SAMPLING THEORY IN CASUALTY INSURANCE

INTRODUCTION AND PARTS I AND II

BY

ARTHUR L. BAILEY

Introduction

The fundamental concept of insurance is that the insured is relieved of any concern, not only as to what is going to happen, but also as to what could happen but probably will not. Of course, at the time the insurance is written, neither the insured nor the insurer knows what is actually going to happen. But, even when the period of the coverage has expired, and the actual events are determined, both parties still should understand that the coverage provided was against what might have happened rather than against the specific events that actually did happen. Thus the losses paid by an insurer never actually reflect the hazard covered, but are always an isolated sample of all of the possible amounts of losses which might have been incurred.

It is this condition, of never being able to determine, even from hindsight, what the exact value of the inherent hazard of the coverage was, that has brought the casualty actuary into being. It becomes his province to make rates and rating plans such that, in the absence of an unprecedented catastrophe, his company will be able to pay the losses incurred in covering a wide group of such unknown inherent hazards and still stay solvent; or preferably, make a profit for its stockholders or pay a dividend to its policyholders. The myriads of compilations of loss experience, classification refinements, and expedients in general, resulting from the actuary's attempts to achieve this goal, need not be lingered on here, being well enough known. A few of the expedients do need to be referred to specifically.

When the loss experience of one group of insureds was first compiled, separately from that of another group, it was found that it was different. The question immediately arose, "Does this mean that the elusive inherent hazard is different for this group of risks, or does it mean that the hazard is the same but the actual losses of the two groups just happen to be different?" Credibility formulae were designed to provide an answer to this question. These have taken a wide variety of forms during the history of casualty insurance. Some have been based on the soundest of theoretical premises, while others have been purely expedients. Whatever the form, these formulae have been applied in recognition of a condition in which the actual observations are only samples of what might have been. It is the hope of the writer that this paper may serve, among other things, to produce more properly applicable credibility formulae.

Shortly after the problem of classification and territory experience had been met, the question of the insurance-minded large risks arose. Such risks knew that their operations were more efficient than those of their competitors; "otherwise they would not have become so large." They also knew that their loss experience was not exactly that contemplated by the manual rates. "Something ought to be done about it." Experience rating plans were developed as an answer and soon became the accepted thing; and a credibility formula was developed that would produce the same increase, or decrease, depending on the selected value of K, in the accuracy of the modified rate as compared with the unmodified rate, irrespective of the size of the risk. In later years this formula has been departed from as the plans have become more complicated. The implied aim has, of course, continued: that of producing a more accurate rate through application of the rating plan, although no means has been available to determine, before the plan is actually applied, whether or not this has been accomplished. The author hopes to provide a basis for the proper evaluation of the constant K.

The most recent development calling for a sound knowledge of sampling distributions is the retrospective rating plan. It is difficult to understand why this form of gambling remained dormant during the speculative twenties, only to break out in the depression thirties. The present epidemic can no doubt be explained as the direct result of the war hysteria. Whatever the cause, certain elements of the insurance industry desire to depart from rates based on expected averages and explore the possibility of rating on the basis of departures from expected averages. The initial essays in this direction have been made only after the application of some good unactuarial horse-sense by the underwriters, in the selection of the risks. Before, or it may be, while, we embark on an all-out retrospective program, it would seem well to seriously investigate the theoretical principles underlying such a course. It is felt by the author that this paper may serve as a foundation for such an investigation.

Another field where a knowledge of the sampling distributions of losses could be used to advantage is that of the rating procedures for deductible and excess coverages. Such procedures are now based on a very broad grouping of classifications, even including the entire line of insurance in many cases. This results in the necessity for a large safety margin in such rates in order to offset the selection against the company that inevitably results from broad classifications. The effect of such a rating procedure is to exclude, through redundant rates, the normal or subnormal risks from electing such coverages. Rating procedures can be developed, however, with a knowledge of the sampling distribution functions, which would give full rate recognition to differences in the hazard of such coverages by classification. These coverages, which include the real elements of insurance as con-

trasted to the chance variations of retrospective rating, could then be made available to the many large risks who need only this type of coverage.

We must recognize that the only data available to us in casualty insurance is in the form of samples of what may occur. From these samples we are required to measure, as well as we may, the inherent hazard of the coverage provided to a particular insured, or group of insureds. In experience rating, and more especially in retrospective rating, we must also measure the probable distribution of these expected losses among the risks. In order to do this we need a rather complete understanding of the theoretical distribution of losses among samples when various causes of variation are present.

The sampling variation due to pure chance fluctuations is always present in our data. Usually, however, our problems are made more complicated by the presence of other types of variation as well as chance. One of the most important of these types of variation is that between the inherent hazards of risks of the same rate classification and territory. Others are those resulting from errors, due to chance or otherwise, in the rate making procedure, or in the rating plan to be applied. Most problems involve the simultaneous consideration of at least two of these types of variation.

In many problems, however, we are only asked to decide whether or not a particular piece of data could reasonably be attributable to chance variation only. Other types of variation may be present in such data but are not involved in the answer to the problem. If the probabilities are greatly against the event occurring as a result of chance only, we may or may not then want to search for the reason. An extreme example of this kind of problem is presented by the $100,000 premium risk with a loss ratio of more than 100% in each of the past three years. The probability that this series of losses arose only from chance is so small that the underwriter himself would cancel the risk, or double the rate. Cases nearer to the borderline definitely come into the province of the actuary and can be answered only on the basis of a knowledge of sampling distributions resulting from chance variation only.

Thus our first step in the development of the theory of sampling distributions will concern itself only with variations due to chance. When these have been fully explored, we will then have to compound the results when one or more of the other types of variation are also present. The first two parts of this paper deal with the theory of purely chance sampling distributions and with methods of numerically approximating such distributions. Later parts will deal with the inclusion of other types of variation, and with the application of the theory to particular kinds of problems such as individual risk underwriting, rate making, experience rating, retrospective rating, and the rating of deductible and excess coverages.

It is hoped that this analysis of the sampling theory as applicable to casualty insurance will help to bring to light any flaws which may now exist in the rate making or rating structures and thereby serve to make them more accurate. It would be expected that the need for such corrections will be greater in the less highly developed rating of deductible and excess coverages and in the highly sensitive retrospective rating plans than in the older and more time-tested rate making and rating procedures.

It is believed that one conclusion will be drawn immediately from a review of the discussion of the Poisson Distribution. This is that the recording and collection of experience on a per accident basis, rather than on a per claim basis, would greatly assist in the interpretation of the data. This should certainly be done for all classifications of hazard involving any appreciable number of multiple-claim accidents.

The writer would appreciate being advised of any algebraic or arithmetic errors which may be found. As it has not been possible to have any independent check on most of this material, the author will have to assume full responsibility for these. That the symbolism used is different from that more recently presented can only be defended on the ground that the paper has been in progress for several years.

I.

Development of Basic Formulae for the Distribution of Casualty Insurance Statistics Due to Chance Fluctuations Only

A. The Poisson Distribution

The number of accidents in casualty insurance is distributed in accordance with the Poisson Distribution. This is not an assumption, but a demonstrable fact. The assumption, which it will later be necessary to make, is that the number of claims is also distributed in the same way.

We have from the Bernoullian Theorem that: if p is the probability of an event occurring and q is the probability of the event failing to occur, then out of s trials the probabilities of the event occurring 0, 1, 2, ..., $s-1$, s times are given by the terms in the expansion of $(q+p)^s$. It can be seen that the Bernoullian Distribution is not applicable to casualty insurance from the fact that the Bernoullian Theorem is dependent on the condition that there are only two possibilities; namely that the event happens, or it fails to happen. In casualty insurance the event (an accident) may not only happen or fail to happen, but it may also happen more than once.

We can, however, approach the conditions of the casualty business with the Bernoullian Distribution. If only one accident could happen each month, the probabilities of 0, 1, 2, ..., 12 accidents occurring during a year would

be given by the terms in the expansion of $\left(q + \frac{p}{12}\right)^{12}$, where p is the average number of accidents per year and q is equal to $1 - \frac{p}{12}$. Similarly, if more than one accident could happen during a month but only one could happen per day, the probabilities of 0, 1, 2, ..., 365 accidents occurring during the year would be the terms in the expansion of $\left(q + \frac{p}{365}\right)^{365}$, where q is equal to $1 - \frac{p}{365}$.

Finally, we could take the limiting case where the year is divided into an infinite number of parts. In this case the probabilities of 0, 1, 2, ..., to infinity accidents occurring during the year would be the terms in the expansion of:

$$\underset{n \to \infty}{\text{Limit}} \left[\left(1 - \frac{p}{n}\right) + \frac{p}{n}\right]^n$$

Only this limiting case would fit casualty insurance, where accidents can happen in very rapid succession although not at exactly the same time without, by definition, being the same accident. This limiting case is the Poisson Distribution; and the probabilities of 0, 1, 2, ..., n, etc. accidents are:

$$\frac{1}{e^p}, \frac{p}{e^p}, \frac{p^2}{\lfloor 2\, e^p}, \ldots\ldots, \frac{p^n}{\lfloor n e^p}, \ldots\ldots, \text{etc.}$$

Having found that the Poisson Distribution provides the probabilities of the occurrence of 0, 1, 2, ..., n, etc. accidents for an individual risk whose hazard remained constant throughout the year, the probabilities can be determined of 0, 1, 2, ..., n, etc. accidents occurring during a year between two risks having different hazards although both remain constant during the year.

Let the probabilities of 0, 1, 2, ..., n, etc. accidents for the first risk be given by:

$$\frac{1}{e^p}, \frac{p}{e^p}, \frac{p^2}{\lfloor 2\, e^p}, \ldots\ldots, \frac{p^n}{\lfloor n e^p}, \ldots\ldots, \text{etc.}$$

and for the second risk by:

$$\frac{1}{e^q}, \frac{q}{e^q}, \frac{q^2}{\lfloor 2\, e^q}, \ldots\ldots, \frac{q^n}{\lfloor n e^q}, \ldots\ldots, \text{etc.}$$

The probability that neither will have an accident is:

$$\frac{1}{e^p} \cdot \frac{1}{e^q} = \frac{1}{e^{(p+q)}}$$

The probability that there will be only one accident is:

$$\frac{1}{e^p} \cdot \frac{q}{e^q} + \frac{p}{e^p} \cdot \frac{1}{e^q} = \frac{p+q}{e^{(p+q)}}$$

The probability that there will be exactly two accidents is:

$$\frac{1}{e^p} \cdot \frac{q^2}{\lfloor 2 \cdot e^q} + \frac{p}{e^p} \cdot \frac{q}{e^q} + \frac{p^2}{\lfloor 2 \cdot e^p} \cdot \frac{1}{e^q} = \frac{q^2 + 2pq + p^2}{\lfloor 2 \cdot e^{(p+q)}} = \frac{(p+q)^2}{\lfloor 2 \cdot e^{(p+q)}}$$

The probability that there will be exactly n accidents is:

$$\frac{1}{e^p} \cdot \frac{q^n}{\lfloor n \cdot e^q} + \frac{p}{e^p} \frac{q^{(n-1)}}{\lfloor n-1 \cdot e^q} + \cdots\cdots + \frac{p^{(n-1)}}{\lfloor n-1 \cdot e^p} \cdot \frac{q}{e^q} + \frac{p^n}{\lfloor n \cdot e^p} \cdot \frac{1}{e^q}$$

$$= \frac{q^n + npq^{(n-1)} + \cdots\cdots + np^{(n-1)}q + p^n}{\lfloor n \cdot e^{(p+q)}} = \frac{(p+q)^n}{\lfloor n \cdot e^{(p+q)}}$$

Thus it is found that the probabilities of the occurrence of 0, 1, 2, ..., n, etc. accidents among two risks having different hazards, although both remain constant during the year, are likewise given by the Poisson Distribution using the combined hazard of both risks, i.e.:

$$\frac{1}{e^{(p+q)}}, \frac{(p+q)}{e^{(p+q)}}, \frac{(p+q)^2}{\lfloor 2 \cdot e^{(p+q)}}, \cdots\cdots, \frac{(p+q)^n}{\lfloor n \cdot e^{(p+q)}}, \text{ etc.}$$

This combination of hazards can obviously be extended to cover any number of risks having any range of individual hazards and also to cover any variation of hazard during the year. For the general case, where c represents the expected total number of accidents for all risks, the probabilities that the total number of accidents will be 0, 1, 2, ..., n, etc. are given by the terms of the Poisson Distribution:

$$\frac{1}{e^c}, \frac{c}{e^c}, \frac{c^2}{\lfloor 2 \cdot e^c}, \cdots\cdots, \frac{c^n}{\lfloor n \cdot e^c}, \cdots\cdots, \text{ etc.}$$

NOTE: Throughout this paper the expected number of accidents or of claims will be indicated either by a small "c" or a capital "C". No difference is intended between these two.

B. Sampling Distribution of the Number of Claims or Claim Frequency

The assumption will be made that the probabilities of the occurrence of 0, 1, 2, ..., n, etc. claims when c are expected are also given by the terms of the Poisson Distribution. This would seem to provide a very close ap-

proximation unless it is known that the claims usually occur in sizable groups.

In dealing with the various sampling distributions it will be found expedient to deal with the ratio of actual to expected values. In the case at hand the exposure element of the claim frequency cancels out in the ratio of actual to expected claim frequencies, and the ratio becomes identical to the ratio of actual number of claims to the expected number of claims. These ratios can take only the values corresponding to 0, 1, 2, ..., n, etc. claims of:

$$\frac{0}{c}, \frac{1}{c}, \frac{2}{c}, \ldots\ldots, \frac{n}{c}, \ldots\ldots, \text{etc.}$$

In order to prepare tables for the practical use of the sampling distributions, it will be necessary to evaluate the mean, variance, and skewness of these distributions. These are obtained as follows from the totals shown in Table 1 (see page 81):

$$\text{Mean} = \frac{\Sigma r \cdot f(r)}{\Sigma f(r)} = \frac{\text{Total of Column (3)}}{\text{Total of Column (2)}} = \frac{1}{1} = 1$$

$$V_{2:r} = \frac{\Sigma r^2 f(r)}{\Sigma f(r)} = \frac{\text{Total of Column (4)}}{\text{Total of Column (2)}} = \frac{1}{c} + 1$$

$$\text{Variance} = U_{2:r} = V_{2:r} - (\text{Mean})^2 = \frac{1}{c}$$

$$V_{3:r} = \frac{\Sigma r^3 f(r)}{\Sigma f(r)} = \frac{\text{Total of Column (5)}}{\text{Total of Column (2)}} = \frac{1}{c^2} + \frac{3}{c} + 1$$

$$U_{3:r} = V_{3:r} - 3\,(V_{2:r})(\text{Mean}) + 2\,(\text{Mean})^3 = \frac{1}{c^2}$$

$$\text{Skewness} = \frac{U_{3:r}}{(U_{2:r})^{3/2}} = \frac{\frac{1}{c^2}}{\left(\frac{1}{c}\right)^{3/2}} = \frac{1}{\sqrt{c}}$$

C. Sampling Distribution of the Total Cost of a Fixed Number of Claims

Before considering the sampling distributions of other statistics, it will be necessary to record certain data concerning the sampling distribution of the total cost of a fixed number of claims. It will be assumed that these claims occur at random out of an infinite number of equally likely possibilities, and that the moments of this infinite population of possible claims can be estimated from the distribution by size of the claims paid in the past.

Before letting the parent population approach the infinite in size, it will

be assumed to consist of N values whose amounts are $x_1, x_2, x_3, \ldots\ldots, x_N$, and from which the following sums are formed:

Σx = the sum of the N values of x.
Σx^2 = the sum of the N values of x^2.
Σx^3 = the sum of the N values of x^3.
Σxx = the sum of the ${}_NC_2$ possible products two at a time.
Σxxx = the sum of the ${}_NC_3$ possible products three at a time.
Σx^2x = the sum of the $2 \cdot {}_NC_2$ possible products of squares and values.

NOTE: The x's in the last three sums must have different subscripts.

From this population of N values of x, all possible combinations of n values will be formed; there being ${}_NC_n$ such combinations, each of which would be equally likely to occur were a single sample drawn. The total cost of these n claims will be designated as t. The required data are the first three moments of t about the origin, $V_{1:t}$, $V_{2:t}$, and $V_{3:t}$, and the second and third moments about the mean, $U_{2:t}$ and $U_{3:t}$.

In each value of t there are n values of x. In the sum of all ${}_NC_n$ possible values of t there are $n \cdot {}_NC_n$ terms of x's; and, as each of the N different values of x are equally frequent, each of the N values must occur $\frac{n}{N} \cdot {}_NC_n$ times. The average of all possible values of t will therefore be:

$$V_{1:t} = \frac{\frac{n}{N} \cdot {}_NC_n \cdot \Sigma x}{{}_NC_n} = n \cdot \frac{\Sigma x}{N} = nV_{1:x}$$

In each value of t^2 there are n terms of x^2 and ${}_nC_2$ terms of xx, each of which has a coefficient of 2. In the total of the ${}_NC_n$ values of t^2 there are $n \cdot {}_NC_n$ values of x^2 and $2 \cdot {}_nC_2 \cdot {}_NC_n$ values of xx. As there are only N different values of x^2 and ${}_NC_2$ different values of xx, then each value of x^2 occurs $\frac{n}{N} \cdot {}_NC_n$ times and each value of xx occurs $2 \cdot \frac{{}_nC_2}{{}_NC_2} \cdot {}_NC_n$ times. The average value of t^2 therefore is:

$$V_{2:t} = \frac{\frac{n}{N} \cdot {}_NC_n \cdot \Sigma x^2 + 2 \cdot \frac{{}_nC_2}{{}_NC_2} \cdot {}_NC_n \cdot \Sigma xx}{{}_NC_n}$$

$$= \frac{n}{N} \cdot \Sigma x^2 + 2 \cdot \frac{{}_nC_2}{{}_NC_2} \cdot \Sigma xx$$

It is noted here that $(\Sigma x)^2 = \Sigma x^2 + 2\Sigma xx$, so that $(\Sigma x)^2 - \Sigma x^2$ may be substituted for $2\Sigma xx$, giving:

$$V_{2:t} = \frac{n}{N} \cdot \Sigma x^2 + \frac{{}_nC_2}{{}_NC_2} \cdot (\Sigma x)^2 - \frac{{}_nC_2}{{}_NC_2} \cdot \Sigma x^2$$

$$= \left[n - \frac{n(n-1)}{N-1} \right] V_{2:x} + \frac{n(n-1)}{1 - \frac{1}{N}} \cdot {V_1}^2{}_{:x}$$

Letting N approach infinity, the limiting value becomes:

$$V_{2:t} = n \cdot V_{2:x} + n(n-1)\, {V_1}^2{}_{:x}$$

and

$$U_{2:t} = V_{2:t} - {V_1}^2{}_{:t} = n \cdot V_{2:x} - n \cdot {V_1}^2{}_{:x} = n \cdot U_{2:x}$$

In each value of t^3 there are n terms of x^3, $2 \cdot {}_nC_2$ terms of x^2x each with a coefficient of 3, and ${}_nC_3$ terms of xxx each with a coefficient of 6. In the total of the ${}_NC_n$ values of t^3 there are then $n \cdot {}_NC_n$ values of x^3, $6 \cdot {}_nC_2 \cdot {}_NC_n$ values of x^2x, and $6 \cdot {}_nC_3 \cdot {}_NC_n$ values of xxx. As there are only N possible values of x^3, $2 \cdot {}_NC_2$ possible values of x^2x, and ${}_NC_3$ possible values of xxx, each value of x^3 occurs $\frac{n}{N} \cdot {}_NC_n$ times, each value of x^2x occurs $3 \cdot \frac{{}_nC_2}{{}_NC_2} \cdot {}_NC_n$ times, and each value of xxx occurs $6 \cdot \frac{{}_nC_3}{{}_NC_3} \cdot {}_NC_n$ times. The average value of t^3 is therefore:

$$V_{3:t} = \frac{\frac{n}{N} \cdot {}_NC_n \cdot \Sigma x^3 + 3 \cdot \frac{{}_nC_2}{{}_NC_2} \cdot {}_NC_n \cdot \Sigma x^2x + 6 \cdot \frac{{}_nC_3}{{}_NC_3} \cdot {}_NC_n \cdot \Sigma xxx}{{}_NC_n}$$

$$= \frac{n}{N} \cdot \Sigma x^3 + 3 \cdot \frac{{}_nC_2}{{}_NC_2} \cdot \Sigma x^2x + 6 \cdot \frac{{}_nC_3}{{}_NC_3} \cdot \Sigma xxx$$

It is noted here that $\Sigma x^2 \cdot \Sigma x = \Sigma x^3 + \Sigma x^2x$ so that $\Sigma x^2x = \Sigma x^2 \cdot \Sigma x - \Sigma x^3$, and that:

$$(\Sigma x)^3 = \Sigma x^3 + 3 \Sigma x^2x + 6 \Sigma xxx$$

$$= 3 \cdot \Sigma x^2 \cdot \Sigma x - 2 \Sigma x^3 + 6 \Sigma xxx$$

so that $6 \Sigma xxx = (\Sigma x)^3 + 2 \Sigma x^3 - 3 \Sigma x^2 \Sigma x$. These values may be substituted to obtain:

$$V_{3:t} = \frac{n}{N} \cdot \Sigma x^3 + 3 \cdot \frac{{}_nC_2}{{}_NC_2} \left[\Sigma x^2 \Sigma x - \Sigma x^3 \right]$$

$$+ \frac{{}_nC_3}{{}_NC_3} \left[(\Sigma x)^3 + 2 \Sigma x^3 - 3 \Sigma x^2 \Sigma x \right]$$

$$= \left[n - \frac{3n(n-1)}{N-1} + \frac{n(n-1)(n-2)}{(N-1)(N-2)} \right] V_{3:x}$$

$$+ 3 \left[\frac{n(n-1)}{1-\frac{1}{N}} - \frac{n(n-1)(n-2)}{\left(1-\frac{1}{N}\right)\left(N-2\right)} \right] V_{2:x} \cdot V_{1:x}$$

$$+ \left[\frac{n(n-1)(n-2)}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)} \right] V_1{}^3{}_{:x}$$

Letting N approach infinity, the limiting value becomes:

$$V_{3:t} = n \cdot V_{3:x} + 3n(n-1)\ V_{2:x} \cdot V_{1:x} + n(n-1)\ (n-2)\ V_1{}^3{}_{:x}$$

and

$$U_{3:t} = V_{3:t} - 3\ V_{2:t}\ \cdot V_{1:t} + 2\ V_1{}^3{}_{:t}$$

$$= n\ [V_{3:x} - 3\ V_{2:x} \cdot V_{1:x} + 2\ V_1{}^3{}_{:x}]$$

$$= n \cdot U_{3:x}$$

This gives as the skewness (Charlier) of the t distribution:

$$\alpha_{3:t} = \frac{n \cdot U_{3:x}}{(n \cdot U_{2:x})^{3/2}} = \frac{a_{3:x}}{\sqrt{n}}$$

D. Sampling Distribution of Total Losses, Pure Premiums, and Loss Ratios

(1) In the form of the ratio of actual to expected value, the exposure divisor of the pure premiums and the premium divisor of the loss ratios cancel out, leaving only the ratio of actual to expected total losses. Thus only a single sampling distribution is required. Furthermore, as the expected total losses is a constant, the moments of this ratio, R, can be obtained directly from the moments of the total losses, T.

The total expectation is the sum, taken over all possibilities, of the product of the probability of an event occurring and the expectation if the event occurs. The average values of the first three powers of the total cost, t, t^2, and t^3, have been obtained in the previous section for a fixed number of claims. The Poisson Distribution will be assumed to give the probabilities of obtaining 0, 1, 2, ..., n, etc. claims. The sum of the products of these will be the first three moments, about the origin, of T, the total cost when c claims are expected.

From the totals on Table 2 (see page 82) are obtained:

$$V_{1:T} = c \cdot V_{1:x}$$

$$V_{2:T} = c \cdot V_{2:x} + c^2 V_1{}^2{}_{:x}$$

$$V_{3:T} = c \cdot V_{3:x} + 3\, c^2 V_{2:x} \cdot V_{1:x} + c^3 V_1{}^3{}_{:x}$$

The moments of R, the ratio of actual to expected losses, are then obtained by dividing these by the first three powers respectively of $c \cdot V_{1:x}$, the expected loss, as:

$$V_{1:R} = 1$$

$$V_{2:R} = \frac{V_{2:x}}{c \cdot V_1{}^2{}_{:x}} + 1$$

$$V_{3:R} = \frac{V_{3:x}}{c^2 V_1{}^3{}_{:x}} + 3\,\frac{V_{2:x}}{c \cdot V_1{}^2{}_{:x}} + 1$$

and:

$$U_{2:R} = \frac{V_{2:x}}{c \cdot V_1{}^2{}_{:x}}$$

$$U_{3:R} = \frac{V_{3:x}}{c^2 V_1{}^3{}_{:x}}$$

with the skewness (Charlier) of:

$$a_{3:R} = \frac{V_{3:x}}{\sqrt{c}\,(V_{2:x})^{3/2}}$$

(2) Under certain conditions we may wish to exclude, or may not have available, the cases for which there were no actual losses. As the proportion of such cases to the total will be e^{-c}, then, designating this select set of ratios by R^1:

$$V_{n:R} = 0\,(e^{-c}) + V_{n:R^1}\,(1 - e^{-c})$$

and

$$V_{n:R^1} = \frac{V_{n:R}}{1 - e^{-c}}$$

Thus the moments of R^1, the ratio of actual to expected losses when cases with no actual losses are excluded are found to be:

$$V_{1:R^1} = \frac{1}{1 - e^{-c}}$$

$$V_{2:R^1} = \frac{1}{1-e^{-c}}\left[\frac{V_{2:x}}{c\,V_1{}^2{}_{:x}} + 1\right]$$

$$V_{3:R^1} = \frac{1}{1-e^{-c}}\left[\frac{V_{3:x}}{c^2\,V_1{}^3{}_{:x}} + 3\,\frac{V_{2:x}}{c\,V_1{}^2{}_{:x}} + 1\right]$$

and

$$U_{2:R^1} = \frac{1}{1-e^{-c}}\left[\frac{V_{2:x}}{c\cdot V_1{}^2{}_{:x}} - \frac{e^{-c}}{1-e^{-c}}\right]$$

$$U_{3:R^1} = \frac{1}{1-e^{-c}}\left[\frac{V_{3:x}}{c^2\cdot V_1{}^3{}_{:x}} - \frac{3\,e^{-c}\cdot V_{2:x}}{(1-e^{-c})\,c\cdot V_1{}^2{}_{:x}} + \frac{e^{-c}\,(1+e^{-c})}{(1-e^{-c})^2}\right]$$

E. The Sampling Distribution of the Average Claim Cost

The first three moments, about the origin, of the total cost, t, of a fixed number of claims were determined in section C as:

$$V_{1:t} = {}_nV_{1:x}$$

$$V_{2:t} = {}_nV_{2:x} + n(n-1)\,V_1{}^2{}_{:x}$$

$$V_{3:t} = {}_nV_{3:x} + 3\,n(n-1)\,V_{2:x}\,V_{1:x} + n(n-1)(n-2)\,V_1{}^3{}_{:x}$$

The first three moments, about the origin, of the average claim cost, a, of a fixed number of claims can be obtained directly from these by dividing respectively by the first three powers of n:

$$V_{1:a} = V_{1:x}$$

$$V_{2:a} = \frac{V_{2:x}}{n} + \frac{n-1}{n}\,V_1{}^2{}_{:x}$$

$$V_{3:a} = \frac{V_{3:x}}{n^2} + 3\,\frac{n-1}{n^2}\,V_{2:x}\,V_{1:x} + \frac{(n-1)(n-2)}{n^2}\,V_1{}^3{}_{:x}$$

The first three moments, about the origin, of the ratio of the actual average claim cost to the expected average claim cost, s, of a fixed number of claims can then be obtained by dividing these respectively by the first three powers of $V_{1:x}$, the expected average claim cost.

$$V_{1:s} = 1$$

$$V_{2:s} = \frac{1}{n}\cdot\frac{V_{2:x}}{V_1{}^2{}_{:x}} + \frac{n-1}{n}$$

$$V_{3:s} = \frac{1}{n^2} \cdot \frac{V_{3:x}}{V_1{}^3{}_{:x}} + 3\,\frac{n-1}{n^2} \cdot \frac{V_{2:x}}{V_1{}^2{}_{:x}} + \frac{(n-1)(n-2)}{n^2}$$

and

$$U_{2:s} = \frac{1}{n}\,\frac{U_{2:x}}{V_1{}^2{}_{:x}}$$

$$U_{3:s} = \frac{1}{n^2}\,\frac{U_{3:x}}{V_1{}^3{}_{:x}}$$

with the skewness (Charlier) of:

$$\alpha_{3:s} = \frac{\alpha_{3:x}}{\sqrt{n}}$$

Combining these with the probabilities of 1, 2, ... , n, etc., claims occurring from the Poisson Distribution, (note that the cases where no losses occur are excluded) the first three moments of the ratio of actual average claim cost to the expected average, for all possible number of claims are obtained as:

$$V_1{:}s = \frac{\sum_{n=1}^{\infty} (1)\left(\frac{c^n}{\lfloor n e^c}\right)}{\sum_{n=1}^{\infty} \frac{c^n}{\lfloor n e^c}} = \frac{1 - \frac{1}{e^c}}{1 - \frac{1}{e^c}} = 1$$

$$V_2{:}s = \frac{\sum_{n=1}^{\infty} \left(\frac{1}{n}\frac{V_{2:x}}{V_1{}^2{}_{:x}} + \frac{n-1}{n}\right)\left(\frac{c^n}{\lfloor n e^c}\right)}{\sum_{n=1}^{\infty} \frac{c^n}{\lfloor n e^c}}$$

$$= \frac{\left(1 - \frac{1}{e^c}\right) + \left(\frac{V_{2:x}}{V_1{}^2{}_{:x}} - 1\right)\left(1 - \frac{1}{e^c}\right)K_{(c)}}{1 - \frac{1}{e^c}} = 1 + K_{(c)}\left(\frac{V_{2:x}}{V_1{}^2{}_{:x}} - 1\right)$$

where $$K_{(c)} = \frac{\sum_{n=1}^{\infty} \frac{c^n}{n \lfloor n\, e^c}}{1 - \frac{1}{e^c}} = \frac{\sum_{n=1}^{\infty} \frac{c^n}{n \lfloor n}}{e^c - 1}$$

NOTE: The only method of determining the values of $K_{(c)}$ and of $G_{(c)}$ below is that of laboriously calculating each term of the series and adding them together. For large values of c an approximation is available as shown in the table at the end of this section.

$$V_{3:s} = \frac{\sum_{n=1}^{\infty} \left(\frac{1}{n^2}\frac{V_{3:x}}{V_1{}^3{}_{:x}} + 3\frac{n-1}{n^2}\frac{V_{2:x}}{V_1{}^2{}_{:x}} + \frac{(n-1)(n-2)}{n^2}\right)\left(\frac{c^n}{\lfloor n e^c}\right)}{\sum_{n=1}^{\infty} \frac{c^n}{ne^c}}$$

$$= \frac{\left(1-\frac{1}{e^c}\right) + 3\left(\frac{V_{2:x}}{V_1{}^2{}_{:x}} - 1\right)\left(1-\frac{1}{e^c}\right)K_{(c)} + \left(\frac{V_{3:x}}{V_1{}^3{}_{:x}} - 3\frac{V_{2:x}}{V_1{}^2{}_{:x}} + 2\right)\left(1-\frac{1}{e^c}\right)G_{(c)}}{1-\frac{1}{e^c}}$$

$$= 1 + 3\left(\frac{V_{2:x}}{V_1{}^2{}_{:x}} - 1\right)K_{(c)} + \left(\frac{V_{3:x}}{V_1{}^3{}_{:x}} - 3\frac{V_{2:x}}{V_1{}^2{}_{:x}} + 2\right)G_{(c)}$$

where

$$G_{(c)} = \frac{\sum_{n=1}^{\infty} \frac{c^n}{n^2 \lfloor n}}{e^c - 1}$$

The moments of this ratio of actual to expected average claim cost about the mean then reduce to:

$$U_2{:}s = K_{(c)} \cdot \left(\frac{V_{2:x}}{V_1{}^2{}_{:x}} - 1\right) = K_{(c)} \cdot \frac{U_{2:x}}{V_1{}^2{}_{:x}}$$

$$U_3{:}s = G_{(c)} \cdot \left(\frac{V_{3:x}}{V_1{}^3{}_{:x}} - 3\frac{V_{2:x}}{V_1{}^2{}_{:x}} + 2\right)$$

and the skewness to:

$$\alpha_{3:s} = \frac{G_{(c)}}{[K_{(c)}]^{3/2}} \cdot \alpha_{3:x}$$

Values of $K_{(c)}$ and $G_{(c)}$

C	$K_{(c)}$	$G_{(c)}$	$\frac{G_{(c)}}{[K_{(c)}]^{3/2}}$
1	.766988	.667235	.993335
4	.329627	.157766	.833642
10	.113021	.015322	.403253
40	.025659	.000677	.165
For larger values of c	$\frac{1}{c-1}$	$\frac{1}{(c-1)(c-2)}$	$\frac{\sqrt{c-1}}{c-2}$

F. A Useful Function of Actual and Expected Losses

In section D of this chapter the moments of the distribution of R (the ratio of actual to expected total losses, pure premiums, or loss ratios) were found to involve the expected number of claims, c, in both second and third moments. In many cases the available data consists of that for individual risks or classifications with c having a different value for each observation. In analyzing such data, the function of actual and expected losses, which is described below, will be found useful. Although its form is such that its first and second moments do not involve the value of c, its value in practical use will not be found to arise from this fact alone. It will be largely due to the effective weighting factor of unity for each observation in the suggested function as contrasted to an effective weighting factor of $\frac{1}{E}$ in the R function, which, therefore, exaggerates the influence of the small risk or small classification experience. The suggested function is:

$$Z = \frac{A-E}{\sqrt{E}} = \left(\frac{A}{E} - 1\right)\sqrt{E} = (R-1)\sqrt{E} = (R-1)\sqrt{c \cdot V_{1:x}}$$

The moments of Z can be determined in terms of the moments of R from the relationship $Z = (R-1)\sqrt{C \cdot V_{1:x}}$ and in terms of the moments of x by substituting the values of the moments of R in terms of those of x as follows:

$$V_{1:Z} = \sqrt{C \cdot V_{1:x}}\,(V_{1:R} - 1) = 0$$

$$V_{2:Z} = C \cdot V_{1:x}\,(V_{2:R} - 2\,V_{1:R} + 1) = \frac{V_{2:x}}{V_{1:x}}$$

$$V_{3:Z} = (C \cdot V_{1:x})^{3/2} \cdot (V_{3:R} - 3\,V_{2:R} + 3\,V_{1:R} - 1) = \frac{V_{3:x}}{\sqrt{C}\,(V_{1:x})^{3/2}}$$

from which:

$$U_{2:Z} = U_{2:R}\,(C \cdot V_{1:x}) = \frac{V_{2:x}}{V_{1:x}}$$

$$\sigma_Z = \sigma_R \sqrt{C \cdot V_{1:x}} = \sqrt{\frac{V_{2:x}}{V_{1:x}}}$$

and

$$U_{3:Z} = U_{3:R}\,(C \cdot V_{1:x})^{3/2} = \frac{V_{3:x}}{\sqrt{C}\,(V_{1:x})^{3/2}}$$

$$\alpha_{3:Z} = \alpha_{3:R} = \frac{V_{3:x}}{\sqrt{C}\,(V_{2:x})^{3/2}}$$

As the moments of R will most often be required in terms of the moments of Z the reverse of these relationships will be given. They are:

$$V_{1:R} = \frac{V_{1:Z}}{\sqrt{C \cdot V_{1:x}}} + 1$$

$$V_{2:R} = \frac{V_{2:Z}}{C\,V_{1:x}} + \frac{2\,V_{1:Z}}{\sqrt{C \cdot V_{1:x}}} + 1$$

$$V_{3:R} = \frac{V_{3:Z}}{(C \cdot V_{1:x})^{3/2}} + \frac{3\,V_{2:Z}}{C\,V_{1:x}} + \frac{3\,V_{1:Z}}{\sqrt{C\,V_{1:x}}} + 1$$

$$U_{2:R} = \frac{U_{2:Z}}{C\,V_{1:x}}$$

$$U_{3:R} = \frac{U_{3:Z}}{(C\,V_{1:x})^{3/2}}$$

$$a_{3:R} = a_{3:Z}$$

It will be noted that, although the first two moments of Z are independent of the amounts of expected losses, that the third moment and $a_{3:Z}$ are still functions of the expected loss ($E = C \cdot V_{1:x}$). As the value of $V_{3:Z}$, as calculated from observations having different values of E, will actually be of the form:

$$V_{3:Z} = \frac{V_{3:X}}{V_{1:x}} \cdot (\text{Average value of } \frac{1}{\sqrt{E}} \text{ in the actual observations})$$

it will be necessary, in order to obtain the value of $V'_{3:Z}$ corresponding to a particular value of E (indicated as E'), to make the adjustment:

$$V'_{3:Z} = \frac{(V_{3:Z} \text{ as calculated from the observations})}{\sqrt{E'} \cdot (\text{Average value of } \frac{1}{\sqrt{E}} \text{ in the actual observations})}$$

It will usually be expedient to make this adjustment directly to the value of $a_{3:Z}$ as:

$$a'_{3:Z} = \frac{(a_{3:Z} \text{ as calculated from the observations})}{\sqrt{E'} \cdot (\text{Average value of } \frac{1}{\sqrt{E}} \text{ in the actual observations})}$$

The "average value of $\frac{1}{\sqrt{E}}$ in the actual observations" would involve a very considerable amount of work to calculate exactly but it can be approximated with an accuracy sufficient for most purposes by the separation of the observations into ten or more groups according to the size of the expected loss, E. For a rough approximation the average $\frac{1}{\sqrt{E}}$ for each group would be assumed to be $\frac{1}{\sqrt{\text{Average } E}}$ and these values weighted by the number of observations in the groups. A closer approximation can be obtained by correcting this estimate, $\frac{1}{\sqrt{\text{Average } E}}$, by the factor $\frac{\sqrt{2r+2}}{1+\sqrt{r}}$, where r is the ratio of the highest value of E in the group to the lowest value of E in the group.

G. The Excess Pure Premium Ratio in Terms of the Loss Ratio Distribution

The excess pure premium ratio (for a loss ratio of B, a premium per risk of P, and a permissible or expected average loss ratio of L) is defined as the ratio of the amount of losses which are expected to be in excess of BP per risk to the total of all expected losses. As the permissible loss ratio is subject to many arbitrary changes, it would seem advisable to eliminate it from the theoretical considerations as well as to construct tables of the excess pure premiums which would be independent of the permissible loss ratio.

This can be done by recognizing the excess pure premium (for a ratio of actual to expected losses of R', and an expected loss per risk of E) as the ratio of the amount of losses which are expected to be in excess of $R'E$ per risk to the total of all expected losses. This can be expressed symbolically as:

$$\chi(R', E) = \frac{\sum_{A=R'E}^{\infty} (A - R'E)}{\sum_{A=0}^{\infty} E}$$

where A represents an actual loss per risk and E the corresponding expected loss per risk.

It should be noted here that there is no specific qualification that the sum of all A's be the same as the sum of all E's. The only conditions necessary to obtain the proper excess pure premium for practical application are that all values of E in this equation are identical and that the values of A in the equation are those which are expected to occur.

As all values of E are the same, we can divide the numerator and denominator by E obtaining:

$$\chi_{(R',\,E)} = \frac{\sum_{A=R'}^{\infty} E\left(\frac{A}{E} - R'\right)}{\sum_{A=0}^{\infty}{}^{(1)}} = \frac{\sum_{R=R'}^{\infty}{}^{(R)} \quad - \quad R' \sum_{R=R'}^{\infty}{}^{(1)}}{\sum_{R=0}^{\infty}{}^{(1)}}$$

Insofar as our expectations are concerned the values that R may take form a continuous function for each value of which the probability that R may take such a value is $F_{(R)}$. Thus, in terms of these probabilities, the excess pure premium ratio is:

$$\chi_{(R',\,E)} = \frac{\int_{R=R'}^{\infty} R \cdot F_{(R)} \cdot dR - R' \cdot \int_{R=R'}^{\infty} F_{(R)} \cdot dR}{\int_{R=0}^{\infty} F_{(R)} \cdot dR}$$

As R may not be negative, we recognize that:

$$\int_{R=0}^{\infty} F_{(R)} \cdot dR = 1,$$

and,

$$\int_{R=R'}^{\infty} R \cdot F_{(R)} \cdot dR = \int_{R=0}^{\infty} R \cdot F_{(R)} \cdot dR - \int_{R=0}^{R'} R \cdot F_{(R)} \cdot dR$$

and as $\int_{R=0}^{\infty} R \cdot F_{(R)} \cdot dR$ is the first moment of R about the origin, then:

$$\int_{R=R'}^{\infty} R \cdot F_{(R)} \cdot dR = V_{1:R} - \int_{R=0}^{R'} R \cdot F_{(R)} \cdot dR$$

Making these substitutions, we have:

$$\chi_{(R',\,E)} = V_{1:R} - \int_{R=0}^{R'} R \cdot F_{(R)} \cdot d\,R - R' \cdot \int_{R=R'}^{\infty} F_{(R)} \cdot d\,R$$

From the formula for integration by parts, we have:

$$\int_{R=0}^{R'} R \cdot F_{(R)} \cdot d\,R = R' \int_{R=0}^{R'} F_{(R)} \cdot d\,R - \int_{R=0}^{R'} \int_{R=0}^{R'} F_{(R)} \cdot d\,R \cdot d\,R$$

which gives us:

$$\chi_{(R',\,E)} = V_{1:R} - R' + \int_{R=0}^{R'} \int_{R=0}^{R'} F_{(R)} \cdot d\,R \cdot d\,R$$

as the actual functional form of the excess pure premium ratio which can then be put in the form of:

$$\chi_{(B,\,P,\,L)} = \frac{V_{1:B}}{L} - \frac{B}{L} + \int_{R=0}^{B/L} \int_{R=0}^{B/L} F_{(R)} \cdot d\,R \cdot d\,R$$

H. The Loss Elimination Ratio in Terms of the Distribution of Individual Losses

The loss elimination ratio, or "K" value, used in determining rates or discounts for deductible insurance, is (for an assured's retention of B dollars) the ratio to total losses of the total of the first $\$B$ of each loss. Thus:

$$K = \frac{(\text{All Losses of less than } \$B) + B \cdot (\text{Number of Losses over } \$B)}{\text{Total of All Losses}}$$

In terms of the distribution of individual losses by size of loss, this becomes:

$$K = \frac{\int_0^B x \cdot f(x) \cdot d\,x + B \int_B^{\infty} f(x) \cdot d\,x}{\int_0^{\infty} x \cdot f(x) \cdot d\,x}$$

which through the substitution of:

$$\int_0^B x \cdot f(x) \cdot dx = B \int_0^B f(x) \cdot dx - \int_0^B \int_0^B f(x) \cdot dx \cdot dx$$

becomes:

$$K = \frac{B\int_0^\infty f(x) \cdot dx - \int_0^B \int_0^B f(x) \cdot dx \cdot dx}{\int_0^\infty x \cdot f(x) \cdot dx}$$

and recognizing $\int_0^\infty x \cdot f(x) \cdot dx$ as $V_{1:x}$ and $\int_0^\infty f(x) \cdot dx$ as unity, we have

$$K = \frac{B - \int_0^B \int_0^B f(x) \cdot dx \cdot dx}{V_{1:x}}$$

as the functional form of the loss elimination ratio.

I. The Fundamentals of Experience Rating

For the purpose of this paper, experience rating will be defined as a procedure to obtain, on the average, better estimates of the inherent hazard of the coverage provided individual risks than that represented by the premium at manual rates. This definition must be recognized as being entirely different from one that would include all methods of partial "self-rating." Many such methods produce premium charges that, on the average, represent poorer estimates of the hazard than the original premium at manual rates.

Obviously, in order for experience rating to be necessary, there must exist either a demonstrable difference in the inherent hazards of risks not ade-

quately measured by the manual rating procedure or appreciable errors in the manual rates. The other basic premise is that the accuracy of an experience rating procedure must be judged on a percentage basis. Otherwise a $100 error on a $1,000 risk could be offset by a $100 error on a $10,000 risk. Because of their simplicity and their firm foundation in practice, the following outline of the experience rating procedure will be based on the linear regression formulae resulting from the methods of least squares.

Figure 1. will help to visualize the experience rating process on this basis. It is representative of all risks for which the premium at manual rates is a specified amount, P, contemplating an expected loss of E'. For such risks the true inherent hazards are represented by $E'(1+m)$, where m varies from risk to risk. The ratio of the true inherent hazard, $E'(1+m)$, to the contemplated hazard, E', is then equal to $(1+m)$, which is measured on the vertical axis. An assumed frequency distribution of risks according to the value of $(1+m)$ is shown along this axis.

The ratio of the actual losses of the risk, A, to E', the expected losses contemplated by the premium at manual rates, is represented by R' and is measured along the horizontal axis. For all risks having a manual premium of P, the frequency distribution of risks according to the value of R' will be a skew distribution such as shown along the horizontal axis (except for very large values of P, when this distribution may be even skew in the other direction). The resulting frequency surface of $(1+m)$ and R' will be approximately as shown by the contour lines.

One very important characteristic of such a frequency surface is that the regression line of R' on $(1+m)$ is always the line: $R' = (1+m)$. This is evident from a consideration of the risks having a particular value of m. For such risks the true inherent hazard is $E'(1+m)$. For risks with such a true inherent hazard the average of the actual losses will be $E'(1+m)$ and the average ratio of the actual losses to the expected losses contemplated by the premium at manual rates will be $(1+m)$. As the regression line of y on x is the straight line, if such a straight line exists, passing through the mean values of y for each particular value of x, then the regression line of R' on $(1+m)$ is the line $R' = (1+m)$.

The regression line of y on x has the formula:

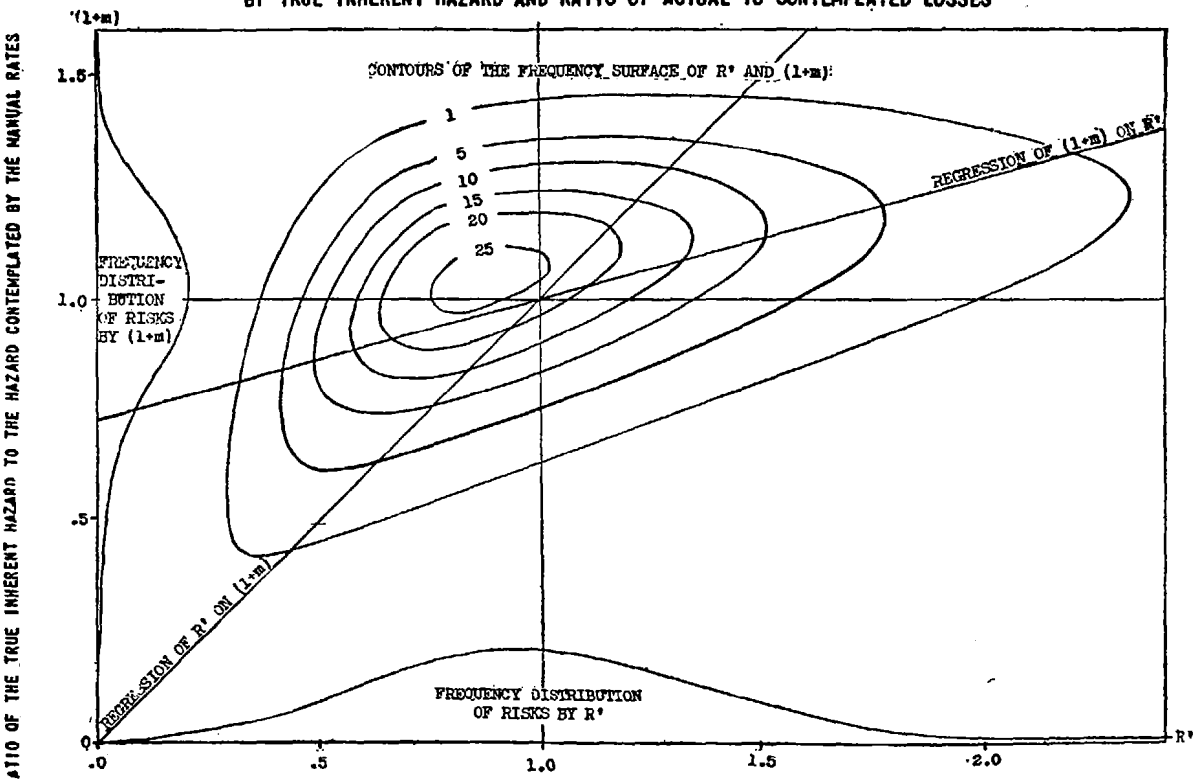
$$y = \left(r_{xy} \cdot \frac{\sigma_y}{\sigma_x}\right) x + \left(V_{1:y} - r_{xy} \cdot \frac{\sigma_y}{\sigma_x} \cdot V_{1:x}\right)$$

in terms of the coefficient of linear correlation between x and y, r_{xy}, and the standard deviations of x and y, σx and σy. This gives us first that

$$r_{R'(1+m)} \cdot \frac{\sigma_{R'}}{\sigma_{(1+m)}} = 1 \text{ or } r_{R'(1+m)} = \frac{\sigma_{(1+m)}}{\sigma_{R'}}$$

FIGURE 1.

DISTRIBUTION OF RISKS, HAVING PREMIUM AT MANUAL RATES OF P CONTEMPLATING LOSSES OF E' BY TRUE INHERENT HAZARD AND RATIO OF ACTUAL TO CONTEMPLATED LOSSES



and secondly that:

$$V_{1:R'} - V_{1:(1+m)} = 0 \text{ or } V_{1:R'} = V_{1:(1+m)}.$$

We can place these values directly in the equation of the other regression line, that of $(1+m)$ on R', of:

$$(1+m) = \left(r_{R'(1+m)}\frac{\sigma_{(1+m)}}{\sigma_{R'}}\right)\cdot R' + \left(V_{1:(1+m)} - r_{R'(1+m)}\frac{\sigma_{(1+m)}}{\sigma_{R'}}\cdot V_{1:R'}\right)$$

to obtain:

$$(1+m) = \left(\frac{U_{2:(1+m)}}{U_{2:R'}}\right)R' + \left(1 - \frac{U_{2:(1+m)}}{U_{2:R'}}\right)V_{1:(1+m)}$$

As $U_{2:(1+m)} = U_{2:m}$ and, if the rate level is assumed to be correct, $V_{1:(1+m)} = 1$, we have as the regression line of $(1+m)$ on R' on a percentage basis:

$$(1+m) = \left(\frac{U_{2:m}}{U_{2:R'}}\right)R' + \left(1 - \frac{U_{2:m}}{U_{2:R'}}\right)$$

and, by multiplying through by E'

$$E'(1+m) = \left(\frac{U_{2:m}}{U_{2:R'}}\right)A + \left(1 - \frac{U_{2:m}}{U_{2:R'}}\right)E'$$

or: Estimated True Inherent Hazard $= Z\cdot A + (1-Z)\cdot E'$, which is recognized as the typical experience rating formula. Furthermore, if we define K as being equal to $E'\left(\frac{U_{2:R'}}{U_{2:m}} - 1\right)$ then $Z = \frac{E'}{E'+K}$

and we have the well established credibility formula originally suggested by Mr. Greene* as a practical approximation to the more complicated formula developed by Mr. Whitney.**

In order to evaluate K let us assume for the moment that, as a result of chance variation only, the actual losses are distributed, for risks for which the true expected losses are E, in such a way that the first two moments about the origin are: $V_{1:T} = E$, and $V_{2:T} = H\cdot E + E^2$. Then, for risks having an inherent hazard of $E'(1+m)$, the corresponding moments would be: $V_{1:T} = E'(1+m)$ and $V_{2:T} = H\cdot E'(1+m) + E'^2(1+m)^2$. The ratios of these actual losses to E', the losses contemplated by the premium at manual rates, would be: $V_{1:R'} = (1+m)$ and $V_{2:R'} = \frac{H\cdot(1+m)}{E'} + (1+m)^2$, for a particular value of m.

* P.C.A.S., Vol. V, page 133.
** P.C.A.S., Vol. IV, page 274.

Averaging these moments for all values of m, but only for risks having losses contemplated by the premium at manual rates of E', we will obtain: $V_{1:R'} = 1$ and $V_{2:R'} = \frac{H}{E'} + 1 + U_{2:m}$ if it is again assumed that the rate level is correct so that $V_{1:m} = 0$. This gives us $U_{2:R'} = \frac{H}{E'} U_{2:m}$ which can be substituted in the formula for K to obtain: $K = \frac{H}{U_{2:m}}$.

Returning to our original supposition involving H we find that:

$$H = \frac{V_{2:T} - E^2}{E} = \frac{V_{2:T} - V_1{}^2{}_{:T}}{E} = \frac{U_{2:T}}{E} = E \cdot U_{2:R} = U_{2:Z}$$

which gives us as the final values of K:

$$K = \frac{U_{2:T}}{E \cdot U_{2:m}} = \frac{E \cdot U_{2:R}}{U_{2:m}} = \frac{U_{2:Z}}{U_{2:m}}$$

It must be definitely understood here that while the second moments of T, R, and Z in the formula for K are the measures of chance variation only. they measure the chance variation of all risks. Thus K is not necessarily a constant but will vary between classifications for at least three reasons: (1) variation in the accuracy of the manual rate, and (2) variation in the diversity of the inherent hazard of risks in the classification, both of which are jointly measured by $U_{2:m}$, and (3) variation in the relative hazards of the classifications as measured by $\frac{V_{2:x}}{V_{1:x}}$. Variation of K by size of E' will also occur as a result of both (2) and (3) as well as a result of variation in loss frequencies within or between classifications. These variations will be studied in subsequent chapters.

For the special case of a group of classifications for which the manual rates are incorrect, but in all of which all risks have the same distribution of losses by size of loss and have the same expected frequency of loss per unit of exposure, we have, by using the values of $U_{2:T}$, $U_{2:R}$, or $U_{2:Z}$ obtained in previous sections of this chapter:

$$K = \frac{V_{2:x}}{V_{1:x}} \cdot \frac{1}{U_{2:m}}$$

as in this case $U_{2:m}$ measures only the errors in the manual rates.

II.

Preparation of Tables of the Normal Sampling Range Due to Chance Fluctuations Only

A. Number of Claims or Claim Frequencies

In other lines of statistical analysis, tables of the normal range of values to be expected as a result of the sampling variation are found to be valuable aids in the interpretation of the significance of observed data. Such a table for the ratio of actual to expected number of claims or claim frequencies would be universally applicable to any line of casualty insurance for which the assumption of the Poisson Distribution is valid.

In Table 3, the probabilities from the Poisson Distribution of obtaining 0, 1, 2, ..., etc., claims are calculated for the values of c, the expected number of claims, of 1, 4, 10, and 40. The ogives of these probabilities are also shown representing the probability of n or less claims occurring. The values of the ratio of actual to expected number of claims corresponding to the .005, .025, and .050 points on these ogives are entered in Table 5.

As c increases, the labor involved in this calculating procedure becomes prohibitive and recourse to an approximation is made. As will be pointed out later, it is believed that this approximation produces the correct result to the number of digits retained in Table 5.

For values of c above 40, the skewness of this sampling distribution $\left(=\frac{1}{\sqrt{c}}\right)$ is comparatively small although significant. For these values of c the Poisson Distribution is closely approximated by the Type III frequency distribution. The ogives of the Type III distribution have been tabulated for 1/10th intervals of skewness.*

In Table 4, the values of the abcissas, measured in standard deviational units from the mean, corresponding to the .005, .025, and .050 points on the ogives, are shown. These are values interpolated from the tables corresponding to the required skewness. The values in Table 5 are calculated directly from these by multiplying by the standard deviation and adding the mean to produce the required results in the scale of the ratio of actual to expected values.

An indication of the accuracy of the approximation of the Poisson Distribution by the Type III distribution is obtained by comparing the values entered

* Note: Although the writer used the tables given in "Introduction to Mathematical Statistics," by J. W. Glover and H. C. Carver, published in mimeograph form in 1926 by Edwards Brothers, Ann Arbor, Michigan, these tables are understood to be available in Volume 2 of the "Annals of Mathematical Statistics" in a paper by L. R. Salvosa.

in Table 5 for c equal to 1, 4, 10, and 40 with the values which would have been obtained from the Type III distribution as calculated in Table 4. For the lower and upper 2½% points, the comparison is:

c	Value of Ratio Corresponding to a Probability of a Lesser Value Occurring of:			
	.025		.975	
	Poisson	Type III	Poisson	Type III
1	.000	—1.000	3.000	3.000
4	.250	.250	2.000	2.000
10	.400	.400	1.700	1.700
40	.700	.700	1.325	1.325

B. Total Losses, Pure Premiums, and Loss Ratios

(1) The various distributions of claims by size of claim are uniform in that they all exhibit a concentration of frequency at the low amounts with a tapering off of the frequencies up to and including very high amounts. This produces a skewness far in excess of that usually encountered in a study of frequency distributions. The only type of theoretical frequency distribution which has been found to fit these distributions of claims by size is the Normal Logarithmic Distribution. Tests of the goodness of fit of this type of distribution have indicated that, except for the concentration of claims at such round-figures values as $50, $100, $500, and $1000, the departures of the actual distributions from the Normal Logarithmic are not greater than would frequently occur in samples of the size tested. (See Table 6 for an example of procedure of fitting such a distribution and the test as to its goodness of fit.)

The only condition necessary to produce a Normal Logarithmic Distribution is that the amount of an observed value be the product of a large number of factors, each of which is independent of the size of any other factor. Reflection as to the conditions entering into the determination of the amount of a claim settlement in casualty insurance, the variations in the seriousness of accidents for which claims are made, and all of the factors eventually recognized in making the final settlement makes it apparent that the necessary condition is at least approximated in the data with which we are concerned. When this condition is met, the logarithms of the observations become the sum of a large number of independent elements, which is the only condition necessary to result in a Normal Distribution. Thus, we shall expect to find the logarithms of the claim amounts normally distributed.

The generalized Normal Logarithmic Distribution, which we shall use, provides an additional degree of freedom in fitting the actual conditions by assuming that only the amount of all observations over and above a fixed amount are distributed in the manner described. Thus, if x represents the

amount of a claim and "a" this fixed amount, it will be assumed that log $(x - a)$ is normally distributed with a mean of l_0 and a standard deviation of σ_e. The quantity:

$$\frac{\log (x - a) - l_0}{\sigma_e}$$

will then be distributed normally with a mean of zero and a standard deviation of unity, permitting the use of available tables of the integral of the normal distribution in fitting this type of distribution to the observed distributions.

Although the necessary transformation from the original scale of observations to the logarithmic distribution is not difficult, the determination of the constants, a, l_0, and σ_e, from the moments of the observed distribution is quite involved. S. D. Wicksell has derived the procedure for the determination of the constants as follows:

If $$s = \frac{-a_3}{2}$$

and $$\eta = \sqrt[3]{-s + \sqrt{s^2 + 1}} + \sqrt[3]{-s - \sqrt{s^2 + 1}}$$

then $$a = M - \frac{\sigma}{\eta},$$

$$l_0 = \log_{10} \frac{(M - a)^2}{\sqrt{U_2 + (M - a)^2}}, \text{ and}$$

$$\sigma_e = \sqrt{2 (\log_{10} e [\log_{10} (M - a) - l_0]}$$

$$= \sqrt{.868589 [\log_{10} (M - a) - l_0]}$$

where: M, σ, U_2, and a_3 represent the Mean, Standard Deviation, the second moment about the mean, and the skewness, respectively, of the distribution to be fitted.

(2) The standard deviation and skewness of the sampling distribution of the ratio of actual to expected losses, pure premiums, and loss ratios corresponding to a particular value of the expected loss, $E = c \cdot V_{1:x}$, are proportional to $\sqrt{\frac{V_{2:x}}{V_{1:x}}}$ and $\frac{V_{3:x}}{V_{2:x}} \div \sqrt{\frac{V_{2:x}}{V_{1:x}}}$ respectively; functions of the distribution of claims by size of claim. The values of these functions vary by line of insurance and may vary by classification or territory. The extent of the variation by line of insurance is shown in the following table, which gives the values for several of the casualty lines calculated from the distribution of claims by size group as reported under the official calls for New York State experience.

Line of Insurance	Coverage	Classifications	$\sqrt{\frac{V_{2:x}}{V_{1:x}}}$	$\frac{V_{3:x}}{V_{2:x}} \div \sqrt{\frac{V_{2:x}}{V_{1:x}}}$
Workmen's Compensation		All	56.89	165.53
Automobile	B.I.	Priv. Pass.	50.34	148.27
"	"	Commercial	55.93	164.32
"	P.D.	"	11.84	60.81
Manufacturers' and Contractors	B.I.	All	70.96	247.34
Manufacturers' and Contractors	P.D.	"	38.86	108.64
Owners', Landlords' and Tenants'	B.I.	Excl. N. Y. C. Apts. and Tenements	36.65	145.44
Product	B.I.	Foodstuffs	15.25	71.93
"	"	All Others	49.17	166.07

To indicate in detail the advocated procedure of calculating the desired table of the normal sampling range, property damage liability coverage on commercial automobiles has been selected as an example. Although the following discussion deals only with this single case, it is believed that the method is equally adaptable to all cases. Comparison of the resulting Table 10 with Table 5 gives a specific comparison of the normal sampling variation in total losses, pure premiums, or loss ratios with that occurring in the number of claims or claim frequencies.

In preparing the desired tables of the normal sampling range of the ratio of actual to expected values of total losses, pure premiums, or loss ratios, the values of M, σ, U_2, and α_3 will be (as found in section D of I):

$$M = 1 \qquad U_2 = \frac{1}{c}\frac{V_{2:x}}{V_1{}^2{}_{:x}}$$

$$\sigma = \frac{1}{\sqrt{c}}\sqrt{\frac{V_{2:x}}{V_1{}^2{}_{:x}}} \text{ and } \alpha_3 = \frac{1}{\sqrt{c}}\frac{V_{3:x}}{(V_{2:x})^{3/2}}$$

For commercial automobile property damage claims we have:

$$\sqrt{\frac{V_{2:x}}{V_1{}^2{}_{:x}}} = 2.050 \text{ and } \frac{V_{3:x}}{(V_{2:x})^{3/2}} = 10.524$$

In fitting available theoretical distributions to this data, we will find three different ranges requiring separate treatment. The first of these will be where the expected number of claims is small (10 or less). In this range, the occurrence of no losses must be recognized as a distinct possibility and the number of such cases set aside before attempting to fit a continuous distribution such as the Normal Logarithmic Distribution to the remaining cases. This pro-

cedure is followed in Table 7, where each successive step has been set out in order to show the algebraic process as well as the arithmetic computation.

The second range is that where the probability of obtaining zero losses is insignificant, although the skewness of the distribution of losses by amount of loss is still a controlling influence. Here the Normal Logarithmic Distribution is fitted directly by omitting steps (2) to (6) inclusive and step (28) as shown in Table 8.

The third range is that for very large values of expected claims. In this range the skewness, although still large enough to preclude the use of the normal distribution, comes to a level recognized by a Type III distribution. The calculation procedure can thus be further reduced to that shown in Table 9.

Table 10 presents the final results of the calculations of Tables 7, 8, and 9 and shows for the ratio of actual to expected total losses, pure premiums, and loss ratios the normal sampling range. This table corresponds for these statistics to Table 5 for the number of claims or claim frequencies.

C. *Average Claim Costs (of a Fixed Number of Claims)*

In actual practice we will usually be concerned only with the sampling variation of the ratio of actual to expected average claim costs for the fixed number of claims that actually occurred.

From section E of part I, we find the necessary statistics to construct the desired table as:

$$M = 1,\ \sigma = \frac{1}{\sqrt{n}} \frac{\sqrt{U_{2:x}}}{V_{1:x}} \text{ and } \alpha_3 = \frac{1}{\sqrt{n}} \alpha_{3:x}$$

which, combined with the values for Commercial Automobiles, P.D., give:

$$M = 1,\ \sigma = \frac{1.7891}{\sqrt{n}} \text{ and } \alpha_3 = \frac{13.972}{\sqrt{n}}$$

In fitting theoretical distributions to this data, it will again be necessary to use the Normal Logarithmic Distribution for the smaller values of n (less than 1440), while the Type III distribution will expedite calculations for larger values of n. The resulting Table 11 is presented, without again showing the details of calculation which are similar to those of Tables 8 and 9.

D. *Average Claim Costs (with c Claims Expected)*

In some few cases we shall be concerned with the sampling variation of the ratio of actual to expected average claim costs when only the expected

number of claims is known. From section E of Part I, we find the necessary statistics from which to calculate the desired table of sampling variation as:

$$M = 1, \qquad \sigma = \sqrt{K_{(c)}} \cdot \frac{\sqrt{U_{2:x}}}{V_{1:x}} \text{ and } a_3 = \frac{G_{(c)}}{K_{(c)}^{3/2}} \cdot a_{3:x}$$

or combined with the values for Commercial Automobiles, P.D.:

$$M = 1, \qquad \sigma = 1.7891 \sqrt{K_{(c)}} \text{ and } a_3 = 13.972 \frac{G_{(c)}}{K_{(c)}^{3/2}}$$

The Normal Logarithmic Distribution will again be found useful in fitting a theoretical distribution for values of c less than 1440. The results are shown in Table 12, where it is found that this table is practically identical with Table 11 for values of c, of 40, or greater.