Statistical Method for Identifying Locations of High Crash Risk to Older Drivers

GARY A. DAVIS AND KONSTANTINOS KOUTSOUKOS

Effective use of finite roadway improvement budgets to accommodate an increasing number of older drivers requires that we be able to identify 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 measure of exposure coupled with the relatively small accident samples available for particular locations makes the standard methods of high-hazard identification inapplicable. The way in which, 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 arising from a particular type of contingency table is described. 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 is described next. The resulting computational procedures are relatively straightforward and can be implemented on a microcomputer. The method is illustrated using accident data for a set of signalized intersections located on a Minnesota highway.

It is a well-established demographic fact that individuals born between 1947 and 1957 constitute a substantial fraction of the current U.S. population, and as these "baby-boomers" age, older drivers will make up 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, so that roadway managers have begun 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 older drivers actually have more difficulty with the existing roadway than do younger drivers and, second, on being able to reliably identify locations that actually cause the difficulty. There is mounting evidence that after about the age of 55 or 60, the accident rate for drivers tends to increase (*1*,*2*). The problem of identifying locations showing increased accident rates for older drivers is the subject of this paper.

G. A. Davis, Department of Civil and Mineral Engineering, University of Minnesota, 500 Pillsbury Drive SE, Minneapolis, Minn. 55455. K. Koutsoukos, North Central Texas Council of Governments, P.O. Box 5888, Arlington, Tex. 76005.

The Minnesota Department of Transportation (MNDOT) has projected that the proportion of older drivers in the state's driving population will increase during the next 20 years and that older drivers appear to be overrepresented 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 authors of this paper began a research project aimed at identifying locations where older drivers were overrepresented in the accident records, 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. In response to this problem, we have been able 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.

CLASSICAL AND EMPIRICAL BAYESIAN HAZARD IDENTIFICATION

Traffic accidents are fortunately "rare events" compared with the amount of travel done by the population generating 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 1 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 is the accident rate at location k and E_k is the exposure of the traveling 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, with standard measures of exposure used in traffic accident analysis being 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

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

"High-hazard" locations can be identified by computing the estimated rate for each location of interest and then selecting those locations where the estimated accident rates are large compared with some regionwide average. Two common methods for high-hazard identification, the accident rate method and the rate quality control method (*4*), are based on the estimator of Equation 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) (*5*). 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 EB statistical techniques, and this work has reached a useful degree of maturity in the EBEST methodology developed at the Texas Transportation Institute (*6*). This methodology begins with a Bayesian 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, 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

$$\lambda_k^* = \left(\frac{E_k}{E_k + \varepsilon}\right)\hat{\lambda}_k + \left(\frac{\varepsilon}{E_k + \varepsilon}\right)\lambda \tag{2}$$

The Bayes estimator is a convex combination of the ML estimator given in Equation 1 and the underlying gamma mean. For those sites with high exposures (and hence lower variances for λ_k) the Bayes estimator tends to weight the ML estimator more heavily, whereas sites with low exposures are "shrunk" more toward the gamma mean. The parameter ε measures the degree of relatedness among the site-specific accident rates, with $\varepsilon = 0$ being the case where the individual λ_k have no relation (so that $\lambda_k^* = \hat{\lambda}_k$), whereas $\varepsilon = \infty$ corresponds to the case $\lambda_k = \lambda$ (i.e., all the individual accident rates are equal to the gamma mean). For intermediate values of ε, the Bayes estimators will tend to be closer (in the mean-square sense) to the true accident rates than will the ML estimator. In most practical situations however, the values of λ and ε will be unknown and also require estimation. EB methods attempt to capitalize on the desirable properties of the Bayes estimator by first replacing ε and λ in Equation 2 with efficient estimates, such as ML estimates, and second, by accounting for the increased uncertainty that results from having less than perfect knowledge of these parameters (*7,8*).

INDUCED EXPOSURE MODEL

In principle, the preceding method of analysis could be extended to the identification of locations where older drivers are overrepresented 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, and
n_{ik} = observed number of accidents for age group i at location k (n_{ik} is 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 overrepresented (indicated by high values of m_{ik}) but also attribute the overrepresentation to overexposure (indicated by high values of E_{ik}), greater risk (indicated by high values of λ_{ik}), or a combination of these effects. This 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, whereas the relationship between ADT estimates and the resulting estimates of accident risk such as Equation 1 are even less clear. 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 that 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, whereas the other is an innocent victim. At-fault drivers are assumed to be subject to accidents according to the above Poisson model, whereas 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 at different types of intersection (*2*). The Michigan methods are essentially deterministic procedures, however, and though they are useful when dealing with large aggregations of accident records, the 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 and λ_{ok} and E_{ok} denoting 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; and
n_k = the total number of two-vehicle accidents at location k.

Under the induced exposure hypothesis, r_k gives the probability that a driver who has an accident "selects" an older driver as the victim, and 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, and 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 two age groups have the same accident rate at a given location reduces to the problem of testing whether two binomial probabilities arising from a cross-classification table are equal. The ML estimators of p_k and r_k are given by

$$\left.\begin{aligned} \hat{p}_k &= \frac{x_k}{n_k} \\ \hat{r}_k &= \frac{y_k}{n_k} \end{aligned}\right\} \qquad (3)$$

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 above 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. 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 a confidence interval for the difference $(p_k - r_k)$ contains the value zero. Before proceeding to methods for computing these confidence intervals, we note that the ratio p_k/r_k can be interpreted as the "involvement ratio" used in other studies using induced exposure methods (*2*). Our preference for the difference $(p_k - r_k)$ stems from the fact that, as will be shown later, the probability distribution of this difference can be readily approximated by a normally distributed random variable, allowing the use of z tables in determining probability values. The distribution of the ratio p_k/r_k, on the other hand, is less tractable.

EB ESTIMATION FOR THE INDUCED EXPOSURE MODEL

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 outcomes of a beta random variable with mean value p and variance $p(1 - p)/(m_1 + 1)$. Given p_k and n_k, x_k is then assumed to be a binomial outcome. If the parameters p and m_1 are known, the Bayes estimator of p_k takes the form

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

Albert then places "noninformative" prior distributions on the parameters p and m_1, producing a three-step pure Bayesian procedure. In principle, all quantities of interest, such as point and interval estimates, can be computed through integration of this Bayesian model's full joint probability distribution, but a computationally simpler approach results by using an EB estimator of the form

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

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 our problem would apply the same degree of shrinkage to each location regardless of the individual accident counts, we propose estimating m_1 as the value that maximizes the function

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

Here $\beta(a, b)$ denotes the beta integral evaluated at the values a and b, and since the function $f(m)$ is proportional to the posterior probability density of m_1 based on Albert's noninformative prior distributions for the parameters p and m_1 (*15*, p. 137), $\hat{m}_1$ is, in fact, 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

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

Formula 7 is used rather than the estimated posterior beta variance $(\tilde{p}_k(1 - \tilde{p}_k)/(\hat{m}_1 + n_k + 1)$ 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 $(x_1, \ldots, x_N)$ as a beta density with mean given by Equation 6 and variance given by Equation 7, and then using a routine that 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 be computed in a manner analogous to the p_k case. An EB estimate of the difference $d_k = (p_k - r_k)$ is then given by

$$\tilde{d}_k = \tilde{p}_k - \tilde{r}_k \approx E[p_k - r_k | x_1, y_1, \ldots, x_N, y_N] \quad (8)$$

and the variance of this estimator is estimated via

$$v_{dk} = v_{pk} + v_{rk} \quad (9)$$

Confidence intervals for the differences d_k could now be computed using the probability distribution for the difference between two beta random variables, but the resulting need for numerical integration and special software can be avoided by exploiting the fact that the difference between two beta random variables is approximately a normal random variable. That is, conditional upon available data $(x_1, y_1, \ldots, x_N, y_N)$, the random variable $d_k = (p_k - r_k)$ is 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 standard normal distribution.

EXAMPLE APPLICATION

To illustrate these methods, we present the following example. Accident records for the years 1988–1990 were obtained from MNDOT for the 29 signalized intersections on Minnesota Trunk Highway (MNTH) 65 running from the city of Columbia Heights northward into Anoka County. 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 data set we selected the records for all two-vehicle accidents for which (*a*) the ages of both drivers were known and (*b*) one driver had one or more contributing factors cited and the other had "no improper driving" cited. The driver with contributing factors cited was then identified as the "at-fault" driver, and 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 or more, and EB estimation methods were used to identify locations of increased risk both for older versus middle drivers and for younger versus middle drivers. All computations were performed using MATHCAD, an interactive formula processing program, on an IBM PS/2 55SX microcomputer. The most computationally demanding task was maximization of the function $f(m)$ to produce MAP estimates of the parameters m_1 and m_2. This was done by using a closed form expression for the ratio of beta integrals appearing in Equation 6:

$$\frac{\beta(mp + x_k, m(1-p) + n_k - x_k)}{\beta(mp, m(1-p))} = \frac{\prod_{i=0}^{x_k - 1} (mp + i) \prod_{i=0}^{n_k - x_k - 1} (m(1-p) + i)}{\prod_{i=0}^{n_k - 1} (m + i)} \quad (10)$$

This permitted use of a univariate numerical integration routine to compute the right-hand side of Equation 6 for any given value of m. Maximization of this expression with respect to m was then accomplished using a dichotomous line-search method.

Before proceeding to the identification of the high-risk locations, we believed it desirable to test whether the assumption that p_k and r_k are generated by beta random variables was in fact plausible for this data set. Following Box (*16*) these tests were 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, the marginal distributions of the random variables x_k are given by

$$P_k(j) = \text{Prob}[x_k = j | m_1, p] = \binom{n_k}{j} \frac{\beta(m_1 + j, m_1(1-p) + n_k - j)}{\beta(m_1 p, m_1(1-p))} \quad (11)$$

whereas the means and variances of the ML estimates $\hat{p}_k = (x_k/n_k)$ are given by

$$\left. \begin{aligned} E[\hat{p}_k | m_1, p] &= p \\ \text{var}[\hat{p}_k | m_1, p] &= \frac{p(1-p)}{(m_1 + 1)} \left(1 + \frac{m_1}{n_k} \right) \end{aligned} \right\} \quad (12)$$

TABLE 1 Parameter Estimates and Goodness-of-Fit for Younger Driver Data

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

Now in principle, the predictive distribution, Equation 11, and its consequences provide a means for checking the adequacy of an underlying statistical model, but, in practice, the theory for model checking is less well understood than that for parameter estimation and hypothesis testing (*17*). Fortunately, it is still possible to provide some rough tests of model adequacy. First, if the underlying beta model is valid, then by replacing the p and m_1 in Equation 12 with estimates, it should be possible to transform the sequence of $\hat{p}_k$ (and similarly the sequence of $\hat{r}_k$) into a sequence of random variables with means equal to 0 and variances equal to 1, and tests for these properties can be performed using standard t and chi-squared statistics (*18*). Second, Box (*16*) has suggested that the adequacy of a beta-binomial model, such as that used here, could be checked using the statistic

$$S_{pk} = \text{Prob}\,\{j{:}P_k(j) \leq P_k(x_k)\} \qquad (13)$$

where $P_k(j)$ is the predictive distribution given in Equation 11. From the definition given in Equation 13, it follows that S_{pk} attains its maximum value of 1.0 when x_k equals the mode of the predictive distribution, and that, under the null hypothesis that x_k is an outcome of the predictive distribution, S_{pk} is its own significance level. For example, $S_{pk} = .05$ can be interpreted as meaning that if x_k actually follows the predictive distribution, the probability of obtaining a value of S_{pk} less than or equal to .05 by chance is equal to .05. Computing the statistic S_{pk} for each location k then allows us to not only assess the general compatibility of the prior distribution with the data but also to identify locations that may be "outliers" with respect to the prior.

Table 1 presents the estimates of p, r, m_1, and m_2 obtained for the younger versus middle data, along with the above described goodness-of-fit tests computed for both the $\hat{p}_k$ and the $\hat{r}_k$. Table 2 gives similar information for older versus middle data. Figures 1 and 2 show the statistics S_{pk} and S_{rk} for the two data sets. The horizontal lines in Figures 1 and 2 correspond to significance levels of $\alpha = .05$. For the most

TABLE 2 Parameter Estimates and Goodness-of-Fit for Older Driver Data

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

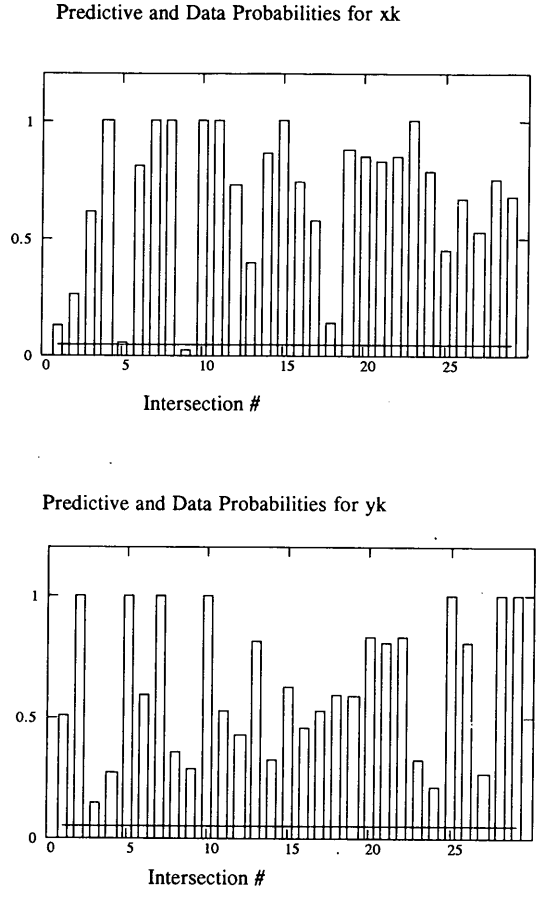

FIGURE 1 Comparison of predictive and data probabilities for younger driver data.

part, it appears that the beta priors placed on the p_k and r_k are tenable, although Intersections 3 and 5 for the older driver data and Intersections 5 and 8 for the younger driver data might be considered atypical compared with the other locations, and thus candidates for a more detailed investigation.

Next, to check the accuracy of the normal approximation used in computing EB confidence intervals, the upper and lower bounds of a nominal 90 percent 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 that Albert's beta approach would assign to these same intervals. Table 3 gives 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 2 percentage points, and we concluded that the normal approximation showed acceptable accuracy.

Finally, Figure 3 shows 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 percent 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 counterintuitive cases from consideration (such as Intersec-

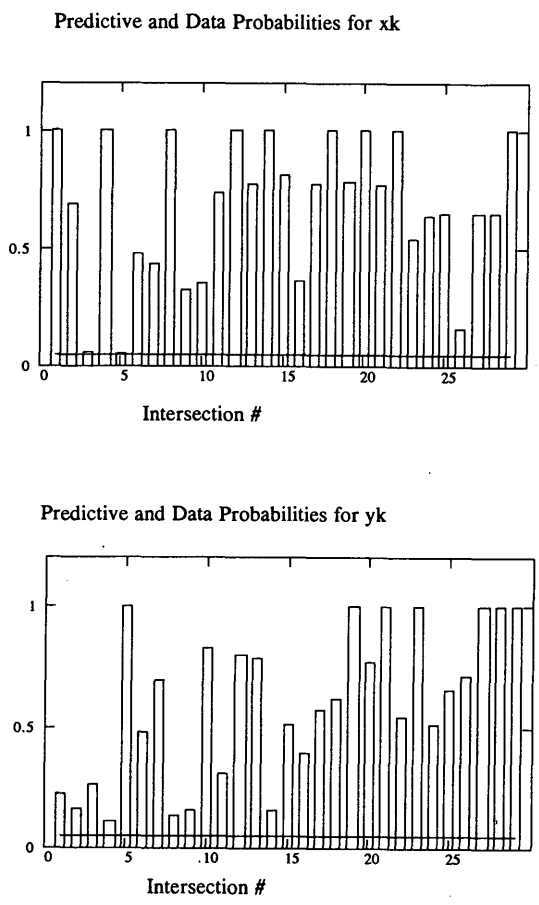

FIGURE 2 Comparison of predictive and data probabilities for older driver data.

TABLE 3 Beta-Derived Confidence Levels for Nominal 90 percent 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

tion 4, where the ML estimate indicates that middle drivers have higher accident rates 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, whereas the increased accident rates for older drivers, if present at all, appear to be localized around Intersections 5 through 7 and Intersections 23 through 25. Assuming that some older driver–oriented improvement of this roadway was desirable, these two sections would be candidates for first consideration.

CONCLUSION

We have presented a statistical method for location-specific testing of the equality of accident rates experienced by two different groups of drivers. To sidestep the need for location- and group-specific measures of exposure, we have based the method on the induced exposure model, and to improve the estimation in the face of the RTM effects inherent in the small samples generally available, we have used an EB estimation procedure. The most computationally demanding feature of our 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 commercially available mathematical spreadsheet software such as MATHCAD. All other computations require no more than a hand calculator. Thus the method should be easy to incorporate in any accident analysis system capable of matching accident records to specific locations and potentially could be used for cost-effective screening of a large number of locations as to their hazard for particular driver groups, such as older drivers.

Before recommending widespread implementation of the method, however, we believe that three issues require further study. First and foremost, the question as to whether the method is robust with respect to different choices for the noninformative prior distributions placed on the hyperparameters p, r, m_1, and m_2 needs investigation. Second, the robustness of the method with respect to different procedures for estimating the hyperparameters should also be investigated. Third, it may be possible to reduce the computational effort required by the current implementation of this method through the use of more efficient search routines such as Golden Section search or less demanding approximations to the numerical integrals used here. Given suitable answers to these questions, the combination of EB statistical methodology with the induced exposure model should provide a useful addition to the safety engineer's analytic toolbox.

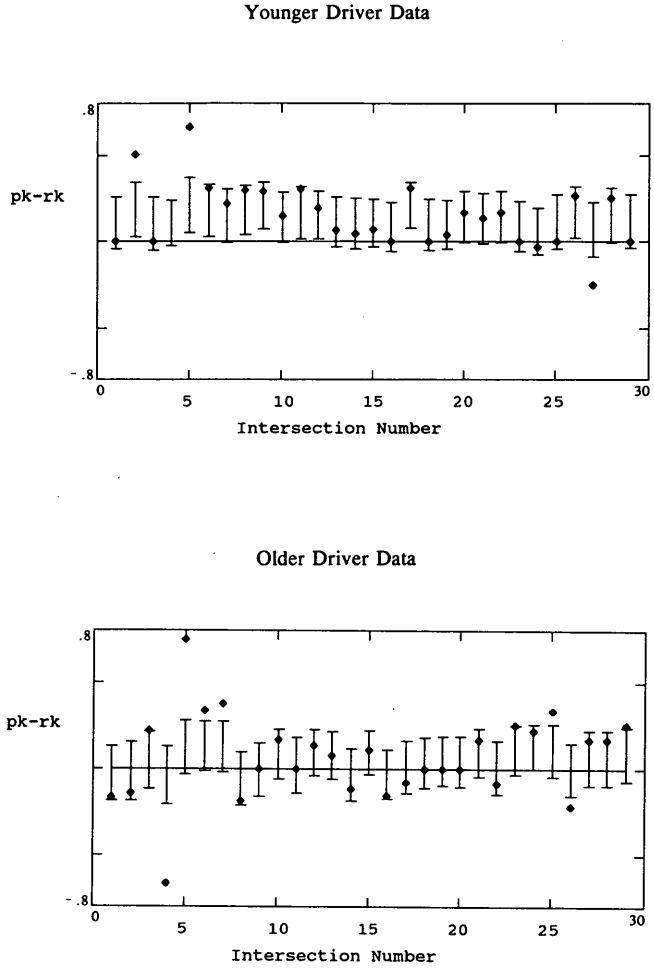

FIGURE 3 Comparison of 90 percent EB confidence intervals (I) and ML point estimates (◆) of the differences $p_k - r_k$.

ACKNOWLEDGMENTS

The authors thank Dave Miller of MNDOT for developing the accident data files used in this research. They also thank Robert Johns of the Center for Transportation Studies at the University of Minnesota for the initial contact work that resulted in this research project.

REFERENCES

1. *Special Report 218: Transportation in an Aging Society*. TRB, National Research Council, Washington, D.C., 1988.
2. F. McKelvey, T. Maleck, N. Stamatiadis, and D. Hardy. Highway Accidents and Older Drivers. In *Transportation Research Record 1172*, TRB, National Research Council, Washington, D.C., 1987, pp. 47–56.
3. *Minnesota Roadway Safety Initiatives for 1990 and Beyond*. Office of Traffic Engineering, Minnesota Department of Transportation, St. Paul, Minn., 1989.
4. C. Zegeer. *Highway Accident Analysis Systems*. TRB, National Research Council, Washington, D.C., 1982.
5. E. Hauer. Bias-by-Selection: Overestimation of the Effectiveness of Safety Countermeasures Caused by the Process of Selection for Treatment. *Accident Analysis and Prevention*, Vol. 12, 1980, pp. 113–117.
6. O. Pendleton and C. Morris. A New Method for Accident Analysis. Presented at 69th Annual Meeting of the Transportation Research Board, Washington, D.C., 1990.
7. C. Morris. Parametric Empirical Bayes Inference: Theory and Applications. *Journal of American Statistical Society*, Vol. 78, 1983, pp. 47–65.
8. P. Carlin and A. Gelfand. Approaches for Empirical Bayes Confidence Intervals. *Journal of American Statistical Association*, Vol. 85, 1990, pp. 105–114.
9. F. Haight. Induced Exposure. *Accident Analysis and Prevention*, Vol. 5, 1973, pp. 111–126.
10. J. Thorpe. Calculating Relative Involvement Rates in Accidents Without Determining Exposure. *Australian Road Research*, Vol. 2, 1964, pp. 25–36.
11. F. Haight. A Crude Framework for Bypassing Exposure. *Journal of Safety Research*, Vol. 2, 1970, pp. 26–29.
12. H. Joksch. A Pilot Study of Observed and Induced Exposure to Traffic Accidents. *Accident Analysis and Prevention*, Vol. 5, 1973, pp. 127–136.
13. M. Koonstra. A Model for Estimation of Collective Exposure and Proneness from Accident Data. *Accident Analysis and Prevention*, Vol. 5, 1973, pp. 157–173.
14. Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass., 1975.
15. J. Albert. Empirical Bayes Estimation of a Set of Binomial Probabilities. *Journal of Statistical Computing and Simulation*, Vol. 20, 1984, pp. 129–144.
16. G. Box. An Apology for Ecumenism in Statistics. In *Scientific Inference, Data Analysis and Robustness* (G. Box et al., eds.), Academic Press, New York, 1983, pp. 51–84.
17. J. Hill. A General Framework for Model-Based Statistics. *Biometrika*, Vol. 77, 1990, pp. 115–126.
18. V. Rohatgi. *An Introduction to Probability Theory and Mathematical Statistics*. Wiley and Sons, New York, 1976.

This research was sponsored by the Minnesota Department of Transportation, but all opinions and conclusions expressed here are solely the responsibility of the authors.

Publication of this paper sponsored by Task Force on Statistical Methods in Transportation.