MN DEPT OF TRANSPORTATION
3 0314 00024 5372

AL METHOD FOR IDENTIFYING AREAS OF HIGH CRASH RISK TO OLDER DRIVERS

Final Report

Prepared by

Gary A. Davis, Ph.D.
Department of Civil and Mineral Engineering
University of Minnesota

Submitted to

Research Administration and Development Section
Office of Materials and Research
Minnesota Department of Transportation

September 1991

CTS
HE
5620
.A24
D38
1991

eport represents the results of research conducted by the author and does not irily reflect the views or policy of the Minnesota Department of Transportation. This does not contain a standard or specified technique.

435 33769

TABLE OF CONTENTS

ABSTRACT	1
INTRODUCTION	2
CLASSICAL AND EMPIRICAL BAYESIAN HAZARD IDENTIFICATION	5
THE INDUCED EXPOSURE MODEL	8
EB ESTIMATION FOR THE INDUCED EXPOSURE MODEL	12
EXAMPLE APPLICATIONS	16
MNTH 65 Data Set	18
MNTH 47 Data Set	20
CONCLUSION AND RECOMMENDATIONS	21
REFERENCES	23
APPENDIX	A-1

ACKNOWLEDGEMENTS

Konstantinos Koutsoukos, a graduate student at the University of Minnesota, performed most of the data preparation and analysis reported here, with some assistance by undergraduate student Bill McNary. Dave Miller of MNDOT developed the accident data files used in this research, while Robert Johns of the Center for Transportation Studies at the University of Minnesota made the initial contacts which resulted in this research project. This research was sponsored by the Minnesota Department of Transportation, but all opinions and conclusions expressed here are solely the responsibility of the author.

ABSTRACT

Effective use of finite roadway improvement budgets to accommodate an increasing number of older drivers requires that we be able to identify those locations where older drivers appear to have a heightened accident risk. Ideally, the accident records from a location (such as a particular intersection) should provide the information needed to assess the risk experienced there by a given group of drivers, but the lack of location and age-specific measures of exposure coupled with the relatively small accident samples available for particular locations makes the standard methods of high-hazard identification inapplicable. In this paper it is first shown how, by using an induced exposure approach, it is possible to test for the equality of group-specific accident rates at a given site by testing for the equality of two binomial probabilities which arise from a particular type of contingency table. It is next shown how an Empirical Bayesian approach to computing point and interval estimates for binomial probabilities, which has appeared in the statistical literature, can be adapted to this problem. The resulting computational procedures are relatively straightforward and can be implemented on a microcomputer. The method is tested using accident data from two Minnesota highways, with the following results. First, the method's underlying probability assumptions appear to be supported by the data sets. Second, a Normal distribution approximation used to compute confidence intervals showed acceptable accuracy. Third, the EB estimation procedure showed a tendency to reduce the scatter characteristic of Maximum Likelihood estimates while still identifying locations showing elevated accident risk for a given driver age group. It is concluded that the method could be a useful tool for safety engineers, and some additional research and development necessary for practical implementation of the method is recommended.

INTRODUCTION

It is a well-established demographic fact that individuals born during the years 1947-1957 constitute a substantial portion of the current U.S. population, and as these "baby-boomers" age, older drivers will comprise an increasingly significant proportion of roadway users. Current road design standards and traffic engineering practice however developed during times when older drivers constituted a small minority of the driving population, leading many transportation professionals to consider whether anticipatory roadway improvements might be needed to block a future increase in traffic accidents *(1)*. The value of such improvements depends first on whether or not older drivers actually have more difficulty with the existing roadway than do younger drivers and, second, on being able to reliably identify those locations which actually cause the difficulty. As to the first point, there is mounting evidence that, after about the age of 55 or 60, the accident rate for drivers tends to increase *(1,2)*, but this evidence is not location specific. As to the second point, the problem of identifying locations showing increased accident rates for older drivers is the subject of this report.

The Minnesota Department of Transportation (MNDOT) has projected that the proportion of older drivers in the state's driving population will increase over the next 20 years, and that older drivers appear to be over-represented in traffic crashes *(3)*. This has led to a proposed program of roadway improvements intended to enhance older driver safety, including increased use of channelization and control at intersections, improved visibility of roadway markings and

signing, and improved positive guidance. Efficient use of limited resources however, requires identifying those areas where older drivers are at greatest risk, and improving these locations first. During the Spring of 1990, the Principal Investigator (PI) for this project began research aimed at identifying locations at which accident records indicated that older drivers were over-represented, but it soon became apparent that statistical identification of such high risk areas from accident records was a nontrivial task for which appropriate statistical tools were not available. A study of this problem indicated that two factors were the causes of these difficulties. First, although it is generally possible, using estimates of average daily traffic (ADT), to estimate the total number of drivers exposed to accidents at some location, a breakdown of the ADT at a location by the ages of drivers is almost always impossible, due to the inability of automatic traffic counters to detect the ages of drivers. This means that standard hazard identification methods based on direct estimation of accident rates cannot be applied to this problem. Second, over a period of one to five years, most specific street locations tend to show statistically small numbers of accidents, so that the large population approximations commonly used in applied statistics for computing confidence intervals and testing hypotheses are not available either. These small accident counts also mean that point estimates of accident parameters will tend to have high variability, and thus be more subject to a biasing phenomenon called regression to the mean (RTM). In response to these problems, this research project shows how to combine the induced exposure model for estimating group-specific accident risk with an Empirical Bayesian (EB) estimation procedure, producing a flexible and computationally tractable statistical tool. The following report first briefly reviews the current state of the art in identifying high-hazard locations, with an emphasis on recent applications of EB methodology. It is then shown that,

under the induced exposure hypothesis, the problem of testing for the equality of two age-specific accident rates can be reduced to a problem of testing for the equality of two binomial probabilities, and also that EB methods can be adapted to solve this problem. This methodology then is tested on two accident data sets taken from Minnesota highways. The report concludes with a summary and recommendations for further research and development.

CLASSICAL AND EMPIRICAL BAYESIAN HAZARD IDENTIFICATION

Traffic accidents are fortunately "rare events" compared to the amount of travel done by the population which generates them, so the Poisson distribution provides the statistical model for much 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 k, n_k is assumed to be a Poisson random variable with mean value $m_k = \lambda_k E_k$, where

λ_k = accident rate at location k,

E_k = exposure of the travelling population at k.

The accident rate λ_k is thus a measure of the proclivity of location k to produce accidents, with locations having higher values of λ_k being more dangerous. The exposure E_k is a measure of the size of the population at risk, and standard measures of exposure used in traffic accident analysis are the total traffic count at a location, used primarily for analysis of intersections, and vehicle-miles of travel, used for roadway 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 time period, the maximum likelihood (ML) estimator of the accident rate is simply

$$\hat{\lambda}_k = \frac{n_k}{E_k} \tag{1}$$

"High hazard" locations can then be identified by computing the estimated accident rate for each location of interest and then selecting those locations where the estimated accident rates are large

compared to some regionwide standard. Two common methods for high hazard identification, the accident rate method and the rate quality control method *(4)* are based on the estimator (1).

It turns out, however, that $\hat{\lambda}_k$ is often a poor predictor of future accident rates due to its failure to correct 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 estimator $\hat{\lambda}_k$ 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 estimates $\hat{\lambda}_k$ will tend to confound 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.

Beginning with Hauer *(5)*, a number of accident researchers have worked at improving the ability to identify high-hazard locations through the application of "shrinkage" or "Empirical Bayesian" statistical techniques, and this work has reached a useful degree of maturity in the EBEST methodology developed at the Texas Transportation Institute *(6)*. This approach begins with a Bayesian statistical model in which accidents are assumed to be generated by a two-step probabilistic process. First, a common underlying Gamma random variable with mean λ and variance (λ/ϵ) generates the accident rates λ_k for each site in a sample, and then the actual accident counts are generated as Poisson outcomes as described above. If, in addition to knowing the exposures E_k and the accident counts n_k, one also knows the values of the Gamma parameters λ and ϵ, it can be shown that the Bayes estimator of the accident rates is given by

$$\lambda^*_k=(\frac{E_k}{E_k+e})\hat{\lambda}_k+(\frac{e}{E_k+e})\lambda \qquad (2)$$

which is a convex combination of the ML estimator given in (1) and the underlying Gamma mean. For those sites with high exposures (and hence lower variances for $\hat{\lambda}_k$) the Bayes estimator tends to weight the ML estimator more heavily, while those sites with low exposures are "shrunk" more toward the Gamma mean. The parameter ϵ measures the degree of relatedness among the individual accident rates, with $\epsilon=0$ being the case where the individual accident rates, λ_k, have no relation (so that $\lambda^*_k=\hat{\lambda}_k$), while $\epsilon=\infty$ corresponds to the case $\lambda_k=\lambda$, (*i.e.* the accident rate at each location is equal to the Gamma mean.) It can also be shown that, for intermediate values of ϵ, the Bayes estimator tends to be closer to the true accident rate than does the ML estimator. In most practical situations however, the values of λ and ϵ will be unknown and also require estimation. Empirical Bayes methods attempt to capitalize on the desirable properties of the Bayes estimator by first replacing the parameters ϵ and λ in (2) with efficient estimates computed from the accident sample and second, by accounting for the increased uncertainty which results from having less than perfect knowledge of these parameters *(7,8)*.

THE INDUCED EXPOSURE MODEL

In principle, the above method of analysis could be extended to the identification of locations where older drivers are over-represented by simply allowing each age group of drivers to have differing accident rates and exposures at each location of interest. Thus we can define

λ_{ik} = accident rate or risk for age group i at location k,

E_{ik} = exposure of age group i at location k,

n_{ik} = observed number of accidents for age group i at location k,

with n_{ik} assumed to be a Poisson random variable with mean $m_{ik}=\lambda_{ik}E_{ik}$. Given observations of n_{ik} and E_{ik} for all groups and locations, one could not only identify those locations where older drivers are over-represented (indicated by high values of m_{ik}) but attribute the over-representation to either overexposure (indicated by high values of E_{ik}), to greater risk (indicated by high values of λ_{ik}) or to a combination of these effects. The above methodology is strictly appropriate however, only when the exposures E_{ik} can be treated as known constants in the analysis. In practice such measures of exposure are derived from a location's Average Daily Traffic (ADT), which in turn is usually estimated from randomly varying traffic count data, so that ADT (and hence exposure) is more properly treated as an additional parameter to be estimated, rather than as a known constant. The current state of the art is such that the statistical properties of various methods for estimating ADT are not well understood, while the relationship between ADT estimates and the resulting estimates of accident risk such as (1), are even less clear. When using accident rate methods to identify locations where older drivers are at risk these statistical questions are academic, however, since disaggregated measures of exposure for single locations

are not generally available, and can only be obtained by the expensive and time-consuming expedient of stopping and sampling vehicles at the location.

The difficulties inherent in obtaining good estimates of exposure have been known for some time, and have motivated a number of researchers to develop measures of exposure which rely only on data contained in the accident records themselves *(9)*. The basic idea of this "induced exposure" method is that for a majority of two-vehicle accidents, one driver can be identified as at fault, while the other is an innocent victim. At-fault drivers are assumed to be subject to accidents according to the above Poisson model, while victims are assumed to be randomly selected in proportion to their exposures at a location. Thus the proportion of an age group in the victim total gives a measure of the relative exposure of that age group at a location, and offers a method for untangling the contributions of risk and exposure for particular age groups. This idea appears to have originated with Thorpe *(10)*, enjoyed intense but brief research interest in the early 1970's *(11-13)* and more recently has been resurrected by researchers at the University of Michigan to investigate the relative hazard for older drivers of different types of intersection *(2)*. The Michigan methods are essentially deterministic procedures however, and though useful when dealing with large aggregations of accident records, their lack of a foundation in statistical theory makes them inappropriate in small sample situations. Fortunately there does exist a natural connection between the induced exposure model and statistical theory, which will now be made explicit.

To simplify the resulting notation, we will treat the case where the population of interest has been divided into only two age groups, "younger" and "older", with λ_{yk} and E_{yk} denoting the risk and exposure for the younger group at location k, while λ_{ok}, E_{ok} denotes the

corresponding quantities for the older group. This covers most cases of interest, but extension to more complicated classifications appears possible through the use of multinomial and Dirichlet random variables in place of the binomial and Beta random variables used here. We then define the following quantities

$r_k = E_{ok}/(E_{yk} + E_{ok})$, the relative exposure of the older group at location k,

$p_k = \lambda_{ok}E_{ok}/(\lambda_{ok}E_{ok} + \lambda_{yk}E_{yk})$, the relative involvement of the older group at location k,

x_k = the total number of two-vehicle accidents at location k for which an older driver was the at-fault driver,

y_k = the total number of two-vehicle accidents at location k for which an older driver was the innocent victim,

n_k = the total number of two-vehicle accidents at location k in the sample.

Under the induced exposure hypothesis, r_k gives the probability that a driver who has an accident "selects" an older driver as the victim, while p_k gives the probability that a given two-vehicle accident has an older driver as the at-fault party. When the two-vehicle accidents at a location are cross-classified by the ages of the drivers involved, the resulting cell counts will also have Poisson distributions, and by exploiting well-known properties of Poisson and multinomial random variables it can be shown that when the accident total n_k is given, the cross-classification counts form a multinomial random vector. The marginal total x_k is now a binomial random variable with parameters n_k and p_k, while the marginal total y_k is binomial with parameters n_k and r_k. Next, from the definitions of p_k and r_k it is straightforward to verify that the condition

$\lambda_{ok}=\lambda_{yk}$ is true if and only if the condition $p_k=r_k$ is also true, so that under the induced exposure model, the problem of testing whether or not two age groups have the same accident rate at a given location reduces to the problem of testing whether or not two binomial probabilities arising from a cross-classification table are equal. The ML estimators of p_k and r_k are given by

$$\hat{p}_k=\frac{x_k}{n_k} \qquad \hat{r}_k=\frac{y_k}{n_k} \tag{3}$$

and if the number of accidents at a site is large (*i.e.* 50 or more) the hypothesis $p_k=r_k$ can be tested using asymptotic likelihood ratio methods (or equivalently, asymptotic methods for contingency table analysis) *(14)*. In practice however, $n_k \geq 50$ is likely to be the rare exception rather than the rule, so that asymptotic methods of hypothesis testing become suspect and the ML estimators $\hat{p}_k$ and $\hat{r}_k$ become subject to more pronounced RTM effects. Since our problem is essentially one of hypothesis testing rather than point estimation, EB procedures such as those described earlier are not directly applicable. In the statistical literature though, Albert *(15)* has presented methods for computing both point and interval EB estimates of binomial probabilities, and this methodology can be adapted to produce not only EB point estimates of the quantity (p_k-r_k) for each location but also approximate EB confidence intervals for these differences. A decision rule for identifying which sites satisfy $p_k=r_k$ (and hence $\lambda_{ok}=\lambda_{yk}$) can then be based on whether or not a confidence interval for the difference (p_k-r_k) contains the value zero.

EB ESTIMATION FOR THE INDUCED EXPOSURE MODEL

In order to illustrate how Albert's formulas can be applied to the problem at hand, we will discuss, in some detail, the problem of estimating the probabilities p_k for a set of locations, and then simply note that estimation of the r_k is exactly parallel. As with the preceding model for accident rates, the EB procedure assumes that the values x_k are generated by a two-step random mechanism, only this time the parameters p_k are assumed to be assigned to locations as the independent outcomes of Beta random variables with expected value $E[p_k]=p$ and variance $var(p_k)=p(1-p)/(m_1+1)$. If p and m_1 were known, the Bayes estimator of p_k would then be given by

$$p_k^* = \left(\frac{n_k}{n_k+m_1}\right)\hat{p}_k + \left(\frac{m_1}{n_k+m_1}\right)p \qquad (4)$$

In practice however, p and m_1 will be unknown, and in order to consider their estimation in a Bayesian context, Albert assumes that *a priori* the uncertainty in these parameter values can be described using noninformative prior distributions, leading to a three-stage, fully Bayesian model. In principle, all quantities of interest, such as point estimates and confidence intervals, could then be computed by appropriate integrations of this Bayesian model's joint probability distribution. However, closed form expressions for the resulting integrals do not exist, and Albert avoids the need for computationally demanding numerical integrations by considering an EB estimator of the form

$$\tilde{p}_k = \left(\frac{n_k}{n_k + \hat{m}_1}\right)\hat{p}_k + \left(\frac{\hat{m}_1}{n_k + \hat{m}_1}\right)\hat{p} \tag{5}$$

where $\hat{p} = (\Sigma x_k / \Sigma n_k)$ is an unbiased estimator of p, and Albert estimates m_1 using an approximate Bayesian procedure. Rather than use Albert's estimator of m_1, which for the case considered here would apply the same degree of shrinkage to all locations regardless of the actual numbers of accidents counted, we propose estimating m_1 as the value which maximizes the function

$$f(m) = \frac{\int \prod_{k=1}^{N} \left[\binom{n_k}{x_k} \frac{\beta(mp + x_k, m(1-p) + n_k - x_k)}{\beta(mp, m(1-p))}\right] \frac{dp}{p(1-p)}}{m} \tag{6}$$

Here $\beta(a,b)$ denotes the Beta integral evaluated at the values a and b. Although the function f(m) described in (6) looks rather complicated, it is essentially the product of the likelihood of the observations $(x_1,..,x_N)$ conditional on m_1, multiplied by the noninformative prior placed on m_1 *(15, p. 137)*. The value of m_1 which maximizes f(m) is, then, a maximum *a posteriori* (MAP) estimator.

Albert also provides an approximation to the posterior variance of p_k given the data x_k which takes the form

$$v_{pk} = \left(\frac{n_k}{n_k + \hat{m}_1}\right)\frac{\tilde{p}_k(1-\tilde{p}_k)}{n_k + 1} + \left(\frac{\hat{m}_1}{n_k + \hat{m}_1}\right)\frac{\hat{p}(1-\hat{p})}{\Sigma n_k + 1} \tag{7}$$

Formula (7) is used rather than the estimated posterior Beta variance $\hat{p}_k(1-\hat{p}_k)/(\hat{m}_1+n_k+1)$ in order to partially account for the added uncertainty incurred by using estimates of p and m_1 instead of their true values. Finally, approximate EB confidence intervals for the p_k can be computed by treating the posterior density of p_k given the data $(x_1,..,x_N)$ as a Beta density with mean given by (6) and variance given by (7), and then using a routine which computes the inverse of a Beta distribution function. Such routines are commonly available in scientific subroutine packages such as IMSL or NAG.

By treating the r_k as outcomes of a Beta random variable with parameters r and m_2, EB estimates for the r_k can also be computed in an manner analogous to the case for the p_k. An EB estimate of the difference $d_k=(p_k-r_k)$ is then given by

$$\tilde{d}_k=\tilde{p}_k-\tilde{r}_k=E[p_k-r_k|x_1,y_1,..,x_N,y_N] \tag{8}$$

and the variance of this estimator is estimated via

$$v_{dk}=v_{pk}+v_{rk} \tag{9}$$

Confidence intervals for the differences d_k could then be computed using the probability distribution for the difference between two Beta random variables, but this would require numerical integrations and specialized software. After studying a number of cases, it was decided for this project to sidestep these difficulties by approximating the difference between two Beta random variables as a Normal random variable. That is, conditional upon available data

$(x_1,y_1,..,x_N,y_N)$, the random variable $\tilde{d}_k=(\tilde{p}_k-\tilde{r}_k)$ is treated as approximately Normally distributed with mean given by equation (8) and variance given by equation (9). Approximate EB confidence intervals can then be computed easily using the tables for the standard Normal distribution. Some results concerning the accuracy of this approximation are presented later in this report.

In the Appendix, a step-by-step description of this estimation procedure is given. The estimation procedure, along with procedures for checking the fit of the Beta prior distribution for the p_k and r_k, and procedures for evaluating the accuracy of the Normal approximation for $\tilde{d}_k$ were all implemented using MATHCAD, a type of interactive mathematical "spreadsheet". The example computations were performed using MATHCAD on an IBM 55SX microcomputer equipped with a floating-point coprocessor.

EXAMPLE APPLICATIONS

To provide a first test of these methods, two accident data sets were obtained from MNDOT. The first set consisted of accident records over the three years 1988, 1989, 1990 for the 29 signalized intersections on Minnesota Trunk Highway (MNTH) 65 running from the city of Columbia Heights northward into Anoka County. The second data set consisted of accident records over the two years 1988 and 1989 for 33 signalized intersections on MNTH 47, running parallel to and about one mile west of MNTH 65. Minnesota's accident reporting form allows the investigating officer to identify, for each driver involved in an accident, one or more actions believed to have contributed to the occurrence of the accident, and so from the above data sets we selected the records for all two-vehicle accidents for which (1) the ages of both drivers were known and (2) one driver had one or more contributing factors cited while the other had "no improper driving" cited. The driver with contributing factors was then identified as the "at-fault" driver while the other was identified as the "innocent victim". The ages of both at-fault and innocent drivers were divided into three groups: "younger" corresponding to ages 15-24, "middle" corresponding to ages 25-54, and "older" corresponding to ages 55 and over. The middle group of drivers was designated the "control" group, and the EB estimation methods described above were used to identify those intersections showing increased accident risk for older versus middle drivers and for younger versus middle drivers.

Before proceeding to the identification of the high risk locations, it was desired to check the assumption that the location specific parameters p_k and r_k were generated by Beta random

variables. Following Box *(16)* this checking was based on the marginal distributions for the data x_k and y_k obtained by integrating out the p_k and r_k from their respective joint distributions. For instance, if the Beta-binomial model actually describes the generation of the x_k, and if the parameters p and m_1 are known, then the marginal distribution of the random variable x_k is given by

$$Prob[x_k=j|m_1,p] = \binom{n_k}{j}\frac{\beta(m_1p+j,m_1(1-p)+n_k-j)}{\beta(m_1p,m_1(1-p))} \tag{10}$$

and the ML estimator $\hat{p}_k=(x_k/n_k)$ is a random variable with mean and variance given by

$$E[\hat{p}_k|m_1,p] = p$$

$$var[\hat{p}_k|m_1,p] = \frac{p(1-p)}{m_1+1}\left(1+\frac{m_1}{n_k}\right)$$

Analogous expressions hold for y_k and $\hat{r}_k$. Equations (10) and (11) suggest three possible checks on the validity of the Beta-binomial model. First, the normalized random variables

$$z_{pk} = \frac{\frac{x_k}{n_k} - p}{\sqrt{var[\hat{p}_k|m_1,p]}} \tag{12}$$

should have an expected value of 0. By replacing p and m_1 in (12) with good estimates it should

be possible to transform the observed $\hat{p}_k$ into a sample with approximately 0-mean, and this property can be tested for using the t-statistic *(17)*. Second, the normalized random variables given in (12) should also have a variance equal to 1.0, and an approximate test for this property can be done using the chi-squared statistic *(17)*. Third, if we define the random variables

S_{pj} = number of locations for which $x_k = j$,

then it is straightforward to show that the expected value for this random count is given by

$$E[S_{pj}|m_1,p] = \sum_{k=1}^{N} Prob[x_k = j|m_1,p] \tag{13}$$

where N is the number of locations in the sample. The histogram formed by computing S_{pj} for j equal to 0,1,2,...,N can then be compared to its expected form either graphically or, if N is suitably large, by using the chi-squared goodness-of-fit test.

MNTH 65 Data Set.

As noted earlier, the actual accident counts for MNTH 65 can be found in the Appendix. Table 1 displays the parameter estimates and goodness-of-fit checks for the MNTH 65 data set, while Figures 1 and 2 display the computed and estimated histograms. In Table 1, Test 1 checks whether or not the normalized scores z_{pk}, z_{rk} have zero mean, Test 2 checks whether or not z_{pk} and z_{rk} have variance equal to 1.0 and Test 3 checks the fit of the histograms S_{pj}, S_{rj} by the expected values given in (13). From Table 1 it can be seen that for both the older driver data and the younger driver data, the normalized scores z_{pk} and z_{rk} were not different from samples with 0-mean and variance equal to 1. The chi-squared tests comparing the computed and

TABLE 1 (MNTH65)

Younger Driver Data

	Involvement Parameters (p_k)		Exposure Parameters (r_k)	
	p=.408	m_1=36.9	r=.274	m_2=39.4
Test 1	t=-0.38	p>.10	t=-0.27	p>.10
Test 2	X^2=23.0	p>.10	X^2=21.4	p>.10
Test 3	X_2=3.61	p>.10	X^2=1.64	p>.10

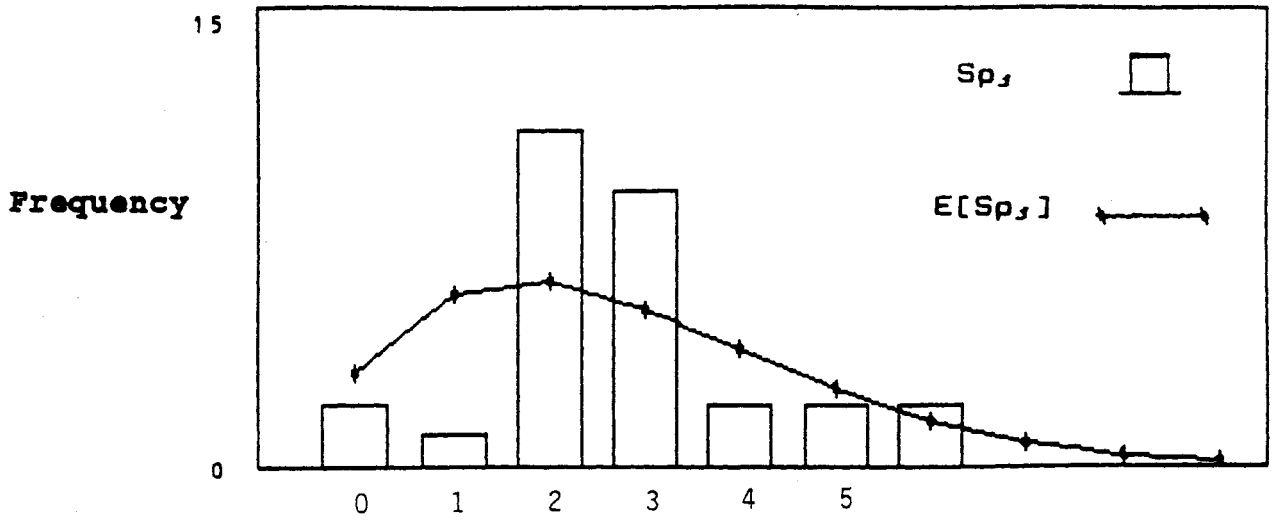
Older Driver Data

	Involvement Parameters (p_k)		Exposure Parameters (r_k)	
	p=.240	m_1=46.5	r=.192	m_2=16.9
Test 1	t=0.39	p>.10	t=-0.07	p>.10
Test 2	X^2=21.4	p>.10	X^2=24.2	p>.10
Test 3	X^2=11.69	p<.01	X^2=1.02	p>.10

FIGURE 1

Observed and Expected Histograms for MNTH65

Older Driver Data



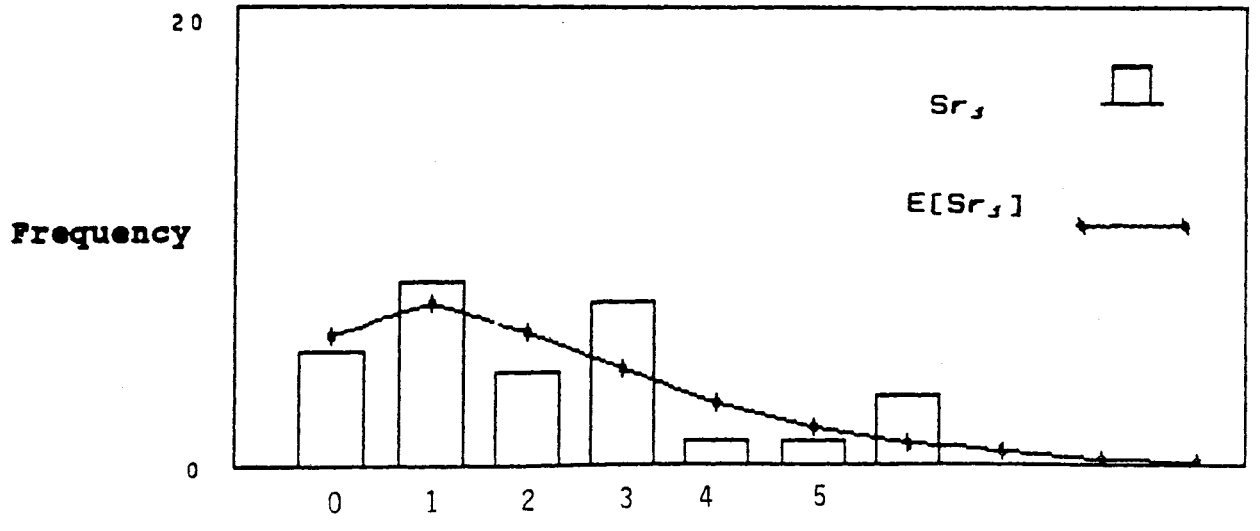
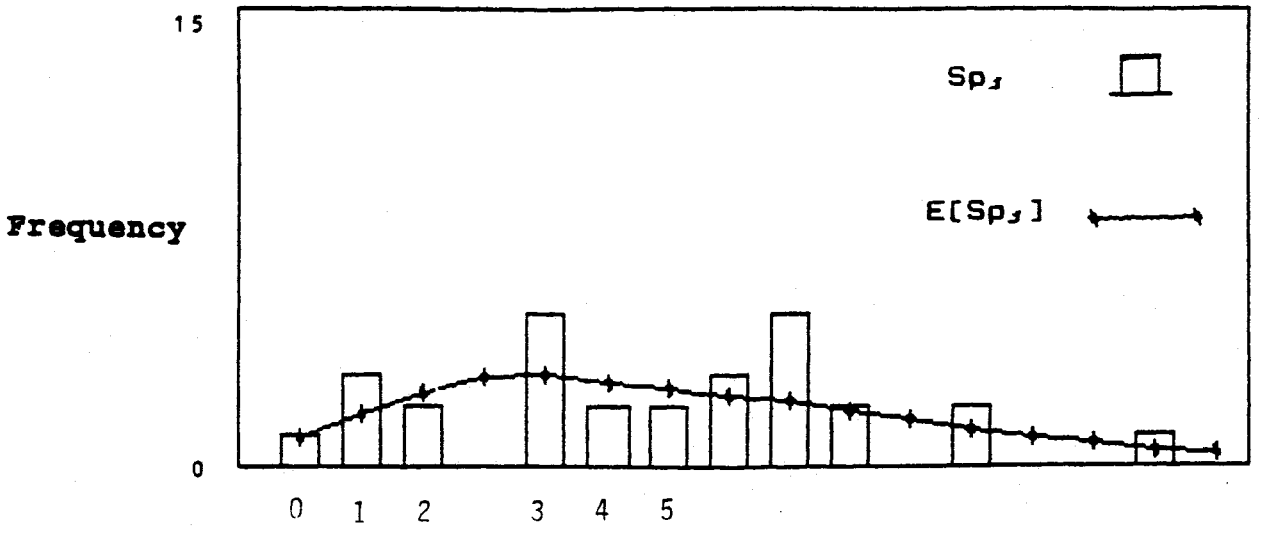
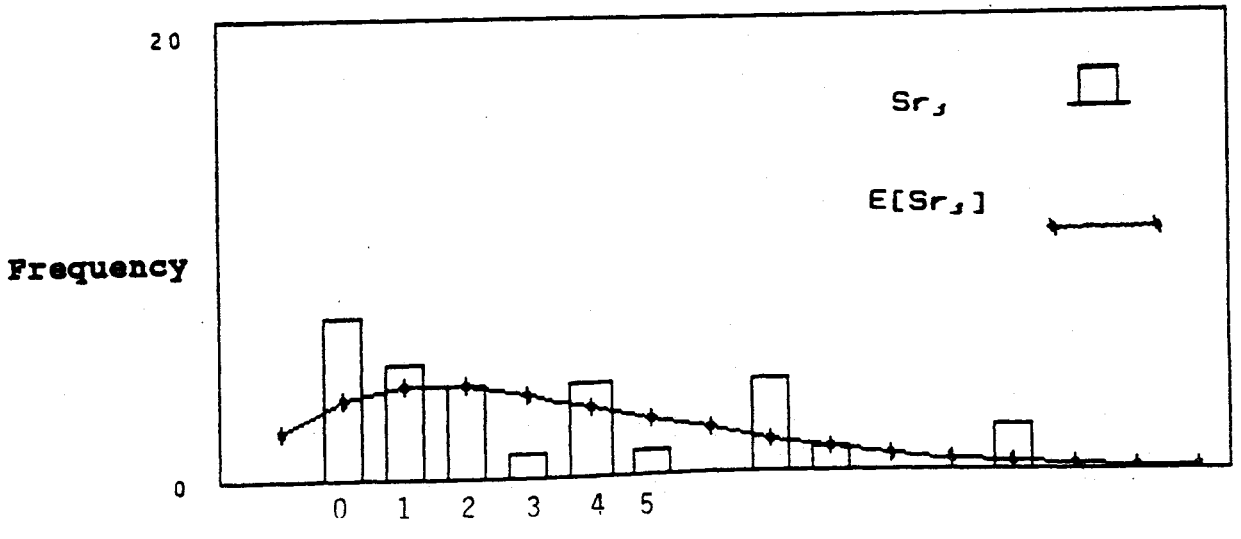


FIGURE 2

Observed and Expected Histograms for MNTH65

Younger Driver Data





estimated histograms should be considered only very approximate tests, due to the small values of N and the tendency of some expected counts to have values less than 5.0 *(17)*. Nevertheless, for three of the four histograms, the difference between the observed and the expected counts appears to be nonsignificant, while there is some lack of fit shown by the x_k obtained from the older driver data. Inspection of Figure 1 indicates that this lack of fit is caused by overcounts at the values 2 and 3 and undercounts at the values 1 and 4, but otherwise the fit is reasonably good, so that overall it appears that the assumption of Beta-generated p_k and r_k is plausible for this data set.

Next, in order to check the accuracy of the Normal approximation used in computing EB confidence intervals, the upper and lower bounds of a nominal 90% confidence interval were computed using the Normal approximation for each intersection and each data set. Using numerical integration, it was then possible to compute the confidence level which Albert's Beta approach would assign to these same intervals. Table 2 displays the computed Beta confidence levels for the two data sets. In almost all cases, the difference between the nominal and computed confidence levels is less than or equal to two percentage points, and it is concluded that the Normal approximation showed acceptable accuracy.

Finally, Figure 3 displays the EB interval estimates for the difference (p_k-r_k) for each intersection along with the ML estimated differences $(\hat{p}_k-\hat{r}_k)$. In all cases, the confidence interval is an approximate 90% interval computed using the Normal approximation. Inspection of Figure 3 shows first that the EB estimates have considerably less scatter than do the ML estimates, and that the EB estimates eliminate certain counter-intuitive cases from consideration (such as intersection 4, where the ML estimate indicates that middle drivers have higher accident rates

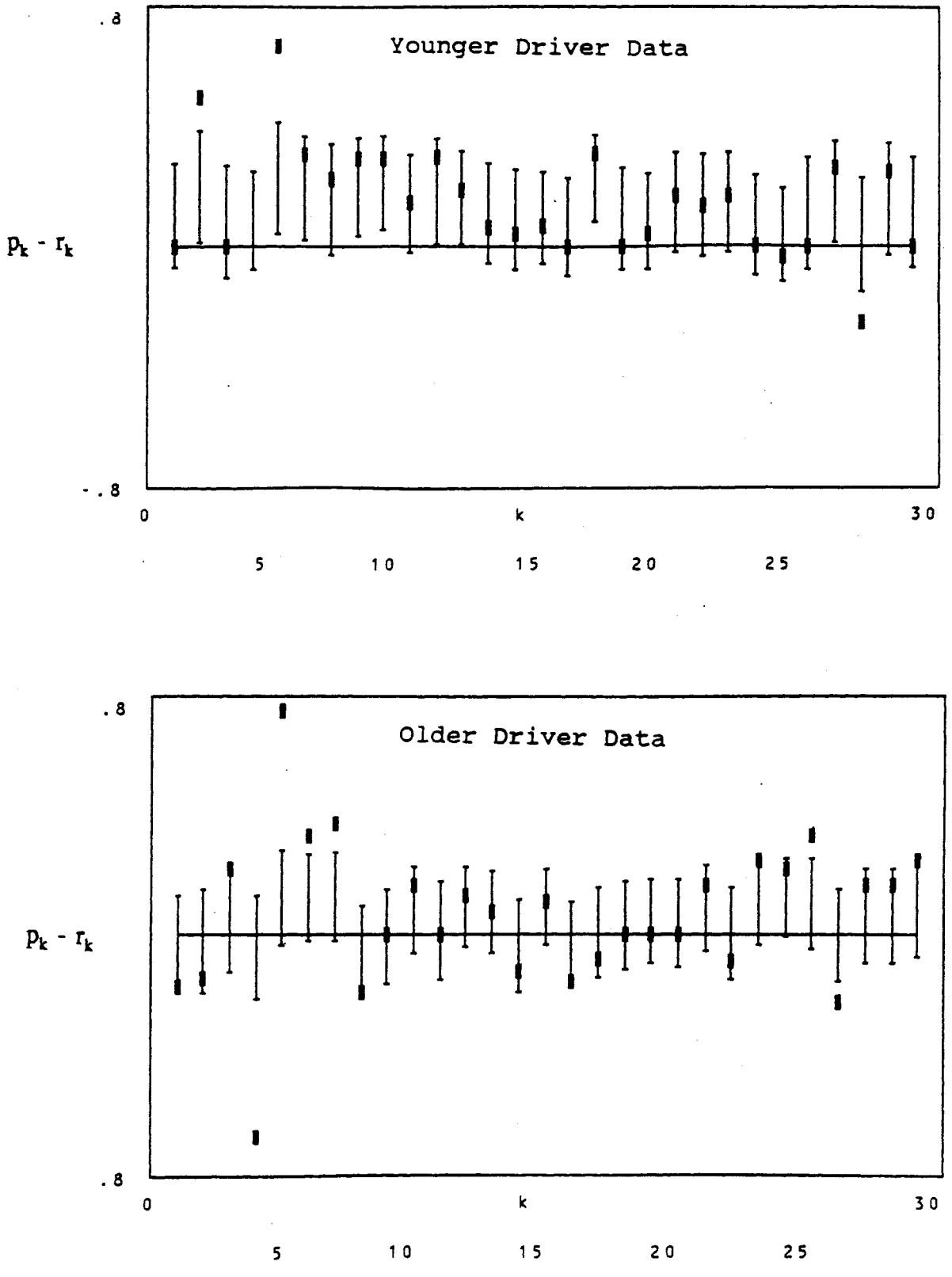
TABLE 2

Beta-Derived Confidence Levels for Nominal 90% Confidence Intervals

Intersection	Younger Driver Data	Older Driver Data
1	.893	.899
2	.901	.897
3	.904	.884
4	.911	.897
5	.901	.908
6	.896	.893
7	.905	.886
8	.911	.907
9	.904	.887
10	.910	.895
11	.913	.915
12	.901	.889
13	.885	.903
14	.906	.912
15	.902	.907
16	.889	.888
17	.914	.894
18	.905	.905
19	.915	.895
20	.902	.903
21	.901	.894
22	.902	.910
23	.883	.891
24	.879	.899
25	.905	.898
26	.898	.888
27	.899	.900
28	.902	.900
29	.882	.892

FIGURE 3

Empirical Bayes and Maximum Likelihood Estimates of $p_k - r_k$ (MNTH65)



than do older drivers). Second, the tendency for younger drivers to have higher accident rates seems to be a somewhat pervasive feature of the entire roadway segment, while the increased accident rates for older drivers, if present at all, appear to be localized around intersections 5-7 and intersections 23-25. Assuming that some older driver-oriented improvement of this roadway was desirable, these two sections would be candidates for first consideration.

MNTH 47 Data Set.

Table 3 displays the parameter estimates and goodness of fit checks computed for both the older driver and the younger driver data set taken from MNTH 47, while Figures 4 and 5 show the observed and estimated frequency distributions for the counts x_k and y_k. Here, the fit provided by the assumption of Beta-generated p_k and r_k is even better than that observed for the MNTH 65 data, with none of the checks providing evidence contradicting the assumption. As with the MNTH 65 data set, the Normal approximation for the estimated differences $\tilde{d}_k$ was checked by comparing the confidence levels computed using Albert's Beta approach to nominal 90% confidence intervals computed using the Normal approximation. Table 4 shows these results, and again, the Normal approximation appears to be quite acceptable. Finally, Figure 6 shows the EB confidence intervals for both the older and younger driver data sets, along with the corresponding ML point estimates. Again, the EB estimates show less scatter than do the ML estimates, and again there appear to be more risky locations for younger drivers than there are for older drivers. In this data set however, no locations show clearly elevated risk for older drivers, and this may primarily be due to the small numbers of older driver accidents actually observed. For the younger drivers, intersections 10, 14, and 19 appear to be the most dangerous.

TABLE 3 (MNTH47)

Younger Driver Data

	Involvement Parameters (p_k)		Exposure Parameters (r_k)	
	p=0.439	m_1=27.46	r=0.3	m_2=23.1
Test 1	t=0.08	p>.10	t=0.77	p>.10
Test 2	X^2=27.9	p>.10	X^2=27.8	p>.10
Test 3	X^2=.20	p>.10	X^2=3.15	p>.10

Older Driver Data

	Involvement Parameters (p_k)		Exposure Parameters (r_k)	
	p=0.226	m_1=21.5	r=0.193	m_2=10.07
Test 1	t=0.09	p>.10	t=-0.9	p>.10
Test 2	X^2=24.5	p>.10	X^2=25.7	p>.10
Test 3	X^2=.46	p>.10	X^2=.31	p>.10

TABLE 4

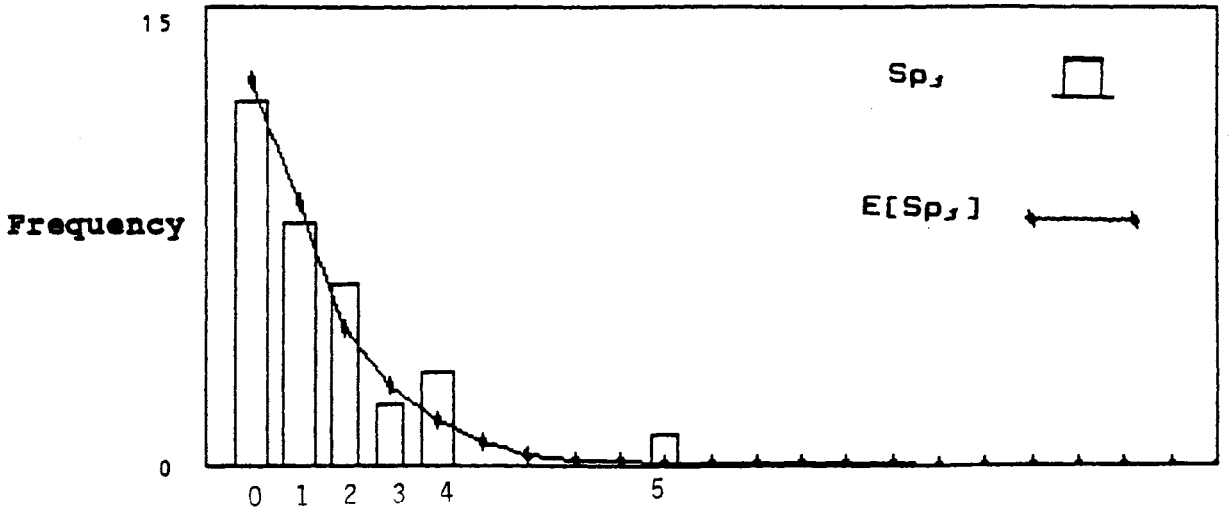
Beta-Derived Confidence Levels for Nominal 90% Confidence Intervals

Intersection	Younger Driver Data	Older Driver Data
1	.892	.909
2	.903	.902
3	.899	.9
4	.905	.892
5	.894	.898
6	.889	.896
7	.884	.913
8	.887	.903
9	.892	.889
10	.912	.906
11	.898	.901
12	.892	.887
13	.891	.903
14	.908	.905
15	.885	.905
16	.899	.893
17	.893	.906
18	.912	.898
19	.9	.914
20	.912	.909
21	.892	.901
22	.904	.909
23	.9	.905
24	.893	.895
25	.908	-
26	.891	.909
27	.904	.898
28	.9	.887
29	.891	.901
30	.885	.887
31	.892	.909
32	.896	.887
33	.891	.909

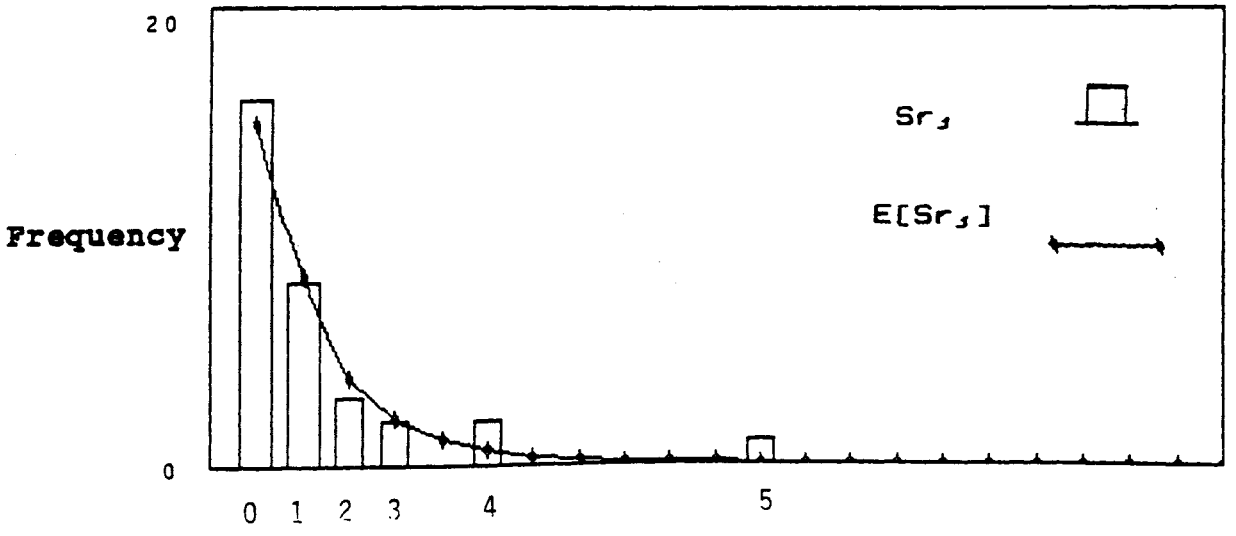
FIGURE 4

Observed and Expected Histograms for MNTH47

Older Driver Data



Intersections with j guilty

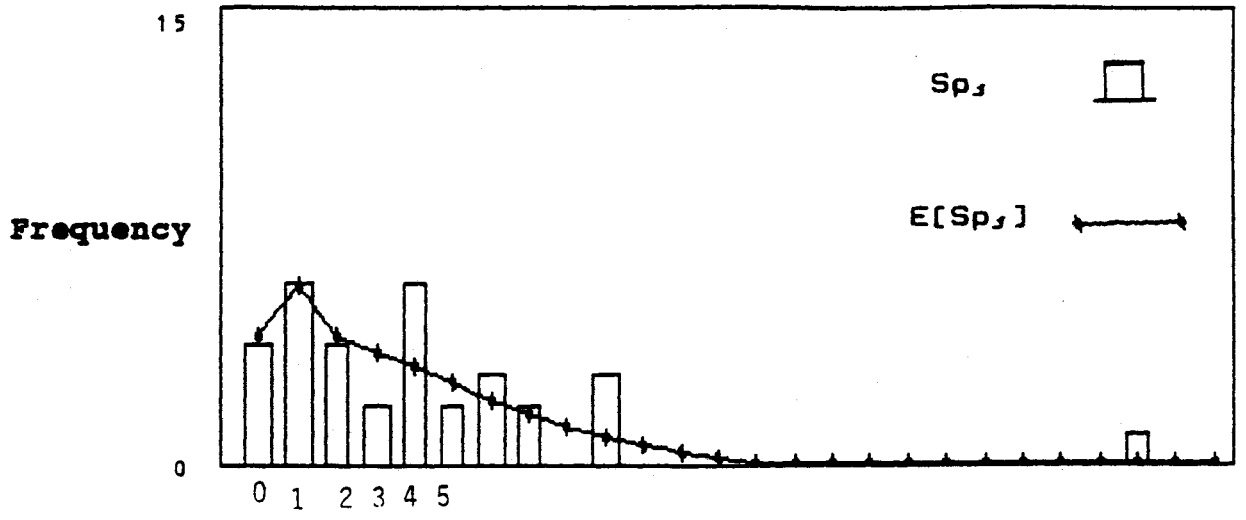


Intersections with j innocent

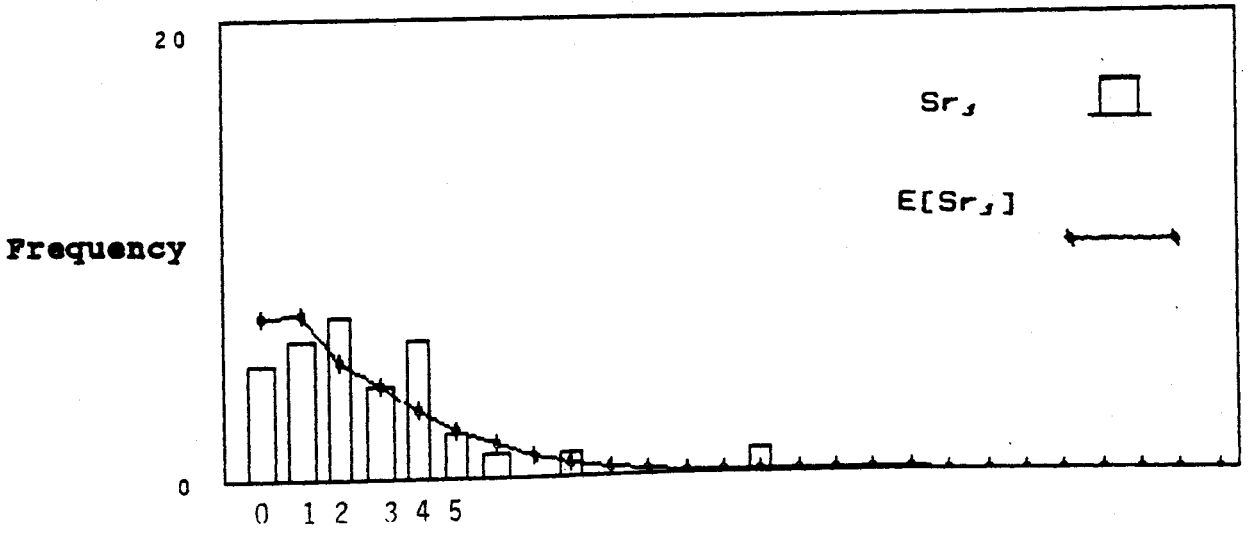
FIGURE 5

Observed and Expected Histograms for MNTH47

Younger Driver Data



Intersections with j guilty

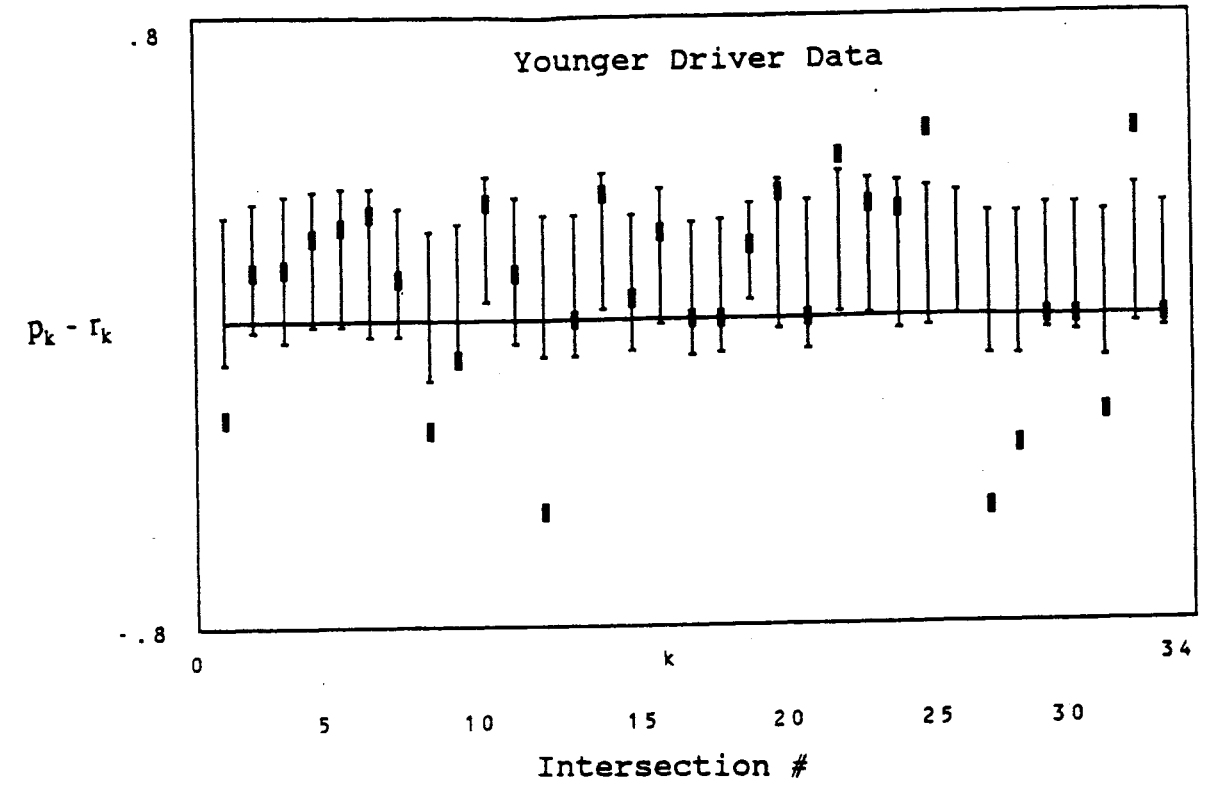


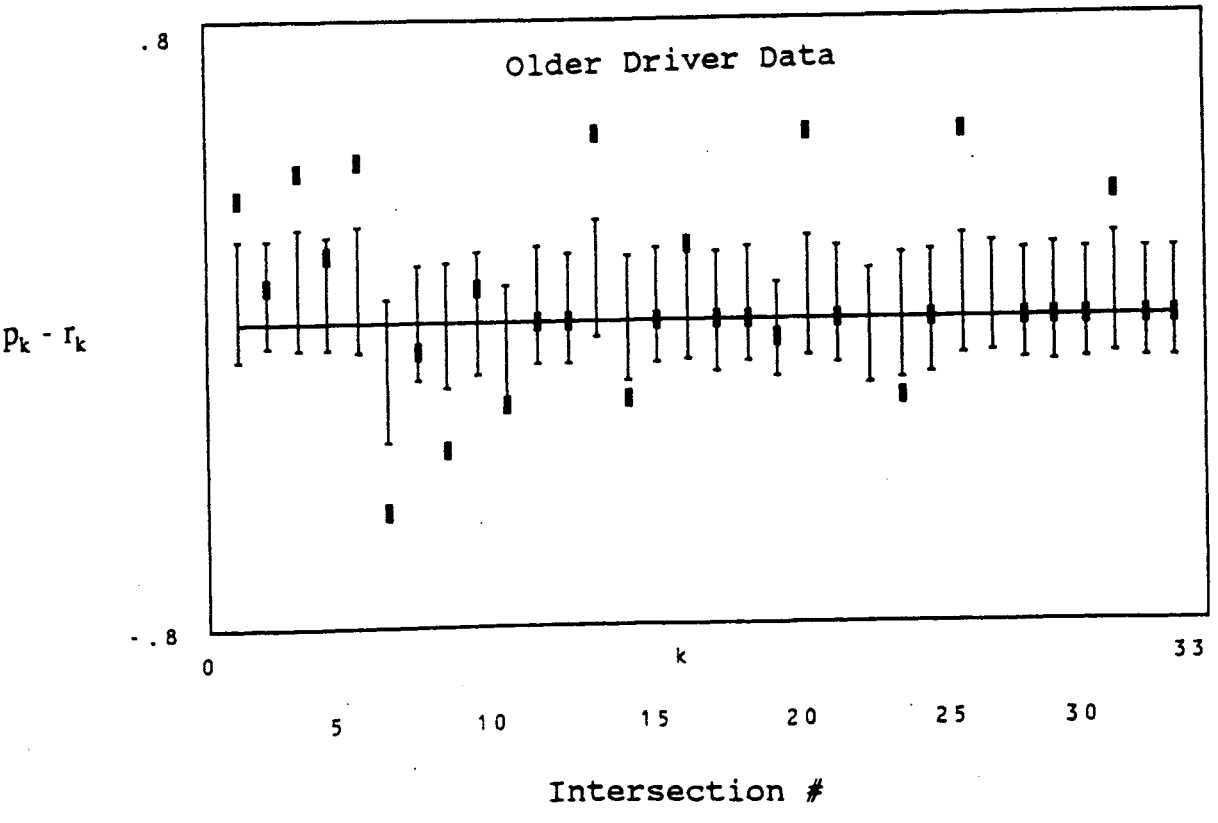
Intersections with j innocent

FIGURE 6

Empirical Bayes and Maximum Likelihood Estimates of p_k-r_k

(MNTH47)





CONCLUSION AND RECOMMENDATIONS

This report presented a statistical method for location-specific testing of the equality of accident rates experienced by two different groups of drivers. In order to side-step the need for location and group-specific measures of exposure, the method employs an induced exposure model, and in order to improve the estimation in the face of the RTM effects inherent in the small samples generally available, it employs an Empirical Bayesian estimation procedure. The most computationally demanding feature of the method is the combined numerical integration and univariate line-search needed to compute MAP estimates of the parameters m_1 and m_2. Since a closed form expression can be given for the ratio of the Beta integrals appearing in equation (6), this optimization problem can be solved on a microcomputer using either commonly available computer languages or by using commercially available "Mathematical Spreadsheet" software such as MATHCAD. All other computations require no more than a hand calculator. Preliminary tests of the method have been completed using two actual data sets, with all computations being performed on a microcomputer. For both data sets, the method's underlying statistical assumptions appeared to be supported and the approximate confidence intervals based on the Normal distribution appeared acceptable. The method itself tended to reduce the scatter shown by maximum likelihood estimates while still identifying those locations showing higher risk for the age group of interest.

Overall, this combination of induced exposure and empirical Bayesian statistics promises to provide a useful addition to the safety engineer's tool kit. Before it receives recommendation

for widespread use however some additional work needs to be done. First, the performance of the method should be tested by using it on several more actual data sets. This would give a better sense for the range of conditions likely to be encountered in practice and also provide a realistic basis for conducting Monte Carlo simulation studies. Second, a combination of analytic and simulation experiments are needed to better delineate the statistical properties of the method. Especially important are robustness properties with respect to deviations from the method's underlying assumptions. Finally, an efficient and user friendly computer code needs to be developed. It should be possible to develop a standalone FORTRAN program implementing the method which would run on IBM-type microcomputers 80386 or better CPUs, and such software would make the method accessible to a much larger group of users than would the current MATHCAD implementation.

REFERENCES

1. Transportation Research Board (1988) *Transportation in an Aging Society*, Special Report 218, Transportation Research Board, Washington, DC.

2. McKelvey, F., Maleck, T., Stamatiadis, N. and Hardy, D. (1987) Highway Accidents and Older Drivers, *Transportation Research Record*, **1172**, pp. 47-56.

3. *Minnesota Roadway Safety Initiatives for 1990 and Beyond*, (1989) Draft Report by Minnesota Dept. of Transportation, Office of Traffic Engineering, St. Paul, MN.

4. Zegeer, C. (1982) *Highway Accident Analysis Systems*, Transportation Research Board, Washington, DC.

5. Hauer, E. (1980) Bias-By-Selection: Overestimation of the Effectiveness of Safety Countermeasures Caused by the Process of Selection for Treatment, *Accident Analysis and Prevention*, **12**, 113-117.

6. Pendleton, O. and Morris, C. (1990) A New Method for Accident Analysis, presentation at 69th Meeting of Transportation Research Board, Washington, DC.

7. Morris, C. (1983) Parametric Empirical Bayes Inference: Theory and Applications, *J. of American Statistical Society*, **78**, 47-65.

8. Carlin, P. and Gelfand, A. (1990) Approaches for Empirical Bayes Confidence Intervals, *J. American Statistical Association*, **85**, pp. 105-114.

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

10. Thorpe, J. (1964) Calculating Relative Involvement Rates in Accidents without Determining Exposure, *Australian Road Research*, **2**, pp. 25-36.

11. Haight, F. (1970) A Crude Framework for Bypassing Exposure, *J. of Safety Research*, **2**,

pp. 26-29.

12. Joksch, H. (1973) A Pilot Study of Observed and Induced Exposure To Traffic Accidents, *Accident Analysis and Prevention,* **5**, pp. 127-136.

13. Koonstra, M. (1973) A Model for Estimation of Collective Exposure and Proneness from Accident Data, *Accident Analysis and Prevention,* **5**, 157-173.

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

15. Albert, J. (1984) Empirical Bayes Estimation of a Set of Binomial Probabilities, *J. of Statistical Computing and Simulation*, **20**, 129-144.

16. Box, G. (1983) An Apology for Ecumenism in Statistics, in Box, G. et al. (Eds.) *Scientific Inference, Data Analysis and Robustness*, New York, Academic Press, 51-84.

17. Rohatgi, V. (1976) *An Introduction to Probability Theory and Mathematical Statistics*, Wiley and Sons, New York.

18. Bazaraa, M. and Shetty, C. (1979) *Nonlinear Programming: Theory and Algorithms,* New York, Wiley and Sons.

APPENDIX

This Appendix describes the development of an estimator for the parameter m_1, with the estimator for m_2 following in an exactly analogous manner. First, given the parameters p and m_1, the data x_k k=1,..,N are assumed to be generated by a two-step process, in which the relative involvement parameters p_k are assigned to locations as the independent outcomes of a Beta random variable parameterized by p and m_1, and then the x_k are the outcomes of binomial random variables with parameters p_k and n_k. Now if one does not know the actual values of the p_k, it is straightforward to show that the likelihood of the sample $x_1,..,x_N$ given m_1, p and n_k is

$$Prob[x_1,..,x_N] = \prod_{k=1}^{N} \binom{n_k}{x_k} \frac{\beta(m_1 p + x_k, m_1(1-p) + n_k - x_k)}{\beta(m_1 p, m_1(1-p))} \qquad \text{(A1)}$$

Next, it is assumed that the prior uncertainty concerning the values of m_1 and p can be described by the noninformative prior

$$g(p,m_1) = \frac{1}{m_1 p(1-p)} \qquad \text{(A2)}$$

Multiplying (A1) by (A2) and then integrating with respect to p gives the posterior distribution of m_1 given the data $x_1,..,x_N$. Replacing the symbol m_1 by m gives

$$f(m) = \frac{\int \prod_{k=1}^{N} \left[\binom{n_k}{x_k} \frac{\beta(mp+x_k, m(1-p)+n_k-x_k)}{\beta(mp, m(1-p))} \right] \frac{dp}{p(1-p)}}{m} \qquad \text{(A3)}$$

which is equation (6) of the text. The value of m which maximizes f(m) is identical to the value of m which minimizes h(m)=-ln(f(m)), and this minimum can be found using the following well-known dichotomous line search algorithm *(18)*:

(0) Let l_1 denote a lower bound for the value of m, let u_1 be an upper bound for the value of m and let e denote the accuracy desired for the final estimate, let k=1;

(1) If $u_k - l_k < 2e$, stop; the minimizing value lies between $[l_k, u_k]$. Otherwise compute

$$\lambda_k = (l_k + u_k)/2 - e; \quad \mu_k = (l_k + u_k)/2 + e$$

(2) if $h(\lambda_k) < h(\mu_k)$, let $l_{k+1} = l_k$, $u_{k+1} = \mu_k$; otherwise let $l_{k+1} = \lambda_k$, $u_{k+1} = u_k$. Let k=k+1 and go to step (1).

If the function h(m) has only one minimum within the initial interval $[l_1, u_1]$, the above algorithm will find it to the specified degree of accuracy. The only difficulty now involves evaluating the function h(m), or equivalently f(m).

The integral shown in (A3) has no closed form, so computation of f(m) requires numerical integration, i.e. numerous evaluations of the integrand. Naive implementation of (A3) as written would in turn require that each evaluation of the integrand would in turn require numerical computation 2N Beta integrals, a task which could prove too demanding for microcomputers. Fortunately, with some algebraic manipulations it can be shown that the ratio of the Beta integrals has a closed-form expression

$$\frac{\beta(mp+x_k, m(1-p)+n_k-x_k)}{\beta(mp, m(1-p))} = \frac{\prod_{i=0}^{x_k-1}(mp+i)\prod_{i=0}^{n_k-x_k-1}(m(1-p)+i)}{\prod_{i=0}^{n_k-1}(m+i)} \qquad \text{(A4)}$$

so that only one univariate numerical integration is needed for each evaluation of f(m).

The MATHCAD implementation of this estimator required around 20 minutes of real-time to compute both estimates m_1 and m_2 for an accident data set. Each estimation required about 20 iterations of the line search to produce an estimate accurate to e=.05 when starting with $[l_1,u_1]=[10,100]$. For an equivalent 386SX computer without a floating-point coprocessor, the computation could require one or more hours. Since the major computational demand is evaluation of the function f(m), any technique which reduces the number of evaluation required will improve performance. In particular, using a more efficient line-search method such as Golden-section or Fibonacci search could yield noticeable improvements. This possibility is currently being investigated.

TABLE A1

Two-Vehicle Accidents At Signalized Intersections Along MNTH 65

Older data for HWY 65

Cross. Streets	X_{mm}	X_{mo}	X_{om}	X_{oo}
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	0
Osborne Rd.	8	5	3	1
81st Ave. NE	14	1	3	0
85th Ave. NE	13	4	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 A2

Two-Vehicle Accidents At Signalized Intersections Along MNTH 65

Younger data for HWY 65

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

TABLE A3

Two-Vehicle Accidents At Signalized Intersections Along MNTH 47

Older data for HWY 47

Cross. Streets	X_{mm}	X_{mo}	X_{om}	X_{oo}
40th Ave. NE	2	0	1	0
44th Ave. NE	9	0	1	0
49th Ave. NE	3	0	2	0
53rd 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
177th Ave. NW	2	0	1	0
177th LN NW	1	0	0	0
CSAH 22	1	0	1	0

TABLE A4

Two-Vehicle Accidents At Signalized Intersections Along MNTH 47

Younger data for HWY 47

Cross. Streets	X_{yy}	X_{ym}	X_{my}	X_{mm}
40th Ave. NE	1	0	1	2
44th Ave. NE	0	4	2	9
49th Ave. NE	1	2	1	3
53rd Ave. NE	0	2	0	7
57th Ave. NE	1	6	3	2
61st Ave. NE	3	3	1	0
Mississippi St.	1	5	3	9
69th Ave. NE	3	0	2	2
73rd Ave NE	0	2	3	5
Osborne Rd	1	8	2	8
81st Ave. NE	2	2	1	3
83rd Ave. NE	0	0	1	1
CR 132 LT	1	4	4	3
USTH 169	2	7	2	4
CSAH 7	7	2	1	7
CSAH 9	3	4	1	5
MN 242	2	4	4	7
Hanson Blvd.	1	3	3	8
Foley Blvd.	8	15	5	23
TH 10 at 610	1	3	1	1
B.N. INCR. ANOKA	1	1	1	2
Industry Blvd.	0	4	1	2
Nowthen Blvd.	0	4	1	5
151st Ave.	2	3	1	1
153rd Ave.	1	1	0	0
156th Ave.	0	1	0	0
160th Ave.	1	0	1	0
167th Ave.	0	0	1	2
170th Ave. NW	0	0	0	1
Green Valley Rd.	1	0	0	0
177th Ave. NW	1	0	1	2
177th LN NW	0	1	0	1
CSAH 22	0	0	0	1