SOME CONSIDERATIONS ON AUTOMOBILE RATING SYSTEMS UTILIZING INDIVIDUAL DRIVING RECORDS

BY

LESTER B. DROPKIN

INTRODUCTION

With the recent introduction of automobile rating systems which modify an otherwise applicable rate by utilizing some form of individual driving record, a number of questions presented themselves. On the one hand it was felt that a mathematical description of a phenomenon—in this case risk distributions by number of accidents—is intrinsically of value and constitutes an advance. The first part of this paper is concerned with the presentation of such a description. A frequency distribution, known as the negative binomial distribution is utilized in these first sections.

Of considerable and immediate importance is the question: What is the probability that an individual rated according to a given "driving record sub-classification" has been correctly classified? The answer to the question as phrased is actually an objective and, as such, is not specifically answered here. Rather, we have utilized a simple type of segregating system, based on the number of traffic violations only without regard to the type of violation involved.[1] In the concluding parts of this paper an analysis of this simple model is made and conclusions drawn. As is there pointed out, this paper has as one of its prime intents, the introduction and utilization of certain approaches to the problem. While an extrapolation of some of these conclusions to the actual rating systems currently being introduced by the rating bureaus and others is made, this paper is by its nature preliminary. It is hoped that the near future will produce more extensive investigations.

THE RATIONALE OF USING THE NEGATIVE BINOMIAL DISTRIBUTION

Of those individuals who have no accidents during an experience period some will be persons with a high loss-causing propensity but have been "lucky", some will be persons with a very low propensity and have seen their "expectations" realized, and conversely. All this we know (or assume). The attempt is here made to unravel some of these threads and to gain a means of approach whereby some of the probabilities involved may be set forth.

In discussions of the distributions of risks by number of accidents it has been traditional to base such discussions on the Poisson fre-

Note 1. For a description of the California study which constitutes the basic data for this paper, see the paper by F. Harwayne entitled "Merit Rating in Private Passenger Automobile Liability Insurance and The California Driver Record Study."

quency function, P(x). That is, if we let n be a random variable (equal to the number of accidents) we have assumed that the probability that n = x, where x = 0, 1, . . . is given by:

(1) P (x) = prob. that n equal x = $(m^x e^{-m}) / x!$

In dealing with a given body of experience the parameter m is set equal to the observed mean because in the Poisson distribution E(x) = m.

A test of goodness of fit by use of the chi-square distribution will, however, often indicate a significant deviation. A much improved fit will often result by considering that n is distributed in accordance with the two parameter frequency function:

(2) $N(x) = \text{prob. that n equals x} = \left(\frac{a}{1+a}\right)^r \binom{-r}{x} \left(\frac{-1}{1+a}\right)^x$

where x = 0,1, . . .

This frequency function is known as the *negative binomial distribution.*[2] For this function $E(x) = r/a$ and $\sigma^2 = (r/a)[(a+1)/a]$ as will be shown subsequently. In fitting observed data to eq. (2) the observed mean and variance are set equal to r/a and $(r/a)[(a+1)/a]$ respectively, whence the parameters r and a can be determined by solving the two equations simultaneously. Upon solving we get that $r = m^2/(\sigma^2 - m)$ and $a = m/(\sigma^2 - m)$. In actually using N(x) with a given body of data it is usual to use the following expanded form in which the values are obtained when $\left(\frac{a}{1+a}\right)^r$ is multiplied by the terms of the sequence:

$$\left\{1, \frac{r}{1+a}, \frac{r(r+1)}{2!(1+a)^2}, \frac{r(r+1)(r+2)}{3!(1+a)^3}, \ldots\right\}$$

That is, the probability that $n = 0$ is $\left(\frac{a}{1+a}\right)^r \cdot 1$;

that $n = 2$ is $\left(\frac{a}{1+a}\right)^r \cdot \frac{r(r+1)}{2(1+a)^2}$

The rationale of the applicability of N(x) to distributions by number of accidents results from the following considerations: If we assume that the parameter m in eq. (1) is itself a (continuous) random variable with the frequency function T(m) then the probability that n takes on any given value x is:

Note 2. See Appendix B for a comparison of the fit achieved by the use of the negative binomial and by the Poisson. The Chi-square test on the Poisson and the very good fit of the negative binomial was called to my attention by F. Harwayne.

$$(3) \qquad \int_0^{\infty} P(x) \cdot T(m)\, dm$$

Without for the moment specifying the form of T(m), the introduction of a variable m can be interpreted as a way of accounting for the variation of risk among the members of a given population. That is, it is assumed that

(a) the individual chances vary from one person to another but (for the given individual) remain constant throughout the experience period, and

(b) these initial propensities are distributed in the population in a simple curve, T(m).

The negative binomial, N(x) results from assigning to T(m) the specific form:

$$(4) \qquad T(m) = \frac{a^r}{\Gamma(r)}\, m^{r-1}\, e^{-am} \qquad \text{(a, r positive)}$$

which is a Pearson Type III. The Type III curve being selected because of its skew form and because it leads to conveniently simple equations for fitting. It is also possible if a frequency is expressible by a Type III curve to express the chance of a variation within a given limit by utilizing Pearson's *Tables of the Incomplete Gamma Function.* This enters into later considerations.

The mathematics of these considerations is given in Appendix A.

THE EFFECT OF SEGREGATING BY DRIVING RECORD

As indicated in the Introduction we have dealt here only with a simple segregation by traffic violation; i.e., we have used only the data appearing in the California Study.

While the average accident involvement generally increases with increasing number of violations (see F. Harwayne, op. cit.) it does appear that for the groups with 5, 6, 7, 8 and 9 or more violations, the mean accident frequencies have become relatively stable. (The respective means are .557, .508, .502, .545 and .656).

The fact that the negative binomial fits the data for the total group indicates that there is a real spread, that is, a distribution, of the probability of having an accident. From the construction of the negative binomial we have seen that this distribution is describable by a Type III curve.

Now it is clearly the function of a segregating system to split up the total heterogeneous group into homogeneous groups. The question is therefore raised as to whether or not, or to what degree, a segregating system based on traffic violations does split up the total group.

If the system we are dealing with here accomplished this purpose totally, then the distributions by number of accidents of the individual

groups should be describable by Poisson curves. Now if the variance of the separate groups were less than the Binomial variance[3], then Poisson curves would indeed be indicated. However, Appendix C shows that this is not the case. In every instance the variance is greater than the Binomial variance. This would seem to indicate that the desired segregation was not achieved.

We can, however, go further. Since a Poisson distribution is not indicated for the distributions by number of accidents, a negative binomial is indicated. But a negative binomial for the distribution by number of accidents is describable by a Type III curve. Now if we can picture these individual Type III curves, we can see in which groups, if any, the probability of having an accident is highly concentrated about the mean probability for that group. In other words, if we can determine what portion of the distribution is within stated deviations from the mean, then we can see how closely a given mean probability (of having an accident) approximates a constant probability and thus how closely the segregating system under consideration achieves its aim.

The required areas (or rather portion of total area) under the various Type III curves can be determined through a utilization of Pearson's *Tables of the Incomplete Gamma Function.* (See Appendix D for details.)

Appendix E sets forth, by individual group, the portion of the distribution within stated deviations from a given mean probability of having an accident. The deviations utilized are plus and minus 20%, 30%, 40%, 50%.

REVIEW, SUMMARY AND CONCLUSIONS

We have in a certain sense conceptually separated this paper into two parts. This was done in order to emphasize what to us seems to be the importance of the negative binomial distribution as a valuable instrument in its own right. It is our belief that this distribution can be an equally useful tool in attacking numerous other actuarial problems. It is also believed that many worthwhile results can flow from a utilization of the general approach illustrated by eq. (3). This equation is typical of the general theory of processes random in time (stochastic processes) and we believe that this theory will come to be of particular value to the actuary.

It is also important to emphasize here that there are two distributions which enter into our considerations. On the one hand there is the distribution of the probability of having an accident. On the other hand there is the distribution of risks by number of accidents. If the first distribution is a constant, then the second is a Poisson. If the first is a Type III, then the second is a negative binomial. Since the two parameters of the negative binomial are also the two para-

Note 3. The Binomial variance is equal to the product of the mean and the complement of the mean.

meters of the component Type III we can use the sample mean and variance to determine them. From a knowledge of the values of these parameters we can determine the spread about the mean probability of having an accident. If there is little spread then the segregating system has performed its function. A review of the figures shown in Appendix E indicates that in no group was there a real concentration about the mean. Thus for the group with 1 violation only about 25% of the group can be expected to lie within plus or minus 20% of the mean, 62% can be expected to fall outside of an interval of plus or minus 30% of the mean, etc. Notice too, that for the group having no violations, which represents 58.7% of the total number of individuals in the study, only a little over 25% of the group can be expected to lie in an interval of plus and minus 40% of the mean.

It is also very instructive to look at the question of overlapping. We see that about 25% of each of the groups having 1, 2, 3 or 4 violations can be expected to have a probability of having an accident greater than or equal to the mean probability for the succeeding group. As examples: The mean probability in group 3 is .354; the portion of group 2 having a probability of .356 (= 1.3 times the mean of group 2) or more is .25 (= 1–.75). The mean probability in group 5 is .553; the portion of group 4 having a probability of .554 (= 1.3 times the mean of group 4) or more is .26 (= 1–.74).

There is, in addition, considerable overlapping in the other direction. Thus, for example, the mean probability in Group 1 is .194; the portion of Group 2 having a probability of .192 (= .7 times the mean of group 2) or less is .36. For Group 2, therefore, about 60% of the group may be expected to have a probability of having an accident which is either less than the mean of the preceding group or greater than that of the following group. Similar figures obtain for other groups.

If in asking these questions we were to think of an interval about the means of the preceding and following groups, the amount of overlapping would of course be greater.

Having now performed these calculations, what are our conclusions? We are, it would seem, to conclude that the segregating system here considered does not function to effectively separate the total into groups sufficiently homogeneous to merit modifications of the rate.

We may well expect, a priori, that a segregating system which is based on only certain violations rather than all violations, that introduces a weighting process for these violations and that includes accident record as well as violation record, will produce a separation into groups more homogeneous than we have seen here. We must, however, also note that the use of 2 years' experience instead of the 3 years which form the data for this study, will act to decrease whatever sharpness of separation the foregoing will presumably introduce.

While it is dangerous to extrapolate, it would appear from the results presented in this paper that two conclusions of general application may be drawn. These are that

(a) after a certain point an increase in the number of violations does not contribute proportionately to an increase in the average number of accidents, and

(b) the effect of segregating according to driving record is less effective than might be heretofore thought.

It is clear that the general area with which this paper is concerned is of current importance and is obviously a fertile field for many future papers. Presentations dealing with models more closely approximating the actual rating systems in use and with utilizations of the negative binomial distribution in other areas are earnestly to be desired.

APPENDIX A

Mathematics of the Negative Binomial

We here display the mathematics of the considerations set forth in the first part of the paper. By substituting in eq. (3) the specific forms P(x) and T(m) given by eqs. (1) and (4) we derive therefrom the equation for N(x) given by eq. (2). Following this we show that:

$$\text{(a)}\quad \sum_{x=0} N(x) = 1$$

$$\text{(b)}\quad E(x) = r/a$$

$$\text{(c)}\quad E(x^2) = \frac{1}{a}\left(\frac{a + r + 1}{a}\right) \text{ and that therefore}$$

$$\text{(d)}\quad \sigma^2 = E(x^2) - [E(x)]^2 = \frac{r}{a}\left(\frac{a + 1}{a}\right)$$

Derivation of N(x)

From eq. (1), $P(x) = (m^x e^{-m})/x!$ and from eq. (4), $T(m) = (a^r m^{r-1} e^{-am})/\Gamma(r)$ we are to derive N(x). We proceed as follows:

$$\text{(5)}\quad N(x) = \int_0^\infty \frac{m^x e^{-m}}{x!} \cdot \frac{a^r m^{r-1} e^{-am}}{\Gamma(r)}\, dm$$

$$\text{(6)}\quad = \int_0^\infty \frac{a^r}{x!\,\Gamma(r)} \cdot m^{(x+r-1)} \cdot e^{-m(1+a)}\, dm$$

$$\text{(7)}\quad = \frac{a^r}{x!\,\Gamma(r)} \cdot \frac{(x + r - 1)!}{(1 + a)^{x+r}} \qquad \text{[see Pierce \#493]}$$

$$\text{(8)}\quad = \left(\frac{a}{1 + a}\right)^r \cdot \frac{1}{(1 + a)^x} \cdot \frac{(x + r - 1)!}{x!\,\Gamma(r)}$$

Now since the last factor in eq. (8) can be transformed as follows:

$$\frac{(x + r - 1)!}{x!\,\Gamma(r)} = \frac{(r + x - 1)!}{x!\,(r - 1)!} = \frac{r(r + 1)\ldots(r + x - 1)}{x!}$$

$$= \frac{(-r)[-(r+1)][-(r+2)]\ldots[-(r+x-1)](-1)^x}{x!}$$

$$\text{(9)}\quad = (-1)^x \binom{-r}{x}$$

we have that

$$(10) \qquad N(x) = \left(\frac{a}{1+a}\right)^r \cdot \frac{1}{(1+a)^x} \cdot (-1)^x \cdot \binom{-r}{x}$$

which is equation

$$(2) \qquad N(x) = \left(\frac{a}{1+a}\right)^r \binom{-r}{x} \left(\frac{-1}{1+a}\right)^x$$

From this it immediately follows that

$$(11) \quad \sum_{x=0} N(x) = \left(\frac{a}{1+a}\right)^r \left(1 - \frac{1}{1+a}\right)^{-r} = \left(\frac{a}{1+a}\right)^r \left(\frac{a}{1+a}\right)^{-r} = 1$$

Derivation of $E(x)$, $E(x^2)$ and σ^2

By definition, $E(x) = \sum_{x=0} x \cdot N(x)$, whence

$$(12) \quad E(x) = \sum_{x=0} x \cdot N(x) = 0 + \sum_{x=1} x \cdot N(x)$$

$$(13) \qquad = \sum_{x=1} \left(\frac{a}{1+a}\right)^r \left(\frac{r}{1+a}\right) \binom{-(r+1)}{x-1} \left(\frac{-1}{1+a}\right)^{x-1}$$

$$(14) \qquad = \left(\frac{a}{1+a}\right)^r \left(\frac{r}{1+a}\right) \left(1 - \frac{1}{1+a}\right)^{-(r+1)}$$

$$(15) \quad E(x) = \left(\frac{a}{1+a}\right)^r \left(\frac{r}{1+a}\right) \left(\frac{a}{1+a}\right)^{-(r+1)} = \frac{r}{1+a} \left(\frac{a}{1+a}\right)^{-1} = \frac{r}{a}$$

Similarly from $E(x^2) = \Sigma x^2 \cdot N(x)$ and $\sigma^2 = E(x^2) - [E(x)]^2$, we have:

$$(16) \quad E(x^2) = \sum_{x=0} x^2 \cdot N(x) = 0 + \left(\frac{a}{1+a}\right)^r \left(\frac{r}{1+a}\right) + \sum_{x=2} x^2 \cdot N(x)$$

By writing $[x(x-1) + x]$ for x^2, we get

$$(17) \quad E(x^2) = \left(\frac{a}{1+a}\right)^r \left(\frac{r}{1+a}\right) + \sum_{x=2} x(x-1)N(x) + \sum_{x=2} x \cdot N(x)$$

But since $\sum_{x=1} x \cdot N(x) = r/a$ it follows that

$$\sum_{x=2} x \cdot N(x) = r/a - \left(\frac{a}{1+a}\right)^r \left(\frac{r}{1+a}\right)$$

Accordingly (17) becomes

$$(18)\quad E(x^2) = \frac{r}{a} + \sum_{x=2} x(x-1)\left(\frac{a}{1+a}\right)^r \binom{-r}{x} \left(\frac{-1}{1+a}\right)^x$$

$$(19)\quad = \frac{r}{a} + \sum_{x=2}\left(\frac{a}{1+a}\right)^r \cdot \frac{r(r+1)}{(1+a)^2} \binom{-(r+2)}{x-2} \left(\frac{-1}{1+a}\right)^{x-2}$$

$$(20)\quad = \frac{r}{a} + \left(\frac{a}{1+a}\right)^r \cdot \frac{r(r+1)}{(1+a)^2} \cdot \left(1 - \frac{1}{1+a}\right)^{-(r+2)}$$

$$(21)\quad E(x^2) = \frac{r}{a} + \frac{r(r+1)}{a^2} = \frac{r}{a}\left(\frac{a+r+1}{a}\right)$$

From this it immediately follows that

$$(22)\quad \sigma^2 = \frac{r}{a}\left(\frac{a+r+1}{a}\right) - \left(\frac{r}{a}\right)^2$$

$$(23)\quad \sigma^2 = \frac{r}{a}\left(\frac{a+1}{a}\right)$$

APPENDIX B

Comparison of Fit by Poisson and Negative Binomial for Total Group

Number of Accidents	*Observed Freq.*		*Theoretical Frequency*			
			Negative Binomial		*Poisson*	
	No.	*%*	*No.*	*%*	*No.*	*%*
0	81714	86.07	81726	86.086	80655	84.959
1	11306	11.91	11273	11.874	13147	13.848
2	1618	1.71	1647	1.735	1072	1.129
3	250	.26	245	.258	58	.061
4	40	.04	37	.039	3	.003
5 or more	7	.01	7	.008	—	—
	94935		94935		94935	

Mean = .163 $\sigma^2 = .193$ Binomial Variance = .136

For fitting the neg. binomial: $r = .8927$; $a = 5.472$; $\frac{a}{1+a} = .8455$

For fitting the Poisson: $e^{-.163} = .84959$.

APPENDIX C

Group (Violations)	*Mean*	*Variance*	*Binomial Variance**
0	.087	.096	.079
1	.194	.207	.156
2	.274	.299	.199
3	.354	.395	.229
4	.426	.501	.245
5 or more	.553	.610	.247
Total	.163	.193	.136

* Equals the product of the mean and its complement.

APPENDIX D

The determination of the ratios of $\int_0^t T(m)\,dm$ to $\int_0^\infty T(m)\,dm$ with T(m) as defined in equation (4), is accomplished by utilizing the *Tables of the Incomplete Gamma Function* prepared under the direction of Karl Pearson in 1922.

The complete gamma function $\Gamma(p+1)$ is defined as $\int_0^\infty e^{-x}x^p dx$ while the incomplete gamma function $\Gamma_x(p+1)$ is defined as $\int_0^x e^{-x}x^p dx$.

If $I(x,p)$ denotes the ratio of the incomplete to the complete gamma function, then $I(x,p)$ gives the portion of the curve to the left of x. However $I(x,p)$ has not been published. Instead a variable $u = x/\sqrt{p+1}$ is used and it is these equivalent tables of $I(u,p)$ which were prepared by Pearson. That is

$$I(u,p) = \frac{\int_0^{u\sqrt{p+1}} v^p e^{-v} dv}{\int_0^{\infty} v^p e^{-v} dv}$$

In order to use the tabulated values of $I(u,p)$ it is necessary to proceed as follows:

We first recall that $\int_0^{\infty} T(m)\,dm = 1$ so that we are looking for values of $\int_0^{t} T(m)\,dm$ and recall that:

$$\int_0^{t} T(m)\,dm = \int_0^{t} \frac{a^r m^{r-1} e^{-am}}{\Gamma(r)}\,dm$$

Now let $v = am$ so that $m = a^{-1}v$ and $dm = a^{-1}dv$. The integral thus becomes:

$$\int_0^{at} \frac{v^{r-1} e^{-v}}{\Gamma(r)}\,dv$$

Now let $p = r-1$; we then have:

$$\int_0^{at} \frac{v^p e^{-v}}{\Gamma(p+1)}\,dv$$

But this is precisely $I(u,p)$ with $at = u\sqrt{p+1}$; from this we get that

$$u = at/\sqrt{p+1} = at/\sqrt{r}$$

Since we know a and r from the data for a given t we have the values of u and p with which to enter the tables. One could for example determine values with $t =$ mean, mean $\pm$ 5%, mean $\pm$ 10%, mean $\pm$ 20%, etc.

APPENDIX E

When the procedures indicated in Appendix D are carried out, for values of $t = 50\%$, 70%, 80%, 120%, 130%, 140% and 150% of the mean, separately for each individual group, the following results are obtained:

PORTION OF CURVE WITHIN INTERVAL SHOWN FOR GROUP SHOWN

	Group (Violations)					
Interval	*0*	*1*	*2*	*3*	*4*	*5 or more*
0 to $.5\bar{x}$	.45	.18	.20	.19	.23	.10
0 to $.6\bar{x}$	.50	.25	.28	.27	.31	.17
0 to $.7\bar{x}$	.54	.32	.36	.35	.39	.26
0 to $.8\bar{x}$	.59	.40	.43	.43	.46	.36
0 to $1.2\bar{x}$	.71	.65	.70	.70	.70	.72
0 to $1.3\bar{x}$	.73	.70	.75	.75	.74	.78
0 to $1.4\bar{x}$	.76	.74	.79	.79	.78	.83
0 to $1.5\bar{x}$	.78	.78	.83	.83	.81	.88
$.5\bar{x}$ to $1.5\bar{x}$	.33	.60	.63	.64	.58	.78
$.6\bar{x}$ to $1.4\bar{x}$	.26	.49	.51	.52	.47	.66
$.7\bar{x}$ to $1.3\bar{x}$	.19	.38	.39	.40	.35	.52
$.8\bar{x}$ to $1.2\bar{x}$	.12	.25	.27	.27	.24	.36