PB2001-108193

DEVELOPMENT AND TESTING OF METHODS FOR ESTIMATING THE IMPACT OF SAFETY IMPROVEMENTS

Final Report

Prepared by:

Gary A. Davis, Ph.D.

Department of Civil Engineering
University of Minnesota
500 Pillsbury Drive SE
Minneapolis, MN 55455

March 2001

Published by:

Minnesota Department of Transportation
Office of Research Services
Mail Stop 330
395 John Ireland Boulevard
St. Paul, MN 55155

This report represents the results of research conducted by the authors and does not necessarily represent the views or policies of the Minnesota Department of Transportation. This report does not contain a standard or specified technique.

REPRODUCED BY: NTIS
U.S. Department of Commerce
National Technical Information Service
Springfield, Virginia 22161

Technical Report Documentation Page

1. Report No. MN/RC – 2001-08	2.	3. Recipients Accession No.

4. Title and Subtitle DEVELOPMENT AND TESTING OF METHODS FOR ESTIMATING THE IMPACT OF SAFETY IMPROVEMENTS	5. Report Date March 2001 6.
7. Author(s) Gary A. Davis, Ph.D.	8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Civil Engineering University of Minnesota 500 Pillsbury Drive SE Minneapolis, MN 55455	10. Project/Task/Work Unit No. 11. Contract (C) or Grant (G) No. c) 74708 wo) 7
12. Sponsoring Organization Name and Address Minnesota Department of Transportation 395 John Ireland Boulevard Mail Stop 330 St. Paul, Minnesota 55155	13. Type of Report and Period Covered Final Report 1996-1999 14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract (Limit: 200 words)

This report describes a Bayesian method for estimating accident rates at individual sites, which takes into account the fact that the total traffic count usually used to measure exposure is generally not known with certainty.

The first step involves deriving an approximation for the probability distribution of total traffic conditioned on a short count sample. This approximation is then used to drive a Bayes estimator of a site's accident rate, conditioned on an accident count, a short count sample, and the total traffic approximation.

The method then uses Gibbs sampling to compute accident rate estimates. Tests based on actual accident and traffic data revealed that accident rate estimates based on a two-week traffic sample area are almost as accurate as estimates based on full traffic counting, but that uncertainty in the estimated accident rates increase by 20 to 50 percent when using a two-day count sample.

17. Document Analysis/Descriptors Traffic counts Accident rates Gibbs sampling	18. Availability Statement No restrictions. Document available from: National Technical Information Services, Springfield, Virginia 22161

19. Security Class (this report) Unclassified	20. Security Class (this page) Unclassified	21. No. of Pages 61	22. Price

ACKNOWLEDGMENTS

The author would like to thank the Office of Transportation Data Analysis and the Office of Traffic Engineering of the Minnesota Department of Transportation, particularly Mark Flinner, Darab Bouzarjomehri and Loren Hill, for providing the data used in this project. The author would also like to thank Chen Wei for processing the accident data, and Shimin Yang for conducting the evaluation of the total traffic predictions.

PROTECTED UNDER INTERNATIONAL COPYRIGHT
ALL RIGHTS RESERVED
NATIONAL TECHNICAL INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE

Reproduced from
best available copy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This report deals with the problem of estimating the accident rate at a specified roadway site when the total traffic at the site is not known with certainty, by must be estimated from a traffic count sample. Although the standard statistical methods for estimating accidents rates implicitly assume that total traffic is known, in almost all practical cases this is not so. Neglecting the additional uncertainty arising from estimating the total traffic can lead to falsely identifying a site as a potential high-hazard location, or to falsely concluding that a safety treatment has had a significant effect on accident occurrence. After briefly reviewing standard practice in accident rate estimation, it is argued that a Bayesian approach to statistical inference more nearly matches the requirements and expectations of safety engineers.

Consideration of accident rate estimation reveals that standard methods in essence use the sample to compute a heuristic prediction of total traffic, but that no formal treatment of this prediction problem has been done. Chapter 2 provides this treatment, and builds on earlier work by the PI to develop a Bayesian method for making such predictions. The accuracy of this predictor was evaluated by using data from 48 Mn/DOT automatic traffic recorders (ATR) , in which 90% confidence intervals for the total 1991 traffic volume at each ATR were computed using traffic samples from 1992, and these intervals were then compared to the actual 1991 traffic totals.

Chapter 3 then describes how the total traffic prediction model can be formally incorporated into accident rate estimation, and how an advanced numerical technique called Gibbs Sampling can be used to estimate accident rates with traffic count samples. Total accident counts for 1992 were prepared for Mn/DOT's outstate ATRs, and for 17 ATR sites accident rate estimates and confidence intervals were computed for full traffic counts, a two-week count sample and a two-day count sample. It was found that the precision obtained from the two-week sample was similar to that obtained from a full count, but that the two-day sample estimates were noticeably less precise. It is recommended that a two-week traffic count be included in the suite of engineering studies conducted when selecting countermeasures for a potentially high-hazard site, and that Bayesian methods be incorporated in network wide analyses, such as identifying potential high-hazard sites, when they become computationally friendly.

CHAPTER 1
INTRODUCTION

This report describes part of a long-term effort by the Principal Investigator to place statistical practice in traffic engineering on a sounder theoretical base, and to bring the state of statistical practice in traffic engineering more nearly in line with the state of the art. Many of the statistical techniques taught to and used by traffic engineers date back to the middle part of the last century, and are based on simplified statistical models adapted to the conceptual and computational constraints of those times. Although in some instances a simple statistical model is perfectly adequate, such as using the normal distribution to describe homogeneous vehicle spot speeds, in other instances the simplified model ignores important problem features, and this can lead to overstating the accuracy of an estimation technique. For example, Davis (1997a, 1997b) pointed out that published statements of the accuracy achieved by two-day traffic counts in estimating annual average daily traffic (AADT) implicitly assumed an unrealistic ability to determine correct seasonal and day-of-week adjustment factors. An alternative method for estimating AADT, employing a Bayesian statistical approach and based on a well-confirmed statistical model for a locations's daily traffic volumes, was described in Davis and Guan (1996) and Davis (1997a).

The estimation of traffic accident rates at individual roadway sites provides another example of how unrealistic assumptions are needed to justify a relatively simple statistical technique. In this case, it is assumed that the total traffic at the site is known with certainty, when in almost all practical cases the total traffic is also a quantity that must be estimated. This uncertainty regarding total traffic makes accident rate estimation more akin to an errors-in-variables regression problem rather than to a standard regression problem, and it is well-known that the statistical properties of standard regression estimates break down when there are unacknowledged errors-in-variables (Seber and Wild, 1989). In this report, we address the estimation of traffic accident rates at individual roadway sites when the total traffic volume using that site is not known with certainty, but must be estimated from a sample of daily traffic counts. The emphasis will be on estimating the accident rates at individual sites, rather than on estimating a common rate for a number of sites, because recent findings concerning the tendency of accident counts to be over-dispersed imply that the statistical models supporting common rates are also over-simplified. In what follows we will often use the traditional term "accident" rather than the currently popular term "crash" , partly out of consistency with past usage, and partly because crashes include events, such as suicides and homicides, whose prevention is not usually affected by safety countermeasures. In this view, the objective of traffic safety engineering is prevention of unintended accidents, rather than deliberate crashes. However, this is primarily a semantic quibble, and we will follow Hauer (1997) in using "crash" and "accident" as synonyms.

In the remainder of this chapter we will first consider how uncertainty in estimates of AADT can affect estimates of accident rate. We will then review the statistical theory for estimating accident rates when the full traffic count is known with certainty, and argue that a Bayesian approach to statistical inference more nearly accommodates the requirements of traffic engineers. In Chapter 2 we will consider the problem of estimating the total traffic volume from a traffic count sample, and develop an approximation to the predictive distribution of the total traffic. We will then present some empirical results supporting the use of this approximation in practice. In Chapter 3 we will take up the problem of estimating accident rates given an accident count and a traffic count sample, and show how a Bayesian approach to this problem can be implemented using a relatively new computational technique called Gibbs sampling. We will then compare the precision achieved in estimating accident rates with a full traffic count, a two-week sample of daily counts, and a two-day sample of daily counts. The comparison will be empirical, using traffic count and accident data from 17 Minnesota outstate automatic traffic recorder (ATR) sites. Finally, Chapter 4 will present our conclusions and recommendations.

ACCIDENT RATES AND AADT

Estimated accident rates are used in several safety engineering activities, such as identification of potential high-hazard sites, predicting the benefits expected from an accident countermeasure, and evaluating the impact of a countermeasure once it is in place. The commonly recommended estimator of a site's accident rate takes the form

$$\hat{\lambda} = \frac{x}{N(AA\hat{D}T)} \tag{1.1}$$

where

$\hat{\lambda}$ = estimated accident rate at the site of interest,

x = the site's accident count over some specified time period,

N = number of days over which the accident count was made,

$\widehat{AADT}$ = estimated AADT for the site.

For example, if over a one-year period a section of road had an observed accident of count of 10, and an estimated AADT of 5500 vehicles/day, the estimated accident rate would be

$\hat{\lambda} = 10/(365*5500) = 4.98\times10^{-6}$ accidents/vehicle.

Because an accident count can vary randomly even though the underlying features which make a site more or less dangerous have not changed, equation (1.1) gives an estimate rather than the (unknown) true accident rate. This means that there will be uncertainty concerning the degree to which the estimate approximates the true rate, and when one attempts to assess this uncertainty it is important to recognize that both the numerator and the denominator in equation (1.1) are random quantities and that both contribute to the overall uncertainty. Ignoring the uncertainty which arises because the traffic total has also been estimated can lead one to overstate the accuracy attached to the accident rate estimate, which in turn could lead to falsely identifying a site as potentially hazardous, or falsely concluding that a safety treatment has had a significant effect. In particular, since an estimate of AADT appears in the denominator of equation (1.1), the issues concerning accurate estimation of AADT carry over to the estimation of accident rates.

In order to determine how uncertainty in the estimate of total traffic can influence uncertainty in the estimated accident rate, it is first necessary to select a measure of how close an estimate tends to be to the (unknown) true rate. One commonly used measure is the mean squared error (MSE), defined as

$$MSE = E[(\hat{\lambda}-\lambda)^2] = VAR[\hat{\lambda}] + (E[\hat{\lambda}]-\lambda)^2 \qquad (1.2)$$

where E[.] denotes the expected value of the random variable appearing inside the square brackets and VAR[.] denotes the corresponding variance. Note that the right hand side of equation (1.2) shows that the MSE can be decomposed into two components, the first term being the variance of the estimator $\hat{\lambda}$ and the second being the square of the estimator's bias. For an unbiased estimator the second term vanishes and MSE is equal to the estimator's variance, so that when an estimator's performance is evaluated in terms of its variance, there is an implicit assumption that the bias of the estimator is negligible. A more intuitive way of presenting MSE is to express the square root of the MSE as a proportion of the true accident rate, a quantity which will be called the root-mean squared proportional error (RMSPE). Formally, this quantity is given by

$$RMSPE = \frac{\sqrt{E[(\hat{\lambda}-\lambda)^2]}}{\lambda} \tag{1.3}$$

and can be interpreted as an expected distance between the estimated rate and the true rate, expressed as a proportion of the true rate.

To illustrate how uncertainty in AADT affects the estimated accident rate, we will consider estimation for the simple case where the analyst has an accident count x on hand, but must estimate the total traffic from only a single day's traffic count, denoted by $\tilde{z}$. This traffic count scenario is chosen primarily for its analytic simplicity, but is not completely unrealistic since some jurisdictions do estimate AADT from 24-hour counts. It is well known that daily traffic volumes vary both throughout the year and within the week, so to capture these effects the expected value of the daily count will be assumed to follow a multiplicative model of the form

$$E[\tilde{z}] = MWz_0 \tag{1.4}$$

where

M = monthly adjustment factor,

W = day of week adjustment factor,

z_0 = AADT.

In this case the adjusted count $\tilde{z}/(MW)$ is an unbiased estimate of the AADT, so this model is consistent with current practice in adjusting traffic count samples. It will also be assumed that the variance of the daily count is given by

$$VAR[\tilde{z}] = (MWz_0)^2 v^2 \tag{1.5}$$

so that v denotes the coefficient of variation for the daily traffic counts. Next, let z_T denote the (unknown) true traffic total at the site, and let the accident count x be the outcome of a Poisson random variable with mean λz_T. Finally, we will assume that the analyst does not know the site's true adjustment terms M and W, but rather must use estimates $\hat{M}$ and $\hat{W}$. The corresponding estimate of the total traffic, which forms the denominator of the estimated accident rate, is simply

$$\hat{z}_T = N\left(\frac{\tilde{z}}{\hat{M}\hat{W}}\right) \tag{1.6}$$

and it is straightforward to verify that

$$\begin{aligned} E[\hat{z}_T] &= (Nz_0)\left(\frac{MW}{\hat{M}\hat{W}}\right) \\ VAR[\hat{z}_T] &= \left((Nz_0)\left(\frac{MW}{\hat{M}\hat{W}}\right)\right)^2 v^2 \end{aligned} \tag{1.7}$$

Hauer (1997) has illustrated how statistical differentials can be used to derive approximate expressions for the expected value and variance of the ratio between two random quantities, and this method can be applied to the accident rate estimate given in equation (1.1) to produce

$$\begin{aligned} E[\hat{\lambda}] &\approx \frac{E[x]}{E[\hat{z}_T]}\left(1+\frac{VAR[\hat{z}_T]}{E[\hat{z}_T]^2}\right) \approx \lambda\left(\frac{\hat{M}\hat{W}}{MW}\right)(1+v^2) \\ VAR[\hat{\lambda}] &\approx \frac{E[x]^2}{E[\hat{z}_T]^2}\left(\frac{VAR[x]}{E[x]^2}+\frac{VAR[\hat{z}_T]}{E[\hat{z}_T]^2}\right) \approx \lambda^2\left(\frac{\hat{M}\hat{W}}{MW}\right)^2\left(\frac{1}{\lambda z}+v^2\right) \end{aligned} \tag{1.8}$$

where the right-most approximations use the fact that $z_0 \approx (z_T/N)$. These approximations can then be substituted into equation (1.3) to give an expression for the RMSPE.

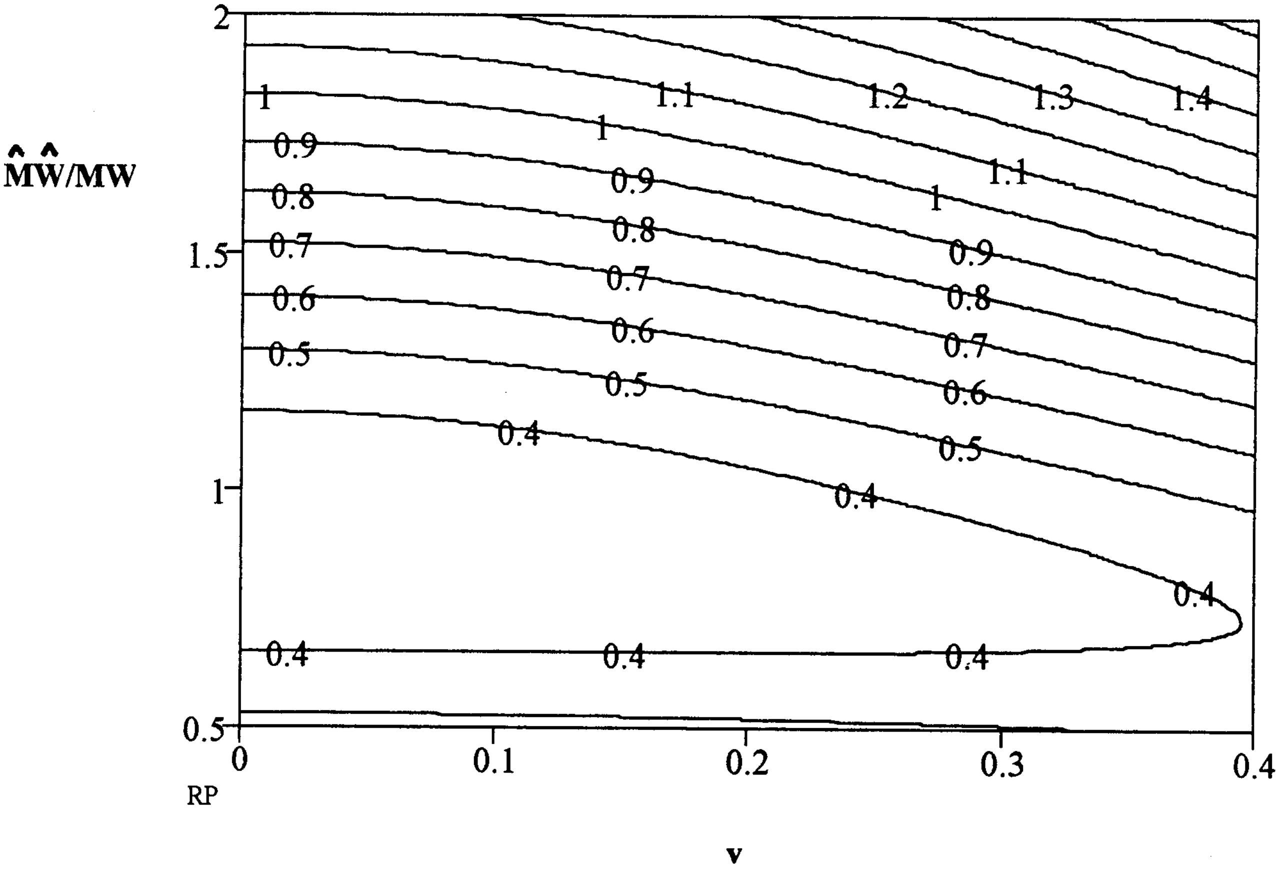
$$RMSPE \approx \sqrt{\left(\left(\frac{\hat{M}\hat{W}}{MW}\right)^2\left(\frac{1}{\lambda z_T}+v^2\right)+\left(\left(\frac{\hat{M}\hat{W}}{MW}\right)(1+v^2)-1\right)^2\right)} \tag{1.9}$$

Equation (1.9) reveals that RMSPE is determined by three quantities, the expected accident count λz_T, the coefficient of variation for the traffic counts, v, and the ratio of the estimated adjustment terms to the actual adjustment terms, $\hat{M}\hat{W}/MW$. To illustrate how these terms interact, Figure 1.1 shows a contour plot of RMSPE as a function of the ratio ($\hat{M}\hat{W}/MW$) and the coefficient of variation v, for the case where the expected accident count λz_T=10. The ranges for ($\hat{M}\hat{W}/MW$) and v were chosen to correspond to those observed in an analysis of daily traffic counts from the 50 Minnesota automatic traffic recorders (ATR) described in (Davis, 1997a). Inspection of Figure 1.1 reveals that although RMSPE increases rather slowly as v increases, it is somewhat more sensitive to the potential biases introduced when estimated adjustment factors must be used to estimate a site's AADT. For example, when $\hat{M}\hat{W}/MW$=1.0, so that an unbiased estimate of AADT is used, and v=0.1, RMSPE=0.33, meaning that on average the estimated accident rate differs from the true rate by 33%. But if one were unlucky in picking the adjustment terms, so that $\hat{M}\hat{W}/MW$=1.5, RMSPE more than doubles, to 72%. It was pointed out in (Davis, 1997b) that in the past the problem of reliably matching a short-count site to a set of adjustment factors tended to be ignored, but more recent work indicates that a site can often be reliably matched to one of a small set of factor groups using a count sample consisting of at least two well-chosen weeks. On the other hand, the standard 24 or 48-hour counts will generally not be sufficient for reliable matching, so that unless an analyst has either good prior knowledge concerning the correct adjustment factors or an atypically large traffic count sample, adjustment bias will introduce uncontrolled and usually unacknowledged error into the accident rate estimate.

COMPARING BAYESIAN AND FREQUENTIST APPROACHES

Statistical methods for accident rates usually begin with the assumption that accidents at a roadway site are "rare events," so that the Poisson distribution can be used to describe the probability of observing a given accident count. Formally, if X denotes a random variable describing the accident count at a site during some specified time period , z_T denotes the total traffic volume at the site during the same time period, and λ denotes the site's accident rate, then

Figure 1.1 Contour Plot Showing RMSPE As a Function of Adjustment Ratio and Coefficient of Variation



$$P[X=x|\lambda,z_T] = \frac{e^{-\lambda z_T}(\lambda z_T)^x}{x!}, \quad ,x=0,1,2,... \tag{1.10}$$

Strictly speaking, the accident rate λ is the expected increase in the accident count resulting when one additional vehicle uses the site, which justifies its use as a measure of the site's accident hazard. Informally, it can be interpreted as the probability an arbitrarily selected vehicle has an accident at the site. When one has an observed accident count and when the traffic volume at the site is known exactly, frequentist estimation of the accident rate begins with the fact that the maximum likelihood estimate of λ is given by

$$\hat{\lambda} = \frac{x}{z_T} \tag{1.11}$$

and that the variance of maximum likelihood estimator is

$$VAR_X\left[\frac{X}{z_T}\right] = \frac{\lambda}{z_T} \approx \frac{x}{z_T^2} \tag{1.12}$$

This estimator has several desirable frequentist properties. In particular, the maximum likelihood estimator is unbiased, that is

$$E_X\left[\frac{X}{z_T}\right] = \lambda \tag{1.13}$$

and efficient, i.e.

$$VAR_X\left[\frac{X}{z_T}\right] \leq VAR_X[\tilde{\lambda}] \qquad (1.14)$$

where $\tilde{\lambda}$ denotes any other estimator of λ. Further, as the expected accident count (λz_T) becomes large, the distribution of the maximum likelihood estimator can be approximated by a normal distribution with mean equal to λ and variance equal to (λ/z_T). An approximate 95% confidence interval for λ can then be computed as

$$\left[\frac{x}{z_T}-(1.96)\sqrt{\frac{x}{z_T^2}}\ ,\ \frac{x}{z_T}+(1.96)\sqrt{\frac{x}{z_T^2}}\right] . \qquad (1.15)$$

We might be tempted to believe at this point that the unbiasedness and efficiency properties of the maximum likelihood estimator permit us to conclude that the estimated rate can be taken as a reasonable approximation of the unknown true rate, and that the confidence interval gives us a range within which the unknown rate lies, but this is not so. In words, an unbiased estimator is one for which, if our data collection experiment were repeated an infinite number of times and the estimate of the accident rate computed for each repetition, the average of these estimates would equal the true rate, while an efficient estimator is one for which the scatter of these estimates about the true rate is minimal. Neither of these properties tells us anything about the relation of a specific estimate to the (unknown) true rate. Similarly, the appropriate frequentist interpretation of the confidence interval is that if we were to compute lower and upper bounds for our infinite number of samples according to equation (1.15), 95% of the time these bounds would catch the unknown true rate, but for any individual confidence interval, the unknown rate is either inside or outside the interval. The frequentist statistical theory can thus be viewed as giving the safety engineer a set of gambling rules, which if followed under appropriate conditions will produce specified results on average over the long run, but no claim is made concerning performance in any particular instance.

In contrast to this rather restricted view of what statistical inference gives us, we might consider what safety engineers think they are getting. As a proxy for a survey of engineers we will look at what a non-random sample of standard references says about confidence intervals. First, in two texts written by statisticians for use by engineers, we find the orthodox frequentist interpretation of the meaning of a confidence interval:

"According to this interpretation, the confidence level 95% is not so much a statement about any particular interval such as (79.3, 80.7) but pertains to what would happen if a very large number of like intervals were to be constructed." (Devore, 1995, p. 279);

"This last probability statement must be interpreted very carefully. It does not mean that the probability of the parameter μ falling into the specified interval equals 2Φ(z)-1; μ is a parameter and either is or is not in the above interval. Rather, the above should be interpreted as follows: 2Φ(z)-1 equals the probability that the random interval

$$(\bar{X}-z\sigma/\sqrt{n}\ ,\ \bar{X}+z\sigma/\sqrt{n})$$

contains μ." (Meyer, 1970, p. 304).

When we turn to materials written by engineers for engineers we find a clear dissatisfaction with the frequentist interpretation:

"A precise definition of a (1-α)100 per cent confidence interval for μ would involve the following statement: 'If in a series of very many repeated experiments an interval such as the one calculated were obtained, we would in the long run be correct (1-α)100 per cent of time in claiming that μ is located in the interval (and wrong 100α percent of the time.)' Unfortunately, such a statement is sometimes difficult to comprehend for the nonstatistician and awkward to interpret operationally. Therefore we shall consider a (1-α)100 per cent confidence interval as a range within which we are (1-α)100 per cent sure that the true parameter is contained,..." (Hahn and Shapiro, 1967, p. 75);

"More properly, then, the interval estimated on the basis of a single sample of size n should be interpreted as follows: 'There is a confidence of (1-α) that the estimated interval contains the unknown μ.'" (Ang and Tang, 1975, p. 234).

Finally, in materials written by traffic engineers for traffic engineers, we do not even find lip service being paid to the frequentist view:

"It can be said with 95% confidence that the true mean speed of all traffic is within the range defined by the observed mean plus or minus twice the standard error" (Homburger and Kell, 1988, p. 6-6);

"This can be read as 'The probability that the true mean lies in the interval

$$\left(\bar{X}-1.96\frac{\sigma}{\sqrt{N}}\ ,\ \bar{X}+1.96\frac{\sigma}{\sqrt{N}}\right)$$

is 0.95'" (McShane and Roess, 1990, p. 129);

"The interpretation of the probability statement for an interval estimate is that the confidence interval contains the population value with a probability that is equal to the confidence coefficient,"(Box and Oppenlander, 1976, p. 198, King, 1994, p. 406);

"This means that it is 95% certain that the true arithmetic mean of the universe lies between 51.01 and 52.99," (Greenshields and Weida, 1952, p. 143, Gerlough and Huber, 1978, p. 101).

If the statements in these texts can be seen as reflecting practice, then it is clear that traffic engineers view the confidence coefficient for a confidence interval as reflecting the degree of certainty attached to a statement that the true parameter is within the interval's bounds. The clear gap between what the frequentist interpretation guarantees from confidence intervals and what traffic engineers think they provide might be taken to imply that engineers are ignorant or being careless. But before fixing on this uncharitable view, let us consider the alternative approach.

Taking a Bayesian approach to the rate estimation problem, we begin as before with the Poisson accident model, but in addition we need to specify a prior probability density characterizing our uncertainty concerning the unknown accident rate λ before obtaining the data (x,z_T). Bayes Theorem is then used to determine how our uncertainty about the site's accident rate should rationally change after the data become available. For example, suppose that our prior uncertainty can be characterized by a gamma probability density of the form

$$f(\lambda) = \frac{\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)\beta^{-\alpha}} \qquad (1.16)$$

It is well known that combining this with the Poisson accident model via Bayes Theorem produces a gamma posterior density, and that a point estimate of the site's accident rate can be given as the posterior expected value

$$E[\lambda|x,z_T] = \frac{\alpha+x}{\beta+z_T} \tag{1.17}$$

The uncertainty concerning this estimate can be summarized by the posterior variance, which for the gamma posterior has the form

$$VAR[\lambda|x,z_T] = \frac{\alpha+x}{(\beta+z_T)^2} \tag{1.18}$$

Comparing equation (1.11) and (1.12) to (1.17) and (1.18) it can be seen that when the parameters α and β, which characterize the gamma prior density, take on small values, the computational outcomes of the Bayesian analysis and of the frequentist analysis are essentially the same. This suggests that when safety engineers perform the likelihood-based computations but then interpret the results as providing information concerning the particular site's accident rate, they are actually conducting a Bayesian analysis, using frequentist formulas to approximate the desired Bayesian quantities. So, rather than suggesting that safety engineers are sloppy frequentists, we think it more appropriate to see them as approximate Bayesians.

We have argued that a Bayesian approach to statistical inference more closely represents engineering practice. We now turn to the situation where the total traffic count z_T is unknown, but where one has a sample of daily counts $\mathbf{z}=[z_1,..,z_n]'$ (the prime symbol $'$ is used to denote the transpose of a vector or matrix.) Taking a Bayesian approach, our objective is to somehow determine the posterior distribution $f(\lambda \mid x,\mathbf{z})$, and in principle this can be expressed in terms of the "full count" posterior $f(\lambda|x,z_T)$ via the expression

$$f(\lambda|x,z) = \int_0^\infty f(\lambda|x,z_T)f(z_T|x,z)dz \tag{1.19}$$

The density $f(z_T \mid x, \mathbf{z})$ is known as the predictive density of the traffic total, and it can be seen that if we could determine a form for this predictive density, the Bayesian estimation problem could in principle be solved by an additional integration. In the Chapter 2 we will consider the somewhat simpler problem of characterizing the predictive distribution $f(z_T \mid \mathbf{z})$, where we are attempting to predict the total traffic volume using a count sample. These results will then be used in Chapter 3 to develop a computationally feasible method for evaluating (1.19).

CHAPTER 2

ESTIMATING TOTAL TRAFFIC VOLUME FROM SAMPLE COUNTS

As noted in Chapter 1, estimating a site's accident rate when we do not have an actual count of the total traffic at that site requires that we estimate the total traffic somehow, usually using a short-count sample of the daily traffic. The objective of this chapter is to develop a useable expression for the predictive distribution of a site's total traffic, conditional on a sample. Before proceeding, let us introduce the following notation

z_t = traffic count on day t, t=1,..,N

$z^N = \sum_t z_t$, the total traffic count over days t=1,..N

z_l = traffic count on the l-th sample day, l=1,..,n

$\mathbf{z} = [z_1,..,z_n]'$, n-dimensional vector containing the sample counts.

In Chapter 1 we introduced the predictive density $f(z^N|\mathbf{z})$, but here it will turn out that approximations and computations are more easily carried out using the corresponding conditional cumulative distribution function $F(z^N \mid \mathbf{z})=P[Z^N \leq z^N \mid \mathbf{z}]$. Since the total traffic count z^N is determined as the sum of the daily counts z_t, the statistical properties of the daily traffic volumes will determine the form of the conditional distribution $F(z^N \mid \mathbf{z})$.

LOGNORMAL APPROXIMATION

We will develop an explicit expression for $F(z^N|\mathbf{z})$ in several steps. The first step is to characterize the statistical properties of the daily traffic volumes z_t, and in a detailed analysis of daily traffic counts from 50 automatic traffic recorders (ATR) in Minnesota, Davis (1997a) showed that the daily traffic volumes could be described using a lognormal regression model of the form

$$y_t = u + \sum_{i=1}^{12} \Delta_{t,i} m_{k,i} + \sum_{j=1}^{7} \delta_{t,j} w_{k,j} + e_t \qquad (2.1)$$

where

$y_t = \log_e(z_t)$, the natural logarithm of a daily count,

u = expected log traffic count on a typical day,

$\Delta_{t,i}$ = 1, if the count z_t was made during month i, i=1,..,12,
0, otherwise;

$m_{k,i}$ = correction term for month i, characteristic of factor group k,

$\delta_{t,j}$ = 1, if the count z_t was made on day-of-week j, j=1,..,7,
0, otherwise,

$w_{k,j}$ = correction term for day-of-week j, characteristic of factor group k,

e_t = random error.

If we let $\beta_k=[m_{k,1},..,m_{k,12}, w_{k,1},..,w_{k,7}]'$ denote a column vector containing the monthly and day-of-week adjustment terms for factor group k, and $x_t=[\ \Delta_{t,1}\ ,...,\ \Delta_{t,12}\ ,\ \delta_{t,1}\ ..,,\ \delta_{t,7}\ ]$, the above equation can be written in a slightly simpler form

$$y_t = u + x_t\beta_k + e_t \tag{2.2}$$

In the above model the mean value of the logarithm of the daily count varies according to month and day of week, and the magnitude of these variations depends on the factor group to which the site of interest belongs. In addition, analyses of the regression residuals indicated that the error terms e_t showed day-to-day dependencies which could be described by a multiplicative autoregressive (AR) model of the form

$$e_t = \phi_1 e_{t-1} + \phi_7 e_{t-7} - \phi_1\phi_7 e_{t-8} + a_t \tag{2.3}$$

where the a_t are independent, identically distributed normal random variables with common variance and mean equal to 0, and ϕ_1 and ϕ_7 are autoregressive coefficients.

The above model is parameterized by u, a mean-value parameter, β, the monthly and day of week adjustment terms, σ^2, the variance of the e_t terms, and the autoregressive coefficients ϕ_1, ϕ_7. In the next step, we will assume we know the values of these parameters, but nothing else about the site. Properties of lognormal random variables can be used to show that the expected value of the total traffic volume is

$$\mu_N = E[z^N] = \sum_{t=1}^{N} \exp(u+x_t\beta_k+\frac{\sigma^2}{2}) \tag{2.4}$$

and the variance of the total traffic volume is

$$\begin{aligned} v_N^2 &= e^{\sigma^2}(e^{\sigma^2}-1)\sum_{t=1}^{N} e^{2x_t\beta_k} \\ &+ (e^{2u+\sigma^2})(\sum_{t=1}^{N} e^{x_t\beta_k}(\sum\nolimits_{s\neq t} e^{x_s\beta_k}(e^{\rho_{t,s}\sigma^2}-1))) \end{aligned} \tag{2.5}$$

where $\rho_{t,s}$ is the correlation between e_t and e_s, which can be computed from the noise covariance parameters ϕ_1, ϕ_7 using standard time-series methods (e.g. Brockwell and Davis,1991). Now since z^N is a sum of lognormal random variables, characterizing its probability distribution turns out to be very difficult even for the case N=2, and no result for general N is known (Johnson, Kotz and Balakrishnan, 1994). In most situations of practical interest however N will be equal to the number of days in one or more years, so that an asymptotic approximation might prove useful. In the Appendix it is shown that the cumulative distribution function of

$$\frac{\mu_N}{v_N}\log_e\left(\frac{z^N}{\mu_N}\right) \tag{2.6}$$

converges to that of a standard normal random variable, implying that for large N, $\log_e(z^N)$ is approximately a normal random variable with mean equal to $\log_e(\mu_N)$ and variance equal to (v_N^2/μ_N^2). This in turn supports using a lognormal approximation for z^N,

$$P[z^N \le \tilde{z} | u,\beta,\phi_1,\phi_7,\sigma^2] = P[\log_e(z^N) \le \log_e(\tilde{z}) | u,\beta,\phi_1,\phi_7,\sigma^2] \approx \Phi\left(\frac{\log_e(\tilde{z}) - \log_e(\mu_N)}{\frac{v_N}{\mu_N}} \right) \qquad (2.7)$$

where $\Phi(.)$ denotes the standard normal probability distribution function.

The sample **z** provides two types of information concerning z^N. On the one hand, it provides information concerning the model parameters, and in principle this information could be summarized by the posterior distribution function $F(u,\beta, \phi_1, \phi_7,\sigma^2 \mid \mathbf{z})$. On the other hand the correlated noise model (2.3) implies that any given daily count z_t is correlated with its neighbors, so that knowing z_t allows us to more accurately predict neighboring values. If the noise model (2.3) is stationary, and if the sample counts are sufficiently separated in time from the counts comprising z^N (such as might occur when trying to predict the total volume for 1999, using a sample taken in 1997) we can take the sample and the total count as being independent of each other, and then the sample provides information on the total only by providing information on the model parameters. The distribution $F(z^N \mid \mathbf{z})$ could then be found as

$$F(z^N|z) = \int F(z^N|u,\beta,\phi_1,\phi_7,\sigma^2)\; dF(u,\beta,\phi_1,\phi_7,\sigma^2|z) \qquad (2.8)$$

When the sample counts are correlated with counts comprising the total z^N, the expression (2.8) will only be approximate, with the approximation deteriorating with increasing overlap between the sample counts and the counts entering into the total. In principle, smoothing algorithms could be used to include dependency on **z** in (2.4) and (2.5), so that a **z**-dependent asymptotic approximation could be developed, but the necessary computational labor appears to be substantial. For the situations commonly encountered in highway and safety engineering, the number of counts entering into the total will be large compared to size of the sample (e.g. one or more years for N, compared to a 48-hour or two one-week counts for n), so that most of the aggregating counts will be separated from suitably accurate approximation, and some empirical evidence supporting this conjecture will be presented later.

The final steps involve characterizing the distribution F(u,β, ϕ_1, ϕ_7,σ^2 | **z**) and then finding a computationally feasible way to evaluate the (multidimensional) integral in (2.8). It turns out however that this problem is very similar to the problem of computing Bayes estimates of mean daily traffic described in Davis (1997a) and Davis and Guan (1996), and a similar solution can be employed here. In this approach, a prior distribution describing our uncertainty concerning the parameter values is combined with the likelihood function of the count sample, and Bayes Theorem is used to derive the posterior distribution for the parameters, given the count sample.

As in Davis (1997a), we will assume that the highway agency has divided its road segments into a set of m factor groups, and that estimates of the adjustment factors for each group, β_k, k=1,..,m, are available. We will further assume that the agency maintains a total of M ATRS, and that for each ATR estimates of the covariance parameters (ϕ_{1p},ϕ_{7p},σ_p), p=1,..,M, are also available. Prior to collecting any data for a site, we will assume that our uncertainty concerning that site's parameters is captured by the prior probability distribution

$$g(u,\beta,\sigma,\phi_1,\phi_7) = \left(\frac{1}{M}\sum_{p=1}^{M} I_{(\sigma_p,\phi_{1p},\phi_{7p})}(\sigma,\phi_1,\phi_7)\right)\left(\frac{1}{m}\sum_{k=1}^{m} I_{\beta_k}(\beta)\right) \qquad (2.9)$$

where $I_b(x)$ denotes a Kronecker delta function

$I_b(x) = 1$, if $x=b$,
0, otherwise.

Basically, this prior assumes that we are completely uncertain as to the value of u in the sense that our prior probability is uniformly distributed on the real line. For the adjustment terms β, we are certain that it takes on one of the values β_1,..,β_m characterizing our factor groups, but that we are equally uncertain as to which of these is correct. Similarly, for (ϕ_1,ϕ_7,σ) we are certain that one of the sets of values estimated from our ATR sites is correct, but prior to collecting data we are equally uncertain as to which.

Because the logarithms of the traffic counts are normal random variables, the likelihood function of the sample is easy to specify. Letting **y** denote the vector containing the logarithms of the sample counts and **V** denote the correlation matrix for the count sample (which can be computed from once one knows the value of the AR parameters ϕ_1 and ϕ_7), then if we knew the site-specific values for the parameters u, β, σ, ϕ_1, and ϕ_7, the likelihood of the sample could be computed using the appropriate multivariate normal density

$$f(y|\beta,u,\sigma,\phi_1,\phi_7)$$
$$(2\pi)^{-n/2}\sigma^{-n}|V|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}(y-X\beta_k-u\mathbf{1}_n)'V^{-1}(y-X\beta_k-u\mathbf{1}_n)\right) \tag{2.10}$$

Here **X** is a matrix, of dimension Nx19, each row having elements equal to 0 or 1 according to the month and day-of-week of the corresponding sample count, while $\mathbf{1}_n$ is an n-dimensional column vector with each element equal to 1.0.

Applying Bayes Theorem to the prior and likelihood to obtain the posterior distribution for the parameters, substituting this into (2.8) and performing the indicated integrations produces, after some tedious algebra

$$P[z^N \le \tilde{z}|z] = P[y^N \le \tilde{y}|z]$$
$$\approx \frac{\sum_{p=1}^{M}\sigma_p^{-(n-1)}|V_p|^{-1/2}(\mathbf{1}_n'V_p\mathbf{1}_n)^{-1/2}\sum_{k=1}^{m}\Phi\left(\frac{\tilde{y}-\hat{y}_{p,k}}{s_{p,k}}\right)\exp\left(-\frac{S_{pk}^2}{2\sigma_p^2}\right)}{\sum_{p=1}^{M}\sigma_p^{-(n-1)}|V_p|^{-1/2}(\mathbf{1}_n'V_p\mathbf{1}_n)^{-1/2}\sum_{k=1}^{m}\exp\left(-\frac{S_{pk}^2}{2\sigma_p^2}\right)} \tag{2.11}$$

where $\mathbf{V}_p$ denotes the sample correlation matrix computed using (ϕ_{1p}, ϕ_{7p}), and

$$
\begin{aligned}
S_{pk}^2 &= (y - X\beta_k - \bar{y}_{pk}\mathbf{1}_n)' V_p^{-1} (y - X\beta_k - \bar{y}_{pk}\mathbf{1}_n) \\
\bar{y}_{pk} &= \frac{\mathbf{1}_n' V_p^{-1} (y - X\beta_k)}{\mathbf{1}_n' V_p^{-1} \mathbf{1}_n} \\
\hat{y}_{pk} &= \bar{y}_{pk} + \frac{\sigma_p^2}{2} + \log_e\left(\sum_{t=1}^{N} e^{x_t \beta_k}\right) \\
s_{pk} &= \left(\frac{\sigma_p^2}{\mathbf{1}_n' V_p \mathbf{1}_n} + \frac{v_{N,kp}^2}{(\mu_{N,kp})^2} \right)^{-1/2}
\end{aligned}
$$

$v^2_{N,kp}$ and $\mu_{N,kp}$ are as defined in (2.4) and (2.5) but evaluated using β_k ϕ_{1p} ϕ_{7p} and σ_p. The distribution given in (2.11) is in essence a finite mixture of normal distributions, where the weights given to the mixture components are the posterior probabilities that the sampled site has adjustment factors and covariance parameters characteristic of each the m factor groups and each of the M ATR sites.

EMPIRICAL EVALUATION OF THE APPROXIMATION

As noted above, the distribution (2.11) approximates the predictive distribution of a total traffic count, the approximation being appropriate when predicting the total of a large number of days (e.g. a year or more) from a small sample (e.g. 2 weeks or less). Before proceeding though it may be helpful to check the accuracy of the approximation using actual traffic data. In an earlier study (Davis, 1997a), daily traffic counts from the year 1992 from 50 ATRs in outstate Minnesota were used to estimate monthly and day-of-week adjustment terms for three factor groups, and covariance parameters for each of the 50 ATRs. These estimates were used to construct the discrete prior distributions for β, and (ϕ_1,ϕ_7,σ), giving m=3 and M=50. In addition, daily counts from the year 1991 were available for 48 ATRS, and for each of these ATRs a sample consisting of a 1-week count from the month of March and a 1-week count from the month of July was drawn. A MATLAB (Mathworks, 1992) program for evaluating (2.11) was written, and then for each of the 48 ATRs, the 5th and 95th percentile points of the predictive distribution of the logarithm of the 1991 total traffic volume was computed by embedding the MATLAB routine inside an equation solver. Finally, the logarithm of the total 1991 traffic volume was also computed for each ATR, and the results of these computations are displayed in Tables 2.1-2.3.

The 5th and 95th percentile points describe the bounds of a 90% confidence interval, and if the approximation (2.11) is acceptably accurate, we would expect the actual count to fall inside the bounds 90% of the time. Inspection of Tables 2.1-2.3 shows that for ATRs 2, 8, 12, 204, 208, 217, 218, and 226 the actual count fell outside out estimated bounds. Since the probability of obtaining eight or more successes in 48 binomial trials with success probability equal to 0.1 is 0.102, this result is consistent with the hypothesis that (2.11) gives a reasonably accurate approximation of the predictive distribution of the total traffic.

Table 2.1. Evaluation of 90% Prediction Intervals: Factor Group 1.

Mn/DOT ATR Number	5%-ile of Predictive Distribution	Log Total Count	95%-ile of Predictive Distribution
2	**12.2954**	**12.5539**	**12.5531**
3	13.9783	14.1189	14.2267
7	12.2483	12.3857	12.4690
8	**10.9598**	**11.3207**	**11.2912**
9	11.8849	11.9756	12.1124
10	12.8934	13.0059	13.0748
12	**13.1947**	**13.1813**	**13.4464**
14	12.9814	13.0870	13.2310
50	13.4775	13.5663	13.6500
54	10.2983	10.5153	10.5339
56	11.4603	11.5531	11.6512
100	15.2954	15.4013	15.5066
102	16.4250	16.4932	16.6343
103	16.0997	16.1653	16.2748
104	15.2014	15.2137	15.2945
110	15.6586	15.7102	15.8464
164	13.8722	13.9628	14.0880
166	13.8270	13.8729	13.9741
170	13.8047	13.8419	13.9555
172	14.9882	15.1130	15.1175
179	13.0413	13.1812	13.2469
197	13.7666	13.8906	13.9826
199	12.4498	12.6153	12.6332
211	13.4673	13.5737	13.6714

Table 2.1 (Continued)

213	14.0399	14.0975	14.1949
216	12.4219	12.5007	12.6338
217	**11.9371**	**12.1674**	**12.1390**
219	13.2525	13.2984	13.4746
225	12.0633	12.2808	12.3005
226	**11.5859**	**11.7921**	**11.7918**

Table 2.2. Evaluation of 90% Prediction Intervals: Factor Group 3.

Mn/DOT ATR Number	5%-ile of Predictive Distribution	Log Total Count	95%-ile of Predictive Distribution
52	11.4197	11.5725	11.6027
175	15.2607	15.3471	15.4842
187	14.9642	15.0832	15.2037
200	15.7115	15.8787	15.9226
204	**14.0265**	**14.2724**	**14.2705**
207	12.8435	12.9302	13.0395
208	**14.8488**	**14.9436**	**14.9377**
215	11.7087	11.8455	11.9711
218	**11.5948**	**11.9728**	**11.8361**
224	13.1007	13.2427	13.3437

Table 2.3. Evaluation of 90% Prediction Intervals: Factor Group 4.

Mn/DOT ATR Number	5%-ile of Predictive Distribution	Log Total Count	95%-ile of Predictive Distribution
1	12.9936	13.0798	13.2428
51	11.4256	11.5018	11.6581
55	12.0286	12.1631	12.1852
57	12.4938	12.7573	12.7840
214	11.5181	11.6925	11.7181
220	13.2472	13.3545	13.4526
221	13.4560	13.4819	13.6290
223	12.8283	12.9375	12.9596

CHAPTER 3

ACCIDENT RATE ESTIMATION WITH TRAFFIC SAMPLES

We return finally to our original problem, estimating a site's accident rate λ when the data at hand consist of an accident count x, and a sample of daily traffic counts **z**=[z$_1$,..,z$_n$]′. This chapter has three main objectives. First, given our argument for the importance of the Bayesian approach in accident rate estimation, we would like to have a computationally feasible method for computing Bayesian estimates of a site's accident rate given an accident count and a traffic count sample. Second, we would like to gain some idea of how uncertainty concerning the accident rate is affected by the size of the traffic count sample, so that informed decisions concerning sample size can be made. Third, given that the Bayesian point estimate of the accident rate is optimal, in the sense of minimizing posterior mean-squared error, we would like to gain some idea of the increase in error which results when more traditional accident rate estimates are used.

As indicated at the end of Chapter 1, the posterior distribution of the accident rate, given the accident count and traffic sample, can be expressed formally as

$$f(\lambda|x,z) = \int_0^\infty f(\lambda|x,z_T)f(z_T|x,z)dz \tag{3.1}$$

There are two probability densities appearing inside the integral. The first can in principle be computed via Bayes Theorem once we have specified the likelihood function for the traffic count and a prior density for λ

$$f(\lambda|x,z_T) = \frac{p(x|\lambda,z_T)f(\lambda)}{\int_0^\infty p(x|\lambda,z_T)f(\lambda)d\lambda} \tag{3.2}$$

Next, if we appoximate the second density f(z$_T$|x,**z**) with a density f(z$_T$|**z**) corresponding to thte distribution function developed in Chapter 2, these expressions could be inserted into (3.1) and we would have our result. In practice though it was found that the computational requirements for evaluating the approximation $f(z_T|\mathbf{z})$ were so burdensome that embedding the approximation inside a multidimensional numerical inegration routine did not appear likely to yield a computationally feasible approach.

The numerical problems arising when attempting to evaluate (3.1) are typical for complex models, especially those containing hidden variables such as our total traffic volume. Similar numerical constraints have limited the practical applicability of Bayesian methods until the early 1990s, when researchers began to show that a set of numerical techniques known as Markov Chain Monte Carlo (MCMC) methods could be profitably applied to a number of hitherto intractable problems (Carlin and Louis, 1996). Standard Monte Carlo (MC) methods have been used for some time to estimate characterisics of a target probability distribution to any required degree of accuracy by simulating, on a computer, a random sample of outcomes from that distribution, and then simply computing the desired estimate by appropriately summing over the elements of the sample. The standard MC methods are limited though to problems with fairly simple structures. In an MCMC method a type of stochastic process called an ergodic Markov chain is constructed which has the target distribution as its stationary distribution, and then a computer is used to generate a simulated realization from the Markov chain. One of the characteristics of an ergodic Markov chain is that after a suitable initialization period, averages computed over a single run of the chain converge to the same limits as averages computed over a random sample from its stationary distribution, so that the output of one (or a small number) of long runs can be used to compute Monte Carlo estimates. The MCMC approach is even more attractive because a software package called BUGS (Speigelhalter, et al. 1995), which implements an MCMC technique called Gibbs sampling, is available for experimental use. So in what follows, BUGS will be used to compute Bayesian estimates of accident rates given traffic count samples.

Our second objective is to gain some idea of how sensitive the Bayesian estimate of accident rate is to traffic sample size. This will be done by comparing estimates based on a "full" traffic count to those based on 2-day and 14-day traffic samples. As in Chapter 2, the daily traffic count data for 50 outstate Minnesota ATRs were used, and these were supplemented by computerized files containing the 1992 accident records for all accidents occurring within about 2.5 miles of each of

Mn/DOT's ATRs. From these accident record files daily accident counts for each day of 1992 were compiled for each of the 50 outstate ATRs. It turned out that the majority of the ATRs had yearly accident counts of zero or one, but a total of 17 ATRs had five or more accidents occurring in their vicinities during 1992, and these were used in our analyses. The locations of these ATRs are shown in Figure 3.1.

To meet the second objective, it was first necessary to compute estimates of accident rates as if we had a "full" traffic count. To accomplish this, for each of the 17 study ATRs those days which Mn/DOT analysts had flagged as containing questionable traffic counts were deleted from the data set, and then total traffic and accident counts were computed for each ATR by summing over the remaining daily traffic and accident counts. In some cases the corresponding accident count was less than the total for that site because one or more accidents occurred on days with bad traffic data. Next, as noted in Chapter 1, computing Bayesian estimates of the accident rate requires first specifying a probability density which expresses the analyst's prior (before data collection) uncertainty concering a site's accident rate, as well as a likelihood function for the accident counts. The Poisson likelihood depicted in (1.10) was used, and the computations were carried out using two different prior densities. The first was the gamma density described in (1.16), with parameters α=0.0005 and β=0.0005, and the second was a lognormal density for which the underlying normal mean was 0.0, and the underlying normal variance was 1,000,000. These priors were chosen because they were both fairly flat (uninformative) over the range from 10^{-6} to 10^{-5}, where most actual accident rates tend to fall. The principle of stable estimation (Edwards, Lindman and Savage, 1962) then implies that the posterior distribution will be dominated by the likelihood function, so that the prior should have little influence on the final result.

As indicated in Chapter 1, closed form expressions are available for the posterior probability density, the posterior expected value and the posterior variance when the gamma prior is used. No such convenient expressions are known for the lognormal prior, and numerical integration must be employed. For this analysis, MATHCAD 6.0+ (Mathsoft, 1995) was used to compute posterior expected values, variances and the 2.5%-ile and 97.5%-ile points for the posterior distributions, for each of the 17 ATR sites. These latter two quantities give lower and upper bounds for a 95% Bayesian confidence interval for the accident rate. Tables 3.1 and 3.2 display the results of these computations for the 17 study sites, along with the total accident count, total traffic count and the number of days during 1992 for which the ATR had good count data. Table 3.3 displays parallel results computed using frequentist procedures.

Comparing Table 3.1 to Tables 3.2, we can see that the gamma and lognormal priors produced nearly identical results, and looking at Table 3.3 we can see that the results of the Bayesian analyses are numerically similar to those resulting from a frequentist analysis. The chief difference between the Bayesian and frequentist results is that the frequentist confidence intervals are symmetric about the point estimates of the accident rate while the Bayesian intervals are skewed. This is because the normal approximation (1.15) was used to compute the frequentist interval while the Bayesian intervals were computed directly from posterior densities. Plots of the likelihood functions for each of the 17 sites showed that in many cases they were skewed, indicating that the expected accident counts were smaller than those needed to justify the normal approximation.

Next, two traffic count samples were taken for each of the study sites. The first sample consisted of daily traffic counts for two consecutive weeks during the month of July, 1992 while the second sample consisted of two consecutive days during the month of July. July was selected as the sample month because (a) an earlier study (Davis, 1997a) indicated that counts from this month tended to be more informative concerning the monthly and day-of-week factors appropriate for a given site, and (b) this month fell in the middle of the normal traffic counting season in Minnesota. It turned out that the distribution of bad traffic count days made it difficult to find a full two weeks for all of the ATR sites, so that some of the so-called "two-week" samples had less than 14 days worth of traffic counts. For the two-day samples, the selected days were always on Tuesdays and Wednesdays .

The computer program BUGS was used to compute posterior means, standard deviations, 2.5%-ile, and 97.5%-ile points for each traffic count sample at each site. It was discovered that BUGS ran more reliably if the accident count was broken down into two subcounts, one of accidents occuring on traffic sample days, and one of accidents occuring on non-sample days. In addition, having BUGS evaluate the variance term (2.5) turned out to be infeasible, so an approximation was used where it is assumed that the autocorrelations are identically equal to 1.0 for all lags. This produces an upper bound for (2.5), and when this bound is used we get the substantial simplification

$$\frac{\nu_N^2}{\mu_N^2} = e^{\sigma^2}-1 \tag{3.3}$$

The distribution of the logarithm of the total traffic count can then be approximated by a normal distribution with mean equal to $\log_e(\mu_N)$ and variance $\exp(\sigma^2)$-1. Listings of example model specifications and output can be found in the Appendix, but an informal description of the statistical model specified for BUGS is as follows:

(1) given an accident rate λ and a traffic count z_T, accident counts were assumed to be Poisson outcomes with mean λz_T;

(2) given the mean value parameter u, the adjustment terms β and standard deviation σ, the logarithm of the total traffic counts was assumed to be normal with mean and variance described in (3.3) above;

(3) given the mean value parameter u, adjustment terms β, and covariance matrix $\sigma^2\mathbf{V}$, the logarithms of the sample counts were assumed to be jointly normal as specified in equation (2.10);

(4) the prior for $\log_e(\lambda)$ was assumed to be normal with mean 0 and variance 1,000,000;

(5) the prior for β was assumed to be a discrete uniform with three components corresponding to the three 1992 factor groups described in Davis (1997a);

(6) the prior for u was taken to be normal with mean equal to 0 and variance equal to 1000;

(7) for the "two-week" sample, the prior for the covariance matrix $\sigma^2\mathbf{V}$ was assumed to be Wishart with expected value being a diagonal matrix with diagonal elements equal to the average variance estimated for the 50 ATR sites;

(8) given σ,ϕ_1, and ϕ_7, the covariance matrix for the two-day sample was 2x2 with diagonal elements equal to σ^2 and off-diagonal elements $\sigma^2\phi_1$.

All BUGS runs consisted of a 2000 iteration burn-in, followed a a 5000 iteration sample. Posterior means, standard deviations and bounds for the 95% Bayesian confidence intervals are shown in Table 3.4.

Four of the ATRs, 52, 208, 172 and 197, had fairly small "two-week" samples, due to the prevalence of bad traffic count days during July, and the BUGS software crashed when we attempted to compute Bayesian accident rate estimates for these sites. This problem did not occur for the two-day samples, which used a different prior distribution for the sample covariance matrix, nor did it occur when we used a lognormal prior for λ with a variance of 1000, instead of 1,000,000.

Comparing the posterior standard deviations for the two sample sizes we see that, as expected, estimates of λ based on the two-day sample showed greater uncertainty than do estimates based on the "two-week" sample. This greater uncertainty is also reflected by the wider confidence intervals for the two-day sample. Interestingly, comparing the variances and confidence intervals for the "two-week" sample to the parallel quantities in Table 3.2 indicates that little additional gain in precision is obtained when moving from the "two-week" sample to the "full-count". The estimates in the Tables 3.2 and 3.4 are not strictly comparable since those in 3.2 assume a time period reflecting only "good" traffic count days, while those in 3.4 assume a time-period of 366 days. Nevertheless, the posterior standard deviations and confidence interval widths indicate that at least for estimating accident rates at these sites, a traffic count sample of about two-weeks duration is sufficient to achieve accuracy similar to that of much longer count samples.

We now turn to our third objective, which is to assess the additional error arising when using more traditional estimates of accident rates. The measure of error we will use is the posterior root-mean squared error (RMSE), defined as

$$RMSE = \sqrt{\sigma_{post}^2 + (\bar{\lambda} - \tilde{\lambda})^2} \qquad (3.4)$$

where

σ_{post} = posterior standard deviation of the accident rate λ,

$\bar{\lambda}$ = posterior mean of the accident rate,

$\tilde{\lambda}$ = an estimate of the accident rate λ.

Informally, posterior RMSE can be seen as the expected distance between an estimate of λ and its true value, and clearly this is minimized when the estimate is taken to be the posterior mean.

Traditional accident rate estimates for 1992 take the form

$$\tilde{\lambda} = \frac{x}{366 \times AADT} \tag{3.5}$$

where AADT means annual average daily traffic. In practice, AADT is often estimated from a 48-hour short count which is then adjusted to correct for seasonal and day-or-week biases. To emulate this practice, we computed two accident rate estimates using the two-day samples described above, both estimates having the form shown in (3.5). The first rate estimate, called the factored estimate, used an estimate of AADT computed by first adjusting each of the two daily counts using the monthly and day-of-week factor computed for that ATR's factor group, and then averaging these adjusted counts. This is the estimate of AADT one would obtain if one knew exactly what adjustment factors were appropriate for the site of interest. The second rate estimate, called the unfactored estimate, used an estimate of AADT computed by simply averaging the two daily counts in the sample without adjusting for month or day-of-week. These estimates, together with two-day sample posterior means and standard deviaitions were then substituted into (3.4) to yield RMSE values, and the results of these computations are shown in Table 3.5.

From Table 3.5 we can see that the dominant component of expected error is the posterior variance, which all rate estimates are subject to. In no case did the percentage increase in expected error incurred when one uses a traditional accident rate estimate exceed 20%, with the unfactored estimates generally showing a greater increase in expected error than did the factored estimates. This suggests that, to the extent that this sample of roadway sites is representative of other sites, the posterior mean is to be preferred as a point estimate of accident rate , but that simpler estimates based on traditional esitmates of AADT, do not appear to incur substantial increases in error.

Figure 3.1. Outstate ATRs with 5 or More Accidents in 1992

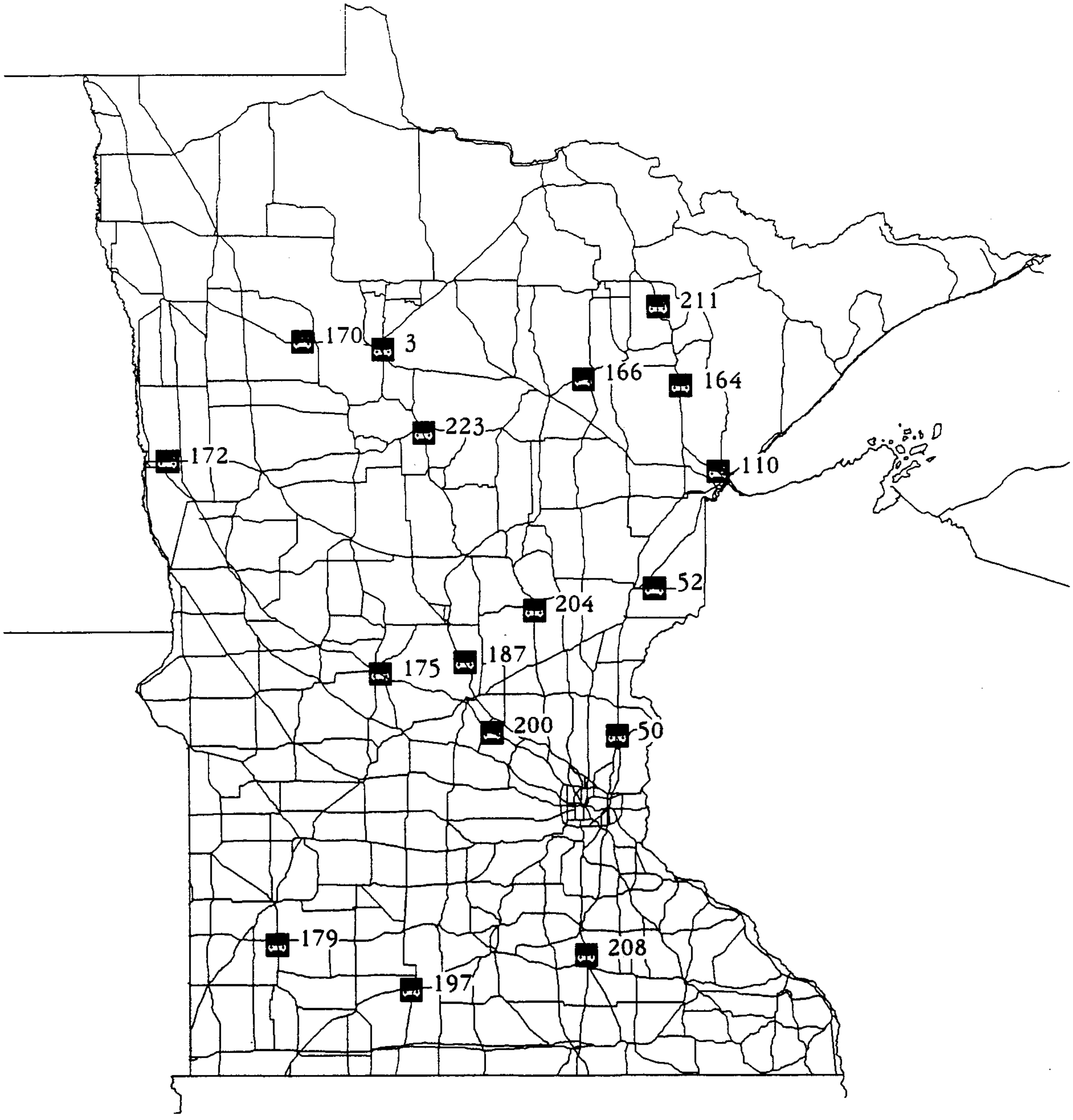


Table 3.1 Bayes Estimation Using "Full-Year" Counts. Gamma Prior for λ.

ATR#	Accident Count (x)	Total Traffic (z)	"Good" Data Days	Posterior Mean $\times 10^6$	Posterior St. Dev.$\times 10^6$	$\lambda_{.025}$ $\times 10^6$	$\lambda_{.975}$ $\times 10^6$
175	11	3633983	238	3.027	0.913	1.511	5.061
52	6	129062	276	46.493	18.980	17.063	90.414
187	2	2031318	181	0.985	0.696	0.119	2.743
200	9	6816272	262	1.320	0.440	0.604	2.313
204	10	1999478	264	5.002	1.582	2.399	8.545
208	21	3781054	205	5.554	1.212	3.438	8.169
223	6	648747	207	9.249	3.776	3.395	17.986
3	5	1357989	306	3.682	1.647	1.196	7.542
50	6	690287	257	8.693	3.549	3.190	16.904
110	38	6851628	336	5.546	0.900	3.925	7.443
164	13	2479466	335	5.243	1.454	2.792	8.454
166	5	1165843	275	4.289	1.918	1.393	8.785
170	6	1091136	355	5.499	2.245	2.018	10.694
172	13	2008653	191	6.472	1.795	3.446	10.436
179	4	552554	354	7.240	3.620	1.973	15.867
197	5	1092392	317	4.578	2.047	1.487	9.376
211	5	1084894	336	4.609	2.061	1.497	9.440

Table 3.2. Bayes Estimation Using “Full-Year” Counts. Lognormal Prior for λ.

ATR#	Accident Count (x)	Total Traffic (z)	“Good” Data Days	Posterior Mean ×10^6	Posterior St. Dev.×10^6	$\lambda_{.025}$ ×10^6	$\lambda_{.975}$ ×10^6
175	11	3633983	238	3.027	0.913	1.511	5.061
52	6	129062	276	46.489	18.980	17.061	90.409
187	2	2031318	181	0.985	0.695	0.119	2.739
200	9	6816272	262	1.318	0.432	0.605	2.482
204	10	1999478	264	5.005	1.558	2.399	8.540
208	21	3781054	205	5.553	1.208	3.439	8.196
223	6	648747	207	9.244	3.778	3.395	18.042
3	5	1357989	306	3.684	1.655	1.196	7.552
50	6	690287	257	8.699	3.536	3.191	17.056
110	38	6851628	336	5.536	1.122	3.926	7.511
164	13	2479466	335	5.230	1.467	2.792	8.453
166	5	1165843	275	4.289	1.916	1.393	8.800
170	6	1091136	355	5.496	2.240	2.019	10.782
172	13	2008653	191	6.686	1.549	3.450	10.806
179	4	552554	354	7.240	3.612	1.972	15.843
197	5	1092392	317	4.577	2.046	1.486	9.391
211	5	1084894	336	4.608	2.060	1.497	9.455

Table 3.3. Maximum Likelihood Estimates, Standard Errors and 95% Confidence Intervals

ATR#	Accident Count (x)	Total Traffic (z)	“Good” Data Days	MLE of Acc. Rate $\times 10^6$	Standard Error$\times 10^6$	$\lambda_{.025}$ $\times 10^6$	$\lambda_{.975}$ $\times 10^6$
175	11	3633983	238	3.027	0.913	1.239	4.816
52	6	129062	276	46.649	18.979	9.291	83.688
187	2	2031318	181	0.985	0.696	-0.380	2.349
200	9	6816272	262	1.320	0.440	0.458	2.183
204	10	1999478	264	5.001	1.582	1.901	8.101
208	21	3781054	205	5.554	1.212	3.179	7.930
223	6	648747	207	9.249	3.776	1.848	16.649
3	5	1357989	306	3.682	1.647	0.455	6.909
50	6	690287	257	8.692	3.549	1.737	15.647
110	38	6851628	336	5.546	0.900	3.783	7.310
164	13	2479466	335	5.243	1.454	2.393	8.093
166	5	1165843	275	4.289	1.918	0.295	8.048
170	6	1091136	355	5.499	2.245	1.099	9.899
172	13	2008653	191	6.472	1.795	2.954	9.990
179	4	552554	354	7.239	3.620	0.145	14.333
197	5	1092392	317	4.577	2.047	0.565	8.589
211	5	1084894	336	4.609	2.061	0.569	8.648

Table 3.4. Performance of Bayesian Estimates of Accident Rates Based on Two-Week and Two-Day Traffic Count Sample (All entries have been multiplied by 10^6).

ATR #	"Two-Week" Sample				Two Day Sample			
	Post. Mean	Post. St. Dev	$\lambda_{.025}$	$\lambda_{.975}$	Post.Mean	Post. St. Dev	$\lambda_{.025}$	$\lambda_{.975}$
175	3.543	0.991	1.876	5.731	4.429	1.433	2.260	7.847
52	---	----	----	----	49.16	24.49	15.69	106.2
187	1.533	0.629	0.575	3.012	1.900	0.897	0.653	4.081
200	0.921	0.345	0.389	1.720	1.218	0.468	0.452	2.259
204	5.571	1.728	2.718	9.377	7.333	2.563	3.484	13.38
208	---	---	---	---	5.059	1.593	2.613	8.771
223	7.066	2.785	2.861	13.36	6.828	3.097	2.440	14.12
3	3.126	1.524	0.943	6.754	3.538	1.833	0.984	9.083
50	6.451	2.926	2.213	13.31	5.868	1.628	3.311	9.673
110	5.390	1.224	3.214	8.075	6.816	3.283	2.147	14.88
164	5.395	1.634	2.680	9.181	5.993	2.177	2.663	11.05
166	3.306	1.605	0.974	7.197	3.512	1.803	0.974	7.862
170	5.186	2.351	1.781	10.82	5.523	2.642	1.823	11.87
172	---	---	---	---	4.439	1.430	2.155	7.675
179	9.170	4.388	2.746	19.67	10.03	5.316	2.910	23.76
197	---	---	---	---	7.448	3.239	2.790	15.14
211	4.418	2.100	1.401	9.355	4.647	2.363	1.331	10.41

Table 3.5. Comparison of Bayesian and Traditional Accident Rate Estimates Using Two-Day Traffic Count Samples. (All entries have been multiplied by 10^6).

ATR #	Posterior Mean	Factored Estimate	Unfactored Estimate	Posterior S.D.	Factored RMSE	Unfactored RMSE
175	4.429	3.948	3.571	1.433	1.512	1.67
52	49.16	43.622	39.455	24.49	25.109	26.345
187	1.900	1.663	1.504	0.897	0.928	0.981
200	1.218	1.006	0.910	0.468	0.514	0.56
204	7.333	6.390	5.780	2.563	2.731	2.997
208	5.059	4.429	4.007	1.593	1.713	1.909
223	6.828	8.398	5.391	3.097	3.472	3.414
3	3.538	3.352	2.800	1.833	1.842	1.976
50	5.868	6.508	5.437	1.628	1.749	1.684
110	6.816	5.653	4.723	3.283	3.483	3.893
164	5.993	5.819	4.862	2.177	2.184	2.453
166	3.512	3.334	2.785	1.803	1.812	1.944
170	5.523	5.265	4.399	2.642	2.655	2.871
172	4.439	4.558	3.911	1.430	1.435	1.524
179	10.03	9.865	8.242	5.316	5.319	5.609
197	7.448	7.108	5.939	3.239	3.257	3.573
211	4.647	4.439	3.709	2.363	2.372	2.542

CHAPTER 4
SUMMARY AND CONCLUSIONS

This report began by pointing out that estimated accident rates play crucial roles in traffic safety programming, and by arguing that a Bayesian approach to statistical inference better supports the type of decisions that safety engineers need to make. It was noted that conventional statistical approaches to estimating accident rates make an assumption that the total traffic volume (or vehicle-miles of travel) at a location is known with certainty, but that in almost all practical instances traffic totals are estimated from short traffic count samples. This means that to some degree the total traffic is uncertain, and that this uncertainty will increase the uncertainty associated with an estimated accident rate. In principle this could in turn lead to falsely identifying a site as a potential high-hazard location, or overestimating the effect of a safety countermeasure in a before-and-after study. Three main objectives were specified for this study:

(1) To develop a method for estimating accident rates that accounted for uncertainty which arises when using estimated traffic totals,

(2) To assess the effect of the size of traffic count sample on the precision of an accident rate estimate, and

(3) To assess the likely increase in error arising when one uses traditional accident rate estimates.

Objective (1) was achieved by first expanding the work in Davis (1997a) and developing a Bayesian approach to estimating traffic volume totals from short-count samples. The resulting relationships between the accident count, the accident rate, the traffic total and the traffic sample were somewhat complicated, but it turned out that a relatively new computational technique called Gibbs sampling could be used to perform the necessary computations, with a suprisingly small amount of additional analytic work. Objective (2)was then achieved by comparing the uncertainty resulting when attempting to estimate accident rates using three different traffic count samples: (i) close to a full year's traffic count data, (ii) approximtely two weeks of traffic count data and (iii) two days of traffic count data. Objective (3) was achieved by comparing the increase in expected error arising when using traditional accident rates estimates based on the two-day traffic count samples.

Table 4.1 displays the expected errors arising when using Bayesian estimates for the three traffic count samples, and when using a factored estimate of AADT to estimate the annual traffic total, to estimate 1992 accident rates at 17 Mn/DOT outstate ATR sites. Because these expected errors are conditional on the actual data they do not tell exactly the same story in all cases, but generally one can see that (1) the error arising when using the two-week count is comparable to the error arising when using the "full-year" count, (2) error increases by 20%-50% when going from the two-week count to the two-day count, and (3) the additional error incurred when using the traditional rate estimate instead of the two-day Bayesian estimate is not substantial.

Our first conclusion is based on what is arguably our most interesting finding, that the precision achieved by full-counting is also achieved by the two-week sample. This means that a safety programming decisions based on well-chosen two-week traffic samples should be as accurate as decisions made using full traffic counts, so that safety programs can be based on traffic sampling without suffering a loss in precision. Two-week count samples are not all that common however, two-day weekday counts being much more typical. If we take the factored AADT estimates based on two-day traffic samples as being representative of current practice, then our second conclusion is that there does not seem to be a substantial difference between the point estimates produced by current practice and the point estimates one would get using a Bayesian procedure on the same data. However, properly interpreting the importance of a point estimate requires some sense of the estimate's precision, and it is here that weaknesses in current practice appear. First, the estimated variance formula (1.3) would normally be used to assess the precision of an estimated accident rate, even if the total traffic count was also estimated using a two-day sample. In practice, this means that the safety engineer would act as if the entries in the "Full-Count" column in Table 4.1 gave the correct assessment of estimator precision when the actual precision is given by the entries in the "2-Day Sample" and the "Factored AADT" columns of Table 4.1. Thus current practice tends to overstate the precision of accident rate estimates based on two-day traffic samples. Second, as illustrated in Table 3.3, current practice treats the distribution of an estimated accident rate as being approximately normal, so that confidence intervals will be symmetric about the point estimate. This normal approximation is commonly used in identifying potential high-hazard locations via the rate quality-control method. However, at least for these data, the likelihood functions tend to be skewed, rather than symmetric, and this is reflected in the shapes of the posterior densities. This means that the bounds defining a confidence interval will be different than those based on the normal approximation, as one can see by comparing Tables 3.1 and 3.2 to Table 3.3. Thus our third conclusion is that for traffic counting programs based on two-day samples, traditional accident rate estimation methods, which implictly assume full counting and normally distributed accident rate estimates, overstate the precision that can actually be provided and this overstatement of precision can possibly lead to inefficienies in safety decision making.

These conclusions suggest two recommendations for improving current safety practice. First, for analyses involving detailed investigation of individual sites, such as when one is attempting to identify countermeasures, or attempting to assess the effect of a countermeasure in a before-and-after study, we recommend that an extended traffic count sample of at least two weeks duration be included as one of the traffic engineering studies done at the site. For before-and-after studies this count should taken in both the before and the after periods. This should reduce the error in estimating accident rates to about what one could expect from full counting. We also note that Davis (1997a) found that at least when considering Mn/DOT's outstate factor groups, a count sample consisting of a one-week count in March and a one-week count in July proved as effective in identifying a site's factor group as did an "optimal" sample design. This suggests that when possible, making two one-week counts in different months could provide some additional precision when estimating total traffic at a site.

Second, we recognize that most agencies do not have the resouces to conduct two-week counts on a network-wide, routine basis. Assuming that the two-day count remains a mainstay of traffic counting programs, we recommend that current safety programming methods be enhanced so as to correctly account for the uncertainty in estimated accident rates. This should be seen as a long-term rather than short term recommendation however, since although promising statistical methods do exist, as demonstrated by the work reported in Chapter 3, they are not available in user friendly implementations, and their correct use requires statistical knowledge beyond that normally taught to traffic engineers. In particular, as the authors of the BUGS software note, "Gibbs sampling can be dangerous!" and at present it is not possible to give a general set of rules for correct application of MCMC methods. Rather each application must be judged on a case-by-case basis. One alternative to MCMC methods would be to perform numerical integration in (3.1). As computers become faster, this approach may become feasible for network wide analyses.

In conclusion, the field of statistics has undergone something of a computational revolution during the past 10 years, and the long-term dependence on simplistic by computationally tractable statistical models has been lessened. One such simplification has been to treat the problem of estimating traffic totals as separate and independent of the problem of estimating accident rates. In principle, it is now possible to treat these problems in a unified manner, and the question of interest is how to make what is possible in principle a practical reality.

Table 4.1 Estimation Error Summary from Different Traffic Samples.

Mn/DOT ATR #	"Full-Count" Stan. Dev.	"2-Week" Sample Stan. Dev.	2-Day Sample Stan. Dev.	Factored AADT RMSE
175	0.913	0.991	1.433	1.512
52	18.98	----	24.49	25.109
187	0.695	0.629	0.897	0.928
200	0.432	0.345	0.468	0.514
204	1.558	1.728	2.563	2.731
208	1.208	----	1.593	1.713
223	3.778	2.785	3.097	3.472
3	1.655	1.542	1.833	1.842
50	3.536	2.926	1.628	1.749
110	1.122	1.224	3.283	3.483
164	1.467	1.634	2.177	2.184
166	1.916	1.605	1.803	1.812
170	2.240	2.351	2.642	2.655
172	1.549	----	1.430	1.435
179	3.612	4.388	5.316	5.319
197	2.046	----	3.239	3.257
211	2.060	2.100	2.363	2.372

REFERENCES

Ang, A., and Tang, W. (1975) *Probability Concepts in Engineering Planning and Design, Volume 1-Basic Principles*, New York, Wiley and Sons.

Box, P., and Oppenlander, J. (1976) *Manual of Traffic Engineering Studies,* Fourth Edition, Washington, DC, ITE.

Carlin, B., and Louis, T. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*, London, Chapman and Hall.

Brockwell, P., and Davis, R. (1991) *Time Series: Theory and Methods*, New York, Springer-Verlag.

Davis, G. (1997a) *Estimation Theory Approach to Monitoring and Updating Average Daily Traffic*, Report 97-05, Minnesota Dept. of Transportation, St. Paul, MN.

Davis, G. (1997b) Accuracy of estimates of mean daily traffic: A review, *Transportation Research Record*, 1593, 12-16.

Davis, G. and Guan, Y, (1996) Bayesian assignment of coverage count locations to factor groups and estimation of mean daily traffic, *Transportation Research Record*, 1542,30-37.

Devore, J. (1995) Probability and Statistics for Engineering and the Sciences, Fourth Edition, Pacific Grove, CA, Brooks/Cole Publishing.

Edwards, W., Lindman, H., and Savage, L. (1963) Bayesian statistical inference for psychological research, *Psychological Review*, 70, 193.

Gallant, R., and White, W. (1988) *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, New York, Basil-Blackwell.

Gerlough, D., and Huber, M. (1975) *Statistics with Applications to Highway Traffic Analysis,* Second Edition, Westport, CT, Eno Foundation.

Greenshields, B., and Weida, F. (1952) *Statistics with Applications to Highway Traffic Analysis*, Saugatuck, CT, Eno Foundation.

Hahn, G., and Shapiro, S. (1967) *Statistical Models in Engineering*, New York, Wiley and Sons.

Hauer, E. (1997) *Observational Before-After Studies in Road Safety*, New York, Elsevier Science.

Homburger, W., and Kell, J. (1988) *Fundamentals of Traffic Engineering*, Twelfth Edition, Berkeley, CA, Institute of Transportation Studies, University of California.

Johnson, N., Kotz, S., and Balakrishnan, N. (1994) *Continuous Univariate Distributions, Volume 1*, Second Edition, New York, Wiley and Sons.

King, L.E. (1994) Statistical analysis, in Robertson, H., et al. (Ed) *Manual of Transportation Engineering Studies*, Englewood Cliffs, NJ, Prentice-Hall.

Marlow, N. (1967) A normal limit theorem for power sums of independent random variables, *Bell Systems Technical Journal*, 46, 2081-2089.

Mathsoft (1995) *Mathcad 6.0+ User's Guide*, Mathsoft, Inc.

Mathworks, (1992) *MATLAB Version 4*, Mathworks, Inc.

McShane, W., and Roess, R. (1990) *Traffic Engineering*, Englewood Cliffs, NJ, Prentice-Hall.

Meyer, P. (1970) *Introductory Probability and Statistical Applications*, (2nd edition), Reading, MA, Addison-Wesley.

Seber, G., and Wild, C. (1989) *Nonlinear Regression*, New York, Wiley and Sons.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1995) *BUGS 0.5 Reference Manual*, Cambridge, UK, MRC Biostatistics Unit, Cambridge University.

APPENDIX A
Asymptotic Approximation for Total Traffic Count

APPENDIX

ASYMPTOTIC APPROXIMATION FOR TOTAL TRAFFIC COUNT

As in Chaper 2, let z_t, t=1,.., N denote a sequence of lognormal random variables, and let $z^N=\sum_t z_t$ denote their sum. $\Phi(x)$ denotes the cumulative normal distribution function. As above, we will assume that the z_t follow the model

$$\log_e(z_t) = u + \boldsymbol{z_t}\beta + e_t$$

and the error terms $\{e_t\}$ follow a stationary p-order autoregressive (AR(p)) process. Marlow (1967) showed that if their existed sequences of positive real number $\{\mu_N\}$ and $\{v_N\}$ such that

$$(a) \quad \lim_{N\to\infty}\frac{v_N}{\mu_N} = 0$$

$$(b) \quad \lim_{N\to\infty}Prob\left[\frac{z^N-\mu_N}{v_N}\le x\right]=\Phi(x)$$

Then

$$\lim_{N\to\infty}Prob\left[\left(\frac{\mu_N}{v_N}\right)\log_e\left(\frac{z_N}{v_N}\right)\le x\right]=\Phi(x)$$

To verify the above conditions, we will impose the restriction that the monthly and day-of-week factors for any given day are bounded from above and also bounded away from zero. That is, there exist constants Δ_1 and Δ_2, such that

$$0 < \Delta_1 \le e^{x_t\beta_k} \le \Delta_2 < \infty$$

for all t and k. It is then possible to show that

$$v_N = \left(e^{\sigma^2}(e^{\sigma^2}-1)\sum_{t=1}^{N} e^{2x_t\beta_k} + (e^{2u+\sigma^2})(\sum_{t=1}^{N} e^{x_t\beta_k}(\textstyle\sum_{s\neq t} e^{x_s\beta_k}(e^{\rho_{t,s}\sigma^2}-1)))\right)^{1/2}$$

$$\leq \left((N)e^{2u+\sigma^2+2\Delta_2}(1+\gamma)\right)^{1/2} = O(N^{1/2})$$

where

$$\gamma = \sum_{k=1}^{\infty}(e^{\rho_k\sigma^2}-1) < \infty$$

whenever the ρ_k are the autocorrelations for a stationary AR(p) process. Similarly,

$$=N(e^{u+\frac{\sigma^2}{2}+\Delta_1}) \leq \mu_N = (e^{u+\frac{\sigma^2}{2}})\sum_{t=1}^{N} e^{x_t\beta} \leq N(e^{u+\frac{\sigma^2}{2}+\Delta_2})$$

so that

$$\frac{v_N}{\mu_N} = O(N^{-1/2})$$

and condition (a) is satisfied.

If the daily counts z_t were independent, we could use the either Lyaponuv or Lindbergh central limit theorems to verify condition (b), as done by Marlowe (1967). A more general central limit theorem, allowing for dependence of the sort generated by the AR(p) model for the errors e_t, is stated in Theorem 5.3 in Gallant and White (1988, p.76). In particular, if we can show that:

(1) $\|z_t - E[z_t]\|_4 \leq \Delta < \infty$, for all t,

(2) supremum$_t$ $(\|z_t - E[z_t | e_{t+m},..,e_t,..,e_{t-m}]\|_2) = O(1/m)$,

(3) the noise process $\{e_t\}$ is "α-mixing" of size -4;

(4) $v_N^{(-2)} = O(1/N)$,

then condition (b) will be satisfied, and we are done.

(1) Let ω=exp(σ^2), and since z_t is lognormal its fourth central moment is known, so that

$$\|z_t - E[z_t]\|_4 = (\omega^2(\omega-1)^2(\omega^2+2\omega^3+3\omega^2-3)e^{4(u+x_t\beta)})^{1/4}$$
$$\leq(\omega^2(\omega-1)^2(\omega^2+2\omega^3+3\omega^2-3)e^{4u}e^{4\Delta_2})^{1/4} = \Delta < \infty$$

(2) This condition is satisfied trivially, since

$$E[z_t | e_{t+m},..,e_t,..,e_{t-m}] = E[z_t | e_t] = z_t$$

implies

$$supremum_t \left(\|z_t - E[z_t | e_{t+m},..,e_t,..,e_{t-m}]\|_2\right) = 0$$

(3) This condition is also satisfied trivially since the fact that the noise process $\{e_t\}$ is a stationary AR(p) process implies that it is α-mixing of all orders.

(4) This follows from the fact, demonstrated above, that v_N^2=O(N).

EXAMPLE MODEL SPECIFICATIONS AND OUTPUT FROM BUGS

ATR #175, Two-Day Sample:

```
Welcome to BUGS on 22 nd Dec 1998  at 13:5:27
BUGS : Copyright (c) 1992 .. 1995 MRC Biostatistics Unit.
All rights reserved.
Version 0.600 for 32 Bit PC.
For general release: please see documentation for disclaimer.
The support of the Economic and Social Research Council (UK)
is gratefully acknowledged.
Bugs>compile("bayrat2.bug")
model bayesrate;
const
        N=2,
        M=3,
        NS=50;
var
   z[N],month[N], day[N], mf[12,M],wf[7,M],
   y[N], R[N,N],T[N,N],  ybar[N], p[M],m[M],ps[NS],sigprior[NS,3],
   u, d[N], yTbar, alg, yT, yS,sig2,taus,sigx,phi,
   groupid, mu1, mu2,yearcnt, loglambda, lambda, N1, N2 ;
data
  mf in "monfac.dat",
  wf in "dayfac.dat",
  sigprior in "sigprior.dat",
  N1, N2 in "rst2175.dat",
  z, month, day in "day2175.dat";
inits in "bayrat.in";
{
m[1]<-370.385; m[2]<-374.451; m[3]<-384.905;
for (i in 1:N) {

 R[i,i] <- 1/(sig2*(1-pow(phi,2)));
 y[i] <- log(z[i]);
 ybar[i] <- u+mf[month[i],groupid]+wf[day[i],groupid];
 for (j in i+1 :N)
  { R[i,j] <- -phi/(sig2*(1-pow(phi,2)));
   R[j,i] <- -phi/(sig2*(1-pow(phi,2)));
   }
 d[i] <- exp(mf[month[i],groupid] + wf[day[i],groupid]);
 }
for (j in 1:M){
 p[j] <- 1/M;
 }
for (j in 1:NS) {ps[j] <- 1/NS;}

y[] ~ dmnorm(ybar[], R[,]);
N1 ~ dpois(mu1);
N2 ~ dpois(mu2);

T[,] <- inverse(R[,]);
alg <- log(m[groupid] - sum(d[]));
yTbar <- alg + u + sig2/2;
yS <- log(sum(z[]));
sig2 <- sigprior[sigx,1];
phi <- sigprior[sigx,2];
taus <- 1/(exp(sig2)-1);
log(mu1) <- yS + loglambda;
log(mu2) <- yT + loglambda;
lambda <- exp(loglambda);
yearcnt <- exp(yT) + sum(z[]);

yT ~ dnorm(yTbar, taus);
u ~ dnorm(0,.001);
loglambda ~ dnorm(0, .001);
groupid ~ dcat(p[]);
sigx ~ dcat(ps[]);
}
```

```
Parsing model declarations.
Loading data value file(s).
Warning -- expected data read before end of file
Warning -- expected data read before end of file
Warning -- expected data read before end of file
Warning -- expected data read before end of file
Warning -- expected data read before end of file
Loading initial value file(s).
Parsing model specification.
Checking model graph for directed cycles.
Generating code.
Generating sampling distributions.
Generating initial values
```

```
Checking model specification.
Choosing update methods.
compilation took  00:00:02
Bugs>
Bugs>init()
Variable not found or bad command syntax
Bugs>
Bugs>data()
     15051        7        3
     15554        7        4

Bugs>
Bugs>update(2000)     time for    2000   updates was  00:00:19
Bugs>
Bugs>monitor(lambda)
Bugs>
Bugs>monitor(u)
Bugs>
Bugs>monitor(sig2)
Bugs>
Bugs>monitor(phi)
Bugs>
Bugs>monitor(R)
Bugs>
Bugs>monitor(taus)
Bugs>
Bugs>monitor(yT)
Bugs>
Bugs>monitor(yearcnt)
Bugs>
Bugs>monitor(groupid)
Bugs>
Bugs>update(5000)     time for    5000   updates was  00:00:48
Bugs>
Bugs>diag(lambda)
             mean       sd       mean       sd        Z       sample
           4.51E-6  2.15E-13  4.51E-6  1.26E-13 -2.30E-2     5000
Bugs>
Bugs>diag(u)
             mean       sd       mean       sd        Z       sample
           9.38      7.94E-3  9.38      3.53E-3  -2.24E-2     5000
Bugs>
Bugs>diag(sig2)
```

```
          mean     sd      mean     sd      Z      sample
        1.84E-2  1.12E-5  2.03E-2  1.36E-5  -1.87      5000
Bugs>
Bugs>diag(phi)
          mean     sd      mean     sd      Z      sample
        4.51E-1  4.05E-4  4.53E-1  1.97E-4  -3.72E-1    5000
Bugs>
Bugs>diag(taus)
          mean     sd      mean     sd      Z      sample
        7.95E+1  6.83E+1  7.75E+1  4.30E+1  9.45E-1    5000
Bugs>
Bugs>diag(yT)
          mean     sd      mean     sd      Z      sample
        1.53E+1  8.71E-3  1.53E+1  5.10E-3  -2.79E-1    5000
Bugs>
Bugs>diag(yearcnt)
          mean     sd      mean     sd      Z      sample
        4.63E+6  1.73E+11  4.70E+6  1.28E+11 -6.19E-1    5000
Bugs>
Bugs>stats(lambda)
          mean      sd     2.5% :  97.5%  CI   median     sample
        4.520E-6  1.464E-6  2.227E-6  7.902E-6  4.314E-6    5000
Bugs>
Bugs>stats(u)
          mean      sd     2.5% :  97.5%  CI   median     sample
        9.384E+0  1.939E-1  9.021E+0  9.716E+0  9.413E+0    5000
Bugs>
Bugs>stats(sig2)
          mean      sd     2.5% :  97.5%  CI   median     sample
        2.006E-2  1.976E-2  5.945E-3  7.592E-2  1.430E-2    5000
Bugs>
Bugs>stats(phi)
          mean      sd     2.5% :  97.5%  CI   median     sample
        4.556E-1  1.394E-1  2.006E-1  6.612E-1  4.556E-1    5000
Bugs>
Bugs>stats(R)
          mean      sd     2.5% :  97.5%  CI   median     sample
[1,1]      1.046E+2  6.437E+1  1.828E+1  2.965E+2  8.739E+1    5000
[1,2]     -5.058E+1  4.135E+1  -1.950E+2  -7.532E+0  -3.608E+1    5000
[2,1]     -5.058E+1  4.135E+1  -1.950E+2  -7.532E+0  -3.608E+1    5000
[2,2]      1.046E+2  6.437E+1  1.828E+1  2.965E+2  8.739E+1    5000
Bugs>
Bugs>stats(taus)
```

```
          mean      sd      2.5% : 97.5% CI   median    sample
       7.796E+1  4.259E+1  1.267E+1  1.676E+2  6.942E+1    5000
Bugs>
Bugs>stats(yT)
          mean      sd      2.5% : 97.5% CI   median    sample
       1.532E+1  2.350E-1  1.488E+1  1.579E+1  1.532E+1    5000
Bugs>
Bugs>stats(yearcnt)
          mean      sd      2.5% : 97.5% CI   median    sample
       4.683E+6  1.202E+6  2.932E+6  7.266E+6  4.560E+6    5000
Bugs>
Bugs>stats(groupid)
          mean      sd      2.5% : 97.5% CI   median    sample
       2.052E+0  8.233E-1  1.000E+0  3.000E+0  2.000E+0    5000
Bugs>
Bugs>q()
```

ATR #175, "Two-Week" Sample

```
Welcome to BUGS on 14 th Dec 1998  at 20:36:4
BUGS : Copyright (c) 1992 .. 1995 MRC Biostatistics Unit.
All rights reserved.
Version 0.600 for 32 Bit PC.
For general release: please see documentation for disclaimer.
The support of the Economic and Social Research Council (UK)
is gratefully acknowledged.
Bugs>compile("bayesrat.bug")
model bayesmdt;
const
        N=13,
        M=3,
        NS=50;
var
   z[N],month[N], day[N], mf[12,M],wf[7,M],
   y[N], R[N,N],T[N,N],  ybar[N], p[M],m[M],ps[NS],sigprior[NS],
   u,nu, d[N], yTbar, alg, yT, yS,sig2,taus,sigx,
   groupid, mu1, mu2,yearcnt, loglambda, lambda, N1, N2 ;
data
  mf in "monfac.dat",
  wf in "dayfac.dat",
  sigprior in "sigprior.dat",
  z, month, day in "samp175.dat",
  N1, N2 in "rest175.dat";
inits in "bayesmdt.in";
{
m[1]<-370.385; m[2]<-374.451; m[3]<-384.905;
for (i in 1:N) {
  R[i,i] <- .028;
  y[i] <- log(z[i]);
  ybar[i] <- u+mf[month[i],groupid]+wf[day[i],groupid];
  for (j in i+1 :N)
   { R[i,j] <- 0;
    R[j,i] <- 0
    }
  d[i] <- exp(mf[month[i],groupid] + wf[day[i],groupid]);
  }
for (j in 1:M){
 p[j] <- 1/M;
 }
for (j in 1:NS) {ps[j] <- 1/NS;}
y[] ~ dmnorm(ybar[], T[,]);
N1 ~ dpois(mu1);
N2 ~ dpois(mu2);
alg <- log(m[groupid] - sum(d[]));
yTbar <- alg + u + sig2/2;
yS <- log(sum(z[]));
nu <- N+2 ;
sig2 <- sigprior[sigx];
taus <- 1/(exp(sig2)-1);
log(mu1) <- yS + loglambda;
log(mu2) <- yT + loglambda;
lambda <- exp(loglambda);
yearcnt <- exp(yT) + sum(z[]);
yT ~ dnorm(yTbar, taus);
u ~ dnorm(0,.001);
loglambda ~ dnorm(0, .001);
T[,] ~ dwish(R[,],nu);
groupid ~ dcat(p[]);
sigx ~ dcat(ps[]);
}

Parsing model declarations.
Loading data value file(s).
Warning -- expected data read before end of file
Warning -- expected data read before end of file
Warning -- expected data read before end of file
Warning -- expected data read before end of file
Warning -- expected data read before end of file
Loading initial value file(s).
Parsing model specification.
Checking model graph for directed cycles.
Generating code.
Generating sampling distributions.
Generating initial values
Checking model specification.
Choosing update methods.
compilation took  00:00:02
Bugs>update(2000)     time for    2000   updates was  00:00:44
Bugs>monitor(lambda)
Bugs>monitor(u)
Bugs>monitor(sig2)
Bugs>monitor(taus)
Bugs>monitor(yT)
Bugs>monitor(yearcnt)
Bugs>monitor(groupid)
Bugs>update(5000)    time for   5000  updates was  00:01:51
Bugs>diag(lambda)
           mean      sd      mean      sd      Z      sample
        3.51E-6  4.92E-14  3.57E-6  2.17E-14 -1.03       5000
Bugs>diag(u)
           mean      sd      mean      sd      Z      sample
        9.59     7.38E-5   9.59     2.68E-5  -4.27E-1    5000
Bugs>diag(sig2)
           mean      sd      mean      sd      Z      sample
        2.87E-2  2.74E-5  2.81E-2  1.48E-5  4.05E-1    5000
Bugs>diag(taus)
           mean      sd      mean      sd      Z      sample
        6.36E+1  5.15E+1  6.41E+1  3.08E+1  -2.47E-1    5000
Bugs>diag(yT)
           mean      sd      mean      sd      Z      sample
        1.55E+1  2.18E-3  1.55E+1  8.88E-4  8.04E-1    5000
Bugs>diag(yearcnt)
           mean      sd      mean      sd      Z      sample
        5.86E+6  9.43E+10  5.80E+6  3.47E+10  8.28E-1    5000
Bugs>stats(lambda)
           mean      sd      2.5% : 97.5% CI    median     sample
        3.569E-6  1.023E-6  1.863E-6  5.847E-6  3.456E-6    5000
Bugs>stats(u)
           mean      sd      2.5% : 97.5%  CI    median     sample
        9.595E+0  2.394E-2  9.550E+0  9.643E+0  9.595E+0    5000
Bugs>stats(sig2)
           mean      sd      2.5% : 97.5%  CI    median     sample
        2.746E-2  2.851E-2  6.261E-3  9.936E-2  1.686E-2    5000
Bugs>stats(taus)
           mean      sd      2.5% : 97.5%  CI    median     sample
        6.483E+1  4.127E+1  9.572E+0  1.592E+2  5.878E+1    5000
Bugs>stats(yT)
           mean      sd      2.5% : 97.5%  CI    median     sample
        1.551E+1  1.720E-1  1.519E+1  1.591E+1  1.550E+1    5000
Bugs>stats(yearcnt)
           mean      sd      2.5% : 97.5%  CI    median     sample
        5.779E+6  1.030E+6  4.187E+6  8.377E+6  5.646E+6    5000
Bugs>q()
```