THE DISTRIBUTION OF AUTOMOBILE ACCIDENTS—ARE RELATIVITIES STABLE OVER TIME?

EMILIO C. VENEZIAN

Abstract

Data on the distribution of automobile accidents typically reject the hypothesis that accident rates are the same for all members of a group. Given these findings, policy analysis is usually based on models that assume that accident proneness differs among individuals of a group and that the differences are stable over time. The analysis presented in this paper is aimed at assessing the validity of these assumptions.

A simple model that allows for variability in both proneness and exposure level is used to estimate the potential contribution of variability in exposure levels to total variability in accident rates. Data indicate that variability of exposures may have a substantial bearing on the variability of accident rates.

Data from several groups of drivers from California and North Carolina are used for direct tests of the stability of relative accident rates. When the California data are used, the tests do not lead to a rejection of the hypothesis that relative accident rates are stable. When the North Carolina data, based on a larger number of observations, are used, the tests clearly reject the hypothesis.

The implications of these findings for economic and policy analysis are discussed.

1. INTRODUCTION

For most types of accidental events, the number of accidents over a period does not exhibit a Poisson distribution, even when data are restricted to a group of individuals who are presumed homogeneous. This finding has been reported in a variety of settings such as automobile accidents [11, 13, 16], health insurance claims [6], and professional liability incidents [9, 15]. This empirical finding is often rationalized by appealing to the notion of differences in "accident proneness" among the individuals in the groups under analysis. The typical model assumes that each individual within the group has an inherent accident rate, and that, for each individual, the number of accidents in a given period has a Poisson distribution with the appropriate parameter. Each individual in the group under study is viewed as having an inherent accident rate, and this rate is assumed to differ among individuals according to some probability distribution. This model, often called the "compound Poisson model"[1] has reportedly been successful in fitting the distribution of the number of accidents or claims observed in a given time interval.

The apparent success of these attempts, especially those based on the assumption that accident proneness has a Gamma distribution, does not provide a sound basis for the formulation of public policy.[2] In the first place, the usual interpretations of the compound Poisson require that the accident rate of a given individual be stable over time, a characteristic for which tests have seldom been performed. Another shortcoming of these methods is that the distribution inferred from this model is identical to the distributions inferred from other models [2, 10]. Moreover, models exist that provide results which are as good as, or better than, those obtained under the assumption that accident proneness differs among individuals; however, these models have very different implications for public policy [10, 14, 17]. It is therefore of interest to examine more closely the relationship between data and hypothesis on one hand and the relationship between hypothesis and policy on the other. This paper examines the issue in the context of automobile insurance.

[1] See, for example, Feller [2], pages 288–293; Seal [10], page 31.

[2] Public policy generally refers to policies adopted by governmental or quasi-governmental entities. In the present context, it includes such diverse areas as licensing, limitation of privileges, and the imposition of premium penalties for past events.

The paper first discusses briefly, in Section 2, the data that will be used in exploring the theoretical issues. Section 3 discusses the compound Poisson model that is often used to justify both private and public initiatives in accident prevention. The paper then consists of three main sections. Section 4 discusses an alternate source of differences in the inferred proneness of individuals. This alternate model leads to a different valuation of the benefits of any policy that restricts driving by individuals who have had relatively large numbers of accidents in the past. There is relatively little that can be done with existing data to discriminate between the models. Since the models lead to different conclusions, data to permit an assessment of the alternatives should be collected if at all possible. Section 5 considers a test of the hypothesis that claim propensities (or more accurately, indices of claim propensity) are constant over time when taken over reasonably long durations. Such constancy is essential if we are to use historical data to implement a policy whose benefits can be asserted to exist only to the extent that past accidents predict future accident propensity for the individual. The available data on automobile accident involvement indicate that constancy is not a reasonable assumption. Section 6 provides a discussion of the findings in the context of economic and policy analysis of insurance issues.

2. DATA FOR ANALYSIS

In order to examine these issues, this article uses two sets of data from available literature. Both of these sets related to the accident records of groups of drivers whose records were followed over a long period of time.

The first body of data relates to a sample of drivers in California. The data used in the present analysis were derived by the author from available tabulations [5]. The data relate to accidents experienced in the years 1969 to 1974 by a sample of California drivers who had licenses active for the period 1961 to 1974. The original paper gives extensive tabulations by sex and by pattern of accidents. The basic data used in this paper were derived from the original data and are presented in three tables in Appendix A. The analysis will be performed separately on the three sets of data: (1) for female drivers, (2) for male drivers, and (3) female and male drivers combined.

The second set of data is available in the form used directly for computations [12]. It relates to accidents experienced in the periods[3] 1967–1968 and 1969–1970 by all North Carolina drivers who were at least 22 years old in November of 1970, and in the twelve-month periods 1969 and 1970 for North Carolina drivers who were 21 years old in November of 1970. For drivers whose age at the end of the study was 22 years or more, the data are available separately for ages 22–25, 26–39, 40–59, and 60 and over. In all cases, the drivers are classified by their age at the beginning of the study period.

3. THE COMPOUND POISSON MODEL AND ITS INTERPRETATION

The compound Poisson model pictures each individual as having an inherent propensity to be involved in an accident. Most models picture that propensity as a fixed number that does not vary over time. Strict constancy from day to day is not necessary as long as it holds over periods of time comparable with those for which data are available. Moreover, the mathematical and statistical analysis would not be affected substantially if that element which is constant were an index of proneness which modifies the average rate for the group as a whole. What is of major importance to the arguments surrounding the compound Poisson model is that this index is immutable for a given individual.

The principal statistical implication of the compound Poisson hypothesis is that individuals with large numbers of accidents are relatively more common than would be predicted by the simple Poisson model. In statistical terms, the consequence of having a compound Poisson distribution is that the variance of the number of claims will be larger than the mean number of claims. In contrast, for the simple Poisson, the mean and the variance of the number of claims are identical.

The usual interpretation of the compound Poisson hypothesis is that individuals with large accident propensities affect the group adversely, leading to a higher average number of claims. This affects the insurability of those members of the group who have low propensity indices. In

[3] The periods do not cover the calendar years. As explained in the original reference, the nominal year 1970, for example, covers the twelve-month period beginning in December 1969.

automobile insurance, two streams of rhetoric have arisen from this interpretation. Some argue that failure to reflect the differences among group members in insurance rates amounts to "guilt by association." Others bemoan the fact that insurance premiums are made to depend on factors that are not controllable by the individual. In the context of medical professional liability, the model elicits the picture that a few "bad apples" are responsible for most of the problems and has resulted in calls to revoke the licenses of these "bad apples" and thus reduce the number of claims.

If this picture is true, an economic analysis of restricting the privilege of driving, either through license restrictions or through the provision of insurance only at high rates, would be useful. The best level of restrictions would be determined by balancing the costs and the benefits of such a decision. The costs arise primarily from curtailing the freedom of some individuals to drive automobiles; they have a monetary component related to the difference in price between driving one's own car and relying on alternate modes of transportation, and a nonmonetary component related to loss of freedom. The benefits arise from the reduced number of accidents, and these also have monetary and nonmonetary elements. For society as a whole, monetary benefits[4] arise from avoiding costs to rectify the consequences of accidents, while the nonmonetary component stems from the reduction in pain and suffering associated with the avoided mishaps. The calculation of the benefits depends very strongly on the exact hypothesis which motivates the compound Poisson model. To explore the extent to which this might affect our thoughts about policy, it is worthwhile to contrast the usual assumption, that the differences in accident experience are due to differences in inherent ability, with a specific alternate hypothesis.

The key parameter in the Poisson distribution is not an "accident propensity" that measures inherent ability, but a weighted measure that recognizes both the ability to perform a dangerous task and how often the task is performed. The expected number of automobile accidents which one individual might have in a year may not be a fixed quantity; it might, for example, depend on the number of miles that the individual

[4] Distributional costs and benefits will also result. Individuals with low accident rates will not have to subsidize individuals in the same group that have higher accident rates.

chooses to drive under various sets of conditions. Similarly, the expected number of claims against an engineer might depend on the number of plants she designs, and the number of claims that a physician might expect could depend on the number of patients that are treated by that physician. Thus variation in the Poisson parameter among drivers does not require that the rate of accidents differs among individuals when the level of activity is identical. It could be explained equally well, from a statisical point of view, if all drivers had exactly the same accident proneness under every given set of driving conditions but they differed in the miles they drove under various conditions.

This alternate hypothesis as to sources of variability suggests a different interpretation of the compound Poisson process. In this picture, all drivers in a group have exactly the same probability of having an accident in each mile they drive; but, they differ from each other inherently in the mileage driven. To keep an exact correspondence to the previous model, it is important that the distance and nature of driving in this version be as immutable as the inherent accident probability in the previous one; these measures of exposure may change only in ways that are strictly coupled with the average for the group as a whole.

Neither of these simple models is likely to be strictly valid. In all likelihood, drivers differ with respect to the probability that they will be involved in an accident under a given set of conditions. In all likelihood, they also differ with respect to the exposure level they chose. Thus a model that recognizes differences in both propensity and activity levels is likely to provide a better explanation of actual experience.[5]

4. INTERPRETATION OF THE EXCESS VARIANCE

When the number of accidents observed for each of many members of the group is analyzed, we expect to see a variance that is approximately equal to the mean if all members have the same probability of having

[5] The combined effect of individual propensity and exposure is particularly important in economic contexts in which the driver has control of the exposure level, at least within broad limits. Unfortunately, this dual determination has seldom been considered. The literature on moral hazard, for example, appeals to a "level of care" which might be selected by the driver, but does not take into account the possible direct choice over the level of exposure by restricting or expanding the mileage driven.

an accident in a unit of time, and a variance greater than the mean if individual members differ in this respect. A positive difference between the variance and the mean, or "excess variance," results from differences in the Poisson parameters of the members of the group.

If we are observing individuals over a period of time T, the Poisson parameter for individual i will be:

$$M_i = k_i p_i T, \qquad (1)$$

where k_i is the number of opportunities for individual i to have an accident and

p_i is the individual's probability of an accident on any given opportunity.

Usually there is a measure of T, but there are no measures of k_i or p_i; so this model does not have an operational meaning.[6] The model adopted in arguing that probability of an accident varies across individuals in the group is equivalent to arguing that k_i is the same for all individuals. The alternate model discussed earlier assumes that k_i varies across individuals but p_i does not. Equation 1 provides a more general formulation and can be made operational if there are measures of the level of activity; for example, the number of miles driven per year for automobile accidents, the number of takeoffs for small aircraft accidents, or the number of specific surgical procedures for medical professional liability. Even in the absence of such measures, formulation is worth considering because it may yield some insight into the process.

If the Poisson parameter, M_i, varies across individuals, it can be proved that the average number of accidents for individuals in a group is given by:

$$\mathrm{E}_i(N) = \mathrm{E}_i(M_i) = T\mathrm{E}_i(k_i p_i), \qquad (2)$$

and

$$\mathrm{Var}_i(N) - \mathrm{E}_i(N) = T^2\mathrm{Var}_i(k_i p_i). \qquad (3)$$

[6] In some contexts it would be possible to obtain information about the level of exposure, even though imperfect. In relation to automobile accidents, the mileage driven per year might serve as a measure of k_i.

In these equations, $\mathrm{E}_i(Z)$ denotes the expected value of Z and $\mathrm{Var}_i(Z)$ denotes the variance of Z, both measured over the population in the group.

If we denote the excess variance, $\mathrm{Var}_i(N) - \mathrm{E}_i(N)$, as $X_i(N)$, the variance of this statistic under the null hypothesis that the Poisson parameter is identical for all members of the population is given by:

$$\mathrm{Var}_i(X_i(N)) = \frac{2}{I}\,\mathrm{E}_i^2(k_ip_i) = \frac{2}{I}\,\mathrm{E}_i^2(N) \tag{4}$$

where I is the total number of individuals observed [14].

It is worth noting that if k_i is the same for all individuals, the excess variance is proportional to the variance of p_i, whereas if p_i is the same for all individuals, then the excess variance is proportional to the variance of k_i. If both k_i and p_i vary, then the excess variance will depend on the joint distribution of p_i and k_i.

Assuming for now that a stable compound Poisson is the proper model, it is of interest to determine whether the data indicate that there is significant heterogeneity in a given group and to interpret the excess variance, if it can indeed be said to be positive. The equations given above can be used for this purpose. A test requires simply computing the observed excess variance and the variance of that quantity under the null hypothesis; this statistic can be estimated by using Equation 4. The sample estimate of the excess variance is the sample estimate of the variance minus the sample estimate of the mean. If the number of observations is large, both these sample estimates are asymptotically normal [1]; it follows that the difference is asymptotically normal, so the ratio of its sample value to the standard deviation should, under the null hypothesis, be distributed as a standard normal deviate. When interest is centered on determining whether there is significant heterogeneity among members of the group, the null hypothesis is that there is no heterogeneity; under those conditions, the distribution of claims would follow a simple Poisson distribution. Table 1 summarizes the data used in assessing the significance of the excess variance. Table 2 presents the main results. It is clear that the excess variance is positive and highly significant for all the groups under consideration, since the ratio of the estimate to its standard deviation is always greater than 15.

TABLE 1

NUMBER OF DRIVERS BY GROUP AND NUMBER OF ACCIDENTS

State and	Number of Accidents								
Group	0	1	2	3	4	5	6	7+	Total
CA Females	19,634	3,573	558	83	19	4	1	0	23,872
CA Males	21,800	6,589	1,476	335	69	16	4	4	30,293
CA All	41,434	10,162	2,034	418	88	20	5	4	54,165
NC 22–25	276,081	69,811	16,770	4,060	967	236	74	26	121,221
NC 26–39	709,649	143,601	29,401	6,658	1,725	447	124	51	891,656
NC 40–59	762,592	138,955	23,580	4,492	1,054	254	74	33	931,044
NC 60+	254,255	47,095	8,159	1,511	344	97	28	24	311,513
NC 21	144,803	25,302	4,007	637	105	15	5	1	174,875

TABLE 2

ANALYSIS OF EXCESS VARIANCE BY GROUP

State and	Number of Accidents		Sample Excess Variance		
Group	Mean[a]	Variance[b]	Value[c]	Std.Dev.[d]	Value/Std.Dev.
CA Females	0.2111	0.2483	0.0372	0.0014	27.22
CA Males	0.3617	0.4435	0.0818	0.0042	19.68
CA All	0.2953	0.3631	0.0678	0.0025	26.72
NC 22–25	0.3294	0.4322	0.1028	0.0011	94.66
NC 26–39	0.2609	0.3439	0.0830	0.0006	155.63
NC 40–59	0.2211	0.2752	0.0541	0.0005	118.05
NC 60+	0.2252	0.2822	0.0570	0.0008	70.63
NC 21	0.2167	0.2353	0.1867	0.0010	180.14

[a] Mean of the number of accidents

[b] Variance of the number of accidents

[c] Variance minus mean of the number of accidents

[d] Calculated as the square root of the variance given by Equation 4

The statistical significance of the excess variance is of importance in examining private policy issues such as merit rating and freedom to underwrite. From this perspective, it is important to know whether the data suggest that the Poisson parameter differs among individual members of the group. The existence of variability among individuals suggests that differential pricing based on experience may be useful in achieving an equitable allocation of future costs. From the point of view of public policy issues such as restricting the ability of individuals to drive, however, this information is not sufficient because the variability may be due to differences in the level of activity of individuals rather than to differences in claim propensity.

While this distinction is not important in dealing with private mechanisms such as classification by individual companies in a market with open competition, it is important in dealing with public mechanisms such as classifications mandated by the state or licensing restrictions. As discussed earlier, if the difference in Poisson parameters arises predominantly from differences in the level of activity, restrictions placed on the privilege of driving by individuals with large numbers of accidents will either restrict their mobility or force them to use alternate drivers who have comparable or higher propensities to have accidents for corresponding exposures. Thus social benefits might not be experienced, but substantial social costs would be incurred.

It is not possible to draw firm conclusions about the relative importance of level of exposure and accident propensity from the available data. The information is sufficient, however, to permit drawing tentative conclusions.[7] The line of inference begins by noting that the excess variance measures the variance of Poisson parameters, as is shown by Equation 3. The ratio of this quantity to the square of the Poisson parameter represents the coefficient of variation of the parameter.

TABLE 3

VARIATION OF POISSON PARAMETER BY GROUP

State and Group	Poisson Parameter[a]	Excess Variance[b]	Coefficient of Variation[c]
CA Females	0.2111	0.0372	0.83
CA Males	0.3617	0.0818	0.63
CA All	0.2953	0.0678	0.78
NC 22–25	0.3294	0.1028	0.95
NC 26–39	0.2609	0.0830	1.22
NC 40–59	0.2211	0.0541	1.11
NC 60+	0.2252	0.0570	1.12
NC 21	0.2167	0.1867	3.98

a. From Table 2, column 2
b. From Table 2, column 4
c. Coefficient of variation of the Poisson parameter

[7] Another possibility that deserves consideration is that the classification system is inadequate.

Table 3 shows the results for the various groups. In most cases, the coefficient of variation of the Poisson parameter among members of a group is very close to one. In the case of North Carolina drivers of 21 years of age, it is almost four.

The interpretation of this number must, unfortunately, rely on the context of the problem because firm data are not available.[8] At one extreme, if the level of exposure is the same for all individuals, the coefficient of variation of the Poisson parameter would approximately equal that of accident proneness. At the other extreme, if the accident proneness were the same for all individuals, this number would equal the coefficient of variation in the exposure level. Any measure of the coefficient of variation of the exposure level will therefore serve to help to place the results in context. In the present case, exposure might be measured by mileage driven in a unit of time [8] and might well exhibit a large coefficient of variation. Rough estimates are discussed in Appendix B; they range from 0.3 to 0.9.

The observed coefficients of variation of the Poisson parameter are generally higher than the corresponding estimates for mileage driven. However, even with the lower estimates for the latter, variation in mileage driven would account for about 25 percent of the variance of Poisson parameters. Thus exposure may play a substantial role in determining the accident rates of individuals. Data relating accident experience and mileage driven by individuals in different time periods could provide better measures of the relative contribution of exposure; even accurate data on the distribution of mileage driven would be useful in assessing the relative effects of exposure and propensity on the Poisson parameter of individuals.

5. A TEST FOR GENERAL COMPOUND POISSON MODELS

The discussion presented above indicates that caution must be exercised in using results from a simple static analysis to guide policy. The usual analyses do not pinpoint the reason for variation in Poisson param-

[8] Even if data were available, it should be remembered that the model used here assumes that individuals select their exposure level without regard to their accident proneness. This may be appropriate when individuals are insured but may be a poor assumption in the absence of insurance.

eters and these reasons may have a bearing on policy issues. For example, even people who would accept the hypothesis that the accident propensity of an individual, p_i, does not change over time might question the hypothesis that the exposure level of the individual, k_i, does not change. Yet the predictability of the Poisson parameter plays a key role in the ideology of classification and merit rating [7]. The relevance of statistical analysis to policy requires analysis of models that are realistic and address the key issues. This is more likely to happen if the public policy issues are examined and statistical tools are developed to analyze the key issues.

One of the important issues in automobile liability is the measurement of the benefits to be derived from restricting the mobility of drivers with several claims.[9] The costs that would be incurred by such restrictions would depend primarily on the number of people on whom restrictions would be placed, not on the model assumed. The benefits, however, may be estimated only in relation to a model. In this context, statistical methods can provide assistance only if they are designed to provide relevant information and if they are valid. A common feature of the two models discussed earlier is the assumption that the likelihood that an individual driver will have an accident is an inherent characteristic of the individual. Statistics are useful in establishing whether this is a valid conclusion.

For the most part, compound Poisson models have been tested by assuming a specific form for the distribution of accident propensities, inferring a theoretical distribution to the number of accidents and performing a goodness of fit test to establish that the fit is adequate in a statistical sense. If they result in a good fit, these statistical procedures can, at best, establish that it is plausible that during a given time period, individuals in a group differ with respect to accident propensity and that the propensities can be characterized as having a distribution similar to the one assumed. The procedures do not test the assumption that the claim propensity of an individual is the same in two different time

[9] The statement is valid whether the restriction occurs by exercise of the power of the state to limit the privilege of driving, or by exercise of economic power to increase the cost faced by certain individuals in order to drive. The discussion will be limited to the former, since analysis of the latter requires knowledge of tradeoffs whose value cannot be estimated readily.

intervals.[10] That assumption is important in most arguments related to either pricing of insurance or to restrictive public policy, since these arguments assume that past experience is a good predictor of future performance for an individual.

The analysis presented above still retains the untested assumption that the Poisson parameter corresponding to i^{th} individual, M_i, is constant over time. Direct tests of this assumption are not feasible since we cannot observe the same individual repeatedly during the same time interval. The literature does provide a method for determining whether this key assumption is correct. The method uses data from a single population studied in successive time periods. It was first suggested by Lundberg [6], who showed that, for a general compound Poisson with the Poisson parameter of each individual being equal to an individual parameter times the average rate for all individuals in any given subinterval of time, the probability that an individual will have m claims in the subinterval t_2, and n claims in subinterval t_1, given that he had $m + n$ claims in the interval $t_1 + t_2$, is given by:

$$\mathrm{P_r}(m,t_2|n,t_1) = \frac{(m+n)!}{m!\,n!}\,\Theta_1^m(1-\Theta_1)^n \qquad (5)$$

where $\Theta_1 = r_1t_1/(r_1t_1 + r_2t_2)$,

r_j is the average accident rate in period j, and

t_j is the duration of period j.

The operational time intervals, r_jt_j contain the average accident rates, r_j, which are not known, along with the calendar time, t_j. In order to provide a valid test, it is necessary to develop measures of the ratios of operational time intervals. Lundberg argues that the ratios may be estimated by the ratio of the number of accidents or claims in each subperiod to the total number of accidents or claims. Once the parameters are known, the conditional distributions for all relevant values of $m + n$ can be computed and compared to the observed data by using a chi-squared test. Lundberg recommends grouping cells so that the expected number of claims is five or more. The degrees of freedom for each value of $m +$

[10] A notable exception is the analysis of Weber [16], who used methods attributed to Greenwood and Yule [3] and Kerrich [4] for the case in which the compounding distribution is the Gamma distribution. The method of Greenwood and Yule actually turns out to be valid for general compounding distributions and is equivalent to the method used here.

n is one less than the number of cells used in the test; the additivity of chi-squared may be used to construct an overall test by adding the contributions to chi-squared and adding the degrees of freedom. Lundberg used this test with data on health insurance claims in Sweden and found the results did not reject the hypothesis of a compound Poisson distribution with stable parameters.

It is worth noting that the method recommended by Lundberg for the estimation of the parameters relies on ratios of the average accident rates. It follows that, if the average accident rates change and the individual accident rates change proportionately, the test will not be affected.[11] Thus the test will be valid if the accident relativities are constant, even though the actual accident rates change. From this perspective, the null hypothesis could be characterized as the assertion that rate relativities are constant over time.

To illustrate the procedure, the data for all drivers in California, given in Table A3 of Appendix A, will be used. The data there were 7,967 accidents for the period 1969–71 and 8,030 accidents for the period 1972–1974. The total number of accidents was 15,997. The best estimate of the needed parameter is $\Theta_1 = 7{,}967/15{,}997 = 0.4980$.

This parameter is used in Equation 5 to estimate the expected fraction of drivers with m accidents in the first period and n accidents in the second period given a total number of $m + n$ accidents in the whole period. These probabilities are multiplied by the observed number of drivers with $m + n$ accidents in the whole period to obtain the expected value of the number of drivers with m accidents in the first period and n accidents in the second period. The observed and predicted numbers of drivers with one, two, three, and four accidents in the period 1969–1974 are shown in Table 4. Tables 5 to 11 provide analogous information for the other groups. Note that in Lundgren's scheme, the data for individuals that had no accidents in either the first or second interval contribute only to the estimation of the frequency of accidents in the two periods, but do not contribute to the value of the test statistic. Accordingly the tables do not show this group.

[11] Lundgren's original application to data on health insurance claims actually involved substantially different claim rates in the first and second period.

TABLE 4

Observed and Predicted Distribution of Female Drivers in California, by Number of Accidents

Number of Accidents in 1969–74		Number of Accidents in 1969–1971						Chi-Squared
		0	1	2	3	4	5	
1	Observed	1,816	1,757					
	Predicted	1,838.5	1,734.8					0.3
2	Observed	164	266	128				
	Predicted	147.7	278.8	131.5				2.5
3	Observed	16	24	33	10			
	Predicted	11.3	32.0	30.2	9.5			4.2
4	Observed		5[a]	8	6[b]			
	Predicted		6.4[a]	7.1	5.5[b]			2.0
Total (8 degrees of freedom)								9.0

a. Drivers with either zero or one accidents in 1969–71

b. Drivers with either three or four accidents in 1969–71

TABLE 5

Observed and Predicted Distribution of Male Drivers in California, by Number of Accidents

Number of Accidents in 1969–74		Number of Accidents in 1969–1971						Chi-Squared
		0	1	2	3	4	5	
1	Observed	3,226	3,363					
	Predicted	3,269.5	3,319.5					1.2
2	Observed	403	692	381				
	Predicted	363.4	738.0	374.6				7.3
3	Observed	44	126	124	41			
	Predicted	40.9	124.7	126.6	42.8			0.4
4	Observed	7	15	26	21[a]			
	Predicted	4.2	17.0	25.9	22.0[a]			2.2
Total (9 degrees of freedom)								11.1

a. Drivers with either three or four accidents in 1969–71

TABLE 6

Observed and Predicted Distribution of All Drivers in California, by Number of Accidents

Number of Accidents in 1969–74		Number of Accidents in 1969–1971						Chi-Squared
		0	1	2	3	4	5	
1	Observed	5,042	5,120					
	Predicted	5,101.0	5,061.0					1.4
2	Observed	567	958	509				
	Predicted	512.5	1,014.0	504.5				9.2
3	Observed	60	150	157	51			
	Predicted	52.9	157.4	156.1	51.6			1.3
4	Observed	9	18	34	23	4		
	Predicted	5.6	22.2	33.0	21.8	5.4		3.3
Total (10 degrees of freedom)								15.3

TABLE 7

Observed and Predicted Distribution of Drivers of Age 22–25 in North Carolina, by Number of Claims

Number of Accidents in 1967–70		Number of Accidents in 1967–1968						Chi-Squared
		0	1	2	3	4	5	
1	Observed	33,615	36,196					
	Predicted	33,847.3	35,963.0					3.1
2	Observed	4,639	7,164	4,967				
	Predicted	3,942.1	8,377.3	4,450.6				358.8
3	Observed	648	1,339	1,358	715			
	Predicted	462.7	1,475.0	1,567.2	555.1			160.7
4	Observed	73	239	313	237	105		
	Predicted	53.4	227.1	362.0	256.4	68.1		35.9
5	Observed	14	41	54	63	36	28	
	Predicted	6.3	40.3	71.4	75.8	40.3	8.6	60.3
6	Observed		9[a]	18	17	18	12[b]	
	Predicted		7.1[a]	16.3	23.1	18.4	9.2[b]	3.2
Total (19 degrees of freedom)								622.0

[a] Drivers with either zero or one accidents in 1967–68

[b] Drivers with either five or six accidents in 1967–68

TABLE 8

Observed and Predicted Distribution of Drivers of Age 26–39 in North Carolina, by Number of Claims

Number of Accidents in 1967–70		Number of Accidents in 1967–1968						Chi-Squared
		0	1	2	3	4	5	
1	Observed	68,827	74,774					
	Predicted	69,682.9	73,918.1					2.3
2	Observed	7,817	13,273	8,311				
	Predicted	6,923.1	14,687.7	7,790.2				92.3
3	Observed	1,064	2,286	2,291	1,017			
	Predicted	760.8	2,421.0	2,568.2	908.1			44.8
4	Observed	149	434	576	398	168		
	Predicted	95.6	405.8	645.8	456.7	121.1		64.9
5	Observed	28	75	115	128	71	30	
	Predicted	12.0	63.8	135.3	143.6	76.1	16.1	40.1
6	Observed		13[a]	28	38	28	17[b]	
	Predicted		11.9[a]	27.3	38.6	30.7	15.4[b]	0.5
Total (19 degrees of freedom)								244.9

[a] Drivers with either zero or one accidents in 1967–68

[b] Drivers with either five or six accidents in 1967–68

TABLE 9

Observed and Predicted Distribution of Drivers of Age 40–59 in North Carolina, by Number of Claims

Number of Accidents in 1967–70		Number of Accidents in 1967–1968						Chi-Squared
		0	1	2	3	4	5	
1	Observed	66,875	72,080					
	Predicted	67,156.4	71,798.6					2.3
2	Observed	5,928	11,048	6,604				
	Predicted	5,507.7	11,776.8	6,295.5				92.3
3	Observed	629	1,540	1,644	679			
	Predicted	507.1	1,626.4	1,738.8	619.7			44.8
4	Observed	76	265	365	261	87		
	Predicted	57.5	245.9	394.4	281.1	75.1		12.9
5	Observed	10	35	85	80	43	11	
	Predicted	7.0	37.2	79.6	85.1	45.5	9.7	2.4
6	Observed		10[a]	20	19	13	12[b]	
	Predicted		7.0[a]	16.2	23.0	18.5	9.3[b]	5.3
Total (19 degrees of freedom)								160.0

[a] Drivers with either zero or one accidents in 1967–68

TABLE 10

Observed and Predicted Distribution of Drivers of Age 60 or more in North Carolina, by Number of Claims

Number of Accidents in 1967–70		Number of Accidents in 1967–1968						Chi-Squared
		0	1	2	3	4	5	
1	Observed	23,367	23,728					
	Predicted	23,252.8	23,842.2					1.1
2	Observed	2,083	3,878	2,198				
	Predicted	1,989.0	4,078.9	2,091.1				17.6
3	Observed	205	529	549	228			
	Predicted	181.9	559.4	573.6	196.1			10.9
4	Observed	23	70	115	90	46		
	Predicted	20.4	83.8	129.0	88.2	22.6		17.6
5	Observed		14[a]	30	29	24[b]		
	Predicted		17.4[a]	29.9	30.7	18.6[b]		2.4
Total (13 degrees of freedom)								49.6

a. Drivers with either zero or one accidents in 1967–68

b. Drivers with either four or five accidents in 1967–68

TABLE 11

Observed and Predicted Distribution of Drivers of Age 21 in North Carolina, by Number of Claims

Number of Accidents in 1967–70		Number of Accidents in 1967–1968						Chi-Squared
		0	1	2	3	4	5	
1	Observed	13,118	12,184					
	Predicted	13,135.0	12,167.0					0.1
2	Observed	1,204	1,767	1,036				
	Predicted	1,079.9	2,000.6	926.6				54.5
3	Observed	113	215	232	77			
	Predicted	89.1	247.6	229.4	70.8			11.3
4	Observed	13	23	33	26	10		
	Predicted	7.6	28.3	39.3	24.2	5.6		9.3
Total (10 degrees of freedom)								75.2

The results indicate that the data from California do not reject the hypothesis that Poisson parameters for each individual bear the same relationship to the aggregate average level in the two periods under consideration. When the two sexes are combined, the value of chi-squared is 15.3 with 10 degrees of freedom, a value that would occur by chance about ten percent of the time if the null hypothesis were valid. This finding is consistent with that of Weber [16], based on California data for a different and shorter period. These data do not provide strong evidence against the hypothesis that relativities are stable.

The results from the North Carolina data, on the other hand, lead to the clear rejection of the null hypothesis. Given the degrees of freedom, the observed values of chi-squared in any one group would correspond to probabilities much lower than one in one thousand if the null hypothesis were true. It is worth noting that large contributions to chi-squared arise from relatively small values of $m + n$. This indicates that the data leading to the rejection of the null hypothesis are not concentrated in the cells corresponding to individuals with a high aggregate accident propensity or to cells with relatively small numbers of observations.

In view of the results with the data from North Carolina, the fact that the California data do not reject the null hypothesis could be attributed to the fact that different mechanisms are operating in the two environments. However, the number of observations in California is much smaller than that in North Carolina. The total number of drivers observed in California is just over 54,000, compared to over 2.5 million in North Carolina. Thus, an alternate explanation of the results is that the power of the test to reject the null hypothesis is so low that the California data cannot attain the conventional levels of confidence. Unfortunately, there are no other formal tests of the hypothesis of interest.[12]

[2] Thyrion [13], among others, has pointed out that there is an interesting recursive relation for the compound Poisson of arbitrary compounding distribution. Statistical tests based on that relation have not been developed.

6. DISCUSSION

The data analysis suggests that driver accident frequencies may be determined jointly by the driver's ability to drive and the exposure that the driver experiences. This finding is important for both economic analysis and policy formulation.

Economic analysis usually assumes that economic agents act rationally, in the sense that they select the options that provide them the greatest level of satisfaction. Analyses of insurance purchasing and the related issue of moral hazard have not, however, considered seriously the possibility that individuals may, in the absence of insurance, select the level of exposure with due regard to the individual's accident propensity and risk aversion. Given the possibility of such effects, the analysis of the insurance purchasing decision may be misleading unless one explicitly recognizes that the utility function depends on both consumption of goods and ability to travel. It may even be important to take into account the relationship between ability to travel and ability to generate income.

This possibility also creates some interesting problems in the analysis of insurance classifications. Given that the individual selects exposure by considering both accident propensity and risk aversion, the net experience of that individual, measured in terms of the expected number of accidents, will reflect a complex interaction of accident propensity and risk aversion. Moreover, this expected number of accidents is not likely to provide much information regarding what its corresponding value after insurance is likely to be, since the existence of insurance coverage may have large effects on the individual's choice of exposure. Also, the experience of an individual under one classification scheme will serve to predict the experience of the same individual under a different classification scheme only to the extent that the new classification scheme will affect neither the individual's propensity to have an accident, nor his selection of a level of exposure, nor his inclination to purchase coverage. In this respect, particular care should be exercised in drawing inferences about plans based on merit rating or bonus-malus systems from corresponding information gathered under classification plans that do not include experience rating.

From the perspective of public policy, the possible effect of variability of exposure suggests that the Poisson parameter of individuals will not be constant from period to period, even if the accident proneness remains fixed. The possibility that individuals will change their exposure level implies that past experience may not be the best predictor of the future experience for an individual. While prediction of average group performance based on the past may be appropriate and necessary for proper functioning of insurance markets, public policy should reflect concern about distributional equity if the past is not a good predictor of the future for an individual. Variation of the Poisson parameter over time implies that the public policy arguments favoring merit rating may not be properly based on fact. If the individual's accident propensity varies from period to period, a rate based on past exposure is not necessarily a good predictor of future experience for the individual. In fact, if propensity varies over time, the issue of whether merit rating is a better predictor of future performance than classification rating must be examined empirically rather than assumed.

Given the likely effect of exposure levels, it is perhaps not surprising that the data do not support the hypothesis that Poisson parameters are constant over time or bear a constant relationship to the group average. At present, the conclusion that the data reject the hypothesis is based largely on the data from North Carolina. The other body of data that is currently available, that of California, does not reject the hypothesis. The California data for both sexes combined is barely consistent with the hypothesis at the ten percent level.[13] It may well be that the hypothesis would be rejected by a larger sample of drivers from this state and period. Additional data with which to probe this question would be valuable, especially if the data base included estimates of the exposure level.

[13] It may be worth noting that if we were to focus our attention on the group with two accidents in the total observation period, the California data would reject the hypothesis. The more comprehensive data do not reject the hypothesis, since the other groups contribute more to the degrees of freedom than they contribute to the chi-square value.

APPENDIX A

SUMMARY OF THE CALIFORNIA DATA

This appendix records the data from California relevant to this study. The data were derived from Table N, Appendix I, of a report prepared by the Department of Motor Vehicles of the State of California [5]. That table gives the distribution of licensed drivers in the California data base by number of accidents in each of the calendar years 1961–63 and 1969–74. The number of accidents in the period 1961–63 was ignored in the present analysis since interest was focused on the accidents in two subintervals of a common length, and because instability of accident rates over the intervening period 1963–1969 could be due to the long gap in information. The tabulations presented below were obtained by grouping all combinations of accident numbers that gave the same total for the years 1969–71, and those that gave the same totals for the years 1972–74.

TABLE A1

NUMBER OF DRIVERS WITH M CLAIMS IN PERIOD 1969–71 AND N CLAIMS IN THE PERIOD 1972–74

CALIFORNIA, FEMALE DRIVERS

1972–74	1969–71 Claims						
Claims	m = 0	m = 1	m = 2	m = 3	m = 4	m = 5	Total
n = 0	19,634	1,757	128	10	1	0	21,530
n = 1	1,816	266	33	5	1	0	2,121
n = 2	164	24	8	1	0	0	197
n = 3	16	3	0	1	0	0	20
n = 4	2	0	0	0	0	0	2
n = 5	1	1	0	0	0	0	2
Total	21,633	2,051	170	16	2	0	23,872

TABLE A2

Number of Drivers with m Claims in Period 1969–71 and n Claims in the Period 1972–74

California, Male Drivers

	1969–71 Claims						
1972–74 Claims	m = 0	m = 1	m = 2	m = 3	m = 4	m = 5	Total
n = 0	21,800	3,363	381	41	3	1	25,589
n = 1	3,226	692	124	18	5	0	4,065
n = 2	403	126	26	5	0	1	561
n = 3	44	15	3	2	0	0	64
n = 4	7	2	2	0	0	1	12
n = 5	0	0	1	1	0	0	2
Total	25,480	4,198	537	67	8	3	30,293

TABLE A3

Number of Drivers with m Claims in Period 1969–71 and n Claims in the Period 1972–74

California, All Drivers

	1969–71 Claims						
1972–74 Claims	m = 0	m = 1	m = 2	m = 3	m = 4	m = 5	Total
n = 0	41,434	5,120	509	51	4	1	47,119
n = 1	5,042	958	157	23	6	0	6,186
n = 2	567	150	34	6	0	1	758
n = 3	60	18	4	2	0	0	84
n = 4	9	2	2	0	0	1	14
n = 5	1	1	1	1	0	0	4
Total	47,113	6,249	707	83	10	3	54,165

APPENDIX B

ESTIMATES OF THE COEFFICIENT OF VARIATION OF EXPOSURE LEVEL

A number of studies have used data on mileage driven per unit of time by individual drivers. Unfortunately, none of these studies presents the essential summary statistics, the mean and variance of the mileage. These summary statistics would be sufficient to estimate the coefficient of variation and provide a standard for comparison. Since data are not available for direct estimation, indirect methods of estimation are needed.

The approach taken in this appendix is to assume that the distribution of mileage driven by members of a population is lognormally distributed. Given this assumption, information on fractiles of the distribution would be sufficient to permit an estimate of the coefficient of variation. Even this, however, is not available directly.

A plausible way of inferring fractiles of the distribution is to assume that published statistics will relate to groups that are of reasonable size. The *California Driver Record Book* for 1976 gives accidents per driver and per mile for drivers using their vehicles for specified annual mileages. The lowest category listed is from zero to 2,250 miles; the highest one is over 100,000 miles. We assume that 2,250 and 100,000 are corresponding fractiles on the left and right tails of the distribution. This assumption leads to an estimate of 15,000 for the median annual mileage driver by California drivers, an estimate that appears acceptable. By postulating which fractile corresponds to these numbers, we obtain estimates of the mean annual mileage and the coefficient of variation in this quantity. The estimates are shown in Table B1. Note that the largest and smallest assumed fractiles are not likely to be correct. The highest, one percent, leads to a coefficient of variation for the exposure level which is comparable to that of the aggregate accident rate; this would imply virtually no variability in the accident propensity per mile driven among individuals. The lowest, one per million, would not allow enough drivers in the extreme groups to provide reliable statistics. Between these extremes, the inferred coefficient of variation is fairly stable. Thus, in spite of the lack of direct data, it is plausible that the coefficient of variation in mileage driven is between one quarter and one half.

TABLE B1

ESTIMATES OF THE MEAN AND COEFFICIENT OF VARIATION OF ANNUAL MILEAGE DRIVEN

	Inferred Value of	
Assumed Fractile	Mean	Coefficient of Variation
1/100	20,900	0.94
1/1,000	18,100	0.46
1/10,000	17,100	0.30
1/100,000	16,600	0.22
1/1,000,000	16,200	0.17

REFERENCES

[1] Cramer, H., *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1974.

[2] Feller, W., *An Introduction to Probability Theory and Its Applications, Volume I*, third edition, John Wiley & Sons, New York, NY, 1968.

[3] Greenwood, M., and G. U. Yule, "An Inquiry into the Nature of Frequency Distributions," *Journal of the Royal Statistical Society*, vol. A83, pp. 255–279, 1920.

[4] Kerrich, J. E., "Accident Statistics and the Concept of Accident Proneness, Part II: The Mathematical Background," *Biometrics*, vol. 7, pp. 391–432, 1951.

[5] Kwong, K. W., J. Kuan, and R. C. Peck, *Longitudinal Study of California Driver Accident Frequencies 1: An Exploratory Multivariate Analysis*, Department of Motor Vehicles, State of California, Sacramento, CA, 1976.

[6] Lundberg, O., *On Random Processes and Their Application to Sickness and Accident Statistics*, Almqvist and Wicksells, Uppsala, Sweden, 1940.

[7] Nelson, D., "Comments on 'Good Drivers and Bad Drivers—A Markov Model of Accident Proneness,'" *PCAS* LXVIII, pp. 86–88, 1981.

[8] Peck, R. C., R. S. McBride and R. S. Coppin, "The Distribution and Prediction of Driver Accident Frequencies," *Accident Analysis and Prevention*, vol. 2, pp. 243–299, 1971.

[9] Rolph, J. E., "Some Statistical Evidence on Merit Rating in Medical Malpractice Insurance," *Journal of Risk and Insurance*, vol. 48, pp. 247–260, 1981.

[10] Seal, H. L., *Stochastic Theory of a Risk Business*, John Wiley and Sons, New York, NY, 1969.

[11] Stanford Research Institute, *The Role of Risk Classifications in Property and Casualty Insurance: A Study of the Risk Assessment Process*, Stanford Research Institute, Menlo Park, CA, 1976.

[12] Stewart, J. R., and Campbell, B. J., *The Statistical Association between Past and Future Accidents and Violations*, Highway Safety Research Center, The University of North Carolina, Chapel Hill, NC, 1972.

[13] Thyrion, P., "Contribution a l'Etude du Bonus pour Non-Sinistre en Assurance Automobile," *ASTIN Bulletin*, vol. I, Part III, pp. 142–162, 1960.

[14] Venezian, E. C., "Good Drivers and Bad Drivers—A Markov Model of Accident Proneness," *PCAS* LXVIII, pp. 65–85, 1981.

[15] Venezian, E. C., B. F. Nye, and A. E. Hofflander, "The Distribution of Medical Malpractice Claims—Some Statistical and Public Policy Aspects," *Journal of Risk and Insurance*, vol, 56, pp. 686–701, 1989.

[16] Weber, D. C., "An Analysis of the California Driver Record Study in the Context of a Classical Accident Model," *Accident Analysis and Prevention*, vol. 4, pp. 109–116, 1972.

[17] Woll, R. G., "A Study of Risk Assessment," *PCAS* LXVI, pp. 84–138, 1979.