A STOCHASTIC APPROACH TO AUTOMOBILE COMPENSATION

DONALD C. WEBER

1. INTRODUCTION

In recent years various automobile compensation plans have been proposed in response to adverse criticism of the existing automobile liability system. This paper is an effort to present a probability model which conceivably could provide the mathematical framework for some future no-fault insurance system. Before proceeding, it is only fair to warn that utilization of the proposed model far ratemaking purposes would have a far-reaching effect upon members of the Casualty Actuarial Society. It would mean counting accident involvements rather than claims. It would mean calculating involvement costs rather than claim costs. And finally, it would mean insuring an individual driver rather than an automobile.

2. THE MODEL

Suppose the discrete random variable $N(t)$ represents the number of accident involvements experienced by a motorist during a time interval of length $t > O$ and $p(n,t)$ denotes the probability that $N(t) = n$. Let the cost of an accident involvement be represented by the non-negative continuous random variable X having distribution function $G(x)$. We shall assume that X is independent of time and of the costs of prior involvements. Denote by $G_n(x)$ the probability that the cost of n accident involvements is less than or equal to x. Then if $X(t)$ is the total cost of involvements over a time span of t units, the relationship between these random variables is given by the analytic expression:

$$F(x,t) = \sum_{n=0}^{\infty} p(n,t)G_n(x) \quad , \quad x \geq 0, t > 0, \tag{1}$$

where:

$$F(x,t) = P_r\{X(t) \leq x\},$$
$$p(n,t) = P_r\{N(t) = n\},$$
$$G_n(x) = P_r\{X_1 + X_2 + \ldots + X_n \leq x\} \quad , \quad n > 1,$$
$$G_0(x) = 1 \quad , \quad G_1(x) = G(x)$$

With the aid of characteristics functions, it is not difficult to show that the mean and variance of the random variable $X(t)$ are:

$$(2) \qquad E[X(t)] = E[N(t)]\,E(X),$$
$$Var\,[X(t)] = E^2(X)\,Var\,[N(t)] + E[N(t)]\,Var\,(X),$$

respectively.

The reader will recognize relation (1) as an adaptation of the basic model employed in collective risk theory as developed by Cramér[1] and others, and discussed by Dropkin[2] in his presentation at the Mathematical Theory of Risk meeting in 1966. The problem now is to obtain realistic and adequate functions for $p(n,t)$ and $G(x)$.

3. THE NEGATIVE BINOMIAL MODEL

In 1920, Greenwood and Yule[3] proposed an accident frequency model which assumes that during a time interval of length t the number of accidents, n, experienced by an individual is a Poisson process with mean and variance λt, i.e.:

$$(3) \qquad p(n,t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \quad , \quad \begin{matrix} n = 0, 1, 2, \ldots, \\ \lambda > 0, t > 0, \end{matrix}$$

and that λ is a value of a random variable having a gamma distribution:

$$(4) \qquad u(\lambda) = \frac{(r/m)^r}{\Gamma(r)} \lambda^{r-1} e^{-(r/m)\lambda} \quad , \quad \lambda > 0, m > 0, r > 0,$$

where:

$$\Gamma(r) = \int_0^\infty y^{r-1} e^{-y} dy.$$

The resulting unconditional distribution for n accidents in time t is the negative binomial:

$$q(n,t) = \int_0^\infty p(n,t)u(\lambda)d\lambda$$

$$(5) \qquad = \frac{\Gamma(n+r)}{n!\,\Gamma(r)} \left(\frac{r}{r+mt}\right)^r \left(\frac{mt}{r+mt}\right)^n \quad , \quad \begin{matrix} n = 0, 1, 2, \ldots, \\ r > 0, m > 0, t > 0, \end{matrix}$$

with mean mt and variance $mt(1 + mt/r)$.

---

[1] Crámer, H., *Collective Risk Theory*, Nordiska bokhandeln, Stockholm, 1955.

[2] Dropkin, L. B., "Loss Distributions of a Single Claim," *PCAS Mathematical Theory of Risk*, 1966.

[3] Greenwood, M., and Yule, G. U., "An inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents," *Journal of the Royal Statistical Society, 83* pp. 255-279, (1920).

Using different notation, Dropkin[4] introduced this model to the Casualty Actuarial Society and successfully applied it to data obtained in the 1958 California Driver Record Study. With reference to the motor vehicle accident scene, we may interpret the parameter λ to be the theoretical accident rate per unit time associated with an individual driver. We then assume that this parameter varies from individual to individual within a population of drivers according to the probability density $u(\lambda)$ which has mean m and variance m^2/r. It follows that $q(n,t)$ is the distribution of accidents within a population, i.e., $100\ q(n,t)$ gives the percentage of individuals in the population involved *in* $n = 0,1,2 \cdots$, accidents during a time period of t units.

The California Department of Motor Vehicles kindly provided this writer with data used in their 1964 California Driver Record Study.[5] For this study a random sample constituting about 2% of the licensed drivers in the state was obtained. Of these, data are available on approximately 148,000 motorists over the full observation period of three years, namely 1961-63. These data include information on certain attributes of the individuals in the sample as well as their driving record in terms of traffic offenses and *reportable* accident involvements. In Table 1 we see the fit, using the method of moments, of the negative binomial model (5) to the empirical accident distributions generated by these 148,000 individuals during the specified time intervals. Notice that due to the time lag between the occurrence of an accident and the processing of the resulting accident report, the 1963 period is estimated to represent a 10½-month interval rather than a full year.

The closeness of fit duplicates the results obtained by Dropkin using the earlier California data. However, it is important to observe that the parameters m and r, as shown by their estimates together with the standard deviation of these estimates, do not seem to remain constant over time. This would suggest that a shifting takes place in the underlying distribution (4), $u(\lambda)$, which implies that the parameter λ of relation (3) is a *function of time.*

In Table 2 we find the negative binomial fitted to the California data by sex. As with the combined data, the negative binomial distribution provides

[4] Dropkin, L. B., "Some Considerations on Automobile Rating Systems Utilizing Individual Driving Records," *PCAS* XLVI, p. 165.

[5] California Department of Motor Vehicles, State of, "The 1964 California Driver Record Study Parts 1-9," Sacramento, California 1964 and 1967.

Table 1

Comparison of Actual and Theoretical (Negative Binomial) Accident Distributions
1964 California Driver Record Study

	No. of Accidents	Actual Distribution	Theoretical Distribution
1961 - 1963			
$\hat{m} = 0.0711 \pm 0.0004$	0	122,593	122,638
$\hat{r} = 1.1400 \pm 0.0378$	1	21,350	21,257
$t = 2.875$	2	3,425	3,457
$\chi^2 = 1.61$, 3 d.f.	3	530	550
	4	89	86
	5+	19	18
	Total	148,006	148,006
1961 - 1962			
$\hat{m} = 0.0709 \pm 0.0005$	0	129,524	129,541
$\hat{r} = 1.0773 \pm 0.0473$	1	16,267	16,236
$t = 2$	2	1,966	1,963
$\chi^2 = 4.90$, 3d.f.	3	211	234
	4	31	28
	5+	7	4
	Total	148,006	148,006
1961			
$\hat{m} = 0.0696 \pm 0.0007$	0	138,343	138,353
$\hat{r} = 1.0691 \pm 0.0894$	1	9,072	9,042
$t = 1$	2	547	571
$\chi^2 = 1.47$, 1d.f.	3+	44	40
	Total	148,006	148,006
1962			
$\hat{m} = 0.0722 \pm 0.0007$	0	138,087	138,094
$\hat{r} = 0.8469 \pm 0.0585$	1	9,211	9,191
$t = 1$	2	650	668
$\chi^2 = 1.00$, 1d.f.	3+	58	53
	Total	148,006	148,006
1963			
$\hat{m} = 0.0715 \pm 0.0008$	0	139,326	139,330
$\hat{r} = 0.8712 \pm 0.0701$	1	8,140	8,133
$t = 0.875$	2	505	509
$\chi^2 = 0.07$, 1d.f.	3+	35	34
	Total	148,006	148,006

Table 2

Comparison of Actual and Theoretical Accident Distributions by Sex
1964 California Driver Record Study

	No. of Accidents	Actual Distribution	Theoretical Distribution
		Males	
1961			
$\hat{m} = 0.0885$	0	79,595	79,606
$\hat{r} = 1.3420$	1	6,638	6,606
$t = 1$	2	451	479
$\chi^2 = 3.19$, 1d.f.	3+	42	35
		86,726	86,726
1962			
$\hat{m} = 0.0925$	0	79,358	79,365
$\hat{r} = 1.0599$	1	6,775	6,752
$t = 1$	2	538	559
$\chi^2 = 1.37$, 1d.f.	3+	55	50
		86,726	86,726
1963			
$\hat{m} = 0.0901$	0	80,369	80,372
$\hat{r} = 1.0648$	1	5,910	5,902
$t = 0.875$	2	415	420
$\chi^2 = 0.07$, 1d.f.	3+	32	32
		86,726	86,726
		Females	
1961			
$\hat{m} = 0.0430$	0	58,748	58,747
$\hat{r} = 1.2423$	1	2,434	2,439
$t = 1$	2	96	91
$\chi^2 = 0.62$, 1d.f.	3+	2	3
		61,280	61.280
1962			
$\hat{m} = 0.0436$	0	58,729	58,726
$\hat{r} = 0.9244$	1	2,436	2,443
$t = 1$	2	112	106
$\chi^2 = 1.16$, 1d.f.	3+	3	5
		61,280	61,280
1963			
$\hat{m} = 0.0451$	0	58,957	58,956
$\hat{r} = 0.9311$	1	2,230	2,232
$t = 0.875$	2	90	88
$\chi^2 = 0.30$, 1d.f.	3+	3	4
		61,280	61,280

a remarkably close fit in every case. Again, the fluctuation in parameter values from one time period to another is apparent. Also, we notice that the distributions of male and female accident involvements are different.

Using the negative binomial model, Arbous and Kerrich,[6] Bates and Neyman[7] and Edwards and Gurland[8] derived various bivariate accident distribution models which differ in certain underlying assumptions. These bivariate models can be used to obtain theoretical distributions of future accidents based upon the number of past accidents. The bivariate negative binomial of Kerrich, which assumes a constant parameter λ (past and future), appeared in the *Proceedings* in a paper by Dropkin[9] and was applied to Canadian data by Hewitt.[10] Actuaries have long recognized, however, that factors other than accident history are related to future automobile accident experience, e.g., age, sex, geographic location, mileage driven, conviction history. In fact, only in recent years has accident experience been incorporated in the ratemaking procedures.

4. ACCIDENT RATE POTENTIAL

For the moment, let us accept the idea that an individual driver's accident frequency over a short period of time is a Poisson process (1). Let us assume that each motorist is characterized by his own particular λ which is a function of accident likelihood variables such as physical, mental and emotional states, attitudes, motor abilities, habits, alertness, environmental driving conditions and amount of driving exposure. In view of the previous section, the parameter λ is a function of time through changing conditions and, therefore, any estimate of this parameter requires frequent updating. However, let us treat λ as a constant over relatively short periods of time in the absence of major changes in the above variables. Thus we may view λ as the result of averaging the individual's accident likelihood variables over

[6] Arbous, A. G. and Kerrich, J. E., "Accident Statistics and the Concept of Accident-Proneness," *Biometrics, 7* pp. 340-432 (1951).

[7] Bates, G. E. and Neyman, J., "Contributions to the Theory of Accident Proneness," *University of California Publications in Statistics I*, pp. 215-276, (1952).

[8] Edwards, C. B., and Gurland, J., "A Class of Distributions Applicable to Accidents," *Journal of the American Statistical Association, 56* pp. 503-517, (1961).

[9] Dropkin, L. B., "Automobile Merit Rating and Inverse Probabilities," *PCAS* XLVII, p. 37.

[10] Hewitt, Jr., C. C., "The Negative Binomial Applied to the Canadian Merit Rating Plan for Individual Automobile Risks," *PCAS* XLVII, p. 55.

the observation period. This "constant" will hereafter be called the *accident rate potential* associated with the individual driver.

Clearly, most of the accident likelihood variables are not directly measurable. We are therefore confronted with the task of trying to estimate an individual's accident rate potential on the basis of available information, information which at best reflects to an unknown degree the actual accident likelihood of the driver. The information used to estimate λ will be called criteria, or criterion variables.

5. THE MODEL FOR ACCIDENT RATE POTENTIAL

On the basis of the analyses appearing in the 1964 California Driver Record Study, this writer chose as criterion variables: sex, marital status, residence, age, conviction history and accident history. Proceeding on the evidence that the distribution of accidents within a population of drivers is negative binomial, if these criteria are truly effective predictors of λ, they should be able to subdivide the California sample into homogeneous groups with respect to accident rate potential, i.e., into "Poisson groups." In an effort to establish the effectiveness of the criterion variables, the 148,000 individuals in the sample were partitioned into 2,880 groups on the basis

TABLE 3

Criterion Variables used to Partition California Sample

Sex	*Marital Status*	*Residence (Counties)*
Male	Married	Area 1: Los Angeles, San Francisco
Female	Single	Area 2: Alameda, Contra Costa, Marin, Orange, Sacramento, San Mateo, Santa Clara
		Area 3: Fresno, San Joaquin, Stanislaus, Yolo
		Area 4: All Other Counties

Age in 1963	*No. of Convictions, 1961-62*	*No. of Accidents, 1961-62*
Less than 21	0	0
21 - 25	1	1
26 - 30	2	2
31 - 40	3	3
41 - 60	4	More than 3
Over 60	More than 4	

of the six chosen criteria. The levels within each variable are presented in Table 3. A computer program printed out the 1963 accident distributions for the 193 groups that contained 100 or more individuals and fitted a Poisson distribution to each such group. In 167 or 86.5% of the cases the hypothesis of a Poisson distribution was acceptable at the .05 level of significance. One may conclude from these results that the six criterion variables did a credible job of classifying the individuals according to negative binomial theory.

On the basis of the above experiment, let us assume that the accident rate potential characterizing an individual is a function of a number of criterion variables, i.e.:

$$\lambda = f(\underline{x} : \underline{\beta}) \tag{6}$$

where $\underline{x}$ represents a vector of criterion variables and $\underline{\beta}$ is a vector of parameters. Our next task is to determine the functional form of f. To do this we turn our attention to the 1964 California Driver Record Study in order to examine the relationship between accident frequency and the selected criterion variables, taken one at a time.

In the partitioning experiment the primary basis used for determining the levels of the area variable was accident rate by drivers residing in a county. During the experience period of the California study, the accident rate per driver is given in Table 4. It reveals that accident rates do indeed

TABLE 4

Accident Rates per driver by Area and County
1964 California Driver Record Study

Area 1 (61,594 cases)	
Los Angeles	0.241
San Francisco	.245
Area 1 Ave.	0.241

Area 2 (37,690 cases)	
Alameda	0.225
Contra Costa	.202
Marin	.201
Orange	.218
Sacramento	.217
San Mateo	.210
Santa Clara	.199
Area 2 Avg.	0.213

Area 3 (7,647 cases)	
Fresno	0.184
San Joaquin	.187
Stanislaus	.172
Yolo	.191
Area 3 Avg.	0.183

Area 4 (40,474 cases)	
All Other	0.147

vary from area to area within the state and, in general, the more populous the area, the higher the accident rate.

To take advantage of the positive correlation between accident rates and population density, it was decided to use the county traffic density index as a criterion variable. This index is defined as the ratio of total registered vehicles in a given county to the total linear miles of roadway in that county. It must be recognized that the use of a countywide index somewhat understates the relationship between accidents and density since the population density within many of the California counties is anything but uniform. In order to fully utilize the predictive power inherent in a traffic density factor with respect to accident frequency an index by geographical area rather than by county lines is needed.

A plot of accident rate versus traffic density index reveals that the mathematical relationship between these two variables is concave downward. The correlation coefficient corresponding to a simple regression analysis of accident rate on the logarithm of traffic density index was 0.85. Accordingly, we will assume that the relationship between mean accident frequency, denoted by y, and the natural logarithm of traffic density index, denoted by x_1, to be:

$$y = a_1 + b_1 x_1$$

where a_1 and b_1 are constants to be estimated.

In Part 5 of the 1964 California Driver Record Study charts are given which visually depict the relationships between accident rates and the personal characteristics (i) sex, (ii) marital status, and (iii) age. These charts again reveal that males and females constitute distinct driving populations, i.e., the relationships are of different character in the two populations. As a consequence, it is necessary to search out a function f in (6) for each of the two sexes.

These charts show that the driving record of married females is better than that of single females at all ages, although the difference is not constant. With the exception of two age groups (under 26 and 56-60), the same statement can be made about male drivers. In order to give recognition to the apparent significant relationship between accident rate, y, and marital status, x_2, we will assume the step function relation:

$$y = a_2 + b_2 x_2$$

where $x_2 = 0$ for a married individual and $x_2 = 1$ for a single person and a_2 and b_2 are constants. Since this assumes a constant difference in mean

accident frequencies between marrieds and unmarrieds, the above formula is acknowledged to be an approximation to the actual situation at best.

Using a transformation on age, it is possible to reasonably express the relationship between accident rates, y, and a function of age, x_3, in the linear form:

$$y = a_3 + b_3 x_3$$

where a_3 and b_3 are parameters to be estimated. Applied to the California data, this writer used $x_3 = 5/(age - 13)$ for males and $x_3 = 125/(age - 13)^3$ for females. The estimates for the parameters obtained from weighted regression analyses are given in Tables 5 and 6 together with the fit to the corresponding empirical rates.

TABLE 5

Weighted Regression of Accident Rates on Transformed Ages (Males)
1964 California Driver Record Study (1961-63)

$y = 0.1823 + 0.3183x_3$
where $x_3 = 5/(age - 13)$

Age Class	*Empirical Accident Rate*	*Theoretical Rate y*
Under 21	0.468	0.459
21 - 25	.332	.341
26 - 30	.290	.288
31 - 40	.253	.253
41 - 60	.229	.226
Over 60	.204	.210

In continuing our search for the functional form of f in (6) we next investigate the possibilities of predicting accident involvement using driver record data. Part 4 of the 1964 California Driver Record Study discusses the relationship between accident and conviction frequencies based upon a three-year experience period involving the 148,000 drivers in the sample. At this point a conviction is defined as a traffic conviction which counts toward an individual's negligent operator point total. This includes all violations involving the safe operation of a motor vehicle as defined in Section 12810 of the California Vehicle Code. In this study, the number of convictions understates the actual number of vehicle code violations in that multiple citations relating to a single incident were counted as one. Also,

TABLE 6

Weighted Regression of Accident Rates on Transformed Ages (Females)
1964 California Driver Record Study (1961-63)

$$y = 0.1191 + 0.1365x_3$$
where $x_3 = 125/(\text{age} - 13)^3$

Age Class	*Empirical Accident Rate*	*Theoretical Rate y*
Under 21	0.209	0.209
21 - 25	.138	.136
26 - 30	.118	.124
31 - 40	.121	.121
41 - 60	.120	.120
Over 60	.121	.119

to avoid a "built-in" correlation between accidents and countable convictions, the number of convictions does not include those resulting from an accident investigation. Harwayne's account[11] in the *Proceedings* on the earlier California study revealed the near linear relationship between accident rates and countable convictions. Accordingly, a weighted regression analysis on conviction counts was performed for each sex. The actual and predicted means are given in Table 7 where y is the mean accident frequency and x_i is the number of conviction counts.

TABLE 7

Weighted Regression of Accident Rates on Number of Countable Convictions
1964 California Driver Record Study (1961-63)

Males: $y = 0.1733 + 0.0953x_4$
Females: $y = 0.0999 + 0.0823x_4$

Number of Convictions (x_4)	*Males*		*Females*	
	Actual	*Theoretical (y)*	*Actual*	*Theoretical (y)*
0	0.17	0.17	0.10	0.10
1	.28	.27	.18	.18
2	.37	.36	.27	.26
3	.45	.46	.37	.35
4	.58	.55	.44	.43
More than 4	.68	.75	.49	.59

[11] Harwayne, F., "Merit Rating in Private Passenger Automobile Liability Insurance and the California Driver Record Study," *PCAS* XLVI, p. 189.

It is necessary to point out explicitly that Table 7 shows a concurrent relationship between accidents and convictions, i.e., the counts for both variables arise from the same experience period. What is of greater interest to us is the predictive nature of past convictions as it concerns future accidents. In this regard Table 8 displays the combined experience of all drivers in the California sample as taken from the tabulation which partitioned the sample into homogeneous groups. There we find that the relationship between 1963 empirical accident rates and 1961-62 conviction counts is dominantly linear by checking the differences in accident rates as we go from one conviction level to the next.

TABLE 8

Observed 1963 Accident Rates by 1961-62 Conviction Counts
1964 California Driver Record Study

No. of Convictions 1961-62	*Empirical Accident Rates 1963*
0	0.0466
1	.0834
2	.1106
3	.1411
More than 3	.1707

Although the relationship in this instance is not as strongly linear as in the concurrent case, let us tacitly assume that the relation:

$$y = a_4 + b_4 x_4$$

also holds when y is defined as future mean accident frequency and x_4 represents number of convictions as it pertains to the prior time interval.

If we accept the tenet that the negative binomial model is at least an approximation to actual automobile experience, we would expect future accident rates to be linearly related to the incidence of past accident involvements on a theoretical basis. See, for example, Dropkin[12] and Hewitt[13]. To confirm this, iterative weighted regression analyses for:

$$y = a_5 + b_5 x_5$$

[12] Dropkin, L. B., *op. cit.*
[13] Hewitt, Jr., C. C., *op. cit.*

were performed using a computer. Here y represents 1963 accident rates and x_5 is the number of 1961-62 involvements. Further discussion of the iterative procedure used appears in the next section. The results of these analyses are given in Table 9:

TABLE 9

Weighted Regression of 1963 Accidents on 1961-62 Accident Counts
1964 California Driver Record Study

Males: $y = 0.07234 + 0.03818x_5$

Females: $y = 0.03686 + 0.03090x_5$

No. of Accidents 1961-62 (x_5)	*Males*		*Females*	
	Actual Rates	*Theoretical (y)*	*Actual Rates*	*Theoretical (y)*
0	0.0721	0.0723	0.0368	0.0369
1	.1112	.1105	.0677	.0678
2+	.1454	.1529	.1008	.1004

A final candidate for a criterion variable is non countable convictions. A non countable conviction is defined as a traffic conviction which does not involve the safe operation of a motor vehicle, e.g., a conviction in connection with certain non moving offenses. The relationship between accidents and non countable convictions was not given separate analysis in the 1964 California Driver Record Study nor did this writer look into the matter. However, in Part 8 of the California study a significant relationship was observed, at least as it concerns concurrent data. Having no reason to believe that the mathematical form of the relationship between accidents and non countable convictions should be different than that between accidents and countable convictions, let us assume the equation:

$$y = a_6 + b_6 x_6$$

where y is the future accident rate and x_6 is the prior non countable conviction count.

On the basis of the linear relationships between accident rates and the investigated criterion variables, let us hypothesize that, in general, the function of f of (6) is given by:

$$\lambda = f(\underline{x}\,;\,\underline{\beta}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{7}$$

where the x_i are the criterion variables which functionally determine the

value of λ and the β_i are the necessary parameters. Now (7) together with (3) permits us to finalize the form of $p(n,t)$ in (1) in terms of the characteristics of the driver, namely:

$$(8) \qquad p(n,t) = \frac{e^{-t\sum_{i=0}^{k}\beta_i x_i}\left(t\sum_{i=0}^{k}\beta_i x_i\right)^n}{n!} \qquad n = 0, 1, 2, \ldots, \quad \sum_{i=0}^{k}\beta_i x_i > 0, t > 0,$$

where; $x_0 = 1$.

Before taking up the problem of estimating λ using the California data, a few comments of the limitations of the data are in order. We have discussed six possible candidates for criterion variables. That does not mean, of course, that these six are the only predictors that have a significant mathematical relationship with accident involvements. For example, miles of driving may be a most significant factor but the California records do not give this information and hence we are unable to *directly* include this variable in our analysis. If at some future date, exposure mileage information by driver were available, it is likely that the relationship between it and accident rate would be found to have a highly significant linear component. Should that be the case, the variable "driving mileage" would take its place as one of the k predictors in relation (7).

At this point it is also appropriate to remind ourselves of other limitations in this study. Recall that our accident count includes only reported accidents, but unreported accidents according to other studies are more numerous than those reported to authorities. Therefore we cannot claim that the relationships derived in this study are applicable when the number of accidents is taken to mean *all* accidents. Also, our estimate of λ in the sections to follow will be based on reported accidents only and so, in terms of all involvements, it will be an understatement. Similarly, our conviction count includes only the incidence of detected violations. Surely, this count is a gross understatement of the number of actual violations and we cannot assume the degree of understatement to be uniform.

6. ESTIMATION OF ACCIDENT RATE POTENTIAL: THEORY

For the sake of simplification but without loss of generality, let $t = 1$ in function (8) in the development that follows. Then the probability that

the j^{th} individual in the sample will be involved in n_j accidents during the next unit of time is given by:

$$(9) \qquad p(n_j) = \frac{e^{-\sum_{i=0}^{k}\beta_i x_{ij}} \left(\sum_{i=0}^{k} \beta_i x_{ij} \right)^{n_j}}{n_j!} \quad , \quad n_j = 0, 1, 2, \ldots , \sum_{i=0}^{k} \beta_i x_{ij} > 0.$$

To obtain the maximum likelihood estimates for the parameters we observe that with respect to a sample of size s, the likelihood function is:

$$L = \prod_{j=1}^{s} \frac{e^{-\sum_{i=0}^{k}\beta_i x_{ij}} \left(\sum_{i=0}^{k} \beta_i x_{ij} \right)^{n_j}}{n_j!} \quad .$$

Taking the natural logarithm, we obtain:

$$ln\, L = -\sum_{j=1}^{s} \sum_{i=0}^{k} \beta_i x_{ij} + \sum_{j=1}^{s} n_j \, ln \left(\sum_{i=0}^{k} \beta_i x_{ij} \right) - \sum_{j=1}^{s} ln\, n_j!$$

Differentiating with respect to β_i, $i = 0, 1, 2, \ldots, k$, we get:

$$\frac{\partial \, ln\, L}{\partial \beta_i} = -\sum_{j=1}^{s} x_{ij} + \sum_{j=1}^{s} \frac{n_j x_{ij}}{\left(\sum_{i=0}^{k} \beta_i x_{ij} \right)} \quad .$$

On setting the $k + 1$ partials equal to zero, the system of maximum likelihood normal equations otained is:

$$(10) \qquad \sum_{j=1}^{s} \frac{n_i x_{ij}}{\left(\sum_{i=0}^{k} \hat{\beta} x_{ij} \right)} = \sum_{j=1}^{s} x_{ij} \quad , \quad i = 0, 1, \ldots , k \, .$$

In a related but slightly different context, Jorgenson[14] showed that a solution to the set of equations (10) can be obtained by using an iterative weighted least squares procedure. If N_j is the random variable having distribution (9), then the parameter λ associated with the j^{th} individual is:

$$\lambda_j = \epsilon(N_j) = \sum_{i=0}^{k} \beta_i x_{ij} = \text{Var}\,(N_j)$$

[14] Jorgenson, D. W., "Multiple Regression Analysis of a Poisson Process," *Journal of the American Statistical Association 56,* pp. 235-245, (1961).

In matrix notation:

$$\underline{\lambda} = \epsilon(\underline{N}) = X\underline{\beta} \quad \text{and} \quad \text{Cov}\,(\underline{N}) = V$$

where the underline of λ and N denote column vectors of dimension s, $\underline{\beta}$ is a vector of $k + 1$ parameters, X is an $s \times (k + 1)$ matrix having the values of the criterion variables as elements and V is an $s \times s$ diagonal matrix with elements $v_j = \sum_{i=0}^{k} \beta_i x_{ij}$. It is well known (e.g., see Goldberger[15] that the minimum variance linear unbiased estimator of $\underline{\beta}$ is:

$$\underline{\hat{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}\underline{N}$$

with:

$$\text{Cov}\,(\underline{\hat{\beta}}) = (X'V^{-1}X)^{-1}$$

The notation X' denotes the transpose of the matrix X and V^{-1} denotes the inverse of V. According to general linear model theory, if $\underline{x_j}$ is the vector of criterion values corresponding to the j^{th} individual, an unbiased estimator for $\epsilon\,(N_j)$ is $\underline{x}_j\underline{\beta}$ with the variance of this estimator being $\underline{x'}_j(X'V^{-1}X)^{-1}\underline{x}_j$.

Unfortunately, since $\underline{\beta}$ is unknown, the matrix V is unknown. Our problem then is to obtain an estimate of V which in turn gives us an estimate of $\underline{\beta}$. Following Jorgenson, we let $\hat{V}_m$ denote the estimate of V obtained on the m^{th} iteration and we let the corresponding estimate of $\underline{\beta}$ be:

$$\underline{b}_m = (X'\hat{V}_m^{-1}X)^{-1}X'\hat{V}_m^{-1}\underline{n}\,.$$

Let $\hat{V}^0$ be the $s \times s$ identity matrix and define:

$$\hat{V}_{m+1} = \text{diag}\,[\underline{x_1}'\,\underline{b}_m\,,\ \underline{x_2}'\,\underline{b}_m\,,\ \ldots\,,\ \underline{x}'_s\,\underline{b}_m]$$

where x_j is defined as before. The iterations are continued until convergence is realized, i.e., $\underline{b}_{m+1} = \underline{b}_m$. Denote this equality vector by $\underline{b}$. Then:

$$\underline{b} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\underline{n} \tag{11}$$

where $\hat{V}$ is the equality matrix $\hat{V}_{m+1} = \hat{V}_m$. As our final estimate of λ_j we may use:

$$\hat{\lambda}_j = \underline{x_j}'\,\underline{b} \tag{12}$$

and as an estimate of the variance of $\hat{\lambda}_j$ we may use:

$$\widehat{\text{Var}}\,(\hat{\lambda}_j) = \underline{x_j}'\,(X'\hat{V}^{-1}X)^{-1}\underline{x_j}\,. \tag{13}$$

[15] Goldberger, A. S., *Econometric Theory* (John Wiley and Sons, Inc., New York, 1964).

Because of having to use $\hat{V}$ instead of V, the estimate (12) is not unbiased and its variance is unknown but Jorgenson[16] points out that it is best asymptotically normal (BA). He also notes that the iterative procedure converges provided that $\hat{V}_m$ and $(X'\hat{V}_m^{-1}X)^{-1}$ are positive definite for all m.

Work by Wald[17] provides a theoretical basis for testing:

$$H_o : L\underline{\beta} = \underline{\gamma}$$

where L is a known $l \times (k+1)$ matrix of rank $l \leq k+1$ and $\underline{\gamma}$ is a specified vector of constants. The appropriate test statistic:

$$(L\underline{b} - \underline{\gamma})'[L(X'V^{-1}X)^{-1}L']^{-1}(L\underline{b} - \underline{\gamma})$$

is asymptotically distributed as chi-square with l degree of freedom. This, of course, can be used to test such hypotheses as:

$$H_o : \beta_i = 0 \quad \text{and} \quad H_o : \lambda = \underline{x}'\underline{\beta} = \lambda_o$$

In this study, the vector $\underline{b}_{m+1}$ was calculated using a standard least squares linear regression program after applying a weight of:

$$\left(\sum_{i=0}^{k} b_{i(m)} x_{ij} \right)^{-1/2}$$

to the data. Here $b_{i(m)}$ is the i^{th} element in the vector $\underline{b}_m$. The usual regression program then obtains $\underline{b}_{m+1}$ by solving the system of $k+1$ equations:

$$\sum_{j=1}^{s} \frac{n_j x_{ij} - x_{ij} \sum_{i=0}^{k} b_{i(m+1)} x_{ij}}{\sum_{i=0}^{k} b_{i(m)} x_{ij}} = 0 \quad , \quad i = 0, 1, \ldots, k \, .$$

It is readily seen that this system reduces to (10) when $b_{i(m+1)} = b_{i(m)}$ for all $i = 0,1,2 \ldots ,k$.

7. ESTIMATION OF ACCIDENT RATE POTENTIAL: EXAMPLES

In this section we illustrate the use of the multiple Poisson regression technique applied to the California data. Recall that in Section 5 we selected

[16] Jorgenson, D. W., *op. cit.*

[17] Wald, A., "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large," *Transactions of American Mathematical Society, 54* pp. 426-482, (1943).

six criterion variables to use as accident rate potential predictors. To review, for any given individual in the California sample, these are:

$x_0 \equiv 1$

x_1 = the natural logarithm of the traffic density index of the county in which the driver resides.

$x_2 = \begin{cases} 0 & \text{, if married,} \\ 1 & \text{, if single.} \end{cases}$

$x_3 = \begin{cases} 5/(\text{age} - 13), & \text{if male,} \\ 125/(\text{age} - 13)^3, & \text{if female.} \end{cases}$

x_4 = the number of countable convictions incurred during years 1961-62.

x_5 = the number of accident involvements incurred during years 1961-62.

x_6 = the number of noncountable convictions incurred during years 1961-62.

Initially, usual least squares analyses were run in order to determine which of the six criterion variables are significant in the presence of the others. The results of these analyses are given in Table 10. Comparing with a critical t value of 1.96 at the .05 significant level, we notice that marital status is a nonsignificant variable in the regression equation for males and two variables, age and noncountable conviction history, are not significant for females. We find that conviction history contributes more to accident prediction than any other variable in both regressions. For males, the degree of contribution to regression by the remaining four significant variables is about equal. The second most significant predictor for females is marital status, while traffic density and accident history provide comparable information in the presence of the other variables.

Table 11 displays the final estimation functions for λ and the estimates of the covariance matrix of the $\underline{\beta}$ estimators. In Table 12, the values of $\hat{\lambda}$ and its estimated standard deviation are given for selected values of the criterion variables. Remember that the estimating equations and the estimates of λ found in the tables reflect a time unit of approximately 10½ months rather than 1 year.

8. DISTRIBUTION OF ACCIDENT INVOLVEMENT COSTS

Because of the rarity of the event of an accident in time and the extreme variability in accident costs, a theoretical distribution of accident costs appli-

Table 10

Unweighted Regression of 1963 Accidents on Six Criterion Variables
1964 California Driver Record Study

Males

Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F Value
Regression	6	108.4692	18.07820	215.19
Residual	86,463	7,263.9349	0.08401	
Total	86,469	7,372.4041		

i	b_i	s_{b_i}	$t = b_i/s_{b_i}$
0	-0.00211		
1	0.01023	0.00117	8.73
2	-0.00081	0.00265	-0.30
3	0.05746	0.00672	8.55
4	0.01980	0.00097	20.37
5	0.02251	0.00224	10.06
6	0.01768	0.00187	9.47

Females

Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F Value
Regression	6	21.7867	3.63112	88.93
Residual	61,118	2,495.6405	0.04083	
Total	61,124	2,517.4272		

i	b_i	s_{b_i}	$t = b_i/s_{b_i}$
0	-0.00857		
1	0.00794	0.00098	8.06
2	0.02099	0.00207	10.16
3	-0.00012	0.00074	-0.17
4	0.01832	0.00141	12.96
5	0.02097	0.00273	7.68
6	0.00356	0.00516	0.69

Table 11

Estimation Function for Accident Rate Potential and Covariance Matrix
1964 California Driver Record Study

Males

$$\hat{\lambda} = 0.00274 + 0.00909x_1 + 0.0532x_3 + 0.0223x_4 + 0.0216x_5 + 0.0169x_6$$

$$(X'\hat{V}^{-1}X)^{-1} = 10^{-4}\begin{bmatrix} 0.1981 & -0.0384 & -0.0754 & 0.0041 & 0.0021 & -0.0024 \\ -0.0384 & 0.0085 & -0.0004 & -0.0012 & -0.0013 & 0.0007 \\ -0.0754 & -0.0004 & 0.4058 & -0.0137 & -0.0057 & -0.0162 \\ 0.0041 & -0.0012 & -0.0137 & 0.0142 & -0.0052 & -0.0050 \\ 0.0021 & -0.0013 & -0.0057 & -0.0052 & 0.0654 & -0.0030 \\ -0.0024 & 0.0007 & -0.0162 & -0.0050 & -0.0030 & 0.0623 \end{bmatrix}$$

Females

$$\hat{\lambda} = -0.00176 + 0.00646x_1 + 0.0209x_2 + 0.0196x_4 + 0.0205x_5$$

$$(X'\hat{V}^{-1}X)^{-1} = 10^{-4}\begin{bmatrix} 0.0991 & -0.0211 & 0.0010 & 0.0016 & 0.0013 \\ -0.0211 & 0.0048 & -0.0015 & -0.0011 & -0.0011 \\ 0.0010 & -0.0015 & 0.0538 & -0.0045 & -0.0044 \\ 0.0016 & -0.0011 & -0.0045 & 0.0355 & -0.0089 \\ 0.0013 & -0.0011 & -0.0044 & -0.0089 & 0.1186 \end{bmatrix}$$

cable to a particular individual cannot be arrived at through the observation of that person's involvement costs over a period of time. Therefore, to gain information about costs applicable to a type of driver it is necessary to look at samples taken from a population of drivers. One such sample is the subject of the study entitled Cost of Motor Vehicle Accidents to Illinois Motorists, 1958,[18] and a subsequent analysis,[19] completed in cooperation with the U.S. Bureau of Public Roads. The passenger car portion is based upon a sample of 2,878 reported and 505 unreported accident involvements. A stratified sampling design was used with the sample size in each stratum determined on the basis of an accuracy level specifying an objective 7% relative error. In terms of a stratum mean $\bar{x}$ and its standard deviation s_x, this implies:

$$s_{\bar{x}}/\bar{x} = 0.07$$

[18] Illinois Department of Public Works and Buildings, State of, "Cost of Motor Vehicle Accidents to Illinois Motorists, 1958," Chicago, 1962.

[19] Billingsley, C. M. and Jorgenson, D. P., "Analyses of Direct Costs and Frequencies of Illinois Motor-Vehicle Accidents, 1958," *Public Roads 32*, pp. 201-213 (1963).

Table 12

Accident Rate Potential Estimates and their Standard Deviations
1964 California Driver Record Study

Sex	Traffic Density	Marital Status	Age	Ct. Conv. History	Accident History	No. Ct. Conv. History	$\hat{\lambda}$	$\sqrt{\widehat{\mathrm{Var}}(\hat{\lambda})}$
Male	10	-	60	0	0	0	0.0293	0.0023
Male	50	-	60	1	0	1	.0831	.0027
Male	150	-	60	1	1	1	.1147	.0034
Male	10	-	40	2	0	0	.0781	.0032
Male	50	-	40	0	1	0	.0697	.0027
Male	150	-	40	0	0	0	.0581	.0011
Male	10	-	20	1	1	0	.1056	.0045
Male	50	-	20	0	0	0	.0763	.0035
Male	150	-	20	3	2	1	.2132	.0059
Female	10	Married	-	0	0	-	.0131	.0017
Female	10	Single	-	0	1	-	.0545	.0043
Female	50	Married	-	1	1	-	.0636	.0036
Female	50	Single	-	1	0	-	.0640	.0027
Female	150	Married	-	0	0	-	.0306	.0009
Female	150	Single	-	2	1	-	.1112	.0047

Since the accident data in the 1964 California Driver Record Study, cited earlier, refers to reported involvements, consistency dictates that we confine our attention to the 2,878 reported cases in the Illinois study. These were comprised of 332 fatal injury, 1,730 nonfatal injury and 816 property damage only cases. After appropriate expansion factors were applied, a "population" of 317,051 reported involvement costs was obtained. The distribution of these costs is given graphically in Figure 1.

In the Illinois study, direct costs are defined as "the money value of damages and losses to persons and property resulting directly from accidents, and which might be saved for the motor vehicle owner by the elimination of accidents."[20] Elements of direct costs include damaged property, injuries to persons, value of time lost, loss of use of vehicle, legal and court costs, and damages awarded in excess of costs. Funeral expenses in connection with a motor vehicle accident were not considered a direct cost since such costs are inevitable; an accident merely fixes the time when they are incurred. In evaluating direct costs in multiple car accidents, only those costs associated with the sample car and its occupants were obtained. However, damage to objects other than another motor vehicle, including pedestrians, was obtained.

The Illinois cost study spotlights the two most outstanding characteristics of accident cost distributions:

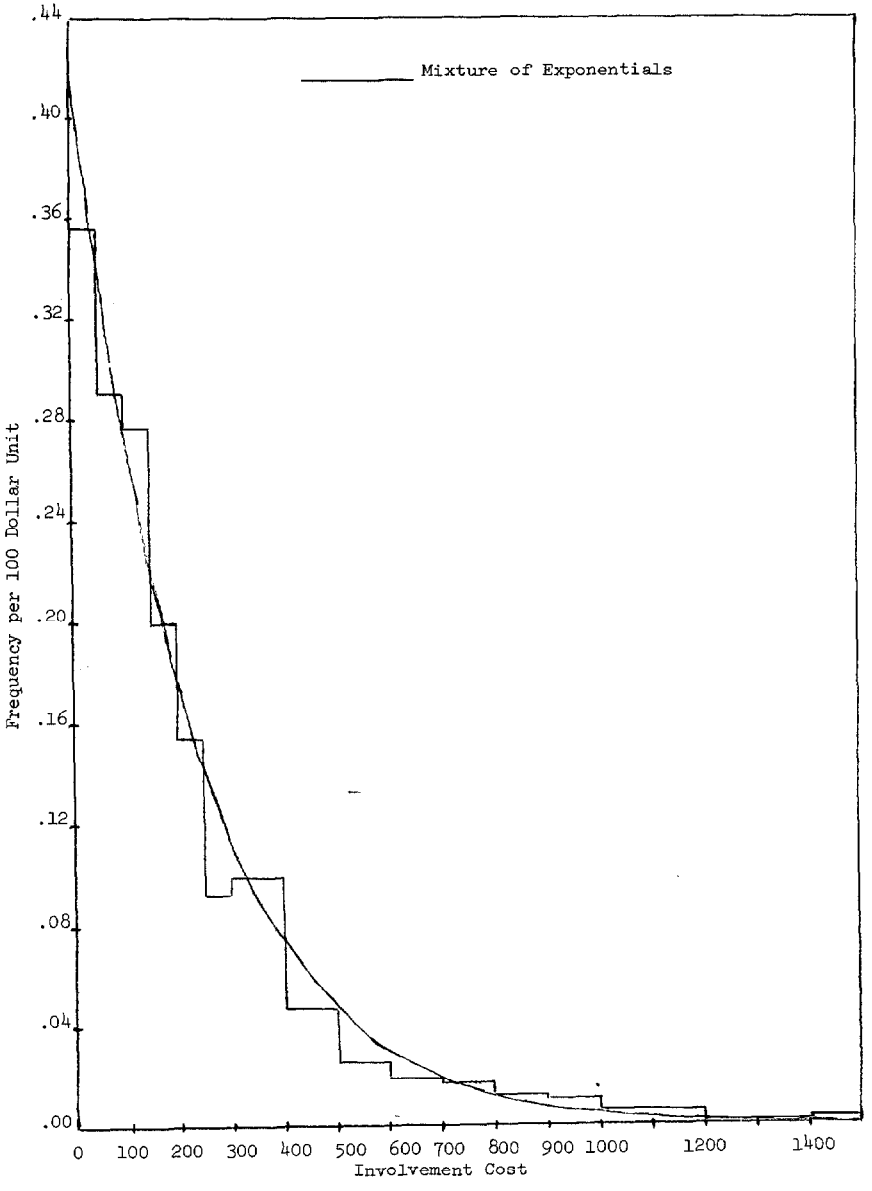
(i) The overall distribution is J-shaped, i.e., low cost accidents are most frequent and high cost accidents least frequent.
(ii) Accident costs depend on where the accident takes place (e.g., urban or rural, divided or undivided highway, intersection or freeway, etc.) and circumstances surrounding the accident (e.g., object struck, number of occupants, speed, etc.).

In an attempt to infer a theoretical distribution of accident costs based on the Illinois data, we must bear in mind that the empirical distribution displayed in Figure 1 is a "dangerous" one and should be taken at somewhat less than "face value." The *constructed* population of involvements is subject to bias through use of incorrect expansion factors in addition to possible large sampling error inherent in such a markedly skewed distribution as manifested by the descriptive statistics: a mean of 471 dollars, a variance of 3,760,963, and a median of 168. Nevertheless, it behooves us

[20] Billingsley, C. M. and Jorgenson, D. P., *ibid.*

Figure 1

Empirical and Theoretical Cost Distributions of a Single Accident Involvement
1958 Illinois Accident Cost Study



to accept the broad characteristics of this constructed distribution as indicative of the true distribution of accident involvement costs in Illinois during the year 1958.

Efforts were made to fit the Illinois data with well-known distributions of non-negative random variables such as the gamma and the lognormal without success. Among the candidates was a mixture of two exponentials:

$$w(x) = \frac{\alpha}{\mu_1} e^{-x/\mu_1} + \frac{(1-\alpha)}{\mu_2} e^{-x/\mu_2} \quad , \quad \begin{array}{l} x \geqq 0,\ \mu_1 > 0,\ \mu_2 > 0, \\ 0 \leqq \alpha \leqq 1 \end{array}$$

Using the method of moments described by Rider[21] and used by Dropkin,[22] we concluded that this distribution also seemed to be unsatisfactory. However, if we equate the sample median with the theoretical median in lieu of equating the unstable third moments, the fit appears to be quite reasonable. The derivation of this modified method of moments procedure is found in Weber.[23] Unfortunately, because of the use of a stratified sampling design, no goodness-of-fit test exists which would reveal whether or not the fit is statistically acceptable. The theoretical adaptation to the empirical distribution is shown graphically in Figure 1 and numerically in Table 13.

TABLE 13

Comparison of Empirical and Theoretical Cost Distributions
1958 Illinois Accident Cost Study.

$$W(x) = 1 - 0.9688e^{-x/231.9} - 0.03119e^{-x/7885.2}$$

x	*Empirical Cumulative*	*Theoretical Cumulative*
50	0.1783	0.1864
100	.3238	.3384
250	.6395	.6395
500	.8328	.8583
1000	.9229	.9595
2500	.9702	.9772
5000	.9874	.9835

[21] Rider, P. R., "The Method of Moments Applied to a Mixture of Two Exponential Distributions," *Annals of Mathematical Statistics 32*, pp. 143-148 (1961).

[22] Dropkin, L. B., "Loss Distributions of a Single Claim," *PCAS Mathematical Theory of Risk*, 1966.

[23] Weber, D. C., "A Stochastic Model for Automobile Accident Experience," Unpublished Ph.D. Dissertation. North Carolina State University, Raleigh, North Carolina, 1970.

Presumably, if reliable estimation procedures were available, we could improve the fit by increasing the number of exponentials in the mixture, i.e., let:

$$w(x) = \sum_{i=1}^{k} \frac{a_i}{\mu_i} e^{-x/\mu_i} \quad , \quad x \geqq 0, \mu_i > 0, a_i \geqq 0, \sum_{i=1}^{k} a_i = 1 \quad ,$$

for some finite integer k. More generally, we can assume the μ's within the population have a continuous distribution in which case $w(x)$ can be viewed as a mixture of an infinite number of exponential distributions in the same manner that the negative binomial (5) is a mixture of Poisson distributions.

As a consequence, let us assume that the distribution of the cost of an accident involvement for an *individual* driver is exponential, i.e., the probability density function of the random variable X is assumed to be:

$$g(x) = \frac{1}{\mu} e^{-x/\mu} \quad , \quad x \geqq 0, \mu > 0. \tag{14}$$

with mean μ and variance μ^2. The corresponding distribution function is:

$$G(x) = 1 - e^{-x/\mu} \quad , \quad x \geqq 0, \mu > 0. \tag{15}$$

It is well-known that the sum of n independently and identically distributed exponential variables is a gamma variable. (See Feller.[24]) Hence, in terms of our model, the probability density function of $X_1 + X_2 + \cdots + X_n$ is:

$$g_n(x) = \frac{x^{n-1}}{\mu^n (n-1)!} e^{-x/\mu} \quad , \quad x \geqq 0, \mu > 0, n = 1, 2, \ldots , \tag{16}$$

so that:

$$G_n(x) = \int_0^x g_n(s) ds \quad , \quad x \geqq 0 \quad , \quad n = 1, 2, \ldots , \tag{17}$$

$$= 1 - e^{-x/\mu} \left[1 + \frac{x/\mu}{1!} + \frac{(x/\mu)^2}{2!} + \ldots + \frac{(x/\mu)^{n-1}}{(n-1)!} \right]$$

and:

$$G_0(x) = \begin{cases} 0 & , \text{ when } x < 0, \\ 1 & , \text{ when } x \geqq 0. \end{cases}$$

[24] Feller, W., *An Introduction to Probability Theory and Its Applications Vol. II.* (John Wiley and Sons, Inc., New York, 1966).

9. AN ESTIMATION PROPOSAL

In this section we will consider the problem of estimating the parameter μ associated with a motorist as a function of his measurable characteristics. Our first inclination on this matter is to construct a model for μ as we did for λ in Section 5. Upon reflection, however, it is not possible to do so, at least with data currently available. Because of the unusually high variability in cost data, sample means have little reliability unless based upon a very large sample. In the face of this variability, the Illinois and similar cost studies are unable to give us concrete information about involvement costs by age and sex of driver, for example. This writer has, therefore, turned to the ratemaking procedures of the casualty insurance industry for a tentative answer to this problem.

As stated in the previous section, if one examines the findings of automobile accident cost studies, it soon becomes apparent that the primary determinant of cost is *location* conditioned by circumstances surrounding the accident. As a case in point, the Illinois study shows that the average involvement cost of an urban accident (one within an incorporate place) to be \$396 as compared to an average of \$931 for one taking place in a rural area. Therefore, to measure the potential cost of an involvement, as it concerns an individual, it is important for us to know where he incurs most of his accidents. Studies indicate, and current ratemaking procedures assume, that generally this is in the immediate vicinity of his residence. Hence, basic to a solution of our estimation problem is the establishment of involvement cost levels by area or territory. The area definitions need to reflect types of highways, population densities, speed limits, geographical and weather conditions, road safety conditions, etc., within a given area. In order to make use of the concept of resident area cost level, we initially assume that all drivers within a given area are characterized by the same μ, say, μ_a. Under this assumption an unbiased estimate for μ_a is given by the statistic $\bar{x}$ with variance μ_a^2/n, where $\bar{x}$ is the mean cost per accident involvement experienced by all drivers residing in the given area and n is the number of involvements upon which $\bar{x}$ is based.

Once an estimate for μ_a is obtained we should be able to assign a μ to each individual driver in that area by applying an appropriate involvement cost index, say, I based upon the characteristics of the motorist. Then the estimate for an individual's μ is given by:

$$\hat{\mu} = \bar{x} I \tag{18}$$

Members of the Casualty Actuarial Society will recognize the index factor I as a class differential. As in present automobile ratemaking procedures, the value of I applicable to a particular type of driver can be developed statistically on the basic of cost experience.

To illustrate the procedure, a New York Department of Motor Vehicles tabulation[25] classifies 477,101 accident involvements by severity class, age, sex, hour of day and day of week. The percentage distributions of severity class by age and sex found in this bulletin are given in Table 14. Using these distributions it is possible to arrive at an accident cost index by age and sex

Table 14

Distributions of Fatal Injury, Non-Fatal Injury and Property Damage Only Accident Involvements by Sex and Age Group
New York Motor Vehicle Bulletin No. 6 (64)

Males

Age Group	FI	NFI	PDO
Under 21	0.81%	52.60%	46.59%
21 - 24	0.70	56.06	43.24
25 - 29	0.64	57.22	42.14
30 - 39	0.56	57.43	42.01
40 - 49	0.48	55.89	43.63
50 - 59	0.53	54.33	45.14
Over 59	0.65	50.98	48.37
All Ages	0.60%	55.37%	44.03%

Females

Age Group	FI	NFI	PDO
Under 21	0.28%	53.59%	46.13%
21 - 24	0.30	55.56	44.14
25 - 29	0.30	58.34	41.36
30 - 39	0.27	57.32	42.41
40 - 49	0.28	54.48	45.24
50 - 59	0.33	52.40	47.27
Over 59	0.64	47.49	51.87
All Ages	0.31%	54.84%	44.85%

[25] New York Department of Motor Vehicles, State of, "Fatal, Non-Fatal, and Property Damage Accidents by Age and Sex, Hour of Day, and Day of Week," *Statistical Bulletin No. 6* (64), Albany, 1964.

if we make assumptions about the relative cost of fatal, nonfatal, and property damage only involvements. Guided by the Illinois cost data, we might use as approximate ratios 30: 6: 1. Applying these weights to the distributions in Table 14 and then converting the results to index form, we find that the patterns displayed in Table 15 emerge.

TABLE 15

Constructed Involvement Cost Indices, I

Age Groups	*Males*	*Females*
Under 21	.985	.959
21 - 24	1.021	.985
25 - 29	1.032	1.021
30 - 39	1.028	1.005
40 - 49	1.003	.970
50 - 59	.986	.947
Over 59	.953	.907
All Ages	1.005	.977

Before taking a second look at our overall model, a few comments are in order. By keeping the proper statistics on cost experience, through the pooling of data (as is done today), the casualty insurance industry could come up with acceptable estimates for μ_a and I and, in time, test assumption (14). Consideration should be given to including factors other than personal characteristics in constructing the index I, e.g. age and make of the insured's automobile. As with parameter λ, the parameter μ is not constant in time. The cost of having accidents is heavily influenced by prevailing medical, material and wage cost levels. Therefore, it will be necessary to frequently update the estimates of the area μ's, and perhaps employ a trend factor when future costs are involved.

10. THE MODEL REVISITED

We established that the distribution function of $X(t)$, the total cost of accident involvements incurred during a time interval of length t, is given by:

$$F(x,t) = \sum_{n=0}^{\infty} p(n,t)\, G_n(x) \quad , \quad x \geq 0, t > 0. \tag{1}$$

In our development, we have assumed that:

$$p(n,t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \quad , \quad \begin{matrix} n = 0, 1, 2, \ldots, \\ \lambda > 0, t > 0 \end{matrix} \tag{3}$$

and:

$$G_n(x) = 1 - e^{-x/\mu}\left[\, 1 + \frac{x/\mu}{1!} + \frac{(x/\mu)^2}{2!} + \ldots + \frac{(x/\mu)^{n-1}}{(n-1)}\,\right] \quad , \quad \begin{matrix} x \geqq 0, \\ n = 1, 2, \ldots, \end{matrix} \tag{17}$$

$$G_0(x) = \begin{cases} 0 & , \text{ when } x < 0, \\ 1 & , \text{ when } x \geqq 0. \end{cases}$$

From (2), together with (3) and (14), we obtain the mean and variance of $X(t)$ as:

$$E(X(t)) = \lambda\mu t \quad \text{and} \quad \text{Var}\,(X(t)) = 2\lambda\mu^2 t. \tag{19}$$

The probability density function of $X(t)$ is given by:

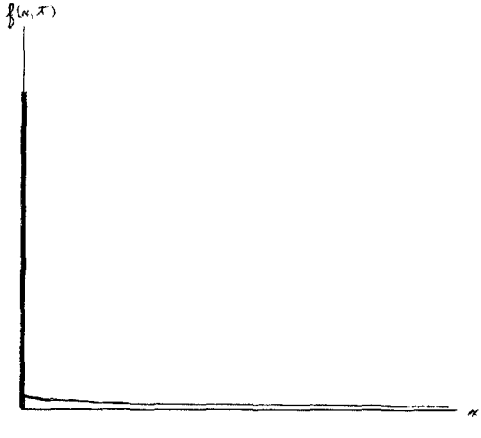
$$f(x,t) = \begin{cases} \sum_{n=1}^{\infty} p(n,t)\, g_n(x) & , \quad x > 0, t > 0, \\ p(o,t) & , \quad x = 0, t > 0, \end{cases} \tag{20}$$

where $p(n,t)$ has the form (3) and:

$$g_n(x) = \frac{x^{n-1}}{\mu^n(n-1)!}\, e^{-x/\mu} \quad , \quad \begin{matrix} x > 0, \mu > 0 \\ n = 1, 2, \ldots. \end{matrix} \tag{16}$$

Figure 2 presents a sketch of $f(x,t)$ when t is small.

Figure 2
The Probability Density Function of X(t) for Small t



The graph illustrates that the random variable $X(t)$ is neither discrete nor continuous but rather is mixed. In the case where t is small much of its mass is concentrated at the point $x = 0$ with the remainder spread over the interval $0 < x < \infty$ according to the continuous function:

$$\sum_{n=1}^{\infty} p(n,t)g_n(x) \quad , \quad x > 0.$$

The sketch is exaggerated in that the plot of this continuous function actually is nearer the x-axis than it appears in Figure 2. Again we view the extreme skewness and dispersion that plagues accident researchers. These are the attributes that make it unrealistic to predict the accident experience of an individual driver in anything other than probabilistic terms.

Under the assumption that $p(n,t)$ is a probability function for a Poisson process, the model (1) represents the distribution function of a compound Poisson process. Because of the general applicability of this particular stochastic process, it is discussed in many recent textbooks dealing with the subject of probability and stochastic processes.[26] An important property of the compound Poisson process distribution is its infinite divisibility. It follows that the sum of independent and identically distributed compound Poisson process variables is also a compound Poisson process variable. For us, this implies that the distribution of accumulated costs during one unit of time for k individuals having a common distribution function is the same as that for one of those individuals over a period of k units of time.

Let us consider the sum:

$$S_k(t) = X_1(t) + X_2(t) + \ldots + X_k(t) \tag{20}$$

for a fixed t. If each $X_i(t)$ has *d.f.* $F(x,t)$ and if the $X_i(t)$ are independent random variables, then the *d.f.* of $S_k(t)$ is $F(x,kt)$. It follows that:

$$\epsilon[S_k(t)] = k\lambda\mu t \quad \text{and} \quad \text{Var}\,[s_k(t)] = 2k\lambda\mu^2 t. \tag{21}$$

This gives us the capability of studying homogeneous groups of drivers as well as individuals. We note that the average cost per driver represented by the random variable $S_k(t)/k$ has mean and variance:

$$\epsilon[S_k(t)/k] = \lambda\mu t \quad \text{and} \quad \text{Var}\,[s_k(t)/k] = 2\lambda\mu^2 t/k. \tag{22}$$

It can be shown that, for a fixed t, the random variable:

$$\frac{S_k(t)/k - \lambda t\mu}{\sqrt{2\lambda\mu^2 t/k}} \tag{23}$$

[26] Feller, W., *op. cit.*

converges in distribution to that of a normal random variable with mean 0 and variance 1 as $k \longrightarrow \infty$. This, of course, is the central limit theorem in the context of our model.

11. ACCIDENT COST POTENTIAL

With respect to an individual driver, the occurrence of an accident involvement is, in general, an infrequent event. This, together with the variability associated with $X(t)$, means that the empirical average annual involvement costs experienced by a motorist, even if computed over a lifetime, do not adequately reflect the individual's driving skill, exposure in terms of mileage, and environmental driving conditions. Thus, we would expect average annual accident costs generated by two drivers of equal skill and identical exposure to be quite different.

O. Lundberg[27] considered the random variable:

$$Z(t) = X(t)/t$$

He found that:

$$\lim_{t \to \infty} F(zt,t) = \begin{cases} 0 & , \quad \text{when } z < \lambda\mu, \\ 1 & , \quad \text{when } z \geqq \lambda\mu. \end{cases}$$

This result may be interpreted that if we were able to observe a driver under the *same conditions* for many, many years, the distribution of his accident costs per unit of time converges in distribution to a constant, $\lambda\mu$. This "constant," which can be associated with each individual driver may be considered his *accident cost potential* and represents a theoretical cost per unit time. The use of quotation marks in the previous sentence emphasizes the point previously made, namely, that $\lambda\mu$ is not a true constant in that it is a function of the individual and his driving environment and, therefore, is subject to change in time. Although the accident cost potential associated with an individual is indicative of his expected accident costs, it does not uniquely characterize him in the sense that the product $\lambda\mu$ does not specify the individual's $F(x,t)$ uniquely. This is obvious from (21).

12. SOME EXAMPLES

In this section we will look at some probability distributions generated by our model for individual drivers and groups of homogeneous drivers. In Table 16 we find theoretical accident cost distributions related to indi-

[27] Lundberg, O., *On Random Processes and Their Application to Sickness and Health Statistics* (Almquist and Wiksells, Uppsala, 1940).

viduals over a time span of one year. The tabulated values correspond to $P\{X(1) \leq x\}$ e.g. the probability that an individual characterized by $\lambda = .08$ and $\mu = 500$ will have total accident costs less than or equal to \$500 during one year's time is 0.9706. This table demonstrates why it is so difficult to distinguish between "good" and "bad" drivers on the basis of experience over a short interval of time. For example, we expect some of the $\lambda = .04$ individuals to suffer accident loss during a year's time (about 4 percent); yet during that time about 85 percent of the individuals having an accident rate potential four times that of the first group will be cost-free.

Table 16

Evaluation of F(x,1) for Specified Values of the Parameters

$\mu = 500$

Total Costs x	$\lambda = .04$	$\lambda = .08$	$\lambda = .12$	$\lambda = .16$
0	0.9608	0.9231	0.8869	0.8521
50	.9645	.9302	.8971	.8652
100	.9678	.9366	.9063	.8771
250	.9760	.9524	.9294	.9068
500	.9853	.9706	.9559	.9413
1,000	.9945	.9888	.9282	.9767
2,500	.9997	.9993	.9990	.9986
5,000	.9999+	.9999+	.9999+	.9999+
E(X(1))	20	40	60	80
$\sqrt{\text{Var}(X(1))}$	141	200	245	283

$\lambda = .12$

Total Costs x	$\mu = 400$	$\mu = 600$	$\mu = 700$	$\mu = 800$
0	0.8869	0.8869	0.8869	0.8869
50	.8995	.8955	.8943	.8934
100	.9106	.9033	.9011	.8995
250	.9372	.9236	.9192	.9157
500	.9652	.9484	.9423	.9372
1,000	.9893	.9765	.9706	.9652
2,500	.9997	.9978	.9961	.9941
5,000	.9999+	.9999+	.9999	.9997
E(X(1))	48	72	84	96
$\sqrt{\text{Var}(X(1))}$	196	294	342	392

Table 17

Distribution Functions of $S_{100}(1)$ for Specified Values of the Parameters

$\mu = 500$

Total Costs $S_{100}(1)$	$\lambda = .04$	$\lambda = .08$	$\lambda = .12$	$\lambda = .16$
0	0.0183	0.0003	0.0000+	0.0000+
2,000	.5717	.1535	.0264	.0034
4,000	.9069	.5503	.2162	.0604
5,000	.9629	.7229	.3748	.1390
6,000	.9863	.8444	.5409	.2539
8,000	.9984	.9610	.8033	.5354
10,000	.9998	.9923	.9352	.7739
12,500	.9999+	.9992	.9880	.9323
E(S)	2,000	4,000	6,000	8,000
$\sqrt{\text{Var}(S)}$	1,414	2,000	2,449	2,828
E(S/100)	20	40	60	80
$\sqrt{\text{Var}(S/100)}$	14.1	20.0	24.5	28.3

$\lambda = .12$

Total Costs $S_{100}(1)$	$\mu = 400$	$\mu = 600$	$\mu = 700$	$\mu = 800$
0	0.0000+	0.0000+	0.0000+	0.0000+
2,000	.0538	.0147	.0090	.0060
4,000	.3748	.1295	.0815	.0538
5,000	.5803	.2407	.1581	.1070
6,000	.7503	.3748	.2589	.1813
8,000	.9352	.6425	.4944	.3748
10,000	.9880	.8337	.7070	.5903
12,500	.9990	.9500	.8790	.7844
E(S)	4,800	7,200	8,400	9,600
$\sqrt{\text{Var}(S)}$	1,960	2,939	3,429	3,919
E(S/100)	48	72	84	96
$\sqrt{\text{Var}(S/100)}$	19.6	29.4	34.3	39.2

Table 18

Distribution Functions of $S_{1000}(1)$ for Specified Values of the Parameters

Evaluated at $\begin{cases} E(S) - 3\sqrt{Var(S)} \\ E(S) - 2\sqrt{Var(S)} \\ E(S) - \sqrt{Var(S)} \\ E(S) \\ E(S) + \sqrt{Var(S)} \\ E(S) + 2\sqrt{Var(S)} \\ E(S) + 3\sqrt{Var(S)} \end{cases}$

$\lambda = .04,\ \mu = 400$	$S_{1000}(1)$	$F(x,1000)$
	5,267	0.0001
$E(S) = 16{,}000$	8,845	.0129
$\sqrt{Var(S)} = 3{,}578$	12,422	.1578
$E(S/1000) = 16$	16,000	.5223
$\sqrt{Var(S/1000)} = 3.58$	19,578	.8420
	23,155	.9690
	26,733	.9962

$\lambda = .08,\ \mu = 500$	$S_{1000}(1)$	$F(x,1000)$
	21,026	0.0003
$E(S) = 40{,}000$	27,351	.0159
$\sqrt{Var(S)} = 6{,}325$	33,675	.1582
$E(S/100) = 40$	40,000	.5158
$\sqrt{Var(S/1000)} = 6.32$	46,325	.8417
	52,649	.9712
	58,974	.9970

$\lambda = .12,\ \mu = 600$	$S_{1000}(1)$	$F(x,1000)$
	44,114	0.0004
$E(S) = 72{,}000$	53,410	.0172
$\sqrt{Var(S)} = 9{,}295$	62,705	.1584
$E(S/1000) = 72$	72,000	.5129
$\sqrt{Var(S/1000)} = 9.30$	81,295	.8416
	90,590	.9723
	99,885	.9973

One of our impressions about Table 16 might be that differences between successive distributions are trivial. Suppose we find the one year probability distributions associated with groups of 100 motorists where each individual within a group has the same λ and μ. In Table 17 we evaluate $S_{100}(1)$ for various combinations of λ and μ, remembering that the evaluation is the same as that for $X(100)$, the total accident costs acquired by an individual over a period of 100 years (assuming unchanging parameters). No longer do the differences between distributions appear inconsequential, but rather distinct differences in performance between groups of like individuals are apparent. The casualty insurance industry has recognized this, of course, through use of classification plans.

In Table 17 we observe that the standard deviation of mean accident costs, $\sqrt{Var(S/100)}$, is quite large relative to average costs, $E(S/100)$. To show how our predictions about average costs become more reliable as k is increased, in Table 18 we find distribution functions of $S_{1000}(1)$, i.e., for $k = 1000$ and $t = 1$. We also see how the distributions are approaching "normality" as indicated by the asymtotic distribution of the standardized $S_k(t)/k$ random variable displayed as (23).

13. RELEVANCY

At a time when proposals for no-fault automobile accident insurance plans have been introduced in the legislatures of New York and other states, perhaps it is time for the Casualty Actuarial Society to consider new techniques in the event of a universal change in state insurance laws. This writer has described a model which he believes is applicable in a no-fault insurance system.

DISCUSSION BY LESTER B. DROPKIN

Don Weber's paper, "A Stochastic Approach to Automobile Compensation," provides us with a most interesting approach to a subject of considerable current concern. If there were those who thought that the problem of pricing a "no-fault" automobile insurance system was still somewhat academic when the paper was presented last May, more recent events will have quickly brought the realization that the problem is now squarely in the forefront.

Whatever the case may have been at one time, today the unmodified term "no-fault" does not uniquely describe a single system. Rather, the