Jornal of Kerbala University , Vol. 5 No.4 Scientific .Decembar 2007

Underlying Rates of Binomial Distributed Traffic Accident Data

المعدلات الأساسية لبيانات الحوادث المرورية ذات التوزيع الثنائي الحدين

Dr. Hussein A. Ewadh[1] Dr. Abdul Hussein H. Habieb[2]

Abstract :

Traffic accident data is considered as an efficient tool to identify the degree of hazard at different locations of highway system. An accurate estimation of underlying true traffic accident rate may lead to efficient and economic safety improvement program. Accident data can be considered as random variables that have Poisson or non-Poisson distributions. A regular variation of accident data may reveal to the appropriateness of the Binomial distribution. A procedure to estimate underlying true accident rate as well as optimum time-period of accident counts for Poisson process is available while it is not for non-Poisson process.This paper proposes a new procedure to estimate the upper & lower limits of underlying accident rates depending on the observed accident rate of accident data having Binomial distribution according to different confidence degrees. The procedure includes testing data for randomness and the appropriate probability distribution that fits the data.The optimum time-period of traffic accident data provides a relatively precise estimation of underlying rate of accidents and minimizes cost of data collection as well as the social-economic losses associated in traffic accidents. A time-period beyond five years shows a relatively small decrease in the proportional uncertainty of the estimated underlying rates. Hence, a time-period of five years is sufficient for the purpose of estimation in case of binomial distributed traffic accident data. The developed procedure is a statistically reliable for purposes of programs identification of hazardous locations that may depend on the true underlying rates rather than the observed rates.

Keywords\ Traffic accident data, Binomial distribution, underlying rates, observed rates, time-period

الخلاصة :

تعد بيانات الحوادث المرورية أداة كفوءة في تشخيص درجة الخطورة على مواقع مختلفة من منظومة المرور. إن التخمين الدقيق للمعدل الأساسي الحقيقي لحوادث المرور يقود إلى برامج تحسين ذات كفاءة ومردود اقتصادي.يمكن اعتبار بيانات الحوادث على أنها متغيرات عشوائية تتوزع بتوزيع بواسون أو غيره. إن التغير المنتظم يفظي إلى أن يكون التوزيع ذي الحدين هو الأنسب. تتوافر طريقة لتخمين المعدل الأساسي في حالة توزيع يواسون, بينما هي غير متوافرة للتوزيعات الأخرى. يقدم البحث طريقة جديدة لتخمين حدود أعلى وأدنى لمعدل الحوادث الأساسي اعتمادا على المعدل المعاين لبيانات تتوزع بتوزيع ذي الحدين وبدرجات ثقة مختلفة. تتضمن الطريقة فحص البيانات فيما يخص العشوائية و التوزيع المناسب.يوفر الزمن الأمثل لفترة بيانات الحوادث المرورية تخمين دقيق نسبيا لمعدل تلك البيانات وكلفة أقل لجمع البيانات وكذلك يقلل فاقد الاجتماعي والاقتصادي الذي يتعلق بتلك الحوادث. إن زمن الفترة المساوي لخمس سنوات فأكثر يظهر نقص قليل نسبيا في نسبة عدم التأكد في تخمين المعدلات الأساسية للحوادث. نستنتج من ذلك, أن زمن الفترة المساوي لخمس سنوات هو المثالي والمناسب لغرض التخمين في حالة بيانات الحوادث ذات التوزيع ذي الحدين. إن الطريقة المستحدثة ذات مصداقية إحصائيا لأغراض برامج تشخيص المناطق الخطرة لكونها تعتمد على المعدلات الحقيقية الأساسية بدلا من المعدل المعاين لبيانات حوادث المرور.

.

Introduction:

Traffic accident represents a worldwide socio-economic problem. A loss of more than 2% of Gross Domestic Product (GDP) for most of countries in the world is only one aspect of that problem. Hence, many efforts are warranted to study the consequences and causalities of traffic accidents.

[1]Assist. Prof. Road & Traffic Engineering, University of Karbala, College of Engineering
[2] Assist. Prof. Statistics, University of Karbala, College of Management & Economy

Identification of hazardous locations is a basic element in the highway-safety improvement system. For identifying hazardous locations, almost all agencies of highway safety rely principally on traffic accident data [10]. Hence, traffic accident data is considered as the corner block of traffic accident studies.Many studies depend on estimating the observed rates to identify the hazardous spots or sections of highways. Also, highway safety improvement safety programs stated a priority and economical feasibility according to the observed rates. There are two limits of underlying true rates embedded in the random occurrence of traffic accidents. Depending on these limits, different decision may be arisen to identify the most hazardous spot or section than that relied on observed rate. Traffic accidents are considered as rare and random events. In consequence, Traffic accidents data represent random variables with a certain probability distribution. Mainly, three probability distributions are supposed to represent the random occurrence of traffic accidents; Poisson distribution, Binomial, and negative Binomial distributions [4]. A procedure to estimate the underlying true accident rates due to Poisson process is available while it is not for non-Poisson process [5].For non Poisson distributions, the five years recommended, that was found suitable to estimate a precise underlying rates in case of Poisson, may be used until a similar procedure is developed [6]. This paper proposes a procedure to estimate the underlying true rates for traffic accident data having a Binomial distribution. Further, a guide is proposed for the time-period sufficient to introduce a relatively precise estimation and minimize the data collection as well as the social-economic losses.

Traffic Accident Count Data:

Thorough studies were accomplished about traffic accident occurrence. Arranged brief conclusions of these studies can be stated as follows:

- Statistical analysis is necessary to study accident count data due to its stochastic nature of occurrence [7].
- In general, a random process governs occurrence of traffic accident [5]. However, Runs test of randomness may be used to ensure randomness before any statistical analysis.
- In many situations related to highway safety, the accident data have a distribution. Analyst should examine accident data for the goodness of fit of the expected theoretical distribution [10].
- In some situation, a Poisson process governs traffic accident counts. Too regular or too irregular variation may reveal for doubting the validity of fitting accident count data to a Poisson distribution [8].
- Accident count data may be examined for the Poisson assumption. A statistical test based on a combinatorial analysis, and the chi-square test may be used for that purpose [8]. If the accident data are Poisson distributed, the expected value of the variance- to- mean ratio is unity. However, if the variance-to-mean ratio lies outside the confidence limits that justify analyzing data according to Poisson, another distribution may be tried.

Underlying Rates of Poisson Distributed Accident Data

A procedure was developed to estimate the Underlying True Accident Rates (UTAR). The following steps summarize the developed procedure [6]:

1. The exact distribution of the sample mean ($\bar{x} = \sum x_i / n$) is governed by:

$P (\bar{x} = c/n) = (e^{-nm} (nm)^{c})/c_{!}$ -------- (1)

Where:

$c = \sum x_i$ ($x_1, x_2, x_3 \ldots x_n$, observed accidents in n years)

$\bar{x}$ = observed accident rate.

m = the underlying true accident rate corresponding to the mean of the Poisson distribution.

The relation between the cumulative sum of the Poisson distribution and the function Q (χ^2 /ν)is given by :

$Q (\chi^2 \mid \nu) = \sum(e^{-m} (m)^{j}) /j_{!}$ ---------- (2)

Where Q (χ^2 | ν) = 1- P (χ^2 | ν)
m = χ^2 /2
c = ν /2

P (χ^2 /ν) is the probability integral of χ^2 with ν degree of freedom (i.e. the cumulative χ^2 distribution)

2. The confidence limits for the underlying true accident rate are given by:

$m_l = ½ \chi^2$ (α | ν=2c)/n ---------------(3)

m_2= ½ χ^2 (1- α | ν=2c+2)/n--------- (4)

That is $m_l \leq m \geq m_2$

With level of confidence ≥ (1- 2 α)

3. Developed graphs can be used to estimate the confidence limits for the underlying true accident rates (UTAR) (m) for various values of the observed value rate x. The developed graphs are matching with the confidence parabolas illustrated by Kendall & Stuart (1979) according to an equation for the 95 percent confidence intervals for the mean of Poisson distribution as follows [1]:

$$\lambda = \left\{ \bar{x} + \frac{1.92}{n} \mp \sqrt{\frac{3.84\bar{x}}{n} + \frac{3.69}{n^2}} \right\} -- (5)$$

For an illustrated example of nine accidents observed in three years, the lower and upper 95% and 90% confidence limits due to Nicholson approach [4.5], are (1.35 and 5.7). This range is approximately similar to the range (1.58 and 5.7) obtained using equation (5) developed by Kendall & Stuart [1].

Derivation of Underlying Rates of Binomial Distributed Accident Data.

In the case of binomial distributed data, there is no general method to estimate the population mean. The sample proportion $\bar{p}$ is used to construct a confidence interval estimate of the population proportion P, then the task is similar to use $\bar{x}$ to estimate μ [3]. The population proportion P is considered as the mean of zero-one binomial population [3]. The following steps illustrate the steps of derivation of underlying rates, which is dependent to the mean of zero-one binomial population of traffic accident data:

1.Consider the following situation: x_1, x_2, -------- x_n accidents have been observed in n years.

2.Data are examined for randomness and the suitable distribution tested to fit the data is the binomial distribution.

3.x is random variable can be considered as :

$$x \sim b \quad (n, P)$$

Where:

P= the mean of zero-one binomial population, (symbol is the capital letter of P).

and the probability function is as follows [2]:

$$f\ (x) = \overset{n}{\underset{x}{C}}\, P^x (1-P)^{n-x} \; ----(x = 0,1,2,.....,n)$$

The Likelihood Function L of a binomial distribution can be stated as:

$$L(x \,|\, P) \; = \overset{n}{\underset{x}{C}}\, P^x (1-P)^{n-x} \; ----(x = 0,1,2,.....,n) --- (6)$$

Where C is constant

4. The confidence intervals for P can be determined according two equations stated as follows [1]:

$$E\left\{ \left(\frac{\partial \ln L}{\partial P} \right)^2 \right\} = -E\left\{ \frac{\partial^2 \ln L}{\partial P^2} \right\} \quad \text{-------------------}(7)$$

where E is the expected values of the specified terms.

and $\Psi = \dfrac{\partial \ln L}{\partial P} \Bigg/ \left[E\left\{ \left(\dfrac{\partial \ln L}{\partial P} \right)^2 \right\} \right]^{\frac{1}{2}}$ -----------------(8)

Where:

Ψ is a standardized normal variate corresponding to 1.96 for (1-α)=0.95.

5.The first derivative of the Likelihood Function is :

$$\frac{\partial \ln L}{\partial P} = \frac{n}{P(1-P)}\left(\frac{x}{n} - P\right)$$

Hence, $\frac{x}{n}$ is the minimum variance bound (MVB) estimator of P with variance $\frac{P(1-P)}{n}$

Let $\frac{x}{n} = p$

Where:

p = the mean of zero-one binomial sample, (symbol is the smaller letter of p).

Then:

$\dfrac{\partial \ln L}{\partial P} = \dfrac{n}{P(\,1-P\,)} \;(p-P)$ ------------------------ (9)

6.The second derivative of the likelihood function is:

$$\frac{\partial^2 \ln L}{\partial P^2} = \frac{-n}{P\,(1-P)} + \frac{n(p-P)(1-2P)}{P^2\,(1-P)^2}$$

and the expected value E of the second derivative can be stated as follows:

$-E\,\dfrac{\partial^2 \ln L}{\partial P^2} = \dfrac{n}{P\,(1-P)}$ ----------------------------(10)

By substitution equations (9) and (10) into (7) and (8) and solving the equation (8) for P, an equation can be stated to estimate mean of zero-one binomial population as follows:

$P = \dfrac{1}{1+\dfrac{\Psi^2}{n}} \left\{ p + \dfrac{\Psi^2}{2n} \mp \Psi \sqrt{\dfrac{p\,(1-p)}{n} + \dfrac{\Psi^2}{4n^2}} \right\}$ ------------------ (11)

Fortunately, equation (11) shows the same trend of confidence limits of a binomial parameters and the probability of obtaining proportion (p) illustrated by Kendall & Stuart (1979) [1] and also similar to the graphs adapted by Mills (1977) for 95% confidence interval of population proportion (P).

7.Taking $p = \frac{\mu_\circ}{n}$

Where:

$\mu_\circ$=observed traffic accident rate for a binomial distributed traffic accident data and, $\mu_{u_{1,2}} = P_{1,2}(n)$ -------------- (12)

Where:

$\mu_{u_{1,2}}$= underlying rates for a binomial distributed traffic accident data

$P_{1,2}$= the upper & lower limit of the mean of zero-one binomial population.

By substituting values of p and $P_{1,2}$ in equation (11), the underlying rates (upper and lower limits) of binomial distributed traffic accident data can be estimated by the following equation:

$$\mu_{u_{1,2}} = \frac{1}{1+\Psi^2/n}\left[\mu_o + \frac{\Psi^2}{2} \pm \Psi\sqrt{\mu_o\left(1-\frac{\mu_o}{n}\right)+\frac{\Psi^2}{4}}\right] \text{ ---------------------- (13)}$$

By no mean, equation (13) can show the variation of underlying rates with time-period. This is because there is a sensitive relation between p and n. As time-period increase, a smaller individual probability is resulted [9]. Hence, any relation of time-period with underlying rate may be examined in relation with the observed rate considering a specified value of p.

Time-Period of Accident Counts

A compatible procedure to predict the accident potential of a location is necessary to avoid much social-economic losses. Authorities of safety in the world used different time-period of traffic accident counts ranged between one to five years. A shorter time-period and relatively precise estimation of accident rate are both necessary for such procedure. Nicholson reported that the longer the time-period the greater the absolute width of confidence interval (Δ) and proportional uncertainty (Δ/ (c/n)) of estimated underlying rates of Poisson distributed data [6]. According to the charts developed by Nicholson, a time-period of five years seemed sufficient for accurate estimation of the UTAR at different sites [6]. The following equation extracted from equation (11), demonstrates the relation between the absolute length of confidence interval of the zero-one binomial population mean and time-period of data:

$$\Delta P = \frac{2\Psi}{1+\frac{\Psi^2}{n}}\left\{\sqrt{\frac{p(1-p)}{n}+\frac{\Psi^2}{4n^2}}\right\} \text{-----------------(14)}$$

Equation (14) reveals that a longer time-period assures a short absolute width of confidence interval for estimating the zero-one binomial population mean for a specified zero-one sample mean. In addition, the proportional uncertainty in the estimation of zero-one population mean can be explained by the following equation:

$$\frac{\Delta P}{p} = \frac{2\Psi}{1+\frac{\Psi^2}{n}}\left\{\sqrt{\frac{(1-p)}{np}+\frac{\Psi^2}{4n^2p^2}}\right\} \text{------------(15)}$$

In order to study the variation of estimation of underlying rates with time-period, proportional uncertainty is statistically more reliable measure than the absolute width of confidence interval. In reference to equation (12), the following relation shows that, the proportional uncertainty of underlying rates equals to that of the zero-one population mean:

$$\frac{\Delta\mu_{u_{1,2}}}{\mu_\circ} = \frac{n*\Delta P}{n*p} = \frac{\Delta P}{p}$$

Hence, equation (15) may explain the variation of the proportional uncertainty of the estimated zero-one population mean and the proportional certainty of the underlying rate, with time-period for any specified (p). Figure (1) demonstrates that variation and shows that as time-period increase, the proportional uncertainty of both estimations, decreases yielding to a relatively constant variation at long time- period of observation. The ranges of p (0.1, 0.2, 0.3, 0.4, 0.5) shown in Figure (1) reveal to the same conclusion for ranges of (0.9, 0.8, 0.7, 0.6, 0.5) respectively.

Figure (1) shows that the knees of the curves in occur at time-period of approximately five years. This may assure that a time-period of five years is an optimum time-period of accident counts for the purpose of the estimation of zero-one population mean as well the underlying rates for Binomial distributed data.

This finding gives a further support to the conclusion drawn by Nicholson [6] that five years is sufficient for the estimation of Poisson distributed accident data. A study of negative-Binomial distributed accident data is appreciated to make a universal vision about the optimum time-period of accident counts.

Illustrative Example of Accident Data Analysis

The procedure of analysis and estimation of underlying rates for traffic accident counts that are expected to be binomial distributed is illustrated for counts of 10 observations at a specified hazardous section of a highway in the following sequence:

7 , 4, 5, 3, 7, 6, 6, 8, 5, 6

Step 1: Test for randomness

In reference to the accident counts of the example, a median of (6), is used as specified value to obtain (4) runs as follows [2,9]:

/ + / -, -, - / +, 0, 0, + / -, 0 /

The range of critical limits using 0.05 degree of significance is 2 to 10. Hence, one cannot reject the null hypothesis and may conclude that the sequence is very probably random.

Step 2: Check for Binomial distribution

Although the accident counts in this example seem to be regular and the ratio of mean to variance is less than unity (0.353), it is intended to check the appropriateness of binomial distribution. Goodness-of- fit test reveals that one cannot reject the binomial assumption for that set of data (χ^2 tabulated = 14.67, χ^2 calculated =2.02).

The test of chi-square was found suitable to test the data for Poisson distribution [8]. As an alternative to the goodness-of-test and for simplicity, it is intended to accomplish chi-square test in the case of binomial distributed data. If the test shows that one may reject the appropriateness of Poisson distribution and the ratio of variance to mean is less than one, it can be concluded that the binomial distribution is the appropriate one. For the case of 57 accidents in 10 years, 90% confident that the variance-to- mean ratio will lies outside the critical range (0.369 to 1.880) due to the χ^2 test. Hence, one may reject the use of the Poisson distribution for analyzing the accident data for that that location and conclude that the binomial distribution is the appropriate alternative one.

Step 3: Estimation of underlying rates

The observed rate of accident data in the example is 5.7 accidents per year. The time-period is more than 5 years. Hence, a precise estimation may result from such set of data. According to the equation (13), the underlying rates for the accident data are 2.89 to 8.12. These rates are approximated to 3 to 8 accidents per year. These limits give a confident vision about the true occurrence traffic accidents at the studied location. Any safety improvement at that location should be studied according the same analysis to ensure the effectiveness of the used countermeasure. If there is no overlap between the confidence intervals of before and after underlying rates, one may assure that the countermeasure is effective.

Conclusions

Estimation of underlying rates of binomial distributed traffic accident data is developed according to a mathematical statistics. The paper proposes an equation to estimate the upper and lower limits of underlying rates due to different degrees of confidence. Estimation of underlying rates is based on finding confidence limits of zero-one population mean of Binomial distribution. The paper demonstrates a typical procedure for estimation of underlying rates of accident counts that is expected to be a binomial distributed.

The optimum time-period of traffic accident data that maximizes the precision of estimation and minimizes cost of data collection as well as the social-economic losses associated in traffic accidents. In view of the research**,** it can be concluded that:

- The proportional uncertainty of underlying rates equals to that of the zero-one population mean.
- As time-period of accident counts (n) increases, the proportional uncertainty of the estimated; zero-one population mean $\frac{\Delta P}{p}$ and underlying rates $\frac{\Delta \mu_{u_{1,2}}}{\mu_{o}}$ for any specified p are decreased yielding to a relatively constant variation at long time- period of observation.
- The knees of the curves of proportional uncertainty occur at time-period of approximately five years. This may assure that a time-period a time-period of five years is an optimum time-period of accident counts for the purpose of the estimation of zero-one population mean as well as the underlying rates for Binomial distributed data.

This paper introduces a procedure of a statistical reliability for the authorities of road safety to identify the true underlying rates rather than the observed rates. The decision makers may make use of that procedure to identify the significance of a proposed countermeasure for safety improvements. No overlap between before and after confidence interval of underlying rates may assure the effectiveness of safety improvement. A procedure for estimation of underlying rates of accident counts that may have a trend of negative-binomial distribution is highly appreciated.

References:

1. Kendall, S.M, and Stuart, A. *The Advanced Theory of Statistics*, Vol.2, 4th edition, Butler& Tanner, U.K., 1979.

2. Mann, P.S. *Introductory Statistics.* 5th edit. John Wiley & Sons, Inc. U.S.A. 2004.

3. Mills, R.L. *Statistics for Applied Economics and Business*. McGraw-Hill Book Co., USA, 1977.

4. Nicholson, A.J. “The Variability of Accident Counts” *Accident Analysis & Prevention”* Vol.17, No.1, pp. 47-56, 1985.

5. Nicholson, A.J. “The Randomness of Accident Counts” *Accident Analysis & Prevention”* Vol.18, No.3, pp. 193-198, 1986.

6. Nicholson, A.J. “The Estimation of Accident Rates and Countermeasure Effectiveness” *Traffic Engineering and Control”* Vol. 28, No.10, pp.518-523, 1986

7. Nicholson, A.J. “Understanding the Stochastic Nature of Accident Occurrence” *Australian Road Research”* , Vol.21, No.1, pp. 30-39, 1991.

8. .Nicholson, A.J., and Wong, Y.D. “Are Accidents Poisson Distributed? A Statistical Test” *Accident Analysis & Prevention”,* Vol.25, No.1, pp. 91-97, 1993.

9. Pelosi, M.K. and Sandifer, T.M. *Elementary Statistics.* John Wiley & Sons, Inc. U.S.A. 2003.

10. Transportation Research Board (TRB). *Methods for Evaluating Highway Safety Improvements*. National Cooperative Highway Research Program Report 162. Washington, D.C. U.S. 1975.

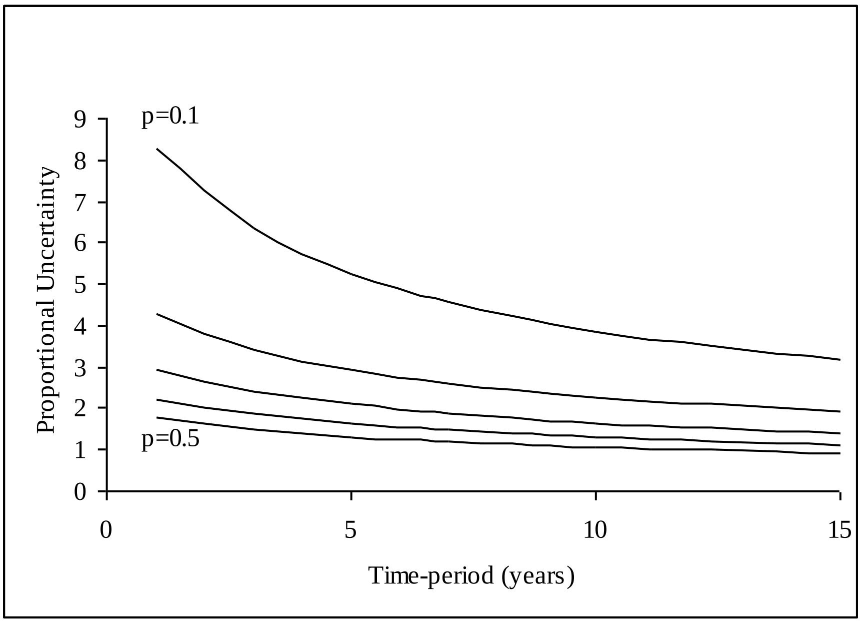

Figure (1) Effect of Time-Period on Proportional Uncertainty of; Zero-One Binomial Population Mean ($\frac{\Delta P}{p}$) & Underlying Rates ($\frac{\Delta \mu_{u_{1,2}}}{\mu_\circ}$).