OPERATIONS RESEARCH CENTER

working paper

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

Interpreting Accident Statistics*

by

Joseph Ferreira, Jr.

OR 004-71

July 1971

*This paper was presented at the 39th National ORSA Meeting, May 6, 1971 in Dallas, Texas. It is based on the author's graduate research at the M.I.T. Operations Research Center, supervised by Professor Alvin W. Drake, and on his work on the staff of the U. S. Department of Transportation Auto Insurance and Compensation Study. Mr. Ferreira's graduate research was supported (in part) by the National Science Foundation under Grants GK-1685 and GK-16471 and (in part) by the M.I.T. Department of Electrical Engineering.

Interpreting Accident Statistics

Table of Contents

Abstract

Accident statistics have often been used to support the argument that an abnormally small proportion of drivers account for a large proportion of the accidents. This paper compares statistics developed from six-year data for 7,800 California drivers with results predicted using compound Poisson models for driver accident involvement that assume specific variations in accident likelihood among drivers. The results indicate that the fraction of drivers accounting for various proportions of all accident involvements is too high to suggest that "chronic" accident repeaters are involved in most accidents.

1.1 Overrepresentation in Accident Statistics

Licensing authorities, State and Federal Legislators and other officials are often advised that a small fraction of drivers account for the majority of accidents and that removing them from the road would reduce the total number of accidents dramatically. This advice is based on a popular notion concerning the role of chance and causal factors in automobile crashes that presumes all drivers are either "good" or "bad" and manifests a belief that for nearly all accidents, one can find at least one "at fault" driver whose behavior was negligent or culpable, and who was thereby instrumental in "causing" the accident. Drivers who "cause" accidents are regarded as accident prone and are thought to be a small, identifiable group that is guilty of hazardous driving, vastly overrepresented in accident statistics, and responsible for the "accident problem." The policy implications of such an attitude are straightforward. If "chance" is thought to play a small role in accident causation, and if driver accident experience is regarded as a good indicator of driver risk, then a compensation system which charges "at fault" drivers for their accident losses is considered equitable, and programs which retrain hazardous drivers and remove hazardous drivers from the road are given priority.

The National Safety Council [14] estimates that motorists in the United States account for fifteen million accidents each year and, as a result incur direct wage losses and medical expenses of 3.7 billion dollars and property damage losses of 3.8 billion dollars. However, these incurred losses are distributed most unevenly among the driving population. Accidents are rare events for individual drivers--one hundred million motorists are licensed but, during any one year, only 2.5% of them are involved in accidents in which an individual is injured.

The rarity of individual involvement in costly accidents complicates the evaluation of overrepresentation arguments. Proponents of the overrepresentation theory cite statistics such as "the 6% of all families with several accidents accounted for not less than 45% of all accidents" [17]. At first glance, such figures are startling. However, the time period included in the accident data is critical--a single week's data would surely indicate that less than 1% of all drivers accounted for all accidents recorded that week. The appropriate basis for comparison of the 6% subgroup of drivers is not the 45% figure but that percentage of accidents for which the 6% would be expected to account if all drivers were equally likely to be involved in accidents, or if specific variations in accident likelihood existed. In this paper, techniques for making such comparisons are developed and specific results are reported.

1.2 Previous Work

The mathematical tools required to make such comparisons are well known among statisticians and students of probability theory, and the problems of identifying high risk drivers has been the subject of many specialized studies. The concept of accident proneness has been investigated for half a century beginning with a study of industrial accidents by Greenwood and Woods in 1919 [9]. Probabilistic accident models have been developed and tested by actuaries and industrial researchers using data indicating the number of drivers involved in 0, 1, 2, ... accidents during each of one or more time periods [2, 13, 16]. Numerous studies have correlated individual driver characteristics and exposure data with accident experience in an effort to predict which drivers are most likely to be involved in accidents [5, 10].

The deceptiveness of the "6% account for 45%" figures has long been recognized by statisticians and students of probability theory and cursory critiques of such figures can be found in the literature. However, traffic safety analysts studying accident data have infrequently made use of the mathematical models of the actuaries and biologists and, as a result, the available mathematical techniques have not been used to make careful interpretations of such statistics that provide results in a form useful to policymakers. Also, accident data covering long time periods have not been available and ambiguities concerning the treatment of multiple-car accidents have complicated such analyses.

In this paper, a sample of six-year driver accident records is used to estimate the fraction of drivers involved in accidents during various time periods. The results are then compared with those that would be expected (a) if all drivers were equally likely to be involved in accidents and (b) if specific differences in accident likelihood among drivers existed. Much of the work reported here was carried out while the author was on the staff of the U. S. Department of Transportation Federal Auto Insurance Study. Many of the results and a brief description of the mathematical methods are also contained in a staff report which this author wrote for the study [6]. To the author's knowledge, the method presented here for examining such "6% account for 45%" figures is original.

2.1 Observations Based on ITTE Data

When estimating the fraction of drivers involved in accidents, one must be careful to distinguish between accidents and accident involvements. Approximately 75% of all auto accidents involve two or more drivers [14]. Department of Motor vehicle driver records indicate the number of accidents in which each individual driver has been involved and, hence, the same accident will often be recorded in the files of two or more individuals. Statistics such as "6% of the drivers accounted for 45% of the accidents" are usually developed from driver record samples. Though the term "45% of the accidents" is used, the data really indicate involvement and "45% of all accident involvements" would be more appropriate. In this report, the percentage of drivers accounting for various portions of all accident involvements is estimated. The sensitivity of the results to this choice will be in Section 2.6.

Since the overall average time between individual involvements in reported accidents is on the order of ten years, accident data covering as long a time period as possible is desired. However, obtaining large sample driver accident data covering a period of more than three years is quite difficult since, until recently, most Motor Vehicle Departments have purged records of information more than three years old, and insurance company records have not been kept in a manner suitable for tabulation of accident data for a controlled sample of drivers.

As a result of the author's association with the Department of Transportation's Federal Auto Insurance and Compensation Study, six-year accident data for a random sample of 7,842 California licensed drivers were made available by Dr. Albert Burg of the Institute for Transportation and Traffic Engineering (ITTE) at U. C. L. A.* This data covered the longest time period of the available data sources and was developed from records of the California Department of Motor Vehicle, widely recognized as having reliable driver accident data that are suitable for research purposes.**

The ITTE data identifies the dates of all state-reported accidents occurring between November, 1959, and February, 1968 and involving at least one of the 7,842 drivers.*** The only exception is that an accident is omitted from a driver's record if his car was stopped and he was not at the wheel at the time of the accident. A total of 3,877 accidents were recorded; the number and percentage of drivers involved in a total of 0, 1, 2, 3, . . . accidents is given in Table 1 (along with other information to be discussed shortly). The final data list developed from the data for use in this report specified the number of drivers out of the total driver population described by each possible six-digit combination (x(1), x(2), . . . , x(6)), where x(i) = the

*For a description of sampling procedures, the reader is referred to the original reports of Dr. Burg [3,4]. The existence of small biases and other minor problems in the data are also considered in Section 3.7 of another report by the author [8].

**For a more complete discussion of the accuracy and availability of driver accident data see Section 1 of Klein and Waller [12].

***State law required reporting of all accidents involving bodily injury or property damage in excess of $100. (Since the data were collected the minimum has been changed to $200).

number of accident involvements during the i^{th} year, i = 1, 2, 3, 4, 5, 6.* A list of 144 such combinations resulted.

For notational convenience, let us define D(X, T) to be the minimum percentage of ITTE drivers whose individual accident records during the first T years include X% of all ITTE accident involvements recorded during that time. For example, D(50, 6) = 10 indicates that half of all accident involvements recorded during six years were accounted for by the 10% of the drivers with the worst record during that time. The 144 accident combinations developed from the ITTE data were used to obtain D(X, T) values. The results are plotted in Figure 1 as a function of T for X = 25%, 75%, and 100%. A smooth curve has been drawn through each set of points as an aid to the reader.

The difference between D(X, T) and X is, in fact, quite large. For example, 20% or 1,537 of the 7,842 drivers accounted for all of the 1,911 accident involvements recorded during the first three years. During the entire six years, 4% of 311 of the drivers accounted for 25% of the 3,877 accidents. However, the observed values of D(X, T) depend strongly on T,

*In the form in which it was made available to the author, the ITTE data specified the date on which each accident occurred. However, the six-year period for each driver was not fixed but centered around one of two possible interview dates. An extensive amount of data analysis was required to appropriately define the number of accidents involving each driver during his six-year period. This analysis was done on two IBM 360/65 systems located at the Federal Highway Administration and at the M. I. T. Information Processing Center and is described in Section 3. 8 of another report by the author [8].

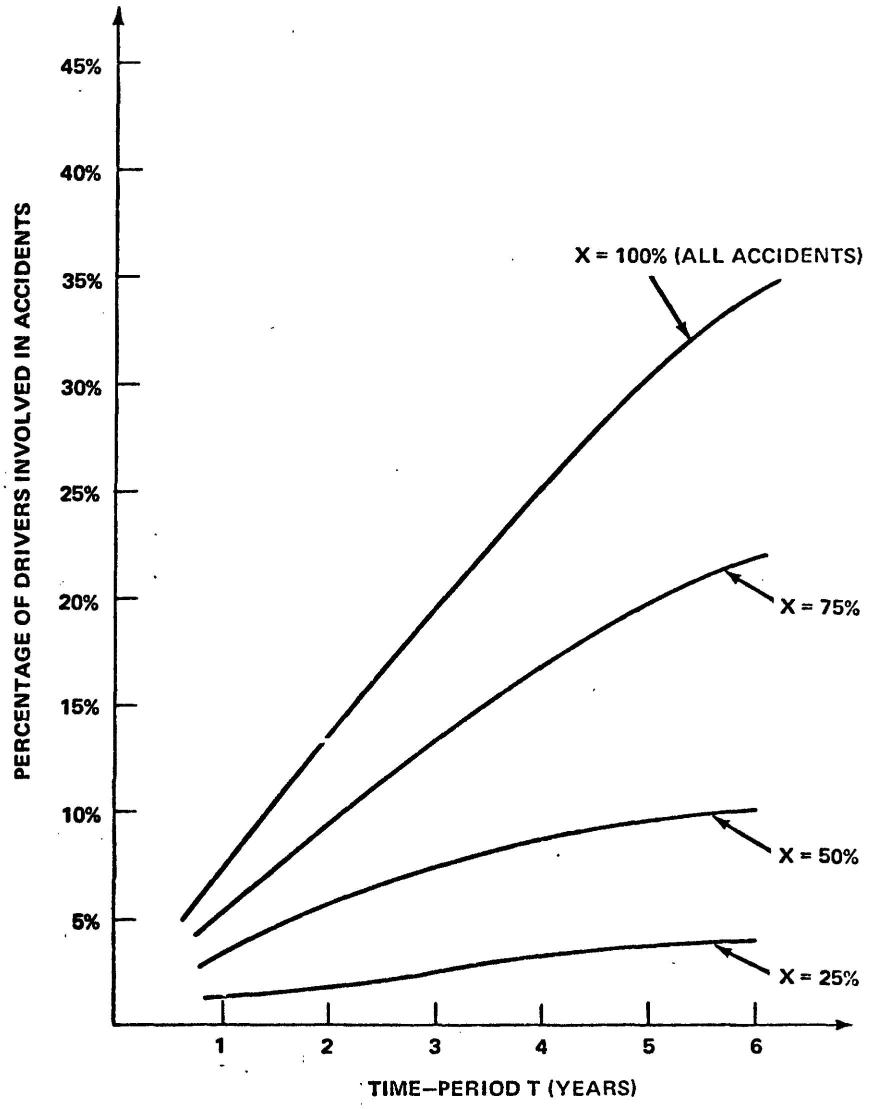


FIGURE 1 ACTUAL PERCENTAGE OF DRIVERS INVOLVED IN X% OF ALL ACCIDENTS OCCURRING DURING T YEARS (BASED ON 6-YEAR ITTE DRIVER RECORD DATA)

an effect that is expected since $T \leq 6$ although the over-all average time per driver between accident involvements was 12 years.

2.2 An "Equally Likely" Accident Model

The next step is to predict the fraction of drivers that would be expected to account for various percentages of all accident involvements if all drivers were equally likely to be involved in accidents. Making such predictions requires explicit consideration of the manner by which individual motorists would be involved in accidents over time if all drivers were "equally likely" to be involved in accidents. The simplest such model is to regard all reported accidents as equivalent, to model individual involvement in accidents as independent renewal processes and to assume that each driver is involved in $i = 0, 1, 2, \ldots$ accidents during T years according to a Poisson probability distribution

$$P_1(i/T, r) = \frac{(rT)^i e^{-rT}}{i\,!} \qquad i = 0, 1, 2, \ldots, \qquad (1)$$

characterized in terms of a single parameter r which is the same for all drivers and may be interpreted as the driver's accident likelihood.*

The Poisson distribution is well known among actuaries, biologists and industrial safety researchers. It is frequently used to model "pure chance" phenomena because of its underlying independent increments,

*The term "accident likelihood" is used rather than "accident proneness" since environmental factors as well as driver skill are included.

homogeneity, and "one event at a time" assumptions [2]. Accordingly, a driver's accident likelihood, r, is assumed unchanged from one year to the next, and the occurrence of an accident at any particular time does not affect the chances of another accident at any other time. Clearly, if a driver is involved in a serious accident he is unlikely to be driving soon after. However, the percentage of reported accidents that involve permanent or fatal injury is about 0.8% [11] and the time period of one year is long when compared with a period of several days during which a driver might be incapacitated as a result of minor injury.

One might suspect that any accident model that assigns the same accident likelihood to all drivers is an oversimplication of reality--indeed we shall find this to be the case. However, it is used in this section to provide base estimates of D(X, T) values--estimates that would result if all drivers were equally likely to be involved in accidents. The contention is that the Poisson accident model provides a plausible description of the driver accident records that would be observed under such circumstances.

2.3 Predictions Based on the Poisson Accident Model

For notational convenience, let us define $D_1(X, T, r)$ as the expected minimum percentage of drivers who would account for X% of all accident involvements reported during T years if accidents were a pure chance phenomenon and all drivers had the same accident likelihood, r. Values of $D_1(X, T, r)$ may be estimated using the Poisson accident model in terms of the right-tail cumulative and right-tail expected value functions of the Poisson distribution, which we denote as $R_1(c)$ and $G_1(c)$ respectively.

According to the Poisson model, the probability $P_1(i/T, r)$ that any individual driver is involved in i accidents during T years may also be interpreted as the fraction of all drivers expected to have a total of exactly i accidents during T years. Thus, the percentage of drivers expected to have at least k accidents each during T years is

$$100\, R_1(k) = \sum_{i=k}^{\infty} P_1(i/T, r).$$

The total number of accidents involving drivers with k or more accidents each during T years is simply the right-tail expected value

$$G_1(k) = \sum_{i=k}^{\infty} i\, P_1(i/T, r).$$

Thus we expect a percentage

$$D_1(X, T, r) = 100\, R_1(k)$$

of all drivers to account for

$$X = \frac{\sum_{i=k}^{\infty} i\, P_1(i/T, r)}{\sum_{i=o}^{\infty} i\, P_1(i/T, r)}$$

$$= \frac{100\, G_1(k)}{G_o(k)}$$

$$= \frac{100\, G_1(k)}{r\, T}$$

per cent of all accident involvements recorded during T years.

Since $R_1(k)$ and $G_1(k)$ are discrete functions for $k = 0, 1, 2, \ldots,$ a k does not always exist such that

$$R_1(k) = \frac{G_1(k)}{G_o(k)} .$$

For example X might be 50% whereas all drivers with two or more accidents account for only 45% of all accident involvements and some, but not all, of the one-accident drivers would have to be included to account for the last 5%. Mathematically, the equation used to obtain $D_1(X, T, r)$ may be adjusted for this case by setting

$$D_1(X, T, r) = 100\, R_1(c) + \frac{X - A(c)}{c - 1}\, rT, \qquad (2)$$

where

$$C = \min_{k=1,2,\ldots} k \qquad \ni \qquad X \geq A(k)$$

and

$$A(k) = \frac{100\, G_1(k)}{G_o(k)} = \frac{100}{rT}\, G_1(k) .$$

In order that $D_1(X, T, r)$ values would be comparable to the results obtained using the ITTE data, r was set equal to 0.0822, the observed average annual accident rate. Since $D_1(X, T, r)$ values were desired for a whole range of X and T values, they were calculated using a short Fortran IV G program run on the CP 67/CMS Time Sharing Facility of the M.I.T. Information Processing Center.

Figure 2 compares the actual values of D(X, T) for the ITTE data with the $D_1(X, T, r)$ values predicted using the Poisson accident model.

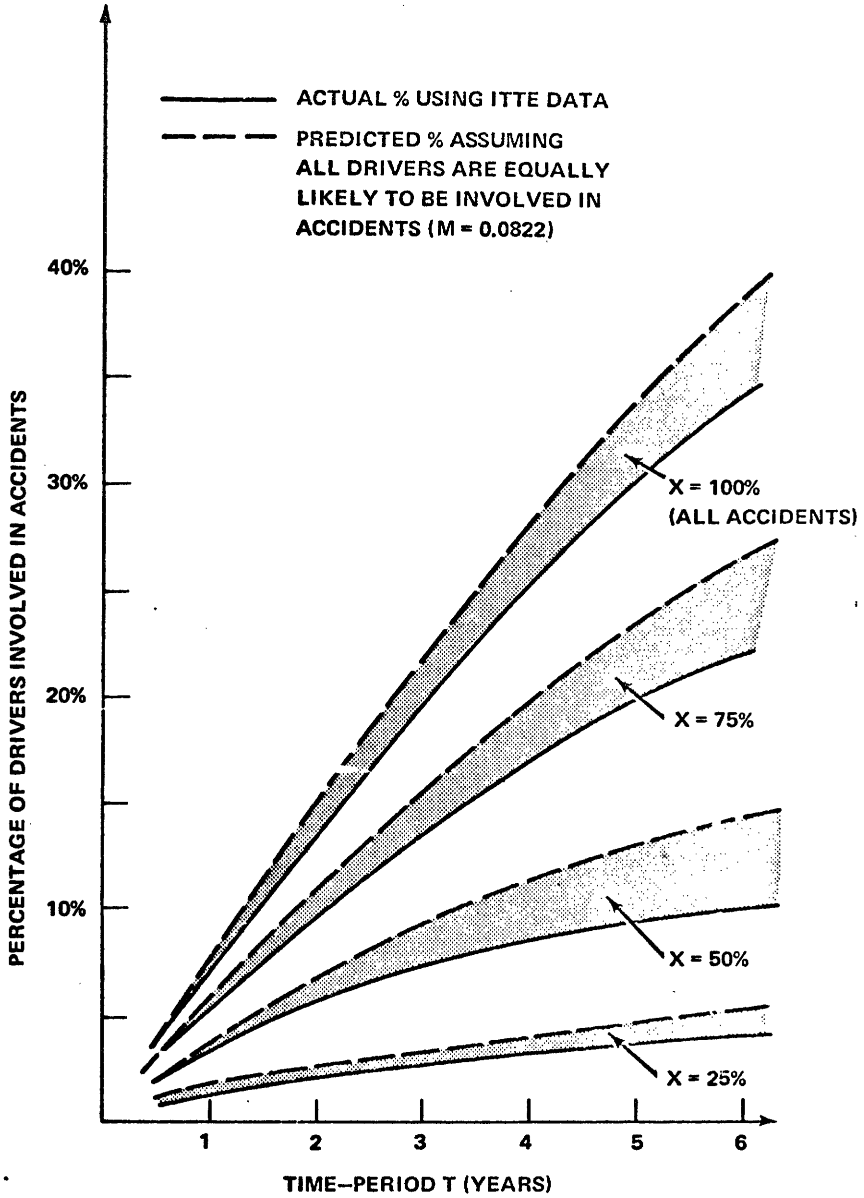


FIGURE 2 ACTUAL AND PREDICTED PERCENTAGE OF DRIVERS INVOLVED IN X% OF ALL ACCIDENTS OCCURRING DURING T YEARS

The shaded area represents the difference between the actual and expected percentage of drivers and must be explained in terms of some variation in accident likelihood among drivers. That is, the fraction of drivers accounting for, say, 75% of the accidents during a five-year period is not expected to be 75%. In fact, the 20% figure reported for the ITTE data differs by only 3.5% from the figure that would be expected if all drivers were equally likely to be involved in accidents.*

At first glance, it is natural to suspect that the equally likely assumption would result in accidents distributed evenly among all drivers. That is, one naively anticipates that, eventually, all drivers would have at least one accident, about 50% of the drivers would be needed to account for 50% of the accidents, and so on. It is obvious from Figure 2, however, that the time period that would be needed before such results would occur is much longer than 10 years.** In fact, $D_1(X, T, r)$ lags behind X even for large T in cases where $X < 100\%$. When T = 50 years, for example, $D_1(50, 50, 0.0822) = 32\%$ and $D_1(75, 50, 0.0822) = 56\%$.

*The statistical significance of these results is considered in Appendix A.

**These results are based on an average accident rate of one state-reportable accident per driver every 12 years. A change in the rate is equivalent to a change in the horizontal scale (time) of the predictions in Figure 2. The National Safety Council estimates a rate of one accident every four years when all accidents, however minor, are counted. The six-year predictions of Figure 2 would then be equivalent to two-year estimates if all accidents were considered. Thus 50% of the drivers would still account for all the accidents occurring during three years. Accidents are still not expected to be evenly distributed among drivers even if accurate data on all accidents were available for the usual one to three years--even if no difference among drivers existed.

2.4 The Negative Binomial Accident Model

The Poisson accident model assumed all drivers were equally likely to be involved in accidents and assigned to each the same value for the accident likelihood r. To allow for differences among drivers, this assumption is relaxed. Individuals are again assumed to be involved in accidents independently according to the Poisson distribution of (1), however, the particular value of the parameter r is now permitted to vary among individual drivers in the population. In particular, the value of r associated with a generic driver is regarded as a random variable the probability density function of which is a gamma-1 function,*

$$f_r(\alpha) = f(\alpha\ /k, m) = \frac{k/m}{\Gamma(k)} \left(\alpha \frac{k}{m}\right)^{k-1} e^{-\alpha \frac{k}{m}}, \qquad \alpha \geq 0, \tag{3}$$

in terms of two non-negative parameters k and m which may be fitted to sample driver accident records [1] [2].** The m parameter of the gamma-1 family may be interpreted as the overall average accident rate

*The use of the gamma-1 family for the distribution of accident likelihood is desirable since it is the natural conjugate family for Poisson sampling and facilitates the computation of values of $P_2(i/T, k, m)$ (to be defined shortly).

**Although this choice of parameters k and m to describe the gamma-1 family differs from those commonly used in the literature, it facilitates their interpretation and also their estimation since estimates of k and m are very nearly independent [1].

(0.082 accidents per year in the ITTE sample). The k parameter affects the shape of the function and results in a simple negative exponential distribution when k = 1.0.

This compound Poisson accident model was first suggested by Greenwood and Woods in 1919 [9] in a study of industrial accidents. It is sometimes referred to as a negative binomial model since the resulting probability that a randomly selected driver is involved in i accidents during T years is now a negative binomial distribution

$$P_2(i/T, k, m) = \int_0^\infty P_i(i/T, r)\ f_r(a)da$$
$$= \frac{\Gamma(k+i)}{i\,!\,\Gamma(k)} \left(\frac{k}{k+mT}\right)^k \left(\frac{mt}{k+mT}\right)^i \quad , \tag{4}$$

where i = 0, 1, 2, . . . ; k ≥ 0 and m ≥ 0.*

In Table 1, the actual number of drivers in the ITTE sample involved in a total of 0, 1, 2, 3, . . . accidents during all six years is compared with predictions based on the simple Poisson model and on the negative binomial model. The parameters are fitted using the method of moments. For the negative binomial model, this method has been shown to be the most efficient for the observed range of values for k and m [1].

Several other tests of the accuracy of the negative binomial model were made. To test the underlying Poisson model for individual accident involvement, the number of drivers, out of all those with a total of two

*The model is also referred to as the "accident proneness" model since it assumes that a driver's accident likelihood does not change from year to year and that some drivers have higher accident likelihoods than others.

Table 1: Theoretical and Actual Accident Distributions

No. of Accidents	Actual No. of Drivers[1]	Poisson Predictions[2]	Negative Binomial Predictions[3]
0	5,147	4,789	5,140
1	1,849	2,362	1,874
2	595	582	586
3	167	96	173
4	54	12	50
5	14	1	14
6+	6	0	5
	7,842	7,842	7,842
		χ^2 = 477 4df.	χ^2 = 0.99 4df.

[1] ITTE accident data for 7,842 drivers over six years.

[2] Predictions using the Poisson accident model assuming all drivers have the same accident likelihood r = 0.0822, the average accident rate observed in the ITTE sample.

[3] Predictions using the negative binomial accident model where the k and m parameters have been fitted to the ITTE data using the method of moments; k = 1.40, m = 0.0822.

accidents, whose accidents fell in particular years were totaled and compared with predictions of a multinomial distribution. Variations in the estimated k and m values were checked using subsets of the entire six-years of data. The "constant accident likelihood" assumption was tested against a "contagious" accident model in which driver's accident likelihood was assumed to depend upon his accident history. Variations in m were within the standard error but k varied between 1.1 and 1.6. Specific results of these tests have been reported elsewhere by the author [8]. They indicated:

(1) The times at which accident repeaters in the ITTE sample were involved in accidents during the six-year period were dispersed in a manner consistent with the Poisson model for individual accident experience.

(2) The compound Poisson hypothesis appeared to be a more appropriate explanation of accident experience than did a "contagious" hypothesis that assumed accident likelihood was a linear function of past accidents.

(3) Fits of the negative binomial distribution to various sample data indicated that variations in the estimates of the k and m parameters were not enough to significantly change the shape of the distribution of accident likelihood.

2.5 Predictions Based on the Negative Binomial Model

On the basis of the arguments and test results given in the previous section, we accept the negative binomial accident model as a plausible first order approximation and return to the problem of explaining the difference between the D(X, T) and D_1(X, T, r) curves of Figure 2. For notational convenience, we define D_2(X, T, k, m) to be the expected minimum percentage of drivers accounting for X% of all accident involvements during T years based on the negative binomial accident model with parameters k and m.

The method used to obtain values for D_2(X, T, k, m) is completely analogous, though computationally more difficult, to that used in calculating D_1(X, T, r) for the Poisson accident model. Once again a short Fortran IV G program was run on the CP 67/CMS Time Sharing Facility of the M.I.T. Information Processing Center to calculate D_2(X, T, k, m). To make the values of D_2(X, T, k, m) comparable to the D(X, T) results obtained from the ITTE data, the fitted values of the k and m parameters from Table 1 were used (k=1.40 and m=0.0822). In Figure 3, these values for D_2(X, T, 1.40, 0.0822) were compared with the observed results D(X, T). Note how well the predicted and actual figures

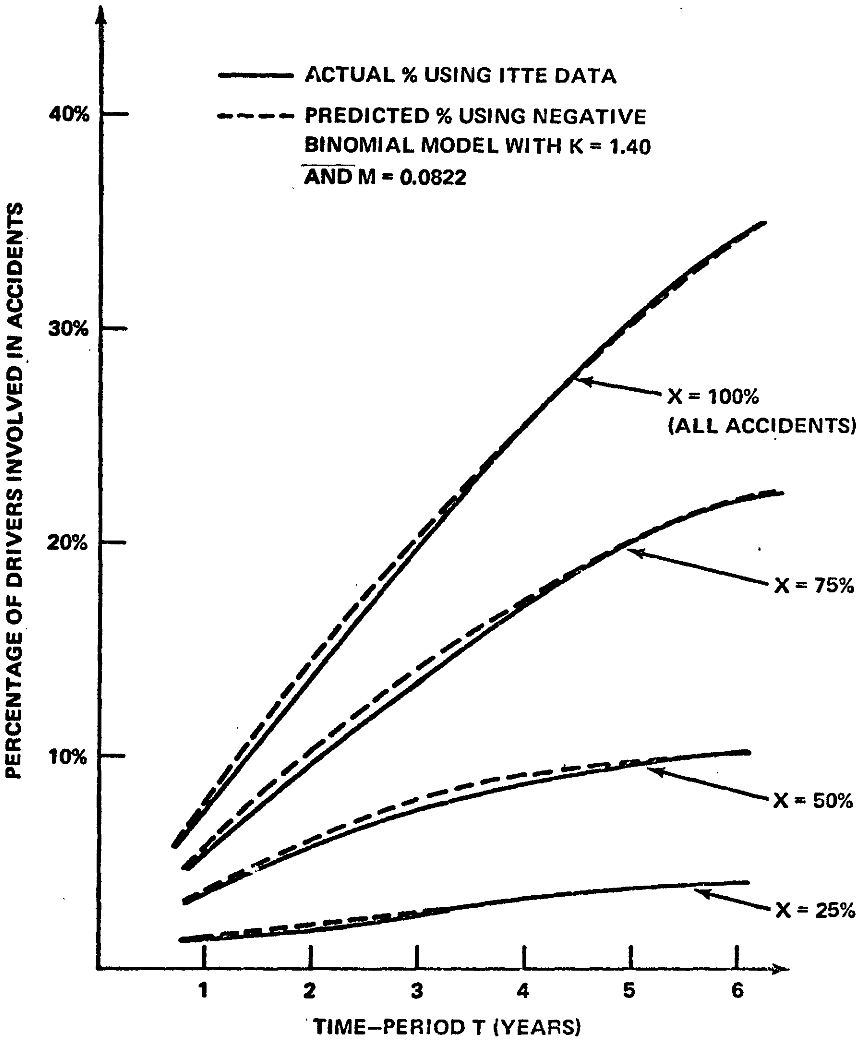


FIGURE 3 ACTUAL AND PREDICTED PERCENTAGE OF DRIVERS INVOLVED IN X% OF ALL ACCIDENTS OCCURRING DURING T YEARS

agree. The curves no longer diverge as T increases as they did in Figure 2 where the equally likely model was used. Smaller values for k result in even smaller predicted values for $D_2(X, T, k, 0.0822)$, indicating that values of k did exist such that the predicted results are substantially less than the observed values--even when m is kept equal to the observed accident rate.*

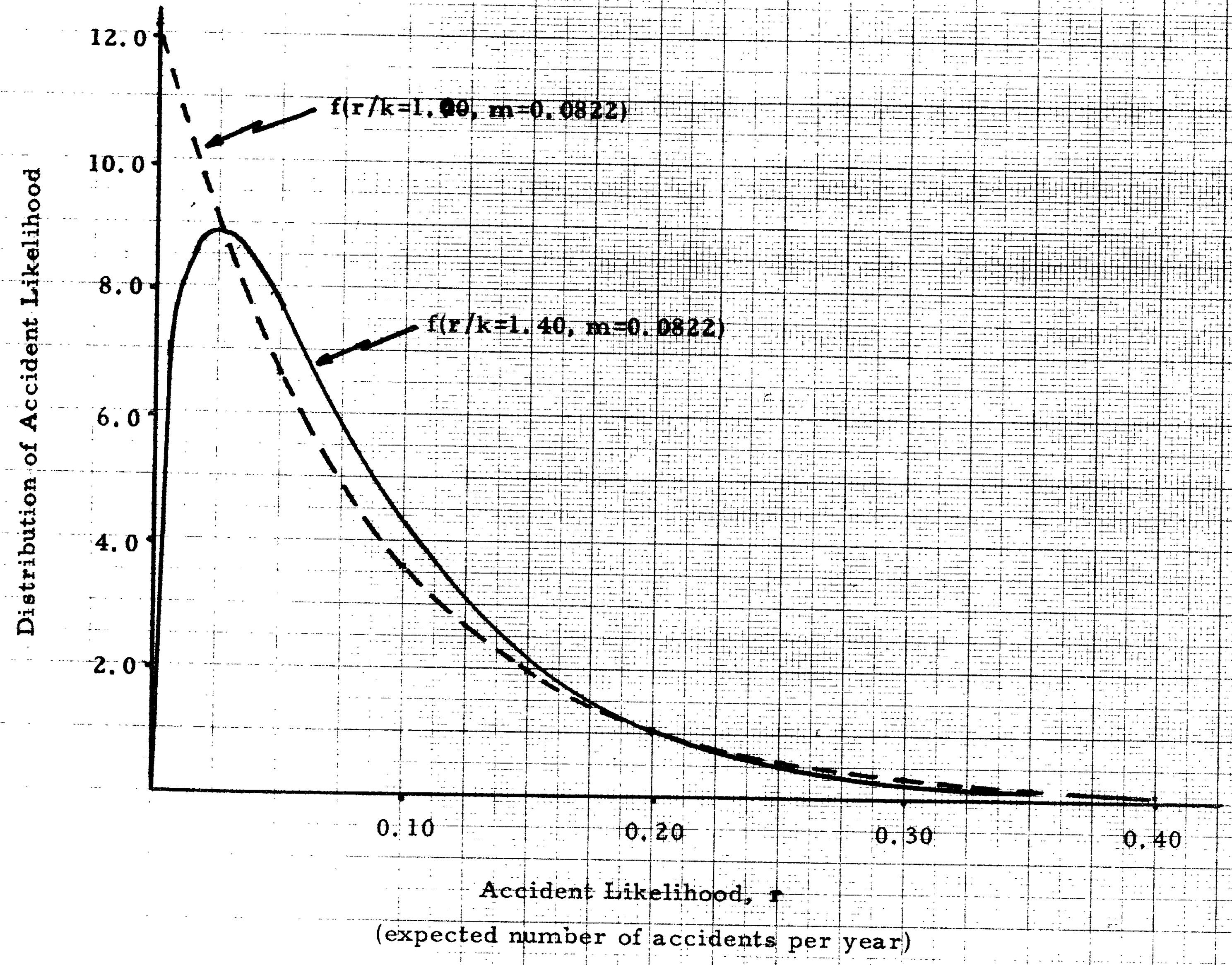
What distribution of accident likelihood among drivers is suggested by the fitted gamma-1 functions? Figure 4 graphs those distributions implied by the k and m values used to obtain the $D_2(X, T, k, m)$ predictions for the fitted k and m values and also for the pure exponential case. Substantial differences in driver accident likelihood are, in fact, predicted; however, the vast majority of drivers have quite low accident likelihoods. For the $k = 1.40$ case, 93% of all drivers are predicted to have an accident likelihood $r \leq 0.20$.* The most common value for r is about 0.025 (one reportable accident every 40 years),

The predicted accident distributions of Figure 5 may also be used to estimate the fraction of all accidents which involve drivers whose accident likelihood falls in a particular interval. Once again we use the right-tail cumulative and right-tail expected value functions, this time for the gamma-1

*The standard error of the predictions and a discussion of what constitutes a significant deviation are given in another report by the author [7].

**That is, $\int_0^{0.2} f(a/k, m)da = 0.93$.

FIGURE 4 ACCIDENT LIKELIHOOD DISTRIBUTIONS BASED ON THE NEGATIVE BINOMIAL ACCIDENT MODEL



distribution f(r/k, m). For notational convenience, we define

$$R_3(c) = \int_d^\infty f(a/k, m)da,$$

and

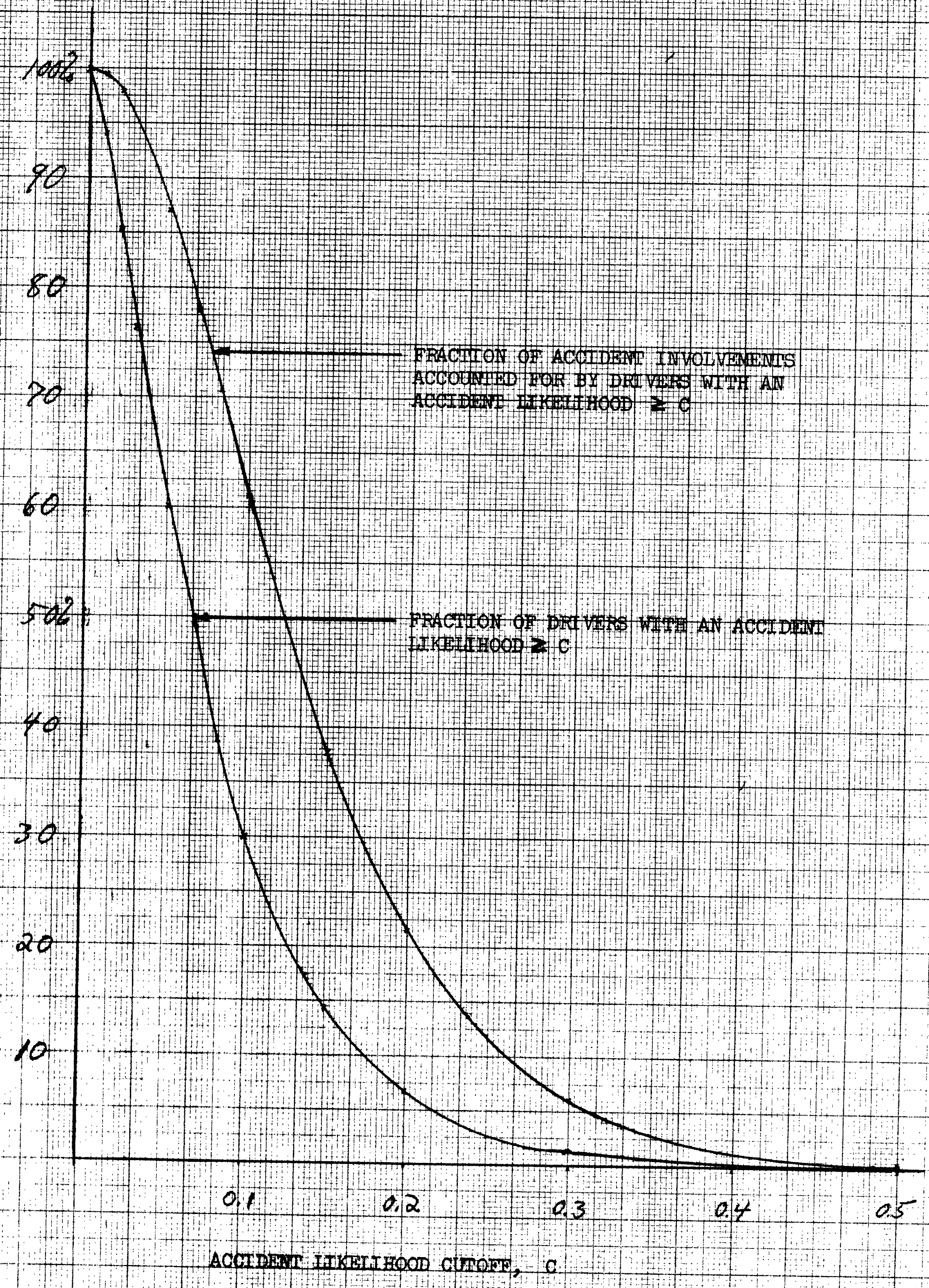
$$G_3(c) = \int_d^\infty a\, f(a/k, m)da,$$

for values $c \geq 0$.

The predicted fraction of drivers with an accident likelihood $r \geq d$ is $R_3(c)$. Similarly, $G_3(c)$ when normalized by dividing by m, the expected number of accidents per driver per unit time, is the fraction of all accident involvements that are expected to involve drivers whose accident likelihood $r \geq d$. By rewriting $G_3(c)$ as another gamma function and using Pearson's Tables [15], $G_3(c)$ may be determined for various values of d. In Figure 5, both $R_3(c)$ and $G_3(c)$ are plotted as a function of c. We see that only 0.06% of the drivers are predicted to have accident likelihoods greater than one accident every two years and these drivers are expected to account for only 0.38% of all accidents. Those 7% of the drivers with $r \geq 0.20$ (one accident every five years) are predicted to account for 22% of all accident involvements. This figure is substantial, although it is less startling than the "6% cause 45%" claim based on data from short periods. More over, an accident likelihood r = 0.20 does not fit one's image of chronic accident repeaters.

The results of Figure 5 may be interpreted as long-term, steady state estimates of the amount of overrepresentation in accidents observed in accident

Figure 5: Theoretical Overrepresentation in Accidents



data. That is, one predicts that D(20, T) will approach 7% as T increases. It is important to note that the "equally likely" model predicts that 10% of the drivers would account for 20% of all accident involvements occurring during a fifty-year period. Thus, a driver's lifetime is still sufficiently short for misrepresentation due to purely chance factors to distort the data on accident involvement by an amount similar to the "7% account for 22%" prediction.

2.6 The Sensitivity of the Results

The calculations in Appendix A suggest that the sample size of 8,000 is sufficient that the differences between D(X, T) and D_1(X, T, r) are significant. However, when T is 1 or 2 years, the distortion of the D(X, T) results, due to the short length of the time period, is so pronounced that D(X, T) is not sensitive to one's assumption about differences among drivers. In such a case, too few drivers are involved in accidents, and almost any assumption about the distribution of accident likelihood will produce figures close to D(X, T). This observation gives us some indication of what results would be obtained if one included only "at fault" accidents in the data used to obtain D(X, T).

As was noted earlier, the results in this paper are developed from driver accident data that includes all accident involvements without regard to fault. If a driver's accident likelihood were highly correlated with his chance of being "at fault" in any particular accident, then elimination of all "innocent" involvements from each driver's record would eliminate a disproportionate number of accident involvements from the records of drivers with low accident likelihoods. In such a case, elimination of "innocent" involvements would result in a smaller observed fraction of drivers accounting for a particular proportion of accidents. At the same time, however, the total number of accident involvements would drop and the time period of the observed data would represent an even smaller fraction of the average time between accidents. Hence the $D_1(X, T, r)$ results predicted using the Poisson model would also drop. The shortness of the time span of the data would account for an even larger part of the difference between D(X, T) and X. An even longer time period would have to be examined before the D(X, T) results would be a meaningful indication of an accident prone driving population.

A more desirable type of accident data for use in the report would be data that include cost considerations. Then, one could examine the fraction of drivers that accounted for various fractions of the total cost. Such data were not available. What is actually used represents a compromise insofar as the state-reported (or insurance company-reported) accidents include only those accidents that involve bodily injury or serious property damage.

One other comment regarding the scope of this analysis is in order. The report treats all reported accidents on an equal basis and does not distinguish between various types of accidents. This choice is made since it is the sweeping generalizations--a few per cent of all drivers account for the majority of all accidents--that are of primary concern. Such figures are most frequently cited and may be grossly misleading. One might suspect that focusing, for example, on only severe accidents and fatalities might produce different results. In fact, alcoholics as a group are consistently found to be involved in a substantial proportion of fatal accidents even though the group represents only a few per cent of all drivers [19]. Such facts cannot be overlooked and are of particular value to those concerned with highway accident prevention. However, as one narrows the type of accidents to be considered, the total number of yearly incidents drops. An estimated 6% of all drivers are alcoholic and approximately one-quarter of all fatal accidents involve alcoholics [18]. But, 6% of all drivers constitutes 6,000,000 drivers, whereas 25% of the fatalities represents a yearly total of about 12,000 accidents involving fatalities. The notion of the same drivers being involved in the accidents year after year remains a misconception even when subgroups of accidents are considered.*

*Of course, programs aimed at preventing driving while under the influence of alcohol remain desirable.

3.1 Conclusions

This paper examined in detail arguments used to support the popular belief that an abnormally small proportion of "high risk" drivers account for a large proportion of all accidents. Specific claims of particular interest to policymakers were studied and observations over various time periods were matched with predictions of analytical models based on explicit assumptions about differences among drivers. The results produced a different picture of the accident involvement process and indicated that:

(1) Statistics such as "6% of all drivers account for 45% of the accidents" are grossly misleading. Examination of even several years of accident data fails to provide a representative picture of the fraction of drivers involved in accidents.

(2) Chance plays a major role in determining who is involved in accidents over time periods of one or two years, and the fraction of drivers involved in accidents provides little insight into the actual distribution of accident likelihood among drivers.

(3) When time periods of more than the usual two or three years were considered, the fraction of drivers involved in accidents is in fact a reliable indicator of differences among drivers. However, the observed values were not sufficiently low to require a group of chronic accident repeaters of sufficient size to account for more than 1 or 2% of all reported accidents.

(4) Frequent, misleading references to the D(X, T) values observed over short time periods of one to three years have incorrectly reinforced the popular notion that a small group of bad drivers account for most of the "accident problem."

Appendix A: The Variance of D(X, T)

Before interpreting the differences between D(X, T) and $C_1(X, T, r)$ or $D_2(X, T, k, m)$, we must comment on the effect of sample fluctuations on the observed values of D(X, T). The size of these fluctuations, will, of course, depend upon the actual process by which individuals are involved in accidents. In this Appendix A, we obtain a rough estimate of the expected magnitude of the variations by calculating for one special case the standard deviation of D(X, T) that would result if the Poisson accident model were valid. In general, the variance of D(X, T) is difficult to determine. For the case where X = 100%, however, it is straightforward.

When X = 100%, D(X, T) is just the percentage of drivers with at least one accident during T years. If the population consists of N drivers and we denote the number of drivers involved in at least one accident during T years by the random variable v, then

$$\begin{aligned} \mathrm{Var}\ [D(100, T) &= \mathrm{Var}\ [100v/N] \\ &= \left[\frac{100}{N}\right]^2 \mathrm{Var}\ (v). \qquad \text{(A-1)} \end{aligned}$$

But v is a binomial random variable since the Poisson model implies that each driver, independent of all other, is involved in no accidents during T years with probability

$$\begin{aligned} P_0 &= P_1(0/T, r) \\ &= e^{-rT} \end{aligned}$$

from (1). Thus, v is simply the number of successes in N trials, and

$$\mathrm{Var}\ (v) = N P_0 (1 - P_0).$$

Substituting this result into (8) above, yields

$$\mathrm{Var}\ [D(100, T) = \left[\frac{100}{N}\right]^2 N P_0(1-P_0).$$

Taking the square root to obtain the standard deviation gives

$$\mathrm{s.d.}\ [D(100, T)] = 100 \sqrt{\frac{e^{-rT}(1-e^{-rT})}{N}} \ . \tag{A-2}$$

This function achieves a maximum when $P_0 = \frac{1}{2}$. For N = 7,842, then, this maximum is 0.56%. With a standard deviation on the order of 0.5%, fluctuations of several standard deviations are easily acceptable without affecting our interpretation of D(X, T) and the expectations of $D_1(X, T, r)$ and $D_2(X, T, k, m)$.

REFERENCES

1. Anscombe, F., "Sampling Theory of the Negative Binomial and Logarithmic Series Distributions," Biometrika, Vol. 37, 1950.

2. Arbous, A., and Kerrich, J., "Accident Statistics and the Concept of Accident-Proneness," Biometrics, December, 1951.

3. Burg, A., The Relationship Between Vision Test Scores and Driving Record: General Findings, Department of Engineering, U.C.L.A., June, 1967.

4. Burg, A., Vision Test Scores and Driving Record: Additional Findings, Department of Engineering, U.C.L.A., December, 1968.

5. California Department of Motor Vehicles, State of; The 1964 California Driver Record Study, Parts 1-9, Sacramento, 1964-67.

6. Ferreira, J., Jr., "Accidents and the Accident Repeater," Driver Behavior and Accident Involvement: Implications for Tort Liability, U. S. Department of Transportation, GPO-1970.

7. Ferreira, J., Jr., "Analytical Aspects of Driver Licensing and Insurance," Ph.D. Thesis, M.I.T., September, 1971.

8. Ferreira, J., Jr., Quantitative Models for Automobile Accidents and Insurance, Report to U. S. Department of Transportation, GPO-1970.

9. Greenwood, M. and Woods, H. M., "The Incidence of Industrial Accidents with Specific Reference to Multiple Accidents," Industrial Fatigue Research Board Report, No. 4, 1919.

10. Hakkinen, S., Traffic Accidents and Driver Characteristics, Finland's Institute of Technology, Scientific Researchers, No. 13, Helsinki, 1958.

11. Harwayne, F., The Relative Cost of Basic Protection Insurance in New York State--An Objective Determination, 1968, p. 40.

12. Klein, D., and Waller, J., Causation, Culpability, and Deterrence in Highway Crashes, Department of Transportation, 1970.

13. Lundberg, O., On Random Processes and Their Application to Sickness and Accident Statistics, Almquist and Wiksell, Uppsala, 1940.

14. National Safety Council, Accident Facts, Chicago, 1969.

15. Pearson, Karl (Editor), Tables of the Incomplete Γ-Function University Press, Cambridge, 1965.

16. Simon L., "An Introduction to the Negative Binomial Distribution and its Applications," Proceedings of the Casualty Actuarial Society, Vol. XLIX; Part 1, No. 91, 1962.

17. Survey Research Center, Public Attitudes Toward Auto Insurance, U. S. Department of Transportation, GPO-1970.

18. U. S. Department of Health, Education and Welfare, Report of the Secretary's Advisory Committee on Traffic Safety, GPO-1968.

19. U. S. Department of Transportation, Alcohol and Highway Safety, GPO-1968.