Report Number 95-23

Research

Empirical Bayes Identification of High Hazard Locations for Older Drivers

UNIVERSITY OF MINNESOTA
CENTER FOR TRANSPORTATION STUDIES

Technical Report Documentation Page

1. Report No.	2.	3. Recipient's Accession No.
MN/RC - 95/23		

4. Title and Subtitle	5. Report Date
EMPIRICAL BAYES IDENTIFICATION OF HIGH HAZARD LOCATIONS FOR OLDER DRIVERS	October 1994
	6.

7. Author(s)	8. Performing Organization Report No.
Gary A. Davis, Ph.D.	

9. Performing Organization Name and Address	10. Project/Task/Work Unit No.
Department of Civil and Mineral Engineering University of Minnesota 500 Pillsbury Dr. SE Minneapolis, Mn. 55455	11. Contract (C) or Grant (G) No. (C) 68879 TOC #72

12. Sponsoring Organization Name and Address	13. Type of Report and Period Covered
Minnesota Department of Transportation 395 John Ireland Boulevard St.Paul Minnesota, 55155	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract (Limit: 200 words)

As part of an emphasis on improving road safety, the Minnesota Department of Transportation seeks to identify the locations where older drivers were over-represented in accident records. This research project reports on the use of three methods to help improve the accuracy of identifying locations where older drivers were at increased risk: a basic statistical model, the Empirical Bayes statistical method and a clustering method.

Overall, the basic statistical model preformed the best. The clustering method and the Empirical Bayes method could both be usefully applied to the traditional task of high-hazard identification--that of automatically screening a large number of accident sites to identify potential candidates for improvement. This information can point the way to areas that may require a more detailed engineering analysis.

17. Document Analysis/Descriptors		18. Availability Statement
Older drivers Empirical Bayes	High hazard locations Induced exposure	No restrictions. Document available from: National Technical Information Services, Springfield, Virginia 22161

19. Security Class (this report)	20. Security Class (this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	54	

EMPIRICAL BAYES IDENTIFICATION OF HIGH HAZARD LOCATIONS FOR OLDER DRIVERS

Final Report

Prepared by

Gary A. Davis, Ph.D.

Department of Civil and Mineral Engineering
University of Minnesota

October, 1994

Published by

Minnesota Department of Transportation
Office of Research Administration
200 Ford Building Mail Stop 330
117 University Avenue
St Paul Minnesota 55155

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

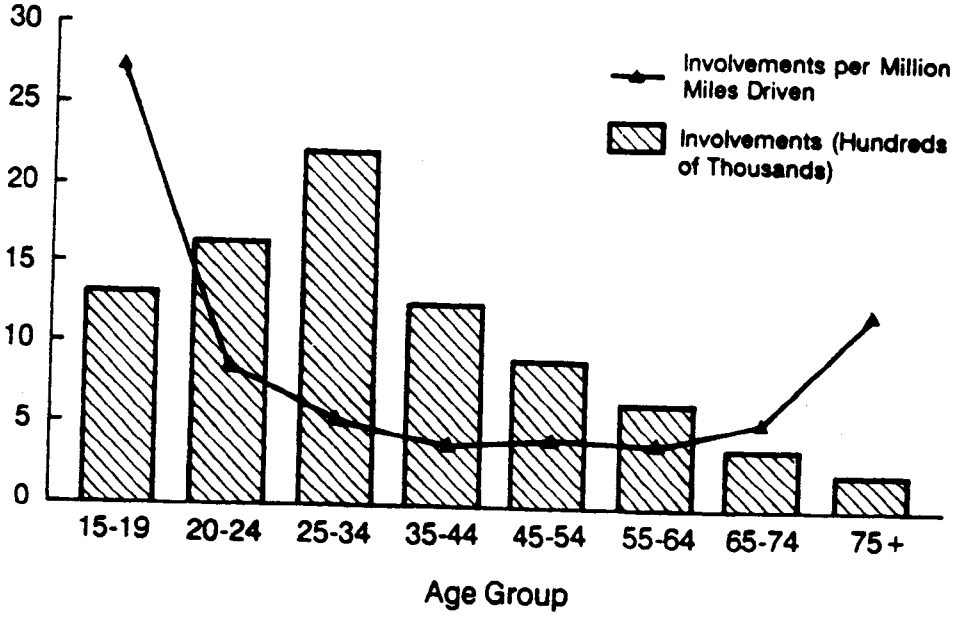
1.1 Introduction

It is a well-known demographic fact that individuals born in the year 1945-1960 constitute a substantial proportion of the population of the United States, while at the same time the life expectancy in the United States continues to increase. As these "baby-boomers" age, older drivers will come to comprise an increasingly significant proportion of individuals using the nation's roadway system. Much of current traffic engineering practice as developed over the last 30 years has tacitly assumed however, that older drivers constitute only a small fraction of the driving population. Thus decisions concerning default perception/reactions times, appropriate sight distances, and warrants for signalization or the provision of turning lanes tend to implicitly assume that only a few of affected drivers will be over age 60. As the population ages, this assumption becomes less tenable, and it is reasonable to ask first whether or not some aspects of traffic engineering practice need change, and second, whether or not it is worthwhile to implement changes in the way we manage our roadways, in anticipation of these demographic changes.

A number of issues relating to the transportation needs of the elderly have been addressed in Special Report 218 by the Transportation Research Board (1988). Figure 1.1 displays a graph from that report, showing accident involvement as a function of age, where accident involvement is estimated as the total number of accidents involving that age group divided by an estimate of the total vehicle-miles of travel by that age group. It can be seen that accident risk is high for younger divers, declines to roughly a constant level for ages 25-60 and then increases for older drivers. The study also emphasized that older drivers differ widely as to at what age their driving skills may be noticeably impaired, and that ready mobility is a necessary component to older persons' quality of life. Since the personal automobile is the dominant mode of transportation in most parts of country, and the only option in many areas, obviously any restrictions on driving must be based on actual impairment, and not on age per se. Generally, it is good for the society to permit older drivers to continue to use the roadway system, as long as they can do it safely.

This leads to questions concerning improvements of the roadway system that would encourage continued safe use by older drivers. Hauer (1988), in a detailed review of information on the safety of older persons, concluded that since older drivers are much more likely to be involved in accidents during the daytime, rather than the nighttime, and since a disproportionate number of older person fatalities occur at street intersections, improvements aimed at making intersections more safe for older drivers should receive a higher priority. He also hypothesized that older drivers would be likely to have greater difficulty with turning movements, especially left turns, although research available at that date was not adequate to answer this question.

Figure 1.1. Driver Involvements in Crashes and Involvement Rates by Age, 1983 (NHTSA and FHWA data) (Source: Special Report 218, TRB, 1988)



1.2 Activities in Minnesota

In 1989, the Minnesota DOT (MNDOT) began series of activities aimed at identifying and correcting safety deficiencies in the state's highway system. This involved several public meetings conducted throughout the state by high state officials, in which public comment concerning highway safety was solicited. One of the dominant concerns identified in this process was the accommodation of older drivers. This in turn led to a proposed target in reduction of highway fatalities for older persons, to be accomplished by a program of roadway improvements, driver education and improved transportation alternatives. Following on this, MNDOT became interested in identifying corridors where older drivers were over-represented in the accident records, and the PI for this project began working on this problem in the Spring of 1990. It soon became clear that although this problem could be viewed as similar to the problem of identifying high-risk locations, the lack information concerning the amount of travel done by older persons at specific locations made actual identification much more difficult. Library research by the project's research assistant, Konstantinos Koutsoukos, indicated that the induced exposure

method, which assesses group specific accident risk by comparing the fraction of drivers in a group which cause two-vehicle accidents to the fraction of drivers in that group which were innocent victims, could be used to sidestep this lack of information. It was also discovered that an Empirical Bayes (EB) statistical method could be used to improve the accuracy of the induced exposure estimates at individual sites, and preliminary tests conducted. This work is described in the report "A Statistical Method for Identifying Areas of High Crash Risk to Older Drivers," submitted to MNDOT in September, 1991.

Although the results of this research were promising, there remained a number of difficulties limited the usefulness of the method. First, although the method could, at least sometimes, identify locations where older drivers were at increased risk, the estimated quantities produced by the method remained difficult to interpret because their connection with the underlying accident model was unclear. Second, it was found that one of the underlying assumptions of the EB method was often violated by actual accident data sets. The preliminary research used an ad hoc method for correcting for this problem, but the robustness of this solution and its effect on the quality of safety decisions remained unclear. Finally, the computational procedures which implemented the method were not in a form that made them useful to practitioners.

To address these difficulties, and MNDOT sponsored an additional research project by this PI, and the results of that work are described in this report. In particular, further thought on the nature of the induced exposure model has led to a new statistic characterizing older driver risk, which can be interpreted as the ratio of the accident rate for older drivers to the accident rate of a comparison group of drivers, usually taken to be drivers aged 25-55. For aggregated data sets such as might be obtained for a highway corridor or a subarea within the metro region, this has led to an extremely simply computational procedure for testing that areas risk to older drivers. In order to provide more guidance as to which specific sites might be more dangerous to older drivers, a clustering procedure was developed which sorted the sites in a data set into two groups, and in each of the tests conducted here, one of these groups appeared to show higher risk to older drivers. Also the EB method developed in the first report was improved. Finally, all statistical procedures developed in this work have been implemented as MATHCAD interactive computational documents, which makes their actual use simple transparent to the user.

Chapter 2 reviews the statistical model common to most accident analysis, and its use in identification of high-hazard locations. This chapter also reviews the induced exposure method and its use in assessing the accident risk to older drivers. Chapter 3 describes the basic statistical model developed in this research, along with tests using three actual accident data sets. Chapter 4 describes the clustering method and its tests, while Chapter 5 describes the EB method. Chapter 6 summarizes the findings of this report and presents its conclusions.

CHAPTER 2

IDENTIFICATION OF HIGH-HAZARD LOCATIONS AND INDUCED EXPOSURE

2.1 High-Hazard Identification

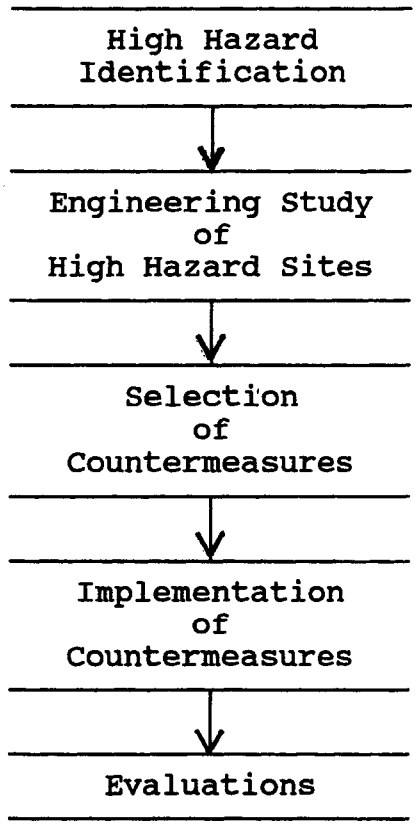
As noted in Chapter 1, the problem of identifying corridors where older drivers are "over-represented" in the accident records can be viewed as similar to the problem of identifying high-hazard locations, but where we are interested in the hazard a location poses to a specific subgroup of drivers, rather than to its hazard to the average driver. In this chapter, a brief review of the problem of identifying high hazard locations is given, with attention to recent Empirical Bayes solutions to this problem. An introduction to the induced exposure method is then given, and the chapter ends with a review of applications of the induced exposure method to the study of the accident risk of older drivers.

Figure 2.1 shows a simplified version of a process by which a safety engineer would identify and treat unsafe sites on a roadway system. Beginning with a large number of sites, a statistical screening process is used to identify those sites showing higher accident risk than is typical for that jurisdiction. A more detailed study of these selected sites, often done using collision diagrams, leads to an identification of the dominant accident type at each site, together with appropriate countermeasures. These countermeasures are then implemented, and ideally, additional study of the accident records is done to determine any resulting reduction in accident risk. Note that of these five steps, only the first and last are naturally accomplished via a statistical analysis of accident records. In principal there is no reason why a similar procedure could not be used to identify and treat sites showing risk to driver subgroups, but in practice this is very difficult, due to the way high-hazard locations are identified. To see the source of this difficulty and its remedy, it is necessary to first review the standard approach to accident analysis.

From a humane point of view, it is fortunate that traffic accidents are "rare events" compared to the amount of travel done by the populations which generates them, and the Poisson distribution provides the basic statistical model for much traffic accident analysis. More formally, if n_k denotes the actual number of accidents counted over some time period (typically one or more years) at a location numbered k, n_k is assumed to be a Poisson random variable with mean value equal given by $m_k = \lambda_k E_k$, where

λ_k = accident rate at location k,
E_k = exposure of the travelling population at location k.

Figure 2.1. Simplified Safety Improvement Procedure



The accident rate is thus a measure of the tendency of location k to produce accidents, with locations having higher accident rates being more dangerous. The exposure E_k is a measure of the size of the population at risk for accidents at locations k, and the standard measures of exposure used in traffic accident analysis are the total traffic count at a locations, used primarily for analysis of the accident risk at intersections, and vehicle-miles of travel, used for highway sections. Given a count of the number of accidents over a period of time at some location and exact knowledge of the exposure during the same period, the maximum likelihood estimator (ML) of the accident rate is simply

$$\hat{\lambda}_k = \frac{n_k}{E_k}$$

and when $\lambda_k E_k > 70$, the probability distribution of the ML estimator can be approximated by a normal distribution with mean equal to λ_k and variance equal to λ_k/E_k. For example, suppose that at an intersection, over one year, 85 accidents were observed, and total number of vehicles entering the intersection from all approaches over that year was 4.5 million vehicles. Then the estimated accident rate for this intersection would be

$$\hat{\lambda} = \frac{85}{4500000} = 1.89\times10^{-5}$$

and the standard error of estimate would be approximately

$$\hat{\sigma} = \sqrt{\frac{1.89\times10^{-5}}{4500000}} = 2.05\times10^{-6}$$

An approximate 95% confidence interval for this intersection's accident rate would then be

$$(\hat{\lambda}-1.96\hat{\sigma},\hat{\lambda}+1.96\hat{\sigma}) = (1.49 \times 10^{-5},2.29 \times 10^{-5})$$

or between 14.9 and 22.9 accidents per million entering vehicles (MEV).

The simplest method of identifying high-hazard locations would then be to simply rank the sites in terms of their estimated accident rates. In practice though the estimated accident rates can often be poor predictors of future accident counts due to failure to account for a statistical phenomenon called regression-to-the mean (RTM). In plain terms, RTM refers to the tendency of extreme random values to be followed by less extreme values, even when no change has occurred in the underlying mechanism generating these values. Since the variance of the ML estimator is inversely proportional to the exposure E_k, the RTM effect will be more pronounced for those locations with low exposures, and a hazard identification method based on a ranking of the estimated accident rates will tend to mix genuinely hazardous sites with locations whose extreme values are due to chance alone, leading to an overemphasis of the hazard at sites with lower exposures.

The most common procedure for sidestepping this difficulty has been the rate-quality control method (e.g. Zeeger, 1982), in which the sites in an analysis sample are assumed to have the same accident rate, except for a few, unknown "outliers". In a procedure analogous to that used to identify defective batches of product in quality control, the mean accident rate for all sites is used to compute a "critical rate" for each site, and if the observed accident rate for a site exceeds its critical rate, it is taken to be an outlier, i.e. a site whose accident rate is not typical, and therefor identified as a high-hazard location.

More recently, a direct attack on accounting for RTM bias has been though the use of Empirical Bayes (EB) statistical methods. The application of EB methods to traffic safety analysis has been an active area of research for the past 10 years, beginning with a paper by Hauer (1985), on correcting for RTM bias in before and after safety studies. Higle and Witkowski (1988) described use of EB estimates of accident rates to identity high hazard locations, and subsequently Pendleton et al. (1990) described a unified method for accomplishing both these tasks, together with a more efficient ML methods for parameter estimation. Hauer (1992) then described how a site's accident rate could be expressed as a function of factors such as traffic volume or geometric properties, and how an ad hoc EB version of regression analysis used to estimate these predictive functions. Christianson et al. (1992) subsequently extended the ML approach of Pendleton et al. (1990) to this problem, and software implementing their methods is currently under development for FHWA.

The EB approach begins with a statistical model which assumes that the accident rates λ_k for the individual sites making up a sample are generated as the outcomes of Gamma random variables with common mean λ and common variance λ/e. That is, the underlying accident rates are randomly distributed across the sites making up the sample. The actual accident counts are then generated by the Poisson mechanism described earlier. If one knows the values of the Gamma parameters λ and e, it can be shown that the Bayes estimates of the individual accident rates is then

$$\lambda^*_k = \left(\frac{E_k}{E_k+e}\right)\hat{\lambda}_k + \left(\frac{e}{E_k+e}\right)\lambda$$

where $\hat{\lambda}_k$ is the individual site ML estimate defined earlier. For those sites with high exposures (and hence lower variances for λ_k) the Bayes estimator tends to weight the ML estimate more heavily, while those sites with low exposure are "shrunk" more toward the mean rate for all sites. The parameter e measures the information concerning the individual accident rates which is contained in the entire sample. When e=0 the entire sample tells us nothing concerning the individual. so that $\lambda^*_k=\hat{\lambda}_k$, while e=∞ corresponds to the case $\lambda^*_k=\lambda$, (i.e. the accident rate at each location is equal to the sample mean). It can also be shown that for intermediate values of e the Bayes estimates tend to be closer to the true accident rates than the ML estimates, and expressions for the variances of the Bayes estimates can also be given. In most practical situations however, the values of λ and e will not be known, and also require estimation. The EB methods described by Higle and Witkowski and Hauer employ method of moments estimates of these hyperparameters, while Pendleton et al. employ the more efficient ML method.

2.2 Induced Exposure Methods

Returning now to the problem of identifying sites where older drivers show increased accident, it is clear that standard methods could be employed if one had available at each of the sites in a sample a count of the accidents involving older drivers together with a measure of the total number of entering vehicles, or total vehicle miles of travel, generated by older drivers.

Since most accident record systems record the ages of involved drivers in the accident report, accident counts for older driver are easy to obtain. A count of the number of older drivers entering an intersection, on the other hand, will be almost impossible to obtain completely, and even sampling strategies would involve stopping vehicles and asking drivers for their licenses. Thus although very aggregated measures of age-specific exposure may be available in national surveys, age-specific exposures broken down by corridor or individual site are for must part nonexistent. Similar difficulties arise when one seeks disaggregated exposures for driver gender (Lyles and Stamatiades, 1991) or vehicle type (Lyles, 1994).

Safety researchers have been aware of the difficulties surrounding disaggregate measures of exposure for at least 30 years, with Thorpe (1964) suggested a method for estimating "relative involvement" of driver subgroups without determining exposures. This "induced exposure" approach is based on the idea that in at least a majority of two-vehicle accidents one driver can be considered to have caused the accident (is "at fault"), while the second driver is an "innocent victim." At fault drivers have accidents according to the above described Poisson model, while the innocent victim is selected randomly from the other drivers using the site. In this way the number of older drivers appearing as innocent victims in the accident records for a particular site is roughly proportional to the exposure of older drivers at that site, while the number of older drivers appearing as at-fault is roughly proportional to the accident rates for older drivers. One can then imagine forming the ratio of the proportion of older drivers in listed as at-fault to the proportion of older drivers listed as innocent, to form an "involvement ratio," with an involvement ratio greater than 1.0 being evidence that older drivers are overrepresented in the accidents at a particular site.

Haight (1970, 1973) made a distinction between the situation where it is possible to establish which driver is guilty or innocent and the situation where no fault is indicated in the accident records. He referred to the latter case as induced exposure analysis, while the former case (where it is possible to identify the at-fault and innocent drivers) he called "quasi-induced exposure." In this report, both these cases will be called induced exposure methods, but the case where it is possible to identify the at-fault and innocent parties will be called a complete classification while the case where such identification is not possible will be called an incomplete classification. This is consistent with statistical practice in contingency table analysis, where the distinction is made between completely observed and incompletely observed tables. Although incomplete induced exposure received a burst of research interest in the early 1970's, (Cerelli, 1973; Koonstra, 1973) it has remained something of an academic curiosity, while in the past seven years complete induced exposure methods have been used in several studies investigating the accident risk of older drivers.

Maleck and Hummer (1987) computed involvement ratios for a range of driver age groups using a large number of accident reports from Michigan Interstate and trunk highways. Figure 2.2 reproduces a figure from their paper, where as noted earlier, a driver group with an involvement ratio greater than 1.0 is more likely to cause an accident than to be involved as an innocent victim. It can be seen that this figure reproduces the U-shaped relationship between accident risk and age shown by the national accident rate estimates in Figure 1.1. Particularly interesting is Figure 2.3, which shows the involvement ratios for left-turn accidents in Michigan. Whereas in Figure 2.1, significant over-representation does not appear until around age 65, for left-turn accidents the increase appeared much younger, at around age 45. In McKelvey et al.

(1988), additional analyses were done using a larger sample of Michigan accident records, and for the most part reproduced the U-shaped relationship between accident risk and driver age. Generally, age 65 seemed to be about the point where over-representation began. Finally Cooper (1990) has replicated McKelvey and Hummer's U-shaped relationship using accident records from British Columbia.

An extensive effort to use induced exposure methods to identify accident characteristics of older drivers at intersections, with an eye towards proposing possible counter-measures, was done by Garber and Srinivasan (1991a, 1991b) for the State of Virginia. Here again, the U-shaped relationship between accident risk and age was found, and overall, age 65 again appears to be the point where noticeable over-representation begins. With regard to angle accidents however, over-representation appears to begin somewhat earlier, definitely around age 60, and possibly earlier. Garber and Srinivasan (1991b) also attempted fitting linear regression equations with an the involvement ratio for older drivers as the dependent variable, and a number of measure relating to the intersection's traffic conditions and control features as dependent variables. Although their models cannot be considered to have identified factors which cause some intersections to have higher older driver involvements than do others, the most reliable predictors were variables related to the presence or absence of protected left-turn phases and left-turn lanes.

Overall, induced exposure methods based on the involvement ratio produce results consistent with findings using more traditional approaches. In particular, the U-shaped relationship between accident risk and age found using accident records, and survey data to estimate age-specific vehicle-miles of travel, is also found by the induced exposure method. The hypothesis that older drivers find left turns especially troublesome also seems to be confirmed by the induced exposure analyses. To date however, all practical applications of induced exposure have used very large samples, with total accident records in the tens or hundreds of thousands, collected over an entire state. Since the effect of sampling variability on the accuracy of estimated quantities declines as the sample size increases, in these cases sampling variability can be ignored, and valid conclusions can be reached without statistical analysis. For smaller spatial units however, such as corridors, sub-areas within an urban region, or individual intersections, the number of accident records will be very much smaller, and statistical methods will be essential for true effects from random noise. There currently seems to be some confusion in the profession concerning what statistical tools to employ in induced exposure analyses, but as the following chapters will show, a rigorous statement of the induced exposure hypothesis together with careful derivation of the consequences of this hypothesis leads naturally to a view that statistical inference concerning induced exposure models as a special case of contingency table analysis, and this in turn leads to methods for applying induced exposure analysis to corridors, for clustering sites within a corridor, or even, in special cases, to individual sites.

Figure 2.2. Relative Accident Involvement by Driver Age for Total and Injury Accidents (Source: Maleck & Hummer, 1987)

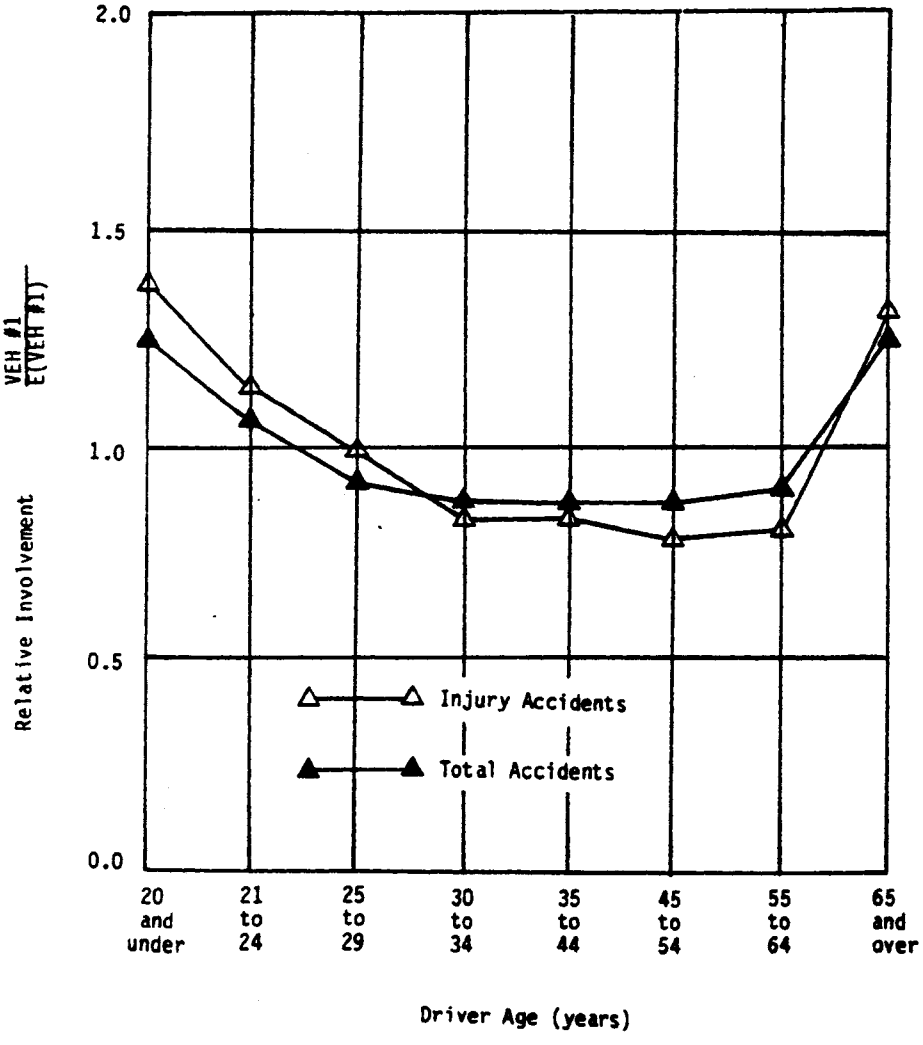
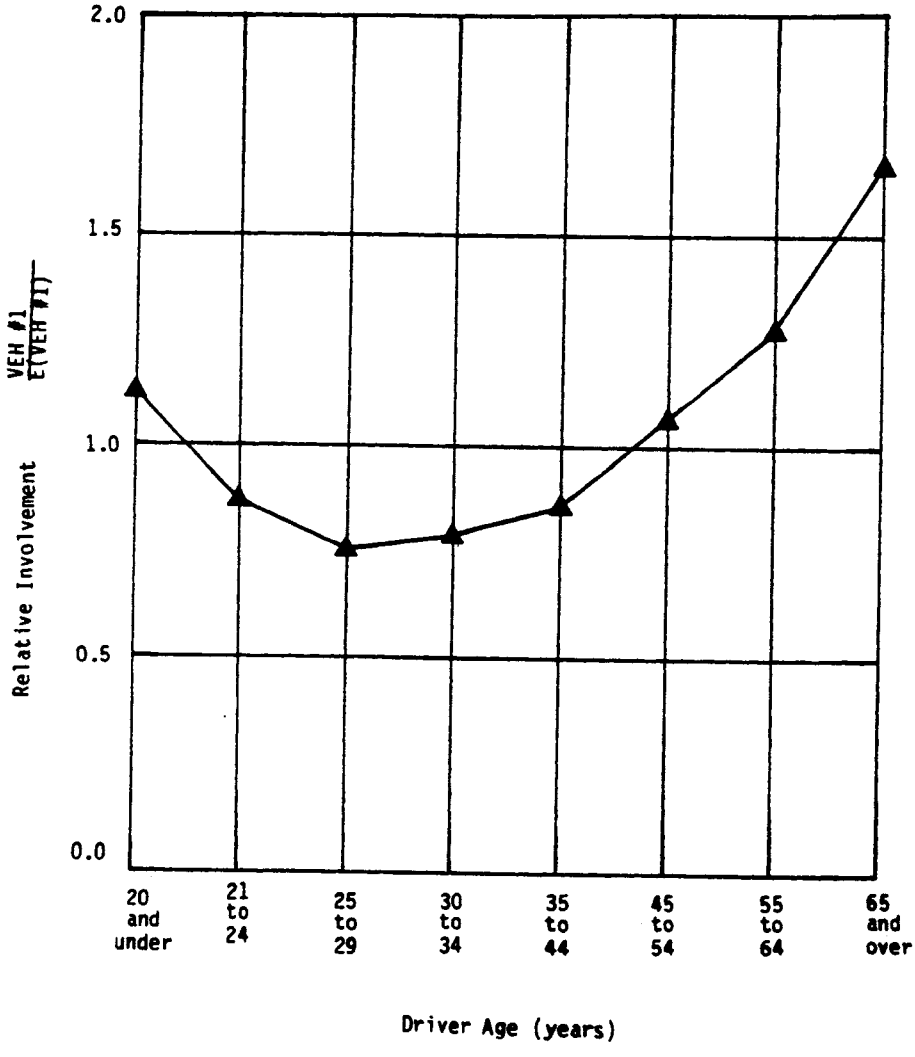


Figure 2.3. Relative Accident Involvement by Driver Age for Left-Turn Accidents (Source: Maleck & Hummer, 1987)



CHAPTER 3

INDUCED EXPOSURE AND CONTINGENCY TABLE ANALYSIS

3.1 Theoretical Development

In this chapter, it will be assumed that the traffic accident records are aggregated over a number of sites, but that the area of aggregation is considerably smaller than an entire state. The methods developed here will thus apply to analysis of urban regions, areas within an urban region, corridors, or relatively long segments of highway. Figures in the preceding chapter showed that the accident risk for drivers is roughly constant for ages 25-55, and then begins increasing at around age 60-65. This pattern suggests that the risk to older drivers can be assessed by comparing their accident rates to the rate for the 25-55 age group, which for want of a better term we call middle-aged. More formally, we will assume that the drivers are divided into two group, "middle-aged" and "older", so that the presence of relatively high-risk younger drivers are not biasing our conclusions. Let n_i (i=1,2) denote the number of accidents involving driver group i over some time interval, with i=1 denoting the older driver group, and i=2 denoting the middle-aged group. Applying the Poisson model described earlier, each n_i is assumed to be the outcome of Poisson random variable, with mean $\lambda_i E_i$, where

λ_i = accident rate for driver group i,
E_i = exposure for driver group i.

As before, if the exposure values E_i are known exactly, the maximum likelihood estimates of the accident rates λ_i are given by

$$\hat{\lambda}_i = \frac{n_i}{E_i} \tag{3.1}$$

and assessment of the relative risk to drivers in each group could be based on these estimated accident rates. But as noted earlier, group-specific measures of exposure are difficult to estimate reliably, so to implement an induced exposure approach, it is first assumed that in a majority of two-vehicle accidents, one driver can be considered to have caused the accident, while the other is assumed to be an innocent victim. The at-fault drivers are assumed to have accidents according the Poisson accident model while the group of the victim is assumed to be selected randomly, with probability of selection being directly proportional to the group's exposure. comparison subgroup. Next, define the following quantities

$r = E_1/(E_1+E_2)$, the probability the victim is an older driver,
$p = \lambda_1 E_1/(\lambda_1 E_1+\lambda_2 E_2)$, the probability the at-fault driver is an older driver,
n_{ij} = number of accidents for which the at-fault driver came from group i while the

victim came from subgroup j,

and taking $n = \Sigma_i\Sigma_j n_{ij}$ as fixed, it then follows that the n_{ij} are outcomes of a multinomial random vector with number of "trials" equal to n. The n_{ij} can be thought of as entries into a cross-tabulation table, where two-vehicle accidents are classified according to the group membership of the at-fault and victim drivers, as illustrated in Table 3.1.

Note in Table 3.1 that the probability a given two-vehicle accident falls in a cell is simply the product of the corresponding row and column marginal probabilities, so that an induced exposure table has the property of statistical independence between its row and column classifications. This structure is a consequence of the assumption that the subgroup of the victim is selected randomly, and the standard tests of independence provide methods for identifying data sets for which this assumption is not valid. For instance, it is well-known (e.g. Agresti, 1991 that under the hypothesis of independence, the log cross-product ratio statistic

$$\hat{\theta} = \log_e\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right) \tag{3.2}$$

has, for large values of the sample size n, approximately a normal distribution with a mean of zero, and a variance which can be estimated by

$$\hat{\sigma}^2_{\theta} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \tag{3.3}$$

Tests for random selection of victims can then be conducted using the standard normal distribution.

Assuming now that the data in an induced-exposure table satisfy the assumption of random selection of victims, our focus turns to the problem of making inferences concerning the accident rates λ_i. To simplify some of the following notation, define the marginal totals

$x = n_{11}+n_{12}$, the number of accidents in which an older driver is at fault,
$y = n_{11}+n_{21}$, the number of accidents in which an older driver is the innocent victim,

and a straightforward application of the maximum likelihood principle yields the maximum likelihood estimators (MLE) for p and r,

$$\hat{p} = \frac{x}{n}$$
$$\hat{r} = \frac{y}{n} \qquad (3.4)$$

TABLE 3.1

EXAMPLE 2x2 INDUCED EXPOSURE TABLE

		Innocent Victim 1	2	
	1	$E[n_{11}]=np_1r_1$	$E[n_{12}]=np_1(1-r_1)$	np_1
At Fault				
Driver	**2**	$E[n_{21}]=n(1-p_1)r_1$	$E[n_{22}]=n(1-p_1)(1-r_1)$	$n(1-p_1)$
		nr_1	$n(1-r_1)$	

Unfortunately, since $\Sigma_i p_i = \Sigma_j r_j = 1$, the probability distribution of the induced exposure table is completely characterized by the three parameters p,r and n, making it impossible to uniquely identify the four parameters λ_1, λ_2, E_1 and E_2. Note however that

$$\frac{p(1-r)}{(1-p)r} = \frac{\lambda_1}{\lambda_2} \qquad (3.5)$$

and defining a log rate-ratio parameter as

$$\Delta = \log_e\left(\frac{\lambda_1}{\lambda_2}\right) \qquad (3.6)$$

it is straightforward to verify that the MLE of Δ can be computed via

$$\hat{\Delta} = \log_e\left(\frac{x(n-y)}{(n-x)y}\right) \tag{3.7}$$

while an application of the delta method yields that for large n, the distribution of Δ is approximately normal, with mean equal to Δ and variance which can be estimated via

$$\hat{\sigma}^2_{\Delta} = \frac{1}{x} + \frac{1}{n-y} + \frac{1}{n-x} + \frac{1}{y} \tag{3.8}$$

This provides a method for testing hypotheses concerning Δ, and for constructing approximate confidence intervals for the rate-ratio λ_1/λ_2. Justifications for the above claims can be found in the Appendix.

Thus, although it is not possible to estimate the individual group accident rates λ_1 and λ_2 from the data in an induced exposure table, it is possible to estimate the ratio of these accident rates, with the interpretation that if the estimated ratio is significantly greater than 1.0, this is evidence that the older drivers have a higher overall accident rate than do middle-aged drivers, for the data in our sample. Correspondingly, if the logarithm of this estimated rate ratio is significantly greater than 0, we again have evidence that the older drivers have significantly higher accident rates. The logarithm of the rate ratio is used here because the normal approximation is generally better for the logarithm of the cross-product ratio statistics. (Agresti, 1990). By means of this normal approximation, the tabulated values for the standard normal distribution, found in most statistics texts, can be used to determine if an estimated log rate-ratio is significantly different from zero.

3.2 Example Applications

The statistical methods developed in this chapter and in the following two will be illustrated using three actual sets of traffic accident data. The first data set consists of the records of traffic accidents occurring at 33 signalized intersections on MNTH 47, running from the intersection of MNTH 47 with 40th Ave NE, in Columbia Heights, to the intersection of MNTH 47 with CSAH 22, in Burns Township. The records in this set were for the years 1988 and 1989. The second data set consists of records for traffic accidents occurring at 29 signalized intersections on MNTH 65, running from the intersection of MNTH 65 with 40th Ave NE, to the intersection of MNTH 65 with Viking Blvd NE, in East Bethel. Both these data sets were obtained from MNDOT. The third data set consists of traffic accident records for 61 signalized intersections along France, Penn, and Portland Avenues, and E. 66th Street, in Edina, Richfield, and Bloomington. These data were provided by Hennepin County. The raw accident records were

processed to identify two-vehicle accidents in which one driver at a least one contributing factor cited by the investigating officer, and the second driver had "no improper driving" cited. The driver with the contributing factor was then identified at the at-fault driver, the other was the innocent victim. These record were then processed again to identify the ages of the at-fault and innocent drivers, with drivers aged 25-55 being classed as "middle-aged" while drivers older than 55 were classed as older, to form induced exposure tables for each intersection. A listing of the intersections and induced exposure data for each data set is given in Appendix 1.

To carry out the computations described above, a MATHCAD 3.0 computational document was constructed, which takes as its input a file of site-specific induced exposure tables, and then automatically computes the log cross-product ratio, the log rate-ratio, their standard errors, and the associated z-statistics. On an IBM-type microcomputer with a 386 SL coprocessor running at 25 MHz, these calculations were essentially instantaneous.

Table 3.2 displays the aggregated induced exposure tables for MNTH 47 and MNTH 65. Checking first to see if the assumption of random victim selection is tenable, for MNTH 47, $\hat{\theta}$=-0.419, z=-0.93, p>.34 while for MNTH 65 $\hat{\theta}$=-0.378, z=-1.08, p>.28. The assumption of random victim selection appears tenable on both highways, so testing for whether or not older drivers have higher accident rates than do middle-aged drivers, the MNTH 47 data gives $\hat{\Delta}$=0.2, z=.84, p>.20, while for MNTH $\hat{\Delta}$=0.28, z=1.50, p<.07. Thus the data from MNTH 47 show no clear evidence for increased accident risk to older drivers, but the data from MNTH 65 give a somewhat tentative suggestion that older drivers have higher accident rates.

Table 3.3 displays the aggregated induced exposure table for the Hennepin County intersections. Testing for random selection of victims, we obtain $\hat{\theta}$=-0.172, z=-0.636, p>0.26, so that the assumption of independent victim selection is also tenable here. Testing for whether or not the older drivers have higher accident rates gives $\hat{\Delta}$=0.82, z=4.94, p<.001, so for this set of intersections there is very strong evidence that older drivers are at increased accident risk.

To summarize, this chapter described two statistical methods. One tests whether or not the age group of the victim in an induced exposure table an be assumed to have been randomly selected and then, given that this is true, the second tests whether or not older drivers have a significantly higher accident rate than do middle-aged drivers. The methods are appropriate for accident data sets created by aggregating over a number of individual sites, but where the number of accident records is considerably smaller than in the state-wide data sets used in earlier studies. In the three example data sets considered here, MNTH 47 showed no evidence of increased accident risk for older drivers, MNTH 65 showed a rather weak indication that older drivers may have higher accident rates and the Hennepin County data showed a very clear indications that older drivers had higher accident rates. If one were programming safety improvements targeted at older drivers, clearly the Hennepin County sites would be more promising than the two north metro highways. However, the methods described here still give no guidance as to which sites might be the more dangerous. In the next chapter, these methods will be extended to handle where the accident sites could be divided into two (or more) subsets, which may differ as to the accident risk they pose to older drivers, but accident rates for the subsets are unknown, as well as the appropriate assignment of sites to subsets.

TABLE 3.2

INDUCED EXPOSURE TABLES FROM TWO MINNESOTA HIGHWAYS

MNTH 47

Innocent Victim

		Middle-Aged	Older
	Middle-Aged	131	34
At Fault			
Driver	Older	41	7

MNTH 65

Innocent Victim

		Middle-Aged	Older
	Middle-Aged	202	52
At Fault			
Driver	Older	68	12

TABLE 3.3

INDUCED EXPOSURE TABLES FROM HENNEPIN COUNTY

		Innocent Victim	
		Middle-Aged	Older
At Fault Driver	Middle-Aged	194	51
	Older	113	25

CHAPTER 4

CLUSTERING SITES WITH RESPECT TO OLDER DRIVER RISK

4.1 Theoretical Development

The statistical model presented in the preceding chapter assumes that the values of the two probability parameters p and r are the same for all sites in the sample, and hence have the same ratio of older driver accident rate to middle-aged driver accident rate. In may be however that a safety engineer could have reason to suspect that the sites in the sample vary as to the risk they pose to older drivers, but has little idea as to which sites these are, or to the accident rate ratios which characterize them. In many cases, it is sufficient to treat the sites as belonging to one of a finite number of homogeneous classes, and the problem becomes one of first sorting the sites into their appropriate classes and then estimating the accident risk posed to older drivers by sites in each class. The sample of sites is then a mixture of sites from each of the classes, and the class identification and estimation problem can be solved using mixture likelihood clustering methods (Titterington, et al., 1986). Thus we treat each site as belonging to one of a finite set of classes, with the induced exposure tables for the sites within a class having the same p_k and r_k values, but with sites in different classes having different values.

In this report, we will only treat the simplest case where it is assumed that the sites a are divided into two classes. An unknown proportion α of the sites in a sample belong to Class 1, while the remaining 1-α sites belong to Class 2. The methods presented here extend readily to problems involving more than two classes, although the computer time needed to estimate the model parameters increases as a function of the number of classes. Given that a site is in Class 1, its induced exposure table is assumed to be generated independently of those of the other sites, as multinomial outcomes with parameters p_1 and r_1, while the tables for those sites in Class 2 are independent multinomial outcomes with parameter p_2 and r_2. This produces a type of statistical model known as a discrete mixture model (Redner and Walker, 1984).

If we knew the parameters α, p_1,r_1,p_2,r_2, but nothing more about the sites, our best estimate of the probability any given site belonged to Class 1 would be α, and the probability it belonged to Class 2 would be 1-α. When we observe and induced exposure table for a site, we obtain information concerning the appropriate Class for that site, and the problem at hand is then how to best update our estimate of the probability a site belongs to Class 1, given the data in the induced exposure table. Using Bayes Theorem, it is can be verified that this updated classification probability is given by

$$\hat{\delta}_k = Prob[site\ k \in Class\ 1 \mid \alpha, p_1, r_1, p_2, r_2, n_k, x_k, y_k]$$

$$= \frac{\alpha p_1^{x_k}(1-p_1)^{n_k-x_k} r_1^{y_k}(1-r_1)^{n_k-y_k}}{\alpha p_1^{x_k}(1-p_1)^{n_k-x_k} r_1^{y_k}(1-r_1)^{n_k-y_k} + (1-\alpha) p_2^{x_k}(1-p_2)^{n_k-x_k} r_2^{y_k}(1-r_2)^{n_k-y_k}} \qquad (4.1)$$

where

x_k = number of two-vehicle accidents with an older driver as the at-fault party, occurring at site k,

y_k = number of two-vehicle accidents with an older driver as the innocent party, occurring at site k,

n_k = total two-vehicle accidents at site k.

In most applications however, the values of the parameters α, p_1,r_1,p_2,r_2, will be unknown, but a practical solution results by simply computing ML estimates of these parameters and substituting the ML estimates into (4.1). Given that we have a total of N sites in our sample, the likelihood function for the induced exposure tables in the sample is

$$L(\alpha,p_1,r_1,p_2,r_2) = \prod_{k=1}^{N} \frac{n_k!}{n_{11,k}!n_{12,k}!n_{21,k}!n_{22,k}!}\left(\alpha p_1^{x_k}(1-p_1)^{n_k-x_k}r_1^{y_k}(1-r_1)^{n_k-y_k} + (1-\alpha)p_2^{x_k}(1-p_2)^{n_k-x_k}r_2^{y_k}(1-r_2)^{n_k-y_k}\right) \tag{4.2}$$

and following Redner and Walker (1984) the an iterative procedure known as the EM algorithm (Dempster et al., 1977) can be used to compute ML estimates of α, p_1, r_1, p_2, r_2. Letting i index the iterations of the EM algorithm, this leads to a recursive calculations of the form

$$\begin{bmatrix} \alpha^{i+1} \\ p_1^{i+1} \\ r_1^{i+1} \\ p_2^{i+1} \\ r_2^{i+1} \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_k \hat{\delta}_k^i}{N} \\ \dfrac{\sum_k x_k \hat{\delta}_k^i}{\sum_k n_k \hat{\delta}_k^i} \\ \dfrac{\sum_k y_k \hat{\delta}_k^i}{\sum_k n_k \hat{\delta}_k^i} \\ \dfrac{\sum_k x_k (1-\hat{\delta}_k^i)}{\sum_k n_k (1-\hat{\delta}_k^i)} \\ \dfrac{\sum_k y_k (1-\hat{\delta}_k^i)}{\sum_k n_k (1-\hat{\delta}_k^i)} \end{bmatrix} \tag{4.3}$$

where Δki is obtained by substituting the iteration i estimates for α, p1, r1, p2, r2 into equation (4.1). In practice, the EM has good global convergence properties, but a very slow rate of convergence. In the implementation developed for this project, a MATHCAD 3.0 computational document uses the EM algorithm for about 50 iterations, and then switches to MATHCAD's internal algorithm to speed up the convergence to a solution of likelihood equations. For the three example data sets considered here, computation of the ML estimates takes around 30 minutes for the MNTH 47 and MNTH 65 data, and about 1 hour for the Hennepin county data, on an IBM-type microcomputer using a 386SL chip at 25 MHZ.

As noted earlier, we seek to first assign sites to their appropriate classes, and then to determine whether the sites in a class show increased accident risk to older drivers. Letting $[\hat{\alpha},\hat{p}_1,\hat{r}_1,\hat{p}_2,\hat{r}_2]$ denote the ML estimates found above, estimates of the probability a site belongs to Class 1 are obtained by substituting the ML estimates into equation (4.1). For computing estimates of the log rate ratio statistics for the two classes, we have two alternatives. Since the site classifications are subject to some degree of uncertainty, the ML estimates for the log rate ratio statistics for the two classes are given by

$$\hat{\Delta}_1 = \log\left[\frac{\hat{p}_1(1-\hat{r}_1)}{(1-\hat{p}_1)\hat{r}_1}\right]$$
$$\hat{\Delta}_2 = \log\left[\frac{\hat{p}_2(1-\hat{r}_2)}{(1-\hat{p}_2)\hat{r}_2}\right] \qquad (4.4)$$

and in Appendix 2 it is shown these estimated log rate ratios are consistent and asymptotically normally distributed, with an expression is given for estimating their covariance matrix. This in turn provides a method for using the standard normal distribution to test whether or not the estimated log rate ratio for a subgroup is significantly different from zero.

4.2 Example Application

The two class clustering method was applied to each of the three data sets described in Chapter 3, and the results of the estimation are displayed in Tables 4.1, 4.2, and 4.3. Looking at the results for MNTH 47, displayed in Table 4.1, be seen that in Class 1 $\hat{p}_1=.241$ and $\hat{r}_1=.104$, suggesting the possibility of over-representation of older drivers at the sites in this class, while for Class 2 $\hat{p}_2=.218$ and $\hat{r}_2=.235$, suggesting that no over-representation is present here. When testing whether or not the log rate ratios in the two classes are significantly different from zero, both tests turn up nonsignificant. Turning to the results for the MNTH 65 data, displayed in Table 4.2, a similar pattern is seen, with Class 1 suggesting the possibility of over-representation of older drivers, but neither class's log rate ratio being significantly different from zero. Finally the result for the Hennepin county data are displayed in Table 4.3. These results suggest that the two classes differ primarily in relative numbers of older drivers making up their

driving populations, i.e. in Class 1 the mix of older drivers and middle-aged drivers is about 22% versus 78%, while in Class 2 the mix is about 8% to 92%. Both classes show estimated log rate-ratios that are greater than zero, suggesting that the accident rates for older drivers is higher in both classes. This interpretation is complicated by the relative rarity of Class 2 sites, which means the parameters for these sites are more difficult to estimate with precision. Thus the standard errors in Table 4.3 are greater for the Class 2 parameters, and the estimated log rate-ratio statistic is not significantly different from zero.

Figures 4.1, 4.2 and 4.3 display the classification probabilities for each of the three data sets. For the MNTH 47 sites displayed in Figure 4.1, a fairly substantial fraction of the sites have classification probabilities in the range of 0.4 to 0.6, indicating that given the data at hand, it is difficult to provide a clearcut assignment of many of the sites. On the other hand, for the MNTH 65 sites shown in Figure 4.2, most sites clearly fall into either Class 1 or Class 2, with relatively few ambiguous cases. Finally, inspection of Figure 4.3 shows that almost of the Hennepin county sites fall in Class 1, with only 4 fairly clear cut outliers.

TABLE 4.1
MAXIMUM LIKELIHOOD ESTIMATES OF MIXTURE MODEL FOR MNTH 47 DATA

Parameter	ML Estimate	Standard Error
α	0.4029	0.9055
p_1	0.2407	0.1095
r_1	0.1038	0.1581
p_2	0.218	0.0949
r_2	0.235	0.1095
Δ_1	1.0069	3.4108
Δ_2	-0.0971	0.9489

TABLE 4.2
MAXIMUM LIKELIHOOD ESTIMATES OF MIXTURE MODEL FOR MNTH 65 DATA

Parameter	ML Estimate	Standard Error
α	0.3686	0.7443
p_1	0.2577	0.0837
r_1	0.092	0.1643
p_2	0.2294	0.0447
r_2	0.2474	0.1304
Δ_1	1.2319	3.9839
Δ_2	-0.0992	0.8099

TABLE 4.3
MAXIMUM LIKELIHOOD ESTIMATES OF MIXTURE MODEL FOR HENNEPIN COUNTY DATA

Parameter	ML Estimate	Standard Error
α	0.882	0.218
p_1	0.385	0.044
r_1	0.217	0.038
p_2	0.204	0.178
r_2	0.076	0.200
Δ_1	0.811	0.049
Δ_2	1.134	6.72

Figure 4.1. MIXTURE ESTIMATES OF PROBABILITY A SITE BELONGS TO CLASS 1 MNTH 47 DATA

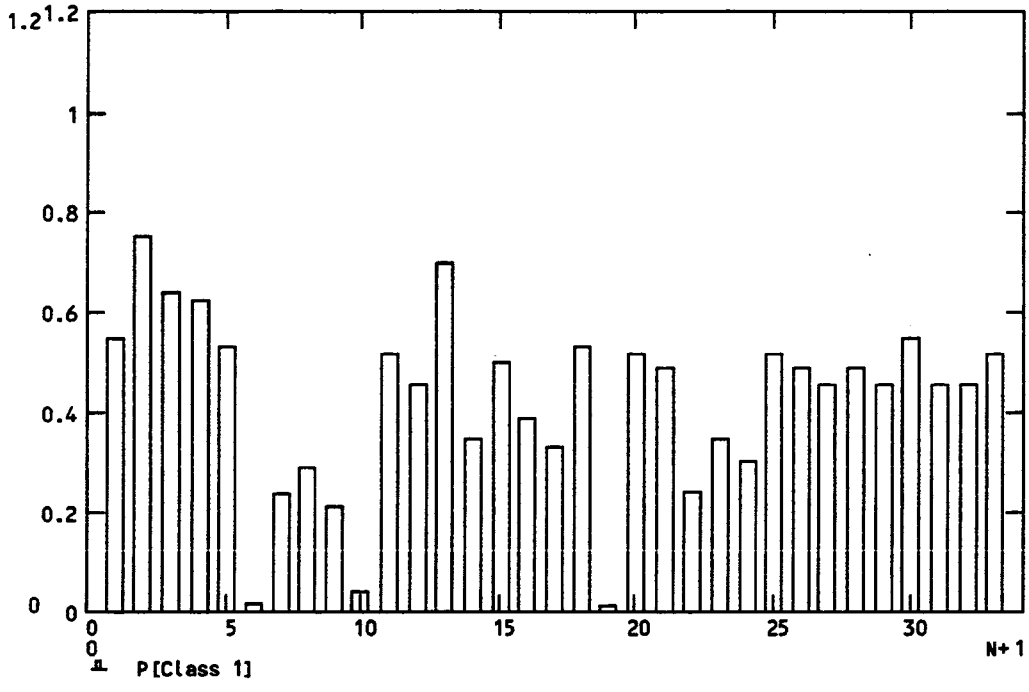


Figure 4.2. MIXTURE ESTIMATES OF PROBABILITY A SITE BELONGS TO CLASS 1 MNTH 65 DATA

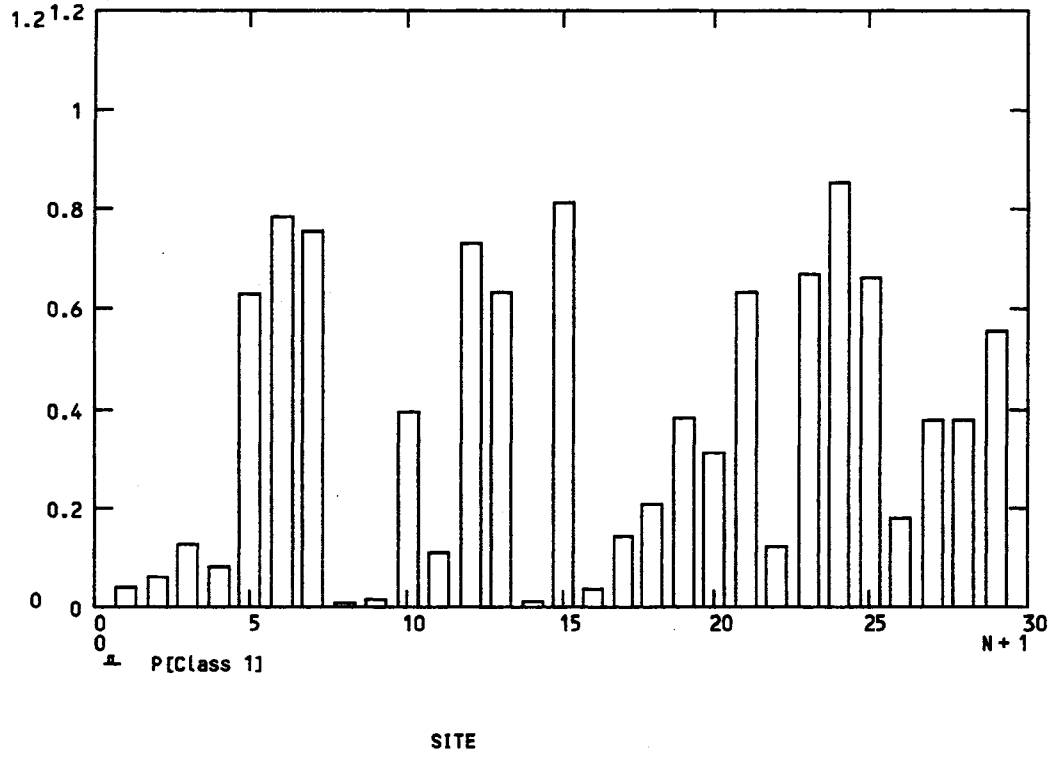
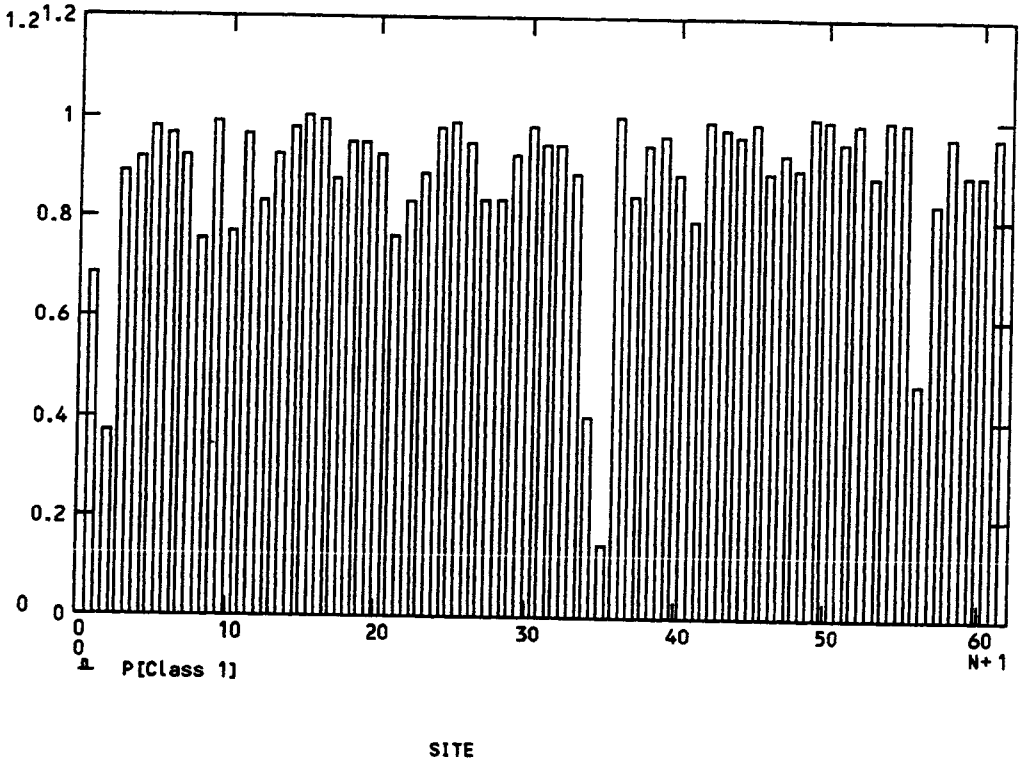


Figure 4.3. MIXTURE ESTIMATES OF PROBABILITY A SITE BELONGS TO CLASS 1 HENNEPIN COUNTY DATA



CHAPTER 5

EMPIRICAL BAYES METHOD FOR IDENTIFYING OLDER DRIVER HIGH RISK SITES

5.1 Theoretical Development

The analyses presented in Chapter 2 indicated a rather weak suggestion that the MNTH 65 corridor might be a candidate for safety improvements targeted to older drivers, but since there were 29 signalized intersections along this corridor, it could very well be that these sites differ as to the risk they pose to older drivers. Such variation in risk could not only explain the weak aggregate effect observed for MNTH 65, but also violates one of the assumptions underlying the statistical tests presented above. Because the numbers of accidents occurring at particular sites over a two or three year period typically tends to be in the range 0-50, the uncertainty attached to site-specific MLEs tends to be high, and application to individual sites of the asymptotic statistical methods described earlier is questionable. Alternatively, identifying high-hazard locations can be viewed as an example of a multi-parameter estimation problem, so that Empirical Bayesian statistical methods might be profitably employed (Higle and Witkowski, 1988; Pendleton, 1991; Hauer, 1992).

As before, let the group 1 denote the older drivers and group 2 denote the middle aged drivers, and assume there is available a 2x2 induced exposure table for each of a set of N sites making up a sample. Let the individual sites be indexed by k=1,..,N, and define the variables

p_k = probability an accident at site k had an older driver as the at-fault driver,
r_k = probability an accident a site k had an older driver as the innocent victim,
n_k = total two-vehicle accidents available for site k,
x_k = number of accidents from site k where an older driver was at-fault,
y_k = number of accidents from site k where an older driver was the innocent victim.

The EB method then assumes the actual accident counts for a site are generated as the outcomes of a two-stage, hierarchical random process where first the probabilities p_k are randomly assigned to sites as the outcomes of independent, identically distributed (iid) Beta random variables, with means and variances given by

$$E[p_k] = p \qquad (5.1)$$
$$Var[p_k] = p(1-p)/(m_1+1),$$

while the r_k are assigned as iid Beta random variables with means and variances

$$E[r_k] = r \qquad (5.2)$$
$$Var[r_k] = r(1-r)/(m_2+1)$$

and second, given p_k, r_k and n_k, the accidents are then assigned to cells in each induced exposure table according to the multinomial model described earlier. The log rate-ratio parameter for site

k becomes

$$\Delta_k = \log_e\left(\frac{p_k(1-r_k)}{r_k(1-p_k)}\right) \tag{5.3}$$

and in the Appendix is shown that the posterior density of Δ_k, given p, m_1, r, m_2 and the data n_k, x_k and y_k is given by

$$f(\Delta_k \mid p, m_1, r, m_2, n_k, x_k, y_k) = \frac{\int \frac{\exp(\Delta_k+z)^{m_1 p + x_k}}{(1+\exp(\Delta_k+z))^{m_1+n_k}} \frac{\exp(z)^{m_2 r + y_k}}{(1+\exp(z))^{m_2+n_k}} dz}{B(m_1 p + x_k, m_1(1-p)+n_k-x_k) B(m_2 r + y_k, m_2(1-r)+n_k-y_k)} \tag{5.4}$$

where B(a,b) denotes the Beta integral evaluated at a and b. The posterior mean and variance of Δ_k are

$$\begin{aligned} E[\Delta_k \mid n_k, x_k, y_k, m_1, p, m_2, r] &= \Psi(m_1 p + x_k) + \Psi(m_2(1-r)+n_k-y_k) - \Psi(m_1(1-p)+n_k-x_k) - \Psi(m_2 r + y_k) \\ Var[\Delta_k \mid n_k, x_k, y_k, m_1, p, m_2, r] &= \Psi'(m_1 p + x_k) + \Psi'(m_1(1-p)+n_k-x_k) + \Psi'(m_2 r + y_k) + \Psi'(m_2(1-r)+n_k-y_k) \end{aligned} \tag{5.5}$$

with $\Psi(x)$ denoting the digamma function, and $\Psi'(x)$ denoting the trigamma function:

$$\begin{aligned} \Psi(x) &= \frac{d\log_e(\Gamma(x))}{dx} \\ \Psi'(x) &= \frac{d\Psi(x)}{dx} \end{aligned} \tag{5.6}$$

Using the probability density (5.4) to find probabilities or confidence intervals will require very laborious computations, but it appears that (5.4) can be approximated by the corresponding normal density with mean and variance given in (5.5), especially for larger values of m_1 and m_2. For instance, Figures 5.1 and 5.2 show typical comparisons of the density (5.4) and the corresponding normal density for low and moderately high values of m_1 and m_2. In this case, given m_1, p, m_2, and r it is straightforward to computes Bayes point and approximate intervals estimates for Δ_k, using the tables for standard normal distribution.

In practice though, the hyperparameters p, m_1, r, m_2 will not be known, and must also be estimated from data. The EB approach proceeds by simply replacing the hyperparameters in (5.5) with these estimates, so that the EB point estimate of Δ_k is

$$\hat{\Delta}_k = \Psi(\hat{m}_1\hat{p}+x_k)+\Psi(\hat{m}_2(1-\hat{r})+n_k-y_k)-\Psi(\hat{m}_1(1-\hat{p})+n_k-x_k)-\Psi(\hat{m}_2\hat{r}+y_k) \qquad (5.7)$$

and the EB estimate of the variance of Δ_k is

$$\hat{\sigma}_k^2 = \Psi'(\hat{m}_1\hat{p}+x_k)+\Psi'(\hat{m}_1(1-\hat{p})+n_k-x_k)+\Psi'(\hat{m}_2\hat{r}+y_k)+\Psi'(\hat{m}_2(1-\hat{r})+n_k-y_k) \qquad (5.8)$$

A naive EB confidence interval with nominal coverage probability $1-\alpha$ would then be

$$(\hat{\Delta}_k - z_{\alpha/2}\,\hat{\sigma}_k,\ \hat{\Delta}_k + z_{\alpha/2}\,\hat{\sigma}_k).$$

where $z_{\alpha/2}$ is the standard normal deviate with cumulative probability $\alpha/2$. Maximum likelihood estimates of the hyperparameters p, m_1, r, m_2 can be found as those values which maximize the marginal distribution

$$L(p,m_1,r,m_2) = \prod_{k=1}^{N}\left[\frac{n_k!}{\prod_{ij} n_{ij,k}!}\,\frac{B(m_1p+x_k,m_1(1-p)+n_k-x_k)}{B(m_1p,m_1(1-p))}\,\frac{B(m_2r+y_k,m_2(1-r)+n_k-y_k)}{B(m_2r,m_2(1-r))}\right] \qquad (5.9)$$

and computation of the ML estimates is simplified by the fact that (5.9) factors into two components, one containing p and m_1 and the other containing r and m_2, so that the maximization problem decomposes into two bivariate problems.

One difficulty which can arise in practice is that the likelihood function (5.9) may be unbounded with respect to either m_1 or m_2, *i.e.* no finite MLE may exist for these parameters.

This was in fact the case for the parameter m_1 for both the MNTH 47 and MNTH 65 data sets, with Figure 5.3 showing the shape of the MNTH 65 likelihoods as functions of m_1 and p, and m_2 and r. One solution to this problem is to assume a plausible noninformative prior distribution for the hyperparameters and then conduct a three-stage hierarchical Bayesian analysis of the problem (Albert, 1984, 1987; Christiansen, et al., 1992), but this leads to the question of choosing an "appropriate" noninformative prior (Eaves, 1982), while tending to obscure the connection between one's choice of a prior and the quality of the ultimate decision. Since an unbounded likelihood for m_1 or m_2 indicates that there is no between site variability in the corresponding values for p_k or r_k, the most defensible solution is to treat these values as constant across the sites. When both m_1 and m_2 have unbounded likelihood, the EB model collapses into the aggregate model described in Chapter 2, but if either m_1 or m_2 have finite ML estimates, the EB method outlined above can still be applied.

FIGURE 5.1.

Comparison of exact posterior density and normal approximation
m1=35, p=.15, m2=30, r=.10, n=10, x=4, y=2

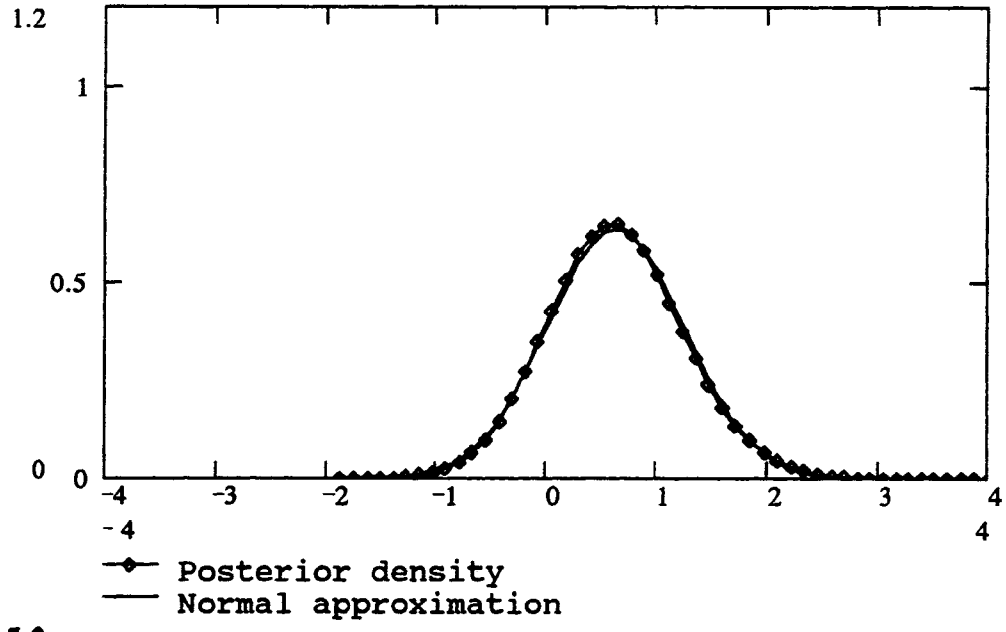


FIGURE 5.2.

Comparison of exact posterior density and normal approximation
m1=150, p=.15, m2=100, r=.10, n=10, x=4, y=2

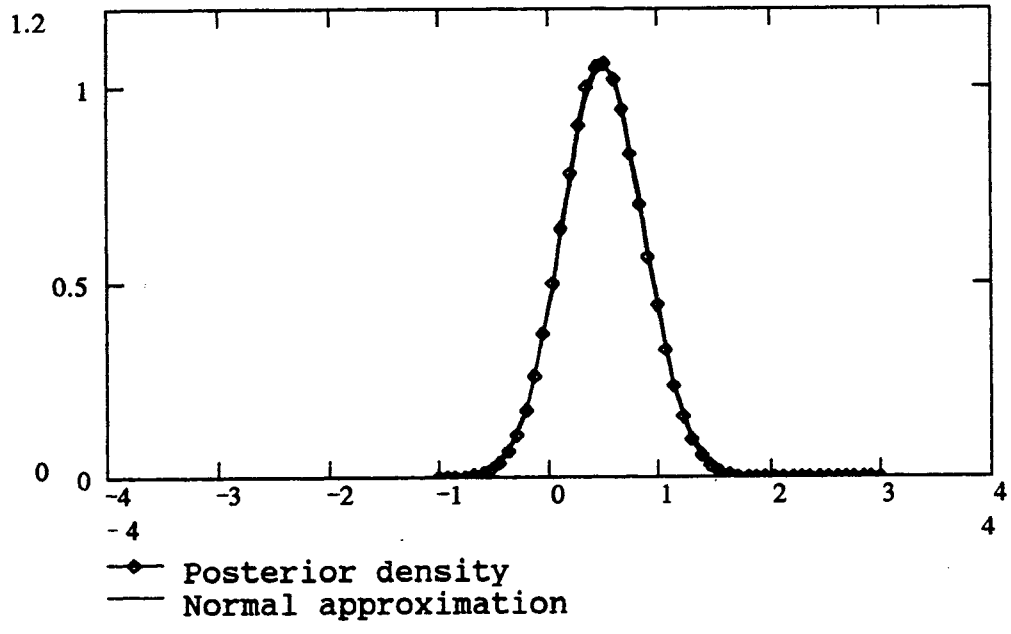
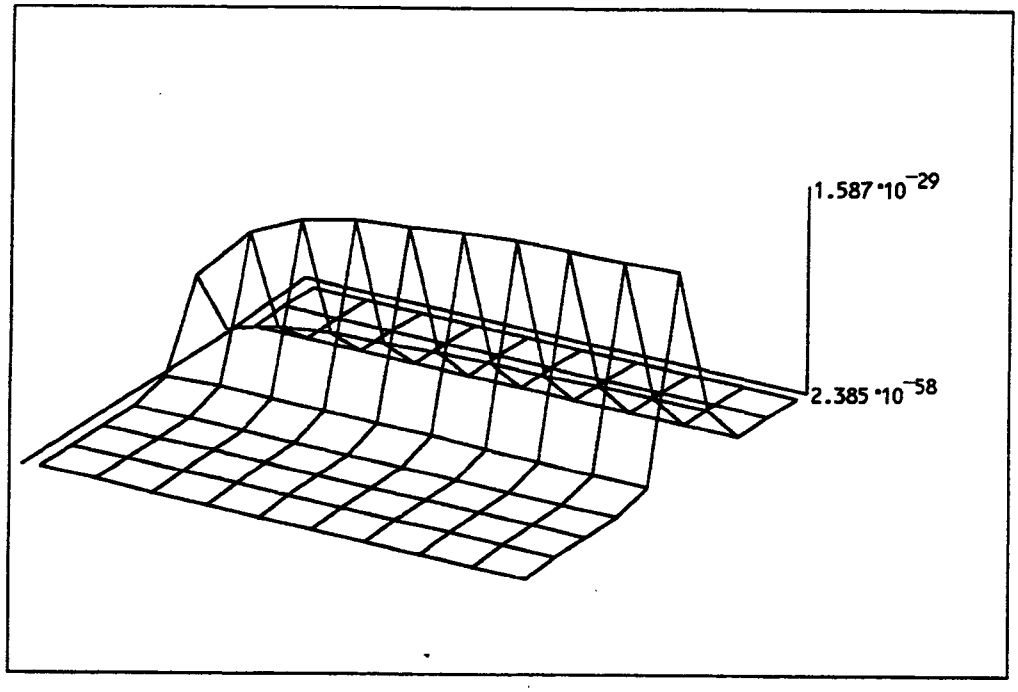
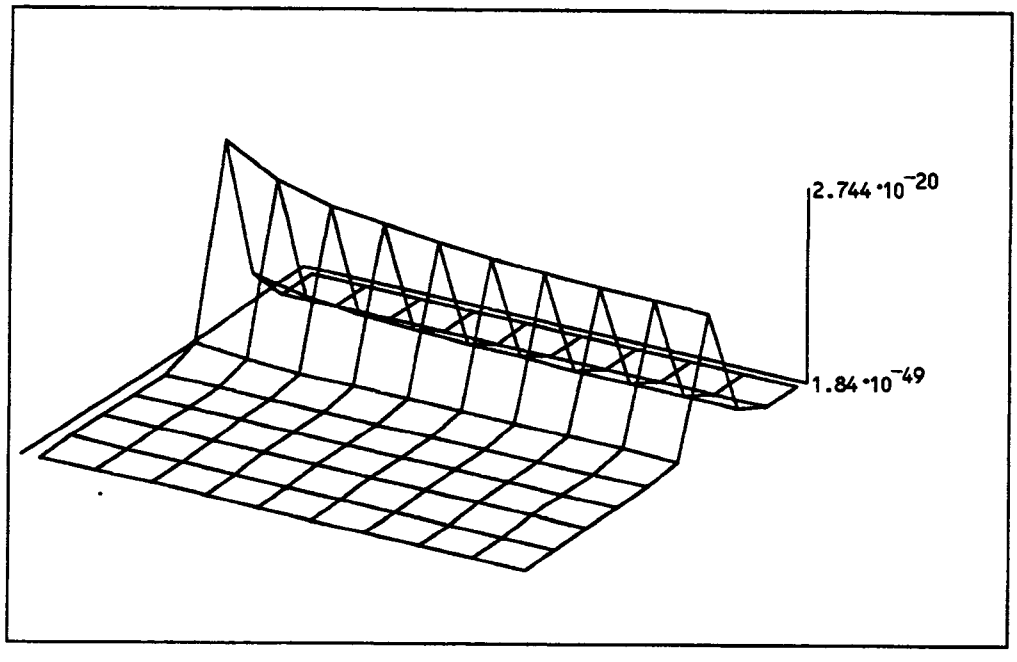


FIGURE 5.3. Likelihood Graphs for Hyperparameters

Likelihood graph for m1 and p



Likelihood graph for m2 and r



5.2 Example Applications

A MATHCAD 3.0 computational document was developed which computed likelihood function graphs for the three data sets described in Chapter 2, and it was found that for both the MNTH 47 and MNTH 65 data, the likelihood function for m1 was unbounded but that for m2 was not. For the Hennepin county data, the likelihood functions for both m1 and m2 were unbounded, making the EB method inapplicable. As an example, Figure 5.3 displays the likelihood graph for the MNTH data. The MATHCAD document also computed ML estimates of the EB parameters p, m_1, r and m_2, where possible, and then wrote these estimates to a text file. An additional MATHCAD document was then developed which took as input the ML estimates and a file of induced exposure tables and then computed the EB point estimates and 90% confidence intervals for the log rate ratio statistics for each site. Figure 5.4 displays these estimates for MNTH 47 while Figure 5.5 displays the estimates for MNBTH 65. Comparing the approximate confidence intervals to the zero value, it can be seen that for MNTH 47, none of the sites can be considered to pose high accident risk to older drivers, since in all cases, the 90% confidence interval also contains the zero values. For the MNTH sites on the other hand, it appears that sites 15 and 24 show evidence that the accident rate for older drivers might be higher there.

FIGURE 5.4. EB Point Estimates of Log Rate-Ratios for 33 Intersections on MNTH 47, Along With Approximate 90% EB Confidence Intervals

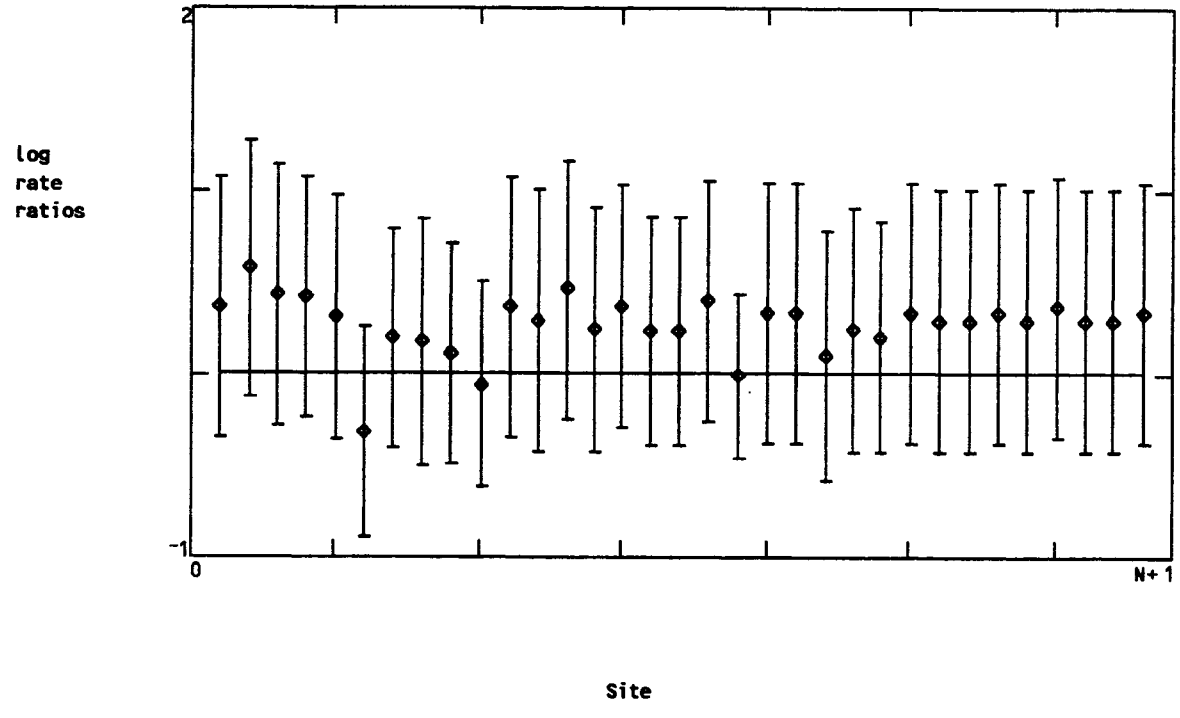
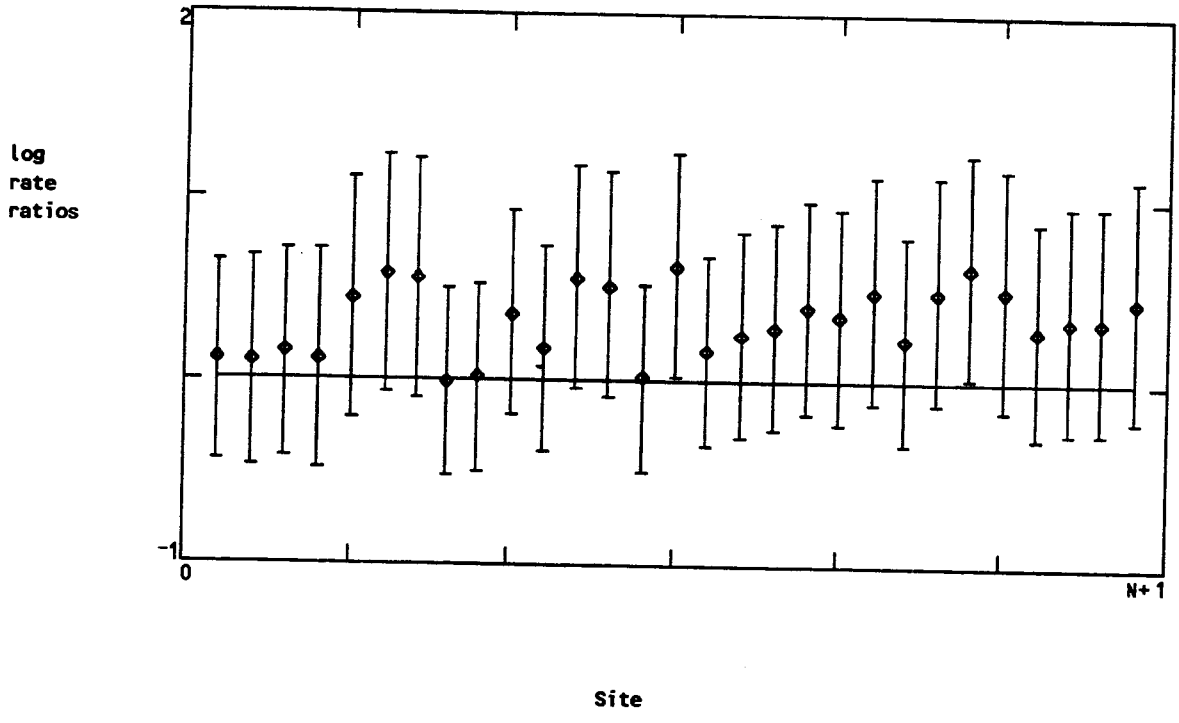


FIGURE 5.5. EB Point Estimates of Log Rate-Ratios for 29 Intersections on MNTH 65, Along With Approximate 90% EB Confidence Intervals



CHAPTER 6

SUMMARY AND CONCLUSIONS

To summarize, Chapter 2 reviewed the identification of high-hazard locations, and concluded that the absence of age-specific measures of exposure, either at the corridor level or the site level, made application of current methods inapplicable. The induced exposure method was then identified as promising a way of side-stepping this difficulty, and it was noted that induced exposure methods have already been employed in several studies investigating the accident risk of older drivers. These studies have all replicated the U-shaped relationship between accident risk and age which was displayed in Chapter 1, but the all were based on highly aggregated data sets, so that specializing their conclusions to corridors or individual sites was impossible. Chapter 3 then presented a formal statement of the induced exposure hypothesis, and showed a count of 2-vehicle traffic accidents, classified by the ages of the at-fault and innocent drivers, could be used to estimate the ratio of the older drivers' accident rate to that of middle-aged drivers. A very simple test for whether or not the older drivers' accident rate was significantly higher than that of the middle-aged drivers was then developed, and tested on three actual accident data sets. In Chapter 4, a clustering procedure for sorting the sites making up an accident data set into two groups was developed, and tested as to its ability to identify high risk sites. Finally in Chapter 5, the Empirical Bayes method developed in a previous project was enhanced and tested.

Generally, the methods developed here performed as they were expected to. The most successful method was that of Chapter 3 clearly identified an area where older drivers appear to be at risk, that of the Hennepin county data, as distinguished from the data from the two Minnesota trunk highways, where older drivers showed little increased risk. This method is very easy to apply, and appropriate for characterizing the accident risk of corridors or subareas, but not that of individual sites.

The two methods for identifying risky sites, the clustering method and the EB method, although somewhat less successful, could still be usefully employed where the circumstances were right. The clustering method seems to require much large data sets than the MNTH 47 and MNTH 65 data sets (with 33 and 29 sites, respectively) in order to achieve a reasonable precision for the estimates of its parameters. Thus although for both MNTH data sets, the clustering method did identify a group of sites where the older drivers' accident rates were higher than those for middle-aged drivers, the difference in the rates was not significant, and this lack of significance is at least in part to the uncertainty of estimating the relative risks and relative exposures when the appropriate class membership of some of the sites is ambiguous. This method is most likely to yield useful results when a large number (50 or more) risky sites need to be identified out of a larger total number (100 or more) of sites.

The EB method could, in principle, identify individual risky sites, but its use in induced exposure modelling seems to be limited by it assumption of over-dispersed data. In the three data sets used here, the MNTH 47 and MNTH 65 data sets showed over-dispersion only in the relative exposure, while the Hennepin county data did not show any over-dispersion. On this basis, it appears that lack of over-dispersion is more likely to be the norm rather than the

exception for induced exposure data, and that the EB method cannot be relied upon. Fortunately it is easy to test whether or not over-dispersion is present in a given data set so that the applicability of the EB method is readily ascertained. It is interesting to note that the MNTH 65 sites identified by the EB method as being risky to older drivers tended to correspond to the sites identified by the clustering method.

Overall, the clustering method and the EB method could both be usefully applied to the traditional task of high-hazard identification, that of automatically screening a large number of accident sites to identify potential candidates for improvement. They do not provide sufficient information for final decisions, which should be based on a more detailed engineering analysis. The simple method of Chapter 3 does appear to provide a reliable way to identify general areas where older drivers are at increased risk, although it does not provide site specific information. This method can be easily used by relatively untrained personnel, while the clustering method and the EB method require some familiarity with their underlying assumptions before they can be used effectively. This also appears to be the case for the EB methods developed by FHWA (Pendleton et al. 1990).

As noted in Chapter 1, the methods developed here have all been implemented as MATHCAD computational documents, and these are available to MNDOT. These documents give the user a capability to conduct the analyses interactively, moving readily between computation and graphical display of results, and to easily document the course of the analysis in a clear and comprehensible manner. To get maximum use of MATHCAD's interactive and graphical capabilities, user should use MATHCAD 3.0 or higher, running on a IBM 386-type microcomputer using Microsoft Windows 3.0 or higher. However, a copy of MATHCAD 2.54, which requires only a DOS operating system, has been purchased by this project for MNDOT use. The PI is prepared to aid the installation of MNDOT in installing this software and the MATHCAD documents produced by this project on a MNDOT computer, and train, support or advise MNDOT personnel on their use.

REFERENCES

Agresti, A. (1990) *Categorical Data Analysis*, New York, Wiley and Sons.

Albert, J. (1984) "Empirical Bayes Estimation of a Set of Binomial Probabilities," *J. Statist. Computer Simul.*, **20**, 129-144.

Albert, J. (1987) "Empirical Bayes Estimation in Contingency Tables," *Commun. Statistics-Theory Method.*, **16**, 2459-2485.

Bishop, Y. Fienberg, S. and Holland, P. (1975) *Discrete Multivariate Analysis*, Cambridge, MA, MIT Press.

Christiansen, C., Morris, C. and Pendleton, O., (1992) "A Hierarchical Poisson Model, with Beta Adjustments for Traffic Accident Analyses," Center for Statistical Science, University of Texas, Austin, Technical Report 103.

Cooper, P. (1990), "Differences in Accident Characteristics among Elderly Drivers and Between Elderly and Middle-Aged Drivers," *Accident Analysis and Prevention*, **22**, 499-508.

Eaves, D., (1980) "On Exchangeable Priors in Lot Acceptance," *J. Royal. Stat. Soc. B*, **42**, 88-93.

Edwards, M. (1992), "Trends in Women's Fatal Crash Involvement", presentation at the 71st Conference of the Transportation Research Board, Washington, DC.

Evans, L. (1991), *Traffic Safety and the Driver*, New York, Van Nostrand.

Garber, N. and Srinivasan, R. (1991), "Risk Assessment of Elderly Drivers at Intersections," *Transportation Research Record*, in press.

Haight, F. (1973), "Induced Exposure," *Accident Analysis and Prevention*, **5**, 111-126.

Hauer, E. (1985), "On Estimation of the Expected Number of Accidents," *Accident Analysis and Prevention*, **18**, 1-12.

Hauer, E. (1992), "Empirical Bayes Approach to the Estimation of Unsafety: The Multivariate Regression Method," *Accident Analysis and Prevention*, **24**, 457-477.

Higle, J, and Witkowski, J. (1988) "Bayesian Identification of Hazardous Locations," *Transp. Res. Record*, **1185**, 24-36.

Hoadley, B. (1971) "Asymptotic Properties of Maximum Likelihood Estimators for the Independent not Identically Distributed Case," *Annals of Mathematical Statistics*, **42**, 1977-1991.

Maleck, T. and Hummer, H. (1987) "Driver Age and Highway Safety," *Transportation Research Record*, **1059**, 6-12.

Maritz, J. (1989), "Empirical Bayes Estimation of the Log Odds Ratio in 2x2 Contingency Tables," *Commun. Statistics-Theory Meth.*, **18**, 3215-3233.

McKelvey, F., Maleck, T., Stamatiades, N., and Hardy, D. (1988), "Highway Accidents and the Older Driver," *Transportation Research Record*, **1172**, 47-57.

Pendleton, O. (1991), *Application of New Accident Analysis Methodologies*, Report FHWA-RD-90-091, FHWA, Washington, DC.

Redner, R., and Walker, H., (1984) "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, **26**, 195-239.

Rohatgi, V. (1977) *Introduction to Probability Theory and Mathematical Statistics*, New York, Wiley and Sons.

Titterington, D., Smith, A. and Makov, U. (1986) *Statistical Analysis of Finite Mixture Data*, Wiley, New York.

Transportation Research Board (1988), *Transportation in an Aging Society: Improving Mobility and Safety of Older Persons*, TRB Special Report 218, National Research Council, Washington, DC.

Wilson, C. and Burtch, T. (1993) "Traffic Accidents and Highway Safety," in J. Pline (ed.) *ITE Traffic Engineering Handbook*, Englewood Cliffs, NJ, Prentice-Hall, 94-116.

APPENDIX A

LISTS OF THE INTERSECTIONS AND INDUCED EXPOSURE DATA

TABLE A1

Two-Vehicle Accidents At Signalized Intersections Along MNTH47

Older Data for HWY 47

Cross. Streets	Xmm	Xmo	Xom	Xoo
40th Ave. NE	2	0	1	0
44th Ave. NE	9	0	1	0
49th Ave. NE	3	0	2	0
53th Ave. NE	7	1	3	0
57th Ave. NE	2	1	4	0
61st Ave. NE	0	4	1	1
Mississippi St.	9	3	2	0
69th Ave. NE	2	1	0	0
73rd Ave. NE	5	2	3	1
Osborne Rd.	8	4	1	1
81st Ave. NE	3	0	0	0
83rd Ave. NE	1	0	0	0
CR 132 LT	3	0	3	0
USTH 169	4	1	0	0
CSAH 7	7	1	1	0
CSAH 9	5	1	3	1
MN 242	7	1	1	1
Hanson Blvd.	8	1	1	0
Foley Blvd.	24	9	7	2
Th 10 at 610	1	0	1	0
B. N. INCR. ANOKA	2	0	0	0
Industry Blvd.	0	1	0	0
Nowthen Blvd.	4	1	0	0
151st Ave.	5	2	2	0
156th Ave.	1	0	1	0
160th Ave.	0	0	1	0
167th Ave.	1	0	0	0
170th Ave. NW	2	0	0	0
Green Valley Rd.	1	0	0	0
177 Ave. NW	2	0	1	0
177th LN NW	1	0	0	0
CSAH 22	1	0	1	0

TABLE A2

Two-Vehicle Accidents At Signalized Intersections Along MNTH 65

Older Data for HWY 65

Cross. Streets	Xmm	Xmo	Xom	Xoo
40th Ave. NE	7	3	1	1
41st Ave. NE	2	3	2	0
44th Ave. NE	3	1	3	2
45th Ave. NE	1	2	0	0
47th Ave. NE	1	0	3	0
49th Ave. NE	6	0	3	0
50th Ave. NE	5	0	3	0
52nd Ave. NE	8	5	2	1
53rd Ave. NE	7	4	4	2
W. Moore Lake Dr.	10	2	5	1
W. Moore Lake Dr. LT	5	2	2	1
Mississippi St.	11	1	3	0
73rd Ave. NE	10	1	2	1
Osborne Rd.	8	5	3	1
81st Ave. NE	14	1	3	0
85th Ave. NE	13	3	1	1
89th Ave. NE	8	3	2	0
91st Ave. NE	9	3	3	0
Clover Leaf Dr.	10	2	2	0
99th Ave. NE	8	2	2	0
105th Ave. NE	8	1	3	0
109th Ave. NE	8	2	1	1
121st Ave. NE	8	0	3	1
MNTH 242	12	1	5	0
129th Ave. NE	4	0	2	0
Bunker Lake Blvd.	7	2	0	0
Constance Blvd. NE	3	1	2	0
Crosstown Blvd. NE	3	1	2	0
Viking Blvd. NE	3	0	1	0

TABLE A3

Two-Vehicle Accidents At Signalized Intersections in Hennepin County

Older Data for Hennepin County

Obs	Location	MIMG	MIOG	OIMG	OIOG
1	CSAH 17 at 89th St.	4	1	0	0
2	CSAH 17 at 84th St.	6	0	0	0
3	CSAH 17 at 90th St.	1	1	0	0
4	CSAH 17 at 98th St.	1	0	1	0
5	CSAH 17 at Minnesota Drive	7	5	3	0
6	CSAH 17 at 76th St.	6	3	3	0
7	CSAH 17 at Parklawn Ave.	6	3	2	0
8	CSAH 17 at Gallagher Drive	4	0	1	0
9	CSAH 17 at Hazelton Road	1	1	3	0
10	CSAH 17 at 70th St.	3	1	0	0
11	CSAH 17 at 69th St./Valley View Rd	7	2	2	1
12	CSAH 17 at Shopping Center Enterance	1	0	0	0
13	CSAH 17 at Shopping Center Exit (79)	0	1	0	0
14	CSAH 17 at CSAH 53	4	2	1	1
15	CSAH 17 at 65th St.	4	8	3	1
16	CSAH 17 at 62ND St.	3	0	3	1
17	CSAH 17 at 60th St.	2	0	1	0
18	CSAH 17 at 58th St.	6	4	2	0
19	CSAH 17 at 54th St.	2	2	1	0
20	CSAH 17 at 51ST St.	0	1	0	0
21	CSAH 17 at CSAH 21	2	0	0	0
22	CSAH 17 at 49 1/2 ST	1	0	0	0
23	CSAH 17 at Sunnyside Ave.	1	1	0	0
24	CSAH 17 at CSAH 20	1	1	0	1
25	CSAH 17 at 89ST St.	1	1	1	1
26	CSAH 17 at 88ST St.	2	2	1	0
27	CSAH 17 at CSAH 3	2	1	0	0
28	CSAH 17 at 94ST St.	2	1	0	0
29	CSAH 17 at 90ST St.	0	1	0	0
30	CSAH 17 at 86ST St.	4	1	2	1
31	CSAH 17 at 84ST St.	1	1	1	0
32	CSAH 17 at 82ND St. LEFT	2	0	0	1

33	CSAH 17 at 82ND St. RIGHT	1	1	0	0
34	CSAH 17 at 80ST St.	13	3	0	1
35	CSAH 17 at 79ST St.	13	2	1	0
36	CSAH 17 at 76ST St.	8	9	2	2
37	CSAH 17 at 75ST St.	3	2	0	0
38	CSAH 17 at 69ST St.	2	0	0	1
39	CSAH 17 at CSAH 53	4	3	2	0
40	CSAH 17 at 64ST St.	1	1	0	0
41	CSAH 17 at 90ST St.	6	4	0	0
42	CSAH 17 at 86ST St.	3	4	1	1
43	CSAH 17 at 84ST St.	1	1	0	1
44	CSAH 17 at 82ND St.	1	0	0	1
45	CSAH 17 at 79ST St.	4	6	2	0
46	CSAH 17 at 76ST St.	6	2	0	1
47	CSAH 17 at 73ST St.	2	3	0	0
48	CSAH 17 at 70ST St.	4	4	0	0
49	CSAH 17 at CSAH 53	13	9	3	3
50	CSAH 17 at Med Center Entrance	0	0	0	2
51	CSAH 17 at CSAH 53	2	4	0	0
52	CSAH 17 at Vincent Ave.	0	2	2	0
53	CSAH 17 at Sheridan Ave.	0	0	0	0
54	CSAH 17 at Logan Ave.	1	1	1	2
55	CSAH 17 at Ray/Lakeshore Drives	5	1	2	2
56	CSAH 17 at Lyndale Ave.	8	1	1	0
57	CSAH 17 at Pillsbury Ave.	1	0	0	0
58	CSAH 17 at CSAH 53	3	2	2	0
59	CSAH 17 at 1ST St.	1	1	0	0
60	CSAH 17 at 12th Ave.	1	1	0	0
61	CSAH 17 at Bloomington Ave.	0	1	1	0

APPENDIX B

DERIVATION OF STATISTICAL FORMULAS

APPENDIX B: DERIVATION OF STATISTICAL FORMULAS

Although the statistical methods presented above are relatively straightforward applications Maximum Likelihood (ML) and Empirical Bayes (EB) theories, the actual formulas are not yet textbook material, so outlines of their derivations are given here. In what follows, the symbol $=>$ will denote the convergence in distribution of two random variables, while $N(\mu,\Sigma)$will denote a normally distributed random vector with mean vector μ and covariance matrix Σ.

First, consider the ML estimate of the log rate ratio statistic given in Equation (3.7) and its estimated variance given in (3.8). By assumption, the total accidents in each row of Table 1 are generated originally as independent Poisson outcomes, so that given the random selection of victim category and the total accident count n, the classification counts depicted in Table 1 become multinomial outcomes, with their log-likelihood function being

$$l(x,y|n,p,r)=\log\left(\frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\right)+x\log(p)+(n-x)\log(1-p)+y\log(r)+(n-y)\log(1-r) \quad \text{(A1)}$$

and the values of p and r which maximize this function are given in (3.4). Standard asymptotic results for multinomials gives the these ML estimates are consistent, and that

$$\sqrt{n}\begin{bmatrix}\hat{p}-p\\ \hat{r}-r\end{bmatrix}\Rightarrow N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}p(1-p) & 0\\ 0 & r(1-r)\end{bmatrix}\right) \quad \text{(A2)}$$

Defining the log rate-ratio as

$$\Delta(p,r) = \log_e\left(\frac{\lambda_1}{\lambda Hx}\right) - \log_e\left(\frac{p(1-r)}{r(1-p)}\right) \quad \text{(A3)}$$

it follows that (3.7) is the ML estimate for Δ (Rohatgi, 1977, p. 383), and since for $0<p,r<1$, Δ is a continuously differentiable function of p and r, it follows that (3.7) is a consistent estimator of $\Delta(p,r)$ (Rohatgi, 1977, p. 245). The delta method then gives

$$\sqrt{n}[\hat{\Delta}(\hat{p},\hat{r})-\Delta(p,r)] \Rightarrow N\left(0,\left[\frac{\partial\Delta}{\partial p},\frac{\partial\Delta}{\partial r}\right]\begin{bmatrix} p(1-p) & 0 \\ 0 & r(1-r)\end{bmatrix}\begin{bmatrix}\frac{\partial\Delta}{\partial p} \\ \frac{\partial\Delta}{\partial r}\end{bmatrix}\right) \tag{A4}$$

(Bishop, et al. 1975, p. 493), and noting that

$$\begin{bmatrix}\frac{\partial\Delta}{\partial p} \\ \frac{\partial\Delta}{\partial r}\end{bmatrix} = \begin{bmatrix}\frac{1}{p(1-p)} \\ \frac{-1}{r(1-r)}\end{bmatrix} \tag{A5}$$

inserting (3.4) for p and r in (A4) and (A5) yields the variance estimated (3.8).

Next, consider the posterior distribution of the site-specific log rate ratios given in Equation (5.4), together with the closed form expressions for the posterior means and covariances, given in (5.5). By assumption, the joint prior distribution for p_k and r_k is the product of their respective Beta distributions, while the likelihood of the induced exposure table for site k has the form given in (A1), with p_k and r_k replacing p and r. The Beta priors are conjugate to this likelihood function, so the posterior distribution of p_k and r_k given x_k, y_k and n_k is simply

$$f(p_k,r_k|x_k,y_k,n_k) = \frac{p_k^{m_1p+x_k-1}(1-p_k)^{m_1(1-p)+n_k-x_k}}{B(m_1p+x_k,m_1(1-p)+n_k-x_k)} \times \frac{r_k^{m_2r+y_k}(1-r_k)^{m_2(1-r)+n_k-y_k}}{B(m_2r+y_k,m_2(1-r)+n_k-y_k)} \tag{A6}$$

Hence the random variables $(p_k|x_k,n_k)$ and $(r_k|y_k,n_k)$ are independent, and the posterior log rate ratio

$$\Delta_k = \log_e\left(\frac{p_k(1-r_k)}{r_k(1-p_k)}\right) = \log_e\left(\frac{p_k}{1-p_k}\right) - \log_e\left(\frac{r_k}{1-r_k}\right) \tag{A7}$$

can be expressed as the difference between two independent random variables, the logits of p_k and r_k. Letting $\pi_k=\log_e[p_k/(1-p_k)]$, $\rho_k=\log_e[r_k/(1-r_k)]$, the posterior density of π_k is then

$$g(\pi_k|x_k,n_k) = \frac{\dfrac{\exp(\pi_k)^{m_1 p+x_k}}{(1+\exp(\pi_k))^{m_1+n_k}}}{B(m_1 p+x_k,\ m_1(1-p)+n_k-x_k)} \tag{A8}$$

while a similar expression can be derived for the posterior distribution of ρ_k. Applying the formula given in Rohatgi (1977, p. 141) for the density of the difference between two independent random variables give (5.4).

Turning now to (5.5), note that

$$\begin{aligned} E[\hat{\Delta}_k|x_k,y_k,n_k] &= E[\pi_k|x_k,y_k,n_k]-E(\rho_k|x_k,y_k,n_k] \\ var[\hat{\Delta}_k|x_k,y_k,n_k] &= var[\pi_k|x_k,y_k,n_k]+var[\rho_k|x_k,y_k,n_k] \end{aligned} \tag{A9}$$

The moment generating function for the posterior density of π_k is

$$M_\pi(s) = \frac{B(m_1 p+x_k+s,m_1(1-p)+n_k-x_k-s)}{B(m_1 p+x_k,m_1(1-p)+n_k-x_k)} \tag{A10}$$

and so that the posterior expected value and variance of π_k can be found by evaluating the first and second derivatives of (A10) with respect to s at the point s=0. An expression similar to (A10) can be given for the moment generating function for the posterior of ρ_k, and after some algebra, (5.5) results.

Finally, to develop the asymptotic ML theory for the two-class mixture model, let $\boldsymbol{\theta}_N=[\alpha,p_1,r_1,p_2,r_2]^T$ denote the ML estimates obtained when induced exposure tables from a total of N sites are available, and $\boldsymbol{\theta}_0=[\alpha,p_1,r_1,p_2,r_2]^T$ denote the true parameter values. If it is possible to show that the sequence of random vectors $\boldsymbol{\theta}_N$ was consistent and asymptotically normally distributed, then the delta method can be employed to show asymptotic normality of the estimated rate ratios given in (4.4). A difficulty arises here because the asymptotic results presented for mixture models in the standard references, such as Redner and Walker (1984) or Titterinton et al. (1986), essentially assume that the observations have been generated by independent, identically distributed random vectors, while the variation of n_k across sites means that the induced exposure tables will be independent but not identically distributed. Hoadley (1971) presents an asymptotic ML theory for independent not identically distributed cases, and if it is assumed that $1\le n_k\le n'$, for some finite upper bound n', then it is fairly straightforward, (but tedious) to first verify that Hoadley's conditions C1, C2, C3', C4' and C5 are satisfied, which are sufficient to imply that $\boldsymbol{\theta}_N$ converges in probability the $\boldsymbol{\theta}_0$, and then to verify that his

conditions N1, N2, N3, N4, N5', N6', N7, N8' and N9 are also satisfied, which together imply that

$$\sqrt{N}[\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0] \rightarrow N\left(0, \Gamma^{-1}(\boldsymbol{\theta}_0)\right) \tag{A11}$$

It then follows that (4.4) gives the ML estimators for Δ_1 and Δ_2, and since the functions given in (4.4) are continuously differentiable with respect to $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, it then follows that the estimators given in (4.4) are consistent, and that

$$\sqrt{N}\begin{bmatrix}\hat{\Delta}_1 - \Delta_1 \\ \hat{\Delta}_2 - \Delta_2\end{bmatrix} \rightarrow N\left(0, J(\boldsymbol{\theta}_0)\Gamma^{-1}(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^T\right) \tag{A12}$$

where

$$J(\boldsymbol{\theta}_0) = \begin{bmatrix} 0 & \dfrac{1}{p_1(1-p_1)} & -\dfrac{1}{r_1(1-r_1)} & 0 & 0 \\ 0 & 0 & 0 & \dfrac{1}{p_2(1-p_2)} & -\dfrac{1}{r_2(1-r_2)} \end{bmatrix} \tag{A13}$$

Letting

$$\pi_{1,k} = \hat{p}_1^{x_k}(1-\hat{p}_1)^{n_k-x_k}\hat{r}_1^{y_k}(1-\hat{r}_1)^{n_k-y_k} \qquad \pi_{2,k} = \hat{p}_2^{x_k}(1-\hat{p}_2)^{n_k-x_k}\hat{r}_2^{y_k}(1-\hat{r}_2)^{n_k-y_k}$$

$$s_k = \frac{1}{\hat{\alpha}\pi_{1,k}+(1-\hat{\alpha})\pi_{2,k}} \begin{bmatrix} \pi_{1,k}-\pi_{2,k} \\ \dfrac{\hat{\alpha}\pi_{1,k}}{\hat{p}_1(1-\hat{p}_1)} \\ \dfrac{\hat{\alpha}\pi_{1,k}}{\hat{r}_1(1-\hat{r}_1)} \\ \dfrac{(1-\hat{\alpha})\pi_{2,k}}{\hat{p}_2(1-\hat{p}_2)} \\ \dfrac{(1-\hat{\alpha})\pi_{2,k}}{\hat{r}_2(1-\hat{r}_2)} \end{bmatrix} \tag{A14}$$

the inverse of the asymptotic covariance matrix $\Gamma(\theta_0)$ can be estimated via

$$\hat{\Gamma} = \frac{1}{N}\sum_{k=1}^{N} \hat{s}_k \hat{s}_k^T \tag{A15}$$

and substituting the estimated values α, p_1, r_1, p_2, r_2 and (A15) into the covariance matrix given in (A12) produces an estimate of the covariance for $\hat{\Delta}_1$ and $\hat{\Delta}_2$. The standard normal distribution can then be used to test whether or not the estimated log rate ratios are significantly different from zero, or to construct confidence intervals.