

# Actuarial I.Q. (Information Quality)

CAS Data Management Educational Materials Working Party

---

## Abstract:

**Motivation.** Provide an introduction to data quality and data management directed at actuaries.

**Method.** Expand on the concepts in Actuarial Standard of Practice No. 23 (Data Quality), then introduce practical methods that actuaries, actuarial analysts, and management can apply to improve their situation, with references for more information.

**Results.** Information quality is about more than coding data: processes affect quality. There are many principles and practices an actuarial department can employ immediately to improve the quality of the information it deals with. Actuaries have a unique role to play in the bigger arena of improving their organizations' information for decision making and it is in their interests to do so.

**Conclusions.** What every actuary should know about data quality and data management.

**Availability.** Code for creating Box Plots in Excel is a link with this paper at <http://www.casact.org/pubs/forum/08wforum/>.

**Keywords.** Actuarial Systems; Data Administration, Warehousing and Design; Data Quality; Data Visualization; Exploratory Data Analysis; Software Testing.

---

## 1. INTRODUCTION

Data quality is a significant concern for most actuaries. In Britain, a GIRO Data Quality working party survey [1] found that about 25% of actuaries' time is expended on data quality issues. The survey also found that about 30% of actuarial analyses are adversely affected by data quality problems. Poor data quality is sometimes viewed as an inescapable fact of life by actuaries and other insurance industry analysts. However, actuaries, as both key consumers and providers of information, are uniquely well-positioned to deal with the pervasiveness of poor data quality in insurance.

Some think data quality is merely the accuracy of data. This paper identifies and discusses other characteristics (such as completeness and timeliness) and then broadens the perspective to information quality, which considers the broader picture of how information is processed and communicated. This includes not only data accuracy but other pitfalls that can result in users

## *Actuarial IQ*

misunderstanding information. Strategically, data quality is more important today, given easy access to an unprecedented level of detail and the proliferation of new tools and analysis techniques. Consequently, actuaries can add value by broadening how they think about data:

1. Data is a corporate asset that needs to be managed and actuaries have a role to play.
2. Data needs to be appropriate for all of its intended uses, not just the analysis at hand.

This paper contains tools, concepts, and references to support and facilitate this expanded perspective in order to help actuaries transform data into more useful information to make better decisions.

### **1.1 Research Context**

The actuarial literature on data quality is relatively sparse. In North America, the Actuarial Standards Board (ASB) Actuarial Standard of Practice No. 23 on Data Quality (ASOP No. 23) [2] provides guidelines to actuaries when selecting data, relying on data supplied by others, reviewing and using data, and making disclosures about data quality.

The Casualty Actuarial Society (CAS) Committee on Management Data and Information and the Insurance Data Management Association (IDMA) produced a white paper on data quality [3]. This CAS committee also promotes periodic calls for papers on data management and data quality which are published in the *CAS Forum*. The CAS online database (DARE) taxonomy can help users narrow their searches to papers on specific topics such as actuarial systems, data organization, and exploratory data analysis.

In response to one such call for papers, Francis [4] provided guidance for specific techniques which can be used to screen data for quality errors. Francis pointed out that 80% or more of time spent on large modeling projects is spent on data issues. However, the focus of the paper was on detecting errors after the fact, and not on techniques for preventing them.

The subject of data quality is also of interest internationally. A working party of the UK General Insurance Research Organization (GIRO) developed recommendations for improving the quality of reserve estimates. The Reserving (GRIT) working party report [5] recommended more focus on data quality and suggested that UK professional guidance notes incorporate standards from ASOP No. 23. Furthermore, the GRIT survey found that many respondents expressed concern over data quality.

In researching this paper, the working party reviewed seven books recommended by the IDMA,

## *Actuarial IQ*

as well as two more recommended by a working party member. Many books talk about data management as a means to achieve data quality, and some deal specifically with data quality. However, these books tend to be written for information technology professionals to apply to any organization. Since our goal is an introduction for actuaries, these texts are only cited occasionally. The collection of reviews of these books was published in the Winter 2007 CAS *Forum* [6].

### **1.2 Objective**

ASOP No. 23 sets standards for data quality that address a number of key areas but there are times when an actuary might want to go further. For example, if a reasonableness check reveals some data shortcomings, ASOP No. 23 outlines the ramifications for the analysis at hand. However, the actuary may be in a position to prevent data quality issues in source databases from arising by advocating improvements in data management and data quality practices. This paper discusses some of the practices and options available.

Other papers published by the CAS tend to focus on particular data management subjects: there is no broad introduction to the subject. Conversely, it is difficult for actuaries to apply nonactuarial texts on data management and data quality since these texts often presume the reader has a working knowledge of related IT concepts and unrestricted access to an organization's data centers.

This paper is a data quality introduction and reference for actuaries and actuarial analysts. As such it attempts to bridge the gap between ASOP No. 23 and the literature available for people in the actuarial profession who want or need more information. It is also the authors' hope that actuaries and actuarial analysts will become advocates for information quality once they see the business value information quality provides in:

- More accurate analyses (and hence smaller margins of error),
- Ability to focus on higher value activities once significant data issues are resolved,
- Increased impact of their analyses by increasing transparency and legibility of results.

### **1.3 Disclaimer**

While this paper is the product of a CAS working party, its findings do not represent the official view of the Casualty Actuarial Society or the employers of the authors. Nor is anything in this paper intended as a standard of practice nor an interpretation or guidance of existing standards. Moreover, while we believe the approaches we describe provide sound guidance on how to address the issue of information quality, we do not claim they are the only acceptable ones. Similarly, we believe the

textbooks and papers cited here are good sources of educational material on data management and data quality issues, but we do not claim they are the only appropriate ones. Finally, we have illustrated various concepts and methods with examples. The particular software used to illustrate examples is not necessarily the only or the best software for the purpose.

## 1.4 Outline

Section 2 will discuss concepts, whereas section 3 will focus on techniques. In brief, section 2 discusses the motivation for data quality and describes characteristics of quality data. It then expands the scope to a discussion of metadata and a common example of metadata in property and casualty insurance: statistical plans. Section 3 begins with techniques for improving the quality of data (exploratory data analysis and data audits) then turns to information quality in processing (models and presentations). It concludes with a discussion of the organizational and management issues: data quality measurement (as a tool to track improvement), improvement strategies, and data management. Section 4 reiterates the main topics of the paper.

## 2. BACKGROUND AND THEORY

Quality issues have come to forefront recently due to several key developments:

- **(Unprecedented) level of detail.** Computerization and cheap data storage along with changes in regulatory requirements have led to extraordinary amounts of data being captured, stored, and provided to actuaries. Consequently, enormous amounts of data can amass enormous numbers of errors and inconsistencies.
- **Availability of new tools.** Recent years have seen the proliferation of powerful data analysis packages and technologies: from XML-enhanced data exchange to object-oriented databases to servers enabled with On Line Analytical Processing.
- **Competition.** Competition encourages pricing techniques to be more and more precise – witness the growth of predictive modeling. In this environment, requirements for quality of data used in pricing algorithms grow immeasurably.
- **The growing data management skill set of actuaries.** Modern actuaries are more technically prepared for the challenges of dealing with huge amounts of data using contemporary tools and techniques. They should be able to tackle data quality issues with aplomb.

## *Actuarial IQ*

In their work, actuaries rely on vast amounts of data: claims loss runs, premium bordereaux, interest rates, and industry statistics, just to name a few. All of these data originate outside of actuarial reach and their collection and accumulation generally occur without actuarial control. Before it reaches actuaries, every piece of data passes through several stages: data are collected by a TPA, MGA, or some other source; then they get transferred to the insurance company system; and, after that, they can be grouped, accumulated, and mapped to a suitable structure. At each of these stages, data are processed and modified by people of different professions and qualifications who inevitably introduce errors into the data. The longer the data pipeline, the more errors accumulate and can compound one another. Multiple data sources also tend to multiply data problems.

As data progresses from input to information to decisions, the actuary's role changes from consumer to provider. This position is almost unique in the insurance data life cycle: indeed, as information providers for decision makers, actuaries are held to the highest standards of work quality; but, as consumers, actuaries depend on someone else. Better than any other professionals in the insurance industry, actuaries can become data quality protectors: they have knowledge of the data content, expertise to develop sophisticated data testing tools, and high stakes in the quality of the data.

Whereas ASOP No. 23 focuses on data's suitability for a particular actuarial analysis, we will present a broader introduction to data and information quality. A schematic overview of the development and usage of insurance data with respect to actuarial work is provided by [Figure 2.0.1](#). The schematic outlines the data life cycle for insurance. The goal of each major step and the function within the organization most responsible are given. These are followed by some examples of the types of errors that can be introduced in each step. The "Topics" column identifies the sections in this paper most pertinent to each step. As such, the figure is a helpful roadmap identifying where to find more information in this paper and providing the general context. Note that metadata (section [2.3](#)), data quality measurement (section [3.5](#)), data quality improvement strategies (section [3.6](#)), and data management (section [3.7](#)) considerations permeate the entire process. The multiple references to actuaries illustrates actuaries' broader opportunity to improve information quality not just for the analysis at hand, but for better decision making in the organization as a whole.

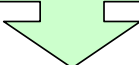
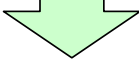



Steps	Purpose	Responsibility	Examples of Errors	Topics
<p>Step 0 <b>Data Requirements</b></p> 	Determine the intended use of data and required data elements	Data managers and actuaries at the source and the destination	Specification errors, granularity mismatches	Data quality (2.1) and its principles (2.2)
<p>Step 1 <b>Data Collection</b></p> 	Collect data and satisfy statistical reporting	Data managers at primary sources: TPAs, MGAs, insurers, statistical agents	Input errors	Statistical plans (2.4), data audits (3.2)
<p>Step 2 <b>Transformations Aggregations</b></p> 	Make data available to users in the necessary format and level of detail	Data managers at the source and the destination	Missing values, duplicate records, mapping errors	EDA (3.1), data audits (3.2)
<p>Step 3 <b>Analysis</b></p> 	Extract useful information from raw data	Actuaries (perhaps with the help of others)	Wrong model choice, censorship, over-fitting, calculation errors	IQ in Models (3.3)
<p>Step 4 <b>Presentation of Results</b></p> 	Help management make right decisions	Actuaries	Inconsistencies, mislabeling, inadequate labeling	Data Presentation (Reports) Quality (3.4)
<p>Final Step <b>Decisions</b></p>	Make profit, customer care, public welfare	Management	Interpretation errors, wrong conclusions	Not addressed in this paper

Fig. 2.0.1

## **2.1 What is Data Quality?**

Generally speaking, something is of high quality if it is particularly appropriate for its purpose. According to ASOP No. 23, “for purposes of data quality, data are appropriate if they are suitable for the intended purpose of an analysis and relevant to the system or process being analyzed” ([2], page 2). ASOP No. 23 advises the actuary to obtain a definition of data elements in the data, to identify questionable values and to compare data to the data used in a prior analysis. The actuary is also advised to judge whether the data is adequate for the analysis, requires enhancement or correction, requires subjective adjustment, or is so inadequate that the analysis cannot be performed. In making this judgment, ASOP No. 23 lists six things actuaries should consider when selecting data (discussed in section [2.2](#) below). ASOP No. 23 is often considered only with respect to the analysis at hand. However, if the analysis is repeated periodically or the same data is used for multiple purposes, it may be advantageous to address some of the recurring data quality issues.

A key component of this bigger picture is the concept of metadata. “Metadata” is simply information about data. As such, it helps determine if particular data are suitable for a particular purpose and insures that it is used appropriately. Metadata can help identify invalid entries, facilitates transferring data among systems, can improve the interpretation of analyses, and can prevent blunders due to misinterpretation of data. It is described more fully in section [2.3](#).

The key idea is that quality data is appropriate for its intended purpose. Note that this makes quality a relative, not absolute, concept: data may be of adequate quality for one analysis while being inappropriate for another purpose. For example, data that is appropriate for an annual overall rate adequacy study may not be appropriate for a relativity analysis or even for a midyear overall rate indication. This is particularly an issue in predictive modeling, where the analyst attempts to find better predictors (of losses, for example): promising variables may not have been coded or processed with the intent of using them for this purpose.

### **2.1.1 Data quality versus information quality**

Everyone has heard the well known IT adage “garbage in – garbage out”: it says that poor quality inputs will lead to poor quality outputs. Put another way, it says that processing or analysis cannot completely correct bad input. This consideration of processing distinguishes information quality from data quality. Dasu and Johnson [7] talk about “end-to-end-data-quality.” That is, there are many stages in the data assembly process where data quality needs to be monitored and improved, such as during data collection, transformation and aggregation, data storage, and data analysis. Their equation:

DATA + ANALYSIS = RESULTS

highlights that quality results depend not only on quality data, but also on quality analysis. The quality of the final product is not only affected by the quality of the data itself, but also by how the data is processed (e.g., how it is transformed, aggregated, analyzed and presented).

This consideration of processing leads to a larger concept of metadata: the initial definition of metadata could be restricted to a particular database, but it can also be expanded to integrate information across applications, as new data is created with each application. Metadata is discussed more fully in section [2.3](#).

Information quality does not have a commonly accepted definition. It is used in this paper to remind readers that data quality is about more than just correct coding: quality is affected by how data is stored, processed, and analyzed, and how results are presented.

From a data manager's perspective, it also includes what facts are captured as data and how they are captured.

## 2.2 Principles of Data Quality

When evaluating the quality of a dataset for a particular analysis, ASOP No. 23 advises actuaries to “select the data with due consideration of the following”:

- **Appropriateness** for the intended purpose of the analysis, including whether the data are sufficiently current;
- **Reasonableness** and **comprehensiveness** of the necessary data elements, with particular attention to internal and external consistency;
- Any known, material **limitations** of the data;
- The cost and feasibility of obtaining **alternative data**, including the availability to obtain the information in a reasonable time frame;
- The benefit to be gained from an **alternative data set** or data source as balanced against its availability and the time and cost to collect and compile it; and
- **Sampling methods**, if used to collect the data. ([2], page 3)

Similarly, the CAS Management Data and Information Committee “White Paper on Data Quality” [3] states that evaluating the quality of data consists of examining the data for:



## Actuarial IQ

- **Validity:** “the value of a given data element is one of all allowable ones” ([3], page 155)
- **Accuracy:** “each data transaction record or code is a true and accurate representation of what it’s intended to represent” ([3], page 156)
- **Reasonableness:** “is the data reasonable compared to our prior and current knowledge?” ([3], page 157), and
- **Completeness:** each record contains all the data necessary for business needs and every step in data collection and processing handles it correctly, without duplication.

The white paper goes on to note that there are three levels of accuracy for usable data:

- **Absolute:** data is 100% correct for every data element and every transaction,
- **Effective:** there are some errors but they should have no material impact on the results of the analysis,
- **Relative:** data is “inaccurate but consistent over time” ([3], page 158).

### 2.2.1 Validity versus accuracy

One misconception is that if data is valid, then it is accurate. To see why this is not true, consider, for example, the ZIP Code. The recorded ZIP Code may be one of the possible ZIP Codes in the state (valid) but it may not be the correct one associated with the particular risk’s address. Standalone edits in policy administration systems can check the validity of the data while more complex relationship edits and audits can be used to check for accuracy.

### 2.2.2 Data quality through data management

Now that data quality is defined, how is it achieved? Section 3 describes a number of options actuarial analysts can pursue to improve their information quality, but the most holistic way is by good data management. This is because good data management broadens the point of view from the data for the analysis at hand to the entire process that gave rise to the data as well as other potential applications and users of the data. There are some additional data quality principles from this broader perspective.

Various authors of data quality literature describe the dimensions of data quality. A comprehensive list is provided in *Data Management: Databases and Organization* [8], by Richard T. Watson. Watson defines eighteen dimensions of data quality. Some of these dimensions are the key principles described above. Others describe ways of storing data such as:

*Actuarial IQ*

Dimension	Conditions for high quality data
Representational consistency	Values for a particular data attribute have the same representation across all tables (e.g., dates)
Organizational consistency	There is one organization-wide table for each data element or entity and one organization-wide data domain for each data attribute
Record consistency	The values in a record are internally consistent (e.g., a home phone number's area code is consistent with a city's location)
Flexibility	The content and format of presentations can be readily altered to meet changing circumstances
Precision	Data values can be conveniently formatted to the required degree of accuracy (e.g., in cents or in thousands)
Granularity	Data are represented at the lowest level necessary to support all uses (e.g., hourly sales)

*Table 2.2.1*

Notice how these dimensions support the key principles of validity, accuracy, reasonableness, and completeness.

Watson's list goes beyond data characteristics to processing and management principles, such as:

Dimension	Conditions for high quality data
Stewardship	Responsibility has been assigned for managing data
Sharing	Data sharing is widespread across organizational units
Timeliness	A value's recentness matches the needs of the most time critical application requiring it. Values remain up to date.
Interpretation	Clients correctly interpret the meaning of data elements

*Table 2.2.2*

Other key concerns for data managers are the proprietary nature of data and the privacy issues. An insurer's data contains much information about its business: who it insures, the premium it charges, the claim it has paid. Many insurers consider this information to be a trade secret. As such, data managers and the users of the data (e.g., actuaries) must be careful to protect the data of their employer or client from being divulged to their competitors. Likewise, insurance data may contain

## *Actuarial IQ*

data elements about an individual person, such as their social security number, FICO scores, and health records that from an ethical and legal perspective should remain confidential.

Data management is discussed more fully in section [3.7](#) below. The next section, metadata, is the key to the interpretation dimension.

### **2.3 Metadata**

#### **2.3.1 What is metadata?**

Metadata is a term used by data management and data quality professionals to denote the data that describes the data, e.g., the documentation of the contents of a database. In addition to information about the data itself, metadata contains information about business rules and data processing. Examples of metadata in insurance are the ISO and NCCI statistical plans.

Good metadata serves as a roadmap to the business processes of the entire organization and as such needs to be shared with the entire organization. As a result, actuaries should take an active role in understanding and developing metadata. The actuary's role in metadata will be discussed in section [2.3.2](#) and the sharing of metadata across an organization will be discussed more in section [2.3.3](#).

At a minimum, metadata will include a listing of all data elements in a database, along with a description of what is contained in each data element. Each data element listed should be defined clearly, and the data that is in the data element described. For example, the data element "pol\_eff\_date" may be defined to contain the policy effective date and should contain only date values. Furthermore, the date format may be specified, such as mm/dd/yyyy. The permissible ranges of the values (e.g., 1/1/2000 to present) should be specified. Any default values (e.g., 1/1/2000) should be documented. Similarly, metadata should define the values in categorical data.

Metadata should also identify when and how a data element is processed. As an example, Table [2.3.1](#) shows seven values in the data for the marital status data element, including a value for the case when marital status is missing. If multiple sources of data are used to populate a database, then the source of the data should be listed. Any transformations done to the data need to be documented as well. The documentation should also describe how frequently the data is updated from the sources.

Marital Status Value	Description
1	Married, data from source 1, straight move of field ms_code
2	Single, data from source 1, straight move of field ms_code
4	Divorced, data from source 1, straight move of field ms_code
D	Divorced, data from source 2, straight move of mstatus
M	Married, data from source 2, straight move of mstatus
S	Single, data from source 2, straight move of mstatus
Blank	Marital status is missing

Table 2.3.1

Metadata can also exist on the compilation or extraction processes. It should include information on such items as fiscal period definitions and how evaluation dates are determined.

Ideally, metadata should also include business rules, such as how reported claims are defined. It should also document interdependencies with other data elements. For example, the date of birth for a driver should be at least 15 years earlier than the date the driver received their license.

The inclusion of documentation on the quality of data can enhance the metadata. For example, to really understand the data, a general narrative on the quality checks and controls of the data is necessary. Other useful metadata include a data quality matrix for each data element. This would describe the quality checks done on the data element, how frequently the checks are done, and where in the process the check occurs.

Better process documentation can also enhance metadata. For example, a high-level data process flow diagram that shows each initial data feed (source) and any data stores (databases) associated with the data will give users and developers better insight into the processing. Another example of enhanced process documentation is a glossary of terms that provides definitions specific to the data and systems under consideration.

Finally, some sort of versioning is helpful to identify when changes take place. For example, when did the claims department change the average reserves? When did rating territory begin being derived from zip code instead of input? When did a new product or alternative distribution channel go live?

## *Actuarial IQ*

A complete description of the contents of a database is important for the appropriate use of the data. Good metadata will assist the analyst in avoiding misunderstandings that result in revisions of the analysis when the contents of a data element or variable are discovered to be other than what it was assumed to be. As a result, metadata is an important tool for actuaries to use when planning their analyses. Problems can arise for actuaries when metadata is either nonexistent or is inaccessible to actuaries. Metadata that is incomplete, inaccurate, or out of date can also lead to problems.

Creating quality metadata at an organizational level is a large undertaking and really requires commitment from all levels of the organization. The next section talks about the actuary's role in metadata and some suggestions that can be used in any organization to get started and perhaps build the necessary commitment.

### **2.3.2 The actuary's role in creating and sharing metadata**

Maintenance of adequate documentation describing data can help avoid problems associated with relying exclusively on people's memories of what is contained in the data. As actuaries, we can help persuade our business and data management partners that system documentation is vital to the actuarial work product.

At the same time, we can employ the same standards of metadata and documentation to the actuarial work product. After all, actuarial work is a source of data and information for others in the insurance industry, so it follows that the same principles of metadata should be applied. Metadata from actuarial projects can be shared with appropriate data management and system colleagues to ensure that the data is being properly used. Sharing of metadata within the user community (actuarial, data management, finance, etc.) is a vital activity for the organization. To quote the *Corporate Information Factory* ([9], page 170): "Metadata is the glue that holds the architecture together. Through metadata, one component of the architecture is able to interpret and make sense of what another component is trying to communicate."

Documenting anything from a basic actuarial project to a complex information system can be a daunting task. The following sets of considerations can be used to help test existing metadata or get started on putting together new metadata.

Minimum considerations:

- Are all the data elements listed?

### *Actuarial IQ*

- Has the source of each data element been provided?
- Is there a special value that is used to indicate missing data?
- Are any transformations being applied to data? (Note: data cleanup such as filling in missing values should be considered a data transformation.)

More advanced considerations:

- Have the contents and use of each data element been properly described?
- Have all the categorical values of each data element been properly described?
- In the case of numeric data, has the range of possible values for each data element been provided?
- Has the valuation date of all data been provided?
- Has a schedule of planned updates to the data been provided?
- Has the business process changed during the experience period?
- Have any of the data definitions changed during the experience period?

As was noted above, a good place to start is with our own actuarial work product. In many instances, we may produce or maintain databases that underlie our analyses. How well documented are these systems? How well understood are the sources that feed the actuarial systems? Once the actuarial systems are understood, one can start to drill back into the source systems. Along the way, missing metadata can be identified. The benefits and costs of producing the metadata can be weighed and ownership could be assigned.

As metadata is developed, it needs to be shared across the organization. That is the topic of the next section.

#### **2.3.3 Sharing metadata across an organization**

Actuaries can also face the problem of access to metadata (or at least to the most up-to-date metadata). Just like data, metadata can exist in multiple forms, such as word processing documents, printed documents, spreadsheets, and databases. It can also be stored in multiple locations, including file servers, paper files and within the documented system itself. Keeping track of and sharing all that metadata can be difficult.

Technology can provide answers to these types of collaboration issues. It is worthwhile for

## *Actuarial IQ*

actuaries to be plugged into the collaboration technologies that are available within their organizations. Examples include intranets, quick places, hyperlinks, comment boxes, and the emerging wiki technologies.

The *Corporate Information Factory* [9] addresses this issue by introducing the concept of autonomous versus shared metadata. The key issue is that “metadata has a need to be shared, and a propensity to be managed and used in an autonomous manner. Unfortunately, these propensities are in direct conflict with each other” ([9], page 170). Consequently, each component of a system, such as a table or database, should have its own metadata and metadata should be split into autonomous and shared groups. Autonomous metadata is only used (or applicable) within the component. “Sharable metadata must be able to be replicated from one architectural component to another” ([9], page 174). Splitting metadata into these groups need to be mutually exclusive and exhaustive. The rule of thumb is that “very commonly used metadata needs to be shared” ([9], page 175).

At the end of the day, access to metadata is as simple (and as difficult) as building and maintaining good relationships between the actuarial and data management communities.

### **2.4 Statistical Plans**

Some of the most widespread examples of metadata are the statistical plans used for the collection of property-casualty insurance statistical data. Regulators in the various jurisdictions are charged with ensuring that rates meet statutory standards – that rates are not inadequate, excessive, or unfairly discriminatory. One of the tools the regulators use to fulfill this function is the collection of data by line of insurance by statistical agents that aggregate the data and report it to regulators. A statistical agent is an organization that helps insurers satisfy legal requirements for reporting data to regulators. The statistical agent processes data submitted by insurers, performs data quality checks on the data, consolidates the data across insurers, and provides aggregate data compilations to state insurance departments on the behalf of the insurers. The well-known statistical agents in the United States are:

- The four that collect data for the major property/casualty lines of insurance, except workers compensation and health. These include the American Association of Insurance Services ("AAIS"), the ISO Data, Inc.<sup>TM</sup> (a wholly owned subsidiary of Insurance Services Office, Inc. or ISO), the Independent Statistical Services ("ISS"), and National Independent Statistical Services ("NISS").
- For workers compensation, the dominant statistical agent is the National Council on

### *Actuarial IQ*

Compensation Insurance (“NCCI”). In some US jurisdictions, workers compensation data is collected by an independent state bureau such as the New York Compensation Insurance Rating Board (“NYCIRB”).

Other statistical agents exist in the United States for more specialized lines of business such crop-hail (National Crop Insurance Services) and surety (Surety & Fidelity Association of America) insurance. In addition there are some state-specific/line-of-insurance-specific agencies that collect industry data. An example of this is the Texas Insurance Checking Office (“TICO”) which collects data for private passenger automobile, residential property, and farm and ranch insurance in Texas under Texas Department of Insurance (“TDI”) statistical plans.

Among these statistical agents, numerous statistical plans have been developed in each of the US jurisdictions. Statistical plans also exist outside of America. In general, the statistical plans are organized around one or more lines of insurance. For example, the ISO has three statistical plans<sup>1</sup> – the personal auto statistical plan (“PASP”), the personal lines statistical plan (other than auto) (“PLSP(OTA)”) and the commercial statistical plan (“CSP”). Each of these plans then has subparts or modules devoted to a particular line of insurance. For workers compensation, the underwriting experience (premiums and losses) is collected through the unit statistical plan (“USP”). Additional unique data collection requirements exist for workers compensation. For a more complete discussion of workers compensation see the study note “NCCI Data Collection Calls and Statistical Plans” by Richard Moncher [10].

In general, the statistical plans contain information or metadata – general reporting requirements and specific, detailed definitions for each data element – that describe the information to be collected. In the sections below, these items are explained further, followed by an example of these instructions and definitions excerpted from the homeowners module of the ISO personal lines statistical plan (other than auto).

#### **2.4.1 Reporting instructions**

Reporting instructions describe the overall scope of the plan such as:

- To which jurisdictions the plan applies,
- To which lines of business the plan applies,

---

<sup>1</sup> ISO also has separate plans for those companies with very limited market share in a line of insurance.



### *Actuarial IQ*

- Instructions on specific situations such as mid-term endorsements to policies and cancellations.

#### **1. GENERAL REPORTING REQUIREMENTS**

##### **A. Premiums**

Premiums must be reported separately for each policy and each unique set of codes in the Coding Section of this module.

When a policy insures more than one dwelling, each dwelling must be reported separately.

When Water Back-Up Damage coverage is attached to a policy, this coverage must be reported separately.

When a policy includes additional coverage which requires coding under a separate module of this Plan, the premium and amount of insurance reported under this module must **not** be increased.

*Table 2.4.1*

#### **2.4.2 Data element definitions**

Each element to be collected on the premium and loss records needs to be defined. In some cases the same data elements are collected on both the premium and loss records. These definitions cover multiple dimensions, including:

- A text description of the element to be collected,
- Field length or field position on the record,

##### **2. Transaction Type Code (Field: Position 5)**

Report the appropriate Transaction Type Code.

- Valid codes or attributes for the data element,

*Actuarial IO*

**TRANSACTION TYPE CODES**

DESCRIPTION	CODE
<b>PREMIUM TRANSACTIONS</b>	
Fully Coded (including "Exception" coded)	1
Limited Coded	2
<b>LOSS TRANSACTIONS</b>	
Paid Losses	6
Outstanding Losses	7
Paid Allocated Loss Adjustment Expense	8+
Outstanding Allocated Loss Adjustment Expense	9+
Salvage (Paid Losses)	4
Subrogation (Paid Losses)	5
+ Applicable to Section II (Liability) Losses only.	

- Record layouts that show the exact position and field length on the statistical plan record,
- Examples of coding and interpretations of the coding,
- Due dates for reporting to the statistical agent,
- Quality requirements.

Quality requirements for the submission would address the error tolerances that may be allowed.

For more information on statistical plans in the United States, the reader should refer to "Statistical Plans for Property/Casualty Insurers," by Virginia R. Prevosto [11], published in the 1997 Casualty Actuarial Society Discussion Paper Program and the study notes by Richard Moncher and Virginia R. Prevosto on the NCCI [10] and the ISO [12] statistical plans, respectively.

### 3. TECHNIQUES AND APPLICATIONS

Section 2 introduced key concepts of information quality. In this section, we present procedures and processes designed to improve information quality.

#### 3.1 Exploratory Data Analysis

A common approach to detecting data quality problems in a dataset is to perform a preliminary screening of the data elements. These data elements are treated as variables for the purpose of statistical analysis. Exploratory data analysis ("EDA") is a family of techniques that use graphs and descriptive statistics to explore the structure of a dataset and to identify outliers. (Data errors are often found by detecting outliers and then investigating the outliers for validity.) These techniques were pioneered and the practice given its name by John Tukey (see "exploratory data analysis" at [www.wikipedia.org](http://www.wikipedia.org)). These techniques are widely accepted in the statistical community as a key

## *Actuarial IQ*

activity within any statistical project, and they are widely implemented in statistical software.

Data quality problems can take several forms including:

- **Missing data and null values**, which impair the analyst's ability to use the affected variables and may render some variables useless for analyses,
- **Data errors** such as a paid amount of \$1,000 coded as \$1,000,000 or the state NY coded as NJ,
- **Default values** may be coded rather than actual values (e.g., for convenience), and
- **Duplicate transactions**: it is not uncommon for duplicates of the same claim, same transaction, etc. to be in a database.

Being mindful of the sources of data errors, one can detect, remediate, and most importantly, prevent them. Dasu and Johnson [7], whose book on data quality and data cleaning is considered a key reference by data mining professionals, detail many mishaps affecting data that create quality problems. Some of the sources of data quality problems are: unreported changes in layout, unreported changes in measurement, and temporary reversion to defaults, missing values, inappropriate default values, and gaps in time series.

The following subsections introduce several EDA techniques to deal with data quality issues in a given dataset. For more information, see Francis [4] or Dasu and Johnson [7].

### **3.1.1 Data cubes**

A data cube is a one-way or multiway summarization of key statistics for the variable(s). Cross-tabulations and pivot tables are examples of data cubes. For instance cross-tabulations or two-way tabulations of the frequencies for two variables are widely used in statistics and most statistical software such as SAS, S-PLUS, SPSS, and Access have the capability of quickly producing cross-tabulations.

For example, one can tabulate the frequency of records in the data containing each value of a categorical variable. Table 3.1.1 displays the frequencies of injuries for each of the 6 injury codes in a Massachusetts Private Passenger Auto database.<sup>1</sup> The table was created using Microsoft Excel's pivot table capability. Note that there are two codes where only a small number of records contain

---

<sup>1</sup> The data was supplied by the Automobile Insurers Bureau of Massachusetts and is from a database used to do fraud research.

### Actuarial IQ

the code. The results from pivot table summaries need to be compared to a document defining which codes are valid values for the data element. These tabulations can be performed over multiple dimensions at once, although it is most common to perform one dimensional (variable by variable) frequency analysis.

Massachusetts Auto PIP	
Injury Type Code	Count of Injury Type Code
1	793
2	197
3	2
4	250
5	151
6	7
<b>Grand Total</b>	<b>1400</b>

Table 3.1.1

#### 3.1.2 Identifying missing data

As noted by Francis [4], missing data is the rule rather than the exception in large insurance databases. Missing data complicates an analysis by reducing the number of data records with completely valid information. At a minimum, the uncertainty about parameter estimates will be increased, even when measures can be taken to adjust the data elements containing missing values. It is not uncommon for the majority of data records to be missing data on variables that are presumably in the database and available to the analyst. If a sufficient percentage of records on a given variable are missing values, that variable may have to be discarded from the analysis. In some extreme circumstances, the missing data problem may be so severe that an analysis cannot be undertaken. Tabulations of missing values should be compiled for each variable in the database.

Analysts must also be alert to missing values they create by their data manipulations. For instance, division by zero will create a missing or not available value that can affect further analyses if not detected. Most statistical software produces a log which records the history of calculations completed and their results. Cody [13] recommends reviewing the logs of the statistical software for statements that missing values are being created as a result of transformations performed.

### Actuarial IQ

Data cubes can be used in the detection of missing values and in screening categorical variables for data glitches ([7] page 74). Table 3.1.2 presents an example of a report that can be produced within most statistical packages. The report displays number of valid, invalid, and missing records for each variable specified:

	<b>Age</b>	<b>Model Year</b>	<b>Incurred Losses</b>	<b>Gender</b>	<b>Marital Status</b>
<b>Valid</b>	41,000	35,000	50,000	45,000	46,000
<b>Invalid</b>	100	1,000	-	500	1,200
<b>Missing</b>	9,000	15,000	-	5,000	4,000

Table 3.1.2

Note that it is not uncommon for missing values to be recorded as blanks. This situation will not be detected by procedures summarizing missing values. However, procedures used to tabulate all the values of a variable (e.g., data cubes, Microsoft Excel's AutoFilter) can be used to summarize the number of blanks on these variables. This is shown in Table 3.1.3:

<b>Value</b>	<b>Gender</b>
M	25,000
F	20,000
	5,000
<b>Total</b>	<b>50,000</b>

Table 3.1.3

Descriptive statistics can also be used to identify the presence of null values in numeric data.

### 3.1.3 Descriptive statistics

Descriptive statistics include such statistics as the mean, median, minimum, maximum, and standard deviation. Table 3.1.4 displays descriptive statistics, produced with the Microsoft Excel Analysis ToolPak, for an illustrative sample of general liability claims. The descriptive statistics summarize key information about the paid allocated expenses in the data. Looking at the minimum and maximum values can quickly inform us as to whether any values appear to be outliers or to have

*Actuarial IQ*

unusual values. In this example, the minimum paid expense is a negative value. The table also indicates that the second smallest value is also negative. Both of these numbers indicate data records that may need to be reviewed further before using in any analysis.

<b>Allocated Loss Adjustment Expenses</b>	
Mean	1,323
Standard Error	252
Median	611
Mode	0
Standard Deviation	8,217
Sample Variance	67,513,031
Kurtosis	207
Skewness	13
Range	170,668
Minimum	(19)
Maximum	170,649
Sum	1,411,246
Count	1,067
Largest(2)	99,206
Smallest(2)	(11)

*Table 3.1.4*

**3.1.4 Box and whisker plots**

A box and whisker plot is a one dimensional visualization of the distribution of a variable. The box plot, a predecessor of the box and whisker plot, can be programmed into Microsoft Excel. It displays a 5-point summary of a variable's distribution. The 5 points are: minimum, 25th percentile, median, 75th percentile, and maximum. A box is placed around the edges encompassing the 25th through 75th percentiles and lines extend from the box to the minimum and maximum values. The box and whisker plot modifies the box plot by displaying lines from the box to a specified distance (e.g., two standard deviations from the mean) from the box and by individually displaying

## Actuarial IQ

observations outside these lines.

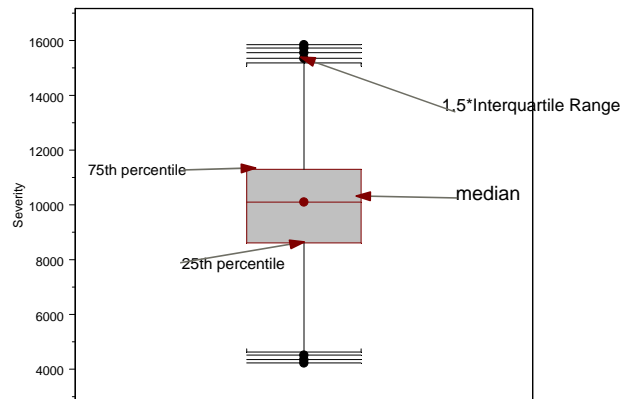


Fig. 3.1.1

Figure 3.1.1 displays a box and whisker plot. The top and bottom of the box are defined by the 75th and 25th percentiles of the distribution plotted. A line through the middle of the box denotes the 50th percentile (i.e., median) value. The width of the box carries no meaning. Lines extend from both the top and bottom of the box. These lines are referred to as the whiskers. For this graph, the lines denote the points 1.5 interquartile ranges<sup>1</sup> above and below the box edges. Points beyond this boundary are individually displayed (the circles with lines through them). These points may be considered outliers; they depict data records that the analyst might want to investigate.

Figure 3.1.2 displays the box and whisker plot for data containing an intentionally introduced error (the first number was replaced with ten times its value):

---

<sup>1</sup> The interquartile range is the difference between the 75th and 25th percentile

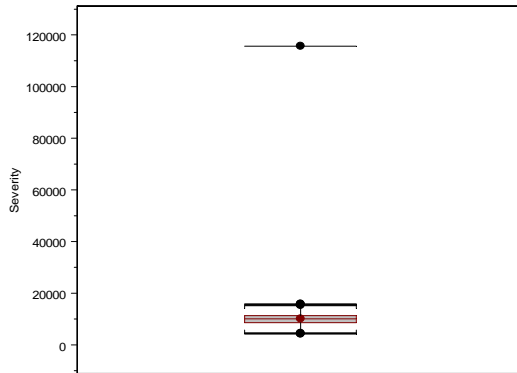


Fig. 3.1.2

In this paper we provide only a basic introduction to the methods of exploratory data analysis. A number of excellent references are available on this topic for those wanting a more thorough exposure to the topic. Hartwig and Dearing [14] provide an easy-to-understand introduction to the methods of exploratory data analysis, and Dasu and Johnson [7] introduce EDA within the context of its application to data cleaning.

### 3.2 Auditing Data

Whereas EDA cleans a dataset, auditing influences the process that generates the data. As such, auditing for data quality is a tool to help both assess and monitor data quality. While ASOP No. 23 does not require actuaries to audit data ([2], sections 1.2 and 3.6), knowing how audits are conducted can improve actuarial practice in at least two ways:

- First, it produces a more informed basis to assess what kind of reliance should be placed on audited versus unaudited data, and
- Second, the procedures and concepts used in auditing can be applied to resolve data issues without having to do a full-scale audit.

The main idea of data auditing is to compare the data intended for use to its original source(s), such as policy applications or notices of loss. This is done using both top-down and bottom-up approaches. The top-down approach is reconciliation: checking that totals from one source match



### *Actuarial IQ*

the totals from another (usually more reliable) source. These totals are usually dollars, but counts and records can also be reconciled. Auditors will often not only do their own reconciliations, but also review an organization's reconciliation procedures. Obviously, making sure totals match is one way to assess the reasonableness and comprehensiveness of a data set, so reconciliation can be useful to actuaries both on its own as well as when it is part of an audit.

The bottom-up approach takes a sample of input records and follows them through all the processing to the final report. Any good sampling textbook should provide the theoretic basis to address sampling issues. One such book is *Elementary Survey Sampling* by Scheaffer, Mendenhall and Ott [15]. Defining accuracy ratios can make results comparable from one audit to the next. An example of an accuracy ratio is the number of occurrences a given data element is correct divided by the number of occurrences reviewed. The number, type and rigor of these statistics are determined by the intended use of the data. Note that ratios of record counts can provide different information than dollar ratios, so sometimes it can be helpful to include both for phenomena of particular interest.

The following summary of major steps in a data quality audit is based on ISO's *Strength in Numbers* pamphlet [16]:

1. **Test the preparation of the data:** Measure how correctly and completely data is coded. Also measure how current it is.
2. **Test the data entry and data transfers:** How much of the data reaches its final destination intact? How much of this takes place in an acceptable period of time?
3. **Test the program controls:** Measure the extent that "only authorized data is entered for processing and that data is processed completely, accurately, and in a controlled environment" ([16], page 6). A controlled processing environment will have procedures and checks to ensure that computer jobs are run in the right order, computer jobs are not accidentally run twice, total outputs equal total inputs, users are aware when software programs end abnormally and so forth.
4. **Test the output controls:** Measure the accuracy, timeliness and correct distribution of reports.
5. **Test error procedures:** Measure the extent that the system detects and corrects errors in a timely manner.

"Performing periodic [data] audits will indicate:

### *Actuarial IQ*

- The accuracy and completeness of the picture... [which the] data gives of the insured risks,
- The timeliness of data processing,
- Any differences between statistical and other insurance data to be reconciled,
- Problems or potential problems related to collecting, coding, and reporting your data” ([16], page 3).

More information on data audits can be found in the Insurance Services Office’s (ISO) *Quality of Data Audit Guide* [17]. Accounting professional organizations may also publish information on auditing.

### **3.3 Information Quality in Models**

We now turn our attention from a strict focus on data to broader information quality issues. With the broader perspective of information quality, it becomes clear that actuaries are active participants in the data life cycle of an organization. They take data as an input, analyze it, and produce output that is used in decision making. The quality of this analytical step is thus a crucial contributor to the overall quality of information used in the company.

Analysis is about building models to explain or predict phenomena. As such, analysis behaves like software in some respects: it is a set of steps to manipulate data. Software quality is a function of design, implementation and testing. Good design decisions may improve not only the functionality and usefulness of the application, but also simplify quality assessment and ensure easy modifications and updates. Testing, especially if integrated with implementation, may improve the quality of the resulting software product. Any actuary involved in design or modification of spreadsheets and other analytical applications will clearly benefit from knowing the main principles of good software design.

#### **3.3.1 Quality design**

To use a manufacturing metaphor, the quality of actuarial work products depends on the choice and quality of the tools actuaries use to process incoming data. The tools should be good and suitable: the actuarial methods used should be appropriate for the data at hand. Quality of the method relies heavily on 1) model selection and validation, 2) model’s parameters estimation and 3) model’s verification (see [18], chapter 2.9 for detailed explanation). To understand the difference between validation and verification one should consider two questions: “did I use the right model?”

## Actuarial IQ

versus “did I use the model right?”

Some actuarial methods are designed only for data with particular properties, i.e., it is assumed that the data satisfy some preliminary conditions. Thus, before applying an actuarial method to a set of data, it would be prudent to test the method’s assumptions on that specific dataset.

Failed assumptions may either indicate inappropriateness of this particular method or uncover hidden data problems. In this sense, assumption testing may also serve as data quality tests.

An aspect of an analysis’s quality is **model performance**. Many actuarial methods for pricing and reserving predict some events that can be observed. Comparison of predicted and actual values may lead to method improvements, recalibration, or even rejection. Note that any of these outcomes leads to improvements in the model’s quality.

### 3.3.2 Implementation (software) quality

In the actuarial toolbox, the spreadsheet occupies a special, quite dominant place. However, while tomes are written about C++ or VBA programming techniques and SQL optimizations, it is very hard to find practical advice on effective spreadsheet design.

The deceptive simplicity of a spreadsheet’s grid makes many users think of a spreadsheet as a single user’s ad hoc advanced calculator that can also chart and print. Users don’t even think there could be design recommendations and do not look for them. Indeed, a spreadsheet created by a user for a single use is quite disposable, but if there are multiple users or repetitive usage, spreadsheets become applications and should be treated as such.

An application is a part of the data flow of an organization and therefore subject to quality control. It has to be well-designed and documented to simplify 1) usage, 2) testing and 3) modification. What can be done on this front? Experience shows that one of the most effective techniques is separation of data and algorithms. Calculations (formulae and VBA code) should be stored in one file or spreadsheet tab (called the **template**), while data should be loaded from an external source (e.g., spreadsheet tab, file, or database). In practice, actuaries usually realize that *input* data like loss triangles, premiums and industry factors do not belong in a calculation template. What they rarely realize is that **output** data such as predicted ultimates or fitted distribution parameters do not belong in the template either: results have to be stored outside just like inputs.

## Actuarial IQ

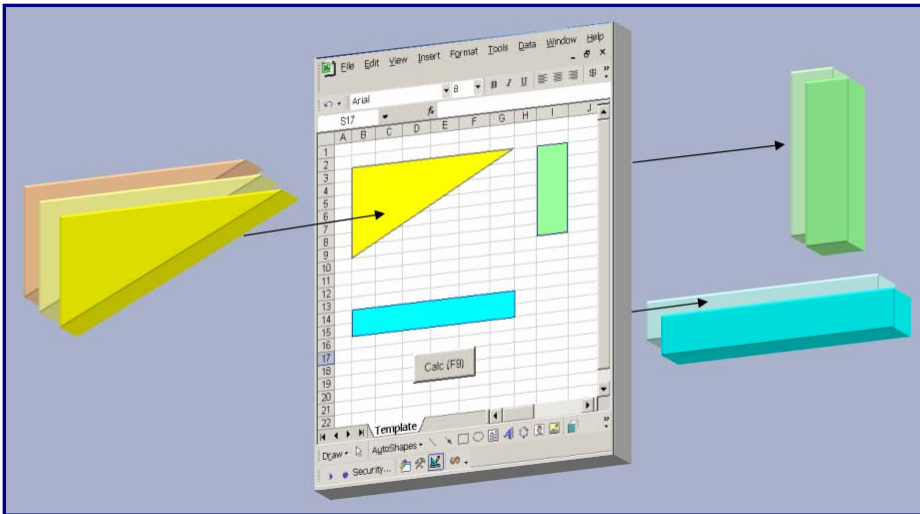


Fig. 3.3.1

Such a setup 1) brings consistency to calculations, 2) simplifies housekeeping, 3) allows versioning, and (combined with access rights) 4) improves control over modifications. An additional benefit for quality pursuers is the fact that separation of data and algorithms facilitates checking calculations with different data samples, thus enormously improving the quality of testing.

Another useful technique which extends the notion of separation is **layering**. Both users and designers may benefit when data (input placeholders), reconciliation, calculation, user interface (scenarios, selections, and assumptions), and presentation (results and charts) layers are located on separate spreadsheet tabs or worksheets. Such a layout not only simplifies navigation, it also shortens the learning curve for users, allows designers to better understand workflow and provides better documentation.

## Actuarial IQ

For example, below is a hypothetical layering scheme for a rate review application:

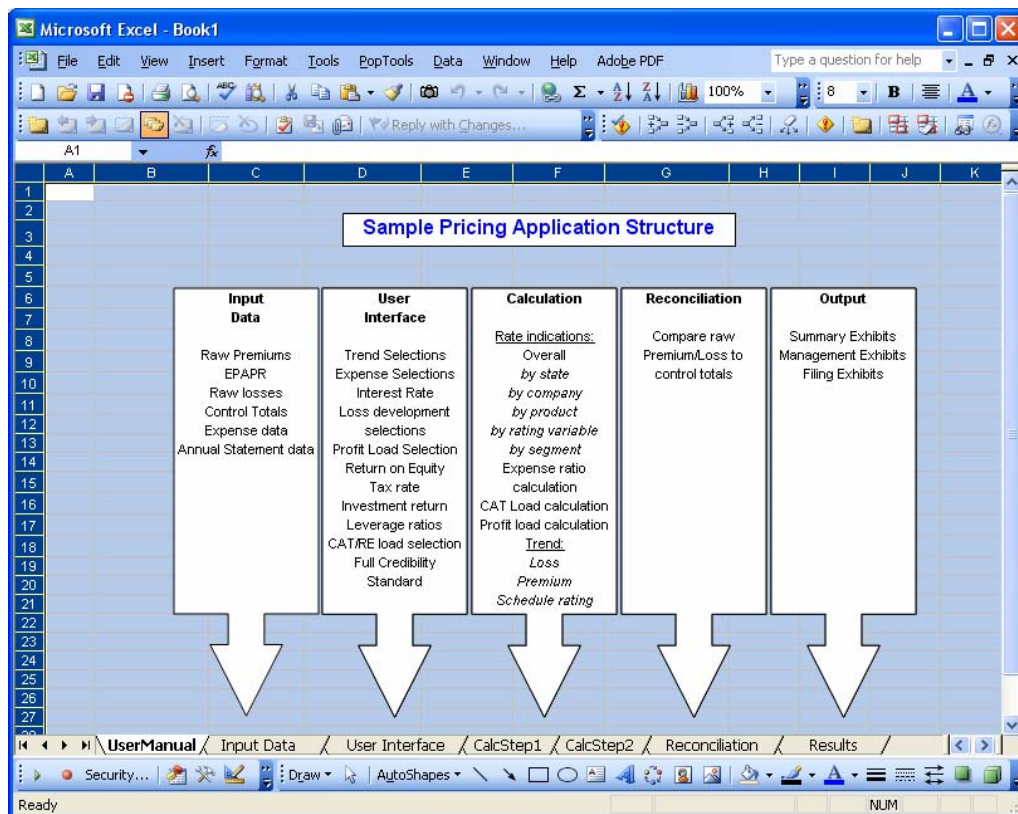


Fig. 3.3.2

Good documentation is a centerpiece of quality design. Every application should have a "header" identifying inputs, outputs, purpose, and contacts. Spreadsheets also generally provide adequate facilities for file versioning, VBA code commentaries and cell comments. As noted in [29], one can use built-in document properties or create custom ones and link them to cells inside spreadsheets. The trick is to remember to update documentation with **every** modification or improvement made to a template.

### 3.3.3 Testing

Testing is critical to the good design of successful applications. Indeed without testing, a spreadsheet (or query or notebook) may never become an application (i.e., a reusable tool) – there

### *Actuarial IQ*

would be no assurance that it could handle different situations correctly. Once an application becomes successful, thus widely used, testing becomes even more important.

The majority of books on testing deal with higher languages (C++ and the like): very few publications give practical advice on spreadsheet development. Some of the main thoughts from these books, however, will be of interest to actuaries.

Testing, according to Edward Kit's *Software Testing in the Real World* [19] should start with **specifications**, end with **final product evaluation**, and should be performed by an independent party. The main testing techniques are verification and validation, i.e., checking the code and examining final product outcomes. In the actuarial paradigm, examples of the final product could be Excel spreadsheets, Mathematica notebooks or Oracle stored procedures. Similarly, specifications could be a reserve test or pricing method, and the "code" could be formulae in cells, VBA subroutines or SQL statements.

Some of Kit's verification testing techniques can be applied to spreadsheets. For example, checking programming code against a list of common mistakes applies mostly to those who use Visual Basic. However, the recent addition of "Formula Evaluation" and "Watch Window" tools make Excel much friendlier for debugging. These tools allow users to validate formulae placed in cells by displaying results of all intermediate calculations and by monitoring values in "watched" cells. They bring debugging power previously available behind the scenes (to VBA coders) to the forefront (to cell formulae designers).

The most common testing technique is validation: checking that calculations produce expected results for different (and not necessarily correct) data. Validation treats an algorithm as a black box, feeding it with different inputs and observing results. On the one hand, validation checks algorithm limitations (e.g., whether it can work with negative amounts, strings, missing values). On the other hand, it also checks the accuracy of calculations on the datasets with known results. In either case, validation feeds the algorithm with different datasets so this process may benefit from separating data from calculations as described above. Indeed, using Excel's "scenarios" functionality, one can create a library of test datasets and recall them by selecting the corresponding scenario. Similarly, with assumption sets, if they are separated from the rest of the application, then it is easier to test algorithm results against various assumption sets.

Testing is very repetitive by nature, so it makes sense to accumulate testing tools for future reuse. It is very easy to build libraries of "bad" and "benchmark" datasets for testing actuarial methods. Testing routines and functions could also be accumulated into a library available to every tester or

designer.

Kit's suggestions that 1) testing should be an integral part of development, and 2) testing should be performed by people outside of the development team should definitely be implemented for any application that is part of a company's data flow.

In conclusion, the keys to quality models are:

- **Good design,**
- **Accurate implementation,** and
- **Thorough testing** of everything: from methods and assumptions to auditing spreadsheet formulae and query results.

### 3.4 Data Presentation (Reports) Quality

If data reaching the presentation stage are accurate, reasonable, complete, and have been analyzed in a high quality model, what can go wrong with the presentation? Unfortunately, a lot:

- Data can be **mislabeled** or incompletely labeled. "Total Loss" may refer to "loss net of recoveries" or "loss and ALAE net of reinsurance" or "unlimited loss before deductibles."
- Data can be **incorrectly related** to other information, producing wrong calculations. Date mismatches in losses and premiums may produce erroneous loss ratios.
- Data may be arranged in such a way that the essential information it is supposed to convey may be **overlooked**. A good report should emphasize the message and guide the reader to the most important information.
- Data can be **misinterpreted** and the message they deliver may be misunderstood. It is not unusual to witness confusion and misuse of such notions as reserve range, expected shortfall, confidence interval, or risk transfer.

To avoid costly mistakes from wrong decisions based on poor data presentation, crucial reports should be prepared with the involvement of someone who understands the data (e.g., an actuary). Therefore actuaries should be familiar with some tools and techniques to improve the quality of reports.

There is an enormous variety of reporting tools of different capabilities and complexities, but the

most versatile, familiar, and readily available is a spreadsheet. Modern spreadsheets provide enough features for building quality reports.

[Appendix A](#) contains some practical solutions to address:

- Unambiguous labeling,
- Consistent calculations,
- Focusing attention, and
- Minimizing misinterpretation.

### 3.5 Measuring Data Quality

The first four subsections of section 3 addressed individual steps of the data life cycle ([Fig. 2.0.1](#)). The remaining three subsections of section 3 address general issues that apply to the entire life cycle.

Many data quality authors (e.g., Redman [20], Dasu and Johnson [7]) are strong proponents of measuring data quality. These authors believe that in order to motivate improvements in data quality, it is imperative that data quality be measured, even when the measures are somewhat subjective. The following is a brief introduction.

A key concept in measuring data quality is the data's "conformance to constraints." Dasu and Johnson describe both static and dynamic constraints ([7], page 131). Static constraints relate to properties of the data itself, such as its validity. For example, for the constraint "value should be present and be only from a fixed list of correct values," the corresponding measure would be "the number or percentage of missing or invalid values in a variable." Dynamic constraints relate to the processes used in the flow of data from its source to the different databases. Examples of dynamic constraints would be 1) "a reserve change is added to prior cumulative reserves (not to cumulative losses)" and 2) "incurred losses can never be less than the sum of the amount paid." Thus, dynamic constraints capture business rules.

Some of the key data quality measures recommended by Dasu and Johnson ([7], pages 131 - 134) are:

1. **Extent of automation:** sample some transactions, follow them through the database creation process, and tabulate the number of manual interventions
2. **Successful completion of end-to-end process:** the number of processes that have the outcome they are expected to have. For instance, a sample of claims can be followed through



### *Actuarial IQ*

closure and it can be determined how soon after the final payment is made that the claim is closed.

3. **Impact on analyses:** measure how many errors in analyses result from errors in the data. Using sampling, the number of analyses adversely affected by data quality problems can be tabulated. Both the frequency and severity of the problems should be measured.
4. **Accessibility:** how easily can the data be accessed? For example, the time between a request for data and access to the data can be measured.
5. **Interpretability:** how understandable is the data? The quality of the metadata determines how interpretable the data is to users. The interpretability of data should be based on 1) the availability of metadata and 2) the extent to which the data adheres to the definitions in the metadata.
6. **Conformance to business rules:** how well does the data adhere to insurance business rules? For instance, how often are negative paid losses recorded in lines where losses should always be positive (i.e., no salvage and subrogation)?
7. **Conformance to structure:** Select important constraints that the data must follow and measure how well the data conforms to those constraints.
8. **Accuracy:** what proportion of the data contains valid values? This can be expensive to measure, so measures based on samples or based on proxies such as complaints or surveys are recommended.
9. **Consistency:** how often do databases at different points in time or data in different databases and tables within the company agree with one another?
10. **Uniqueness:** certain data elements should only have one observation in the dataset. For instance, a claimant level database should have only one record for each claimant. Measuring this amounts to identifying duplicates, which is discussed in section [3.1](#).
11. **Timeliness:** how often is the data updated and what proportion of it is available on schedule? Dasu and Johnson also mention that data should have an accurate time stamp.
12. **Completeness:** to what extent does the data contain all the data elements relevant to the analyses and reports a company undertakes? Thus, a database that is accurate and timely may be of low quality because it contains only a few variables or only a few years of history.

The different metrics are weighted together into an overall data quality index using business

### Actuarial IQ

considerations and the analysts' goals to develop weights. For example, if improvement in the database itself is considered most important, the static measures (e.g., accuracy, completeness, timeliness) might be given greater weight than dynamic measures (e.g., successful completion of end-to-end processes).

Table 3.5.1 illustrates a simple data quality measurement for a company beginning a data quality initiative (i.e., these are simple not comprehensive measures). All audits, sample findings, and survey results have been converted into scores between one and ten, where one is low and ten is high. Weights have been assigned subjectively.

Measure	Score	Weight
Extent of Automation	4	0.1
Accuracy	3	0.2
Glitches in Analyses	3	0.2
Completeness	6	0.2
Interpretability	7	0.3
<b>Total</b>	4.9	

Table 3.5.1

The data quality variables can be measured periodically after a data quality initiative is undertaken. Over time, the score should improve. Dasu and Johnson note that when used as a tool for quality improvement, it is the **direction** of the data quality measure over time that is of interest ([7], page 134). A number of other authors (e.g., Loshin [21], Redman [20]) offer additional advice as well as some alternative measures of data quality.

### 3.6 Data Quality Improvement Strategies

Two strategies to improve data quality are data cleansing and re-engineering. The objective of the first strategy is to take defective data and correct, reformat, consolidate, and standardize it so that standards are met and maximum value can be achieved from the data. The objective of re-engineering is to proactively eliminate the causes of poor quality data by changing processes. Note that data cleansing is an ongoing cost-added process. The overall objective is to eliminate the need to perform error correction, but the most effective approach is to couple data cleansing with re-engineering. While the former attacks specific defects in the data, the latter focuses on the root

## Actuarial IQ

causes of the defects. Note, however, that the profiling process advocated by Olsen would often require process changes such as a full-time team dedicated to data quality, as well as recommended changes resulting from the team's investigations.

What follows are some alternative strategies based on *Data Quality, the Accuracy Dimension* [22] and *Improving Data Warehouse and Business Information Quality* [23]. Note that the costs of applying a particular data improvement technique need to be weighed against the benefits. For example, it may be too expensive to correct lost, missing, or incorrect data if the source data is not readily accessible.

### 3.6.1 Data cleansing

The objective of data cleansing is to improve the data quality in existing files to maximize its value and to minimize the cost due to poor quality information. This includes correcting wrong data, standardizing nonstandard data values, filling in for missing data, and consolidating duplicate occurrences.

In *Data Quality, the Accuracy Dimension*, Olsen introduces a proactive data quality assurance program for detecting and addressing data inaccuracies throughout the many databases used by an organization. The system has two basic approaches, denoted **inside out** (a ground up, detailed, data-dependent approach) and **outside-in** (essentially an outcomes-based, business-driven approach

The **inside-out** approach can be summarized as follows:

1. Build the organization's metadata to have a complete and correct set of rules that define data accuracy for a particular dataset,
2. Gather inaccurate data evidence, i.e., collect facts about data shortcomings,
3. Aggregate the inaccurate data evidence into issues,
4. Analyze the issues to determine the external impact,
5. Set the priority of each issue based on its external impact, and then
6. Rectify the issues.

The **inside-out** approach can detect many data inaccuracies that are routinely missed by users working with aggregated data.

In contrast, the **outside-in** method "identifies facts that suggest that data quality problems are having an impact on the business" ([22], page 73). Such facts might be reworks, returned merchandise, or customer complaints, for example. The facts are then evaluated to determine the

### *Actuarial IQ*

degree of culpability attributable to defects in the data. The advantage of the outside-in approach is that it automatically focuses on issues that have a noticeable external impact. One of its disadvantages is that it may miss issues with still larger, but unnoticed, impacts. That is, by using the outside-in approach alone, only those data quality problems that have already manifested as business issues will be detected. It is also less likely to discover the full scope of related issues that interact to produce the observed impact. This approach requires the participation of business analysts along with a dedicated data quality analyst. Olsen's recommendation is that both approaches need to be applied.

The **outside-in** approach can be summarized as follows:

1. Identify information indicating a data quality problem
  - a. investigate customer complaints
  - b. investigate business user complaints
  - c. interview users of data to assess their level of satisfaction
2. Determine the extent to which data accuracy issues contribute to the problem.

The following chart summarizes the two approaches to data quality improvement programs:

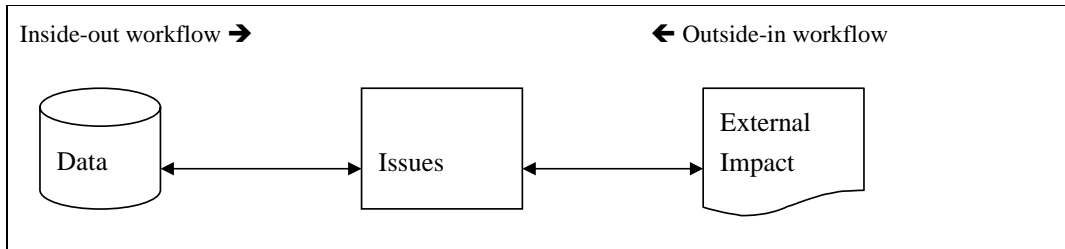


Fig. 3.6.1

For comparison, English's general steps for data cleansing are outlined in [Appendix B](#).

### 3.6.2 Re-engineering, a.k.a. process improvements

This strategy improves business processes by eliminating the causes of poor quality data.<sup>1</sup> It is a proactive method that analyzes the cause of problems and eliminates them. The rationale is that it is much less expensive in the longer term to prevent errors than to repeatedly screen for them and repair them in different databases every period. Therefore the long-term solution to data quality is not to fix the data but to fix the processes that produce the defective data. Data cleansing fixes the problems after they have occurred, whereas process improvements eliminate the causes. English ([23], pp. 285-310) also provides a data defect prevention approach, which is described in [Appendix C](#).

## 3.7 Actuarial Data Management

Actuaries are among the most prominent users of an organization's data. Thus, they have a natural vested interest in ensuring that their organization's data is of the highest quality. Over time, data management has evolved as a unique specialty within the actuarial community. In some insurers, though, the role of data manager is not held by an actuary, but by a person trained in this field that understands the data needs of the actuary and other users of data.

In performing any analysis, the actuary must consider many things, but the starting point for the analysis is the historical premium, exposure, loss and expense experience for the type of insurance under review. "This experience is relevant if it provides a basis for developing a reasonable indication of the future. Other relevant data may supplement historical experience. These other

---

<sup>1</sup> Note that the data cleansing process as describes by Olsen [22] and Redman [20] is intended to also affect the processes generating the errors once the errors are uncovered and thus may entail some re-engineering.

### *Actuarial IQ*

data may be external to the company or to the insurance industry and may indicate the general direction of trends in insurance claim costs, claim frequencies, expenses, and premiums” ([24], page 7).

The data management actuary provides a bridge between those who are responsible for the collection and repository of the organization's data and the pricing or reserving actuary who will use the data in analyses. Thus, two critical areas for the actuarial data manager are:

- The appropriateness of the collected data elements for the analysis to be done, and
- The quality of the collected statistical experience for the analysis to be done.

Some of the activities performed by the data management actuary include:

- Reviewing the various data compilations for reasonableness. This includes comparing the current data compilation against the previous data compilation to ensure that the change in the data for overlapping years is as expected. For example, the losses as of 24 months versus as of 12 months have grown as expected for the line of business under review.
- Reviewing the growth patterns by year within a compilation.
- Reviewing the distribution of data within a data element. For example, reviewing the written premium distribution by geographic location to make sure it accurately reflects the book of business in the compilation and that it has not been erroneously coded to one location.
- Ensuring that any definitional changes in the data elements are accounted for and notifying the actuary who will use the report of this situation.
- Reviewing the data compilation for completeness – that is, only the data that was supposed to be included is included and it is included only once.

As data management actuaries grow in responsibility, they should also take a more proactive role in understanding the data processing stream from source through transformations, data base repositories, and data extraction and compilation; ensuring that the organization is following data management best practices at every step in the process. Thus, they will have a complete understanding of the data that has been extracted and will ensure its proper use in analyses done by themselves or another actuary to whom they are supplying the data.

The insurance data management profession has established a set of guiding principles and best practices for data management [25]. Some of these key principles and practices are listed in bold

### *Actuarial IQ*

below, followed by further explanation and where appropriate, a cross-reference to ASOP No. 23 [2].

1. **Data must be fit for the intended business use.** This principle is in sync with ASOP No. 23 that states "For purposes of data quality, data are appropriate if they are suitable for the intended purpose of an analysis and relevant to the system or process being analyzed" ([2], page 2). Data should be collected in the level of detail (breadth and depth of the data elements) and at a level of quality that are sufficient for the intended applications or analyses to be performed.
2. **Data should be obtained from the authoritative and appropriate source.** Data should flow from the underlying business process, whether it is the underwriting and rating of the risk or other processes such as claim reserving, accounting of payments received or claim paid out, or litigation metrics. For example, insurance statistical data for a risk related to the premiums charged should be collected in a level of detail consistent with how the risk is underwritten and rated. That is, which data elements are collected and the depth of the detail (or attributes) within the data elements should be consistent with how the risk is underwritten or rated. The actuary using data received from others is required by ASOP No. 23 to "take into account the extent of any checking, verification, or auditing that has already been performed on the data, the purpose and nature of the assignment, and relevant constraints" ([2], page 4). It is also important that data be supplied by a source that understands the data. For example, detailed data regarding the nature of an injury should be supplied by the health care provider who understands the nature of the injury rather than a claims coder.
3. **Common data elements must have a single documented definition and be supported by documented business rules.** As ASOP No. 23 notes "The actuary should make a reasonable effort to determine the definition of each data element used in the analysis" ([2], page 4).
4. **Metadata must be readily available to all authorized users of the data.** The actuarial data manager should ensure that data, systems, and reporting mechanisms are designed and maintained in a manner that promotes good data management and data quality. This includes a robust, comprehensive business data dictionary that provides a clear, unambiguous definition of each data element that is consistent with the underlying business process.
5. **Data standards are key building blocks of data quality.** To promote consistency in the data collected, increase efficiency of the data collection process, and maximize utility of the data, organizations must foster the development and adoption of data standards and data quality standards. Industry standards must be consulted and reviewed before a new data element is

created.

6. **Data should have a steward** responsible for defining the data, identifying and enforcing the business rules, reconciling the data to the benchmark source, assuring completeness, and managing data quality.
7. **Data should be input only once and edited, validated, and corrected at the point of entry.** Data quality should be managed as close to the source as possible. This includes defining the data quality standards for the data to be collected. Processing steps between the data source and the data capture increase the likelihood there will be errors and often increase the cost of correcting those errors.
8. **Data should be captured and stored as informational values, not codes.** For example, if age of driver is a desired date element, the birth date of the driver should be captured and stored rather than the driver's age. By following this principle, misinterpretation of the data will be reduced, and serious errors in business decisions can be avoided. The data will also be more complete and more likely to be useful in answering unanticipated questions. Following this principle also facilitates reviews of the data for reasonableness and consistency.
9. **Data must be readily available to all appropriate users and protected against inappropriate access and use.** Insurance statistical data is the life blood of the property-casualty insurance industry and much of the data is considered a trade secret or is highly personal in nature (see 2.2.2). Data managers must balance access to data against inappropriate access or use. The actuarial data manager should ensure the actuaries' repository data base meets current and future business and analytical needs by partnering with the IT professionals in designing it.

For more information regarding data management best practices, see the Insurance Data Management Association website, <http://idma.org/productsDMBestPractices.htm>.

#### 4. CONCLUSIONS

Data quality is a core issue affecting the quality and usefulness of the actuarial work product. Data quality is often perceived as a mundane issue with less recognition and attention devoted to it than other issues, such as actuarial models and methodologies. However, data exists to fulfill a need: the need for optimal decisions. To the authors' knowledge, this is the first paper to provide a general introduction to data quality and data management directed specifically at actuaries since the CAS Committee on Data Management and Information White Paper of 1997.



#### **4.1 Pragmatically**

[Figure 2.0.1](#) outlines the steps in the insurance data life cycle, the kinds of errors that can occur in each, and references to relevant sections of this paper. As such, figure 2.0.1 forms a handy reference both to trace where a particular error may be occurring and which section of this paper may be most relevant.

Several tools to help actuaries improve their information quality are:

1. Exploratory data analysis to identify outliers and explore the structure of a dataset ([3.1](#)),
2. Improving the quality of actuarial models ([3.3](#)),
3. Improving actuarial presentations and reports ([3.4](#)),
4. Measuring data quality to track progress ([3.5](#)) and awareness of quality audits ([3.2](#))
5. Strategies to improve data quality ([3.6](#)), and
6. Guiding principles and best practices ([3.7](#)).

Each section has references to books and/or CAS papers for readers who need more information.

## 4.2 Conceptually

We began by drawing attention to the increased importance of data quality given easy access to an unprecedented level of detail and the proliferation of new tools and techniques to analyze such data. The actuarial frame of reference (2.1) was broadened beyond the scope of ASOP No. 23 in three ways:

1. **Data is a corporate asset that needs to be managed and actuaries can play a role.** Actuaries have the ability and motivation to influence the processes that give rise to the data they use (3.7)
2. **Data needs to be appropriate for all of its intended uses.** Actuaries have a unique role to play in achieving this goal here too: actuaries can expand their concerns for data beyond the analysis at hand. Finally
3. **Expansion of data quality principles (2.2 and 2.3)** to support these broader perspectives.
4. It should be noted that these expansions are those of the working party; not interpretations of the standard.

Data quality is not just about how data is coded: we have coined the phrase “information quality” to emphasize the impact of processes on the quality of the final product(s). Metadata (2.3), information about the data, is critical to actuaries correctly interpreting their data and the glue that holds an organization’s data structures together. Statistical plans (2.4) were introduced as a form of metadata. Data management best practices (3.7) embrace and support all of the above.

Ultimately, empowering actuaries to improve the quality of information in their organizations can increase the efficiency, effectiveness and impact of actuaries on their organizations by turning data into more useful information to make better decisions.

### Acknowledgment

The working party thanks our IDMA liaisons for their feedback and support throughout this project.

The working party also thanks the Insurance Services Office, Inc., for the use of excerpts from their homeowners module of the ISO personal lines statistical plan (other than auto).

### Supplementary Material

Code for creating Box Plots in Excel (described in section 3.1.4) can be found at [www.data-mines.com](http://www.data-mines.com).

Presentation template with live charts (described in Appendix A, section 4) can be downloaded from [www.casact.org/research/drmwp/DRM%20presentation.ppt](http://www.casact.org/research/drmwp/DRM%20presentation.ppt).

**Appendix A: Practical solutions for addressing some problems with presentation quality (expansion on section 3.4)**

**1. Practical solutions: unambiguous labeling**

Unambiguous labeling requires first that the label is consistent with the content and, second, that the label is descriptive enough to avoid ambiguity.

Consistency of labels and content can be achieved by examining every transformation<sup>1</sup>, transfer<sup>2</sup>, and calculation data goes through while keeping track of data sources and formulae applied to the data. Spreadsheets provide some assistance with this: a table created by importing external data keeps query information available for examining and editing. This SQL text helps to identify sources and clarify the nature of the extracted data. A spreadsheet’s ability to name ranges gives users an ability to create readable and, thus, traceable calculations. The next logical step in readability is to use labels within formulas. The “using labels in formulas” feature allows the user to create a quite traceable expression like “=Case Reserves + IBNR” using field names “Case Reserves” and “IBNR” in the formula for “Reserves.” Another useful facility in spreadsheets is “commenting:” a descriptive tag attached to an upper left corner of the triangle will “travel” with the data during copy and paste operations and will help the user to avoid obvious errors, such as making sure that paid losses

AY\Age		36	48	60
1994	\$ ...	107,847	\$ 115,288	\$ 124,592
1995	\$ Shape --> <b>Triangle</b>	110,271	\$ 112,562	
1996	\$ Amount--> <b>Losses</b>	104,029		
1997	\$ Cumulative- <b>True</b>			
1998	\$ 105,647			

wouldn’t end up in a calculation intended for claim counts.

The second type of labeling problem, which we will call “disambiguation of labels,” presents a different challenge. Readability and aesthetics considerations advocate short labels, while the need for quality and precision requires labels to be quite detailed and relatively long. The solution seems to be in hiding less necessary details until needed. The user would still need to be able to display the

<sup>1</sup> Data transformation step – edits, rearrangements and conversions from one format to another.

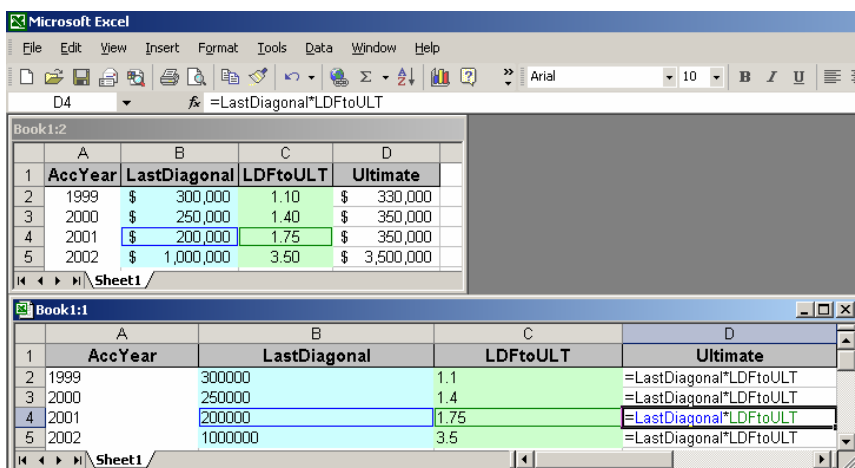
<sup>2</sup> Data transfer step – extraction from one system, transportation and upload to another system.

## Actuarial IQ

detailed labels on demand. Spreadsheet “comments” satisfy such design requirements: they are hidden until the computer’s mouse moves over a cell with a label in question. This technique, while convenient, is not very reliable because it doesn’t firmly relate short and long labels. More reliable, but much more involved, is SmartTag technology that allows a spreadsheet to recognize certain labels, lookup for their longer descriptions in the metadata table, and display long labels on demand. SmartTags may ensure enterprise-wide label consistency, but may not help in the creation of an ad-hoc report with the new labels. Another hide-display technique available to spreadsheet users is the use of an outline. Short labels can be placed on a higher outline level and additional (clarifying) meta-information can be placed on a lower outline level and collapsed. A collapsible outline view is a convenient arrangement for other meta-information regarding reports: for example, lists of formulae used in reports or lists of data sources and analysis methods.

### 2. Practical solutions: calculation consistency

It is very hard to prevent users from adding “doctors” to “hospital beds” to obtain “total exposure,” but some precautions could be made to prevent embarrassing mistakes. One can borrow



	A	B	C	D
1	<b>AccYear</b>	<b>LastDiagonal</b>	<b>LDFtoULT</b>	<b>Ultimate</b>
2	1999	\$ 300,000	1.10	\$ 330,000
3	2000	\$ 250,000	1.40	\$ 350,000
4	2001	\$ 200,000	1.75	\$ 350,000
5	2002	\$ 1,000,000	3.50	\$ 3,500,000

	A	B	C	D
1	<b>AccYear</b>	<b>LastDiagonal</b>	<b>LDFtoULT</b>	<b>Ultimate</b>
2	1999	300000	1.1	=LastDiagonal*LDFtoULT
3	2000	250000	1.4	=LastDiagonal*LDFtoULT
4	2001	200000	1.75	=LastDiagonal*LDFtoULT
5	2002	1000000	3.5	=LastDiagonal*LDFtoULT

an idea from programming languages that enforce so-called “strong typing”: every piece of data has a type associated with it and no operations between incongruent types are allowed. To mimic “strong typing” in a spreadsheet situation one has to keep “type” information associated with data elements, bring it (“type” information) to the report along with the data and use it for “type checking” in the formulae. For example, the formula for a loss ratio should first check that both the numerator and denominator belong to the same year *and the same kind of year* in order to avoid “underwriting year” vs. “accident year” mismatch. Spreadsheets don’t have built-in “typing”

## Actuarial IQ

enforcement tools; however, they provide mechanisms that may help avoid some simple errors. In a columnar report, one can write formulae with column labels rather than with nondescriptive cell references. Assuming that labels correspond to column content, this is a much more reliable way to refer to particular data. Additional information on this topic and implementation ideas can be found in [26].

More accurate solutions would involve storing results of the actuarial analysis in a well-designed relational database and creating reports from it. Assuming that database integrity is intact, the database engine would ensure proper relationships between data elements from different tables.

### 3. Practical solutions: focusing attention

There are many techniques for attracting a report reader's attention to important information. We will mention just three of them: adaptive reporting, visualization, and alarm systems. All three are within a reach of any spreadsheet user and can be used to improve the informational value of reports (see [27]).

**Adaptive or data-driven** are reports whose size, shape, and format adapts to the data. Placing these reports in an interactive environment such as a spreadsheet allows the user to interact dynamically with the report (effectively creating a whole family of reports rather than a single one), shaping it to the level of detail that suits the user.

A partial list of data-driven implementations found in spreadsheets includes:

- **Filtering** (reduces amount of data displayed).
- **Outlining** (hierarchically organizes data with an ability to hide and display data on different levels of the hierarchy).

1	2	3	4	5	A	B	C	D	E
	1				<b>Reinsured</b>	<b>LOB</b>	<b>State</b>	<b>AccYear</b>	<b>UltNetLoss</b>
	2				ABC	WC	NY	1996	1,712,201
	3				ABC	WC	NY	1997	1,730,918
	4				<b>ABC WC NY Total</b>				3,443,119
	5				ABC	WC	CT	1996	1,944,502
	6				ABC	WC	CT	1997	1,975,489
	7				<b>ABC WC CT Total</b>				3,919,991
	8				ABC	WC	NJ	1996	2,172,041
	9				ABC	WC	NJ	1997	2,227,708
	10				<b>ABC WC NJ Total</b>				4,399,750
	11				<b>ABC WC Total</b>				11,762,860
	14				<b>ABC GL Total</b>				14,245,270
	17				<b>ABC AL Total</b>				7,249,632
	18				<b>ABC Total</b>				33,257,762
	48				<b>XYZ Total</b>				32,809,931
	49				<b>Grand Total</b>				<b>66,067,693</b>

### Actuarial IQ

- **Sorting** (does not reduce amount of data displayed, but brings the most important information to the top or bottom).
- **Conditional Formatting** (defines color, font, size and other formatting attributes of a cell as a function of the values in it or in other cells).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108				
2	1992	2.441	1.423	1.140	1.239	1.101	1.087	1.069	1.085				
3	1993	2.373	1.406	1.429	1.110	1.060	1.091	1.061					
4	1994	2.387	1.567	1.143	1.133	1.091	1.074						
5	1995	2.420	1.356	1.138	1.112	1.087							
6	1996	2.322	1.374	1.162	1.166								
7	1997	2.365	1.310	1.198									
8	1998	2.237	1.368										
9	1999	2.371											

- **OLAP-enabled tools** (provide an ability to display cross-sections or aggregations of multi-dimensional data in 2-D). OLAP-enabled tool (such as Excel's Pivot Table) is the ultimate adaptive reporting mechanism which supports filtering, sorting, outlining and conditional formatting and as such should become a preferred choice for any report designer.

Comment [d1]: Should be bold?

**Alarm system** is a technological solution whose purpose is to warn about undesired development. Alarm system usually triggers some action when the problem is found. The actions range from passive (paint some cells differently in the report) to interactive (display a warning dialog, send an e-mail requiring a response) to autonomous (launch a software program to fix the problem). At different stages of the data workflow, alarm messages can be aimed at different recipients: data integrity issues could be addressed to data managers, model's assumption test failures should be directed to actuaries, and sudden reserve increases should be presented to the management. Correspondingly, determination of which events under what conditions trigger an alarm is up to professionals responsible for the information quality on every given stage. In particular, actuaries should define what is acceptable and (on the other hand) what constitutes error or warning for data

## *Actuarial IQ*

suitability testing, actuarial analysis and presentation of actuarial results.

**Visualization** is the process of exploring, transforming, and viewing data as images to gain understanding and insight into the data. Images have unparalleled power to convey information and ideas. Consequently, visualization is a primary tool for communicating complex and/or voluminous information.

There exist a multitude of visualization approaches: mapping scalars to colors, contouring (iso-surfaces), glyphs (arrows of different color, length, direction), warping (display of different stages in the motion), displacement plots, time animations, streamlines (particle traces), and tensor algorithms. For the majority of actuaries, the most convenient and familiar visualization tool is a chart. From a presentation quality perspective, the report designer should be most concerned with the chart type, axis scaling, and the clarity and accuracy of the legend.

While there are numerous **chart types** available in spreadsheets, their add-ins, and other reporting packages, only a few are usually suitable for displaying each particular type of data. Percentages and shares are best presented by a pie chart, while XY-scatter is better suited for dependencies or comparisons (i.e., “Risk vs. Return”). Discrete values (i.e., “Total Premium per year”) are easy to present as a bar chart, while continuous variables (i.e., “Payment pattern”) are better displayed as lines. One shouldn’t use stacked bars for nonadditive values (i.e., “Incurred Loss” stacked on top of “Paid Loss”) or radar chart other than for comparison of several sets of data in multiple categories (criteria).

**Axis scaling** is very important for the readability of the chart, especially when displaying several data series. Sometimes in situations when one set of data (i.e., “Premium in dollars”) dwarfs another (i.e., “Exposure in number of cars”) it is necessary to create a second axis (with different scale) for the second set of data. Choosing an axis to be “time-scaled” automatically adds a capability to display monthly and annual aggregations of the data by selecting corresponding axis step (so called “axis base unit”). Occasionally, automatic scaling provided by a spreadsheet makes a wrong guess or is not as illustrational as desired. Sometimes data are better viewed in a logarithmic scale or in reverse order or with preset maximum and minimum. For example, displaying “Inception-to-date payments” on a chart with the maximum preset to “Aggregate Limit” could be more informative than just using automatically scaled axis<sup>1</sup>.

---

<sup>1</sup> To set up Maximum Value for the Axis click on the chart, right-click on the Axis and select “Format Axis...” menu option. In the “Format Axis” dialogue on the “Scale” tab uncheck “Maximum” checkbox and type desired value in the corresponding edit box.

## Actuarial IQ

The importance of **clarity and accuracy** of the chart title and legend cannot be overstated. Even the most primitive chart needs a precise description of the data displayed. Even more important, the user should provide accurate axes definitions and data series descriptions for the charts with multiple data series, dual axes, or of mixed types (i.e., bar and line on one chart). Without that, a chart may become a source of confusion instead of being a source of information.

### 4. Practical solutions: fighting misinterpretation

Actuaries deal with more and more sophisticated notions that are easy to misinterpret, misunderstand, and misuse. Three of the most difficult notions (as identified in [28]) for decision makers, regulators, accountants, and auditors are uncertainty, development, and multidimensional ranking. Attempts to explain and illustrate these concepts can result in confusion and wrong decisions. The problem is fundamental, given that accountants, performance measurers, and lawmakers operate with numbers rather than with distributions of random values. For example, an attempt to represent the distribution of possible aggregate losses with just one (“Reserve”) or two (“Reserve Range”) numbers inevitably leads to shortcuts in understanding and may create the impression that any value within a range is equally probable. The misinterpretation may be reinforced by a chart with reserve ranges shown as solid bars.

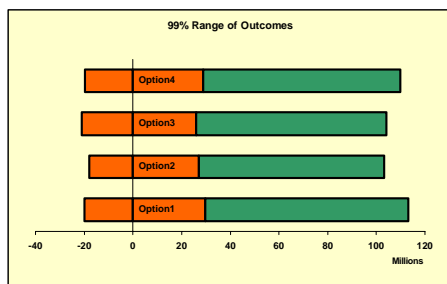


Fig. A.4.1

In reality, aggregate losses are not uniformly distributed and deserve more sophisticated graphical representation.



## Actuarial IQ

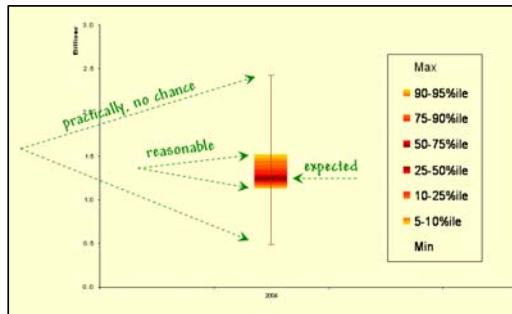


Fig. A.4.2

Some visual cues like gradients<sup>1</sup> or properly shaded areas should assist in visualizing uncertainty. Indeed, vanishing color is supposed to emphasize diminishing probabilities of extreme outcomes. Thus, Fig. A.4.3 may give better representation of the reserve ranges than Fig. A.4.1.

---

<sup>1</sup> To set up chart attributes such as gradient and borders, right-click on the chart element (i.e., bar), choose “Format %chart element%...” menu option and select tab “Pattern”. For gradient click on “Fill Effects...” button in the “Area” section, for borders make proper selections in the “Borders” section of the dialog.

## Actuarial IQ

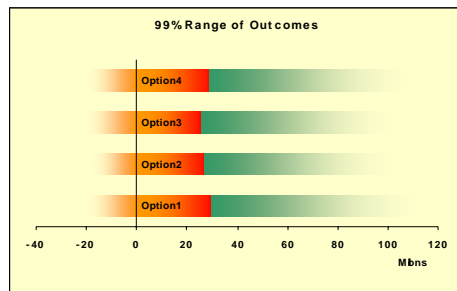


Fig. A.4.3

Another hurdle for the end users of reports to overcome is the concept of **development**. Combined with uncertainty (which itself changes over time), it creates a lot of opportunities for misunderstanding. Numbers in a spreadsheet or on a printed page do little to demystify trends, while standard chart options produce misleading results. With some effort, however, it is possible to illustrate development of random values in a spreadsheet chart (for example, charts on the Fig. A.4.4 below utilize vertical dimension (width of the curve in one case and height of the line in another) to illustrate the size of uncertainty which changes over time).

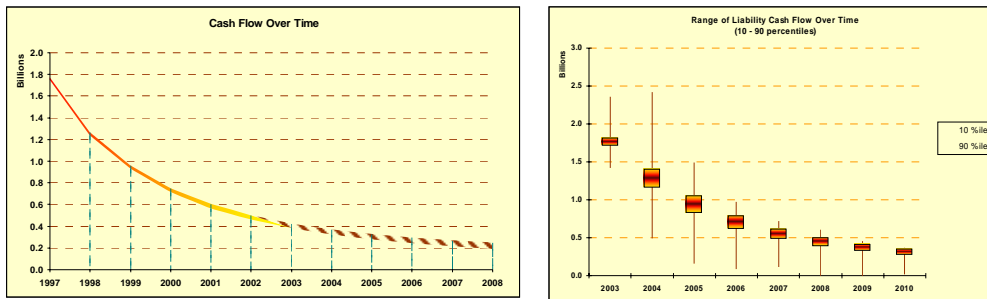


Fig. A.4.4

Decision makers rarely have their options conveniently ranked for them in one numeric dimension (i.e., “Net Profit”); they usually have to take into account multiple considerations (i.e., “Profit vs. Risk”), attempting to do **multidimensional ranking**. Geometrically speaking, their challenge is to say which one of the several points on the 2-D plane is “the best” one.

## Actuarial IQ

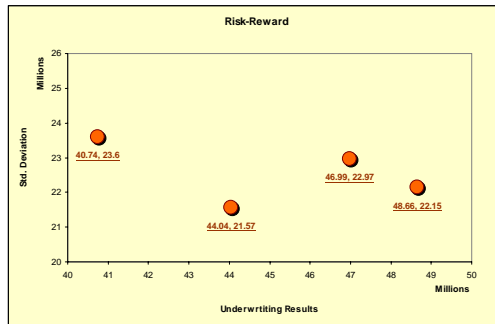


Fig. A.4.5

If the decision maker can formalize his preferences and express them as a so-called “goal function” (i.e., “the goal is to maximize risk to return ratio”), then display of the data can be optimized for that goal-seeking purpose. Taking a cue from a geographical map, the report designer may draw isolines (where goal function remains constant) and shade areas in between differently (for different values of these constants).

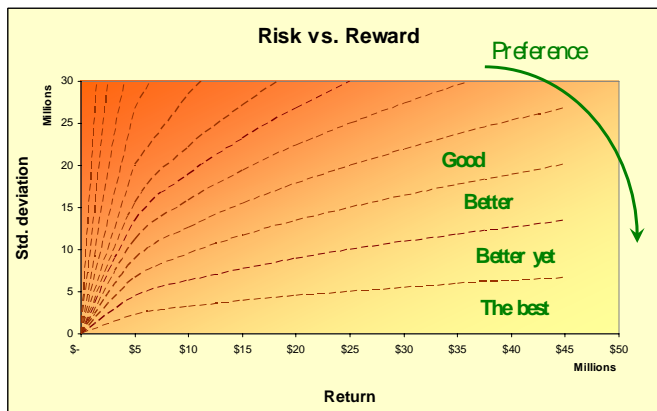


Fig. A.4.6

Placing 2-D points on such a map may significantly assist in selecting “the best” option.

## Actuarial IQ

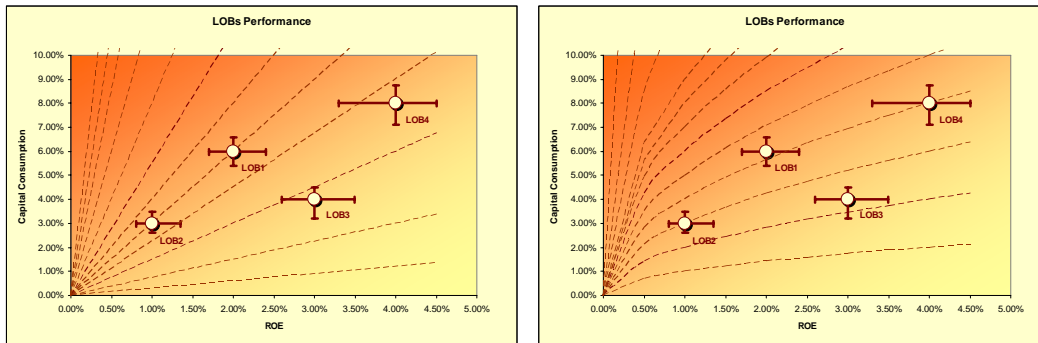


Fig. A.4.7

Fig. A.4.7 places points from the Fig A.4.5 on the grid of iso-lines from Fig. A.4.6. With a visual cue from the gradient (reddish areas are bad, yellowish areas are better) it is evident that lower right point labeled “LOB3” should be ranked #1 for a given choice of goal functions.

While use of visualization techniques requires some effort from the report designers and some training for the report readers, the payoff in interpretation quality (and, consequently, in decision accuracy) is considerable.

### Appendix B: Data cleansing steps (addendum to section 3.6.1)

The following is based on *Improving Data Warehouse and Business Information Quality* ([23], pp. 237-282).

1. **Identify data sources** and select the most authoritative source (i.e., company policy database, company claims database, company bill review database, etc.). Note that this may not be as easy as selecting a single file: the most reliable version of different data elements could come from different files. Similarly, different sources may be more authoritative for a single data element in different circumstances. Best data are coming from the sources and processes that have the largest stake in the correctness of data (e.g., the accounting department may treat payments info more accurately than their claims management colleagues and vice versa for the case reserves data). Frequency and timeliness of updates may also serve as indicator of reliability (more recently updated data probably has been looked at and corrected, so it could be more accurate). Metadata (2.3) can help to identify the most authoritative files.
2. Extract and **analyze** source data **for anomalies**.

### *Actuarial IQ*

- a. Analyze the meaning of the data with source data subject matter experts. For example, confirm that “AccDate” is indeed a date of the accident and find out what “GrossNetPrem” means.
  - b. Document the definitions, domain value sets, and business rules for each data element as used in its source file.
  - c. Extract a representative data sample and analyze it to confirm that the actual data is consistent with its definition and to discover any anomalies in how the data was used and what these incorrect entries mean. The objective is to discover undocumented values and their meanings.
3. **Standardize** the contents of data attributes: the definition and domain value sets for each standardized data attribute become the authoritative enterprise definition. Format nonstandardized data into standardized data elements with standardized domain value sets. For example, if certain files were using “2” for “married” but the enterprise definition is “M” for “married,” then replace the “2”s with “M”s.
4. **Correct and Complete Data.** Improve the quality of the existing data by correcting inaccurate or nonstandard data values and finding and capturing missing data values. The objective is to improve the quality of the data to the highest level.
- a. Identify missing data and obviously incorrect or suspect data (using, for example, EDA techniques described in [3.1](#))
  - b. Prioritize data to be cleansed based on value of correct data compared to correction costs.
  - c. Determine how to handle suspect data. The most efficient approach for simple but massive cleansing is to use automated transformation routines that can modify data according to business rules. However, where the suspect data is critical and the investigation is economically feasible, the best handling of suspect data is investigation and request for correct data from the source. Alternatives:
    - i. Reject the data.
    - ii. Accept the data without change.
    - iii. Accept the data without change but document that it is suspect.
    - iv. Accept the data but estimate the correct or approximate values based on other

### *Actuarial IQ*

related attributes. In this case, make sure that data are flagged as “estimated” and that impact on the intended use of data is tolerable.

- v. The best handling of suspect data is investigation and request for correct data from the source.
  - d. Implement the selected approach(s) for cleansing.
  - e. Document what was done and why.
5. **Eliminate duplicates.**
- a. Establish criteria to identify duplicate data records.
  - b. Determine impact of incorrectly consolidating multiple different records into one.
  - c. Determine matching techniques to use.
  - d. Look for intra-file duplicate records.
  - e. Look for inter-file duplicate records.
  - f. Investigate duplicates to make sure they are in fact duplicate records.
  - g. Document the matching and merge rules in the data map of source to target.
  - h. Establish a control mechanism to cross-reference duplicate occurrences in multiple files when primary key cannot be kept identical across files.
  - i. Examine and re-relate data related to the old records being consolidated to the new record.
  - j. Maintain an archive of the original source data for an appropriate length of time for error recovery purposes.
6. Analyze data for **patterns of errors**. The objective here is to leverage the knowledge of the data cleansing work to discover patterns of data errors and eliminate the most significant problems caused by data errors, as well as the most significant causes of errors. Analyze results to understand the kinds of errors, frequencies, and the cost impacts of the errors on the business.
- a. List and analyze examples of various kinds of data anomalies.
  - b. List two or three representative examples of each type of data defect.
  - c. Categorize the information quality problems and patterns.

### *Actuarial IQ*

- d. Estimate the frequency of each information quality problem.
  - e. Estimate the relative costs or impacts of each information quality problem, if possible.
  - f. Summarize the impact by data defect type.
7. **Map** the corrected data into its data file. Prepare the data for loading into the warehouse or target database, and include converting or formatting the cleansed consolidated data into the new data architecture. This step requires:
- a. Cleansed and standardized data.
  - b. Data from external information sources for integration with internal data.
  - c. Business rules governing the source data.
  - d. Business rules governing the target data warehouse data.
  - e. Transformation rules governing the transformation process.
8. **Optimize** data warehouse performance by determining and storing derived data (like triangles or other pre-aggregations) for the most frequently asked queries requiring complex calculations.
9. **Audit** and control data extraction, transformation, and loading. **Update** these procedures as necessary. Once the above steps are completed, this step is a matter of implementing procedures to assure the processes are performed as specified and kept up to date. See [3.2](#) for more information on data audits.

#### **Appendix C: Data defect prevention (addendum to section [3.6.2](#))**

The following is based on *Improving Data Warehouse and Business Information Quality* ([23], pp. 285-310). The systematic approach for preventing data defects from recurring contains 6 organizational steps:

0. Analysis and identification of all data processes and procedures in the company with particular focus on those processes associated with defective data.
1. Selecting a particular process for improvement.
2. Brainstorming and developing an improvement plan.

### *Actuarial IQ*

3. Implementing improvements in a controlled manner and confirming that improvements do solve a real problem.
4. Evaluation of impact against preset success criteria.
5. Rollout of improvement through the entire company along with training and documentation.

As an example, consider a process of building loss triangles for actuarial analysis of a company's reserves using accurate data.

Step 1 would first involve identifying and selecting those processes associated with the most significant payoff based on the impact of the data errors. Then identify the data sources (company's warehouse) and data owners (data management team) along with data consumers (reserving actuaries). This step ends with an assignment of improvement project sponsor and a team accountable for process changes.

Step 2 is the essence of the defect prevention activity. It requires identification of the root of the problem (for example, miscoding of "Line of Business" attribute), a feasible technical solution (integrity check, link to line-subline table, update of incorrect values), a plan to implement a solution (build line-subline table, write SQL queries, test on a sample of data, deploy), a measure of success (% of incorrect records) and costs associated with fixing the problem versus not fixing it (programming time versus errors in reserves and insolvency).

Step 3 includes not just implementation itself but also testing, documentation of the changes, and training of the personnel.

Step 4 consists of measurement of success defined in Step 2 or analysis of failure with possible repeat of Steps 2 and 3.

Step 5 involves generalization of the improvement with an attempt to apply it to all applicable areas in the company (fixing triangles used for pricing).

Making improvements in a systematic rather than haphazard manner will help to prevent more errors more effectively and more successfully with less effort and lower cost.



## 5. REFERENCES

- [1] Campbell, R.; Francis, L.; Prevosto, V.; Rothwell, M; Sheaf, S., “Report of the Data Quality Working Party” 2006, <http://www.actuaries.org.uk/files/pdf/proceedings/giro2006/Francis.pdf>
- [2] Actuarial Standards Board. *Actuarial Standard of Practice No. 23: Data Quality, revised edition*. Schaumburg, Illinois: American Academy of Actuaries, 2004.
- [3] CAS Committee on Management Data and Information. “White Paper on Data Quality.” *CAS Winter 1997 Forum*. 145-168.
- [4] Francis, Louise A. “Dancing with Dirty Data: Methods for Exploring and Cleaning Data.” *CAS Forum Winter 2005*: 198-254.
- [5] Copeman, P.; Gibson, L; Jones, T.; Line, N.; Lowe, J.; Martin, P.; Mathews, P.; Powell, D., “A Change Agenda for Reserving: A Report of the General Insurance Reserving Issues Task Force” 2006, [www.actuaries.org.uk](http://www.actuaries.org.uk)
- [6] CAS Data Management Educational Materials Working Party. “Survey of Data Management and Data Quality Texts.” *CAS Winter 2007 Forum*. 273-306.
- [7] Dasu, Tamraprni and Theodore Johnson. *Exploratory Data Mining and Data Cleaning*. Wiley, 2003.
- [8] Watson, Richard T. *Data Management: Databases and Organization, fifth edition*. New Jersey: Wiley, 2005.
- [9] Immon, William and Claudia Imhoff and Ryan Sousa. *Corporate Information Factory, second edition*. New Jersey: Wiley, 2000.
- [10] Moncher, Richard B. “Study Note: NCCI Data Collection Calls and Statistical Plans.” *CAS Exam Study Note Casualty Actuarial Society - Arlington, Virginia*, 1-17
- [11] Prevosto, Virginia R. “Statistical Plans for Property/Casualty Insurers.” *Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society - Arlington, Virginia*, May 1997, 201-216
- [12] Prevosto, Virginia R. “Study Note: ISO Statistical Plans.” *CAS Exam Study Note Casualty Actuarial Society - Arlington, Virginia*, 1997, 1-21
- [13] Cody, R. *Cody's Data Cleaning Techniques Using the SAS Software*, SAS Institute, 1999.
- [14] Hartwig, Frederick and Brian E. Dearing. *Exploratory Data Analysis*. Beverly Hills: Sage Publications, 1979.
- [15] Richard Scheaffer, William Mendenhall III and Lyman Ott. *Elementary Survey Sampling, Fifth Edition*. Wadsworth. 1996.
- [16] Insurance Services Office, Inc. *Strength in Numbers – A Total Data Quality Audit Program for Your Company*. Jersey City, New Jersey, 1989.
- [17] Insurance Services Office, Inc. *Quality of Data Audit Guide*. Jersey City, New Jersey, 1978.
- [18] Stuart A. Klugman, Harry H. Panjer, Gordon E. Willmot. *Loss Models: from data to decisions*. Wiley, 1998
- [19] Kit, Edward. *Software Testing in the Real World*. New York: Addison-Wesley, 1995.
- [20] Redman, Thomas C. *Data Quality, the Field Guide*. Boston: Digital Press, 2001.
- [21] Loshin, David. *Enterprise Knowledge Management*. Morgan Kaufman, 2001.
- [22] Olson, Jack E. *Data Quality: the Accuracy Dimension*. Morgan Kaufman, 2003.
- [23] English, Larry P. *Improving Data Warehouse and Business Information Quality*. New York: Wiley, 1999.
- [24] Casualty Actuarial Society. *Statement of Principles Regarding Property and Casualty Insurance Ratemaking*. 1988.
- [25] Insurance Services Office, Inc. *Data Management Best Practices*. Jersey City, New Jersey, 2003.
- [26] Popelyukhin, Aleksey S. “On Hierarchy of Actuarial Objects: Data Processing from the Actuarial Point of View”, *CAS Forum Winter 1999*: 219-237.
- [27] Popelyukhin, Aleksey S. “Let Me See: Visualizing Actuarial Information.” *CAS Forum Winter 2001*: 399-425.
- [28] Popelyukhin, Aleksey S. “Presenting DRM Results: Helping Executives Make Sense of DRM.” A report of CAS Working Party on Executive Level Decision Making Using DRM, 2005.
- [29] Popelyukhin, Aleksey S. “Rainy Day: Actuarial Software and Disaster Recovery.” *CAS Forum Winter 2003*: 55-73.

Abbreviations and Glossary

**ASB**, Actuarial Standard Board

**ASOP**, Actuarial Standard of Practice

**ASOP No. 23**, Actuarial Standard of Practice No. 23

**CAS**, Casualty Actuarial Society

**Categorical data**, (as opposed to numerical data) data whose values correspond to a specific category or label. Examples include alphanumeric data such as claimant state or NCCI injury code.

**Data attribute** is a characteristic of an object or an observation. Data attribute consists of a name and a value and is usually stored in a field in a data record. For example, attribute's name: "Date of Accident", value: "January 1, 2000".

**Data cube**: a multi-dimensional representation of the data. Dimensions are usually constructed from the categorical data, while cube content is usually some aggregate function (sum, count, max) of numerical data. For example, Excel's pivot table is a 2-dimensional projection of the data cube.

**Data domain**, the set of values valid for a given data element. For example, data domain for the "Gender" data element is a pair {"Male"; "Female"}.

**Data element** or **data entity**, the smallest unit of data record that has meaning to a knowledgeable worker. Data element is usually a value of a data attribute or a reference to another record in a (more detailed) table. For example, a loss record may contain the following data elements: values of the "Date of Accident" and "Line of Business" attributes and "Policy ID" reference to a record in "Policies" table.

**Data record** or **database record** is a (structured) row in a database table that represents a single object or observation as a collection of related data elements (stored as fields). For example, a record for insurance policy may consist of "Policy ID," "Inception Date," "Expiration Date," and "Premium" data elements.

**EDA**, Exploratory Data Analysis

**Field**, a column in a database table that stores a value of a single data attribute or a reference (key) to a record in another table.

**GIRO**, General Insurance Research Organization

**GRIT**, General insurance Reserving Issues Taskforce

**IDMA**, Insurance Data Management Association

**IT**, Information Technology

**ISO**, Insurance Services Office, Inc.

**MDDDB**, Multi-dimensional Database

**MGA**, Managing General Agency

**NAII**, National Association of Independent Insurers

**NCCI**, National Council on Compensation Insurance

**OLAP**, On-Line Analytical Processing, a mechanism for efficient analytical queries. OLAP heavily relies on data cubes as data structure and pre-aggregations as a way to speed up queries.

**Regulator**, Insurance is regulated by state insurance departments. Financial statements, rates, licenses to write business, etc. are monitored by regulators, including actuaries, who work for insurance departments.

**SQL**, Structured Query Language, a computer language to retrieve (place and modify) information from a (relational) database.

**Statistical Agent**, an organization that helps insurers satisfy legal requirements for reporting data to regulators. The statistical agent processes data submitted by insurers, performs data quality checks on the data, consolidates the data across insurers, and provides aggregate data compilations to state insurance departments on the behalf of the insurers.

**TPA**, Third Party Administrator, a company managing insurance claims, one of main sources of actuarial data.

**VBA**, Visual Basic for Applications, a programming language implemented in many applications, most notably in Microsoft Office.

**XML**, eXtensible Markup Language, a language that combines text with descriptive information about that text. For example, XML would store Excel's cell value along with the formula that generated that value.

## **Biographies of Working Party Contributors**

**Keith Allen** is the associate actuary for United Educators and is responsible for underwriting duties within the public school sector and general corporate actuarial issues. Allen has 13 years of experience in the insurance industry as an underwriter, claims adjuster, and actuary. Keith previously worked for Tillinghast-Towers Perrin as an actuarial specialist where he did reserving, pricing, and forecasting for various public and private entities. Prior to that, Allen worked as a claims adjuster and underwriter for State Farm Insurance where he helped develop the “Reinspection Program” used to assess coastal risks. Before joining the insurance industry, Allen was a teacher at Bellaire High School in Houston, TX. Allen holds a bachelor’s degree in mathematics from the University of Texas and is an Associate of the Casualty Actuarial Society.

**Robert Campbell** is Assistant Vice President, Actuarial Services at Lombard Canada in Toronto, Canada. He has a Bachelor of Mathematics in Business Administration from the University of Waterloo. He is a Fellow of the CAS and a Fellow of the Canadian Institute of Actuaries. He is chair of the Data Management Educational Materials working party, participates on the CAS Committee on Data Management and Information, and was a participant on the 2006 GIRO Data Quality working party.

**Louise Francis** is a Consulting Principal at Francis Analytics and Actuarial Data Mining, Inc. She is involved in data mining projects as well as conventional actuarial analyses. She has a BA degree from William Smith College and an MS in Health Sciences from SUNY at Stony Brook. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She serves on several CAS committees /working parties and is a frequent presenter at actuarial and industry symposia. She is a four-time winner of the Data Quality, Management and Technology call paper prize, including one for “Dancing with Dirty Data: Methods for Exploring and Cleaning Data (2005).”

**Dave Hudson** is an Actuary for The Travelers in Hartford, CT. He has a MS degree in Mathematics from Washington State University in Pullman, WA. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. He is also a member of the CAS Committee on Data Management and Information.

**Gary W. Knoble** is a consultant for insurance data management and professional education. He serves as a senior advisor to the Insurance and Finance Professional Education Consulting (Beijing) Co. Ltd . (Bao Rong), and the US Asia Business and Financial Services group assisting them in providing educational services to the Chinese insurance industry. Prior to his retirement in January 2006 he was a Vice President of The Hartford Financial Services Group where he directed the Data Management Division of the Actuarial Department. He is a past president of the Insurance Accounting and Systems Association (IASA). He is a founding member and past president of the Insurance Data Management Association and currently serves on the Board as Vice President of Professionalism. He has served on the ACORD P&C Steering Committee, the International Association of Industrial Accident Boards and Commissions (IAIABC) Electronic Data Interchange Council and Associate Member Council, ISO’s Ad Hoc Data Reporting Group, and for many years chaired the Surety Association of America’s Statistical Committee. He serves as a Visiting Professor at Nanjing Audit University and as an advisor to the Actuarial Alumni Association of the University of Science and Technology of China in Hefei. Mr. Knoble is a recipient of several distinguished industry awards including the IASA President’s award in 2002, an outstanding achievement award from the IAIABC in 2004, and a special award for creativity, diplomacy, fidelity and vision from the IDMA in 2005. In 2005, IDMA announced the creation of the Gary Knoble Award that will be given periodically to an individual who has made an outstanding career contribution to the field of Data Management. A native of Salt Lake City, Utah, Mr. Knoble is a graduate of Yale University with a major in International Political and Economic Institutions.

**Rudy Palenik** is the Commercial Actuary at Westfield Insurance Group in Westfield Center, Ohio. He is responsible for the development of rates for all the commercial lines of business. He has a degree in Math from Marquette University in Milwaukee, Wisconsin and is a Fellow of the Casualty Actuarial Society and a member of the American Academy of Actuaries. Rudy participates on a number of CAS committees including: Data Management and Information, Actuarial Education and Research Foundation, Research Paper Classifier and University Liaison.

**Aleksey Popelyukhin** is a Vice-President of Information Systems with the 2 Wings Risk Services and a Head of

### *Actuarial IQ*

Quantitative Analytics Group with the Wall Street North Consulting in Stamford, Connecticut. He holds a Ph.D. in Mathematics and Mathematical Statistics from Moscow University (1989). Aleksey actively participates in CAS research and is frequent presenter on CAS conferences. CAS recognized Aleksey's contributions by awarding him the very first prize in "Data Management" papers competition and inviting him to the very first Working Party (on presentation of DFA/DRM results). In addition to numerous publications Aleksey helps to advance actuarial science by building convenient software tools for actuaries such as Triangle Maker®, Affinity and Actuarial Toolchest™. For those actuaries having troubles explaining statistics to the management, Aleksey built a DRM presentation template available from CAS website. And for those who have troubles fitting clean models to dirty data Aleksey developed advanced data quality service called Data Quality Shield<sup>SM</sup>. Aleksey is currently developing an integrated pricing/reserving/DRM computer system for reinsurance called "SimActuary" and also an action/adventure computer game tentatively called "Actuarial Judgement."

**Virginia R. Prevosto** is a Vice President at Insurance Services Office, Inc. Ms. Prevosto is a Phi Beta Kappa graduate of the State University at Albany with a Bachelor of Science degree in Mathematics, *summa cum laude*. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She serves as General Officer of the CAS Examination Committee and as liaison to various other CAS admission committees. She also serves on the CAS Committee on Management Data and Information. In the past, Ms. Prevosto also served on the Data Quality Task Force of the Specialty Committee of the Actuarial Standards Board that wrote the first data quality standard of practice. Virginia has been a speaker at the Casualty Loss Reserve Seminar on the data quality standard and to various insurance departments on data management and data quality issues. Ms. Prevosto authored the paper "Statistical Plans for Property/Casualty Insurer" and "Study Note: ISO Statistical Plans" and co-authored "For Want of a Nail the Kingdom was Lost – Mother Goose was right: Profit by Best (Data Quality) Practices" for the IAIDQ.