Statistical Methods To Support Induced Exposure Analyses of Traffic Accident Data

GARY A. DAVIS AND YIHONG GAO

When it is possible to identify the drivers involved in two-vehicle accidents as either at fault or innocent, induced exposure methods offer a way to assess the relative accident risk of driver subgroups, even when group-specific measures of exposure are unavailable. A cross tabulation of two-vehicle accidents by group membership of the at-fault and victim drivers forms a contingency table, and statistical methods derived from contingency table analysis can be used to make inferences concerning the variables arising in the induced exposure model. It is shown how the standard contingency table test for independence of row and column classifications provides a test of the assumption that the victims are sampled randomly and how an odds ratio statistic can be used to estimate the ratio of the accident rates between two driver subgroups. This estimator is asymptotically normally distributed, and a formula is given for estimating its standard error. An Empirical Bayes method for identifying sites where one driver subgroup has a significantly higher accident rate than does another is then presented. These procedures are illustrated using several actual accident data sets.

Over the past several years, the traffic safety community has shown an increased interest in assessing the accident risk of particular driver subgroups. The main emphasis has been on older drivers (*1*), but recently attention has also been given to younger drivers (*2*) and to mounting evidence for an increasing number of accidents involving women drivers (*3*). Unfortunately, the study of such problems is made difficult by the fact that an increase in accidents for some driver subgroup can be attributed either to an increase in the tendency of that group to have accidents (its accident rate), to an increase in that group's opportunity to be involved in accidents (its exposure), or to some interaction between these factors. Using accident counts to make inferences concerning a subgroup's accident rate will generally require knowledge of that group's exposure, but measures of exposure are difficult to define in a completely satisfactory way (*2*) and are even more difficult to obtain in disaggregated forms. For instance, even if we can agree that a variable, such as vehicle kilometers of travel (VKT), is the appropriate exposure measure for older drivers, estimating the VKT for this subgroup usually requires asking a sample of drivers to estimate the number of kilometers they have driven during the past year. And even when such data are available, they usually tell us little about the exposure of a subgroup on smaller areal units, such as a highway corridor or a single intersection.

Safety researchers have been aware of these difficulties for at least 25 years, and in the early 1970s induced exposure methods were presented as providing at least a partial solution to this problem (*4*). Following some intense initial interest, induced exposure methods appear to have suffered a period of neglect, but recently several papers have used these ideas to investigate accident risk to older drivers (*5–8*). The induced exposure model assumes that in a majority of two-vehicles accidents, one driver can be identified as the at-fault driver, whereas the other is treated as an innocent victim. Innocent victims are assumed to be "selected" by the at-fault driver randomly from the pool of available drivers, with the probability that the innocent victim is the member of a given subgroup being directly proportional to that subgroup's exposure at the accident site. Thus the same measure of exposure reflects both a subgroup's opportunity to cause accidents and its opportunity to be involved as victims. From comparisons of the proportion of accidents that a subgroup causes with the proportion in which it is involved as innocent victims, it is possible to identify subgroups that have accident rates higher or lower than the average for all groups (*9*).

But before the promise of induced exposure methods can be fully realized, it is necessary to answer several questions related to the statistical properties of induced exposure measures. First, the assumption of random selection of victims appears somewhat controversial (*10*), although studies investigating its validity have tended to support it (*5,10*). Still, it would be useful if a test of the tenability of this assumption could be conducted on any given data set. Second, recent studies have used the induced exposure method in essentially a deterministic manner, treating data-dependent quantities as if they were known with certainty, with no attempts made to estimate likely ranges of error. When one has very large data sets, it may be possible to invoke the law of large numbers to justify ignoring random effects, but many, if not most, applications of induced exposure will likely involve more modest data sets. Here it would be useful to have procedures for determining confidence bounds and testing hypotheses for induced exposure estimates. Finally, many safety engineers are ultimately responsible for deciding on particular safety improvements for particular locations. If these improvements are targeted at a specific driver subgroup, it may be necessary to identify specific locations where that subgroup is at heightened risk. Aggregated data will not generally provide this

Department of Civil and Mineral Engineering, University of Minnesota, 500 Pillsbury Drive SE, Minneapolis, Minn. 55455.

level of detail, necessitating an extension of induced exposure ideas to the problem of identifying high-hazard locations.

In what follows, we will first show a natural correspondence between statistical inference using induced exposure ideas and more standard methods of contingency table analysis. This will lead first to a straightforward test as to whether the assumption of random selection of victims is tenable for a given data set and then to a method for computing maximum likelihood estimates of the ratio of the accident rates for two driver subgroups. We point out that the natural logarithm of this estimated rate-ratio is approximately normally distributed and give a formula for estimating its standard error. These methods are then illustrated using two actual traffic data sets. We next turn to the problem of identifying high-hazard locations and derive a Bayes estimator for the log rate-ratio, along with its posterior standard error. Given data from a number of sites, we then show how an Empirical Bayes approach can be used to compute point estimates and approximate confidence intervals for the log rate-ratios for each site. The paper ends with conclusions and recommendations for further research.

INDUCED EXPOSURE AND CONTINGENCY TABLES

We begin with a more formal statement of the induced exposure model. Suppose the driver population has been divided into m subgroups, and let n_i denote the number of accidents involving Driver Subgroup i in some area (such as a city) over some time interval (such as a year). Assume n_i to be the outcome of a Poisson random variable, with mean $\lambda_i E_i$, where λ_i is the accident rate for Driver Subgroup i and E_i is the exposure for Driver Subgroup i. If the exposure values E_i are known exactly, the maximum likelihood estimates of the accident rates λ_i are given by

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

and assessment of the relative risk to driver subgroups can be based on these estimated accident rates. But as noted earlier, group-specific measures of exposure are difficult to estimate reliably. 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, whereas the other is assumed to be an innocent victim. The at-fault drivers are assumed to cause accidents according to the Poisson accident model, whereas the subgroup of the victim is assumed to be selected randomly, with probability of selection being directly proportional to the group exposures. Defining $r_i = E_i/\Sigma_k E_k$, the probability the victim is chosen from Subgroup i, and n_{ij} = number of accidents for which the at-fault driver came from Subgroup i while the victim came from Subgroup j, it follows that the n_{ij} are the outcomes of independent Poisson random variables with mean values $r_j \lambda_i E_i$. By taking the total number of accidents in the sample, $n = \Sigma_i \Sigma_j n_{ij}$, as fixed and defining $p_i = \lambda_i E_i / \Sigma_k \lambda_k E_k$, it can be shown that the n_{ij} are now the outcomes of a multinomial random vector with number of "trials" equal to n and the probability of a given two-vehicle accident having an at-fault driver from Subgroup i and a victim from Subgroup j being simply $p_i r_j$. The n_{ij} can be thought of as entries in a cross-tabulation table, where two-vehicle accidents are classified according to the group membership of the at-fault and victim drivers. For example, Table 1 gives the expected cell counts (the expected values of the n_{ij}) and marginal probabilities for the case where only two subgroups, denoted by 1 and 2, are of interest.

In Table 1 the probability that a given two-vehicle accident falls in a cell is simply the product of the corresponding row and column marginal probabilities, so that the table shows 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. As an example, for a 2 × 2 table such as that given in Table 1, it is well known (*11*) that under the hypothesis of independence, the log cross-product ratio statistic

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

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

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

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

Assuming now that the data in an induced-exposure table satisfy the assumption of random selection of victims, we turn to the problem of making inferences concerning the accident

TABLE 1 Example 2 × 2 Induced Exposure Table

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

rates λ_i. To simplify some of the following notation, we define the marginal totals

$$x_i = \sum_k n_{ik}$$

$$y_j = \sum_k n_{kj}$$

A straightforward application of maximum likelihood methods yields the ML estimators of p_i and r_j,

$$\hat{p}_i = \frac{x_i}{n}$$

$$\hat{r}_j = \frac{y_j}{n}$$

Unfortunately, since $\Sigma_i p_i = \Sigma_j r_j = 1$, the induced exposure table is completely characterized by $2(m - 1) + 1$ parameters, making it impossible to uniquely identify the $2m$ parameters λ_i and E_j. It is possible, however, to estimate and compare relative quantities, p_i for example being the ratio of the expected number of accidents caused by Subgroup i to the total expected number of accidents, whereas r_j is the ratio of the exposure for Subgroup j to the total exposure. The measure that has appeared most often in the literature is the involvement ratio (*5–10*)

$$IR_i = p_i/r_i$$

with $IR_i = 1.0$ being taken as evidence that the accident rate for Group i is typical of the whole population. This interpretation follows by noting that the accident rates should be independent of the exposures, so that $IR_i = 1.0$ and $E_j = E$ for each j implies

$$\lambda_i = (1/m) \sum_j \lambda_j$$

When only two subgroups are available (i.e., $m = 2$), $IR_1 = 1.0$ is equivalent to $\lambda_1 = \lambda_2$. Alternatively, as elsewhere (*12*), one could consider the difference $p_i - r_i$, with $p_i - r_i = 0$ having the same interpretation as $IR_i = 1.0$.

The involvement ratio allows the analyst to identify which subgroups have accident rates that exceed the populationwide average but does not provide readily interpretable information concerning the magnitude of this discrepancy, nor does it provide a means for comparing the relative accident risks of two different subgroups. However, the ratio of two involvement ratios has the form of an odds ratio statistic and is equal to the ratio of the respective accident rates:

$$\frac{IR_i}{IR_j} = \frac{p_i r_j}{p_j r_i} = \frac{\lambda_i}{\lambda_j}$$

If we define the log rate-ratio statistic as

$$\Delta_{ij} = \log_e\left(\frac{\lambda_i}{\lambda_j}\right)$$

it is straightforward to verify that the ML estimate of Δ_{ij} can be computed via

$$\hat{\Delta}_{ij} = \log_e\left(\frac{x_i y_j}{x_j y_i}\right)$$

and an application of the delta method yields that, for large n, the distribution of Δ_{ij} is approximately normal, with mean equal to $\hat{\Delta}_{ij}$ and variance that can be estimated via

$$\hat{\sigma}^2_{\Delta} = \frac{1}{x_i} + \frac{1}{y_j} + \frac{1}{x_j} + \frac{1}{y_i}$$

This last result is a consequence of the fact that, given row and column independence, the likelihood function of the induced exposure table factors into two components, one being proportional to the marginal likelihood of the row totals and the other being proportional to the marginal likelihood of the column totals. This provides a method for testing hypotheses concerning Δ_{ij} and for constructing approximate confidence intervals for the rate-ratio λ_i/λ_j.

As an illustration of the utility of these methods, first consider the data given in Table 2, originally presented by Lyles et al. (*10*). Here we have two 2×2 induced-exposure tables, with the driver subgroups being male and female. The upper table gives the cross tabulation for non–rush hour daytime interstate accidents in Michigan for 1988, and the lower table is a similar cross tabulation of nighttime interstate accidents. Testing first whether the assumption of random victim selection is tenable for these tables (i.e., that the estimated log cross product ratio, $\hat{\theta}$, is not significantly different from zero), we obtain for the upper table $\hat{\theta} = -0.039$, $z = -0.502$, $p > .6$, whereas for the lower table we obtain $\hat{\theta} = 0.055$, $z = 0.65$, $p > .5$. In both cases, independence of row and column classifications appears tenable. Next, we consider whether the accident rate for males is greater than that for females by testing the null hypothesis $\lambda_m = \lambda_f$ against the one-sided al-

TABLE 2 Example Induced Exposure Tables from Lyles et al. (*10*)

		Day Time Non-rush Hour	
		Innocent Victim	
		Male	Female
	Male	1810	941
At Fault			
Driver	Female	678	339

		Night Time	
		Innocent Victim	
		Male	Female
	Male	2232	894
At Fault			
Driver	Female	605	256

ternative $\lambda_m > \lambda_f$. The log rate-ratio provides an appropriate test statistic, and for the upper table we obtain $\hat{\Delta}_{mf} = 0.33$, $z = 6.57$, $p < .001$, and for the lower table we obtain $\hat{\Delta}_{mf} = 0.386$, $z = 7.43$, $p < .001$, indicating that, in both cases, the accident rate for males is significantly higher than that for females. For the upper table, an approximate 90 percent confidence interval for the rate-ratio λ_m/λ_f would be (1.28, 1.51), whereas a similar confidence interval for the lower table would be (1.35, 1.60). For both tables, it appears that the accident rate for male drivers is around 40 percent higher than that for female drivers.

As a second example, consider the data presented in Table 3. Here, drivers are divided into two subgroups according to age, with Group 1 being middle-aged drivers (ages 25 to 55) and Group 2 being "older" drivers (ages 56 and over). The upper table presents a cross tabulation of two-vehicle accidents occurring at the signalized intersections along a section of Minnesota Trunk Highway (MNTH) 47 during 1988–1989. The lower table presents a similar cross tabulation for MNTH 65, which runs about 1.5 km east and parallel to MNTH 47. Checking first to see whether the assumption of random victim selection is tenable for these two tables, we obtain for MNTH 47 $\hat{\theta} = -0.419$, $z = -0.93$, $p > .34$. For MNTH 65 we obtain $\hat{\theta} = -0.378$, $z = -1.08$, $p > .28$. Again, the random selection assumption appears acceptable. Testing next for whether older drivers have higher accident rates than do middle-aged drivers, we obtain for MNTH 47 $\hat{\Delta} = 0.2$, $z = .84$, $p > .20$, and for MNTH 65 we obtain $\hat{\Delta} = 0.28$, $z = 1.50$, $p < .07$. Thus the data from MNTH 47 show no 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. This sort of information could be useful to a safety engineer responsible for programming safety improvements.

TABLE 3 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

EMPIRICAL BAYES IDENTIFICATION OF HIGH-HAZARD LOCATIONS

The second example presented above suggested that the MNTH 65 corridor might be a candidate for safety improvements targeted at older drivers. But since there are 29 signalized intersections providing data for that example, it could very well be that these sites differ in the risk they pose to older drivers. Because the numbers of accidents occurring at particular sites over a 2- or 3-year period typically tend to be in the range 0 to 50, the uncertainty attached to site-specific ML estimates tends to be high, and application of the asymptotic statistical methods described earlier to individual sites is questionable. Alternatively, identifying high-hazard locations can be viewed as an example of a multiparameter estimation problem, so that Empirical Bayesian (EB) statistical methods might profitably be employed (*13,14*); in fact Davis and Koutsoukos (*12*) have described an EB approach for estimating the difference $p_i - r_i$. Here, we describe how EB estimates and confidence intervals can be computed for the log rate-ratio statistic defined above. To simplify the presentation of some of the following equations, we will restrict our attention to the case in which only two driver subgroups are of interest.

Let the two driver subgroups of interest be denoted by 1 and 2 and assume that there is available a 2 × 2 induced-exposure table for each of a set of N sites making up our sample. Let the individual sites be indexed by $k = 1, \ldots, N$, and define the variables

p_k = probability that an accident at Site k had a driver from Subgroup 1 as the at-fault driver,
r_k = probability that an accident at Site k had a driver from Subgroup 1 as the innocent victim,
n_k = total two-vehicle accidents available for Site k,
x_k = number of accidents from Site k where the at-fault driver was from Subgroup 1, and
y_k = number of accidents from Site k where the innocent victim was from Subgroup 1.

The EB model assumes that the actual accident counts for a site are generated by a two-stage random process. 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

$$E[p_k] = p$$

and

$$Var[p_k] = p(1 - p)/(m_1 + 1)$$

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

$$E[r_k] = r$$

and

$$Var[r_k] = r(1 - r)/(m_2 + 1).$$

Given p_k, r_k, and n_k, the accidents are then assigned to cells in the induced exposure table according to the multinomial

model described earlier. The log rate-ratio statistic for Site k becomes

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

In a manner similar to that used by Maritz (*15*), it can be shown that if the underlying prior parameters p, m_1, r, and m_2 are known in advance, the posterior means and variances of the Δ_k are given by

$$E[\Delta_k|n_k, x_k, y_k, m_1, p, m_2, r] = \Psi(m_1p + x_k) + \Psi[m_2(1 - r) + n_k - y_k] - \Psi[m_1(1 - p) + n_k - x_k] - \Psi(m_2r + y_k)$$

$$Var[\Delta_k|n_k, x_k, y_k, m_1, p, m_2, r] = \Psi'(m_1p + x_k) + \Psi'[m_1(1 - p) + n_k - x_k] + \Psi'(m_2r + y_k) + \Psi'[m_2(1 - r) + n_k - y_k] \quad (1)$$

where $\Psi(x)$ denotes the digamma function and $\Psi'(x)$ denotes the trigamma function:

$$\Psi(x) = \frac{d\log_e[\Gamma(x)]}{dx}$$

$$\Psi'(x) = \frac{d\Psi(x)}{dx} \quad (2)$$

The expression for the posterior variance of Δ_k follows from the fact that the joint posterior distribution of p_k and r_k factors into two components, one containing m_1, p, and x_k and the other containing m_2, r, and y_k. When numerical software for evaluating these functions is not available, they can be approximated using the first-order terms of their asymptotic expansions (*16*)

$$\Psi(x) \approx \log_e(x) - \frac{1}{2x}$$

$$\Psi'(x) \approx \frac{1}{x} \quad (3)$$

Furthermore, the posterior distribution of the Δ_k is well approximated by a normal distribution with means and variances given in Equation 1, so that if the prior parameters m_1, p, m_2, and r are known, point and interval estimates of the Δ_k can be computed using either Equation 1 or Equation 3.

In practice, though, the prior parameters p, m_1, r, and m_2 will not be known and must also be estimated from data. The EB approach proceeds by simply replacing the prior parameters in Equation 1 with these estimates, so that the EB estimate of Δ_k is

$$\hat{\Delta}_k = \Psi(\hat{m}_1\hat{p} + x_k) + \Psi[\hat{m}_2(1 - \hat{r}) + n_k - y_k] - \Psi[\hat{m}_1(1 - \hat{p}) + n_k - x_k] - \Psi(\hat{m}_2\hat{r} + y_k) \quad (4)$$

and the EB estimate of the variance of Δ_k is

$$\hat{\sigma}_k^2 = \Psi'(\hat{m}_1\hat{p} + x_k) + \Psi'[\hat{m}_1(1 - \hat{p}) + n_k - x_k] + \Psi'(\hat{m}_2\hat{r} + y_k) + \Psi'[\hat{m}_2(1 - \hat{r}) + n_k - y_k] \quad (5)$$

An EB confidence interval with approximate coverage probability $1 - \alpha$ would then be $(\hat{\Delta}_k - z_{\alpha/2}\,\hat{\sigma}_k,\ \hat{\Delta}_k + z_{\alpha/2}\,\hat{\sigma}_k)$.

Maximum likelihood estimates of the parameters p, m_1, r, and m_2 can be found as values maximizing the marginal distribution

$$L(p, m_1, r, m_2) = \prod_{k=1}^{N}\left\{\frac{n_k!}{\prod_{ij} n_{ij,k}!}\frac{B[m_1p + x_k, m_1(1 - p) + n_k - x_k]}{B[m_1p, m_1(1 - p)]} \times \frac{B[m_2r + y_k, m_2(1 - r) + n_k - y_k]}{B[m_2r, m_2(1 - r)]}\right\} \quad (6)$$

Here $B(a, b)$ denotes the Beta integral evaluated at a and b. Computation of the estimates is simplified by the fact that Equation 6 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 problem that can arise in practice is that the likelihood function (Equation 6) 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 65 and the MNTH 47 data sets. The simplest solution to this problem is to constrain the parameters m_1 and m_2 to be less than some appropriately large value and use this bound as the MLE in those situations where the MLE is unbounded. To arrive at a plausible upper bound, recall that the objective of this method is to identify locations where the accident rates for Groups 1 and 2 satisfy $\lambda_1 > \lambda_2$. Using the formulas in Equation 1 coupled with the normal approximation of the posterior distribution of Δ_k, it is possible to express the posterior probability that $\lambda_1 > \lambda_2$ as a function of m_1 and m_2. By inserting the MLE for m_2 into this function and then plotting this probability as a function of m_1, it is possible to gain an idea of the sensitivity of the final decision to the choice of an upper bound. Figure 1 shows such plots for four typical intersections selected from MNTH 65. In each

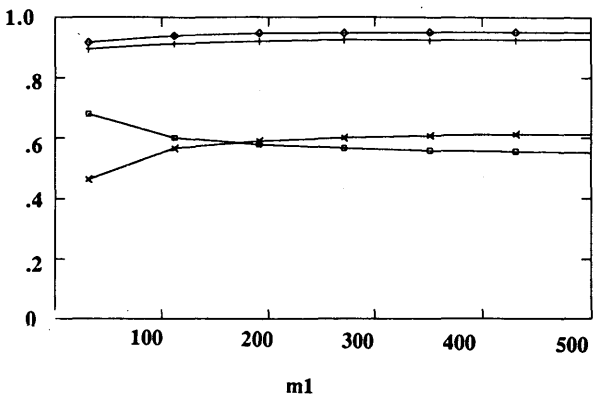

FIGURE 1 Approximate posterior probability that $\lambda_1 > \lambda_2$ as a function of m_1, for four intersections on MNTH 65.

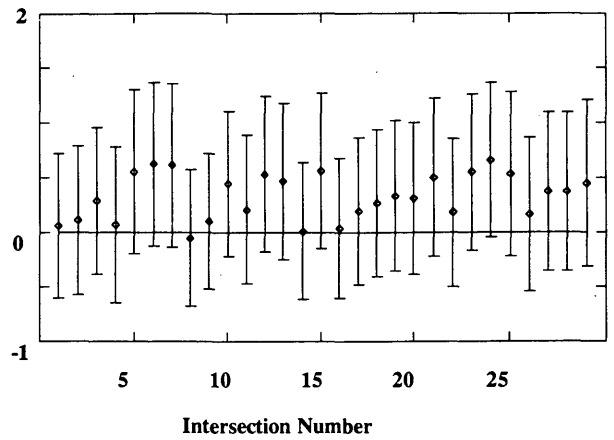

FIGURE 2 EB point estimates of log rate-ratios for 29 intersections on MNTH 65, along with approximate 90 percent EB confidence intervals. Upper bound for m_1 set at 100.

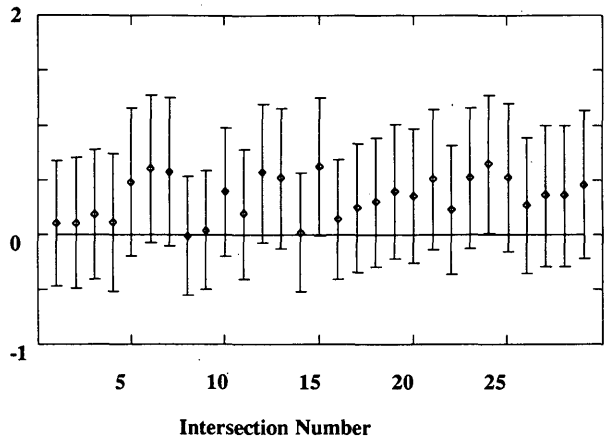

FIGURE 4 EB point estimates of log rate-ratios for 29 intersections on MNTH 65, along with approximate 90 percent EB confidence intervals. Upper bound for m_1 set at 500.

of these cases, when $m_1 > 200$ the posterior probability tends to stabilize into a slowly monotonic function of m_1. This pattern was present in each of the sites included in this study, with most of the change in posterior probability tending to occur for m_1 less than 500 and values of m_1 beyond 500 tending to produce fairly small changes in posterior probability.

To illustrate this EB approach, we return to the MNTH 65 example presented earlier. There were a total of 29 signalized intersections along this section of MNTH 65, and the induced exposure data presented in Table 3 were disaggregated according to the intersection where the accidents took place. Two MATHCAD 3.0 computational documents were developed. The first computed bounded ML estimates of p, m_1, r, and m_2 via Equation 6, with upper bounds being user-specified inputs, and then wrote these estimates to a file. The second document read these estimates, computed the EB estimates for Δ_k, σ_k, and approximate 90 percent confidence intervals for Δ_k, for each of the 29 intersections, and then created the graphs shown in Figures 2 through 4. The upper bound for m_1 was set at 100 in Figure 2, at 200 in Figure 3, and at 500 in Figure 4. Since an estimated $\hat{\Delta}_k$ that is not significantly different from zero indicates a site where the accident rate for older drivers is not significantly different from that of middle-aged drivers, inspection of Figures 2 through 4 indicates that the risk to older drivers is not evenly distributed along the roadway. If present at all, it appears concentrated on two segments, one containing Intersections 5, 6, and 7 and the other containing Intersections 23, 24, and 25. This qualitative identification appears to be robust with respect to the upper bounds placed on m_1. These results are similar to those presented for the same data set, but a somewhat different computational method, elsewhere (*12*).

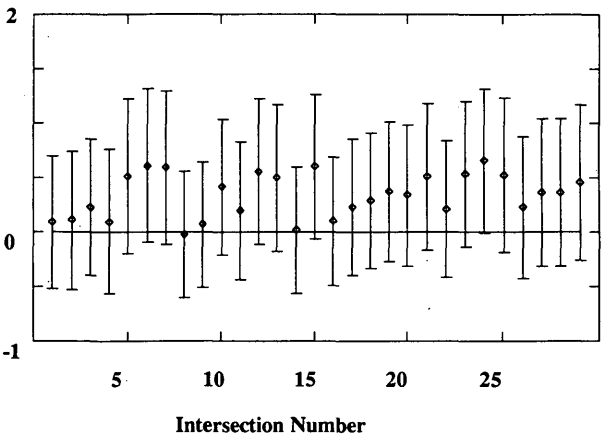

FIGURE 3 EB point estimates of log rate-ratios for 29 intersections on MNTH 65, along with approximate 90 percent EB confidence intervals. Upper bound for m_1 set at 200.

CONCLUSION

In this paper we have formalized some of the relationships between induced exposure and contingency table analyses, used these results to identify a test for the random selection of accident victims, and then developed an estimator for the ratio between the accident rates for two different driver subgroups. An Empirical Bayes approach was then presented for estimating these rate-ratios for each of a number of accident sites and using approximate confidence intervals around these estimates to identify locations where a given driver subgroup might be at increased risk. The utility of these procedures was illustrated using actual traffic accident data.

Although certainly not a panacea, the induced exposure model offers a promising approach for estimating the differential in accident risk experienced by subgroups of drivers, and it is hoped that the statistical methods described here will facilitate a wider use of and research into induced exposure methods. Of particular interest would be an extension of this approach to multiway cross-tabulation tables, permitting the analyst to assess the effect of possible causal factors on accident rate differentials. A special case would be the problem of assessing the impact of safety countermeasures, using before and after data. Finally, user-friendly implementations of these methods are probably needed to facilitate their widespread adoption.

ACKNOWLEDGMENTS

The authors would like to thank Susan Scharenbroich and Dave Miller of the Minnesota Department of Transportation (MNDOT) for their assistance in conducting this research. This project was supported by MNDOT.

REFERENCES

1. *Special Report 218: Transportation in an Aging Society: Improving Mobility and Safety of Older Persons.* TRB, National Research Council, Washington, D.C., 1988.
2. L. Evans. *Traffic Safety and the Driver.* Van Nostrand, New York, 1991.
3. M. Edwards. Trends in Women's Fatal Crash Involvement. Presented at the 71st Annual Meeting of the Transportation Research Board, Washington, D.C., 1992.
4. F. Haight. Induced Exposure. *Accident Analysis and Prevention*, Vol. 5, 1973, pp. 111–126.
5. T. Maleck and H. Hummer. Driver Age and Highway Safety. In *Transportation Research Record 1059*, TRB, National Research Council, Washington, D.C., 1987, pp. 6–12.
6. F. McKelvey, T. Maleck, N. Stamatiades, and D. Hardy. Highway Accidents and the Older Driver. In *Transportation Research Record 1172*, TRB, National Research Council, Washington, D.C., 1988, pp. 47–57.
7. N. Garber and R. Srinivasan. Risk Assessment of Elderly Drivers at Intersections. In *Transportation Research Record 1325*, TRB, National Research Council, Washington, D.C., 1991.
8. P. Cooper. Differences in Accident Characteristics Among Elderly Drivers and Between Elderly and Middle-Aged Drivers. *Accident Analysis and Prevention*, Vol. 22, 1990, pp. 499–508.
9. E. Cerelli. Driver Exposure: The Indirect Approach for Obtaining Relative Measures. *Accident Analysis and Prevention*, Vol. 5, 1973, pp. 147–156.
10. R. Lyles, N. Stamatiades, and D. Lighthizer. Quasi-Induced Exposure Revisited. *Accident Analysis and Prevention*, Vol. 23, 1991, pp. 275–285.
11. A. Agresti. *Categorical Data Analysis.* Wiley and Sons, New York, 1990.
12. G. Davis and K. Koutsoukos. Statistical Method for Identifying Locations of High Crash Risk to Older Drivers. In *Transportation Research Record 1375*, TRB, National Research Council, Washington, D.C., 1992.
13. C. Morris. Parametric Empirical Bayes Inference: Theory and Application. *Journal of the American Statistical Association*, Vol. 78, 1983, pp. 47–65.
14. O. Pendleton. *Application of New Accident Analysis Methodologies.* Report FHWA-RD-90-091. FHWA, U.S. Department of Transportation, 1991.
15. J. Maritz. Empirical Bayes Estimation of the Log Odds Ratio in 2 × 2 Contingency Tables. *Commun. Statistics-Theory Meth.*, Vol. 18, 1989, pp. 3215–3233.
16. M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions* (9th ed.). Dover, New York, 1970.

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.