# Risk management for traffic safety control

Malwina J Luczak*· Jonathan Wylie†

**Abstract**

This paper offers a range of modelling ideas and techniques from mathematical statistics appropriate for analysing traffic accident data for the East region operation of CLP Power Hong Kong Limited and for the Hong Kong population in general. We further make proposals for alternative ways to record and collect data, and discuss ways to identify the major contributing factors behind accidents. We hope that our findings will enable the design of effective accident prevention strategies for CLP.

## 1 Introduction

CLP Power Hong Kong Limited is Hong Kong's largest energy supplier. The East Region operation of the company owns a large fleet comprising of 84 company vehicles of various types that are driven by 273 authorised drivers, the majority of whom are tradesman drivers. The majority of CLP drivers are not trained as professional drivers. CLP believe that their accident rate may be too high and are seeking objective methods to determine an acceptable rate.

Our aim is to study road accident data for CLP. If the rate turns out too high, according to some reasonable criteria, then further study should be carried out to identify the chief causes of accidents. Statistical tools should then be applied to test whether any new safety measures bring significant improvements.

The costs incurred by CLP as a result of accidents are relatively low, approximately HK$100 per driver per year. Therefore it is crucial that any accident prevention strategy be highly targeted and cost-efficient for it to be worthwhile. Certain strategies are relatively straightforward to implement. For instance, one could isolate the drivers with the largest number of accidents and subject them to appropriate training. However, to be sure that any prevention strategy is well-focused, we must have a high degree of confidence that our conclusions are not influenced by random fluctuations. The confidence in test results will necessarily increase together with quantity of data used to perform these tests. Occasionally, the amount of data at hand is not satisfactory, and then it is a major challenge to find reliable ways of testing for significance.

The data presently at our disposal consists of the total number of traffic accidents per year, plus the total number of kilometres driven by the entire fleet during the years 1998 - 2001. Let us point out that the term 'accident rate' may be defined in a number of ways; it could mean the number of collisions per, say, 1000 kilometres, the number of collisions in a certain fixed number of days, or the number of collisions in a fixed number of trips. Tests may lead to very different conclusions, depending on which definition is adopted. Given the nature of the data in question, and the generally held belief that the mileage is the primary factor influencing accident rates, our focus will be on the average number of collisions per certain fixed number of kilometres. However, any one of the other parameters could be estimated via exactly the same techniques.

* Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, UK. e-mail: M.J.Luczak@statslab.cam.ac.uk

† Department of Mathematics, City University of Hong Kong, Hong Kong.
e-mail: wylie@math.cityu.edu.uk

The natural target for the CLP accident rate is the expectation of the accident rate for the entire Hong Kong's population. Relevant data that can be used to estimate this expectation is available from the Hong Kong Transport Department. However, several difficulties present themselves when one attempts to draw a comparison. Firstly, the definition of an 'accident' is bound to differ between the two data sets. The CLP criteria are almost certainly more stringent than the Hong Kong Transport Department ones. This is because minor crashes involving members of the public may be settled without police involvement, whereas any such incidents would have to be reported by CLP drivers, so that vehicles involved may be repaired. Further, driving conditions experienced across the two sets are not believed to be the same. For example, CLP drivers drive much less frequently at night than a member of the public, and CLP alcohol rules are far more stringent than those in place for ordinary Hong Kong residents. Despite such limitations, a comparison of this kind might still be a useful and instructive exercise, particular when examining year-on-year trends. Nevertheless, it would be far more informative to look at the performance of drivers in some other Hong Kong companies with similar profile and similar needs if such data could be found.

One could easily start questioning the usefulness of analysis based on rather crude data, such as described above. Any findings will be extremely unlikely to help pinpoint major causes of accidents involving company drivers. For this reason it is important to adopt a systematic and detailed method of collecting and recording data. In fact, motor insurance companies have already carried out extensive research on the causes of road accidents using various statistical techniques to analyse vast quantities of carefully collected data. Typically, the following factors are found to affect the probability of a driver being involved in a road accident: his or her age, vehicle type, time elapsed since the driving test, mileage, history of accidents and injuries, history of repair costs, occupation, primary usage of vehicle, and the history of driving convictions. Of these criteria, occupation and usage will be the same for most CLP drivers. Marital status and driving convictions might be inappropriate due to privacy and legal issues. Additionally, one should factor in circumstances such as road and weather conditions, time of day, work patterns, and the distance driven during the day.

With such data in hand it is feasible to come up with statistical tests to determine the key factors that increase the rate of accidents or lead to the highest repair costs. not just the result of chance events. The information obtained through the tests should then enable design of an effective accident reduction strategy. Without the knowledge based on analysis of statistical data, any strategy would necessarily have to rely on guesswork. On the other hand, identification of drivers with a significantly higher rate of accidents would make it possible to institute additional training where required. Determining problem vehicles might lead to re-examination of decisions on purchases or leases. Identification of problem work conditions or habits would afford CLP the opportunity to educate drivers and raise their awareness of work-related risks.

A few common-sense ideas also come to mind. For instance, perhaps a reduction in the number of accidents could be achieved through improved scheduling of tasks. One could try matching drivers who are statistically safest to vehicles and tasks that are statistically most dangerous. Certain tasks could be scheduled so as to enable drivers to avoid rush hours, etc. Naturally, any such precautions will be limited by numerous practical constraints. It would be wise to first estimate their expected benefits if a well-informed management decision is to be taken.

Additionally, we recommend examining the temporal homogeneity of the data. The rate of accidents may well vary over time. For instance, drivers might be more vigilant following a serious crash, then after a certain period grow complacent, and hence be more likely to be involved in an accident. Similarly, weather and road conditions change fairly frequently. However, in order to test such variations more detailed data is required than that so far made available to us.

Different tests require different form and structure of data. For example, there are only a small number of vehicle types, and it is straightforward to test for significant variations in accident rates among the different types. On the other hand, the age of the drivers varies continuously. We may either divide them into discrete age groups, or instead choose to use a non-parametric technique to analyse the accident rate as a function of driver age.

The following sections contain the technical details of the study. We first carry out tests suited to the type of data that we have been provided with, namely the number of accidents within a fixed distance driven. However, we stress that increasing the level of detail in gathering data increases the range of techniques available for use in analysis. We outline some of these techniques at the end.

## 2 Probability model

Our first task is to determine whether the mean accident rate for CLP drivers differs significantly from the mean rate for Hong Kong overall. We remarked in the introduction that the accident rate can be defined in more than one way. It can mean the number of collisions within a fixed distance driven, the number of collisions during a fixed length time period, or the number of collisions in a fixed number of trips. Here we consider the number of collisions per fixed distance driven.

Since there is far more data on road accidents available for Hong Kong overall than for CLP drivers, we feel that the following approach will be most appropriate. We adopt the same probability model for both sets of data, involving some unknown parameter whose value is not necessarily the same in both. We then estimate the value of the parameter for the Hong Kong model, and test whether the corresponding value for the CLP model equals our estimate. Intuitively, the amount of data for the Hong Kong public at large is so much larger than the corresponding amount for CLP, that if we were to perform some kind of two-sample test, the information concerning CLP would get 'lost' among the information pertaining to all of Hong Kong. This approach might also be considered in comparing the accident rate for a particular type of vehicle in the CLP fleet with the rate for the rest of the fleet.

We start by obtaining a reliable estimate of the accident rate $\mu_0$ per specified fixed distance driven for Hong Kong overall. Subsequently, we shall carry out a hypothesis test for the CLP data. This data is assumed to come from the same distribution, possibly with a different parameter $\mu$. Thus we test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$ (two-sided test) or $H_1 : \mu > \mu_0$ (one-sided test).

Let us start by making a general preliminary comment about goodness of fit. The Poisson distribution has often been used to model the frequency of occurrence of events such as accidents in fixed intervals of time or distance. Previous studies [7, 6, 4] of sequences of industrial and road accidents have demonstrated it is usually a good fit at least to a rough approximation. Nevertheless, we should test to see the validity of the Poisson assumption for our data. If significant discrepancies from a Poisson model were found either for CLP or for Hong Kong, it would be necessary to look for another model. Possible reasons why the Poisson model might occasionally be unsuitable would be variations of accident rate over time or strong correlations among drivers.

The usual approach is to draw a probability plot. The general technique works as follows. Suppose we have a random sample $X_1, \dots, X_n$, assumed to come from a distribution $F$. Let $X_{(1)}, \dots, X_{(n)}$ be the corresponding order statistics. We plot ordered data values $X_{(i)}$ against $F^{-1}(i/(n+1))$. If the assumed model is correct, then the plot should be an approximate straight line. There are also ways to test the homogeneity of data using the likelihood ratio statistic. This is all based on standard theory, which can be found for instance in [5] and in many other sources, such as undergraduate and graduate textbooks.

## 3 Parametric tests involving the Poisson distribution

We can model the number of accidents in a given fixed distance interval, say, 1000 kilometres, by a random variable $X$ which has a Poisson distribution with mean $\mu$. If $Z$ is the number of accidents in $n$ days, or in $1000n$ kilometres, then $Z$ has a Poisson distribution with mean $n\mu$, provided that accidents in any two disjoint intervals are independent. Now suppose we have a random variable

$$Z_X = \sum_{i=1}^{n} X_i,$$

where $X_i$ are independent Poisson random variables, and we want to test the hypothesis:

$$H_0 : \mu_Z = n\mu_0$$

against the hypothesis

$$H_1 : \mu_Z > n\mu_0,$$

where $\mu_Z$ is the expected value of $Z$, that is $\mu_Z = \mathbf{E}[Z]$.

In our case $\mu_0$ will be the expected number of accidents per 1000 kilometres for the Hong Kong public. The test assumes that we have a reliable estimate of that expectation, which, as remarked above, is a separate problem. However, one that is not too difficult to deal with, since the amount of data for Hong Kong is large enough to enable us to find a fairly precise confidence interval for the parameter in question.

Generally, in a one-parameter model with log-likelihood $l(\theta)$, and observations $x_1, \dots, x_n$ (these are realizations of the random variables $X_1, \dots, X_n$), the observed information is

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}.$$

The expected information or Fisher's information (see [1, 5]) is

$$I(\theta) = -\mathbf{E}\left[\frac{\partial^2 l(\theta)}{\partial \theta^2}\right].$$

When observations are independent, then the likelihood $L(\theta)$ is a product of densities, so

$$J(\theta) = -\sum_i \frac{\partial^2}{\partial \theta^2} \log f(x_i, \theta).$$

Let $\hat{\theta}$ be the value that maximises $L(\theta)$, that is the maximum likelihood estimate. We have for $\theta$ near $\hat{\theta}$,

$$l(\theta) \approx l(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2 J(\hat{\theta}).$$

When the number of observations $n$ is large, then $\hat{\theta}$ is approximately normal. To be precise, under certain regularity conditions we have

$$\hat{\theta} \to N(\theta, I(\theta)^{-1}),$$

where the convergence is understood in distribution. Thus an asymptotic approximate $(1-\alpha)\%$ confidence interval is $\hat{\theta} \pm \Phi^{-1}(1-\alpha/2)I(\hat{\theta})^{-1/2}$ or $\hat{\theta} \pm \Phi^{-1}(1-\alpha/2)J(\hat{\theta})^{-1/2}$, where $\Phi$ is the distribution of a standard normal $N(0,1)$ random variable (with zero mean and unit variance). When $X_1, \dots, X_n$ are independent Poisson with mean $\theta$, then the maximum likelihood estimator is the sample mean:

$$\hat{\theta} = \bar{x} = \frac{1}{n}\sum_i x_i.$$

The observed information and Fisher's information are $J(\theta) = n\bar{x}/\theta^2$, and $I(\theta) = n/\theta$. Thus it is straightforward to calculate the maximum likelihood estimator for Hong Kong data, and then use it to test the CLP data as described below.

## 3.1 Uniformly most powerful test

We start by stating the Neyman-Pearson Lemma [5], see also standard statistics textbooks.

**Lemma 3.1.** *Let* $\mathbf{X} = X_1, \dots, X_n$ *be a random sample from a distribution with parameter* $\theta$, *where* $\theta \in \Theta = \{\theta_0, \theta_1\}$, *and let* $L(\mathbf{x}, \theta)$ *be the likelihood function. If there exists a test at significance level* $\alpha$ *such that, for some positive constant* $k$,

1. $L(\mathbf{x},\theta_0)/L(\mathbf{x},\theta_1) \le k$, *for each* $\mathbf{x} \in C_1$ *(that is, for all* $\mathbf{x}$ *in the critical region),*

2. $L(\mathbf{x},\theta_0)/L(\mathbf{x},\theta_1) > k$, *for each* $\mathbf{x} \in C_0$ *(that is, for all* $\mathbf{x}$ *outside the critical region),*

*then this test is most powerful at significance level* $\alpha$ *for testing the null hypothesis* $H_0 : \theta = \theta_0$ *against the alternative hypothesis* $H_1 : \theta = \theta_1$.

The Neyman-Pearson Lemma deals with simple $H_0$ versus simple $H_1$. However, we can try to find a uniformly most powerful test of size $\alpha$ (that is, a test that is most powerful for each simple alternative hypothesis in $H_1$). Let $\mu_1$ be an arbitrary point in $(\mu_0, \infty)$ and consider testing $H_0 : \mu_Z = n\mu_0$ versus $H_1 : \mu_Z = n\mu_1$. The likelihood ratio is

$$\frac{f_1(\mathbf{x},\mu_1)}{f_0(\mathbf{x},\mu_0)} = \frac{e^{-n\mu_1}\prod_i {\mu_1}^{x_i}/x_i!}{e^{-n\mu_0}\prod_i {\mu_0}^{x_i}/x_i!} = e^{-n(\mu_1-\mu_0)}\left(\frac{\mu_1}{\mu_0}\right)^{\sum_i x_i}.$$

If $\mu_1 > \mu_0$, then this is monotonic increasing in $\bar{x} = \frac{1}{n}\sum_i x_i$ for any fixed $\mu_1$ and $\mu_0$, and so the likelihood ratio critical region whereby $H_0$ is rejected in favour of $H_1$ if $f_1(\mathbf{x},\mu_1)/f_0(\mathbf{x},\mu_0) > k_\alpha$, say, is equivalent to the region that rejects $H_0$ when $n\bar{x} > c_\alpha$ for a suitable $c_\alpha$. This critical region is most powerful for any $\mu_1 > \mu_0$ and so is uniformly most powerful.

For an $\alpha$ size test we require

$$\mathbf{Pr}_{\mu_0}(n\bar{x} > c_\alpha) = \alpha.$$

Now by the above, $Z$ is Poisson with mean $n\mu_0$ under $H_0$, and so we want

$$1 - \varphi_{n\mu_0}(c_\alpha) = \alpha = 0.05,$$

say, where $\varphi_\mu(k) = \sum_{x=0}^{k} \mu^x e^{-\mu}/x!$. The probability of type II error when $\mu_Z = n\mu_1$ is

$$\beta = \mathbf{Pr}_{n\mu_1}(n\bar{x} \le c_\alpha) = \varphi_{n\mu_1}(c_\alpha),$$

and the power of the test is equal to $1-\beta$. Note that the power depends on the actual mean $\mu_Z$.

When $n$ is large, then by the Central Limit Theorem, under the null hypothesis $Z$ is approximately normally distributed, $Z \sim N(n\mu_0, n\mu_0)$. So we could compare the statistic $(n\mu_0)^{-1/2}(\sum X_i - n\mu_0)$ against statistical tables for a standard normal $N(0,1)$.

## 3.2 Two-sample approximate $t$-test

It is possible to carry out a two-sample test, which does not entail first estimating the mean for Hong Kong. However, for reasons given above, this might turn out not to be very informative. We divide the distance driven by CLP drivers into $n$ fixed length intervals, and $X_i$ will be the number of accidents in the $i$-th interval. We divide the period or distance in miles driven by all Hong Kong drivers into $m$ fixed length intervals, and $Y_i$ will be the number of accidents in the $i$-th interval. We assume that $X_i$ and $Y_i$ are all independent Poisson with mean $\mu_X$ and $\mu_Y$ respectively. We wish to test the hypothesis $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X \neq \mu_Y$. We can do this using the asymptotic approximate normality of $\sum_i X_i$ and $\sum_i Y_i$.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and let $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ be the sample means for the $X_i$ and $Y_i$. Let $S_{XX} = \sum_i(X_i - \bar{X})^2$ and $S_{YY} = \sum_i(Y_i - \bar{Y})^2$ be the sample variances. When $n$ and $m$ are large, then the statistic

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_{XX}+S_{YY}}{n+m-2}(\frac{1}{n} + \frac{1}{m})}}$$

has an approximate $t$-distribution with $n + m - 2$ degrees of freedom. Thus we can calculate its value for our data and compare against $t$-distribution tables.

# 4 The likelihood ratio test

The theory presented here is based on the approach in [5], but can also be found in numerous statistics textbooks. Suppose we want to test in a situation where the adopted probability model involves several unknown parameters (in this case, the mean number of accidents per, say, 1000 kilometres for Hong Kong, and the corresponding mean number for CLP Power). Let $\Theta$ be the parameter space and let $\Theta_0, \Theta_1$ be subsets of $\Theta$. We want to test the null hypothesis $\theta \in \Theta_0$ against the alternative $\theta \in \Theta_1$. We use the likelihood ratio, $\lambda(\mathbf{x})$ defined as

$$\lambda(\mathbf{x}) = \frac{\sup\{L(\mathbf{x},\theta) : \theta \in \Theta_0\}}{\sup\{L(\mathbf{x},\theta) : \theta \in \Theta_1\}}, \qquad \mathbf{x} \in \mathbb{R}^n_X.$$

Here the numerator and denominator represent the likelihood of seeing what we have seen under the null and alternative hypotheses respectively. In particular, if $\Theta_1 = \Theta$, then for a realization $\mathbf{x}$, we determine its best chance of occurrence under $H_0$ and its best chance overall. The ratio can never exceed unity, but, if small, would constitute evidence for rejection of the null hypothesis.

Since the function $-2\log\lambda(\mathbf{x})$ is decreasing in $\lambda(\mathbf{x})$, it follows that the critical region of the likelihood ratio test can also be expressed in the form

$$C_1 = \{\mathbf{x} : -2\log\lambda(\mathbf{x}) \geq c\}.$$

The statistic $\Lambda(\mathbf{x}) = -2\log\lambda(\mathbf{x})$ is called the likelihood ratio statistic. There is a general asymptotic result for the likelihood ratio statistic, namely that under certain regularity conditions $\Lambda$ converges in distribution to a random variable $\chi^2_p$, where $p = \dim\Theta - \dim\Theta_0$. This follows, as one might expect, from the Central Limit Theorem. We omit the proof details, as they are quite involved and technical in nature.

For our data, we can use the likelihood ratio test in a number of ways. First, just as before we could test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. This form of test relies once again on having a good estimate of the mean number of accidents per fixed length interval for Hong Kong. Here, the dimensions of $\Theta_0$ and $\Theta$ are zero and unity respectively, so that $p = 1$. Also, it is not difficult to check that

$$\Lambda(\mathbf{x}) = 2n[\mu_0 - \bar{x} + \bar{x}\log(\bar{x}/\theta_0)].$$

Alternatively, we could proceed as follows. Let $X_1, \ldots, X_n$ represent the data corresponding to CLP Power, that is $X_i$ is the number of accidents in a fixed interval involving company drivers. We assume that $X_1, \ldots, X_n$ are independent Poisson each with mean $\mu_X > 0$. Let $Y_1, \ldots, Y_m$ be the data corresponding to Hong Kong, that is $Y_i$ is the number of accidents in a fixed interval (of the same length as before) for Hong Kong overall. We assume that $Y_1, \ldots, Y_m$ are independent Poisson each with mean $\mu_Y > 0$. We can take

$$\Theta_0 = \{(\mu_X, \mu_Y) : \mu_X = \mu_Y\}.$$

The likelihood ratio here is given by

$$\lambda(\mathbf{X}, \mathbf{Y}) = \frac{(n+m)^{-(\sum_i X_i + \sum_i Y_i)} (\sum_i X_i + \sum_i Y_i)^{\sum_i X_i + \sum_i Y_i}}{n^{-\sum_i X_i} m^{-\sum_i Y_i} (\sum_i X_i)^{\sum_i X_i} (\sum_i Y_i)^{\sum_i Y_i}}.$$

For computational purposes, it is convenient to rewrite the above expression as

$$\lambda(\mathbf{X}, \mathbf{Y}) = \left( \frac{1 + \frac{\sum_i Y_i}{\sum_i X_i}}{1 + \frac{m}{n}} \right)^{\sum_i X_i} \left( \frac{1 + \frac{\sum_i X_i}{\sum_i Y_i}}{1 + \frac{n}{m}} \right)^{\sum_i Y_i}.$$

Additionally, the likelihood ratio statistic can also be used in testing for variations in rate of accidents over time and space. Thus it can settle the issue of the homogeneity of the data mentioned in section 2.

## 5 Analysis of time intervals between accidents

The range of statistical methods available to us is greater if data is collected with more detail and care. For instance, it would be useful to record the date and time of each accident. Then the basic data consists of a sequence of intervals of varying length. Analysis may therefore be applied to the time intervals (or distance intervals) between accidents as well as to the frequencies of accidents occurring in successive fixed intervals. If we can assume that accidents are taking place at random in time and at a constant average rate, we shall obtain the same estimate of this rate either from the average number of accidents occurring in successive fixed intervals or from the average length of the varying interval between accidents. However, if we require more details, such as in testing changes in time, then methods of analysis based on accurate interval data will be more powerful than those often employed using only the accident frequencies in relatively long fixed intervals. We now present various methods available for use in analysis. Methods of this kind were used in [6] and [4] to analyse the records of time intervals between explosions in coal mines during the period from 15 March 1851 to 22 March 1962. The reader is referred to these two papers for more details.

### 5.1 Goodness of fit

First of all we can test for goodness of fit by drawing a probability plot, as described earlier, or drawing a histogram and fitting an exponential curve. If the fit appears good, we can proceed to carry out an analysis which relies on the properties of the exponential distribution. For more details about goodness of fit testing the reader is referred to [6, 4, 2]. The last reference contains a brief analysis of time intervals between the failure of air-conditioning equipment in aircraft.

### 5.2 The distribution of the sample mean of intervals

If interval lengths $T_1, \dots, T_n$ are independent exponential random variables with mean $\lambda$, and $\bar{T} = \frac{1}{n}\sum_i T_i$, then $2n\lambda T$ is $\chi^2$ with $2n$ degrees of freedom. Then, with obvious notation, $[\chi^2_{\alpha/2}/2n\bar{t}, \chi^2_{1-\alpha/2}/2n\bar{t}]$ is a $(1-\alpha)\%$ confidence interval for the unknown parameter $\lambda$, and should give an accurate estimate when $n$ is large. We could use this to estimate the mean separation $\lambda_0$ in time or space between two accidents involving Hong Kong drivers. We could then calculate the value of that statistic for data concerning CLP drivers, and test the null hypothesis $H_0 : \lambda = \lambda_0$ against the alternative $H_1 : \lambda \neq \lambda_0$ (two-sided test), or against $H_1 : \lambda > \lambda_0$ (one-sided test).

### 5.3 Ratio of two sample means

If $\bar{T}_1$ and $\bar{T}_2$ are sample means for independent samples of $n_1$ and $n_2$ intervals, then $\lambda_1\bar{T}_1/\lambda_2\bar{T}_2$ will have an $F_{n_1,n_2}$ distribution. Thus to test whether $\lambda_1 = \lambda_2$, we may refer $\bar{t}_1/\bar{t}_2$ to the tables of the $F$ distribution. In this way we could test for significant differences between the accident risk for CLP drivers and Hong Kong drivers, assuming that intervals between accidents are independent exponential random variables.

### 5.4 Extreme observations in samples from an exponential distribution

The application of the above tests depends on the assumption of homogeneity. Tests for homogeneity may be based on the distribution of extreme intervals. Let $T_{(n)}$ and $T_{(1)}$ be the longest and shortest among $n$ independent intervals. We could carry out tests using statistics $T_{(n)}$, $T_{(n)}/n\bar{T}$, $T_{(n)}/T_{(1)}$, $T_{(n)} - T_{(1)}$. Something else that we can do is break the whole sequence of intervals into $k$ groups of $m$ successive intervals and test for a significant difference between the $k$ means, each of which is an estimate of $\lambda$. In this way, we could detect variations in the process over time, for instance a decrease in accident rate.

One may wonder what to do if we happen to have some unusually long or short intervals: are these simply outliers resulting from random fluctuations or do they constitute significant evidence against the probability model adopted? In such cases, whether the exponential distribution should still be the basis of future application depends on the purpose. If occurrences of very short or very long intervals are of concern, then it would be unwise to use it. However, if the main interest lies in the mean, the standard deviation or simply in rough extrapolation, then the exponential model would still be sensible.

## 6 Non-parametric methods

In parametric hypothesis testing the distribution under the null hypothesis is either completely specified or is given except for a finite number of unknown parameters. Consider a situation in which the null hypothesis involves explicitly or implicitly, arbitrary and usually unknown densities. The words *distribution-free* and *nonparametric* are used broadly for the resulting techniques.

### 6.1 Median test

One simple nonparametric test is the so-called median test. Suppose we have a sample $X_1, \dots, X_n$ from an unknown distribution $F$. Let $X_{(1)}, \dots, X_{(n)}$ be the corresponding order statistics. Let $x_M$ denote the median of $F$, that is $x_M = F^{-1}(1/2)$. Then for $0 \leq r < s \leq n$, we have

$$\Pr(X_{(r)} < x_M < X_{(s)}) = \sum_{k=r}^{s-1} \binom{n}{k} \left(\frac{1}{2}\right)^n .$$

If the above expression takes value $1-\alpha$ for a suitable $\alpha$, then $(X_{(r)}, X_{(s)})$ is a $(1-\alpha)\%$ confidence interval for $x_M$. Given a realisation $x_{(1)}, \dots, x_{(n)}$ of the order statistics, we obtain such a confidence interval $(x_{(r)}, x_{(s)})$, and then for any value $x$ within that interval we cannot reject the null hypothesis that the median equals $x$ at the $(1-\alpha)\%$ significance. Now for many common probability distributions (generally, ones that are highly concentrated around the mean), the mean and median are very close, and so we can also make statements about the mean of such a distribution on the basis of the median test.

We refer the reader to [1] for a description of some other nonparametric tests. In general, these are rather more involved than parametric tests, and their use involves a certain loss of efficiency and power. On the whole, provided data are screened for outliers, results of nonparametric tests are not very different from those of analogous parametric tests. Also, their main emphasis lies on avoiding assumptions about the distribution; in many applications, however, the most critical assumptions are those of independence.

function $g(a) = \mathbf{E}(X|A=a)$. model and make use of the Maximum Likelihood Estimate technique as explained above. is to use non-parametric methods.

## 7 A related study

In this section we briefly discuss a related report, namely [3]. The main purpose of that project was to conduct a detailed statistical investigation into the effects of safety cameras on the reduction of accident numbers. Other goals were a comparison of the road safety for Cambridgeshire and the whole of Britain, and also a detailed analysis of the effects of factors like speed and seasonality on the distribution and severity of accidents.

Cambridgeshire appears to have a poor accident record if we consider the number of injury accidents per heads of population. However, in [3] that comparison was shown to be unfair. In fact, the level of traffic in Cambridgeshire is higher than the national average, and the analysis showed that the overall safety record for Cambridgeshire is just as good as the safety record for all of Great Britain. The main difficulty lay in lack of reliable data on the number of miles travelled in Cambridgeshire per year; the data available pertained only to the annual increase in traffic.

The analysis of accident causes used contingency tables, probability plots and generalised linear models. Clear relationships emerged between factors like speed, weather and seasonality and the severity of injuries suffered in accidents.

The effects of safety cameras are extremely hard to measure. For instance, it is not feasible to fit a sensible model at a single site where the average number of accidents is less than one a month due to too much random fluctuations in the data. This excludes application of techniques such as time series. However, a new method developed in [3] managed to demonstrate that safety cameras have indeed had a positive effect on road safety.

The thesis also considers different ways of modelling accident data, such as time series, moving averages procedures and exponential smoothing techniques. Such models can then be used to predict future observations (such as the number of accidents per month) as an estimate of the current level of the series which is a weighted average of past observations up to the current point in time.

**Acknowledgement** We would like to thank all other people who have worked on this problem: S. Ben Hariz, A. Fitt, Z. Hao, J. Ockendon, K. Tam, L. Zehui.

## References

[1] Cox, D.R. and Hinkley, D.V. (1979). *Theoretical Statistics.* London: Chapman and Hall.

[2] Cox, D.R. and Snell, E.J. (1981). *Applied Statistics: Principles and Examples.* London: Chapman and Hall

[3] Hess, S. (2002). *A statistical analysis of the effects of safety cameras on traffic accident rates in Cambridgeshire.* M.Phil. Thesis. University of Cambridge.

[4] Jarrett, R.G. (1979). A note on the intervals between coal-mining disasters. *Biometrika* **66**, 191-3.

[5] Lunn, D. (1996). Lecture Notes for b8 Statistics. University of Oxford.

[6] Maguire, B.A., Pearson, E.S., and Wynn, A.H.A. (1952). The time intervals between industrial accidents. *Biometrika* **39**, 168-80.

functions. *Ann. Math. Statist.* **27** 832.

[7] Whitworth, W.A. (1901). *Choice and Chance.* Cambridge: Deighton Bell and Co.