Comparison of Accident Rates Using the Likelihood Ratio Testing Technique

ALI AL-GHAMDI

Comparing transportation facilities (i.e., intersections and road sections) in terms of traffic accident occurrences is among the interests of most traffic safety analysts. Traditionally, traffic accidents are represented as occurrences of events per certain unit, such as time and vehicle miles; this representation is consistent with Poisson nature. The Poisson distribution is used to describe the distribution of traffic accident occurrences. The objective is to develop a test statistic to enable traffic analysts to compare traffic accident rates in various transportation facilities. The obtained test statistic is simple and requires minimal data to perform the comparison.

Traffic safety improvements have been the concern of traffic engineers lately. The growth in the number of both motorists and traffic accidents is behind this concern. As a result, accident data have been used to analyze traffic accidents as well as to find appropriate techniques to understand the nature of such accidents. This understanding may help traffic accident analysts to draw realistic conclusions regarding the causes or frequencies of accidents and to make appropriate decisions to prevent these causes or reduce the frequencies. This paper uses a hypothesis-testing technique, the likelihood ratio testing technique, to develop a closed form of test statistic to assist traffic analysts in comparing the significance of traffic accident occurrences among different transportation facilities in a transportation system. The occurrences of traffic accidents follow Poisson phenomenon (*1–3*). Hence, the Poisson distribution function is used herein as a basis for deriving the test statistic. The test statistic requires minimal data and can be easily computerized.

STATISTICAL BACKGROUND

Since the approach of this study is based on a statistics technique, the likelihood ratio testing technique, a brief theoretical background of this technique will be given. A general review of statistical distributions and hypothesis testing is given first.

Distributions and Hypothesis Testing

Statistical distributions are useful in interpreting a wide variety of phenomena where randomness is present. In traffic studies the most important distributions are discrete distributions—usually known as counting distributions. Such distributions are useful in describing the occurrence of events that can be counted, such as the number of accidents and the number of arrivals at a certain location.

Two types of discrete distributions are widely used in traffic. They are the binomial and Poisson distributions (*1,3,4*). These distributions have been used in several traffic studies, including studies of speeds, gap acceptances in traffic flow, and accidents. For example, Gerlough and Huber state:

> Counting the number of cars arriving during an interval of time is the easiest and oldest measurement of traffic. When counts from a series of equal time intervals are compared, they appear to form a random series. This led early traffic engineers to investigate distributions as a means of describing the occurrence of vehicle arrivals during an interval. (*1*)

Binomial Distribution

The binomial distribution is formed from a sequence of independent Bernoulli trials, in which the number of successes of a certain number of trials is the quantity of interest. The expansion of the binomial $(q + p)^N$ forms the basis of the binomial distribution function. If N is a positive integer, the $(k + 1)$th term in this expansion is

$$\binom{N}{k} p^k q^{N-k} = \frac{N!}{k(N-k)!} p^k q^{N-k}$$

With parameters N and p, the binomial distribution of a random variable X is defined as

$$\Pr[X = k] = \begin{cases} \binom{N}{k} p^k q^{N-k} & k = 0, 1, 2, \ldots, N \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$ and $q = 1 - p$. The mean and variance of X are Np and Npq, respectively.

This distribution has been used in several traffic applications. In congested traffic (in the case where the ratio of the observed variance/mean is substantially less than 1), for instance, the binomial distribution can be used to describe the distribution of traffic arrivals. When n is very large and p is very small the binomial distribution is approximated by the Poisson distribution.

Poisson Distribution

A random variable X is said to have a Poisson distribution with parameter θ if it has discrete pdf of the form

$$\Pr[X = k] = \frac{e^{-\theta}\theta^k}{k!} \qquad k = 0, 1, 2, \ldots; \theta > 0$$

Department of Civil Engineering, King Saud University, Riyadh, Saudi Arabia.

The random variable X has the same mean and variance. Along with exponential distribution, the pdf of which is defined below, the Poisson distribution has been applied in traffic studies, particularly in studies involving simulation applications. A continuous random variable X has the exponential distribution with parameter $\theta > 0$ if its pdf has the following form:

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The application of the Poisson distribution to traffic studies has been in existence since the 1930s (*3*). This distribution has been used in fitting traffic accidents and vehicle arrivals at certain locations.

Hypothesis Testing

Hypothesis testing can be defined as the process of making a decision about the truth or the falsehood of a particular hypothesis on the basis of experimental evidence. Generally, experimental outcomes are subject to random error, so any decision made is subject to error too. Occasional decision errors cannot be avoided; however, it is possible to construct tests so that such errors occur infrequently and at some prespecified rate.

For example, suppose our past experience with the traffic accident rate at a specific location indicates that the mean of this rate is 4 if a certain type of traffic control is present, and the mean of such rate may be greater than 4 if that type of control is not present. On the basis of a random sample of size n vehicle accidents in our experiment, we would try to decide which case is true. That is, our test would be the null hypothesis $\mu = 4$ versus the alternative hypothesis $\mu > 4$.

To test a specific hypothesis, a certain critical region is required. The critical region is the subset of the sample space that corresponds to rejecting the null hypothesis. In our example, the sufficient statistic for μ is $\bar{X}$; therefore, we can represent our critical region in terms of the univariate variable—the test statistic. According to the alternative hypothesis, we write our critical region in the following form:

$$C = \{(X_1, \ldots, X_n) | \bar{X} \geq c\}$$

for some appropriate constant c (this constant can be obtained on the basis of the distribution of the random variable in the left-hand side of the inequality). In other words, we will reject the null hypothesis if $\bar{X} \geq c$, and we will accept it if $\bar{X} < c$.

Two possible errors can be made under this testing procedure. The first one is called the Type I Error—rejecting a true H_o. Failing to reject H_o when H_o is false is known as the Type II Error. The objective is to keep both types of error as small as possible. That is, we hope that the selected test statistic and its critical region will yield a small probability of making these two errors. The common notations for these error probabilities are as follows:

$$P[\text{Type I Error}] = \alpha$$

$$P[\text{Type II Error}] = \beta$$

An increase in the sample size will reduce α and β simultaneously. In practice, by selecting a small α we ensure that β will be small too, especially when the sample size is large enough and thus there is no need to specify a value for β. The traditional levels of significance are .01, .05, and .1.

Generalized Likelihood Ratio Test

Suppose $X_1, \ldots, X_n$ have joint pdf $f(x, \theta)$ for $\theta \in \Omega$, and we test the hypothesis $H_o: \theta \in \Omega_o$ versus $H_a: \theta \in \Omega - \Omega_o$. The generalized likelihood ratio (GLR) is defined by

$$\lambda(x) = \frac{\max_{\theta \in \Omega_o} f(x; \theta)}{\max_{\theta \in \Omega} f(x; \theta)} = \frac{f(x; \hat{\theta}_o)}{f(x; \hat{\theta})}$$

where $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ and $\hat{\theta}_o$ is the MLE under a true H_o. That is, $\hat{\theta}$ and $\hat{\theta}_o$ are determined by maximizing $f(x; \theta)$ over the general parameter space Ω and the restricted parameter space Ω_o. The numerator represents the likelihood function under the null hypothesis (i.e., a subspace of the general parameter space), and the denominator represents the same function but over the general parameter space. The generalized likelihood ratio test is to reject H_o if $\lambda(x) \leq k$, where k is based on the size of significance. In other words, the value of k can be determined to satisfy

$$P[\lambda(x) \leq k | \text{under } H_o] = \alpha$$

It is obvious that if $\lambda(x)$ is a valid statistic (i.e., free of parameters), it will be possible to obtain the exact critical value k. Yet, in many cases the distribution of $\lambda(x)$ is a function of unknown parameters, and thus the critical region cannot be defined. To solve this dilemma, an approximation can take place. That is, MLEs are asymptotically normally distributed. (A distribution dependent on a parameter n, usually a sample number, is said to be asympotatically normal if, as n tends to infinity, the distribution tends to the normal form.) Then it can be proven that the asymptotic distribution of $\lambda(x)$ is free of parameters, and an approximate test will be available to determine the critical region (*5*). In particular, if $x \sim f(X; \theta_1, \ldots, \theta_k)$, then under H_o: $(\theta_1, \ldots, \theta_r) = (\theta_{1o}, \ldots, \theta_{ro})$, $r < k$, for large n, the following approximation holds:

$$-2 \log \lambda(x) \sim \chi_r^2$$

Thus, H_o is rejected if

$$-2 \log \lambda(x) \geq \chi_{1-\alpha, r}^2$$

ANALYSIS

The analysis in this study consists of two stages. First, real accident data were used for four types of highways in Ohio (Table 1) to develop a test statistic for comparing their accident rates. This test statistic was generalized to be applicable for different types of data.

TABLE 1 Accident Rate Data for Ohio (6)

Highway Type	Characteristics		Accident Rate
	All Accidents	AMVM	
Scenic	3,621	1,021	3.55
Other 2-lane	36,752	11,452	3.21
Multi-lane	20,348	6,290	3.23
Interstate	10,460	9,412	1.11
Total	71,181	28,177	2.53

Derivation of the Test Statistic Based on Real Data

Accident data from the Ohio Department of Transportation are shown in Table 1 (6). The table presents 1-year accidents for different highway types, including scenic, other 2-lane, multilane, and Interstate. The last column of the table presents the accident rate for each type. This rate is the total number of accidents divided by the annual million vehicle miles (AMVM).

Our interest is to find out whether accident rates among these types are different. In other words, the numbers suggest some differences among accident rates, but the question can be asked whether such differences are true differences or the result of randomness.

Since we are dealing with the occurrence of number of accidents (events) per AMVM (unit of exposure), it is worthwhile to assume that the number of accidents (X) is Poisson distributed. Thus X_i is Pois(μ_i), $i = 1, 2, 3, 4$ (i represents a highway type), and

$$f(x_i, \mu_i) = \frac{e^{-\mu_i}\mu_i^{x_i}}{x_i!}$$

where $x_i = 0, 1, 2, \ldots$; $i = 1, 2, 3, 4$; and $\mu_i > 0$. In addition,

$$\mu_i = \lambda_i t_i$$

where λ_i is the accident rate for highway type i and t_i is the annual vehicle miles for highway type i.

To test whether such rates are unequal, we need to develop our hypotheses. The null and alternative hypotheses are H_o: $\lambda_1 = \lambda_2 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda$ and H_a:$\lambda_i \neq \lambda_k$ for some i, k (i and k are two different types of highways).

The likelihood ratio technique is used to test the above hypothesis. The joint pdf of X_1, X_2, X_3, and X_4, also called the likelihood function, is

$$L = \prod_{i=1}^{4} \frac{(\lambda_i t_i)^{x_i}}{x_i!} e^{-\lambda_i t_i} \tag{1}$$

By taking the log of Equation 1, it can be simplified to

$$\log L = \sum_{i=1}^{4} [x_i \log(\lambda_i t_i) - \log(x_i!) - \lambda_i t_i]$$

$$= \sum_{i=1}^{4} [x_i \log \lambda_i + x_i \log t_i - \log(x_i!) - \lambda_i t_i] \tag{2}$$

Under H_o the derivative of Equation 2 is obtained, set to equal 0, and solved for the parameter λ:

$$\frac{\partial \log L}{\partial \lambda} = \frac{\sum_{i=1}^{4} x_i}{\lambda} - \sum_{i=1}^{4} t_i \overset{\text{set}}{=} 0$$

$$\rightarrow \frac{\sum_{i=1}^{4} x_i}{\lambda} = \sum_{i=1}^{4} t_i$$

$$\rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{4} x_i}{\sum_{i=1}^{4} t_i} = \frac{\bar{x}}{\bar{t}} \tag{3}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i}{4}$$

and

$$\bar{t} = \frac{\sum_{i=1}^{4} t_i}{4}$$

This is the MLE (this solution maximizes the likelihood function under the null hypothesis). Thus, under H_o the maximum of likelihood function equals

$$L_o = \prod_{i=1}^{4} \frac{(\hat{\lambda} t_i)^{x_i}}{x_i!} e^{-\hat{\lambda} t_i}$$

$$= \prod_{i=1}^{4} \frac{\left(\frac{\bar{x} t_i}{\bar{t}}\right)^{x_i}}{x_i!} e^{-(\bar{x} t_i/\bar{t})}$$

$$= \exp\left(-\frac{\bar{x} \sum_{i=1}^{4} t_i}{\bar{t}}\right) \prod_{i=1}^{4} \frac{\left(\frac{\bar{x}_i}{\bar{t}}\right)^{x_i}}{x_i!}$$

$$= e^{-4\bar{x}} \prod_{i=1}^{4} \frac{\left(\frac{\bar{x} t_i}{\bar{t}}\right)^{x_i}}{x_i!} \tag{4}$$

Equation 4 will be used, shortly, as the numerator in the likelihood ratio function. Over the general parameter space where H_a:$\lambda_i \neq \lambda_k$ for some i and k, by taking the derivative of Equation 2 with respect to each λ_i we obtain

$$\frac{\partial \log L}{\partial \lambda_1} \overset{\text{set}}{=} 0 \Rightarrow \hat{\lambda}_1 = \frac{x_1}{t_1}$$

$$\frac{\partial \log L}{\partial \lambda_2} \overset{\text{set}}{=} 0 \Rightarrow \hat{\lambda}_2 = \frac{x_2}{t_2}$$

$$\frac{\partial \log L}{\partial \lambda_3} \overset{\text{set}}{=} 0 \Rightarrow \hat{\lambda}_3 = \frac{x_3}{t_3}$$

$$\frac{\partial \log L}{\partial \lambda_4} \overset{\text{set}}{=} 0 \Rightarrow \hat{\lambda}_4 = \frac{x_4}{t_4} \tag{5}$$

which is the MLE under the alternative hypothesis. Under the alternative hypothesis, where λ_i's are not the same, the maximum of likelihood function becomes

$$L_a = \prod_{i=1}^{4} \frac{\left(\frac{x_i}{t_i} t_i\right)^{x_i}}{x_i!} e^{-(x_i/t_i)} t_i$$

$$= e^{-4\bar{x}} \prod_{i=1}^{4} \frac{(x_i)^{x_i}}{x_i!} \tag{6}$$

The ratio of L_o to L_a is called the likelihood ratio and is denoted by

$$\psi = \frac{L_o}{L_a} = \frac{\prod_{i=1}^{4} \frac{(\hat{\lambda} t_i)^{x_i}}{x_i!} e^{-\hat{\lambda} t_i}}{e^{-4\bar{x}} \prod_{i=1}^{4} \frac{(x_i)^{x_i}}{x_i!}}$$

$$= \frac{\prod_{i=1}^{4} \left(\frac{\bar{x} t_i}{\bar{t}}\right)^{x_i}}{\prod_{i=1}^{4} (x_i)^{x_i}}$$

$$= \prod_{i=1}^{4} \left(\frac{\bar{x} t_i}{\bar{t} x_i}\right)^{x_i}$$

$$= \prod_{i=1}^{4} \left(\frac{\bar{x} t_i}{x_i \bar{t}}\right)^{x_i} \tag{7}$$

Hence, the test statistic, approximately, for large n is $-2 \ln \psi$, which has chi-square distribution. Specifically,

$$-2 \ln \psi = -2 \sum_{i=1}^{4} x_i [\ln(\bar{x} t_i) - \ln(\bar{t} x_i)]$$

$$= -2 \left[\sum_{i=1}^{4} x_i \ln(\bar{x} t_i) - \sum_{i=1}^{4} x_i \ln(\bar{t} x_i) \right] \tag{8}$$

is chi-square with three degrees of freedom, and the approximate size test is to reject H_o if

$$-2 \ln \psi \geq \chi^2_{1-\alpha}(4 - 1)$$

Thus, the critical region for the above test can be defined through the following form:

$$P[-2 \ln \psi \geq \chi^2_{1-\alpha}(4 - 1)] = \alpha$$

Generalization of the Test Statistic

The test statistic reached in the solution of the data given in the previous section can be generalized to cover more applications as long as the setup for Table 1 is unchanged. The general setup for this table is presented in Table 2.

In this table the unit of exposure could be any type of units used when accidents were observed, such as vehicle miles, hours, or days. The facility type refers to the place where the accidents take place. In the field of transportation people are served by a variety of facilities, including highways, intersections, local streets, and parking lots. The variables listed in this table are as follows:

TABLE 2 Accident Rate Data, General Setup

Highway Facility Type	*Characteristics*		Accident Rate
	All Accidents	Unit of Exposure	
1	x_1	t_1	λ_1
2	x_2	t_2	λ_2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
j	x_j	t_j	λ_j
Total	$\sum_{i=1}^{j} x_i$	$\sum_{i=1}^{j} t_i$	

x_i = the total number of accidents that occur in Facility i,
t_i = the unit during which accidents occur in Facility i (exposure), and
λ_i = the accident rate at Facility $i = x_i/t_i$.

Notice that the analyst could use any unit of exposure on the basis of the available data.

The variable of interest x_i is assumed to have Poisson distribution with parameter λ_i. Thus, the test statistic derived in the previous section can be slightly modified to take the following general form:

$$-2 \ln \psi = -2 \left[\sum_{i=1}^{j} x_i \ln(\bar{x} t_i) - \sum_{i=1}^{j} x_i \ln(\bar{t} x_i) \right] \tag{9}$$

where

$\bar{x}$ = mean of x_i's, $i = 1, 2, \ldots, j$;
$\bar{t}$ = mean of t_i's;
i = Facility i; and
j = number of facilities under consideration.

This test statistic can be used to detect the difference among accident rates for any type of facilities provided that the above table setup is satisfied. This test statistic is chi-square distributed with $j - 1$ degrees of freedom. Notice that this test statistic requires only the number of accidents occurring at Facility i and the desired unit of exposure during which these accidents occur. This general form is applicable for any number of transportation facilities (j).

Recall $\lambda_j = x_j/t_j$. Accident rates, in literature, are usually compared in terms of their quantities. In other words, of two locations, the one with the higher accident rate is considered to be more severe. Unfortunately, this is misleading. That is, it may be inconclusive statistically since the difference may be due to chance. The test statistic developed in this paper, however, can detect whether such a difference is significant or due to chance.

APPLICATION

In the previous sections we went step by step through the likelihood ratio technique to test the hypothesis of equal ac-

TABLE 3 Pairwise Comparisons for the Four Highway Types Presented in Table 1

The Highway Pair*	Significance at 5% level
1&2	significant
1&3	significant
1&4	significant
2&3	insignificant
2&4	significant
3&4	significant

* The numerical codes: 1 for Scenic, 2 for Other 2-lane, 3 for Multi-lane, 4 for Interstate.

cident rates given in Table 1, and we ended with the general form of a test statistic defined in Equation 9 to perform the hypothesis testing. In this section we apply this test statistic to the data given in that table. Moreover, a computer program was written to perform the computations of the test statistic.

In Table 1, four types of highways are presented. Therefore, j is 4 in our general form and we have the following hypothesis test: H_o: $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda$ and H_a: $\lambda_i \neq \lambda_k$ for some i, k. The test statistic in Equation 9 is

$$-2 \ln \psi = -2\left[\sum_{i=1}^{4} x_i \ln(\bar{x}t_i) - \sum_{i=1}^{4} x_i \ln(\bar{t}x_i)\right]$$

A comparison of the four groups indicated that the value of the test statistic is greater than 7.81 (the critical value at 0.05 level of significance), indicating that the null hypothesis is rejected and the difference among the accident rates for these four highway types is significant. In fact, this result was expected, since Table 1 indicates such differences, particularly between scenic highway and Interstate highway types.

If we decide to make pairwise comparisons, different results are obtained. For example, when we compare other two-lane with multilane ($j = 2$ in this case), the value of the test statistic is very small, namely 0.9375, which in turn means that the difference between accident rates for these two types is not significant. Table 3 presents the pairwise comparisons of the highway types given in Table 1. Figure 1 shows the pairwise comparisons. The insignificant pair is underlined in Figure 1.

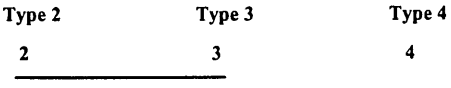

FIGURE 1 Graphical representation of pairwise comparisons presented in Table 3.

CONCLUSION

The finding of this paper was a test statistic for the comparison of accident rates in several transportation facilities. This finding was based on the assumption that such accidents were Poisson distributed. The likelihood ratio statistical technique was used to develop the test statistic. With minimal data this statistic can be adopted by traffic analysts to detect whether accident rates at several locations in a transportation system are significantly different. To show the applicability of the derived test statistic, traffic accident data from Ohio were used to compare accident rates for four highway types, including scenic, other two-lane, multilane, and Interstate. These rates were found to be significantly different. Pairwise comparisons for these types indicated that there is no significant difference between the accident rates for the other two-lane type and the multilane type. The results of this study have shown the applicability of the developed test statistic.

ACKNOWLEDGMENTS

The author wishes to thank the officials of the Ohio Department of Transportation, Columbus, Ohio, for their help in making available the data used in this study. The valuable suggestions and comments made by Ramey Rogness of the Department of Civil Engineering at The Ohio State University during the course of this research are also appreciated.

REFERENCES

1. Gerlough and Huber. *Statistics with Applications to Highway Traffic Analyses*. Eno Foundation for Transportation, Inc., Westport, Conn., 1978.
2. Gerlough and Huber. *Traffic Flow Theory: A Monograph*. TRB, National Research Council, Washington, D.C., 1975.
3. Gerlough and Barnes. *Poisson and Other Probability Distributions in Highway Traffic*. Eno Foundation for Transportation, Inc., Westport, Conn., 1971.
4. Taylor and Young. *Traffic Analysis: New Technology New Solutions*. Hargreen Publishing Co., Maryborough, Victoria, 1988.
5. Lee and Engelhardt. *Introduction to Probability and Mathematical Statistics*. Duxbury Press, Boston, 1987.
6. *Highway Safety Improvement Programs: Progress and Evaluation Report, Fiscal Year 1990*. Ohio Department of Transportation, Sept. 1990.

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