A STOCHASTIC MODEL
FOR AUTOMOBILE ACCIDENT EXPERIENCE

by

Donald Chester Weber

Institute of Statistics
Mimeograph Series No. 651
January, 1970

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

1. INTRODUCTION

1.1 The Problem

The National Safety Council's 1969 edition of Accident Facts gives the following estimates on motor vehicle accidents occurring in 1968:

Deaths	55,200
Disabling injuries	2,000,000
Costs	$11,300,000,000
Drivers involved	26,000,000

These figures bear out the fact that accidents resulting from the use of automobiles are among the major health hazards confronting the American people today.

In this dissertation we shall consider the problem of constructing a mathematical model for automobile accident experience applicable to an individual and to groups of individuals. The model will feature two aspects of accidents--incidence and severity. The incidence component will be measured by an accident involvement; the component of severity will be measured by "direct costs" associated with an involvement. After we have derived the functional form of our model, we will consider procedures for estimating its parameters so that different parameter values reflect differences between individual drivers and between homogeneous groups of drivers. It is hoped that the model developed will provide additional insights into the accident phenomenon as it relates to measurable characteristics of drivers. In particular, the model will be used to investigate the possibility of predicting future accident experience and therefore should be of interest to driver-licensing authorities.

Because of the cost component, this study would seem to have overtones for the casualty insurance industry. Any possible impact, however, is diminished by differences in the counting of incidences and the valuation of costs. To conform with the tort liability systems in effect in all states, the insurance industry counts claims rather than involvements and it values costs on the basis of liability rather than actual money value of damages or losses. As a consequence, when an insured is involved in an accident, this may mean to his insuror no claim, one claim or several claims depending upon degree of fault and upon the number of persons involved. A further departure from insurance methods is that in this study we will be focusing our attention on the driver of the automobile whereas insurance coverage follows the vehicle.

Currently discussion in this country revolves about reforming the present automobile insurance system to one not primarily predicated on fault. One such proposal is the Keeton-O'Connell (1965) plan which basically would require the driver's insurance company to compensate victims of an accident, including the motorist himself, regardless of fault. It is conceivable that the accident model developed in this thesis could provide the theoretical framework for some future no fault insurance system.

1.2 The Data

In order to ensure that the model will be relevant to the real world, its essential features must harmonize with the properties of pertinent and reliable accident data. Therefore the empirical data

from which inferences are drawn and which support the model conclusions must be considered an integral part of this work. In fact, in some instances the data sources dictate the definitions that are adopted for reasons of compatibility. The analysis relating to frequency of involvement will be based upon data generated from the driver record files of the California Department of Motor Vehicles for its 1964 California Driver Record Study. The analysis relating to the cost of an involvement will be largely based upon data collected for the study entitled Cost of Motor Vehicle Accidents to Illinois Motorists, 1958 (Illinois Department of Public Works and Buildings, State of, 1962).

The California study consists of nine parts. Each part is a report which analyzes a particular phase of accident involvement as it relates to the attributes and driving record of approximately 225,000 individuals selected at random from the population of licensed California drivers. This sample constituted about 2 percent of the drivers residing in the state. Of these, data are available on about 148,000 individuals over the full observation period of three years, namely, 1961-63. These data include information on such items as county, sex, marital status, age, accident involvement and conviction counts, all of which will play a role in this dissertation. It is important to note that the accident count in the California study includes only accidents reported to the Department of Motor Vehicles during the years in question. According to Part 2 of the report (California Department of Motor Vehicles, State of, 1964-67, Part 2, pp. 3-4),

> this basically includes all fatal and injury accidents, all accidents investigated by or reported to the California Highway Patrol, either by individuals or other enforcement agencies, and all property damage accidents reported in compliance with California's Financial Responsibility Law (those in excess of $100.00 damage).

Therefore, to conform with the data, throughout this work an accident involvement is taken to mean a reported involvement. To get an idea of the proportion of the accidents that are of the "reportable" type, findings in the Illinois Cost Study showed that about 75 percent of the total number of accident involvements were attributed to unreported incidents. In the State of Illinois statutes, the definition of a reportable accident is essentially the same as in California.

The Illinois Motor Vehicle Cost Study was completed by the Illinois Division of Highways 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 percent relative error. In terms of a stratum mean $\bar{x}$ and its standard deviation $s_{\bar{x}}$, this implies

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

The 2,878 cases of interest to us were comprised of 332 fatal injury, 1,730 nonfatal injury and 816 property damage only cases. Appropriate expansion factors were then applied to each case in order to obtain a "population" of 317,051 involvements.

In the Illinois study, direct costs are defined as (Illinois Department of Public Works and Buildings, State of, 1962, p. 133)

> 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.

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. In the Illinois study, 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 valuating 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, were obtained.

Occasionally reference will be made to a more recent cost study, The Washington Area Motor Vehicle Accident Cost Study (Smith and Associates, 1966), completed in 1966. Since the scope of this study was limited to a metropolitan area, it was felt that the Washington Area data is not as appropriate to this investigation as the state-wide Illinois data. These two studies are not exactly comparable since funeral expenses and loss of future earnings were considered elements of direct costs in the more recent study. Inclusion of these two elements increases the skewness of the resulting cost distributions. Currently, a similar cost study conducted in Ohio is in the final stages of completion but it has not yet been made available to this writer.

2. LITERATURE REVIEW

2.1 Composite Models

Because this study is concerned with two elements of accident experience--incidence and cost--literature pertaining to this research appears in three areas, namely, (i) composite models which recognize both elements, (ii) accident frequency distributions, and (iii) accident cost distributions. For that reason the literature in each of these areas will be reviewed separately. To begin, two composite models are cited, one deterministic and the other probabilistic. The first model features predictive factors which can be used to arrive at expected accident costs according to varying conditions. The second is a formulation upon which the model of this thesis is patterned.

For the purpose of ratemaking, the casualty insurance industry uses a procedure that combines frequency and cost into a single unit known as pure premium. It is defined as the average incurred loss cost per unit of earned exposure (Stern, 1965). In essence the model for automobile liability basic limits pure premium is given by the formula

$$Y_{ijk} = BT_iC_jM_k \ .$$

Y_{ijk} is the pure premium for a car insured in territory i, driver class j and merit rating class k. T_i, C_j and M_k are territory, driver class and merit rating class differential factors applied to B, the pure premium for a car garaged in the base territory and driven by individuals assigned to the base driver class. That is,

$$B = \frac{\text{loss experience for cars in } T_B \text{ and } C_B}{\text{earned car-years in } T_B \text{ and } C_B} \ .$$

This deterministic model differs in many respects from the stochastic model which is the subject of this study. As already noted in Chapter 1, these include accident experience by vehicle rather than by driver and costs based on legal liability instead of involvement.

The model used in this thesis is a variation of the collective risk theory model which has interested Scandinavian actuaries since the early part of this century. An account of this theory and its results is given by Cramér (1954, 1955). This theory is concerned with the business of an insurance company regarded as a game of chance between the company and its collection of policyholders. Using the notation of the next chapter, the basic collective risk theory model is given by the expression

$$F(x,t) = \sum_{n=o}^{\infty} p(n,t)\, G_n(x) \quad .$$

Here $p(n,t)$ is defined as the probability that n claims have occurred when the expected number of claims is t and $G_n(x)$ is the n-fold convolution of the distribution function of the amount of a single claim. Hence $F(x,t)$ is the distribution function of a random variable that represents the total amount of claims paid by the company up to the moment when t is the expected number of claims. It should be noted that the parameter t represents operational time rather than real time since the number of policyholders of an insurance company is continually changing.

Bohman and Esscher (1963) report the results of research done by a committee interested in collective risk theory. The object of their

investigation was to explore approximation procedures in evaluating two functions of interest, one of which is $F(x,t)$. The necessity to consider approximation methods becomes apparent when it is observed that $G_n(x)$ is an n-fold convolution which makes the mathematical calculation of $F(x,t)$ by direct methods overwhelmingly complex and often impossible. Information on the precision of various methods was obtained for the case when $p(n,t)$ is either a Poisson or a negative binomial probability function. These two functions occupy much space in accident research literature as will be noted in the next section.

2.2 Accident Frequency Models

Dating back to the mathematical models of Greenwood and Yule (1920), the topic of accident distributions has frequently appeared in research literature. No effort will be made here to review all articles on the subject; rather, it will be sufficient to cite certain works that when taken together provide the reader with an overall view of what has been done in the area of accident frequency modeling.

Variations of two of the three models considered by Greenwood and Yule are being utilized in this thesis. The first is the Poisson model

$$p(n) = \frac{e^{-\lambda}\lambda^n}{n!}, \qquad n = 0, 1, 2, \ldots, \quad \lambda > 0 .$$

If N is a random variable that represents the number of accidents sustained by an individual, $p(n)$ can be interpreted as the probability

that he will be involved in n accidents when the expected number is λ. The second relevant model is the negative binomial model derived as a compound Poisson distribution. If we assume that λ is a value of a random variable having a gamma distribution

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

then the negative binomial model is given by the expression

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

$$= \frac{\Gamma(n+r)}{n!\Gamma(r)} \left(\frac{c}{c+1}\right)^r \left(\frac{1}{c+1}\right)^n, \qquad n = 0, 1, 2, \ldots,\ c > 0,\ r > 0.$$

Greenwood and Yule (1920) and later Newbold (1927) interpreted the distribution $u(\lambda)$ to represent the different degrees of "accident liabilities" among individuals, i.e., they assumed individuals differ in their likelihood of having accidents. This premise subsequently gave rise to the concept of "accident proneness" which is ably discussed in a critical review by Arbous and Kerrich (1951). Basically, this notion assumes that each individual has an inherent personal quality that influences his susceptibility toward accidents and that this accident propensity remains constant in time.

Arbous and Kerrich point out that a good fit of accident data by a negative binomial distribution does not prove the hypothesis of accident proneness. First of all, the probability of having an accident depends on other factors besides inherent proneness. These include personal qualities such as age, health and experience, and

environmental factors. Secondly, the negative binomial can arise in a number of ways. Kerrick, in his part of the paper, derives the negative binomial as a "contagious" distribution by assuming that the initial susceptibility of individuals to an accident is the same, but that for each individual each accident increases the probability of another accident. Thus the diametrically opposed hypotheses of constant individual liability and varying liability lead to the same univariate model. This was recognized earlier by Lundberg (1940) and discussed by Feller (1943) and (1949). Kerrich then goes on to show that the concepts of proneness and contagion are distinguishable in the bivariate case. He constructs a bivariate compound Poisson model and a bivariate contagious model, both of which have negative binomial marginal distributions, and "fits" them to data that has been divided into two consecutive time periods. Use will be made of the bivariate compound Poisson distribution in Chapter 4 so that the essential features of Kerrick's derivation are given in Appendix 9.1.

To parallel Greenwood and Yule's mathematically cumbersome "biased" distribution, Kerrich constructs another contagious distribution which he calls the "burnt fingers" distribution. It assumes that until an individual has had his first accident, the probability that he will have an accident is constant, but as soon as he experiences that first accident this probability decreases and thereafter remains constant no matter how many accidents he subsequently suffers.

Bates and Neyman (1952) generalize the bivariate Poisson model of Kerrich by assuming that the Poisson parameters of two time periods are λ_1 and λ_2, respectively, such that $\lambda_2 = k\lambda_1$, where k is a constant.

In this framework Kerrich's bivariate negative binomial is the special case, k = 1. As a further generalization, Bates and Neyman visualize that λ_1 and λ_2 characterize the proneness of an individual to two kinds of accidents, say light and severe accidents, so that the number of light accidents in one period can be used to predict the number of severe accidents in a subsequent period. As an extension of these notions, they go on to derive a multivariate negative binomial distribution of s dimensions where the random variable X_i represents the number of accidents of the i^{th} kind.

In Appendix 9.1, it is seen that in his derivation of the bivariate compound Poisson distribution, Kerrich assumes that the Poisson distributions for the two non-overlapping periods of time are independent; i.e., the number of accidents experienced by an individual in the two intervals are uncorrelated. Edwards and Gurland (1961), in their extended proneness model, assume a correlated bivariate Poisson distribution which implies a constant non-negative correlation between the occurrence of accidents involving an individual in one time period and the occurrence of accidents involving that same individual in another period. Their derived distribution is referred to as a compound correlated bivariate Poisson.

In their book, Cresswell and Froggatt (1963) challenge the concept of accident proneness. As opposed to the "proneness" hypothesis, they present a "Long" distribution and a "Short" distribution constructed on the basis of "random" hypotheses. The Long distribution is predicated on the premise that every driver is liable to a "spell" and that no accident can occur outside a spell. (During a spell his

driving performance is sub-standard.) The number of spells per driver is assumed to be a Poisson variable with parameter λ, the same for all drivers, and the number of accidents per spell is another Poisson variable with parameter θ, also the same for all drivers. The Short distribution differs from the Long in that it admits the possibility of accidents outside of spell periods. These accidents are assumed to have a Poisson distribution with parameter ϕ. Cresswell and Froggatt's Long distribution is a Neyman Type A distribution derived by Neyman (1939), and their Short distribution is the convolution of a Neyman Type A with a simple Poisson. Like Arbous and Kerrich (1951), Cresswell and Froggatt (1963) include throughout their work an extensive historical and literature review of accident frequency research and findings up to the time of publication.

An alternative approach to accident frequency distributions is the distribution of time intervals between accidents. Shaw and Sichel (1961) and Sichel (1965) make use of the fact that the interarrival times of a Poisson process are independent identically distributed random variables obeying an exponential probability law. (See Parzen (1962).) Therefore, if a person's accident frequency follows a Poisson distribution with parameter ν, then the time interval between successive accidents is given by the exponential probability density

$$h(t) = \nu e^{-\nu t}, \qquad t \geq 0, \quad \nu > 0 .$$

In his development of this model, Sichel gives the statistical estimate of ν, a confidence interval for ν, and tests to establish the credibility of the model.

Dropkin (1959) excited the casualty insurance industry when he demonstrated that a negative binomial provided a close fit to the accident frequency distribution of a sample of 94,935 California drivers over the three years ending with 1958. The ability to withstand large sample scrutiny gave credence to the appropriateness of the negative binomial as an accident model.

In conclusion it should be mentioned that much of the impetus for accident research during the last half century has been supplied by psychologists in their efforts to identify the personal factors which, when taken together, constitute an individual's "proneness toward accidents". The literature abounds with studies performed to measure such qualities as intelligence, mechanical aptitude, psychomotor abilities, emotional stability, social adjustment, personality, attitudes, health, etc., in relation to accident involvement. A relatively recent example of such a study is a doctoral dissertation by Häkkinen (1958), which includes a survey of the previous studies in this area together with an extensive bibliography.

2.3 Accident Cost Models

The only proposed theoretical model for the total cost of a motor vehicle accident involvement found by this writer in his search through the literature appears in a book by Leimkuhler (1963). In his analysis of the consequences of truck accidents, on the basis of a random sample of 200 accidents involving tractor-semitrailers, he concluded that vehicle damage sustained in accidents involving one or more trucks follows a lognormal distribution, i.e., the logarithm of

vehicle damage is normally distributed. In an effort to substantiate his theory, Leimkuhler found that the total direct costs for automobile reported and unreported accidents derived in a Massachusetts study in 1953 seemed to follow a lognormal distribution. From this he hypothesized the relationship between total accident cost (C) and vehicle damage (D) to be

$$\log C = \log a + b \log D$$

or

$$C = aD^b$$

where a and b are constants and log D and consequently log C are normally distributed. The above relationship gave him a method for estimating total accident costs from the amount of vehicle damage as recorded in accident reports.

3. THE MODEL

3.1 Fundamental Assumptions

Basic to this study is the notion that an automobile accident is a "chance" event but that individual drivers differ in their susceptibility to being involved in an accident, and when involved, differ in the probable amount of damage suffered. We shall therefore assume that the probability of accident occurrence and the probability of accident costs are related to certain characteristics of the driver. In this context, characteristics is used in the broad sense to include not only personal traits but also exposure in terms of annual mileage driven, driving experience, the roads used, the traffic density encountered, weather and light conditions while driving, the age and horsepower of the vehicle, the number of passengers, and all other factors which might determine to some degree the likelihood of accident involvement and severity. Personal qualities include such variables as physical characteristics and health, mental and emotional state, personality, attitudes, social adjustment, habits (e.g., alcohol, drugs), motor abilities, mental alertness and intelligence. These accident determinants are not found in public records; in fact, many of these characteristics can only be measured imperfectly through psychological tests and questionnaires, if at all. Even the clearly measurable characteristic, annual mileage, becomes nebulous when several persons in a family drive the same car. We are therefore confronted with the task of trying to predict an individual's accident experience on the basis of information found in Motor Vehicle

Department records, information which at best reflects to an unknown degree the actual accident determinants of the driver. Let us look at some of this recorded information and see how it can be used as a vehicle for accident prediction.

Conceivably, age reflects a person's sense of responsibility, attitudes (e.g., "daring of youth" vs "conservatism of old age"), alertness, motor ability, amount of driving experience, time during day and week of driving and driving mileage. Marital status supposedly reflects sense of responsibility, attitudes, social adjustment, number of passengers and driving mileage. Geographical area reflects road conditions, types of highways, traffic density, weather conditions and speed limits. Violation history reflects attitudes, mental and emotional states and driving mileage. Accident history presumably reflects driving skill, alertness, traffic density and driving mileage.

From the nature of these predictor variables it is evident not only that accident likelihood differs from person to person but also accident likelihood changes with time for the same person. That is, we expect an individual's driving performance and exposure at age 60 to differ from what it was at age 16. Again, if a person moves from a rural village to a large city we expect a change in his probability of accident occurrence and in the probable cost if and when involved in an accident.

The foregoing discussion can be summarized in three fundamental assumptions upon which this work is based. They are:

(3.1)

(a) Accidents occur by chance and therefore are subject to probability laws.

(b) The probability distributions characterizing individuals have a specific form but differ from driver to driver in parameter values.

(c) The parameter values associated with an individual driver are functions of certain factors which (in themselves) are dependent on time.

The implications of these fundamental assumptions are: (a) a stochastic model is appropriate, (b) population and individual probability distributions differ in form, and (c) estimates of parameter values require frequent updating because the parameters are time dependent.

3.2 Derivation of the Model

Following Cramer (1955), suppose the random variable N(t) represents the number of accident involvements during a time interval of length $t > 0$ and $p(n,t)$ denotes the probability that $N(t) = n$. Let the cost of an accident involvement be represented by the non-negative random variable X having distribution function G(x). We shall assume that X is independent of time and of the costs of prior involvements. Consequently, the X_i are independently and identically distributed random variables, each having the d.f. G(x). Accordingly, the distribution function of the sum $X_1 + X_2 + \cdots + X_n$ is the n-fold convolution of G(x) with itself, which we shall denote by $G_n(x)$. $G_n(x)$, then, is the d.f. of the conditional total accident cost under the hypothesis that n involvements have occurred.

Finally, let the random variable X(t) having distribution function F(x,t) represent the unconditional cost of accident involvements during a time interval of length t. Since the events that exactly 0, 1, 2, ..., accident involvements have occurred are mutually exclusive, it follows that

$$
\begin{aligned}
F(x,t) &= \Pr\{X(t) \leq x\} \\
&= \sum_{n=o}^{\infty} p(n,t) G_n(x) , \qquad x \geq 0, \quad t > 0 ,
\end{aligned}
\tag{3.2}
$$

where

$$
G_n(x) = \int_0^{\infty} G_{n-1}(x-y)\, dG(y) , \qquad n > 1 ,
$$

and

$$
G_o(x) = \begin{cases} 0 , & \text{when } x < 0 , \\ 1 , & \text{when } x \geq 0 , \end{cases}
$$

$$
G_1(x) = G(x) .
$$

What may be a typical sample function of the X(t) process is graphically represented in Figure 3.1. It illustrates the course of the total cost of accidents, X(t), as a function of time. A jump occurs when the individual is involved in an accident, the frequency of which is determined by p(n,t) while the size of the jump is governed by G(x).

To obtain the moments of X(t), let f(h) and g(h) be the characteristic functions corresponding to F(x,t) and G(x), respectively.

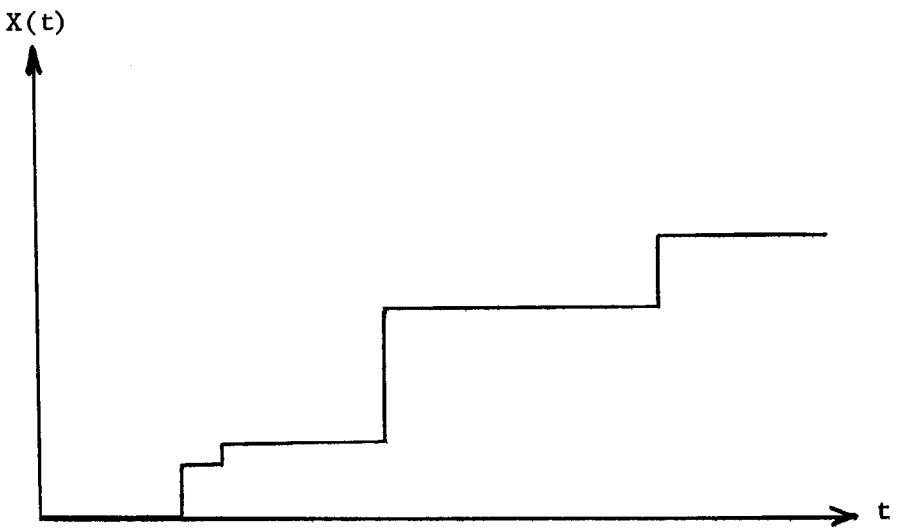

Figure 3.1 The model as a stochastic process

Then

$$f(h) = E(e^{ihX(t)}) = \int_{-\infty}^{\infty} e^{ihx} dF(x,t)$$

$$= \sum_{n=o}^{\infty} p(n,t) \int_{-\infty}^{\infty} e^{ihx} dG_n(x) \qquad (3.3)$$

$$= \sum_{n=o}^{\infty} p(n,t) [g(h)]^n$$

where

$$g(h) = \int_{-\infty}^{\infty} e^{ihx} dG(x) \ .$$

Differentiating with respect to h we obtain

$$f'(h) = \sum_{n=o}^{\infty} np(n,t)g'(h)[g(h)]^{n-1}$$

$$f''(h) = \sum_{n=o}^{\infty} np(n,t)g''(h)[g(h)]^{n-1}$$

$$+ \sum_{n=o}^{\infty} n(n-1)p(n,t)[g'(h)]^2[g(h)]^{n-2} .$$

Setting $h = 0$, noting that $g(0) = 1$, that the k^{th} moments of $X(t)$ and X are given by $f^{(k)}(0)/i^k$ and $g^{(k)}(0)/i^k$, respectively, and that the k^{th} moment of $N(t)$ is defined as $\sum_{n=o}^{\infty} n^k p(n,t)$, it follows that

$$E(X(t)) = E(N(t))E(X)$$

(3.4)

$$E(X^2(t)) = E(N(t))E(X^2) + E^2(X)[E(N^2(t)) - E(N(t))] .$$

From (3.4) we immediately obtain

(3.5) $$Var(X(t)) = E^2(X)Var(N(t)) + E(N(t))Var(X) .$$

3.3 Assumptions Relative to p(n,t)

Recall that $p(n,t)$ is the probability that n accident involvements have occurred during a time interval of length t. Looking at $p(n,t)$ from the standpoint of an individual driver, we shall make the following basic assumptions:

(i) The probability of an accident involvement during a small time interval from t to $t + \Delta t$ does not depend on t and is proportional to Δt, or

$$p(1, \Delta t) = \lambda \cdot \Delta t + o(\Delta t) , \qquad \lambda > 0 ,$$

where

$$\lim_{\Delta t \to o} \frac{o(\Delta t)}{\Delta t} = 0 .$$

(ii) The probability of more than one accident involvement during such a time interval is negligible, i.e.,

$$\sum_{n=2}^{\infty} p(n, \Delta t) = o(\Delta t) .$$

(iii) The numbers of accidents in nonoverlapping intervals are stochastically independent.

The above assumptions are immediately recognized as the postulates of the Poisson process and described in many textbooks, e.g., Feller (1968). The fulfillment of these conditions leads to a system of differential equations having the solution

$$(3.6) \qquad p(n,t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} , \qquad n = 0, 1, 2, \ldots ,$$
$$\lambda > 0, \; t > 0 .$$

We observe that in this case

$$(3.7) \qquad E(N(t)) = \lambda t \quad \text{and} \quad Var(N(t)) = \lambda t .$$

In reference to the first section of this chapter, we shall assume that each individual driver is characterized by his own particular λ which is a function of the individual's physical and

emotional characteristics, attitudes, environmental driving conditions and amount of driving exposure. In view of (3.1c), the parameter λ is a function of time through changing conditions. However, we shall treat this parameter as a constant over relatively short periods of time, say one or two years, 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 the observation period.

According to fundamental assumption (3.1b), during a period of time the parameter λ differs from individual to individual within the population of drivers. Therefore we shall assume that the λ's are values of a random variable Λ having distribution function $U(\lambda)$. Using the formula for conditional probabilities we obtain the expression

$$(3.8) \qquad q(n,t) = \int_0^\infty p(n,t)\,dU(\lambda) , \qquad \begin{array}{l} n = 0,\ 1,\ 2,\ \ldots, \\ \lambda > 0,\ t > 0 . \end{array}$$

The probability density function $q(n,t)$ is the distribution of accident involvements for individuals within a population of drivers characterized by the d.f. $U(\lambda)$. When $p(n,t)$ is of the form (3.6), the class of $q(n,t)$ distributions is called a compound Poisson distribution. In the previous paragraph we observed that λ in reality is a function of many factors which change with time. Accordingly, for a fixed population, the parameters of $U(\lambda)$ are time dependent mirroring the changes taking place among the individuals within that population.

To summarize, we will make use of two probability distributions for the random variable $N(t)$. The Poisson model, (3.6), is applicable

when we are focusing our attention on an individual for whom the characterizing λ is known or "essentially" known. When referring to an individual selected at random from a population of drivers, the appropriate model is (3.8).

3.4 The Model as a Compound Poisson Process

If the expression for $p(n,t)$ in the overall model (3.2) is the Poisson probability function (3.6), then the model represents the d.f. of a compound Poisson process. Throughout this work the word "process" will be used to distinguish between the two types of compound Poisson distributions as represented by (3.2) and (3.8), respectively. Because of the general applicability of this particular stochastic process, it is discussed in many of the recent textbooks dealing with the subject of probability and stochastic processes. Two such works are Feller (1966) and Parzen (1962). It is pointed out in these references that a compound Poisson process has stationary independent increments. This means that the k random variables

$$X(t_1) - X(t_o), \ldots, X(t_k) - X(t_{k-1})$$

are independent and $X(s + t) - X(s)$ depends only on the length of the interval t but not on s.

Starting with (3.3), it is a simple matter to obtain the characteristic function of the compound Poisson process variable, X(t). Denoting its c.f. by f(h) we get

$$f(h) = \sum_{n=o}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} [g(h)]^n$$

$$(3.9) \qquad = e^{-\lambda t} \sum_{n=o}^{\infty} \frac{[\lambda t g(h)]^n}{n!}$$

$$= e^{\lambda t[g(h)-1]}$$

where $g(h) = \int_{-\infty}^{\infty} e^{ihx} dG(x)$.

Since we may write

$$f(h) = (e^{\frac{\lambda}{k} t [g(h)-1]})^k$$

for every k = 1, 2, 3, ..., the compound Poisson distribution is infinitely divisible. It follows that the sum of independent identically distributed compound Poisson process variables is also a compound Poisson process variable. An important implication for this study is that the distribution of accumulated costs during one unit of time for k individuals having a common d.f. is the same as that for one of those individuals over a period of k units of time.

Let us consider the sum

$$(3.10) \qquad S_k(t) = X_1(t) + X_2(t) + \ldots + X_k(t)$$

for a fixed t. Then if the $X_i(t)$ are independently and identically distributed compound Poisson process variables, the d.f. of $S_k(t)$ is

given by

$$F(x,kt) = \sum_{n=o}^{\infty} \frac{e^{-k\lambda t}(k\lambda t)^n}{n!} G_n(x)$$

for which

$$E(N(t)) = k\lambda t \quad \text{and} \quad Var(N(t)) = k\lambda t \ .$$

Assuming finite first and second moments for G(x), denoted by μ and μ_2', respectively, we obtain from (3.4) and (3.5)

$$(3.11) \qquad E(S_k(t)) = k\lambda tu \quad \text{and} \quad Var(S_k(t)) = k\lambda t\mu_2' \ .$$

As a consequence of the central limit theorem, for a fixed t the random variable

$$\frac{S_k(t)/k - \lambda t\mu}{\sqrt{\lambda t\mu_2'/k}}$$

converges in distribution to that of a normal variable with mean 0 and variance 1 as $k \to \infty$. Analogous to this result, Cramér (1937) first showed that if X(t) is a compound Poisson process variable, then

$$\frac{X(t) - \lambda t\mu}{\sqrt{\lambda t\mu_2'}}$$

is asymptotically normal N(0,1) as $t \to \infty$.

3.5 The Concept of Accident Cost Potential

As the next chapter will clearly indicate, when we focus our attention on an individual driver, the occurrence of an accident involvement is, in general, an infrequent event. This fact, together

with the variability associated with X(t), means that the empirical average annual involvement costs experienced by an individual, even if computed over a lifetime, does not adequately reflect the individual's driving skill, exposure in terms of mileage, and environmental driving conditions. That is to say, we would expect average annual accident costs generated by two drivers of equal skill and identical exposure to be quite different. In an effort to minimize the fluctuations in average annual costs as accumulated with time, we may very well ask the hypothetical question: "Suppose we were able to observe a driver under certain representative conditions for thousands of years. What is the asymptotic distribution of his average accident costs?"

To answer the posed question, we will consider the random variable $Z(t) = X(t)/t$. We note that Z(t) represents the accident involvement cost per unit of time. Using the characteristic function technique, it follows from (3.9) that

$$E(e^{ihZ(t)}) = e^{\lambda t[g(h/t)-1]} .$$

If we assume that $E(X) = \mu < \infty$, we may write

$$g(h/t) = 1 + i\mu \frac{h}{t} + \frac{h}{t} \epsilon(h/t)$$

where $\epsilon(\gamma) \to 0$ as $\gamma \to 0$.

On substitution we immediately see that

$$\lim_{t\to\infty} E(e^{ihZ(t)}) = e^{i\lambda\mu h} .$$

Now $e^{i\lambda\mu h}$ is the characteristic function of a degenerate probability

having all of its mass at the point $z = \lambda\mu$. By the uniqueness theorem for characteristic functions, the corresponding distribution function is

$$\lim_{t\to\infty} F(zt,t) = \begin{cases} 0 , & \text{when } z < \lambda\mu , \\ 1 , & \text{when } z \geq \lambda\mu , \end{cases} \tag{3.12}$$

since $\Pr\{Z(t) \leq z\} = \Pr\{X(t) \leq zt\} = F(zt,t)$.

This result may be interpreted as follows: If we were able to observe a driver under the same conditions for many, many years, the distribution of his accident costs per unit time converges in distribution to a constant, $\lambda\mu$. This "constant", which can be associated with each individual driver, will hereafter be called his accident cost potential and represents a theoretical accident 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 likely to change with time.

The derivation demonstrated in the previous paragraph is a result that follows from a general theorem obtained by O. Lundberg (1940). He concerned himself with the model

$$F(x,t) = \sum_{n=o}^{\infty} p(n,t)\, G_n(x) , \qquad t > 0 , \tag{3.2}$$

and assumed $p(n,t)$ to be of the form

$$(3.13)\qquad p(n,t) = \int_0^\infty \frac{e^{-st}(st)^n}{n!}\, dU(s) , \qquad n = 0, 1, 2, \ldots ,$$

$$s > 0, \quad t > 0 .$$

He showed that if G(x) is a d.f. with finite mean $\mu > 0$, then for all real values of x

$$(3.14)\qquad \lim_{t\to\infty} F(zt,t) = U(z/\mu) .$$

Since (3.13) becomes (3.6) when

$$U(s) = \begin{cases} 0 , & \text{when } s < \lambda , \\ 1 , & \text{when } s \geq \lambda , \end{cases}$$

we see that (3.12) is an expected consequence of (3.14).

The result for the random variable $Z(t) = X(t)/t$ can also be easily demonstrated for the random variable $N(t)/t$, where N(t) has the Poisson distribution (3.6). Since the characteristic function of N(t) is

$$e^{\lambda t[e^{ih}-1]}$$

it follows that

$$E(e^{ihN(t)/t}) = e^{\lambda t[e^{ih/t} - 1]}$$

$$= e^{\lambda t[i\frac{h}{t} - \frac{h^2}{2t^2} - \frac{h^3}{6t^3} + \ldots]} .$$

Hence

$$\lim_{t\to\infty} E(e^{ihN(t)/t}) = e^{i\lambda h}$$

so that we obtain the expected result that the distribution of accident frequency per unit time converges in distribution to the "constant" λ assumed to be characteristic of the particular driver. Henceforth, λ will be known as the individual's accident rate potential.

Although the concept of accident rate potential essentially follows the rationale of the accident proneness phenomenon, in view of our fundamental assumptions we do not accept propensity as the sole determinant of accident susceptibility for an individual. Instead, our contention is that proneness, in the sense of an inherent tendency to accident, plays only a minor role in accident rates as compared to personal factors such as experience, environmental factors such as traffic density, and exposure in terms of annual mileage. However, we assume that although these personal, environmental and exposure factors change from week to week, an individual's mode of living dictates that over the course of a year they "average out" so that in this sense his accident rate potential is thought to be relatively constant from year to year. This naturally does not hold true if a person moves from one community to another quite different in character, or if he drives 5,000 miles one year and 10,000 the next. In such circumstances, by definition, the accident rate potential associated with the individual would change.

3.6 Assumptions Relative to G(x)

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

costs applicable 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 cost distributions applicable to a type of driver it is necessary to look at samples taken from a population of drivers.

The Illinois and Washington cost studies referred to in Chapter 1 spotlight 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. (See Figure 5.1.)

(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 Chapter 5, where we analyze accident cost data, we find that a mixture of exponential distributions of the form

$$w(x) = \alpha_1\theta_1 e^{-\theta_1 x} + \alpha_2\theta_2 e^{-\theta_2 x}, \qquad x \geq 0, \quad \theta_1 > 0, \quad \theta_2 > 0,$$
$$\alpha_1 \geq 0, \quad \alpha_1 + \alpha_2 = 1,$$

graduates the empirical population distribution reasonably well. Presumably, if reliable estimation procedures were available, we could improve the fit by increasing the number of exponentials in the mixture, i.e., let

$$w(x) = \sum_{i=1}^{k} \alpha_i \theta_i e^{-\theta_i x}, \qquad x \geq 0, \quad \theta_i > 0, \quad \alpha_i \geq 0,$$

$$\sum_{i=1}^{k} \alpha_i = 1, \quad i = 1, 2, \ldots, k,$$

for some finite k. More generally, we can assume that the θ'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 (3.8) is a mixture of Poisson distributions. This conforms with our fundamental assumption (3.1b) of which characteristic (ii) above is an expression.

As a consequence, we shall assume that the distribution of the cost of an accident involvement for an individual driver is exponential, i.e., the distribution function of the random variable X is assumed to be

$$(3.15) \qquad G(x) = 1 - e^{-\theta x}, \qquad x \geq 0, \quad \theta > 0.$$

The corresponding probability density function is

$$(3.16) \qquad g(x) = \theta e^{-\theta x}, \qquad x \geq 0, \quad \theta > 0,$$

with mean and variance

$$(3.17) \qquad E(X) = \theta^{-1} \quad \text{and} \quad Var(X) = \theta^{-2}.$$

For convenience, hereafter we shall denote

$$(3.18) \qquad E(X) = \mu \quad \text{and} \quad Var(X) = \mu^2.$$

We further assume that θ is a particular value of the random variable Θ having d.f. $V(\theta)$ where $\theta > 0$. It follows from these

assumptions that the p.d.f. of accident involvement costs for the population as a whole is a compound exponential,

$$(3.19) \qquad w(x) = \int_0^\infty \theta e^{-\theta x}\, dV(\theta) , \qquad x \geq 0, \quad \theta > 0 .$$

It is well-known that the sum of n independently and identically distributed exponential variables is a gamma variable. (See Feller (1966).) Hence, in terms of our model, the p.d.f. of $X_1 + X_2 + \ldots + X_n$ is

$$(3.20) \qquad g_n(x) = \frac{\theta^n}{(n-1)!} x^{n-1} e^{-\theta x} , \qquad x \geq 0, \quad \theta > 0 , \quad n = 1, 2, 3, \ldots,$$

so that

$$(3.21) \qquad G_n(x) = \int_0^x g_n(s)\,ds , \qquad x \geq 0, \quad n = 1, 2, 3, \ldots,$$

$$= 1 - e^{-\theta x}\left(1 + \frac{\theta x}{1!} + \frac{(\theta x)^2}{2!} + \ldots + \frac{(\theta x)^{n-1}}{(n-1)!}\right) ,$$

and $$G_o(x) = \begin{cases} 0 , & \text{when } x < 0, \\ 1 , & \text{when } x \geq 0 . \end{cases}$$

We are now in a position to express the mean and variance of X(t) in terms of the parameters of the assumed distributions. From (3.7) and (3.18), the expressions (3.4) and (3.5) become

$$(3.22) \qquad E(X(t)) = \lambda\mu t \quad \text{and} \quad Var(X(t)) = 2\lambda\mu^2 t .$$

4. THE ANALYSIS OF ACCIDENT FREQUENCY

4.1 Introduction

In Section 3 of the previous chapter the assumptions concerning the probability density function $p(n,t) = Pr\{N(t) = n\}$ were stated and discussed. Recall that these assumptions resulted in the expression

$$(3.6) \qquad p(n,t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} , \qquad n = 0, 1, 2, \ldots, \quad \lambda > 0, \quad t > 0 ,$$

where λ is called the accident rate potential of the individual. The parameter λ is assumed to be a function of the personal, environmental and driving exposure characteristics of the individual and consequently it varies with time. In addition, it is assumed that λ is a value of a random variable Λ having d.f. $U(\lambda)$ which is the distribution of accident rate potentials within a population of drivers. Consequently, the distribution of $N(t)$ with respect to individuals within a population is given by

$$(3.8) \qquad q(n,t) = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!}\, dU(\lambda) , \qquad n = 0, 1, 2, \ldots, \quad \lambda > 0, \quad t > 0 .$$

The product $100q(n,t)$ gives the percentage of individuals in the population involved in $n = 0, 1, 2, \ldots,$ accidents during a time period of t units.

In this chapter we shall investigate the historical candidate for $U(\lambda)$ first proposed by Greenwood and Yule (1920) and later ably

discussed in all of its ramifications by Arbous and Kerrich (1951). We shall look into the adequacy of the negative binomial model with respect to the California Driver Record Study (California Department of Motor Vehicles, 1964-67). Finally, a presentation will be made on proposed procedures for estimating the value λ for an individual based upon his personal characteristics, driving record and location using the California data for illustrative purposes.

4.2 The Negative Binomial Model

If we assume that Λ is a random variable having a gamma distribution with $U(\lambda)$ of the form

$$(4.1) \qquad U(\lambda) = \frac{1}{\Gamma(r)} \int_0^{\frac{r}{m}\lambda} y^{r-1}e^{-y}dy , \qquad y > 0, \quad r > 0, \quad m > 0 ,$$

where $\Gamma(r)$ is the gamma function defined as

$$\Gamma(r) = \int_0^{\infty} x^{r-1}e^{-x}dx , \qquad r > 0 ,$$

then the probability density function of Λ is given by

$$(4.2) \qquad u(\lambda) = \frac{(r/m)^r}{\Gamma(r)} \lambda^{r-1}e^{-(r/m)\lambda} , \qquad \lambda > 0, \quad m > 0, \quad r > 0 .$$

From (4.2) we find that the mean and variance of Λ are

$$(4.3) \qquad E(\Lambda) = m \qquad \text{and} \qquad Var(\Lambda) = \frac{m^2}{r} .$$

The function (4.2) is a variation of the gamma p.d.f. used by Greenwood and Yule in their classic 1920 paper.

We are now in a position to establish the frequency of accident involvements for individuals within a population according to formula (3.8). Using the functional forms (3.6) and (4.2) for $p(n,t)$ and $u(\lambda)$, respectively, we obtain

$$(4.4) \qquad q(n,t) = \int_0^\infty p(n,t)\,u(\lambda)\,d\lambda$$

$$= \frac{\Gamma(n+r)}{n!\,\Gamma(r)} \left(\frac{r}{r+mt}\right)^r \left(\frac{mt}{r+mt}\right)^n , \qquad n = 0,\ 1,\ 2,\ \ldots ,$$

$$r > 0, \quad m > 0, \quad t > 0 .$$

The form of $q(n,t)$ is that of a negative binomial p.d.f. having mean and variance

$$(4.5) \qquad E(N(t)) = mt \qquad \text{and} \qquad Var(N(t)) = mt\left(1 + \frac{mt}{r}\right) .$$

The function (4.4), then, is the negative binomial model which occupies so much space in accident research literature.

From (4.5) we see that the parameter m is the mean number of accidents per individual per unit of time. It follows from (4.3) that the parameter r is a measure in the inverse sense of the degree of variability in the random variable Λ. That is, if r is small the heterogeneity of the population with respect to accident rate potential is great whereas, correspondingly, a large r is indicative of homogeneity. In fact, if r is very large the form of $q(n,t)$ tends to that of a Poisson distribution. This can be seen by finding the limiting form of the characteristic function of $N(t)$. Here

$$E(e^{ihN(t)}) = \sum_{n=o}^{\infty} e^{ihn}\, q(n,t)$$

$$= \left\{ \frac{r}{r + mt - mte^{ih}} \right\}^{r}$$

$$= \left\{ 1 - \frac{mt}{r} (e^{ih} - 1) \right\}^{-r} .$$

Hence

$$\lim_{r\to\infty} E(e^{ihN(t)}) = e^{mt(e^{ih}-1)} ,$$

the right-hand side of which is the characteristic function of a Poisson random variable having parameter mt.

At this point it is instructive to verify that $U(\lambda)$ is in fact the distribution of accident rate potentials within the population of drivers as defined in Section 3.5. From the above calculation we know immediately that the characteristic function of $N(t)/t$ is

$$E(e^{ihN(t)/t}) = \left\{ 1 - \frac{mt}{r} (e^{ih/t} - 1) \right\}^{-r}$$

$$= \left\{ 1 - \frac{mt}{r} \left(i \frac{h}{t} - \frac{h^2}{2t^2} - i \frac{h^3}{6t^3} + \ldots\right) \right\}^{-r} .$$

Taking the limit as $t \to \infty$ we get

$$\lim_{t\to\infty} E(e^{ihN(t)/t}) = \left\{ 1 - i \frac{mh}{r} \right\}^{-r} .$$

The expression on the right-hand side is the characteristic function of a gamma random variable having a d.f. of the precise form (4.1).

Therefore by the characteristic function uniqueness theorem, $U(\lambda)$ is the distribution function of $N(t)/t$ as $t \to \infty$.

In order to determine the appropriateness of this model in explaining accident distributions, it is necessary to estimate the parameters m and r from empirical data. Bliss (1953) and Sichel (1951) discuss this problem. If n_i denotes the number of accidents in time t suffered by the i^{th} individual in a sample of size k, the method of moments yields the following estimators for m and r:

$$(4.6) \qquad \hat{m} = \frac{\bar{n}}{t} \qquad \text{where} \qquad \bar{n} = \frac{1}{k} \sum_{i=1}^{k} n_i$$

and

$$(4.7) \qquad \hat{r} = \frac{\bar{n}^2}{s^2 - \bar{n}} \qquad \text{where} \qquad s^2 = \frac{1}{k-1} \sum_{i=1}^{k} (n_i - \bar{n})^2 .$$

The variances of these estimators are

$$\mathrm{Var}(\hat{m}) = \frac{m}{kt} \left(1 + \frac{mt}{r}\right)$$

and (for k large)

$$\mathrm{Var}(\hat{r}) \doteq \frac{2r(r+1)}{k} \left(1 + \frac{r}{mt}\right)^2 .$$

The latter is the asymptotic variance of r. The method of maximum likelihood yields the same estimator for m, (4.6). A complex iterative procedure is required, however, to obtain the maximum likelihood estimator for r. A chart included in the Sichel article reveals that in the case of the California data the method of moment's estimator

as compared to the maximum likelihood estimator varies in efficiency from 95 percent when applied to data for one year to 85 percent when applied to three-year data. As a consequence, it was felt that the method of moments yields a sufficiently accurate estimate for r and so it is the procedure used in the next section where the model is fitted to the California data.

4.3 Fit of Negative Binomial Model to California Data

A description of the 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67) appears in Chapter 1 of this work. Recall that the California study provides data on approximately 148,000 drivers over the three-year experience period 1961-63. Actually, due to the time lag between the occurrence of an accident and the accident report, it is estimated that the 1963 period represents about a 10½-month interval rather than a full 12 months. For that reason, from the standpoint of a period of time, 1963 is considered to be 7/8 of a year.

In Table 4.1 we see the fit of the negative binomial model (4.4) to the empirical accident distributions generated by these 148,000 individuals during the specified time intervals. In each case the fit is amazingly close. At first glance this would seem to indicate that the negative binomial is indeed a valid model. However, we 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 $u(\lambda)$ distribution contrary to the "accident proneness" assumption of the earlier studies. This assumption of constant parameter

Table 4.1 Comparison of actual and theoretical (negative binomial) accident distributions, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

	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
$\chi^2 = 4.90$, 3 d.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
$\chi^2 = 1.47$ 1 d.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
$\chi^2 = 1.00$, 1 d.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
$\chi^2 = 0.07$, 1 d.f.	3+	35	34
	Total	148,006	148,006

values from one time period to the next is clearly implied in Appendix 9.1.1 where a variation of Kerrich's bivariate negative binomial derivation is given.

To show that differences in parameter values do differ significantly, we can perform statistical tests. Because the 1963 effective time period is not the same as that for 1961 and 1962, we cannot reasonably perform an analysis of variance to test for the equality of the three means. However, since 1961 and 1962 represent intervals of equal length, we can test the hypothesis $m_1 = m_2$, where m_1 and m_2 are the respective population means, by using the test statistic

$$z = \frac{\hat{m}_1 - \hat{m}_2}{s_{\hat{m}_1 - \hat{m}_2}} .$$

Here $\hat{m}_1 = \bar{n}_1$ and $\hat{m}_2 = \bar{n}_2$. If we assume independent samples, then

$$s_{\hat{m}_1 - \hat{m}_2} = \sqrt{s^2_{\bar{n}_1} + s^2_{\bar{n}_2}} ,$$

and we get $z = -2.58$ indicating a significant difference at the .01 level. It may be argued that the samples are not independent since both samples are composed of the same individuals. Appendix 9.1.3 shows that the samples are indeed positively correlated. According to the bivariate negative binomial model this correlation theoretically is on the order of 0.065. The California Driver Record Study (California Department of Motor Vehicles, 1964-67, Part 6, Table 7) shows positive correlations of a magnitude slightly less than this. In any event, admitting the samples to be correlated increases the

numerical value of z since in that case

$$s_{\hat{m}_1-\hat{m}_2} = \sqrt{s^2_{\bar{n}_1} + s^2_{\bar{n}_2} - 2\hat{\rho}s_{\bar{n}_1}\, s_{\bar{n}_2}}$$

and $2\hat{\rho}s_{\bar{n}_1}\, s_{\bar{n}_2}$ is positive.

In like manner, to "test" the hypothesis $r_1 = r_2$ for the years 1961 and 1962 we may improvise the statistic

$$w = \frac{\hat{r}_1 - \hat{r}_2}{s_{\hat{r}_1-\hat{r}_2}} \quad .$$

If we assume independence we obtain w = 2.08 which is likely evidence of unequal parameters, although the distribution of w is not known. In conclusion, the California data would lead us to believe that the negative binomial model is appropriate for a given observational period but that the parameters, through a changing distribution of accident cost potentials, vary over time. This latter phenomenon reinforces our fundamental assumption (3.1c). As to the former, this is not to be expected in general. However, it can be shown that the sum of two negative binomial variables is negative binomial <u>if the parameters are the same</u>. See Appendix 9.1.2.

In Part 6 of the California Study (California Department of Motor Vehicles, 1964-67) the accident distributions by sex for each of the three years is given. Using the method of moments we find the negative binomial fitted to the observed data in Table 4.2. As with the combined data, the negative dinomial distribution provides a remarkably close fit in every case. This decomposition of a negative

Table 4.2 Comparison of actual and theoretical accident distributions by sex, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

	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
$\chi^2 = 3.19$, 1 d.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
$\chi^2 = 1.37$, 1 d.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
$\chi^2 = 0.07$, 1 d.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
$\chi^2 = 0.62$, 1 d.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
$\chi^2 = 1.16$, 1 d.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
$\chi^2 = 0.30$, 1 d.f.	3+	3	4
		61,280	61,280

binomial distribution into a mixture of two negative binomial distributions is not true in general but is an obvious characteristic of these data. 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.

4.4 Predictive Aspects of the Negative Binomial Model

In order to consider the element of prediction present in the negative binomial model we again turn to the derivations found in Appendix 9.1. A selection of the functions found there can be interpreted within the framework of accident occurrence as follows:

1. $q(n_1,t_1;\ n_2,t_2)$ is the probability that an individual will have n_1 accidents during a period of t_1 units of time and n_2 accidents during a nonoverlapping time period of t_2 units.
2. $q(n_2,t_2|n_1,t_1)$ is the probability that an individual who has had n_1 accidents during time t_1 will have n_2 accidents during time t_2.
3. $u(\lambda|n_1,t_1)$ is the probability density function of the random variable Λ corresponding to those individuals observed to have had n_1 accidents during t_1 units of time.

In applying these functions to the problem of prediction we must bear in mind that their derivations are based on the assumption that the negative binomial distributions corresponding to two time periods have common parameters m and r. We have seen in the previous section that in the case of the California data this assumption seems to be untenable. Accordingly, when we apply the derived expressions to the

available data we need to acknowledge that in these particular examples some of the model's inherent predictive power has been lost. Later in this thesis we will attempt to predict 1963 results on the basis of the combined prior years 1961-62. With that in mind we will look at the negative binomial predictions with respect to those two time periods.

Because the estimates of the parameters, particularly for r, are not consistent in the two time periods, we are confronted with the decision of choosing a common m and r. The weighted averages for $\hat{m}$ and $\hat{r}$, together with the computed intervals $\hat{m} \pm 2s_{\hat{m}}$ and $\hat{r} \pm 2s_{\hat{r}}$, indicate that the values m = 0.0710 and r = 1.0000 are reasonable conjectures. It is to be noted that both of these hypothesized values fall within the above intervals for both time periods. In Table 4.3 we see the fit of the theoretical bivariate negative binomial $q(n_1,2;\ n_2,7/8)$, within parentheses, to the observed California data for the periods 1961-62 and 1963. Although the observed and predicted numbers for the joint distribution are "in the same ball park," they are not compatible according to a chi-square goodness of fit test. Curiously, however, the marginal distributions for the observed and expected are in statistical agreement as evidenced by a chi-square value of 2.81 with 3 degrees of freedom for the 1963 marginal and by a chi-square value of 8.64 with 4 degrees of freedom for the 1961-62 marginal.

As previously noted, the function $q(n_2,t_2|n_1,t_1)$, derived in Appendix 9.1.4 reveals the probability of a future number of accidents corresponding to those individuals having had n_1 accidents in the past. Table 4.4 gives the values of this function along with the empirical

Table 4.3 Observed and (expected) bivariate accident distributions in periods 1961-62 and 1963, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

$m = 0.0710, \quad r = 1.0000$

No. of Accidents in 1961-62 (n_1)	No. of Accidents in 1963 (n_2)				Marginal Distr. 1961-62
	0	1	2	3+	
0	122,716 (123,062)	6,558 (6,349)	364 (328)	22 (19)	129,660 (129,758)
1	14,831 (14,512)	1,344 (1,497)	119 (116)	11 (9)	16,305 (16,134)
2	1,728 (1,711)	217 (265)	22 (27)	0 (2)	1,967 (2,005)
3	177 (202)	31 (42)	4 (5)	0 (1)	212 (250)
4+	29 (27)	6 (7)	1 (1)	2 (0)	38 (35)
Marginal Distr. 1963	139,481 (139,514)	8,156 (8,160)	510 (477)	35 (31)	148,182 (148,182)

results in the California sample of 148,000 drivers. Again we see that the theoretical and sample results are of the same order of magnitude. Perhaps of greater interest is the expected number of future accidents for those individuals who have been involved in 0, 1, 2, ... past accidents. This information is contained in the conditional mean of $N_2(t_2)$ given $N_1(t_1) = n_1$, namely,

$$(4.8) \qquad E(N_2(t_2) \mid n_1, t_1) = \frac{mt_2(n_1+r)}{r+mt_1} \quad .$$

The conditional variance of $N_2(t_2)$ is

Table 4.4 Comparison between conditional theoretical probabilities and observed proportions, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

$m = 0.0710, \quad r = 1.0000$

No. of Accidents 1963	Theoretical 1963	Observed 1963
	N_0 Accidents during 1961-62	(129,660 cases)
0	0.9484	0.9464
1	.0489	.0506
2	.0025	.0028
3+	.0002	.0002
	One Accident during 1961-62	(16,305 cases)
0	0.8995	0.9096
1	.0928	.0824
2	.0072	.0073
3+	.0005	.0007
	Two Accidents during 1961-62	(1,967 cases)
0	0.8531	0.8785
1	.1320	.1103
2	.0136	.0112
3+	.0013	.0000
	Three Accidents during 1961-62	(212 cases)
0	0.8091	0.8349
1	.1670	.1462
2	.0215	.0189
3+	.0024	.0000

$$(4.9) \qquad \mathrm{Var}(N_2(t_2)\,|\,n_1,t_1) = \frac{mt_2(n_1+r)}{r+mt_1}\left(\frac{r+mt_1+mt_2}{r+mt_1}\right) .$$

In the language of statistics the function (4.8) is called the regression of $N_2(t_2)$ on $N_1(t_1)$. It is important to note that the relationship between mean number of future accidents and the number of past accidents is linear according to the negative binomial model. In Table 4.5 we have a comparison between the expected number and the

Table 4.5 Comparison of 1963 actual mean number of accidents and expected number of accidents for drivers having had n_1 accidents during 1961-62, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

$m = 0.0710, \quad r = 1.0000$

n_1	Actual Mean	$E(N_2(7/8) \mid n_1, 2)$	$Var(N_2(7/8) \mid n_1, 2)$
0	0.0567	0.0544	0.0574
1	.0991	.1088	.1147
2	.1327	.1632	.1721
3	.1840	.2176	.2294
> 0	0.1040	0.1165	0.1233

actual mean number of accidents suffered by the California drivers in 1963 classified according to their 1961-62 accident record. The figures for $n_1 > 0$ are included so that a comparison can be made between those who had at least one accident during the prior two-year period and those who were accident-free during that period.

At first glance the theoretical and corresponding sample means displayed in Table 4.5 appear to be close in value. However, if we construct the five intervals

$$E(N_2(7/8) \mid n_1, 2) \pm 2 \sqrt{Var(N_2(7/8) \mid n_1, 2)/k_{n_1}}$$

for $n_1 = 0, 1, 2, 3$ and $n_1 > 0$ we find that in every case the actual mean falls outside of the corresponding interval. Here k_{n_1} is the number of drivers observed to have had n_1 accidents. Again the sample results appear to statistically belie the hypothesized bivariate negative binomial model. Nevertheless, we can conclude from

Tables 4.4 and 4.5 that as far as groups of individuals are concerned accident history is a predictor of future accidents.

Finally, let us investigate the information that can be derived from the negative binomial model in regard to the accident rate potentials, λ. First of all, Table 4.6 contains values of the theoretical distribution function $U(\lambda)$ as determined from the California data using the assumed parameter values. Table 4.6 was compiled with the aid of the Pearson (1957) tables. Similar tables could be prepared separately for the male drivers as a group and the female drivers as a group. It should be noted, however, that in order to be consistent with the trait observed near the end of Section 4.3, the relationships between the parameters of the combined groups and the parameters of the mixture in the proportions α_1, $\alpha_2 (\alpha_1 + \alpha_2 = 1)$ are necessarily

Table 4.6 Theoretical distribution of accident rate potentials, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

$m = 0.0710, \quad r = 1.000$

λ	$U(\lambda) = Pr\{\Lambda \leq \lambda\}$
0.02	0.245
.04	.430
.06	.570
.071	.632
.08	.676
.10	.755
.12	.815
.15	.879
.20	.940

$$m = \alpha_1 m_1 + \alpha_2 m_2$$

and

$$r = \frac{(\alpha_1 m_1 + \alpha_2 m_2)^2}{\alpha_1 m_1^2 (1+r_1^{-1}) + \alpha_2 m_2^2 (1 + r_2^{-1}) - (\alpha_1 m_1 + \alpha_2 m_2)^2} \quad .$$

This follows when we equate the first two theoretical moments of a negative binomial with those of a mixture of two negative binomial distributions.

The question might be asked, what information does the model give us in regard to the accident rate potential associated with a particular individual? To answer this question, first recall that the function $u(\lambda|n_1, t_1)$ is the distribution of Λ for those individuals who have had n_1 accidents during a time period of length t_1. As noted in Appendix 9.1.4(a), the transformed random variable $2(r/m + t_1)\Lambda$ is a chi-square variable with $2(n_1 + r)$ degrees of freedom. This permits us to obtain confidence intervals for λ based upon the individual's history of accident involvements. Assuming an $m = 0.0710$ and an $r = 1.0000$ and letting $t_1 = 2.875$ we find in Table 4.7 the 90 percent confidence limits for the theoretical λ as it applies to an individual who experienced n_1 accidents during the period of the California study.

From the standpoint of using accident records to discriminate between high and low predisposition to future accidents the results in Table 4.7 are indeed disappointing. First of all, the intervals are too wide to indicate meaningfully a λ value based upon past accidents,

Table 4.7 Confidence limits for an individual's accident rate potential, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

$m = 0.0710, \quad r = 1.0000, \quad r_1 = 2.875$

No. of Accidents n_1	90 Percent Confidence Limits for λ, Given n_1
0	0.003 - 0.177
1	.021 - .280
2	.048 - .371
3	.081 - .457
4	.116 - .540
5	.154 - .620
6	.194 - .698

at least for $t_1 = 2.875$. We may note that the intervals decrease in length as t_1 increases since

$$\mathrm{Var}(\Lambda|n_1,t_1) = (n_1 + r)\left(\frac{m}{r + mt_1}\right)^2 .$$

Secondly, the intervals overlap to such an extent that it is risky to claim that an accident-free driver has in fact a lower accident rate potential than one who has been involved in many accidents. For example, according to the confidence limits it is quite possible that a person who had been involved in four accidents actually possesses a smaller λ than one who had no accidents during the experience period. In conclusion, we learn from Table 4.7 that as far as an individual is concerned, the number of past accidents alone is not an efficient predictor of future accident involvements.

Let us summarize what we have learned in applying the negative binomial model to the California data. When applied to a given

interval of time, the model fits the data exceedingly well. But when we look at two or more time intervals, whether they are overlapping or not, the assumption of constant parameters over time is violated. This would indicate that the accident rate potentials are time dependent, at least to a degree. A consequence of this phenomenon of "changing" parameters is a decrease in prediction accuracy implicit in the model. We have seen that accident history is the only variable that the model admits as a predictor of future accidents. This variable is less than adequate where prediction involving a single individual is concerned but it provides approximate if not statistically acceptable estimates relative to a group of individuals. Later in this chapter we shall find that there are better predictors than past accidents available. According to the model there is a linear relationship between number of past accidents and the expected number of future accidents. We will use this property of linearity to our advantage in sections to follow.

Although the review in the last paragraph indicates the negative binomial model is not entirely tenable, it certainly qualifies as an "approximate" model for explaining the California data. The applicability of the other models which have appeared in the literature and which were cited in Section 2 of Chapter 2 have not been investigated. It is difficult to imagine obtaining better fits to the California data in the univariate case than those displayed in Tables 4.1 and 4.2. As to the bivariate case, it is highly probable that the Bates-Neyman model would be superior to Kerrich's bivariate compound Poisson model since it admits a change in the accident rate potential from one

period to the next. Their model assumes the relationship $\lambda_2 = k\lambda_1$, where the value of k is the same for all persons under observation. This also is a rather strong assumption and it is doubtful that it is ever completely satisfied in the real world. In a recent study on accident repeatedness among children, Mellinger et al. (1965) applied the Bates-Neyman model and found that it too failed to adequately discriminate between λ values based on accident incidence alone. In any event, since the parameter k must be estimated from the data in two observational periods, any prediction the model makes about "future" accidents is really an after-the-fact type of prediction. It should be noted that the regression of future accidents on number of past accidents according to the Bates-Neyman model is linear, the same result as obtained in Kerrich's model.

In an effort to overcome the inadequacies of the earlier models, the remainder of this chapter is devoted to the development of a model for estimating individual λ's based upon not only accident history but also on other factors such as age, marital status, conviction history and principal driving environment. As mentioned in Chapter 1 and again in Chapter 3, the set of accident predictors in this study is necessarily limited to those recorded in Motor Vehicle Department files. Therefore this writer is the first to concede that in the work which follows important variables have not been included and so no claim about a final answer can or should be made. Rather, the purpose of this work is to present an approach.

4.5 The Model for Accident Rate Potential

Proceeding on the evidence that automobile accident distributions are approximately negative binomial, the question is, how does the negative binomial distribution arise? It was pointed out in Chapter 2 that the negative binomial can be generated in several ways. By allowing assumptions (3.8) and (4.1) we have committed ourselves to the following interpretation: The population of drivers is comprised of a large number of groups homogeneous with respect to accident rate potential, i.e., each individual within a given group is characterized by the same parameter, λ. Hence the probability that a particular individual belonging to this group will have n accidents during exposure time t is given by the Poisson distribution $p(n,t)$. But the Poisson parameter varies from group to group according to the distribution function $U(\lambda)$. Now if we select an individual at random from the population of drivers, prior to selection $q(n,t)$ is the probability that this individual will experience exactly n accidents during t units of time. We have seen that $q(n,t)$ is the negative binomial distribution if $U(\lambda)$ is the d.f. of a gamma variable.

In an effort to verify empirically the above explanation, the 148,000 individuals in the California sample were partitioned into 2,880 groups on the basis of six criterion variables. The intent of the classification was to subdivide the California sample into homogeneous groups with respect to accident rate potential as measured by the personal characteristics of the individual and his driving record. The variables are sex, marital status, age, area of residence, conviction history, and accident history. The levels within each

variable are presented in Table 4.8. These six factors and how they relate to accidents will be discussed later in this section. At that time it will become evident as to why these variables were chosen as vehicles for assigning drivers to groups. Care had to be exercised in determining the number of criteria and their levels for if the resulting partition is too fine, the number of cases in each category would be too few to yield reliable information. On the other hand, if the partitioning is too coarse, the resulting groups would likely be heterogeneous rather than homogeneous. If in fact the groups are homogeneous, then, according to the theory, the distribution of future accidents within each group should be Poisson.

Table 4.8 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	

The computer program which partitioned the members of the California sample into the stated 2,880 groups printed out the 1963 accident distributions for the 193 groups that contained 100 or more individuals. In addition, the program calculated the maximum likelihood estimate for the Poisson parameter, namely the sample mean, for each of the 193 groups and fitted the corresponding theoretical Poisson distribution to the observed within group distribution. At the same time the value of the chi-square goodness of fit test statistic was calculated and printed out. In 167 or 86.5 percent of the 193 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 in segregating the individuals into Poisson groups. We would expect that a finer partitioning of the sample through additional variables and/or additional levels would have resulted in groups even more homogeneous. The outcome of this experiment substantiates our assumption of a compound Poisson model (3.8) as being reasonable.

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

$$(4.10) \qquad \lambda = f(\underline{x}; \underline{\beta})$$

where $\underline{x}$ represents the vector of criterion variables and $\underline{\beta}$ is the 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 (California Department of Motor Vehicles, 1964-67)

and proceed to examine the relationship between accident frequency and a selection of driver variables. First, let us investigate the influence of residence on accident rates.

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 4.9. It reveals that accident rates do indeed 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

Table 4.9 Accident rates per driver by area and county, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

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

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

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 4 (40,474 cases)	
All Other	0.147

using 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. As a case in point, the northwest corner of Riverside County is a population center whereas the eastern half of the county is a desert region. It follows that, to fully utilize the predictive power with respect to accident frequency inherent in the traffic density factor, an index by geographical area rather than by county lines is needed. However, we are obliged to use a countywide index since the residence of an individual in the California sample is specified by county alone.

A plot of accident rate versus traffic density index revealed that the mathematical relationship between these two variables may be concave downward, i.e., of decreasing slope rather than of constant slope. To verify this conjecture, two simple regression analyses were performed; the first, accident rate on traffic density index and the second, accident rate on the logarithm of traffic density index. The corresponding correlation coefficients were 0.79 and 0.85, respectively. This means that the logarithmic function provided a slightly better fit to the data. 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.

Next we shall seek out possible relationships between accident rates and the personal characteristics (i) sex, (ii) marital status,

and (iii) age. Figures 4.1 and 4.2, reproduced from Part 5 of the 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67), give us a visual representation of these relationships. Again we see that males and females constitute different driving populations, i.e., the relationships between accident rate and age and between accident rate and marital status are of different character in the two populations. As a consequence, we will be required to find a function f of (4.10) for each of the two sexes.

We notice 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 a few age groups, the same statement can be made about male drivers. Some bias undoubtedly appears in these accident rates since the data was not coded to reflect a change in marital status during the experience period. 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.

The line graphs in Figure 4.1 clearly show a curvilinear relationship between accidents and the age of male drivers. They suggest

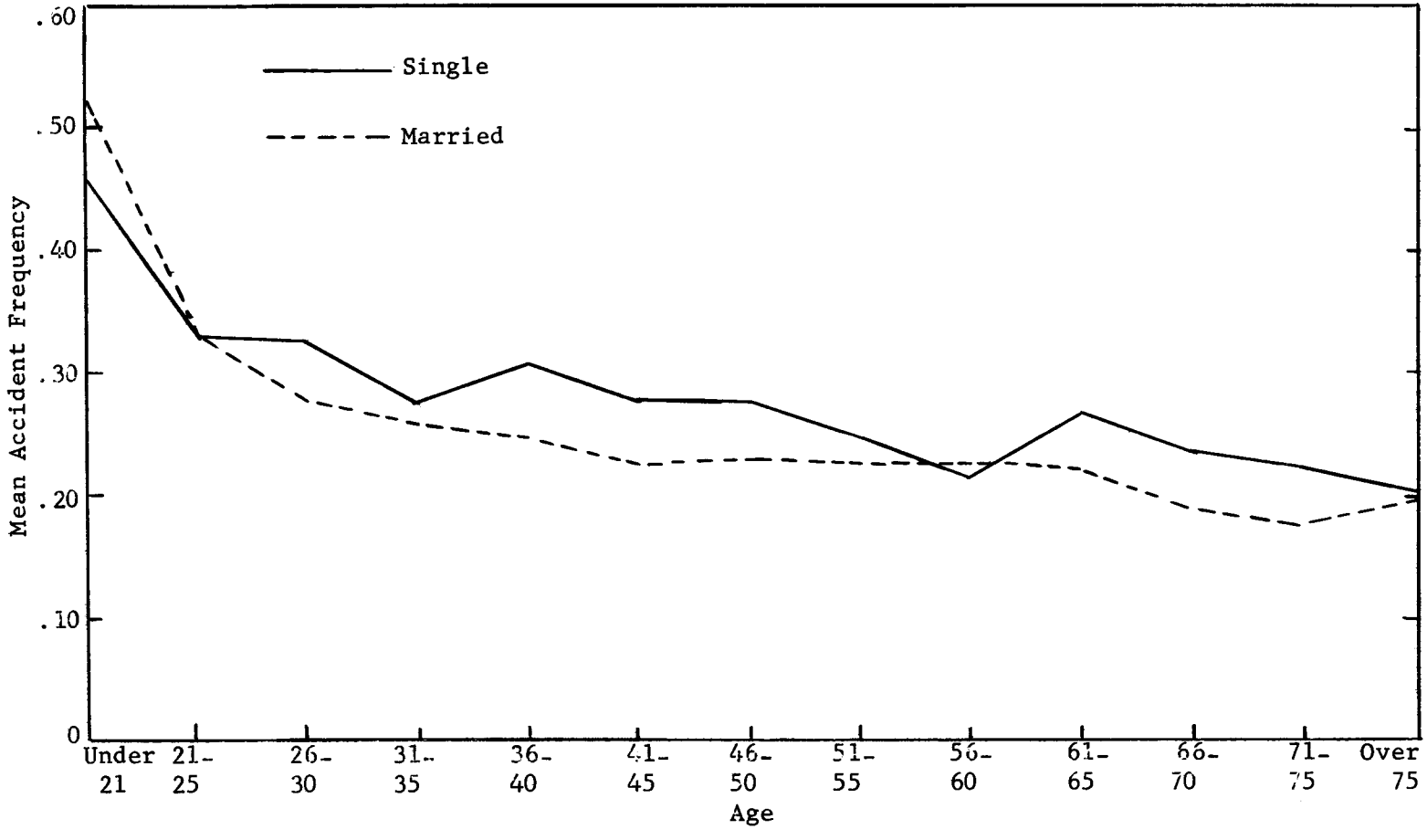

Figure 4.1 Male mean accident frequency by age and marital status, 1964 California Driver Record Study (1961-63) (California Department of Motor Vehicles, 1964-67)

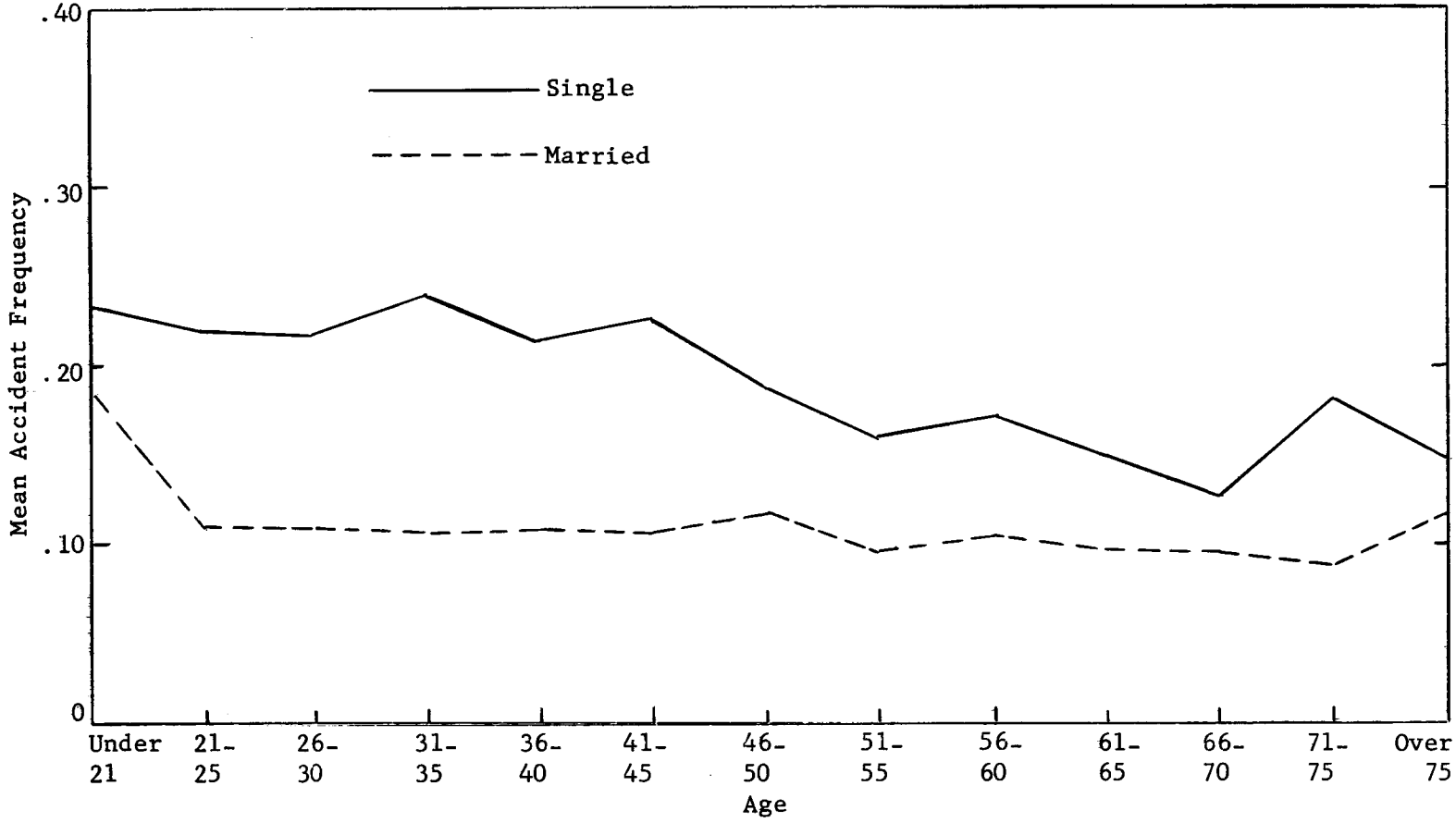

Figure 4.2 Female mean accident frequency by age and marital status, 1964 California Driver Record Study (1961-63) (California Department of Motor Vehicles, 1964-67)

a decreasing function for the mean accident frequency, y, of the type

$$y = a_3 + \frac{b_3}{z^k}$$

where z is some increasing function of age, and a_3, b_3 and k are positive constants. Assuming the intrinsically linear form above and setting z = (age in years - 13)/5, a series of regression equations were calculated for k = ½, 1, 3/2, and 2. Since y is the mean of a negative binomial variable whose variance is a function of its mean, the data calls for a weighted least squares analysis. This was done using as weights the corresponding reciprocal value of the estimate of y obtained in a first run usual (unweighted) regression analysis.

The equation which provided the best fit to the empirical means was the one derived for k = 1. The comparison between the empirical accident rates and the corresponding functional value of y is shown in Table 4.10. On the basis of these calculations we shall assume that

Table 4.10 Weighted regression of accident rates on transformed ages (males), 1964 California Driver Record Study (1961-63) (California Department of Motor Vehicles, 1964-67)

$$y = 0.1823 + 0.3183x_3 \quad \text{where} \quad x_3 = 5/(\text{age} - 13)$$

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

the relationship between accident rates, y, and age for male drivers is

$$y = a_3 + b_3 x_3$$

where $x_3 = 5/(\text{age} - 13)$, and a_3 and b_3 are parameters.

In comparing the curves in Figures 4.1 and 4.2 we find that the relationship between accident rates and the age of female drivers is less curvilinear by a considerable degree than that which exists for males. In terms of the assumed functional form in z, this observation suggests a value for k somewhat larger than that obtained for males. Following the same procedure as outlined above it was found that k = 3 provided the equation of best fit for the female drivers. The fit as displayed in Table 4.11 is seen to be quite satisfactory. Consequently we shall assume that the relationship between mean accident frequency, y, and age of female drivers is

Table 4.11 Weighted regression of accident rates on transformed ages (females), 1964 California Driver Record Study (1961-63) (California Department of Motor Vehicles, 1964-67)

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

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

given by

$$y = a_3 + b_3 x_3$$

where $x_3 = 125/(\text{age} - 13)^3$, and a_3 and b_3 are constants.

In continuing our search for the functional form of f, we will next investigate the possibilities of predicting accident involvement using driver record data. In Part 4 of the 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67) we find the relationship between accident and conviction frequencies based upon a three-year experience period involving 148,000 drivers. 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, to avoid a "built-in" correlation between accidents and countable convictions, the number of convictions does not include those resulting from an accident investigation. Bearing in mind these qualifying conditions on the counts, we find in Figure 4.3 a graphical representation of the empirical relationship between concurrent accidents and total countable convictions by sex. The actual accident rates are plotted against the conviction counts of 0, 1, 2, 3 and 4. Corresponding to a conviction count of 6 we find an accident rate based on the weighted average obtained for conviction counts of 5 or more. This was done because the number of

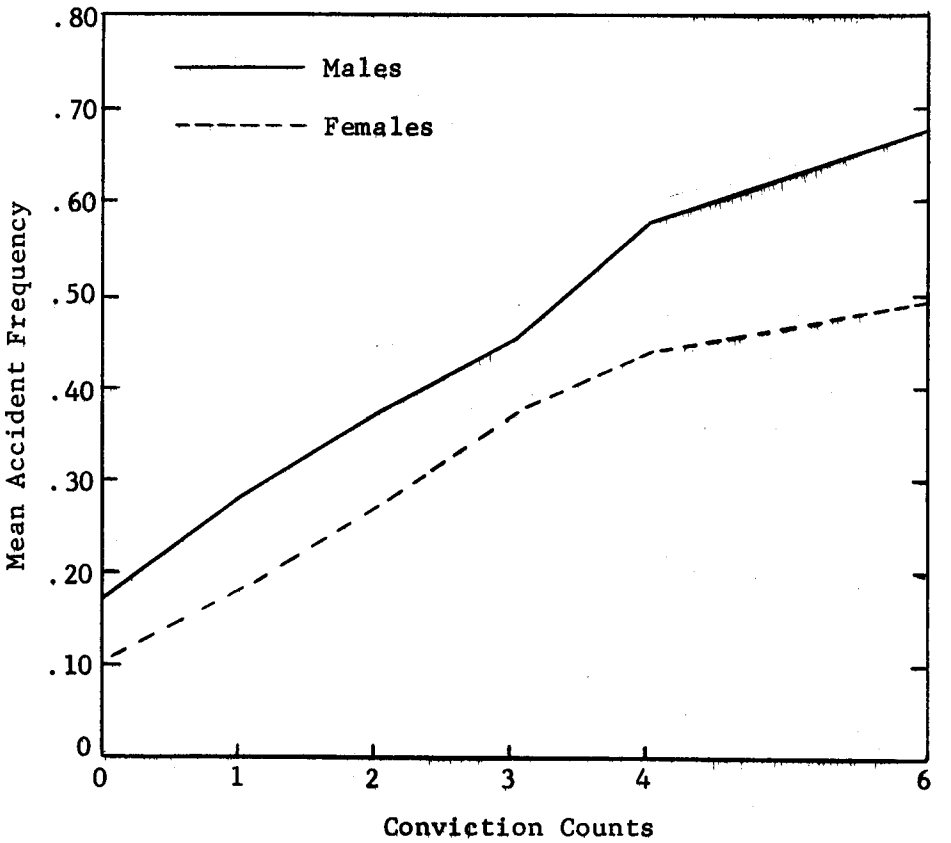

Figure 4.3 Mean accident frequencies by number of countable convictions, 1964 California Driver Record Study (1961-63) (California Department of Motor Vehicles, 1964-67)

drivers having counts of 5, 6, 7, ..., were too few per category to yield a reliable mean.

The line graphs in Figure 4.3 reveal a near linear relationship between accident means 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 4.12. On the basis of these calculations we shall hypothesize that the

Table 4.12 Weighted regression of accident rates on number of countable convictions, 1964 California Driver Record Study (1961-63) (California Department of Motor Vehicles, 1964-67)

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

mathematical relationship between mean accident frequency, y, and number of conviction counts, x_4, is

$$y = a_4 + b_4x_4$$

where a_4 and b_4 are parameters to be estimated.

It is necessary to point out explicitly that Figure 4.3 and Table 4.12 show 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, however, is the predictive nature of past convictions as it concerns future accidents. In this regard Table 4.13 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 the relationship between 1963 empirical accident rates and 1961-62

Table 4.13 Observed 1963 accident rates by 1961-62 conviction counts, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

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

conviction counts is dominantly linear by checking the differences in accident rates as we go from one conviction level to the next. Although the relationship in this instance is not as strongly linear as in the concurrent case, we shall 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.

In the previous section we concluded that the negative binomial model is at least an approximation to actual automobile accident experience. If we accept this tenet, (4.8) is the theoretical foundation for assuming that future accident rates are linearly related to the incidence of past accident involvements. To check the validity of this assumption with respect to the California data, an iterative weighted regression analysis for

$$y = a_5 + b_5 x_5$$

was performed using a computer. Here y represents 1963 accident rates and x_5 is the number of 1961-62 involvements. The weights used were the reciprocal values of the estimate of y obtained from the previous regression equation. The procedure was terminated when the estimates of the parameters stabilized to eight decimal places. Further discussion of this iterative procedure appears in Section 4.6. The results of this analysis, by sex, are given in Table 4.14. The linear relationship is apparent and so we shall accept this linearity property as a fact. Notice the greater degree of accuracy achieved in using the empirical regression line (Table 4.14) as compared to the theoretical regression line in Table 4.5.

Table 4.14 Weighted regression of 1963 accidents on 1961-62 accident counts, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

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 noncountable convictions. A noncountable 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 nonmoving offenses. The

relationship between accidents and noncountable convictions was not given separate analysis in the 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67) 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 noncountable convictions should be different than that between accidents and countable convictions we shall assume the equation

$$y = a_6 + b_6 x_6$$

where y is the future accident rate, x_6 is the prior noncountable conviction count, and a_6 and b_6 are constants.

Recall that λ is defined as the number of accident involvements per unit time based on an experience period of an infinite number of time units. Also, since the Poisson distribution, $p(n,t)$, is infinitely divisible, the experience generated during unit time by k identical individuals (with respect to λ) is equivalent to that generated by one of those individuals during k units of time. It follows that if y is defined as the mean accident frequency per unit time and if it is estimated from a large sample of homogeneous drivers, then y is a reasonable estimate of λ. On the basis of empirical evidence we have seen that a linear or near linear relationship exists between y and each investigated criterion variable, or a transform thereof. Having assumed the relationships to be in fact linear we are now in a position to write down an expression for f of (4.10). It is

$$(4.11) \qquad \lambda = f(\underline{x};\ \underline{\beta}) \equiv \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_s x_s$$

where the x_i are the criterion variables which functionally determine the value of λ and the β_i are the necessary parameters. Now (4.11) together with (3.6) permits us to finalize the form of $p(n,t)$ in terms of the characteristics of the driver, namely,

$$(4.12) \qquad p(n,t) = \frac{e^{-t\sum_{i=o}^{s}\beta_i x_i}\left(t\sum_{i=o}^{s}\beta_i x_i\right)^n}{n!}, \quad n = 0,\ 1,\ 2,\ \ldots,$$

$$\sum_{i=o}^{s}\beta_i x_i > 0,\ t > 0,$$

where $x_o = 1$. Here

$$(4.13) \qquad E(N(t)) = t\sum_{i=o}^{s}\beta_i x_i = Var(N(t)) .$$

In the following sections we shall take up the problem of estimating λ in (4.11) using the California data. Before doing so, a few comments on the limitations of the data are in order. We have discussed six possible candidates for criterion variables. That does not mean that these six are the only predictors that have a significant mathematical relationship with accident involvements. Rather, as pointed out in Chapter 3, perhaps the most significant factor of all in accident prediction is not among the six, namely, miles of driving exposure. The California records do not give this information and hence we are unable to directly include this variable in our analysis. However, as suggested in the early part of Chapter 3, some of the

variables under consideration do reflect driving mileage so that its influence does enter into our consideration indirectly. If at some future date, exposure mileage information by driver were available, it is likely that the relationship between it and accident rates 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 s predictors in relation (4.11).

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 committed and so again we are cautioned against making inferences beyond the data.

Finally, it is erroneous to conclude that females and older people are more skillful drivers than other groups within the driving population because of their comparatively low accident rates in time. Rather, the superior driving records of these two groups probably can be accounted for by differences in driving exposure. For example, in a study dealing with accidents on main rural highways, Solomon (1964) reports that the accident rate in terms of number of involvements per

100 million miles was somewhat higher for females than for males. Also, his data indicates that older drivers, as well as younger drivers, have a higher involvement rate per mile than middle-aged drivers in both sexes. Again, relating to driving skill and caution, McFarland et al. (1964) conclude that middle-aged drivers are less likely to be at fault or to blame when involved in an accident than those in the other age groups.

4.6 Poisson Regression

Our objective in this section is to provide the theoretical framework for estimating the value of an individual's accident rate potential, λ, as a linear function of s criterion variables. Comparing (4.11) and (4.13) we see that

$$\lambda = E(N(t)/t) \quad . \tag{4.14}$$

For the sake of simplification, let t = 1 and denote N(1) by N. Then, according to (4.12), the probability that the jth individual in a sample will be involved in n_j accidents during the next unit of time is given by

$$p(n_j) = \frac{e^{-\sum_{i=o}^{s} \beta_i x_{ij}} \left(\sum_{i=o}^{s} \beta_i x_{ij}\right)^{n_j}}{n_j!} \quad , \tag{4.15}$$

$$n_j = 0, 1, 2, \ldots,$$

$$\sum_{i=o}^{s} \beta_i x_{ij} > 0, \quad x_{oj} = 1,$$

$$j = 1, 2, \ldots, k \; .$$

Since

$$\lambda_j = E(N_j) = \sum_{i=o}^{s} \beta_i x_{ij} = Var(N_j)$$

we find ourselves in a generalized linear regression setting. The departure from classical linear regression is that here the variance of N_j is a function of its expected value. In matrix notation,

(4.16) $\underline{\lambda} = E(\underline{N}) = X\underline{\beta}$ and $Cov(\underline{N}) = V$,

where the underline denotes a column vector of dimension k, X is a k x (s + 1) matrix having the values of the criterion variables as elements, and V is a k x k diagonal matrix with diagonal elements $v_j = \sum_{i=o}^{s} \beta_i x_{ij}$, j = 1, 2, ... k. The off-diagonal elements of V are taken to be zero on the assumption that the observed accident involvements are statistically independent.

Let us denote the transpose of matrix A by A' and its inverse by A^{-1}. It is well known (e.g., see Goldberger (1964)) that

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

is the minimum variance linear unbiased estimator of $\underline{\beta}$. The variance-covariance matrix of the vector $\hat{\underline{\beta}}$ is

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

If we wish to estimate $\lambda = E(N)$ for a given set of criterion values, $\underline{x}' = (1, x_1, x_2, \ldots, x_s)$, we may use the relation

(4.19) $\hat{\lambda} = \underline{x}'\hat{\underline{\beta}}$

for which

$$(4.20) \qquad E(\hat{\lambda}) = \underline{x}'\underline{\beta} \qquad \text{and} \qquad Var(\hat{\lambda}) = \underline{x}'(X'V^{-1}X)^{-1}\underline{x} \ .$$

Unfortunately, since $\underline{\beta}$ is unknown, the matrix V is unknown. In order to obtain an estimate of V, and hence an estimate of λ, the following iterative procedure is proposed. Let $\hat{V}_m$ denote the estimate of V obtained on the m^{th} iteration and 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 $\hat{V}_o$ be the k x k identity matrix and define

$$\hat{V}_{m+1} = \text{diag}\,[\underline{x}_1'\underline{b}_m, \ \underline{x}_2'\underline{b}_m, \ \ldots, \ \underline{x}_k'\underline{b}_m]$$

where $\underline{x}_j$ is the vector of criterion variables corresponding to the j^{th} individual. The iterations are continued until convergence is realized, i.e., until $\underline{b}_{m+1} = \underline{b}_m$. Denote this equality vector by $\underline{b}$. Then

$$(4.21) \qquad \underline{b} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\underline{n}$$

where $\hat{V}$ is the equality matrix $\hat{V}_{m+1} = \hat{V}_m$. As our final estimate of λ_j relation (4.19) dictates that we use

$$(4.22) \qquad \hat{\hat{\lambda}}_j = \underline{x}_j'\underline{b}$$

and from (4.20) we may use

$$(4.23) \qquad Var(\hat{\hat{\lambda}}_j) = \underline{x}_j'(X'\hat{V}^{-1}X)^{-1}\underline{x}_j$$

as an estimate of the variance of $\hat{\hat{\lambda}}_j$. Because of having to use $\hat{V}$ instead of V, the estimate (4.22) is not unbiased and its variance is unknown. However, in a related but slightly different context, Jorgenson (1961), points out that $\hat{\hat{\lambda}}$ is best asymptotically normal (BAN). 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.

In practice, the above iterative scheme can be carried out using a standard least squares linear regression program. The vector $\underline{b}_{m+1}$ can be calculated by applying a weight of

$$\left(\sum_{i=o}^{s} b_{i(m)} x_{ij}\right)^{-\frac{1}{2}}$$

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 s + 1 equations

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

Hence the vector $\underline{b}' = (b_o, b_1, \ldots, b_s)$, as defined in (4.21), is a solution set to the system

$$(4.24) \qquad \sum_{j=1}^{k} \frac{x_{ij} n_j}{\sum_{i=o}^{s} b_i x_{ij}} = \sum_{j=1}^{k} x_{ij} \ , \qquad i = 0, 1, \ldots, s \ .$$

In his 1961 article, Jorgenson showed that $\underline{b}$ is a solution to the maximum likelihood normal equations for estimating $\underline{\beta}$. From (4.15), the

likelihood function is

$$(4.25) \qquad L = \prod_{j=1}^{k} \frac{e^{-\sum_{i=o}^{s} \beta_i x_{ij}} \left(\sum_{i=o}^{s} \beta_i x_{ij} \right)^{n_j}}{n_j!} .$$

Taking the natural logarithm we obtain

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

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

$$\frac{\partial \ln L}{\partial \beta_i} = - \sum_{j=1}^{k} x_{ij} + \sum_{j=1}^{k} \frac{x_{ij} n_j}{\sum_{i=o}^{s} \beta_i x_{ij}} .$$

On setting the $s + 1$ partials equal to zero, the system of maximum likelihood normal equations obtained is

$$\sum_{j=1}^{k} \frac{x_{ij} n_j}{\sum_{i=o}^{s} \hat{\beta}_i x_{ij}} = \sum_{j=1}^{k} x_{ij} , \qquad i = 0, 1, 2, \ldots, s .$$

Notice that this system is identical to that of (4.24). Since (4.24) is a nonlinear system, it is quite possible that more than one solution set exists. Whether or not the weighted least squares solution described in this section provides the absolute maximum of the likelihood function (4.25) probably depends on the data.

For the purpose of making statistical inferences, work by Wald (1943) provides a theoretical basis for testing hypotheses when the sample size is large. Consider the hypothesis formulation

(4.26) $H_o: \; L\underline{\beta} = \underline{\gamma}$

where L is a known ℓ x (s + 1) matrix of rank $\ell \leq s + 1$ and $\underline{\gamma}$ is a specified vector of constants. Then the statistic

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

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

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

4.7 Regression Results

In this section we illustrate the use of the Poisson regression technique applied to the California data. Recall that in Section 4.5 we selected 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.

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

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

(4.28)

x_4 = number of countable convictions incurred during past experience period.

x_5 = number of accident involvements incurred during past experience period.

x_6 = number of noncountable convictions incurred during past experience period.

As applied to the California data, x_1, x_2, x_3 refer to 1963 conditions and x_4, x_5, x_6 refer to combined 1961 and 1962 counts. The variable n is, of course, the number of accident involvements sustained during 1963.

Initially unweighted 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 4.15. The "F" values are of a magnitude that indicates both regressions are highly significant. (Compare with $F_{.01} = 2.80$ with 6 and ∞ degrees of freedom under normal regression.) Since the b_i's are linear combinations of independent Poisson variables, the unusually large sample sizes in both analyses assure us that the t-tests for

$$H_o: \beta_i = 0 \qquad \text{vs} \qquad H_1: \beta_i \neq 0 , \qquad i = 1, 2, \ldots, 6,$$

are valid. Comparing with a critical t value of 1.96 at the .05 significant level, we notice that marital status is a non-significant 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.

On the basis of these preliminary analyses we select as our final regression models

Table 4.15 Unweighted regression of 1963 accidents on six criterion variables, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

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

$$(4.29) \qquad \lambda \equiv E(N) = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 \qquad \text{for males}$$

and

$$(4.30) \qquad \lambda \equiv E(N) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 \qquad \text{for females.}$$

In applying the procedure of Poisson regression described in the previous section, convergence to seven decimal places in the $\underline{b}$ vector was achieved on the fifth iteration of the computation for males. To obtain the same degree of convergence for the female regression function nine iterations were required. The results of this estimation technique are given in Table 4.16. So that the reader may get an idea about the order of magnitude of the estimate of λ and its variance, values of $\hat{\hat{\lambda}}$ and $\sqrt{\widehat{\text{Var}}(\hat{\hat{\lambda}})}$ are given in Table 4.17 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.

As we expect from the preliminary analyses displayed in Table 4.15, all criterion variables in models (4.29) and (4.30) are highly significant according to Wald's asymptotic test statistic (4.27). To illustrate, suppose we test the hypothesis

$$H_o: \quad \beta_5 = 0$$

in model (4.29). Referring to statement (4.26), here

$$L = (0,0,0,0,1,0) \qquad \text{and} \qquad \underline{\gamma} = 0 \ .$$

Hence from Table 4.16 we have

Table 4.16 Estimation function for accident rate potential and covariance matrix, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

Males

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

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

Females

$$\hat{\hat{\lambda}} = -0.0176 + 0.00646x_1 + 0.0209x_2 + 0.0196x_4 + 0.0205x_5$$

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

Table 4.17 Accident rate potential estimates and their standard deviations, 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67)

Sex	Traffic Density	Marital Status	Age	Ct. Conv. History	Accident History	No Ct. Conv. History	$\hat{\hat{\lambda}}$	$\sqrt{\widehat{\mathrm{Var}}(\hat{\hat{\lambda}})}$
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

$$W = \frac{(2.16 \times 10^{-2})^2}{0.0654 \times 10^{-4}} = 71.35 \ .$$

Comparing with 6.64, the .01 point of the $\chi^2(1)$ distribution, we conclude that $\beta_5 \neq 0$.

To close this chapter, some remarks on the interaction of criterion variables are in order. In Figure 4.1 we have a visual representation of interaction between age and marital status. Regression analyses were performed which included all of the two-factor interactions making a total of 21 "independent" variables in the equations. For these interaction models, the regression sums of squares were 114.2104 for males and 23.0605 for females. This compares with 108.4692 and 21.7867, respectively, obtained for the models assuming no interaction (Table 4.15). The additional sum of squares is significant in both cases. (Calculated "F" = 4.56 and 2.08 with 15 and ∞ degrees of freedom.) However, any gain in accuracy achieved through inclusion of interaction terms is counteracted by complexity and interpretation difficulties. To this writer, the objective of simplicity is more important than a minor increase in precision.

5. THE ANALYSIS OF ACCIDENT COSTS

5.1 Introduction

Recall that the random variable X in our model represents the cost incurred in an accident involvement by an individual driver. That is, if the individual is involved in a multi-car accident, X represents the cost applicable to that individual only and is not the cost of the accident in total. In Chapter 3 we assumed the cost distribution of a single accident involvement associated with an individual to be exponential, i.e.,

$$(3.15) \qquad G(x) = 1 - e^{-\theta x}, \qquad x \geq 0, \quad \theta > 0,$$

is assumed to be the distribution function of X. The parameter θ characterizes the driving conditions that influence the probable cost of an accident when it occurs. Under this concept, θ is a function of such factors as types of highways, speed limits, weather and road conditions, and number of occupants.

It is natural to expect individuals to differ in the potential consequences of an accident involvement. For example, a person who drives almost exclusively in a metropolitan area is likely to be involved in less costly accidents than one who does most of his driving on rural roads. Therefore we have assumed that θ is a value of a random variable Θ having d.f. $V(\theta)$ so that the compound exponential function

$$(3.19) \qquad w(x) = \int_0^\infty \theta e^{-\theta x}\, dV(\theta), \qquad x \geq 0, \quad \theta > 0,$$

represents the theoretical distribution of accident involvement costs within a population of drivers.

In this chapter we shall explore the sample results of the Illinois Motor Vehicle Accident Cost Study (Illinois Department of Public Works and Buildings, 1962) which provides us with an empirical counterpart to $w(x)$. We shall see how it supports our conjectures (3.15) and (3.19). Also we shall see how accident costs are related to the characteristics of the involved driver and to accident location. Finally, a suggested procedure for estimating the parameter θ will be given.

5.2 Distribution of Accident Involvement Costs

Billingsley and Jorgenson (1963) provide a summary of the findings of the 1958 Illinois Accident Cost Study which includes data on unreported involvements as well as those reported to traffic authorities. Since the counts in the California Driver Record Study (California Department of Motor Vehicles, 1964-67) are based on reported incidences only, it is imperative that we confine our attention to that phase of the Illinois study. Recall that involvement costs in the Illinois study are defined as those directly attributable to an accident which represent the use of resources that would have been available for other purposes had the accident not occurred. Some of the elements of these direct costs are given in Chapter 1.

In Figure 5.1 we find the histogram of reported passenger car accident involvement costs. This distribution is based upon a stratified sample of 332 fatal injury, 1,730 nonfatal injury, and

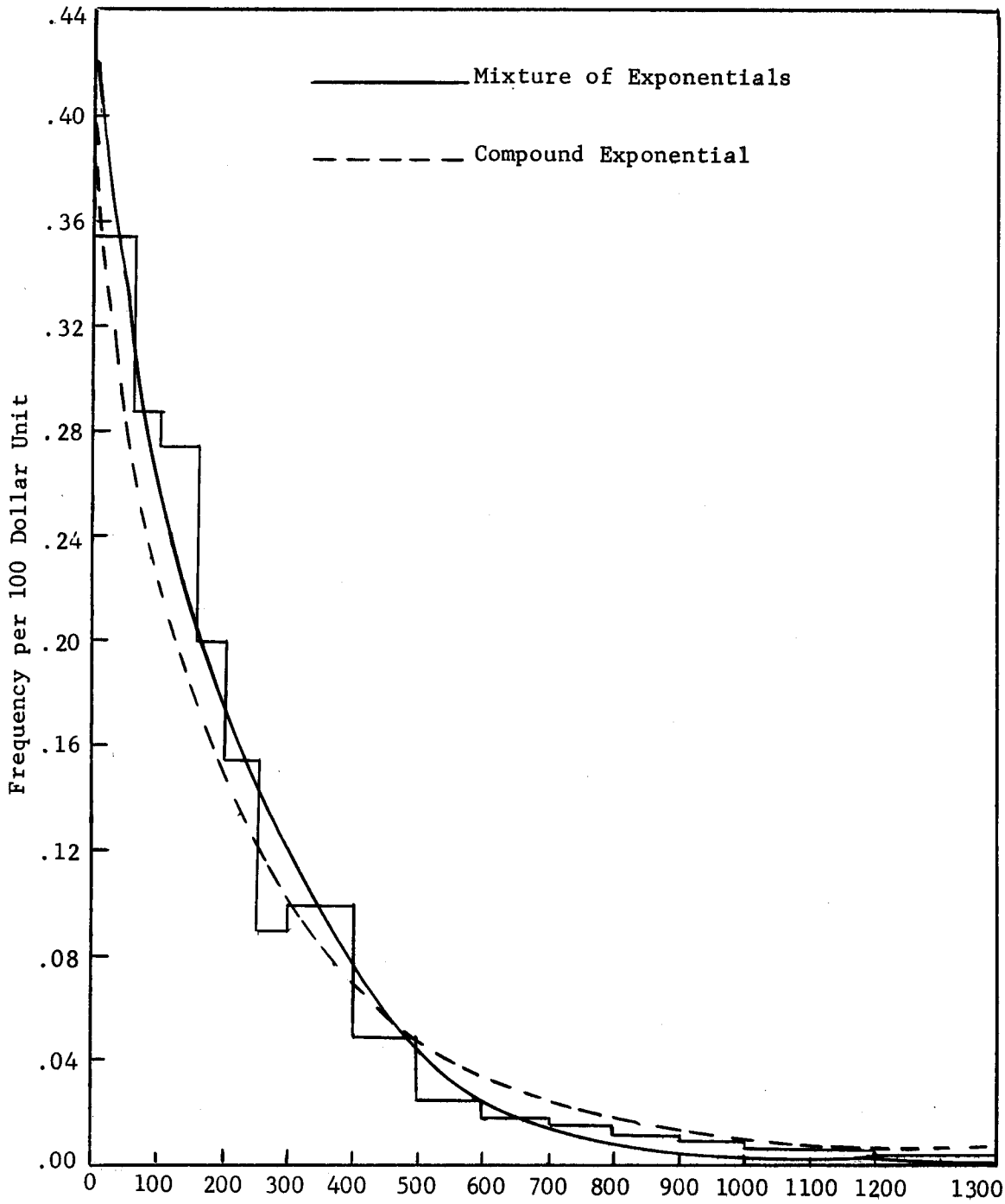

Figure 5.1 Empirical and theoretical cost distributions of a single accident involvement, 1958 Illinois Accident Cost Study (Billingsley and Jorgenson (1963))

816 property damage only cases. These 2,878 cases were then expanded to a "population" of 317,051 involvements by applying expansion factors to the sample cases in each of the three severity classes. The factors employed were 4.7, 54.4 and 260.2, respectively. Hence the constructed population of involvements is subject to bias through use of incorrect factors in addition to possible large sampling errors inherent in such a markedly skewed distribution. The skewness of this cost distribution is not only obvious from the visual representation in Figure 5.1 but is manifested by the descriptive statistics: a mean of 471 dollars, a variance of 3,760,963 and a median of 168. In short, the expanded distribution is a "dangerous" one and should be taken at somewhat less than "face value". However, it behooves us to accept the broad characteristics of this constructed distribution as indicative of the true distribution of accident involvement costs in Illinois during 1958. When evaluating costs on the same basis, an expanded distribution very similar in nature to this one was obtained in the 1966 Washington Area Motor Vehicle Accident Cost Study.

To arrive at a theoretical distribution of accident costs, 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,

$$(5.1) \qquad w(x) = \alpha\theta_1 e^{-\theta_1 x} + (1 - \alpha)\theta_2 e^{-\theta_2 x}, \quad x > 0, \quad \theta_1 > 0, \quad \theta_2 > 0, \quad 0 \leq \alpha \leq 1 .$$

Using the method of moments described by Rider (1961) for estimating

the three parameters, 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 Appendix 9.2. 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 5.1 and numerically in Table 5.1.

Table 5.1 Comparison of empirical and theoretical cost distributions, 1958 Illinois Accident Cost Study (Billingsley and Jorgenson, 1963)

Mixed Exponential: $W(x) = 1 - 0.9688e^{-\frac{x}{0.4928(470.6)}} - 0.03119e^{-\frac{x}{16.7557(470.6)}}$

Compound Exponential: $W(x) = 1 - (1 + x/529.4)^{-2.125}$

x	Empirical Cumulative	Mixed Exponential	Compound Exponential
50	0.1783	0.1864	0.1738
100	.3238	.3384	.3066
250	.6395	.6395	.5596
500	.8328	.8583	.7565
1000	.9229	.9595	.8950
2500	.9702	.9772	.9754
5000	.9874	.9835	.9932

Presumably, a better fit could be obtained by increasing the number of exponential distributions in the mixture. Hence it is

convenient to assume the existence of a continuous distribution function, $V(\theta)$, such that expression (3.19) is, in effect, the distribution which results when we mix together an infinite number of exponentials. Each exponential distribution in the mixture corresponds to a class of individuals in the population characterized by their particular θ. Knowledge of the exact form of the hypothesized $V(\theta)$ is not essential to this study. However, for illustration purposes only, suppose we choose

$$(5.2) \qquad V(\theta) = \frac{c^k}{\Gamma(k)} \int_0^\theta y^{k-1} e^{-cy} dy , \qquad \theta > 0, \quad c > 0, \quad k > 0 .$$

Then

$$(5.3) \qquad w(x) = \int_0^\infty \theta e^{-\theta x} dV(\theta)$$

$$= \frac{k}{c(1 + x/c)^{k+1}} , \qquad x \geq 0, \quad c > 0, \quad k > 0 .$$

The general characteristics of distribution (5.3) are those of the type in which we are interested, namely, J-shaped and highly skewed to the right. Using the method of moments, a fit of (5.3) to the Illinois data was made with the results shown in Figure 5.1 and Table 5.1. On comparing the two fits, we find that our particular choice for V failed to yield an improvement over the mixed exponential.

5.3 An Estimation Proposal

We have assumed that with respect to an individual driver, the random variable X has p.d.f.

(3.16) $g(x) = \theta e^{-\theta x}$, $x > 0$, $\theta > 0$,

with

(3.17) $E(X) = \theta^{-1}$ and $Var(X) = \theta^{-2}$.

The present problem is to find a procedure for estimating the parameter θ. Our first thought on this matter is to construct a model

$$E(X) = \theta^{-1} = h(\underline{z}; \underline{\gamma})$$

just as we did for the estimation of λ in the preceding chapter. Here, as before, $\underline{z}$ represents a vector of criterion variables which are used to predict the expected cost to the driver, given an accident, and $\underline{\gamma}$ denotes a parameter vector for the function h. It is evident, however, that this approach is not practical in terms of data currently available to us. Because of the unusually high variability in cost data, sample means have little reliability unless based upon a very large sample. For example, the Illinois study shows a range of costs per involvement from zero dollars to $136,800 and a coefficient of variation of 412 percent. As a consequence, the Illinois and Washington Area studies, based on 2,878 and 5,845 cases, respectively, give us "meaningless" information as respects to such refinements as costs by age and sex of driver. The difficulties arising from this variability in the data oblige us to turn to the crude but effective ratemaking procedures of the casualty insurance industry for an answer to our problem. These methods were reviewed briefly in Section 1 of Chapter 2.

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. That is, we would expect accidents occurring on four-lane controlled access highways to have a different average cost than those occurring in city 25 mile-per-hour zones, or those occurring on two-lane rural roads, or those occurring in downtown parking lots. As a case in point, the Illinois study shows the average involvement cost of an urban accident (one within an incorporate place) to be $396 as compared to an average of $931 for those 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 does most of his driving. In general, this is in the immediate vicinity of his residence. For example, the Washington Area study reports that 92.7 percent of the total costs of traffic accidents involving owners of motor vehicles registered in the Washington Metropolitan Area were suffered in occurrences within the boundaries defined by that area.

In view of these facts, basic to a solution of our estimation problem, is the establishment of involvement cost levels by area analogous to the territorial rate levels used in automobile insurance ratemaking. The area definitions need to reflect types of highways, population densities, speed limits, geographical and weather conditions, road safety conditions, etc., within the given areas. Once the areas have been established, we can proceed as follows. 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 θ. Denote this area θ by θ_a. Under this assumption an unbiased estimate for θ_a is given by

(5.4) $\quad (n - 1)/n\bar{x}$ with variance $\theta_a^2/(n - 2)$,

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. Since the derivations of this estimate and its variance are not generally found in textbooks on mathematical statistics, they are produced in Appendix 9.3.

A note of caution. As with the parameter λ, the parameter θ is not constant in time. The cost of having accidents is heavily influenced by prevailing medical, material and wage cost levels. Therefore, as in casualty ratemaking, it will be necessary to frequently update the estimates of the area θ's, and perhaps employ a trend factor when future costs are involved.

As mentioned previously, the Illinois and Washington samples are too small to shed any light on costs by driver characteristics. Nevertheless, two recent, published tabulations indicate that cost differences by sex and age are likely to exist. First, in a summary of 32,387 drivers involved in injury-producing passenger car accidents, Campbell (1966) shows that the distribution of accident types (struck object, head-on, sideswipe, etc.) differ by sex and by age; and it is well-known that some types of accidents are more costly than others. (See Illinois s t u d y (Billingsley and Jorgenson, 1963), and Washington s t u d y (Smith and Associates, 1966).) Second, in a New York Department of Motor Vehicles (1964) tabulation, 477,101 accident

involvements are classified 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 5.2 below. Using these distributions it is possible to arrive at an accident cost index by age and sex if we make assumptions about the relative costs of fatal, nonfatal, and property damage only involvements.

Table 5.2 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) (New York Department of Motor Vehicles, 1964)

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%

The Illinois and Washington Area studies indicate that average costs of fatal, nonfatal and property damage only involvements are in the approximate ratios 30 : 6 : 1. Applying these weights to the distributions in Table 5.2 and then converting the results to index form, a definite pattern emerges. The constructed involvement cost index is produced in Table 5.3. This would indicate that for both sexes, drivers in the 25 - 29 age category suffer greater loss on average when in an accident than those at other ages. For males, however, the difference between this age group and the 30 - 39 class is negligible. Somewhat surprising are the relatively low indices associated with the very young and the very old since their higher fatal accident rates have been widely publicized.

Table 5.3 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

After having obtained the area estimate, θ_a, we are able to assign a θ to each individual driver in that area by applying the appropriate involvement cost index. Denote this index by I and, as before, let $\mu = \theta^{-1}$. Then the estimate for an individual's μ is given by

(5.5) $$\hat{\mu} = \left(\frac{n\bar{x}}{n-1}\right) I$$

and

(5.6) $$\hat{\theta} = 1/\hat{\mu} \ .$$

Having given estimation procedures for the parameters involved in the distribution of the random variable X(t), we can now turn our attention to our overall model, F(x,t).

6. FULL MODEL DISTRIBUTIONS

6.1 Distributions Related to Individuals

We will begin this chapter with a brief review of the assumptions, concepts and mathematical results associated with the random variable X(t), the total cost of accident involvements incurred during a time interval of length t. If the parameters, λ and θ, characterizing an individual are known, then the distribution function of X(t) is given by

$$(3.2) \qquad F(x,t) = \sum_{n=o}^{\infty} p(n,t)\, G_n(x) \ , \qquad x \geq 0, \quad t > 0 \ ,$$

where

$$(3.6) \qquad p(n,t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \ , \qquad n = 0, 1, 2, \ldots,$$
$$\lambda > 0, \quad t > 0 \ ,$$

and

$$(3.21) \qquad G_n(x) = 1 - e^{-\theta x}\left(1 + \frac{\theta x}{1!} + \ldots + \frac{(\theta x)^{n-1}}{(n-1)!}\right) \ ,$$
$$n = 1, 2, 3, \ldots,$$
$$x \geq 0 \ ,$$

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

The mean and variance of X(t) are

$$(3.22) \qquad E(X(t)) = \lambda\mu t \qquad \text{and} \qquad Var(X(t)) = 2\lambda\mu^2 t$$

where $\mu = \theta^{-1}$.

The parameter λ is called the accident rate potential of the individual and the product $\lambda\mu$ is known as his accident cost potential.

Figure 6.1 presents a sketch of the probability density function of X(t) for t small given by the relations

$$f(x,t) = \begin{cases} \sum_{n=1}^{\infty} p(n,t)\, g_n(x) & , \quad x > 0, \quad t > 0 , \\ p(0,t) & , \quad x = 0, \quad t > 0 , \end{cases} \tag{6.1}$$

where p(n,t) has the form (3.6) and

$$g_n(x) = \frac{\theta^n x^{n-1}}{(n-1)!} \quad , \qquad n = 1, 2, 3, \ldots , \quad x > 0, \quad \theta > 0 . \tag{3.20}$$

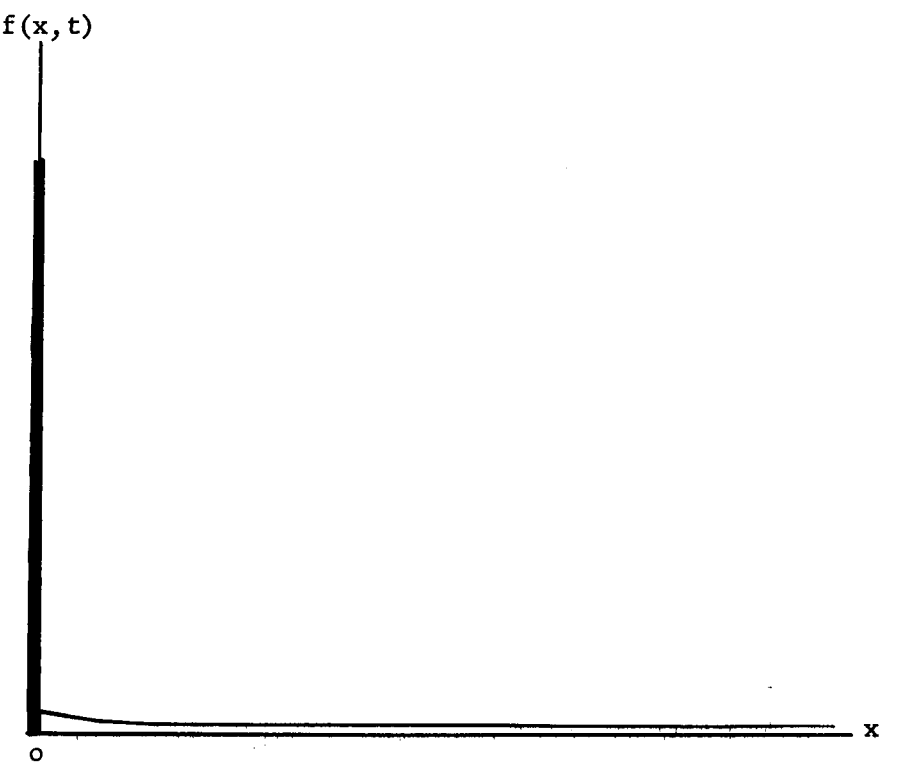

Figure 6.1 The probability density function of X(t) for small t

The graph illustrates that the random variable X(t) is neither discrete nor continuous but rather it 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

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

The sketch is exaggerated in that the plot of this continuous function actually is nearer the x-axis than it appears in Figure 6.1. 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.

Table 6.1 illustrates the effect on X(1) (t = 1 year) as we vary the accident cost potential by changing one parameter value at a time. The tabulated values correspond to $\Pr\{X(1) \leq x\}$; e.g., the probability that an individual characterized by $\lambda = .08$ and $\mu = 500$ will have total accident costs less than or equal $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 (4 percent) of the $\lambda = .04$ individuals to suffer accident loss during a year's time; yet during that time 85 percent of the individuals having an accident rate potential four times that of the first group will be cost-free.

Table 6.1 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	.9828	.9767
2,500	.9997	.9993	.9990	.9986
5,000	.9999+	.9999+	.9999+	.9999+
E(X(1))	20	40	60	80
$\sqrt{Var(X(1))}$	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	.8895	.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
$\sqrt{Var(X(1))}$	196	294	342	392

One of our impressions about Table 6.1 might be that differences between successive distributions are trivial. Suppose we find the cost distributions for k individuals having the same characteristic parameters. That is, let us investigate the distribution of

$$(3.10) \qquad S_k(t) = X_1(t) + X_2(t) + \ldots + X_k(t)$$

where the $X_i(t)$'s are independently and identically distributed. Since the d.f. of $S_k(t)$ is $F(x,kt)$ we know that the random variable $S_k(t)$ has the same d.f. as the random variable $X(kt)$, i.e.,

$$(6.2) \qquad S_k(t) \equiv X(kt) \ .$$

It follows that the mean and variance of $S_k(t)$ are

$$(3.11) \qquad E(S_k(t)) = k\lambda\mu t \qquad \text{and} \qquad Var(S_k(t)) = 2k\lambda\mu^2 t \ .$$

In Table 6.2 we find the distributions of $S_{100}(1)$ displayed for various combinations of λ and μ. Remember, 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. We note that the average cost per driver, i.e., the random variable $S_k(t)/k$ has mean and variance

$$(6.3) \qquad E(S_k(t)/k) = \lambda\mu t \qquad \text{and} \qquad Var(S_k(t)/k) = 2\lambda\mu^2 t/k \ .$$

Relative to the mean the standard deviation, $\sqrt{Var(S_k(t))/k}$, is still large when $k = 100$. However, possibilities for predictions about average costs for groups of individuals begin to emerge.

To carry this one step further, suppose we examine some distributions of $S_k(1)$ when $k = 1000$. As we vary the parameters the resulting $S_k(1)$ distributions are so divergent for a given set of total costs that a specific set is required for each $S_k(1)$ distribution in order to display something meaningful. This is done for three distributions

Table 6.2 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
$\sqrt{Var(S)}$	1,414	2,000	2,449	2,828
E(S/100)	20	40	60	80
$\sqrt{Var(S/100)}$	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	.5803
12,500	.9990	.9500	.8790	.7844
E(S)	4,800	7,200	8,400	9,600
$\sqrt{Var(S)}$	1,960	2,939	3,429	3,919
E(S/100)	48	72	84	96
$\sqrt{Var(S/100)}$	19.6	29.4	34.3	39.2

in Table 6.3, each evaluated at integral multiples of its standard deviation from its mean. Observe that the standard deviation of $S_k(1)/k$ for $k = 1000$ is small relative to the mean which implies predictions about the average accident costs of 1000 individuals can be made with some reliability. Also we find that the distributions of $S_k(t)$ for $k = 1000$ are tending toward normality in the sense of symmetry and areas over the $E(S_k(1)) \pm c\sqrt{Var(S_k(1))}$ intervals. This phenomenon, of course, is not surprising since it is a consequence of the central limit theorem as discussed in Section 4 of Chapter 3. Graphically, a $S_k(t)$ distribution for fixed t and large k might look something like that sketched in Figure 6.2. This compares with the decidedly asymmetrical distribution plotted in Figure 6.1 where $k = 1$. Here the discrete mass at $x = 0$ is negligible.

Although the accident cost potential, $\lambda\mu$, associated with an individual is indicative of his expected accident costs, it does not uniquely characterize him; i.e., the product $\lambda\mu$ does not specify the individual's $F(x,t)$ uniquely. This is obvious from (3.22). To demonstrate this fact, we find in Table 6.4 two pairs of group distributions having the same accident cost potential. Because their variances are unequal a marked difference in $F(x,t)$ distributions within pairs results.

6.2 An Example Relating to a Population

Suppose the random variable $X(t)$ represents the total cost of accident involvements incurred by an individual selected at random from a population of drivers. Then, in terms of previously introduced

Table 6.3 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

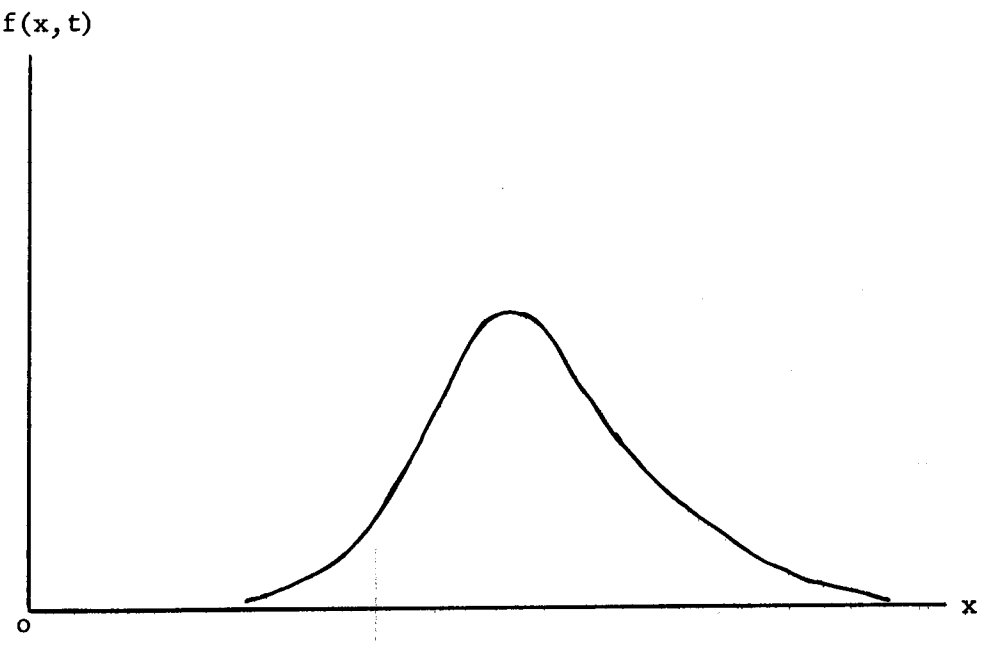

Figure 6.2 Example of a $S_k(t)$ distribution for large k

notation, the distribution function of X(t) is given by

$$F(x,t) = \sum_{n=o}^{\infty} q(n,t) W_n(x) \quad . \tag{6.4}$$

In this case the function F(x,t) is the distribution of cost experience within the population during an observation period of length t. Alternately, F(x,t) multiplied by 100 gives us the percentage of individuals incurring $\leq$ x dollars in total costs during t units of time. In Chapter 4 we observed that during a given observational period q(n,t) is a negative binomial distribution, i.e.,

Table 6.4 Distribution functions corresponding to the same accident cost potential

$\lambda\mu = 48$

x	F(x,100) $\lambda = .12,\ \mu = 400$	$\lambda = .08,\ \mu = 600$
0	0.0000	0.0003
2,000	.0538	.1036
4,000	.3748	.4144
5,000	.5803	.5822
6,000	.7503	.7229
8,000	.9352	.8988
10,000	.9880	.9698
12,500	.9990	.9947

$\lambda = .12,\ \mu = 400$		$\lambda = .08,\ \mu = 600$	
x	F(x,1000)	x	F(x,1000)
29,410[a]	0.0004	25,232[a]	0.0003
35,606	.0172	32,821	.0159
41,803	.1584	40,411	.1582
48,000	.5129	48,000	.5158
54,197	.8416	55,589	.8417
60,394	.9723	63,179	.9712
66,590	.9973	70,768	.9970

[a] F(x,1000) evaluated at integral multiples of the standard deviation from the mean 48,000.

$$q(n,t) = \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^n}{n!}\, dU(\lambda)$$

(4.4)

$$= \frac{\Gamma(n+r)}{n!\Gamma(r)} \left(\frac{r}{r+mt}\right)^r \left(\frac{mt}{r+mt}\right)^n, \qquad n = 0, 1, 2, \ldots,$$

$$r > 0,\ m > 0,\ t > 0,$$

where

$$U(\lambda) = \frac{1}{\Gamma(r)} \int_0^{\frac{r}{m}\lambda} y^{r-1} e^{-y}\, dy, \qquad \lambda > 0,\ r > 0,\ m > 0.$$

From (3.14) we can write down the distribution of accident cost potential within the population immediately. It is

$$\lim_{t\to\infty} F(zt,t) = \frac{1}{\Gamma(r)} \int_0^{\frac{r}{m\mu}z} y^{r-1} e^{-y}\, dy\,, \quad z>0,\ r>0,\ m>0,\ \mu>0\,, \tag{6.5}$$

which we recognize as the gamma distribution having mean and variance

$$E(Z) = m\mu \quad \text{and} \quad \mathrm{Var}(Z) = \frac{(m\mu)^2}{r}\,. \tag{6.6}$$

Here m is the mean number of accidents per individual per unit time and μ is the mean cost per accident involvement in the population.

In Chapter 5 we did not pursue our investigation to the point where we could reasonably assume an explicit mathematical form for W(x). Even if W(x) were known, however, it is quite likely that the convolution integrals, $W_n(x)$, would be intractable so that we would still be unable to evaluate F(x,t) exactly. In the event of intractability, approximation methods would have to be devised.

Suppose we agree that

$$W(x) = 1 - \alpha e^{-\theta_1 x} - (1-\alpha) e^{-\theta_2 x}, \qquad x \geq 0,\ \theta_1 > 0,\ \theta_2 > 0,\ 0 \leq \alpha \leq 1$$

is at least a first approximation to the d.f. of X. Then it is possible to provide an illustration of (6.4) for a particular F(x,t). The convolution integrals $W_n(x)$ for n = 0, 1, 2, 3, 4 have been worked out and appear in Appendix 9.4. If t = 1, then q(n,t) is

negligible for $n > 4$. Since $W_n(x)$ is bounded, i.e., $0 \leq W_n(x) \leq 1$ for all x and n, the evaluation of (6.4) can be achieved to four decimal places by calculating the first five terms of $F(x,t)$ for each x. The results of such a calculation are given in Table 6.5 for the specified parameter values. From this table we read that 95.45 percent of the drivers in the corresponding population will experience accident costs totaling less than or equal to \$100 in one year's time.

Table 6.5 Example of distribution of costs within a population of drivers

$r = 1.0000, \quad m = 0.07, \quad t = 1,$

$\mu = E(X) = 500, \quad Var(X) = 4,250,000, \quad Med(X) = 180$

x	F(x,1)
0	0.9346
50	.9455
100	.9545
250	.9733
500	.9884
1,000	.9967
2,500	.9984
5,000	.9988
10,000	.9993

7. SUMMARY AND SUGGESTIONS FOR RESEARCH

7.1 Summary

Assumption A: Accidents occur by chance and therefore are subject to probability laws.

Assumption B: The probability distributions characterizing individuals have a specific form but differ from driver to driver in parameter values.

Assumption C: The parameter values associated with an individual driver are functions of certain factors which (in themselves) are dependent on time.

Starting with these fundamental assumptions we have constructed a stochastic model which predicts in probabilistic terms the automobile experience of an individual driver. In so doing we have considered separately and in coalition three random variables: N(t), X, and X(t). The discrete random variable N(t) represents the number of accident involvements during a time period of t units; the continuous random variable X represents the cost of a single involvement; the mixed random variable X(t) represents the total cost of accident involvements over a time span of t units. If x is a particular cost value and $X_1, X_2, \ldots, X_n$ is a set of independently and identically distributed X variables, the relationship between these random variables is given by the analytic expression

$$F(x,t) = \sum_{n=o}^{\infty} p(n,t)\, G_n(x)$$

where

$$p(n,t) = \Pr\{N(t) = n\} ,$$

$$G_n(x) = \Pr\{X_1 + X_2 + \ldots + X_n \leq x\} ,$$

$$F(x,t) = \Pr\{X(t) \leq x\} .$$

In our analysis of N(t) we made liberal use of data collected on 148,000 drivers for the 1964 California Driver Record Study (California Department of Motor Vehicles, 1964-67). In harmony with many of the accident frequency distributions cited in the literature since 1920, we have seen that the number of accidents by driver within this sample during a fixed period of time has a negative binomial distribution. The underlying parameters, however, were seen to vary with time contrary to the historical concept of accident proneness. To verify a mathematical derivation for the genesis of the negative binomial, we were able to partition the California sample into Poisson groups on the basis of personal characteristics and driving record. This accomplishment confirmed our assumption that the incidence of accidents with respect to an individual follows a Poisson process, i.e.,

$$p(n,t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!} , \qquad n = 0, 1, 2, \ldots,$$

$$\lambda > 0, \quad t > 0 .$$

The parameter λ, known as the individual's accident rate potential, is his mean number of accidents per unit of time measured over an infinite time span. This single quantity uniquely determines the theoretical accident frequency distribution associated with the individual. From theoretical considerations, we expect the

distribution of accident rate potential within a population of drivers to be that of a gamma variable. To estimate λ we have proposed a least squares multiple Poisson regression procedure which gives us a solution compatible with the maximum likelihood system of equations. Using the California data, we took the opportunity to calculate estimates of λ and the variances of these estimates for hypothetical individuals having a specified set of values for six criteria: resident traffic density, marital status, age, countable conviction history, accident history, and noncountable conviction history. Up to the present time estimates of λ, as reported in research journals, have been based on accident history alone; we have found that, of the six criterion variables listed above, the most effective predictor of accidents is countable conviction history for both males and females. Because of different response patterns to some of the criterion variables, it was necessary to treat men and women drivers as belonging to two different populations. Also, we have observed that since the selected criterion variables vary over time, λ is a function of time through changing conditions and therefore it is necessary to periodically update this "constant." This means that an estimate for an individual's accident rate potential is valid for only a limited period of time.

Guided by the constructed distributions found in the Illinois and Washington Area motor vehicle cost studies (Billingsley and Jorgenson (1963), and Smith and Associates (1966)), we have assumed that the random variable X has an exponential distribution, i.e.,

$$G(x) = 1 - e^{-\theta x}, \qquad x \geq 0, \quad \theta > 0,$$

so that

$$G_n(x) = 1 - e^{-\theta x}\left(1 + \frac{\theta x}{1!} + \ldots + \frac{(\theta x)^{n-1}}{(n-1)!}\right),$$

$$n = 1, 2, 3, \ldots,$$

$$x \geq 0,$$

$$G_o(x) = \begin{cases} 1 & , \quad \text{when } x \geq 0, \\ 0 & , \quad \text{when } x < 0. \end{cases}$$

In order to obtain a reliable estimate for θ from data characterized by high variability, we have proposed a procedure analogous to that used by the casualty insurance industry which, by its very nature, is continually confronted with this problem. The method calls for the establishment of an overall cost level by area which in turn is adjusted to an individual basis by applying an appropriate index factor. As with λ, θ in reality is a function of time requiring frequent updating.

Since N(t) is a Poisson process variable, X(t) is a compound Poisson process variable. In our situation, X(t) is a mixed random variable with its probability density function having a discrete mass point at x = 0. We have seen that the sum of k independent and identically distributed X(t) variables is an X(kt) random variable having distribution function F(x,kt). This gives us the capability of studying homogeneous groups of drivers as well as individuals.

Denoting $\mu = E(X)$, the product $\lambda\mu$ is called the individual's accident cost potential and is defined as his mean accident cost per

unit of time evaluated over an infinite time span. Although the accident cost potential characterizes the individual by a single number, it is not unique in the sense that it does not uniquely determine the theoretical accident cost distribution associated with the individual. From theoretical considerations we infer that the distribution of accident cost potential within a population of drivers follows that of a gamma variable.

$F(x,t)$ and $F(x,kt)$ have been evaluated for specific values of λ, μ, t and k. From these illustrations we have seen the effect of varying one of these parameters at a time; we have seen how $X(kt)$ tends to that of a normal variable as k gets large; we have seen how similar $F(x,1)$ distributions appear, but how dissimilar $F(x,100)$ and $F(x,1000)$ distributions are for different values of λ and μ. Most significant of all, these illustrations display the overwhelming variability characteristic of accident distributions. Three of the practical consequences of this distinguishing trait are given below.

(1) The measure of an individual's accident cost potential, even if accurately estimated, does not guarantee some level of future performance. For example, a person of high potential may compile a better accident record than one of low potential over a given period of time; or, for that matter, even over their life-times.

(2) It is virtually impossible to give a meaningful prediction of accident costs with respect to an individual in terms of an absolute number.

(3) We can speak reliably of average accident costs only if k, the number of individuals in the group, is very large.

7.2 Suggestions for Future Research

Future research suggested by this study falls into two distinct categories: (i) the improvement of and extensions to the model developed in this dissertation, and (ii) the unsolved problems in statistical theory and methodology encountered in its development. First, the model itself.

Our model is "incomplete" in the sense that we have restricted our attention to reported rather than to all accident involvements. This limitation was imposed upon us by the available data. There is no doubt that if unreported accidents were included in our counts, the parameter values of the distributions would change significantly. It remains to be seen whether such an inclusion would alter the forms of the probability functions.

Again, because of data limitations, we were not able to include logical criterion variables in our estimation function for λ. Perhaps the most important omission from our function is the mileage exposure variable. Among other variables which might be significant in accident prediction are: number of years of driving experience, alcohol and drug consumption, educational level attained, and vocation.

An investigation should be made using various "before" and "after" time combinations in estimating the accident rate potential, λ. In the illustration of this study, expected accident rates for 7/8 of a year were calculated based upon a previous two-year driving record. The question might be asked, how much more reliable is an accident

rate estimate based upon a 3 years - 1 year, or a 3 years - 2 years, or a 5 years - 1 year experience combination?

We cannot claim the involvement cost indices constructed in this dissertation to be "the final figures". We can only say that our methods were able to induce a pattern which appears promising. Actually, our whole proposed procedure for estimating θ is rather crude, having been forced upon us by the variability problem. Can a more mathematically elegent and appealing procedure be found in the face of this variability?

Next we turn to unsolved problems for statistical research. More inquiry into multiple Poisson regression beyond that referred to and presented in this work is needed. Gart (1964) and Roberts and Coots (1965) give estimation and hypothesis testing procedures only for the simplest Poisson regression situation, namely,

$$p(n_j) = \frac{e^{-\beta x_j}(\beta x_j)^{n_j}}{n_j!}, \qquad n_j = 0, 1, 2, \ldots,$$
$$\beta x_j > 0, \quad j = 1, 2, \ldots, k .$$

However, complexity sets in as soon as another parameter is included in the model. A question yet to be answered: does the iterative scheme used in this study for obtaining estimates of the regression coefficients yield the maximum likelihood estimates?

The most promising theoretical distribution for involvement costs in a population of drivers is a mixture of k exponentials,

$$w(x) = \sum_{i=1}^{k} \alpha_i \theta_i e^{-\theta_i x}, \qquad x \geq 0, \quad \theta_i > 0, \quad \alpha_i \geq 0,$$
$$\sum_{i=1}^{k} \alpha_i = 1 .$$

In this study we had to be content with $k = 2$ because no reliable procedure for estimating the parameters is currently available when $k > 2$. It is true that we can always use the method of moments, but in doing so difficulty often arises in that the higher empirical moments are very unstable. As a result, one or more of the θ_i estimates may turn out to be negative. For us, the method of moments was inadequate even in the $k = 2$ case so that we equated medians in lieu of third moments.

A third statistical area for investigation suggested by this dissertation is a goodness-of-fit test when data is obtained through the use of a stratified sampling design. To the knowledge of this writer, no such test exists today.

8. LIST OF REFERENCES

Arbous, A. G. and J. E. Kerrich. 1951. Accident statistics and the concept of accident-proneness. Biometrics 7:340-432.

Bates, G. E. and J. Neyman. 1952. Contributions to the theory of accident proneness. University of California Publications in Statistics, I:215-276.

Billingsley, C. M. and D. P. Jorgenson. 1963. Analyses of direct costs and frequencies of Illinois motor-vehicle accidents, 1958. Public Roads 32:201-213.

Bliss, C. I. 1953. Fitting the negative binomial distribution to biological data. Biometrics 9:176-200.

Bohman, H. and F. Esscher. 1963. Studies in risk theory with numerical illustrations concerning distribution functions and stop loss premium, Part I. Skandinavisk Aktuarietidskrift 46: 173-225.

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

Campbell, B. J. 1966. Driver age and sex related to accident time and type. Traffic Research Review 10(2):36-43.

Cramér, H. 1937. Random Variables and Probability Distributions. Cambridge Tracts in Mathematics and Mathematical Physics, No. 36. Cambridge University Press, Cambridge.

Cramér, H. 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton.

Cramér, H. 1954. On some questions connected with mathematical risk. University of California Publications in Statistics, II: 99-125.

Cramér, H. 1955. Collective Risk Theory. Nordiska bokhandeln, Stockholm.

Cresswell, W. L. and P. Froggatt. 1963. The Causation of Bus Driver Accidents: An Epidemiological Study. Oxford University Press, London.

Dropkin, L. B. 1959. Some considerations on automobile rating systems utilizing individual driving records. Proceedings of the Casualty Actuarial Society XLVI:165-176.

Edwards, C. B. and J. Gurland. 1961. A class of distributions applicable to accidents. Journal of the American Statistical Association 56:503-517.

Feller, W. 1943. On a general class of "contagious" distributions. Annals of Mathematical Statistics 14:389-399.

Feller, W. 1949. On the theory of stochastic processes with particular reference to applications, pp. 403-432. In J. Neyman (ed.), Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley.

Feller, W. 1966. An Introduction to Probability Theory and Its Applications, Vol. II. John Wiley and Sons, Inc., New York.

Feller, W. 1968. An Introduction to Probability Theory and Its Applications, Vol. I (3rd Ed.). John Wiley and Sons, Inc., New York.

Gart, J. J. 1964. The analysis of Poisson regression with an application in virology. Biometrika 51:517-521.

Goldberger, A. S. 1964. Econometric Theory. John Wiley and Sons, Inc., New York.

Greenwood, M. and G. U. Yule. 1920. An unquiry 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 83:255-279.

Hãkkinen, S. 1958. Traffic accidents and driver characteristics. Finland's Institute of Technology, Scientific Researches No. 13, Helsinki.

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

Jorgenson, Dale W. 1961. Multiple regression analysis of a Poisson process. Journal of the American Statistical Association 56: 235-245.

Keeton, R. E. and J. O'Connell. 1965. Basic Protection for the Traffic Victim - A Blueprint for Reforming Automobile Insurance. Little, Brown and Company, Boston.

Leimkuhler, F. 1963. Trucking of Radioactive Materials: Safety vs Economy in Highway Transport. The John Hopkins Press, Baltimore.

Lundberg, O. 1940. On Random Processes and Their Application to Sickness and Health Statistics. Almqvist and Wiksells, Uppsala.

McFarland, R. A., G. S. Tune and A. T. Welford. 1964. On the driving of automobiles by older people. Journal of Gerontology 19:190-197.

Mellinger, G. D., D. L. Sylvester, W. R. Gaffey and D. I. Manheimer. 1965. A mathematical model with applications to a study of accident repeatedness among children. Journal of the American Statistical Association 60:1046-1059.

National Safety Countil. 1969. Accident Facts, 1969 Edition. Chicago.

New York Department of Motor Vehicles, State of. 1964. Fatal, non-fatal, and property damage accidents by age and sex, hour of day and day of week, 1963. Statistical Bulletin No. 6 (64). Albany.

Newbold, E. M. 1927. Practical applications to the statistics of repeated events, particularly to industrial accidents. Journal of the Royal Statistical Society 90:487-547.

Neyman, J. 1939. On a new class of contagious distributions applicable in entomology and bacteriology. Annals of Mathematical Statistics 10:35-37.

Parzen, E. 1962. Stochastic Processes. Holden-Day, Inc., San Francisco.

Pearson, K. 1957. Tables of the Incomplete Gamma Function. University Press, Cambridge.

Rider, P. R. 1961. The method of moments applied to a mixture of two exponential distributions. Annals of Mathematical Statistics 32:143-148.

Roberts, E. A. and G. G. Coots. 1965. The estimation of concentration of viruses and bacteria from dilution counts. Biometrics 21:600-615.

Shaw, L. and H. S. Sichel. 1961. The reduction of accidents in a transport company by the determination of the accident liability of individual drivers. Traffic Safety Research Review 5(4):2-12.

Sichel, H. S. 1951. The estimation of the parameters of a negative binomial distribution. Psychometrika 16:107-127.

Sichel, H. S. 1965. The statistical estimation of individual accident liability. Traffic Safety Research Review 9(1):8-15.

Smith, W. and Associates. 1966. A Report on the Washington Area Motor Vehicle Accident Cost Study. Clearinghouse, Springfield, Virginia.

Solomon, D. 1964. Accidents on Main Rural Highways Related to Speed, Driver, and Vehicle. United States Government Printing Office, Washington, D. C.

Stern, P. K. 1965. Ratemaking procedures for automobile liability insurance. Proceedings of the Casualty Actuarial Society LII: 139-202.

Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions American Mathematical Society 54:426-482.

Wilks, S. S. 1962. Mathematical Statistics. John Wiley and Sons, Inc., New York.

9. APPENDICES

9.1 Bivariate Compound Poisson Distribution

9.1.1 Derivation of Bivariate Negative Binomial as a Bivariate Compound Poisson

Suppose $N_1(t_1)$ and $N_2(t_2)$ are two independent Poisson random variables having probability density functions

$$p_1(n_1, t_1|\lambda) = \frac{e^{-\lambda t_1}(\lambda t_1)^{n_1}}{n_1!}, \qquad n_1 = 0, 1, 2, \ldots, \quad \lambda > 0, \quad t_1 > 0,$$

and

$$p_2(n_2, t_2|\lambda) = \frac{e^{-\lambda t_2}(\lambda t_2)^{n_2}}{n_2!}, \qquad n_2 = 0, 1, 2, \ldots, \quad \lambda > 0, \quad t_2 > 0.$$

Let the common parameter λ be a value of a random variable Λ having p.d.f.

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

Then

$$q_1(n_1, t_1) = \int_0^\infty p_1(n_1, t_1|\lambda)\, u(\lambda)\, d\lambda$$

$$= \frac{\Gamma(n_1+r)}{n_1!\,\Gamma(r)} \left(\frac{r}{r+mt_1}\right)^r \left(\frac{mt_1}{r+mt_1}\right)^{n_1},$$

$$n_1 = 0, 1, 2, \ldots,$$

$$r > 0, \; m > 0, \; t_1 > 0$$

and

$$q_2(n_2, t_2) = \int_0^\infty p_2(n_2, t_2|\lambda)\, u(\lambda)\, d\lambda$$

$$= \frac{\Gamma(n_2+r)}{n_2!\Gamma(r)} \left(\frac{r}{r+mt_2}\right)^r \left(\frac{mt_2}{r+mt_2}\right)^{n_2},$$

$$n_2 = 0, 1, 2, \ldots,$$

$$r > 0,\ m > 0,\ t_2 > 0.$$

We note that $q_1(n_1, t_1)$ and $q_2(n_2, t_2)$ are the p.d.f.'s of univariate negative binomial variables with

$$E(N_i(t_i)) = mt_i \quad \text{and} \quad Var(N_i(t_i)) = mt_i\left(1 + \frac{mt_i}{r}\right), \quad i = 1, 2.$$

The joint distribution of $N_1(t_1)$ and $N_2(t_2)$ is given by

$$p(n_1, t_1;\ n_2, t_2) = p_1(n_1, t_1|\lambda) \cdot p_2(n_2, t_2|\lambda)$$

$$= \frac{(\lambda t_1)^{n_1} (\lambda t_2)^{n_2} e^{-(t_1+t_2)\lambda}}{n_1! n_2!},$$

$$n_1 = 0, 1, 2, \ldots,$$

$$n_2 = 0, 1, 2, \ldots,$$

$$\lambda > 0,\ t_1 > 0,\ t_2 > 0.$$

Then the bivariate negative binomial distribution is defined as

$$q(n_1,t_1;\ n_2,t_2) = \int_0^\infty p(n_1,t_1;\ n_2,t_2|\lambda)u(\lambda)d\lambda$$

$$= \frac{\Gamma(n_1+n_2+r)}{n_1!n_2!\Gamma(r)} \left(\frac{r}{r+mt_1+mt_2}\right)^r \left(\frac{m}{r+mt_1+mt_2}\right)^{n_1+n_2} t_1^{n_1} t_2^{n_2},$$

$$n_1 = 0,1,2,\ldots,\ n_2 = 0,1,2,\ldots,$$

$$r > 0,\ m > 0,\ t_1 > 0,\ t_2 > 0.$$

9.1.2 Derivation of the Distribution for $N(t) = N_1(t_1) + N_2(t_2)$

Suppose $N_1(t_1)$ and $N_2(t_2)$ are two negative binomial variables having the p.d.f.'s $q_1(n_1,t_1)$ and $q_2(n_2,t_2)$ of Section 9.1.1, respectively. Let

$$N(t) = N_1(t_1) + N_2(t_2)$$

where t_1 and t_2 represent units of time over two non-overlapping time intervals such that $t = t_1 + t_2$. Let

$$n = n_1 + n_2$$

so that

$$n_1 = n - n_2 \qquad \text{where} \qquad 0 \leq n_2 \leq n .$$

Then from Section 9.1.1 we have as the joint distribution of $N(t)$ and $N_2(t)$,

$$q(n,n_2) = \frac{\Gamma(n+r)}{\Gamma(r)} \left(\frac{r}{r+mt_1+mt_2}\right)^r \left(\frac{m}{r+mt_1+mt_2}\right)^n \frac{t_1^{n-n_2} t_2^{n_2}}{(n-n_2)!\,n_2!} ,$$

$$n = 0,1,2,\ldots,\ 0 \le n_2 \le n ,$$

$$r > 0,\ m > 0,\ t > 0 .$$

The p.d.f. of N(t) is

$$q(n) = \sum_{n_2=o}^{n} q(n,n_2)$$

$$= \frac{\Gamma(n+r)}{n!\Gamma(r)} \left(\frac{r}{r+mt_1+mt_2}\right)^r \left(\frac{m}{r+mt_1+mt_2}\right)^n (t_1+t_2)^n$$

$$= \frac{\Gamma(n+r)}{n!\Gamma(r)} \left(\frac{r}{r+mt}\right)^r \left(\frac{mt}{r+mt}\right)^n , \qquad n = 0,\ 1,\ 2,\ \ldots ,$$

$$r > 0,\ m > 0,\ t > 0 .$$

We note that N(t) has a negative binomial distribution for which

$$E(N(t)) = mt \qquad \text{and} \qquad Var(N(t)) = mt\left(1 + \frac{mt}{r}\right) .$$

9.1.3 Derivation of Correlation Coefficient, ρ

The moment generating function of the joint distribution of the bivariate negative binomial distribution is

$$E(e^{\theta_1 N_1(t_1) + \theta_2 N_2(t_2)}) = \sum_{n_1=o}^{\infty} \sum_{n_2=o}^{\infty} e^{n_1\theta_1 + n_2\theta_2} q(n_1,t_1;\ n_2,t_2)$$

$$= \left(\frac{r}{r+mt_1+mt_2}\right)^r \sum_{n_1=o}^{\infty} \sum_{n_2=o}^{\infty} \frac{\Gamma(n_1+n_2+r)}{n_1!\,n_2!\,\Gamma(r)} \left(\frac{mt_1 e^{\theta_1}}{r+mt_1+mt_2}\right)^{n_1} \left(\frac{mt_2 e^{\theta_2}}{r+mt_1+mt_2}\right)^{n_2}$$

$$= \left(1 + \frac{mt_1}{r} + \frac{mt_2}{r} - \frac{mt_1 e^{\theta_1}}{r} - \frac{mt_2 e^{\theta_2}}{r}\right)^{-r} .$$

Hence

$$\frac{\partial^2 E(e^{\theta_1 N_1(t_1)+\theta_2 N_2(t_2)})}{\partial\theta_1\partial\theta_2}$$

$$= \frac{(r+1)}{r} mt_1 e^{\theta_1} mt_2 e^{\theta_2} \left(1 + \frac{mt_1}{r} + \frac{mt_2}{r} - \frac{mt_1 e^{\theta_1}}{r} - \frac{mt_2 e^{\theta_2}}{r}\right)^{-(r+2)} .$$

Setting $\theta_1 = 0$ and $\theta_2 = 0$ we get

$$E(N_1(t_1)\cdot N_2(t_2)) = \frac{(r+1)}{r} m^2 t_1 t_2 .$$

$$\therefore \text{Cov}(N_1(t_1), N_2(t_2)) = E(N_1(t_1)\cdot N_2(t_2)) - E(N_1(t_1))E(N_2(t_2))$$

$$= \frac{m^2 t_1 t_2}{r} .$$

We note that $\text{Cov}(N_1(t_1), N_2(t_2)) > 0$. Hence

$$\rho^2 = \frac{\text{Cov}^2(N_1(t_1), N_2(t_2))}{\text{Var}(N_1(t_1))\cdot \text{Var}(N_2(t_2))}$$

$$= \frac{m^2 t_1 t_2}{(r+mt_1)(r+mt_2)}$$

$$= \frac{1}{(1 + \frac{r}{mt_1})(1 + \frac{r}{mt_2})} .$$

We note that if $t = t_1 + t_2$ where t is fixed,

$$\max \rho = \frac{mt}{2r + mt}$$

is achieved when $t_1 = t_2 = \frac{1}{2}t$.

9.1.4 Derivation of the Conditional Distributions

(a) Conditional distribution of Λ.

$$u(\lambda|n_1,t_1) = \frac{p_1(n_1,t_1|\lambda)u(\lambda)}{\int_0^\infty p_1(n_1,t_1|\lambda)u(\lambda)\,d\lambda}$$

$$= \left(\frac{r}{m} + t_1\right)^{n_1+r} \frac{1}{\Gamma(n_1+r)} \lambda^{n_1+r-1} e^{-(r/m + t_1)\lambda},$$

$$n_1 = 0, 1, 2, \ldots,$$

$$\lambda > 0,\ r > 0,\ m > 0,\ t_1 > 0 .$$

We note that $u(\lambda|n_1,t_1)$ is a gamma p.d.f. so that $2(r/m + t_1)\lambda$ is a value of a chi-square variable with $2n_1 + 2r$ degrees of freedom.

(b) Conditional distribution of the negative binomial variable, $N_2(t_2)$.

$$q(n_2,t_2|n_1,t_1) = \frac{q(n_1,t_1;\ n_2,t_2)}{q_1(n_1,t_1)}$$

$$= \frac{\Gamma(n_1+n_2+r)}{n_2!\Gamma(n_1+r)} \left(\frac{r+mt_1}{r+mt_1+mt_2}\right)^{n_1+r} \left(\frac{mt_2}{r+mt_1+mt_2}\right)^{n_2},$$

$$n_1 = 0,1,2,\ldots,\ n_2 = 0,1,2,\ldots,$$

$$r > 0,\ m > 0,\ t_1 > 0,\ t_2 > 0.$$

We note that the conditional distribution of $N_2(t_2)$ given that $N_1(t_1) = n_1$ is that of a negative binomial distribution. Hence

$$E(N_2(t_2)\,|\,n_1,t_1) = \frac{mt_2(n_1+r)}{r+mt_1}$$

and

$$Var(N_2(t_2)\,|\,n_1,t_1) = \frac{mt_2(n_1+r)}{r+mt_1}\left(\frac{r+mt_1+mt_2}{r+mt_1}\right) .$$

(c) Special cases: (i) $n_1 = 0$, (ii) $n_1 > 0$.

(i) Suppose $n_1 = 0$, then

$$q(n_2,t_2\,|\,0,t_1) = \frac{\Gamma(n_2+r)}{n_2!\,\Gamma(r)}\left(\frac{r+mt_1}{r+mt_1+mt_2}\right)^r\left(\frac{mt_2}{r+mt_1+mt_2}\right)^{n_2} ,$$

$$n_2 = 0,\ 1,\ 2,\ \ldots,$$

$$r > 0,\ m > 0,\ t_1 > 0,\ t_2 > 0 .$$

Hence

$$E(N_2(t_2)\,|\,0,t_1) = \frac{rmt_2}{r+mt_1}$$

and

$$Var(N_2(t_2)\,|\,0,t_1) = \frac{rmt_2}{r+mt_1}\left(\frac{r+mt_1+mt_2}{r+mt_1}\right) .$$

(ii) Suppose $n_1 > 0$, then $p_1(>0,t_1\,|\,\lambda) = 1 - p_1(0,t_1\,|\,\lambda)$

and

$$u(\lambda|>0,t_1) = \frac{p_1(>0,t_1|\lambda)u(\lambda)}{\int_0^\infty p_1(>0,t_1|\lambda)u(\lambda)d\lambda}$$

$$\frac{u(\lambda) - (\frac{r}{r+mt_1})^r u(\lambda|0,t_1)}{1 - (\frac{r}{r+mt_1})^r}$$

$$\therefore\ q(n_2,t_2|>0,t_1) = \int_0^\infty p_2(n_2,t_2|\lambda)u(\lambda|>0,t_1)d\lambda$$

$$= \frac{q_2(n_2,t_2) - (\frac{r}{r+mt_1})^r q(n_2,t_2|0,t_1)}{1 - (\frac{r}{r+mt_1})^r} .$$

From this conditional distribution of $N_2(t_2)$ given $N_1(t_1) = n_1 > 0$ we get

$$E(N_2(t_2)|>0,t_1) = \frac{mt_2\{1 - (\frac{r}{r+mt_1})^{r+1}\}}{1 - (\frac{r}{r+mt_1})^r}$$

and

$$E(N_2^2(t_2)|>0,t_1) = \frac{A - (\frac{r}{r+mt_1})^r B}{D}$$

where

$$A = \frac{mt_2}{r}(r + mt_2 + rmt_2)$$

$$B = \frac{rmt_2}{(r+mt_1)^2}\,(r + mt_1 + mt_2 + rmt_2)$$

$$D = 1 - \left(\frac{r}{r+mt_1}\right)^r \quad .$$

Thus

$$\mathrm{Var}(N_2(t_2) \mid >0, t_1) = E(N_2^2(t_2) \mid >0, t_1) - E^2(N_2(t_2) \mid >0, t_1) \quad .$$

9.2 Estimating the Parameters of a Mixed Exponential Distribution

Suppose we wish to find estimates for the parameters of a mixed exponential distribution

$$(9.1) \qquad w(x) = \alpha\theta_1 e^{-\theta_1 x} + (1-\alpha)\theta_2 e^{-\theta_2 x}, \qquad x \geq 0, \quad \theta_1 \neq \theta_2 ,$$
$$\theta_1 > 0, \quad \theta_2 > 0 ,$$
$$0 < \alpha < 1 ,$$

where α, θ_1 and θ_2 are unknown. Denote the estimates of these parameters by

$$(9.2) \qquad \hat{\alpha} = a, \qquad \hat{\theta}_1 = \frac{1}{b_1} , \qquad \hat{\theta}_2 = \frac{1}{b_2} \quad .$$

Following is a modified method of moments procedure for obtaining values of a, b_1 and b_2 based upon a random sample from (9.1).

Let m_1, m_2 denote the first and second sample moments about zero and let p_k denote the k^{th} percentile point of the sample. Observe that

$$E(X) = \frac{\alpha}{\theta_1} + \frac{(1-\alpha)}{\theta_2} ,$$

$$E(X^2) = 2\left(\frac{\alpha}{\theta_1^2} + \frac{(1-\alpha)}{\theta_2^2}\right) ,$$

$$Pr\{X \leq x\} = 1 - \alpha e^{-\theta_1 x} - (1-\alpha) e^{-\theta_2 x} .$$

If we equate the theoretical values to the corresponding sample values we obtain the three equations:

$$(9.3) \qquad ab_1 + (1-a)b_2 = m_1$$

$$(9.4) \qquad ab_1^2 + (1-a)b_2^2 = \tfrac{1}{2}m_2$$

$$(9.5) \qquad ae^{-p_k/b_1} + (1-a)e^{-p_k/b_2} = 1 - k .$$

From (9.3) we find that

$$(9.6) \qquad a = \frac{m_1 - b_2}{b_1 - b_2} .$$

Substituting (9.6) into (9.2) leads to the equation

$$(m_1 - b_2)(b_1 + b_2) = \tfrac{1}{2}m_2 - b_2^2 .$$

Solving for b_1 we obtain

$$(9.7) \qquad b_1 = \frac{m_2 - 2m_1 b_2}{2(m_1 - b_2)} .$$

Next substituting (9.7) into (9.6) we get

$$(9.8) \qquad a = \frac{2(m_1 - b_2)^2}{m_2 - 4m_1 b_2 + 2b_2^2}$$

and

$$(9.9) \qquad 1 - a = \frac{m_2 - 2m_1^2}{m^2 - 4m_1 b_2 + 2b_2^2} \quad .$$

At this point we have a and b_1 expressed in terms of the first two sample moments and b_2. On substituting (9.7), (9.8) and (9.9) into (9.5), we obtain a functional equation in b_2 as follows.

$$(9.10) \qquad f(b_2) = 2(m_1-b_2)^2 e^{-\frac{2(m_1-b_2)p_k}{m_2-2m_1b_2}} + (m_2-2m_1^2)e^{-\frac{p_k}{b_2}}$$

$$- (m_2 - 4m_1 b_2 + 2b_2^2)(1 - k) = 0 \ .$$

The first derivative of $f(b_2)$ with respect to b_2 is

$$(9.11) \qquad f'(b_2) = -4(m_1 - b_2)e^{-\frac{2(m_1-b_2)p_k}{m_2-2m_1b_2}}$$

$$+ 4(m_1 - b_2)^2 \frac{(m_2-2m_1b_2-2m_1^2+2b_2)p_k e^{-\frac{2(m_1-b_2)p_k}{m_2-2m_1b_2}}}{(m_2 - 2m_1b_2)^2}$$

$$+ \frac{(m_2 - 2m_1^2)p_k e^{-\frac{p_k}{b_2}}}{b_2^2} + (4m_1 - 4b_2)(1 - k) \ .$$

We are now able to solve equation (9.10) by using Newton's method whereby successive approximations to b_2 are obtained by the recursion formula

$$b_2^{(i+1)} = b_2^{(i)} - \frac{f(b_2^{(i)})}{f'(b_2^{(i)})}, \qquad i = 0, 1, 2, \ldots,$$

where the superscript denotes the number of the iteration.

In particular, if $k = \frac{1}{2}$, then p_k is the sample median in which case (9.10) and (9.11) become

$$(9.12) \qquad f(b_2) = 4(m_1-b_2)^2 e^{-\frac{2(m_1-b_2)p_{\frac{1}{2}}}{m_2-2m_1b_2}} + 2(m_2 - 2m_1^2) e^{-\frac{p_{\frac{1}{2}}}{b_2}}$$

$$- m_2 + 4m_1b_2 - 2b_2^2 = 0$$

and

$$(9.13) \qquad f'(b_2) = -8(m_1 - b_2) e^{-\frac{2(m_1-b_2)p_{\frac{1}{2}}}{m_2-2m_1b_2}}$$

$$+ 8(m_1-b_2)^2 \frac{(m_2-2m_1b_2-2m_1^2+2b_2)p_{\frac{1}{2}} e^{-\frac{2(m_1-b_2)p_{\frac{1}{2}}}{m_2-2m_1b_2}}}{(m_2 - 2m_1b_2)^2}$$

$$+ \frac{2(m_2 - 2m_1^2)p_{\frac{1}{2}} e^{-\frac{p_{\frac{1}{2}}}{b_2}}}{b_2^2} + 4m_1 - 4b_2 .$$

Remarks: (i) According to (9.1) and (9.2), b_1 and b_2 are necessarily positive. If the data being fitted is such that a mixed exponential distribution is a "reasonable" assumption, then the equation

$$f(b_2) = 0 \qquad (9.10) \text{ or } (9.12)$$

will have exactly two positive roots. Depending on the starting value, Newton's iterative procedure will search out one or the other. It is immaterial whether the smaller or larger root is found first since (9.7) and (9.8) will then yield the second root and the proper value for $a = \hat{\alpha}$, respectively.

(ii) The reliability of the estimators is unknown. To provide some idea of their variances, Rider (1961) derives expressions for the asymptotic variances of b_1 and b_2 assuming α is known.

(iii) To facilitate computation, this writer applied the transformation $z = x/\bar{x}$ to the data, where x is a sample observation and $\bar{x} = m_1$ is the sample mean.

9.3 Estimators for the Parameter of the Exponential Distribution

9.3.1 Maximum Likelihood Estimator

If X is an exponential random variable having p.d.f.

$$g(x) = \theta e^{-\theta x}, \qquad x \geq 0, \quad \theta > 0,$$

it is well-known that the maximum likelihood estimator for θ is

$$\hat{\theta} = 1/\bar{X} \qquad \text{where} \qquad \bar{X} = \sum_{i=1}^{n} X_i/n .$$

To calculate the variance of $\hat{\theta}$, we note that the moment generating function of $\bar{X}$ is given by

$$M_{\bar{X}}(h) = (1 - \frac{h}{n\theta})^{-n} , \qquad \theta > h ,$$

so that $\bar{X}$ has a gamma distribution with parameters n and $1/n\theta$. Hence

$$f(\bar{x}) = \frac{(n\theta)^n}{\Gamma(n)} \bar{x}^{n-1} e^{-n\theta\bar{x}} , \qquad \bar{x} \geq 0 .$$

It follows that

$$E(\hat{\theta}) = E(\frac{1}{\bar{X}}) = \int_0^\infty \frac{1}{\bar{x}} f(\bar{x}) d\bar{x}$$

$$= \frac{n\theta}{n-1}$$

and

$$E(\hat{\theta}^2) = E(\frac{1}{\bar{X}^2}) = \int_0^\infty \frac{1}{\bar{x}^2} f(\bar{x}) d\bar{x}$$

$$= \frac{(n\theta)^2}{(n-1)(n-2)} .$$

Therefore

$$Var(\hat{\theta}) = E(\hat{\theta}^2) - E^2(\hat{\theta})$$

$$= \frac{(n\theta)^2}{(n-1)^2(n-2)} .$$

9.3.2 Unbiased Estimator

From the preceding development we immediately observe that

$$T = \frac{n-1}{n\bar{X}}$$

is an unbiased estimator for θ. Also

$$\mathrm{Var}(T) = \frac{(n-1)^2}{n^2}\,\mathrm{Var}\left(\frac{1}{\bar{X}}\right)$$

$$= \frac{\theta^2}{n-2} \quad .$$

Notice that $\mathrm{Var}(T) < \mathrm{Var}(\hat{\theta})$ for all $n > 2$.

9.4 Distribution Functions of Sums of Mixed Exponential Random Variables

9.4.1 Derivation of Distribution Functions

Suppose X_1 is a mixed exponential random variable having probability density function

$$w(x_1) = \alpha_1\theta_1 e^{-\theta_1 x_1} + \alpha_2\theta_2 e^{-\theta_2 x_1}, \qquad x_1 \geq 0, \quad \alpha_1 + \alpha_2 = 1 ,$$
$$\theta_1 > \theta_2 > 0 .$$

Then the moment generating function of X_1 is

$$M_{X_1}(h) = E(e^{hX_1}) = \int_0^\infty e^{hx_1}\, w(x_1)\,dx_1$$

$$= \alpha_1\left(1 - \frac{h}{\theta_1}\right)^{-1} + \alpha_2\left(1 - \frac{h}{\theta_2}\right)^{-1} , \qquad \theta_1 > \theta_2 > h .$$

Next let $X = X_1 + \ldots + X_n$ represent the sum of n independently and identically distributed mixed exponential variables. The m.g.f. of X is

$$M_X(h) = M_{X_1}^n(h) = \{\alpha_1(1 - \frac{h}{\theta_1})^{-1} + \alpha_2(1 - \frac{h}{\theta_2})^{-1}\}^n$$

$$= \alpha_1^n(1 - \frac{h}{\theta_1})^{-n} + \binom{n}{1}\alpha_1^{n-1}\alpha_2(1 - \frac{h}{\theta_1})^{-n+1}(1 - \frac{h}{\theta_2})^{-1}$$

$$+ \binom{n}{2}\alpha_1^{n-2}\alpha_2^2(1 - \frac{h}{\theta_1})^{-n+2}(1 - \frac{h}{\theta_2})^{-2} + \ldots + \alpha_2^n(1 - \frac{h}{\theta_2})^{-n}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} .$$

Now the expression

$$(1 - \frac{h}{\theta})^{-n} , \qquad \theta > h ,$$

where n is an integer ≥ 1 is the m.g.f. of a gamma random variable Y having p.d.f.

$$f_o(y;\theta) = \frac{\theta^n}{(n-1)!} y^{n-1} e^{-\theta y} , \qquad y \geq 0, \quad \theta > 0 .$$

The expression

$$(1 - \frac{h}{\theta_1})^{-n+h}(1 - \frac{h}{\theta_2})^{-k} , \qquad k = 1, 2, \ldots, n-1,$$

is the m.g.f. of the sum of two independent gamma variables, $Z = Y_1 + Y_2$, where the p.d.f.'s of Y_1 and Y_2 are

$$g_1(y_1;\theta_1) = \frac{\theta_1^{n-k}}{(n-k-1)!}\, y_1^{n-k-1}\, e^{-\theta_1 y_1}, \qquad y_1 \geq 0,\ \theta_1 > 0,$$

$$g_2(y_2;\theta_2) = \frac{\theta_2^{k}}{(k-1)!}\, y_2^{k-1}\, e^{-\theta_2 y_2}, \qquad y_2 \geq 0,\ \theta_2 > 0.$$

Hence the p.d.f. of $Z = Y_1 + Y_2$ is given by the convolution integral

$$f_k(z;\theta_1,\theta_2) = \int_0^z g_1(y_1)\, g_2(z-y_1)\, dy_1 .$$

It follows that the p.d.f. of $X = X_1 + X_2 + \ldots + X_n$ is

$$w_n(x) = \alpha_1^n f_o(x;\theta_1) + \binom{n}{1}\alpha_1^{n-1}\alpha_2 f_1(x;\theta_1,\theta_2)$$

$$+ \binom{n}{2}\alpha_1^{n-2}\alpha_2^2 f_2(x;\theta_1,\theta_2) + \ldots + \alpha_2^n f_o(x;\theta_2) .$$

By definition, the d.f. of X is

$$W_n(x) = \int_0^x w_n(s)\, ds \qquad \text{if } n = 1, 2, 3, \ldots,$$

$$= 1 \qquad \text{if } n = 0.$$

9.4.2 Distribution Functions, $W_n(x)$, for n = 0, 1, 2, 3, 4

Assume $\theta_1 > \theta_2 > 0$. Then

$$W_o(x) = 1$$

$$W_1(x) = 1 - \alpha_1 e^{-\theta_1 x} - \alpha_2 e^{-\theta_2 x}$$

$$W_2(x) = 1 - \alpha_1^2(1 + \theta_1 x)e^{-\theta_1 x} - \alpha_2^2(1 + \theta_2 x)e^{-\theta_2 x}$$

$$- \frac{2\alpha_1\alpha_2}{\theta_1 - \theta_2}(\theta_1 e^{-\theta_2 x} - \theta_2 e^{-\theta_1 x})$$

$$W_3(x) = 1 - \alpha_1^3(1 + \theta_1 x + \frac{\theta_1^2 x^2}{2})e^{-\theta_1 x}$$

$$- \alpha_2^3(1 + \theta_2 x + \frac{\theta_2^2 x^2}{2})e^{-\theta_2 x} - \frac{3\alpha_1^2\alpha_2\theta_1^2}{(\theta_1 - \theta_2)^2} e^{-\theta_2 x}$$

$$- \frac{3\alpha_1\alpha_2^2\theta_2^2}{(\theta_1 - \theta_2)^2} e^{-\theta_1 x} - \frac{3\alpha_1^2\alpha_2\theta_2^2}{(\theta_1 - \theta_2)^2}(1 + \theta_1 x)e^{-\theta_1 x}$$

$$+ \frac{3\alpha_1^2\alpha_2\theta_1\theta_2}{(\theta_1 - \theta_2)^2}(2 + \theta_1 x)e^{-\theta_1 x} - \frac{3\alpha_1\alpha_2^2\theta_1^2}{(\theta_1 - \theta_2)^2}(1 + \theta_2 x)e^{-\theta_2 x}$$

$$+ \frac{3\alpha_1\alpha_2^2\theta_1\theta_2}{(\theta_1 - \theta_2)^2}(2 + \theta_2 x)e^{-\theta_2 x}$$

$$W_4(x) = 1-\alpha_1^4(1+\theta_1 x+\frac{\theta_1^2x^2}{2}+\frac{\theta_1^3x^3}{6})e^{-\theta_1 x}-\alpha_2^4(1+\theta_2 x+\frac{\theta_2^2x^2}{2}+\frac{\theta_2^3x^3}{6})e^{-\theta_2 x}$$

$$-\frac{4\alpha_1^3\alpha_2\theta_1^3}{(\theta_1-\theta_2)^3}e^{-\theta_2 x}+\frac{4\alpha_1\alpha_2^3\theta_2^3}{(\theta_1-\theta_2)^3}e^{-\theta_1 x}+\frac{12\alpha_1^3\alpha_2\theta_1^2\theta_2}{(\theta_1-\theta_2)^3}e^{-\theta_1 x}$$

$$-\frac{12\alpha_1\alpha_2^3\theta_1\theta_2^2}{(\theta_1-\theta_2)^3}e^{-\theta_2 x}-\frac{12\alpha_1^3\alpha_2\theta_1\theta_2^2}{(\theta_1-\theta_2)^3}(1+\theta_1 x)e^{-\theta_1 x}$$

$$+\frac{12\alpha_1\alpha_2^3\theta_1^2\theta_2}{(\theta_1-\theta_2)^3}(1+\theta_2 x)e^{-\theta_2 x}+\frac{8\alpha_1^3\alpha_2\theta_1^3\theta_2 x}{(\theta_1-\theta_2)^3}e^{-\theta_1 x}$$

$$-\frac{8\alpha_1\alpha_2^3\theta_1\theta_2^3 x}{(\theta_1-\theta_2)^3}e^{-\theta_2 x}+\frac{4\alpha_1^3\alpha_2\theta_2^3}{(\theta_1-\theta_2)^3}(1+\theta_1 x)e^{-\theta_1 x}$$

$$-\frac{4\alpha_1\alpha_2^3\theta_1^3}{(\theta_1-\theta_2)^3}(1+\theta_2 x)e^{-\theta_2 x}+\frac{2\alpha_1^3\alpha_2\theta_1^2\theta_2 x^2}{\theta_1-\theta_2}e^{-\theta_1 x}$$

$$-\frac{2\alpha_1\alpha_2^3\theta_1\theta_2^2 x^2}{\theta_1-\theta_2}e^{-\theta_2 x}-\frac{18\alpha_1^2\alpha_2^2\theta_1\theta_2^2}{(\theta_1-\theta_2)^3}e^{-\theta_1 x}$$

$$+\frac{18\alpha_1^2\alpha_2^2\theta_1^2\theta_2}{(\theta_1-\theta_2)^3}e^{-\theta_2 x}-\frac{6\alpha_1^2\alpha_2^2\theta_1^2\theta_2^2 x}{(\theta_1-\theta_2)^3}e^{-\theta_1 x}$$

$$+\frac{6\alpha_1^2\alpha_2^2\theta_1^2\theta_2^2 x}{(\theta_1-\theta_2)^3}e^{-\theta_2 x}+\frac{6\alpha_1^2\alpha_2^2\theta_2^3}{(\theta_1-\theta_2)^3}(1+\theta_1 x)e^{-\theta_1 x}$$

$$-\frac{6\alpha_1^2\alpha_2^2\theta_1^3}{(\theta_1-\theta_2)^3}(1+\theta_2 x)e^{-\theta_2 x}\ .$$