model yields the estimates shown in Table 10. This model indicates a sprinklered factor of $e^{-0.7167} = 0.488$. The coefficient for log(AOI) indicates that the log of the mean frequency increases 0.416 for each unit increase in log(AOI)—or, equivalently, the frequency response to AOI (in real terms) is proportional to the power curve $AOI^{0.4161}$.

We now wish to test whether the AOI curve should be different for sprinklered versus non-sprinklered properties. To do so, we specify that we would like to add the interaction of sprinklered and log(AOI) to our model. The GLM fitting software adds a column to our design matrix that is the product of the indicator column for "sprinklered:Yes" and log(AOI). Thus, the resulting new predictor is 0 where sprinklered = No, and the log of AOI otherwise.

Running this GLM yields the output shown in Table 11. For this model, the coefficient for the log(AOI) main effect yields the AOI curve for risks of the base class

Table 10. A Model with Occupancy Class, Sprinklered Status and AOI as Main Effects

	Estimate	Std. Error	*p*-Value
(Intercept)	–8.8431	0.1010	<0.0001
occupancy:2	0.2909	0.0248	<0.0001
occupancy:3	0.3521	0.0267	<0.0001
occupancy:4	0.0397	0.0266	0.1353
sprinklered:Yes	–0.7167	0.0386	<0.0001
log(AOI)	0.4161	0.0075	<0.0001

Table 11. Adding the Interaction of AOI and Sprinklered Status

	Estimate	Std. Error	*p*-Value
(Intercept)	–8.9456	0.1044	<0.0001
occupancy:2	0.2919	0.0247	<0.0001
occupancy:3	0.3510	0.0266	<0.0001
occupancy:4	0.0370	0.0265	0.1622
sprinklered:Yes	0.7447	0.3850	0.0531
log(AOI)	0.4239	0.0078	<0.0001
sprinklered:Yes, log(AOI)	–0.1032	0.0272	0.0001

of "sprinklered" (that is, risks for which sprinklered = "No"). The model also estimates a coefficient of –0.1032 for the interaction term, which indicates that the rise of frequency in response to AOI is less steep for sprinklered properties than for non-sprinklered properties. The *p*-value indicates that this difference in curves is significant.

The positive coefficient estimated for "sprinklered:Yes" in this model may be a bit startling at first. Does this mean that a *premium* should now be charged for having a sprinkler?

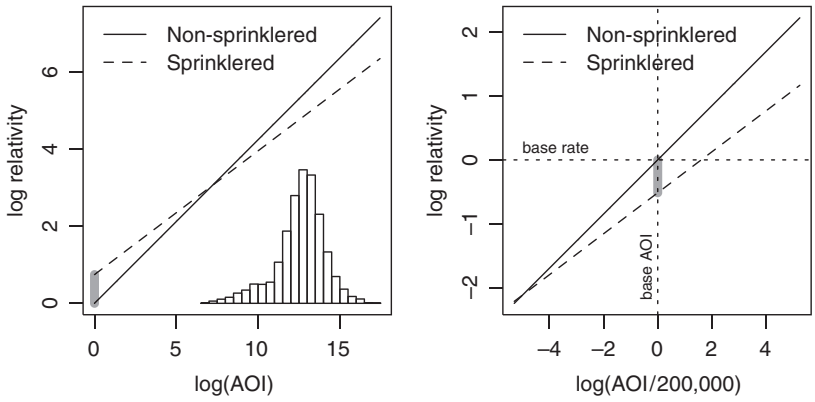
Not quite. In interpreting the meaning of this, it is important to recognize that the model includes another variable that is non-zero for sprinklered properties—the interaction term, which captures the difference in the AOI *slope* between sprinklered and non-sprinklered risks. Thus, the main sprinklered effect may be thought of as an adjustment of the *intercept* of the AOI curve, or the indicated sprinklered relativity where log(AOI) = 0.

Of course, where log(AOI) is zero, AOI is $1—a highly unlikely value for AOI. The left panel of Figure 14 shows a graphical interpretation of this model's indicated effects of AOI and sprinklered status. The *x*-axis is the log of AOI, and *y*-axis shows the (log) indicated relativity to the hypothetical case of a non-sprinklered property with an AOI of $1. The two lines show the effect of AOI on log mean frequency: the slope of the "sprinklered" line is less steep than that of "non-sprinklered," due to the negative coefficient of the interaction term.

The vertical gray stripe at the bottom left highlights what the main sprinklered effect coefficient refers to: it raises the sprinklered AOI curve at *log*(AOI) = 0. However, as the AOI histogram overlaid on the graph shows, *log*(AOI) = 0 is way out of the range of the data, and so this "sprinklered surcharge" is just a theoretical construct, and no actual policy is likely to be charged such a premium.

An alternate way of specifying this model—one that leads to better interpretation—is to divide the AOI by the base AOI prior to logging and including it in the model. Supposing our chosen base AOI (which would receive a relativity of 1.00 in our rating plan) is $200,000, we use log(AOI/200,000) as the predictor in our model. The resulting estimates are shown in Table 12.

Figure 14. Illustration of the Effect of the Interaction of Sprinklered and Amount of Insurance (*left panel*) and the Same Model After Dividing AOI by Its Base Amount (*right panel*)



This model is equivalent to the prior model; that is, they will both produce the same predictions. The sprinklered coefficient (now negative) still refers to the specific case where the value of the AOI predictor is zero—however, with the AOI predictor in this form it has a more natural interpretation: it is the (log) sprinklered relativity for a risk with the *base* AOI.

The right panel of Figure 13 illustrates the output of this model. (The x-axis in that panel spans only the values actually present in the data.) The gray stripe at center shows the main effect for sprinklered status, which is to lower the mean response at $x = 0$ (the base AOI) by 0.5153 for sprinklered risks.

Note that in all this discussion, we described this interaction as "the slope of the AOI curve varying based on the sprinklered status." Of course, it is just as valid to characterize it as "the sprinklered discount varying based on AOI." Which way it is presented in the rating plan is a matter of preference.

Table 12. The Model of Table 11 with log AOI Centered at the Base AOI

	Estimate	Std. Error	*p*-Value
(Intercept)	–3.7710	0.0201	<0.0001
occupancy:2	0.2919	0.0247	<0.0001
occupancy:3	0.3510	0.0266	<0.0001
occupancy:4	0.0370	0.0265	0.1622
sprinklered:Yes	–0.5153	0.0635	<0.0001
log(AOI/200000)	0.4239	0.0078	<0.0001
sprinklered:Yes, log(AOI/200000)	–0.1032	0.0272	0.0001

As an aside, note that this last model form, with AOI centered at the base AOI, has an additional benefit: the intercept term (after exponentiating) yields the indicated frequency at the base case (i.e., when all variables are at their base levels). In general, for a GLM to have this property, all continuous variables need to be divided by their base values prior to being logged and included in the model.

5.6.3. Interacting Two Continuous Variables

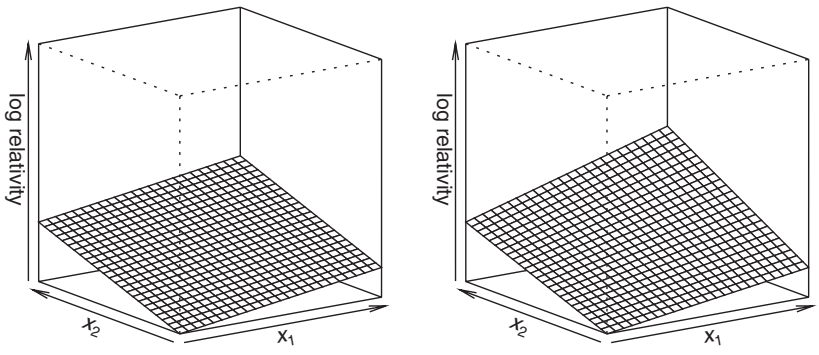
To understand the third type of interaction—a continuous variable with another continuous variable—it is useful to visualize the combined effects of the variables on the log mean response as perspective plots, with the two variables shown along the *x*- and *y*-axes, and the relative log mean shown along the *z*-axis.

The left panel of Figure 15 graphs a scenario where two variables, x_1 and x_2, are included in a model as main effects only, and the GLM indicates coefficients for them of 0.02 and 0.04, respectively. The response curve slopes for those two variables can be seen by following the front edge of the plane along the *x*- and *y*-axes; clearly, x_2 has a steeper slope than x_1, which is due to its coefficient being larger. However, for any given value of x_2, the x_1 curve, while in a different position vertically, has the same slope, and vice versa. As such, the effect of the two variables are independent of each other.

If we believe the slope for each variable should depend on the value of the other variable, we may include an interaction term. This term takes the form of the *product* of the two predictors. The right panel illustrates the case where the main effect coefficients are the same as before, but an added interaction term has a coefficient of 0.005. The slope of x_1 where $x_2 = 0$ (the front edge of the plane) is the same as in the left panel graph. However, moving inward, as x_2 increases, we see the slope of x_1 becomes more steep. Similarly, the slope of x_2 steepens as x_1 increases.

As with the earlier interaction types, the *p*-value estimated for the interaction term guides us in our determination of whether this effect is significant, or whether the variables should be left independent.

Figure 15. Perspective Plots of the Log Mean Response to Two Continuous Variables, both Without (*left*) and With (*right*) an Interaction Term



6. Model Refinement

6.1. Some Measures of Model Fit

GLM software provides a number of statistical measures of how well the model fits the training data, which are useful when comparing candidates for model specifications and assessing the predictive power of individual variables. The most important such measures are *log-likelihood* and *deviance.*

6.1.1. Log-Likelihood

For any given set of coefficients, a GLM implies a probabilistic mean for each record. That, along with the dispersion parameter and chosen distributional form, implies a full probability distribution. It is therefore possible to calculate, for any record, the probability (or probability density) that the GLM would assign to the actual outcome that has indeed occurred. Multiplying those values across *all* records produces the probability of all the historical outcomes occurring; this value is called the **likelihood**.

A GLM is fit by finding the set of parameters for which the likelihood is the highest. This is intuitive; absent other information, the best model is the one that assigns the highest probability to the historical outcomes. Since likelihood is usually an extremely small number, the log of likelihood, or **log-likelihood**, is usually used instead to make working with it more manageable.

Log-likelihood by itself can be difficult to interpret. It is therefore useful to relate the log-likelihood to its hypothetical upper and lower bounds achievable with the given data.

At the low end of the scale is the log-likelihood of the **null model**, or a hypothetical model with no predictors—only an intercept. Such a model would produce the same prediction for every record: the grand mean.

At the other extreme lies the **saturated model**, or a hypothetical model with an equal number of predictors as there are records in the dataset. For such a model, Equation 2 becomes a system of equations with n equations and n unknowns, and therefore a perfect solution is possible. This model would therefore perfectly "predict" every historical outcome. It would also be, most likely, useless; overfit to the extreme, it is essentially nothing more than a complicated way of restating the historical data. However, since predicting each record perfectly is the theoretical best a model can possibly do, it provides a useful upper bound to log-likelihood for this data.

While the null model yields the lowest possible log-likelihood, the saturated model yields the highest; the log-likelihood of your model will lie somewhere in between. This naturally leads to another useful measure of model fit: deviance.

6.1.2. Deviance

Scaled deviance for a GLM is defined as follows:

$$scaled\ deviance = 2 \times \left(ll_{saturated} - ll_{model} \right) \tag{15}$$

where $ll_{saturated}$ is the log-likelihood of the saturated model, and ll_{model} is the log-likelihood of the model being evaluated. This may be more formally stated as follows (with scaled deviance denoted as D^*):

$$D^* = 2 \times \sum_{i=1}^{n} \ln f\left(y_i \middle| \mu_i = y_i \right) - \ln f\left(y_i \middle| \mu_i = \mu_i \right) \tag{16}$$

The first term after the summation sign is the log of the probability of outcome y_i given that the model's predicted mean is y_i—the mean that would be predicted by the saturated model. The second term is the log probability assigned to the outcome y_i by the actual model. The difference between those two values can be thought of as the magnitude by which the model missed the "perfect" log-likelihood for that record. Summing across all records and multiplying the result by 2 yields the scaled deviance.

Multiplying the scaled deviance by the estimated dispersion parameter ϕ yields the *unscaled deviance*.[12] The unscaled deviance has the additional property of being independent of the dispersion parameter and thereby being useful for comparing models with different estimates of dispersion.

However, irrespective of the type of deviance measure (i.e., scaled or unscaled), note that the fitted GLM coefficients are those that minimize deviance. Recall that the previous section stated that the GLM is fit by maximizing log-likelihood, and in fact those two statements are equivalent: maximizing log-likelihood is also minimizing deviance. It is easy to see that by examining Equation 15 above. The first term inside the parentheses, $ll_{saturated}$, is constant with respect to the model coefficients, as it is purely a function of the data and the selected distribution. Since the log-likelihood of our model is subtracted from it, the coefficients yielding the maximum log-likelihood also yield the minimum deviance.

The deviance for the saturated model is zero, while the deviance for the null model can be thought of as the total deviance inherent in the data. The deviance for your model will lie between those two extremes.

---

[12] See Anderson, et al. § 1.154-1.158 for a more formal and generalized definition of the unscaled deviance. Further note that there is some discrepancy in terminology among various GLM texts, as some (e.g., Dobson & Barnett [2008]) use the term "deviance" to refer to the measure presented here as "scaled deviance," and use "scaled deviance" to refer to that measure multiplied by the estimated dispersion parameter (i.e., the "unscaled deviance" in this text). We have followed the terminology used in Anderson et al [2007] and McCullaugh and Nelder [1989].

6.1.3. Limitations on the Use of Log-Likelihood and Deviance

The next section discusses some statistical tests that can be used to compare the fits of different models using these measures. However, at the outset, it is important to note the following caveats:

Firstly, when comparing two models using log-likelihood or deviance, the comparison is valid only if the datasets used to fit the two models are exactly identical. To see why, recall that the total log-likelihood is calculated by summing the log-likelihoods of the individual records across the data; if the data used for one model has a different number of records than the other, the total will be different in a way that has nothing to do with model fit.

This, by the way, is something to look out for when adding variables to an existing model and then comparing the resulting model with the original. If the new variable has missing values for some records, the default behavior of most model fitting software is to toss out those records when fitting the model. In that case, the resulting measures of fit are no longer comparable, since the second model was fit with fewer records than the first.

For any comparisons of models that use deviance, in addition to the caveat above, it is also necessary that the assumed distribution must be identical as well. This restriction arises from deviance being based on the amount by which log-likelihood deviates from the "perfect" log-likelihood; changing any assumptions other than the coefficients would alter the value of the "perfect" log-likelihood as well the model log-likelihood, muddying the comparison.

6.2. Comparing Candidate Models

As described above, the process of building and refining a GLM is one that takes place over many iterations; frequent decisions need to be made along the way, such as: which predictors to include; the appropriate transformations, if any, to be applied to continuous variables; and the groupings of levels for categorical variables. This section presents several statistical tests, based on the measures of model fit just discussed, that can be used to compare successive model runs and guide our decision making.

6.2.1. Nested Models and the F-Test

Where a model uses a subset of the predictors of a larger model, the smaller model is said to be a **nested model** of the larger one. Comparisons of nested models frequently occur when considering whether to add or subtract predictors. We may have one model that includes the extra predictors being considered, and one that does not but includes all the other variables. We then wish to compare the model statistics to answer the question: is the larger model, with the added variables, better than the smaller one? In other words, do the added predictors enhance the predictive power of the model?

We can do that by comparing the deviance (subject to the caveats noted above). However, in doing so we must consider a basic fact: adding predictors to a model *always* reduces deviance, whether the predictor has any relation to the target variable

or not. This is because more predictors—which means more parameters available to fit—gives the model fitting process more freedom to fit the data, and so it *will* fit the data better. At the extreme end of that is the saturated model, where the model fitting process has enough freedom to produce a perfect fit—even if the predictors are purely random and have no predictive power at all; with n unknowns and n equations, a perfect fit is always mathematically possible.

Therefore the meaningful question when comparing deviances is: did the added predictors reduce the deviance *significantly more* than we would expect them to if they are *not* predictive? One way to answer that is through the ***F*-test**, wherein the ***F*-statistic** is calculated and compared against the ***F* distribution**.

The formula for the *F*-statistic is

$$F = \frac{D_S - D_B}{(\#\ of\ \ added\ \ parameters) \times \hat{\phi}_B} \tag{17}$$

In Equation 17, the symbol "D" refers to the *unscaled* deviance, and the subscripts "S" and "B" refer to the "small" and "big" models, respectively. The numerator is the difference in the unscaled deviance between the two models—that is, the amount by which the unscaled deviance was reduced by the inclusion of the additional parameters. As described above, this value is positive, since deviance always goes down.

The $\hat{\phi}_B$ in the denominator is the estimate of the dispersion parameter for the big model. As it happens, this is also a good estimate of the amount by which we can expect unscaled deviance to go down for each new parameter added to the model—simply by pure chance—if the parameter adds no predictive power. Multiplying this value by the number of added parameters gives us the total expected drop in deviance. For the added predictors to "carry their weight," they must reduce deviance by significantly more than this amount. (If some of the added predictors are categorical, note that a categorical variable with m levels adds $m - 1$ parameters—one for each level other than the base level.)

Thus, the ratio in Equation 17 has an expected value of 1. If it is significantly greater than 1, we may conclude that the added variables do indeed improve the model.

How much greater than 1 is significant? Statistical theory says that the *F*-statistic follows an *F* distribution, with a numerator degrees of freedom equal to the number of added parameters and a denominator degrees of freedom equal to $n - p_B$, or the number of records minus the number of parameters in the big model. If the percentile of the *F*-statistic on the *F* distribution is sufficiently high, we may conclude that the added parameters are indeed significant.

As an example, suppose the auto GLM we built on 972 rows of data with 6 parameters yields an unscaled deviance of 365.8 and an estimated dispersion parameter of 1.42. We wish to test the significance of an additional potential predictor: rating territory, a categorical variable with 5 levels.

We run the GLM with the inclusion of rating territory, thereby adding $5 - 1 = 4$ parameters to the model. Suppose the unscaled deviance of the resulting model is 352.1, and its estimated dispersion parameter is 1.42.

Using this information and Equation 17, we calculate the *F*-statistic.

$$\frac{365.8 - 352.1}{4 \times 1.42} = 2.412$$

To assess the significance of this value, we compare it against an *F* distribution with 4 numerator degrees of freedom and 972 − 10 = 962 denominator degrees of freedom. An *F* distribution with those parameters has 2.412 at its 95.2 percentile, indicating a 4.8% probability of a drop in deviance of this magnitude arising by pure chance. As such, rating territory is found to be significant at the 95% significance level.

6.2.2. Penalized Measures of Fit

The *F*-test of the prior section is only applicable to nested models. Frequently, though we would want to compare non-nested models—that is, models having different variables, where one does not contain a subset of the variables of the other. As described above, deviance alone can not be used, since adding parameters always reduces deviance, and so selecting on the basis of lowest deviance gives an unfair advantage to the model with more parameters, which can lead to over-fitting.

A practical way to avoid the problem of over-fitting is to use a *penalized measure of fit.* While deviance is strictly a measure of model goodness of fit on the training data, a penalized measure of fit also incorporates information about the model's complexity, and so becomes a measure of model quality. Using one of these measures, one can compare two models that have different numbers of parameters. The two most commonly used measures of deviance are **AIC** and **BIC**.

AIC, or the **Akaike Information Criterion**, is defined as follows:

$$AIC = -2 \times log\text{-}likelihood + 2p \tag{18}$$

where p is the number of parameters in the model. As with deviance, a smaller AIC suggests a "better" model. The first term in the above equation declines as model fit on the training data improves; the second term, called the *penalty term,* serves to increase the AIC as a "penalty" for each added parameter. (The rationale for using twice the number of parameters as the penalty is grounded in information theory and out of the scope of this monograph.) Using this criterion, models that produce low measures of deviance but high AICs can be discarded.

Note that the first additive term of equation 18 is the same as the formula for scaled deviance in Equation 15 but without the $ll_{saturated}$ term, which is constant with respect to the model predictions. As such, the AIC can also be thought of as a penalized measure of *deviance,* when using it to compare two models. (AIC has little meaning outside of the context of a comparison anyway.) As a matter of fact, some statistical packages occasionally take advantage of this equivalence and substitute deviance for $-2 \times log\text{-}likelihood$ where it would simplify the calculation.

BIC, or the **Bayesian Information Criterion**, is defined as $-2 \times \textit{log-likelihood} + p \log(n)$, where p is once again the number of parameters, and n is the number of data points that the model is fit on. As most insurance models are fit on very large datasets, the penalty for additional parameters imposed by BIC tends to be much larger than the penalty for additional parameters imposed by AIC.

Most statistical packages can produce either of these measures. In practical terms, the authors have found that AIC tends to produce more reasonable results. Relying too heavily on BIC may result in the exclusion of predictive variables from your model.

6.3. Residual Analysis

One useful and important means of assessing how well the specified model fits the data is by visual inspection of the *residuals,* or measures of the deviations of the individual data points from their predicted values. For any given record, we can think of the residual as measuring the magnitude by which the model prediction "missed" the actual value. In our GLM, this is assumed to be the manifestation of the *random* component of the model—that is, the portion of the outcome driven by factors other than the predictors, which our model describes using Equation 1 and our assumed distribution. Therefore, it is natural to inspect these values to determine how well our model actually does at capturing this randomness.

The simplest kind of residual is the **raw residual** which is just the difference between actual and expected, or $y_i - \mu_i$. For GLMs, however, two measures of deviation of actual from predicted that are much more useful are the **deviance residual** and the **working residual**. These measures have many useful properties for assessing model fit, and are discussed in the following sections.

6.3.1. Deviance Residuals

The square of the deviance residual for any given record is defined as that record's contribution to the unscaled deviance. The deviance residual takes the same sign as actual minus predicted. Look back at Equation 16 (on page 63); the deviance residual for any record i is the square root of: twice the term to the right of the summation sign multiplied by the scale parameter. We use the negative square root where actual (y_i) is less than expected (μ_i), and the positive square root where $y_i > \mu_i$.

Intuitively, we can think of the deviance residual as the residual "adjusted for" the shape of the assumed GLM distribution, such that its distribution will be approximately normal if the assumed GLM distribution is correct.

In a well-fit model, we expect deviance residuals to have the following properties:

- **They follow no predictable pattern.** Remember, the residuals are meant to be the random, or unpredictable, part of the data. If we discover any way the residuals can be predicted, then we are leaving some predictive power on the table and we can probably improve our model to pick it up.
- **They are normally distributed, with constant variance.** The *raw* residuals are certainly not expected to follow a normal distribution (assuming we selected a distribution other than normal); furthermore, the variance of the raw residual of any

record would be dependent on its predicted mean, due to the variance property of Equation 3. However, as the deviance residuals have been adjusted for the shape of the underlying distribution, they are expected to be normal and with constant variance. (The latter property is called **homoscedasticity**.) Any significant deviations from normality or homoscedasticity may indicate that the selected distribution is incorrect.

Figure 16 shows examples of two ways we might assess the normality of deviance residuals for a model of claim severity built using the gamma distribution. The left panel shows a histogram of the deviance residuals, with the best normal curve fit super-imposed. If the random component of the outcome indeed follows a gamma distribution, we would expect the histogram and the normal curve to be closely aligned. In this case, however, the histogram appears right-skewed relative to the normal curve, which suggests that the data exhibits greater skewness than what would be captured by a gamma distribution.

Another means of comparing the deviance residual distribution to normal is the *q-q* plot, shown on the right panel of Figure 16. In this plot, the theoretical normal quantile for each point is plotted on the *x*-axis, and the empirical (sample) quantile of the deviance residual is plotted on the *y*-axis. If the deviance residuals are indeed normal, the points should follow a straight line; a line passing through the 25 and 75 theoretical quantiles is shown for reference. We observe that at the edges of the distribution, the points lie above the line; in particular the right-most points deviate significantly upward, which means that there are many more high-valued deviance residuals than would be expected under a normal distribution. This indicates that the distribution of deviance residuals is more skewed than normal—and by extension, the data is more skewed than gamma—confirming what we observed in the histogram.

Based on these results, we may suppose that an inverse Gaussian distribution, which assumes greater skewness, may be more appropriate for this data than the gamma distribution. Figure 17 shows the diagnostic plots for the same model but with the

Figure 16. Graphical Comparisons of Deviance Residuals of a Gamma Model with the Normal Distribution

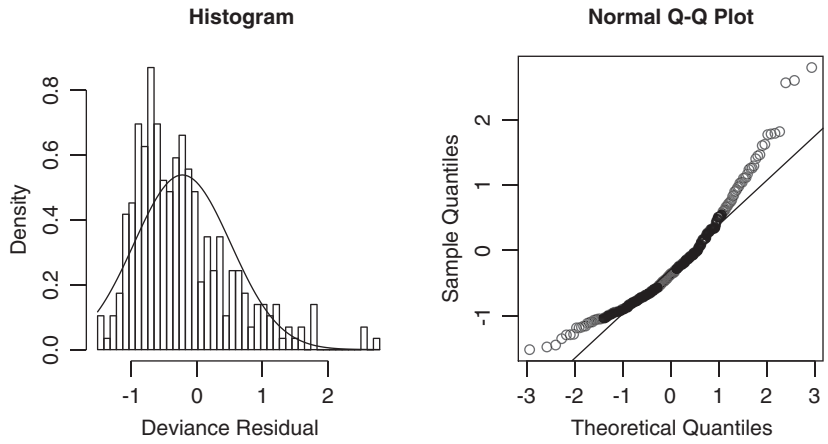
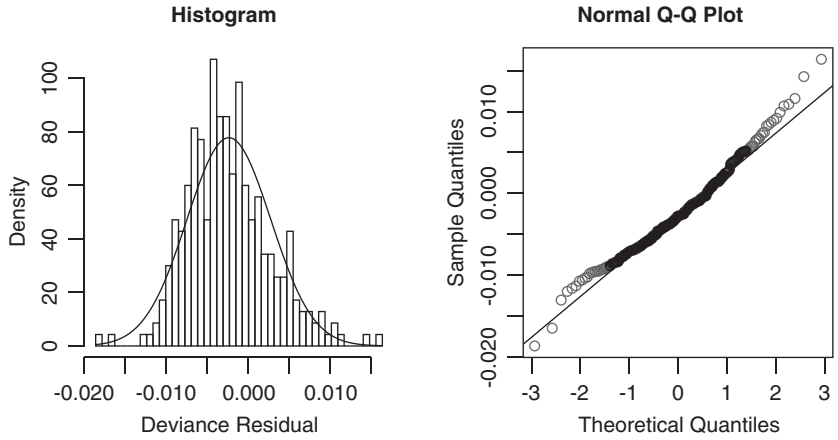


Figure 17. Graphical Comparisons of Deviance Residuals of the Inverse Gaussian Model with the Normal Distribution



assumption of inverse Gaussian as the underlying distribution. For this model, the histogram more closely matches the normal curve and the *q-q* plot better approximates the straight line, confirming that this model is indeed a better fit.

Discrete Distributions. For discrete distributions (such as Poisson or negative binomial) or distributions that otherwise have a point mass (such as Tweedie, which has a point mass at zero), the deviance residuals will likely *not* follow a normal distribution. This is because while the deviance residuals factor in the shape of the distribution, they do not adjust for the discreteness; the large numbers of records having the same target values cause the residuals to be clustered together in tight groups. This makes deviance residuals less useful for assessing the appropriateness of such distributions.

One possible solution is to use *randomized quantile residuals,* which have similar properties as deviance residuals, but add random jitter to the discrete points so that they wind up more smoothly spread over the distribution. Randomized quantile residuals are described in detail in Dunn and Smyth (1996).[13] Another possible solution is to use binned working residuals, as described in the next section.

Where discretely-distributed data is highly aggregated, such as for claims data where a single record may represent the average frequency for a large number of risks, the target variable will take on a larger number of distinct values, effectively "smoothing out" the resulting distribution. This causes the distribution to lose much of its discrete property and approach a continuous distribution, thereby making deviance residuals more useful for such data.

[13] In R, randomized quantile residuals are available via the qresiduals() function of the "statmod" package. Note, however, that for the Poisson distribution, randomized quantile residuals can only be calculated for the "true" Poisson distribution (with $\phi = 1$) but not the overdispersed Poisson; this diminishes their usefulness for most insurance data where "true" Poisson is unlikely to yield a good fit.

6.3.2. Working Residuals

Another useful type of residual is called the **working residual**. Most implementations of GLM fit the model using the Iteratively Reweighted Least Squares (IRLS) algorithm, the details of which are beyond the scope of this monograph. Working residuals are quantities that are used by the IRLS algorithm during the fitting process. Careful analysis of the working residuals is an additional tool that can be used to assess the quality of model fit.

Working residuals are defined as follows:

$$wr_i = (y_i - \mu_i) \cdot g'(\mu_i)$$

For a log link model, this simplifies to:

$$wr_i = \frac{y_i - \mu_i}{\mu_i}$$

For a logistic model, the working residual formula simplifies to:

$$wr_i = \frac{y_i - \mu_i}{\mu_i \cdot (1 - \mu_i)}$$

The main advantage of working residuals is that they solve a key problem that arises when visually inspecting the residuals via graphical methods, such a scatterplot. Such graphical plots are a highly useful means of detecting misspecifications or other shortcomings of a model. As noted above in the discussion of deviance residuals, any predictable pattern observed in the residuals indicates that the model could (and should) be improved, and a graphical analysis is an effective means of looking out for such patterns. However, most insurance models have thousands or even millions of observations, and the quantity being modeled is usually highly skewed. It can be difficult to identify predictable patterns in the dense clouds of skewed individual residuals, even for models with gross errors in specification.

Therefore, for such models, it is critical to **bin** the residuals before analyzing them. That is, we group the residuals by similar values of the *x*-axis of our intended plot, and aggregate (by averaging) both the *x*-axis values and the residuals prior to plotting. Binning the residuals aggregates away the volume and skewness of individual residuals, and allows us to focus on the signal. The advantage of *working* residuals is that they can be aggregated in a way that preserves the common properties of residuals – that is, they are unbiased (i.e., have no predictable pattern in the mean) and homoscedastic (i.e., have no pattern in the variance) for a well-fit model.[14]

[14] See the Appendix for the mathematical derivation of these properties.

To accomplish this, the working residuals are aggregated into bins, where each bin has a (roughly) equal sum of **working weights**. Working weights are defined as:[15]

$$ww_i = \frac{\omega_i}{V(\mu_i) \cdot [g'(\mu_i)]^2}$$

For each bin, the **binned working residual** is calculated by taking the weighted average of the working residuals of the individual observations within the bin, weighted by the working weights. Mathematically, for bin b, binned residual br_b is defined as:

$$br_b = \frac{\sum_{i \in b} wr_i \cdot ww_i}{\sum_{i \in b} ww_i}$$

If, in the course of graphically analyzing the working residuals over the different dimensions of the data, we are able to find a way to sort the working residuals into bins such that binned residuals appear to be predictably biased or "fanning out", then we have identified a shortcoming in the model specification. The following are several examples of binned working residual scatterplots that may be useful in revealing flaws in the model.

Plotting Residuals over the Linear Predictor. Plotting the residuals over the value of the linear predictor may reveal "miscalibrations" in the model—that is, areas of the prediction space where the model may be systematically under- or over-predicting. Figure 18 shows examples of such plots for two example models.

Both plots use binned working residuals; the underlying models have thousands of observations, but we have binned the working residuals into 100 bins prior to plotting. Thus, for example, the left-most point of each plot represents those observations with the lowest 1% of linear predictor values, and the x-axis and y-axis values for that point are the average linear predictor and average working residual for those observations, both averages weighted by the working weights as described above.

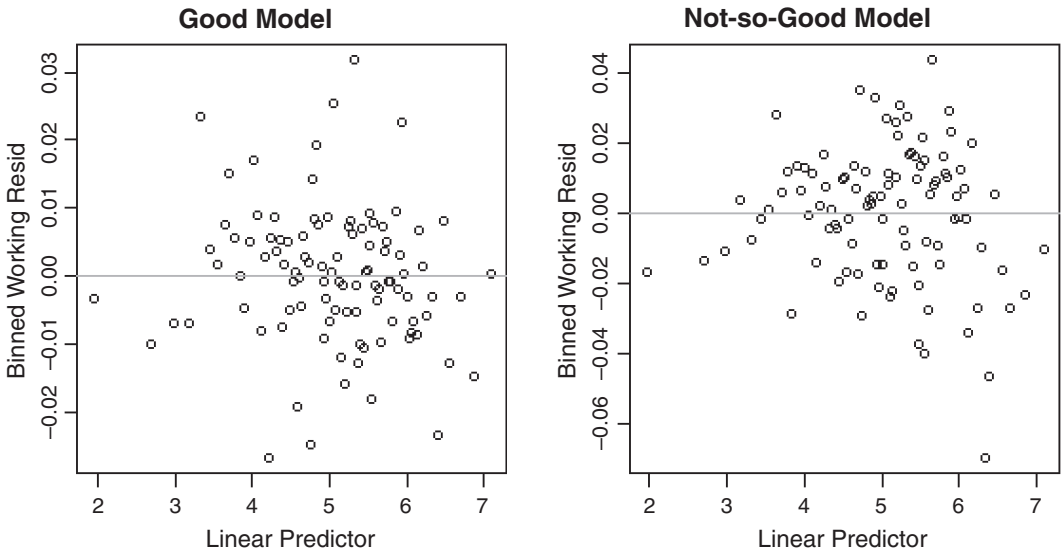
The left-hand plot of Figure 18 reveals no structural flaws in the model. The plot points form an uninformative cloud with no apparent pattern, as they should for a well-fit model.

The right-hand plot, on the other hand, shows signs of trouble. The residuals in the left region tend to be below the zero line, indicating that the model predictions for those

[15] The following table shows the simplification of this formula for several common model forms:

Distribution	Link function	Working Weights
Poisson	Log	$\omega_i \cdot \mu_i$
Gamma	Log	ω_i
Tweedie	Log	$\omega_i \cdot \mu_i^{2-p}$
Binomial	Logit	$\omega_i \cdot \mu_i \cdot (1 - \mu_i)$

Figure 18. Plotting Residuals over Linear Predictor



observations are higher than they should be. The model then seems to under-predict part of the middle region, and then once again over-predict for the highest-predicted observations. This may be caused by a non-linear effect that may have been missed, and the issue may be made clearer with plots of residuals over the various predictors, as discussed below.

Plotting Residuals over the Value of a Predictor Variable. While it is good practice to check partial residual plots (discussed in section 5.4.1) during the modeling process to understand the effect of responses and adjust as necessary, plots of residuals over the various predictors in the model may also reveal non-linear effects that may have been missed or not properly adjusted for.

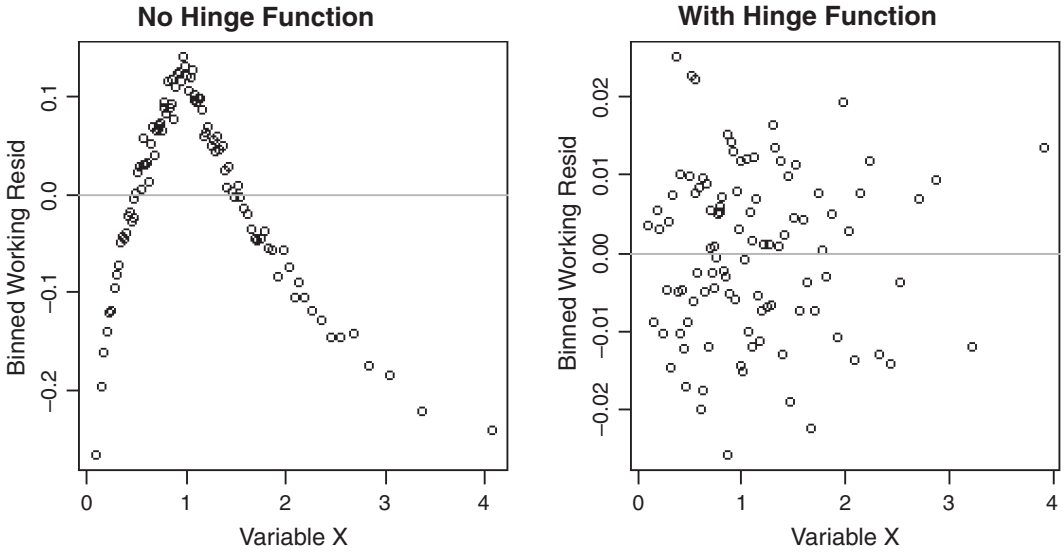
Figure 19 shows binned working residual plots over one of the predictor variables (labeled "Variable X") for two example models.

The left-hand plot clearly reveals that Variable X has a non-linear relationship with the target variable that is not being adequately addressed. The right-hand shows the plot that results after this issue had been fixed with a hinge function.

Plotting Residuals over the Weight. A plot of residuals over the weight variable used in the model (or over a variable that could potentially be a good choice of weight in the model) may reveal information about the appropriateness of the model weight (or lack thereof). Figure 20 shows plots of residuals over the number of exposures.

The model that generated the left-hand plot of Figure 20 did not use exposure as a weight in the model. This shows a "fanning out" effect on the left side, which violates the expectation of homoscedasticity, i.e., no pattern in the variance. Specifically, the lower-exposure records show more variance, and the higher-exposure records show less variance. This might be expected if no weight is specified; observations based on

Figure 19. Plotting Residuals over the Value of a Predictor Variable



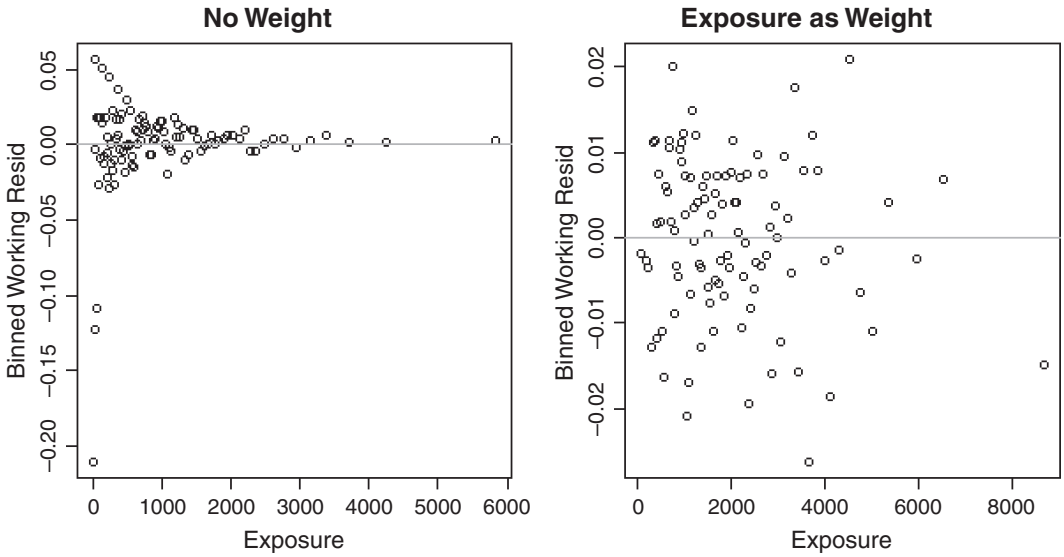
larger volume of exposure tend to be more stable (i.e., exhibit lower variance in the outcome) than lower-volume records, as discussed in Section 2.5.

The right-hand plot results after this issue is rectified by adding exposure as the weight in the model. In this model, the expectation of lower variance with higher exposure has already been assumed by the model, and so the residuals have this effect adjusted out, forming a homoscedastic cloud.

6.4. Assessing Model Stability

Model stability refers to the sensitivity of a model to changes in the modeling data. We assume that past experience will be a good predictor of future events, but small

Figure 20. Plotting Residuals over the Number of Exposures



changes in the past that we've observed should not lead to large changes in the future we predict. The classic example of this occurring is an unusually large loss experienced by an insured in a class with few members. A model run on all of the data may tell us with a high degree of confidence that this class is very risky. But if we remove the large loss from the dataset, the model may tell us with the same degree of confidence that the class is very safe. The model is not very stable with respect to the indication for this class, and so we may not want to give full weight to its results.

In the example above, the large loss is a particularly influential record, in that its removal from the dataset causes a significant change to our modeled results. Influential records tend to be highly weighted outliers. Assessing the impact of influential records is a straightforward way to assess model stability. A common measure of the influence of a record in GLM input data, calculable by most statistical packages, is **Cook's distance**. Sorting records in descending order of Cook's distance will identify those that have the most influence on the model results—a higher Cook's distance indicates a higher level of influence. If rerunning the model without some of the most influential records in the dataset causes large changes in some of the parameter estimates, we may want to consider whether or not those records or the parameter estimates they affect should be given full weight.

Another way to assess model stability is via cross validation, as described in Section 4.3.4 above. In that section, we presented cross validation as a means of testing the out-of-sample model performance. However, looking at the *in-sample* parameter estimates across the different model runs can yield important information about the stability of the model as well. The model should produce similar results when run on separate subsets of the initial modeling dataset.

Still another way to assess model stability is via **bootstrapping**. In bootstrapping, the dataset is randomly sampled with replacement to create a new dataset with the same number of records as the initial dataset. By refitting the model on many bootstrapped versions of the initial dataset, we can get a sense of how stable each of the parameter estimates are. In fact, we can calculate empirical statistics for each of the parameter estimates—mean, variance, confidence intervals, and so on—via bootstrapping. Many modelers prefer bootstrapped confidence intervals to the estimated confidence intervals produced by statistical software in GLM output.

7. Model Validation and Selection

Before diving into this section, some explanation is in order. As described above, the process of model refinement is really a process of creating two candidate models and comparing them. To put it another way, all model refinement involves model selection. But sometimes model selection can be used for goals other than model refinement. Sometimes a decision needs to be made between a number of alternate final models. If the best efforts of two modelers working independently are not identical (and they will never be), how is one to choose between them? The techniques discussed below are suitable for making this decision, while the techniques discussed in Chapter 6 are not. There are two key reasons for this:

First, one or more of the alternate models may be proprietary. Any rating plan is a model, and rating plans can come from all sorts of places: subsidiaries, consultants, competitor filings, rating bureaus, and so on. Most of the time, the data used to build these rating plans will not be available and neither will the detailed form of the underlying model. Even if this information is available, it might be impractical to evaluate it—the rating plan need not have been created using a GLM! The techniques discussed in Chapter 6 cannot be used under these circumstances, but the techniques below can. In order for the techniques below to be used, one only needs a database of historical observations augmented with the predictions from each of the competing models. The process of assigning predictions to individual records is called **scoring**.

Second, while the model refinement process is entirely technical, choosing between two final models is very often a business decision. Those responsible for making the final decision may know nothing about predictive modeling or even nothing about actuarial science. The techniques below compare the performance of competing models in a way that is accessible.

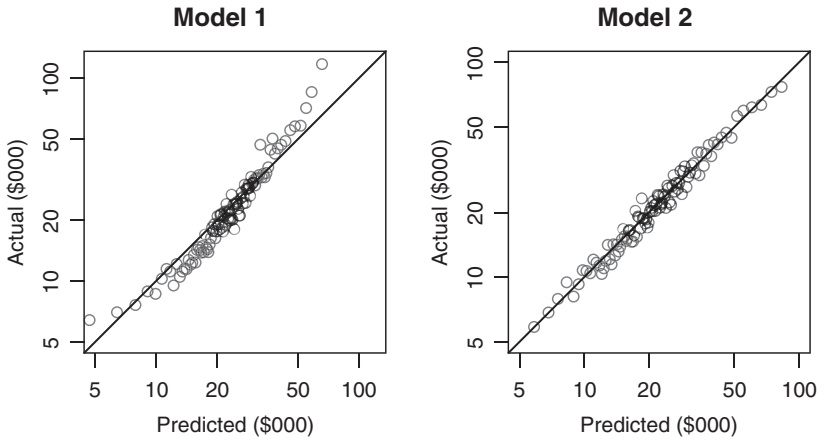
Some of the techniques below can also be used for model refinement, to the extent that they produce new data or insights that can be acted on.

7.1. Assessing Fit with Plots of Actual vs. Predicted

A very simple and easily understandable diagnostic to assess and compare the performance of competing models is to create a plot of the actual target variable (on the y-axis) versus the predicted target variable (on the x-axis) for each model. If a model fits well, then the actual and predicted target variables should follow each other closely.

Consider Figure 21, which shows plots of actual vs. predicted target variables for two competing models.

Figure 21. Actual vs. Predicted Plots for Two Competing Models



From these charts, it is clear that Model 2 fits the data better than Model 1, as there is a much closer agreement between the actual and predicted target variables for Model 2 than there is for Model 1.

There are three important cautions regarding plots of actual versus predicted target.

First, it is important to create these plots on holdout data. If created on training data, these plots may look fantastic due to overfitting, and may lead to the selection of a model with little predictive power on unseen data.

Second, it is often necessary to aggregate the data before plotting, due to the size of the dataset. A common approach is to group the data into percentiles. The dataset is first sorted based on the predicted target variable, then it is grouped into 100 buckets such that each bucket has the same aggregate model weight. Finally, the averages of the actual and predicted targets within each bucket are calculated and plotted, with the actual values on the *y*-axis and the predicted values on the *x*-axis.

Third, it is often necessary to plot the graph on a log scale, as was done in Figure 20. Without this transformation, the plots would not look meaningful, since a few very large values would skew the picture.

7.2. Measuring Lift

Broadly speaking, model lift is the economic value of a model. The phrase "economic value" doesn't necessarily mean the profit that an insurer can expect to earn as a result of implementing a model, but rather it refers to a model's ability to prevent adverse selection. The lift measures described below attempt to visually demonstrate or quantify a model's ability to charge each insured an actuarially fair rate, thereby minimizing the potential for adverse selection.

Model lift is a relative concept, so it requires two or more competing models. That is, it doesn't generally make sense to talk about the lift of a specific model, but rather the lift of one model over another.

In order to prevent overfitting, model lift should always be measured on holdout data.

7.2.1. Simple Quantile Plots

Quantile plots are a straightforward visual representation of a model's ability to accurately differentiate between the best and the worst risks. Assume there are two models, Model A and Model B, both of which produce an estimate of the expected loss cost for each policyholder. Simple quantile plots are created via the following steps:

1. Sort the dataset based on the Model A predicted loss cost (from smallest to largest).
2. Bucket the data into quantiles, such that each quantile has the same volume of exposures. Common choices are quintiles (5 buckets), deciles (10 buckets), or vigintiles (20 buckets).
3. Within each bucket, calculate the average predicted pure premium (predicted loss per unit of exposure) based on the Model A predicted loss cost, and calculate the average actual pure premium.
4. Plot, for each quantile, the actual pure premium and the pure premium predicted by Model A.
5. Repeat steps 1 through 4 using the Model B predicted loss costs. There are now two quantile plots—one for Model A and one for Model B.
6. Compare the two quantile plots to determine which model provides better lift.

In order to determine the "winning" model, consider the following 3 criteria:

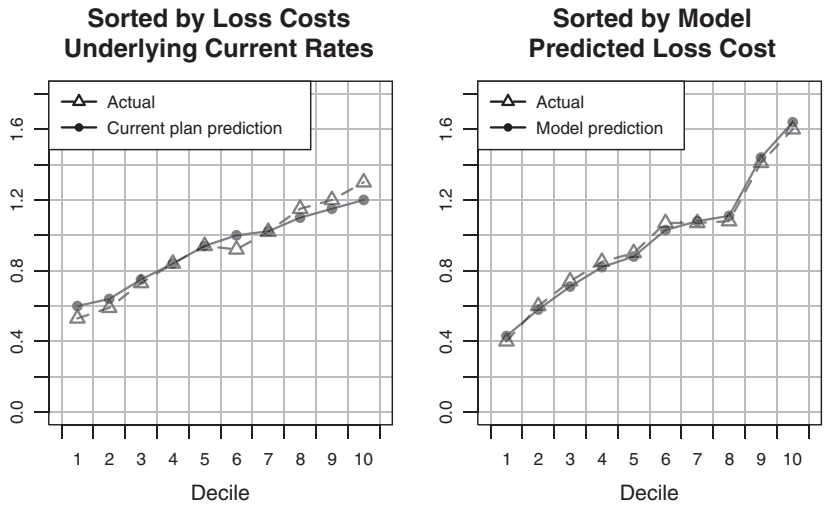
1. **Predictive accuracy.** How well each model is able to predict the actual pure premium in each quantile.
2. **Monotonicity.** By definition, the predicted pure premium will monotonically increase as the quantile increases, but the actual pure premium should also increase (though small reversals are okay).
3. **Vertical distance between the first and last quantiles.** The first quantile contains the risks that the model believes will have the best experience, and the last quantile contains the risks that the model believes will have the worst experience. A large difference (also called "lift") between the actual pure premium in the quantiles with the smallest and largest predicted loss costs indicates that the model is able to maximally distinguish the best and worst risks.

Figure 22 shows simple decile plots for an example comparison between the current rating plan (left panel) and a newly-constructed plan (right panel). For ease of interpretation, both the actual and predicted loss costs for each graph have been divided by the average model predicted loss cost.

In both plots, the solid line is the predicted loss cost (either by the current rating plan or by the new model) and the broken line is the actual loss cost. Which model is better?

1. *Predictive accuracy*—for the right panel graph, the plotted loss costs correspond more closely between the two lines than for the left panel graph, indicating that the new model seems to predict actual loss costs better than the current rating plan does.
2. *Monotonicity*—the current plan has a reversal in the 6th decile, whereas the model has no significant reversals.

Figure 22. Simple Decile Plots for Both the Current Rating Plan (*left panel*) and for a Newly-Constructed Plan (*right panel*)



3. *Vertical distance between the first and last quantiles*—the spread of actual loss costs for the current manual is 0.55 to 1.30, which is not very much. That is, the best risks have loss costs that are 45% below the average, and the worst risks are only 30% worse than average. The spread of the proposed model though is 0.40 to 1.60.

   Thus, by all three metrics, the new plan outperforms the current one.

7.2.2. Double Lift Charts

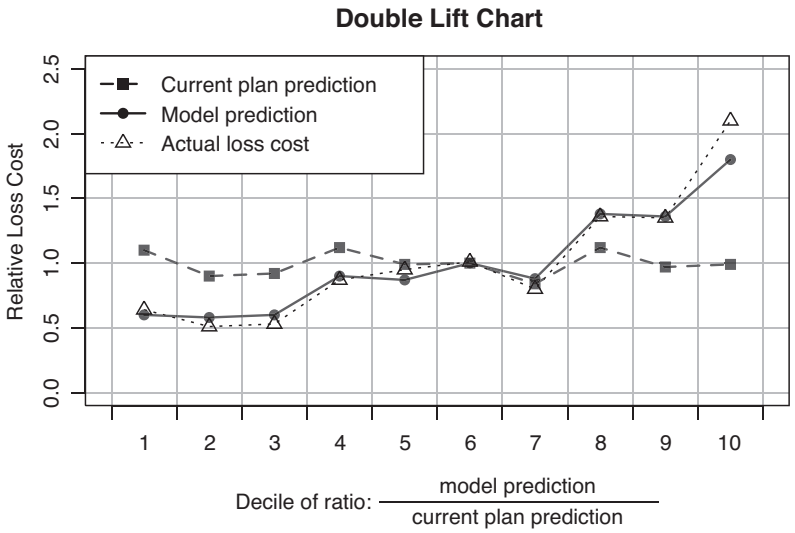
A double lift chart is similar to the simple quantile plot, but it directly compares two models. Assume that there are two models, Model A and Model B, both of which produce an estimate of the expected loss cost for each policyholder. A double lift chart is created via the following steps:

1. For each record, calculate Sort Ratio = (Model A Predicted Loss Cost)/(Model B Predicted Loss Cost).
2. Sort the dataset based on the Sort Ratio, from smallest to largest.
3. Bucket the data into quantiles, such as quintiles or deciles.
4. Within each bucket, calculate the Model A average predicted pure premium, the Model B average predicted pure premium, and the actual average pure premium. For each of those quantities, divide the quantile average by the overall average.
5. For each quantile, plot the three quantities calculated in the step above.

In a simple quantile plot, the first quantile contains those risks which Model A thinks are best. In a double lift chart, the first quantile contains those risks which Model A thinks are best *relative to Model B.* In other words, the first and last quantiles contain those risks on which Models A and B disagree the most (in percentage terms).

In a double lift chart, the "winning" model is the one that more closely matches the actual pure premium in each quantile. To illustrate this, consider the example double

Figure 23. A Sample Double Lift Chart



lift chart in Figure 23, in which we use a double lift chart to compare a proposed rating model to the current rating plan.

The solid line shows the loss costs predicted by the model, the thick broken line shows the loss costs in the current rating plan, and the dotted line shows the actual loss costs. The sort order for this graph is the model prediction divided by the current plan prediction, and the data is segmented into deciles.

It is clear that the proposed model more accurately predicts actual pure premium by decile than does the current rating plan. Specifically, consider the first decile. It contains the risks that the model thinks are best relative to the current plan. As it turns out, the model is correct. Similarly, in the 10th decile, the model more accurately predicts pure premium than does the current plan.

As an alternate representation of a double lift chart, one can plot two curves—the percent error for the model predictions and the percent error for the current loss costs, where percent error is calculated as (Predicted Loss Cost)/(Actual Loss Cost) − 1. In this case, the winning model is the one with the flatter line centered at $y = 0$, indicating that its predictions more closely match actual pure premium.

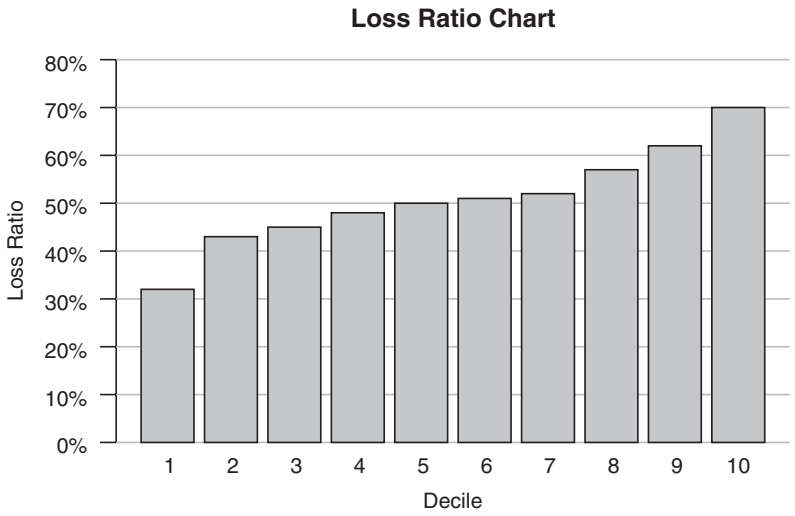
7.2.3. Loss Ratio Charts

In a loss ratio chart, rather than plotting the pure premium for each bucket, the loss ratio is instead plotted. The steps for creating a loss ratio chart are very similar to those for creating a simple quantile plot:

1. Sort the data based on the predicted loss ratio (= [predicted loss cost]/premium).
2. Bucket the data into quantiles, such that each quantile has the same volume of exposures.
3. Within each bucket, calculate the *actual loss ratio* for risks within that bucket.

Ideally, the model is able to identify deficiencies in the current rating program by segmenting the risks based on loss ratio. To illustrate this, consider Figure 24. If a

Figure 24. A Sample Loss Ratio Chart



rating plan is perfect, then all risks should have the same loss ratio. The fact that this model is able to segment the data into lower and higher loss ratio buckets is a strong indicator that it is outperforming the current rating plan.

The advantage of loss ratio charts over quantile plots and double lift charts is that they are simple to understand and explain. Loss ratios are the most commonly-used metric in determining insurance profitability, so all stakeholders should be able to understand these plots.

7.2.4. The Gini Index

The Gini index, named for statistician and sociologist Corrado Gini, is commonly used in economics to quantify national income inequality.

The national income inequality Gini index is calculated as follows:

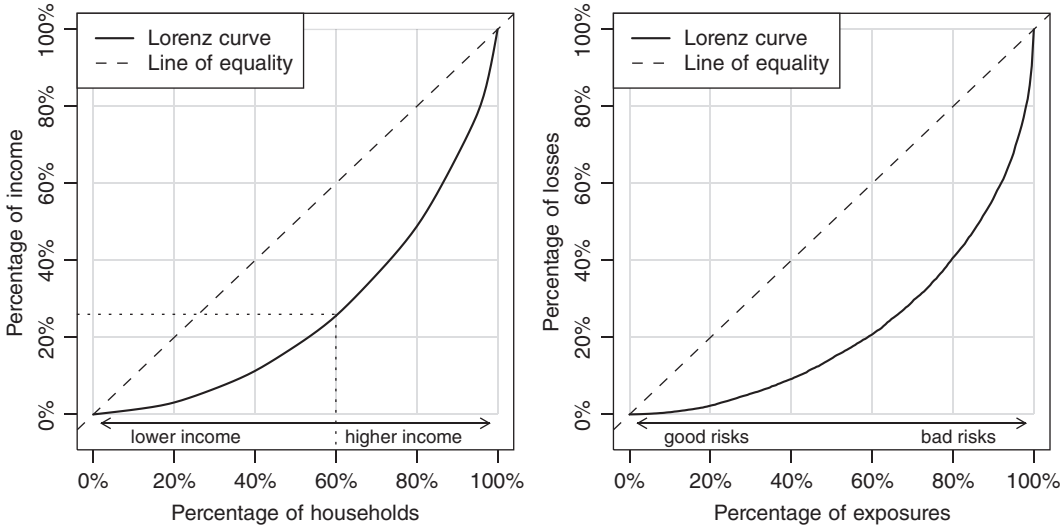
1. Sort the population based on earnings, from those with the lowest earnings to those with the highest earnings. (This could also be done based on wealth rather than earnings.)
2. The *x*-axis is the cumulative percentage of earners.
3. The *y*-axis is the cumulative percentage of earnings.

The locus of points created by plotting the cumulative percentage of earnings against the cumulative percentage of earners is called the **Lorenz curve**. The left panel of Figure 25 plots the Lorenz curve for year 2014 household income in the United States.[16]

The 45-degree line is called the line of equality, so named because, if every person earned the same exact income, then the Lorenz curve would be the line of equality. In that hypothetical scenario, if there are 100 people in the society, then each represents 1% of the population and would earn 1% of the income. Everyone doesn't earn the same income, though, so the Lorenz curve is bow-shaped. As the graph shows, the

[16] Source: https://www.census.gov/hhes/www/income/data/historical/household/

Figure 25. Gini Index Plots of United States 2014 Household Income (*left panel*) and a Sample Pure Premium Model (*right panel*)



poorest 60% of households earn roughly 25% of the total income. The Gini index is calculated as twice the area between the Lorenz curve and the line of equality. (In 2014, that number was 48.0%).

The Gini index can also be used to measure the lift of an insurance rating plan by quantifying its ability to segment the population into the best and worst risks. The Gini index for a model which creates a rating plan is calculated as follows:

1. Sort the dataset based on the model predicted loss cost. The records at the top of the dataset are then the risks which the model believes are best, and the records at the bottom of the dataset are the risks which the model believes are worst.
2. On the *x*-axis, plot the cumulative percentage of exposures.
3. On the *y*-axis, plot the cumulative percentage of losses.

The locus of points is the Lorenz curve, and the Gini index is twice the area between the Lorenz curve and the line of equality.

The right panel of Figure 25 plots a sample Lorenz curve for a sample pure premium model. As can be seen, this model identified 60% of exposures which contribute only 20% of the total losses. Its Gini index is 56.1%.

Note that a Gini index does not quantify the profitability of a particular rating plan, but it does quantify the ability of the rating plan to differentiate the best and worst risks. Assuming that an insurer has pricing and/or underwriting flexibility, this will lead to increased profitability.

7.3. Validation of Logistic Regression Models

For logistic regression models (discussed in Section 2.8), the GLM yields a prediction of the probability of the occurrence of the modeled event. Many of the model validation diagnostics discussed in the previous sections can be applied to such models as well. For example, a quantile plot can be created by bucketing records of the

holdout set into quantiles of predicted probability and graphing the actual proportion of positive occurrences of the event within each quantile against the average predicted probability; a good model will yield a graph exhibiting the properties of accuracy, monotonicity and vertical distance between first and last quantiles, as described in Section 7.2.1. Similarly, a Lorenz curve can be created by sorting the records by predicted probability and graphing cumulative risks against cumulative occurrences of the event, and a Gini index can be computed from the resulting graph by taking the area between the curve and the line of equality.

For such models, a diagnostic called the *receiver operating characteristic* curve, or *ROC* curve, is commonly used due its direct relation to how such models are often used in practice, as discussed in the following section.

7.3.1. Receiver Operating Characteristic (ROC) Curves

While a logistic model predicts the *probability* of an event's occurrence, for many practical applications that probability will need to be translated into a binary prediction of occurrence vs. non-occurrence for the purpose of deciding whether to take a specific action in response. For example, suppose we build a model to detect claims fraud; for each new claim, the model yields a probability that it contains fraud. Based on this prediction, we will need to decide whether or not to assign a team to further investigate the claim.

We can make such a determination by choosing a specific probability level, called the *discrimination threshold*—say, 50%—above which we will investigate the claim and below which we will not. This determination may be thought of as the model's "prediction" in a binary (i.e., fraud/no fraud) sense.

Under this arrangement, for any claim, the following four outcomes are possible:

1. The model predicts that the claim contains fraud (that is, $\mu_i > 0.50$), and the claim is indeed found to contain fraud. This outcome is called a *true positive.*
2. The model predicts fraud, but the claim does not contain fraud (i.e., a *false positive*).
3. The model predicts no fraud (i.e., $\mu_i < 0.50$), but the claim contains fraud (i.e., a *false negative*).
4. The model predicts no fraud, and the claim does not contain fraud (i.e., a *true negative*).

Outcome #1—the true positive—clearly represents a success of the model, as it correctly identifies a fraudulent claim, thus preventing unnecessary payment and saving the company money. Outcome #4, the true negative, while not as dramatic, similarly has the model doing its job by not sending us on a wild-goose chase.

Outcomes #2 and #3—the false positive and false negative—are failures of the model, and each comes with a cost. The false negative allows a fraudulent claim to slip by undetected, resulting in unnecessary payment. The false positive also incurs a cost in the form of unnecessary resources expended on a claims investigation as well as possible impairment of goodwill with the insured.

If the model were perfect—that is, it would predict a probability of 0% for each non-fraud and 100% for each fraud—then the true positive and true negative would be the only possible outcomes, regardless of the threshold chosen. For real-life models, on the other hand, false negatives and false positives are possible, and selection of the

discrimination threshold involves a trade-off: a lower threshold will result in more true positives and fewer false negatives than a higher threshold, but at the cost of more false positives and fewer true negatives.

We can assess the relative likelihoods of the four outcomes for a given model and for a specified discrimination threshold using a test set. We use the model to score a predicted probability for each test record, and then convert the predictions of probability into binary (yes/no) predictions using the discrimination threshold. We then group the records by the four combinations of actual and predicted outcomes, and count the number of records falling into each group. We may display the results in a 2×2 table called a *confusion matrix.* The top panel of Table 13 shows an example confusion matrix for a claims fraud model tested on a test set that contains 813 claims, using a discrimination threshold of 50%.

The ratio of true positives to total positive events is called the **sensitivity**; in this example, that value is 39/109 = 0.358. This ratio, also called the *true positive rate* or the *hit rate,* indicates that with a threshold of 50%, we can expect to catch 35.8% of all fraud cases.

The ratio of true negatives to total negative events is called the **specificity**, and is 673/704 = 0.956 in our example. The complement of that ratio, called the *false positive rate,* is 1 − 0.956 = 0.044. This indicates that the hit rate of 35.8% comes at the cost of also needing to investigate 4.4% of all non-fraud claims.

We may wish to catch more fraud by lowering the threshold to 25%. The bottom panel of Table 13 shows the resulting confusion matrix. As can be seen, the hit rate under this arrangement improves to 75/109 = 68.8%—but it comes at the cost of an increase in the false positive rate to 103/704 = 14.6%.

Table 13. Confusion Matrices for Example Fraud Model With Discrimination Thresholds of 50% (*top*) and 25% (*bottom*)

Discrimination Threshold: 50%

	Predicted				
Actual	Fraud		No Fraud		Total
Fraud	*true pos.:*	39	*false neg.:*	70	109
No Fraud	*false pos.:*	31	*true neg.:*	673	704
Total		70		743	813

Discrimination Threshold: 25%

	Predicted				
Actual	Fraud		No Fraud		Total
Fraud	*true pos.:*	75	*false neg.:*	34	109
No Fraud	*false pos.:*	103	*true neg.:*	601	704
Total		178		635	813

A convenient graphical tool for evaluating the range of threshold options available to us for any given model is the **receiver operating characteristic curve**, or **ROC curve**, which is constructed by plotting the false positive rates along the *x*-axis and the true positive rates along the *y*-axis for different threshold values along the range [0,1]. Figure 26 shows the ROC curve for our example claims fraud model.

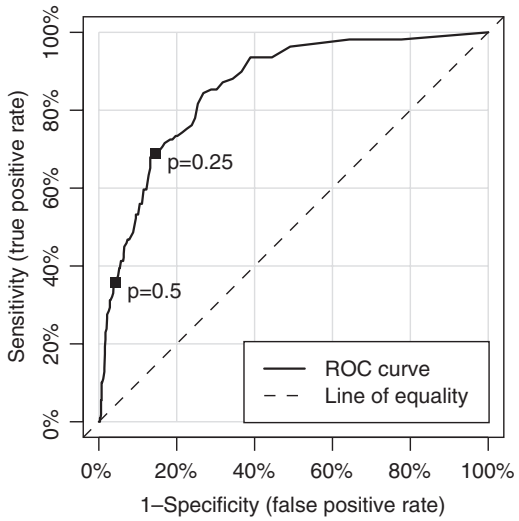
The (0, 0) point of this graph represents a threshold of 100%, with which we catch no fraudulent claims (but investigate no legitimate claims either). Moving rightward, we see that lowering the threshold and thereby incurring some false positives yields large gains in the hit rate; however, those gains eventually diminish for higher false positive rates. The two example thresholds detailed in Table 13 are plotted as points on the graph.

The ROC curve allows us to select a threshold we are comfortable with after weighing the benefits of true positives against the cost of false positives. Different thresholds may be chosen for different claim conditions—for example, we may choose a lower threshold for a large claim where the cost of undetected fraud is higher. Determination of the optimal threshold is typically a business decision that is out of the scope of the modeling phase.

The level of accuracy of the model, though, will affect the severity of the trade-off. A model that yields predictions that are no better than random will yield true positives and false positives in the same proportions as the overall mix of positives and negatives in the data, regardless of the threshold chosen. Therefore, for such a model, the ROC curve will follow the line of equality. A model with predictive power will yield true positives at a higher rate than false positives, resulting in a ROC curve that is higher than the line of equality. Improved accuracy of the model will move the ROC curve farther from equality, indicating that the model allows us a better hit rate for any level of false positive cost.

The model accuracy as indicated by the ROC curve can be summarized by taking the area under the curve, called the **AUROC** (for "area under ROC"). A model with no

Figure 26. ROC Curve for Example Fraud Model



predictive power will yield an AUROC of 0.500. The ROC curve of the hypothetical "perfect" model described earlier will immediately rise to the top of the graph (as any threshold below 100% would correctly identify all fraud cases and trigger no false positives), thereby yielding an AUROC of 1.000. The ROC curve of our example model plotted in Figure 26 yields an AUROC of 0.857.

Note, however, that the AUROC measure bears a direct relationship to the Gini index discussed in the previous section, such that one can be derived from the other.[17] As such, AUROC and the Gini index should not be taken as separate validation metrics, since an improvement in one will automatically yield an improvement in the other.

[17] Specifically, the AUROC is equal to 0.5 × *normalized Gini* + 0.5, where *normalized Gini* is the ratio of the model's Gini index to the Gini index of the hypothetical "perfect" model (where each record's prediction equals its actual value).

8. Model Documentation

8.1. The Importance of Documenting Your Model

Model documentation is important enough, and overlooked enough, that it deserves its own section. This section comes with some unsolicited career advice which we hope will be helpful even for those of you who don't build models as part of your day job.

Model documentation serves at least three purposes:

- To serve as a check on your own work, and to improve your communication skills
- To facilitate the transfer of knowledge to the next owner of the model
- To comply with the demands of internal and external stakeholders

If you're a credentialed actuary working in the United States, all of the documentation you produce should comply with ASOP 41 on Actuarial Communications.

8.2. Check Yourself

Actuarial work tends to be complex; modeling work, even more so. You're going to make mistakes. No matter how smart you are, no matter how experienced you are, no matter how brilliant or elegant your work product appears to be—it's more likely than not that you've overlooked something. We're all just human here and that's just how it goes. As an actuary, you're obliged to own up to the mistakes that you make. The first time that someone discovers you sweeping a mistake under the rug is the last time that anyone will trust you to do anything. The better you are at identifying and correcting mistakes you've made in your own work, the easier your life will be. If you want to succeed in your career you'd be well-served to internalize this.

So how are you supposed to find mistakes in your own work? One of the best ways is to *try to explain what you've done in writing*. When you write down what you've done in a way that allows someone else to understand it, you're forced to revisit your work in full detail, and to think critically about all of the decisions you made along the way. This has a way of bringing errors (especially conceptual errors) to the surface. This is especially true when you share your documentation with others. It may be easy for a peer to identify a conceptual error in a narrative that they would not have been able to detect in a package of code.

Another benefit of documentation is that it serves to reinforce your understanding of the work that you're documenting. It's been said that "to teach is to learn twice over".

This is true! The level of understanding required to document or explain or teach a topic is greater than the level of understanding required to simply execute. When you start from a foundation of deeper understanding, your subsequent work product will be of higher quality, and will stand up better to scrutiny. This means that you should *document your work as you go*. Documentation isn't a task for the end of the project, so that you discover mistakes when you no longer have time to address them. On the contrary, it's a task for *right this minute*, so that in your very next project meeting, you'll be able to field questions that no one else has even thought of yet.

A final benefit of documentation for you, personally, is that it serves to improve your communication skills. There is nothing more important to an actuary than their ability to communicate. Our *work* may involve any number of complex statistical analyses, but our *work product* is always a report to someone else that details the work we've done. Your ability to communicate will become more important as you progress through your career, as you will find yourself increasingly responsible for presenting to stakeholders who are not also actuaries. The Casualty Actuarial Society doesn't have an exam on communication. If you'd like to improve in this area, you're going to have to find a way to do it yourself. An easy way to do this is to force yourself to document the things that you do in such a way that a non-technical person can follow along.

Nothing in this section is hypothetical. The authors of this monograph are actuaries, just like you, not too many years removed from taking exams. This monograph is a form of documentation and we've become better actuaries by writing it. (And yes, we've made our share of errors as well.)

8.3. Stakeholder Management

Every modeling project you work on will eventually come to an end, but as discussed in Section 3.9, models will need to be maintained and rebuilt. The tasks of maintaining and rebuilding the model may fall to someone else, or they may fall to you. In either case, good documentation will make everyone's lives easier. Even if you retain ownership of the model forever, we can tell you from experience that it's easy to forget important details of a project after only a few months of not working on it. Creating good documentation now will make life easier for you in the future.

Others may develop an interest in the models that you build, either now or in the future. Insurance is a highly-regulated field, and there's a good chance that regulators will have questions for you, either during the filing process or during a regular examination. Internal and outside auditors and risk managers tend to have a keen interest in models and their documentation. And in a large organization, any number of internal stakeholders—including executives, underwriters, claims adjusters, other actuaries, and IT personnel—may eventually come calling with detailed questions on work that may have been done months or years ago. In all of these cases, we can tell you from experience that it's easier to have good documentation on hand ready to send to anyone who asks for it than it is to try to answer questions from first principles when your memory of what you've done may be a little fuzzy.

To meet the needs of these stakeholders, your documentation should:

- Include everything needed to reproduce the model from source data to model output
- Include all assumptions and justification for all decisions
- Disclose all data issues encountered and their resolution
- Discuss any reliance on external models or external stakeholders
- Discuss model performance, structure, and shortcomings
- Comply with ASOP 41 or local actuarial standards on communications

8.4. Code as Documentation

Your model code serves as a form of documentation. Your code should be clearly written and commented. Moreover, it should be easy to differentiate the "production" version of your code from any draft work that led up to it. If you use R, you should use the "tidyverse" package and adhere to the tidyverse style guide.[18] Even if you don't use R, we recommend that you give this style guide a read, as the philosophies that it espouses are more or less universal can be applied to work done on any platform.

[18] We recognize that other packages, such as data.table, may be more appropriate than tidyverse packages in some situations. However, it is generally not advisable to use base R for functionality that has been implemented in more advanced packages such as tidyverse.

9. Other Topics

9.1. Modeling Coverage Options with GLMs (Why You Probably Shouldn't)

The policy variables included in a rating plan can be broadly categorized into two types: characteristics of the insured or insured entity, such as driver age or vehicle type for auto liability, building construction type or territory for homeowners insurance, or industry classification for general liability; and options selected by the insured, such as deductible, limit, or peril groups covered.

When using GLMs to formulate such rating plans, it is tempting to try and estimate factors for coverage options by simply throwing those variables in with the rest in the GLM—only to sometimes discover that GLM produces seemingly counterintuitive results. For example, consider the case of the deductible factor. When including deductible as a categorical variable in a pure premium GLM—setting the basic deductible as the base level—it is not uncommon for the GLM to produce a positive coefficient (indicating a factor above unity) for a deductible *higher* than the base deductible. This result—and a highly significant one, to boot!—would seem to indicate that more premium should be charged for less coverage, and may leave actuaries scratching their heads. What gives?[19]

The answer may lie in the basic statistical truth that correlation does not imply causation. A coefficient estimated by a GLM need not be the result of a causal effect between the predictor and the target; there may be some latent variables or characteristics not captured by our model that may correlate with the variable in question, and those effects may influence the model result. In the case of deductible, there may be something systematic about insureds with higher deductibles that may make them a worse risk relative to others in their class. Possibilities of how this may arise are:

- The choice of high deductible may be the result of a high risk appetite on the part of an insured, which would manifest in other areas as well.
- The underwriter, recognizing an insured as a higher risk, may have required the policy to be written at a higher deductible.

[19] We note that while a positive indication for a higher deductible may be considered counterintuitive in a frequency or pure premium model, in a severity model it is to be expected. This is because despite the deductible eliminating a portion of each loss, thereby lowering the numerator of severity, the deductible also eliminates many small claims, lowering the denominator of severity. As the latter effect is usually stronger than the former, the total effect of deductible on severity is most often positive.

Thus, the coefficients estimated by the GLM may be reflecting some of this increased risk due to such selection effects.

Counterintuitive results such as these have led some to believe that GLMs "don't work" for deductibles. This may not be a fair characterization; the factors estimated by the GLMs may very well be predictive—*if* the goal is to predict loss for an *existing* set of policies. But that isn't usually our objective; rather, we are trying to estimate the pricing that would make sense for policies sold in the future.

To be sure, for most *other* variables, potential correlation with a latent variable is not a bad thing; if a variable we have collected also yields some information about one we haven't, all the better.[20] However, where the variable in question relates to a policy option selected by the insured, having its factor reflect anything other than pure loss elimination would not be a good idea. Even if the indicated result is not something as dramatically bad as charging more premium for less coverage, to the extent that the factor differs from the pure effect on loss potential, it will affect the way insureds choose coverage options in the future. Thus, the selection dynamic will change, and the past results would not be expected to replicate for new policies.

For this reason it is recommended that factors for coverage options—deductible factors, ILFs, peril group factors and the like—be estimated outside the GLM, using traditional actuarial loss elimination techniques. The resulting factors should then be included in the GLM as an offset.

9.2. Territory Modeling

Territories are not a good fit for the GLM framework. Unlike other variables you might consider in your model, which are either continuous or can easily be collapsed into a manageable number of levels, you may have hundreds or thousands or hundreds of thousands of territories to consider—and aggregating them to a manageable level will cause you to lose a great deal of important signal.

So the solution is to use other techniques, such as spatial smoothing, to model territories. Discussion of these techniques is beyond the scope of this monograph. But in creating a classification plan, you must still be aware of and have access to the output of these models. Since there are usually many complicated relationships between territory and other variables, your GLM should still consider territory. This is accomplished by including territory in your model as an offset. Offsetting for territory only requires populating policy records with their indicated territory loss cost (taken from the standalone model). This way, your classification plan variables will be fit after accounting for territorial effects, and so will not proxy for them.

But, of course, it's a two-way street. Just as your classification plan model should be offset for territory loss costs, so too should the territory model be offset for the classification plan. So the process is iterative—both models should be run, using the other as an offset, until they reach an acceptable level of convergence. In theory this can

[20] An important exception is where a variable included in a model may correlate with a protected class or any other variable that may not be rated on. In such instances, the actuary must take care to ensure that the model is in accordance with all regulatory requirements and actuarial standards of practice.

be done in one pass, but in practice these models may be updated at different times and by different groups of people, so convergence may only set in over a period of years.

9.3. Ensembling

Consider this scenario: your company assigns its two top predictive modelers, Alice and Bob, to develop a Homeowners pure premium model, and advises them to each work independently off the same set of data.

They get to work, and, after some time, each proposes their finished model. Naturally—since there is no one "right" way to build a model—the models are somewhat different: each has variables selected for inclusion that the other does not have; some continuous variables have been bucketed in one model while having been transformed using polynomial functions in the other; and so on. However, when testing the models, they both perform equally well: the loss ratio charts and double lift charts both show the same improvement over the existing plan, and calculating Gini indices on the holdout set and in cross validation yields very similar results between the two. We now need to make a decision: which model is better—Alice's or Bob's?

The answer, most likely, is: both. Combining the answers from both models is likely to perform better than either individually.

A model that combines information from two or more models is called an **ensemble** model. There are many strategies for combining models, and a full treatment of the subject is beyond the scope of this text. However, a simple, yet still very powerful, means of ensembling is to simply take the straight average of the model predictions.[21] Two well-built models averaged together will almost always perform better than one, and three will perform even better—a phenomenon known as the *ensemble effect.* Generally, the more models the better, though subject to the law of diminishing returns. In fact, ensembling is one notable exception to the parsimony principle in modeling (i.e., the "keep it simple" rule); adding more models to an ensemble—thereby increasing the complexity—will rarely make a model worse.

An interesting example of the ensemble effect in the real world is the "guess the number of jelly beans in the jar" game sometimes used for store promotions. In this game, any individual's guess is likely to be pretty far off from the right answer; however, it is often observed that taking the *average* of all the submitted guesses will yield a result that is very close to correct. As individuals, some people guess too high and some guess low, but *on average* they get it right.

Predictive models, like people, each have their strengths and weaknesses. One model may over-predict on one segment of the data while under-predicting on another; a different model is not likely to have the same flaws but may have others. Averaged together, they can balance each other out, and the gain in performance can be significant.

---

[21] If both models are log-link GLMs, the multiplicative structure of the resulting ensemble can be preserved by taking the *geometric* average of the predictions. Equivalently, one can construct multiplicative factor tables that use the geometric averages of the individual model factors. (When doing so, for any variable present in one model but absent in the other, use a factor of 1.00 for the model in which it is absent.)

One caveat though—for the ensemble effect to work properly, the model errors should be as uncorrelated as possible; that is, the models shouldn't all be systematically missing in the same way. (Much as the averaged jelly bean guesses would not work well if everyone guessed similarly.) Thus, if ensembling is to be employed as a model-building strategy, it is best if the models are built by multiple people or teams working *independently,* with little or no sharing of information. Done properly, though, ensembles can be quite powerful; if resources permit, it may be worth it.

10. Variations on the Generalized Linear Model

As we have seen in the preceding sections, the GLM is a flexible, robust and highly interpretable model that can accommodate many different types of target variables and covariate relationships. However, it does have a number of shortcomings, most notably:

- Predictions must be based on a linear function of the predictors. Certainly, there are workarounds to handle non-linearity (such as polynomials or hinge functions) but those must be explicitly specified by the modeler.
- GLMs exhibit instability in the face of thin data or highly correlated predictors.
- Full credibility is given to the data for each coefficient, with no regard to the thinness on which it is based.
- GLMs assume the random component of the outcome is uncorrelated among risks.
- The exponential family parameter ϕ must be held constant across risks.

Many of the more advanced predictive modeling techniques used by data scientists in other disciplines, such as *neural nets, random forests* or *gradient boosting machines,* do not have these flaws, and are therefore able to produce stronger models that yield more accurate predictions. However, using those methods would entail a huge loss of interpretability, which, for many actuarial applications, is as great a necessity as predictive accuracy, if not greater.

Fortunately, a number of extensions to GLMs have been developed that address some of the limitations noted above. We *briefly* discuss some of them in this section. As each of the models presented here is either based on the GLM framework or something very similar, using them sacrifices little or no loss in interpretability, while potentially yielding increased flexibility, robustness and accuracy.

We caution that the discussions below are meant to be brief overviews of these models, and are intended to introduce the reader to them and motivate further learning. Each has many nuances and complexities not covered here, and the reader is urged to refer to other statistical texts that cover these methods in greater detail prior to attempting to use them in a real business scenario.

10.1. Generalized Linear Mixed Models (GLMMs)

In a standard GLM, the randomness of the outcome is considered to be the only source of randomness in the model; the coefficients themselves are assumed to be fixed values. To be sure, from *our* perspective, where the coefficients are unknown

and will need to be estimated from random data, the estimates of those parameters are random. (This is the randomness that statistics such as the standard error are meant to describe.) However, the underlying model assumes that *some* fixed set of values exist that always describe the relationship between the predictors and the expected value of the target variable. To see this, take a look back at Equation 2: the equals sign indicates a deterministic relationship involving fixed values; the only tilde (denoting randomness) appears in Equation 1.

The practical effect of this is that in seeking to maximize likelihood, the fitting procedure "moves" the coefficients as close to the data as possible, even for those where the data is thin. In other words, it gives the data full credibility, since we have not supplied it with any information to signal that the coefficients should behave otherwise.

A useful extension to the GLM is the **generalized linear mixed model**, or GLMM, which allows for some of the coefficients to be modeled as random variables themselves. In this context, predictors with coefficients modeled as random variables are called **random effects**; parameters modeled as having fixed values are called **fixed effects**. In practice, random effects would be estimated for categorical variables with many levels that lack the credibility for their coefficients to be estimated fully from their own data.

To illustrate, we present a simple example of an auto severity model with three predictors: driver age (a continuous variable), marital status (coded as 0 = unmarried, 1 = married) and territory, a categorical variable with 15 levels. Driver age and marital status will be designated as fixed effects in our model; territory, with many of its levels sparse and lacking credibility, will be designated as a random effect.

We denote driver age as x_1 and marital status as x_2. The territory variable is transformed to 15 dummy-coded (0/1) variables of a design matrix, where 1 indicates membership in that territory.[22] Rather than denote those 15 predictors as $x_3 \ldots x_{17}$, we will use a new symbol—namely, z—to distinguish random effects from fixed effects, and so the territory variables are denoted $z_1 \ldots z_{15}$. The coefficients for the fixed effects are denoted β_1,β_2, and the coefficients for the random effects are denoted $\gamma_1 \ldots \gamma_{15}$.

A typical setup for this model might be as follows:

$$g(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 z_1 + \cdots + \gamma_{15} z_{15} \tag{19}$$

$$y \sim gamma(\mu_i, \phi) \tag{20}$$

$$\gamma \sim normal(\nu, \sigma) \tag{21}$$

Equations 19 and 20 are the familiar fixed and random components of a regular GLM. Equation 21 introduces a probability distribution for the fifteen γ parameters, which are taken to be independent and identically distributed random variables in this model. (The normal distribution is used here for illustrative purposes; depending on the implementation, a different distribution may be used.)

[22] For random effects we do not designate a base level, and so all 15 levels get a column in the design matrix.

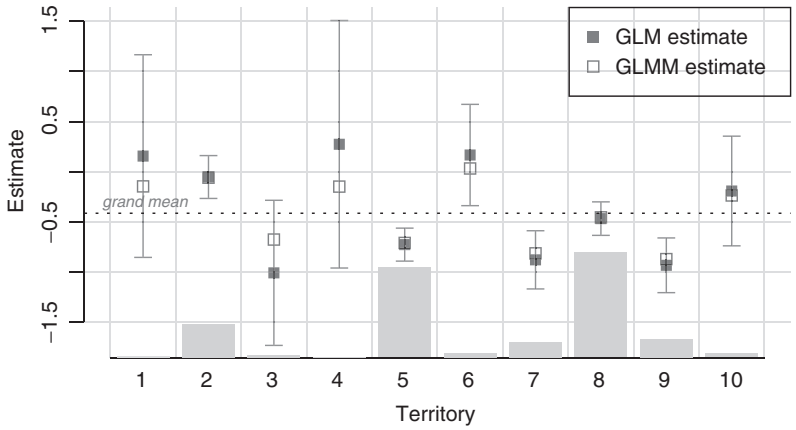
In maximizing likelihood for this setup, we now have *two* probability distributions to simultaneously consider: the distribution of outcomes y of Equation 20, and the distribution of random coefficients γ of Equation 21. Moving any of the γ coefficients close to the data raises the likelihood of y, while moving it away from the mean of the other γs lowers the likelihood of γ. In being forced to balance those two opposing forces, the model will produce territory relativities that are somewhere between the full-credibility estimates of a GLM and the grand mean for territory. The less data available for a territory, the closer its estimate will be to the mean. This effect is referred to as **shrinkage**.

Figure 27 shows an example of the estimates produced by a GLMM compared with those estimated by a standard GLM. The dotted line shows the grand mean log relativity across all territories. For territories where the data is the sparsest—and the standard errors the widest—the GLMM estimates move farther from the GLM indications and closer toward the mean.

In practice, GLMMs are estimated as a two-step process. First, estimates of all the "fixed" parameters underlying the model are produced. For the fixed effects, this stage would produce actual estimates of the coefficient; for the random effects, on the other hand, this stage produces estimates related to the probability distribution that their coefficients follow. The second stage produces estimates for all levels of categorical variables that were specified as random effects. These estimates use a Bayesian procedure that factors in the estimated randomness of the parameter as estimated by the first step as well as the volume of data available at each level.

In our example, the initial fitting procedure produces estimates for the following parameters: the intercept, β_0; the coefficients for the fixed effects, β_1 and β_2; the

Figure 27. A comparison of GLM and GLMM estimates. The filled squares show the GLM estimates, and the error bars indicate the 95% confidence intervals around those estimates. The unfilled squares show the GLMM estimates. The vertical bars are proportional to the volume of data within each territory.



dispersion parameter, ϕ; and the parameters related to the *distribution* of the γ coefficients—namely, ν and σ. Note that at this stage, the γ coefficients themselves have not been estimated; we've only estimated their distribution.

A second stage will produce the estimates of the γ coefficients. Rather than basing the estimate for each territory entirely on its own data—as a regular GLM would do—the GLMM estimates will incorporate several pieces of information: the observed severity within the territory; the estimated distribution of the γ parameters; and the estimated variance of y. Generally, estimates for more dense levels will be closer to those indicated by the data, while estimates for more sparse levels are driven closer to the overall mean.

If any of this seems eerily similar to Bühlmann-Straub credibility, that's because it is. In fact, the variance of the γ distribution—denoted σ above—is analogous to the familiar credibility concept of "between-variance" among the theoretical means; residual variance in the model—represented by ϕ—corresponds to the "within-variance." The estimated γ for each territory will in effect be a blend between the grand mean severity among territories (ν) and the territory's own observed severity, with the weighting determined based on the expected "within-variance" given the volume of data in the territory, relative to the "between-variance." Thus, the GLMM is a useful means of introducing classical credibility concepts into a GLM for multi-level categorical variables.[23]

Correlation Among Random Outcomes. In addition to allowing for credibility, the GLMM is also a means of inducing correlation into a model. Consider the case where a multi-year dataset may contain multiple renewals of the same policy. If we are concerned that the correlation among policy records is large enough so as distort the GLM results, we may wish to include policy ID as a random effect in a GLMM. In this instance, although the GLMM will produce an estimate for each policy ID, those are probably not of interest to us.

10.2. GLMs with Dispersion Modeling (DGLMs)

Recall that a constraint built into GLMs is that the dispersion parameter of the exponential family (ϕ) must be held constant for all records. An extension to the GLM that loosens up this restriction is a GLM with a **dispersion modeling** component, which allows for each record to have a unique ϕ as well as μ, controlled by a linear combination of coefficients and predictors. Those predictors may be the same as those that predict the μ parameter, or they may be different. This type of model is sometimes called a **double-generalized linear model** (or **DGLM**).[24]

The mathematical specification of a DGLM is as follows:

$$y_i \sim Exponential\left(\mu_i, \phi_i\right) \tag{22}$$

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{23}$$

[23] See Klinker (2011a) for a more detailed discussion on the relationship between classical credibility and GLMMs.
[24] Smyth and Jørgensen (2002).

$$g_d(\phi_i) = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_n z_{ip} \tag{24}$$

Equation 22 is similar to Equation 1, with a subtle difference: the ϕ parameter now has a subscript i attached to it, indicating that it may vary by record. Equation 23 is identical to Equation 2.

The chief innovation of the DGLM is Equation 24, which specifies the relationship between the dispersion parameter and the predictors $z_1 \ldots z_p$, which may or may not be the same as the μ predictors, $x_1 \ldots x_p$. Coefficients for $z_1 \ldots z_p$—denoted here as $\gamma_1 \ldots \gamma_p$—are estimated by the model. The linear combination of those predictors and coefficients equals the dispersion parameter transformed by a link function, denoted here as $g_d(.)$. The subscript d is added to distinguish it from the link function applied to μ in Equation 23, since those two need not be the same; in practice, though, it is common to use a log link for both.

Implementation. DGLMs are implemented in the "dglm" package available for both the R and S-Plus statistical languages. However, where the distribution is a member of the Tweedie family (that is, either the normal, Poisson, gamma, inverse Gaussian or Tweedie distribution), the DGLM parameters can be closely approximated using any GLM estimation software with the following iterative procedure:[25]

1. Begin by assigning a value of 1 to all ϕ_i.
2. Run a GLM to estimate the β coefficients as usual, but with one modification: the weight variable should be the inverse of the dispersion parameter for each record—that is, $1/\phi_i$. If we wish to use a weight in the model, we must divide it by ϕ (i.e., set the weight variable to ω_i/ϕ_i).
3. Using the predictions generated by the model estimated in step 2, calculate the *unit deviance* for each record. The unit deviance is defined as:

$$d_i = 2\phi_i\left[\ln f(y_i|\mu_i = y_i) - \ln f(y_i|\mu_i = \mu_i)\right]$$

   Note that this formula is the record's contribution to the total unscaled deviance described in Section 6.1.2.[26]
4. Run a GLM specified as follows:
   - The target variable is the unit deviance calculated in step 3.
   - The distribution is gamma.
   - As predictors, use whatever variables we believe may affect dispersion. These are the z variables of Equation 24, which may or may not be the same as the main GLM predictors.

---

[25] Smyth and Jørgensen (2002).

[26] For the Tweedie distribution, that formula works out to be the following:

$$d_i = 2\omega_i\left(y_i\frac{y_i^{1-p} - \mu_i^{1-p}}{1-p} - \frac{y_i^{2-p} - \mu_i^{2-p}}{2-p}\right)$$

where ω denotes the weight variable.

5. Set the dispersion parameters ϕ_i to be the predictions generated by the model of step 4.
6. Repeat steps 2 through 5 until the model converges (that is, the model parameters cease to change significantly between iterations).

Where to Use It. In a general sense, using a DGLM rather than a GLM may produce better predictions of the mean, particularly in cases where certain classes of business are inherently more volatile than others. Allowing the dispersion parameter to "float" will in turn allow the model to give less weight to the historical outcomes of the volatile business, and more weight to the stable business whose data is more informative—thereby ignoring more noise and picking up more signal.

The following are particular scenarios where using a DGLM rather than a GLM may provide additional benefit:

- For some actuarial applications, the full distribution of the outcome variable, rather than just the mean, is desired. In such scenarios, a GLM with constant dispersion may be too simplistic to adequately describe the distribution. The DGLM, on the other hand, models two distributional parameters for each risk and thereby has greater flexibility to fit the distributional curves.
- GLMs that use the Tweedie distribution to model pure premium or loss ratio, by keeping the dispersion parameter constant, contain the implicit assumption that all predictors have the same directional effect on frequency and severity. (See Section 2.7.3 for further discussion on this.) The DGLM, on the other hand, by allowing the dispersion parameter to vary, provides the flexibility for the model to mold itself to the frequency and severity effects observed in the data.

10.3. Generalized Additive Models (GAMs)

As noted in the introduction to this chapter, a hallmark assumption of the GLM is linearity in the predictors. While non-linear effects can be accommodated by adding various transformations of the predictors into the linear equation, those are workarounds that must be specified manually.

The **generalized additive model** (GAM) is a GLM-like model that handles non-linearity natively. The mathematical specification of a GAM is as follows:

$$y_i \sim Exponential\left(\mu_i, \phi\right) \tag{25}$$

$$g(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p\left(x_{ip}\right) \tag{26}$$

Equation 25 is identical to equation 1. GAMs, like GLMs, assume the random component of the outcome to follow an exponential family distribution.

Equation 26 is similar to equation 2, but with an important twist: the addends making up the linear predictor are no longer linear functions of the predictors—rather, they are

any arbitrary functions of the predictors. Those functions, denoted $f_1(.) \ldots f_n(.)$ specify the effects of the predictors on the (transformed) mean response as smooth curves. The shapes of these curves are estimated by the GAM software.

Note that the "additive" of "generalized additive model" refers to the fact that the linear predictor is a series of additive terms (though free from the constraint of linearity). As with a GLM, we can specify a log link, which would turn the model multiplicative.

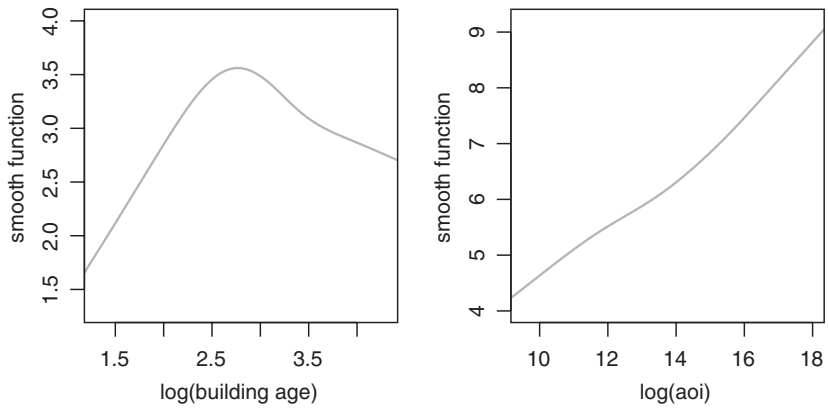
Unlike in a GLM, where the effect of a variable on the response can be easily determined by examining its coefficient, for a GAM we are provided no such convenient numeric description of the effect. As such, predictor effects must be assessed graphically. Figure 28 shows examples of such graphs, using the example severity model discussed back in Section 5.4. For this illustration, two continuous variables—building age and amount of insurance, both logged—are included in a log link GAM, and their estimated smooth functions are graphed in the left and right panels, respectively.

For building age, the GAM estimated a clearly non-linear function, with mean severity first rising, reaching a peak at around building age $e^{2.8} = 16$ years, then declining. (Compare this to Figures 10 and 11.) For amount of insurance, on the other hand, although the GAM was free to fit any arbitrary function, the one it estimated is nearly linear (albeit with some curvature), indicating that a linear fit would probably suffice for this variable.

The GAM allows us to choose from among several different methods for estimating the smooth functions; we will not delve into those details here. Each of those methods allows us to specify parameters that control the degree of smoothness for each function. Those parameters must be fine-tuned carefully, as allowing for too-flexible a function runs the risk of overfitting.

Implementation. GAMs are available through the R packages "gam" or "mgcv," or through PROC GAM in SAS.

Figure 28. Graphical Display of GAM Smoother Functions for Log of Building Age (*left panel*) and Log of Amount of Insurance (*right panel*)



10.4. MARS Models

Another GLM variant that is great at handling non-linearities is **multivariate adaptive regression splines**, or MARS. Rather than fit smooth functions for the predictors, as does the GAM discussed in the preceding section, MARS models operate by incorporating piecewise linear functions, or *hinge functions,* into a regular GLM. These hinge functions are the same as those discussed in Section 5.4.4. However, in that section we manually created the functions and determined cut points by eyeballing partial residual plots; MARS models create the functions and optimize the cut points automatically.

To illustrate, we continue with our example severity model of the previous section. This time, we will use a MARS model to capture potential non-linearity in the building age and amount of insurance variables. Table 14 shows the portion of the resulting coefficient table relating to those two variables.

In the output below, the function $h(.)$ refers to the "hinge function" discussed in Section 5.4.4. For example, "h(log(AoC)−1.94591)" is defined as max(log(AoC)−1.94591,0).

Looking at the three hinge functions for building age, notice that this handling of that variable is fairly similar to the piecewise linear functions we set up in Section 5.4.4, which had cut points at 2.75 and 3.5. The MARS model also found another cut point at 1.95. MARS did not include the unaltered log(AoC) term in this model, meaning that the response curve for log(AoC) below 1.95 is flat. (In a practical sense, that means this model would not differentiate between buildings of ages 1 to 7 years.)

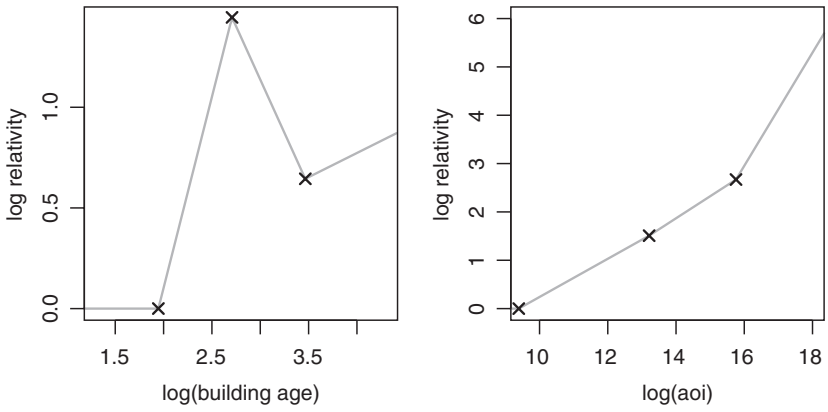
Figure 29 graphs the response curves indicated by this model for those two variables. The ×'s mark the locations of the cut points. Compare those to the curves indicated by the GAM output of the previous section.

As with GAMs, MARS has tuning parameters to control the flexibility of the fit. A more flexible model will create more cut points, allowing for finer segmentation. Of course, with that additional flexibility comes the risk of chasing noise.

Table 14. Partial Output of MARS Coefficient Table

Parameter	Estimate	Std. Error	*p*-Value
. . .	. . .	. . .	. . .
'h(log(AoC)–1.94591)'	1.8977	0.1976	<0.0001
'h(log(AoC)–2.70805)'	–2.9557	0.2598	<0.0001
'h(log(AoC)–3.46574)'	1.2980	0.3457	0.0002
'h(log(AOI)–9.39124)'	0.3949	0.0359	<0.0001
'h(log(AOI)–13.2124)'	0.0611	0.0657	0.3526
'h(log(AOI)–15.7578)'	0.7151	0.2263	0.0016
. . .	. . .	. . .	. . .

Figure 29. Graphical display of MARS indicated relativities for log of building age (*left panel*) and log of amount of insurance (*right panel*). The ×'s mark the locations of the cut points.



In addition to its natural ability to handle non-linearities, MARS has a number of additional highly useful features, including:

- It performs its own variable selection. Unlike a GLM—which will generate a coefficient for each predictor input by the user—MARS will keep only those that are significant. (Tuning parameters are available to control how many variables are retained.)
- It can also search for significant interactions. It is quite flexible in this regard; in addition to the 2-way interactions discussed in Section 5.6, it can search for 3-way (or higher degree) interactions, as well as interactions among the piecewise linear functions.

Even where we require our final model to be in the form of a standard GLM, MARS may still be a very valuable tool in the model refinement process: we can run a MARS model on the data, examine its output—hinge functions it created, interactions it discovered, and so on—and copy whichever terms we like into our GLM. Consider the output shown in Table 14; it is very easy to simply replicate those same hinge functions in our GLM, and get the same benefit of the non-linear fit.

Used in this way, MARS may uncover non-linear transformations or interactions we may not have thought to try. Great care needs to be taken, though, as such a "deep search" through the data can easily turn up spurious effects.

Implementation. MARS is available as commercial software from Salford Systems. Implementations of the same procedure (not called *MARS,* due to Salford Systems' trademark on the name) are available through the "earth" package in R and PROC ADAPTIVEREG in SAS (beginning with SAS/STAT version 13.1).

10.5. Elastic Net GLMs

When modeling in situations where there are a large number of potential predictor variables, overfitting can be a real concern for GLMs. GLMs make full use of all the

predictors fed into them to fit the training data as best as possible—that is, it will find coefficients for all predictors such that the deviance of the training set is minimized. Including too many predictors will cause the model to pick up random noise in the training data, yielding a model that may perform poorly on unseen data. In such a scenario, variable selection—choosing the right variables to include in the model while omitting the others—can be quite challenging.

Elastic net GLMs provide a powerful means of protecting against overfitting even in the presence of many predictors. Elastic nets GLMs are, at the core, identical to GLMs in their mathematical specification. The chief difference is in the method by which the coefficients are fit. Rather than aggressively minimizing deviance on the training set—as a regular GLM would—elastic nets enable you to constrain the fit, by minimizing a function that is deviance subject to a penalty term for the size and magnitude of the coefficients. This penalty term can be fine-tuned to allow you to find the right balance where the model fits the training data well—but at the same time, the coefficients of the model are not too large.

The function minimized by elastic nets is as follows:[27]

$$Deviance + \lambda\left(\alpha\sum|\beta| + (1-\alpha)\frac{1}{2}\sum\beta^2\right) \tag{27}$$

The first additive term of the above expression is just the GLM deviance; if this were a regular GLM, we'd be minimizing just that. The elastic net adds the part following the plus sign, called the *penalty term.* Let's examine that closely.

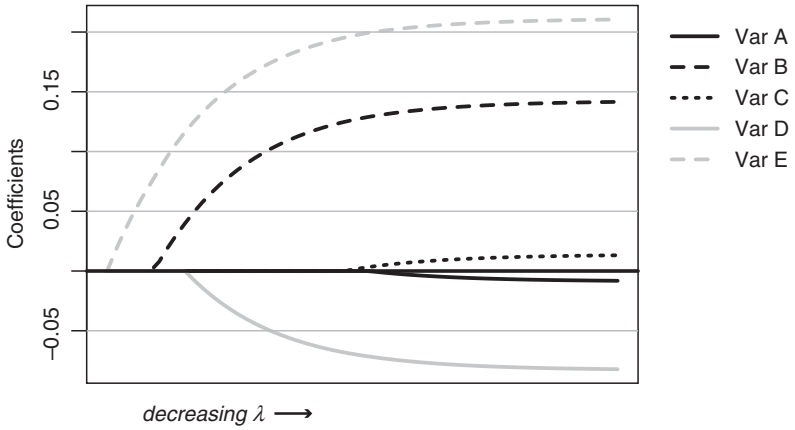
Inside the parentheses is a weighted average of the sum of the absolute values of the coefficients and the halved sum of squared coefficients, with the weights determined by α, a parameter between 0 and 1 that we control. This use of a weighted average is primarily due to the fact that this model is a generalization of two earlier variations on this same concept: the **lasso** model, which uses absolute value of coefficients, and **ridge** model, which uses squared coefficients. The important thing to recognize, though, is that the terms inside the parentheses yield an increasing function of the *magnitude* of the coefficients, or the degree by which the coefficients deviate from zero. Thus, a greater penalty is applied for larger coefficients.[28]

The more important tuning parameter in Equation 27 is the λ that sits outside the parentheses. This allows us to control the severity of the penalty that gets applied. The practical effect of raising λ is that it forces coefficients to shrink closer to zero, to compensate for the increased penalty, in minimizing Equation 27. Under certain

[27] In Equation 27, the vector of coefficients represented by β does not include β_0, the intercept term, which does not contribute to the penalty.

[28] In elastic net models, all predictor variables are automatically centered and scaled prior to running the model. This way, the resulting β coefficients are on similar scales, and so the magnitude of deviation from zero means roughly the same thing for all variables, regardless of the scales of the original variables. Note, however, that most implementations of elastic nets will return the coefficients on the scales of the original variables, so this standardization that happens behind the scenes poses no obstacle to implementation of the resulting model.

Figure 30. An Illustration of the Effect of Varying λ on Elastic Net Coefficients



conditions, some less-important predictors will be assigned coefficients of zero (effectively removing them from the model entirely).

In Figure 30 we illustrate this effect for a simple model that has five predictors, which we name A through E. Each predictor is represented by a different curve. For each, the value that the coefficient assigns to the predictor is plotted on the *y*-axis for different values of λ, with λ *decreasing* from left to right along the *x*-axis.

At the far left of the graph—where λ is at its highest—the penalty for coefficient size is severe, and so no variables make it in with a non-zero coefficient. As we move rightward, dialing down λ and thereby easing up on the penalty, Variable E—clearly the most significant variable here—enters our model and grows in influence as λ declines. Moving farther to the right, more variables make their way in and their coefficients grow—eventually converging toward the maximum likelihood estimates that a regular GLM would give them.

In practice, the λ parameter is usually fine-tuned through cross validation. Doing so produces a model that is likely to perform better on unseen data than would a regular GLM. After all, a GLM is just a special case of the elastic net (where $\lambda = 0$) and so the fine-tuning procedure has the flexibility to produce a standard GLM if in fact it is the best model. Usually, though, the model can be improved by setting a non-zero penalty.

As we have seen, a non-zero penalty causes the model parameters to exhibit the shrinkage effect that is characteristic of actuarial credibility models as well as GLMMs discussed above. In fact, it has been shown that elastic nets bear direct relationships to many classical credibility models.[29] Thus, as with GLMMs discussed above, elastic nets provide a convenient means of incorporating familiar credibility concepts into the GLM framework.

[29] See Miller (2015) for further discussion on this equivalence and its derivation.

Elastic nets also have the advantage of being able to perform automatic variable selection, as variables that are not important enough to justify their inclusion in the model under the penalty constraint will be removed.

Furthermore, elastic nets perform much better than GLMs in the face of highly correlated predictors. The penalty term provides protection against the coefficients "blowing up" as they might in a GLM. Rather, one or two variables of a group of correlated predictors will typically be selected, and they will be assigned moderate coefficients.

The main disadvantage of elastic nets is that they are much more computationally complex than standard GLMs. The computational resources and time needed to fit elastic nets and optimize λ may make elastic nets impractical for large datasets.

Implementation. Elastic nets are implemented in the "glmnet" package in R.[30] It is also available in SAS (beginning with SAS/STAT version 13.1) using PROC GLMSELECT.

[30] As of this writing, the glmnet package does not support the gamma or Tweedie distributions. Fortunately, the "HDtweedie" package provides an implementation of glmnet for the Tweedie distribution; the gamma distribution is accessible through this package by setting the Tweedie p parameter to be 2.

Bibliography

Several of the items in this section reference chapters of *Predictive Modeling Applications in Actuarial Science: Vol. 1,* edited by Jed Frees, Richard Derrig and Glenn Meyers. In addition to providing more detailed and in-depth technical discussions of GLMs and other models discussed in this monograph, that book also provides several insurance datasets on which the reader can test and practice those models and other techniques described in this text.

Anderson, Duncan, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi. 2007. *A Practitioner's Guide to Generalized Linear Models.* https://www.casact.org/pubs/dpp/dpp04/04dpp1.pdf.

Antonio, Katrien and Yanwei Zhang. 2014. "Nonlinear Mixed Models." *Predictive Modeling Applications in Actuarial Science: Vol. 1.* Chap. 16. New York: Cambridge University Press.

Brockett, Patrick L., Shuo-Li Chuang and Utai Pitaktong. 2014. "Generalized Additive Models and Nonparametric Regression." *Predictive Modeling Applications in Actuarial Science: Vol. 1.* Chap. 15. New York: Cambridge University Press.

Clark, David R. and Charles A. Thayer. 2004. *A Primer on the Exponential Family of Distributions.* https://www.casact.org/pubs/dpp/dpp04/04dpp117.pdf.

Dean, Curtis Gary. 2014. "Generalized Linear Models." *Predictive Modeling Applications in Actuarial Science: Vol. 1.* Chap. 5. New York: Cambridge University Press.

Dunn, Peter K. and Gordon K. Smyth. 1996. "Randomized Quantile Residuals." *Journal of Computational and Graphical Statistics* 5:236–244.

Frees, Edward W., Glenn Meyers and David A. Cummings. 2011. "Predictive Modeling of Multi-Peril Homeowners Insurance." *Variance* 6:1, pp. 11–31.

Frees, Edward W., Glenn Meyers and David A. Cummings. 2014. "Insurance Ratemaking and a Gini Index." *Journal of Risk and Insurance* 81(2) pp. 335–366, 2014.

Frees, Edward W. 2014. "Frequency and Severity Models." *Predictive Modeling Applications in Actuarial Science: Vol. 1.* Chap. 6. New York: Cambridge University Press.

Harrell, Frank E., Jr. 2015. *Regression Modeling Strategies*. Second Ed. New York: Springer.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

de Jong, Piet and Gillian Z. Heller. 2008. *Generalized Linear Models for Insurance Data.* New York: Cambridge University Press.

Klinker, Fred. 2011a. "Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting." Casualty Actuarial Society *E-Forum* (Winter):2

Klinker, Fred. 2011b. "GLM Invariants." Casualty Actuarial Society *E-Forum*, Summer 2011.

Kuhn, Max and Kjell Johnson. *Applied Predictive Modeling.* 2013. New York: Springer.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models.* 2nd ed. London: Chapman & Hall.

Miller, Hugh. 2015. *A Discussion on Credibility and Penalised Regression, with Implications for Actuarial Work.* Presented to the Actuaries Institute 2015 ASTIN, AFIR/ERM and IACA Colloquia.

Smyth, Gordon K. and Bent Jørgensen. 2002. "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling." *ASTIN Bulletin* 32(1):143–157.

Werner, Geoff and Claudine Modlin. 2010. *Basic Ratemaking.* 2nd ed. Casualty Actuarial Society.

Yan, Jun, James Guszcza, Matthew Flynn, and Cheng-Sheng Peter Wu. 2009. "Applications of the Offset in Property-Casualty Predictive Modeling." Casualty Actuarial Society *E-Forum* (Winter 2009):366–385.

Appendix

In section 6.3.2, which discusses binned working residuals, we noted two properties that such residuals hold for a well-specified model, which makes them highly useful for performing residual analysis on models built from large datasets: (1) They follow no predictable pattern, as the mean of these residuals is always zero; and (2) they are homoscedastic, i.e., their variance is constant. In this appendix we show the derivation of these properties.

Given a model with n observations, let $i = 1, \ldots, n$ be the index of the observations. We divide the observations into m bins; let $b = 1, \ldots, m$ be the index of the bins.

Define *working residual* as

$$wr_i = (y_i - \mu_i) \cdot g'(\mu_i)$$

Define *working weight* as

$$ww_i = \frac{\omega_i}{V(\mu_i) \cdot [g'(\mu_i)]^2}$$

Define *binned working residual* as

$$br_b = \frac{\sum_{i \in b} wr_i \cdot ww_i}{\sum_{i \in b} ww_i}$$

We assign observations to bins such that all bins have equal sums of working weights, i.e., $\sum_{i \in b} ww_i = k$. It follows that

$$\sum_{i=1}^{n} ww_i = SWW = m \cdot k \rightarrow k = \frac{SWW}{m}$$

For a properly specified model, the following holds:

$$E(y_i) = \mu_i,$$

$$Var(y_i) = \frac{\phi \cdot V(\mu_i)}{\omega_i},$$

$$E\left(\omega_i \frac{(y_i-\mu_i)}{V(\mu_i)\,g'(\mu_i)}\right)=0.^{1}$$

Property 1: $E(br_b)=0$.

$$\begin{aligned}E(br_b) &= E\left(\frac{\sum_{i\in b} wr_i \cdot ww_i}{\sum_{i\in b} ww_i}\right)\\ &= \frac{1}{k}E\left(\sum_{i\in b} wr_i \cdot ww_i\right)\\ &= \frac{1}{k}E\left(\frac{(y_i-\mu_i)\cdot g'(\mu_i)\cdot \omega_i}{V(\mu_i)\cdot[g'(\mu_i)]^2}\right)\\ &= \frac{1}{k}E\left(\omega_i \frac{(y_i-\mu_i)}{V(\mu_i)\,g'(\mu_i)}\right)=0\end{aligned}$$

Property 2: $Var(br_b)=Constant=\dfrac{\phi\cdot m}{SWW}$

$$\begin{aligned}Var(br_b) &= Var\left(\frac{\sum_{i\in b} wr_i \cdot ww_i}{\sum_{i\in b} ww_i}\right)\\ &= \frac{1}{k^2}Var\left(\sum_{i\in b} wr_i \cdot ww_i\right)\end{aligned}$$

Assume that the working residuals are independent. Therefore,

$$\frac{1}{k^2}Var\left(\sum_{i\in b} wr_i \cdot ww_i\right)=\frac{1}{k^2}\sum_{i\in b} Var\left(wr_i \cdot ww_i\right)$$

Let's simplify the $Var\,(wr_i \cdot ww_i)$ term:

$$Var\left(wr_i \cdot ww_i\right)=Var\left(\omega_i \frac{y_i-\mu_i}{V(\mu_i)\cdot g'(\mu_i)}\right)$$

---

[1] See Klinker (2011b), who demonstrates that $\sum \omega_i \frac{(y_i-\mu_i)}{V(\mu_i)\,g'(\mu_i)}=0$, both over the entire GLM training data as well as over any subset with the same level of a categorical variable. In a well-fit model there is no predictable pattern in the residuals, and so the expected value $E\left(\omega_I \frac{(y_i-\mu_i)}{V(\mu_i)\,g'(\mu_i)}\right)=0$ for any individual observation as well.

$$= \frac{\omega_i^2}{V(\mu_i)^2 \cdot [g'(\mu_i)]^2} Var(y_i - \mu_i)$$

$$= \frac{\omega_i^2}{V(\mu_i)^2 \cdot [g'(\mu_i)]^2} Var(y_i)$$

$$= \frac{\omega_i^2}{V(\mu_i)^2 \cdot [g'(\mu_i)]^2} \cdot \frac{\phi V(\mu_i)}{\omega_i}$$

$$= \frac{\omega_i \cdot \phi}{V(\mu_i) \cdot [g'(\mu_i)]^2} = \phi \cdot ww_i$$

Plugging this simplified term back into the original equation,

$$Var(br_b) = \frac{1}{k^2} \sum_{i \in b} Var(wr_i \cdot ww_i) = \frac{1}{k^2} \sum_{i \in b} \phi \cdot ww_i$$

$$= \frac{\phi}{k^2} \sum_{i \in b} ww_i = \frac{\phi \cdot k}{k^2} = \frac{\phi}{k} = \frac{\phi \cdot m}{SWW}$$

A good rule of thumb is to select the number of bins m such that $Var(br_b) \leq 0.01$.

ABOUT THE SERIES:

CAS monographs are authoritative, peer-reviewed, in-depth works focusing on important topics within property and casualty actuarial practice. For more information on the CAS Monograph Series, visit the CAS website at www.casact.org.

Expertise. Insight.
Solutions.

www.casact.org

Clarification and Errata to *Catastrophe Modeling: A New Approach to Managing Risk* (Grossi, P. and Kunreuther, H., Editors)

Casualty Actuarial Society Syllabus Committee *

March 29, 2021

Abstract

This notes presents an errata and clarifying remarks to Section 2.4 Derivation and Use of an Exceedance Probability Curve of *Catastrophe Modeling: A New Approach to Managing Risk*.

1 Clarification

The use of the phrase "exceedance probability" in Section 2.4 is ambiguous. Specifically, "exceedance probability" can be used in one of three ways:

Occurrence Exceedance Probability (OEP) The OEP is the probability that at least one loss exceeds the specified loss amount.

Aggregate Exceedance Probability (AEP) The AEP is the probability that the sum of all losses during a given period exceeds some amount.

Conditional Exceedance Probability (CEP) The CEP is the probability that the amount on a single event exceeds a specified loss amount; this is equal to 1-CDF of the severity curve as used by actuaries in other contexts.

For actuaries who have not worked with catastrophe models, the OEP may be a new concept. Actuaries usually think of severity distributions, which correspond to the CEP - not the OEP. In Section 2.4, the term "exceedance probability" refers to the **Occurrence Exceedance Probability (OEP)**. The *OEP* is the distribution of the largest loss in the period and is based on the theory of order statistics.

*This note was originally prepared by Rajesh Sahasrabuddhe, FCAS, MAAA, CAS Syllabus Committee Chairperson in 2013. It has been revised based on similar comments provided contemporaneously by Josh Taub, FCAS and Matthew M. Iseler, FCAS.

2 Errata

- The end continued paragraph at the top of page 30 is corrected as follows:

  A list of ~~15~~ 14[1] such events is listed in Table 2.1 ranked in descending order of the amount of loss. ~~In order to keep the example simple and the calculations straightforward, these events were chosen so the set is exhaustive (i.e., sum of probabilities for all events equals one).~~

- The first complete paragraph on page 30 is corrected as follows:

  The events listed in Table 2.1 are assumed to be independent Bernoulli random variables ~~., each with a~~. It is assumed that each event only occurs at most once with the probability mass function defined as: …

- The second complete paragraph on page 30 is corrected as follows:

  If an event E_i does not occur, the loss for that event is 0. …

- The fourth complete paragraph on page 30 is corrected as follows:

  Assuming that during a given year, at most only one of each disaster occurs, the OEP ~~exceedance probability~~ for a given level of loss, $OEP(L_i)$, can be determined by calculating: …

- The first sentence of the fifth complete paragraph on page 30 is corrected as follows:

  The resulting OEP is the probability that at least one loss exceeds a given value ~~exceedance probability is the annual probability that the loss exceeds a given value~~.

- The upper limit of the product in the last equation on page 30 is corrected from i to $i-1$ as follows:

$$OEP(L_i) = 1 - \prod_{j=1}^{i} \prod_{j=1}^{i-1} (1 - p_i)$$

[1] Editor's note: The definition of E_i includes events that "*could* damage a portfolio of structures" (emphasis added). We assume that event #15 in the original Table 2.1 would have met this standard (e.g. a hurricane that turns away from land). We have removed event #15 in order to emphasize that the probabilities need not sum to 1.000.

- Table 2.1 is replaced with the following:

Table 2.1: Events, Losses and Probabilities

Event (E_i)	Annual Probability of Occurrence (p_i)	Loss (L_i)	Occurrence Exceedance Probability $[OEP(L_i)]$	$E[L] = p_i \times L_i$
1	0.002	$25,000,000	0.0000	$50,000
2	0.005	15,000,000	0.0020	75,000
3	0.010	10,000,000	0.0070	100,000
4	0.020	5,000,000	0.0169	100,000
5	0.030	3,000,000	0.0366	90,000
6	0.040	2,000,000	0.0655	80,000
7	0.050	1,000,000	0.1029	50,000
8	0.050	800,000	0.1477	40,000
9	0.050	700,000	0.1903	35,000
10	0.070	500,000	0.2308	35,000
11	0.090	500,000	0.2847	45,000
12	0.100	300,000	0.3490	30,000
13	0.100	200,000	0.4141	20,000
14	0.100	100,000	0.4727	10,000
Total		Average Annual Loss (AAL)		760,000

AN EXAMPLE OF CREDIBILITY AND SHIFTING RISK PARAMETERS

HOWARD C. MAHLER

Abstract

In this paper, the won-lost record of baseball teams will be used to examine and illustrate credibility concepts. This illustrative example is analogous to the use of experience rating in insurance. It provides supplementary reading material for students who are studying credibility theory.

This example illustrates a situation where the phenomenon of shifting parameters over time has a very significant impact. The effects of this phenomenon are examined.

Three different criteria that can be used to select the optimal credibility are examined: least squares, limited fluctuation and Meyers/Dorweiler. In applications, one or more of these three criteria should be useful.

It is shown that the mean squared error can be written as a second order polynomial in the credibilities with the coefficients of this polynomial written in terms of the covariance structure of the data. It is then shown that linear equation(s) can be solved for the least squares credibilities in terms of the covariance structure.

The author wishes to thank Julie Jannuzzi and Gina Brewer for typing this paper.

1. INTRODUCTION

In this paper, the won-lost record of baseball teams will be used to examine and illustrate credibility concepts. This illustrative example is analogous to the use of experience rating in insurance. The mathematical details are contained in the appendices.

One purpose of this paper is to provide supplementary reading material for students who are studying credibility theory. However, this paper also contains a number of points which should prove of interest to those who are already familiar with credibility theory.

Of particular interest is the effect of shifting risk parameters over time on credibilities and experience rating. This example illustrates a situation where the phenomenon of shifting parameters over time has a very significant impact.

The general structure of the paper is to go from the simplest case to the more general. The mathematical derivations are confined to the appendices.

Section 2 briefly reviews the use of credibility in experience rating.

Section 3 describes the data sets from baseball that are used in this paper in order to illustrate the concepts of the use of credibility in experience rating.

Section 4 is an analysis of the general structure of the data. It is demonstrated that the different insureds (baseball teams) have significantly different underlying loss potentials. It is also shown that for this example a given insured's relative loss potential does shift significantly over time.

Section 5 states the problem whose solution will be illustrated. One wishes to estimate the future loss potential using a linear combination of different estimates.

Section 6 discusses simple solutions to the problem presented in Section 5.

Section 7 discusses three criteria that can be used to distinguish between solutions to the problem in Section 5.

Section 8 applies the three criteria of Section 7 to the forms of solution presented in Section 6. The results of applying the three different criteria are compared. The reduction in squared error and the impact of the delay in receiving data are both discussed.

Section 9 discusses more general solutions to the problem than those presented in Section 6.

Section 10 applies the three criteria of Section 7 to the forms of the solution presented in Section 9.

Section 11 shows equations for Least Squares Credibility that result from the covariance structure assumed.

Section 12 discusses miscellaneous subjects.

Section 13 states the author's conclusions.

2. CREDIBILITY AND EXPERIENCE RATING

Experience rating and merit rating modify an individual insured's rate above or below average. From an actuarial standpoint, the experience rating plan is using the observed loss experience of an individual insured in order to help predict the future loss experience of that insured. Usually this can be written in the form:

$$\text{New Estimate} = (\text{Data}) \times (\text{Credibility}) + (\text{Prior Estimate}) \times (\text{Complement of Credibility})$$

For most experience rating plans, the prior estimate is the class average. However, in theory the prior estimate could be a previous estimate of the loss potential of this insured relative to the class average. This paper will treat both possibilities.

2.1 Shifting Parameters Over Time

There are many features of experience rating plans that are worthy of study by actuaries. Meyers [1], Venter [2], Gillam [3], and Mahler [4] present examples of recent work. The example in this paper will deal with only one aspect, that is, how to best combine the different years of past data.

The author, in a previous paper [5], came to the following conclusion concerning this point:

"When there are shifting parameters over time, older years of data should be given substantially less credibility than more recent years of data. There may be only a minimal gain in efficiency[1] from using additional years of data."

3. THE DATA SETS

This paper will examine two very similar sets of data in order to illustrate certain features of credibility. Each set of data is the won-lost record for a league of baseball teams.[2] One set is for the so-called National League while the other is for the American League.[3] Each set of data covers the sixty years from 1901 to 1960. During this period of time each league had eight teams.

For each year, called a season in baseball, for each team, we have the losing percentage, i.e., the percentage of its games that the individual team lost.

3.1 Advantages of this Data

This example has a number of advantages not to be found using actual insurance data. First, over a very extended period of time there is a constant set of risks (teams). In insurance there are generally insureds who leave the data base and new ones that enter.

Second, the loss data over this extended period of time are readily available, accurate and final. In insurance the loss data are sometimes hard to compile or obtain and are subject to possible reporting errors and loss development.

Third, each of the teams in each year plays roughly the same number of games.[4] Thus the loss experience is generated by risks of roughly equal "size." Thus, in this example, one need not consider the dependence of credibility on size of risk.

---

[1] Meyers [11] defines the efficiency of an experience rating plan as the reduction in expected squared error accomplished by the use of the plan. The higher the efficiency the smaller the expected squared error.

[2] Appendix A gives some relevant features of the sport of baseball.

[3] These two leagues are referred to as the major leagues. They generally contain the best players in North America. The data for the two leagues are independent of each other, since no inter-league games are included in the data.

[4] Over the 60 years in question, teams usually played about 150 games per year.

4. ANALYSIS OF THE GENERAL STRUCTURE OF DATA

The loss experience[5] by risk (team) by year are given in Table 1 for the National League and Table 2 for the American League.[6]

4.1 Is There an Inherent Difference Between Teams?

The first question to be answered is whether there is any real difference between the experience of the different teams, or is the apparent difference just due to random fluctuations. This is the fundamental question when considering the application of experience rating.

It requires only an elementary analysis in order to show that there is a non-random difference between the teams. The average experience for each team over the whole period of time differs significantly from that of the other teams. If the experience for each team were drawn from the same probability distribution, the results for each team would be much more similar. The standard deviation in losing percentage over a sample of about 9000 games[7] would be .5%.[8] Thus if all the teams' results were drawn from the same distribution, approximately 95% of the teams would have an average losing percentage between 49% and 51%.[9]

The actual results are shown on Table 3. In fact, only 3 of 16 teams have losing percentages in that range. The largest deviation from the grand mean is 15 times the expected standard deviation if the teams all had the same underlying probability distribution.

---

[5] For each of 60 years, the percentage of games lost is given for each team. The data are from *The Sports Encyclopedia* [6].

[6] For the National League the teams are in order: Brooklyn, Boston, Chicago, Cincinnati, New York, Philadelphia, Pittsburgh and St. Louis. For the American League the teams are in order: Boston, Chicago, Cleveland, Detroit, New York, Philadelphia, St. Louis and Washington. In both cases, the city given is that in which the team spent the majority of the data period.

[7] About 150 games for a team each year times 60 years.

[8] A binomial distribution with a 50% chance of losing, for 9000 games, has a variance of $9000(1/2)(1 - 1/2) = 2250$. This is a standard deviation of 47 games lost, or $47 \div 9000 = .5\%$ in losing percentage.

[9] Using the standard normal approximation, 95% of the probability is within two standard deviatioins of the mean which in this case is 50%.

TABLE 1

National League Losing Percentages

	NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
1901	.500	.419	.619	.626	.620	.407	.353	.457
1902	.467	.457	.504	.500	.647	.591	.259	.582
1903	.580	.485	.406	.468	.396	.637	.350	.686
1904	.641	.634	.392	.425	.307	.658	.431	.513
1905	.669	.684	.399	.484	.314	.454	.373	.623
1906	.675	.566	.237	.576	.368	.536	.392	.653
1907	.608	.561	.296	.569	.464	.435	.409	.660
1908	.591	.656	.357	.526	.364	.461	.364	.682
1909	.706	.641	.320	.497	.399	.516	.276	.645
1910	.654	.584	.325	.513	.409	.490	.438	.588
1911	.709	.573	.403	.542	.353	.480	.448	.497
1912	.660	.621	.393	.510	.318	.520	.384	.588
1913	.543	.564	.425	.582	.336	.417	.477	.660
1914	.386	.513	.494	.610	.455	.519	.552	.471
1915	.454	.474	.523	.539	.546	.408	.526	.529
1916	.414	.390	.562	.608	.434	.405	.578	.608
1917	.529	.536	.519	.494	.364	.428	.669	.461
1918	.573	.548	.349	.469	.427	.553	.480	.605
1919	.590	.507	.464	.314	.379	.657	.489	.606
1920	.592	.396	.513	.464	.442	.595	.487	.513
1921	.484	.493	.582	.542	.386	.669	.412	.431
1922	.654	.506	.481	.442	.396	.627	.448	.448
1923	.649	.506	.461	.409	.379	.675	.435	.484
1924	.654	.403	.471	.458	.392	.636	.412	.578
1925	.542	.556	.558	.477	.434	.556	.379	.497
1926	.566	.536	.468	.435	.510	.616	.451	.422
1927	.610	.575	.444	.510	.403	.669	.390	.399
1928	.673	.497	.409	.487	.396	.717	.441	.383
1929	.636	.542	.355	.571	.444	.536	.425	.487
1930	.545	.442	.416	.617	.435	.662	.481	.403

TABLE 1

(CONTINUED)

	NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
1931	.584	.480	.455	.623	.428	.571	.513	.344
1932	.500	.474	.416	.610	.532	.494	.442	.532
1933	.461	.575	.442	.618	.401	.605	.435	.464
1934	.483	.533	.430	.656	.392	.624	.507	.379
1935	.752	.542	.351	.556	.405	.582	.438	.377
1936	.539	.565	.435	.519	.403	.649	.455	.435
1937	.480	.595	.396	.636	.375	.601	.442	.474
1938	.493	.537	.414	.453	.447	.700	.427	.530
1939	.583	.451	.455	.370	.490	.702	.556	.399
1940	.572	.425	.513	.346	.526	.673	.494	.451
1941	.597	.351	.545	.429	.516	.721	.474	.366
1942	.601	.325	.558	.500	.441	.722	.551	.312
1943	.556	.471	.516	.435	.641	.584	.481	.318
1944	.578	.591	.513	.422	.565	.601	.412	.318
1945	.559	.435	.364	.604	.487	.701	.468	.383
1946	.385	.471	.464	.565	.604	.552	.591	.372
1947	.390	.442	.552	.526	.474	.597	.597	.422
1948	.455	.405	.584	.582	.494	.571	.461	.448
1949	.370	.513	.604	.597	.526	.474	.539	.377
1950	.422	.461	.582	.569	.442	.409	.627	.490
1951	.382	.506	.597	.558	.376	.526	.584	.474
1952	.373	.582	.500	.552	.403	.435	.727	.429
1953	.318	.403	.578	.558	.545	.461	.675	.461
1954	.403	.422	.584	.519	.370	.513	.656	.532
1955	.359	.448	.529	.513	.481	.500	.610	.558
1956	.396	.403	.610	.409	.565	.539	.571	.506
1957	.455	.383	.597	.481	.552	.500	.597	.435
1958	.539	.403	.532	.506	.481	.552	.455	.532
1959	.436	.449	.519	.519	.461	.584	.494	.539
1960	.468	.429	.610	.565	.487	.617	.383	.442

TABLE 2

American League Losing Percentages

	AL1	AL2	AL3	AL4	AL5	AL6	AL7	AL8
1901	.419	.390	.599	.452	.489	.456	.650	.545
1902	.438	.448	.493	.615	.638	.390	.426	.551
1903	.341	.562	.450	.522	.463	.444	.532	.686
1904	.383	.422	.430	.592	.391	.464	.572	.748
1905	.487	.395	.506	.484	.523	.378	.647	.576
1906	.682	.384	.418	.523	.404	.462	.490	.633
1907	.604	.424	.441	.387	.527	.393	.546	.675
1908	.513	.421	.416	.412	.669	.556	.454	.559
1909	.417	.487	.536	.355	.510	.379	.593	.724
1910	.471	.556	.533	.442	.417	.320	.695	.563
1911	.490	.490	.477	.422	.500	.331	.704	.584
1912	.309	.494	.510	.549	.671	.408	.656	.401
1913	.473	.487	.434	.569	.623	.373	.627	.416
1914	.405	.545	.667	.477	.545	.349	.536	.474
1915	.331	.396	.625	.351	.546	.717	.591	.444
1916	.409	.422	.500	.435	.481	.765	.487	.503
1917	.408	.351	.429	.490	.536	.641	.630	.516
1918	.405	.540	.425	.563	.512	.594	.525	.437
1919	.518	.371	.396	.429	.424	.743	.518	.600
1920	.529	.377	.364	.604	.383	.688	.503	.553
1921	.513	.597	.390	.536	.359	.654	.474	.477
1922	.604	.500	.494	.487	.390	.578	.396	.552
1923	.599	.552	.464	.461	.355	.546	.513	.510
1924	.565	.569	.562	.442	.414	.533	.513	.403
1925	.691	.487	.545	.474	.552	.421	.464	.364
1926	.699	.471	.429	.487	.409	.447	.597	.460
1927	.669	.542	.569	.464	.286	.409	.614	.448
1928	.627	.532	.597	.558	.344	.359	.468	.513
1929	.623	.612	.467	.545	.429	.307	.480	.533
1930	.662	.597	.474	.513	.442	.338	.584	.390

TABLE 2

(CONTINUED)

	AL1	AL2	AL3	AL4	AL5	AL6	AL7	AL8
1931	.592	.634	.494	.604	.386	.296	.591	.403
1932	.721	.675	.428	.497	.305	.390	.591	.396
1933	.577	.553	.503	.513	.393	.477	.636	.349
1934	.500	.651	.448	.344	.390	.547	.559	.566
1935	.490	.513	.464	.384	.403	.611	.572	.562
1936	.519	.464	.481	.461	.333	.654	.625	.464
1937	.474	.442	.461	.422	.338	.642	.701	.523
1938	.409	.561	.434	.455	.349	.651	.638	.503
1939	.411	.448	.435	.474	.298	.638	.721	.572
1940	.468	.468	.422	.416	.429	.649	.565	.584
1941	.455	.500	.513	.513	.344	.584	.545	.545
1942	.388	.554	.513	.526	.331	.643	.457	.589
1943	.553	.468	.464	.494	.364	.682	.526	.451
1944	.500	.539	.532	.429	.461	.532	.422	.584
1945	.539	.523	.497	.425	.467	.653	.464	.435
1946	.325	.519	.558	.403	.435	.682	.571	.506
1947	.461	.545	.481	.448	.370	.494	.617	.584
1948	.381	.664	.374	.494	.390	.455	.614	.634
1949	.377	.591	.422	.435	.370	.474	.656	.675
1950	.390	.610	.403	.383	.364	.662	.623	.565
1951	.435	.474	.396	.526	.364	.545	.662	.597
1952	.506	.474	.396	.675	.383	.487	.584	.494
1953	.451	.422	.403	.610	.344	.617	.649	.500
1954	.552	.390	.279	.558	.331	.669	.649	.571
1955	.455	.409	.396	.487	.377	.591	.630	.656
1956	.455	.448	.429	.468	.370	.662	.552	.617
1957	.468	.416	.503	.494	.364	.614	.500	.643
1958	.487	.468	.497	.500	.403	.526	.516	.604
1959	.513	.390	.422	.506	.487	.571	.519	.591
1960	.578	.435	.506	.539	.370	.623	.422	.526

TABLE 3

AVERAGE LOSING PERCENTAGES (1901-1960)

Risk (Team)	NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
National League	53.4	49.9	47.3	51.8	44.7	56.5	47.8	48.8

Risk (Team)	AL1	AL2	AL3	AL4	AL5	AL6	AL7	AL8
American League	49.5	49.4	47.0	48.5	42.6	52.9	56.4	53.5

Thus there can be no doubt that the teams actually differ.[10] It is therefore a meaningful question to ask whether a given team is better or worse than average.

A team that has been worse than average over one period of time is more likely to be worse than average over another period of time. If this were not true, we would not have found the statistically significant difference in the means of the teams.

Thus if we wish to predict the future experience of a team, there is useful information contained in the past experience of that team. In other words, there is an advantage to experience rating.

4.2 Shifting Parameters Over Time

A similar, but somewhat different question of interest is whether for a given team the results for the different years are from the same distribution (or nearly the same distribution). In other words, are the observed different results over time due to more than random fluctuation? The answer is yes. This is a situation where the underlying parameters of the risk process shift over time.

[10] The situation here is somewhat complicated by the fact that one team's loss is another team's win. Thus the won-loss records of seven teams determine that of the remaining team. However, the author confirmed with a straightforward simulation that in this case this phenomenon would not affect the conclusion. For 8 teams each with the 50% loss rate playing 9000 games each, in 32 out of 600 cases (5%) a team had a winning percentage lower than 49% or more than 51%. In none of the 600 cases did a team have a winning percentage as low as 48% or as high as 52%.

As discussed in Section 2.1, the extent to which risk parameters shift over time has an important impact on the use of past insurance data to predict the future.

Whether the risk parameters shift over time can be tested in many ways. Two methods will be demonstrated here. These methods can be applied to insurance data as well as the data presented here.

The first method of testing whether parameters shift over time uses the standard chi-squared test. For each risk, one averages the results over separate 5 year periods.[11] Then one compares the number of games lost during the various 5 year periods. One can then determine by applying the chi-squared test that the risk process could not have the same mean over this entire period of time. The results shown in Table 4 are conclusive for every single risk. Even the most consistent risk had significant shifts over time.

In the second method of testing whether parameters shift over time, one computes the correlation between the results for all of the risks for pairs of years. Then one computes the average correlation for those pairs of years with a given difference in time. Finally, one examines how the average correlation depends on this difference. The results in our case are displayed in Table 5.

Observed values of the correlation different from zero are not necessarily statistically significant. For this example, a 95% confidence interval around zero for the correlation is approximately plus or minus .10.[12] Thus, for this example, the correlation decreases as the difference in time increases until about ten years when there is no longer a significant correlation between results.[13]

[11] The data were grouped in five year intervals for convenience. Other intervals could also have been used.

[12] For larger distances between the years, we have fewer observations to average, so the confidence interval expands to approximately plus or minus .12. The confidence intervals were determined via repeated simulation in which the actual data for each year were separately assigned to the individual risks at random; thus for the simulated data any observed correlation is illusory.

[13] For a difference of between 15 and 20 years there is again a small but significant positive correlation. The author has no explanation for this long term cycle.

TABLE 4

RESULTS OF CHI-SQUARED TEST OF SHIFTING PARAMETERS OVER TIME

For each risk (team) its experience over the 60 year period was averaged into 12 five-year segments. (The simplifying assumption was made of 150 games each year; this did not affect the results.) Then for each risk separately, the chi-square statistic was computed in order to test the hypothesis that each of the five year segments was drawn from a distribution with the same mean. The resulting chi-square values are:

NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
107	45	98	35	39	73	114	119
AL1	**AL2**	**AL3**	**AL4**	**AL5**	**AL6**	**AL7**	**AL8**
114	69	34	30	97	162	53	65

For example, for the risk (team) NL2 the data by five-year segments are as follows:

	'01 – '05	'06 – '10	'11 – '15	'16 – '20	'21 – '25	'26 – '30	'31 – '35	'36 – '40	'41 – '45	'46 – '50	'51 – '55	'56 – '60
(1) Games Lost*	402	451	412	357	370	389	391	386	326	344	354	310
(2) Expected Games Lost**	374	374	374	374	374	374	374	374	374	374	374	374
(3)=[(1)−(2)]^2/(2)	2.1	15.9	3.9	.8	0	.6	.8	.4	6.2	2.4	1.1	11.0

The sum of row (3) is 45, which is the chi-square value for this risk.

For each risk there is less than a .2% chance that the different five-year segments were drawn from distributions with the same mean.*** Thus we reject the hypothesis that the means are the same over time; we accept the hypothesis of shifting risk parameters over time.

*Assuming 150 games per year, and the observed losing percentage for the five year segment.

**Assuming 150 games per year, and the observed losing percentage for the whole 60 years.

***For 11 degrees of freedom, there is a .16% chance of having a chi-square value of 30 or more. There is a .004% chance of having a chi-square value of 40 or more.

TABLE 5

AVERAGE CORRELATIONS OF RISKS EXPERIENCE OVER TIME (1901-1960)

Difference Between Pairs of Years of Experience	Correlation	
	NL	AL
1	.651	.633
2	.498	.513
3	.448	.438
4	.386	.360
5	.312	.265
6	.269	.228
7	.221	.157
8	.190	.124
9	.135	.078
10	.100	.090
11	.083	.058
12	.103	.063
13	.154	.101
14	.176	.104
15	.180	.141
16	.246	.178
17	.278	.166
18	.219	.198
19	.176	.219
20	.136	.225
21	.090	.159
22	.065	.125
23	.055	.093
24	.004	.048
25	−.024	.006

TABLE 5

(CONTINUED)

Difference Between Pairs of Years of Experience	Correlation	
	NL	AL
26	−.028	.010
27	−.095	−.002
28	−.128	−.013
29	−.107	−.032
30	−.062	.006
31	−.061	−.019
32	−.028	.027
33	−.015	.002
34	.017	.088
35	.038	.143
36	−.014	.156
37	−.024	.214
38	−.012	.238
39	−.017	.138
40	−.095	.093
41	−.174	.055
42	−.216	.028
43	−.332	−.043
44	−.423	−.018
45	−.363	−.035
46	−.332	.066
47	−.324	.069
48	−.373	.136
49	−.423	.075
50	−.475	.145

The correlation between years that are close together is significantly greater than those further apart. This implies that the parameters of the risk process are shifting significantly over time. If the parameters were reasonably constant over time, the correlations would not depend on the length of time between the pair of years.

On the other hand, there is a significant correlation between the results of years close in time. Thus recent years can be usefully employed to predict the future.

5. STATEMENT OF THE PROBLEM

Let X be the quantity we wish to estimate. In this case, X is the expected losing percentage for a risk.

Let Y_1, Y_2, Y_3, etc., be various estimates for X. Then one might estimate X by taking a weighted average of the different estimates Y_i.

$$X = \sum_{i=1}^{n} Z_i Y_i,$$

where X = quantity to be estimated,
Y_i = one of the estimates of X,
Z_i = weight assigned to estimate Y_i of X.

Here only linear combinations of estimators are being considered. In addition, the estimators themselves will be restricted to a single year of past data for the given risk or to the grand mean (which is 50% in this case).[14] No subjective information or additional data beyond the past annual losing percentages will be used.[15] In other words, this is a situation analogous to (prospective) experience rating. This is not a situation analogous to schedule rating.

[14] In other words, in this case, Y either equals the observed losing percentage for the risk in one year or equals the grand mean of 50%. Credibility methods can be applied to more general estimators.

[15] The use of information on the retirement of players or acquisition of new players might enable a significant increase in the accuracy of the estimate. The breakdown of the data into smaller units than an entire year might enable a significant increase in the accuracy of the estimate.

The problem to be considered here is what weights Z_i produce the "best" estimate of future losing percentage. In order to answer that question, criteria will have to be developed that allow one to compare the performance of the different methods to determine which is better. In the example being dealt with in this paper, it is easy to get unbiased estimators. Since all of the estimators being compared will be unbiased, the question of which method is better will focus on other features of the estimators.

Usually the weights Z_i are restricted to the closed interval between 0 and 1. In the most common situation we have two estimates, i.e., $i = 2$. In that case we usually write:

$$X = Z \cdot Y_1 + (1 - Z) \cdot Y_2$$

where Z is called the credibility and $(1 - Z)$ is called the complement of credibility. However, it is important to note that the usual terminology tempts us into making the mistake of thinking of the two weights and two estimates differently. The actual mathematical situation is symmetric.

6. SIMPLE SOLUTIONS TO THE PROBLEM

In this section, various relatively simple solutions to the problem will be presented.

6.1 Every Risk is Average

The first method is to predict that the future losing percentage for each risk will be equal to the overall mean of 50%. This method ignores all the past data; i.e., the past data are given zero credibility. While this is not a serious candidate for an estimation method in the particular example examined in this paper, it is a useful base case in general.

6.2 The Most Recent Year Repeats

The second method is to predict that the most recent past year's losing percentage for each risk will repeat. This is what is meant by giving the most recent year of data 100% credibility.

6.3 Credibility Weight the Most Recent Year and the Grand Mean

In the third method, one gives the most recent year of data for each risk weight Z, and gives the grand mean, which in this case is 50%, weight $1 - Z$.

When $Z = 0$, one gets the first method; when $Z = 1$, one gets the second method. Since each of these is a special case of this more general method, by the proper choice of Z one can do better than or equal to either of the two previous methods. This is an important and completely general result. It does not depend on either the criterion that is used to compare methods or the means of deciding which value of Z to use.

6.4 Determining the Credibility

When employing the third method, the obvious question is how does one determine the value of credibility to use. Ideally one would desire a theory or method that would be generally applicable, rather than one that only worked for a single example. There have been many fine papers on this subject in the actuarial literature.

Generally, the credibility considered "best" is determined by some objective criterion. This will be discussed later.

Using either Bühlmann/Bayesian credibility methods or classical/limited fluctuation credibility methods, one determines which credibility will be expected to optimize the selected criterion in the future. One can also empirically investigate which credibility would have optimized the selected criterion if it had been used in the past; i.e., one can perform retrospective tests. This will be discussed in more detail later.

6.5 Equal Weight to the N *Most Recent Years of Data*

In the fourth method, one gives equal weight to the N most recent years of data for each risk, and gives the grand mean, which in this case is 50%, weight $1 - Z$. This method gives each of the N most recent

years weight of Z/N.[16] When $N = 1$ this reduces to the previous method. Thus this method will perform at least as well as the previous method, with the proper choices of N and Z.

7. CRITERIA TO DECIDE BETWEEN SOLUTIONS

In this section, we will discuss *three* criteria that can be used to distinguish between solutions. These criteria can be applied in general and not just to this example.

7.1 Least Squared Error

The first criterion involves calculating the mean squared error of the prediction produced by a given solution compared to the actual observed result. The smaller the mean squared error, the better the solution.

The Bühlmann/Bayesian credibility methods attempt to minimize the squared error; i.e., they are least squares methods. Minimizing the squared error is the same as minimizing the mean squared error.

7.2 Small Chance of Large Errors

The second criterion deals with the probability that the observed result will be more than a certain percent different than the predicted result. The less this probability, the better the solution.

This is related to the basic concept behind "classical" credibility which has also been called "limited fluctuation" credibility [7]. In classical credibility, the full credibility criterion is chosen so that there is a probability, P, of meeting the test that the maximum departure from expected is no more than k percent.

The reason the criterion is stated in this way rather than the way it is in classical credibility is that, unlike the actual observations, one cannot observe directly the inherent loss potential.[17] However, the two concepts are closely related, as discussed in Appendix G.

---

[16] In later methods, the weights given to the different years of data will be allowed to differ from each other.

[17] It has been shown that the loss potential varies for a risk over time. Thus it cannot be estimated as the average of many observations over time.

7.3 Meyers/Dorweiler

The third criterion has been taken from Glenn Meyers' paper [1]. Meyers in turn based his criterion upon the ideas of Paul Dorweiler [8].

This criterion involves calculating the correlation between two quantities. The first quantity is the ratio of actual losing percentage to the predicted losing percentage. The second quantity is the ratio of the predicted losing percentage to the overall average losing percentage. The smaller the correlation between these two quantities, i.e., the closer the correlation is to zero, the better the solution.

To compute the correlation, the Kendall τ statistic is used.[18] This is explained in detail in Appendix B. The relation of this criterion as used here and as it is used by Meyers to examine experience rating is also discussed in that appendix.

8. THE CRITERIA APPLIED TO THE SIMPLE SOLUTIONS

In this section the three criteria in Section 7 will be applied to the simple solutions given in Section 6. More knowledgeable readers may wish to skip to Section 8.4 which compares the results of applying the three different criteria. Section 8.5 discusses the reduction in squared error. Section 8.6 examines the impact of a delay in receiving data.

8.1 The Two Base Cases

The two simplest solutions either always use as the estimate the overall mean ($Z = 0$), or always use as the estimate the most recent observation ($Z = 1$). While neither of these solutions is expected to be chosen, they serve as the base cases for testing the other solutions.

[18] Meyers in [1] used the Kendall τ statistic. In the example here, any other reasonable measure of the correlation could be substituted.

The first criterion is the smallest mean squared error. For the two data sets the results are:

	Mean Squared Error	
	NL	AL
$Z = 0$	.0091	.0095
$Z = 1$	.0059	.0068

The second criterion is to produce a small probability of being wrong by more than k percent. For the two data sets the results are as follows:

	Percent of time that the estimate is in error by more than 5%		Percent of time that the estimate is in error by more than 10%		Percent of time that the estimate is in error by more than 20%	
	NL	AL	NL	AL	NL	AL
$Z = 0$	82.2%	80.3%	64.8%	63.8%	29.0%	31.4%
$Z = 1$	75.8%	72.9%	52.3%	55.7%	19.1%	22.0%

The third criterion is to have a correlation as close to zero as possible between the ratio of the actual to estimated and the ratio of estimated to the overall mean. For the two data sets the results are as follows:

	Correlation (Kendall τ)	
	NL	AL
$Z = 0$*	.48	.46
$Z = 1$	−.24	−.27

* Limit as Z approaches zero.

8.2 Applying Credibility to the Latest Year of Data

The third prediction method, explained in Section 6.3, uses credibility to combine the latest year of data and the grand mean. The mean squared error depends on the credibility. As shown in Table 6, the mean squared error is a minimum for Z between 60% and 70%.[19] The probability of having errors of 20% or more is displayed in Table 7. Based on this second criterion, the optimal Z is between 50% and 80%.[20] This criterion does not distinguish very sharply between the different values of credibility.

The correlations used in the third criterion are displayed in Table 8. Based on the third criterion the optimal Z is approximately 70%.[21]

8.3 Applying Credibility to the Latest N *Years of Data*

The fourth method, explained in Section 6.4, uses credibility to combine the grand mean with the latest N years of data (giving each year of data the same weight.)

The results of applying the first criterion are shown in Table 6. Based on most actuarial uses of credibility, an actuary would expect the optimal credibilities to increase as more years of data are used. In this example they do not. In fact, using more than one or two years of data does an inferior job according to this criterion.

This result is to be expected, since the parameters shift substantially over time. Thus the use of older data (with equal weight) eventually leads to a worse estimate.[22]

---

[19] For the NL data set, the minimum occurs when $Z = 68\%$. For the AL data set, the minimum occurs when $Z = 66\%$. Also, it should be noted that the squared errors for $Z = 0$ vary somewhat with the number of years of data used, solely due to the differing periods of time over which the test can be performed.

[20] For the NL data set, the optimal Z is 75%. For the AL data set, the optimal Z is 55%. It should be noted that, given the limited number of observations, two values of Z can produce identical results for this criterion.

[21] For the NL data set, the correlation is closest to zero for $Z = 71\%$. For the AL data set, the correlation is closest to zero for $Z = 66\%$.

[22] The number of years of data to use to get the best estimate will depend on the particular example. This general subject was explored in Mahler [5].

TABLE 6

MEAN SQUARED ERROR (.0001)

Z	N = 1	N = 2	N = 3	N = 4	N = 5	N = 7	N = 10	N = 15	N = 20	N = 25
					NL					
0	91	90	90	89	87	84	80	80	80	80
.10	80	80	80	80	79	77	74	75	76	77
.20	70	72	72	72	71	71	70	72	72	74
.30	62	65	65	65	65	66	67	69	69	71
.40	56	59	60	60	60	62	64	67	67	69
.50	52	55	56	56	57	59	63	66	65	68
.60	50	53	53	53	55	57	62	65	64	68
.70	49	52	52	52	53	56	63	65	63	68
.80	51	53	52	52	54	57	64	66	64	69
.90	54	55	53	53	55	58	66	68	65	70
1.00	59	59	56	56	57	61	70	70	66	72
					AL					
0	95	96	96	95	95	95	95	92	91	95
.10	84	85	86	86	88	89	90	88	87	91
.20	75	77	78	79	81	83	86	85	83	87
.30	67	69	71	73	75	79	82	82	80	83
.40	61	64	66	68	71	75	80	81	78	80
.50	58	60	62	64	68	73	78	79	76	78
.60	56	57	60	62	66	71	78	79	74	76
.70	56	56	59	61	66	71	78	79	73	74
.80	58	57	59	62	66	72	79	79	73	73
.90	62	59	61	64	68	74	81	81	73	73
1.00	68	63	64	67	71	77	84	83	74	73

TABLE 7

PERCENT OF TIME THAT THE ESTIMATE IS IN ERROR BY MORE THAN 20%

					NL					
Z	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$	$N = 7$	$N = 10$	$N = 15$	$N = 20$	$N = 25$
0	29	29	29	28	28	27	25	25	25	26
.10	25	26	25	25	25	25	25	24	24	25
.20	23	23	23	23	24	24	23	24	23	26
.30	19	21	22	22	21	22	23	23	23	24
.40	18	19	20	20	22	23	22	22	23	25
.50	17	19	18	19	21	21	21	22	24	25
.60	17	18	18	18	18	21	21	22	24	25
.70	17	18	18	19	19	21	21	22	26	26
.80	17	17	18	19	20	22	23	24	25	27
.90	18	19	17	18	21	22	24	25	25	27
1.00	19	20	18	20	21	23	25	26	25	28
					AL					
0	31	31	32	32	32	32	33	32	32	34
.10	27	27	27	27	27	28	28	28	29	30
.20	23	24	25	25	25	27	27	27	28	30
.30	21	21	22	23	24	25	27	27	27	29
.40	19	19	21	23	24	26	27	26	26	27
.50	18	19	21	22	23	24	28	26	26	27
.60	18	18	21	21	22	25	27	25	26	26
.70	20	18	19	20	22	25	26	25	25	27
.80	19	19	18	20	21	26	27	26	25	26
.90	20	21	19	22	23	27	26	27	25	27
1.00	22	23	22	24	26	28	29	28	27	26

TABLE 8

CORRELATION (KENDALL τ)

NL

Z	N = 1	N = 2	N = 3	N = 4	N = 5	N = 7	N = 10	N = 15	N = 20	N = 25
0*	.48	.45	.46	.45	.43	.39	.31	.28	.29	.25
.10	.44	.41	.42	.41	.40	.35	.27	.24	.26	.22
.20	.38	.36	.37	.36	.35	.30	.23	.20	.22	.18
.30	.32	.30	.31	.31	.30	.25	.18	.16	.18	.14
.40	.25	.24	.25	.25	.24	.20	.12	.11	.14	.10
.50	.17	.17	.19	.19	.18	.14	.07	.06	.10	.06
.60	.09	.09	.12	.12	.12	.08	.02	.02	.05	.02
.70	.01	.02	.04	.05	.05	.02	−.04	−.03	.01	−.02
.80	−.08	−.06	−.03	−.02	−.02	−.04	−.09	−.07	−.03	−.06
.90	−.16	−.13	−.01	−.09	−.08	−.10	−.14	−.12	−.08	−.10
1.00	−.24	−.21	−.17	−.16	−.15	−.16	−.19	−.16	−.12	−.14

*Limit as Z approaches zero

CORRELATION (KENDALL τ)

AL

Z	N = 1	N = 2	N = 3	N = 4	N = 5	N = 7	N = 10	N = 15	N = 20	N = 25
0*	.46	.45	.44	.41	.38	.34	.28	.24	.27	.30
.10	.42	.41	.40	.38	.35	.30	.25	.21	.25	.28
.20	.36	.36	.35	.33	.30	.26	.21	.17	.21	.25
.30	.29	.30	.30	.27	.25	.21	.16	.13	.18	.22
.40	.22	.24	.23	.22	.19	.16	.12	.09	.15	.19
.50	.14	.16	.17	.15	.13	.11	.07	.05	.11	.16
.60	.05	.08	.10	.08	.07	.05	.02	.01	.07	.12
.70	−.03	.00	.02	.02	.00	−.01	−.03	−.03	.03	.09
.80	−.11	−.07	−.05	−.05	−.06	−.07	−.07	−.07	−.01	.05
.90	−.19	−.15	−.12	−.12	−.12	−.12	−.12	−.11	−.05	.01
1.00	−.27	−.22	−.19	−.18	−.18	−.17	−.16	−.15	−.09	−.02

*Limit as Z approaches zero

The results of applying the second criterion are displayed in Table 7. This criterion does not sharply distinguish between the different values of credibility. There is a broad range of credibilities all of which do reasonably well.[23] This is particularly true for larger values of N. Again the use of more years of data eventually leads to an inferior estimate.

The results of applying the third criterion are displayed in Table 8. Again the optimal credibility does not increase as N increases. Unlike the other criteria, the third criterion cannot be used to distinguish between values of N. For each N, there is a Z, such that the correlation is zero. Thus each value of N performs as well as all the others.

Meyers points out that the distribution of Kendall's τ can be used to obtain a confidence interval for the credibility. As explained in Appendix B, for this example a 95% confidence interval for τ around zero has a radius of about .07.

For example, using 10 years of data, the optimal credibility using the Meyers/Dorweiler criterion for the NL set of data is 63%. However, this point estimate for the credibility is actually an estimate of an interval of credibilities that correspond to τ between plus and minus .07. The optimal credibility is $63\% \pm 13\%$.[24]

8.4 Comparison of the Results of the Three Criteria

In Table 9 the optimal credibilities are displayed as determined by the three criteria for various values of N. Note that the listed values of credibility are those that happened to work best over the period of time observed. Values close to these values would also work well over this period of time.

One should think of the point estimates listed in Table 9 as the centers of interval estimates. This is illustrated when one compares the different estimates obtained by analyzing the NL and AL data sets. There is no inherent difference in the two data sets. Thus one would expect the credibilities from the two analyses to be the same. They are similar,

[23] This is true to a lesser extent for the first criterion. This subject is explored in Mahler [9].

[24] For $Z = 63.3\%$, $\tau = 0$. For $Z = 50.1\%$, $\tau = .07$. For $Z = 76.4\%$, $\tau = -.07$.

but far from identical. This indicates that the peculiarities of the specific observed values are sufficient to affect the answers somewhat. There is some lack of precision in the estimates in Table 9.

TABLE 9

OPTIMAL CREDIBILITY

Number of Years of Data Used	NL			AL		
	Criterion #1	Criterion #2	Criterion #3	Criterion #1	Criterion #2	Criterion #3
1	68%	75%	71%	65%	55%	66%
2	71	80	72	70	56	70
3	74	87	76	72	77	73
4	76	57	77	72	69	72
5	74	61	77	70	70	71
7	71	64	73	67	51	68
10	60	49	63	62	70	64
15	63	43	64	65	69	62
20	71	40	73	81	82	77
25	64	30	64	97	61	94

Criterion #1: Least Squares (Section 7.1)
Criterion #2: Small Chance of Large Errors (Section 7.2)
Criterion #3: Meyers/Dorweiler (Section 7.3)

This can be illustrated further by reversing the time arrow and analyzing the data sets going backwards in time rather than forwards. For example, one could use data from years 1902 to 1911 to "predict" 1901. This analysis is equally valid for determining optimal credibilities in this example as was the original anlysis.

For $N = 10$, one gets the following optimal credibilities for the different data sets, where NLR and ALR represent respectively the NL and AL data sets reversed in time.

	Optimal Credibilities (N = 10)				
	NLR	NL	ALR	AL	Average
Criterion #1	72%	60%	57%	62%	63%
Criterion #2	58	49	42	70	55
Criterion #3	77	63	58	64	65

The optimal credibilities differ between the four data sets. The amount of variation provides some idea of the imprecision of the different estimates. While the optimal credibilities differ between the three criteria, the differences do not appear to be sufficiently large to allow one to draw any definitive conclusions.

In this case, the use of any value of credibility between 50% and 70% would perform reasonably well according to all three criteria for all four data sets. As a practical matter, the difference in the predictions will not vary that much depending on which value of credibility is chosen in that range.[25]

In most applications of credibility, values for the credibility that differ somewhat from optimal perform reasonably well and the choice between these values has a relatively small practical impact.

8.5 Putting the Reduction in Squared Error in Context

The first criterion used to determine the optimal credibility is to minimize the squared error. Using the optimal credibility based on this criterion will reduce the squared error between the observed and predicted result. What should be considered a significant reduction in squared error?

[25] The maximum difference in any prediction for N = 10 between using 50% and 70% credibility is 3.3% in the losing percentage. In most cases it is much smaller. On average it would make about a 1% difference.

Let us examine an example. For the NL data, using one year of data, the optimal credibility is 68% as shown in Table 9. As shown in Table 6 the mean squared errors are:

Z	Mean Squared Error
0	.0091
68%	.0049
100%	.0059

In this case, by the use of credibility, the squared error has been reduced from .0059 if the data were relied upon totally, or .0091 if the data were totally ignored, to .0049. In this case, the squared error has been reduced to 83% (.0049/.0059) of its previous value.[26]

As discussed in Appendix E, in the current case, the best that can be done using credibility to combine two estimates is to reduce the mean squared error between the estimated and observed values to 75% of the minimum of the squared errors from either relying solely on the data or ignoring the data.[27]

The reduction of the squared error to 83% of its previous value appears significant in light of the maximum possible reduction to 75%.[28]

8.6 Effect of Delay in Receiving Data

It has been shown previously for the data set examined in this paper that the further apart in time two years are, the lower the correlation between them. Thus if there is a delay before the data are available for use in experience rating, the resulting estimate of the future will be less accurate.

---

[26] The "previous" value of the squared error is considered to be the minimum of the squared errors that result from either ignoring the data entirely or relying on the data entirely.

[27] When using more than two or more years of data, the reduction in squared error depends on the impact of shifting parameters over time. However, in the absence of shifting parameters over time, for N years with the same weight applied to each year, the maximum possible reduction is $1 \div (2(N + 1))$.

[28] The maximum reduction is possible when the squared errors for $Z = 0$ and $Z = 1$ are equal.

As is shown in Table 10, as the delay increases, the squared error increases significantly. The increase in squared error is particularly significant as one goes from a situation of having the data from the most recent year available to predict the coming year to a situation of having

TABLE 10

MINIMUM SQUARED ERROR (.0001)

Time Between Latest Data Point and Future Prediction	NL				
	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
1	49	52	51	51	53
2	66	62	60	60	60
3	69	66	65	64	65
4	73	71	69	69	70
5	77	73	73	72	72
6	76	75	75	73	74
7	78	77	75	75	75
8	79	77	77	76	75
9	78	78	77	76	75
10	78	78	76	75	75

Time Between Latest Data Point and Future Prediction	AL				
	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
1	56	56	59	61	66
2	71	70	71	74	76
3	78	77	80	81	83
4	83	85	85	87	88
5	89	89	90	91	91
6	91	91	92	93	93
7	93	93	94	93	94
8	95	94	94	94	93
9	95	94	94	93	93
10	94	94	93	93	94

only the next most recent year available. Unfortunately, the latter situation is more common in insurance than is the former.

As is shown in Table 11, the optimal credibility (as determined using the least squares criterion) decreases as the delay increases. Less current information is less valuable for estimating the future.

TABLE 11

OPTIMAL CREDIBILITY (CRITERION #1, LEAST SQUARES)

Time Between Latest Data Point and Future Prediction	NL				
	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
1	68	71	74	76	74
2	51	59	64	64	63
3	47	53	55	56	55
4	40	45	47	47	45
5	33	38	40	39	36
6	30	33	34	32	30
7	24	26	26	25	24
8	19	20	21	21	20
9	14	16	17	18	20
10	11	13	15	18	21

Time Between Latest Data Point and Future Prediction	AL				
	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
1	65	70	72	72	70
2	51	57	58	57	56
3	42	47	46	46	45
4	35	36	37	36	36
5	25	28	28	28	25
6	21	22	22	19	18
7	15	16	14	14	13
8	11	9	10	10	9
9	6	7	8	7	9
10	7	7	7	9	10

9. MORE GENERAL SOLUTIONS

In Section 6, four relatively simple forms of a solution were given. In this section, more general forms of a solution will be given.

9.1 Combine Previous Estimate and Most Recent Data

In the fifth method, one gives the latest year of *data* weight Z, and gives the previous *estimate* weight $1 - Z$. Of course, one has to choose an initial estimate.[29] In this case, for each risk the initial estimate will be taken as the grand mean of 50%.[30] Once this estimation method has been used for several years, the initial estimate has very little weight.

For example, let us assume $Z = 60\%$. Then the weights assigned to the given years of data used in estimating the result for the year 1911 would be as follows:

Year of Data	Weight in Estimate of 1911
1910	$Z = 60\%$
1909	$Z(1 - Z) = 40\% \times 60\% = 24\%$
1908	$Z(1 - Z)^2 = 40\% \times 40\% \times 60\% = 9.6\%$
1907	$Z(1 - Z)^3 = 9.6\% \times 40\% = 3.84\%$
1906	$Z(1 - Z)^4 = 3.84\% \times 40\% = 1.54\%$
1905	$Z(1 - Z)^5 = 1.54\% \times 40\% = .61\%$
1904 and Prior	$(1 - Z)^6 = .41\%$

The above assumes that the latest year of data is always given 60% weight, while the current estimate is given 40% weight.

Thus in this case, one gets a geometrically decreasing weight. This procedure is called (single) exponential smoothing [10]. It is an example of what mathematicians call a "filter."[31] Once the process of exponential

[29] This is precisely analogous to choosing a "seed" value in exponential smoothing.

[30] One could use subjective judgement to choose the initial estimate for each risk. Also one could use data from the period prior to that displayed in this paper; this has been avoided for the sake of simplicity.

[31] Morrison [11] gives this as an example of a "fading-memory polynomial filter."

smoothing gets "up to speed," it is equivalent to a weighted least squares regression, where the fitted line is horizontal,[32] and where the weights are geometrically decreasing as the data get less recent.

9.2 More General Varying Weights

In Section 9.1, one gave geometrically decreasing weight to years of data further in the past. More generally one can make the estimate:

$$F = \Sigma Z_i X_i + (1 - \Sigma Z_i)M$$

where the weights Z_i depend on how far in the past are the data X_i. For years for which data are not available (presumably because they are too far in the past) one uses the grand mean M instead of the data. This method is a generalization of the methods in Sections 6 and 9.1.

Unfortunately, calculating or empirically determining the optimal values of the weights Z_i becomes difficult as more years of data are used. Also, there are many vectors of Z_i that are very close to optimal; i.e., the n-dimensional volume of values $Z_1, \ldots, Z_n$ that produce close to optimal results is relatively large.

10. THE CRITERIA APPLIED TO THE MORE GENERAL SOLUTIONS

In this section the three criteria in Section 7 will be applied to the more general solutions to the problem given in Section 9. For simplicity, the results will be shown for the situation where there is no delay in obtaining the data for use in making the next estimate. In Section 8.6, an example was given of the results of such a delay in receiving data. The same general pattern would apply here.

10.1 Geometrically Decreasing Weights

In Section 9.1, weight Z is applied to the latest available year of data, while weight $1 - Z$ is applied to the previous estimate.

---

[32] Double exponential smoothing, sometimes called linear exponential smoothing, would be equivalent to a weighted linear least squares regression, with geometrically decreasing weights as the data got less recent.

Table 12 gives the mean squared errors for various values of Z. The optimal values of Z, using criterion #1 (least squares), are all close to 55%.[33] This results in weights to the various years of data very similar to those in the example in Section 9.1.

TABLE 12

MEAN SQUARED ERRORS* (.0001) THAT RESULT FROM APPLYING Z TO LATEST YEAR OF DATA AND $1 - Z$ TO PREVIOUS ESTIMATE

Z	NL	NLR**	AL	ALR**
0	79	97	95	96
.1	61	70	72	78
.2	56	63	65	71
.3	52	60	60	67
.4	50	57	57	64
.5	49	56	55	63
.6	50	56	55	63
.7	50	57	55	64
.8	52	58	56	66
.9	54	59	58	69
1.0	57	62	61	73

* First 10 years are not included in the computation of the squared errors in order to eliminate the calibration period.

** Data reversed in time.

In this case there is no significant reduction in squared error beyond what was previously obtained by applying credibility to the latest available year.[34]

Table 13 displays the results of applying criterion #2, limited fluctuation. Values of the credibility between 40% and 80% generally perform well.

[33] For the NL data set the optimal credibility is 53%. For the NLR data set, it is 58%. For the AL data set it is 60%. For the ALR data set it is 54%.

[34] Compare the results in Table 6 for $N = 1$ with those in Table 12.

Table 14 displays the results of applying criterion #3, Meyers/Dorweiler.[35] Unlike the previous two cases, the optimal credibilities are close to zero; 5% to 10% credibility produces correlations close to zero. The use of such small credibilities is approximately the same as using 10 to 20 years of data as the basis for the estimate, since the geometrically decreasing weights decline only slowly.

TABLE 13

PERCENT OF TIME* THAT THE ESTIMATE IS IN ERROR BY MORE THAN 20%
APPLYING Z TO LATEST YEAR OF DATA AND $1 - Z$ TO PREVIOUS ESTIMATE

Z	NL	NLR**	AL	ALR**
0	25	31	33	31
.1	23	23	25	26
.2	21	22	24	25
.3	18	21	21	23
.4	16	21	20	21
.5	16	20	19	22
.6	16	20	19	21
.7	17	19	20	22
.8	18	19	19	22
.9	18	20	18	23
1.0	19	21	19	26

* First 10 years are not included in the computation in order to eliminate the calibration period.

** Data reversed in time.

[35] In this case, the results of the first 20 years were excluded from the computation, in order to eliminate the calibration period. Twenty years were used, rather than ten years as in the previous two tables, since in this case smaller credibilities are optimal and smaller credibilities require a longer calibration period.

TABLE 14

CORRELATIONS* (KENDALL TAU) THAT RESULT FROM APPLYING Z TO LATEST YEAR OF DATA AND $1 - Z$ TO PREVIOUS ESTIMATE

Z	NL	NLR**	AL	ALR**
0***	.11	.16	.28	.14
.1	−.03	.01	.00	−.09
.2	−.05	−.05	−.04	−.10
.3	−.08	−.09	−.07	−.12
.4	−.10	−.12	−.10	−.13
.5	−.13	−.15	−.12	−.15
.6	−.15	−.18	−.14	−.17
.7	−.18	−.20	−.16	−.20
.8	−.20	−.23	−.18	−.22
.9	−.23	−.25	−.21	−.24
1.0	−.26	−.27	−.23	−.28

* First 20 years are not included in the computation of the correlations in order to eliminate the calibration period.

** Data reversed in time.

*** Limit as Z approaches zero.

10.2 More General Varying Weights

In Section 9.2, varying weights Z_i are applied to the most recent N years, while the remaining weight is given to the grand mean. This method will only be examined using criterion #1, least squares. One can solve numerically for the set of weights which produce the least squared error, using a given number of years of data.[36] The results are as follows:

[36] Unfortunately, as the number of years increases, the amount of computer time required also increases.

Using Most Recent Two Years of Data ($N = 2$, $\Delta = 1$)

	Credibility		
	Second Most Recent Year	Most Recent Year	Mean Squared Error (.0001)
NL	9.6%	61.1%	48
AL	13.1%	56.9%	54

Using Most Recent Three Years of Data ($N = 3$, $\Delta = 1$)

	Credibility			
	Third Most Recent Year	Second Most Recent Year	Most Recent Year	Mean Squared Error (.0001)
NL	16.4%	1.1%	59.0%	45
AL	8.1%	9.1%	55.7%	53

Most of the credibility is assigned to the most recent year. The complement of credibility, which is assigned to the grand mean, is about 25 to 35 percent, decreasing as N increases.

	Complement of Credibility		
	$N = 1$*	$N = 2$	$N = 3$
NL	32%	29%	24%
AL	35%	30%	27%

* One minus the optimal credibility from Table 9.

The mean squared error is reduced from that using only the latest year of data.[37]

[37] Since the use of fewer years of data is just a special case, the least squared error using more years of data must be less than or equal that using fewer years of data.

	Mean Squared Error (.0001)		
	$N = 1$*	$N = 2$	$N = 3$
NL	49	48	45
AL	56	54	53

* From Table 6.

11. EQUATIONS FOR LEAST SQUARES CREDIBILITIES

In Section 11.2 are equations to solve for the least squares credibility. These equations follow from the assumed covariance structure discussed in Section 11.1. In Section 11.3 the equations in Section 11.2 are modified to constrain them to place no weight on the grand mean. Section 11.4 compares the mean squared errors that result from different credibilities. Section 11.5 briefly discusses the validity of the results derived in this paper.

11.1 The Covariance Structure

By analyzing the covariance structure, one can set up matrix equations to solve for the credibilities that minimize the squared error. These matrix equations are discussed in the next section.

As shown in Appendix D, the variance of the data can be broken down into two pieces. There is the variance between the risks.[38] There is also the variance within the risks.[39] These two variances add up to the total variance.

	Between Variance	Within Variance	Total Variance[40]
NL	.001230	.007892	.009121
AL	.001619	.007875	.009494

[38] This has been denoted as τ^2.

[39] This has been denoted as $\delta^2 + \zeta^2$. δ^2 is what is usually termed process variance, while ζ^2 is the variance due to shifting parameters over time.

[40] May differ slightly from the sum of the other two variances due to rounding.

Also of interest is the covariance between the years of data. It is assumed that this is a function of the number of years separating the data. The observed values are given in Table 15. As was seen in Table 5, the covariance decreases as the years of data are further apart. After about 6 years the covariances are relatively close to zero.

TABLE 15

COVARIANCE (.0001)

Years Separating Data	NL	AL
0*	7892	7875
1	4919	4527
2	3416	3175
3	3128	2411
4	2541	1766
5	1810	780
6	1566	383
7	955	−99
8	387	−561
9	−74	−1068
10	−394	−878
11	−558	−980
12	−389	−1092
13	3	−737
14	59	−814
15	212	−453
16	603	−39
17	786	−139
18	302	214
19	47	279
20	−268	415

*Equal by definition to the within variance.

It is possible to divide the within variance into two parts. The first part is the process variance excluding the effect of shifting parameters over time.[41] The second part is that portion of the within variance due to shifting parameters over time.[42] While this division may aid our understanding, it is not necessary for the calculation of the least squares credibilities. Not coincidentally, this division cannot be performed based solely on the reported data contained in Tables 1 and 2. This subject is discussed in more detail in Appendix D.

11.2 Matrix Equations for Least Squares Credibilities

Using the estimation method described in Section 9.2:

$$F = \sum_{i=1}^{N} Z_i X_i + (1 - \sum_{i=1}^{N} Z_i)M \tag{11.1}$$

As derived in Appendix C, one gets the following expression for the expected squared error between the observation and prediction:

$$\begin{aligned} V(Z) = & \sum_{i=1}^{N} \sum_{j=1}^{N} Z_i Z_j \, (\tau^2 + C(|i - j|)) \\ & - 2 \sum_{i=1}^{N} Z_i \, (\tau^2 + C(N + \Delta - i)) \\ & + \tau^2 + C(0) \end{aligned} \tag{11.2}$$

In equation (11.2) we have used the following quantities defined in Appendix D.

τ^2 = between variance
$C(k)$ = covariance for data for the same risk, k years apart
= "within covariance"
$C(0)$ = within variance
Δ = the length of time between the latest year of data used and the year being estimated

[41] This has been denoted as δ^2.
[42] This has been denoted as ζ^2.

Equation 11.2 shows that the squared error is a second order polynomial in the Z_i.[43] This equation is the fundamental result for analyzing least squares credibility.

One can differentiate equation 11.2 in order to get N linear equations in N unknowns, which can be solved for the optimal credibilities.

$$\sum_{j=1}^{N} Z_j(\tau^2 + C(|i - j|)) = \tau^2 + C(N + \Delta - i) \quad i = 1, 2, \ldots N \qquad (11.3)$$

The set of equations 11.3 can be solved on a computer relatively easily using the usual methods from matrix theory. The results of doing so for $\Delta = 1$, using the average of the variances and covariance determined from the NL and AL data separately,[44] are shown in Table 16.

TABLE 16

LEAST SQUARES CREDIBILITIES, SOLUTIONS OF MATRIX EQUATIONS 11.3 ($\Delta = 1$)

Number of Years of Data Used (N)	Years Between Data and Estimate									
	1	2	3	4	5	6	7	8	9	10
1	66.0%	—	—	—	—	—	—	—	—	—
2	57.7	12.6	—	—	—	—	—	—	—	—
3	56.1	4.8	13.5	—	—	—	—	—	—	—
4	55.6	4.6	11.5	3.5	—	—	—	—	—	—
5	55.7	5.1	11.7	6.0	−4.4	—	—	—	—	—
6	55.9	4.9	11.3	5.8	−6.6	3.9	—	—	—	—
7	56.0	4.7	11.5	6.2	−6.5	5.9	−3.5	—	—	—
8	56.0	4.7	11.4	6.2	−6.3	5.9	−2.8	−1.2	—	—
9	56.1	4.9	11.0	6.6	−6.7	5.3	−3.1	−4.3	5.6	—
10	55.9	5.0	11.2	6.4	−6.4	5.1	−3.4	−4.5	3.6	3.5

The complement of credibility is applied to the grand mean.

First column is the credibility applied to the most recent year, second column is the credibility applied to the next most recent year, etc.

Note: Based on the average of the variances and covariances determined from the NL and AL data separately; however, assumes that for a separation of eight years or more, the covariance is zero.

[43] When $N = 1$, the squared error is a parabola as a function of the credibility. This has been noted before, for example in Appendix B of Meyers [12].

[44] It is assumed that for a separation of eight years or more, the covariance is zero.

The results conform reasonably well to those determined in Section 10.2.

The credibilities applied to the most recent year quickly converge to about 56%. The credibilities for the less recent years are much smaller. However, these credibilities do not monotonically decline as the years get less recent. There is a complicated pattern of weights determined by the covariance matrix. Some of the weights are even less than zero.[45]

The optimal credibilities are uniquely determined given the covariance structure. However, there are many other sets of credibilities which produce expected squared errors very close to minimal. The precise values of the credibilities are not particularly important, although the general range of credibilities that perform well might be instructive.

One can apply equation (11.2) to the method discussed in Sections 6.5 and 8.3 of applying equal weight Z_i to the latest N years of data, where

$Z_i = Z/N$ for $i = 1, \ldots, N$

As shown in Appendix C, the least squares credibility in this case is given by:

$$Z = N \frac{N\tau^2 + \sum_{i=1}^{N} C(N + \Delta - i)}{N^2\tau^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} C(|i - j|)} \tag{11.4}$$

The results of applying this equation for $\Delta = 1$, using the average of the variances and covariances determined from the NL and AL data separately,[46] are shown in Table 17.

Table 17 can be compared to Table 11.

The results in Table 11 conform reasonably well to those determined empirically for each data set (for $\Delta = 1$).

[45] Giving negative weight to some years allows a larger weight to be given to other years. The net effect is to reduce the expected squared error.

[46] It is assumed that for a separation of eight years or more, the covariance is zero.

TABLE 17

LEAST SQUARES CREDIBILITY, SOLUTION TO EQUATION 11.4 ($\Delta = 1$)

Number of Years of Data Used (N)	Z	$Z \div N$
1	66.0%	66.0%
2	70.3	35.2
3	72.9	24.3
4	73.6	18.4
5	72.2	14.4
6	71.3	11.9
7	69.9	10.0
8	68.2	8.5
9	67.3	7.5
10	66.9	6.7

Equal weight Z/N is applied to each of the N most recent years of data. The complement of credibility, $1 - Z$, is applied to the grand mean.

Note: Based on the average of the variances and covariances determined from the NL and AL data separately; however, assumes that for a separation of eight years or more, the covariance is zero.

11.3 Placing No Weight on the Grand Mean

Once the estimation method described in Sections 9.1 and 10.1 gets "up to speed," the initial estimate, which was taken as the grand mean, has very little weight. For all intents and purposes each risk is estimated based on its own past data, without relying on the data of other risks, in particular the grand mean.[47]

[47] The covariance structure is herein estimated using the data for all risks. This in turn is used to estimate the optimal credibilities. However, the credibilities are applied to the data for the particular risk we are estimating.

One can constrain the credibilities used in equation 11.1, so that they add to unity, thus giving no weight to the grand mean. Equation 11.1 then becomes

$$F = \sum_{i=1}^{N} Z_i X_i \tag{11.5}$$

with the constraint

$$\sum_{i=1}^{N} Z_i = 1. \tag{11.6}$$

The least squres credibilities for equations 11.5 and 11.6 are derived in Appendix C using the method of Lagrange Multipliers. The result is a set of $N + 1$ linear equations in $N + 1$ unknowns, the Z_i for $i = 1, \ldots, N$, and λ, the Lagrange Multiplier. There is the single constraint equation 11.6, plus the N equations 11.7.

$$\sum_{j=1}^{N} Z_j \, C(|i - j|) = C(N + \Delta - i) + \frac{\lambda}{2}, \quad i = 1, 2, \ldots, N \tag{11.7}$$

The set of equations 11.6 and 11.7 can be solved on a computer relatively easily using the usual methods from matrix theory. The results of doing so for $\Delta = 1$, using the average of the variances and covariances determined from the NL and AL data separately,[48] are shown in Table 18.

11.4 Mean Squared Errors

The mean squared errors that result from using the credibilities in Tables 16, 17, and 18 are displayed in Table 19.

When applying general weights to the latest N years of data, giving the most remote year of data no weight is equivalent to the case of using the latest $N - 1$ years of data. Since using the latest $N - 1$ years of data is a special case of using the latest N years of data, we expect the squared errors to decline, or remain constant.

This is what we observe for the credibilities from Table 16. They decline until $N = 6$, where the point of diminishing returns is reached.

[48] It is assumed that for a separation of eight years or more, the covariance is zero.

TABLE 18

LEAST SQUARES CREDIBILITIES, SOLUTIONS OF EQUATIONS 11.6 AND 11.7 ($\Delta = 1$)

Number of Years of Data Used (N)	Years Between Data and Estimate									
	1	2	3	4	5	6	7	8	9	10
1	100.0%	—	—	—	—	—	—	—	—	—
2	72.6	27.4	—	—	—	—	—	—	—	—
3	66.1	10.3	23.6	—	—	—	—	—	—	—
4	63.5	9.1	16.0	11.4	—	—	—	—	—	—
5	63.1	8.7	15.8	9.5	2.9	—	—	—	—	—
6	62.8	7.6	14.1	8.6	−3.9	10.8	—	—	—	—
7	62.5	7.7	13.8	8.2	−4.1	9.0	2.9	—	—	—
8	62.3	7.3	14.0	7.7	−4.8	8.6	−0.2	5.1	—	—
9	61.8	7.3	13.0	8.3	−5.7	7.0	−1.1	−1.9	11.2	—
10	60.8	7.5	13.1	7.7	−5.2	6.3	−2.2	−2.5	6.1	8.4

The credibilities are constrained to sum to unity.

First column is the credibility applied to the most recent year, second column is the credibility applied to the next most recent year, etc.

Note: Based on the average of the variances and covariances determined from the NL and AL data separately; however, assumes that for a separation of eight years or more, the covariance is zero.

Applying the same weight to each year is a special case of using the general weights. Thus the squared errors that result from using the credibilities from Table 17 should be greater than or equal to those that result from the credibilities from Table 16. This is the case, as shown in Table 19. Also, as was observed in Section 8.3, using more years of data leads in this case to larger squared errors.

Applying no weight to the grand mean is a special case of using the general weights. Thus the squared errors that result from using the credibilities from Table 18 should be greater than or equal to those that result from the credibilities from Table 16. As is shown in Table 19, the squared errors are substantially greater, with the gap narrowing as the number of years increases.

TABLE 19

Number of Years of Data Used (N)	Mean Squared Errors (.0001)*		
	Using the Credibilities From Table 16**	Using the Credibilities From Table 17***	Using the Credibilities From Table 18****
1	52	52	63
2	51	54	58
3	49	55	54
4	48	57	52
5	48	60	52
6	47	61	51
7	47	64	51
8	47	66	51
9	47	68	51
10	47	70	50

* Mean squared error using the stated credibilities to predict for the NL and AL data sets.

** The complement of credibility is given to the grand mean.

*** Equal weight to N years, with the complement of credibility given to the grand mean.

**** The credibilities add up to one, and thus no weight is given to the grand mean.

11.5 Validity of Results

The credibilities determined in Sections 10 and prior are all determined empirically by directly working with the data. In this section equations for the least squares credibilities have been introduced along with an assumed covariance structure.

The credibilities resulting from the use of the equations in this section are comparable to those determined in the previous sections empirically. As is shown in Appendix F, the observed pattern of squared errors is comparable to that derived from the assumed covariance structure.

Therefore, the results of this section are an appropriate means of estimating least squares credibilities for this example. How well these results would apply to another situation would depend on the covariance structure that underlies the particular data set.

12. MISCELLANEOUS

Section 12.1 contrasts the Meyers/Dorweiler Criterion vs. the other criteria. Section 12.2 discusses a somewhat artificial ratemaking example. It is intended to point the way towards applying these or similar methods to practical situations. Section 12.3 compares the baseball example to typical insurance applications. Section 12.4 shows that the estimates that result herein from the use of the credibilities are in balance. Section 12.5 discusses the question of what estimation method to select for predicting the future loss record of baseball teams. It is included in order to complete the illustrative example used throughout this paper.

12.1 Contrasting the Meyers/Dorweiler Criterion vs. the Other Criteria

Section 10.1 provides a good example of how criterion #3, Meyers/Dorweiler, differs on a basic conceptual level from the first two criteria. Both of the other criteria are concerned with eliminating large errors.[49] Criterion #1, least squares, does this since even a few large errors will

[49] Mahler [7] compares the credibilities that result from the application of the Bühlmann/least squares criterion and the credibilities that result from the application of the classical/limited fluctuation criterion.

greatly increase the sum of squared errors. Criterion #2, limited fluctuation, does this directly by minimizing the number of errors larger than the selected size.

In contrast, criterion #3, Meyers/Dorweiler, is concerned with the pattern of the errors. Large errors are not a problem, as long as there is no pattern relating the errors to the experience rating modifications. For example, consider the following two situations. In each case, for simplicity, only four risks are assumed.

Situation #1	
Modification	Error
1.20	+30%
1.20	−30%
.80	+40%
.80	−40%

Situation #2	
Modification	Error
1.30	+2%
1.10	+1%
.90	−1%
.70	−2%

Situation #2 with its small errors is preferable under the first two criteria. Situation #1 with its lack of a pattern of errors is preferable under the Meyers/Dorweiler criterion. Most actuaries would prefer Situation #2.

This example is not meant to discourage use of the Meyers/Dorweiler criterion. Rather it is meant to point out the potential hazards of relying solely on any single criterion, as well as the importance of understanding exactly what is being tested by any criterion that is being used.

12.2 A Ratemaking Example

Assume for a given line of insurance that the five most recent annual loss ratios are being combined to calculate a rate level indication.[50] Assume that it is three years from the latest year of data to the average date of loss under the proposed new rates.[51] A weighted average of the annual loss ratios will be used to estimate the future loss ratio.

If we assume a given covariance structure, equations 11.6 and 11.7 can be used to calculate the optimal least squares set of weights, Z_i, such that

$$\sum_{i=1}^{5} Z_i = 1.$$

Assume the covariance of the loss ratios separated by a given number of years is as follows:[52]

Separation in Years	Covariance in Loss Ratios (.00001)
0	130
1	60
2	55
3	50
4	45
5	40
6	35
7	30

Then the optimal weights are: 11.6%, 13.4%, 17.3%, 23.8%, 33.9%, with the more recent data receiving more weight. It is interesting to note that these weights can be reasonably approximated by the weights used in Walters [13], i.e., 10%, 15%, 20%, 25%, and 30%.

This example is for illustrative purposes only. It should not be taken as a derivation of the correct weights to use in any real world application. Unfortunately, in order to apply this idea to real world applications one

[50] The loss ratios for the separate years are presumed to have been adjusted for trend, development, and any other factors such as law changes.

[51] This period will vary, but $\Delta = 3$ is not uncommon.

[52] This would be produced by $\delta^2 = .0004$, $\zeta^2 = .0009$, $\ell(1) = .667$, $\ell(2) = .611$, $\ell(3) = .556$, $\ell(4) = .500$, $\ell(5) = .494$, $\ell(6) = .389$, $\ell(7) = .333$, where the quantities are defined as in Appendix D.

has to estimate the covariance matrix. This will be affected by shifting parameters over time. It will also be affected by the varying quantity of data available in each year.[53] It will be affected by the uncertainty in the trend estimates and loss development estimates applied to each year. These complications are beyond the scope of this paper.

12.3 Baseball Example vs. Typical Insurance Applications

In many typical insurance applications, credibility is used in the process of determining relativities. For example, credibility is used to determine the rate for a class or territory relative to the overall rate level. Credibility is also used in experience rating, where the rate for an individual risk is adjusted relative to an average.

In these situations, where a class, territory, or individual risk is compared relative to an average, the result depends on the other classes, territories, or risks which make up the average. An automobile territory with a low relativity in Massachusetts could have a higher loss potential than a automobile territory with a high relativity in Vermont. A workers compensation insured could have a credit experience modification simply because of the bad loss experience of several other employers in the same business in the same state. An insured with a .9 experience modification could have a higher loss potential than another risk with 1.1 modification in a different business or in a different state. The baseball example has this same feature. A team is being compared relative to the average in the league. The losing percentage only has relevance to rank teams in a single league relative to the average for that league. The difference in this example is that the average is a known constant. The grand mean is always .500.

In baseball if somebody loses, then somebody else wins. Thus the win-loss records of seven teams determine that of the remaining team in an eight team league.[54]

---

[53] In this paper, each year of baseball data represented a comparable number of games, so this aspect was not important.

[54] The win-loss record of teams in the same league should be negatively correlated by an amount proportional to the number of games the two teams have played.

This could have had a major impact on the analysis of this example. However, each team played each other team in the league approximately the same number of times each year and each team played approximately the same number of games in total.[55] Thus no team had its results distorted by playing a weaker or stronger schedule.

12.4 Estimates in Balance

The estimation methods used herein are always in balance.

The most general estimation method considered herein was given by equation 9.1, where the subscript j has been added to identify team j:

$$F_j = \sum_{i=1}^{N} Z_i X_{ij} + \left(1 - \sum_{i=1}^{N} Z_i\right) M$$

Then the average of the estimates F_j for all the teams in the league is given by:

$$\begin{aligned} \frac{1}{8} \sum_{j=1}^{8} F_j &= \sum_{i=1}^{N} Z_i \left(\frac{1}{8} \sum_{j=1}^{8} X_{ij} \right) + \left(1 - \sum_{i=1}^{N} Z_i\right) M \\ &= \left(\sum_{i=1}^{N} Z_i \right) M + \left(1 - \sum_{i=1}^{N} Z_i\right) M \\ &= M \end{aligned}$$

Note that for a given year i, the credibility Z_i assigned to each team's experience X_{ij} for that year is the same for all teams. Also note the fact that the grand mean is the same for all years.

That the estimates are in balance can be verified directly for the example given in Table 20. The predicted losing percentages for each year average to .500, subject to rounding.

12.5 Choice of a Prediction Method

The example in this paper is for illustrative purposes only; the purpose of this paper was not to predict baseball teams' win-loss records. Nevertheless, it may be of interest to choose a reasonable prediction method

[55] The schedule was exactly balanced, but a few scheduled games are sometimes not played.

TABLE 20

National League, Predictions of Losing Percentages

	NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
1904	.541	.479	.461	.495	.469	.575	.379	.606
1905	.582	.568	.432	.456	.398	.610	.423	.534
1906	.615	.613	.424	.480	.368	.504	.408	.588
1907	.627	.568	.334	.533	.389	.531	.421	.598
1908	.594	.559	.351	.544	.448	.463	.426	.616
1909	.578	.598	.375	.529	.408	.476	.405	.631
1910	.633	.599	.366	.508	.427	.498	.354	.614
1911	.614	.576	.371	.509	.426	.492	.430	.581
1912	.651	.563	.411	.524	.400	.490	.443	.522
1913	.624	.582	.414	.511	.376	.508	.425	.557
1914	.561	.555	.438	.550	.377	.454	.471	.596
1915	.458	.526	.478	.570	.441	.504	.515	.509
1916	.468	.493	.505	.541	.504	.443	.517	.529
1917	.437	.438	.536	.574	.464	.440	.551	.559
1918	.503	.506	.519	.511	.423	.442	.603	.492
1919	.534	.519	.425	.493	.440	.512	.514	.565
1920	.560	.512	.467	.394	.413	.584	.509	.565
1921	.567	.448	.488	.458	.449	.573	.490	.528
1922	.509	.486	.543	.501	.419	.618	.449	.474
1923	.592	.492	.499	.469	.426	.596	.461	.466
1924	.596	.503	.485	.448	.412	.626	.450	.479
1925	.615	.448	.478	.462	.418	.605	.440	.536
1926	.553	.522	.525	.474	.441	.562	.418	.505
1927	.556	.516	.485	.458	.488	.583	.452	.465
1928	.571	.550	.472	.497	.441	.610	.422	.436
1929	.613	.509	.441	.487	.434	.648	.452	.418
1930	.603	.530	.406	.539	.449	.558	.442	.471
1931	.556	.472	.430	.570	.448	.614	.476	.434
1932	.564	.487	.452	.586	.448	.559	.498	.403
1933	.513	.478	.441	.584	.504	.520	.467	.492

TABLE 20

(CONTINUED)

	NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
1934	.487	.537	.455	.588	.442	.564	.460	.468
1935	.487	.523	.447	.609	.434	.578	.492	.433
1936	.633	.534	.405	.558	.427	.568	.460	.417
1937	.545	.543	.442	.532	.426	.603	.470	.440
1938	.518	.563	.421	.582	.412	.579	.457	.467
1939	.498	.536	.436	.490	.449	.635	.450	.507
1940	.543	.486	.456	.437	.477	.641	.518	.445
1941	.547	.458	.494	.398	.508	.635	.495	.466
1942	.569	.406	.522	.433	.510	.659	.491	.411
1943	.572	.381	.538	.477	.472	.661	.525	.378
1944	.551	.452	.519	.457	.573	.590	.492	.368
1945	.559	.530	.515	.451	.544	.586	.455	.363
1946	.546	.470	.428	.543	.513	.629	.472	.399
1947	.450	.487	.468	.538	.562	.559	.538	.400
1948	.434	.459	.511	.531	.495	.579	.559	.433
1949	.453	.439	.548	.554	.504	.554	.497	.451
1950	.413	.492	.571	.564	.511	.502	.527	.419
1951	.440	.470	.564	.556	.470	.454	.570	.477
1952	.414	.501	.572	.548	.429	.503	.563	.472
1953	.411	.542	.518	.541	.428	.458	.646	.457
1954	.375	.455	.553	.543	.503	.475	.627	.469
1955	.416	.456	.554	.521	.423	.497	.626	.507
1956	.395	.454	.532	.515	.481	.497	.594	.531
1957	.419	.434	.572	.453	.521	.523	.566	.512
1958	.451	.421	.567	.482	.533	.504	.571	.471
1959	.507	.425	.538	.492	.501	.532	.492	.512
1960	.464	.451	.523	.509	.482	.551	.502	.518

Note: Using latest three years of data, with weights of 10%, 10%, 55% (55% weight to the most recent year; 25% weight to grand mean).

for this particular problem. Assume that $\Delta = 1$, i.e., 1910 data are available to predict 1911, etc.

Based on Table 19, the credibilities in Table 16 work well.

The author would recommend avoiding using many years of data unless it substantially improved the accuracy. It is better to keep things simple. For this particular problem, based on Table 19, there seems little advantage to using more than 3 years of data. For example, giving 55% weight to the most recent year, 10% weight to the next most recent year, 10% weight to the third most recent year, and the remaining 25% weight to the grand mean works reasonably well.[56]

The predictions that result from this method of estimation applied to the National League data are shown in Table 20.[57] The errors are shown in Table 21.

The mean squared error is .0046.[58] There is a 14% chance of an error of more than 20%. The correlation used in the Meyers/Dorweiler criterion is .02, not significantly different from zero. Thus according to all three criteria this prediction method works well.

13. CONCLUSIONS

The data from baseball used in this paper provide a useful way to examine and illustrate credibility concepts.

The methods and concepts illustrated here can be applied to problems actuaries deal with in insurance. However, this paper is only a first step; there is further work required to apply these general concepts to any specific practical situation.

---

[56] Many other choices would also work reasonably well. This illustrates the typical situation where once the general form of the weights is determined, there is a range of weights that work well. Usually, the specific choice of weights within that range has relatively little impact on the final result.

[57] For example, the 1904 entry under NL2: .479 = (.10)(.419) + (.10)(.457) + (.55)(.485) + (.25)(.500), where the first three values are from Table 1, and .500 is the grand mean.

[58] The mean squared error is .0049 when the method is applied to both the AL and NL data sets. This is a standard deviation of 10½ losses out of a season of 150 games; the process standard deviation is about 6 losses out of a season of 150 games.

TABLE 21

National League, Errors of Predictions in Table 20

	NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
1904	−.100	−.155	.069	.070	.162	−.083	−.052	.093
1905	−.087	−.116	.033	−.028	.084	.156	.050	−.089
1906	−.060	.047	.187	−.096	.000	−.032	.016	−.065
1907	.019	.007	.038	−.036	−.075	.096	.012	−.062
1908	.003	−.097	−.006	.018	.084	.002	.062	−.066
1909	−.128	−.043	.055	.032	.009	−.040	.129	−.014
1910	−.021	.015	.041	−.005	.018	.008	−.084	.026
1911	−.095	.003	−.032	−.033	.073	.012	−.018	.084
1912	−.009	−.058	.018	.014	.082	−.030	.059	−.066
1913	.081	.018	−.011	−.071	.040	.091	−.052	−.103
1914	.175	.042	−.056	−.060	−.078	−.065	−.081	.125
1915	.004	.052	−.045	.031	−.105	.096	−.011	−.020
1916	.054	.103	−.057	−.067	.070	.038	−.061	−.079
1917	−.092	−.098	.017	.080	.100	.012	−.118	.098
1918	−.070	−.042	.170	.042	−.004	−.111	.123	−.113
1919	−.056	.012	−.039	.179	.061	−.145	.025	−.041
1920	−.032	.116	−.046	−.070	−.029	−.011	.022	.052
1921	.083	−.045	−.094	−.084	.063	−.096	.078	.097
1922	−.145	−.020	.062	.059	.023	−.009	.001	.026
1923	−.057	−.014	.038	.060	.047	−.079	.026	−.018
1924	−.058	.100	.014	−.010	.020	−.010	.038	−.099
1925	.073	−.108	−.080	−.015	−.016	.049	.061	.039
1926	−.013	−.014	.057	.039	−.069	−.054	−.033	.083
1927	−.054	−.059	.041	.052	.085	.086	.062	.066
1928	−.102	.053	.063	.010	.045	−.107	−.019	.053
1929	−.023	−.033	.086	−.084	−.010	.112	.027	−.069
1930	.058	.088	−.010	−.078	.014	−.104	−.039	.068
1931	−.028	−.008	−.025	−.053	.020	.043	−.037	.090
1932	.064	.013	.036	−.024	−.084	.065	.056	−.129
1933	.052	−.097	−.001	−.034	.103	−.085	.032	.028

TABLE 21

(CONTINUED)

	NL1	NL2	NL3	NL4	NL5	NL6	NL7	NL8
1934	.004	.004	.025	−.068	.050	−.060	−.047	.089
1935	−.265	−.019	.096	.053	.029	−.004	.054	.056
1936	.094	−.031	−.030	.039	.024	−.081	.005	−.018
1937	.065	−.052	.046	−.104	.051	.002	.028	−.034
1938	.025	.026	.007	.129	−.035	−.121	.030	−.063
1939	−.085	.085	−.019	.120	−.041	−.067	−.106	.108
1940	−.029	.061	−.057	.091	−.049	−.032	.024	−.006
1941	−.050	.107	−.051	−.031	−.008	−.086	.021	.100
1942	−.032	.081	−.036	−.067	.069	−.063	−.060	.099
1943	.016	−.090	.022	.042	−.169	.077	.044	.060
1944	−.027	−.139	.006	.035	.008	−.011	.080	.050
1945	.000	.095	.151	−.153	.057	−.115	−.013	−.020
1946	.161	−.001	−.036	−.022	−.091	.077	−.119	.027
1947	.060	.045	−.084	.012	.088	−.038	−.059	−.022
1948	−.021	.054	−.073	−.051	.001	.008	.098	−.015
1949	.083	−.074	−.056	−.043	−.022	.080	−.042	.074
1950	−.009	.031	−.011	−.005	.069	.093	−.100	−.071
1951	.058	−.036	−.033	−.002	.094	−.072	−.014	.003
1952	.041	−.081	.072	−.004	.026	.068	−.164	.043
1953	.093	.139	−.060	−.017	−.117	−.003	−.029	−.004
1954	−.028	.033	−.031	.024	.133	−.038	−.029	−.063
1955	.057	.008	.025	.008	−.058	−.003	.016	−.051
1956	−.001	.051	−.078	.106	−.084	−.042	.023	.025
1957	−.036	.051	−.025	−.028	−.031	.023	−.031	.077
1958	−.088	.018	.035	−.024	.052	−.048	.116	−.061
1959	.071	−.024	.019	−.027	.040	−.052	−.002	−.027
1960	−.004	.022	−.087	−.056	−.005	−.066	.119	.076

Note: Predicted Losing Percentage minus Actual Losing Percentage

When shifting parameters over time is an important phenomenon, older years of data should be given substantially less credibility than more recent years of data. The more significant this phenomenon, the more important it is to minimize the delay in receiving the data that is to be used to make the prediction.

In this paper three different criteria were examined that can be used to select the optimal credibility: least squares, limited fluctuation, and Meyers/Dorweiler. In applications, one or more of these three criteria should be useful. While the first two criteria are closely related, the third criterion can give substantially different results than the others.

Generally the mean squared error can be written as a second order polynomial in the credibilities. The coefficients of this polynomial can be written in terms of the covariance structure of the data. This in turn allows one to obtain linear equation(s) which can be solved for the least squares credibilities in terms of the covariance structure.

REFERENCES

[1] G. G. Meyers, "An Analysis of Experience Rating," *PCAS* LXXII, 1985, p. 278.

[2] G. G. Venter, "Experience Rating–Equity and Predictive Accuracy," *NCCI Digest,* Volume II, Issue I, April 1987, p. 27.

[3] W. R. Gillam, "Parameterizing the Workers' Compensation Experience Rating Plan," Casualty Actuarial Society Discussion Paper Program, May 1990, p. 857.

[4] H. C. Mahler, "An Actuarial Analysis of the NCCI Revised Experience Rating Plan," *Casualty Actuarial Society Forum,* Winter 1991, p. 35.

[5] H. C. Mahler, Discussion of [1], *PCAS* LXXIV, 1987, p. 119.

[6] D. S. Neft and R. M. Cohn, *The Sports Encyclopedia: Baseball,* St. Martin's Press, 1987.

[7] G. G. Venter, "Classical Partial Credibility with Application to Trend," *PCAS* LXXIII, 1986, p. 27.

[8] P. Dorweiler, "A Survey of Risk Credibility in Experience Rating," *PCAS* XXI, 1934, p. 1.

[9] H. C. Mahler, "An Actuarial Note on Credibility Parameters," *PCAS* LXXIII, 1986, p. 1.

[10] S. C. Wheelwright and S. Makridakis, *Forecasting Methods for Management,* John Wiley and Sons.

[11] N. Morrison, *Introduction to Sequential Smoothing and Prediction,* McGraw-Hill, 1969.

[12] G. G. Meyers, "Empirical Bayesian Credibility for Workers' Compensation Classification Ratemaking," *PCAS* LXXI, 1984, p. 96.

[13] M. A. Walters, "Homeowners Insurance Ratemaking," *PCAS* LXI, 1974, p. 15.

[14] M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Volume 2, Macmillan, 1979.

[15] F. De Vylder, *Introduction to the Actuarial Theory of Credibility*. English translation by Charles A. Hachmeister, 1980.

[16] H. U. Gerber and D. A. Jones, "Credibility Formulas of the Updating Type," *Credibility: Theory and Applictions*, edited by P. M. Kahn, Academic Press, 1975, pp. 89–105.

APPENDIX A

SOME RELEVANT FEATURES OF BASEBALL

Baseball is a competitive sport involving a combination of luck and skill. Two teams play against each other in a game; the team that scores the most "runs" wins the game, the other team loses.[1]

Each team has nine players in the game at a time.[2] Players may be substituted for, but once they leave the game they cannot return. Over this period of time each team had 20 to 25 players on its roster.[3] The individual skills of the players, as well as how their skills complement each other, has a direct impact on the quality of the team.

In addition to the players, a team has coaches and a field manager. By supervising the players' training and conditioning, providing advice, deciding who plays, and by various decisions throughout the game, these people have some effect on the percentage of games lost or won by the team.

Each team has an owner(s) and other office personnel.[4] By developing new players, trading for players with other teams, etc., management has some effect on the percentage of games lost or won by the team.

All of these elements that affect the quality of the team shift over time. A team's roster of players typically changes a little during the course of a single year; over the course of several years the changes are substantial. It is unusual for a player to be with a single team for more than 10 years, although on very rare occasions a player has played for a single team for 20 years.

Even if the identity of the players were to stay the same, the skill level of individual players changes over time. The most important effect is aging; as a player gets older he generally improves until he reaches a

[1] While it is possible for a baseball game to end in a tie, such games are ignored in major league standings.

[2] Currently the American League has added a tenth player, the designated hitter.

[3] Of the players on the roster, about half get most of the playing time, while the remainder see much less playing time.

[4] During the latter half of this period a team had a general manager.

peak and then declines. Injuries can have a profound impact on a player's skill; sometimes that impact is temporary while sometimes it is permanent.

The field managers and coaching staff also change over time.[5] In addition, the owner(s) and upper management change, but much less frequently.

Finally, a team occasionally relocates to another city.

It might be useful to think of the following analogy to a workers compensation risk. The baseball players correspond to the workers in the factory. The field manager corresponds to the plant manager. The baseball upper management corresponds to the corporation's upper management.

[5] Quite often the departure of the field manager will be related to the poor record of the team.

APPENDIX B

MEYERS/DORWEILER CRITERION AND KENDALL'S TAU

If an experience rating plan works properly, then after the application of experience rating, an insurer should be equally willing to write debit and credit risks. In other words, the modified loss ratio of expected losses to modified premiums should be the same for debit and credit risks.

Mathematically, we desire that the correlation between the experience modification and the modified loss ratio be zero.[1]

In the example in this paper, the experience modification corresponds to the ratio of predicted losing percentage to the grand mean losing percentage.[2] For example, a predicted losing percentage of 60% is equivalent to an experience modification of 60% ÷ 50% = 1.2. The modified loss ratio corresponds to the ratio of the actual losing percentage and the predicted losing percentage.[3] The third criterion used in this paper is that the correlation between these two ratios be zero. This corresponds to the criterion used by Meyers.

Meyers [1] uses the Kendall τ to measure correlation.

Let X and Y be two vectors of length n.[4] Kendall's τ can be calculated as follows [14]. Suppose Y is arranged in its natural order. Assume that the corresponding ranks of X are $X_1, X_2, \ldots X_n$, a permutation of 1, 2, ... n. Let Q be the number of inversions in $X_1, X_2, \ldots X_n$.[5] Then let

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

[1] If the correlation is positive, then insurers would prefer to write credit risks. The credits and debits given are on average too small, i.e., the credibility assigned to the experience is too small. The situation is reversed for a negative correlation.

[2] The predicted losses are equal to the experience modification times the expected losses for an average risk in the class. In a more general situation one would have classifications of risks; in this example we have only one such classification and thus use the grand mean rather than the class mean.

[3] In general the modified loss ratio is equal to the expected loss ratio times the actual losses over the predicted losses. In this example, the expected loss ratio can be thought of as unity.

[4] In our case, X would be the experience modifications and Y would be the corresponding modified loss ratios.

[5] For example, in the X-ranking 3214 for $n = 4$, there are 3 inversions of order.

τ is symmetrically distributed on the range $[-1, +1]$. As is usual for measures of correlation, $+1$ signifies complete agreement and -1 signifies complete disagreement.

As shown in Kendall and Stuart [14],

$$\text{Var } \tau = \frac{2(2n + 5)}{9n(n - 1)}$$

As n approaches infinity the distribution of τ approaches the normal distribution.

In the examples in this paper, the variance of τ varies from .0009 to .0016.[6] The standard deviation of τ goes from .031 to .040. Thus an approximate 95% confidence interval around zero for τ has a radius of approximately .07, about two standard deviations.

[6] $n = 8$ teams times 59 years = 472 to $n = 8$ times 35 = 280.

APPENDIX C

MATRIX EQUATIONS FOR LEAST SQUARES CREDIBILITY

In this appendix, equations 11.2, 11.3, 11.4, and 11.7 in the main text are derived. The squared error is written as a second order polynomial in the credibilities, with the coefficients depending on the covariance structure discussed in Appendix D. This squared error is minimized by setting the partial derivative(s) with respect to the credibilities equal to zero.

Assume an estimate for year $N + \Delta$, using N years of data, is given by:

$$F = \sum_{i=1}^{N} Z_i X_i + (1 - \Sigma Z_i)M$$

where X_i is the data for year i, and M is the grand mean.[1] Let $Z_0 = 1 - \Sigma Z_i$. Write Z for the vector $Z_0, Z_1, \ldots, Z_N$.

Then the mean squared error between the prediction and the observation is given by the expected value of the squared difference between F and $X_{N+\Delta}$.

$$\begin{aligned} V(Z) &= \mathrm{E}[(F - X_{N+\Delta})^2] \\ &= \mathrm{E}\left[\left\{\sum_{i=1}^{N} Z_i \left(X_i - X_{N+\Delta}\right) + Z_0 \left(M - X_{N+\Delta}\right)\right\}^2\right] \\ &= \sum_{i=1}^{N} Z_i^2 \mathrm{E}[(X_i - X_{N+\Delta})^2] \\ &\quad + \sum_{i=1}^{N} \sum_{j \neq i} Z_i Z_j \, \mathrm{E}[(X_i - X_{N+\Delta})(X_j - X_{N+\Delta})] \\ &\quad + 2 \sum_{i=1}^{N} Z_0 Z_i \, \mathrm{E}[(X_i - X_{N+\Delta})(M - X_{N+\Delta})] \\ &\quad + Z_0^2 \, \mathrm{E}[(M - X_{N+\Delta})^2] \end{aligned}$$

[1] It is assumed that the grand mean is known. This is the case in this paper. It is the case whenever one is only concerned with relativities compared to the overall average.

From Appendix D we have,[2]

$$\mathrm{E}[(X_i - X_{N+\Delta})^2] = 2\delta^2 + 2\zeta^2(1 - \ell(N + \Delta - i))$$

$$\mathrm{E}[(X_i - X_{N+\Delta})(X_j - X_{N+\Delta})] = \delta^2 + \zeta^2(1 + \ell(|i - j|) - \ell(N + \Delta - i) - \ell(N + \Delta - j))$$

$$\mathrm{E}[(X_i - X_{N+\Delta})(M - X_{N+\Delta})] = \delta^2 + \zeta^2(1 - \ell(N + \Delta - i))$$

$$\mathrm{E}[(M - X_{N+\Delta})^2] = \delta^2 + \zeta^2 + \tau^2$$

Therefore

$$V(Z) = \sum_{i=1}^{N} Z_i^2(2\delta^2 + 2\zeta^2(1 - \ell(N + \Delta - i))) + \sum_{j=1}^{N} \sum_{j \neq i} Z_i Z_j \left[\delta^2 + \zeta^2(1 + \ell(|i - j|) - \ell(N + \Delta - i) - \ell(N + \Delta - j)\right] + 2Z_0 \sum_{i=1}^{N} Z_i(\delta^2 + \zeta^2(1 - \ell(N + \Delta - i)) + Z_0^2(\delta^2 + \tau^2 + \zeta^2)$$

$$V(Z) = \delta^2 \left[\sum_{i=0}^{N} \sum_{j=0}^{N} Z_i Z_j + \sum_{i=1}^{N} Z_i^2 \right] + Z_0^2 \tau^2 + \zeta^2 \left[\sum_{i=0}^{N} \sum_{j=0}^{N} Z_i Z_j + \sum_{i=1}^{N} \sum_{j=1}^{N} Z_i Z_j(\ell(|i - j|) - \ell(N + \Delta - i) - \ell(N + \Delta - j)) - 2Z_0 \sum_{i=1}^{N} Z_i \ell(N + \Delta - i) \right]$$

but

$$\sum_{i=0}^{N} Z_i = Z_0 + \sum_{i=1}^{N} Z_i = (1 - \sum_{i=1}^{N} Z_i) + \sum_{i=1}^{N} Z_i = 1$$

[2] In Appendix D, $X(\theta,t)$ = the observation for risk θ at time t. Since in this appendix none of the calculations are performed for individual risks, the θ has been suppressed in order to simplify the notation.

Therefore

$$V(Z) = \delta^2 + Z_0^2\tau^2 + \zeta^2 + \delta^2 \sum_{i=1}^{N} Z_i^2$$

$$+ \zeta^2 \sum_{i=1}^{N} \sum_{j=1}^{N} Z_i Z_j (\ell(|i-j|) - \ell(N+\Delta-i) - \ell(N+\Delta-j))$$

$$- \zeta^2 Z_0 2 \sum_{i=1}^{N} Z_i \ell(N+\Delta-i)$$

$$V(Z) = \delta^2 + \zeta^2 + \tau^2 + \tau^2 \left[\sum_{i=1}^{N} Z_i \right]^2 - 2\tau^2 \sum_{i=1}^{N} Z_i$$

$$+ \delta^2 \sum_{i=1}^{N} Z_i^2 + \zeta^2 \sum_{i=1}^{N} \sum_{j=i}^{N} Z_i Z_j (\ell(|i-j|) - \ell(N+\Delta-i)$$

$$- \ell(N+\Delta-j)) - 2\zeta^2 \sum_{i=1}^{N} Z_i \ell(N+\Delta-i)$$

$$+ 2\zeta^2 \sum_{i=1}^{N} \sum_{j=1}^{N} Z_i Z_j \ell(N+\Delta-i)$$

$$V(Z) = \delta^2 + \zeta^2 + \tau^2 + \tau^2 \sum_{i=1}^{N} \sum_{j=1}^{N} Z_i Z_j - 2\tau^2 \sum_{i=1}^{N} Z_i$$

$$+ \delta^2 \sum_{i=1}^{N} Z_i^2 + \zeta^2 \sum_{i=1}^{N} \sum_{j=i}^{N} Z_i Z_j (\ell(|i-j|) - \ell(N+\Delta-j))$$

$$- 2\zeta^2 \sum_{i=1}^{N} Z_i \ell(N+\Delta-i)$$

$$V(Z) = \sum_{i=1}^{N} \sum_{j=1}^{N} Z_i Z_j (\delta^2 \delta_{ij} + \tau^2 + \zeta^2 \ell(|i-j|))$$

$$- 2 \sum_{i=1}^{N} Z_i (\tau^2 + \zeta^2 \ell(N+\Delta-i)) + \delta^2 + \zeta^2 + \tau^2$$

This is equation 11.2 in the main text, with $\delta^2\delta_{ij} + \zeta^2\ell(|i-j|) = C(|i-j|)$ the covariance between data for a given risk $|i-j|$ years apart. It is left as an exercise to the reader to verify that the formula for the mean squared error compared to the underlying mean rather than the observed value would be exactly δ^2 less.

In order to minimize this squared error, one sets the partial derivatives with respect to Z_i equal to zero. This yields the following set of N equations.

$$2Z_i(\delta^2 + \tau^2 + \zeta^2) + \sum_{j \neq i} 2Z_j(\tau^2 + \zeta^2\ell(|i-j|)) - 2(\tau^2 + \zeta^2\ell(N + \Delta - i)) = 0, \quad i = 1, \ldots, N$$

$$\sum_{j=1}^{N} Z_j(\delta^2\delta_{ij} + \tau^2 + \zeta^2\ell(|i-j|)) = \tau^2 + \zeta^2\ell(N + \Delta - i), \quad i = 1, \ldots, N$$

This is equation 11.3 in the main text, again with

$$\delta^2\delta_{ij} + \zeta^2\ell(|i-j|) = C(|i-j|).$$

It is worth noting that equation 11.3 is very similar to the usual general matrix equation for optimal least squares credibilities:

$$\vec{Z} = \frac{\text{COV}[\vec{X}, Y]}{\text{COV}[\vec{X}, \vec{X}]}$$

where $\vec{X}$ is the vector of observations, and Y is the quantity to be estimated.[3] Here in equation 11.3, there is an additional term of τ^2, the between variance, added to the covariances. This is due to the application of the complement of credibility to the grand mean.

In the absence of shifting parameters over time ($\zeta^2 = 0$), the squared error is given by:

$$V(Z) = \delta^2 \left(1 + \sum_{i=1}^{N} Z_i^2\right) + \tau^2 \left(1 - \sum_{i=1}^{N} Z_i\right)^2$$

[3] See, for example, Theorem 3.3 in Chapter III of De Vylder [15].

The optimal credibilities are given by the solution to the equations:

$$\sum_{j=1}^{N} Z_j(\delta^2\delta_{ij} + \tau^2) = \tau^2, \quad i = 1, \ldots, N$$

The solution has all the credibilities equal:

$$Z_i = \frac{\tau^2}{N\tau^2 + \delta^2}, \quad i = 1, \ldots, N$$

$$\sum_{i=1}^{N} Z_i = \frac{N\tau^2}{N\tau^2 + \delta^2} = \frac{N}{N + \delta^2/\tau^2}$$

This is the familiar expression for the least squares credibility in the absence of shifting parameters over time.

If we set $Z_i = Z/N$ for $i = 1, \ldots, N$ then equation 11.2 becomes:

$$V(Z) = \frac{Z^2}{N^2}\left\{N\delta^2 + N^2\tau^2 + \zeta^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \ell(|i - j|)\right\}$$

$$- 2\frac{Z}{N}\left\{N\tau^2 + \zeta^2 \sum_{i=1}^{N} \ell(N + \Delta - i)\right\} + \delta^2 + \zeta^2 + \tau^2$$

Setting the derivative of $V(Z)$ equal to zero gives the least squares credibility:

$$Z = N \frac{N\tau^2 + \zeta^2 \sum_{i=1}^{N} \ell(N + \Delta - i)}{N^2\tau^2 + N\delta^2 + \zeta^2 \sum_{i=1}^{N}\sum_{j=1}^{N} \ell(|i - j|)}$$

This is equation 11.4 in the main text, with $C(|i - j|) = \delta^2\delta_{ij} + \zeta^2\ell(|i - j|)$.

We can minimize $V(Z)$ in equation 11.2, given the constraint $\Sigma_{i=1}^{N} Z_i = 1$, by using Lagrange Multipliers.

We set the partial derivatives with respect to Z_i of

$V(Z) - \lambda\left(\sum_{i=1}^{N} Z_i - 1\right)$ equal to zero.

This produces the following N equations:

$$\sum_{j=1}^{N} Z_j(\delta^2\delta_{ij} + \zeta^2\ell(|i-j|)) = \zeta^2\ell(N + \Delta - 1) + \frac{\lambda}{2} \quad i = 1, \ldots, N$$

This is equation 11.7 in the main text. It is worth noting the absence from the above equation of τ^2, the between variance. This follows logically from the fact that the grand mean is given no weight and each risk is estimated solely from its own data.

APPENDIX D

COVARIANCE STRUCTURE

In this appendix, the covariance structure for the data sets in Tables 1 and 2 will be analyzed. As discussed in Section 11.1, the variance is the sum of three pieces, the between variance, the variance due to shifting parameters over time, and the process variance excluding the effect of shifting parameters over time. The analysis herein will define these three pieces.

Let $X(\theta,t)$ be the observation for risk θ at time t.

Let $\mu(\theta,t)$ be the expected value for risk θ at time t.

$\mu(\theta,t) = E[X(\theta,t)]$.

Let $\mu(\theta) = E_t[X(\theta,t)]$.

Let M be the grand mean.

$M = E[\mu(\theta,t)] = E_\theta[\mu(\theta)]$.

In our case, θ and t are both discrete rather than continuous variables. We can observe X. M is known since we are dealing with relativities compared to the overall average. On the other hand $\mu(\theta,t)$ is unknown and can never be observed directly.

We can observe the squared error that results from using different estimations. This squared error can be usefully expressed in another form. To do so, we split the variance of X into various pieces. Define

$$\delta^2 = E_\theta[E_t[E[(X(\theta,t) - \mu(\theta,t))^2|\theta,t]]]$$

$$\zeta^2 = E_\theta[E_t[(\mu(\theta,t) - \mu(\theta))^2|\theta]]$$

$$\zeta^2\ell(s) = E_\theta[E_t[COV[X(\theta,t),X(\theta,t+s)]|\theta]]$$

$$= E_\theta[E_t[COV[\mu(\theta,t),\mu(\theta,t+s)]|\theta]]$$

$$\tau^2 = VAR_\theta[E_t[\mu(\theta,t)]] = VAR_\theta[\mu(\theta)]$$

Then δ^2 is the process variance excluding any impact of shifting risk parameters over time. ζ^2 is the variance due to shifting parameters over time. $\ell(s)$ is a correlation measuring how much the risk parameters shift over time. $\ell(0) = 1$. $\ell(s) \leq 1$ for $s > 0$.[1] τ^2 is the parameter variance, the variance between the different risks.

For later convenience of notation define

$$\delta^2(\theta,t) = E[(X(\theta,t) - \mu(\theta,t))^2|\theta,t]$$
$$\delta^2(\theta) = E_t[\delta^2(\theta,t)]$$
$$\zeta^2(\theta) = E_t[(\mu(\theta,t) - \mu(\theta))^2|\theta]$$
$$\ell(s,\theta) = E_t[COV[\mu(\theta,t), \mu(\theta,t+s)]] + \zeta^2(\theta)$$

then

$$\delta^2 = E_\theta[\delta^2(\theta)] = E_{\theta,t}[\delta^2(\theta,t)]$$
$$\zeta^2 = E_\theta[\zeta^2(\theta)]$$
$$\ell(s)\zeta^2 = E_\theta[\ell(s,\theta)\zeta^2(\theta)]$$

It is useful to rearrange the definitions of the variances in the usual manner so as to express the expected value of a quantity squared as the sum of a squared mean and a variance.

$$E[X^2(t,\theta)|t,\theta] = \mu^2(t,\theta) + \delta^2(\theta,t)$$
$$E_t[\mu^2(t,\theta)] = \mu^2(\theta) + \zeta^2(\theta)$$
$$E_\theta[\mu^2(\theta)] = M^2 + \tau^2$$

A similar expression can be derived from the definition of the covariance.

$$E_t[\mu(t,\theta)\mu(t+s,\theta)] = \mu^2(\theta) + \ell(s,\theta)\zeta^2(\theta)$$

For the formula for the expected value of the squared error of the estimate from the observation, one needs to express various expected values in terms of the variances and correlations defined above.

[1] One should note that it is an assumption that this correlation depends only upon the separation of the two years in question. Whether or not this is a reasonable approximation to reality is an empirical question which depends on the particular application.

$$\begin{aligned}
E_{t,\theta}[X^2(t,\theta)] &= E_\theta[E_t[E[X^2(t,\theta)|t,\theta]]] \\
&= E_\theta[E_t[\mu^2(t,\theta) + \delta^2(\theta,t)]] \\
&= E_\theta[\mu^2(\theta) + \zeta^2(\theta) + \delta^2(\theta)] \\
&= M^2 + \tau^2 + \zeta^2 + \delta^2
\end{aligned}$$

$$\begin{aligned}
E_{t,\theta}[X(t,\theta)X(t+s,\theta)] &= E_\theta[E_t[E[X(t,\theta)X(t+s,\theta)|t,\theta]]] \\
&= E_\theta[E_t[\mu(t,\theta)\mu(t+s,\theta)]] \\
&= E_\theta[\mu^2(\theta) + \ell(s,\theta)\zeta^2(\theta)] \\
&= M^2 + \tau^2 + \ell(s)\zeta^2
\end{aligned}$$

$$E_{t,\theta}[MX(t,\theta)] = ME[X(t,\theta)] = M^2$$

Then it follows that:

$$\begin{aligned}
E_{t,\theta}[(X(t,\theta) - X(t+s,\theta))^2] &= E_{t,\theta}[X^2(t,\theta)] + E_{t,\theta}[X^2(t+s,\theta)] \\
&\quad - 2E_{t,\theta}[X(t,\theta)X(t+s,\theta)] \\
&= M^2 + \tau^2 + \zeta^2 + \delta^2 + M^2 + \tau^2 \\
&\quad + \zeta^2 + \delta^2 - 2(M^2 + \tau^2 + \ell(s)\zeta^2) \\
&= 2\delta^2 + 2\zeta^2(1 - \ell(s))
\end{aligned}$$

$$\begin{aligned}
&E_{t,\theta}[(X(t,\theta) - X(t+s,\theta))(X(t+u,\theta) - X(t+s,\theta))] \\
&\quad = E_{t,\theta}[X(t,\theta)X(t+u,\theta)] + E_{t,\theta}[X^2(t+s,\theta)] \\
&\qquad - E_{t,\theta}[X(t+s,\theta)X(t+u,\theta)] - E_{t,\theta}[X(t,\theta)X(t+s,\theta)] \\
&\quad = M^2 + \tau^2 + \ell(u)\zeta^2 + M^2 + \tau^2 + \zeta^2 + \delta^2 - (M^2 + \tau^2 \\
&\qquad + \ell(s-u)\zeta^2) - (M^2 + \tau^2 + \ell(s)\zeta^2) \\
&\quad = \delta^2 + \zeta^2(1 + \ell(u) - \ell(s-u) - \ell(s))
\end{aligned}$$

$$\begin{aligned}
&E_{t,\theta}[(X(t,\theta) - X(t+s,\theta))(M - X(t+s,\theta))] \\
&\quad = M^2 - M^2 + (M^2 + \tau^2 + \zeta^2 + \delta^2) \\
&\qquad - (M^2 + \tau^2 + \ell(s)\zeta^2) \\
&\quad = \delta^2 + \zeta^2(1 - \ell(s))
\end{aligned}$$

$$\begin{aligned} E_{t,\theta}[(M - X(t,\theta))^2] &= M^2 - 2M^2 + (M^2 + \tau^2 + \zeta^2 + \delta^2) \\ &= \delta^2 + \zeta^2 + \tau^2 \end{aligned}$$

These results are used in Appendix C.

It is of interest to note that variance of $X = E_{t,\theta}[(M - X(t,\theta))^2] = \delta^2 + \zeta^2 + \tau^2$. This is the split of the variance of X into three pieces that was discussed above.

Let $C(s)$ = Covariance for data for the same risk, Δ years apart. Then for $s > 0$

$$C(s) = E[(X(t,\theta) - \mu(\theta))(X(t+s,\theta) - \mu(\theta))]$$

$$\begin{aligned} C(s) &= E[X(t,\theta)X(t+s,\theta)] - E[X(t,\theta)\mu(\theta)] - E[X(t+s)\mu(\theta)] \\ &\quad + E[\mu^2(\theta)] \\ &= M^2 + \tau^2 + \ell(s)\zeta^2 - (M^2 + \tau^2) - (M^2 + \tau^2) + M^2 + \tau^2 \\ &= \ell(s)\zeta^2 \end{aligned}$$

$$\begin{aligned} C(0) &= E[(X(t,\theta) - \mu(\theta))^2] = E(X^2(t,\theta)] - 2E[X(t,\theta)\mu(\theta)] \\ &\quad + E[\mu^2(\theta)] \\ &= M^2 + \tau^2 + \zeta^2 + \delta^2 - 2(M^2 + \tau^2) + M^2 + \tau^2 \\ &= \zeta^2 + \delta^2 \end{aligned}$$

It is worth noting that the covariance structure assumed herein differs from that in Gerber and Jones [16]. The covariance structure which in Gerber and Jones is shown to give credibility formulas of the updating variety[2] can be written as:

$$COV[X_i,X_j] = \begin{cases} W_i & i < j \\ W_i + V_i & i = j \end{cases}$$

That covariance structure would assume for example that the covariance of the 1940 data with the data for each of the years earlier than

[2] Credibilities of the updating variety are such that new estimate = (prior estimate × complement of credibility) + (new data × credibility). This is the form of the estimate discussed in Section 9.1.

1940 is the same. In fact we observe that the distance between the years has an extremely significant impact on the covariance between the years.

The covariance structure assumed here can be written as:

$$\mathrm{COV}[X_i,X_j] = \begin{cases} \ell(j-i)\zeta^2 & i < j \\ \zeta^2 + \delta^2 & i = j \end{cases}$$

Thus the optimal least squares credibilities that result from the matrix equations that are given in Appendix C will generally not be of the updating variety.[3]

We can directly estimate only the following quantities from the data: τ^2, C(0), C(1), C(2), etc. Not coincidentally, these are the quantities that enter into the formula in Appendix C for the squared error. Thus, these are also the quantities that enter into the calculation of the optimal credibilities.

Thus, it is not necessary to estimate δ^2 by itself. However, if one does so, the values for ζ^2 and $\ell(i)$ follow. We will estimate δ^2 here solely in order to aid our understanding; it does not affect any of the calculated values of the credibilities.[4]

For a binomial process, with a success rate of .4 or .6, the variance is $.24n$.[5] This is approximately the variance for the average risk in this example, with $n = 150$.[6] The resulting variance of games lost is (150)(.24). The variance in losing percentage is $(150)(.24)/(150)^2 = .0016$.

Thus a reasonable approximate value for δ^2 is .0016. The values for the variances and correlations are shown in Table D1. It should be noted that as the difference in years increases, the correlations get close to zero.

For example, the observed value for the NL data for $\delta^2 + \zeta^2 = .007892$. Thus since we assume $\delta^2 = .001600$, we estimate $\zeta^2 = .006292$. The observed value of $\zeta^2\ell(1) = .004919$. Thus we estimate $\ell(1) = .004919/.006292 = .782$. For this example, the observed value of $\tau^2 = .001230$.

[3] They will be of the updating variety when $\ell(s) = 1$ for all s.

[4] In general if something cannot be observed in the squared errors, then it is not needed to calculate the optimal least squares credibilities.

[5] The variance is $p(1-p)n$.

[6] Teams played about 150 games per year over this period of time.

It is important to note that the total variance of the observations is equal to $\delta^2 + \zeta^2 + \tau^2 = .009122$. Thus, what has been done here is just an analysis of variance, breaking the variance into its various sources. For this example, about 13.5% of the variance of the observation is due to the differences between the risks, about 17.5% is due to the process variance, and about 69.0% is due to shifting parameters over time.

One can verify that the observed pattern in the covariance structure in Table D1 is not due solely to random chance. One can rearrange the data in random fashion, and observe the covariances.

TABLE D1

COVARIANCE STRUCTURE

	NL	AL
τ^2	.001230	.001619
δ^2*	.001600	.001600
ζ^2**	.006292	.006275
$\ell(0)$***	1.000	1.000
$\ell(1)$	.782	.721
$\ell(2)$	.543	.506
$\ell(3)$	.497	.384
$\ell(4)$	.404	.283
$\ell(5)$	.288	.124
$\ell(6)$	.249	.061
$\ell(7)$	.158	−.016
$\ell(8)$	.062	−.089
$\ell(9)$	−.012	−.170
$\ell(10)$	−.063	−.140

* δ^2 estimated as .001600 based on an assumed binomial process.

** ζ^2 is based on the assumed value of δ^2 and the observed value of $\zeta^2 + \delta^2$.

*** $\ell(0)$ is unity by definition.

First one can rearrange the entries in each row of Table 1; for each row separately, assign each entry in that row to a randomly selected column. Similarly one can rearrange the entries in each column of Table 1; for each column separately, assign each entry in that column to a randomly selected row. The resulting covariances that are computed for these two "scrambled" data sets are shown in Table D2. All of the covariances $\ell(i)$, $i > 0$ are close to zero. Therefore, one can conclude that there is a significant pattern being displayed in Table D1.

TABLE D2

COVARIANCE STRUCTURE, SCRAMBLED DATA

	NL Entries in Each Row Rearranged	NL Entries in Each Column Rearranged
τ^2	.000191	.001230
δ^2*	.001600	.001600
ζ^2**	.007330	.006292
$\ell(0)$***	1.000	1.000
$\ell(1)$	.010	−.117
$\ell(2)$	−.009	.021
$\ell(3)$	.008	−.070
$\ell(4)$	−.084	−.035
$\ell(5)$	−.025	−.039
$\ell(6)$	−.020	−.006
$\ell(7)$	−.030	−.053
$\ell(8)$	−.058	.082
$\ell(9)$	.049	.091
$\ell(10)$	.042	−.019

* δ^2 estimated at .001600 based on an assumed binomial process.
** ζ^2 is based on the assumed value of δ^2 and the observed value of $\zeta^2 + \delta^2$.
*** $\ell(0)$ is unity by definition.

APPENDIX E

PUTTING THE REDUCTION IN SQUARED ERROR IN CONTEXT

The first criterion used to determine the optimal credibility is to minimize the squared error. Using the optimal credibility based on this criterion will reduce the squared error between the observed and predicted result. What should be considered a significant reduction in squared error?

Let us examine an example. For the NL data set, using one year of data, the optimal credibility is 68% as shown in Table 9. As shown in Table 6 the mean squared errors are:

Z	Mean Squared Error
0	.0091
68%	.0049
100%	.0059

In this case, by the use of credibility, the squared error has been reduced from .0059 if the data were relied upon totally, or .0091 if the data were totally ignored, to .0049. In this case, the squared error has been reduced to 83% (.0049/.0059) of its previous value.[1]

All of these squared errors include the variation of the observed results around the expected value.[2] The use of credibility does not affect this source of variation. Thus credibility methods cannot reduce the squared error between the observed value and the estimated/predicted value to as great an extent as they reduce the squared error between the true mean and the estimated/predicted mean.[3]

It is shown in Mahler [9] that the best that can be done using credibility to combine two estimates is to halve the mean squared error between the estimated and theoretical true underlying mean. However,

[1] The "previous" value of the squared error is considered to be the minimum of the squared errors that result from either ignoring the data entirely or relying on the data entirely.

[2] This random variation is usually referred to as process risk.

[3] It should be noted that the former squared error is concrete and easily observed, while the latter squared error is theoretical and difficult if not impossible to observe.

in this paper the squared error being examined is between the estimated/predicted and the observed result, rather than the true underlying mean. This squared error is inherently larger due to the random variation in the observed result. Also the result derived in Mahler [9] was derived in the absence of shifting parameters over time.

It turns out that, in the current case, the best that can be done using credibility to combine two estimates is to reduce the mean squared error between the estimated and observed values to 75% of the minimum of the squared errors from either relying solely on the data or ignoring the data.[4] One can think of half[5] of the squared error as being due to two sources: the inherent process variance associated with comparing to observed results, and the presence of shifting parameters over time. This portion of the squared error is independent of the value chosen for the credibility. The remainder of the squared error can be thought of as that which is affected by the choice of the value of credibility; as stated above this can be at most cut in half by the use of credibility methods. If half of the squared error is cut in half, this reduces the total squared error to 75% of what it was.

Assume one is estimating the future by credibility weighting together a single year of data and the grand mean.[6] Let $V(0)$ be the squared error between the predicted and observed results for $Z = 0$. Let $V(1)$ be the squared error between the predicted and observed results for $Z = 1$. Then as is shown in Appendix F:

Z	Squared Error Between Predicted and Observed
0	$V(0)$
Optimal	$V(1)\left(1 - \frac{V(1)}{4V(0)}\right)$
100%	$V(1)$

with the optimal credibility given by: Z optimal $= 1 - V(1)/2V(0)$.

[4] When using more than two or more years of data, the reduction in squared error depends on the impact of shifting parameters over time. However, in the absence of shifting parameters over time, for N years with the same weight applied to each year, the maximum possible reduction is $1/(2(N + 1))$.

[5] This is only a half for the case when the squared errors for $Z = 0$ and $Z = 1$ are equal. However, this is the case when one gets the maximum reduction in squared error.

[6] The formula given below does not hold when using several years of data.

In the example above, we had $V(0) = .0091$, $V(1) = .0059$. Using these values in the above formula gives Z optimal = 68%, equal to the empirically determined 68%. The formula for the minimum squared error gives a value of .0049, which is equal to the empirical minimum squared error. The reduction of the squared error to 83% of its previous value appears significant in light of the maximum possible reduction to 75%.[7]

[7] The maximum reduction is possible when the squared errors for $Z = 0$ and $Z = 1$ are equal.

APPENDIX F

SQUARED ERRORS

In Appendix C, the fundamental formula for the squared error was derived:

$$V(Z) = \sum_{i=1}^{N} \sum_{j=1}^{N} Z_i Z_j (\delta^2 \delta_{ij} + \tau^2 + \zeta^2 \ell(|i - j|)) - 2 \sum_{i=1}^{N} Z_i (\tau^2 + \zeta^2 \ell(N + \Delta - i)) + \delta^2 + \zeta^2 + \tau^2.$$

One can actually check this result against the observed squared errors.[1] For example, let $N = 2$ and $\Delta = 3$. Then

$$\begin{aligned} V(Z_1, Z_2) = {} & Z_1^2(\delta^2 + \tau^2 + \zeta^2) + 2Z_1Z_2(\tau^2 + \zeta^2\ell(1)) \\ & + Z_2^2(\delta^2 + \tau^2 + \zeta^2) - 2Z_1(\tau^2 + \zeta^2\ell(4)) \\ & - 2Z_2(\tau^2 + \zeta^2\ell(3)) + \delta^2 + \zeta^2 + \tau^2 \end{aligned}$$

Using the average of the NL and AL values in Table D1 for the covariance structure:

$$\tau^2 = .001425 \qquad \delta^2 + \zeta^2 = .007884$$

$$\zeta^2\ell(1) = .004723 \qquad \zeta^2\ell(3) = .002770 \qquad \zeta^2\ell(4) = .002158$$

$$\begin{aligned} V(Z_1, Z_2) = {} & Z_1^2(.009309) + Z_1Z_2(.012296) + Z_2^2(.009309) \\ & - Z_1(.007166) - Z_2(.008390) + .009309 \end{aligned}$$

Table F1 contains the results of the test for various values of Z_1 and Z_2. (Z_1 is the credibility applied to the less recent year of the two.) The mean squared errors are a close match to those given by the equation.[2]

---

[1] The covariances were estimated from the same data as is being used to test the equation for the squared error. Thus, the magnitude of the covariances is not being tested. However, the validity of the assumed form of the covariance structure as well as the validity of the derivation of the equation for $V(Z)$ are being tested.

[2] The differences are largely due to the fact that at the two ends of the data period there are either no predictions or no actual observation to enter into the computation of an error.

When $N = 1$, one gets the following parabola for $V(Z)$:

$$V(Z) = Z^2(\delta^2 + \tau^2 + \zeta^2) - 2Z(\tau^2 + \zeta^2\ell(\Delta)) + \delta^2 + \zeta^2 + \tau^2$$

$V(0) = \delta^2 + \tau^2 + \zeta^2$ = squared error ignoring the data

$V(1) = 2\delta^2 + 2\zeta^2(1 - \ell(\Delta))$ = squared error relying solely on the data

$$Z \text{ optimal} = \frac{\tau^2 + \zeta^2\ell(\Delta)}{\tau^2 + \delta^2 + \zeta^2} = \frac{V(0) - V(1)/2}{V(0)} = 1 - \frac{V(1)}{2V(0)}$$

$$V(Z \text{ optimal}) = -\frac{(\tau^2 + \zeta^2\ell(\Delta))^2}{\tau^2 + \delta^2 + \zeta^2} + \delta^2 + \zeta^2 + \tau^2$$

$$= -\frac{(V(0) - V(1)2)^2}{V(0)} + V(0)$$

$$= -V(0) + V(1) - \frac{V(1)^2}{4V(0)} + V(0)$$

$$= V(1)\left(1 - \frac{V(1)}{4V(0)}\right)$$

This is the result referred to in Appendix E. The reduction in mean squared error is greatest when $V(1) = V(0)$; then the squared error is reduced to 75% of the minimum of the squared errors that result from relying solely on the data or ignoring the data.

In the absence of shifting parameters over time,[3] the estimate improves as one uses more and more years of data. For large N, relying solely on the data produces a very good estimate; this is reflected in the fact that the optimal credibility approaches 1 as N gets large. Thus for large N, one cannot reduce the squared error significantly by using credibility.

[3] In the presence of shifting parameters over time the situation is much more complicated.

TABLE F1

MEAN SQUARED ERRORS (.0001)

Z_1	Z_2	Observed	Estimated by 2nd Order Polynomial
0	0	9,182	9,309
0	.25	7,592	7,793
.25	0	7,963	8,099
0	.5	7,202	7,441
.15	.35	7,087	7,293
.25	.25	7,172	7,352
.5	0	7,949	8,053
0	.75	8,011	8,253
.25	.5	7,581	7,769
.5	.25	7,957	8,057
.75	0	9,140	9,171
0	1	10,020	10,228
.25	.75	9,189	9,349
.5	.5	9,165	9,260
.75	.25	9,947	9,961
1	0	11,536	11,452
.75	.75	15,162	15,031
1	1	25,162	24,667

Note: Mean Squared Errors in estimating NL and AL data. $N = 2$, $\Delta = 3$. Estimate uses data from the fourth and third years prior to the estimation period with weights Z_1 and Z_2, respectively, and the complement of credibility applied to the grand mean. $Z_1 = 15\%$ and $Z_2 = 35\%$ is the solution to equation 11.3 for the least squares credibility.

The exact behavior can be derived using the results of Appendix C. In the absence of shifting parameters over time ($\zeta^2 = 0$), and applying equal weight Z/N to each of N years, based on the result in Appendix C, the squared error is given by:

$$V(Z) = Z^2\left(\frac{\delta^2}{N} + \tau^2\right) - 2Z\tau^2 + \delta^2 + \tau^2$$

$$V(0) = \delta^2 + \tau^2$$

$$V(1) = \delta^2\left(\frac{N+1}{N}\right)$$

$$Z \text{ optimal} = \frac{N\tau^2}{N\tau^2 + \delta^2} = \frac{(N+1)V(0) - NV(1)}{(N+1)V(0) - (N-1)V(1)}$$

$$V(Z \text{ optimal}) = \delta^2 + \tau^2 - \frac{\tau^4 N}{N\tau^2 + \delta^2}$$

$$= V(1)\left(1 - \frac{V(1)}{(N+1)^2V(0) - (N^2-1)V(1)}\right)$$

The maximum reduction in squared error compared to the minimum of $V(0)$ and $V(1)$ occurs when $V(0) = V(1)$. For this case

$$Z \text{ optimal} = 1/2$$

$$V(Z \text{ optimal}) = V(1)\left(1 - \frac{1}{2(N+1)}\right)$$

As N gets large, there is no significant reduction in squared error due to using credibility (in the absence of shifting parameters over time).

APPENDIX G

THE SECOND CRITERION AND LIMITED FLUCTUATION CREDIBILITY

The second criterion in Section 7 deals with the probability that the observed result will be more than a certain percent different than the predicted result. The less this probability, the better the solution.

This is related to the basic concept behind "classical" credibility which has also been called "limited fluctuation" credibility [7]. In classical credibilty, the full credibility criterion is chosen so that there is a probability, P, of meeting the test, that the maximum departure from expected is no more than k percent.

The reason the criterion is stated in this way rather than the way it is in classical credibility is that, unlike the actual observations, one cannot observe directly the inherent loss potential.[1]

However, the two concepts are closely related. If there is a small chance of the estimate differing by a large amount from the true value of the inherent loss potential, then, since the observed values are distributed about the true value, the chance of the estimate differing by a large amount from the observed value will be smaller than it would otherwise be.

For example, assume the inherent loss potential is .550 and that the observed values are distributed approximately normally with a standard deviation of .050. Then there is approximately a 95% probability that the observed value will be between .452 and .648.[2]

Assume the estimated values are also approximately normally distributed about the inherent loss potential.[3] Assume a standard deviation of .028. Then there is a 95% chance that the estimate will be between .495 and .605, i.e., within 10% of the true inherent loss potential.

---

[1] It has been shown that the loss potential varies for a risk over time. Thus, it cannot be estimated as the average of many observations over time.

[2] The mean plus or minus 1.96 standard deviations.

[3] An unbiased estimator has the same expected value as the inherent loss potential.

The difference between the estimated value and the observed value will also be approximately normally distributed about zero.[4] The standard deviation is .057.[5] Thus, there would be a 95% chance that the absolute difference between the estimated and observed value will be less than .112. This corresponds to about a 95% chance that the estimated value will be within ±20% of the observed value.[6]

In a particular example, the result would depend on the relative size of the variances of the observations and the estimates. However, the smaller the variance in the estimates, the smaller the variance in the difference between the estimates and the observations. Thus the smaller the probability that the estimate and the true mean differ by a large amount, the smaller the probability that the estimate and the observation differ by a large amount.

[4] The sum or difference of two normal distributions is also a normal distribution. The new mean is the difference of the two means.

[5] The new variance is the sum of the two variances.

[6] .112 ÷ .550 = .204.

NCCI's 2007 Hazard Group Mapping

by John P. Robertson

ABSTRACT

Excess loss factors, which are ratios of expected losses excess of a limit to total expected losses, are used by the National Council on Compensation Insurance (NCCI) in class ratemaking (estimating the expected ratio of losses to payroll for individual workers compensation classifications) and are used by insurance carriers to determine premiums for certain retrospectively rated policies (on policies for which claims used in the premium determination are subject to a per-claim limitation). Collections of workers compensation classifications that use the same expected excess loss factors are called hazard groups. At the beginning of 2007, NCCI implemented a new seven-hazard-group system, replacing the previous four-hazard-group system. This paper describes the analysis that led to the assignment of classes to the new seven hazard groups.

KEYWORDS

Hazard group, excess loss factor, excess ratio, cluster analysis, weighted k-means algorithm, standardization, credibility, injury type

1. Introduction

In the United States, most private employers are required to provide workers compensation coverage to pay employees injured on the job lost wages and medical costs arising from the work injury. Often employers provide this coverage by purchasing workers compensation insurance. For many insureds, premiums are based, in part, on the payroll classification of the employer, which is based on the type of business and operations performed by employees. For example, there is a classification for roofing businesses, and another classification for professional employees of hospitals. Currently there are about 800 different classifications in use in states for which NCCI provides ratemaking services (although the exact number used in any given state varies).

For various individual risk-rating purposes, for use in NCCI ratemaking, and for other reasons, it is useful to have tables of excess loss factors. An *excess ratio* or *excess loss factor* (*ELF*)[1] is the ratio of the expected amount of claims excess of a given limit to total expected claims. Because the probability that a loss is large, given that a loss occurs, varies by class, it is useful to have ELFs that vary by class.

A *hazard group* is a collection of workers compensation classifications that have relatively similar expected excess loss factors over a broad range of limits. NCCI periodically publishes tables of ELFs for states where NCCI provides ratemaking services. Generally these tables are updated annually, and give ELFs (or closely related factors) by hazard group for selected limits.

At the beginning of 2007, NCCI implemented a new seven-hazard-group system, replacing the previous four-hazard-group system. That is, under the new system, each classification is assigned to one of seven hazard groups. The seven new hazard groups are not simply a subdivision of the previous four; they are a substantially different mapping of classes to hazard group. This article describes the analysis that led to the assignment of classes to the new seven hazard groups.

Table 1. Distribution of classes by prior hazard group

NCCI Hazard Group	Number of Classes	Premium (billions)	Percent of Total Premium
I	38	$1.3	0.9%
II	428	$67.2	45.6%
III	318	$75.3	51.1%
IV	86	$3.6	2.5%

Under the previous NCCI four-hazard-group system, the bulk of workers compensation (WC) exposure in NCCI states was concentrated in two hazard groups, as can be seen in Table 1.

In our analysis, we considered whether a finer delineation would be possible, and what might be the optimal number of hazard groups. Apart from those considerations, hazard group assignments should be reviewed periodically because of changes over time in the insurance industry, technology, workplaces, and the evolution of the classification system and workers compensation infrastructure. The previous review had been done in 1993.

NCCI defines hazard groups on a country-wide basis. That is, the grouping of classes into hazard groups does not vary by state. Most workers compensation classes apply in every state where NCCI provides ratemaking services, although there are a few classes known as "state specials" that apply in only one state or a few states. NCCI takes the view, as it does in class ratemaking, that classes are homogeneous with respect to operations of the insureds, and therefore that the relative mix of injuries within a class should not vary much from state to state.

[1] In published tables, what we denote here as ELFs are often called Excess Loss Pure Premium Factors, or ELPPFs. And in published tables, ratios of excess loss to premium are often called Excess Loss Factors, or ELFs. Some published tables give ratios of excess loss plus allocated loss adjustment expense to either premium or loss plus allocated loss adjustment expense. We are concerned only with ratios of excess losses to total losses.

1.1. Prior work

The prior NCCI hazard groups were developed by first identifying seven variables based on relative claim frequency, severity, and pure premium, which were thought to be indicative of excess loss potential (NCCI 1993). These variables were the ratios of class average to statewide weighted average:

1. serious[2] to total claim frequency ratio
2. serious indemnity severity[3]
3. serious medical severity
4. serious severity, including medical
5. serious to total indemnity pure premium[4] ratio
6. serious medical to total medical pure premium ratio
7. serious pure premium to total pure premium ratio

Because of the correlations among these seven variables, the seven variables were grouped into three subsets based on an examination of the partial correlation matrix. A principal components[5] analysis was then done to determine a single representative variable from each of the three subsets and the linear combination of these representative variables that maximized the proportion of the total variance explained. The representative variables selected were the first, second, and last variables. The linear combination so identified is called the first principal component and is the single variable that was used to sort classes into hazard groups. Determination of the optimal number of hazard groups was outside the scope of that study and so the number of hazard groups remained unchanged at four.

A very different approach was employed by the Workers Compensation Insurance Rating Bureau of California (WCIRB 2001). The WCIRB's objective was to group classes with similar loss distributions. They used two statistics to sort classes into hazard groups. The first statistic was the percentage of claims excess of $150,000. This statistic was thought of as a proxy for large loss potential. The second statistic measured the difference between the class loss distribution and the average loss distribution across all classes. The different hazard groups corresponded to different ranges of these two statistics. The results were checked by using cluster analysis on these two variables.

1.2. Overview

Our approach owes much to the prior work on the subject, yet it is quite distinct. We sorted classes into hazard groups based on their excess ratios rather than proxy variables. As shown in Corro and Engl (2006), a distribution is characterized by its excess ratios and so there is no loss of information in working with excess ratios rather than with the size of loss density or distribution function. Section 2 describes how we computed class-specific excess ratios.

Section 3 describes how we used cluster analysis to group classes with similar excess ratios, and how we determined that seven is the optimal number of hazard groups. In Section 4 we compare the new hazard group assignments with the prior assignments.

Following the analytic determination of hazard groups, the tentative assignments were reviewed by several underwriters, and, based on this input, NCCI changed some assignments; we describe this in Section 5.

Finally, Section 6 recaps the key ideas of this study and the key features of the new assignments.

[2] A *serious claim* is one for which at least one of the following benefits for lost wages is paid or is expected to be paid:
a. Fatal (death)
b. Permanent Total (injured worker not expected to ever be able to work)
c. Permanent Partial (able to work after recovery period, but with a permanent injury, such as loss of a limb) and benefits for lost wages exceed certain thresholds that vary by state and year.

[3] *Severity* is the average claim cost. *Indemnity* is benefits for lost wages. *Medical* is benefits for medical costs.

[4] *Pure premium* is the ratio of expected losses to payroll in $100s.

[5] See Johnson and Wichern (2002) for a discussion of principal components.

2. Class excess ratios

Gillam (1991) describes in detail the NCCI procedure for computing excess ratios by hazard group for individual states. In the NCCI procedure, each ELF for a state and hazard group is a weighted average of ELFs by injury type specific to the state and hazard group. The ELFs for an injury type for a state and hazard group are derived from ELFs for the injury type in the state, adjusted to the estimated mean loss in the hazard group in the state. Injury types used by NCCI are Fatal, Permanent Total, Permanent Partial, Temporary Total, and Medical Only.

To put this in mathematical terms, let X_i be the random variable giving the amount of loss for injury type i in the state, and let X_i have density function $f_i(x)$ and mean μ_i. Let S_i be the normalized state excess ratio function for injury type i; that is

$$S_i(r) = E\left[\max\left(\frac{X_i}{\mu_i} - r, 0\right)\right] = \int_r^\infty (t-r) g_i(t)\,dt,$$

where $g_i(x) = \mu_i f_i(\mu_i x)$ is the density function of the normalized losses X_i/μ_i and $r \geq 0$ can be interpreted as an *entry ratio*, i.e., the ratio of a loss amount to the mean loss amount. For hazard group j, the overall excess ratio $R_j(L)$ at limit L is

$$R_j(L) = \sum_i w_{i,j} S_i(L/\mu_{i,j}), \qquad (1)$$

where $w_{i,j}$ is the percentage of losses due to injury type i in hazard group j (so $\sum_i w_{i,j} = 1$), and $\mu_{i,j}$ is the average cost per case for injury type i in hazard group j.

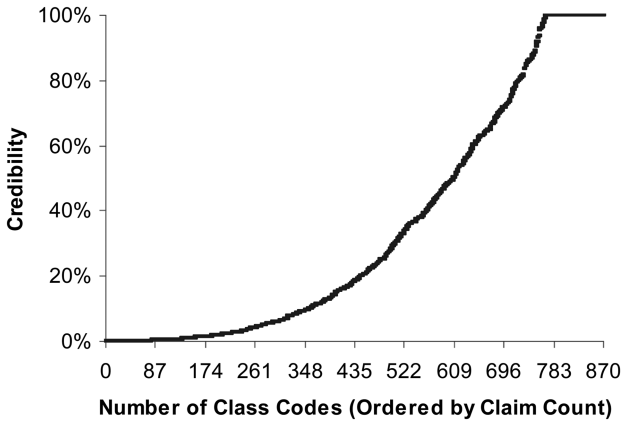
In the same way we can compute countrywide excess ratios for a given class by just knowing the weights and average costs per case by injury type for a class. These excess ratios were based on the most recent five years of data, as of April 2005, and included claim counts and losses by injury type for the states where NCCI collects such data. Losses were developed, trended, and brought on-level to reflect changes in workers compensation benefits. With some minor state exceptions, the same classes apply in all states. As such, we could estimate class excess ratios on a countrywide basis. Thus for each class, c, we had a vector

$$R_c = (R_c(L_1), R_c(L_2), \ldots, R_c(L_n))$$

of excess ratios at certain loss limits $L_1, L_2, \ldots, L_n$.

The credibility to assign to each class excess ratio vector is considered in the next subsection, and selection of the loss limits to use in the analysis is discussed in Section 3.

Figure 1. Class code credibility



2.1. Credibility

In the prior review, the credibility given to a class was

$$z = \min\left(\frac{n}{n+k} \times 1.5, 1\right), \qquad (2)$$

where n is the number of claims in the class and k is the average number of claims per class. This gives a class with the average number of claims 75% credibility and a class with at least twice the average number of claims full credibility. Figure 1 shows the credibility produced by this formula by size of class. The fully credible classes have over 70% of the total premium, as can be seen in Table 2. A few classes have most of the claims, as can be seen in Figure 2, where the classes with the greatest number of claims are to the left. Indeed, the distribution of claims per class is very

Table 2. Distribution of classes by credibility

Credibility Range	Claims per Year	Number of Classes	Percent of Premium
$0 \le z < 10\%$	0–237	355	1.2%
$10 \le z < 20\%$	238–511	89	1.3%
$20 \le z < 30\%$	512–831	61	1.6%
$30 \le z < 40\%$	832–1209	56	2.7%
$40 \le z < 50\%$	1210–1662	46	2.5%
$50 \le z < 60\%$	1663–2216	34	2.5%
$60 \le z < 70\%$	2217–2909	46	4.8%
$70 \le z < 80\%$	2910–3799	35	4.3%
$80 \le z < 90\%$	3800–4987	29	4.0%
$90 \le z < 100\%$	4988–6649	18	3.2%
$z = 100\%$	≥ 6650	101	71.8%
Total		870	100.0%

Figure 2. Distribution of classes by claim count

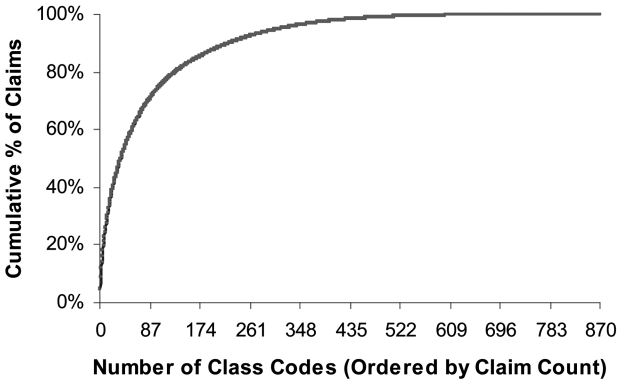
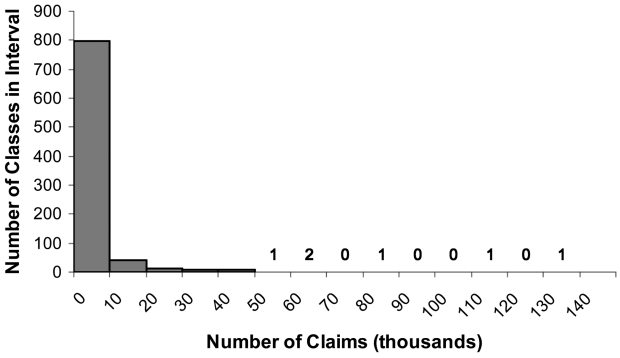


Figure 3. Histogram of number of claims by class



highly skewed, as can be seen in Figure 3. Figure 4 expands the first bar in Figure 3, and shows the persistency of the skewness. And Figure 5 further expands the first bar in Figure 4, revealing the same general pattern. The average number of claims per class is nearly ten times the median. We thus considered using the median rather than

Figure 4. Detail of histogram of number of claims by class

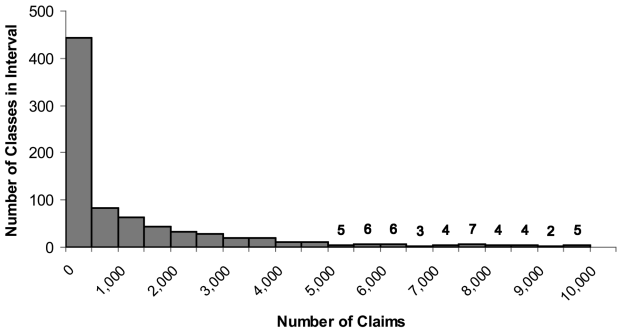


Figure 5. Detail of histogram of number of claims by class

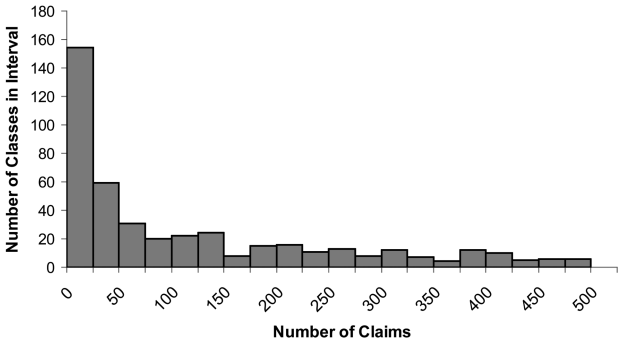
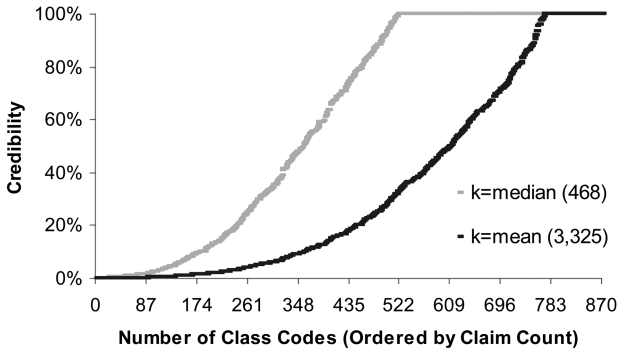


Figure 6. Comparison of credibility formulas



the mean for k in Formula 2. This would have resulted in a very large increase in credibility, as shown in Figure 6. We considered several other variations on Formula 2 as well. Because Medical Only claims have almost no impact on the ELFs at the published limits, we considered excluding all Medical Only claims. Taking that idea a step further, we looked at including only Serious claims. We also considered taking k in Formula 2

to be the mean number of claims over only those classes with some minimal number of claims.

In addition, we considered basing credibility on various square root rules. We considered a simple square root rule of the form

$$z = \sqrt{\frac{n}{384}},$$

where n is the number of claims in a class, and z is capped at 1. The full credibility standard of 384, given in Hossack, Pollard, and Zehnwirth (1983, p. 159), corresponds to a 95% chance of the actual number of claims being within 10% of the expected number of claims. For the determination of ELFs, serious claims (Fatal, Permanent Total, and major Permanent Partial) are more important than nonserious claims, so we looked at the following variation on the square root rule

$$z = \frac{N_F\sqrt{\frac{n_F}{384}} + N_M\sqrt{\frac{n_M}{384}} + N_m\sqrt{\frac{n_m}{384}}}{N_F + N_M + N_m},$$

where

n_F = the number of fatal claims in the class;
N_F = the number of fatal claims in all classes;
n_M = the number of permanent total and major permanent partial claims in the class;
N_M = the number of permanent total and major permanent partial claims in all classes;
n_m = the number of minor permanent partial and temporary total claims in the class;
N_m = the number of minor permanent partial and temporary total claims in all classes.

We also considered varying the full credibility standard by injury type with the following credibility formula

$$z = \frac{N_s\sqrt{\frac{n_s}{175}} + (N - N_s)\sqrt{\frac{n - n_s}{384}}}{N}$$

where

n_s = the number of serious claims in the class;
N_s = the number of serious claims in all classes;
n = the total number of claims in the class;
N = the total number of claims in all classes.

In the end, none of the alternatives considered seemed compelling enough to warrant a change and the results did not seem to depend heavily on the credibility formula; consequently we retained Formula 2 for computing credibility.

For the complement of credibility we used the excess ratios corresponding to the current hazard group of the class. More precisely, for each class c we have a vector of excess ratios

$$R_c = (R_c(L_1), R_c(L_2), \ldots, R_c(L_n))$$

and a credibility z. We also have a vector of excess ratios for the hazard group HG containing the class c (which can be determined, as above, as a loss weighted sum over vectors for classes in HG)

$$R_{HG} = (R_{HG}(L_1), R_{HG}(L_2), \ldots, R_{HG}(L_n)).$$

We now associate to each class a credibility-weighted vector of excess ratios

$$zR_c + (1 - z)R_{HG}.$$

It is these credibility-weighted vectors of excess ratios that we use in the cluster analysis described in the next section.

3. Analytic determination of the new hazard groups

The fundamental analytic method used to determine the new hazard groups is Cluster Analysis. It is a way to group classes with similar ELFs and is described in this section.

3.1. Selection of loss limits

The class excess ratio is a function of the loss limit, so it was necessary to select the limits to use in the analysis. We used limits of 100, 250, 500, 1000, and 5000, in thousands of dollars. Because excess ratios at different limits were highly correlated, five limits were thought to be sufficient. We considered using fewer limits but decided that it was better to use five limits to cover the range commonly used for retrospective rating.

Table 3. Correlations among excess ratios at selected limits

Limit	100,000	250,000	500,000	1,000,000	5,000,000
100,000		0.992	0.973	0.935	0.824
250,000			0.994	0.969	0.879
500,000				0.990	0.925
1,000,000					0.968
5,000,000					

Table 4. Correlations of ELFs for pairs of limits

Limits not Selected	Most Correlated Limit of the Five Selected	Correlation Coefficient
25,000	100,000	0.9882
30,000	100,000	0.9907
35,000	100,000	0.9926
40,000	100,000	0.9942
50,000	100,000	0.9965
75,000	100,000	0.9993
125,000	100,000	0.9996
150,000	100,000	0.9985
175,000	250,000	0.9987
200,000	250,000	0.9995
750,000	1,000,000	0.9982
2,000,000	1,000,000	0.9919

We began by considering the 17 limits for which NCCI published excess loss factors before 2005. These limits, in thousands of dollars, were: 25, 30, 35, 40, 50, 75, 100, 125, 150, 175, 200, 250, 300, 500, 1000, 2000, and 5000. We modified this list by dropping $300,000 and adding $750,000. We reduced this to the five selected limits based primarily on two considerations:

- ELFs at any pair of excess limits are highly correlated across classes, especially when the ratio of the limits is close to 1.
- Limits below $100,000 are heavily represented in the list of 17 limits.

The correlations were computed using only the 162 classes with at least 75% credibility. Classes with small credibility have estimated ELFs close to those for the prior overall hazard group. Including the low-credibility classes would skew the correlations towards those of the overall hazard groups.

Even among the five selected limits, correlations between ELFs for pairs of limits are very high, as can be seen in Table 3.

Each of the 12 limits not used has a correlation coefficient of at least 0.9882 with a limit that was used, as can be seen in Table 4.

Although we ultimately used five limits, we experimented by clustering with different limits. We found that the hazard group assignments resulting from five limits were quite similar to those resulting from 17. When mapping the classes to seven hazard groups, only 68 out of 870 classes were assigned to different hazard groups and these accounted for just 5.5% of the total premium.

To see whether five limits were more than needed for the analysis, we tried clustering the classes using only a single limit. In one instance we used $100,000 and in another we used $1,000,000. Figures 7 and 8 compare those single limit assignments with clustering using the five-limit approach. In both cases, the results differed from the five-limit case, markedly so when $1,000,000 was used. This indicates that too much information is lost by dropping down to one limit. Retrospectively rated policies are purchased over a range of limits and no single limit captures the full variability in excess ratios.

We used principal components analysis to enhance the clustering investigation. The first two principal components of the five limits retained over 99% of the variation in the data. While this might suggest that fewer limits could have been used, we decided to use five limits in order to cover the range of limits commonly used in retrospective rating. The distance between two classes in principal components space does not have the same simple interpretation as it does in excess ratio space. However principal components analysis allows one to project a five-dimensional plot onto two dimensions. Clustering using the five limits and plotting the resulting hazard group assignments using the first two principal compo-

Figure 7. Clustering using $100,000 limit compared to five selected limits (the number of classes that moved is shown above each bar)

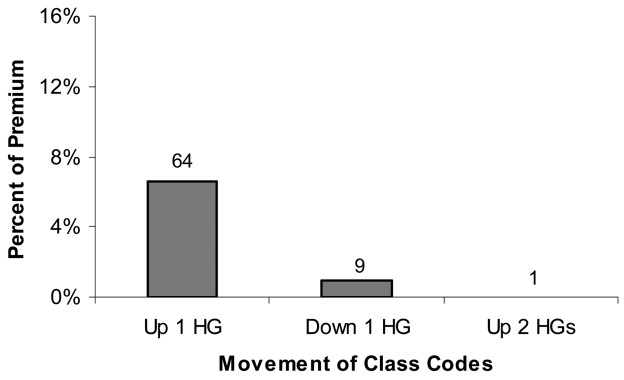
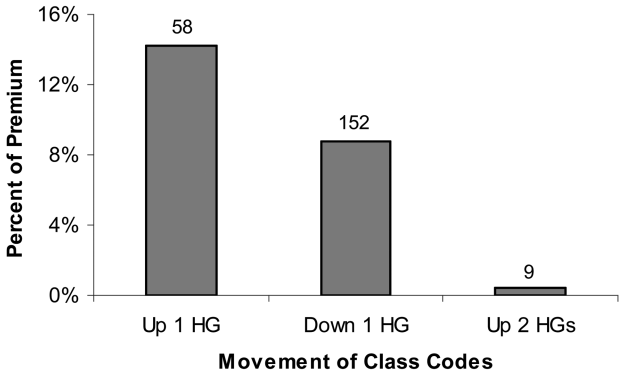


Figure 8. Clustering using $1,000,000 limit compared to five selected limits (the number of classes that moved is shown above each bar)



nents showed that the clusters were well separated and that outliers were easily identified. In our view, this confirmed the success of the five-dimensional clustering.

3.2. Metrics

The objective of assigning classes to hazard groups is to group classes with similar vectors of excess ratios. This raises the question of how to determine how similar or "close" two vectors are. The usual approach is to measure the distance between the vectors. If

$$x = (x_1, x_2, \ldots, x_n) \quad \text{and} \quad y = (y_1, y_2, \ldots, y_n)$$

are two vectors in $\mathbb{R}^n$, then the usual Euclidean, or L^2, distance between x and y is specified as

$$\|x - y\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

This metric is used extensively in statistics and is what we used. In linear regression this metric penalizes large deviations. That is, one big deviation is seen to be worse than many small deviations.

There are many other metrics. Perhaps the second most common distance function is the L^1 metric which specifies

$$\|x - y\|_1 = \sum_{i=1}^{n} |x_i - y_i|.$$

Here a large deviation in one component gets no more weight than many small deviations. The intuitive rationale for using this metric is that it minimizes the relative error in estimating excess premium. If $R_c(L)$ is the hypothetically correct excess ratio at a limit of L for a class c and the premium on the policy is P then the excess premium is given by $P \cdot PLR \cdot R_c(L)$, where PLR denotes the permissible loss ratio. But in practice the class excess ratio is approximated by the hazard group excess ratio $R_{HG}(L)$. The relative error in estimating the excess premium is then

$$\frac{|P \cdot PLR \cdot R_{HG}(L) - P \cdot PLR \cdot R_c(L)|}{P} = PLR \cdot |R_{HG}(L) - R_c(L)|.$$

If we assume that each loss limit is equally likely to be chosen by the insured, then the expected relative error in estimating the excess premium is given by

$$\sum_{i=1}^{n} \frac{PLR}{n} |R_{HG}(L_i) - R_c(L_i)| = \frac{PLR}{n} \|R_{HG} - R_c\|_1,$$

which is proportional to the L^1 distance between the two excess ratio vectors.

Our analysis was not very sensitive to whether the L^1 or L^2 metric was used and we preferred the more traditional L^2 metric.

3.3. Standardization

When clustering variables are measured in different units, standardization is typically applied to prevent a variable with large values from exerting undue influence on the results. Standardization ensures that each variable has a similar impact on the clusters. Duda and Hart (1973) point out that standardization is appropriate when the spread of values in the data is due to normal random variation, however "it can be quite inappropriate if the spread is due to the presence of subclasses. Thus, this routine normalization may be less than helpful in the cases of greatest interest."

We considered two common approaches to standardization. The usual approach is to subtract the mean and divide by the standard deviation of each variable. For example, if $x_1, x_2, \ldots, x_n$ are the sample values of some random variable, with sample mean $\bar{x}$, and sample standard deviation s, then the standardized values are given by

$$z_i = \frac{x_i - \bar{x}}{s}.$$

An alternative standardization method depends on the range of observations. Under this approach we would take

$$z_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}.$$

We conducted two cluster analysis trials in which we standardized according to the approaches described above. In each case we clustered the classes into seven hazard groups. Both trials resulted in hazard groups that were not very different from those produced without standardization.

Further, two issues were apparent with regard to standardizing in our particular analysis. First, excess ratios at different limits have a similar unit of measure, which is dollars of excess loss per dollar of total loss. That is, excess ratios share a common denominator. Any attempt to standardize would have resulted in new variables without a common unit interpretation. Second, all excess ratios are between zero and one. Some standardization approaches would have resulted in standardized observations outside this range.

Another consideration is the greater range of excess ratios at lower limits. Without standardization, the excess ratios at lower loss limits have more influence on the clusters than do those at higher limits. This result is not undesirable because excess ratios at lower limits are based more on observed loss experience than on fitted loss distributions (see Corro and Engl 2006). Even on a nationwide basis, there are few claims with reported losses above $5,000,000, but there are many more claims above $100,000. Greater confidence can be placed in the relative accuracy of excess ratios at lower limits because they are based on a greater volume of data.

In summary, the determination was made not to standardize because standardization would have eliminated the common denominator and it would have led to increased emphasis on higher limits. Our clustering algorithm used the L^2 metric and unstandardized credibility-weighted class excess ratios at the five selected loss limits: $100,000, $250,000, $500,000, $1,000,000 and $5,000,000. Premium weights were used to cluster the classes, as will be discussed in the next section.

3.4. Cluster analysis

Given a set of n objects, the objective of cluster analysis is to group similar objects. In our case, we wanted to group classes with similar vectors of excess ratios, where similarity is determined by the L^2 metric. At this stage the number of clusters is taken as given. Typically partitions of the objects into $1, 2, 3, \ldots, n$ clusters are considered. Non-hierarchical cluster analysis simply seeks the best partition for any given number of clusters. In hierarchical cluster analysis the partition with $k + 1$ clusters is related to the partition with k clusters in that one of the k clusters is simply subdivided to get the $k + 1$ element parti-

tion. Thus if two objects are in different clusters in the k cluster partition then they will be in different clusters in all partitions with more than k elements. This places a restriction on the clusters that can be sensible in some contexts. Our approach was non-hierarchical.

3.5. Optimality of k-means

The clustering technique we used is called k-means. For a given number, k, of clusters, k-means groups the classes into k hazard groups so as to minimize

$$\sum_{i=1}^{k} \sum_{c \in HG_i} \|R_c - \bar{R}_i\|_2^2, \tag{3}$$

where the centroid

$$\bar{R}_i = \frac{1}{|HG_i|} \sum_{c \in HG_i} R_c$$

is the average excess ratio vector for the ith hazard group and $|HG_i|$ denotes the number of classes in hazard group i. Theoretically there is a difference between the hazard group excess ratio vector, R_{HG_i}, computed using (1), and the hazard group centroid, $\bar{R}_i$, but in practice this difference is very small.

There is a commonly used algorithm to determine clusters, known as the *k-means algorithm* (Johnson and Wichern 2002). To start, some assignment to clusters is made. The algorithm then has two steps, performed iteratively until the clustering stabilizes. The first step is to compute the centroid of each cluster. The second step is to find the centroid closest to each class, and assign the class to that cluster. If any classes have been reassigned from one cluster to another during the second step, return to the first step. If no classes have been reassigned, then the algorithm terminates.

Commercial software for clustering is also available. We computed clusters using the SAS FASTCLUS routine.[6]

[6]We used SAS software, Version 8.2 of the SAS System for a SunOS 5.8 platform.

Hazard groups determined by k-means have several desirable optimality properties. First, they maximize the following statistic

$$1 - \frac{\sum_{i=1}^{k} \sum_{c \in HG_i} \|R_c - \bar{R}_i\|_2^2}{\sum_c \|R_c - \bar{R}\|_2^2}, \tag{4}$$

where

$$\bar{R} = \frac{1}{C} \sum_c R_c$$

is the overall average excess ratio vector, with $C = \sum |HG_i|$ being the total number of classes. Formula (4) is analogous to the R^2 statistic in linear regression. It gives the percentage of the total variation explained by the hazard groups.

A second way to evaluate hazard groups is based on the traditional concepts of within and between variance. We would like the hazard groups to be homogeneous and well separated. Thus we would like to minimize the within variance and maximize the between variance; using k-means accomplishes both.

Instead of considering a single excess ratio for each class, we have a vector of excess ratios, one excess ratio for each of several fixed loss limits. Thus we do not have a single random variable corresponding to an excess ratio at a single loss limit, but rather a random vector, with one random variable, the excess ratio, for each loss limit, from which we get a variance-covariance matrix. If X_i is the random variable for the excess ratio function at the ith loss limit, L_i, across classes c, then the observed values are the $R_c(L_i)$. The variance-covariance matrix of the random vector $X = (X_1, X_2, \ldots, X_n)$ is given by

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix},$$

where

$$\sigma_{ik} = E[(X_i - \mu_i)(X_k - \mu_k)]$$

is the covariance of X_i and X_k and $\mu_i = E[X_i]$. If we regard X as a $1 \times n$ matrix then

$$\Sigma = E[(X-\mu)^T(X-\mu)],$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ and $(X-\mu)^T$ is the transpose of $(X-\mu)$.

In practice the variance-covariance matrix is not known, but must be estimated from the data, i.e., the vectors

$$R_c = (R_c(L_1), R_c(L_2), \ldots, R_c(L_n)).$$

Let

$$\bar{x}_j = \frac{1}{C}\sum_c R_c(L_j),$$

where C is the total number of classes, and let

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n).$$

Then the sample covariance of the ELFs at L_i and L_k is

$$s_{ik} = \frac{1}{C}\sum_c (R_c(L_i) - \bar{x}_i)(R_c(L_k) - \bar{x}_k),$$

and the sample variance-covariance matrix is given by

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix}$$
$$= \frac{1}{C}\sum_c (R_c - \bar{x})^T(R_c - \bar{x}).$$

One way to generalize the concept of variance to the multivariate context is to consider the *trace* of S, the sum of the main diagonal of S

$$\text{trace}(S) = s_{11} + s_{22} + \cdots + s_{nn}.$$

This is just the sum of the sample variances of each variable and is called the *total sample variance*.

We let

$$T = CS = \sum_c (R_c - \bar{x})^T(R_c - \bar{x}).$$

The matrix T is proportional to the variance-covariance matrix for the whole data set. It is called the *dispersion matrix*, and is the matrix of sums of squares and cross products. We can proceed similarly within each hazard group and define

$$W_i = \sum_{c \in HG_i} (R_c - \bar{x}_i)^T(R_c - \bar{x}_i).$$

If we let

$$B_i = |HG_i|(\bar{x}_i - \bar{x})^T(\bar{x}_i - \bar{x}),$$

then it can be shown (see Späth 1985) that

$$\sum_{c \in HG_i} (R_c - \bar{x})^T(R_c - \bar{x}) = B_i + W_i.$$

We then let

$$W = \sum_{i=1}^{k} W_i.$$

This is the pooled within group dispersion matrix. For the between variance we let

$$B = \sum_{i=1}^{k} B_i.$$

This is the weighted between group dispersion matrix. We then have

$$T = B + W.$$

This means, roughly that the total variance is the sum of the between variance and the within variance. Taking the trace we get

$$\text{trace}(T) = \text{trace}(B) + \text{trace}(W).$$

Thus the total sample variance is the sum of the between and within sample variance. Because trace(T) is constant, maximizing trace(B) is equivalent to minimizing trace(W), which is what k-means cluster analysis accomplishes.

3.6. Weighted k-means

As observed in Section 2, some classes are much larger than others. To avoid letting the small classes have an undue influence on the analysis, we weighted each class by its premium. In simplest terms, this amounts to counting a

class twice if it has twice as much premium as the smallest class. So instead of minimizing the expression in (3), we instead minimized

$$\sum_{i=1}^{k} \sum_{c \in HG_i} w_c \|R_c - \bar{R}_i\|_2^2,$$

where w_c is the percentage of the total premium in class c. We used the premium-weighted centroids as well, that is

$$\bar{R}_i = \frac{\sum_{c \in HG_i} w_c R_c}{\sum_{c \in HG_i} w_c}.$$

3.7. Optimal number of hazard groups

So far, we have discussed the task of determining clusters when the number of clusters is given. We now address how to tell whether one number of clusters performs better than another, e.g., whether seven clusters works better than six or eight.

Various test statistics can be used to help determine the optimal number of clusters. The procedure is to compute the test statistic for each number of clusters under consideration and then identify the number of clusters at which the chosen statistic reaches an optimal value (either a minimum or a maximum, depending on the particular test statistic being used). Milligan and Cooper (1985) and Cooper and Milligan (1988) tested such procedures to determine which statistics were the most reliable.

Milligan and Cooper (1985) performed a simulation to test 30 procedures. The simulated clusters were well separated from each other and they did not overlap. For each simulated data set, the true number of clusters was known, and they computed the number of clusters indicated by each method of determining the optimal number of clusters. The methods were ranked according to the number of times that they successfully indicated the correct number of clusters.

They noted that their simulation was idealized but that "It is hard to believe that a method that fails on the present data would perform better on less defined structures" (1985, p. 161). Hence, although the hazard group data had both noise and overlap, it was useful to refer to Milligan and Cooper (1985) to determine which methods to rule out.

In a later study, Cooper and Milligan (1988) conducted tests that were more relevant to our application because random errors were added to the simulated data. That study found that the two best performing methods in the error-free scenario were also the best with errors (Cooper and Milligan 1988, p. 319). The best performing method is due to Calinski and Harabasz. Milligan and Cooper (1985, p. 163) define the Calinski and Harabasz statistic as

$$\frac{\text{trace}(B)/(k-1)}{\text{trace}(W)/(n-k)}$$

where n is the number of classes and k is the number of hazard groups, B is the between cluster sum of squares and cross product matrix, and W is the within cluster sum of squares and cross product matrix. Higher values of this statistic indicate better clusters because that corresponds to higher between clusters distances (the numerator) and lower within cluster distances (the denominator). This test is also known as the Pseudo-F test due to its resemblance to the F-test of regression analysis, often used to determine whether the explanatory variables as a group are statistically significant.

Another test that ranked high in the Milligan and Cooper testing was the Cubic Clustering Criterion (CCC). This test compares the amount of variance explained by a given set of clusters to that expected when clusters are formed at random based on data sampled from the multidimensional uniform distribution. If the amount of variance explained by the clusters is significantly higher than expected then a high value of the CCC statistic will result, indicating a high-performing set of clusters. An optimum number of clusters is identified when the test statistic

reaches a maximum (Milligan and Cooper 1985, p. 164).

Milligan and Cooper (1985) found that the Calinski and Harabasz test produced the correct number of clusters for 390 data sets out of 432. The CCC test produced the correct value 321 times. We could not use some of the other methods that ranked high because they were only applicable to hierarchical clustering, or for other reasons.

In a SAS Institute technical report, Sarle (1983) noted that the CCC is less reliable when the data is elongated (i.e., variables are highly correlated). Excess ratios are correlated across limits, so we gave the CCC results less weight than the Calinski and Harabasz results.

We performed cluster analyses for four to nine hazard groups. There were four hazard groups in the prior NCCI system, and we saw no reason to consider any smaller number. Implementing ten or more hazard groups would be substantially more difficult than implementing nine or fewer, because having 10 or more requires an additional digit for coding hazard groups. Testing up to nine was appropriate because the Workers Compensation Insurance Rating Bureau of California uses nine hazard groups (WCIRBC 2001).

In the first phase of our cluster analysis, we assigned classes and calculated the two test statistics for each number of groups under consideration. Figure 9 shows that the Calinski and Harabasz statistic indicated that the best number of hazard groups was seven. Figure 10 shows that the CCC statistic suggested nine hazard groups.

But nine hazard groups produced crossover, meaning that at some high loss limit the hazard group excess ratio for a higher hazard group was lower than the hazard group excess ratio for a lower hazard group. While crossover is possible in principle (from a purely mathematical standpoint, it is easy to specify two loss distributions so that one has higher ELFS at low limits and the other has higher ELFs at high limits),

Figure 9. Indicated number of hazard groups

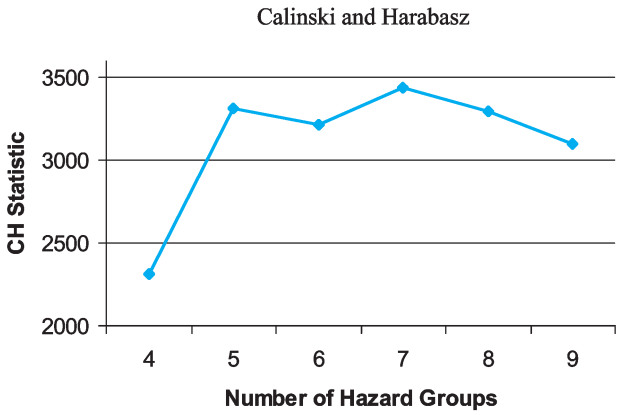
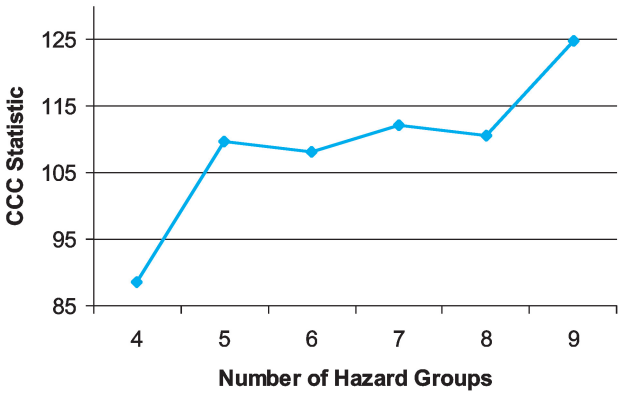


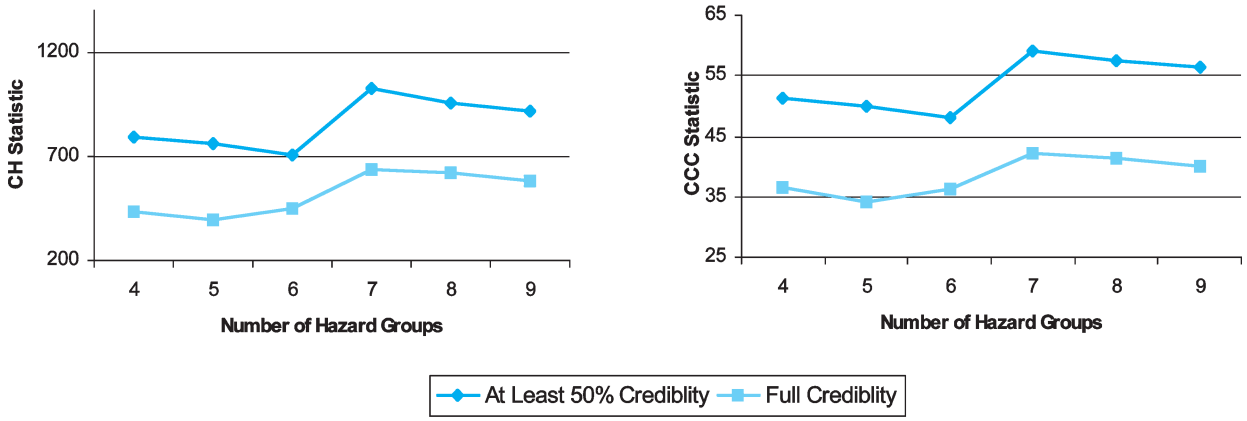
Figure 10. Indicated number of hazard groups, cubic clustering criterion



we don't think the data provided strong evidence for crossover, and one of our guiding principles was that there would be no crossover in the final hazard groups. In our opinion, the crossover that occurred with the clustering into nine hazard groups suggested that nine is more clusters than can accurately be distinguished.

As can be seen in Table 2, most of the premium is concentrated in the largest classes with the highest credibility. We were concerned that the indicated number of hazard groups in the analysis could have been distorted by the presence of hundreds of non-credible classes. In the second phase of our cluster analysis, we applied the tests to determine the optimal number of clusters using large classes only.

Figure 11. Statistics for various numbers of hazard groups, only classes with at least 50 percent credibility



In one scenario, we applied the Calinski and Harabasz and CCC tests using only those classes with credibility greater than or equal to 50 percent. In a second scenario, we applied the tests using only fully credible classes. As shown in Figure 11, the indicated number of hazard groups was seven for both tests in both scenarios.

In summary, we used two test statistics in three scenarios for a total of six tests. Seven hazard groups was the indicated optimal number in five of these six tests. The exception was the scenario in which all classes were included, where the CCC test indicated that nine hazard groups were optimal. There are four reasons why this exception received little emphasis:

- Milligan and Cooper (1985) and Cooper and Milligan (1988) found that the Calinski and Harabasz procedure outperformed the CCC procedure.
- The CCC procedure deserves less weight when correlation is present, which was the case in all of our scenarios.
- The selection of the optimal number of clusters ought to be driven by the large classes where most of the experience is concentrated. The large classes have the highest credibility and so the most confidence can be placed in their excess ratios.
- There is crossover in the nine hazard groups, and we had a guiding principle that there would not be crossover.

We concluded that seven hazard groups were optimal. These are denoted A to G, with Hazard Group A having the smallest ELFs and Hazard Group G having the largest.

3.8. Alternate mapping to four hazard groups

We recognized that some insurers would not be able to adopt the seven hazard group system immediately because they needed additional time to make the necessary systems changes. Therefore we produced a four hazard group alternative to supplement the seven hazard group system. We chose to collapse the seven hazard groups into four by combining Hazard Groups A and B to form Hazard Group 1, combining C and D to form 2, combining E and F to form 3, and letting Hazard Group 4 be the same as G. Having an alternate mapping to four hazard groups simplifies comparisons between the prior and new mappings as well.

Prior to choosing this simple scheme we considered other alternatives. We tried using k-means cluster analysis to map the seven hazard group centroids into four. This approach resulted in a hazard group premium distribution that was not homogeneous enough. Another approach we

considered was using cluster analysis to group the classes directly into four hazard groups. That approach yielded reasonable results, but it resulted in a non-hierarchical collapsing scheme, i.e., the seven hazard groups were not a result of subdividing the four hazard groups. The hierarchical collapsing scheme we chose has this feature, which allows users to know which of the four hazard groups a class is in based on knowing that class' assignment in the seven hazard group system.

The new four hazard group system is intended to be temporary. The four hazard group system is in place only to ensure that all carriers have sufficient time to make the transition to seven hazard groups.

4. Comparison of new mapping with old

4.1. Distribution of classes and premium

The bulk of the exposure was concentrated in two of the hazard groups prior to our review. Hazard Groups I and IV contained a small percentage of the total premium. Hazard Groups II and III, on the other hand, contained 97 percent of the total premium (see Table 1). We knew that a more homogeneous distribution of premium by hazard group would improve pricing accuracy. When discussing the new hazard groups in this section we will focus on the mapping that resulted directly from the statistical analysis. Later on, as will be discussed in the underwriting review subsection, numerous classes were reassigned among the groups based on feedback gathered in our survey of underwriting experts. These changes are not reflected in Figures 12 to 20.

Figures 12 and 13 compare the prior mapping to the collapsed new mapping based on the distribution of classes and premium. Hazard Group 1 has a large number of classes and a substantial portion of total premium in contrast to Hazard

Figure 12. Prior mapping vs. collapsed new mapping, number of classes per hazard group

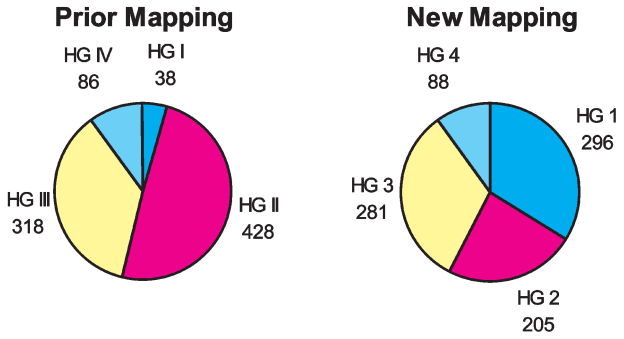
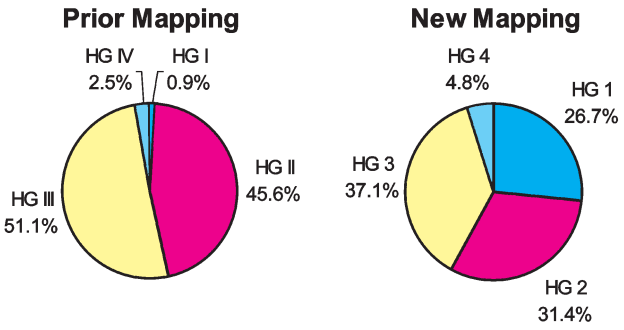


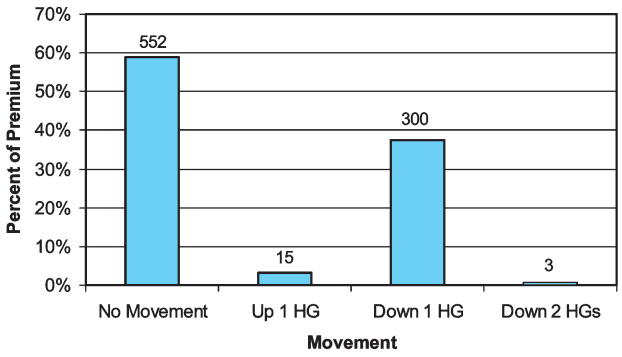
Figure 13. Prior mapping vs. collapsed new mapping, percent of premium by hazard group



Group I. Hazard Groups 2 and 3 have become slightly smaller than before although they are still large. In the prior mapping Hazard Groups II and III each had over 45 percent of the premium, but in the new mapping, none of the four groups has as much as 40 percent. This refinement allows for improved homogeneity of classes within each hazard group. Hazard Group 4 has retained a similar number of classes but it has more premium than Group IV.

Figure 14 shows that most of the classes and premium remained in the same hazard group when assigned to the new four Hazard Groups. Among those classes that did move, the great majority (300 classes and 37 percent of the premium) moved down one hazard group. Most of this movement was from Hazard Group II to 1. The movements of classes and premium are detailed in Table 5. The table can be read vertically. For instance, among the 428 classes in Hazard Group II, 255 were mapped into Haz-

Figure 14. Comparison of old with new assignment to four hazard groups (the number of classes that moved is shown above each bar)



ard Group 1, 164 into Hazard Group 2, nine into Hazard Group 3, and none into Hazard Group 4. The 255 classes that moved from Hazard Group II into Hazard Group 1 comprised 25.4% of the total premium. A significant number of classes and amount of premium moved from Hazard Group III to 2. Three classes moved from III to 1. Just 15 classes moved up by one hazard group, making up three percent of the premium. Hazard Group 1 is so large primarily because of classes that entered it from Hazard Group II. Hazard Group 2 is quite different than Hazard Group II because many of the classes in 2 originated in III and many of the classes that were in II have moved into 1.

The new seven hazard group assignment has a fairly homogenous distribution of classes and

Table 5. Comparison of distributions of classes between prior and new hazard group assignments

Prior Mapping					
Hazard Group	I	II	III	IV	Total
Number of Classes	38	428	318	86	870
% Premium	0.9%	45.6%	51.1%	2.5%	100%
Hazard Group					
1	38	255	3	0	296
	0.9%	25.4%	0.5%	0.0%	26.7%
2	0	164	41	0	205
	0.0%	19.6%	11.8%	0.0%	31.4%
3	0	9	268	4	281
	0.0%	0.6%	36.3%	0.2%	37.1%
4	0	0	6	82	88
	0.0%	0.0%	2.6%	2.2%	4.8%

premium, as shown in Figure 15. This distribution is a marked improvement over the prior mapping. In terms of premium, Hazard Group A is 11 times larger than Hazard Group I was. Hazard Group G is twice as large as Hazard Group IV was.

Table 6 shows the distribution of classes to hazard groups based on their level of credibility. Overall there were 162 classes with at least 75 percent credibility and 708 classes with lower credibility. Generally, within each hazard group most of the premium is due to highly credible classes but most of the classes have lower credibility. Hazard Groups D and G are exceptions. Hazard Group D has nearly equal numbers of

Figure 15. Number of classes and percent of premium in each hazard group

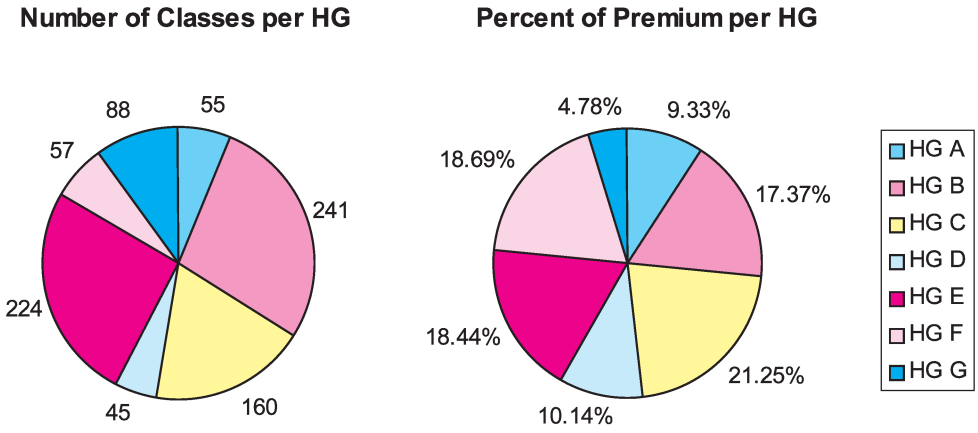


Table 6. Number of classes with given credibility by hazard group

	162 Classes with Credibility ≥ 75%		708 Classes with Credibility < 75%	
Hazard Group	Number of Classes	% Premium	Number of Classes	% Premium
A	18	8.4%	37	0.9%
B	40	14.5%	201	2.8%
C	41	17.6%	119	3.6%
D	22	8.9%	23	1.3%
E	22	14.0%	202	4.4%
F	15	15.0%	42	3.7%
G	4	2.4%	84	2.4%
Total	162	80.9%	708	19.1%

high and low-credibility classes. In Hazard Group G, high and low-credibility classes have similar premium percentages.

Although Hazard Groups B and E have far more classes than the other hazard groups, they do not have far more premium. The reason that they have the most classes with credibility less than 75 percent is that the complement of credibility is the prior hazard group excess ratio. For instance, the excess ratio of Hazard Group III at $100,000 was 0.451 which is close to the excess ratio of Hazard Group E. Given a small class in Hazard Group III, the credibility-weighted excess ratio was likely to be close to the excess ratio of Hazard Group E.

4.2. Range of excess ratios

In Figure 16 each horizontal bar represents the range of credibility-weighted excess ratios within a particular hazard group. The vertical line within each bar represents the overall excess ratio for the hazard group. Among the classes in Hazard Group I, the excess ratios at $100,000 ranged from 0.254 to 0.315. In Hazard Group II, the excess ratios at $100,000 ranged from 0.223 to 0.451. Thus the range of Hazard Group I excess ratios was contained within that of Hazard Group II, indicating that Hazard Groups I and II were not as well separated as might be desired.

Figure 16. Prior mapping excess ratio ranges at $100K

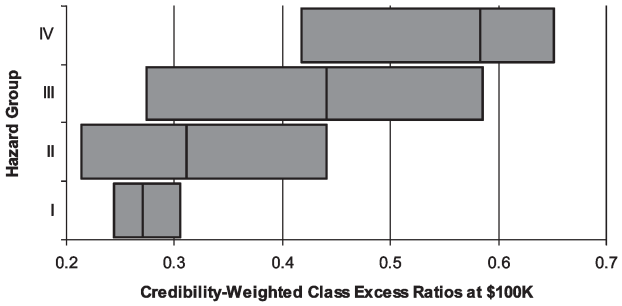


Figure 17. New mapping excess ratio ranges at $100K

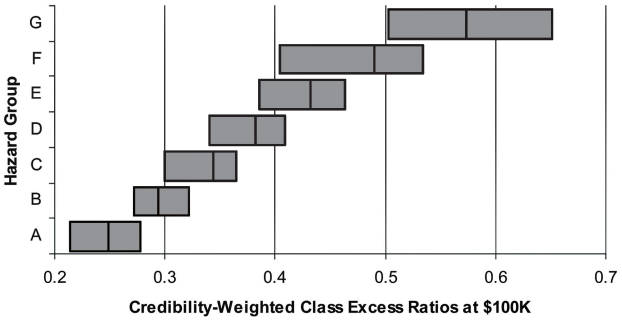
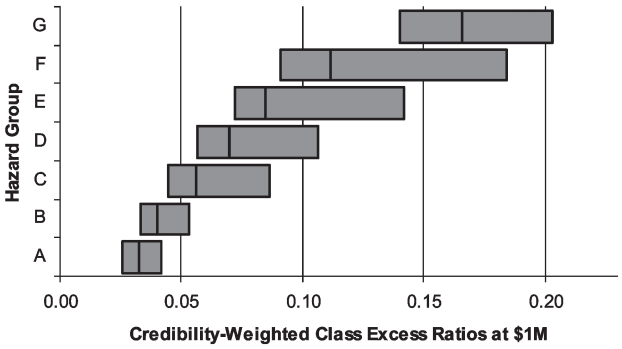


Figure 18. New mapping excess ratio ranges at $1M



The same behavior was observed at $1,000,000 as well.

As shown in Figure 17, *k*-means clustering resulted in well separated hazard groups. Because five dimensions were used, we could not avoid overlap in each dimension, but the excess ratio distribution is a noticeable improvement over the prior mapping. The new mapping also shows a well-separated excess ratio distribution at $1,000,000 as shown in the Figure 18.

Figure 19. New mapping excess ratio ranges at $100K, classes with at least 75% credibility

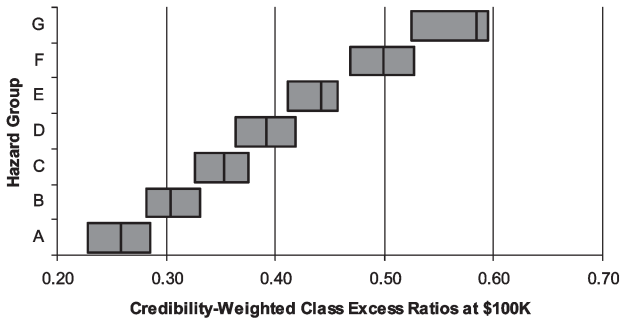
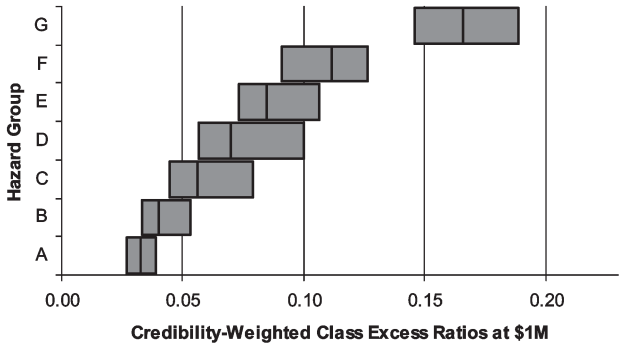


Figure 20. New mapping excess ratio ranges at $1M, classes with at least 75% credibility



Most of the exposure is concentrated in the largest classes, and so the hazard group excess ratios are highly sensitive to the placement of large classes. In Figures 16–18, the range of excess ratios for each hazard group is calculated using all of the classes in that hazard group.

Figures 19 and 20 show that if ranges are computed using only those classes with at least 75 percent credibility, then the separation of hazard groups by excess ratios is quite strong at both $100,000 and $1,000,000.

5. Underwriting review

After completing the cluster analysis, we conducted a survey of underwriters to solicit their comments on the proposed new mapping. The survey was sent to all members of NCCI's Underwriting Advisory List (UAL), and included the draft mapping that resulted from the analytic determination of the hazard groups. The survey asked the underwriters to judge the hazardousness of each class based on the likelihood that a given claim would be a serious claim. We also pointed out that if the mix of operations in two classes was very similar then the two classes should probably be in the same hazard group.

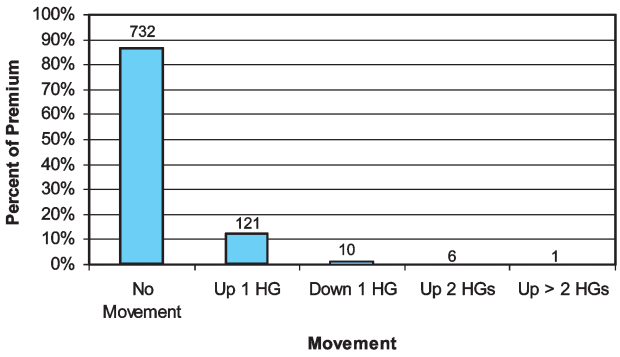
Members of the UAL recommended changes in the hazard group assignment for a third of the classes. We also received feedback from two underwriters on staff at NCCI. After the survey comments were compiled, a team consisting of NCCI actuaries and underwriters reviewed the comments from UAL members and decided on the final assignment for each class. When deciding whether to reassign a class, we considered whether the feedback on that class was consistent. We considered the credibility of each class and placed more weight on the cluster analysis results for those classes with a large volume of loss experience. For each class we compared the excess ratios to the overall hazard group excess ratios and identified the nearest two hazard groups.

Class 0030 illustrates the process used at NCCI to decide on the hazard group for each class. This class is for employees in the sugar cane plantation industry and is only applicable in a small number of states. This class

- had 12% credibility,
- was in Hazard Group III under the prior mapping, and
- was assigned to Hazard Group E under the cluster analysis.

An underwriter pointed out that Class 0030 has operations similar to Class 2021, which is for employees who work at sugar cane refining. Insureds in either class can have both farming and refining operations, their class being determined by which operation has the greater payroll. Also, both farming and refining involve use of heavy

Figure 21. Percent of premium that moved during the underwriting review (the number of classes that moved is shown above each bar)



machinery. Class 2021

- applies nationally,
- had 31% credibility,
- was in Hazard Group II under the prior mapping,
- was assigned to Hazard Group C under the cluster analysis, and
- prior to credibility weighting had excess ratios close to the overall excess ratios for Hazard Group D.

Credibility weighting had reduced Class 2021's excess ratios so that they were between the overall excess ratios of Hazard Groups C and D, because the prior assignment of Class 2021 had been to Hazard Group II.

We concluded that Hazard Group D was the best choice for 2021 based on its excess ratios prior to credibility weighting and its mix of operations. We determined that 0030 should be assigned to the same hazard group as 2021, so we also assigned Class 0030 to Hazard Group D.

Underwriters made several other types of comments besides those comparing one class to another. For instance, they commented on the degree to which employees in a given class are prone to risk from automobile accidents. They commented on the extent to which heavy machinery is used in various occupations and how much exposure there is to dangerous substances.

Figure 21 displays the movements of premium and classes during the underwriting review under the collapsed new mapping. It shows that the overall effect of the underwriting review was to move a significant number of classes up to a higher hazard group. The majority of the classes that moved up one hazard group, 78 of them, moved from Hazard Group 1 to 2, while 20 classes moved from Hazard Group 2 to 3, and 23 classes moved from Hazard Group 3 to 4.

6. Conclusion

Our approach to remapping the hazard groups was founded on three key ideas.

1. Computing excess ratios by class

The data is too sparse to directly estimate excess ratios by both class and state. But countrywide excess ratios can be computed by class in the same way that hazard group excess ratios are computed. This does not require separate loss distributions for each class. The existing loss distributions by injury type can be used along with the usual scale assumption. Thus all that is needed is average costs per case by injury type and injury type weights for each class.

2. Sorting classes based on excess ratios

Rather than using indirect variables to capture the amorphous concept of "excess loss potential," we used excess ratios directly because hazard groups are indeed used to separate classes based on excess ratios. Because a loss distribution is in fact characterized by its excess loss function, this approach involves no loss of information. By sorting classes based on excess ratios we achieve the goal of sorting classes based on their loss distributions as well.

3. Cluster analysis

Problems involving sorting objects into groups are not unique to actuarial science. We were thus able to make use of a large statistical literature on cluster analysis. This provided an objective criterion for determining the hazard groups as

well as the optimal number of hazard groups. Our approach to determining the seven hazard groups was non-hierarchical because we wanted the best seven group partition and because hypothetical partitions into six hazard groups are not relevant in this context.

As a result of our analysis the number of NCCI hazard groups was increased from four to seven. The distribution of both premium and classes is much more even across the new hazard groups. The highest hazard group is still relatively small. The new seven hazard groups collapse naturally and hierarchically into four hazard groups. Comparing the new four hazard groups with the old, over two-thirds of the classes, with nearly 60% of the premium, did not move at all. This stability was largely a result of the fact that we used the old hazard group as a complement of credibility and there were a large number of classes with very little premium. Of the classes that did move, the overwhelming majority moved down one hazard group.

The new mapping was filed in mid-2006 to be effective with the first rate or loss cost filing in each state on or after January 1, 2007. The filing (Item Filing B-1403) was approved prior to the end of 2006 in all states in which NCCI files rates or loss costs.

Acknowledgments

Many staff at NCCI contributed to this paper, including Greg Engl, Ron Wilkins, and Dan Corro. We also thank the NCCI Retrospective Rating Working Group and the NCCI Underwriting Advisory List for their input.

References

Cooper, M. C., and G. W. Milligan, "The Effect of Measurement Error on Determining the Number of Clusters in Cluster Analysis," in Gaul, W., and M. Schader (eds.), *Data, Expert Knowledge, and Decisions: An Interdisciplinary Approach With Emphasis on Marketing Applications*, Berlin: Springer-Verlag, 1988, pp. 319–328.

Corro, D. and G. Engl, "The 2004 NCCI Excess Loss Factors," Casualty Actuarial Society *Forum*, Fall 2006, pp. 513–571, http://www.casact.org/pubs/forum/06fforum/517.pdf.

Duda, R. O., and P. E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1973.

Gillam, W. R., "Retrospective Rating: Excess Loss Factors," *Proceedings of the Casualty Actuarial Society* 78, 1991, pp. 1–40, http://www.casact.org/pubs/proceed/proceed91/91001.pdf.

Hossack, I. B., J. H. Pollard, and B. Zehnwirth, *Introductory Statistics with Applications in General Insurance*, New York: Cambridge University Press, 1983.

Johnson, R. A., and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Upper Saddle River, NJ: Prentice Hall, 2002.

Milligan, G. W., and M. C. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika* 50, 1985, pp. 159–179.

NCCI (National Council on Compensation Insurance), "Revised Hazard Group Assignments," Actuarial Committee Agenda, Item ACT-92-57, April 20, 1993.

Sarle, W. S., Cubic Clustering Criterion (Technical Report A-108), Cary, NC: SAS Institute, 1983.

Späth, H., *Cluster Dissection: Theory, FORTRAN Programs, Examples*, New York: Halsted Press, 1985.

WCIRB (Workers Compensation Insurance Rating Bureau of California), "Retrospective Rating Methodologies—Hazard Group Assignment," Actuarial Committee Agenda, Item AC01-05-02, May 30, 2001.