THE UNIVERSITY OF MANITOBA

IMPLEMENTATION OF THE BAYESIAN PARADIGM

IN MULTIPARAMETER DISTRIBUTIONS

by

JEFFREY ALAN SLOAN

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

WINNIPEG, MANITOBA

JUNE, 1991

IMPLEMENTATION OF THE BAYESIAN PARADIGM IN

MULTIPARAMETER DISTRIBUTIONS

BY

JEFFREY ALAN SLOAN

A thesis submitted to the Faculty of Graduate Studies of
the University of Manitoba in partial fulfillment of the requirements
of the degree of

DOCTOR OF PHILOSOPHY

© 1991

IMPLEMENTATION OF THE BAYESIAN PARADIGM
IN MULTIPARAMETER DISTRIBUTIONS

TABLE OF CONTENTS

ABSTRACT

In this dissertation, the difficulties encountered in implementing the Bayesian paradigm are explored under a number of multiparameter distributional settings. A combination of statistical theory, mathematical approximations and numerical solutions are applied to each problem.

Chapter one contains a precis of the philosophical discussion that has surrounded the implementation of the Bayes paradigm since it was first suggested. The remaining chapters deal with a variety of multiparameter situations where implementation of the Bayesian paradigm is undertaken and compared to results for the classical or likelihood approach.

Chapters two, three and four compare the performance of the classical and Bayesian methods in application to a variety of distributions most commonly used in reliability and lifetesting situations. Two approximations of the Bayes estimator, due to Lindley (1980) and Tierney & Kadane (1986) are compared to assess the degree of success achieved in circumventing the intractability problems common to the Bayes ratio of integrals problem. Improved numerical methods for obtaining maximum likelihood estimators are proposed. Chapter two compares the two Bayes methods to the maximum likelihood estimator in application to the two-parameter normal and two-parameter inverse gaussian distribution. Chapter three continues to detail the applicability of the approximations to three commonly used three-parameter distributions in reliability and lifetesting - the Weibull, gamma and lognormal distributions. Chapter four gives a similar presentation involving the five-parameter estimation task necessary in working with the bivariate normal distribution.

Chapters five and six discuss estimation in two special mixture distribution settings. Both censored and uncensored sampling environments are explored. Chapter five details the work for a mixture of exponentials distributions while Chapter six deals with a mixture of Weibull's distribution. The use of predictive intervals for both distributions is explored.

Chapter seven investigates two settings for Bayesian alternatives to traditional quality control techniques. The first section proposes alternatives for the traditional p-chart control limits by taking advantage of the inherent updating of information available through the Bayesian paradigm. The second portion presents a modification of the standard p-chart based on the predictive distribution.

ACKNOWLEDGEMENTS

ES&D

CHAPTER 1: INTRODUCTION

"All models are incorrect."   George E. P. Box

"A rose by any other name would smell as sweet."   William Shakespeare

1.1 Motivation

Objectives of the work contained herein are twofold. Of primary importance is the advancement of the theory and application of that segment of statistical methodology commonly referred to as Bayesian inference. Much has been built upon the framework as set out by Reverend Thomas Bayes in 1763. More than two hundred years later, Bayesian methodology remains a viable alternative to the classical or likelihood sampling theory approach.

This text undoubtedly will be read by proponents of both schools of thought with an eye towards gathering ammunition for further contributions to the literary debate. It is due to this fact that it is necessary to state, in the most emphatic terms possible, that this dissertation is not written for this purpose.

Bayes and classical methodology are merely different platforms upon which to build a framework for attacking statistical problems. Since, as the opening quote of one of this century's foremost statisticians infers, all such platforms enjoy certain advantages while suffering other shortcomings, absolute superiority by either school of thought is impossible. Philosophical arguments pervade the literature which attempt to ridicule either philosophy. The most recent example with lengthy discussants by well known authors on both sides of the issue can be found in Lindley (1990). Such self-indulgent exercises are merely a sign of academic immaturity bordering on religious fanaticism.

The primary goal of any statistical analysis is to find a quantifiable answer to the problem at hand. The methods used to attain that answer are, for the most part, irrelevant, providing a satisfactory answer is obtained. Moreover, using more than one method to attack the same problem often gains the researcher added insight. As Berger (1990) points out, it is often sensible "to carry along separate models until the end, hoping that the answer will turn out to be insensitive to the various models". Whether the modelling approaches used are referred to as Bayes, Fisherian or a hybrid procedure, the ultimate assessment should be based only on the quality of the answer. Furthermore, since in many situations competing methods produce the same answer, the name of the process is irrelevant.

The motivation of this dissertation, therefore, is to examine the relative strengths and weaknesses of the two doctrines under various distributional settings. As will be seen, neither philosophy can claim uniform superiority. Rather, each approach has its own arena of enhanced performance. Many of these areas were heretofore unexplored from a Bayesian perspective. As the reader progresses through the dissertation, it is desirable to keep this intent of the author in mind. This is not a Bayesian dissertation, nor is it a dissertation of a Bayesian statistician. It is, instead, a Statistics dissertation, exploring alternative solution paths to complex estimation problems.

A persistent criticism of the Bayesian approach is that it often leads one through a morass of algebra culminating in a blank wall of mathematical intractability. As extensively detailed by Smith et al (1987), the advent of modern computer technology has had a major impact on the intractability problem. Advances in numerical analysis techniques allow for numerical solutions or at least approximate solutions when closed

form ones are unobtainable. Along the path to new results in Bayesian methods, one is required to produce innovative algorithmic solutions to the numerical difficulties encountered. "One of the main obstacles to the routine implementation of Bayesian methods has been the absence of efficient algorithms for carrying out the computational tasks implicit in the Bayesian approach" (Smith et al, 1987). Hence the second contribution of this thesis will be the development of algorithms designed to circumvent the mathematical roadblocks.

A great deal of research has been done in estimation theory involving single parameter distributions from both a classical and Bayesian perspective, and is available from many sources in a unified format (Johnson and Kotz, 1970 for example). A considerably smaller battery of work is available for situations in which more than one parameter is of interest, especially when more than two parameters are involved. Algebraic manipulations become rather complex, particularly under a Bayesian framework.

Recently some results have been found to be useful in obtaining approximations to the true Bayes estimators for multiparameter distributions. The application and properties of these approximation techniques will be scrutinized under various distributional settings. Wherever possible, closed form solutions will be obtained.

1.2 The Bayesian Paradigm

The phrase "Bayesian Paradigm" was coined by D. V. Lindley and has come into common usage in the literature. Just as the exact definition of a paradigm is the grammatical collection of a verb's conjugation, so does a paradigm of Bayes

methodology attempt to represent any and all statistical environments which incorporate a Bayesian framework.

Lindley (1990) states that this Bayesian framework relies upon one basic premise: that we "make judgements about as yet unobserved data y on the basis of observed data x". He further refers to the Bayesian paradigm as that field of inference (inductive logic) which utilizes only the calculus of conditional probability. Smith (1986) gives a more intuitive overview of the Bayesian paradigm by calling it a "form of disciplined uncertainty accounting".

Although it may be argued that the classical or sampling theory approach may also be formulated in such a way as to mimic the above definition of the Bayesian format, there remains a fundamental difference - the inferential starting point. While the classical approach defines a static parameter θ and discusses the "likelihood" of various data outcomes in terms of the value of the parameter θ, the Bayesian framework requires that the data remain the starting point and concerns itself instead with the distribution $P(\theta|x)$ of the parameter given the data results. Thus, a key feature of the Bayesian paradigm is that the parameter θ is given the status of an unknown random variable. Another way of expressing the difference would be to say that the classical approach is to be suspicious about the sample results for a given value of θ while the Bayesian perspective involves suspicions about the parameter for given sample results.

The Bayesian approach to estimation may be presented as an alternative to the classical techniques when one considers θ, the quantity to be estimated, as a random rather than a static entity. One assumes that θ, which may be single or vector-valued such that $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_m)$, has an a priori or PRIOR probability density function (pdf)

g(θ) which represents all the information, objective and subjective, that is known regarding the behaviour of the environmental characteristic θ. Throughout this dissertation the notation θ and $\underline{\theta}$ will be used to represent the parameter of interest, be it single or vector-valued. In explaining the works of other authors, the notation used for θ will be consistent with that of the original author. This results in a somewhat uneven representation, but it maintains consistency with the published literature. It will be generally understood that unless explicitly stated otherwise, the parameter of interest is a vector-valued entity.

It is not unreasonable to assume that for a fixed point in time $\underline{\theta}$ will be static and hence the classical school of thought could be employed as well for the same situation. In this light it can be seen that the two inferential methods may be applied to a single experimental setting and are not necessarily contradictory. The decision may be reduced to a question of whether or not for convenience, be it practical or mathematical, it is reasonable to interpret the entity under scrutiny as a fixed constant or variable quantity.

On a basic level, many lay people involved in statistical applications use the Bayesian paradigm instinctively. The so-called "fudge factor" utilised to adjust a result which from the expert individual's experience (prior information), seems too extreme, is in fact an ad hoc Bayesian analysis. For example, in estimating the binomial parameter p involved in a shop floor application, an investigator will often consider results too close to either zero or one for p as biased or improbable from an intuitive standpoint and want to adjust the results accordingly. Mathematically, this is analogous to replacing the informative prior distribution most often used for the binomial parameter p, which is $g(p)=p(1-p)$, by something more in line with past experience, such as

$[p(1-p)]^{0.5}$, so as to reduce or "dampen" the importance of the endpoint values.

Besides basic estimation theory settings, the Bayesian paradigm has gained wide acceptance among industrial settings such as assembly line production investigations. Classical methodology has difficulty incorporating a worker's vital knowledge of the line gained through twenty years of experience into the analysis. Furthermore, the classical framework would require either a complex sequential analysis or the bending of systemic assumptions such as independent sampling to address the time element inherent in the process. The incorporation of today's results in tomorrow's investigation is inherent to the Bayesian paradigm. Today's results merely become tomorrow's prior.

The mechanism of constructing Bayes estimators may be described as follows. It is assumed that X, the data quantity to be observed is a random variable with a probability function $f(x|\theta)$, which involves the environmental characteristic θ. Further assume that θ is a random variable with a prior distribution $g(\theta)$ and that a random sample of n observations is taken from the population with the aforementioned probability function and denoted by $\underline{x} = (X_1, X_2, \ldots, X_n)$.

Define Y to be a statistic that is a function of $\underline{x}$, say $Y = u(X_1, X_2, \ldots, X_n)$. Fisher's likelihood function is meant to represent the information that the sample has to offer on the likelihood of various true values for the parameter θ. It is given by

$$L(\theta|\underline{x}) = \prod_{i=1}^{n} f(x_i|\theta) \quad . \tag{1.1}$$

Finding Bayesian estimators is equivalent to finding the conditional distribution of the statistic Y given the prior information about θ. This conditional distribution will be proportional to the product of the likelihood of the sample and the prior information

known about θ. The conditional distribution of θ given the sample information Y, the so-called POSTERIOR distribution of θ is

$$\pi(\theta|Y) = K(Y,\theta)/C = L(\theta|\underline{x})g(\theta)/C \qquad (1.2)$$

where C^{-1} is a normalizing constant. More specifically, C is given by

$C = \int L(\theta|\underline{x})g(\theta)d\theta$. The posterior distribution of θ represents the idea that the prior information about θ has been adjusted by what has been observed through sampling. A natural bridge between the two methods is the fact that if g(θ) is a uniform density function, then the posterior distribution (1.2) is equivalent to the likelihood function.

Once obtained, the posterior distribution gives rise to a number of alternative Bayes estimators of θ and related characteristics, depending upon the way in which the investigator believes loss will be incurred by an inaccurate estimator. In general, the Bayes estimator is that which minimizes the average or expected loss incurred. The loss incurred by the use of the Bayes estimator θ^* as a guess at θ is denoted by $\ell(\theta^*,\theta)$ and the choice of θ^* is obtained by seeking the minimum of

$$E\left[\ell(\theta^*,\theta)\right] = \int \ell(\theta^*,\theta)\pi(\theta|Y)d\theta$$

The most commonly-used loss function is the quadratic or squared error loss of the form $\ell(\theta^*,\theta) = (\theta^* - \theta)^2$. The expected loss is minimized under squared error loss if θ^* is the mean of the posterior distribution of θ.

Hence the Bayes estimator under squared error loss is defined as the posterior expectation of θ, i.e.

$$\theta^* = E[\theta \mid \underline{x}] = \frac{\int \theta g(\theta) L(\theta \mid \underline{x}) d\theta}{\int g(\theta) L(\theta \mid \underline{x}) d\theta} \quad . \tag{1.3}$$

Although other loss functions are used, unless otherwise specified this text will only be concerned with squared error loss models. Howlader (1982) gives a detailed discussion of the effects of various loss functions.

Recall that θ may be a vector-valued parameter and thus produce a multivariate posterior distribution. In such a case, it is necessary to find the marginal posterior for each element of θ to construct the corresponding Bayes estimators.

Clearly, the Bayesian paradigm is more than merely constructing Bayes estimators. Examination of the posterior distribution itself is but one of many points of interest in such a framework. Rather than delineate all possible applications and methodology, the Bayesian paradigm description will be left at this point. Other aspects will be introduced as required in the text to follow.

1.3 Implementation Considerations for the Bayesian Paradigm

The previous section infers that implementation of the Bayesian paradigm is straightforward. Clearly, the degree of difficulty encountered will be directly linked to the nature of the prior and likelihood functions. If they comprise a natural "conjugate" pair, typically of the exponential family, then the integrations required in (1.3) to produce the Bayes estimator are able to be performed analytically and produce a closed form solution. Press (1989) summarizes the use of the more commonly seen natural

conjugate priors. Unfortunately in the majority of cases, as Naylor and Smith (1982) point out, "the forms of likelihoods and/or prior densities do not permit such a tractable analysis and the required integrations must either be performed numerically, or analytic approximations found". It is this basic difficulty with the mathematics of calculus that has hampered the application of the Bayesian paradigm virtually since its inception.

One of the approaches taken to solve this problem involves the use of mathematically simplistic prior distributions when little is known regarding the behaviour of the parameter of interest. It has been argued that if Bayes methods perform well under this "worst case" scenario, performance will only be enhanced with improved prior information. Most notable among efforts in the development of "useful" prior distributions are Jeffreys (1961) and Hartigan (1964) in producing the so-called noninformative prior. Howlader (1982, Section 1.4) provides a detailed discussion of alternative prior distributions and their relative merits. More recent work has been done by Zellner (1986), Efron et al (1986) and Berger & Bernardo (1989).

Many statisticians have addressed the evaluation of (1.3) under the title "the ratio of integrals problem", when analytic closed form solutions are unavailable. Several authors have attacked the problem by drawing on a wide variety of mathematical theories to approximate the integral ratio in (1.3). Most notable are approximations due to Lindley (1980) and Tierney/Kadane (1986). Reilly (1976), Smith et al (1985), Van Dijk, Hop & Louter (1987), Leonard et al (1989) and Gelfand & Smith (1990) also propose methods to approximate the posterior distribution.

Virtually all of these methods would be unusable except over a very narrow range of problems if it were not for the availability of considerable computing power.

Although these approaches bring the most modern technology to bear on what has been a longstanding problem, they also introduce the pitfalls of this new methodology to the process. Just as it is a folly to suggest that constructing the posterior distribution formula (1.3) means it is simple to carry out, so is it easy to forget that involving the computer in the analytic loop introduces some new difficulties that must be overcome. These computer-based problems can be split into two categories: method and accuracy.

There are various competing methods to attack any numerical problem, each one having its own strengths and weaknesses. Numerical search routines form a large part of the body of computer science literature. Due to the variety and complexity of Bayes problems, it is not possible to find an algorithm to suit all cases. The most commonly known Newton-Raphson search technique, for example, performs well for some problems while failing at others. Rice (1988) provides an extensive discussion of alternative methods.

Any mathematical exercise on the computer has a built in error. Such errors compound directly with the number of calculations performed. In Statistics we deal with mathematical truisms such as the sum of a probability distribution being one. The computer, no matter how precise it may be, cannot incorporate such rules. Further complications exist for the Bayesian approach in that the very form of (1.3) typically involves calculation of extremely large or extremely small quantities. These produce overflow/underflow complexities which must be addressed with great care so as to not lose too much accuracy for the final result. Hence in using a computer to produce the results of a mathematically designed approximation to the ratio of integrals problem, we are actually performing an approximation to the approximation. Good programming

techniques minimize the error of the computer approximation, but it is important to point out the issue so as to forewarn the practitioner.

1.4 Attempts to Address Implementation Problems

Many authors have attempted to address the problems of implementing the Bayesian paradigm. This section will describe in detail two such attempts which will be used extensively in this dissertation. Other alternatives will be described briefly.

1.4.1 Lindley's Approximation

Lindley (1980) proposed a method to tackle the general problem of evaluating the ratio of two integrals and obtaining the Bayes estimator $u^*(\theta)$, where $u(\theta)$ is a function of the vector-valued parameter θ, defined as

$$u^*(\theta) = \frac{\int u(\theta) g(\theta) \mathcal{L}(\theta \mid \underline{x}) \, d\theta}{\int g(\theta) \mathcal{L}(\theta \mid \underline{x}) \, d\theta} \tag{1.4}$$

where $u(\theta)$ is an arbitrary function of the parameter of interest, $g(\theta)$ is the prior distribution of θ and $\mathcal{L}(\theta \mid \underline{x})$ is the likelihood function. Lindley actually utilised $L(\theta \mid \underline{x})$, the log of the likelihood function $\mathcal{L}(\theta \mid \underline{x})$ in his paper so that (1.4) becomes

$$u^*(\theta) = \frac{\int u(\theta) g(\theta) \exp[L(\theta \mid \underline{x})] \, d\theta}{\int g(\theta) \exp[L(\theta \mid \underline{x})] \, d\theta} \tag{1.5}$$

which is the Bayes estimator of $u(\theta)$ under squared-error loss.

Consider the numerator of (1.5) and expand both $w(\theta) = u(\theta) g(\theta)$ and $L(\theta \mid \underline{x})$ by Taylor series expansion about the maximum likelihood estimator (mle) of θ designated

as $\hat{\theta}$. For the sake of brevity $L(\theta|\underline{x})$ will be written simply as L and all functions involving θ assumed to be evaluated at the mle.

The Taylor series expansion of L about $\hat{\theta}$ is

$$L(\theta) = L(\hat{\theta}) + \sum_{i=1}^{m} L_i(\theta_i - \hat{\theta}_i) + \frac{1}{2!}\sum_{i,j=1}^{m} L_{ij}(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

$$+ \frac{1}{3!}\sum_{i,j,k=1}^{m} L_{ijk}(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)(\theta_k - \hat{\theta}_k) + \ldots \qquad (1.6)$$

where

$$L_i = \frac{\partial L}{\partial \theta_i}, \quad L_{ij} = \frac{\partial^2 L}{\partial \theta_i \partial \theta_j}, \quad L_{ijk} = \frac{\partial^3 L}{\partial \theta_i \partial \theta_j \partial \theta_k}$$

Using θ_i to represent the deviation term $(\theta_i - \hat{\theta}_i)$, the expression simplifies to

$$L(\theta) = L(\hat{\theta}) + \sum_{i=1}^{m} L_i\theta_i + \frac{1}{2!}\sum_{i,j=1}^{m} L_{ij}\theta_i\theta_j + \frac{1}{3!}\sum_{i,j,k=1}^{m} L_{ijk}\theta_i\theta_j\theta_k + \ldots$$

where all summations run over all subscripts from 1 to m, the dimensionality of θ.

Applying the Taylor series expansion to $w(\theta) = u(\theta)g(\theta)$ gives

$$w(\theta) = w(\hat{\theta}) + \sum_{i=1}^{m} w_i\theta_i + \frac{1}{2!}\sum_{i,j=1}^{m} w_{ij}\theta_i\theta_j + \frac{1}{3!}\sum_{i,j,k=1}^{m} W_{ijk}\theta_i\theta_j\theta_k + \ldots$$

Hence

$$\int w(\theta)\exp L(\theta)\,d\theta = \int \left[w(\hat\theta) + \sum w_i\theta_i + \frac{1}{2}\sum w_{ij}\theta_i\theta_j + \frac{1}{6}\sum w_{ijk}\theta_i\theta_j\theta_k + \ldots \right]$$

$$\cdot \exp\left[L(\hat\theta) + \sum L_i\theta_i + \frac{1}{2}\sum L_{ij}\theta_i\theta_j + \frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + \ldots \right] d\theta \ .$$

To simplify, factor out the w(θ) term and define

$$W_i = \frac{w_i}{w(\hat\theta)} \ , \quad W_{ij} = \frac{w_{ij}}{w(\hat\theta)} \ , \quad W_{ijk} = \frac{w_{ijk}}{w(\hat\theta)} \ , \quad and\ so\ on.$$

All terms of the form $L_i\theta_i$ are zero since the functions are evaluated at the mle. This leads us to the simplified form

$$\int w(\theta)\exp[L(\theta)]\,d\theta = w(\hat\theta)e^{L(\hat\theta)}$$

$$\cdot \int \left[1 + \sum W_i\theta_i + \frac{1}{2}\sum W_{ij}\theta_i\theta_j + \frac{1}{6}\sum W_{ijk}\theta_i\theta_j\theta_k + \ldots \right]$$

$$\cdot \exp\left[\frac{1}{2}\sum L_{ij}\theta_i\theta_j + \frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + \ldots \right] d\theta \ . \qquad (1.7)$$

Lindley then uses the relation $e^{a+x} = e^a[1 + x + x^2/2 + x^3/6 + \ldots]$ and sets $a = (1/2)\Sigma L_{ij}\theta_i\theta_j$ to get

$$\int w(\theta)\exp[L(\theta)]\,d\theta = w(\hat\theta)e^{L(\hat\theta)}$$

$$\int \left[1 + \sum W_i\theta_i + \frac{1}{2}\sum W_{ij}\theta_i\theta_j + \frac{1}{6}\sum W_{ijk}\theta_i\theta_j\theta_k + \ldots \right]$$

$$\cdot \exp\left[\frac{1}{2}\sum L_{ij}\theta_i\theta_j \right] \cdot \left[1 + \left(\frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + \frac{1}{24}\sum L_{ijkl}\theta_i\theta_j\theta_k\theta_l + \ldots \right) \right.$$

$$\left. + \left(\frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + \frac{1}{24}\sum L_{ijkl}\theta_i\theta_j\theta_k\theta_l + \ldots \right)^2 + \ldots \right]d\theta \ .$$

Expanding the latter part of the equation produces

$$\int w(\theta) \exp[L(\theta)] d\theta = w(\hat{\theta}) e^{L(\hat{\theta})} \int \exp\left[\frac{1}{2}\sum L_{ij}\theta_i\theta_j\right]$$

$$\cdot [1 + \sum W_i\theta_i + \frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + \frac{1}{2}\sum W_{ij}\theta_i\theta_j$$

$$+ \left(\sum W_i\theta_i\right)\frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + \frac{1}{24}\sum L_{ijkl}\theta_i\theta_j\theta_k\theta_l + \ldots] d\theta .$$

Lindley (1980) notes that L, and all its derivatives are O(n) and further that all terms involving any segment of the multiparameter θ are $O(n^{-1/2})$. He considers the terms only to $O(n^{-1})$, assuming that terms of higher order tend towards zero as n→∞. Using this assumption, the numerator of the integral ratio (1.5) becomes

$$\int w(\theta) \exp[L(\theta)] = w(\hat{\theta}) e^{L(\hat{\theta})} \int \exp\left[\frac{1}{2}\sum L_{ij}\theta_i\theta_j\right]$$

$$\cdot [1 + \sum W_i\theta_i + \frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + \frac{1}{2}\sum W_{ij}\theta_i\theta_j$$

$$+ \left(\sum W_i\theta_i\right)\frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + R] d\theta \qquad (1.8)$$

where R contains terms involving only L, and not W or its first derivatives.

The term $\exp[(1/2)\Sigma L_{ij}\theta_i\theta_j]$ is in the form of the kernel of the multivariate normal distribution. The mean vector is $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$. The variance-covariance matrix for this multivariate normal will thus be the inverse of the matrix with the $-L_{ij}$ as elements. Write this as $\Sigma = [\sigma_{ij}] = [-L_{ij}]^{-1}$. The expression (1.8) then involves integration of moments of the multivariate normal distribution. Using the fact that the

mle θ is asymptotically unbiased for θ and the following results (Anderson, 1958; page 39) for the multivariate normal distribution,

$$E(\theta_i - \hat{\theta}_i) = 0$$

$$E(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) = \sigma_{ij}$$

$$E(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)(\theta_k - \hat{\theta}_k) = 0$$

$$E(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)(\theta_k - \hat{\theta}_k)(\theta_l - \hat{\theta}_l) = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \quad .$$

Equation (1.8) will yield

$$\int w(\theta)\exp[L(\theta)]d\theta = w(\hat{\theta})e^{L(\hat{\theta})}(2\pi)^{m/2}|\sum|^{1/2}$$

$$\cdot E[1 + \sum W_i\theta_i + \frac{1}{6}\sum L_{ijk}\theta_i\theta_j\theta_k + \frac{1}{2}\sum W_{ij}\theta_i\theta_j$$

$$+ \frac{1}{6}(W_i\theta_i)\sum L_{ijk}\theta_i\theta_j\theta_k + R] \quad . \tag{1.9}$$

Note that the second last term may be rewritten as

$$\frac{1}{6}(W_i\theta_i)\sum L_{ijk}\theta_i\theta_j\theta_k = \frac{1}{6}\sum_{i,j,k,l} L_{ijk}W_l\theta_i\theta_j\theta_k\theta_l$$

which involves the fourth moment. Applying the results for the multivariate normal distribution reduces (1.9) to

$$\int w(\theta)\exp[L(\theta)]d\theta = w(\hat{\theta})e^{L(\hat{\theta})}(2\pi)^{m/2}|\sum|^{1/2} \cdot [1 + \frac{1}{2}\sum W_{ij}\sigma_{ij}$$

$$+ \frac{1}{6}\sum L_{ijk}W_l(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}) + R^*] \tag{1.10}$$

where R* is the appropriate result of integrating all the remaining terms, not involving w. Recall that L_{ijk}, being a partial derivative term, has the property that the order of

differentiation is irrelevant to the final answer (for example $L_{123} = L_{312}$). Hence the three-term permutation of the σ subscripts in the last term of (1.10) are redundant as the three summations will be equivalent.

Thus, the basic result is

$$\int w(\theta) \exp[L(\theta)] = w(\hat{\theta}) e^{L(\hat{\theta})} (2\pi)^{m/2} |\Sigma|^{1/2}$$

$$\cdot \; [1 + \frac{1}{2}\sum W_{ij}\sigma_{ij} + \frac{1}{2}\sum L_{ijk}W_{l}\sigma_{ij}\sigma_{kl} + R^{*}] \quad .$$

If this algebraic route is repeated for the denominator of (1.5), the result is identical except for the w's being replaced by v's(say), where in the present application $v(\theta) = g(\theta)$. Furthermore, in this application the likelihood function of interest in the numerator and denominator will be the same since the denominator is always merely a normalizing constant on the same posterior distribution. It follows directly that the variance-covariance matrix Σ will be the same for the numerator and denominator of (1.5). Hence there will be considerable cancellation once the ratio is constructed, leaving the result

$$u^{*}(\theta) = \frac{w(\hat{\theta})\left[1 + \frac{1}{2}\sum W_{ij}\sigma_{ij} + \frac{1}{2}\sum L_{ijk}W_{l}\sigma_{ij}\sigma_{kl} + R^{*}\right]}{v(\hat{\theta})\left[1 + \frac{1}{2}\sum V_{ij}\sigma_{ij} + \frac{1}{2}\sum L_{ijk}V_{l}\sigma_{ij}\sigma_{kl} + R^{*}\right]}$$

$$u^{*}(\theta) = \frac{w(\hat{\theta})}{v(\hat{\theta})}\left[1 + \frac{1}{2}\sum W_{ij}\sigma_{ij} + \frac{1}{2}\sum L_{ijk}W_{l}\sigma_{ij}\sigma_{kl}\right] \qquad (1.11)$$

$$\cdot \left[1 - \left(\frac{1}{2}\sum V_{ij}\sigma_{ij} + \frac{1}{2}\sum L_{ijk}V_{l}\sigma_{ij}\sigma_{kl}\right) + (\ldots)^{2} - (\ldots)^{3} + \ldots\right]$$

so that

$$u^*(\theta) = \frac{w(\hat{\theta})}{v(\hat{\theta})}\left[1 + \frac{1}{2}\sum (W_{ij}-V_{ij})\,\sigma_{ij} + \frac{1}{2}\sum (W_I-V_I)\,L_{ijk}\sigma_{ij}\sigma_{kl} + \ldots\right] \; .$$

Recall that for this special application, v(θ) was defined to be the prior distribution g(θ) and w(θ) was set to be the product of the prior and the parameter of interest, producing by definition the posterior mean of u(θ) which is the Bayes estimator under squared-error loss. Translating back into the original notation then, it is seen that

$$W_i - V_i = \frac{w_i}{w} - \frac{v_i}{v}$$

becomes

$$W_i - V_i = \frac{\dfrac{\partial u(\theta)g(\theta)}{\partial\theta_i}}{u(\theta)g(\theta)} - \frac{\dfrac{\partial u(\theta)g(\theta)}{\partial\theta_i}}{g(\theta)}$$

which, after differentiation and using the established algebra becomes

$$W_i - V_i = \frac{g(\theta)u_i(\theta) + u(\theta)g_i(\theta)u(\theta)}{u(\theta)g(\theta)} = \frac{u_i(\theta)}{u(\theta)} \; . \qquad (1.12)$$

Note that all functions are evaluated at the maximum likelihood estimator $\hat{\theta}$.

Similarly, it can be shown that

$$W_{ij} - V_{ij} = \frac{u_{ij}(\theta)}{u(\theta)} + \frac{u_i(\theta)g_j(\theta) + u_j(\theta)g_i(\theta)}{u(\theta)g(\theta)} \; . \qquad (1.13)$$

Substituting the results (1.12) and (1.13) into (1.11) gives

$$u^*(\theta) = u(\hat{\theta}) + \frac{1}{2}\sum_{i,j}\left[u_{ij}(\theta) + \frac{2u_{ij}(\theta)g_j(\theta)}{g(\theta)}\right]\sigma_{ij}$$

$$+ \frac{1}{2}\sum_{i,j,k,l} u_l(\theta)L_{ijk}\sigma_{ij}\sigma_{kl} \quad . \tag{1.14}$$

For convenience, Lindley (1980) assigns $\rho(\theta) = \log[g(\theta)]$ so that (1.14) becomes

$$u^*(\theta) = u(\hat{\theta}) + \frac{1}{2}\sum_{i,j}\left[u_{ij}(\theta) + 2u_i(\theta)\rho_j(\theta)\right]\sigma_{ij} + \frac{1}{2}\sum_{i,j,k,l} L_{ijk}u_l(\theta)\sigma_{ij}\sigma_{kl}$$

$$(1.15)$$

evaluated at $\theta = \hat{\theta}$, the maximum likelihood estimator. This is the basic result. One

of the major drawbacks to this approach is that it requires evaluation of the third

derivative of the likelihood function, which for certain distributional settings may be

onerous. Lindley (1980) describes an alternate route by using the logarithmic posterior

distribution $\Lambda(\theta)=\log[g(\theta)L(\underline{x}\,|\,\theta)]$ in the expansion, which is maximized at the posterior

mode. The form of the final approximation then becomes

$$u^{**}(\theta) = u(\theta'') + \frac{1}{2}\sum_{i,j}u_{ij}\tau_{ij} + \frac{1}{2}\sum_{i,j,k,l}\Lambda_{ijk}u_l\tau_{ij}\tau_{kl} \tag{1.16}$$

where θ'' is the posterior mode, the point at which all functions in (1.16) are evaluated,

and $\tau_{ij} = -\Lambda_{ij}$. The advantage of this approach over the use of the mle expansion will be

for distributions that have unruly forms for the derivatives of the likelihood functions.

In such situations, Lindley suggests that the method of differencing be used whereby the

appropriate derivative function is approximated by the same order differenced function.

This allows the researcher to be able to compute the Bayes estimator without having to refer to an iterative computer search technique to evaluate the derivatives of the likelihood function. It does involve, however, the evaluation of the log-posterior over a surface grid surrounding the posterior mode θ''. Howlader and Weiss (1987a) give an excellent exposition on the details of carrying out the differencing technique.

Howlader (1982) has used this technique for one- and two-parameter estimation problems. Howlader and Weiss (1987(a) and (b)) discuss some of the advantages and problems in the implementation of the procedure in these specific settings for the Cauchy distribution. Their work provides detailed explanation of the implementation of Lindley's technique using both the posterior evaluated at the mle and the log-posterior evaluated at the mode approaches.

To facilitate the implementation of Lindley's method, a computer program was developed to produce the general expanded form of Lindley's approximation equation for any number of parameters under consideration. The equation can be formulated in terms of the third partial derivatives (L_{ijk}'s), elements of the inverse Hessian matrix (σ_{ij}'s) and partial derivatives of the log-prior (ρ_j's) and the parameter to be estimated (u_i's and u_{ij}'s). This algorithm, written in the BASIC language, was included in Press' (1989) review of algorithms available for Bayesian inference under the name LINDLEY.BAS. The program produces the algebraic form after expansion and collection of like terms.

Algebraic forms for one and two parameters have been given by several authors (Sinha, 1986; Howlader and Weiss, 1988 for example). Forms for three and four parameters are given in Output #1 and Output #2. Although the program does produce results for as many as seven parameters, the equation is too long to be of practical use.

The output equation for five parameters, for example, covers seven pages. Once input to a symbolic algebra program, however, the production of Lindley's expansion for any number of parameters becomes a manageable task by simply entering the appropriate elements (L_{ijk}'s, σ_{ij}'s the ρ_j's and the u_i's, u_{ij}'s) and allowing the program to perform the tedious algebra.

OUTPUT #1

LINDLEY'S METHOD ALGEBRAIC EXPANSION

To estimate the vector-parameter $\underline{\theta}$ and related functions,

where $\underline{\theta}$ is made up of 3 separate parameters, Lindley's equation is

$$\underline{\theta}_{\text{Bayes}} = \underline{\theta}_{\text{MLE}} + (1/2)\text{PART1} + (1/2)\text{PART2}$$

where PART1 and PART2 are given by:

PART1

$$\sigma_{11}(u_{11}+2u_1\rho_1)+2\sigma_{12}(u_{12}+u_1\rho_2+u_2\rho_1)+2\sigma_{13}(u_{13}+u_1\rho_3+u_3\rho_1)+\sigma_{22}(u_{22}+2u_2\rho_2)$$

$$+2\sigma_{23}(u_{23}+u_2\rho_3+u_3\rho_2)+\sigma_{33}(u_{33}+2u_3\rho_3)$$

PART2

$$L_{111}(u_1\sigma_{11}^2+u_2\sigma_{11}\sigma_{12}+u_3\sigma_{11}\sigma_{13})+L_{112}(3u_1\sigma_{11}\sigma_{12}+u_2\sigma_{11}\sigma_{22}+2u_2\sigma_{12}^2+u_3\sigma_{11}\sigma_{23}+2u_3\sigma_{12}\sigma_{13})$$

$$+L_{113}(3u_1\sigma_{11}\sigma_{13}+u_2\sigma_{11}\sigma_{23}+2u_2\sigma_{12}\sigma_{13}+u_3\sigma_{11}\sigma_{33}+2u_3\sigma_{13}^2)$$

$$+L_{122}(u_1\sigma_{11}\sigma_{22}+2u_1\sigma_{12}^2+3u_2\sigma_{12}\sigma_{22}+2u_3\sigma_{12}\sigma_{23}+u_3\sigma_{13}\sigma_{22})$$

$$+L_{123}(2u_1\sigma_{11}\sigma_{23}+4u_1\sigma_{12}\sigma_{13}+4u_2\sigma_{12}\sigma_{23}+2u_2\sigma_{13}\sigma_{22}+2u_3\sigma_{12}\sigma_{33}+4u_3\sigma_{13}\sigma_{23})$$

$$+L_{133}(u_1\sigma_{11}\sigma_{33}+2u_1\sigma_{13}^2+u_2\sigma_{12}\sigma_{33}+2u_2\sigma_{13}\sigma_{23}+3u_3\sigma_{13}\sigma_{33})+L_{222}(u_1\sigma_{12}\sigma_{22}+u_2\sigma_{22}^2+u_3\sigma_{22}\sigma_{23})$$

$$+L_{223}(2u_1\sigma_{12}\sigma_{23}+u_1\sigma_{13}\sigma_{22}+3u_2\sigma_{22}\sigma_{23}+u_3\sigma_{22}\sigma_{33}+2u_3\sigma_{23}^2)$$

$$+L_{233}(u_1\sigma_{12}\sigma_{33}+2u_1\sigma_{13}\sigma_{23}+u_2\sigma_{22}\sigma_{33}+2u_2\sigma_{23}^2+3u_3\sigma_{23}\sigma_{33})+L_{333}(u_1\sigma_{13}\sigma_{33}+u_2\sigma_{23}\sigma_{33}+u_3\sigma_{33}^2)$$

OUTPUT #2

LINDLEY'S METHOD ALGEBRAIC EXPANSION

To estimate the vector-parameter $\underline{\theta}$ and related functions,

where $\underline{\theta}$ is made up of 4 separate parameters, Lindley's equation is

$$\underline{\theta}_{Bayes} = \underline{\theta}_{MLE} + (1/2)PART1 + (1/2)PART2$$

where PART1 and PART2 are given by:

PART1

$\sigma_{11}(u_{11}+2u_1\rho_1)+2\sigma_{12}(u_{12}+u_1\rho_2+u_2\rho_1)+2\sigma_{13}(u_{13}+u_1\rho_3+u_3\rho_1)+2\sigma_{14}(u_{14}+u_1\rho_4+u_4\rho_1)$

$+\sigma_{22}(u_{22}+2u_2\rho_2)+2\sigma_{23}(u_{23}+u_2\rho_3+u_3\rho_2)+2\sigma_{24}(u_{24}+u_2\rho_4+u_4\rho_2)+\sigma_{33}(u_{33}+2u_3\rho_3)$

$+2\sigma_{34}(u_{34}+u_3\rho_4+u_4\rho_3)+\sigma_{44}(u_{44}+2u_4\rho_4)$

PART2

$L_{111}(u_1\sigma_{11}^2+u_2\sigma_{11}\sigma_{12}+u_3\sigma_{11}\sigma_{13}+u_4\sigma_{11}\sigma_{14})$

$+L_{112}(3u_1\sigma_{11}\sigma_{12}+u_2\sigma_{11}\sigma_{22}+2u_2\sigma_{12}^2+u_3\sigma_{11}\sigma_{23}+2u_3\sigma_{12}\sigma_{13}+u_4\sigma_{11}\sigma_{24}+2u_4\sigma_{12}\sigma_{14})$

$+L_{113}(3u_1\sigma_{11}\sigma_{13}+u_2\sigma_{11}\sigma_{23}+2u_2\sigma_{12}\sigma_{13}+u_3\sigma_{11}\sigma_{33}+2u_3\sigma_{13}^2+u_4\sigma_{11}\sigma_{34}+2u_4\sigma_{13}\sigma_{14})$

$+L_{114}(3u_1\sigma_{11}\sigma_{14}+u_2\sigma_{11}\sigma_{24}+2u_2\sigma_{12}\sigma_{14}+u_3\sigma_{11}\sigma_{34}+2u_3\sigma_{13}\sigma_{14}+u_4\sigma_{11}\sigma_{44}+2u_4\sigma_{14}^2)$

$+L_{122}(u_1\sigma_{11}\sigma_{22}+2u_1\sigma_{12}^2+3u_2\sigma_{12}\sigma_{22}+2u_3\sigma_{12}\sigma_{23}+u_3\sigma_{13}\sigma_{22}+2u_4\sigma_{12}\sigma_{24}+u_4\sigma_{14}\sigma_{22})$

$+L_{123}(2u_1\sigma_{11}\sigma_{23}+4u_1\sigma_{12}\sigma_{13}+4u_2\sigma_{12}\sigma_{23}+2u_2\sigma_{13}\sigma_{22}+2u_3\sigma_{12}\sigma_{33}+4u_3\sigma_{13}\sigma_{23}$
$\qquad +2u_4\sigma_{12}\sigma_{34}+2u_4\sigma_{13}\sigma_{24}+2u_4\sigma_{14}\sigma_{23})$

$+L_{124}(2u_1\sigma_{11}\sigma_{24}+4u_1\sigma_{12}\sigma_{14}+4u_2\sigma_{12}\sigma_{24}+2u_2\sigma_{14}\sigma_{22}+2u_3\sigma_{12}\sigma_{34}+2u_3\sigma_{13}\sigma_{24}$
$\qquad +2u_3\sigma_{14}\sigma_{23}+2u_4\sigma_{12}\sigma_{44}+4u_4\sigma_{14}\sigma_{24})$

$$+L_{133}\left(u_1\sigma_{11}\sigma_{33}+2u_1\sigma_{13}^2+u_2\sigma_{12}\sigma_{33}+2u_2\sigma_{13}\sigma_{23}+3u_3\sigma_{13}\sigma_{33}+2u_4\sigma_{13}\sigma_{34}+u_4\sigma_{14}\sigma_{33}\right)$$

$$+L_{134}\left(2u_1\sigma_{11}\sigma_{34}+4u_1\sigma_{13}\sigma_{14}+2u_2\sigma_{12}\sigma_{34}+2u_2\sigma_{13}\sigma_{24}+2u_2\sigma_{14}\sigma_{23}+4u_3\sigma_{13}\sigma_{34}\right.$$
$$\left.+2u_3\sigma_{14}\sigma_{33}+2u_4\sigma_{13}\sigma_{44}+4u_4\sigma_{14}\sigma_{34}\right)$$

$$+L_{144}\left(u_1\sigma_{11}\sigma_{44}+2u_1\sigma_{14}^2+u_2\sigma_{12}\sigma_{44}+2u_2\sigma_{14}\sigma_{24}+u_3\sigma_{13}\sigma_{44}+2u_3\sigma_{14}\sigma_{34}+3u_4\sigma_{14}\sigma_{44}\right)$$

$$+L_{222}\left(u_1\sigma_{12}\sigma_{22}+u_2\sigma_{22}^2+u_3\sigma_{22}\sigma_{23}+u_4\sigma_{22}\sigma_{24}\right)$$

$$+L_{223}\left(2u_1\sigma_{12}\sigma_{23}+u_1\sigma_{13}\sigma_{22}+3u_2\sigma_{22}\sigma_{23}+u_3\sigma_{22}\sigma_{33}+2u_3\sigma_{23}^2+u_4\sigma_{22}\sigma_{34}+2u_4\sigma_{23}\sigma_{24}\right)$$

$$+L_{224}\left(2u_1\sigma_{12}\sigma_{24}+u_1\sigma_{14}\sigma_{22}+3u_2\sigma_{22}\sigma_{24}+u_3\sigma_{22}\sigma_{34}+2u_3\sigma_{23}\sigma_{24}+u_4\sigma_{22}\sigma_{44}+2u_4\sigma_{24}^2\right)$$

$$+L_{233}\left(u_1\sigma_{12}\sigma_{33}+2u_1\sigma_{13}\sigma_{23}+u_2\sigma_{22}\sigma_{33}+2u_2\sigma_{23}^2+3u_3\sigma_{23}\sigma_{33}+2u_4\sigma_{23}\sigma_{34}+u_4\sigma_{24}\sigma_{33}\right)$$

$$+L_{234}\left(2u_1\sigma_{12}\sigma_{34}+2u_1\sigma_{13}\sigma_{24}+2u_1\sigma_{14}\sigma_{23}+2u_2\sigma_{22}\sigma_{34}+4u_2\sigma_{23}\sigma_{24}+4u_3\sigma_{23}\sigma_{34}\right.$$
$$\left.+2u_3\sigma_{24}\sigma_{33}+2u_4\sigma_{23}\sigma_{44}+4u_4\sigma_{24}\sigma_{34}\right)$$

$$+L_{244}\left(u_1\sigma_{12}\sigma_{44}+2u_1\sigma_{14}\sigma_{24}+u_2\sigma_{22}\sigma_{44}+2u_2\sigma_{24}^2+u_3\sigma_{23}\sigma_{44}+2u_3\sigma_{24}\sigma_{34}+3u_4\sigma_{24}\sigma_{44}\right)$$

$$+L_{333}\left(u_1\sigma_{13}\sigma_{33}+u_2\sigma_{23}\sigma_{33}+u_3\sigma_{33}^2+u_4\sigma_{33}\sigma_{34}\right)$$

$$+L_{334}\left(2u_1\sigma_{13}\sigma_{34}+u_1\sigma_{14}\sigma_{33}+2u_2\sigma_{23}\sigma_{34}+u_2\sigma_{24}\sigma_{33}+3u_3\sigma_{33}\sigma_{34}+u_4\sigma_{33}\sigma_{44}+2u_4\sigma_{34}^2\right)$$

$$+L_{344}\left(u_1\sigma_{13}\sigma_{44}+2u_1\sigma_{14}\sigma_{34}+u_2\sigma_{23}\sigma_{44}+2u_2\sigma_{24}\sigma_{34}+u_3\sigma_{33}\sigma_{44}+2u_3\sigma_{34}^2+3u_4\sigma_{34}\sigma_{44}\right)$$

$$+L_{444}\left(u_1\sigma_{14}\sigma_{44}+u_2\sigma_{24}\sigma_{44}+u_3\sigma_{34}\sigma_{44}+u_4\sigma_{44}^2\right)$$

1.4.2 Tierney-Kadane Approximation

Tierney and Kadane (1986) attacked the ratio of integrals problem from a perspective different from that of Lindley, employing the Laplace method for integrals. Their technique is closely related to that of Lindley however, in that the major difference is merely the means by which the complicated integral equation (1.3) is evaluated. Rather than expanding about the maximum likelihood estimators, Tierney & Kadane use the point which maximizes the value of the integral, which in this case turns out to be a function of the posterior mode. Because the numerator and denominator are of different forms, it is to be expected that the point about which each integral is maximized will differ in the two cases. Hence this method uses a different point about which to expand each integral in a Taylor series. They then use a result involving the Laplace method for integrals to evaluate the integrals and, subsequently, the ratio. In competition with the Lindley approximation, the Tierney-Kadane (T-K) method has been seen to be more accurate in terms of estimating posterior variance and numerically more convenient in some cases, although does suffer when the sample size n is small (Howlader and Weiss (1987)). The T-K method represents an alternative to Lindley's method so that given suitable conditions it is left to the researcher to decide which of the two techniques is most appropriate for the particular situation under study.

Specifically, the method begins by reexpressing the integral ratio (1.5) as

$$u^*(\theta) = \frac{\int u(\theta)g(\theta)\exp[L(\theta|\underline{x})]d\theta}{\int g(\theta)\exp[L(\theta|\underline{x})]d\theta} = \frac{\int \exp[nL_*(\theta_*)]d\theta}{\int \exp[nL_o(\theta_o)]d\theta}$$

(1.17)

where

$$L_o(\theta) = \frac{\left(\log[g(\theta)] + L(\theta|\underline{x})\right)}{n}$$

$$L_*(\theta) = \frac{\left(\log[u(\theta)] + \log[g(\theta)] + L(\theta|\underline{x})\right)}{n} = L_o + \frac{\log[u(\theta)]}{n}$$

and $L(\theta|\underline{x})$ is the logarithmic-likelihood function. The points θ_o and θ_* maximize the L_o and L_* functions respectively. Tierney and Kadane then proceed to construct asymptotic approximations for the numerator and denominator separately. They first consider the simpler of the two functions, the denominator and assume that L_o is essentially unimodal. They expand L_o about the modal value θ_o, approximating it by $L_o(\theta_o)=L_o(\theta-\theta_o)'\Sigma_o L_o(\theta-\theta_o)$ where Σ_o is the inverse of the matrix with the negatives of the second partial derivatives of L_o with respect to θ as elements, evaluated at θ_o. They then make use of a result due to Laplace, described in detail by DeBruijn (1961). Known as the Laplace method for integrals, it is based on the premise that if one were to integrate over the entire real line for any reasonably unimodal distribution, most of the value of the integration will come from a relatively compact area around the distribution's modal value (Laplace, 1776). Stigler (1986) gives a translated and modernized account of Laplace's approach. The mathematical justification for applying Laplace's work to the T-K method is detailed by Kass, Tierney and Kadane (1990). Specifically, to integrate $\tau(x,t)$ over the real line with respect to x only, if t is large, one need only consider the value of the integration around the mode of $\tau(x,t)$ to get a reasonable approximation to the value of the entire integral. For example, (DeBruijn (1961)) consider the problem of integrating the function $\int \exp[-tx^2]\log(1+x+x^2)dx$. Large values of t cause the integral to be dispersed rather

closely around zero. If one were to do this integration for large t over the interval [-0.5,0.5], the result would be a reasonable approximation of the overall integral. Furthermore, if necessary, it may be possible to find a close approximation to the function required over the newly restricted area of integration which is easier to integrate. Applying this idea to the L_o integral produces

$$\int \exp[nL_o]d\theta \approx \exp[nL_o]\int \exp\left[-nL_o(\theta-\theta_o)'\sum_o^{-1}L_o(\theta-\theta_o)\right]d\theta \qquad (1.18)$$

This integral is in the form of a multivariate normal distribution, so that (1.18) becomes

$$\int \exp[nL_o]d\theta \approx \exp[nL_o]\left(\frac{2\pi}{n}\right)^{m/2}|\sum_o|^{1/2} \qquad (1.19)$$

Similarly, the numerator integral, evaluated at its maximization point θ_* becomes

$$\int \exp[nL_*(\theta_*)]d\theta \approx \exp[nL_*(\theta_*)]\left(\frac{2\pi}{n}\right)^{m/2}|\sum_*|^{1/2} \qquad (1.20)$$

Combining (1.19) and (1.20), the Tierney-Kadane approximation for the ratio of integrals (1.5) is found to be

$$u^*(\theta) \approx \left[\frac{|\sum_*|}{|\sum_o|}\right]^{1/2}\exp[n\{L_*(\theta_*)-L_o(\theta_o)\}] \qquad (1.21)$$

The error of this approximation is $O(n^{-2})$, the same degree of accuracy as that of Lindley's method. Furthermore, in estimating the posterior variance via the relation $Var[u(\theta)] = E\{[u(\theta)]^2\} - \{E[u(\theta)]\}^2$, the authors state that due to fortuitous cancellation of terms, the error of approximation is again of order $O(n^{-2})$ whereas Lindley's method has error $O(n^{-1})$ for estimating the posterior variance. They then state, however, that in practice they have encountered difficulties with numerical precision because the variance

calculation involves taking the difference of two very large numbers, which naturally could lead to the loss of significant digits. Furthermore the results are strictly asymptotic and they do not expect great accuracy in results for small sample size. They note that it is possible to obtain negative variance estimates or nonsingular matrices for small n.

The primary advantage of this approach over Lindley's is that one does not have to derive any third partial derivatives. However two sets of second partial derivatives must be derived for L_o and L_*. Furthermore one must separately maximize L_o and L_* to find the evaluation points θ_o and θ_* respectively. The authors suggest that the maximum likelihood estimators are an appropriate starting point for maximization since each of the functions is related to the likelihood function. Newton-Raphson iteration from that point has proven to be convenient with convergence usually appearing after one or two iterations. Clearly, for some distributional settings, this method will be superior to Lindley's in terms of numerical convenience, but not in general.

One of the major restrictions imposed on this technique is that the function $u(\theta)$, the parameter of interest, be nonnegative. The reason for this assumption is involved in the cancellation of error terms in approximating the numerator and denominator integrals. If $u(\theta)$ is nonnegative, then the two integrals are roughly of the same shape. If $u(\theta)$ can take on negative values, then the overall shape of the numerator integral may be markedly different from that of the denominator. Tierney and Kadane (1986) soften this restriction by stating that as long as the posterior distribution of $u(\theta)$ is "concentrated almost entirely" on the positive side of the origin, the approximation technique should be applicable. For the applications considered in this text, this assumption is met.

An updated version of the T-K method was presented in Tierney, Kass and

Kadane (1989) in order to address the restricted range space problem. They attack the problem by using the fact that the integrals involved in (1.5) are actually forms of moment-generating functions. As such, they can estimate the integrals by approximating the moment-generating function and differentiating to get means and variances.

This "new" method covers both "positive" and "negative" parameter spaces. The method is mathematically equivalent to simply adding a large constant to the function, using their previously published method and then subtracting the same constant from the result. Furthermore, the new method is seen to be equivalent to Lindley's method. Since this updated T-K method is redundant with Lindley's method, it will not be discussed separately here.

1.4.3 Relative Merits of the Two Procedures

Tierney and Kadane (1986) state that they proposed their approximation because "it would be useful to have approximations that are more accurate than the normal approximation, yet not as computationally intensive as numerical integration methods". They refer to Lindley's technique as "being accurate enough" but being rather tedious in application due to the required construction of the third partial derivatives of the log-likelihood function. In fact, for estimating posterior means (Bayes estimators), the two techniques have the same error $O(n^{-2})$ and it is only for posterior variance estimates that the T-K method becomes more accurate than that of Lindley.

A technical detail often omitted from published papers dealing with Bayes methods is de-emphasized by Tierney and Kadane (1986), although it has implications for use of their technique in practical situations. The T-K method retains the basic

structure of the Bayes problem - that being the ratio of two integrals. In practical situations this often involves taking the ratio of a very large number to a very small number so that the differences almost balance. Typically involving large powers, these calculations are prone to overflow/underflow problems as well as numerical accuracy considerations.

Although Tierney and Kadane mention that caution should be used if n is very large, they fail to address this problem directly. Lindley's method, however, does not suffer from this problem simply because of the way it attacks the ratio of integrals problem. Whereas T-K separately approximate the numerator and denominator, Lindley approaches the ratio as a whole and produces a single numerical result. The T-K method literally requires setting up two separate functions, L_o and L_*, and obtaining separate approximations for each. Lindley's approach is more convenient in that once the functional structure has been achieved it can be utilised repeatedly for various parameter functions without much additional work.

Due to asymptotic foundations, small sample sizes produce theoretical problems for both procedures. This in turn causes practical problems for producing variance estimates. In estimating $V(\theta)$, by estimating $E(\theta^2)$ and using the relation $V(\theta)=E(\theta^2)-[E(\theta)]^2$, there is a risk of negative variance estimates due to overestimating $E(\theta)$ and/or underestimating $E(\theta^2)$, both in Lindley's and the T-K method.

Howlader and Weiss (1987) report such problems and note that the T-K method evidences instability for small n. The reason for this is the same as that for the negative variance estimates. By separately estimating each part of the integral ratio, it incorporates the risk of over/underestimating one or both parts, thereby exaggerating the

error.

In comparing numerical ease, Tierney and Kadane only address the issue of computational ease. They state that their method is superior to Lindley's because it requires only second order partial derivatives of the likelihood function whereas Lindley's demands construction of the third order partials. This is correct but misleading. It is true that for an m-dimensional parameter one will need to construct $m(m+1)/2$ second derivatives and $m(m+1)(m+2)/6$ third partials to implement Lindley's method. What is omitted is that once constructed these partials become constants with respect to the parameter being estimated so that they need never be recalculated or modified. Changing the parameter of interest involves only a change in the $u(\theta)$ function and its partials $u_i(\theta)$, $u_{ij}(\theta)$. This means that once $u(\theta)=\theta_i$ has been estimated, for example, θ^2, $\log(\theta)$, etc. can be estimated by merely recalculating the partials of $u(\theta)$ and substituting them into the static Lindley expansion. The T-K method requires a separate reconstruction of the second partial derivatives for every parameter estimated. This is due to the fact that the method requires L_o and L_* functions be constructed, leading to the Hessian matrices Σ_o and Σ_* respectively. Although L_o remains constant across all estimators and therefore need only be constructed and maximized once, L_* changes with every change in $u(\theta)$. This induces a need for the reconstruction of the Σ_* matrix, involving $m(m+1)/2$ second partial derivatives.

These facts point to a general conclusion regarding the two techniques. If numerical maximization and partial differentiation routines are readily available, the T-K method will be just as simple to implement as Lindley's approximation. If, however, one wishes to produce closed form algebraic solutions, or perform estimation for numerous

functions of the parameter space, Lindley's method is both more convenient and efficient.

As originally proposed, the T-K method was applicable only to smooth positive functions on the parameter space. This problem does not really hamper the technique's applicability. The addition of a large constant to a negative parameter before approximation, followed by the subtraction of the same constant from the resultant estimate produces the desired quantity without violating the procedural assumptions.

Both methods do rely on the existence of the maximum likelihood estimator and the unimodality of the likelihood function. In applications where no unique global MLE exists, local maxima are usable. Regularity conditions have been discussed in Tierney and Kadane (1986) and Kass, Tierney & Kadane (1990), but largely assumed to be derived from the works of LeCam (1970), Johnson (1970) and Walker (1969). Collectively, these works state that the following regularity conditions are assumed for the asymptotic expansions involved in the approximations to be valid:

1) the posterior distribution is jointly measurable in $(\underline{x}, \underline{\theta})$ and has continuous partial derivatives of the first and second order with respect to $\underline{\theta}$.

2) the maximum likelihood estimate is strongly consistent and is asymptotically normal.

3) the prior density is continuous and nonnegative.

The main function of these assumptions is to make the second order partial derivatives of the likelihood function "behave sufficiently smoothly" (Walker, 1969) near the mode.

For the limiting distribution of the posterior to be asymptotically normal, Walker (1969) demands that the data be independent of the parameter space, which is not the case for several models that involve multiparameter settings (such as the three-parameter

Weibull). He then goes on to say that it is not strictly needed as long as the continuity conditions seen above are satisfied. He does state, however, that if θ is involved in the range space of the data, then another assumption typically fails. The assumption regarding the differentiability of the likelihood function is usually drawn into question. This problem is circumvented by use of local maximum likelihood estimators.

A final issue of comparison is of relevance mainly to statistical theorists. Although the T-K method is convenient with the aid of numerical maximization routines, it does not lend itself well to algebraic closed form solutions. Lindley's method on the other hand routinely produces closed form estimators that are simple modifications of the maximum likelihood estimator. Both methods produce estimators that are asymptotically indistinguishable from the MLE. Closed form solutions arrived at via the T-K approach tend to be more mathematically complex than those arrived at via Lindley's method. As will be seen in examples given later in the text, the T-K closed form typically involve functions raised to the nth power followed by taking the nth root of the result. This makes mathematical work with T-K closed forms inconvenient at best. Dealing with location parameters is especially difficult algebraically in the T-K approach.

In summary, the two methods provide reasonable alternatives depending upon the distributional setting and the parameters to be estimated. The T-K method was designed to be used in a numerical setting using readily available computing facilities. Although Lindley's method can also be applied under such a setting, it has clear advantages if mathematical work is desirable.

1.4.4 Other Approaches

A considerable amount of work has been done to complement the works by Lindley (1980) and Tierney & Kadane (1986) addressing the general problem of evaluating the inconvenient ratio of integrals. R. A. Johnson (1967 and 1970) and Naylor & Smith (1982, 1983) have presented an optional route by approximating the ratio by a multivariate normal distribution multiplied by a power series, which is generally expanded about the mle's. Johnson (1967) lays the groundwork and his 1970 work describes the strengths and weaknesses of the technique. When successful, the technique has compared favorably to other methods (Naylor & Smith 1982, 1983). Two major difficulties arise, however. The primary problem lies in the assumption that the posterior may be represented roughly by a normal distribution, which not only limits the scope of its application, but also may cause anomalous results. Secondly, Johnson (1970) states that the method is only applicable in situations where the mle's are strongly consistent and asymptotically normal. It is not known at this time what the weakest regularity conditions for the posterior are for the expansion to work nor how large a sample is actually sufficient to ensure a reasonable degree of integrity in the resultant estimates. Naylor and Smith have applied modified forms of this process with success in the biomedical field. They make a substantial contribution to solving many of the numerical problems encountered in the applications.

A summary of the approximation techniques, mainly involving numerical quadrature, has been given by Smith et al (1985). Not only do they describe each approximation technique, but they also detail some of the practical solutions crucial to the success of their implementation. Clearly, the entire area of integral ratio

approximations is presently a major topic in Bayesian inference and will lead to a wider application of the Bayesian paradigm. This will be especially true in the more complex mathematical settings where classical techniques involving the maximum likelihood principle encounter tractability problems that have been a roadblock of the Bayesian approach since its inception.

In recent years the computer has taken the foreground in Bayesian estimation theory. Various techniques involving computer-intensive techniques for estimating the posterior distributions have been put forward. Bagchi and Kadane (1991) categorize the alternate methods of attack by the type of numerical approximation used. Aside from methods already discussed, they cite Van Dijk, Hop and Louter (1987) as a good representation of methods involving Monte Carlo integration for approximating the posterior distribution. Finally, the most recent form of approximation makes use of Gibbs sampling as exemplified by Gelfand and Smith (1990).

1.5 Mathematical Intractability (Simulation Versus Derivation)

A good portion of the work in this dissertation relies on the results of Monte Carlo simulations. Wherever possible, algebraic results have been sought after and obtained, so that all-encompassing distributional properties may be explored. As has been discussed, however, it is the very nature of Bayesian methodology that makes such closed form analytic results the exception rather than the rule.

As a result the general thrust in Bayesian methodology has evolved towards an interactive, applied environment. Several works (notably Kass et al, 1988; Lindley, 1988; Smith et al, 1987) have addressed the basic issue, as presently developed Bayesian

methodology requires an interactive and powerful computing environment to be effective.

This fact has caused some authors to express concern over a movement away from theoretical derivation of analytic results towards too much reliance upon the computer. Keenly aware of this dichotomy in the literature, this dissertation attempts to strike a balance. Some of the work contained herein takes a purely analytic approach, others rely on computer simulation. In all segments, however, justification for and advantages of each approach are given. It is true that computers have been a boon to the amount of work that is attainable in statistical theory. It is not meant to replace the calculus.

1.6 Outline of the Dissertation

This Chapter lays out the philosophy behind the dissertation work. It may be considered a precis of the philosophical discussion that has surrounded the implementation of the Bayes paradigm since it was first suggested. The remaining chapters deal with a variety of multiparameter situations where implementation of the Bayesian paradigm is undertaken.

Chapters two, three and four present the application of the Lindley and Tierney-Kadane approximations to a variety of distributions most commonly used in reliability and lifetesting situations. Chapter two describes the merits and disadvantages of the two methods in application to the two-parameter normal and two-parameter inverse Gaussian distribution. Chapter three continues to detail the applicability of the approximations to the three most commonly used three-parameter distributions in reliability and lifetesting. The Weibull, gamma and lognormal distributions are studied. Chapter four gives a

similar presentation involving the five-parameter estimation task necessary in working with the bivariate normal distribution.

Chapters five and six discuss estimation in two of the most commonly seen mixture distributions. Both censored and uncensored sampling environments are explored. Chapter five details the work for a mixture of exponentials distributions while Chapter six deals with a mixture of Weibull distributions.

Chapter seven investigates two settings for Bayesian alternatives to traditional quality control techniques. The first section proposes alternatives for the traditional p-chart control limits by taking advantage of the inherent updating of information available through the Bayesian paradigm. The second portion presents a modification of the standard p-chart based on the predictive distribution.

Three papers concerning work contained in this dissertation have been published in refereed journals. The material published pertains to work involving the three-parameter Weibull distribution (section 3.1), the mixture of exponential distributions (Chapter five) and the mixture of Weibull distributions (Chapter six). Three other papers are in the refereeing process at the time of writing this dissertation. These papers pertain to the three-parameter gamma work (section 3.2), the work in Chapter four dealing with the bivariate normal and the predictive p-chart material (Chapter seven). The exact references for papers published at the time of this writing are noted at the beginning of the appropriate sections.

CHAPTER 2: BAYES APPROXIMATIONS FOR TWO-PARAMETER DISTRIBUTIONS

2.1 Introduction

This chapter examines two Bayesian approximation methods, due to Lindley (1980) and Tierney/Kadane (1986), for distributions commonly used for reliability and lifetesting situations involving two systemic parameters. These approximations are examined in relation to maximum likelihood estimators and, wherever possible, exact Bayes estimators. Closed form solutions for the approximations are explored.

As mentioned in Chapter 1, one of the basic difficulties encountered in implementing the Bayesian approach is evaluating the ratio of integrals involved in producing the Bayes estimator as the mean of the posterior distribution. Several authors have turned their attention to the problem, typically leading to an approximation for the exact Bayes estimator, often necessitating the use of computer maximization routines. This means that more work is generally involved than most other estimation techniques and, in a practical sense, loses its appeal. Evaluation of these approximation methods is undertaken in the hopes of developing, under various distributional settings, convenient Bayes estimators that perform reasonably well.

Little work has been done in applying the generalized approximation techniques to specific applications so that the algebraic manipulations required be kept to a minimum. Most often, in fact, the authors of such techniques recommend the construction of interactive computing environments to handle the approximations. This not only leaves the researcher at the mercy of the accuracy of computer technology and methodology employed, but also ignores the possibility that a closed-form solution may exist. If such a solution exists, it is of much greater value than a computational

approximation since distributional characteristics may be derived and studied.

2.2 Normal Distribution

The normal distribution $N(\mu,\sigma)$ with probability density function

$$f(x|\underline{\theta}) = f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad -\infty<\mu<\infty, \ \sigma>0$$

provides a convenient forum to test the approximations since the exact Bayes estimator is known to be (Sinha, 1986)

$$\mu^* = \overline{X} \quad and \quad \sigma^* = \frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{2}}$$

for the mean and standard deviation respectively. Sinha (1986) assumes a vague prior (Jeffreys, 1964) for $\underline{\theta}=(\mu,\sigma)$ of the form $\pi(\underline{\theta}) \propto \sigma^{-1}$.

This is an exemplary setting because of the inherent normality of functions of variables involving the normal distribution. Both Lindley and Tierney/Kadane used normal distribution theory to develop their approximations. If the approximations perform poorly under these most favourable conditions, it would suggest that difficulties would most likely arise in their applicability to more complex settings.

2.2.1 Lindley's Approximation

As described in Chapter 1, Lindley's approximation to the Bayes estimator is

$$u^*(\underline{\theta}) = u(\hat{\underline{\theta}}) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\left[u_{ij}(\hat{\underline{\theta}}) + 2u_i(\hat{\underline{\theta}})\right]\sigma_{ij} + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{\ell=1}^{p}L_{ijk}u_{\ell}\sigma_{ij}\sigma_{k\ell}$$

where

p is the number of parameters to be estimated

$\underline{\theta}$ is the parameter vector, evaluated at the maximum likelihood estimator

u_i, u_{ij} are the first and second partial derivatives of the parameter function being estimated

ρ_j is the first partial derivative of the log-prior

σ_{ij} is the ijth element of the variance-covariance matrix

L_{ijk} is the third partial derivative of the log-likelihood function.

Now for estimating $\underline{\theta}=(\mu,\sigma)$ from the normal distribution

$$f(x|\underline{\theta}) = \frac{1}{\sqrt{2\Pi}\sigma}\exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right] \qquad -\infty < (X,\mu) < \infty, \quad \sigma > 0$$

it can be seen that

$$u_i = \begin{bmatrix} 1 & \textit{if index matches} \\ 0 & \textit{otherwise} \end{bmatrix} \qquad \textit{and} \quad u_{ij}=0, \forall_{ij}$$

and calculate

$$\rho(\underline{\theta}) = -\log\sigma, \quad \rho_1(\underline{\theta}) = \frac{\partial\rho(\underline{\theta})}{\partial\mu} = 0, \quad \rho_2(\underline{\theta}) = \frac{\partial\rho(\underline{\theta})}{\partial\sigma} = -\frac{1}{\sigma^2}$$

It is further found that

$$L(\underline{\theta}|\underline{X}) = C - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

so that the partial derivatives are found as follows

$$L_1 = \frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{I=1}^{n}(X_i-\mu)$$

$$L_2 = \frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$L_{11} = -\frac{n}{\sigma^2}$$

$$L_{12} = -\frac{2}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)=0, \quad \text{evaluted at the mle } \underline{\theta} = \left[\overline{X}, \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]$$

$$L_{22} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{2n}{\sigma^2}, \quad \text{at the mle}$$

and the variance-covariance matrix, evaluated at the mle is

$$\sigma_{ij} = [-L_{ij}]^{-1} = \begin{bmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{2n}{\sigma^2} \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{\sigma^2}{2n} \end{bmatrix} .$$

The third partial derivatives are

$$L_{111} = 0$$

$$L_{112} = \frac{2n}{\sigma^3}$$

$$L_{122} = \frac{6}{\sigma^4} \sum_{i=1}^{n} (X_i - \mu) = 0 \quad \text{at the mle}$$

$$L_{222} = -\frac{2n}{\sigma^3} + \frac{12}{\sigma^5} \sum_{i=1}^{n} (X_i - \mu)^2 = \frac{10n}{\sigma^3} \quad \text{at the mle.}$$

Substituting the zero elements into Lindley's approximation formula, the equation reduces

to

$$u^*(\underline{\theta}) = u(\underline{\hat{\theta}}) + \sum_{i=1}^{2} \sum_{j=1}^{2} u_i(\underline{\hat{\theta}}) \, \rho_j(\underline{\hat{\theta}}) \, \sigma_{ij} + \frac{1}{2} L_{112} u_2 \sigma_{11} \sigma_{22} + \frac{1}{2} L_{222} u_2 \sigma_{22}^2 \, .$$

For estimating the normal mean μ, set $u(\underline{\theta}) = \mu$. It is found that $u_1(\underline{\theta}) = 1$, $u_2(\underline{\theta}) = 0$.

Hence all terms in the above formula have a zero element so that

$$u^*(\underline{\theta}) = u(\underline{\hat{\theta}}) + 0 = u(\underline{\hat{\theta}}) \, .$$

So Lindley's approximation for the normal mean parameter turns out to be exactly equal

to the true Bayes estimator and the mle.

Now, for the standard deviation, set $u(\underline{\theta}) = \sigma$ to get $u_1(\underline{\theta}) = 0$, $u_2(\underline{\theta}) = 1$.

Combined with other zero elements, Lindley's approximation equation reduces to

$$u^*(\underline{\theta}) = \hat{\sigma} + \rho_2 \sigma_{22} + \frac{1}{2} L_{112} \sigma_{11} \sigma_{22} + \frac{1}{2} L_{222} \sigma_{22}^2$$

and substituting from the above calculations, Lindley's approximation to the Bayes estimator of σ turns out to be

$$u^*(\underline{\theta}) = \hat{\sigma} - \frac{\hat{\sigma}}{2n} + \frac{\hat{\sigma}}{2n} + \frac{5\hat{\sigma}}{4n}$$

$$= \hat{\sigma}\left(1 + \frac{5}{4n}\right)$$

So after an amount of tedious algebra, an easily calculable closed form for Lindley's approximation to the Bayes estimator of σ is produced. From a classical theory perspective the relationship can be examined between the posterior variance for Lindley's approximation and that of the mle. Here

$$Var(u^*(\underline{\theta})) = Var\left[\hat{\sigma}\left(1 + \frac{5}{4n}\right)\right] = \left[1 + \frac{5}{4n}\right]^2 Var(\hat{\sigma}) \ .$$

Since the constant is clearly greater than one for all values of n, it is true that $Var[u^*(\underline{\theta})] > Var(\hat{\sigma})$.

2.2.2 Tierney-Kadane Approximation

As described in Chapter one, the Tierney-Kadane approximation procedure can be detailed as a five-point process as follows:

1. Construct $L(\underline{\theta}|\underline{x})$, the log-likelihood function of the pdf under study. Define $\pi(\underline{\theta})$ to be the prior distribution of the multivariate parameter $\underline{\theta}$ and

    $u(\underline{\theta})$ to be a function of the multivariate parameter to estimate.

2. Construct two modified log-likelihood functions L_0 and L_* where

    $L_0 = (\log[\pi(\underline{\theta})] + L)/n$ and

    $L_* = (\log[u(\underline{\theta})] + \log[\pi(\underline{\theta})] + L)/n = L_0 + \log[u(\underline{\theta})]/n.$

3. Find points $\underline{\theta}_o$ and $\underline{\theta}_*$ which maximize L_o and L_*.

4. Construct two matrices, Σ_o and Σ_*, the negative of the inverse of the second order partial derivatives of L_o and L_* respectively, and evaluated at the maximization points of $\underline{\theta}_o$ and $\underline{\theta}_*$ respectively.

5. Calculate the estimator by the formula

$$u^*(\underline{\theta}) = \sqrt{\frac{|\Sigma_*|}{|\Sigma_o|}} \cdot \exp\{n[L_*(\underline{\theta}_*) - L_o(\underline{\theta}_o)]\}$$

Once again, consider the normal distribution and follow the five-step algorithm.

1. The log-likelihood function and prior distribution are as above.

2. Construct the modified maximum-likelihood functions L_o and L_* as

$$L_o = \frac{1}{n}\log\left(\frac{1}{\sigma}\right) - \log\sigma - \frac{1}{2\sigma^2 n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$= -\left(1 + \frac{1}{n}\right)\log\sigma - \frac{1}{2n\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

and L_* will differ for each of the parameters μ and σ. As such the discussion will be split into two parts.

First, for estimating σ the population standard deviation, construct L_*

$$L_* = \frac{1}{n}\log u(\underline{\theta}) + L_o = -\log\sigma - \frac{1}{2n\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Next, the points $\underline{\theta}_o$ and $\underline{\theta}_*$ that maximize L_o and L_* respectively must be found. The partial derivatives of the functions are

$$\frac{\partial L_o}{\partial \mu} = \sum_{i=1}^{n} \frac{(X_i - \mu)}{n\sigma^2} \qquad (2.2)$$

$$\frac{\partial L_o}{\partial \sigma} = -\frac{\left(1 + \frac{1}{n}\right)}{\sigma} + \frac{1}{n\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2 \qquad (2.3)$$

Setting these expressions equal to zero and solving finds $\underline{\theta}_o$ to be

$$\underline{\theta}_o = (\mu_o, \sigma_o^2) = \left[\overline{X}, \frac{1}{n+1} \sum_{i=1}^{n} (X_i - \overline{X})^2\right] .$$

Similarly for L_*, the partial derivatives are

$$\frac{\partial L_*}{\partial \mu} = \sum_{i=1}^{n} \frac{(X_i - \mu)}{n\sigma^2}$$

$$\frac{\partial L_*}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{n\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2$$

and solving the system of equations produces the maximizing point

$$\underline{\theta}_* = (\mu_*, \sigma_*^2) = \left[\overline{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_*)^2\right] .$$

Step 4 requires Σ_o and Σ_*, the negative inverse of the matrix of second derivatives, to be evaluated at the two maximization points $\underline{\theta}_o$ and $\underline{\theta}_*$ respectively.

Using (2.2) and (2.3) produces

$$\frac{\partial^2 L_o}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

$$\frac{\partial^2 L_o}{\partial \sigma^2} = \frac{\left(1+\frac{1}{n}\right)}{\sigma^2} - \frac{3}{n\sigma^4}\sum_{i=1}^{n} (X_i-\mu)^2$$

$$\frac{\partial^2 L_o}{\partial \mu \partial \sigma} = -\frac{2\sum_{i=1}^{n} (X_i-\mu)}{n\sigma^3}$$

The second partials are then evaluated at the maximization point by substituting μ_o and σ_o for μ and σ respectively so that

$$\frac{\partial^2 L_o}{\partial \sigma^2} = \frac{\left(1+\frac{1}{n}\right)}{\sigma_o^2} - \frac{3(n+1)}{n\sigma^4}\sigma_o^2 = -\frac{2\left(1+\frac{1}{n}\right)}{\sigma_o^2}$$

$$\frac{\partial^2 L_o}{\partial \mu \partial \sigma} = 0.$$

Hence

$$\Sigma_o = -\begin{bmatrix} -\dfrac{1}{\sigma_o^2} & 0 \\ 0 & -\dfrac{2\left(1+\frac{1}{n}\right)}{\sigma_o^2} \end{bmatrix}^{-1} = \begin{bmatrix} \sigma_o^2 & 0 \\ 0 & \dfrac{\sigma_o^2}{2\left(1+\frac{1}{n}\right)} \end{bmatrix}$$

The matrix construction process is then repeated for Σ_*. The second partials of L_* are

$$\frac{\partial^2 L_*}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

$$\frac{\partial^2 L_*}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{3}{n\sigma^4} \sum_{i=1}^{n} (X_i - \mu)^2$$

$$\frac{\partial^2 L_*}{\partial \mu \partial \sigma} = -\frac{2}{n\sigma^3} \sum_{i=1}^{n} (X_i - \mu)$$

and the maximization points μ_* and σ_* are substituted so that

$$\frac{\partial^2 L_*}{\partial \sigma^2} = \frac{2}{\sigma_*^2}$$

$$\frac{\partial^2 L_*}{\partial \mu \partial \sigma} = 0 .$$

Finally step 5 is reached where the pieces are put together to evaluate

$$u^*(\underline{\theta}) = \sqrt{\frac{|\Sigma_*|}{|\Sigma_o|}} \exp\{n[L_*(\underline{\theta}_*) - L_o(\underline{\theta}_o)]\} \qquad (2.4)$$

The relation $\sigma_*^2 = \left(1 + \frac{1}{n}\right)\sigma_o^2$ is used to evaluate the first part of (2.4)

$$\sqrt{\frac{|\Sigma_*|}{|\Sigma_o|}} = \sqrt{\frac{\frac{\sigma_*^4}{2}}{\frac{\sigma_o^4}{2\left(1+\frac{1}{n}\right)}}} = \sqrt{\frac{\sigma_o^2\left(1+\frac{1}{n}\right)^3}{\sigma_o^2}} = \left(1+\frac{1}{n}\right)^{3/2}$$

The second part of (2.4) can be evaluated as

$$\exp\{n[L_*(\underline{\theta}_*) - L_o(\underline{\theta}_o)]\} = \left\{ n \left[-\log\sigma_* - \frac{1}{2n\sigma_*^2} \sum_{i=1}^{n} (X_i - \mu_*)^2 \left(1 + \frac{1}{n}\right)\log\sigma_o \right. \right.$$

$$\left. \left. + \frac{1}{2n\sigma_o^2} \sum_{i=1}^{n} (X_i - \mu_o)^2 \right] \right\}$$

Using the relation above and the fact that $\mu_* = \mu_o$, it is found that

$$\exp\{n[L_*(\underline{\theta}_*) - L_o(\underline{\theta}_o)]\} = \exp\left\{ n\left[\frac{1}{2n} - \log\sqrt{1 + \frac{1}{n}} + \frac{1}{n}\log\sigma_o \right] \right\}$$

$$= \exp\left[\frac{1}{2} - n\log\sqrt{1 + \frac{1}{n}} + \log\theta_o \right]$$

$$= \sigma_o \sqrt{\frac{e}{\left(1 + \frac{1}{n}\right)^n}}$$

Combining the two parts of (2.4), a closed form solution for $\sigma_{T\text{-}K}^*$ the Tierney-Kadane approximation to the Bayes estimator for the population standard deviation is produced to be

$$\sigma_{T\text{-}K}^* = \sqrt{\left(1 + \frac{1}{n}\right)^3 \frac{e}{(n+1)\left(1 + \frac{1}{n}\right)^n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\sigma_{T\text{-}K}^* = \sqrt{\frac{e}{(n+1)\left(1 + \frac{1}{n}\right)^{n-3}} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Further, write

$$\sigma_{T\text{-}K}^{*} = \sqrt{\left(\frac{n}{n+1}\right)\frac{e}{\left(1+\frac{1}{n}\right)^{n-3}}\hat{\sigma}^{2}}$$

and since $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n} = e$, asymptotically $\sigma_{T\text{-}K}$ tends to the maximum likelihood

estimator, as does Lindley's approximation.

Because a closed form is obtained, the variance properties of the approximation may be

examined to find

$$Var\left(\sigma_{T\text{-}K}^{*}\right) = \left[e\left(\frac{n}{n+1}\right)^{n+1}\right]Var(\hat{\sigma})$$

and since the constant is clearly less than one for all values of the sample size n, one can

see that the variance of the Tierney-Kadane approximation is less than that of the mle.

Hence

$$Var(\sigma_{T\text{-}K}^{*}) \le Var(\hat{\sigma}) < Var(\sigma_{Lindley}^{*})$$

Now, returning to the population mean parameter, set $u(\underline{\theta})=\mu$ and begin by

creating L_{o} and L_{*} again. The first function L_{o} is the same regardless of the parameter

being estimated so that the maximization point θ_{o} and matrix Σ_{o} will be the same. The

L_{*} function does change for each different parameter, and is found here to be

$$L_{*} = \frac{1}{n}\log\mu - \left(1+\frac{1}{n}\right)\log\sigma - \frac{1}{2n\sigma^{2}}\sum_{i=1}^{n}(X_{i}-\mu)^{2}$$

Take partial derivatives of L_* to find the maximization point θ_* and obtain

$$\frac{\partial L_*}{\partial \mu} = \frac{1}{n\mu} + \frac{1}{n\sigma^2} \sum_{i=1}^{n} (X_i - \mu)$$

$$\frac{\partial L_*}{\partial \sigma} = -\frac{\left(1 + \frac{1}{n}\right)}{\sigma} + \frac{1}{n\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2$$

Producing a closed-form solution for the maximization point θ_* from the above system of equations requires use of the quadratic formula to obtain the roots of $ax^2 + bx + c = 0$. This, it turns out produces a very messy result. Substituting this result into Σ_* and attempting to solve the Tierney-Kadane approximation equation becomes intractable. This is the first of several distributional applications where this phenomenon appeared. Typically the Tierney-Kadane approximation works out algebraically in a fairly convenient fashion, as seen above, when working with scale parameters not involving the sample space. When applying the technique for a threshold or location parameter, however, the addition of the log term involving the threshold parameter to L_* prevents the convenient cancelling of terms when the partial derivatives are taken. When working with scale parameters, L_* typically involves other terms with the scale parameter in the denominator, as seen above. Thus, when partial derivatives are taken, convenient collecting of terms and cancellation are possible. Such is not the case with a location parameter because it produces an L_* equation as above with terms involving the threshold parameter alternately in the numerator and denominator, effectively ruling out the possibility of a convenient closed-form solution. Numerical optimization routines become necessary to achieve the Tierney-Kadane approximation when dealing with such threshold parameters on a case-by-case basis. This fact holds important implications for the

comparative success in the application of the two approximation techniques.

2.2.3 Example

Using data from an N(μ,σ) example in Sinha (1986b), the two approximation methods are examined and compared against the exact Bayes estimator. Sinha (1986b) also gave forms for the estimated posterior variance of the exact Bayes estimator. For the approximations, as well as the exact Bayes estimator, an estimate of the posterior variance can be obtained by use of the relation

$$Var[u(\underline{\theta}|\underline{X})]=E[u^2(\underline{\theta}|\underline{X})]-\left[E[u(\underline{\theta}|\underline{X}]\right]^2$$

Estimating each part of the variance equation produces an estimate for the estimator's variance. Performing estimation in this manner does bring into the possibility of negative variance estimates. Such estimates are typically observed in small sample cases. These estimators are useful for comparing the relative precision of the various techniques. The Sinha (1986b) sample dealt with 30 component lifetimes. Maximum likelihood estimates were found to be $\hat{\mu}=\bar{X}=19.978$ and $\hat{\sigma}=5.588$. The Bayes methods results are in Table 2.1.

Table 2.1: Bayes Estimation Results for Sinha (1986b) Sample

Parameter	Exact Bayes	$\hat{V}$(Exact)	Lindley	$\hat{V}$(Lindley)	Tierney -Kadane	$\hat{V}$(T-K)
μ	19.978	1.160	19.978	1.160	19.974	1.159
σ	5.836	2.564	5.821	2.782	5.822	2.755

The example produces results that seem to favour neither of the approximation

techniques in terms of closeness to the exact Bayes estimator. The Tierney-Kadane estimator does exhibit smaller estimated variance, however. The results for σ are more interesting than those for μ since the Lindley approximation is algebraically equivalent to the exact Bayes estimator.

2.2.4 Monte Carlo Simulation

To assess the distributional assumptions, Monte Carlo simulations were run for both small and moderate sample sizes. Again recall that results for μ will not be as interesting as those for σ, since Lindley's method produces the exact Bayes estimator. For this reason, the results for the scale parameter only are given.

Two runs were performed. First 5000 samples of size 10 from N(50,3) were generated, followed by a run of 1000 samples of size 100, both with parameter settings of $\underline{\theta} = (\mu, \sigma) = (50.0, 3.0)$. The mean value for the simulation sample estimates and mean square error are given in Table 2.2 for the three Bayes estimators as well as the maximum likelihood estimator.

Table 2.2: Simulation Results for Sampling from N(μ, σ)=N(50,3)

Estimator	n=10 5000 samples		n=100 1000 samples	
	Mean	MSE	Mean	MSE
Exact Bayes	3.2040	.6251	3.0177	.0468
Lindley	3.1329	.5698	3.0177	.0467
Tierney	3.1347	.5718	3.0177	.0467
MLE	2.7837	.4833	2.9804	.0457

The first notable result is that sample size has a profound impact on results. This

is not surprising as the authors of both approximation techniques state their methods rely on asymptotic results. Neither technique has been adequately explored for implications of small sample size. From Table 2.2 it is clear that for a moderate sample size, there is little to choose between the Bayes approximations. Both the Lindley and T-K method produce results identical to that of the exact Bayes estimator. The Bayes estimators are certainly competitive alternatives to the mle, especially considering that the work is performed in a noninformative prior situation.

Results for the small sample size run reveal some differences between the approximations. Lindley's approximation reveals a slightly less accurate estimator on average, with marginally better mse than the T-K estimator. Both approximations still do perform remarkably well considering the sparse sample information.

Figures 2.1 through 2.4 provide the sampling distributions of the four estimators for the smaller sample size run. As can be seen they are almost identical. Also notable is the bell-shaped consistency. Given that all the estimators are merely a constant multiplier of the mle, which is known to follow a normal distribution, one would expect this. The Shapiro-Wilk test of normality was applied to all these distributions and produced a nonrejection of the normality hypothesis with a typical p-value of 0.80.

The fact that there is little to choose between the approximations in this very simplistic setting has implications for choice of technique. The Lindley method would have to be preferred for its algebraic simplicity over the Tierney-Kadane approach. Lindley's method produces closed-form results for both parameters whereas the T-K method becomes intractable in working with the location parameter μ. Even in the case of estimating σ, the form of the Lindley approximation is much simpler than that of the

Figure 2.1   **ESTIMATION FOR NORMAL DISTRIBUTION STANDARD DEVIATION**
RESULTS FOR 1000 GENERATED SAMPLES
N=10, μ=50, σ=3.0

PERCENTAGE

MAXIMUM LIKELIHOOD ESTIMATOR
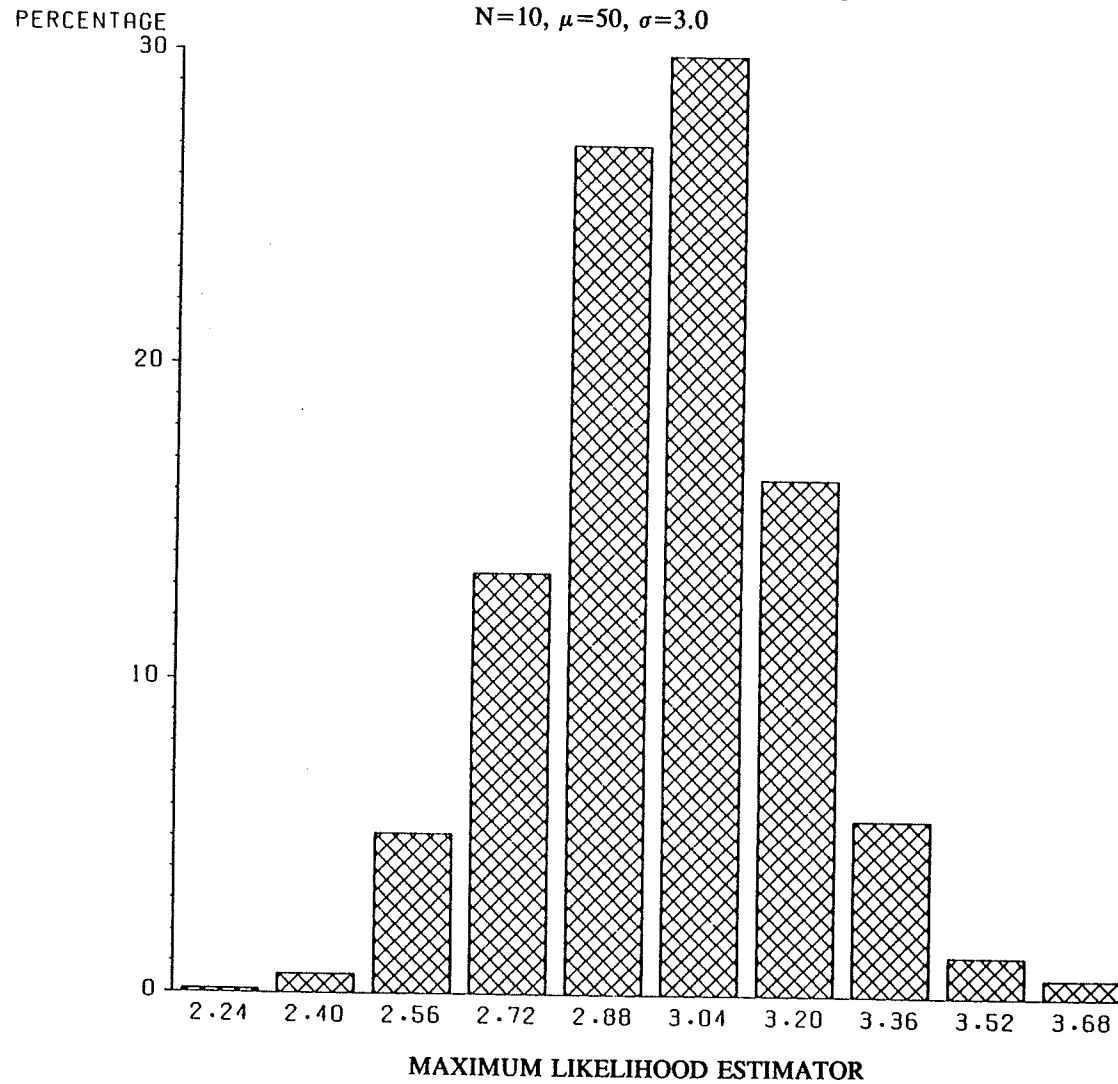
Figure 2.2   **ESTIMATION FOR NORMAL DISTRIBUTION STANDARD DEVIATION**
RESULTS FOR 1000 GENERATED SAMPLES
N=10, μ=50, σ=3.0
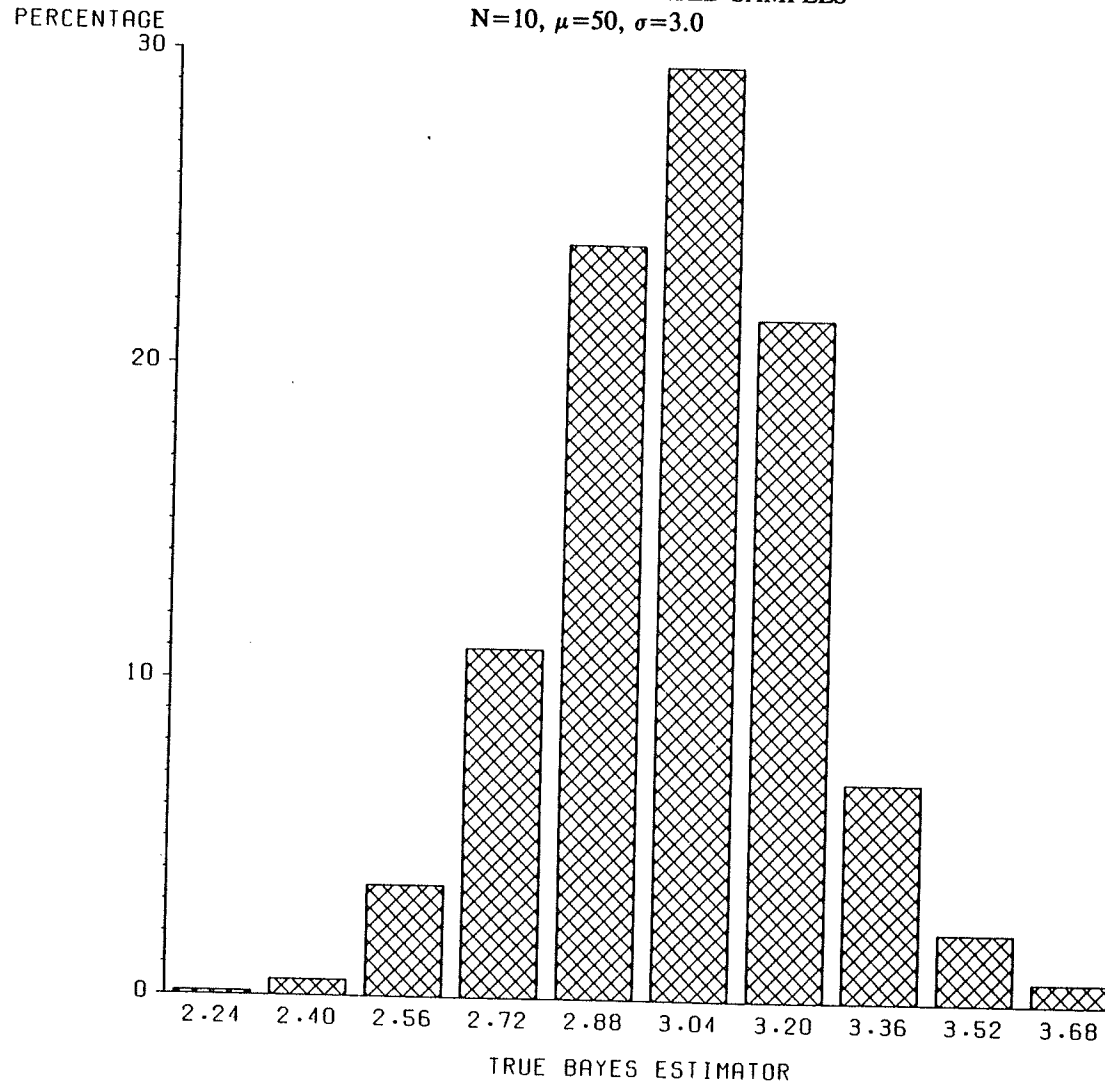
Figure 2.3

ESTIMATION FOR NORMAL DISTRIBUTION STANDARD DEVIATION
RESULTS FOR 1000 GENERATED SAMPLES
$N=10$, $\mu=50$, $\sigma=3.0$

PERCENTAGE

LINDLEY ESTIMATOR
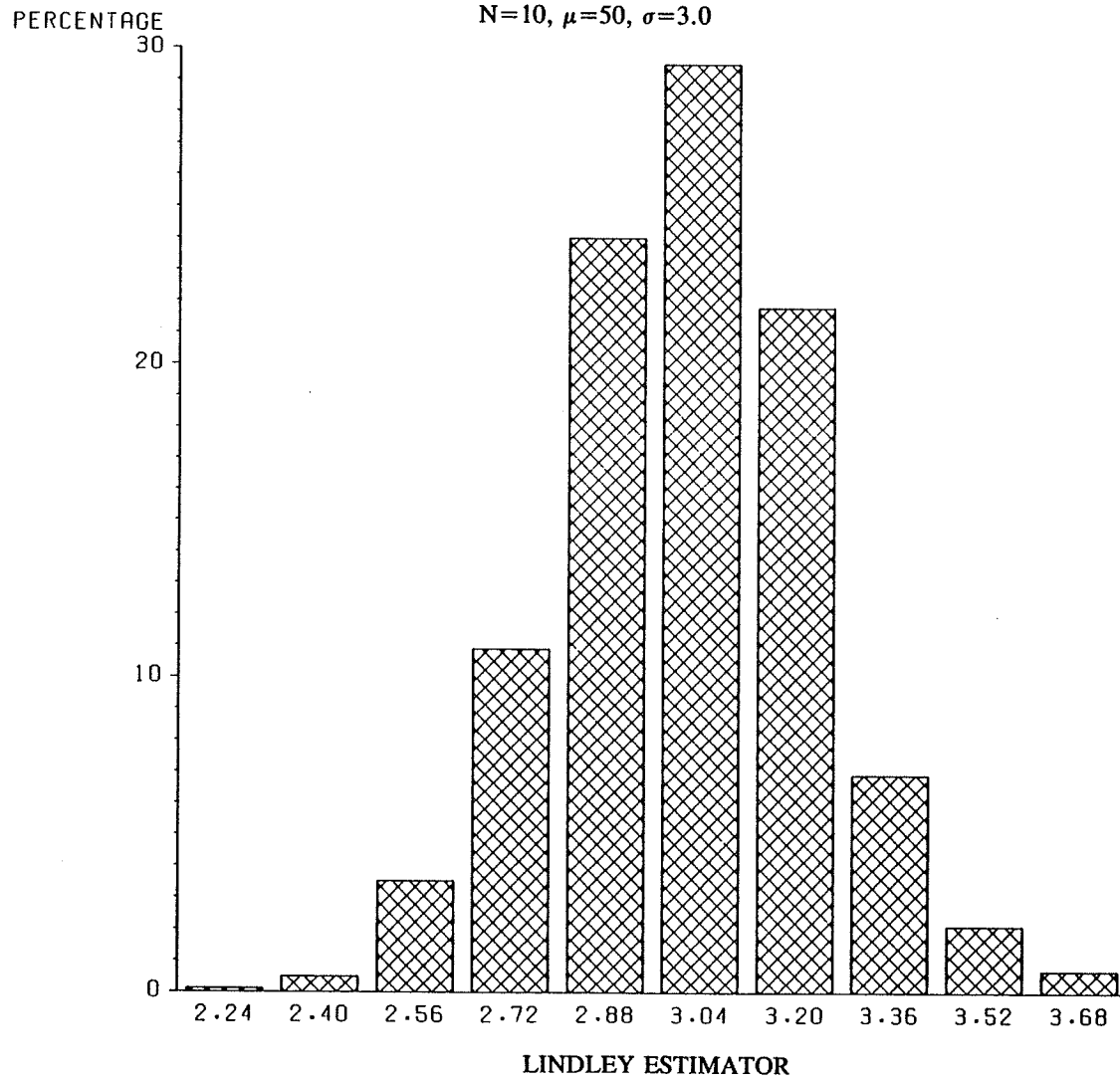
Figure 2.4

ESTIMATION FOR NORMAL DISTRIBUTION STANDARD DEVIATION

RESULTS FOR 1000 GENERATED SAMPLES

$N=10$, $\mu=50$, $\sigma=3.0$

TIERNEY-KADANE ESTIMATOR
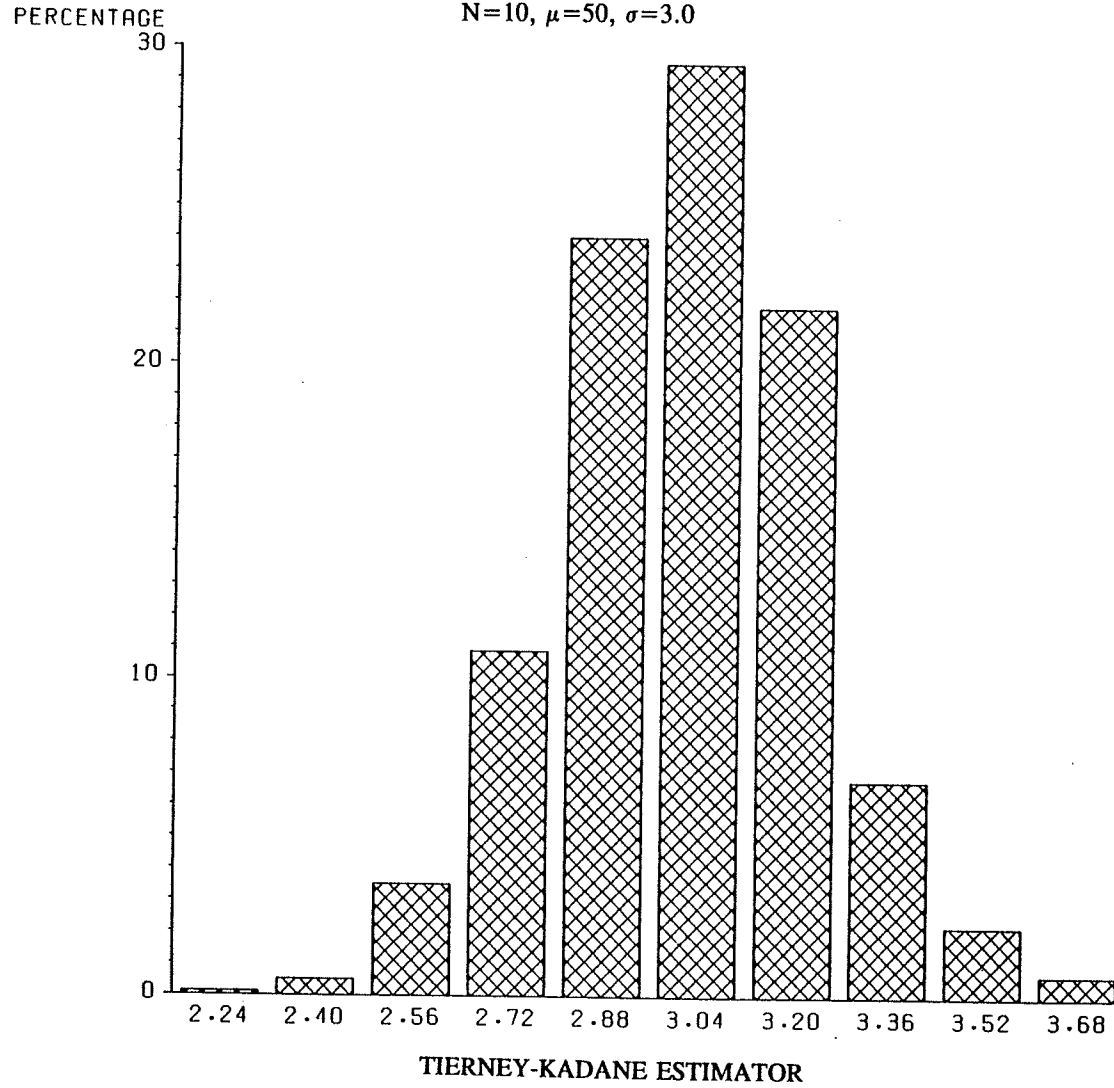
T-K technique. The T-K technique will always involve some form of the constant e raised to a power, divided by a cancelling constant. This produces comparable results to the Lindley method, albeit by a roundabout algebraic manipulation. It seems clear that the T-K method is better suited to computer application environments while the Lindley method lends itself more readily to algebraic investigation. As will be seen, these basic findings will persist through a variety of distributional settings.

2.3 Inverse Gaussian Distribution

A close relative of the normal distribution, the inverse Gaussian distribution is often used in reliability and lifetime studies. The same pattern of analysis is followed as with the normal distribution, making comparisons among the two approximations, the exact Bayes estimator and the maximum likelihood estimator.

Consider a distribution for the lifetime X of a component which arises from the two-parameter inverse Gaussian IG(μ,λ) which is defined by

$$f(x \mid \lambda, \mu) = f(x \mid \underline{\theta}) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left[\frac{-\lambda(x-\mu)^2}{2x\mu^2}\right] \quad x, \mu, \lambda > 0 \ .$$

The maximum likelihood estimators of the parameters are well known as

$$\hat{\mu} = \overline{X}, \quad \hat{\lambda} = \frac{n}{\displaystyle\sum_{i=1}^{n}\left[\frac{1}{X_i} - \frac{1}{\overline{X}}\right]} \quad .$$

The vector parameter $\underline{\theta} = (\mu, \lambda)$ is assumed to follow a diffuse prior distribution (Box and

Tiao, 1973) so that

$$\pi(\underline{\theta}) \propto \frac{1}{\lambda}, so\ that\ \rho(\underline{\theta}) = -\log\lambda \quad .$$

The same procedures can be carried out as was done with the normal distribution in section 2.

2.3.1 Lindley's Approximation

Sinha (1986) produced closed forms for Lindley's approximation under the inverse Gaussian distribution following the same process seen in section 2.2.1, finding

$$\mu_{Lindley}^{\bullet} = \hat{\mu} + \frac{3\hat{\mu}^2}{n\hat{\lambda}}, \quad \lambda_{Lindley}^{\bullet} = \left(\frac{n-1}{n}\right)\hat{\lambda} \quad .$$

The variance of the Lindley approximations for both parameters is smaller than that of the mle although for moderate to large sample sizes the differences tend toward zero. Similar to the closed-form solution for the normal distribution, Lindley's approximation produces easily calculable estimators that are simple functions of the maximum likelihood estimator.

2.3.2 Tierney-Kadane Approximation

Following the five step algorithm described in section two, construct the log-likelihood function

$$L(\underline{\theta}|\underline{X}) = C + \frac{n}{2}\log\lambda - \frac{\lambda}{2\mu^2}\left[n\bar{X} - 2n\mu + \mu^2\sum_{i=1}^{n}\frac{1}{X_i}\right]$$

where C is a normalizing constant, irrelevant to calculations.

Further,

$$L_o = (\log[\pi(\underline{\theta}]+L)/n = C+\left(\frac{1}{2}-\frac{1}{n}\right)\log\lambda-\frac{\lambda}{2\mu^2}\left[\overline{X}-2\mu+\frac{\mu^2}{n}\sum_{i=1}^{n}\frac{1}{X_i}\right]$$

and the partial derivatives of L_o are

$$\frac{\partial L_o}{\partial \mu} = \frac{\lambda}{\mu^3}(\overline{X}-\mu)$$

$$\frac{\partial L_o}{\partial \lambda} = \frac{\left(\frac{1}{2}-\frac{1}{n}\right)}{\lambda}-\frac{1}{2\mu^2}\left[\overline{X}-2\mu+\frac{\mu^2}{n}\sum_{i=1}^{n}\frac{1}{X_i}\right] \quad .$$

The maximizing point $\underline{\theta}_o$ is obtained by setting the partials to zero and solving so that

$$\underline{\theta}_o = (\mu_o,\lambda_o) = \left[\overline{X}, \frac{2\overline{X}\left[\frac{1}{2}-\frac{1}{n}\right]}{-1+\frac{\overline{X}}{n}\sum_{i=1}^{n}\frac{1}{X_i}}\right] \quad .$$

The L. function changes for each parameter function to be estimated so that for $u(\underline{\theta})=\lambda$, it is found to be

$$L_* = C+\frac{1}{2}\log\lambda-\frac{\lambda}{2\mu^2}\left[\overline{X}-2\mu+\frac{\mu^2}{n}\sum_{i=1}^{n}\frac{1}{X_i}\right]$$

$$\frac{\partial L_*}{\partial \mu} = \frac{\lambda}{\mu^3}(\overline{X}-\mu)$$

$$\frac{\partial L_*}{\partial \lambda} = \frac{1}{2\lambda}-\frac{1}{2\mu^2}\left[\overline{X}-2\mu+\frac{\mu}{n}\sum_{i=1}^{n}\frac{1}{X_i}\right]$$

producing the result

$$\underline{\theta}_* = \left[\overline{X}, \ \frac{\overline{X}}{-1+\dfrac{\overline{X}}{n}\displaystyle\sum_{i=1}^{n}\dfrac{1}{X_i}} \right] .$$

Next construct the two matrices Σ_o and Σ_*, comprised of partial derivatives of the L_o and L_* functions respectively, evaluated at the maximization points $\underline{\theta}_o$ and $\underline{\theta}_*$. The partials of L_o are

$$\frac{\partial^2 L_o}{\partial \mu^2} = \frac{\lambda}{\mu^4}(2\mu - 3\overline{X}) = -\frac{\lambda}{\overline{X}^3} \ since \ \mu_o = \overline{X}$$

$$\frac{\partial^2 L_o}{\partial \lambda^2} = -\frac{\left(\dfrac{1}{2}-\dfrac{1}{n}\right)}{\lambda^2}$$

$$\frac{\partial^2 L_o}{\partial \mu \partial \lambda} = \frac{1}{\mu^3}(\overline{X}-\mu) = 0 \ since \ \mu_o = \overline{X}$$

so that Σ_o is

$$\Sigma_o = \begin{bmatrix} \dfrac{-\lambda_o}{\overline{X}^3} & 0 \\[3ex] 0 & \dfrac{-\left(\dfrac{1}{2}-\dfrac{1}{n}\right)}{\lambda_0^2} \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{\overline{X}^3}{\lambda_o} & 0 \\[3ex] 0 & \dfrac{\lambda_o^2}{\left(\dfrac{1}{2}-\dfrac{1}{n}\right)} \end{bmatrix} .$$

Similarly, for Σ_o the second partial derivatives evaluated at the maxima are

$$\frac{\partial^2 L_*}{\partial \mu^2} = -\frac{\lambda}{\overline{X}^3}$$

$$\frac{\partial^2 L_*}{\partial \lambda^2} = -\frac{1}{2\lambda^2}$$

$$\frac{\partial^2 L_*}{\partial \mu \partial \lambda} = 0$$

so that Σ_* is

$$\Sigma_* = -\begin{bmatrix} -\dfrac{\lambda}{\overline{X}^3} & 0 \\ 0 & -\dfrac{1}{2\lambda^2} \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{\overline{X}^3}{\lambda} & 0 \\ 0 & 2\lambda^2 \end{bmatrix} \ .$$

Evaluate the Tierney-Kadane approximation formula (2.4) in two pieces as before so that the first part is

$$\sqrt{\dfrac{|\Sigma_*|}{|\Sigma_o|}} = \sqrt{\dfrac{2\lambda_*\overline{X}^3}{\dfrac{\lambda_o\overline{X}^3}{\left(\dfrac{1}{2}-\dfrac{1}{n}\right)}}} = \sqrt{\dfrac{2\left(\dfrac{1}{2}-\dfrac{1}{n}\right)\lambda_*}{\lambda_o}}$$

but it can be seen that $\lambda_o = 2\left(\dfrac{1}{2}-\dfrac{1}{n}\right)\lambda_*$, so the first part becomes

$$\sqrt{\dfrac{|\Sigma_*|}{|\Sigma_o|}} = \sqrt{\dfrac{\lambda_o}{\lambda_o}} = 1 \ .$$

So the Tierney-Kadane approximation for the Bayes estimator of λ is

$$\lambda_{T\text{-}K}^* = 1 \cdot \exp\left\{n\left[L_*(\underline{\theta}_*) - L_o(\underline{\theta}_o)\right]\right\}$$

which becomes

$$\lambda_{T\text{-}K}^* = \exp\left[C + \dfrac{1}{2}\log\lambda_* - \dfrac{\lambda_*}{2\overline{X}}\left(-1 + \dfrac{\overline{X}}{n}\sum_{i=1}^{n}\dfrac{1}{X_i}\right) - C - \left(\dfrac{1}{2}-\dfrac{1}{n}\right)\log\lambda_o + \dfrac{\lambda_o}{2\overline{X}}\left(-1 + \dfrac{\overline{X}}{n}\sum_{i=1}^{n}\dfrac{1}{X_i}\right)\right] \ .$$

After substituting the calculated values and simplifying, the result is

$$\lambda^*_{T-K} = \exp\left[-1-\left(\frac{n}{2}-1\right)\left[\log 2 + \log\left(\frac{1}{2}-\frac{1}{n}\right)\right]+\log\lambda_*\right] .$$

The final form of the approximation is then

$$\lambda^*_{T-K} = \frac{\lambda_*}{e^{\left[2\left(\frac{1}{2}-\frac{1}{n}\right)\right]^{\frac{n}{2}-1}}} = \hat{\lambda}\left[\frac{1}{e^{\left[2\left(\frac{1}{2}-\frac{1}{n}\right)\right]^{\frac{n}{2}-1}}}\right] .$$

Once again one may notice that asymptotically λ^*_{T-K} becomes indistinguishable from the mle.

Now turn to complete the approximation for the location parameter μ. Recall that L_o, Σ_o and the maxima $\underline{\theta}_o$ have already been obtained. Reconstruct L_* for μ so that $u(\underline{\theta}) = \mu$ and

$$L_* = \frac{1}{n}\log\mu + C + \left(\frac{1}{2}-\frac{1}{n}\right)\log\lambda - \frac{\lambda}{2\mu^2}\left[\overline{X}-2\mu+\frac{\mu^2}{n}\sum_{i=1}^{n}\frac{1}{X_i}\right] .$$

To maximize L_*, obtain the two partial derivatives

$$\frac{\partial L_*}{\partial\mu} = \frac{1}{n\mu}+\frac{\lambda}{\mu^3}(\overline{X}-\mu)$$

$$\frac{\partial L_*}{\partial\lambda} = \frac{\left(\frac{1}{2}-\frac{1}{n}\right)}{\lambda}-\frac{1}{2\mu^2}\left[\overline{X}-2\mu+\frac{\mu^2}{n}\sum_{i=1}^{n}\frac{1}{X_i}\right] .$$

Unfortunately, the same problem appears that was encountered with estimation of the location parameter in the normal distribution. The added μ term in the denominator of the partial derivative with respect to μ requires that the quadratic formula be used, producing a very unwieldy algebraic form for the maximization point. At this point a closed-form solution for $\mu^*_{(T-K)}$ becomes untenable, forcing reliance upon

computer maximization routines.

2.3.3 Example

Bannerjee and Bhattacharyya (1979) present a method to find the exact Bayes estimator in terms of the solution of a system of equations and give an example dataset. The sample dataset comprised nine lifetimes. The results of estimation are given in Table 2.3.

Table 2.3: Estimation Results for Bannerjee and Bhattacharyya Sample

Parameter	Exact Bayes	Lindley	Tierney	MLE
μ	not given	4.1178	4.0930	2.6064
λ	1.6229	1.6228	1.6216	1.6589

Lindley's approximation more closely estimates the true Bayes estimator than the T-K method.

2.3.4 Monte Carlo Simulation

Two simulation runs were performed, as with the normal distribution, generated from an inverse Gaussian distribution with $\mu=5$ and $\lambda=4$. Since both parameters μ and λ displayed interesting results for the example dataset, results are presented for both parameters in Tables 2.4 and 2.5. In looking at the individual samples it was seen that the T-K estimator was always less than the Lindley estimator, although the magnitude of difference was relatively small.

Table 2.4: Simulation Results for Estimating μ from an IG(5,4) Distribution

	n=10 5000 samples		n=100 1000 samples	
Estimator	Mean	MSE	Mean	MSE
Lindley	5.6395	15.2415	5.2698	0.5731
Tierney	5.6102	15.1507	5.2695	0.5729
MLE	6.0231	18.6680	5.3771	0.6621

Table 2.5: Simulation Results for Estimating λ from an IG(5,4) Distribution

	n=10 5000 samples		n=100 1000 samples	
Estimator	Mean	MSE	Mean	MSE
Lindley	5.8525	20.1326	4.3722	0.6730
Tierney	5.8404	20.0192	4.3721	0.6730
MLE	6.5028	26.8824	4.4163	0.7187

Findings for both the scale and location parameters in terms of estimator performance are basically the same. The reduced variability of the T-K approach is evidenced, although there is little to choose between the two approximation methods in terms of precision.

As with the normal distribution, the impact of sample size is notable. For the smaller sample size, the discrepancy between the two approximation techniques is greater. For the larger sample size, the discrepancy between the two techniques becomes negligible. Figures 2.5 through 2.7 contain sampling distributions for the three estimators of λ involved for the (n=100) larger sample size. The two Bayes approximations produce identical plots while the mle does differentiate itself slightly.

Figure 2.5 ESTIMATION FOR INVERSE GAUSSIAN DISTRIBUTION λ
RESULTS FOR 1000 GENERATED SAMPLES
N=100, μ=5.0, λ=4.0

Figure 2.6

ESTIMATION FOR INVERSE GAUSSIAN DISTRIBUTION λ
RESULTS FOR 1000 GENERATED SAMPLES
N=100, μ=5.0, λ=4.0

PERCENTAGE

LINDLEY ESTIMATOR

Figure 2.7

ESTIMATION FOR INVERSE GAUSSIAN DISTRIBUTION λ

RESULTS FOR 1000 GENERATED SAMPLES

N=100, μ=5.0, λ=4.0

PERCENTAGE

TIERNEY-KADANE ESTIMATOR
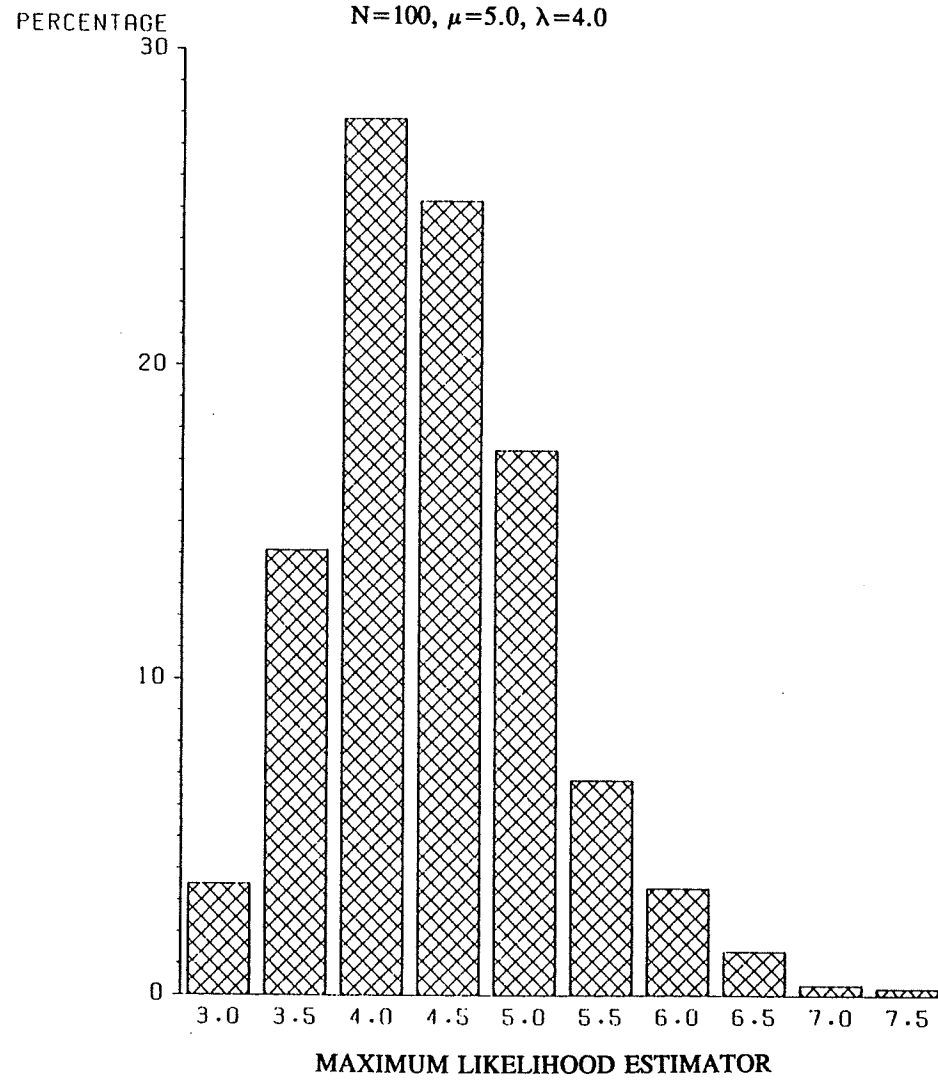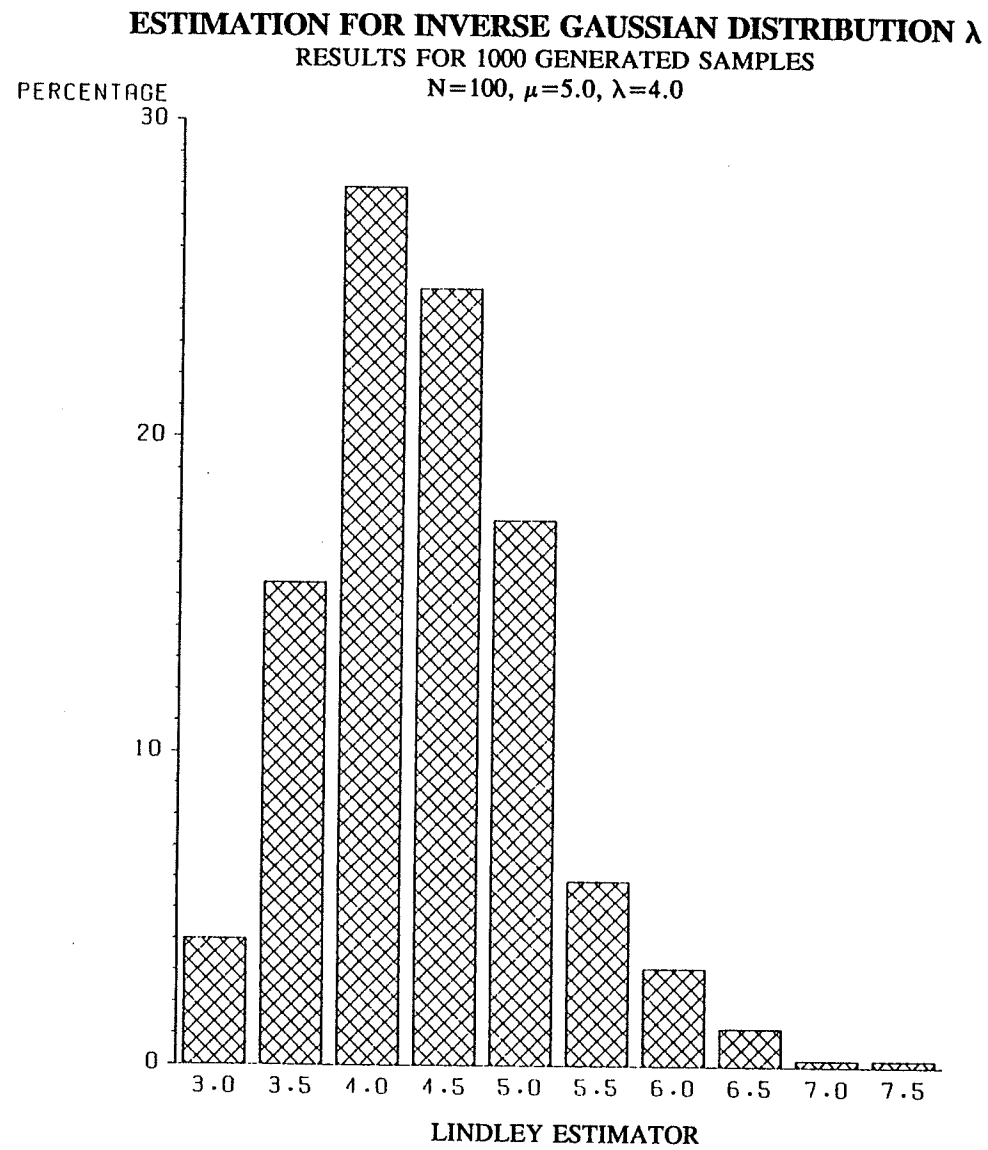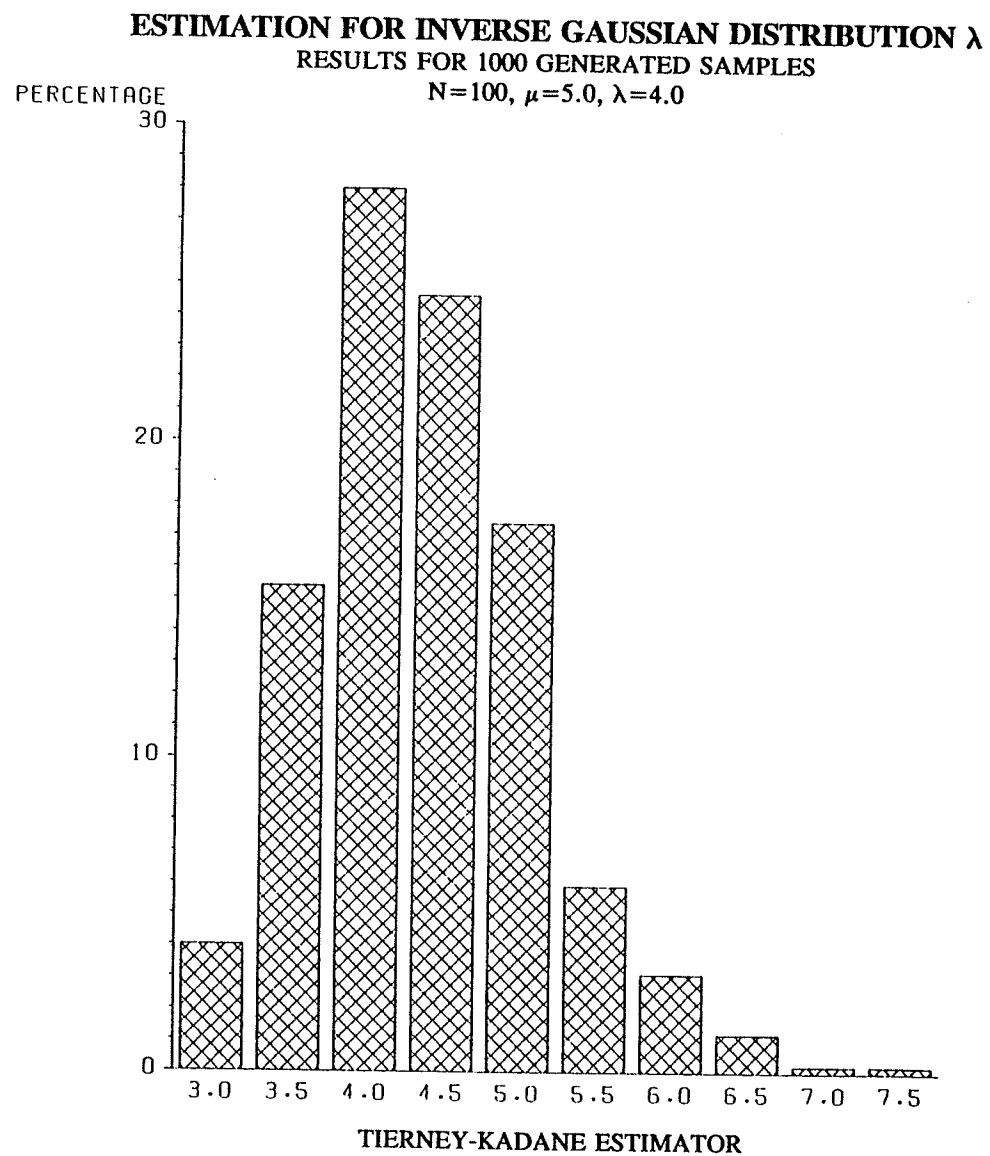
2.4 Summary

In their paper, Tierney and Kadane (1986) assert that their method is computationally easier than Lindley's technique because of the need for the second partial derivatives only to be calculated, while Lindley requires third partials. As has been seen, however, the T-K method requires such derivations for each parameter estimated while Lindley's method demands such algebraic manipulations only once. Furthermore, each technique requires matrix inversion. The Lindley approach needs such inversion only once, whereas the T-K approach needs a new matrix construction and inversion for each parameter to be estimated.

From these two distributional examples, it is clear that the Tierney-Kadane approximation lends itself to closed-form solutions only in the case of scale-type parameters and becomes algebraically intractable for threshold/location parameters. This has largely to do with the choice of prior, of course, and the fact that in this setting priors were used that involve only the scale parameter. This is not an unusual situation. Lindley's method on the other hand does not seem to have any mathematical tendency towards either scale or location parameters and produces closed forms with an equal amount of manipulation.

In terms of the closed-form solutions themselves, the Lindley estimators are much simpler, avoiding the exponential form of the Tierney-Kadane solutions. Lindley's method would be preferred for its simplicity at a negligible cost of accuracy.

Attention will now be turned to more complex distributional settings where even obtaining a numerical solution is difficult.

CHAPTER 3: BAYESIAN APPROXIMATIONS FOR THREE-PARAMETER DISTRIBUTIONS

This chapter continues the work of the previous section, expanding the work to now deal with more complicated estimation environments involving three systemic parameters. Three distributions used routinely in reliability and lifetesting work are the Weibull, gamma and lognormal distributions. Estimation procedures from both the classical and Bayesian perspective are examined. While in many applied situations these three distributions may be thought of as interchangeable, each one has its own special mathematical characteristics. This individuality leads to complicating factors in any attempt at generalization across the three distributions.

3.1 Weibull Distribution

In this section, the three-parameter Weibull distribution will be used to demonstrate the efficacy of the two Bayes approximation techniques relative to the maximum likelihood estimator. Work in this section has been published in IEEE Transactions on Reliability by Sinha and Sloan (1988).

Of the three-parameter distributions commonly had in life testing situations presented in this chapter, the Weibull distribution has seen the most attention in the literature. Named for the Swedish physicist who applied it to material breaking strength testing during the second world war (Weibull, 1939), the Weibull distribution is particularly useful since the reliability and hazard functions can be expressed in simple, closed forms and if the shape parameter is greater than one, the Weibull distribution aptly takes into account failure due to aging and fatigue.

The density function for the three-parameter Weibull distribution is

$$f(x \mid \mu, b, c) = \frac{c}{b^c} (x-\mu)^{c-1} \exp\left[-\left(\frac{x-\mu}{b}\right)^c\right]$$

where c is a shape parameter, b is a scale parameter and μ is the location or threshold parameter representing minimum time until wear out.

A considerably larger body of publications deal with the Weibull than with the alternative life testing distributions of the gamma and lognormal. Citing that there is no clear theoretical reason why the Weibull is so popular a choice over alternative competitors, Johnson and Kotz (1970) provide an extensive list of applications.

The reliability function at a given time t is $\Pr(X > t) = R(t)$ and is given by

$$R(t) = \exp\left[-\left(\frac{t-\mu}{b}\right)^c\right], \quad t > \mu .$$

Although the idea of a minimum "guaranteed" time until failure is intuitively appealing, the addition of the μ parameter creates numerous mathematical problems. As such, work in the literature until recently concerned itself more often with the two-parameter incarnation of the Weibull distribution, assuming that μ was zero.

The more difficult three parameter situation is of interest for use with the Bayes approximations for a number of reasons. First, the problem of finding Bayes estimators for the systemic parameters does not produce closed form solutions. Second, since the Bayes approximations are both based on the mle's which are also unavailable in closed form, some insight to the problems can be gained. Finally, in practical applications, if numerical work is required to produce mle's there is no reason to prefer them over any competing estimator, except for possibly desirable asymptotic properties.

3.1.1 Estimation Problems

The value of c, the shape parameter, is critical for the production of estimators for the three-parameter Weibull. If $c < 1$, the pdf of a Weibull takes the shape of an exponential distribution. The mle's are regular and have the usual desirable properties if and only if $c > 2$ (Cohen and Whitten, 1982). When $c = 1$, the Weibull is a two-parameter exponential variate. Johnson and Haskell (1983) discuss the application of the Weibull distribution for measuring lumber strength, a situation which demands the shape parameter lie between one and two, the nonregular situation. They then demonstrate that the mle's are in fact consistent with the traditional regularity properties as long as c is greater than one. They further give an intuitive interpretation for the shape parameter c. If c is greater than one, the items under test must be wearing out over time, whereas if $c < 1$ the components can be thought of having to go through an initial "breaking in" period.

Due to the extraordinary eccentricity of their behaviour, maximum likelihood estimators of the parameters of this distribution have been studied by several authors. They do not satisfy the usual regularity conditions, due in part because the threshold parameter is a function of the observational range space. As a result for some combinations of parameters and sample sizes the mle may not exist or may lead to inconsistent estimators (Cohen and Whitten, 1982). Smith (1985) demonstrates that the mle's are asymptotically efficient and have a normal distribution only if $c > 2$. Further, if $c < 1$, the maximum likelihood estimators may not exist at all. At best they are difficult to find by numerical methods.

The likelihood function (assuming a sample of size n, namely $\underline{x} = (x_1, x_2, \ldots, x_n)$

from the three-parameter Weibull distribution) is

$$l(c,b,\mu\,|\,\underline{x}) = \left[\frac{c}{b^c}\right]^n \left[\prod_{i=1}^{n} (x_i-\mu)\right]^{c-1} \exp\left[-\sum_{i=1}^{n} \left[\frac{x_i-\mu}{b}\right]^c\right]$$

and the log-likelihood function is

$$L(c,b,\mu\,|\,\underline{x}) = n(\log c - c\log b) + (c-1)\sum_{i=1}^{n} \log(x_i-\mu) - \sum_{i=1}^{n} \left[\frac{x_i-\mu}{b}\right]^c .$$

To produce maximum likelihood estimators, L must be maximized with the restriction that $\mu \le x_{[1]}$. The system of partial derivative equations that would produce mle's is

$$L_1 = \frac{\partial L}{\partial c} = \frac{n}{c} - n\log b + \sum_{i=1}^{n} \log(x_i-\mu) - \sum_{i=1}^{n} \left[\frac{x_i-\mu}{b}\right]^c \log\left[\frac{x_i-\mu}{b}\right] = 0$$

$$L_2 = \frac{\partial L}{\partial b} = -\frac{nc}{b} + \frac{c}{b^{c+1}}\sum_{i=1}^{n} (x_i-\mu)^c = 0$$

$$L_3 = \frac{\partial L}{\partial \mu} = (1-c)\sum_{i=1}^{n} \left[\frac{1}{x_i-\mu}\right] + \frac{c}{b}\sum_{i=1}^{n} \left[\frac{x_i-\mu}{b}\right]^{c-1} = 0 .$$

Unfortunately, this system cannot be solved in closed form so that a number of alternative approaches are possible. The true global maximum for the log-likelihood function occurs at $(\hat{c},\hat{b},\hat{\mu}) = (0,0,x_{[1]})$. Since this is an impractical mle, a local maximum must be found. Rockette, Antle and Klimko (1974) provide the definitive work on mle's with the three-parameter Weibull model. They demonstrate that either no point of inflection will exist for the above system of equations on the log-likelihood surface (hence no mle), or that along with a local maximum a saddle point will exist.

In the former case, the mle is found to be $(\hat{c},\hat{b},\hat{\mu}) = (1, \frac{1}{n}\sum_{i=1}^{n} (x_i-\hat{\mu}), x_{[1]})$. In the

latter case, the value of the likelihood function at the local maximum found through iterative search procedures should be compared with this so-called "corner point". It is possible that the log-likelihood function will be greater at this corner point than at the local maximum. After presenting this result, however, Rockette et al fail to mention that the relatively involved calculation of the log-likelihood function becomes trivial at the corner point.

Specifically,

$$L(\underline{\theta}|\underline{x})_{[corner]} = -n \log(\hat{b}+1) \ .$$

This bit of information economizes a considerable amount of CPU time in a Monte Carlo simulation, as an added pass through the data is unnecessary.

Numerical approaches for solving the mle system of equations abound in the literature. Zanakis (1977, 1979a and 1979b) presents extensive work on various alternatives. He also notes that the log-likelihood function is incredibly flat in the region around the local maximum so that great care must be taken to ensure that the numerical routine produces accurate results. Harter and Moore (1965) suggest solving the equations in a cyclical fashion, slowly zeroing in on the true local maxima, which is CPU intensive. Other numerical recipes are given by Haan and Beer (1967), Wingo (1972), Lemon (1974) and Archer (1980). A detailed comparison of these methods is given by Amin (1981).

The nature of difficulties with the system of mle equations has led to several modified maximum likelihood procedures. Cohen and Whitten (1982a) and Cohen, Whitten & Ding (1984) replace the partial derivative equation for μ in the system above with a relation involving other moments of the distribution or the first order statistic and

solve this new modified system of equations via an iterative numerical routine.

Because of the difficulty in producing maximum likelihood estimators, a number of authors have presented alternative estimation approaches. Zanakis (1979) proposed a set of simple estimators based on sample percentiles which always exist and have at least comparable properties to the mle's. Amin (1981) describes a method based on spacings between successive observations that also produce competitive results. Adatia and Chan (1985) present an approach based on order statistics. Any of these estimators are useful as starting points in the iterative numerical routines necessary to find the mle.

In this section, an adapted Newton-Raphson algorithm is used to locate the local mle's, using either the moment estimators or Zanakis-type estimators for starting the iterative search. Cohen (1965) gives moment estimators of the parameters. Convergence was typically obtained after three to ten iterations. As pointed out by Cohen and Whitten (1982) and Zanakis (1979), it is possible that an iterative numerical process could converge to a saddle point, or to an erroneous local maxima. To circumvent this problem, a 3x3 dimensional grid of the log-likelihood function was constructed around the final estimates produced by the iterative algorithm. If the grid was not uniformly decreasing around the local maximum, the direction of ascent was followed until a further convergence was achieved. If convergence was not achieved, a further set of initial points was used. The estimates reached by use of the second set of initial estimates and the first were then used to bracket the search process. The grid process was then repeated. In this way, it was easy to verify that estimates produced by the iterative process were indeed at a local maximum of the log-likelihood function.

The mle for the reliability function is found simply by substituting the mle's for

the systemic parameters into the reliability function. Despite the difficulties in producing the mle for the three parameters, Thoman, Bain and Antle (1970) show that the maximum likelihood estimator of the system reliability is unbiased for even moderate sample sizes and has desirable variance properties.

3.1.2 Lindley's Approximation

For the Bayesian framework, a state of "in-ignorance" about the parameters a priori is assumed so that the vague priors

$$g(b,c) \propto \frac{1}{bc} \quad and \quad h(\mu) = constant$$

as per Jeffreys (1961) would be appropriate. It is reasonable to believe that μ is distributed independently of c and b since any a priori idea one might have about μ is not likely to be very much influenced by one's knowledge about the values of the other parameters (Box and Tiao, 1973). The joint prior distribution for $\underline{\theta}=(c,b,\mu)$ can then be written as the product of the marginals so that

$$g(c,b,\mu) \propto \frac{1}{bc}$$

and the joint log-prior is

$$\rho(\underline{\theta}) = \log[g(\underline{\theta})] = -\log c - \log b$$

with partial derivatives

$$\rho_1(\underline{\theta}) = \frac{\partial \rho(\underline{\theta})}{\partial c} = -\frac{1}{c}, \quad \rho_2(\underline{\theta}) = \frac{\partial \rho(\underline{\theta})}{\partial b} = -\frac{1}{b}, \quad \rho_3 = \frac{\partial \rho(\underline{\theta})}{\partial \mu} = 0 .$$

Although the mle's are difficult to obtain for the three-parameter Weibull, the

approximation to the Bayes estimator as proposed by Lindley (1970) encounters no further complications. The formulation for the expansion seen in chapter one was utilised and for brevity will not be repeated here. The necessary quantities are derived below for estimating the systemic parameters by setting $u(\underline{\theta})$ in turn to each of the elements of $\underline{\theta} = (c, b, \mu)$ and running through the machinations.

The second partial derivatives, required to construct the information matrix are

$$L_{11} = \frac{\partial^2 L}{\partial c^2} = -\frac{n}{c^2} - \sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^c \left[\log\left[\frac{x_i - \mu}{b}\right]\right]^2$$

$$L_{22} = \frac{\partial^2 L}{\partial b^2} = \frac{nc}{b^2} - \frac{c(c+1)}{b^{c+2}} \sum_{i=1}^{n} (x_i - \mu)^c$$

$$L_{33} = \frac{\partial^2 L}{\partial \mu^2} = -(c-1)\left[\sum_{i=1}^{n} (x_i - \mu)^{-2} + \frac{c}{b^2}\sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^{c-2}\right]$$

$$L_{12} = \frac{\partial^2 L}{\partial c \partial b} = -\frac{n}{b} + \frac{1}{b}\left[\sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^c \left[c\log\left[\frac{x_i - \mu}{b}\right] + 1\right]\right]$$

$$L_{13} = \frac{\partial^2 L}{\partial c \partial \mu} = -\sum_{i=1}^{n} (x_i - \mu)^{-1} + \frac{1}{b}\sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^{c-1} \left[c\log\left[\frac{x_i - \mu}{b}\right] + 1\right]$$

$$L_{23} = \frac{\partial^2 L}{\partial b \partial \mu} = -\frac{c^2}{b^2}\sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^{c-1}$$

The third partial derivatives (L_{ijk}'s) are

$$L_{111} = \frac{\partial^3 L}{\partial c^3} = \frac{2n}{c^3} - \sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^c \left[\log\left[\frac{x_i - \mu}{b}\right]\right]^3$$

$$L_{222} = \frac{\partial^3 L}{\partial b^3} = \frac{c}{b^3} \left[(c+1)(c+2) \sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^c - 2n\right]$$

$$L_{333} = \frac{\partial^3 L}{\partial \mu^3} = (c-1) \left[\frac{c(c-2)}{b^3} \sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^{c-3} - 2 \sum_{i=1}^{n} (x_i - \mu)^{-3}\right]$$

$$L_{112} = \frac{\partial^3 L}{\partial c^2 \partial b} = \frac{1}{b} \sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^c \left[c\left[\log\left[\frac{x_i - \mu}{b}\right]\right]^2 + 2\log\left[\frac{x_i - \mu}{b}\right]\right]$$

$$L_{122} = \frac{\partial^3 L}{\partial c \partial b^2} = \frac{n}{b^2} - \frac{1}{b^2} \left[\sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^c \left[(c^2 + c) \log\left[\frac{x_i - \mu}{b}\right] + 2c + 1\right]\right]$$

$$L_{113} = \frac{\partial^3 L}{\partial c^2 \partial \mu} = \frac{1}{b} \sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^{c-1} \left[2\log\left[\frac{x_i - \mu}{b}\right] + c\left[\log\left[\frac{x_i - \mu}{b}\right]\right]^2\right]$$

$$L_{133} = \frac{\partial^3 L}{\partial c \partial \mu^2} = -\sum_{i=1}^{n} (x_i - \mu)^{-2} - \frac{1}{b^2} \sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^{c-2} \left[c + (c-1)\left[c\log\left[\frac{x_i - \mu}{b}\right] + 1\right]\right]$$

$$L_{233} = \frac{\partial^3 L}{\partial b \partial \mu^2} = \frac{c^2(c-1)}{b^3} \sum_{i=1}^{n} \left[\frac{x_i - \mu}{b}\right]^{c-2}$$

Note that for the three-parameter Weibull distribution, one cannot algebraically simplify the L_{ijk} functions as was done for the simpler two-parameter distributions seen earlier in chapter one because the mle's are not available in closed form.

Now for estimating the reliability function, R(t), write the parameter to be estimated as

$$u(\underline{\theta}) = R(t) = \exp\left[-\left(\frac{t-\mu}{b}\right)^c\right], \quad t>\mu$$

and find the first and second partial derivatives of the function to be estimated (the u_i's and u_{ij}'s) as follows

$$u_1 = \frac{\partial u(\underline{\theta})}{\partial c} = -n\left(\frac{t-\mu}{b}\right)^{c+1}R(t)$$

$$u_2 = \frac{\partial u(\underline{\theta})}{\partial b} = \frac{c}{b}\left(\frac{t-\mu}{b}\right)^c R(t)$$

$$u_3 = \frac{\partial u(\underline{\theta})}{\partial \mu} = u_2/\left(\frac{t-\mu}{b}\right)$$

$$u_{11} = \frac{\partial^2 u(\underline{\theta})}{\partial c^2} = n^2\left(\frac{t-\mu}{b}\right)^{c+2}\left[\left(\frac{t-\mu}{b}\right)^c-1\right]R(t)$$

$$u_{22} = \frac{\partial^2 u(\underline{\theta})}{\partial b^2} = \frac{c}{b^2}\left(\frac{t-\mu}{b}\right)^c\left[c\left(\frac{t-\mu}{b}\right)^c-c-1\right]R(t)$$

$$u_{33} = \frac{\partial^2 u(\underline{\theta})}{\partial \mu^2} = \frac{c}{b^2}\left(\frac{t-\mu}{b}\right)^{c-2}\left[c\left(\frac{t-\mu}{b}\right)^c-c+1\right]R(t)$$

$$u_{13} = \frac{\partial^2 u(\underline{\theta})}{\partial c\partial \mu} = \frac{1}{b}\left(\frac{t-\mu}{b}\right)^{c-1}\left[nc\left(\frac{t-\mu}{b}\right)\left[1-\left(\frac{t-\mu}{b}\right)^c\right]+1\right]R(t)$$

$$u_{12} = \frac{\partial^2 u(\underline{\theta})}{\partial c\partial b} = \left(\frac{t-\mu}{b}\right)u_{13}$$

$$u_{23} = \frac{\partial^2 u(\theta)}{\partial b \partial \mu} = \frac{c^2}{b^2}\left(\frac{t-\mu}{b}\right)^{c-1}\left[\left(\frac{t-\mu}{b}\right)^c - 1\right]R(t)$$

The prior distribution and its partials remain the same, as do the σ_{ij}'s and the L_{ijk}'s.

Using the Lindley algebra from Chapter one, Sinha and Sloan (1988) give some simplified algebraic equations for the three-parameter Weibull. These forms serve to reduce CPU time. Lindley's equation (1.15) for estimating the three systemic parameters becomes

$$c^* = c - \frac{\sigma_{11}}{c} - \frac{\sigma_{12}}{b} + \frac{1}{2}(A\sigma_{11} + B\sigma_{12} + D\sigma_{13})$$

$$b^* = b - \frac{\sigma_{21}}{c} - \frac{\sigma_{22}}{b} + \frac{1}{2}(A\sigma_{21} + B\sigma_{22} + D\sigma_{23})$$

$$\mu^* = \mu - \frac{\sigma_{31}}{c} - \frac{\sigma_{32}}{b} + \frac{1}{2}(A\sigma_{31} + B\sigma_{32} + D\sigma_{33})$$

where

$$A = \sigma_{11}L_{111} + 2\sigma_{12}L_{121} + 2\sigma_{13}L_{131} + 2\sigma_{23}L_{231} + \sigma_{22}L_{221} + \sigma_{33}L_{331}$$

$$B = \sigma_{11}L_{112} + 2\sigma_{12}L_{122} + 2\sigma_{13}L_{132} + 2\sigma_{23}L_{232} + \sigma_{22}L_{222} + \sigma_{33}L_{332}$$

$$D = \sigma_{11}L_{113} + 2\sigma_{12}L_{123} + 2\sigma_{13}L_{133} + 2\sigma_{23}L_{233} + \sigma_{22}L_{223} + \sigma_{33}L_{333}$$

Recall that $\sigma_{ij} = -\{L_{ij}\}^{-1}$ and that all parametric functions are evaluated at the maximum likelihood estimator.

The Lindley approximation to the Bayes estimator of $R(t)$ is found in a similar manner to be

$$R^*(t) = \hat{R}(t) + \lambda_1 U_1 + \lambda_2 U_2 + \lambda_3 U_3 + a_4 + a_5 \ .$$

where

$$\lambda_1 = \frac{A}{2} - \frac{1}{c}, \quad \lambda_2 = \frac{B}{2} - \frac{1}{b}, \quad \lambda_3 = \frac{D}{2}$$

$$U_j = u_1 \sigma_{j1} + u_2 \sigma_{j2} + u_3 \sigma_{j3}, \quad j = 1, 2, 3$$

$$\hat{a}_4 = u_{12} \sigma_{12} + u_{13} \sigma_{13} + u_{23} \sigma_{23}$$

$$a_5 = \frac{1}{2} (u_{11} \sigma_{11} + u_{22} \sigma_{22} + u_{33} \sigma_{33})$$

and the u function partial derivatives are based on the setting of $u(\underline{\theta}) = R(t)$. Note that in this case the partial derivatives of u are all nonzero and contribute to the approximation, whereas in estimating the three elements of $\underline{\theta} = (c, b, \mu)$, all but one of the partial derivatives was zero.

The posterior variance can be estimated using Lindley's approximation. In Lindley's algebra this means, to obtain the posterior variance of systemic parameter c, setting $u(\underline{\theta}) = c^2$ and reconstructing the u_i and u_{ij} functions.

After some algebra it is seen that

$$E(c^2 | \underline{x}) = \hat{c}^2 - 2\hat{c} \left[\frac{\sigma_{11}}{\hat{c}} + \frac{\sigma_{12}}{\hat{b}} \right] + \sigma_{11} + \hat{c} (A\sigma_{11} + B\sigma_{12} + D\sigma_{13})$$

and substituting into the variance relation form, the posterior variance for c is found to be

$$Var(c | \underline{x}) = E(c^2 | \underline{x}) - (c^*)^2 = \sigma_{11} - \left[\frac{1}{2} (A\sigma_{11} + B\sigma_{12} + D\sigma_{12}) - \frac{\sigma_{11}}{c} - \frac{\sigma_{12}}{b} \right]^2$$

Similarly, for the other systemic parameters

$$Var(b|\underline{x}) = \sigma_{22} - \left[\frac{1}{2}(A\sigma_{21}+B\sigma_{22}+D\sigma_{23}) - \frac{\sigma_{11}}{c} - \frac{\sigma_{12}}{b}\right]^2$$

$$Var(\mu|\underline{x}) = \sigma_{33} - \left[\frac{1}{2}(A\sigma_{31}+B\sigma_{32}+D\sigma_{33}) - \frac{\sigma_{31}}{c} - \frac{\sigma_{32}}{b}\right]^2 \quad .$$

Again it is important to note that all functions are evaluated at the mle. This algebra reinforces the finding of Sinha (1987) which proved that by including terms up to $O(n^{-2})$, the posterior variance is less than the estimated asymptotic variance of the mle for the same parameter.

In estimating the posterior variance of the reliability estimator, the same route is followed so that

$$E(R^2(t)|\underline{x}) = \hat{R}^2(t) + 2\hat{R}(t)(\lambda_1 U_1 + \lambda_2 U_2 + \lambda_3 U_3 + a_4 + a_5)$$

$$+ u_1^2\sigma_{11} + u_2^2\sigma_{22} + u_3\sigma_{33} + 2u_1u_2\sigma_{12} + 2u_1u_3\sigma_{13} + 2u_2u_3\sigma_{23}$$

and finally

$$Var[R(t)|\underline{x}] = V\hat{a}r[\hat{R}(t)] - \left[\lambda_1 U_1 + \lambda_2 U_2 + \lambda_3 U_3 + a_4 + a_5\right]^2 \quad .$$

Again it is seen that the posterior variance will be less than the estimated asymptotic variance of the mle if terms to order $O(n^{-2})$ are included.

3.1.3 Tierney-Kadane Approximation

To produce a closed form solution for the Tierney-Kadane approximation to the Bayes estimator, the L_o and L_* functions must first be constructed and maximized, producing the points θ_o and θ_* respectively where $\theta = (c, b, \mu)$.

The L_o function, constant for any parameter estimated, is

$$L_o = \frac{1}{n}\left[\log(g(\underline{\theta}))+L(\underline{\theta}|\underline{x})\right]$$

$$= \frac{1}{n}\left[-\log b-\log c+n\log c-nc\log b+(c-1)\sum_{i=1}^{n}\log(x_i-\mu)-\sum_{i=1}^{n}\left[\frac{x_i-\mu}{b}\right]^c\right]$$

and the first partial derivatives with respect to $\underline{\theta}$ necessary to produce the maximization point θ_o are

$$\frac{\partial L_o}{\partial c} = \frac{1}{n}\left[\frac{n-1}{c}-n\log b+\sum_{i=1}^{n}\log(x_i-\mu)-\sum_{i=1}^{n}\left[\frac{x_i-\mu}{b}\right]^c\log\left[\frac{x_i-\mu}{b}\right]\right]$$

$$\frac{\partial L_0}{\partial b} = -\frac{nc+1}{nb}+\frac{c}{nb^{c+1}}\sum_{i=1}^{n}(x_i-\mu)^c$$

$$\frac{\partial L_o}{\partial \mu} = \frac{1}{n}\left[(1-c)\sum_{i=1}^{n}\left[\frac{1}{x_i-\mu}\right]+\frac{c}{b}\sum_{i=1}^{n}\left[\frac{x_i-\mu}{b}\right]^{c-1}\right]$$

These functions are slightly modified functions of the log-likelihood, and as such any maximization attempt suffers from the same problems seen in constructing the mle. Hence, closed form solutions are not achievable for the Tierney-Kadane approximation either and the computer must be used instead.

In a similar fashion the L_* functions that are required to estimate the elements of $\underline{\theta} = (c, b, \mu)$ are

$$L_{*c} = \frac{1}{n} (\log c) + L_o$$

$$L_{*b} = \frac{1}{n} (\log b) + L_o$$

$$L_{*\mu} = \frac{1}{n} (\log \mu) + L_o$$

and again they cannot be maximized algebraically.

To produce posterior variances, the $u(\underline{\theta})$ function is reset, as was done in the previous chapter, to estimate $E(c^2 | \underline{x})$ and substitute this value into the variance relation form. Although L_o remains constant, three more L_* functions are necessary.

$$L_{*c^2} = \frac{1}{n} (2 \log c) + L_o$$

$$L_{*b^2} = \frac{1}{n} (2 \log b) + L_o$$

$$L_{*\mu^2} = \frac{1}{n} (2 \log \mu) + L_o$$

Once the maximization points are achieved numerically, the remaining steps of the T-K method, involving the production of the Σ_o and Σ_* matrices, must be carried out. Recall these matrices are constructed using the second partial derivatives of the L_o and L_* functions respectively. These turn out to be slight modifications of the L_{ij} functions presented in the previous section and so will not be repeated here.

To produce an estimate for the reliability function R(t), reset

$$u(\underline{\theta}) = R(t) = e^{-\left(\frac{t-\mu}{b}\right)^c}$$

and reconstruct

$$L_{*R(t)} = -\frac{1}{n}\left(\frac{t-\mu}{b}\right)^{c} + L_{o} \ .$$

Again the maximization must be done numerically.

As was seen earlier in work done with other distributions, the two approximation methods take very different algebraic routes toward the same goal. However, except for very small samples ($n < 10$), it was found that results for the two approximations are virtually identical. For sake of brevity, then, results in the subsequent sections will be presented under the heading "Lindley" only, yet can be considered to be representative of the T-K approximation as well. The "Lindley" heading is used to be consistent with the published paper in IEEE. This reduces the amount of redundancy in the results for example datasets and Monte Carlo simulation.

In terms of computing time, the T-K method did prove to be slightly advantageous over Lindley's method for estimating the systemic parameters. The maximization performed on the L_{o} and L_{*} functions in the T-K method was equivalent to the time required to produce the maximum likelihood estimators necessary to Lindley's approach. The subsequent matrix construction and inversion needed to complete the T-K approximation took less time than the calculation of the L_{ijk}'s needed to finish Lindley's procedure.

Once the added task of estimating R(t) was considered, however, the CPU time required to perform the T-K method overtook that for Lindley's. This was due to the T-K approach requiring a further maximization of the new $L_{*R(t)}$ seen above, as opposed to Lindley's method where the mle for R(t) was found simply by substituting the

parameter mle's into the R(t) equation.

As mentioned previously, the choice between the two approximations becomes largely a matter of personal preference and availability of computer software. This is especially true for the three-parameter Weibull case. The algebraic advantages of Lindley's method seen for the simpler two-parameter distributions fail to produce closed form solutions. The computer must be relied upon for both approximation techniques for the production of estimates.

3.1.4 Examples

In searching the literature for real data examples, one comes across several situations where the mle's are non-existent (Steen and Stickler, 1976, for example). Lindley's approximation cannot be used in such a situation since the formula is based on an adjustment to the value of the maximum likelihood estimator. In efforts to implement the Tierney-Kadane method on such samples, the same problems facing the mle were encountered and so were not obtainable either. Although this does not constitute a proof that the T-K method will be unable to produce estimates whenever the environment is so eccentric so as to exclude the possibility of mle's, it is the author's opinion that this is likely the case.

Real life examples of three-parameter Weibull datasets in the literature seem to specialize in producing troublesome samples, either due to small sample size or deviant shape parameter. If the Bayes approximations can perform adequately using a noninformative prior setting under such conditions, one can only expect them to improve under less adverse conditions.

Three real data examples are provided by Englehardt & Lee (1979, E&L), Dumonceaux & Antle (1973, D&A) and Cohen & Whitten (1982, C&W) involving ball-bearing lifetimes, flood levels and component lifetimes respectively. Although none of the parameters are known, it is clear that the shape parameter is between one and two, which is the nonregular case described at the beginning of this section. These are also good illustrative samples as the sizes range from very small (n=10) for the E&L sample to moderately large (n=100) for the C&W sample.

Table 3.1 presents the results of parameter estimation for the three given application datasets. Once the mle's were obtained, empirical goodness of fit tests as per D'Agostino and Stephens (1986) were performed to ensure that the model assumption of a three-parameter Weibull distribution was appropriate for each sample.

The approximation to the Bayes estimators are very similar to the mle's of c,b and μ for the D&A and C&W samples and could adequately be described as competitive. For the very small sample of E&L, however, the Bayes approximations are markedly different from the mle's. The mle for the shape parameter c, suggests the nonregular case by an estimate below two whereas the Bayes approach produces a value of $c^* = 3.67$, suggesting that the sample arose from a more regular Weibull process. Bayes estimates for the other two parameters based on the E&L sample are also terribly discordant with the mle's.

Table 3.1: Parameter Estimation Results For Application Samples

Source	n	c mle	b mle	μ mle	c^*	b^*	μ^*
E&L	10	1.74	53.34	86.57	3.67	103.90	43.91
D&A	20	1.24	0.17	0.26	1.32	0.15	0.26
C&W	100	1.50	0.97	0.09	1.58	1.02	0.05

The posterior variance estimates also show stark contrast to the estimated asymptotic variance of the mle in Table 3.2. The forms for the estimated asymptotic variance of the mle were taken from Cohen and Whitten (1982). For the two smaller samples (E&L and D&A) negative posterior variance estimates are observed. As discussed previously, this phenomenon is common when n is small, due to the estimation process being composed of two separate estimates subtracted from one another. Even though the D&A sample produces a posterior variance for c of 0.002, it is undoubtedly underestimated and not to be given much credence. The C&W sample demonstrates that for moderate sample sizes, the negative estimates disappear and results are comparable to the mle. The posterior variance values are less than the estimated asymptotic variances of the mle's, a result found algebraically by Sinha (1987).

Table 3.2: Variance Estimation Results For Application Samples

Source	n	c mle	b mle	μ mle	c^*	b^*	μ^*
E&L	10	1.45	674.45	343.44	neg	neg	neg
D&A	20	0.13	0.0016	0.0001	0.002	neg	neg
C&W	100	0.02	0.006	0.0006	0.01	0.004	0.0005

Note: neg = negative variance estimate

Estimation results for the reliability function are given for the Englehardt and Lee

(E&L) sample in Table 3.3. Inspection of the table reveals that the problems seen in estimating the parameters seemingly disappear in estimating R(t). The outcome is quite surprising considering how poorly the Bayes parameter estimates were in relation to the mle's. The estimated R(t) functions are given in relation to the true value in Figure 3.1 for a more parsimonious display. The Bayes approximation is clearly closer to the true R(t) function throughout most of the range than to the mle. The mle function seems to better parallel the R(t) function however. From Table 3.3, it can be seen that the problem with negative variance estimates is absent for this very small sample (n=10). The posterior variance is consistently smaller than the estimated asymptotic variance.

Table 3.3: R(t) Estimates For Englehardt & Lee (1979) Sample(n=10)

t	True R(t)	MLE R(t)	$R^*(t)$	Estd V(^)	Posterior V(*)
100	0.9726	0.9130	0.8247	0.00439	0.00341
110	0.8948	0.7871	0.7664	0.01516	0.01473
120	0.7788	0.6415	0.7024	0.02514	0.02143
130	0.6412	0.4968	0.6037	0.02511	0.01368
140	0.4994	0.3669	0.4749	0.01894	0.00727
150	0.3679	0.2590	0.3395	0.01303	0.00654
160	0.2564	0.1751	0.2214	0.00953	0.00739
170	0.1690	0.1136	0.1336	0.00745	0.00705
180	0.1054	0.0708	0.0775	0.00564	0.00560
190	0.0622	0.0424	0.0465	0.00385	0.00383
200	0.0347	0.0245	0.0313	0.00231	0.00226

Figure 3.1 — THREE−PARAMETER WEIBULL DISTRIBUTION
R(T) ESTIMATES FOR N=10 ENGLEHARDT & LEE SAMPLE
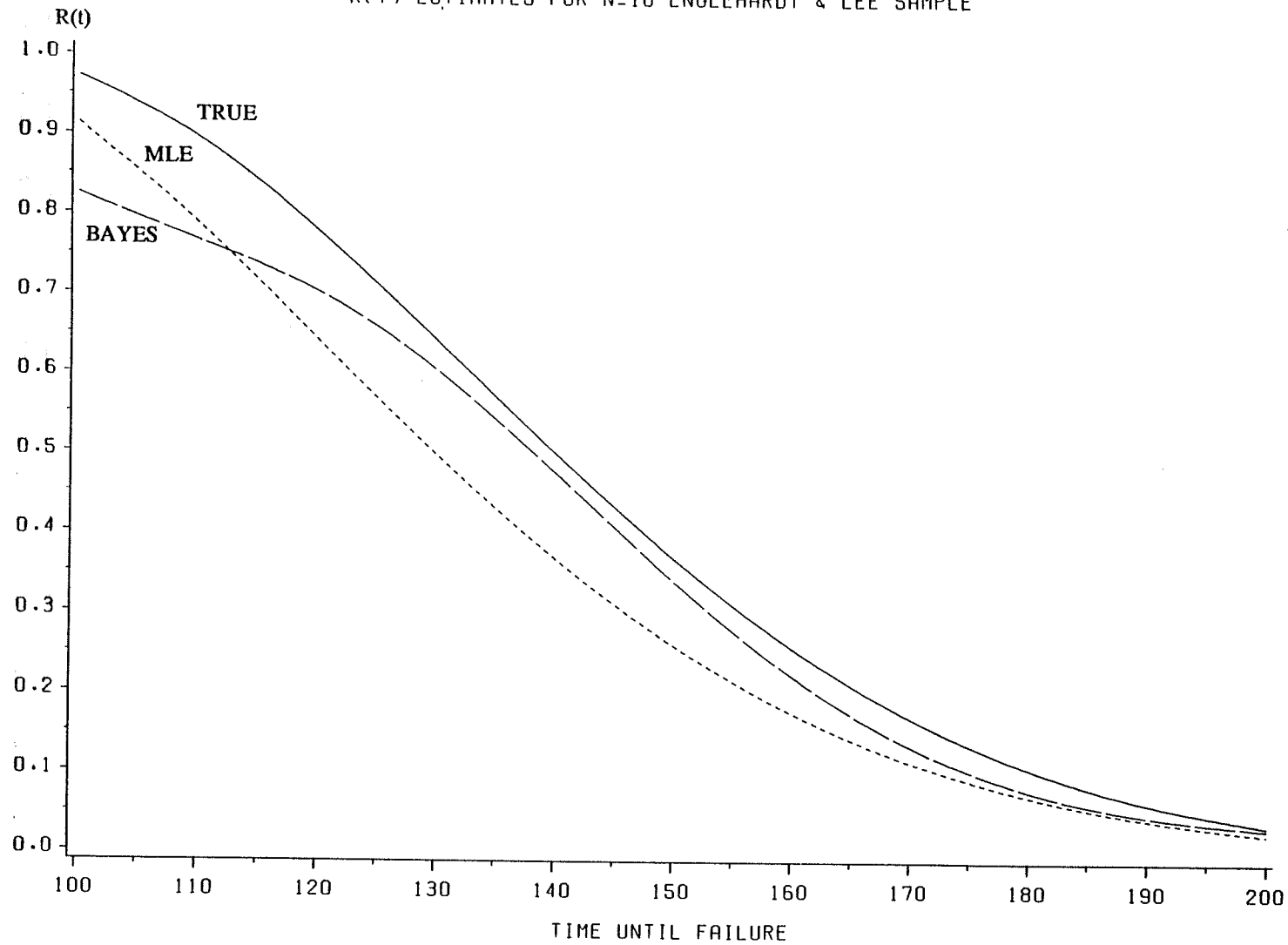
3.1.5 Monte Carlo Simulation

To explore more fully the effect of sample size, a series of samples were generated from a three-parameter Weibull process. An interesting side issue arose in this sample generation process.

Mathematically, the well known inverse transformation approach proceeds as follows

1) Let u_i be a point generated from a Uniform [0,1] distribution

2) Transform u_i to obtain $x_i = b[-\log(1-u_i)]^{1/c} + \mu$

The resultant x_i will follow a three-parameter Weibull distribution as required. In beta-testing the samples generated from the above algorithm, it became apparent via empirical goodness of fit testing that a larger proportion of samples were failing to fit the model than would otherwise have been expected. In graphing some of the generated distributions, it was found that an inordinate number of large-valued observations were being generated, causing the distribution to have either an extension or a "bump" in the upper tail. Upon closer inspection it became clear that numerical problems associated with the use of the logarithmic function were the cause. If the u_i generated were small, raising the logarithm of this value to the prescribed power resulted in a loss of accuracy so that several different u_i's would produce the same x_i value. As an alternative, the following algorithm was used

1) Let g_i be a point generated from an Exponential distribution with mean one.

2) Transform to obtain $x_i = b(g_i)^{1/c} + \mu$.

This algorithm performed much better than the previous one, with no evidence of "bumps" in the resultant empirical distribution.

There are no added complications for generating observations from an exponential distribution as opposed to a uniform variate, since there are numerous accurate and efficient algorithms available. The IMSL (1975) routines GGUBS and GGEXP were used for the uniform and exponential variates respectively. All simulations were produced via the PL/I language on an Amdahl 470 mainframe.

Parameter settings of $(c,b,\mu)=(3.0, 100.0, 30.0)$ were used so that the lack of regularity of the environment would not influence estimation results. Sample size was varied from $n=10$ to $n=400$ with results contained in Table 3.4. The method of moment estimators, used to initiate the search routine for the mle's are also given for comparison.

Table 3.4: Estimation Results Based On Varying Sample Sizes

With Parameters $(c,b,\mu)=(3,100,30)$

Parameter	mle	Lindley	Moment	MLE Var	Post Var
n=10					
c	2.44	4.73	2.98	2.67	neg
b	44.69	85.78	53.52	564.5	neg
μ	69.42	31.36	61.29	423.0	neg
n=20					
c	1.66	2.14	1.87	1.25	0.02
b	45.68	58.53	50.26	406.7	neg
μ	76.91	66.89	73.25	363.1	neg
n=40					
c	3.02	3.73	2.77	0.58	0.06
b	90.09	111.78	83.47	331.2	neg
μ	40.03	20.00	46.54	265.8	neg

n=100					
c	3.26	3.72	3.56	0.53	32.1
b	99.16	112.77	171.29	348.0	163.2
μ	30.44	17.42	-34.96	200.7	131.5
n=200					
c	2.96	3.13	1.55	0.16	0.13
b	91.48	96.75	1.00	100.5	72.73
μ	36.59	31.61	633.22	82.09	57.49
n=300					
c	2.99	3.12	3.28	0.16	0.10
b	92.53	96.75	101.90	77.19	62.46
μ	34.74	31.61	25.95	63.79	50.66
n=400					
c	2.78	2.87	3.09	0.07	0.06
b	93.43	96.23	103.15	47.70	39.90
μ	37.51	34.90	28.41	37.68	30.87

The impact of sample size is clearly illustrated as the Bayes estimates go from wildly different values in comparison with the mle's to practically identical values. Interestingly enough, however, the convergence seen between the Bayes and mle approaches in the two-parameter distributions earlier in this chapter was much more rapid than is the case for the three-parameter Weibull distribution. By n=100 for the two-parameter case, the competing estimates were identical to significant digits. Here in Table 3.5, even at n=400 there are observable differences in the estimates. This is due likely to the more difficult estimation environment produced by the introduction of a threshold parameter.

The Bayes estimators of c and b are consistently larger than the corresponding mle. There is no clear winner between the two approaches, although both are clearly superior to the method of moments estimators. The smaller samples naturally evidence the greatest fluctuation. In agreement with the literature, it is apparent that if the shape parameter is not adequately estimated, the other two parameters are poorly estimated. The n=10 sample size results show marked differences between the Bayes and mle c estimates, producing quite different results for the other two parameters. Stability with the Bayes estimators does not seem to arrive until the n=200 sample, whereas the mle results look decent at the n=100 size.

The negative posterior variance estimates for the examples from the previous section reappear. Not until the sample size reaches n=200 do the posterior variance results look believable from a practical standpoint. In other simulations the author has observed negative posterior variance estimates for samples as large as 200 observations. For the larger samples, however, the Bayes approximations do demonstrate lower variance than the asymptotic estimates of the mle.

To further investigate the surprising results of reliability estimation in the previous section, the process was repeated for the simulation series. Running the sample size from n=40 to n=400 produced a monotone effect as n increased. As intermediate results along a continuum, the figures for each sample size will not be given. Representative information can be gleaned from inspection of the n=40 sample results given in Table 3.5 and the n=400 sample size in Table 3.6. Graphical displays of the estimated functions in relation to the true R(t) function are given in Figures 3.2 and 3.3 respectively.

Figure 3.2

THREE-PARAMETER WEIBULL DISTRIBUTION
RELIABILITY ESTIMATES FOR N=40 GENERATED SAMPLE

R(t)

MLE

BAYES

TRUE

TIME UNTIL FAILURE
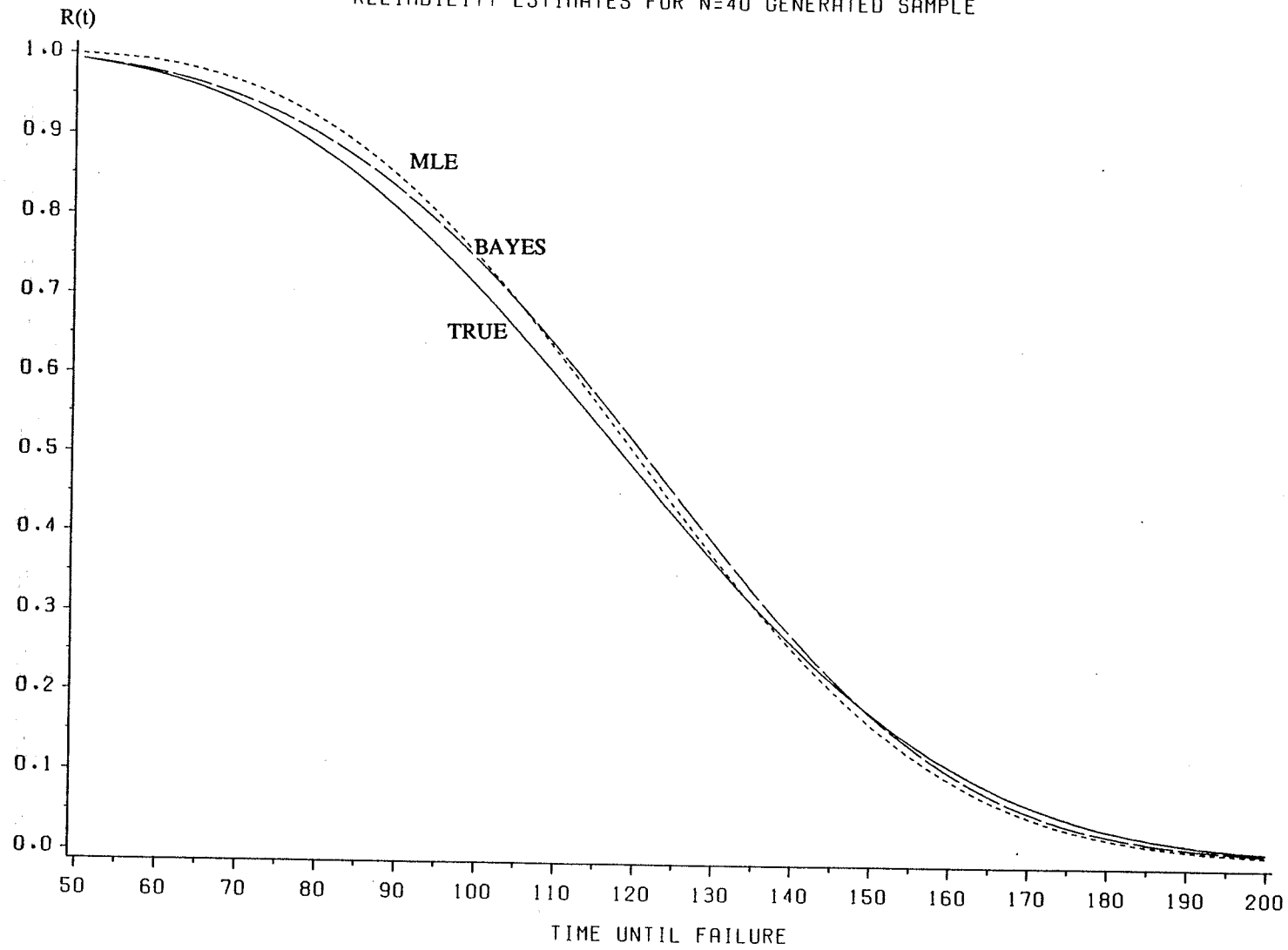
Figure 3.3

THREE−PARAMETER WEIBULL DISTRIBUTION
RELIABILITY ESTIMATES FOR N=400 GENERATED SAMPLE

TIME UNTIL FAILURE
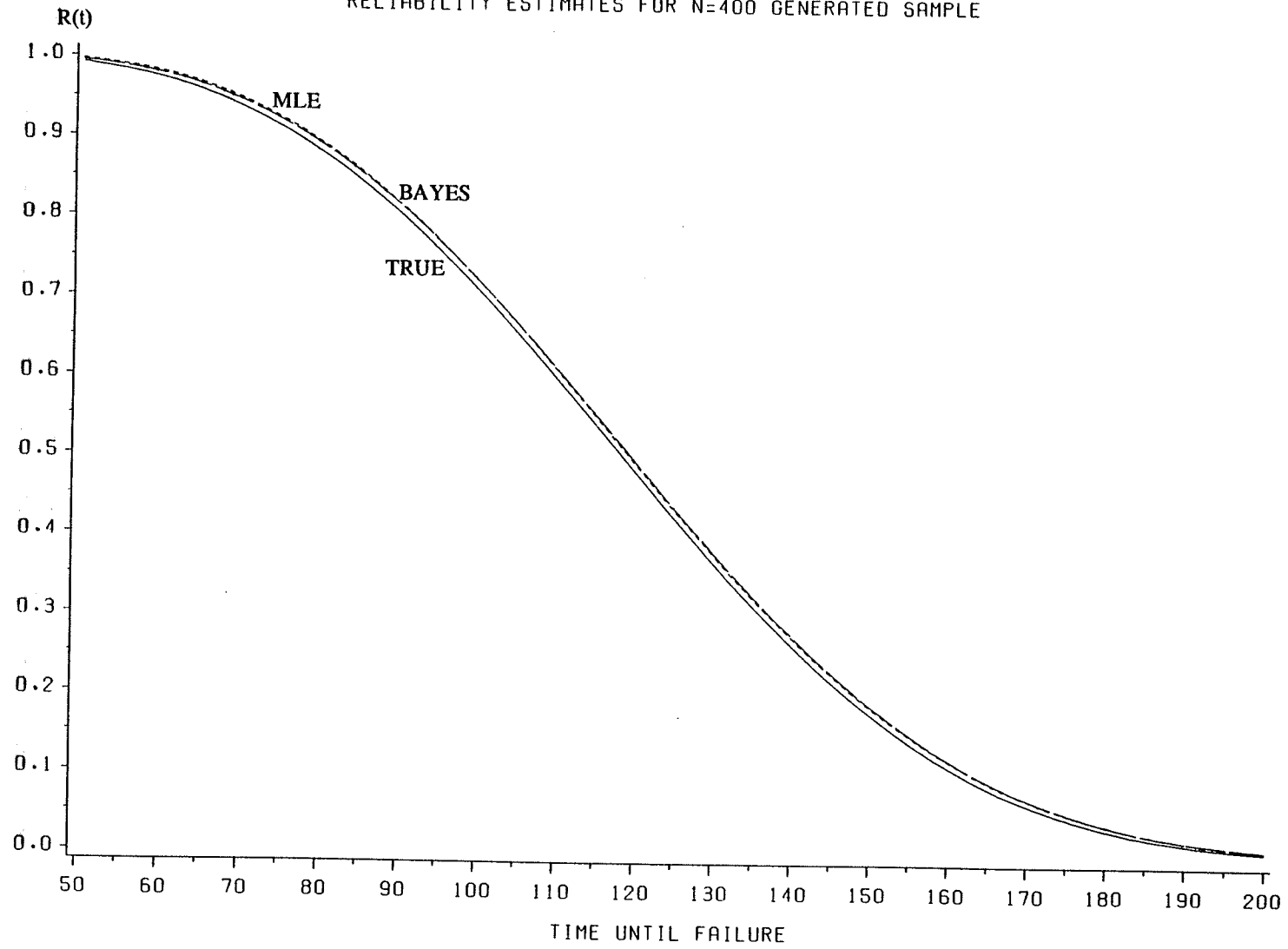
Table 3.5: Reliability Estimates For The n=40 Generated Sample

t	True R(t)	MLE R(t)	R*(t)	Estd V(^)	Posterior V(*)
50	0.9920	0.9988	0.9920	0.00001	neg
60	0.9734	0.9899	0.9764	0.00012	neg
70	0.9380	0.9655	0.9469	0.00040	0.00006
80	0.8825	0.9192	0.8998	0.00098	0.00060
90	0.8057	0.8470	0.8323	0.00197	0.00175
100	0.7096	0.7492	0.7437	0.00319	0.00316
110	0.5993	0.6309	0.6361	0.00418	0.00415
120	0.4824	0.5015	0.5151	0.00446	0.00427
130	0.3679	0.3731	0.3907	0.00402	0.00371
140	0.2642	0.2577	0.2744	0.00317	0.00289
150	0.1776	0.1638	0.1767	0.00224	0.00207
160	0.1111	0.0950	0.1038	0.00139	0.00132
170	0.0643	0.0499	0.0555	0.00073	0.00070
180	0.0342	0.0235	0.0273	0.00031	0.00030
190	0.0166	0.0098	0.0126	0.00010	0.00009
200	0.0074	0.0036	0.0055	0.00003	0.00002

Convergence of results between the competing estimation procedures is much more rapid for estimating R(t) than it was for the systemic parameters. The n=40 results show the two methods equally advantageous, depending upon the value of t. The n=40 case seems to suggest that the mle does a better job in the middle range of t while the Bayes approach exhibits less bias in the tails. By the time n reaches 400, there is little to choose between the two methods as the estimated functions almost coincide with the true R(t).

The Bayes procedure still evidences negative variance estimates in Table 3.5. As might be expected, they tend to appear in the tails of the R(t) function when the true value is close to zero or one. The larger sample size results of Table 3.6 reinforce the idea of the Bayes estimator having smaller variance.

Table 3.6: Reliability Estimates For The n=400 Generated Sample

t	True R(t)	MLE R(t)	$R^*(t)$	Estd V(^)	Posterior V(*)
50	0.9920	0.9963	0.9950	0.00001	0.00001
60	0.9734	0.9812	0.9792	0.00003	0.00003
70	0.9380	0.9485	0.9462	0.00007	0.00006
80	0.8825	0.8944	0.8925	0.00013	0.00013
90	0.8057	0.8180	0.8169	0.00024	0.00024
100	0.7096	0.7215	0.7215	0.00036	0.00036
110	0.5993	0.6105	0.6117	0.00045	0.00045
120	0.4824	0.4931	0.4951	0.00047	0.00046
130	0.3679	0.3782	0.3805	0.00042	0.00041
140	0.2642	0.2742	0.2762	0.00033	0.00033
150	0.1776	0.1870	0.1885	0.00025	0.00025
160	0.1111	0.1194	0.1203	0.00017	0.00017
170	0.0643	0.0711	0.0716	0.00011	0.00011
180	0.0342	0.3926	0.0395	0.00006	0.00006
190	0.0166	0.0200	0.0202	0.00003	0.00003
200	0.0074	0.0094	0.0096	0.00001	0.00001

3.2 Gamma Distribution

In this section, Bayes estimators of the parameters and reliability function for the three-parameter gamma distribution are derived and compared with those of the maximum likelihood approach. Methods due to Lindley (1980) and Weiss (1988) are presented. Separate results for the Tierney-Kadane approach are omitted because, as has been seen earlier with other distributions, the two Bayes approximation methods produce almost indistinguishable results. Some difficulties encountered due to the inherent problem of working with the origin unknown are encountered which are similar to those seen with the three-parameter Weibull distribution.

The gamma distribution, also referred to as a Pearson's Type III distribution, can take on a wide diversity of shapes ranging from a near-normal to an extraordinarily positively skewed distribution. It is seen more commonly in its two-parameter form, due largely to the mathematical complications that arise with the addition of a threshold parameter. The generalization to three parameters was first proposed by Stacy and Mihram (1965). Johnson and Kotz (1980) give an extensive list of references which can be supplemented by a definitive text on estimation with the gamma distribution by Bowman and Shenton (1988). In life testing situations, the three-parameter gamma distribution is used typically when the exponentiality of a process is in doubt or as a substitute for the three-parameter lognormal distribution.

A particular member of the three parameter gamma family is specified by the triad of (α,β,c) where α is the shape parameter, β is the scale parameter and c is the threshold or origin. The concept of a threshold parameter has direct application in reliability theory as it typically refers to a minimum lifetime past which a component is

guaranteed to survive. The density function is given by

$$f(x|\alpha,\beta,c) = \frac{1}{\Gamma(\alpha)\beta}\left[\frac{x-c}{\beta}\right]^{\alpha-1}\exp\left[-\frac{(x-c)}{\beta}\right] \qquad (3.1)$$

where $\alpha,\beta > 0$ and $x > c$.

If $\alpha = 1$, the resultant distribution is exponential and for positive values of α the distribution is often referred to as the Erlang. As α is increased the relative normality of the distribution is augmented. Larger values of α also tend to make the task of parameter estimation more tenable.

In this section the relative merit of the maximum likelihood approach is compared with that of the Bayesian paradigm. Both the systemic parameters and reliability function are studied. Single sample estimation results are presented and a Monte Carlo simulation study is undertaken.

3.2.1 Estimation Problems

No text dealing with the three-parameter gamma distribution can avoid mentioning the myriad of problems that arise in attempting to estimate the three systemic parameters. The problems are of such a nature as to be a likely cause of the relative lack of attention given to the three-parameter case in favour of the more convenient two parameter case. The addition of the threshold parameter, while intuitively appealing for a wide variety of applications, complicates any estimation process considered, involving polygama and incomplete gamma functions which do not lend themselves readily to closed form algebraic solutions.

Bain (1978) mentions that the maximum likelihood estimation(MLE) approach

produces "probably the best estimates available", but concedes that the method is particularly difficult to implement for the three-parameter gamma distribution. Johnson and Kotz (1980) suggest that the mle's only be used for $\alpha > 2.5$ as the mle approach becomes impractical otherwise due to frequently unobtainable solutions. If $\alpha < 1$, the likelihood tends towards $-\infty$ as $\hat{c} \to x_{min}$ and the estimators fail to exist.

Smith (1985) shows that for this and other distributions of a similar form (notably the three-parameter Weibull distribution) if the shape parameter (here α) is greater than two, the maximum likelihood estimators have the same properties as the regular cases (i.e. efficient and asymptotically normal). For shape parameters less than two the situation degenerates into a case-by-case inspection, but in general does not have the desirable properties.

Moment estimators that perform reasonably well in the two-parameter case are subject to extreme variability when the threshold parameter is added. This is due to the required third component in the system of equations that produces moment estimators involving the third moment. Both the third and fourth moments of the distribution are unstable (Cohen and Whitten, 1988). Although sometimes usable as starting points in iterative searching for the mle, the moment estimators have little practical value.

Most notable in the development of estimation procedures for the three-parameter gamma distribution are the works of Harter and Moore (1965, 1967) which was expanded in the collected works of Harter (1969). Moment, mle and modified moment estimators complete with algorithmic instructions can be found from these sources.

Cohen and Whitten (1988) apply such modified methods of both the moment and maximum likelihood approaches by replacing one of the systemic equations with a

simpler form. The modified maximum likelihood estimators of Cohen and Whitten (1988) have been seen to be quite usable when $\alpha < 2$, which circumvents the problems with the mle's. Kappenman (1985) substituted an equation involving convenient percentile relations to produce estimators that always exist and seem to compare favourably with the mle's.

3.2.2 Maximum Likelihood Estimation

In this section the maximum likelihood approach is explored and an extension of the method by Cohen and Whitten (1988) for finding the mle's is proposed. This does not address the problem that the mle's do fail to exist for some samples, but ensures that if they do exist, they are obtainable.

The log-likelihood function corresponding to the three-parameter gamma pdf (3.5) is

$$L(\alpha, \beta, c \mid \underline{x}) = -n\log[\Gamma(\alpha)] - n\alpha\log\beta + (\alpha-1)\sum_{i=1}^{n} \log(x_i - c) - \sum_{i=1}^{n} \frac{(x_i - c)}{\beta}$$

and the system of first partial derivatives which must be solved to obtain the maximum likelihood estimators is

$$L_1 = \frac{\partial L}{\partial \alpha} = -n\psi(\alpha) + \sum_{i=1}^{n} \log\left[\frac{x_i - c}{\beta}\right]$$

$$L_2 = \frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{(x_i - c)}{\beta^2} - \frac{n\alpha}{\beta} \qquad (3.2)$$

$$L_3 = \frac{\partial L}{\partial c} = \frac{n}{\beta} - (\alpha - 1)\sum_{i=1}^{n} \frac{1}{(x_i - c)}$$

where $\psi(\alpha)$ is the digamma function, defined as

$$\psi(\alpha) = \frac{\partial[\log\Gamma(\alpha)]}{\partial\alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

The above system of equations (3.2) does not produce closed form algebraic solutions. Several authors have suggested graphical techniques be used to find initial values to feed into an iterative root-finding algorithm. Such an approach is practical only if a single sample is being considered as it demands a further referral to a graph for each sample.

Analogous to procedures used for the three-parameter lognormal distribution, a reasonable approach is to obtain a feasible estimate for the threshold parameter c and continue in a cyclic fashion to estimate the other two parameters (Stacy and Mihram, 1965). Cohen and Whitten (1988) give such an algorithm for producing mle's while Bowman and Shenton (1988) compare the relative merits of five competing algorithms. All methods eliminate one parameter at a time typically estimating the threshold parameter c first and then solving for the other two parameters directly from one of the

partial derivative equations of (3.2).

An irritating aspect of these algorithms is the requirement of a starting point estimate for the threshold parameter. As one needs to be very close to the final value for the mle of c, it puts up another barrier to finding the mle's. The primary difficulty lies in finding a reasonable first estimate for c because such an estimate typically involves the third sample moment. As mentioned above, the third and fourth sample moments are quite unstable, especially for small samples, and hence have a tendency to produce extreme initial estimates for the threshold parameter.

The method proposed is a minor extension of the method proposed by Cohen and Whitten (1988). Their method in brief is to:

1) find a reasonable first estimate of c, say c_1

2) Calculate α_1 by substituting c_1 into

$$\alpha_i = \left[1 - \frac{n}{(\overline{x}-c_i) \sum_{i=1}^{n} (x_i-c_i)^{-1}} \right]^{-1}$$

3) Calculate β_1 by substituting (α_1, c_1) into $\quad \beta_i = \dfrac{\overline{x}-c_i}{\alpha_i}$

4) Substitute (α_1, β_1, c_1) into the L_1 equation of (3.2) above.

5) Iterate via linear interpolation to solve $L_1 = 0$.

This method merely negates the need for an initial estimate of c, which is a potential stumbling block. This is accomplished by examining the L_1 function over the entire range $0 \leq c < x_{min}$ and using adaptive linear interpolation to find the point of solution. The other values are then substituted into the above formulae.

This method accomplishes two things. First, it allows for the investigation into whether or not a solution for a given sample is available. Bowman and Shenton (1988) provide a table of simulation results showing that as many as 50% of simulated samples will fail to produce a solution to the maximum likelihood equations (3.2), depending upon sample size and parametric settings. Second, it guards against the possibility that a malformed likelihood function will produce a local maximum instead of the correct mle.

Cohen and Whitten (1988) discuss how the L_1 function may not have a zero value over the range $0 \leq c < x_{min}$. In such cases the mle fails to exist. They do mention however that there may well be more than one zero point, or a very flat surface that could point to erroneous values for c due to a poor starting value.

Figure 3.4 provides L_1 functions that are representative of the type of cases possible. The horizontal axis has been scaled as a percentage of each sample's minimum value so that comparisons across samples is possible. Types one and three are examples of small samples that provide a solution to $L_1 = 0$ and hence produce mle's. Note that although the type one case L_1 function is well-behaved and should provide no problems for any numerical search routine, the type three case would cause many routines to head towards $c = 0$ and not produce a solution unless the starting value used was close enough to x_{min} so that L_1 is positive.

The type two case function illustrates the situation where no solution to L_1 exists. Here the L_1 function remains above zero. The type four curve, the large sample case, is monotone increasing and provides an easily found solution.

Figure 3.4

THREE-PARAMETER GAMMA ESTIMATION
BEHAVIOR OF L(1) FUNCTION VERSUS THRESHOLD

L(1)

% OF MINIMUM VALUE

TYPE ✳–✳–✳ 1   ◇–◇–◇ 2   +–+–+ 3   ⊟–⊟–⊟ 4
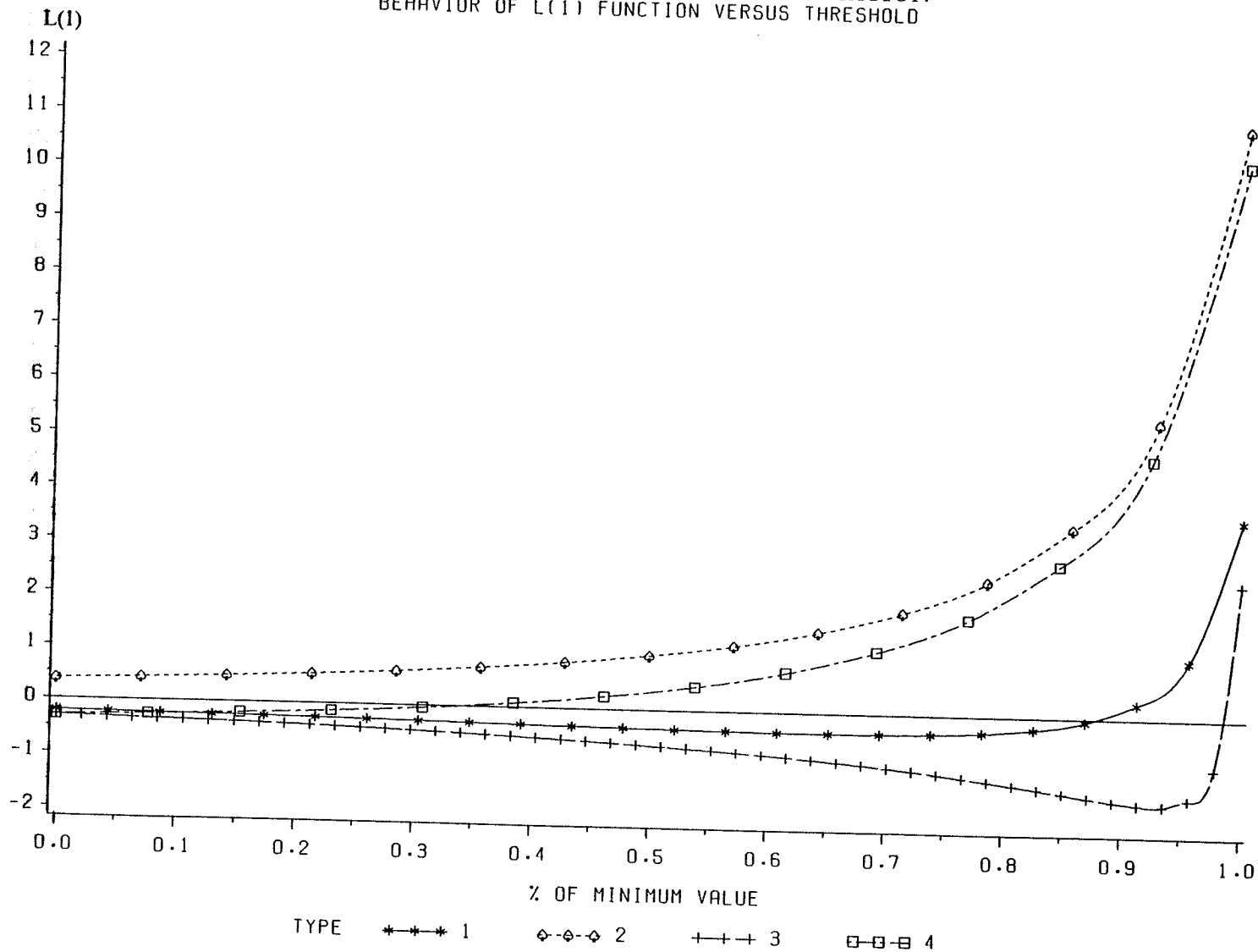
True asymptotic variances and covariances for the mle's are given in Parr and Webster (1965). The second partial derivatives, needed for producing the sample information matrix are

$$L_{11} = \frac{\partial^2 L}{\partial \alpha^2} = -n\psi'(\alpha)$$

$$L_{22} = \frac{\partial^2 L}{\partial \beta^2} = \frac{n\alpha}{\beta^2} - 2\sum_{i=1}^{n} \frac{(x_i - c)}{\beta^3}$$

$$L_{33} = \frac{\partial^2 L}{\partial c^2} = -(\alpha - 1)\sum_{i=1}^{n} \frac{1}{(x_i - c)^2}$$

$$L_{12} = \frac{\partial^2 L}{\partial \alpha \partial \beta} = -\frac{n}{\beta}$$

$$L_{13} = \frac{\partial^2 L}{\partial \alpha \partial c} = -\sum_{i=1}^{n} \left[\frac{1}{(x_i - c)} \right]$$

$$L_{23} = \frac{\partial^2 L}{\partial \beta \partial c} = -\frac{n}{\beta^2}$$

where $\psi'(\alpha)$ is the trigamma function. Several authors have noted that the sample information matrix is very unstable, especially for small sample sizes. Since Lindley's method relies heavily on this information, it could cause the approach to be more variable than desired.

3.2.3 Lindley's Approximation

As this is the fourth distribution for which Lindley's method is applied in this dissertation, the more technical aspects will be described in brief. For a more complete

discussion of the mechanics of the approximation the reader is referred to earlier sections.

Jeffreys' invariant prior approach is employed, analogous to that seen for the three-parameter Weibull, to produce the joint prior distribution

$$\pi(\alpha, \beta, c) \propto \frac{1}{\beta} \ .$$

Hence the log-prior function is $\rho(\alpha, \beta, c) = -\log\beta$, with partial derivatives

$$\rho_1 = \frac{\partial \rho}{\partial \alpha} = 0, \ \rho_2 = \frac{\partial \rho}{\partial \beta} = -\frac{1}{\beta}, \ \rho_3 = \frac{\partial \rho}{\partial c} = 0 \ .$$

An attempt was made to produce simplified algebra for the Lindley equation as was seen in the section on the three-parameter Weibull. Unfortunately the formulae involved for the three-parameter gamma distribution do not lend themselves well to such a solution. Although a final form was achieved, it was longer than the original approximation formula and as such impractical. Work was also done on the T-K approximation, with the algebra even more intractable. Once again the computer must be used to produce estimates on a case by case basis.

The L_{ijk} terms necessary for the Lindley expansion are

$$L_{111} = \frac{\partial^3 L}{\partial \alpha^3} = -n\psi''(\alpha)$$

$$L_{222} = \frac{\partial^3 L}{\partial \beta^3} = 6\sum_{i=1}^{n} \frac{(x_i - c)}{\beta^4} - \frac{2n\alpha}{\beta^3}$$

$$L_{333} = \frac{\partial^3 L}{\partial c^3} = -2(\alpha - 1)\sum_{i=1}^{n} \frac{1}{(x_i - c)^3}$$

$$L_{122} = \frac{\partial^3 L}{\partial \alpha \partial \beta^2} = \frac{n}{\beta^2}$$

$$L_{223} = \frac{\partial^3 L}{\partial \beta^2 \partial c} = \frac{2n}{\beta^3}$$

$$L_{133} = \frac{\partial^3 L}{\partial \alpha \partial c^2} = -\sum_{i=1}^{n} \frac{1}{(x_i - c)^2}$$

where $\psi''(\alpha)$ is the tetragamma function. All other L_{ijk} terms are zero. Specifically, these are L_{112}, L_{113}, L_{233}, L_{123} and their permuted subscripts.

The final piece to Lindley's approximation are the u_i and u_{ij} functions. For estimating the systemic parameters three sets of u functions are needed, namely $u(\alpha, \beta, c) = \alpha$, β and c respectively. In doing so one may notice that for any one of these functions all of the second partial derivatives (u_{ij} functions) will be zero. Further, the u_i functions will be zero for all values of i except where the index matches the parameter to be estimated, in which case the u_i function equals one.

In implementing the approximation techniques, one of the hurdles to surmount was the evaluation of the di-, tri- and tetragamma functions. After researching the

literature of the time, it was found that no standard algorithm existed for evaluation of the polygamma functions. The IMSL set of routines does contain algorithms for the log-gamma and the digamma functions, however no algorithms are provided for the trigamma or higher order functions (see Rice (1983)). The genesis of such an algorithm had to be undertaken before Lindley's method could be implemented.

Abramowitz and Stegun (1965) present an asymptotic approximation to the polygamma functions on page 260 of their reference as

$$\psi^{(n)}(z) \approx (-1)^{n-1}\left[\frac{(n-1)!}{z^n} + \frac{n!}{2z^{n+1}} + \sum_{k=1}^{\infty} B_{2k}\frac{(2k+n-1)!}{(2k)!\,z^{2k+n}}\right]$$

where B_i is the ith Bernoulli number. Unfortunately the approximation is not terribly accurate if $|z| < \pi$. To construct an all-encompassing algorithm, the recurrence relation for the polygamma functions is used, namely,

$$\psi^{(n)}(z+1) = \psi^{(n)}(z) + (-1)^n n!\, z^{-n-1}$$

which can also be found in Abramowitz and Stegun. Testing of the algorithm revealed that using the first ten terms of the recurrence relation gave more than sufficient accuracy for most practical applications. To ensure a degree of accuracy that would be usable under virtually any conditions, the final algorithm uses the first fifteen terms.

The POLYGAMA algorithm was then tested extensively for various values of z. The algorithm successfully duplicated, to the displayed number of significant digits, Tables 6.1 through 6.5 in Abramowitz and Stegun. While preparing a paper concerning this algorithm, the author discovered that Balakrishnan, working on an entirely different problem, had arrived at the same result (personal communication, March 1986) and

submitted a paper for publication. More recently Cohen and Whitten (1988) present the same algorithm in their text.

The reliability function R(t) for the three-parameter gamma distribution is

$$R(t) = Pr(T \geq t) = \int_t^{\infty} \frac{1}{\Gamma(\alpha)\beta} \left(\frac{x-c}{\beta}\right)^{\alpha-1} \exp\left[-\frac{(x-c)}{\beta}\right] dx \ .$$

Maximum likelihood estimation of R(t) is straightforward using the principle of invariance and merely substituting the mle's of the systemic parameters into the reliability function.

Bayes estimation is once again more complicated due to mathematical intractability. An approximation may be used as before as well as the possibility of a nontraditional approach proposed by Weiss (1988) which states that although the Bayes estimators may not be invariant, substituting the Bayes estimators analogous to the mle does produce estimators that are Bayesian in nature. This approach has been criticized as being "quasi-Bayesian" and is somewhat controversial. As stated at the outset of the dissertation, however, the goal here is to explore different roads toward the same destination, not to question the quality of the asphalt.

Setting $y = (x-c)/\beta$, R(t) becomes

$$R(t) = \int_{\frac{t-c}{\beta}}^{\infty} \frac{1}{\Gamma(\alpha)} y^{\alpha-1} \exp(-y) \, dy, \quad \alpha > 1 \ .$$

In implementing Lindley's method, for convenience the u(θ) function is written as follows

$$u(\alpha,\beta,c) = \int\limits_0^{\frac{t-c}{\beta}} \frac{1}{\Gamma(\alpha)} y^{\alpha-1} \exp(-y) \, dy$$

so that Lindley's approximation to the Bayes estimator will be $R^*(t)=1-u(\alpha,\beta,c)$.

To construct the appropriate u_i and u_{ij} functions, the following relation from Protter and Morrey (1966) is used whereby if

$$I = \int\limits_{v(x)}^{w(x)} f(x,t) \, dt$$

then

$$\frac{\partial I}{\partial x} = f(x,w) \frac{\partial w}{\partial x} - f(x,v) \frac{\partial v}{\partial x} + \int\limits_v^w \frac{\partial f(x,t)}{\partial x} \, dt \quad .$$

Now set $I=u(\alpha,\beta,c)$ to construct the u_i functions

$$u_1 = \frac{\partial u}{\partial \alpha} = \frac{1}{\Gamma(\alpha)} \int\limits_0^{\frac{t-c}{\beta}} y^{\alpha-1} \exp(-y) \log(y) \, dy - \psi(\alpha) u$$

Note that the part of u_1 that is an integral is the incomplete digamma function.

$$u_2 = \frac{\partial u}{\partial \beta} = - \frac{1}{\Gamma(\alpha)\beta} \left(\frac{t-c}{\beta}\right)^{\alpha} \exp\left(\frac{c-t}{\beta}\right)$$

$$u_3 = \frac{\partial u}{\partial c} = \left(\frac{\beta}{t-c}\right) u_2$$

and the u_{ij} functions

$$u_{11} = \frac{\partial^2 u}{\partial \alpha^2} = -2\psi(\alpha)u_1 - \left([\psi(\alpha)]^2 + \psi'(\alpha)\right)u$$

$$+ \int_0^{\frac{t-c}{\beta}} \frac{1}{\Gamma(\alpha)} y^{\alpha-1} \exp(-y) [\log(y)]^2 dy \; .$$

Note that the integral part of u_{11} is the incomplete trigamma function.

$$u_{22} = \frac{\partial^2 u}{\partial \beta^2} = \frac{u_2}{\beta}\left[\frac{t-c}{\beta} - \alpha - 1\right]$$

$$u_{33} = \frac{\partial^2 u}{\partial c^2} = \frac{u_2}{(t-c)}\left[1 - \frac{\beta(\alpha-1)}{t-c}\right]$$

$$u_{12} = \frac{\partial^2 u}{\partial \alpha \partial \beta} = u_2\left[\log\left(\frac{t-c}{\beta}\right) - \psi(\alpha)\right]$$

$$u_{13} = \frac{\partial^2 u}{\partial \alpha \partial c} = \left(\frac{\beta}{t-c}\right)u_{12}$$

$$u_{23} = \frac{\partial^2 u}{\partial \beta \partial c} = \frac{u_2}{t-c} + \left(\frac{\beta}{t-c}\right)u_{22}$$

3.2.4 Examples

Application data for the three-parameter gamma distribution is absent in the literature. Harter (1969), Kappenman (1985) and Cohen and Whitten (1988) all have exemplary samples in their presentations, but they are all the result of computer simulation. Harter's so-called G3 sample was used for comparison of the Bayes approximation with the previously published results.

Generated from $\Gamma(3,50,20)$, the sample of size 40 does pass the Anderson-Darling (1954) empirical frequency distribution(edf) goodness of fit test, but is a typically problematic sample from a parameter estimation viewpoint in that there is a seeming outlier as the smallest sample value. This problem is a result of the skewed nature of the ditribution and causes difficulty in estimating the threshold parameter in that the L_1 function will be of case type two from Figure 3.4. Harter proposes to set $\hat{c}=0$ under such circumstances. The mle and Lindley estimates of the parameters are presented in Table 3.7 below.

Table 3.7: Estimation Results For Harter's (1969) G3 Sample

PARAMETER	TRUE	MLE	LINDLEY
α	3.00	3.59	5.36
β	50.00	48.87	40.71
c	20.00	0.00	-33.05

Clearly the Bayes estimates are unacceptable. The information given by the Bayes estimates suggest a totally different parametric environment from that of the mle results. Variance approximations for the Lindley estimators were also negative for α and c, similar to what was seen for the three-parameter Weibull.

Reliability estimation proceeds remarkably well despite the above results for the systemic parameters, analogous to the three-parameter Weibull. Figure 3.5 indicates that over a large range of the sample space, Lindley's approximation is as close to the true value as the mle. All approaches overestimate the true reliability for the given sample, with Weiss' estimate being highest. At no point do the mle and Lindley estimates differ by more than 1.5%.

Figure 3.5   THREE–PARAMETER GAMMA RELIABILITY ESTIMATION

HARTER"S G3 PROBLEM SAMPLE
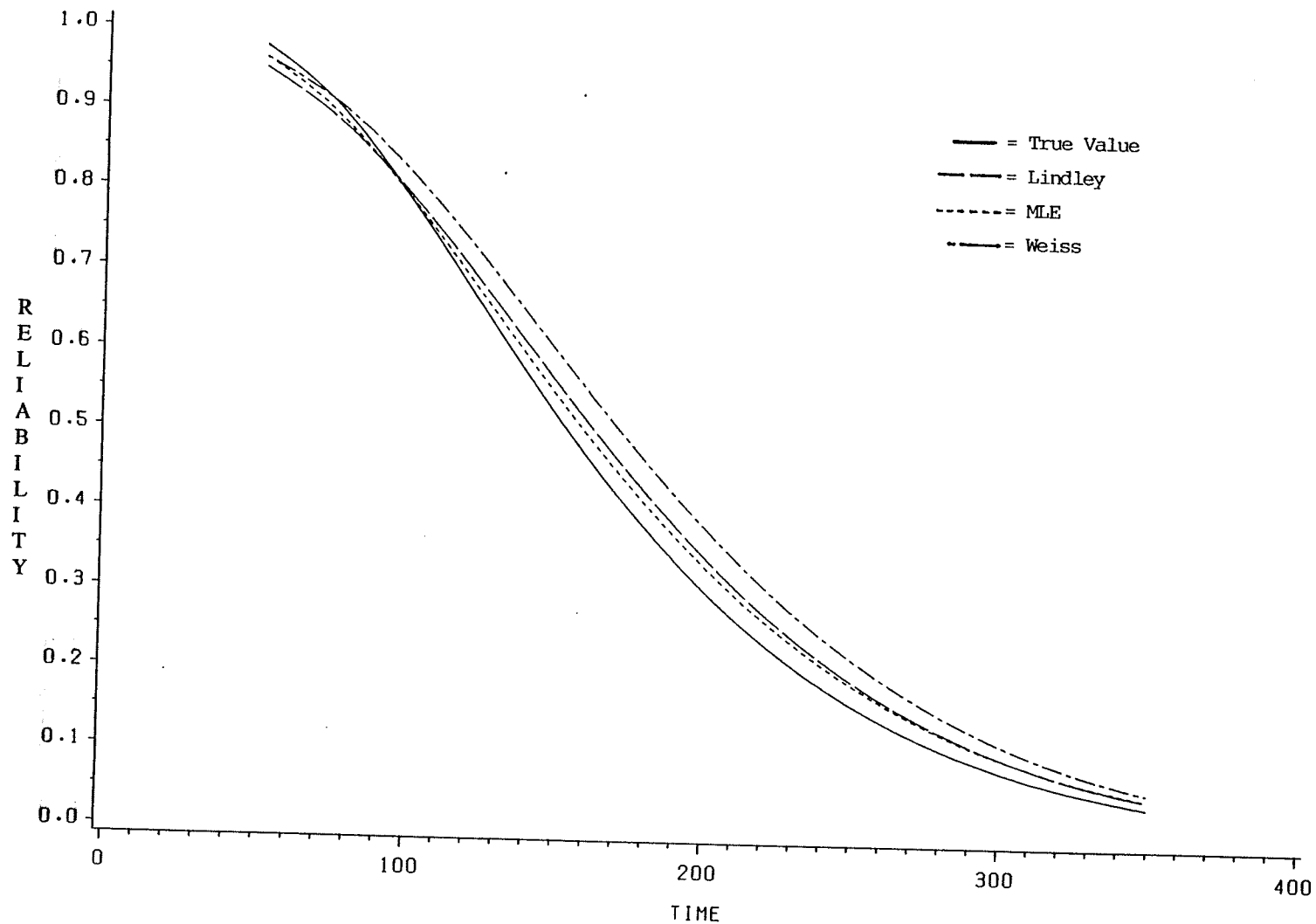
The above sample was put forward as a "problem sample" to demonstrate the difficulties in estimation with the three-parameter gamma distribution. Acknowledging such difficulties exist, the goal here is to demonstrate that for "non problem" samples the Bayes approximation is a usable approach. As such the simulations presented in the following discussion use a large shape parameter value of $\alpha=6.0$ in the hopes that the regularity of the environment will allow for a clearer and fairer comparison of the competing estimation procedures. It is a given premise that "non problem" samples may exist for the three-parameter gamma distribution under very specific parametric circumstances.

Sample generation was accomplished by the IMSL routine GGAMR which produces one-parameter gamma variates and then transforms them into three-parameter gamma variates. The following sample of fifty observations was drawn from $\Gamma(\alpha,\beta,c)=(6,50,20)$

118.25	150.19	155.92	161.80	163.09	170.74	174.86
180.34	187.24	190.16	193.88	202.45	204.33	217.63
220.39	220.97	221.79	230.43	237.58	240.61	242.04
256.69	274.04	280.60	288.24	290.41	296.39	303.57
307.57	314.68	327.70	328.28	338.63	339.32	347.27
369.41	378.04	378.68	402.65	422.58	445.03	446.07
467.23	470.51	476.74	477.06	501.78	553.49	558.70
647.12						

The sample proved to give a good fit to the prescribed model using the Anderson-Darling (1954) and Cramer-von Mises (D'Agostino and Stevens (1986)) edf statistics. Estimation

results are given in Table 3.8.

Table 3.8 Estimation Results For Generated Γ(6,50,20) Sample

| PARAMETER	TRUE	MLE	LINDLEY
α	6.00	2.55	4.09
β	50.00	80.85	57.96
c	20.00	101.17	58.35

Even for this "non problem" sample marked differences exist between the mle and approximate Bayes estimators, but note that neither approach looks particularly preferable. The variance estimates for the Bayes approximations are once again negative for α and c and as such are useless.

Figure 3.6 reveals that the reliability estimates are once again reasonable despite what was seen for the systemic parameters. All estimation approaches here underestimate the true reliability for most of the sample range, with the Weiss estimator consistent in its underestimation. The mle and Bayes methods compete across the range much the same way as was seen for the three-parameter Weibull. The Weiss estimator is clearly the weakest of the three approaches, but may be practical as an extremist benchmark in applications.

The effect of sample size is more pronounced for the three-parameter gamma than for other distributions seen to this point. This is due mainly to the poor information a small sample gives about the threshold parameter. Once this parameter is missed, results for the other two parameters are of dubious merit as seen in the above examples. To give a more detailed picture a series run for the above generated sample was undertaken to see how long it would take before the estimates stabilized. Results are given in Table

Figure 3.6

THREE–PARAMETER GAMMA RELIABILITY ESTIMATION
RESULTS FOR GENERATED SAMPLE
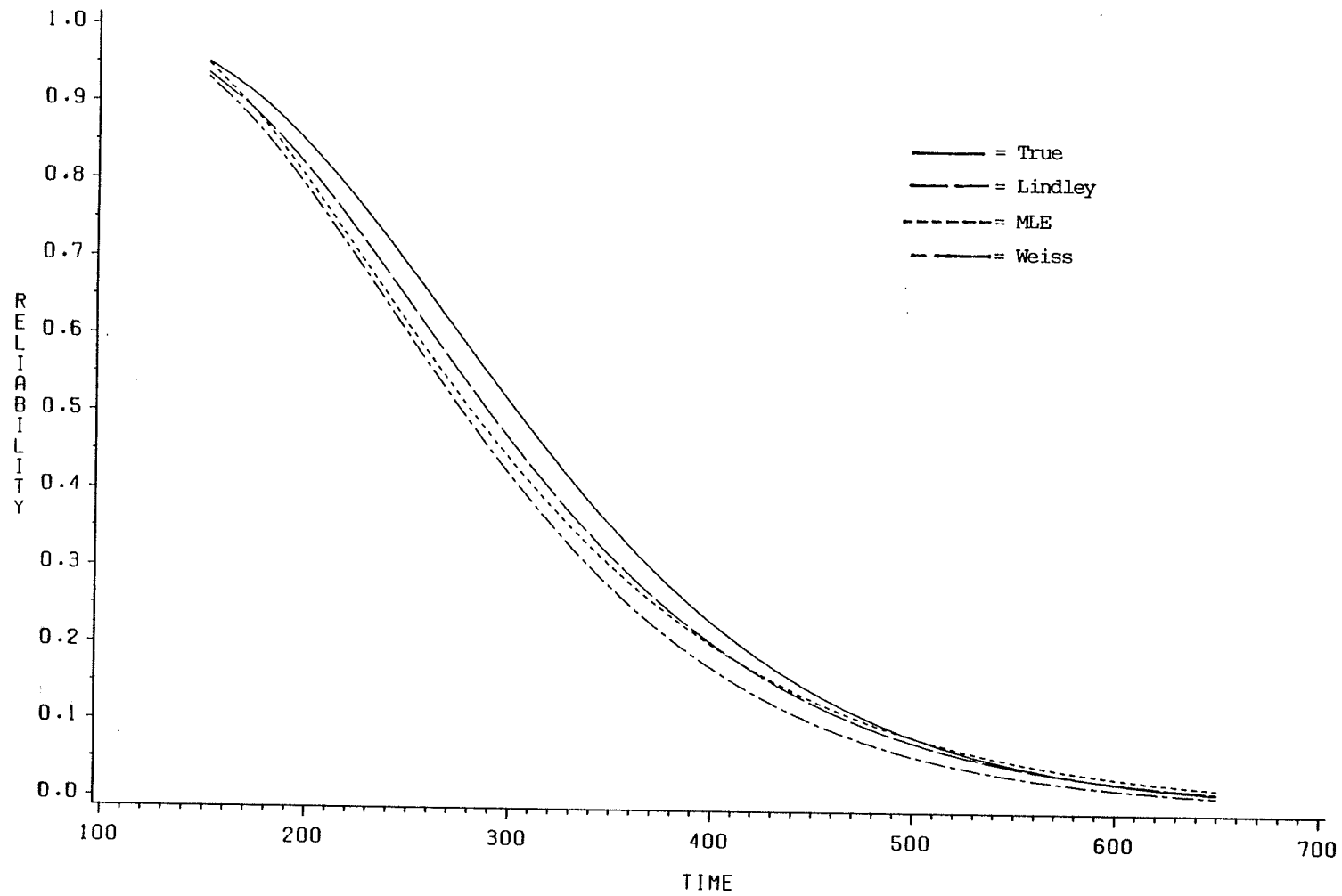
N=50, (ALPHA,BETA,C) = (6,50,20)

3.9 for sample size ranging as large as n=2000.

Table 3.9: Three-parameter Gamma Single Sample Series Run

Parameter Estimates

	ALPHA(SHAPE)			BETA(SCALE)			C(THRESHOLD)		
N	MME	MLE	BAY	MME	MLE	BAY	MME	MLE	BAY
50	8.18	2.55	4.09	43.13	80.85	57.96	-45.47	101.2	58.40
100	10.8	3.12	4.76	37.69	73.49	56.10	-91.89	84.95	40.77
200	7.62	3.45	4.36	46.52	71.36	63.06	-28.70	79.82	55.14
500	8.66	5.39	6.30	41.27	53.09	49.51	-37.13	34.05	14.30
1000	7.05	6.04	6.54	45.52	49.50	47.87	-0.63	21.38	11.05
2000	6.72	5.93	6.16	46.81	50.07	49.30	8.55	26.30	21.51

Negative threshold estimates persist for the method of moments estimator up to n=1000. Differences between the Lindley and mle results remain even for n as large as 2000, which is different from what has been seen previously in using Bayes approximations for other distributions. Clearly the three-parameter gamma distribution is unique in its degree of estimation difficulty. Reliability estimates (not shown for single sample) converge more readily to their true values, reinforcing the previous conclusions drawn from the generated sample.

Negative values for the posterior variances of the parameters are a problem encountered in applying Bayes approximations, especially for small sample sizes. As was seen earlier, by increasing the sample size above n=200 these estimates also stabilize. As can be seen in Table 3.10, however, the negative estimates remain at n=200 and are clearly underestimated for n=500. It is not until 1000 observations are taken that the Bayes posterior variance estimates compare reasonably with the asymptotic variances of the maximum likelihood estimators.

Table 3.10: Asymptotic Versus Estimated Posterior Variance

Effect Of Sample Size Using Generated $\Gamma(6,50,20)$ Samples

	α(SHAPE)		β(SCALE)		C(THRESHOLD)	
N	MLE	BAY	MLE	BAY	MLE	BAY
50	15.18	neg.	402.94	53.74	6647.07	neg.
100	7.56	neg.	201.47	60.14	3323.53	neg.
200	3.78	neg.	100.74	104.37	1661.77	neg.
500	1.51	0.02	40.29	44.70	664.71	382.28
1000	0.76	0.74	20.15	21.98	332.35	333.70
2000	0.38	0.38	10.07	10.08	161.18	161.19

It is important to keep in mind that the discrepancy between the mle and Bayes approaches should not be interpreted as being due to a lack of accuracy in the approximation methodology. Rather, it is indicative that under a noninformative a priori state the Bayes estimator of the systemic parameters does not perform as well as the mle for the three-parameter gamma distribution. Improving prior knowledge may bring the Bayes estimators' performance into line.

3.2.5 Monte Carlo Simulation

To more fully investigate the distributional properties of the Bayes approximation, 1000 samples of size n=100 were run with parameter settings $\Gamma(4,20,50)$. As discussed in previous sections, because the maximum likelihood estimators fail to exist for some samples, screening rules had to be incorporated into the simulation run. A sample was rejected if:

1) the sample skewness was negative (Bowman and Shenton, 1988)

2) the L_1 function was ill-behaved (case type two from Figure 3.4)

3) the Bayes estimator for c was negative

Bowman and Shenton note that for small n and/or large α the rejection rate for generated samples is high. They consider any $n < 100$ to be small. As such the simulations were run with $n = 100$ to investigate the Bayes estimators' relative merit.

A total of 1147 samples had to be generated to produce 1000 usable samples. Of the 147 rejected samples 120 were due to negative Bayes estimators while the remaining 20 had malformed L_1 functions. The negative Bayes estimators often coincided with low p-values for the Anderson-Darling goodness of fit test, so it is uncertain whether or not sampling is the true problem in these cases. These phenomena are similar to previously published results, although other authors did not specify the criteria for sample rejection.

From a practical perspective, if faced with a sample that failed the above criteria, the researcher has a dilemna. Even if only the reliability function is of interest, the Bayes approximate estimator is unobtainable because it is based on the mle of R(t). Given that the other closely related distributions seen earlier in this chapter do provide estimators more readily, it may be advisable to use the lognormal or Weibull distribution as a substitute for the gamma. Such a switch would have to be predicated on a sound theoretical base from the literature that dealt with the particular application under study. Naive switching between theoretical models is not recommended in general.

Results are given in two sections. First, Table 3.11 compares the parametric estimators via the maximum likelihood(MLE) and Lindley approaches. The method of moments estimator(MME) is also given as it is a useful reference point. Second, Table 3.12 presents reliability estimation results for the mle, Lindley and Weiss estimators.

Table 3.11: Simulation Results For 1000 Valid Γ(4,20,50) Samples

Sampling Distribution Mean And Mean Square Error

PARAMETER	MME		MLE		LINDLEY	
	MEAN	MSE	MEAN	MSE	MEAN	MSE
α	5.69	17.09	3.62	1.36	5.50	7.33
β	19.46	50.09	22.20	28.30	17.88	28.46
C	39.82	765.08	54.39	96.29	40.17	338.07

Table 3.12: R(t) Estimates Sampling Distribution Mean And MSE

TIME	TRUE	MLE		LINDLEY		WEISS	
		MEAN	MSE	MEAN	MSE	MEAN	MSE
75	.9617	.9657	.00024	.9572	.00022	.9542	.00056
100	.7576	.7527	.0014	.7637	.0012	.7498	.0023
125	.4838	.4771	.0018	.4926	.0020	.4745	.0034
150	.2650	.2628	.0013	.2681	.0014	.2544	.0026
175	.1303	.1319	.0007	.1294	.0008	.1215	.0013
200	.0591	.0623	.0003	.0578	.0003	.0537	.0005
225	.0253	.0283	.0001	.0247	.0001	.0226	.0001
250	.0103	.0126	.00004	.0104	.00003	.0093	.00004
275	.0041	.0055	.00001	.0044	.00001	.0037	.00001
300	.0016	.0024	.000005	.0019	.000003	.0015	.000002

In terms of parameter estimates, it is not surprising that the method of moments estimator approach is the worst of the three. It is surprising however that the mle results are clearly superior to Lindley's, both in terms of deviation from the true value and mean square error. Lindley's is especially poor in estimating the shape and threshold parameters.

Entirely different findings are evident for the reliability estimation. Here once again the Bayes and mle approach compete across the range space, with neither clearly superior. The Bayes method seems slightly better when estimating in the upper tail of the sample space in terms of bias. Mean square errors are virtually equal. The Weiss estimator is consistently smaller than both the true reliability and the other estimators.

Reliability results for the mean of the sampling distributions at each time point are given in Figure 3.7. Again the Bayes estimator and the mle compete well in terms of estimating the true reliability. The Weiss estimator reliability curve is uniformly below the true reliability and the other two estimators.

The threshold parameter's mle and Lindley estimators' sampling distributions are depicted in Figures 3.8 and 3.9. Although similar in shape, the distribution of the approximation to the Bayes estimator is clearly more variable. From these pictures one can see that it is a general, consistent failure of the Bayes estimator in that the adjustment it makes to the mle is overly large. It is not merely a few cases that have a special characteristic about them that causes the Bayes approximation to produce a markedly different result from the mle.

Results for Bayes approximations applied to the three-parameter gamma distribution differ somewhat from what was seen for the other distributions in this chapter. Often in life-testing studies, models involving the three-parameter gamma, lognormal and Weibull distributions will be used interchangeably because the parametric settings can be arranged to mimic one another. Given the poor performance of the approximation seen above, the Bayesian approach cannot be suggested as a viable option for estimating the three systemic parameters. It can be used successfully to estimate the

Figure 3.7   THREE—PARAMETER GAMMA RELIABILITY ESTIMATION

ESTIMATOR AVERAGES FOR 1000 SAMPLES

Figure 3.8

ESTIMATION FOR THREE-PARAMETER GAMMA THRESHOLD C
RESULTS FOR 1000 GENERATED SAMPLES
N=100, α=4.0, β=20.0, c=50

PERCENTAGE

MAXIMUM LIKELIHOOD ESTIMATOR
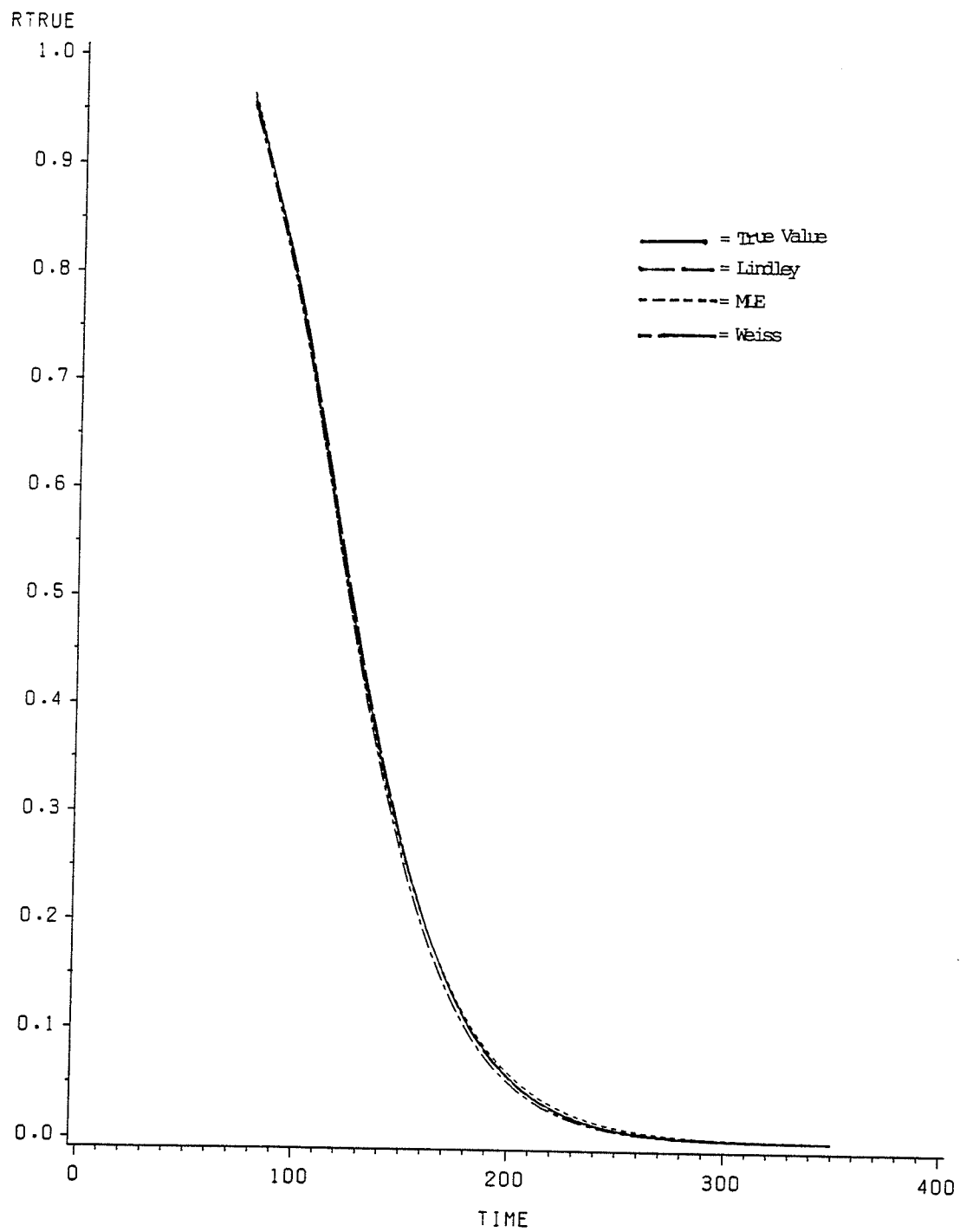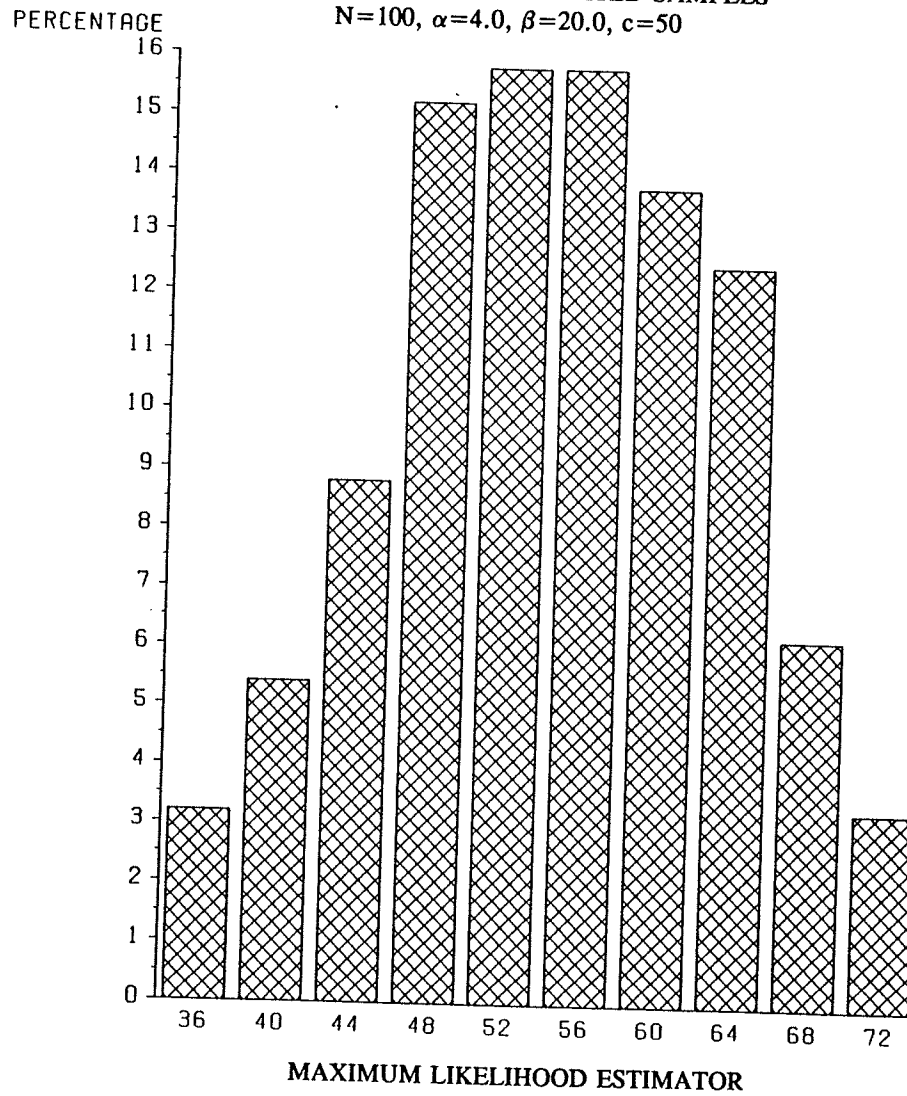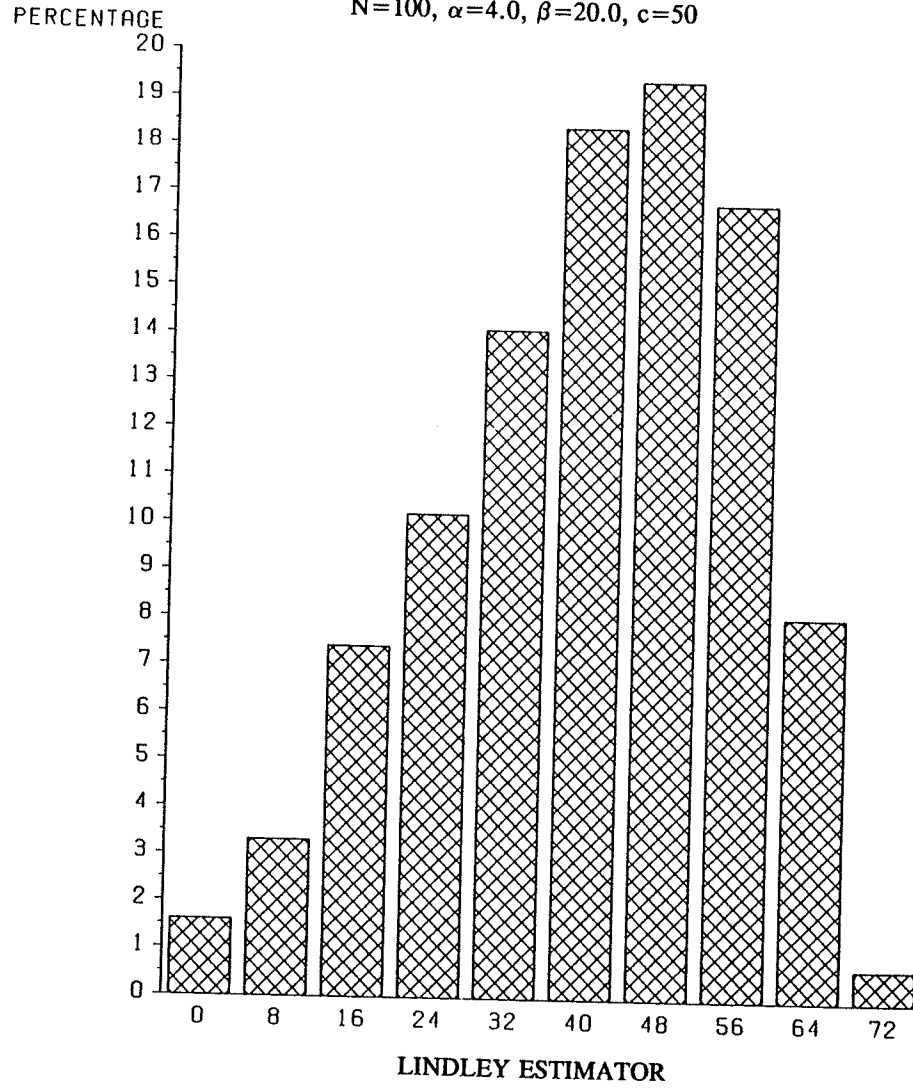
Figure 3.9

ESTIMATION FOR THREE-PARAMETER GAMMA THRESHOLD C
RESULTS FOR 1000 GENERATED SAMPLES
$N=100$, $\alpha=4.0$, $\beta=20.0$, $c=50$

LINDLEY ESTIMATOR

system reliability, where it compares favourably with the maximum likelihood approach.

Again it should be stressed that the Bayes approach was examined under the worst possible conditions by assuming a noninformative prior. As such it is possible that results would improve for situations where good prior information is available. It is important to keep in mind, however, that such conditions were present in applying this approximation to other life-testing distributions with much better results.

The problem seems inherent with the Bayes estimator itself as the T-K method, of approximating the Bayes estimator encounters the same difficulties. Negative values for systemic parameter and variance estimates have been observed in other situations.

3.3 Lognormal Distribution

In this section, Bayes estimators of the parameters and the reliability function for the three-parameter lognormal distribution are derived and compared with the maximum likelihood estimators. The Tierney-Kadane approach is used for the systemic parameter estimates, while Lindley's approach is used to estimate R(t).

Separate results for Lindley's approach applied to estimating the parameters of the lognormal distribution $\theta = (k,\mu,\sigma)$ are omitted because, as has been seen earlier with other distributions, the two Bayes approximation methods produce indistinguishable findings. Furthermore, work was recently published by Lye, Sinha and Booy (1988) using Lindley's approximation for flood data assumed to follow a three-parameter lognormal distribution. Work with Lindley's approximation and the lognormal distribution by the author predates the work of Lye et al. Their paper was published after consultation with the author and the paper in fact refers to the author's prepublished work with the three-parameter Weibull distribution. Lye et al cover the application of Lindley's method sufficiently well so that any attempt to describe it here would be superfluous.

In any event, the Lye et al paper does not address the Tierney-Kadane approach, although as seen in previous sections, the results are basically equivalent except for very small samples. Further, in this section the estimation of the reliability function is undertaken.

Papers dating back to 1879 include discussion of a distribution whose logarithm is normally distributed. The typical genesis of the distribution is attributed to life testing environments where the degree of variability in the data is related to the value observed

(Cohen, 1951). A popular application of the distribution is incumbent upon regression theory where an appropriate stabilizing or normalization transformation is to work with the logarithms of the observed values. Other popular areas of application besides life testing range from agriculture to economics. Extensive application lists are available from the books by Crow and Shimizu (1988), Aitchison and Brown (1957) and in Chapter 14 of Johnson and Kotz (1970).

The three-parameter lognormal pdf is

$$f(x|k,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma(x-k)}\exp\left[-\frac{1}{2\sigma^2}[\log(x-k)-\mu]^2\right]$$

such that $x > k$, $\sigma > 0$ and $-\infty < \mu < \infty$.

The distribution parameters are typically described in this way to indicate that x-k is a variate whose logarithm is $N(\mu,\sigma)$ where μ and σ are the usual location and shape parameters. In the lognormal pdf, however, μ is a scale parameter and σ is a shape parameter. Larger values of produce a more spread out distribution. Very small values of σ generate an almost exponential-like curve whereas as σ increases the distribution's skewness becomes greater. For this section then our vector systemic parameter is $\underline{\theta} = (k,\mu,\sigma)$.

3.3.1 Estimation Problems

There is a long history of difficulty recorded in the literature of solving the system of likelihood equations to obtain the mle's of k, μ and σ. As with the previously seen Weibull and gamma distributions, work with the lognormal distribution is complicated when a third systemic parameter representing a threshold or minimum time

until failure is added to the mathematical environment. Much of the literature deals with the simpler situations of having one or more of the parameters known a priori. The remainder of the published materials dedicates itself to circumventing the intractability difficulties due to the added threshold parameter. As was seen with the other three-parameter distributions earlier in this chapter, there is as yet no really satisfying estimation procedure for the three-parameter lognormal that works in all cases for all parametric values.

The primary difficulty once again lies in the fact that an absolute maximum for the likelihood function may or may not exist for a given sample. Hill (1963) shows that iterative solutions arrived at from some numerical process are actually only local maximums. As was seen for the three-parameter Weibull, there is a path which exists for the mle of k to tend towards infinity while the likelihood function remains bounded and roughly constant (Hill, 1963). It turns out that the true value that maximizes the likelihood occurs at $(\hat{k}, \hat{\mu}, \hat{\sigma}) = (x_{[1]}, -\infty, \infty)$. Clearly the true mle would be an absurd estimator to use in practice.

The use of a local maximum as the mle has become standard practice and has been shown to have most of the well known asymptotic properties of the mle (Harter and Moore, 1966). From a Bayesian perspective, Hill (1963) notes that a local maximum will suffice because the goal is to find a "region of high posterior probability". The Bayes approach would hence downplay the classical true maximum as an extreme and therefore unlikely case in terms of posterior probability.

Hill also describes in some detail how the problem of finding even the local maximum is complicated in small samples, especially if the threshold parameter k is not

estimated with care. In essence it occurs that the local maximum is attained at a point with a k value very close to the sample minimum $x_{[1]}$. If the iterative procedure used begins by underestimating k, the process will overshoot the interval of convergence and send the estimate off towards the global maximum at $k = -\infty$.

This result sparked a flurry of activity for a method which would circumvent this problem. Richards (1961) suggests a method whereupon one fixes the value of one parameter and then proceeds in a cyclical fashion to solve for the other parameters. Unfortunately, the definition of what constitutes a "suitable" estimate for the troublesome threshold parameter is not obtained so that one is left with a hunt and peck approach for any given sample. Harter and Moore (1966) discover that for some samples not even a local maximum will exist for the likelihood function. In such situations, they suggest the method of censoring the first order statistic and using its value as an upper bound for the iterative process used to find the estimate for k.

Calitz (1973) makes some very useful contributions regarding the choices of initial values for the iteration process. He suggests setting $k=0$ as an appropriate starting point and bracketing the search with $k = \max(0.6x_{[1]}, x_{[1]}-.1)$.

Giesbrecht and Kempthorne (1976) provide another alternative involving the discretization of the sample. This allows for the approximation of the lognormal by the multinomial distribution as an alternative form of the likelihood function. Unfortunately, this method also suffers from the possibility that no set of maximizing values will be obtainable.

Cohen and Whitten (1980) apply their modified moment and modified mle method used for various distributions. Again they substitute a simpler equation for one of the

partial derivatives of the log-likelihood function. In doing so, estimators will always exist and have most of the desirable properties of the mle. Unfortunately, these estimators may also be hard to find numerically. In general they found that if the lognormal sample was symmetric or slightly negatively skewed, the numerical procedure would fail to converge. The more positively skewed the sample, the more regular the results became. Since the lognormal distribution is put into practical use mainly where data are extraordinarily positively skewed (Johnson and Kotz, 1970), a sample which evidences a lack of skewness or slight negativism of skew could be regarded as an anomaly and would be better represented by a distribution other than the lognormal. This a posteriori model decision-making process limits the scope of this procedure.

Amin (1981) give an excellent review of the difficulties encountered in implementation of the mle process and discusses various means of attacking the problems. He also suggests a method of surmounting the computational difficulties in a modification of Cohen's (1951) technique. Unfortunately, his technique also runs into the problem of needing a very good initial estimate for the threshold parameter k to avoid heading off towards $-\infty$.

To further demonstrate the basis of numerical problems encountered in maximizing the likelihood function, Figures 3.10, 3.11 and 3.12 present the log-likelihood surfaces in the area around the maximum likelihood estimates for the given sample used in subsequent subsections. In all three figures the actual local maximum (the mle) is located on or near the centre of the surface grid. The problem in searching for the maximization point involving the dimension of k is immediately evident from Figures 3.10 and 3.11. The drop towards negative infinity as k increases begins very

THREE−PARAMETER LOGNORMAL LOG−LIKELIHOOD FUNCTION
K VERSUS MU MAXIMIZATION

LOG(L)

46.02

31.34

16.66

1.98

6.07

5.98

5.89

5.80

K

-1.14

-1.26

MU

-1.38

-1.50
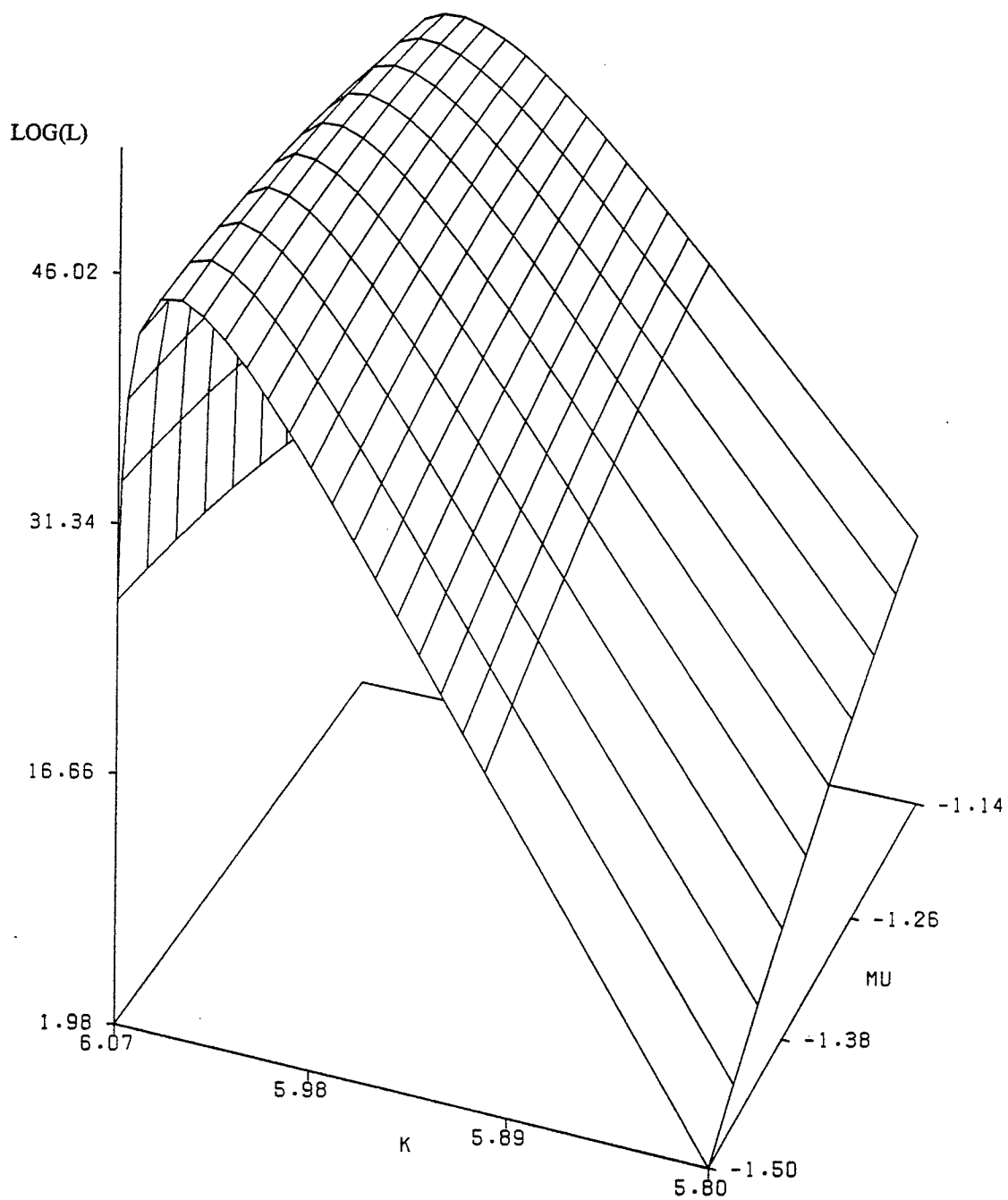
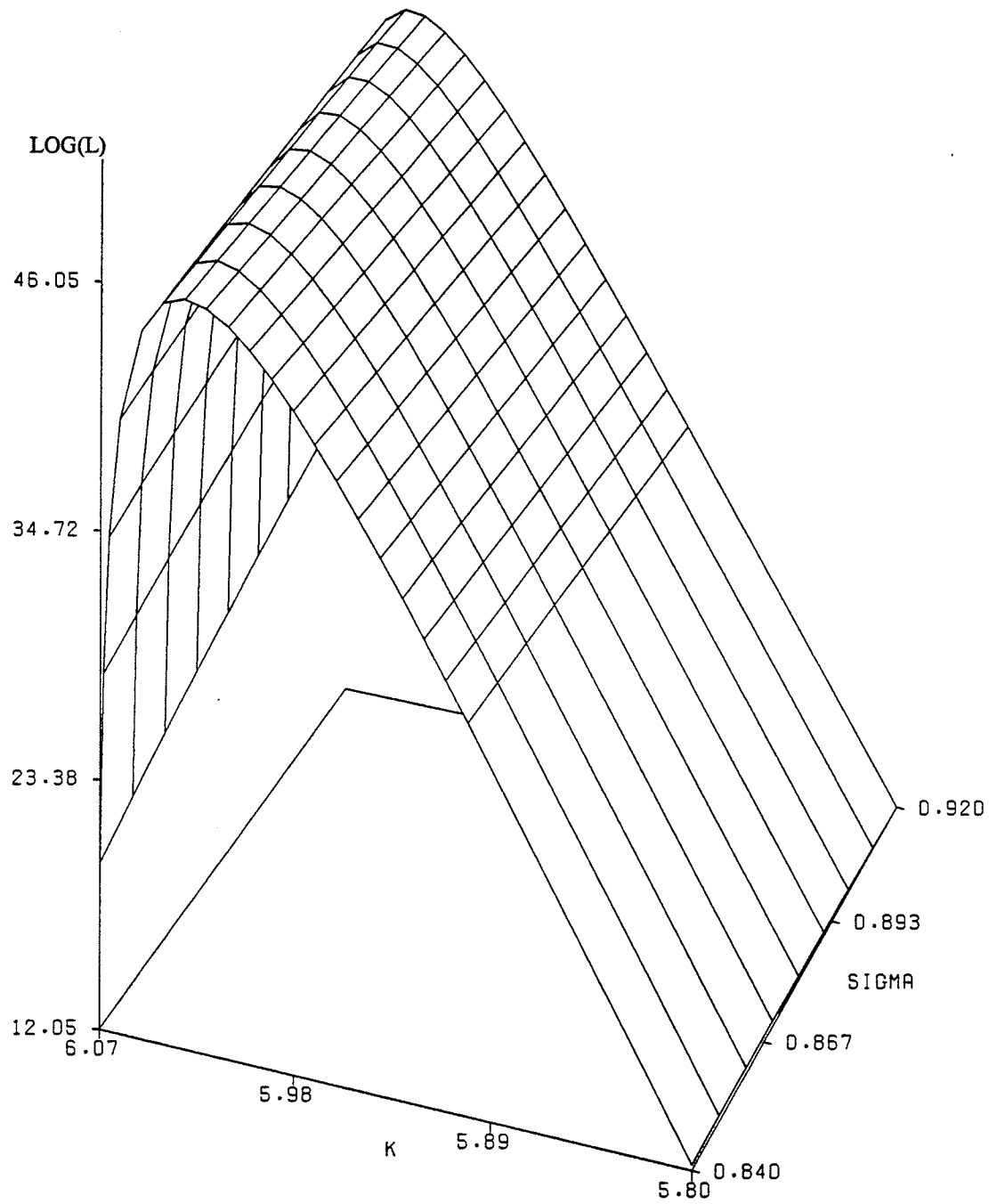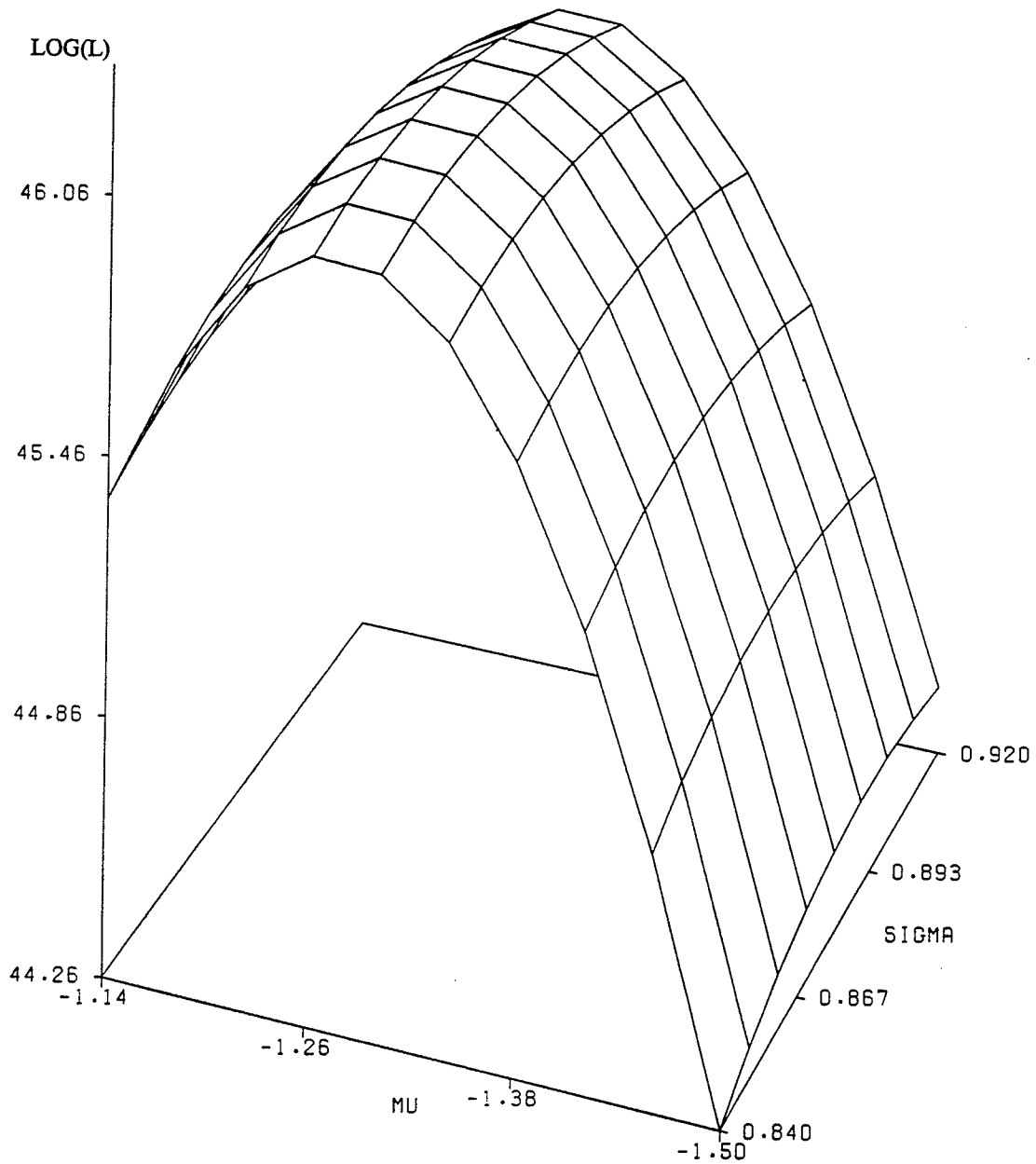NOTE: SIGMA SET TO MLE

THREE−PARAMETER LOGNORMAL LOG−LIKELIHOOD FUNCTION

K VERSUS SIGMA MAXIMIZATION

NOTE: MU SET TO MLE

Figure 3.12
THREE-PARAMETER LOGNORMAL LOG-LIKELIHOOD FUNCTION
MU VERSUS SIGMA MAXIMIZATION

NOTE: K SET TO MLE

close to the minimum of the observed values. This illustrates the need for an initial estimate of k to be very close to the first order statistic and yet still on the appropriate side of the log-likelihood surface for the iterative process to converge.

Figure 3.10 reveals that the seeming regularity of the surface along the μ axis is complicated by the cliff-like face of the k axis and certainly is much flatter. Similar results are seen in Figure 3.11, which portrays the interaction between k and σ. The μ, σ grid in Figure 3.12 illustrates a regular parabolic surface over which to maximize. This indicates why without the involvement of the troublesome threshold parameter, finding the mle for the lognormal distribution is a simple task. Graphical techniques such as these would suffice in many practical applications but would not be feasible as a routine process for a large number of samples.

For obtaining the mle in this study, a combination of the Cohen (1951) and Calitz (1973) method was used. The algorithm was constructed to check for the appropriateness of the two techniques for each sample and then a choice was made based upon the shape of the log-likelihood function along the k axis. Once convergence was obtained for the k parameter, the other parameters were found by simple substitution. The final result was checked via the 3x3 grid method used for the three-parameter Weibull distribution to ensure a local maximum had been obtained.

The mle search method is based on solving the equation $\lambda(k)=0$ such that

$$\lambda(k) = \sum_{i=1}^{n} \frac{\log(x_i-k)}{x_i-k} + \sum_{i=1}^{n} \frac{1}{x_i-k} \left[\frac{1}{n}\sum_{i=1}^{n} \log^2(x_i-k) - \frac{1}{n}\sum_{i=1}^{n} \log(x_i-k) \right.$$
$$\left. - \frac{1}{n^2}\left[\sum_{i=1}^{n} \log(x_i-k)\right]^2 \right] .$$

Calitz (1973) and Amin (1981) mention that searching this function may be difficult and demonstrate that the function is very steep in the neighbourhood of the solution. Calitz proposes that a Newton-Raphson approach could be used. This necessitates the construction of the derivative of $\lambda(k)$

$$\lambda'(k) = \sum_{i=1}^{n} \frac{\log(x_i-k)-1}{(x_i-k)^2}$$

$$+ \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(x_i-k)^2} \left[\sum_{i=1}^{n} \log^2(x_i-k) - \sum_{i=1}^{n} \log(x_i-k) - \frac{1}{n}\left[\sum_{i=1}^{n} \log(x_i-k)\right]^2 \right]$$

$$+ \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(x_i-k)} \left[\sum_{i=1}^{n} \frac{1}{x_i-k} - 2\sum_{i=1}^{n} \frac{\log(x_i-k)}{x_i-k} + \frac{2}{n}\left[\sum_{i=1}^{n} \frac{1}{x_i-k}\right]\left[\sum_{i=1}^{n} \log(x_i-k)\right] \right]$$

Searching for a root along this functional surface is also somewhat tenuous however and Newton-Raphson may not converge if the interval is not chosen appropriately. Figure 3.13 displays the form of the $\lambda'(k)$ function for the Cohen (1951) sample. One must ensure that the search interval must be along the steep cliff near the sample minimum and yet be on either side of zero. It was found that the smaller the sample the more steep the cliff became. This would cause many search algorithms to miss the zero point.

Previous algorithmic suggestions involve taking a percentage of the sample minimum for the search interval starting point. This is not ideal, however, because depending upon the parametric settings, such a point could be either on the plateau or cliff area of Figure 3.13, with a great deal of inconsistency across samples.

A new algorithm is proposed whereby the search is split into two parts. First, the $\lambda'(k)$ function is searched to find the zero point by "creeping" along the plateau area as seen in Figure 3.13 until the cliff edge is reached so that $\lambda'(k)$ becomes negative.

Figure 3.13

THREE—PARAMETER LOGNORMAL

LAMBDA(K) FUNCTION EXPLORATION
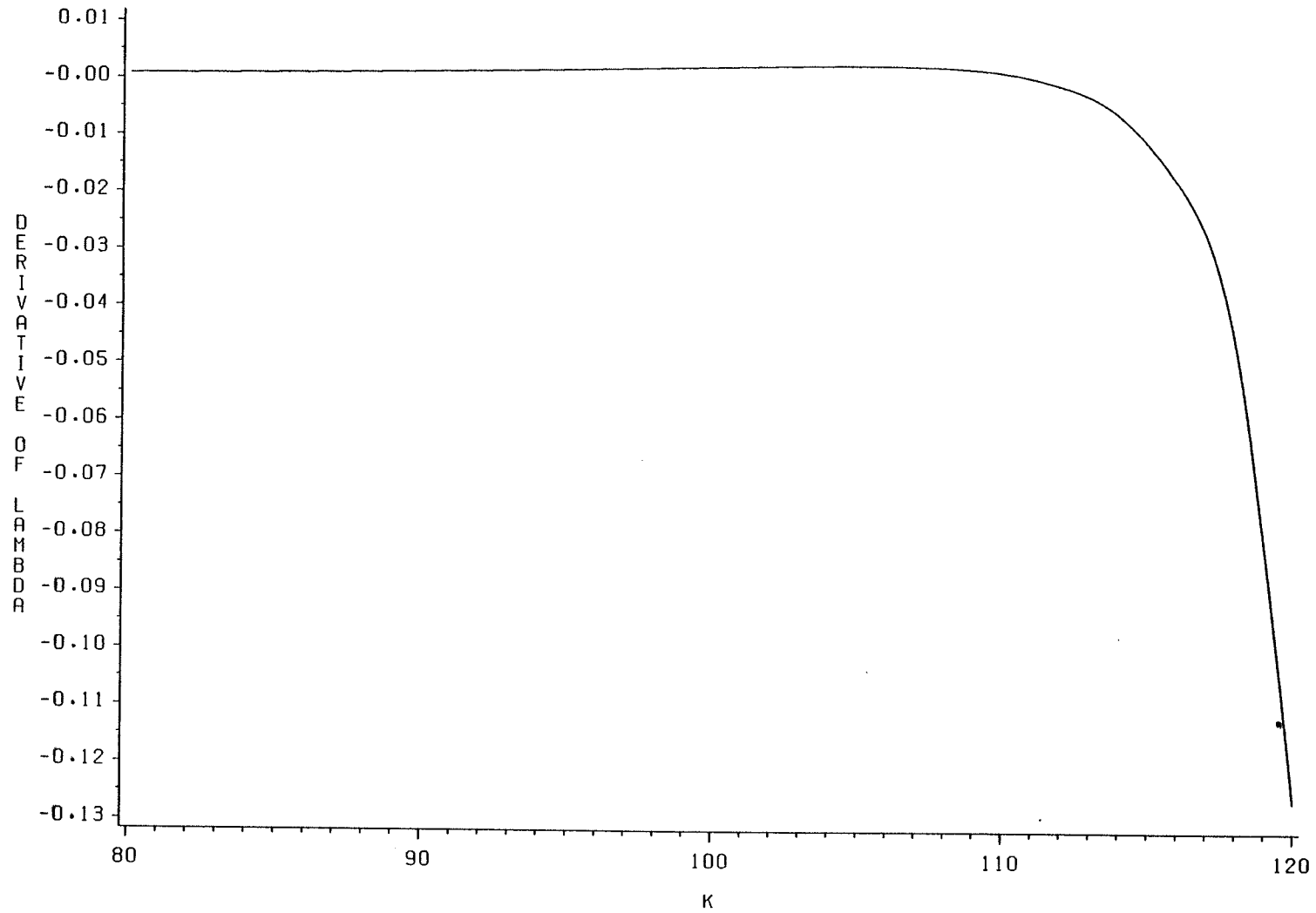
This point and $x_{[1]}-\epsilon$ are then used as boundaries for a linear search routine of the lambda function. With epsilon and the size of the successive "creeping" steps chosen as being very small, the search routine is guaranteed to find the solution. Numerically, the routine is not optimal as it typically takes 30 iterations for each of the two parts to produce a solution. Unlike other algorithms in the literature, however, results are universally obtainable.

3.3.2 Tierney-Kadane Approximation

Jeffreys' invariant prior approach is once again used to represent a prior lack of knowledge about the parameters. Hill (1963) first investigated the use of this prior with the lognormal distribution. Many authors have used Bayes methods for estimation of lognormal distributions, but strictly for the two-parameter case. Padgett (1988) provides an extensive list of references. Although he refers to the prior as "unrealistic", Hill finds that the results seen with Jeffreys' prior yield "nearly the same posterior density" as a more informative model. Using this approach, assuming prior independence of the three systemic parameters, the joint prior is

$$g(k,\mu,\sigma) \propto \frac{1}{\sigma^2}$$

and the log-likelihood function is

$$L(k,\mu,\sigma|\underline{X}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \sum_{i=1}^{n}\log(x_i-k) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[\log(x_i-k)-\mu]^2 \ .$$

The T-K method requires the construction of one L_0 function and an L_* function for each parameter to be estimated. These modified log-likelihood functions here are calculated

to be

$$L_o = -\frac{1}{2}\log(2\pi) - \frac{n+2}{n}\log\sigma - \frac{1}{n}\sum_{i=1}^{n}\log(x_i-k) - \frac{1}{2n\sigma^2}\sum_{i=1}^{n}[\log(x_i-k) - \mu]^2$$

and, for estimating the three systemic parameters, set $u(\theta)=\theta_i$ in turn to produce three L_* equations where

$$L_{*_i} = \frac{1}{n}\log(\theta_i) + L_o \quad.$$

To produce the maximization points θ_o and θ_* for the L_o and L_* functions, the first partial derivatives are necessary. These are

$$\frac{\partial L_o}{\partial k} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(x_i-k)} + \frac{1}{n\sigma^2}\sum_{i=1}^{n}\frac{\log(x_i-k) - \mu}{x_i-k}$$

$$\frac{\partial L_o}{\partial \mu} = \frac{1}{n\sigma^2}\sum_{i=1}^{n}[\log(x_i-k) - \mu]$$

$$\frac{\partial L_o}{\partial \sigma} = -\frac{n+2}{n\sigma} + \frac{1}{n\sigma^3}\sum_{i=1}^{n}[\log(x_i-k) - \mu]^2$$

for L_o, and again for L_* the relationship between the two functions can be used so that the partials of L_* are easily found by

$$\frac{\partial L_{*_i}}{\partial \theta_i} = \frac{\partial L_o}{\partial \theta_i} + \frac{\partial}{\partial \theta_i}\left[\frac{1}{n}\log(\theta_i)\right] \quad.$$

Unfortunately, like the mle system of equations, the L_o and L_* functions do not produce closed forms for the maximization points θ_o and θ_* respectively. Numerical routines as described in previous sections were used, with the mle's as a starting point for iterative searches. Convergence is typically obtained within 15 iterations.

To produce the Σ_o and Σ_* matrices, the second derivatives of L_o and L_* must be construced and evaluated at the maximizing points θ_o and θ_* respectively. These are found to be

$$L_{o11} = \frac{\partial^2 L_o}{\partial k^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(x_i-k)^2} + \frac{1}{n\sigma^2}\sum_{i=1}^{n}\frac{\log(x_i-k)-\mu-1}{(x_i-k)^2}$$

$$L_{o22} = \frac{\partial^2 L_o}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

$$L_{o33} = \frac{\partial^2 L_o}{\partial \sigma^2} = \frac{n+2}{n\sigma^2} - \frac{3}{n\sigma^4}\sum_{i=1}^{n}[\log(x_i-k)-\mu]^2$$

$$L_{o12} = \frac{\partial^2 L_o}{\partial k \partial \mu} = -\frac{1}{n\sigma^2}\sum_{i=1}^{n}\frac{1}{x_i-k}$$

$$L_{o13} = \frac{\partial^2 L_o}{\partial k \partial \sigma} = -\frac{2}{n\sigma^3}\sum_{i=1}^{n}\frac{\log(x_i-k)-\mu}{x_i-k}$$

$$L_{o23} = \frac{\partial^2 L_o}{\partial \mu \partial \sigma} = -\frac{2}{n\sigma^3}\sum_{i=1}^{n}[\log(x_i-k)-\mu] \quad .$$

The partials of L_* follow directly via the relation

$$\frac{\partial^2 L_{*_i}}{\partial \theta_j \partial \theta_k} = \frac{\partial^2 L_o}{\partial \theta_j \partial \theta_k} + \frac{\partial^2}{\partial \theta_j \partial \theta_k}\left[\frac{1}{n}\log(\theta_i)\right] \quad .$$

Once obtained, the remaining steps in the T-K algorithm are carried out by numerical methods.

Calitz (1973) mentions that the value of the parameters should have little impact on competing estimation methods' relative efficiency. To circumvent the difficulty of positive-valued parameters without loss of generality, positive parameter settings were

used for the large scale simulation runs. For the single sample case which did involve negative parameters, the approach of using large additive constants as described by Tierney and Kadane (1985) was used.

3.3.3 Reliability Estimation

The reliability function for the three-parameter lognormal distribution is

$$R(t) = Pr(T \geq t) = \int_{t}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma(x-k)} \exp\left[-\frac{1}{2\sigma^2}[\log(x-k)-\mu]^2\right] dx$$

To estimate R(t) using Bayesian methods, the alternatives are Lindley's or the Tierney-Kadane approximations. Once the three systemic parameters have been estimated by Lindley's approach, only the u_i and u_{ij} functions need to be reconstructed. Using the T-K method, however, the L_o and L_* functions and their respective first and second partial derivatives are necessary to obtain the maximization points. Furthermore, a numerical maximization process must still be relied upon to produce the T-K approximation which could be subject to further convergence problems seen with the mle's. With Lindley's method the L_{ijk}'s do not need to be reconstructed. As such, the Lindley approximation is easier to implement in this situation.

Set $u(k,\mu,\sigma)=R(t)$ as above and let

$$y = \frac{\log(x-k)-\mu}{\sigma} \quad so \quad that \quad dy = \frac{dx}{\sigma(x-k)}$$

so that

$$R(t) = \int_{\frac{\log(t-k)-\mu}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}y^2\right] dy$$

or, taking the complement

$$R(t) = 1 - \Phi\left(\frac{\log(t-k)-\mu}{\sigma}\right)$$

where Φ is the standard normal distribution function.

For estimation purposes, the work simplifies by setting

$$u(\underline{\theta}) = \int_{-\infty}^{\frac{\log(t-k)-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}y^2\right] dy = I$$

Recall, as with the gamma $R(t)$ function, the relation that if

$$I = \int_{v(x)}^{w(x)} f(x,t) dt$$

then

$$\frac{\partial I}{\partial x} = f(x,w)\frac{\partial w}{\partial x} - f(x,v)\frac{\partial v}{\partial x} + \int_{v(x)}^{w(x)} \frac{\partial f(x,t)}{\partial x} dt$$

Using this result, the u_i functions are

$$u_1 = \frac{\partial I}{\partial k} = -\frac{1}{\sqrt{2\pi}\,\sigma\,(t-k)}\exp\left[-\frac{1}{2}\left[\frac{\log(t-k)-\mu}{\sigma}\right]^2\right]$$

$$u_2 = \frac{\partial I}{\partial \mu} = -\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2}\left[\frac{\log(t-k)-\mu}{\sigma}\right]^2\right]$$

$$u_3 = \frac{\partial I}{\partial \sigma} = -\frac{\log(t-k)-\mu}{\sqrt{2\pi}\,\sigma^2}\exp\left[-\frac{1}{2}\left[\frac{\log(t-k)-\mu}{\sigma}\right]^2\right]$$

The u_{ij} functions can now be found by straightforward differentiation.

$$u_{11} = \frac{\partial^2 u(\underline{\theta})}{\partial k^2} = \frac{u_1}{(t-k)}\left[1+\frac{\log(t-k)-\mu}{\sigma^2}\right]$$

$$u_{22} = \frac{\partial^2 u(\underline{\theta})}{\partial \mu^2} = \frac{u_2}{\sigma^2}[\log(t-k)-\mu]$$

$$u_{33} = \frac{\partial^2 u(\underline{\theta})}{\partial \sigma^2} = \frac{u_3}{\sigma}\left[\frac{[\log(t-k)-\mu]^2}{\sigma}-2\right]$$

$$u_{12} = \frac{\partial^2 u(\underline{\theta})}{\partial k \partial \mu} = \frac{u_1}{\sigma^2}[\log(t-k)-\mu]$$

$$u_{13} = \frac{\partial^2 u(\underline{\theta})}{\partial k \partial \sigma} = \frac{u_1}{\sigma}\left[\frac{[\log(t-k)-\mu]^2}{\sigma^2}-1\right]$$

$$u_{23} = \frac{\partial^2 u(\underline{\theta})}{\partial \mu \partial \sigma} = \frac{u_2}{\sigma}\left[\frac{[\log(t-k)-\mu]^2}{\sigma^2}-1\right]$$

Lindley's approximation to the Bayes estimator is then constructed by simple substitution into the formula.

3.3.4 Examples

Several application datasets are available in the literature, mainly to demonstrate the difficulty in estimation for the three-parameter lognormal. Three such examples are given by Cohen (1951), Dumonceaux and Antle (1973, D&A) and McCool (1974). The Cohen sample has been used by a number of other authors as a test sample of mle search routines. The three samples are interesting because they are all small and give any numerical routine problems. They also reflect very different parametric settings.

The McCool (1974) sample is of special interest because, as noted by Crow and Shimizu (1988), the sample was a good example where the mle was unobtainable. It turned out that this sample was a good example of the very steep cliff face mentioned in the earlier section on estimation problems. Although it is true that the sample required a very narrow search interval, results for the sample were obtained and verified as the mle.

Each sample was tested via the empirical goodness of fit tests described earlier. The Cohen (1951) sample failed the test (p-value = .0001), which is surprising seeing that it has been used extensively in the literature. Estimation results for the three samples are given in Table 3.13.

Table 3.13: Parameter Estimation Results For Lognormal Samples

Source	n	k hat	μ hat	σ hat	k^*	μ^*	σ^*
McCool	10	144.12	3.94	0.91	124.02	4.41	0.65
D&A	20	0.18	-1.56	0.51	0.05	-0.98	0.30
Cohen	20	117.72	3.37	0.60	89.58	4.32	0.16

The two methods give markedly different parameter estimates for these small

samples. Researchers using the Bayes approximation would arrive at a much different result in terms of the assumed parent distribution than those using the mle.

In terms of variance estimation, Cohen (1951) provides forms for the asymptotic variances of the mle for the systemic parameters. The posterior variance estimates, arrived at in the same manner as for the other distributions in this chapter, once again exhibit negative values and are as such inadmissible.

Reliability estimation was subsequently undertaken with results for the McCool (1974) sample are presented in Table 3.14 and displayed in Figure 3.14.

Table 3.14: R(t) Estimates For McCool (1974) Sample(n=10)

t	MLE R(t)	$R^*(t)$
155	0.9559	0.7065
185	0.5989	0.6808
205	0.4257	0.5076
235	0.2651	0.3073
265	0.1732	0.1928
295	0.1180	0.1297
325	0.0831	0.0938
355	0.0602	0.0723
385	0.0446	0.0552
405	0.0370	0.0521
445	0.0259	0.0426

From Figure 3.14, it can be seen that the Bayes approximate estimator displays a disturbing tendency for the lower t values. There is actually a part of the t range space where an increase in t causes an increase in $R^*(t)$, which is theoretically impossible.

Figure 3.14   **THREE-PARAMETER LOGNORMAL RELIABILITY ESTIMATION**
McCOOL(1974), N=10 SAMPLE
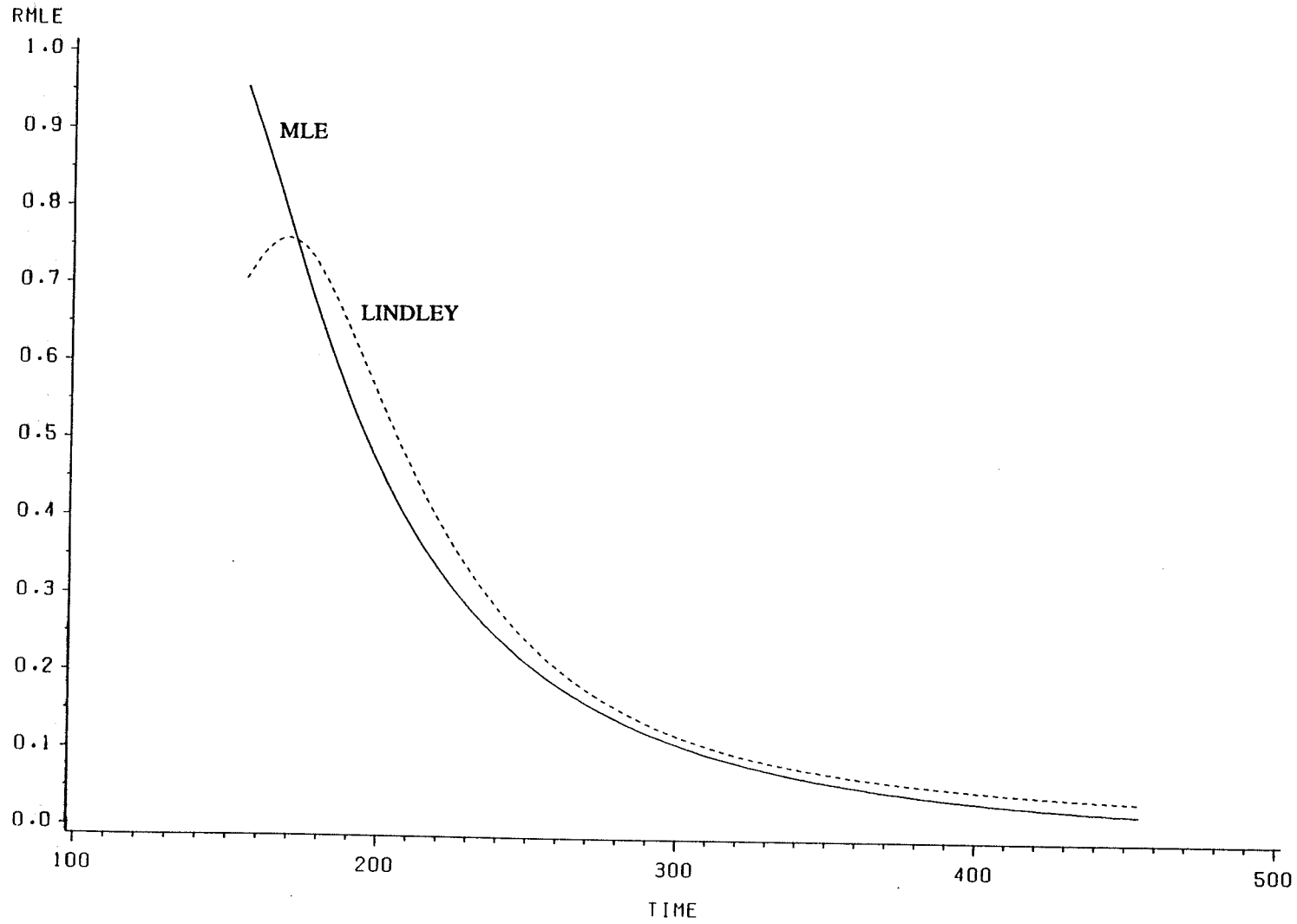
Further, the Bayes approximation is consistently larger than the mle once t is taken to be outside the "disturbing" range. After the initial problem, the Bayes results are within 1% of the mle estimate. The abnormal result for small values of t makes the use of this approach, for small samples at least, untenable. Posterior variance estimates were produced as well, but the presence of negative values makes a comparison with the asymptotic variance of the mle meaningless. It should be noted at this point the sample size is extremely small, so the results are not all that surprising.

To establish sample size impact limitations, three generated samples (n=25,50,100) with the parametric settings $(k,\mu,\sigma)=(25,2,0.8)$ were used for R(t) estimation. The use of generated samples also allows for a comparison against the true R(t) value which was not possible for the McCool sample. As mentioned by Calitz (1973), these settings provide a good environment for illustration without loss of generality. The estimators along with the true parametric values are given in Tables 3.15, 3.16 and 3.17 for the small, moderate and large sample sizes respectively. Figures 3.15, 3.16 and 3.17 display the resultant R(t) functions.

Figure 3.15   THREE-PARAMETER LOGNORMAL RELIABILITY ESTIMATION
GENERATED SAMPLE, N=25
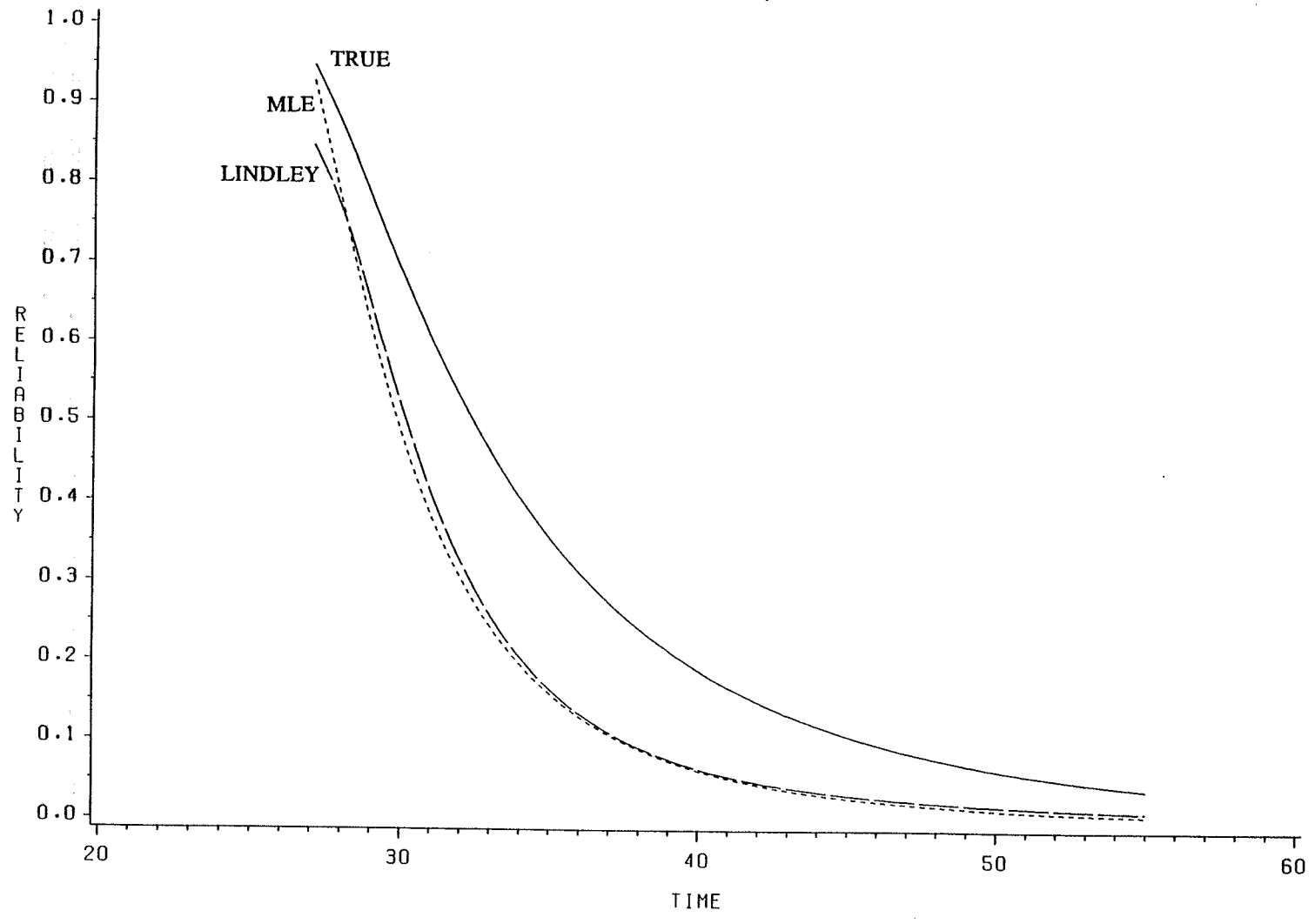
Figure 3.16   **THREE-PARAMETER LOGNORMAL RELIABILITY ESTIMATION**
GENERATED SAMPLE, N=50

Figure 3.17 THREE-PARAMETER LOGNORMAL RELIABILITY ESTIMATION
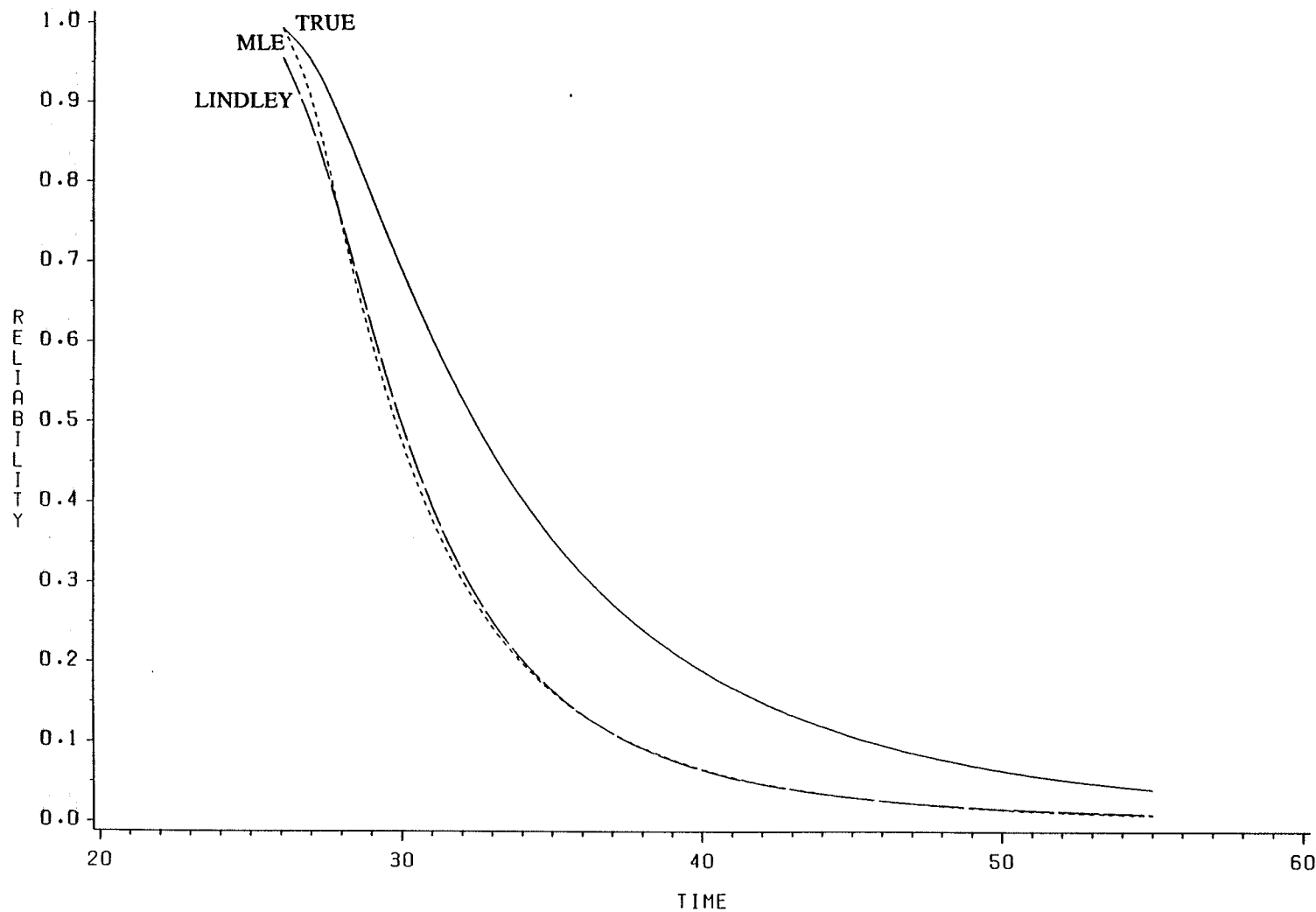GENERATED SAMPLE, N=100
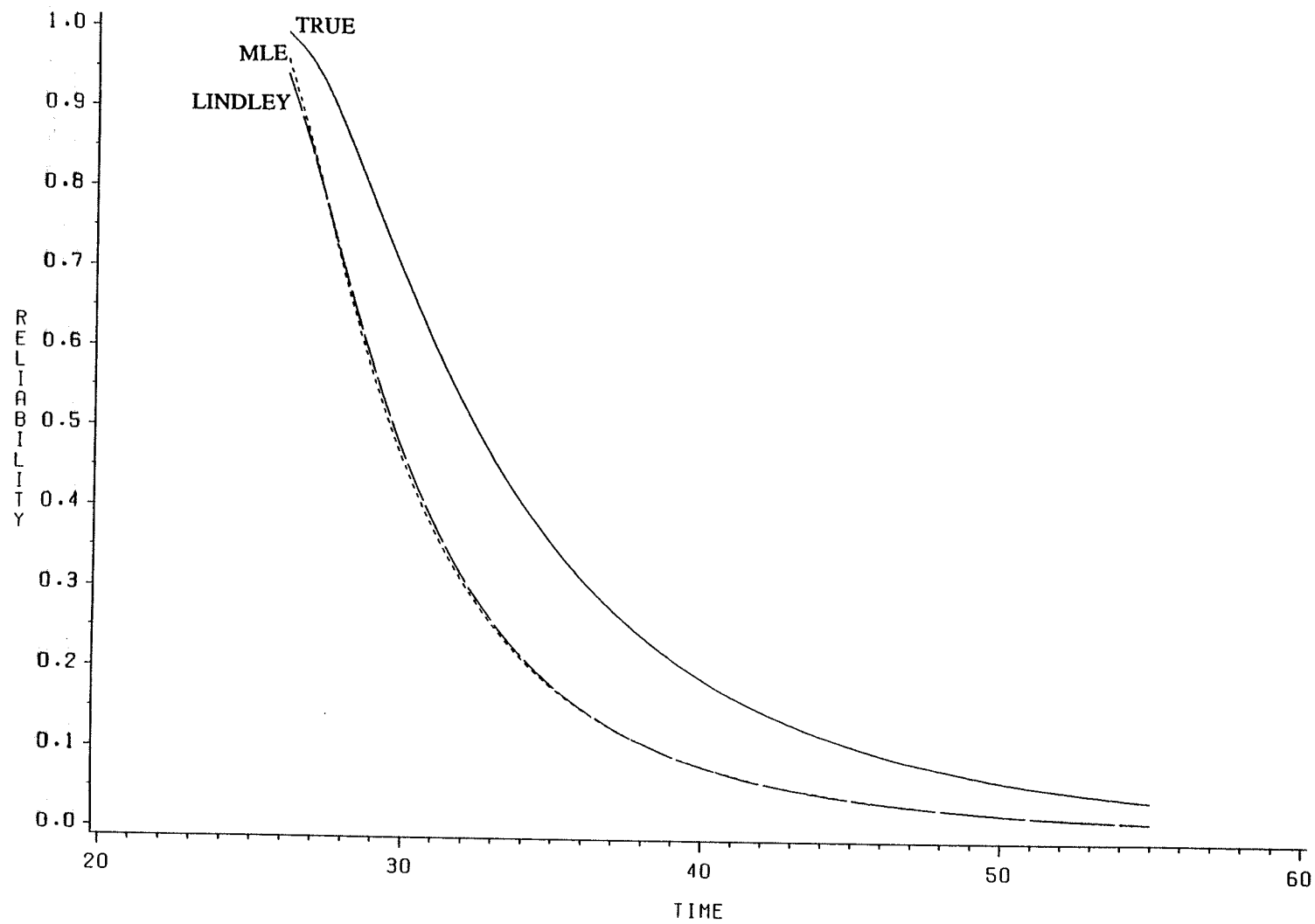
Table 3.15: R(t) Estimates For Generated Sample(n=25)

t	True R(t)	MLE R(t)	R*(t)
27	0.9448	0.9286	0.8480
30	0.6873	0.4819	0.5156
33	0.4604	0.2410	0.2548
36	0.3095	0.1294	0.1331
39	0.2122	0.0743	0.0763
42	0.1488	0.0450	0.0477
45	0.1066	0.0284	0.0321
48	0.0779	0.0187	0.0228
51	0.0579	0.0126	0.0170
54	0.0437	0.0088	0.0130

Both estimation techniques consistently underestimate the true R(t) value for all three sample sizes. In Figure 3.15, the Bayes estimates are closer to the true value for small values of t but less than the mle for larger values of t. This result agrees with what was obtained for R(t) results for the three-parameter gamma distribution. The margin of error for the estimates is as large as 6%.

Table 3.16: R(t) Estimates For Generated Sample(n=50)

t	True R(t)	MLE R(t)	$R^*(t)$
26	0.9938	0.9941	0.9566
29	0.7785	0.5942	0.6135
32	0.5270	0.3013	0.3129
35	0.3526	0.1614	0.1634
38	0.2400	0.0921	0.0913
41	0.1671	0.0555	0.0546
44	0.1189	0.0349	0.0347
47	0.0863	0.0228	0.0231
50	0.0638	0.0153	0.0161
53	0.0479	0.0106	0.0116

Results for n=50 in Table 3.16 and Figure 3.16 indicate a converging of the two estimation processes, unfortunately to a point distant from the true R(t) value. The underestimation by the Bayes estimator continues. Due to the larger sample size, the deviant behavior of the Bayes estimator seen in the McCool sample is absent from these results.

Table 3.17: R(t) Estimates For Generated Sample(n=100)

t	True R(t)	MLE R(t)	R*(t)
26	0.9938	0.9606	0.9419
29	0.7785	0.5581	0.5688
32	0.5270	0.3052	0.3115
35	0.3526	0.1767	0.1785
38	0.2400	0.1082	0.1082
41	0.1671	0.0694	0.0689
44	0.1189	0.0463	0.0458
47	0.0863	0.0318	0.0316
50	0.0638	0.0225	0.0224
53	0.0479	0.0163	0.0164

Even with n=100, R(t) is consistently underestimated, with the margin as much as 20%. Figure 3.17 indicates that the two estimation process do indeed virtually overlap at this point, unfortunately, it is not at R(t).

3.3.5 Monte Carlo Simulation

Sample generation for the three-parameter lognormal distribution is straightforward using a normal sample generation procedure such as the IMSL (1975) routine RNOR and the relation $x_i = \exp\left[\sigma(RNOR_i + \mu)\right] + k$. Empirical goodness of fit testing as was done with the previous distributions in this chapter was performed to ensure sample veracity. Parameter settings of $\underline{\theta} = (k, \mu, \sigma) = (25, 2, 0.8)$ were once again used. Three separately seeded simulation runs of 1000 samples each were used to examine the impact of sample size on the estimation results (n=25, 50, 100). This allows

for a discussion of small versus moderate and large samples. Convergence for the maximum likelihood estimators was achieved in all cases. It was found, however, that the search interval for the small sample size had to be extremely narrow to achieve the 100% success rate. The 3x3 grid approach was used to ensure that a local maximum was indeed the point of convergence. The method of moments estimator (Cohen, 1951) was included for comparison purposes as it is recognized as being poor estimates relative to the mle.

Table 3.18 presents the mean and mean square error for parameter estimates using the 1000 samples produced for each sample size.

Table 3.18: Estimation Results Based On Varying Sample Sizes

With Parameters $(k, \mu, \sigma) = (25, 2, 0.8)$

	METHOD OF MOMENTS		MLE		TIERNEY-KADANE	
Parameter	Mean	MSE	Mean	MSE	Mean	MSE
n=25						
k	18.883	63.628	25.137	1.282	23.388	18.249
μ	2.377	0.3029	1.506	0.3578	1.8440	0.2965
σ	0.464	0.1419	0.8718	0.0694	0.6849	0.0899
n=50						
k	21.234	20.94	25.119	0.4268	24.525	1.4657
μ	2.165	0.1280	1.548	0.2492	1.683	0.1592
σ	0.553	0.0866	0.838	0.0268	0.762	0.0278
n=100						
k	22.434	9.802	25.072	0.1790	24.809	0.3092
μ	2.021	0.0731	1.573	0.2029	1.635	0.1554
σ	0.618	0.0556	0.820	0.0114	0.784	0.0114

The impact of sample size is impressive, but not uniform for estimating the three parameters. The mle's exhibit smaller bias and mse for estimating k and σ. The Bayes estimator of μ is closer on average to the parametric value with smaller variability than the mle. The mle and approximate Bayes estimators are clearly superior to the method of moments estimators at all sample sizes.

In estimating the location parameter for small samples, the Bayes approximation is subject to much greater variability than its mle counterpart. The mle results are satisfactory considering the difficulty required to produce them.

Once enough sample information is available, the Bayes approximations begin to converge towards the mle as expected. At $n = 100$ it is arguable which estimator would be preferred. The mle exhibits smaller mse for k than the Bayes approach, but larger mse for μ and indistinguishable results for σ.

Posterior variance estimators once again are useless for the small $n = 25$ samples. Only 1% of the posterior variance estimates for k were greater than zero. Over half of the posterior variance estimates for μ were negative. Surprisingly, only 10% of the σ posterior variance estimates were negative.

At $n = 50$ the proportion of negative variance estimates decreases (40%, 2% and 1% for k, μ and σ respectively), but are still present so as to bring their validity into question. The negative estimates are absent when $n = 100$ and so Table 3.19 compares the posterior variances to the asymptotic variances of the mle.

Table 3.19: Average Estimated Posterior VS Asymptotic Variance

Parameter	Bayes	MLE
k	0.0820	0.0850
μ	0.0140	0.0141
σ	0.0093	0.0163

The above table substantiates Sinha's (1987) claim that the posterior variance of the Bayes approximations is less than the asymptotic variance of the mle.

The empirical distributions of the estimates produced for the three simulation runs provide further insight. The stark contrast between the success of the two methods is seen in comparing the n=25 sampling distributions in Figures 3.18 to 3.23.

In comparing the sampling distributions for estimators of the location parameter k (Figures 3.18 and 3.21), the Bayes approach seems to be succeptible to wildly inaccurate estimates, albeit rarely. While both sampling distributions are markedly negatively skewed, the Bayes estimates have a few negative estimates. Clearly the asymptotic properties which underlie this approach are absent at n=25.

The pictures for estimating μ are quite different (Figures 3.19 and 3.22). The mle's sampling distribution appears almost symmetric while the Bayes estimator is noticeably postively skewed. Here also the tendency to have an occasional outlying estimate in the Bayes approach is evident, while absent for the mle. The Bayes sampling distribution does have a greater frequency of estimates close to the true μ value, however.

Figures 3.20 and 3.23 depict the sampling distributions for estimating σ. Here the results look much more regular, although the mle distribution is positively skewed

Figure 3.18   ESTIMATION OF THREE-PARAMETER LOGNORMAL THRESHOLD

RESULTS FOR 1000 SAMPLES

N=25, K=25.0, μ=2.0, σ=0.8
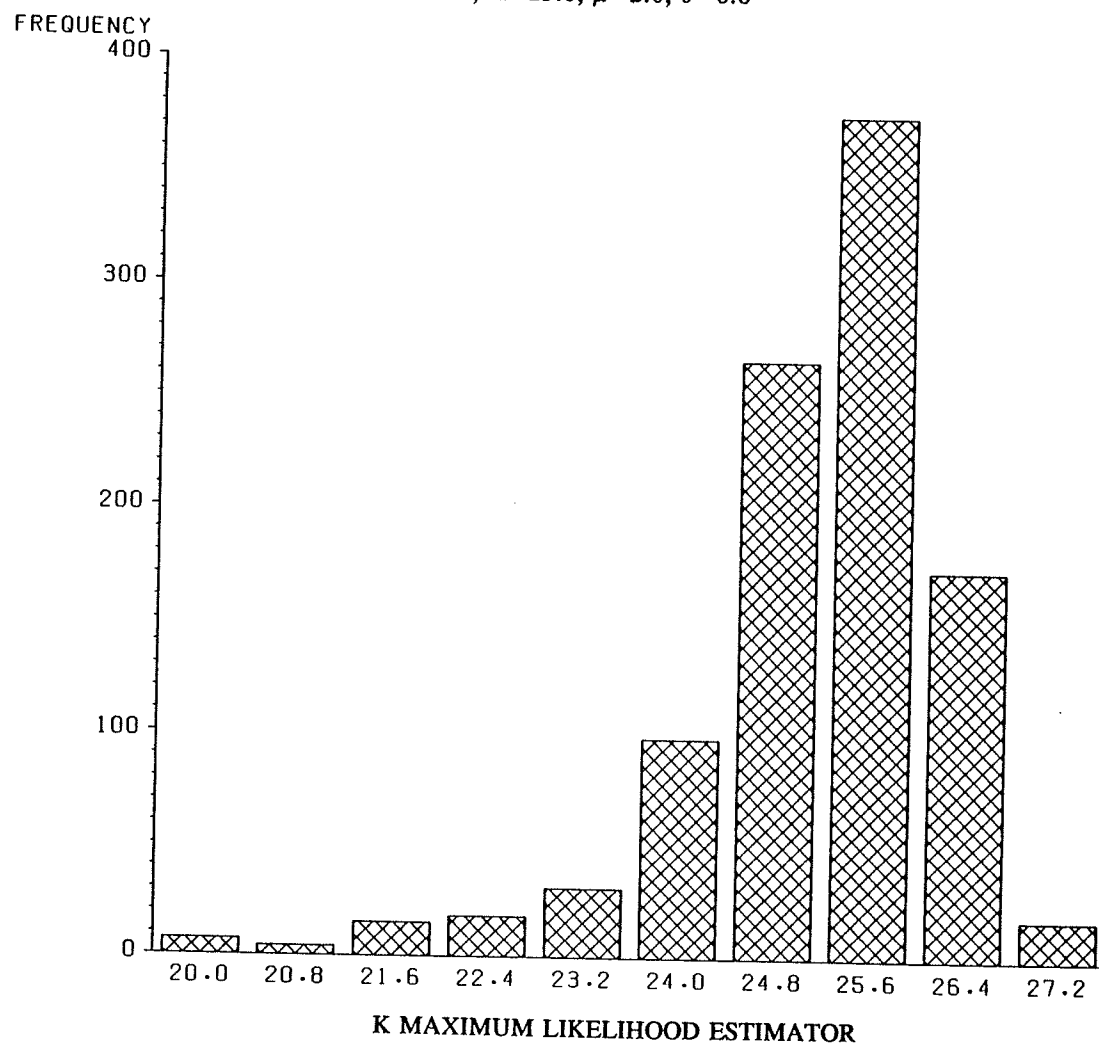

K MAXIMUM LIKELIHOOD ESTIMATOR

Figure 3.19 **ESTIMATION OF THREE-PARAMETER LOGNORMAL μ PARAMETER**
RESULTS FOR 1000 SAMPLES
N=25, K=25.0, μ=2.0, σ=0.8

FREQUENCY

μ MAXIMUM LIKELIHOOD ESTIMATOR

Figure 3.20   ESTIMATION OF THREE-PARAMETER LOGNORMAL σ PARAMETER
RESULTS FOR 1000 SAMPLES
N=25, K=25.0, μ=2.0, σ=0.8

σ MAXIMUM LIKELIHOOD ESTIMATOR

Figure 3.21   **ESTIMATION OF THREE-PARAMETER LOGNORMAL THRESHOLD**
RESULTS FOR 1000 SAMPLES
N=25, K=25.0, μ=2.0, σ=0.8

Figure 3.22   **ESTIMATION OF THREE-PARAMETER LOGNORMAL μ PARAMETER**
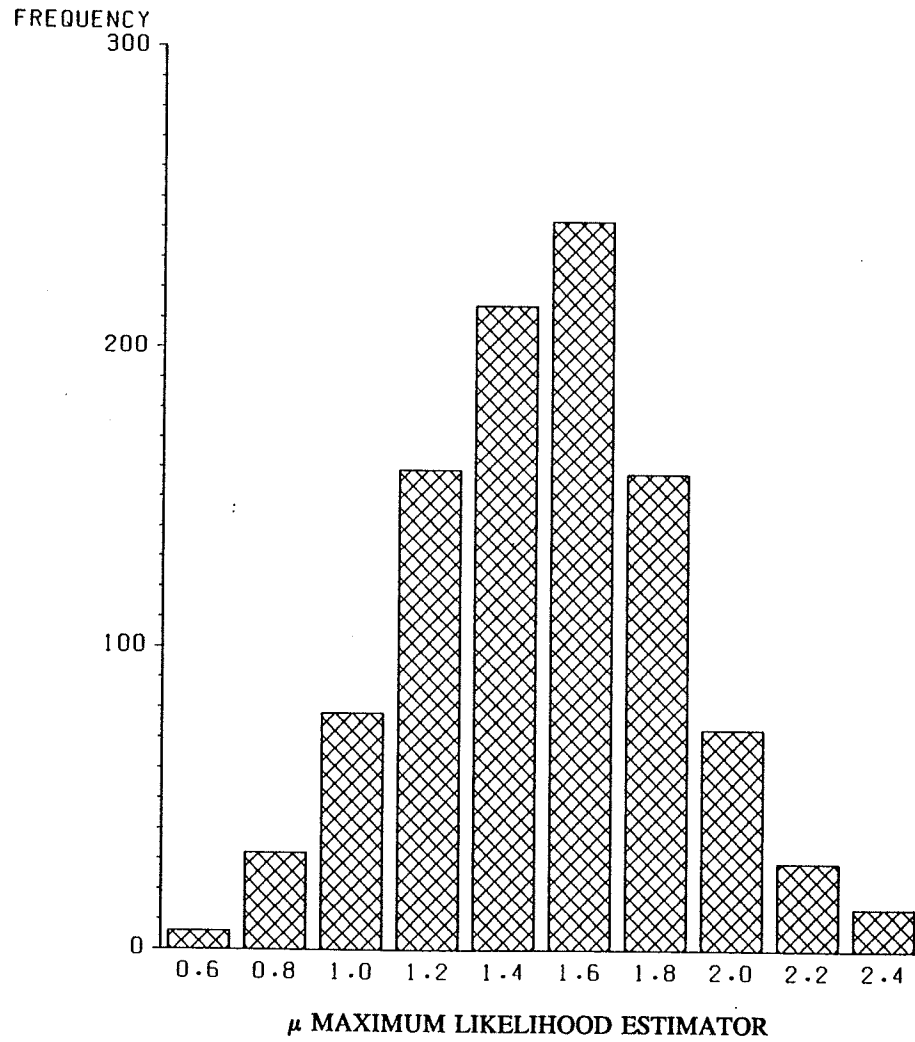RESULTS FOR 1000 SAMPLES
N=25, K=25.0, μ=2.0, σ=0.8

μ TIERNEY-KADANE ESTIMATOR

Figure 3.23   **ESTIMATION OF THREE-PARAMETER LOGNORMAL σ PARAMETER**
RESULTS FOR 1000 SAMPLES
N=25, K=25.0, μ=2.0, σ=0.8

σ TIERNEY-KADANE ESTIMATOR

and the Bayes negatively skewed. In this case at least the two distributions look like those of competing estimators, unlike those for the other two parameters where the mle is clearly superior. Closer inspection reveals, however, that negative estimates for σ are obtained which are inadmissible.

Once again when n becomes moderately large the distributions converge so that Figures 3.24 to 3.26 present the Bayes sampling distributions, although they are representative of the mle results as well. The negative estimates seen for the smaller sample size are absent and the sampling distributions are much more well-behaved.

The location parameter's sampling distribution is still negatively skewed with a tendency to produce inordinately small estimates. This is likely a fact of working with a location parameter and not the result of the estimation process itself. In observing some of the individual samples it became clear that the location parameter was extremely sensitive to outlying observations.

The pictures of the sampling distributions for μ and σ (Figures 3.25 and 3.26) indicate a much more well behaved estimation process due to the larger sample size. The skewness is very slight in both cases and the distributions are appropriately centered around the parametric values.

Results would tend to suggest that for estimating the parameters of the three-parameter lognormal, the maximum likelihood approach is superior for small samples (n < 50), especially if it is important to accurately estimate the location parameter. For moderately large samples, the Bayes approach is competitive, even given the noninformative a priori environment.

In estimating the reliability function R(t), neither approach is particularly
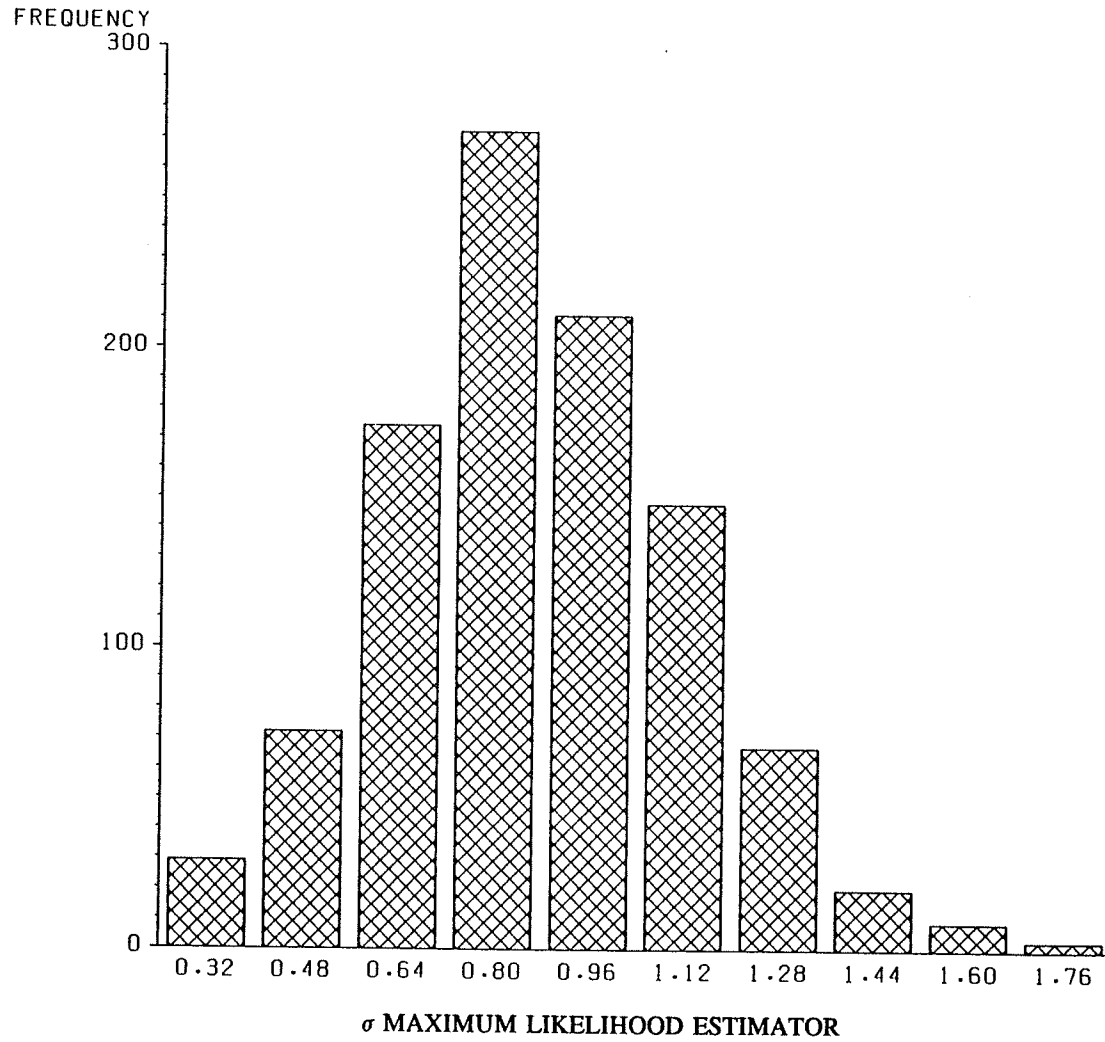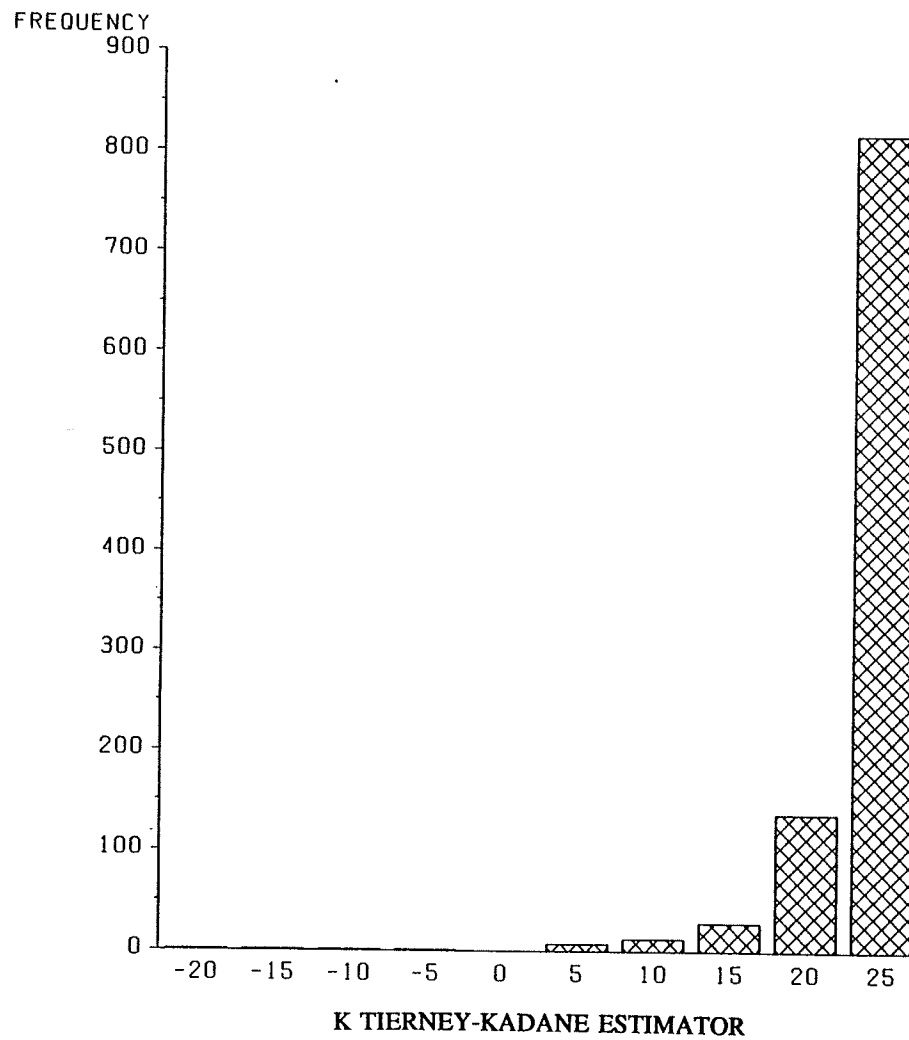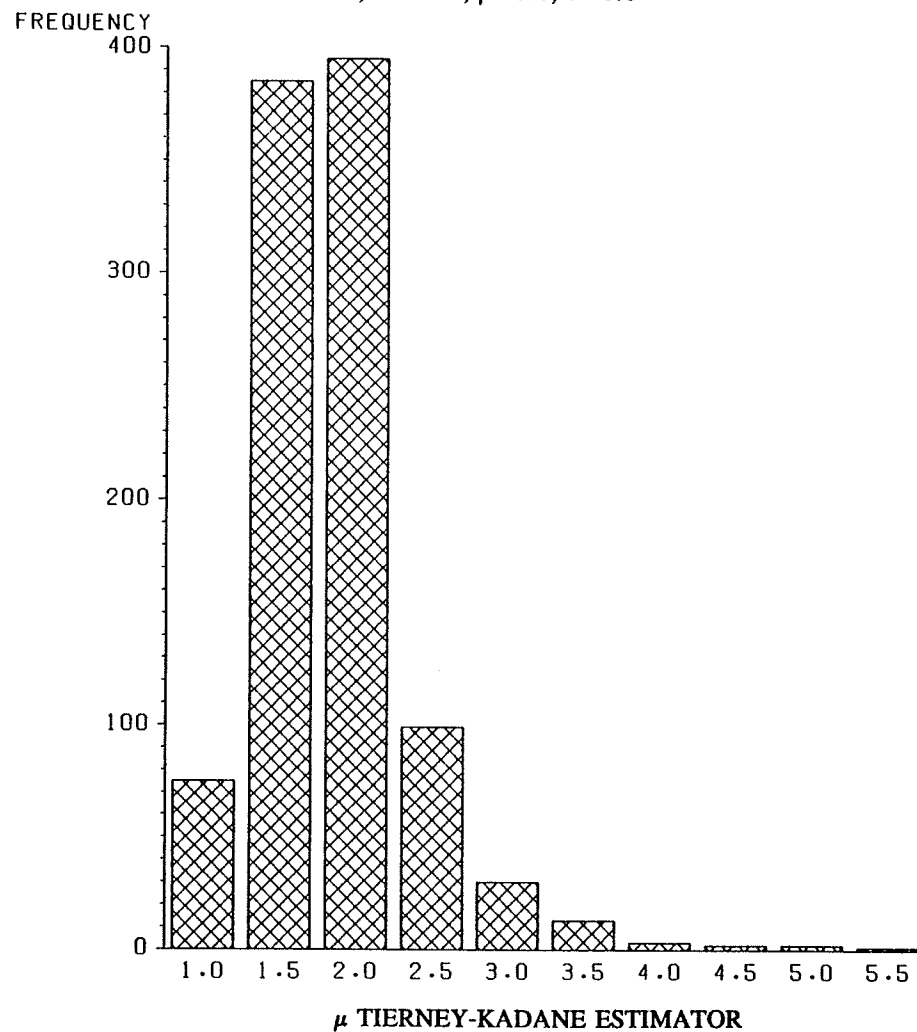
Figure 3.24   ESTIMATION OF THREE-PARAMETER LOGNORMAL THRESHOLD
RESULTS FOR 1000 SAMPLES
N=100, K=25.0, μ=2.0, σ=0.8

Figure 3.25   **ESTIMATION OF THREE-PARAMETER LOGNORMAL μ PARAMETER**
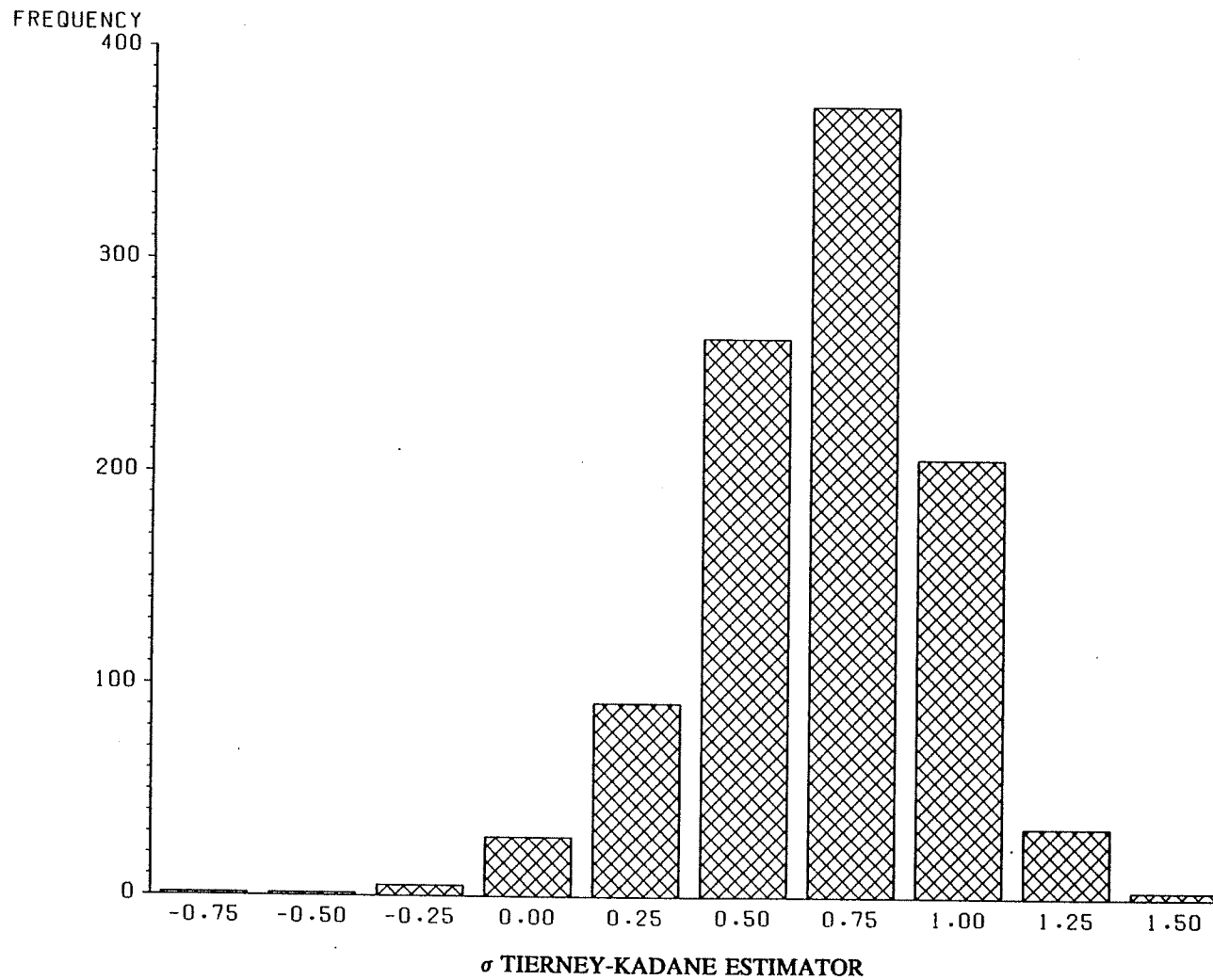RESULTS FOR 1000 SAMPLES
N=100, K=25.0, μ=2.0, σ=0.8

μ TIERNEY-KADANE ESTIMATOR

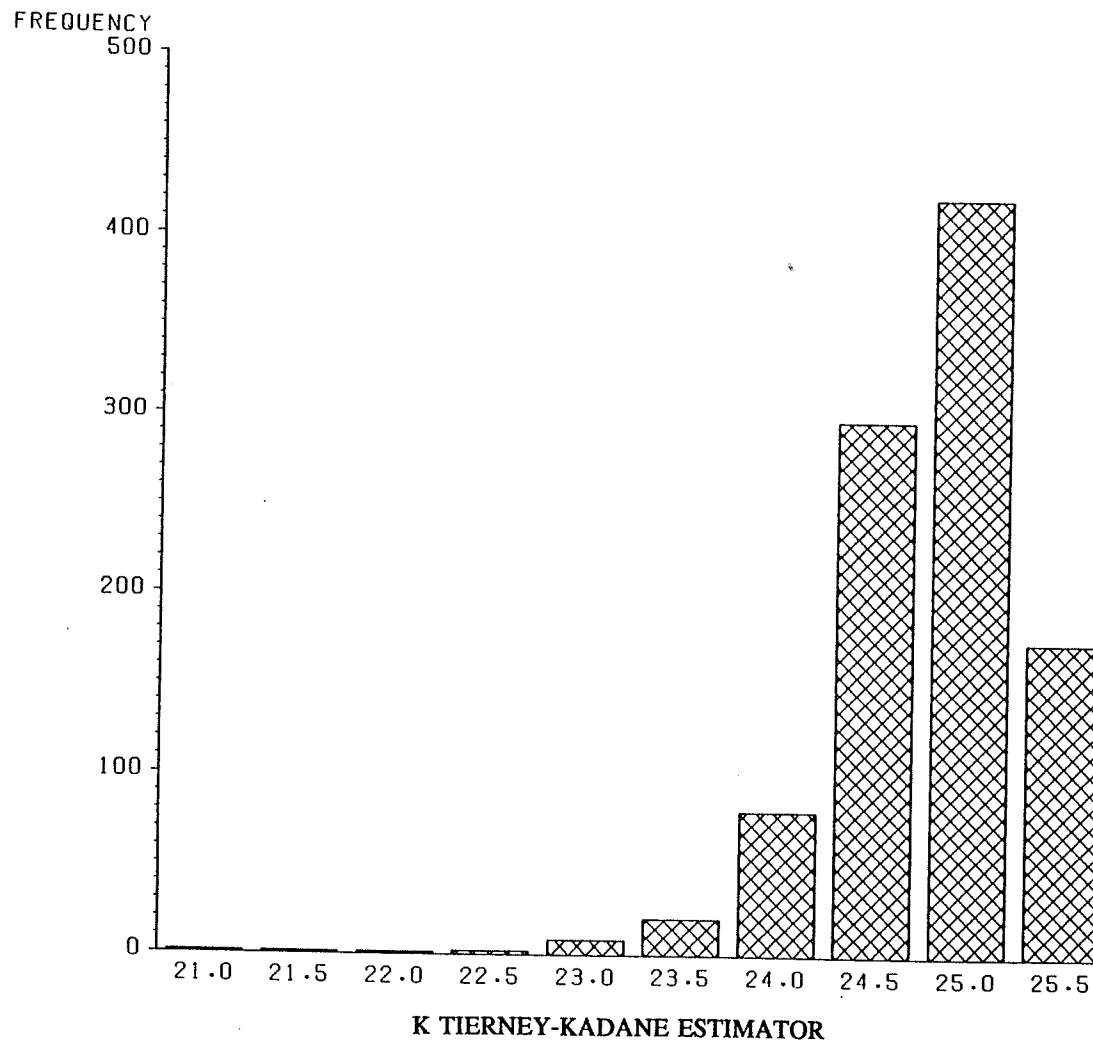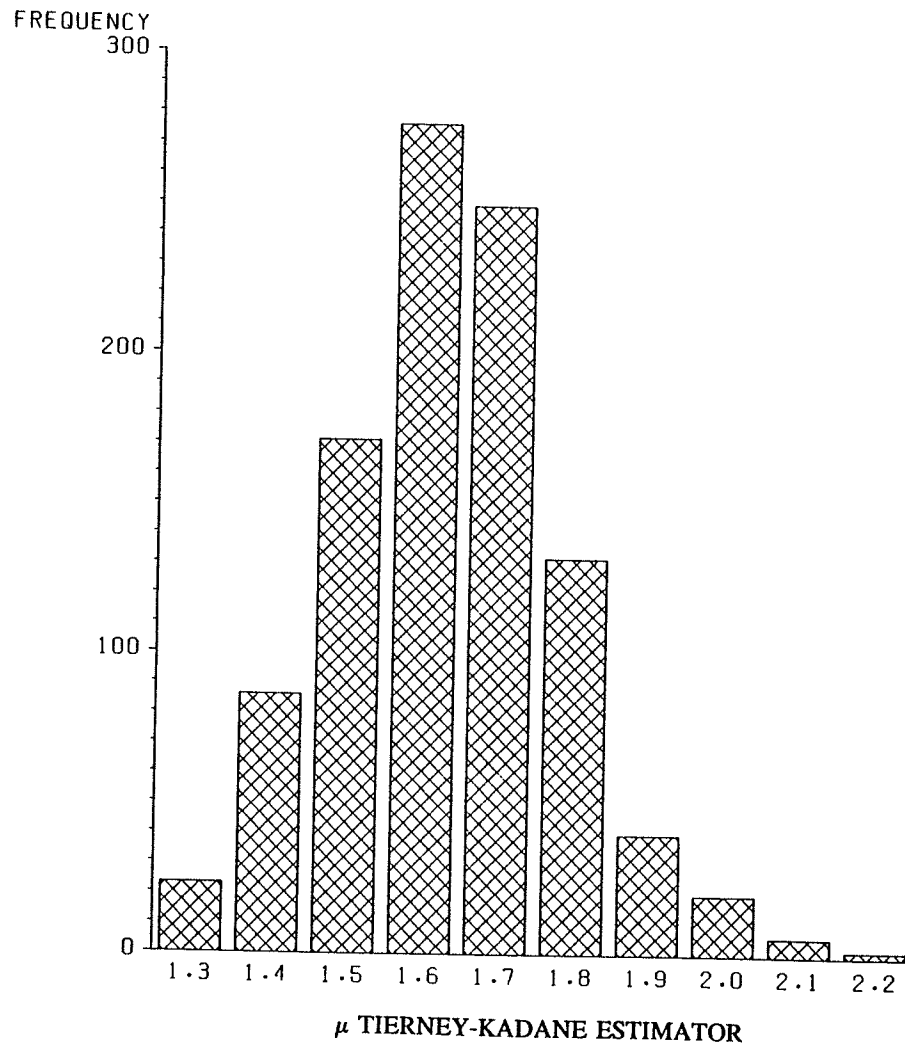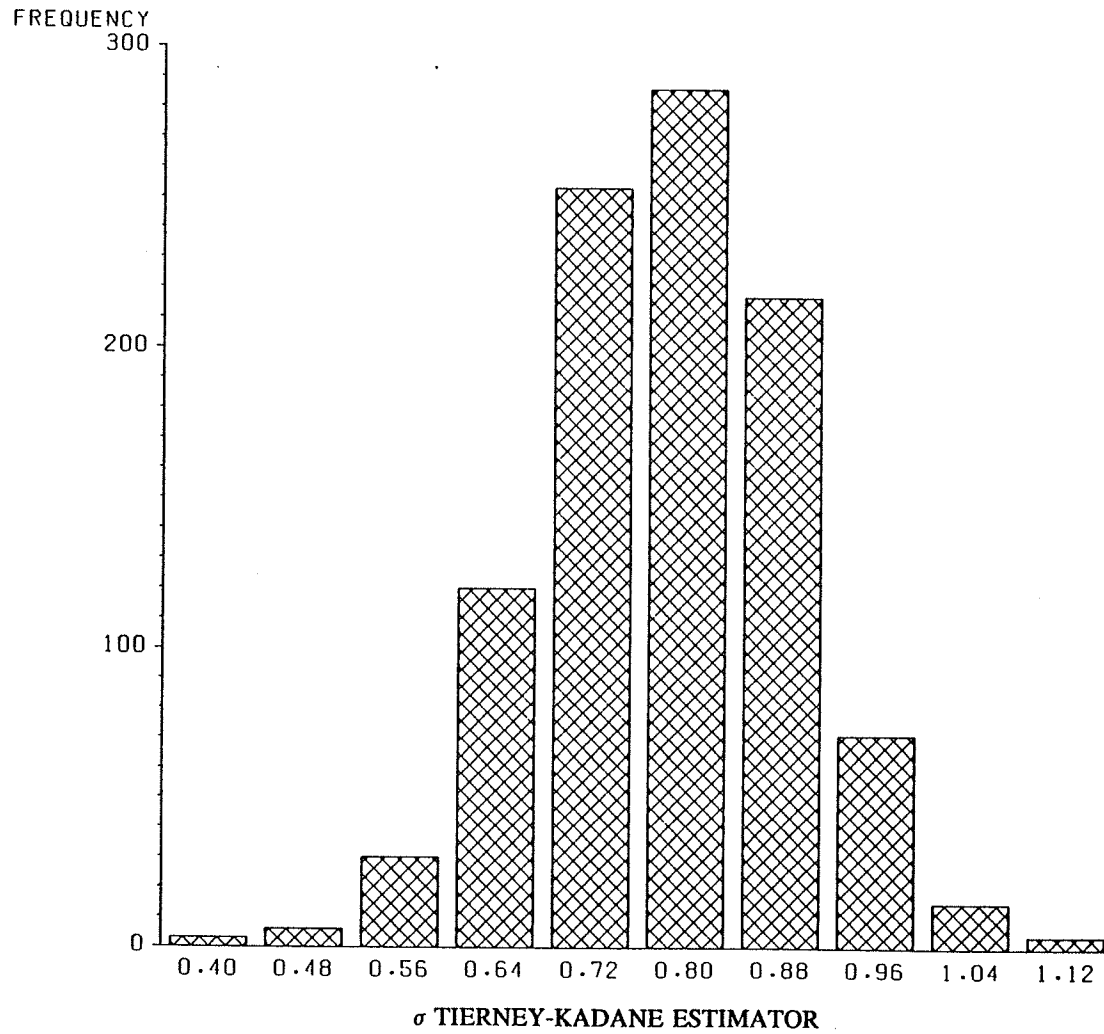**Figure 3.26**

ESTIMATION OF THREE-PARAMETER LOGNORMAL σ PARAMETER
RESULTS FOR 1000 SAMPLES
N=100, K=25.0, μ=2.0, σ=0.8

σ TIERNEY-KADANE ESTIMATOR

successful. For the small sample sizes especially, results of estimation are markedly different from the true R(t). The two methods do converge quite well, unfortunately they converge to a consistent underestimation of the true reliability.

Some of the results of this section contrast the Lye et al (1988) paper. The marked discrepancies between the mle and the Bayes estimators are largely ignored. Further although they mention the negative variance estimates, they shrug them off as a minor nuisance. In doing so they fail to notice that as with previous distributions in this chapter, the variance estimates cannot really be trusted below n=100 whereas all samples presented in the paper are smaller than this.

3.4 Summary

As the work in this chapter covered detailed estimation procedures for three different distributions, a summary of the findings in terms of the estimation process is useful. For some distributions and sampling environments, the mle is clearly superior whereas for others, the Bayes estimators have attractive properties. A choice of which approach to use for a particular application really boils down to:

1) how much is known about the environment (prior and distributional

options)

2) how many estimation processes are expected to be needed

3) is bias or variance more important to the application

4) what access is available to advanced numerical routines

The use of good a priori information in most sampling environments will produce superior estimators. In the absence of such information, the mle is likely the better

choice.

If options for distributional choice are allowable, often there are clearly superior estimation environments. For example, the Weibull with shape less than two is less preferable to a lognormal in terms of difficulties in producing estimators. Further, the lognormal distribution seems to be much more unstable than the other two distributions for estimation processes.

In terms of a choice between the two Bayes approaches, there is no clear choice either. If many distributional parameters and related functions need to be estimated under varying prior information, Lindley's approach will be less work in the long run, assuming the mle's are obtainable. Otherwise the Tierney-Kadane method is less arduous.

Depending upon the distribution and estimator to be studied, the competing schools of thought have both advantages and disadvantages in terms of bias/variance. For example, a small sample size will suggest the use of the mle for variance estimation because although they may not be terribly accurate, at least they produce a believable positive estimate.

Access to a standard powerful computer workstation with advanced numerical software will suggest the T-K method and its reduced CPU time. Otherwise, Lindley's method should be preferred as it is more likely to produce a closed form algebraic solution and once the L_{ijk}'s are constructed, prior distributional comparisons are easier.

CHAPTER 4: ESTIMATION FOR THE BIVARIATE NORMAL DISTRIBUTION

4.1 Introduction

In this chapter, two methods for the approximation to the Bayes estimators due to Lindley (1980) and Tierney & Kadane (1986) are compared to the maximum likelihood estimator of the distributional parameters for the bivariate normal. Jeffreys' invariant prior is developed.

The bivariate normal distribution has received a great deal of attention in the literature (see Johnson and Kotz, 1970 for an extensive list) because of its wide range of applicability and the mathematical challenge it presents incorporating five systemic parameters. The majority of work has been done with assumptions regarding the parameters to make the mathematics easier. Several texts only consider situations where some parameters are known so as to make the mathematics tractable and focus on a subset of the parameters, such as Lindley (1965). In this chapter, the approximations will be applied assuming all five systemic parameters are unknown and a state of ignorance a priori exists about the parameters.

The pdf of the bivariate normal random variable $\underline{X} = (X_1, X_2)$ with the five element vector system parameter $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ is given by

$$f(x_1, x_2 | \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\right.$$

$$\left.\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right]$$

where

$$-\infty < x_1, x_2 < \infty, \quad -\infty < \mu_1, \mu_2 < \infty, \quad \sigma_1, \sigma_2 > 0, \quad -1 \le \rho \le 1.$$

The maximum likelihood estimator (mle) of the systemic parameters is

$$\hat{\mu}_i = \overline{x}_i, \quad \hat{\sigma}_i^2 = \frac{1}{n}\sum_{j=1}^{n} (x_{ij} - \overline{x}_i)^2, \quad i = 1, 2; \quad j = 1, 2, \dots, n.$$

$$\hat{\rho} = \frac{1}{n}\sum_{j=1}^{n} \frac{(x_{1j} - \hat{\mu}_1)(x_{2j} - \hat{\mu}_2)}{\hat{\sigma}_1 \hat{\sigma}_2}$$

Having closed form expressions for the mle will allow for the investigation of closed forms for the Bayes approximations which were not obtainable for the three-parameter distributions seen previously. Obtaining closed forms allow for further insight into the distributional properties of the Bayes approximations as well as removing the need for reliance upon the computer for the mathematics.

4.2 Prior Distributions

Lindley (1965) gives a noninformative prior for the bivariate normal vector parameter $\underline{\theta}$ as

$$g(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)}$$

based on the assumptions that the priors of $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ are independent, with the means' priors uniform in nature and the ρ prior nonvanishing.

A well known alternative method for constructing invariant priors is due to Jeffreys. This approach had not been used for the bivariate normal distribution. Jeffreys' approach requires that the prior be proportional to the square root of the

determinant of the information matrix (Box and Tiao, 1973). More specifically, if Σ_θ is the asymptotic variance-covariance matrix for $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ then Jeffreys' prior is

$$g^*(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \propto \sqrt{\left| \frac{\Sigma_\theta^{-1}}{n} \right|}$$

The information matrix has as its elements the negatives of the expected values of second partial derivatives of the well known log-likelihood function. The second partial derivatives with respect to $\underline{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ are:

$$L_{11} = \frac{\partial^2 L}{\partial \mu_1^2} = -\frac{n}{\sigma_1^2(1-\rho^2)}$$

$$L_{12} = \frac{\partial^2 L}{\partial \mu_1 \partial \mu_2} = \frac{n\rho}{\sigma_1 \sigma_2(1-\rho^2)}$$

$$L_{13} = \frac{\partial^2 L}{\partial \mu_1 \partial \sigma_1} = -\frac{n}{(1-\rho^2)} \left[\frac{2(\overline{X}_1 - \mu_1)}{\sigma_1^3} - \frac{\rho(\overline{X}_2 - \mu_2)}{\sigma_1^2 \sigma_2} \right]$$

$$L_{14} = \frac{\partial^2 L}{\partial \mu_1 \partial \sigma_2} = \frac{n\rho(\overline{X}_2 - \mu_2)}{\sigma_1 \sigma_2^2(1-\rho^2)}$$

$$L_{15} = \frac{\partial^2 L}{\partial \mu_1 \partial \rho} = \frac{n}{\sigma_1(1-\rho^2)^2} \left[\frac{2\rho(\overline{X}_1 - \mu_1)}{\sigma_1} - \frac{(1+\rho^2)(\overline{X}_2 - \mu_2)}{\sigma_2} \right]$$

$$L_{22} = \frac{\partial^2 L}{\partial \mu_2^2} = -\frac{n}{\sigma_2^2(1-\rho^2)}$$

$$L_{23} = \frac{\partial^2 L}{\partial \mu_2 \partial \sigma_1} = \frac{n\rho(\overline{X}_1 - \mu_1)}{\sigma_1^2 \sigma_2(1-\rho^2)}$$

$$L_{24} = \frac{\partial^2 L}{\partial \mu_2 \partial \sigma_2} = \frac{n}{\sigma_2^2(1-\rho^2)}\left[\frac{\rho\,(\overline{X}_1-\mu_1)}{\sigma_1} - \frac{2\,(\overline{X}_2-\mu_2)}{\sigma_2}\right]$$

$$L_{25} = \frac{\partial^2 L}{\partial \mu_2 \partial \rho} = \frac{n}{\sigma_2(1-\rho^2)^2}\left[\frac{2\rho\,(\overline{X}_2-\mu_2)}{\sigma_2} - \frac{(1+\rho^2)\,(\overline{X}_1-\mu_1)}{\sigma_1}\right]$$

$$L_{33} = \frac{\partial^2 L}{\partial \sigma_1^2} = \frac{1}{\sigma_1^2(1-\rho^2)}\left[n\,(1-\rho^2) - \frac{3\sum_{j=1}^{n}(X_{1j}-\mu_1)^2}{\sigma_1^2} + \frac{2\rho\sum_{j=1}^{n}(X_{1j}-\mu_1)\,(X_{2j}-\mu_2)}{\sigma_1\sigma_2}\right]$$

$$L_{34} = \frac{\partial^2 L}{\partial \sigma_1 \partial \sigma_2} = \frac{\rho\sum_{j=1}^{n}(X_{1j}-\mu_1)\,(X_{2j}-\mu_2)}{\sigma_1^2\sigma_2^2(1-\rho^2)}$$

$$L_{35} = \frac{\partial^2 L}{\partial \sigma_1 \partial \rho} = \frac{1}{\sigma_1(1-\rho^2)^2}\left[\frac{2\rho\sum_{j=1}^{n}(X_{1j}-\mu_1)^2}{\sigma_1^2} - \frac{(1+\rho^2)\sum_{j=1}^{n}(X_{1j}-\mu_1)\,(X_{2j}-\mu_2)}{\sigma_1\sigma_2}\right]$$

$$L_{44} = \frac{\partial^2 L}{\partial \sigma_2^2} = \frac{1}{\sigma_2^2(1-\rho^2)}\left[n\,(1-\rho^2) - \frac{3\sum_{j=1}^{n}(X_{2j}-\mu_2)^2}{\sigma_2^2} + \frac{2\rho\sum_{j=1}^{n}(X_{1j}-\mu_1)\,(X_{2j}-\mu_2)}{\sigma_1\sigma_2}\right]$$

$$L_{45} = \frac{\partial^2 L}{\partial \sigma_2 \partial \rho} = \frac{1}{\sigma_2(1-\rho^2)}\left[\frac{2\rho\sum_{j=1}^{n}(X_{2j}-\mu_2)^2}{\sigma_2^2} - \frac{(1+\rho^2)\sum_{j=1}^{n}(X_{1j}-\mu_1)\,(X_{2j}-\mu_2)}{\sigma_1\sigma_2}\right]$$

$$L_{SS} = \frac{\partial^2 L}{\partial \rho^2} = \frac{1}{(1-\rho^2)^2}\left[n(\rho^2+1) - \left[\frac{3\rho^2+1}{1-\rho^2}\right]\left[\frac{\sum_{j=1}^n (X_{1j}-\mu_1)^2}{\sigma_1^2} + \frac{\sum_{j=1}^n (X_{2j}-\mu_2)^2}{\sigma_2^2}\right] + \left[\frac{2\rho(3+\rho^2)}{1-\rho^2}\right]\left[\frac{\sum_{j=1}^n (X_{1j}-\mu_1)(X_{2j}-\mu_2)}{\sigma_1\sigma_2}\right] \right]$$

Taking expectations, and after some algebra the inverse of Σ_θ is found to be

$$\Sigma_\theta^{-1} = \begin{bmatrix} \dfrac{n}{\sigma_1^2(1-\rho^2)} & \dfrac{-n\rho}{\sigma_1\sigma_2(1-\rho^2)} & 0 & 0 & 0 \\[2mm] \dfrac{-n\rho}{\sigma_1\sigma_2(1-\rho^2)} & \dfrac{n}{\sigma_2^2(1-\rho^2)} & 0 & 0 & 0 \\[2mm] 0 & 0 & \dfrac{n(2-\rho^2)}{\sigma_1^2(1-\rho^2)} & \dfrac{-n\rho^2}{\sigma_1\sigma_2(1-\rho^2)} & \dfrac{-n\rho}{\sigma_1(1-\rho^2)} \\[2mm] 0 & 0 & \dfrac{-n\rho^2}{\sigma_1\sigma_2(1-\rho^2)} & \dfrac{n(2-\rho^2)}{\sigma_2^2(1-\rho^2)} & \dfrac{-n\rho}{\sigma_2(1-\rho^2)} \\[2mm] 0 & 0 & \dfrac{-n\rho}{\sigma_1(1-\rho^2)} & \dfrac{-n\rho}{\sigma_2(1-\rho^2)} & \dfrac{n(1+\rho^2)}{(1-\rho^2)^2} \end{bmatrix}$$

Jeffreys invariant prior is the square root of the determinant of

$$\frac{\Sigma_\theta^{-1}}{n} = \frac{1}{(1-\rho^2)}\begin{bmatrix} \dfrac{1}{\sigma_1^2} & \dfrac{-\rho}{\sigma_1\sigma_2} & 0 & 0 & 0 \\[2mm] \dfrac{-\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2} & 0 & 0 & 0 \\[2mm] 0 & 0 & \dfrac{2-\rho^2}{\sigma_1^2} & \dfrac{-\rho^2}{\sigma_1\sigma_2} & \dfrac{-\rho}{\sigma_1} \\[2mm] 0 & 0 & \dfrac{-\rho^2}{\sigma_1\sigma_2} & \dfrac{2-\rho^2}{\sigma_2^2} & \dfrac{-\rho}{\sigma_2} \\[2mm] 0 & 0 & \dfrac{-\rho}{\sigma_1} & \dfrac{-\rho}{\sigma_2} & \dfrac{1+\rho^2}{1-\rho^2} \end{bmatrix}$$

so that

$$\det\left[\frac{\sum_{\theta}^{-1}}{n}\right] = \det\left|\frac{1}{(1-\rho^2)}\begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} & 0 & 0 & 0 \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{2-\rho^2}{\sigma_1^2} & \frac{-\rho^2}{\sigma_1\sigma_2} & \frac{-\rho}{\sigma_1} \\ 0 & 0 & \frac{-\rho^2}{\sigma_1\sigma_2} & \frac{2-\rho^2}{\sigma_2^2} & \frac{-\rho}{\sigma_2} \\ 0 & 0 & \frac{-\rho}{\sigma_1} & \frac{-\rho}{\sigma_2} & \frac{1+\rho^2}{1-\rho^2} \end{bmatrix}\right|$$

$$\det\left[\frac{\sum_{\theta}^{-1}}{n}\right] = \frac{1}{(1-\rho^2)^5}\det\left|\begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} & 0 & 0 & 0 \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{2-\rho^2}{\sigma_1^2} & \frac{-\rho^2}{\sigma_1\sigma_2} & \frac{-\rho}{\sigma_1} \\ 0 & 0 & \frac{-\rho^2}{\sigma_1\sigma_2} & \frac{2-\rho^2}{\sigma_2^2} & \frac{-\rho}{\sigma_2} \\ 0 & 0 & \frac{-\rho}{\sigma_1} & \frac{-\rho}{\sigma_2} & \frac{1+\rho^2}{1-\rho^2} \end{bmatrix}\right|$$

which after some algebra becomes

$$\det\left[\frac{\sum_{\theta}^{-1}}{n}\right] = \frac{1}{(1-\rho^2)^5}\left[\frac{4(1-\rho^2)}{\sigma_1^4\sigma_2^4}\right] = \frac{4}{\sigma_1^4\sigma_2^4(1-\rho^2)^4}$$

and so Jeffreys' invariant prior is

$$g^*(\underline{\theta}) \propto \sqrt{\frac{4}{\sigma_1^4 \sigma_2^4 (1-\rho^2)^4}} \propto \frac{2}{\sigma_1^2 \sigma_2^2 (1-\rho^2)^2}$$

which is basically the square of the noninformative prior. This result may have implications for other Bayesian methods, but for the two approximations under consideration in this endeavour the log of the prior is the important quantity. As such, the two log-priors differ only by an multiplicative constant and are therefore equivalent for our purposes.

4.3 Lindley's Approximation

Recall Lindley's estimator $u^*(\underline{\theta})$ is an approximation to the Bayes estimator of the vector parameter $u(\underline{\theta})$, taking the form

$$u^*(\underline{\theta}) = u(\hat{\underline{\theta}}) + \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}[u_{ij}(\hat{\underline{\theta}}) + 2u_i(\hat{\underline{\theta}})\rho_j(\hat{\underline{\theta}})]\sigma_{ij} + \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\sum_{k=1}^{P}\sum_{\ell=1}^{P} L_{ijk}u_\ell\sigma_{ij}\sigma_{k\ell}$$

<div align="right">(4.1)</div>

where

$\underline{\theta}$ = maximum likelihood estimator

u_i, u_{ij} = first and second partial derivatives of $u(\underline{\theta})$

ρ_j = first partial derivative of the log-prior

σ_{ij} = element of the asymptotic variance-covariance matrix

L_{ijk} = third partial derivative of the log-likelihood function

p = dimensionality of $\underline{\theta}$.

Recall that all functions are evaluated at the mle. For easier reading, the mle "hats" will

be omitted from our formulations except where ambiguity requires clarification. For estimating any of the systemic parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, the u_i's will all be zero except for the partial pertaining to the appropriate parameter and the u_{ij}'s will all be zero.

For the prior distribution

$$g(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{k}{\sigma_1 \sigma_2 (1 - \rho^2)}$$

the log prior function is, apart from the constant k,

$$\rho(\underline{\theta}) = \log[g(\underline{\theta})] = -\log\sigma_1 - \log\sigma_2 - \log(1 - \rho^2)$$

and the partial derivatives (ρ_j's) are

$$\rho_1(\underline{\theta}) = \frac{\partial \rho(\underline{\theta})}{\partial \mu_1} = 0$$

$$\rho_2(\underline{\theta}) = \frac{\partial \rho(\underline{\theta})}{\partial \mu_2} = 0$$

$$\rho_3(\underline{\theta}) = \frac{\partial \rho(\underline{\theta})}{\partial \sigma_1} = -\frac{1}{\sigma_1}$$

$$\rho_4(\underline{\theta}) = \frac{\partial \rho(\underline{\theta})}{\partial \sigma_2} = -\frac{1}{\sigma_2}$$

$$\rho_5(\underline{\theta}) = \frac{\partial \rho(\underline{\theta})}{\partial \rho} = \frac{2\rho}{(1 - \rho^2)}$$

The σ_{ij}'s are found via the inverse of the matrix Σ seen earlier whose elements are the negative of the second partial derivatives, evaluated at the mle.

To find the inverse of the information matrix, it is first important to note that the matrix is of the form

$$\Sigma_{\underline{\theta}}^{-1} = C \begin{bmatrix} A_{2x2} & 0_{2x3} \\ 0_{3x2} & B_{3x3} \end{bmatrix}$$

whose inverse may be found from Anderson (1958), page 342, by

$$\Sigma_{\underline{\theta}} = C^{-1} \begin{bmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{bmatrix}$$

where C is a constant.

For the given matrix,

$$C = \frac{n}{(1-\rho^2)}$$

$$A = \begin{bmatrix} \dfrac{1}{\sigma_1^2} & \dfrac{-\rho}{\sigma_1\sigma_2} \\ \dfrac{-\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2} \end{bmatrix}$$

$$B = \begin{bmatrix} \dfrac{2-\rho^2}{\sigma_1^2} & \dfrac{-\rho^2}{\sigma_1\sigma_2} & \dfrac{-\rho}{\sigma_1} \\ \dfrac{-\rho^2}{\sigma_1\sigma_2} & \dfrac{2-\rho^2}{\sigma_2^2} & \dfrac{-\rho}{\sigma_2} \\ \dfrac{-\rho}{\sigma_1} & \dfrac{-\rho}{\sigma_2} & \dfrac{1+\rho^2}{1-\rho^2} \end{bmatrix}$$

so that the inverses of A and B are obtainable.

This produces the required matrix as

$$
\Sigma_{\underline{\theta}} = \frac{(1-\rho^2)}{n}
\begin{bmatrix}
\dfrac{\sigma_1^2}{1-\rho^2} & \dfrac{\rho\sigma_1\sigma_2}{1-\rho^2} & 0 & 0 & 0 \\[2ex]
\dfrac{\rho\sigma_1\sigma_2}{1-\rho^2} & \dfrac{\sigma_2^2}{1-\rho^2} & 0 & 0 & 0 \\[2ex]
0 & 0 & \dfrac{\sigma_1^2}{2(1-\rho^2)} & \dfrac{\rho^2\sigma_1\sigma_2}{2(1-\rho^2)} & \dfrac{\rho\sigma_1}{2} \\[2ex]
0 & 0 & \dfrac{\rho^2\sigma_1\sigma_2}{2(1-\rho^2)} & \dfrac{\sigma_2^2}{2(1-\rho^2)} & \dfrac{\rho\sigma_2}{2} \\[2ex]
0 & 0 & \dfrac{\rho\sigma_1}{2} & \dfrac{\rho\sigma_2}{2} & 1-\rho^2
\end{bmatrix}
$$

or more simply

$$
\Sigma_{\underline{\theta}} = \frac{1}{n}
\begin{bmatrix}
\sigma_1^2 & \rho\sigma_1\sigma_2 & 0 & 0 & 0 \\[2ex]
\rho\sigma_1\sigma_2 & \sigma_2^2 & 0 & 0 & 0 \\[2ex]
0 & 0 & \dfrac{\sigma_1^2}{2} & \dfrac{\rho^2\sigma_1\sigma_2}{2} & \dfrac{\rho\sigma_1(1-\rho^2)}{2} \\[2ex]
0 & 0 & \dfrac{\rho^2\sigma_1\sigma_2}{2} & \dfrac{\sigma_2^2}{2} & \dfrac{\rho\sigma_2(1-\rho^2)}{2} \\[2ex]
0 & 0 & \dfrac{\rho\sigma_1(1-\rho^2)}{2} & \dfrac{\rho\sigma_2(1-\rho^2)}{2} & (1-\rho^2)^2
\end{bmatrix}
$$

Lindley's approximation formula (4.1) simplifies in application to the bivariate normal because of a considerable number of zero terms, when evaluated at the mle. Substituting the u_i, u_{ij} and ρ_j terms

$$
u^*(\underline{\theta}) = u(\underline{\hat{\theta}}) + \sum_{j=3}^{5} \rho_j(\underline{\hat{\theta}})\,\sigma_{ij} + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} L_{ijk}\sigma_{ij}\sigma_{kl} \tag{4.2}
$$

Note that in the second term the summation goes from $j=3$ to $j=5$ rather than from 1 to 5 because ρ_1 and ρ_2 are zero. Further, the l subscript represents which of the five systemic parameters is being estimated and is therefore a constant for each approximation.

At this point the approximation formula 4.1 simplifies to varying degrees depending upon which systemic parameter is being estimated. For each parameter, it is a straightforward, albeit tedious, problem of algebraically evaluating each of the 125 terms in equation 4.2 to produce a closed-form for Lindley's approximation to the Bayes estimator. Of the 125 terms, only 35 are actually distinct because the subscripts of L_{ijk} can be transposed. The population means will be dealt with first, followed by the standard deviations and finally the correlation coefficient.

In estimating the population means, inspection of the Σ_θ matrix reveals $\sigma_{13}=\sigma_{14}=\sigma_{15}=\sigma_{23}=\sigma_{24}=\sigma_{25}=0$ so that the third term of (4.2) will be zero for all $k>2$. For $k=1$ and $k=2$ it can be seen that most of the L_{ijk} terms are zero when evaluated at the mle. Going through an elimination process, it can be shown that all of the terms in the triple summation of 4.2 are zero. As such, Lindley's approximation for the Bayes estimator of the population mean is merely

$$\mu_i^* = \hat{\mu}_i = \overline{X}_i \quad for \; i=1,2$$

Next consider the population standard deviations so that

$$u^*(\underline{\theta}) = \sigma_i \quad for \; i=1,2$$

For this exercise $u(\theta)=\sigma_1$ will be used throughout without loss of generality. Starting from 4.2, the second term is

$$\sum_{j=3}^{p} \rho_j(\underline{\theta})\,\sigma_{3j} = -\frac{1}{\sigma_1}\left[\frac{\sigma_1^2}{2n}\right] - \frac{1}{\sigma_2}\left[\frac{\rho^2\sigma_1\sigma_2}{2n}\right] + \frac{2\rho}{1-\rho^2}\left[\frac{\rho\sigma_1(1-\rho^2)}{2n}\right] = -\frac{\sigma_1(1-\rho^2)}{2n}$$

The fact that several L_{ijk} and σ_{ij} terms are zero when evaluated at the mle causes only 17 of the 35 distinct terms from the 125 term triple summation to be nonzero. Specifically,

the approximation formula 4.2 becomes

$$\sigma_1^* = \hat{\sigma}_1 - \frac{\hat{\sigma}_1(1-\hat{\rho}^2)}{2n} + \frac{1}{2}L_{113}\sigma_{11}\sigma_{33} + \frac{1}{2}L_{115}\sigma_{11}\sigma_{53} + L_{123}\sigma_{12}\sigma_{33} + L_{124}\sigma_{12}\sigma_{43} + L_{125}\sigma_{12}\sigma_{53}$$

$$+ \frac{1}{2}L_{224}\sigma_{22}\sigma_{43} + \frac{1}{2}L_{225}\sigma_{22}\sigma_{53} + \frac{1}{2}L_{333}\sigma_{33}^2 + \frac{3}{2}L_{334}\sigma_{33}\sigma_{43} + \frac{3}{2}L_{335}\sigma_{33}\sigma_{53}$$

$$+ \frac{1}{2}L_{344}(2\sigma_{34}^2 + \sigma_{33}\sigma_{44}) + \frac{1}{2}L_{345}(4\sigma_{34}\sigma_{35} + 2\sigma_{33}\sigma_{45}) + \frac{1}{2}L_{355}(2\sigma_{35}^2 + \sigma_{33}\sigma_{55})$$

$$+ \frac{1}{2}L_{444}\sigma_{44}\sigma_{43} + \frac{1}{2}L_{445}(2\sigma_{34}\sigma_{45} + \sigma_{44}\sigma_{35}) + \frac{1}{2}L_{455}(2\sigma_{45}\sigma_{35} + \sigma_{55}\sigma_{34}) + \frac{1}{2}L_{555}\sigma_{55}\sigma_{53}$$

which, after substituting the third partial derivatives and σ_{ij} elements from above becomes

$$\sigma_1^* = \hat{\sigma}_1 - \frac{\hat{\sigma}_1(1-\hat{\rho}^2)}{2n} + \frac{\hat{\sigma}_1}{2n}\left(\frac{9}{2} - 2\hat{\rho}^2\right)$$

$$= \hat{\sigma}_1 + \frac{\hat{\sigma}_1}{4n}(7 - 2\hat{\rho}^2)$$

Similarly results are obtained for σ_2 merely by substitution.

The closed form solution for Lindley's approximation with regard to the population standard deviation is quite easy to calculate and use in practice. As is expected, asymptotically the Bayes estimator converges to the mle and will have a normal posterior distribution. The exact distribution for the estimator under small samples is more complex to derive and is perhaps a direction for future work. The posterior variance of the approximation will be close to the variance of the mle for large samples.

The final systemic parameter for which a closed form approximation is needed is the traditionally difficult ρ. Once again the starting point is equation 4.2.

The first part of the equation becomes

$$\sum_{j=3}^{5} \rho_j(\underline{\hat{\theta}})\, \sigma_{5j} = -\frac{1}{\sigma_1}\left[\frac{\rho\sigma_1(1-\rho^2)}{2n}\right] - \frac{1}{\sigma_2}\left[\frac{\rho\sigma_2(1-\rho^2)}{2n}\right] + \frac{2\rho}{1-\rho^2}\left[\frac{(1-\rho^2)^2}{n}\right]$$

$$= \frac{\rho(1-\rho^2)}{n}$$

The next step involves examination of the 125-term triple summation. Zero terms and redundancies reduce the form substantially to involve 13 distinct nonzero terms. Specifically,

$$\rho^* = \hat{\rho} - \frac{\hat{\rho}(1-\hat{\rho}^2)}{n} + \frac{1}{2}L_{113}\sigma_{11}\sigma_{35} + \frac{1}{2}L_{115}\sigma_{11}\sigma_{55} + L_{123}\sigma_{12}\sigma_{35} + L_{124}\sigma_{12}\sigma_{45} + L_{125}\sigma_{12}\sigma_{55}$$

$$+ \frac{1}{2}L_{224}\sigma_{22}\sigma_{45} + \frac{1}{2}L_{225}\sigma_{22}\sigma_{55} + \frac{1}{2}L_{333}\sigma_{33}\sigma_{35} + \frac{1}{2}L_{334}(2\sigma_{34}\sigma_{35} + \sigma_{33}\sigma_{45})$$

$$+ \frac{1}{2}L_{335}(2\sigma_{35}^2 + \sigma_{33}\sigma_{55}) + \frac{1}{2}L_{344}(2\sigma_{34}\sigma_{45} + \sigma_{44}\sigma_{35}) + \frac{1}{2}L_{345}(4\sigma_{35}\sigma_{45} + 2\sigma_{34}\sigma_{55})$$

$$+ \frac{3}{2}L_{355}\sigma_{35}\sigma_{55} + \frac{1}{2}L_{444}\sigma_{44}\sigma_{45} + \frac{1}{2}L_{445}(2\sigma_{45}^2 + \sigma_{44}\sigma_{55}) + \frac{3}{2}L_{455}\sigma_{45}\sigma_{55} + \frac{1}{2}L_{555}\sigma_{55}^2$$

Now putting the pieces together

$$\rho^* = \hat{\rho} + \frac{\hat{\rho}(1-\hat{\rho}^2)}{n} - \frac{5\hat{\rho}(1-\hat{\rho}^2)}{2n}$$

$$= \hat{\rho} - \frac{3\hat{\rho}(1-\hat{\rho}^2)}{2n}$$

Again the Lindley approximation will tend towards the mle as n increases.

As mentioned in Chapter one, Lindley's approximation formula lends itself to algebraic closed-form solutions much more readily than does the Tierney-Kadane method. Although the Lindley method requires construction and evaluation of the third partial derivatives, once derived that section of the approximation formula remains constant regardless of the prior distribution. This is an important point for algebraic work since

it means altering the prior does not necessitate starting the algebraic process from the beginning. An altered prior demands only a reconstruction of the first summation term of 4.2, which is not tedious.

For example, suppose the a priori state of knowledge about the population standard deviations was such that a natural conjugate prior as suggested by Press (1989) was appropriate. In this case one would assume independent inverted gamma priors for the two σ's so that

$$g_i(\sigma_i) \propto \frac{1}{\sigma_i^{b_i+1}} \exp\left[-\frac{a_i}{\sigma_i}\right] \quad \text{for } i=1,2$$

This new assumption would cause the joint prior distribution to be

$$g(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = \frac{1}{\sigma_1^{b_1+1}\sigma_2^{b_2+1}(1-\rho^2)} \exp\left[-\frac{a_1}{\sigma_1}-\frac{a_2}{\sigma_2}\right]$$

and the log-prior is

$$p(\underline{\theta}) = \log[g(\underline{\theta})] = -\log(1-\rho^2) - (b_1+1)\log\sigma_1 - (b_2+1)\log\sigma_2 - \frac{a_1}{\sigma_1} - \frac{a_2}{\sigma_2} .$$

The partial derivatives of the log-prior are as before except now

$$p_3(\underline{\theta}) = \frac{\partial p(\underline{\theta})}{\partial \sigma_1} = \frac{a_1 - (b_1+1)\sigma_1}{\sigma_1^2}$$

$$p_4(\underline{\theta}) = \frac{\partial p(\underline{\theta})}{\partial \sigma_2} = \frac{a_2 - (b_2+1)\sigma_2}{\sigma_2^2}$$

and, after substitution into the first part of 4.2

$$\sum_{j=3}^{p} \rho_j(\underline{\theta}) \sigma_{3j} = \left[\frac{a_1 - (b_1+1)\sigma_1}{\sigma_1^2}\right]\left[\frac{\sigma_1^2}{2n}\right] + \left[\frac{a_2 - (b_2+1)\sigma_2}{\sigma_2^2}\right]\left[\frac{\rho^2\sigma_1\sigma_2}{2n}\right] + \frac{2\rho}{1-\rho^2}\left[\frac{\rho\sigma_1(1-\rho^2)}{2n}\right]$$

$$= \frac{\sigma_1}{2n}\left[a_1 - b_1 - 1 + \rho^2\left[\frac{a_2}{\sigma_2} - b_2 + 1\right]\right]$$

so that the revised Bayes estimator is

$$\sigma_1^* = \hat{\sigma}_1 + \frac{\hat{\sigma}_1}{2n}\left[a_1 - b_1 - 1 + \hat{\rho}^2\left[\frac{a_2}{\hat{\sigma}_2} - b_2 + 1\right]\right] + \frac{\hat{\sigma}_1}{2n}\left(\frac{9}{2} - 2\hat{\rho}^2\right)$$

$$= \hat{\sigma}_1 + \frac{\hat{\sigma}_1}{2n}\left[\frac{7}{2} - \hat{\rho}^2\left[1 + b_2 - \frac{a_2}{\hat{\sigma}_2}\right] + a_1 - b_1\right]$$

or, one can express the relationship between the estimators produced under the noninformative (NI) and conjugate (CON) prior conditions as

$$\sigma_{1_{CON}}^* = \sigma_{1_{NI}}^* + \frac{\sigma_1}{2n}\left[a_1 - b_1 - \left[b_2 - \frac{a_2}{\hat{\sigma}_2}\right]\hat{\rho}^2\right]$$

Similarly, with indices reversed as above, for σ_2.

For ρ, again the only part of 4.2 that changes is the second term to be

$$\sum_{j=3}^{5} \rho_j(\underline{\theta}) \sigma_{5j} = \left[\frac{a_1 - (b_1+1)\sigma_1}{\sigma_1^2}\right]\left[\frac{\rho\sigma_1(1-\rho^2)}{2n}\right] + \left[\frac{a_2 - (b_2+1)\sigma_2}{\sigma_2^2}\right]\left[\frac{\rho\sigma_2(1-\rho^2)}{2n}\right]$$

$$+ \frac{2\rho}{1-\rho^2}\left[\frac{(1-\rho^2)^2}{n}\right]$$

$$= \frac{\rho(1-\rho^2)}{2n}\left[\frac{a_1}{\sigma_1} + \frac{a_2}{\sigma_2} - b_1 - b_2 + 2\right]$$

so that under the inverted gamma prior assumptions, Lindley's approximation to the

Bayes estimator of ρ is

$$
\begin{aligned}
\overset{\bullet}{\rho}_{CON} &= \hat{\rho} + \frac{\hat{\rho}(1-\hat{\rho}^2)}{n}\left[\frac{a_1}{\hat{\sigma}_1} + \frac{a_2}{\hat{\sigma}_2} - b_1 - b_2 + 2\right] - \frac{5\hat{\rho}(1-\hat{\rho}^2)}{2n} \\
&= \hat{\rho} - \frac{\hat{\rho}(1-\hat{\rho}^2)}{2n}\left[3 - \frac{a_1}{\hat{\sigma}_1} - \frac{a_2}{\hat{\sigma}_2} + b_1 + b_2\right] \\
&= \overset{\bullet}{\rho}_{NI} + \frac{\hat{\rho}(1-\hat{\rho}^2)}{2n}\left[\frac{a_1}{\hat{\sigma}_1} + \frac{a_2}{\hat{\sigma}_2} - b_1 - b_2\right].
\end{aligned}
$$

It is a straightforward task therefore for a practitioner in an applied setting to investigate algebraically the impact of various a priori assumptions via Lindley's approximation. One needs only to reconstruct the first part of the approximation formula to adjust for different prior information.

4.4 Tierney-Kadane Approximation

In this section the Tierney-Kadane (1986) approximation to the Bayes estimator will be applied to the bivariate normal distribution. As described in Chapter one, the Tierney-Kadane approximation to the Bayes estimator is

$$
u^{*}(\theta) = \sqrt{\frac{\det\sum_*}{\det\sum_0} \cdot \exp\left[n\{L_*(\theta_*) - L_0(\theta_0)\}\right]}
$$

where

$$
L_0(\theta) = \frac{\left(\log[g(\theta)] + L(\theta|\underline{x})\right)}{n}
$$

$$
L_*(\theta) = L_0 + \frac{\log[u(\theta)]}{n}
$$

and the points θ_0 and θ_* are maximizing points of the L_0 and L_* functions respectively. The Σ_0 and Σ_* matrices are the inverses of matrices that are comprised of elements of the

negative of the second partial derivatives of the L_0 and L_* functions respectively.

For the bivariate normal distribution, assuming the noninformative prior situation

$$L_0 = -\left(\frac{1}{n}+1\right)(\log\sigma_1+\log\sigma_2) -\left(\frac{1}{n}+\frac{1}{2}\right)\log(1-\rho^2)-\log(2\Pi)$$

$$-\frac{1}{2n(1-\rho^2)}\sum_{j=1}^{n}\left[\frac{(X_{1j}-\mu_1)^2}{\sigma_1^2}+\frac{(X_{2j}-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(X_{1j}-\mu_1)(X_{2j}-\mu_2)}{\sigma_1\sigma_2}\right]$$

which remains constant regardless of the parameter being estimated. For estimating each

element of the vector parameter $\underline{\theta}$, a separate L_* function is needed. For the sake of

brevity, these may be represented as

$$L_{*_i} = \frac{1}{n}\log(\theta_i)+L_0 \quad for\ i=1,2,\ldots,5\ .$$

Six separate maximization points must be found (θ_0 and five θ_*'s). As an example, to

find the maximization point algebraically for L_0 one must take five partial derivatives to

produce a system of five equations with five unknowns.

$$\frac{\partial L_0}{\partial\mu_1} = -\frac{1}{n(1-\rho^2)}\sum_{j=1}^{n}\left[\frac{\rho(X_{2j}-\mu_2)}{\sigma_1\sigma_2}-\frac{(X_{1j}-\mu_1)}{\sigma_1}\right]$$

$$\frac{\partial L_0}{\partial\mu_2} = -\frac{1}{n(1-\rho^2)}\sum_{j=1}^{n}\left[\frac{\rho(X_{1j}-\mu_1)}{\sigma_1\sigma_2}-\frac{(X_{2j}-\mu_2)}{\sigma_2}\right]$$

$$\frac{\partial L_0}{\partial\sigma_1} = -\left(\frac{1}{n}+1\right)\frac{1}{\sigma_1}-\frac{1}{n\sigma_1(1-\rho^2)}\sum_{j=1}^{n}\left[\frac{\rho(X_{1j}-\mu_1)(X_{2j}-\mu_2)}{\sigma_1\sigma_2}-\frac{(X_{1j}-\mu_1)}{\sigma_1^2}\right]$$

$$\frac{\partial L_0}{\partial\sigma_2} = -\left(\frac{1}{n}+1\right)\frac{1}{\sigma_2}-\frac{1}{n\sigma_2(1-\rho^2)}\sum_{j=1}^{n}\left[\frac{\rho(X_{1j}-\mu_1)(X_{2j}-\mu_2)}{\sigma_1\sigma_2}-\frac{(X_{2j}-\mu_2)}{\sigma_2^2}\right]$$

$$\frac{\partial L_0}{\partial \rho} = \frac{2\rho}{1-\rho^2}\left(\frac{1}{n}+\frac{1}{2}\right)+\frac{1}{n(1-\rho^2)}\sum_{j=1}^{n}\left[\frac{(X_{1j}-\mu_1)\,(X_{2j}-\mu_2)}{\sigma_1\sigma_2}\right]$$

$$-\frac{\rho}{n(1-\rho^2)^2}\sum_{j=1}^{n}\left[\frac{(X_{1j}-\mu_1)^2}{\sigma_1^2}+\frac{(X_{2j}-\mu_2)^2}{\sigma_2^2}-\frac{2\rho\,(X_{1j}-\mu_1)\,(X_{2j}-\mu_2)}{\sigma_1\sigma_2}\right]$$

The solution to this system of equations is mathematically inconvenient in a fashion similar to what was seen for the univariate normal distribution in Chapter two. Even if this solution were obtained, the point θ_0 would have to be substituted into a 5x5 matrix and inverted algebraically. To produce closed form solutions for the five systemic parameters, a similar exercise would have to be accomplished for each L_* function. Again, the solutions to these five equation systems in five unknowns are inconvenient to the extent that working with them algebraically quickly becomes intractable. As such, it is evident that the T-K algorithm is not as well suited for algebraically closed form solutions as Lindley's method. Furthermore, even if the closed form solution were obtained for the noninformative setting, the entire exercise would have to be repeated to duplicate the effort for a change in the prior distribution. This is due to the fact that the prior distribution is directly involved with the L_* function. The maximization point θ_* will change for any but the most rudimentary change in the prior distribution, necessitating a reconstruction of the Σ_* matrices. The comparative work done in the previous section with the inverted gamma priors would be intractable via the T-K algorithm algebraically.

This is not to say that the T-K algorithm is inappropriate for application to the bivariate normal distribution. Rather it illustrates that the work must be done by numerical approximation, the means by which the method was designed to produce the

approximation to the Bayes estimator. As was seen in Chapter two, even in the simplest applications, the T-K method is algebraically inconvenient. In terms of numerical computation however, the method has some distinct advantages over Lindley's. Newton's method, is typically all that is required to produce the maximization points. Once again, the issue of applicability boils down to the availability of computer technology and the degree to which algebraic results are seen to be desirable.

4.5 Example

As an illustration, data from Pope, Lehrer and Stevens (1980) are used. Bivariate measures on reading skills were taken on 26 children with the Woodcock reading test and were demonstrated to be applicable to the bivariate normal model. Table 4.1 presents the estimation results. The estimated posterior variances were computed by the same method in Chapters one and two. This method involved separately estimating the two pieces of the posterior variance formulation.

Table 4.1: Estimation Results for Pope et al (1980) Sample

Parameter	MLE	LINDLEY	T-K	∇(MLE)	∇(LINDLEY)	∇(T-K)
μ_1	6.404	6.404	6.378	.1292	.1631	.1208
μ_2	6.869	6.869	6.797	.2908	.2673	.2656
σ_1	2.059	2.179	2.181	.0702	.0672	.0669
σ_2	2.636	2.790	2.794	.1586	.1101	.1093
ρ	0.688	0.667	0.670	.0234	.0102	.0088

Both approximations produce similar results. Note that Lindley's approximation for the population means is exactly equal to the mle whereas the T-K approximate is slightly different. The Bayes estimators for the population standard deviations are both

larger than the mle and have smaller estimated posterior variance than the mle. As noted by Tierney and Kadane (1986), however, when n is small it is easy for the estimate for the variance to be small and even to be negative if either of the two parts of the variance formula is poorly estimated. In previous work, notably with the three-parameter Weibull distribution, the author has noted negative variance estimates for even samples of moderately large size. The observed smaller variance is likely due to this phenomenon. The Bayes estimators for the correlation coefficient are both smaller than the mle and again present smaller variance estimates.

The key point to note here is that Lindley and the T-K Bayes approximate estimators under noninformative prior conditions produce reasonable estimates, relative to the mle. Lindley's is certainly no more difficult to calculate than the mle, since a closed form has been produced. The T-K estimate is gained only after considerable numerical work. The next section will explore these issues a bit deeper with the use of Monte Carlo simulation.

4.6 Monte Carlo Simulation

To compare the closed form estimators with the T-K numerical approximations, Monte Carlo simulation was undertaken. Samples were generated via the IMSL routine GGNSM (1975) and validated through the use of the empirical distribution function tests of Anderson and Darling found in D'Agostino and Stephens (1986). Any samples that failed to meet the criterion were rejected and another sample generated. Algorithms were constructed via the PL/I language on an Amdahl 470 mainframe at the University of Manitoba.

Four simulation runs of 1000 samples each were performed using samples of size 10, 25, 50 and 100 respectively with parameter settings of $\mu_1 = 14.0$, $\mu_2 = 26.0$, $\sigma_1 = 3.0$, $\sigma_2 = 6.0$ and $\rho = 0.60$. Different random seeds were used for each run.

Bayes estimators via Lindley's approximation and the Tierney-Kadane method were constructed for each sample as well as the mle's for the five systemic parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ. Estimates for the posterior variances were calculated for each parameter by the piecewise method described in Chapter one. For each estimator the root mean square error based on the 1000 samples was calculated as a relative measure of precision. The results for estimating the two population standard deviations and the population correlation coefficient are summarized in Tables 4.2 through 4.4. Results for the population means are not presented because the mle's and Lindley's approximation are identical while the T-K approximation produces near identical results.

Table 4.2: Simulation Results for $\sigma_1 = 3.0$ (1000 Samples)

	ESTIMATE AVERAGE			ROOT MSE			AVG EST VARIANCE		
n	MLE	LINDLEY	T-K	MLE	LINDLEY	T-K	MLE	LINDLEY	T-K
10	2.71	3.13	3.16	.7130	.7456	.7509	0.35	0.21	0.20
25	2.88	3.06	3.07	.4374	.4480	.4503	0.14	0.14	0.13
50	2.88	2.97	2.98	.3250	.3099	.3108	0.07	0.08	0.08
100	2.90	2.94	2.94	.2302	.2160	.2161	0.04	0.04	0.04

Table 4.3: Simulation Results for $\sigma_2=6.0$ (1000 Samples)

n	ESTIMATE AVERAGE			ROOT MSE			AVG EST VARIANCE		
	MLE	LINDLEY	T-K	MLE	LINDLEY	T-K	MLE	LINDLEY	T-K
10	5.49	6.34	6.38	1.39	1.50	1.53	1.40	0.84	0.81
25	5.78	6.14	6.16	0.85	0.88	0.90	0.56	0.55	0.53
50	5.78	5.96	5.97	0.62	0.59	0.60	0.28	0.30	0.29
100	5.77	5.86	5.86	0.49	0.46	0.46	0.14	0.16	0.16

Table 4.4: Simulation Results for $\rho=0.60$ (1000 Samples)

n	ESTIMATE AVERAGE			ROOT MSE			AVG EST VARIANCE		
	MLE	LINDLEY	T-K	MLE	LINDLEY	T-K	MLE	LINDLEY	T-K
10	.5849	.5396	.5488	.2626	.2379	.2418	.0301	.0392	.0375
25	.5910	.5698	.5706	.1350	.1378	.1396	.0120	.0165	.0159
50	.5967	.5857	.5860	.0959	.0965	.0971	.0060	.0082	.0080
100	.5975	.5919	.5920	.0679	.0683	.0684	.0030	.0041	.0040

Supplementary runs of $n=200$ and $n=400$ were also performed. The results for the three estimators become indistinguishable for these sample sizes, however, and are therefore not worth presenting. The interesting point is that for moderate to large sample sizes, any one of the three techniques will give the same estimator. This may have implication for practitioners of both schools of thought.

In estimating $\sigma_1=3.0$ in Table 4.2, the three methods provide quite different results for the smaller sample sizes. Although the average of the mle estimate over 1000 samples was below the true parametric value, both Bayes approximations overestimated σ_1, with the T-K method very slightly higher than Lindley's. The positive difference

between the mle and the Bayes approximations remains for n as large as 100, although by then the two approximations become indistinguishable from one another. The posterior variance is clearly underestimated at the small sample size and should not be given any credence. In terms of precision, the mle is a clear winner for the small sample sizes although the root mean square error for the Bayes approximations becomes smaller than that of the mle for the larger samples.

Results for σ_2 in Table 4.3 are comparable to those for σ_1. The Bayes approximations actually look to be closer on average to the true parametric value of $\sigma_2 = 6.0$ than the mle, but again for larger samples the results converge.

In estimating $\rho = 0.60$, the Bayes approximations exhibit greater precision (in terms of root mean square error) than the mle for the n=10 samples. On average, the mle is closer to the true value, however. The underestimation of the posterior variance for the Bayes approximations is not as clear for the smaller samples as it was in estimating the population standard deviations. This information is important to those who might use the approximation and assume that as long as the posterior variance estimate is positive, it must be appropriate. Traditionally ρ is the most difficult parameter to estimate for the bivariate normal. This can be reinforced by the observation that the results for the three procedures do not converge by the n=100 sample size. The supplementary runs did show, however, that the results are identical for n=200 and larger. For the smaller sample sizes the Bayes approximations tend to markedly underestimate ρ relative to the mle.

The posterior sampling distributions for the two Bayes approximations are virtually identical, even for the small sample sizes, so only the pictures of the Lindley

estimator and the mle has been given. Figures 4.1 and 4.2 provide the histograms for the mle and Lindley approximation of σ_1 respectively for the 1000 samples of size n=10. Both distributions are positively skewed with Lindley's estimator evidencing the greater variability seen in Table 4.2. The Shapiro-Wilk test supports a lack of normality for both posterior sampling distributions.

Figures 4.3 and 4.4 give the histograms for the same estimators with the 1000 samples of size n=100. These two distributions now appear normal, a hypothesis which is supported by Shapiro-Wilk testing. The difference in variability between the two estimators seen for the small sample size is no longer evident.

Figures 4.5 through 4.8 give the same figures for the σ_2 parameter. Surprisingly, the smaller sample size histograms do achieve normality (again by Shapiro-Wilk). Other than this surprising result, the comments regarding σ_1 are applicable.

The histograms in Figures 4.9 through 4.12 deal with the results for the two estimation methods on ρ for n=10 and n=100 sample sizes. Most noticeable is the shift in shape from the low to moderate sample size from an extremely negatively skewed distribution to ones approaching normality by the time n=100. The distribution is still significantly different from that of a normal distribution (p-value of 0.0001). The instability of estimating ρ is evident in the fact that negative estimates are found even at the larger sample size for the parametric value of ρ=0.60. The superior precision of the Bayes approximations for the smaller sample size mentioned in Table 4.4 is evident in that Lindley's method actually seems to produce estimates closer to the true ρ value more often than the mle for the 1000 simulated samples from Figures 4.9 and 4.10.

The n=25 and n=50 histograms are merely gradients along the continuum from

Figure 4.1

ESTIMATION OF THE BIVARIATE NORMAL σ_1 PARAMETER
RESULTS FOR 1000 SAMPLES
$N = 10$, $\mu_1 = 14.0$, $\mu_2 = 26.0$, $\sigma_1 = 3.0$, $\sigma_2 = 6.0$, $\rho = 0.60$
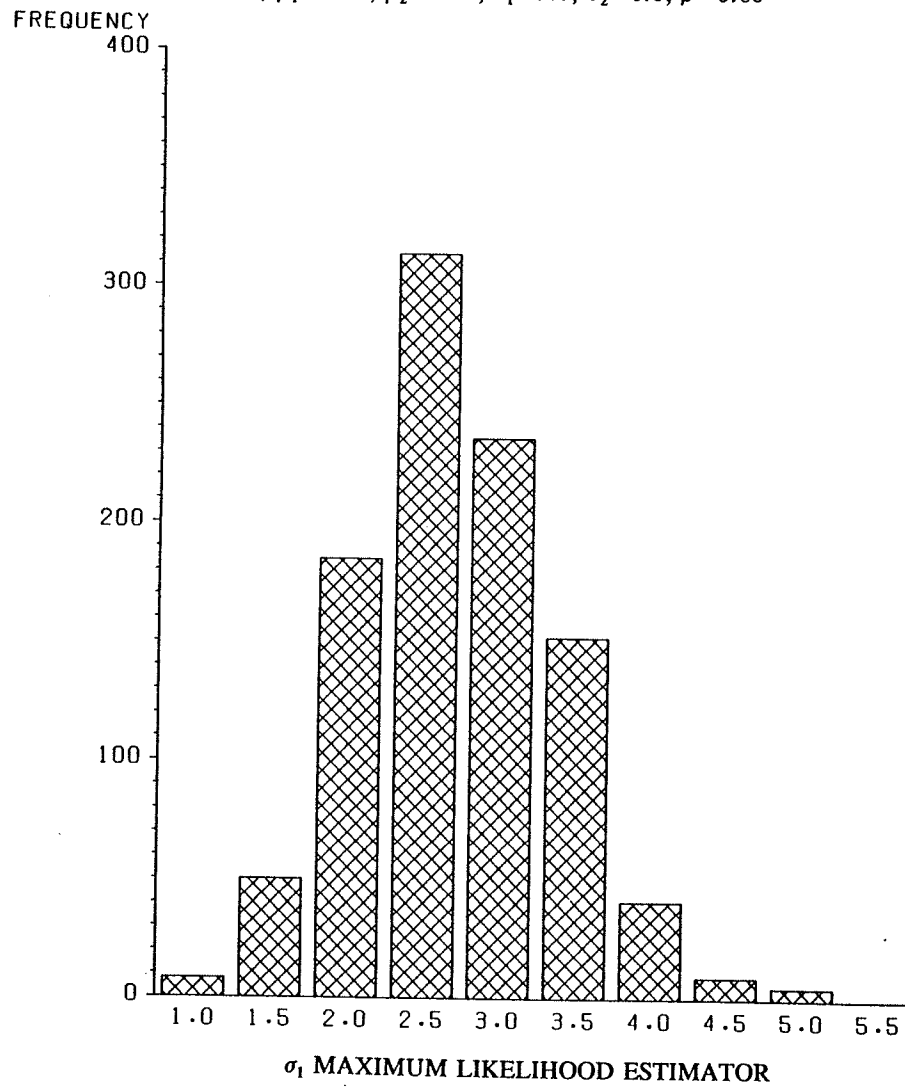
σ_1 MAXIMUM LIKELIHOOD ESTIMATOR

Figure 4.2

ESTIMATION OF THE BIVARIATE NORMAL σ_1 PARAMETER
RESULTS FOR 1000 SAMPLES
$N=10$, $\mu_1=14.0$, $\mu_2=26.0$, $\sigma_1=3.0$, $\sigma_2=6.0$, $\rho=0.60$
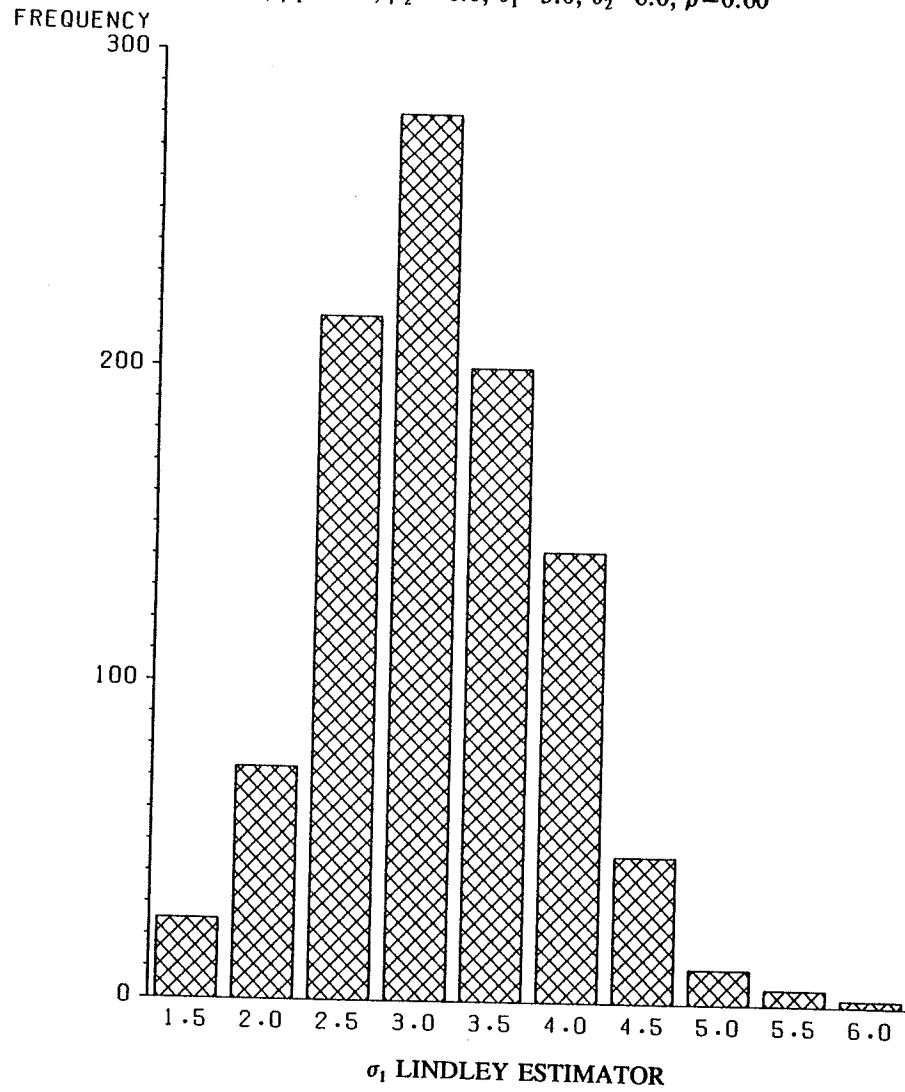
FREQUENCY

σ_1 LINDLEY ESTIMATOR

Figure 4.3   **ESTIMATION OF THE BIVARIATE NORMAL σ_1 PARAMETER**
RESULTS FOR 1000 SAMPLES
$N=100$, $\mu_1=14.0$, $\mu_2=26.0$, $\sigma_1=3.0$, $\sigma_2=6.0$, $\rho=0.60$

σ_1 MAXIMUM LIKELIHOOD ESTIMATOR

Figure 4.4   **ESTIMATION OF THE BIVARIATE NORMAL σ_1 PARAMETER**
RESULTS FOR 1000 SAMPLES
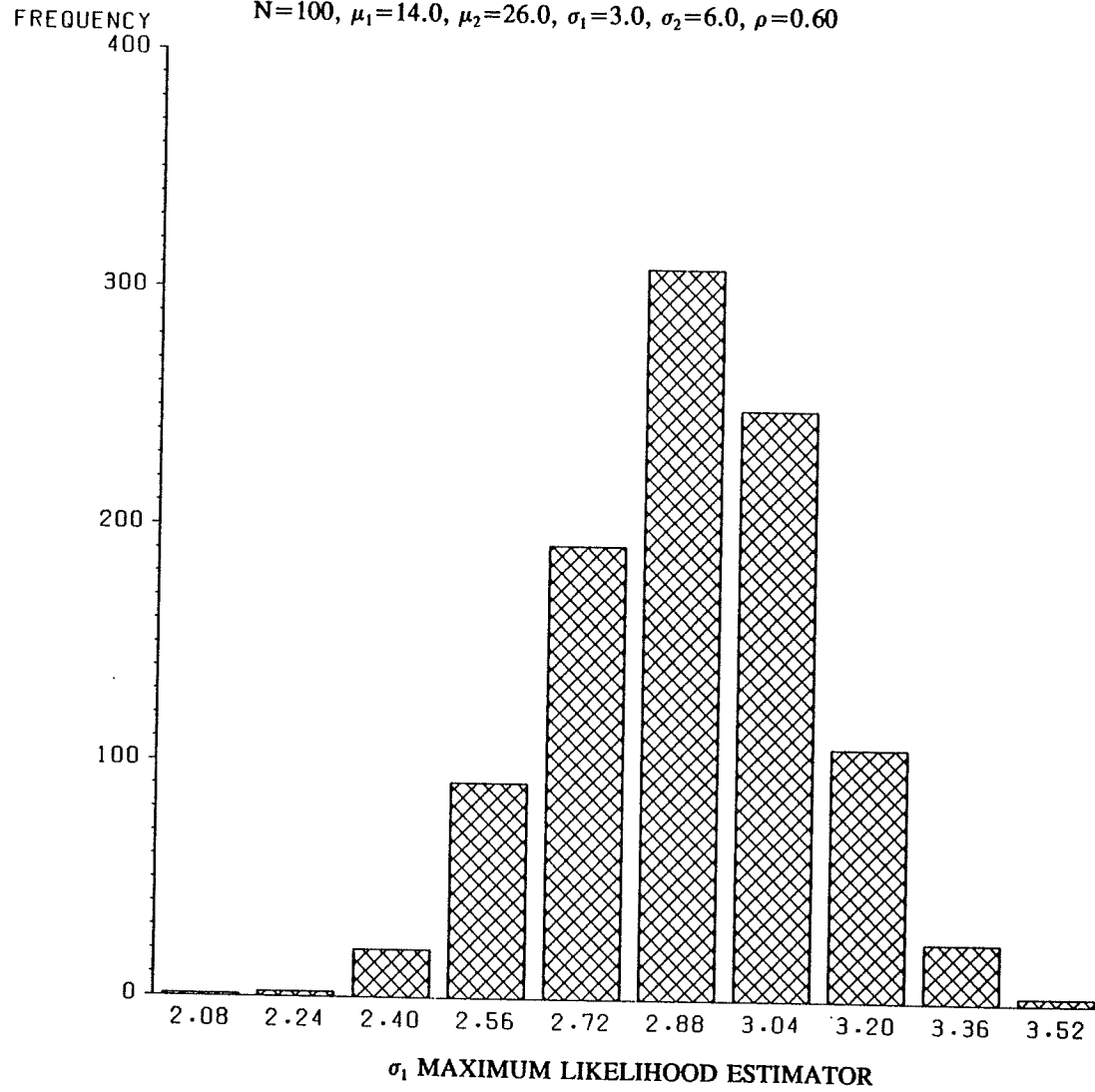$N=100$, $\mu_1=14.0$, $\mu_2=26.0$, $\sigma_1=3.0$, $\sigma_2=6.0$, $\rho=0.60$

Figure 4.5

ESTIMATION OF THE BIVARIATE NORMAL σ_2 PARAMETER
RESULTS FOR 1000 SAMPLES
$N=10$, $\mu_1=14.0$, $\mu_2=26.0$, $\sigma_1=3.0$, $\sigma_2=6.0$, $\rho=0.60$
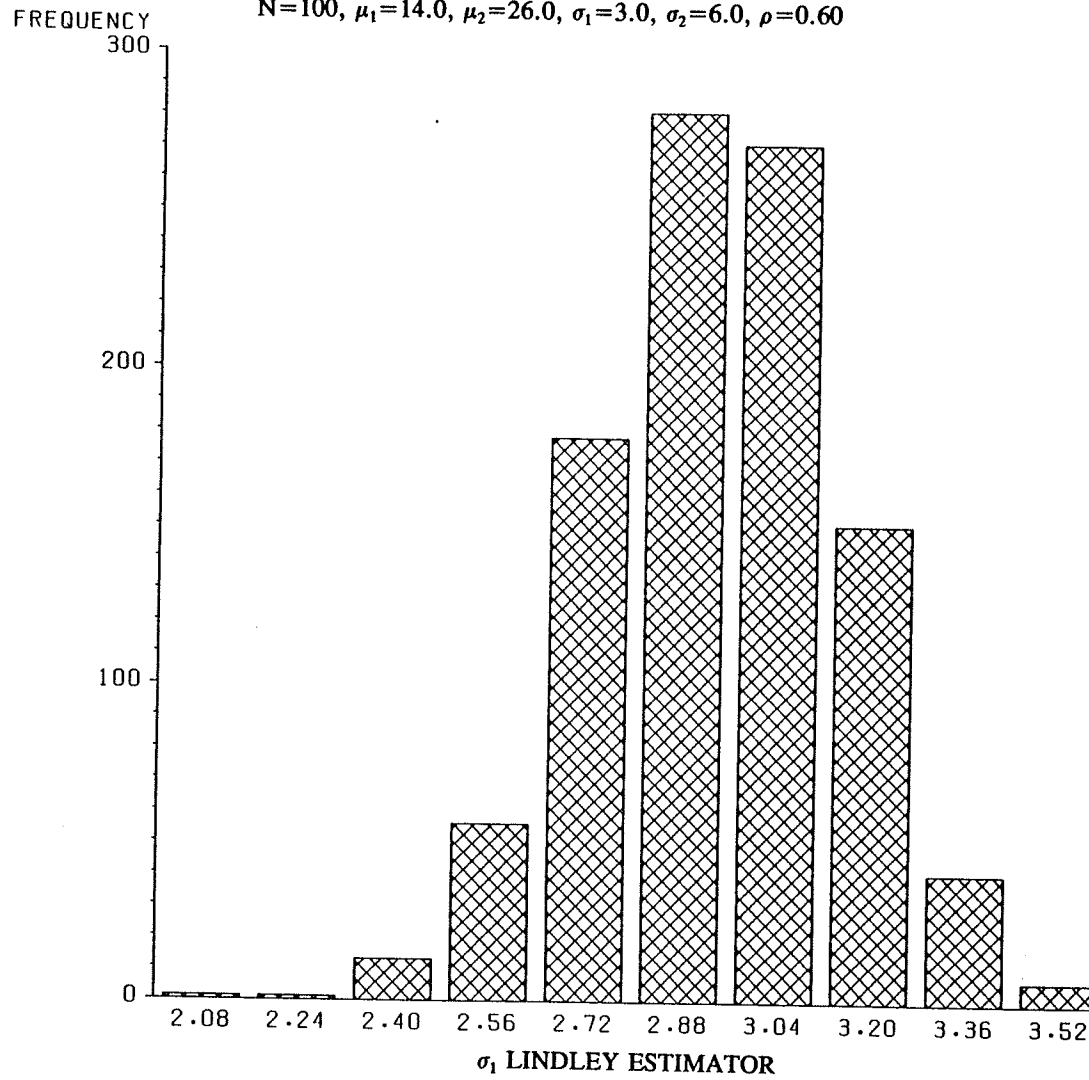
σ_2 MAXIMUM LIKELIHOOD ESTIMATOR

Figure 4.6   ESTIMATION OF THE BIVARIATE NORMAL σ_2 PARAMETER
RESULTS FOR 1000 SAMPLES
N=10, μ_1=14.0, μ_2=26.0, σ_1=3.0, σ_2=6.0, ρ=0.60

Figure 4.7   ESTIMATION OF THE BIVARIATE NORMAL σ_2 PARAMETER
RESULTS FOR 1000 SAMPLES
N=100, μ_1=14.0, μ_2=26.0, σ_1=3.0, σ_2=6.0, ρ=0.60
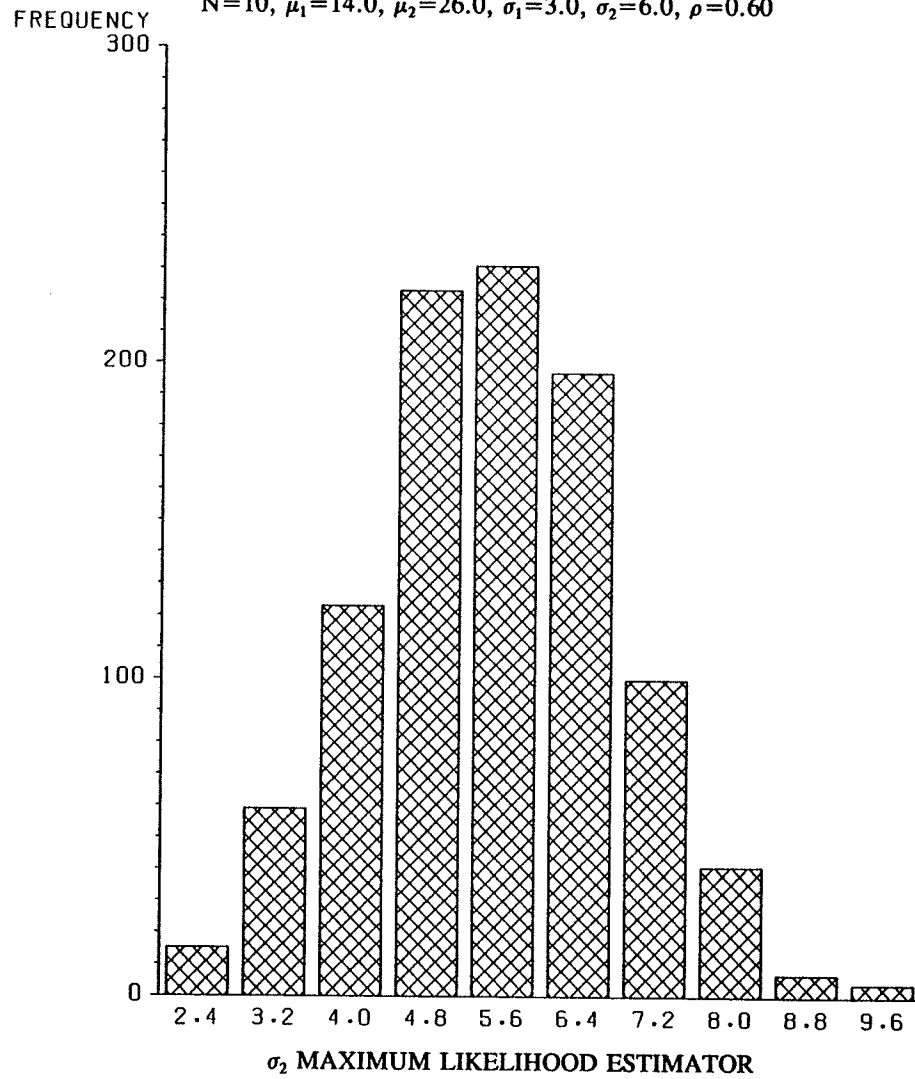
FREQUENCY

σ_2 MAXIMUM LIKELIHOOD ESTIMATOR

Figure 4.8

ESTIMATION OF THE BIVARIATE NORMAL σ_2 PARAMETER
RESULTS FOR 1000 SAMPLES
N=100, μ_1=14.0, μ_2=26.0, σ_1=3.0, σ_2=6.0, ρ=0.60
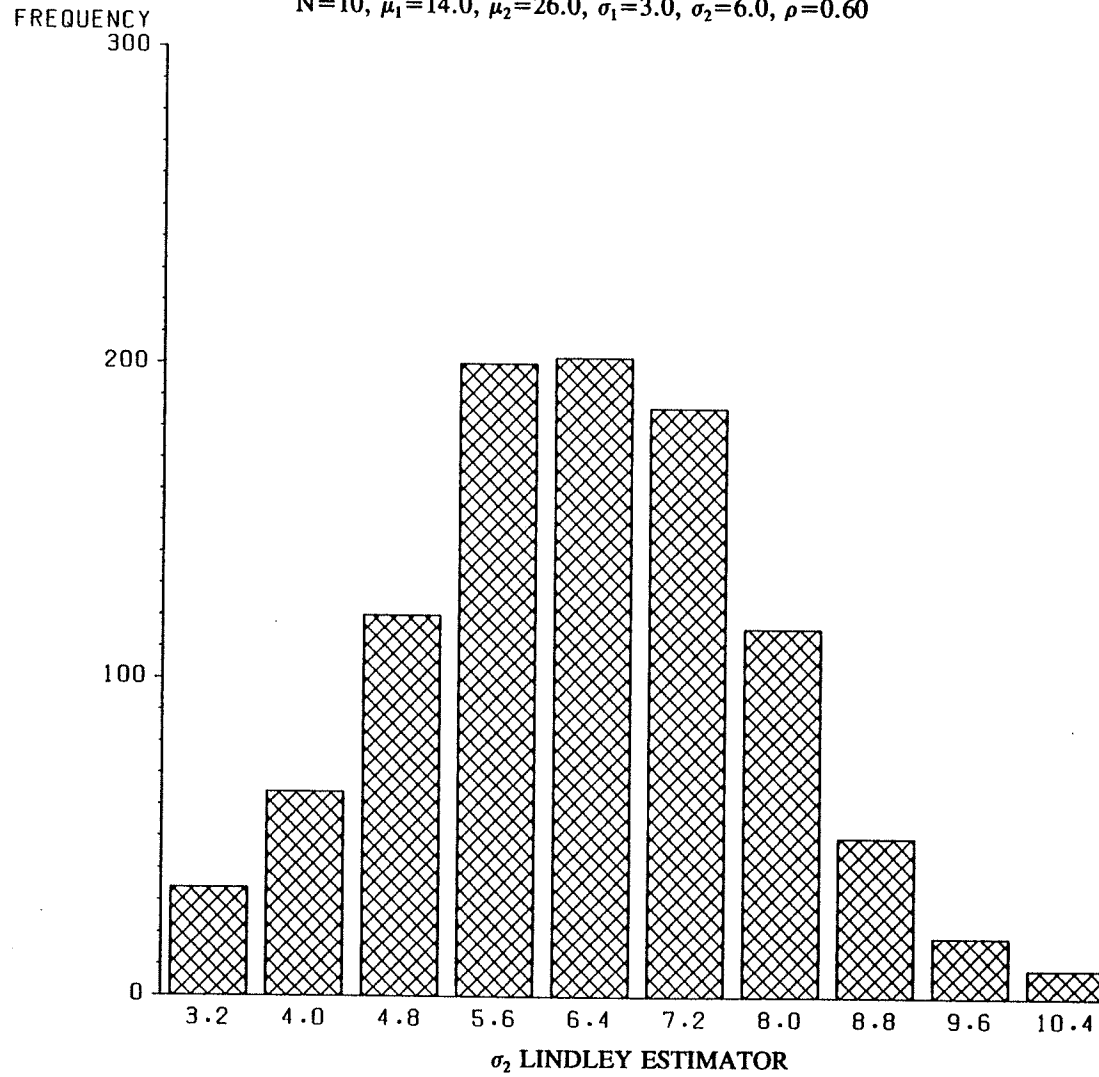

σ_2 LINDLEY ESTIMATOR

Figure 4.9 ESTIMATION OF THE BIVARIATE NORMAL ρ PARAMETER
RESULTS FOR 1000 SAMPLES
N=10, μ_1=14.0, μ_2=26.0, σ_1=3.0, σ_2=6.0, ρ=0.60

ρ MAXIMUM LIKELIHOOD ESTIMATOR

Figure 4.10

ESTIMATION OF THE BIVARIATE NORMAL ρ PARAMETER
RESULTS FOR 1000 SAMPLES
$N=10$, $\mu_1=14.0$, $\mu_2=26.0$, $\sigma_1=3.0$, $\sigma_2=6.0$, $\rho=0.60$
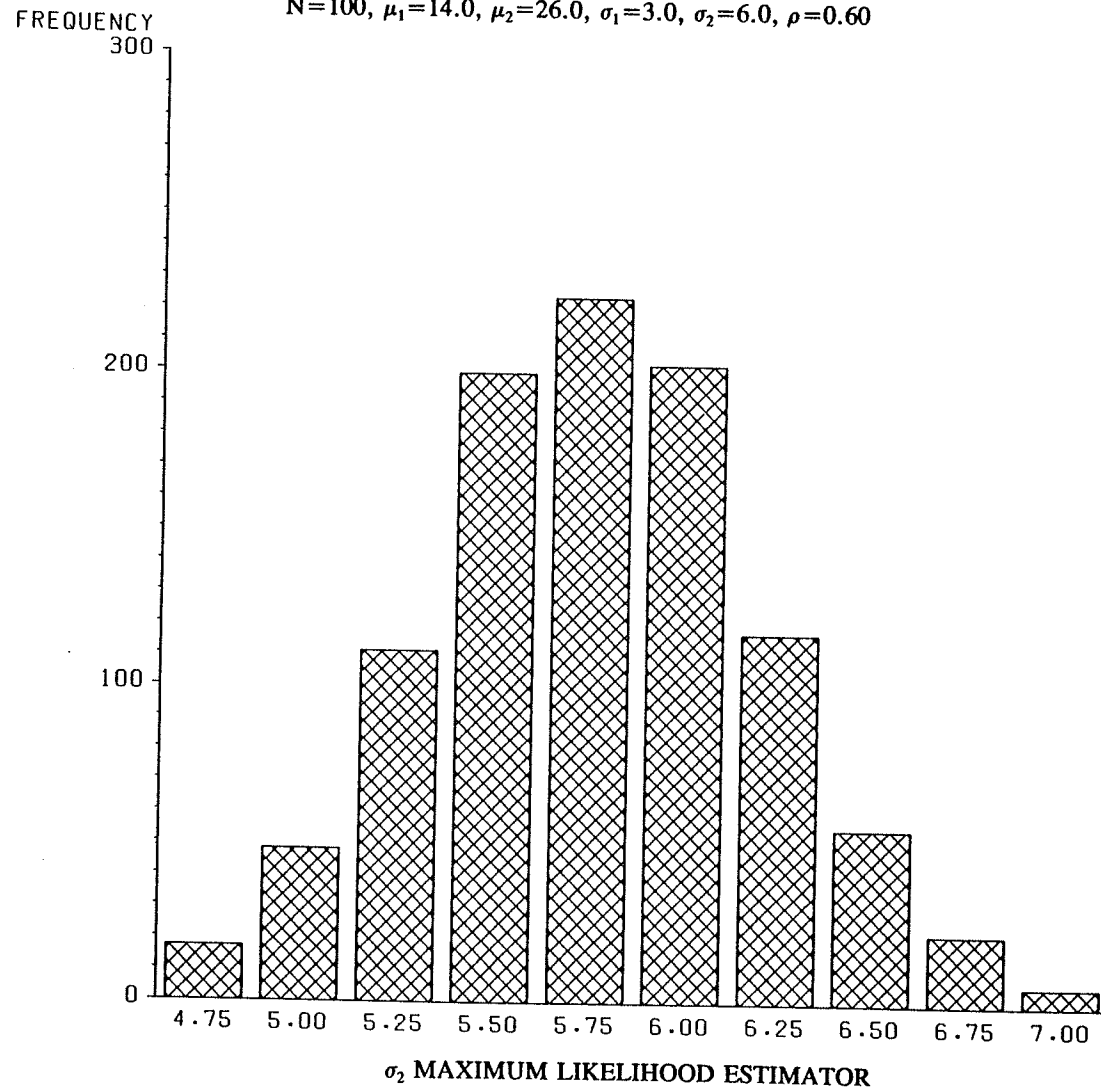

ρ LINDLEY ESTIMATOR

Figure 4.11

ESTIMATION OF THE BIVARIATE NORMAL ρ PARAMETER
RESULTS FOR 1000 SAMPLES
N=100, μ_1=14.0, μ_2=26.0, σ_1=3.0, σ_2=6.0, ρ=0.60

FREQUENCY

ρ MAXIMUM LIKELIHOOD ESTIMATOR

Figure 4.12

ESTIMATION OF THE BIVARIATE NORMAL ρ PARAMETER
RESULTS FOR 1000 SAMPLES
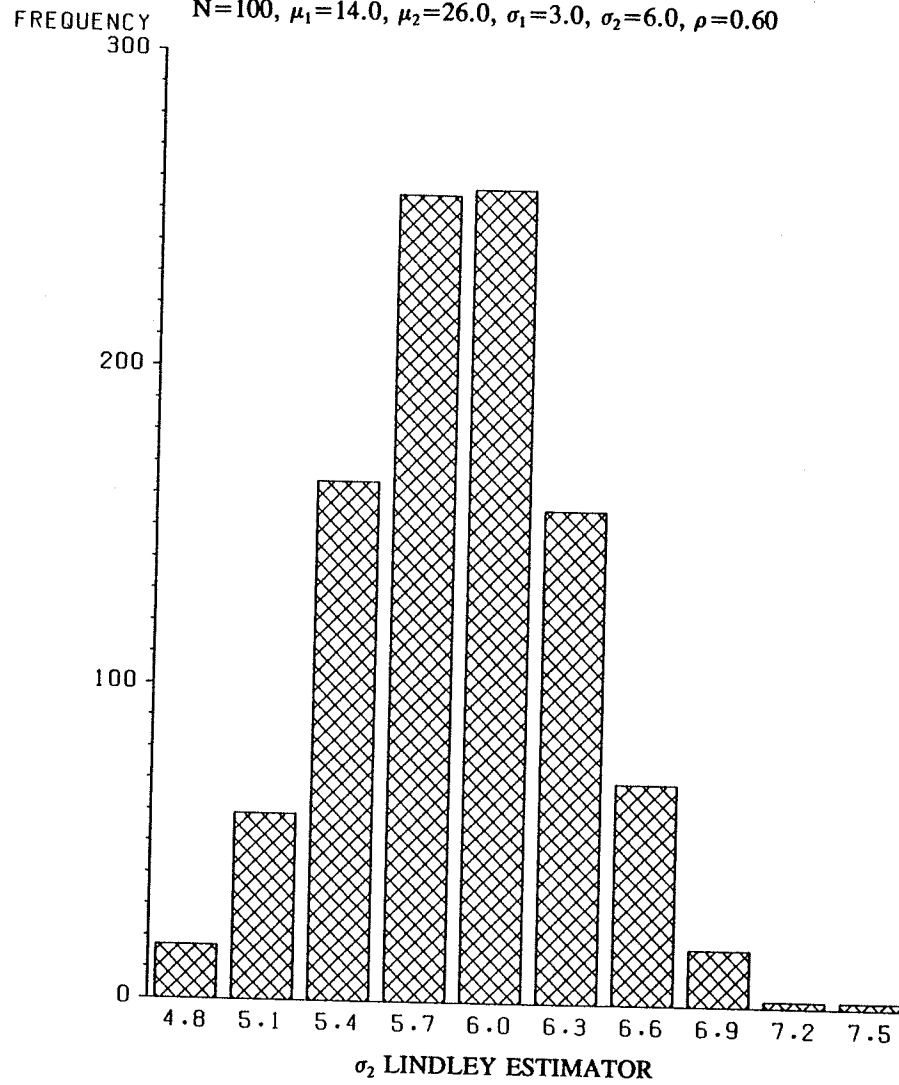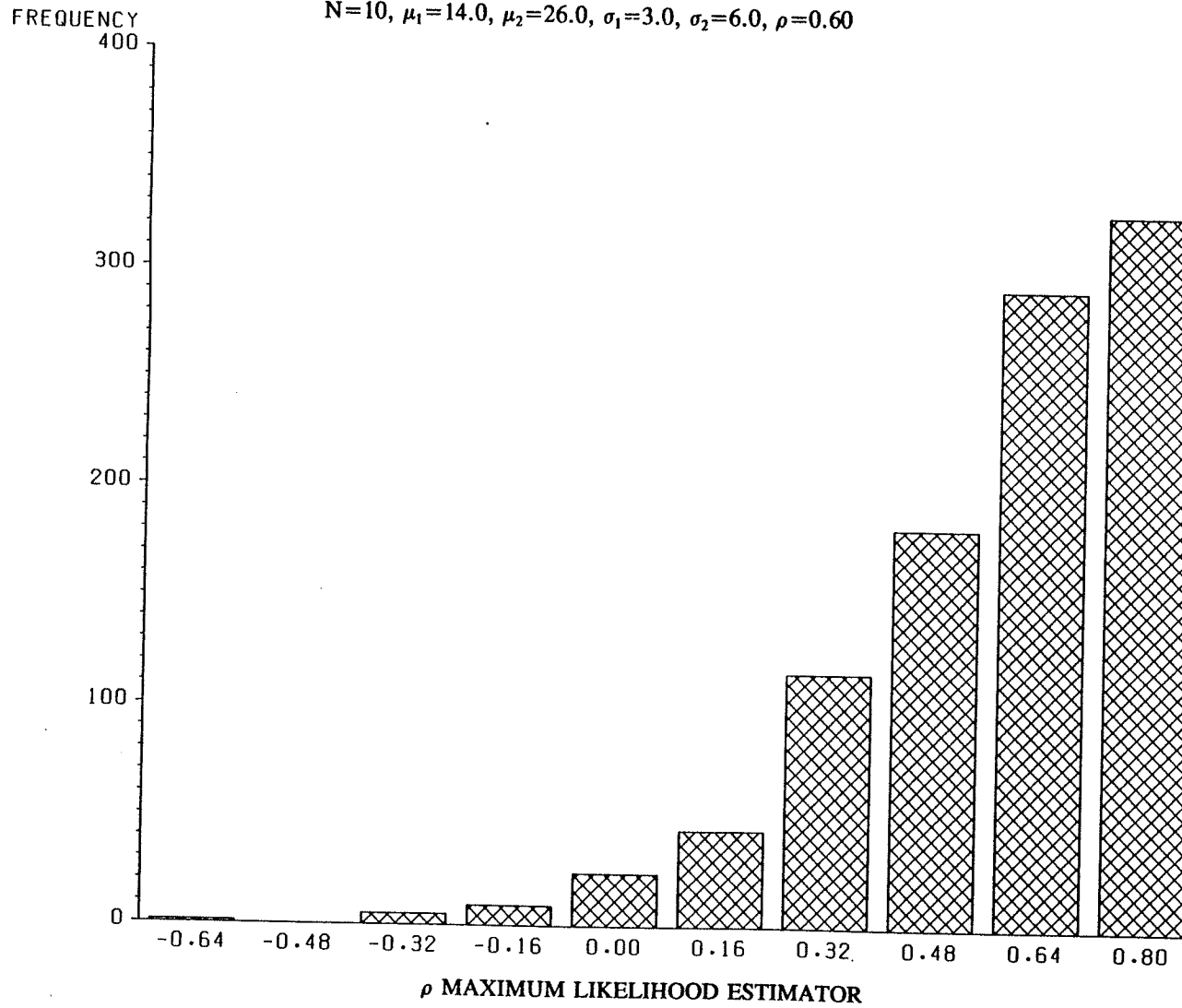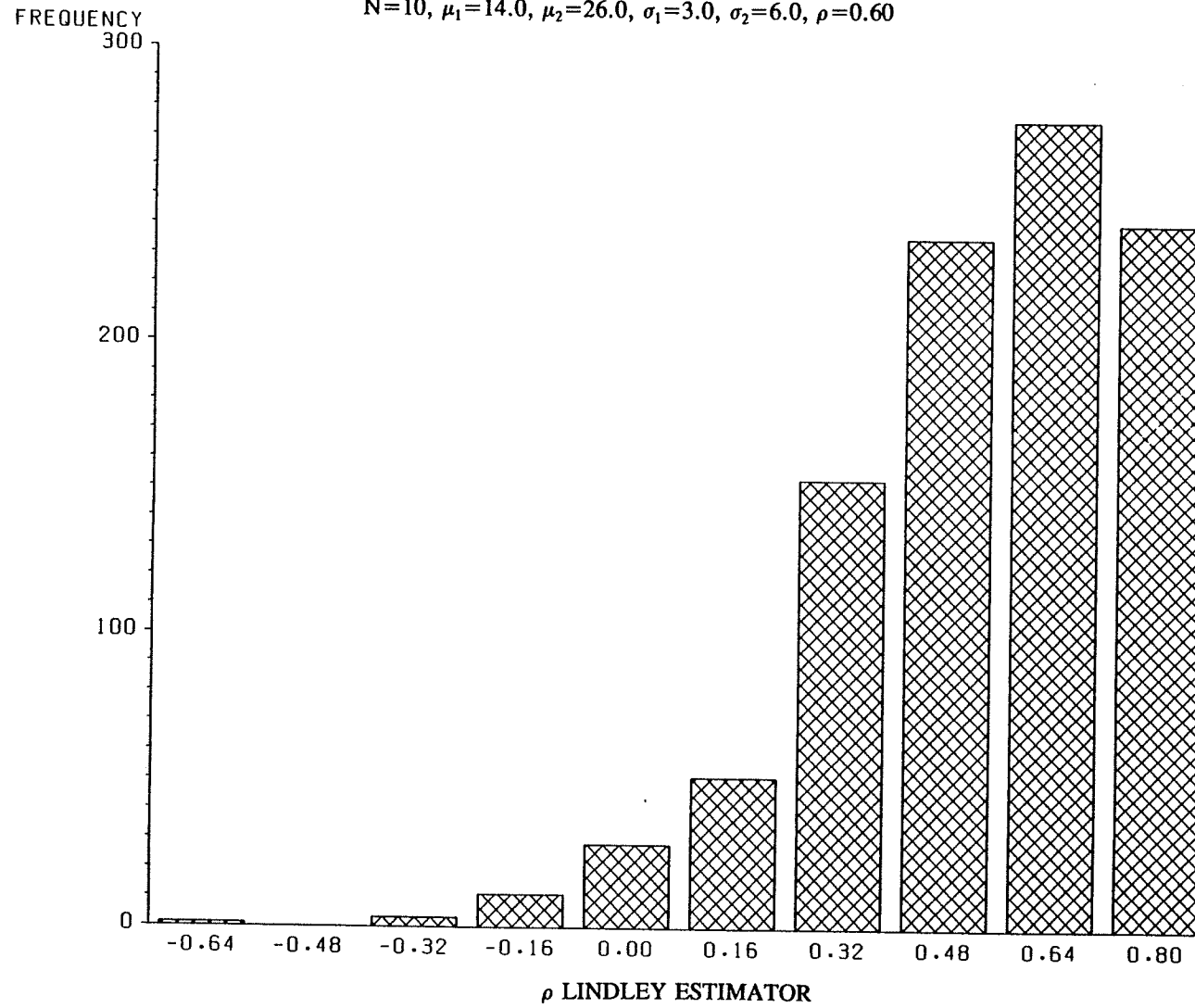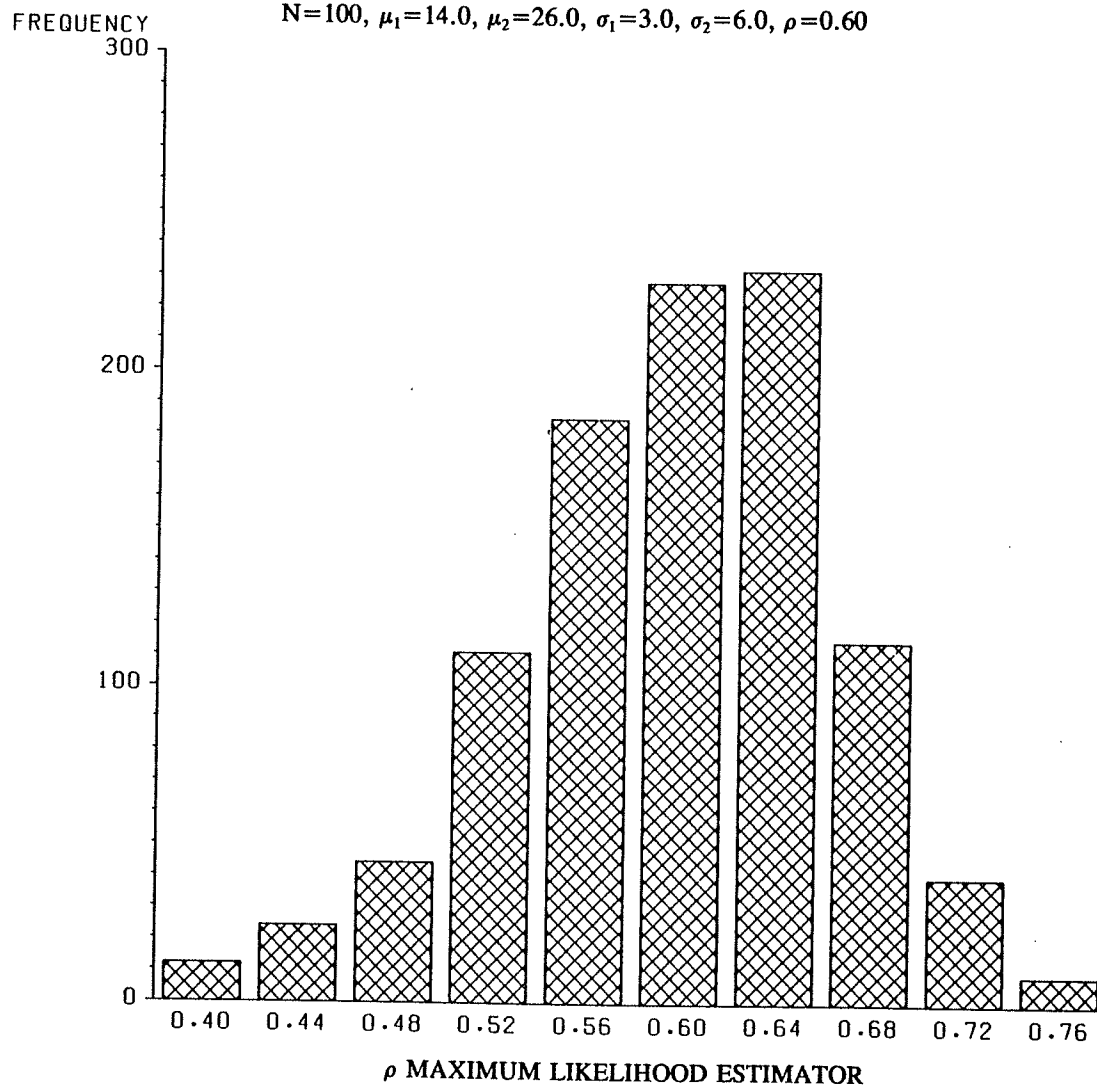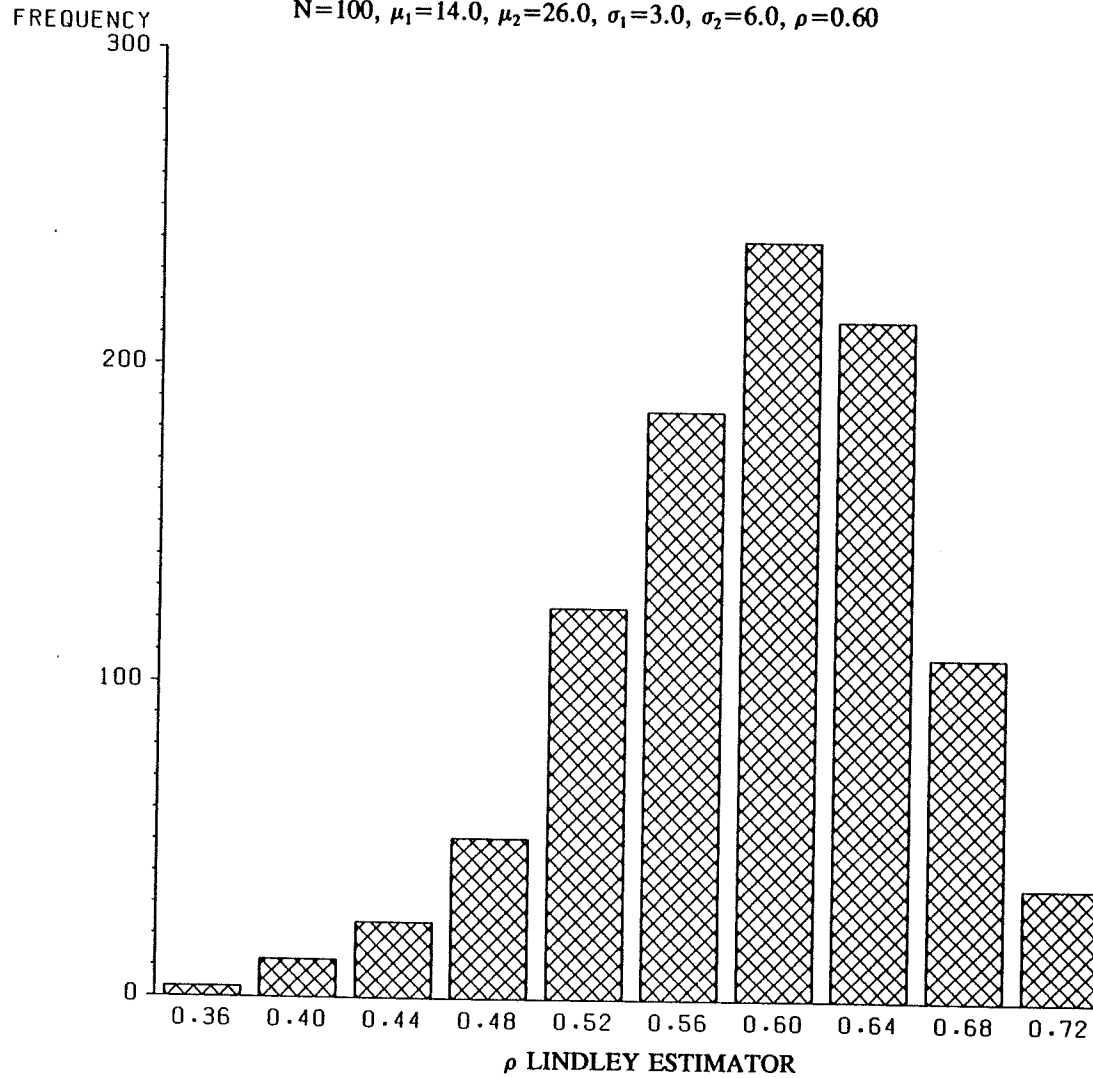$N=100$, $\mu_1=14.0$, $\mu_2=26.0$, $\sigma_1=3.0$, $\sigma_2=6.0$, $\rho=0.60$

ρ LINDLEY ESTIMATOR

n=10 to n=100 and so for the sake of brevity are not presented here.

4.7 Summary

In this chapter, the most general case of the bivariate normal distribution was studied and the two Bayes approximations of Lindley and Tierney-Kadane compared to the mle. The Bayes estimators were set in an a priori environment of noninformation and so are as hampered as they could possibly be. If accurate prior information is available, it is reasonable to expect that the Bayes approximations would perform even better.

There is no clear winner between the two schools of estimation for the bivariate normal distribution. For some of the parameters and certain sample sizes, the mle has some advantageous properties. For others, the Bayes approximations are seemingly more precise. A most encouraging finding is that for samples of n=200 and larger, the two methodologies converge so that no matter which path is chosen , the destination reached will be the same in both cases.

There is little to choose between the two Bayes approximations in terms of precision and accuracy. As was demonstrated, however, Lindley's approach opens the door to algebraically closed form solutions for the Bayes estimator whereas the Tierney-Kadane method becomes arduous and intractable. The end result would seem to favour Lindley's approach for those practitioners who do not have ready access to powerful numerical routines.

CHAPTER 5: PREDICTIVE INTERVALS FOR A MIXTURE OF EXPONENTIAL FAILURE-TIME DISTRIBUTIONS

5.1 Introduction

In this chapter, predictive intervals of a future observation for a mixture of exponential distributions with time-censored sampling are studied assuming inverted gamma priors. Effects of the prior information and sample size on the predictive interval are discussed. Distributional properties of Monte Carlo sampling distributions of the predictive intervals are examined and Pearsonian curves fitted. A portion of this work has been published in Statistics and Probability Letters by Sloan and Sinha (1991).

Consider a mixture distribution $F(t) = pF_1(t) + qF_2(t)$ where $q = 1-p$ and

$$F_i(t) = 1 - \exp\left[-\frac{t}{\theta_i}\right] \quad t \geq 0, \ \theta_i > 0, \ i = 1, 2 \qquad (5.1)$$

and p is an unknown mixture proportion for the two component distributions. Situations where this model applies arise often in life-testing problems. For example, a manufacturer may produce a brand name line as well as a generic product, each with a different life expectancy.

Sinha (1983) provides an extensive list of such applications. Cheng, Fu and Sinha (1985) use an empirical procedure to estimate the systemic parameters and reliability function. Titterington, Smith and Makov (1988) include a considerable amount of work on the mixture of exponentials model from a theoretical perspective in their monograph on mixture distributions. They deal not only with various estimation approaches for finite mixture models, but also give algorithms for identifying the number of component distributions present.

The model dealt with in this chapter is slightly more complicated than simply

drawing observations at random from the mixture density 5.1. The mathematical complications introduced are a result of the timing and type of information that is available in an applied sampling environment. The model which is the focus of this chapter first appeared in Mendenhall and Hader (1958) involving observed failure times of transmitter-receivers until a specified time T had elapsed. When brought in for maintenance, the items were further checked and classified into two groups - confirmed and unconfirmed failures. Thus, only once a failure had occurred could an item be attributed to the appropriate subpopulation. Time censoring occurred due to policy considerations which set an upper limit on the time until replacement became mandatory.

We assume that n units from the Mendenhall and Hader (1958) mixture model (5.1) are subjected to some life-testing experiment and let $\underline{x} = (x_1, x_2, \ldots, x_n)$ be the failure times of these units. We further assume that n components of the same kind are to be put into future use and let $\underline{y} = (y_1, y_2, \ldots, y_n)$ be the future failure times of these components. The predictive distribution and the corresponding prediction interval of the failure time y are studied on the basis of previously observed life test data $\underline{x}$ whose parent population is the mixture density (5.1) using a Bayesian approach.

5.2 Algebraic Preliminaries

If the experiment involving the n components subjected to the life-testing experiment were allowed to continue until all items failed, the mixture in the sample of the two subpopulations (i.e., n_1 and n_2) would be known. In most situations, however, this approach is impractical as the experiment could continue for a very long time. Typically, censoring is introduced after some predetermined time T. At time T, only r

of the n items will have failed, leaving n-r items still working. The r items that fail by time T can be attributed to their respective subpopulations. Hence it is known that the r failed items can be separated into r_1 and r_2 items from the respective subpopulations. When the remaining n-r items will fail and the mixture of these items is unknown. As such there is no way of knowing n_1 and n_2, beyond the relation $0 \leq n_1, n_2 \leq n-r$.

The likelihood function must then take into account all possible combinations of the subpopulation mixture for the remaining n-r items. Let t_{ij} denote the failure time for the jth unit belonging to the ith subpopulation, $j = 1,2,...,r_i$; $i = 1,2$ and $t = \{t_{ij}|j = 1,2,...,r_i; i = 1,2\}$ so that $r_1 + r_2 = r$ censored failure times are observed.

Mendenhall and Hader (1958) show that in such a sampling situation, the likelihood is given by

$$L(p, \theta_1, \theta_2 | t_{11}, ..., t_{1r_1}; t_{21}, ..., t_{2r_2}, T) \propto$$

$$\frac{1}{\theta_1^{r_1}\theta_2^{r_2}}\sum_{k=0}^{n-r}\binom{n-r}{k}p^{n-r_2-k}q^{r_2+k}\exp\left(-\frac{1}{\theta_1}[r_1\bar{t}_1+T(n-r-k)]-\frac{1}{\theta_2}[r_2\bar{t}_2+Tk]\right)$$

where

$$\bar{t}_i = \sum_{j=1}^{r_i}\frac{t_{ij}}{r_i} \qquad i=1,2 \qquad\qquad (5.2)$$

The complex likelihood function (5.2) is due to the fact that the sampling environment involved in this chapter (and chapter six) is not simply drawing a random sample from a parent population that is a mixture distribution. The environment is, in fact, complicated by censoring and additional information regarding attributability to the appropriate subpopulation. The following detailed derivation of the likelihood function

involving a mixture of exponentials parent population applies without loss of generality to the Weibulls situation in Chapter six. This discussion follows the logic of Mendenhall and Hader (1958) and Chapter four of Sinha (1986).

Assume n items subjected to some lifetesting experiment which is terminated after some predetermined time T. Consider the two subpopulations sp_1 and sp_2 mixed in unknown proportion p and define q=1-p. Further assume that $f_1(x|\theta_1)$ and $f_2(x|\theta_2)$ are exponential density functions such that

$$f_i(x|\theta_i) = \frac{1}{\theta_i}\exp\left[-\frac{x}{\theta_i}\right], \quad \theta_i > 0, \quad i=1,2. \tag{5.3}$$

The population cumulative distribution function and density function are defined as

$$F(t) = pF_1(t) + qF_2(t) \quad and \quad f(t) = pf_1(t) + qf_2(t) \tag{5.4}$$

respectively.

Suppose that after the experimental termination time T, r units are observed to have failed, leaving n-r units still working at time T. It is assumed that once an item has failed it can be attributed to the appropriate subpopulation sp_1 or sp_2. Thus, further information is gained from this sampling environemnt in that it is known that the r units which failed before time T are comprised of r_1 units from subpopulation sp_1 and r_2 units from subpopulation sp_2 such that $r=r_1+r_2$.

Let t_{ij} denote the failure time for the jth unit belonging to the ith subpopulation sp_i such that $i=1,2$; $j=1,2,...,r_i$. The observed sample is hence summarized as

$$\underline{t} = \{t_{ij} | i=1,2; \; j=1,2,...,r_i\} = \{t_{11}, t_{12}, ..., t_{1r_1}; \; t_{21}, t_{22}, ..., t_{2r_2}\}$$

As per Mendenhall and Hader (1958), the additional information about the attributability of each failed item to the appropriate subpopulation produces a multinomial sampling environment. Specifically, given a sample of n units, the probability of r_1 units from subpopulation sp_1 failing, r_2 units from subpopulation sp_2 failing where the failure times are less than or equal to T, and n-r units surviving is derived from the multinomial distribution to be

$$Pr(r_1, r_2, n-r | n) = \frac{n!}{r_1! r_2! (n-r)!} [pF_1(T)]^{r_1} [qF_2(T)]^{r_2} [1-F(T)]^{n-r} .$$

Furthermore, the conditional density of obtaining r_i ordered failure times from a particular subpopulation sp_i , given r_i and the fact that all the units failed before the censoring time T is

$$Pr(t_{i1}, t_{i2}, ..., t_{ir_i} | r_i; \; t_{ij} \le T) = \frac{r_i! \prod_{j=1}^{r_i} f_i(t_{ij})}{[F_i(T)]^{r_i}} \quad , \; i=1,2 .$$

Combining this information produces the Mendenhall and Hader (1958) likelihood for the sample based on the mixture population and the information from this specialized sampling environment as

$$L(p, \theta_1, \theta_2 | \underline{t}, T) = \frac{n!}{(n-r)!} p^{r_1} q^{r_2} \prod_{j=1}^{r_1} f_1(t_{1j}) \prod_{j=1}^{r_2} f_2(t_{2j}) [1-F(T)]^{n-r} .$$

More specifically, this can be written as

$$L(p,\theta_1,\theta_2|\underline{t},T) \propto p^{r_1}q^{r_2}\prod_{j=1}^{r_1}\left[\frac{1}{\theta_1}\exp\left(-\frac{t_{1j}}{\theta_1}\right)\right]\prod_{j=1}^{r_2}\left[\frac{1}{\theta_2}\exp\left(-\frac{t_{2j}}{\theta_2}\right)\right]$$

$$\cdot\left[\int_T^\infty\left[\frac{p}{\theta_1}\exp\left(-\frac{t}{\theta_1}\right)+\frac{q}{\theta_2}\exp\left(-\frac{t}{\theta_2}\right)\right]dt\right]^{n-r}$$

which simplifies to (Sinha, 1986, page 104)

$$L(p,\theta_1,\theta_2|\underline{t},T) \propto \frac{p^{r_1}q^{r_2}}{\theta_1^{r_1}\theta_2^{r_2}}\exp\left[-\frac{\sum_{j=1}^{r_1}t_{1j}}{\theta_1}\right]\exp\left[-\frac{\sum_{j=1}^{r_2}t_{2j}}{\theta_2}\right]$$

$$\cdot\left[p\exp\left(-\frac{T}{\theta_1}\right)+q\exp\left(-\frac{T}{\theta_2}\right)\right]^{n-r} .$$

Expanding and defining $\overline{t}_i = \dfrac{\sum_{j=1}^{r_i}t_{ij}}{r_i}$, for $i=1,2$ produces

$$L(p,\theta_1,\theta_2|\underline{t},T) \propto \frac{p^{r_1}q^{r_2}}{\theta_1^{r_1}\theta_2^{r_2}}\exp\left[-\frac{r_1\overline{t}_1}{\theta_1}-\frac{r_2\overline{t}_2}{\theta_2}\right]$$

$$\cdot\sum_{k=0}^{n-r}\left[\binom{n-r}{k}q^k\exp\left[-\frac{kT}{\theta_2}\right]p^{n-r-k}\exp\left[-\frac{T(n-r-k)}{\theta_1}\right]\right] .$$

Finally, collecting like terms together under the summation sign results in

$$L(p,\theta_1,\theta_2|\underline{t},T) \propto$$

$$\frac{1}{\theta_1^{r_1}\theta_2^{r_2}}\sum_{k=0}^{n-r}\binom{n-r}{k}p^{n-r_2-k}q^{r_2+k}\exp\left[-\frac{1}{\theta_1}\left[r_1\overline{t}_1+T(n-r-k)\right]-\frac{1}{\theta_2}\left[r_2\overline{t}_2+Tk\right]\right]$$

which is the likelihood function (5.2).

The choice of a prior distribution must now be addressed. Assume a uniform prior distribution for the unknown mixture proportion p. Raiffa and Schlaifer (1961) suggest an inverted gamma distribution as an appropriate prior for the exponential parameters as follows

$$g_i(\theta_i) \propto \frac{1}{\theta_i^{b_i+1}} \exp\left(-\frac{a_i}{\theta_i}\right) \quad \theta_1 > 0; \ a_i, b_i > 0 \quad for \ i=1,2 \tag{5.5}$$

Titterington, Smith and Makov (1988) support the use of such a conjugate prior (5.5) without providing a specific recommendation for the mixture of exponentials problem.

Further, assuming prior independence of the three parameters, the joint prior distribution for $\underline{\theta} = (p, \theta_1, \theta_2)$ is

$$g(p, \theta_1, \theta_2) \propto \frac{\exp\left[-\left(\frac{a_1}{\theta_1} + \frac{a_2}{\theta_2}\right)\right] \cdot}{\theta_1^{b_1+1} \theta_2^{b_2+1}} \tag{5.6}$$

The joint posterior distribution is given by

$$\Pi(p, \theta_1, \theta_2 | \underline{t}) \propto$$

$$\sum_{k=0}^{n-r} \binom{n-r}{k} p^{n-r_2-k} q^{r_2+k} \frac{\exp\left[-\frac{1}{\theta_1}[r_1\bar{t}_1 + T(n-r-k) + a_1] - \frac{1}{\theta_2}[r_2\bar{t}_2 + Tk + a_2]\right]}{\theta_1^{r_1+b_1+1} \theta_2^{r_2+b_2+1}}$$

This posterior distribution will be asymptotically normal (Titterington, Smith and Makov, 1988) although only for considerably large samples. The Bayes predictive density of a future observation y is defined as the posterior expectation of the density function of y

and is given by

$$h(y|\underline{t}) \propto \iiint f(y|p,\theta_1,\theta_2)\Pi(p,\theta_1,\theta_2|\underline{t})\,dp\,d\theta_1\,d\theta_2 \qquad (5.7)$$

which, on using (5.4) and after some algebra reduces to

$$h(y|t) = C\sum_{k=0}^{\infty}\binom{n-r}{k}\left[\frac{(r_1+b_1)B(n-r_2-k+2,r_2+k+1)}{[r_1\overline{t}_1+(n-r-k)T+a_1+y]^{r_1+b_1+1}(r_2\overline{t}_2+kT+a_2)^{r_2+b_2}}\right.$$

$$\left. + \frac{(r_2+b_2)B(n-r_2-k+1,r_2+k+2)}{[r_1\overline{t}_1+(n-r-k)T+a_1]^{r_1+b_1}(r_2\overline{t}_2+kT+a_2+y)^{r_2+b_2+1}}\right] \qquad (5.8)$$

where

$$C^{-1} = \sum_{k=0}^{n-r}\binom{n-r}{k}\frac{B(n-r_2-k+1,r_2+k+1)}{[r_1\overline{t}_1+(n-r-k)T+a_1]^{r_1+b_1}(r_2\overline{t}_2+kT+a_2)^{r_2+b_2}}$$

The $100(1-\alpha)\%$ predictive limits (L,U) are solutions to the equations

$$\int_0^L h(y|\underline{t})\,dy = \int_U^{\infty} h(y|\underline{t})\,dy = \frac{\alpha}{2} \qquad (5.9)$$

Using (5.8) and (5.9), we obtain equations for the predictive limits L and U as

$$\frac{\alpha}{2} = C \sum_{k=0}^{n-r} \binom{n-r}{k}$$

$$\left[\frac{B(n-r_2-k+2,r_2+k+1)}{[r_2\bar{t}_2+kT+a_2]^{r_2+b_2}} \left(\frac{1}{[r_1\bar{t}_1+(n-r-k)T+a_1]^{r_1+b_1}} - \frac{1}{[r_1\bar{t}_1+(n-r-k)T+a_1+L]^{r_1+b_1}} \right) \right.$$

$$\left. + \frac{B(n-r_2-k+1,r_2+k+2)}{[r_1\bar{t}_1+(n-r-k)T]^{r_1+b_1}} \left(\frac{1}{[r_2\bar{t}_2+kT+a_2]^{r_2+b_2}} - \frac{1}{[r_2\bar{t}_2+kT+a_2'+L]^{r_2+b_2}} \right) \right]$$

and   (5.10)

$$\frac{\alpha}{2} = C \sum_{k=0}^{n-r} \binom{n-r}{k} \left[\frac{B(n-r_2-k+2,r_2+k+1)}{[r_2\bar{t}_2+kT+a_2]^{r_2+b_2}[r_1\bar{t}_1+(n-r-k)T+a_1+U]^{r_1+b_1}} \right.$$

$$\left. + \frac{B(n-r_2-k+1,r_2+k+2)}{[r_1\bar{t}_1+(n-r-k)T+a_1]^{r_1+b_1}[r_2\bar{t}_2+kT+a_2+U]^{r_2+b_2}} \right]$$

which can be solved by an iterative linear search technique.

5.3 Sample Generation

Sinha (1983) obtained Bayes estimators of the parameters and reliability function for the mixtures of exponentials model and used Mendenhall and Hader's (1958) aircraft components data as an example. The dataset had $n=369$ observations censored at time $T=630$ hours of operation, producing $r_1=218$ confirmed failures and $r_2=107$ unconfirmed failures. Inspection of the sample histograms raised some question as to the fit for the assumed model to this data.

To test the validity of the mixture of exponentials model producing such a sample, a goodness of fit testing was initiated. The power of the usual chi-square test is known to be suspect for small samples, so empirical distribution function or EDF tests as described by D'Agostino and Stephens (1986) were used. The Anderson-Darling (1954) statistic and the Cramer-von Mises W^2 statistic have been recommended for testing exponentiality from a power standpoint. Further, if we can totally specify the pdf of our mixture population (5.1), we can perform an omnibus test on the combined sample to investigate the assumption of our hypothesized model. Thus, we can directly test whether or not it is likely that, as a single entity, the combined samples did arise from a population comprised of a mixture of two such exponential subpopulations. Note that the omnibus test will only be possible for computer generated samples due to having specified the model parameters for sample generation. These tests were conducted via algorithmic implementations in BASIC on an IBM PC microcomputer. These programs, collectively named EDFIT, are obtainable from the authors on request.

Using the Mendenhall and Hader (1958) data, the test for the unconfirmed failures sample found that the exponential model was supportable, producing a p-value of .143. The test on the unconfirmed failures data resulted in the rejection of the exponential null hypothesis (p-value of 0.0004). Because we cannot completely specify the mixture pdf parameters for this sample, we cannot perform the omnibus test precisely. Using the parameter estimates given by Sinha (1983), we can, however, carry out the process. This results in further evidence that the mixture of exponentials model is somewhat questionable for the Mendenhall and Hader dataset, producing a p-value of 0.006.

To investigate this anomaly a bit further, Figure 5.1 depicts the individual

subpopulation histograms for the Mendenhall and Hader dataset. The confirmed failures' histogram does evidence marked departure from an exponential distribution and is no doubt the source of the significant result indicating a lack of model fit. The unconfirmed failures' histogram seems much more likely to have arisen from an exponential process.

This potentially confounding effect makes it necessary to generate samples from a mixture distribution to ensure our results are not compromised by the lack of model fit. We will produce results for the Mendenhall and Hader dataset for completeness.
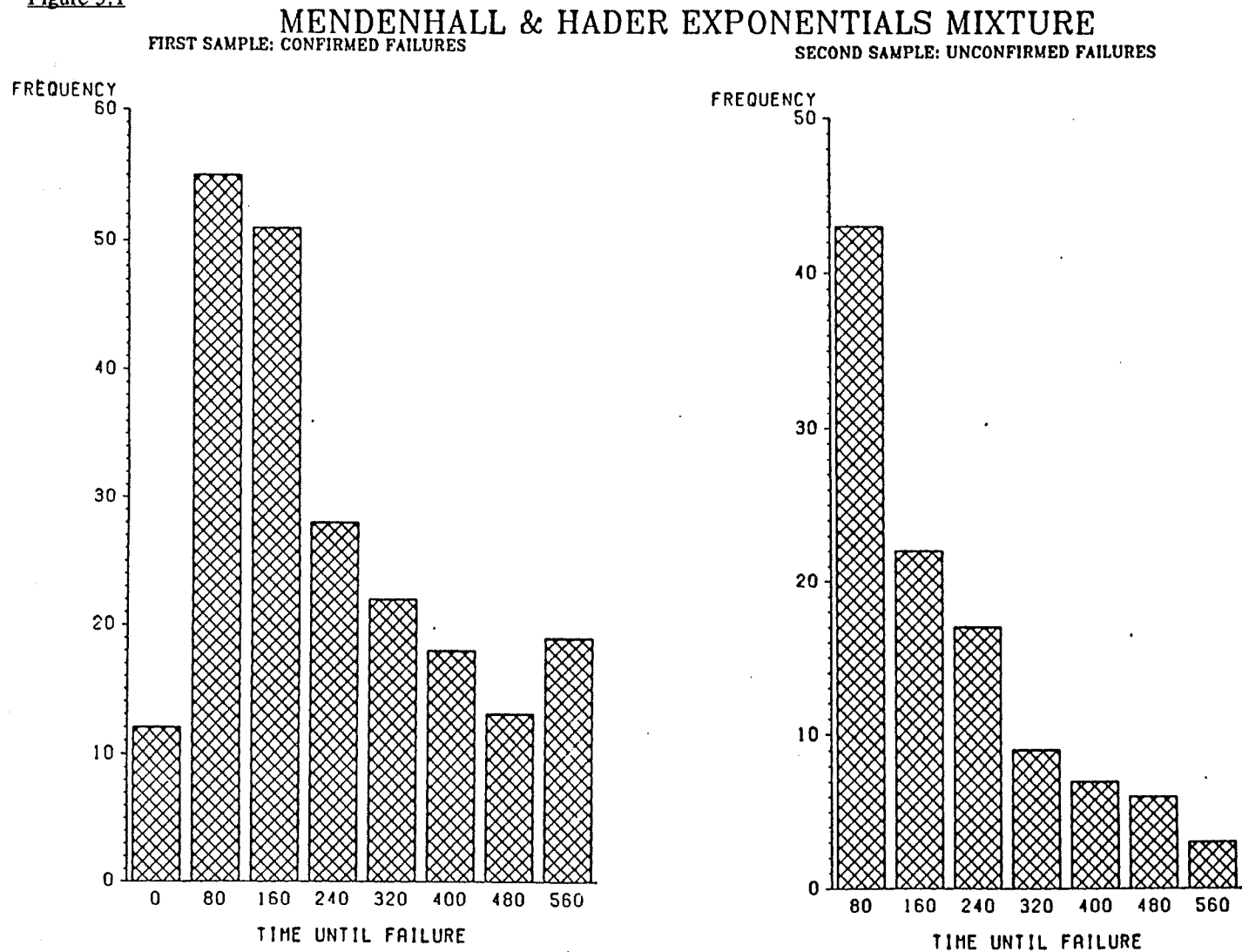
The generation of samples from a mixture of distributions was outlined by Marsaglia (1961). It proceeds as follows:

1. Generate u from $U(0,1)$.

2. If $u < p$ then generate a deviate from the first exponential

   population by $t_{ij} = \theta_1(\log p - \log u)$;

   otherwise, generate a deviate from the second exponential

   population by $t_{ij} = \theta_2(\log(1-p) - \log(u-p))$.

The use of the inverse cumulative density function approach does provide a mechanism for generating random variables from a distribution with pdf (5.1). For further details see Kennedy and Gentle (1980), pages 72-75.

The results of this simulation approach may be more representative than those seen in Sinha (1983) because the subpopulation sample sizes are not fixed. It is not realistic to assume that the experimenter would know, in advance, the makeup of the sample in terms of subpopulation representation. If this were the case, there would be no need to model the situation as a mixture distribution and instead it would be more efficiently dispatched by individual estimation on each subpopulation.

Figure 5.1

MENDENHALL & HADER EXPONENTIALS MIXTURE

FIRST SAMPLE: CONFIRMED FAILURES   **SECOND SAMPLE: UNCONFIRMED FAILURES**

217

Using the above algorithm, test samples of size n = 100 and n = 25 were generated with θ_1 = 50, θ_2 = 40, p = 0.75. The pdf of the population is given in Figure 5.2, having the appearance of an exponential distribution with an elongated tail. The goodness of fit tests described above were used on both the subsamples and combined samples, producing p-values in excess of 0.6 in all cases. The uncensored samples are given below. Again it is important to stress that although the complete sample is generated, it is an artificial construct of the generation process. The sampling environment described at the beginning of this chapter is still assumed to be the process under which the sample information is obtained.

Uncensored Sample #1 (n=100)

First Subsample:

0.2087	2.2476	2.8014	3.0100	3.8054	4.7702	5.1446
6.4945	8.6456	9.6344	9.9463	10.4170	10.6014	11.4011
11.6312	13.6771	13.8988	14.9665	15.2257	15.6106	18.4381
21.5347	23.5108	24.0437	24.5634	27.3509	27.7656	31.5811
33.7771	34.4975	34.7668	35.5825	40.3650	41.4672	43.1305
43.6357	45.1505	45.5953	48.2052	48.6234	48.9956	49.0108
50.7063	51.5283	56.7423	60.7091	64.2133	64.7181	65.1220
66.2278	71.6285	79.6639	80.3535	82.7512	88.6533	90.6015
99.4277	99.5481	103.4692	104.8867	111.6388		
124.9026	131.3938	157.7569	200.8856			

Second Subsample:

1.1076	2.5084	2.5325	3.0741	4.2815	5.1013	5.9654
6.9616	7.7776	8.5718	9.8603	12.8670	13.2814	13.6507
15.0578	15.7602	16.0200	20.2314	20.7096	21.8975	23.2716
24.2991	29.9614	30.6392	33.9987	44.9577	45.8027	47.4253
67.7229	67.8506	74.2290	81.6555	93.3770	99.8087	184.8992

The computer produces n_1 = 65, n_2 = 35, $\tilde{t}_1$ = 47.314, $\tilde{t}_2$ = 33.060, information that is considered as unknown in our sampling environment. Censoring at time T=100, the

Figure 5.2

MIXTURE OF EXPONENTIALS PDF
WITH P=0.75, θ_1=50.0, θ_2=40.0
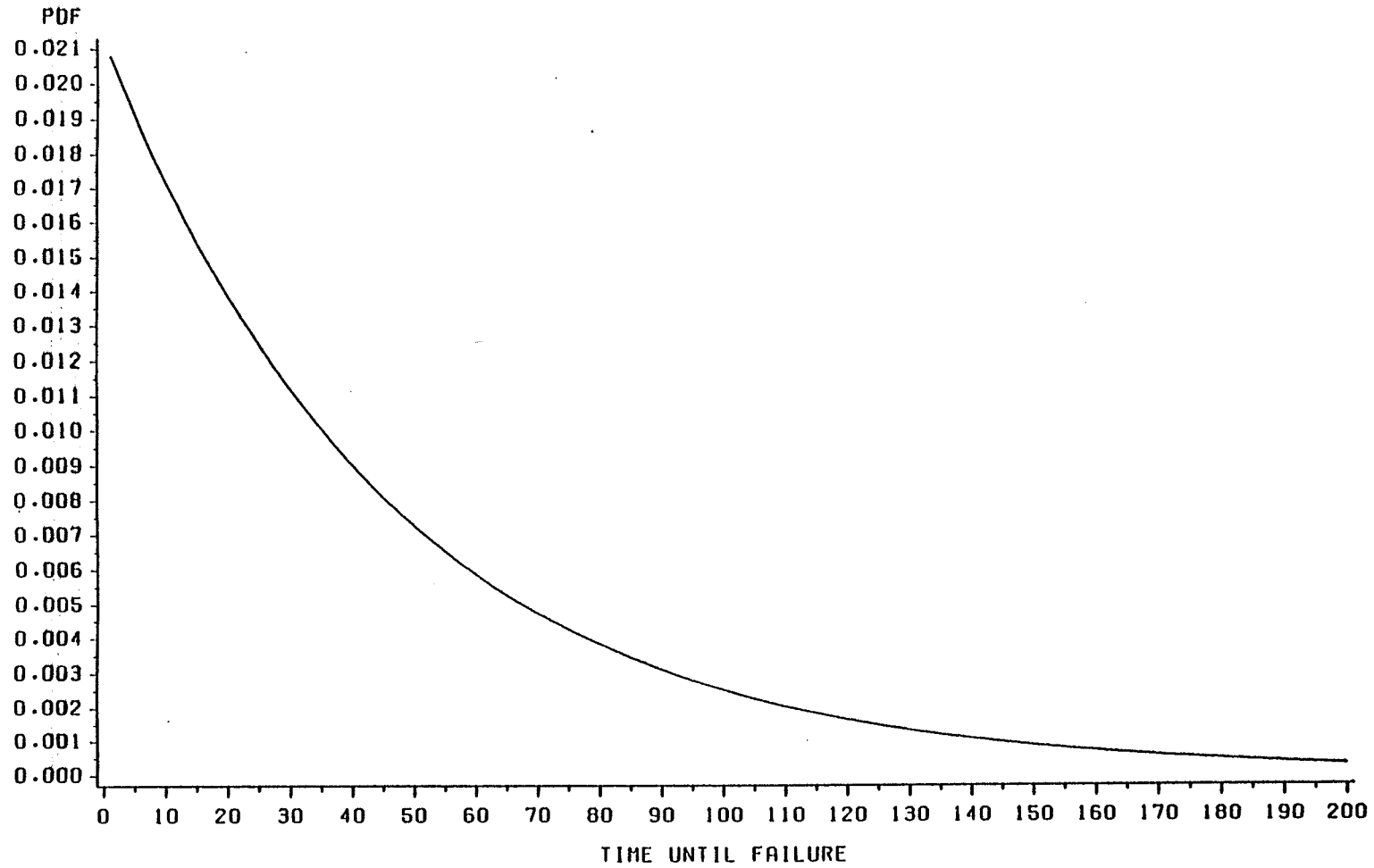
PARAMETERS USED TO GENERATE SAMPLE DATA

sample information obtained is $r_1 = 58$, $r_2 = 34$, $\bar{t}_1 = 36.905$, $\bar{t}_2 = 28.595$

<p style="text-align:center"><u>Uncensored Sample #2 (n=25)</u></p>

First Subsample:

6.1886	8.7778	9.1949	9.7326	11.5059	13.0639	13.1145
26.3979	27.2176	28.7843	30.0532	31.0214	32.5078	36.1116
38.4326	46.7884	56.2416	91.1947	101.0766	110.0240	131.0561
208.2173						

Second Subsample:

12.4904	49.0183	84.0717

The computer produces the additional (and considered to be unknown) information of $n_1 = 22$, $n_2 = 3$, $\bar{t}_1 = 48.487$, $\bar{t}_2 = 48.527$. Censoring once again at T=100, the sample information actually obtained is $r_1 = 18$, $r_2 = 3$, $\bar{t}_1 = 28.686$, $\bar{t}_2 = 48.527$.

Once appropriateness of model selection is established for each sample, censoring then takes place by adding a third algorithmic rule as follows:

3. if t_{ij} produced is in the interval [0,T] then increment r_i by 1,

otherwise we consider this observation to be censored.

5.4 Effect Of Prior Distribution

An important aspect of Bayesian inference is the degree to which the prior information impacts on the resultant posterior distribution. In this paper we are more interested, however in the related impact on the predictive distribution. We examine the effects of altering the inverted gamma prior (5.4) parameters (a_1, b_1, a_2, b_2) for our moderate n = 100 size sample and our small n = 25 size sample in Figures 5.3 and 5.4 respectively, censoring once again at T=100. Each plot has six predictive density functions, produced by varying the prior parameter from zero to 50 in steps of 10. The

Figure 5.3

MIXTURE OF EXPONENTIALS: EFFECT OF PRIOR
FUTURE OBSERVATION DENSITY(n=100)
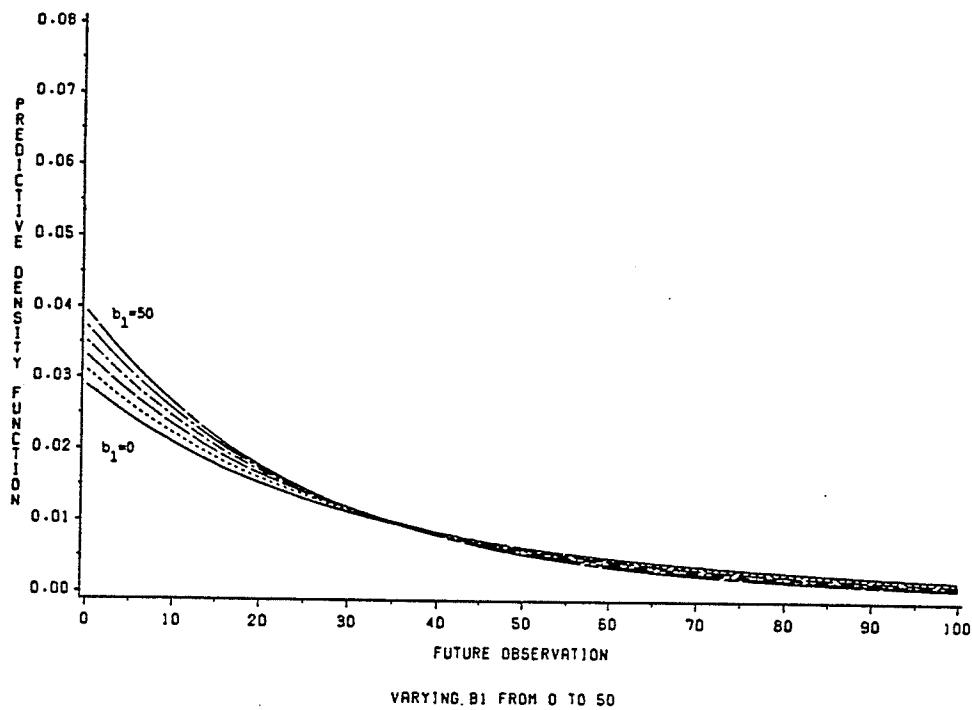
VARYING B1 FROM 0 TO 50

MIXTURE OF EXPONENTIALS: EFFECT OF PRIOR
FUTURE OBSERVATION DENSITY (n=100)

VARYING A1 FROM 0 TO 50
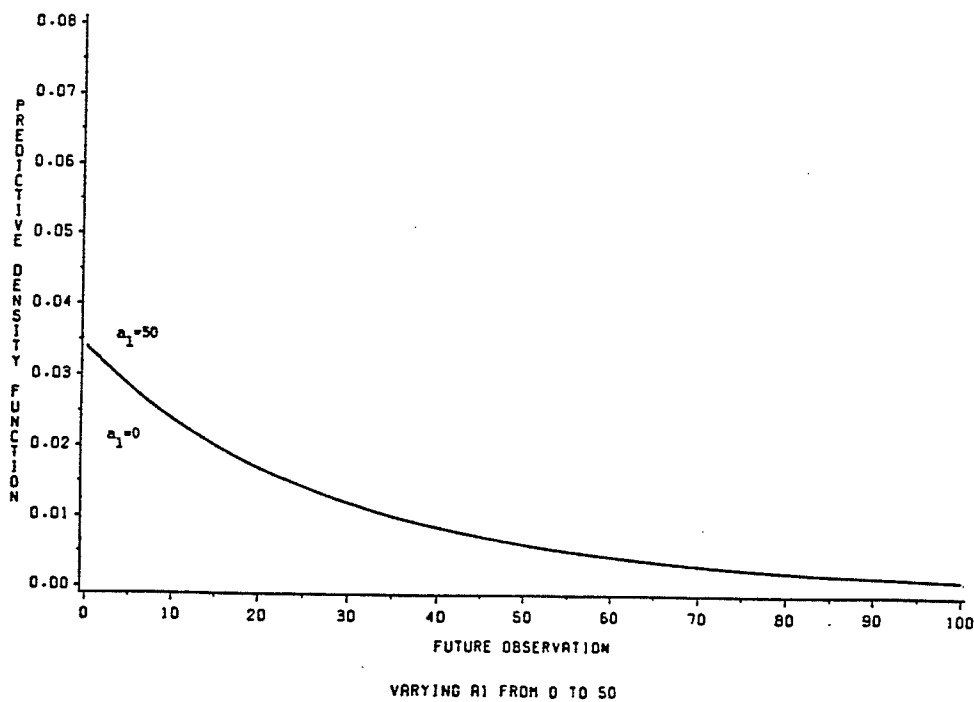
Figure 5.4 — MIXTURE OF EXPONENTIALS: EFFECT OF PRIOR
FUTURE OBSERVATION DENSITY(n=25)
VARYING B1 FROM 0 TO 50
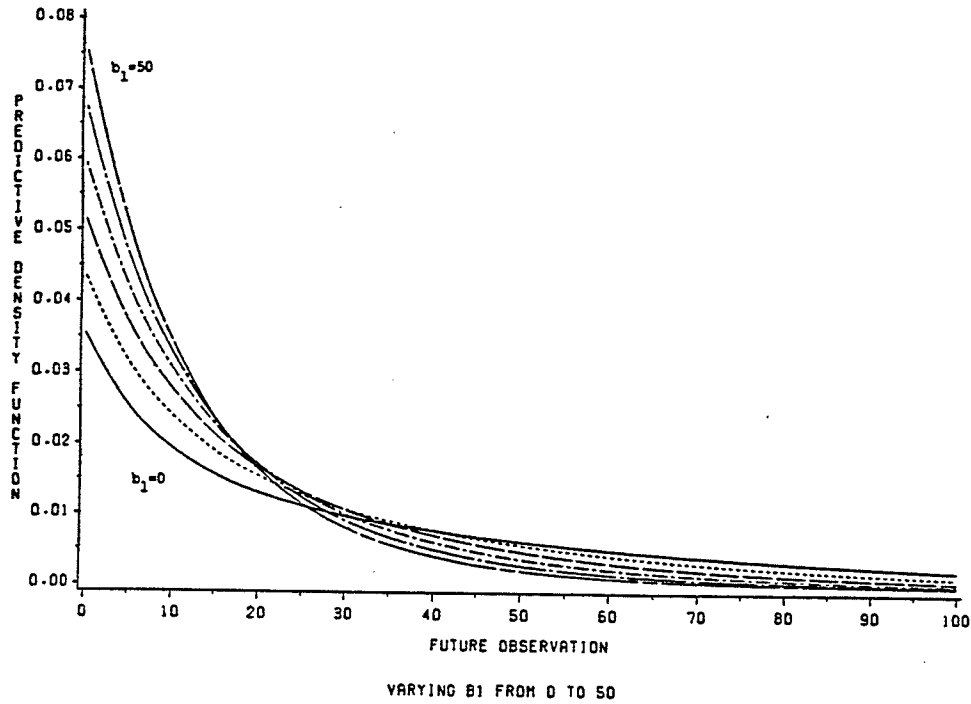
MIXTURE OF EXPONENTIALS: EFFECT OF PRIOR
FUTURE OBSERVATION DENSITY (n=25)
VARYING A1 FROM 0 TO 50
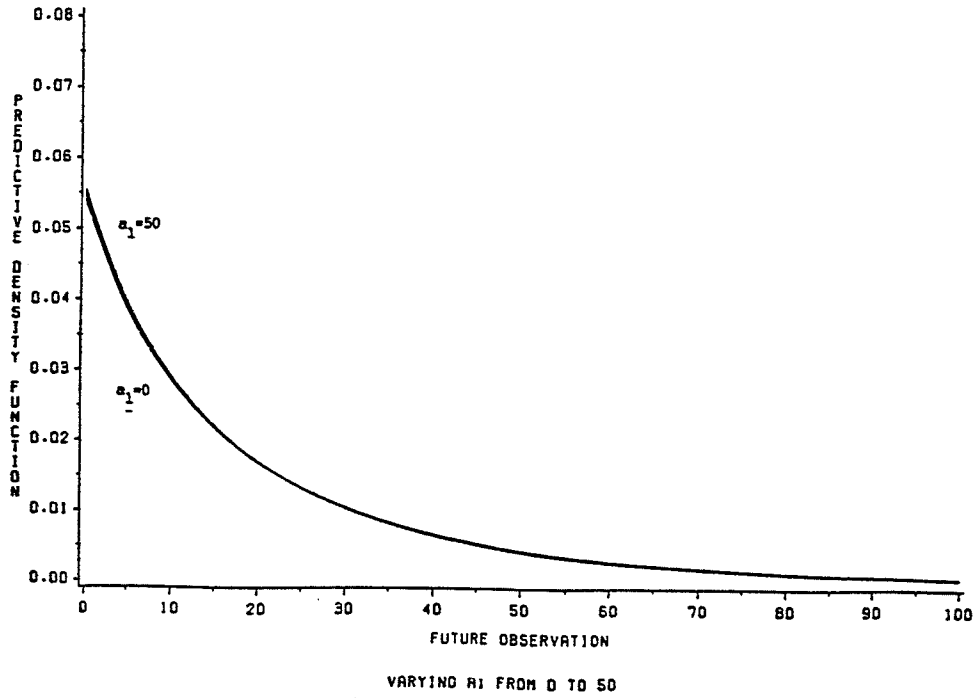
vertical axes on the four plots are uniformly scaled so that comparisons can be made. Although the impact of the prior distribution is more pronounced for the smaller sample size, the effects are the same. Altering the a_i parameters has very little effect on the predictive distribution. From the prior formula it can be seen that any change in the a_i parameters is dampened by the size of θ parameters. The impact of the b_i prior specifications is clearly demonstrated in the figures as a result of being in the exponent of the denominator. Increasing the value of the b_i parameters causes the predictive distribution to become increasingly steeper. Without doing any calculations, it is clear that the predictive interval will become shorter as the b_i's are increased. With this in mind, it is important to have meaningful prior knowledge of the parameters, especially the b_i's. If no information is available, it is perhaps advisable to set the prior parameters all to zero, in which case the usual uniform prior is the result (Jeffreys, 1983). This could be said to be a state of ignorance, or a noninformative prior.

A modified regula falsi approach (Rice (1983), page 222) was used to solve the system of equations (5.8) and obtain 95% predictive intervals for a future observation, under the various prior parameter settings for the two given samples. Typically, ten iterations were necessary for algorithmic convergence. Results are presented in Tables 5.1 and 5.2.

Table 5.1: 95% Predictive Intervals And Width δ For Various Prior Parameters

(N=100, T=100)

(L,U) δ	(10,10)	(20,20)	(b_1,b_2) (30,30)	(40,40)	(50,50)
(a_1, a_2) =(10,10)	(.86, 140.59) 139.73	(.74, 122.64) 121.90	(.64, 108.97) 108.33	(.57, 98.13) 97.56	(.51, 89.29) 88.78
(20,20)	(.87, 141.14) 140.27	(.74, 123.10) 122.36	(.64, 109.36) 108.72	(.57, 98.47) 97.90	(.51, 89.60) 89.09
(30,30)	(.87, 141.68) 140.81	(.74, 123.56) 122.82	(.65, 109.76) 109.11	(.57, 98.82) 98.25	(.52, 89.92) 89.40
(40,40)	(.88, 142.23) 141.35	(.75, 124.02) 123.27	(.65, 110.15) 109.50	(.58, 99.17) 98.59	(.52, 90.23) 89.71
(50,50)	(.88, 142.78) 141.90	(.75, 124.48) 123.73	(.66, 110.55) 109.89	(.58, 99.52) 98.94	(.52, 90.55) 90.03

Table 5.2: 95% Predictive Intervals And Width δ For Various Prior Parameters

(N=25, T=100)

(L,U) δ	(10,10)	(20,20)	(b_1,b_2) (30,30)	(40,40)	(50,50)
(a_1, a_2) =(10,10)	(.67, 125.62) 124.95	(.46, 94.37) 93.91	(.35, 75.60) 75.25	(.28, 63.05) 62.77	(.24, 54.08) 53.84
(20,20)	(.69, 126.82) 126.13	(.48, 95.25) 94.77	(.36, 76.30) 75.94	(.29, 63.64) 63.35	(.25, 54.58) 54.33
(30,30)	(.71, 128.03) 127.32	(.49, 96.13) 95.64	(.37, 77.00) 76.63	(.30, 64.23) 63.93	(.25, 55.09) 54.84
(40,40)	(.73, 129.25) 128.52	(.50, 97.01) 96.51	(.38, 77.71) 77.33	(.31, 64.81) 64.50	(.26, 55.59) 55.33
(50,50)	(.74, 130.48) 129.74	(.51, 97.89) 97.38	(.39, 78.41) 78.02	(.32, 65.40) 65.08	(.27, 56.09) 55.82

As expected, shifts in the b_i parameters cause much more movement in the predictive interval than shifts in the a_i parameters. Because the predictive distribution

is unimodal and becomes steeper with larger b_i values, the length of the predictive interval decreases. Furthermore, the upper limit of the interval (U) moves more readily than the lower limit (L) because of the shape of the distribution.

As mentioned previously, setting the prior parameters to zero is equivalent to a state of ignorance of the parameters. This produces extremely wide predictive intervals for both the $n=100$ and $n=25$ samples of (1.04, 168.24) and (1.10, 214.26) and interval widths of 167.20 and 213.16 respectively. The effect of sample size on the interval length is clearly evident here.

The potential effect of prior information can be seen to be equally as important by setting the b_i parameters to the large values of $b_1 = 100$ and $b_2 = 100$ which produces extraordinarily small intervals for the $n=100$ and $n=25$ samples of (.38, 71.64) and (0.16, 41.75) with widths of 71.26 and 41.59 respectively. The priors are more concentrated about zero as the b_i's increase, since the prior means are $a_i/(b_i-1)$, . Thus, choice of the prior parameters has an enormous impact on the steepness of the resultant predictive distribution and hence the predictive interval. We see dominance of the prior distributional parameters even in the face of considerable information from the data. Clearly, this points towards a need for careful selection of the prior parameters for results to have any relevance to the application at hand.

The predictive density for the Mendenhall and Hader (1958) dataset was obtained for varying prior parameters, with results analogous to those for our given samples. Once again the influence of the b_i parameters was profound. A lack of prior information, represented by setting all prior parameters to zero, produced an enormously wide interval of (L,U) = (7.493, 1157.227). Using the values of $(a_1, b_1) = (10, 50)$ and

$(a_2, b_2) = (20, 40)$, we get $(L, U) = (5.693, 960.205)$ for an interval width of $\delta = 954.512$. Using b_i parameters to values assumed to be close to the theta parameters from the Mendenhall and Hader paper, we get $(a_1, b_1) = (10, 230)$ and $(a_2, b_2) = (20, 190)$ and produce a much narrower interval of $(3.060, 583.040)$ with a width of $\delta = 579.98$.

5.5 Monte Carlo Simulation

To more fully investigate the distributional properties of the length of the predictive interval (δ), Monte Carlo simulation of samples of size $n = 25$ was undertaken using prior parameter settings of $(a_1, b_1) = (10, 40)$ and $(a_2, b_2) = (10, 40)$ with censoring at time $T = 100$ because these gave reasonable results for the single sample case. Each sample was tested using the two EDF tests mentioned previously as well as the omnibus hypothesis of a mixture of exponential populations.

Over the 1000 samples, censoring at time $T = 100$ typically censored three or four of the $n = 25$ observations. This represented a moderate censoring between 12% and 16%. The predictive intervals averaged $\delta = 55.03$ in width with a standard deviation of 6.77 units. The smallest interval observed was $(0.144, 22.176)$ while the largest was $(0.208, 97.435)$ and a typical (median) interval was $(0.176, 54.413)$. The empirical frequency distribution is illustrated in Figure 5.5.

The interval produced for any particular sample was largely a result of the size and makeup of the smaller subsample. Because less information was available from the second subpopulation (the mixture proportion, p, was set at 0.75), r_2 could be as small as a single data point. Even if r_2 was bigger (averaging 6 points over the 1000 generated samples), a single relatively large value would inflate the censored sample average and

Figure 5.5

MIXTURE OF EXPONENTIALS PREDICTIVE INTERVAL
RESULTS FOR 1000 GENERATED SAMPLES
N=25, P=0.75, θ_1=50.0, θ_2=40.0, T=100.0
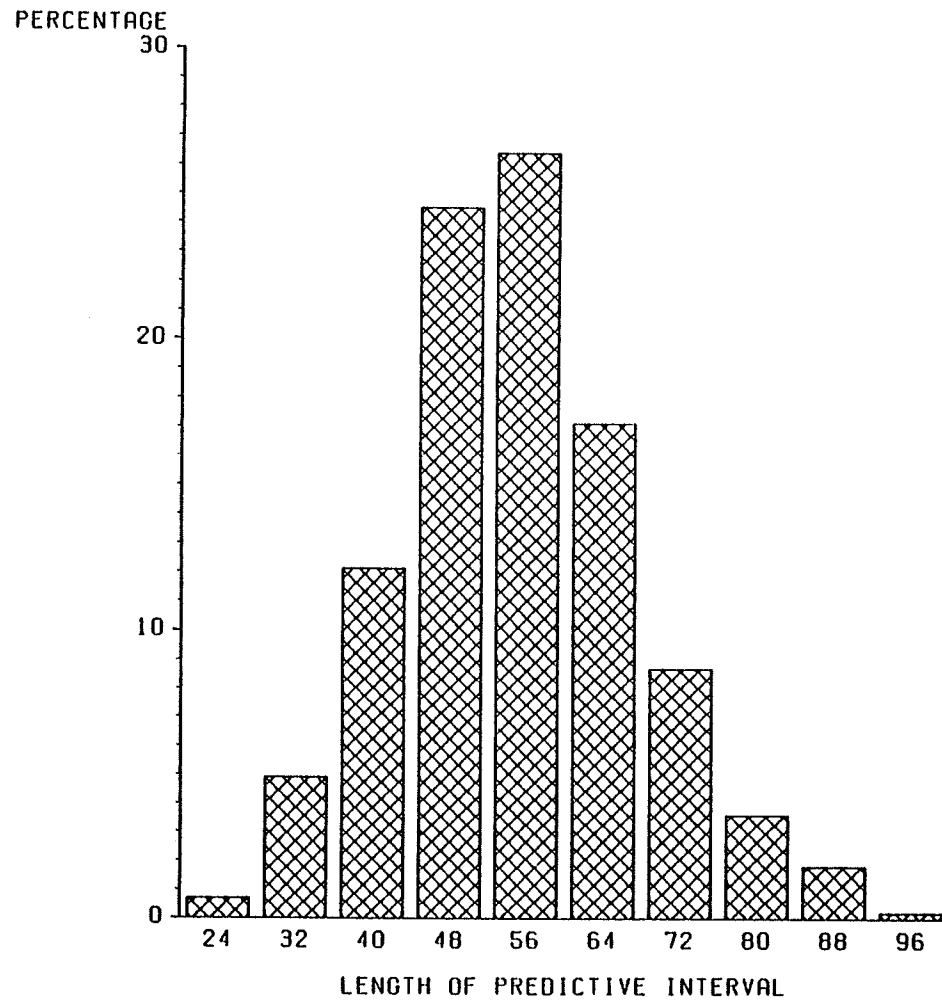
PRIOR PARAMETERS: (a_1, b_1)=(10, 40), (a_2, b_2)=(10, 40)

hence produce an unusually large predictive interval. This is evidenced by the fact that the empirical frequency distribution of the endpoints of the predictive intervals are markedly different. The lower endpoint's empirical frequency distribution follows that of a normal distribution while the upper endpoint's distribution is markedly and positively skewed (Shapiro-Wilk test for normality p-values of .2619 and .0001 respectively).

The Shapiro-Wilk test for normality (p-value = 0.0012) revealed that a normal curve does not adequately represent the sampling distribution of the predictive interval length (δ). It is desirable to be able to find a relatively simple closed-form approximation to the true pdf because the exact distribution cannot be obtained in closed form due to mathematical intractability.

Pearson compiled a family of probability density functions and established a methodology, based on observed momental constants, to select a family member curve from the system that would adequately fit the shape of an empirical frequency distribution (Elderton and Johnson (1969)). The type of Pearsonian curve chosen for a given empirical distribution depends upon the sample measures of skewness and kurtosis, typically denoted as β_1 and β_2 respectively. Elderton and Johnson (1969) provide a simple table to choose the appropriate Pearsonian curve based on the values of β_1 and β_2. Alternatively, one may use a nomograph from Pearson and Hartley's (1966) Biometrika Tables for Statisticians.

The observed values of $\beta_1 = .1069$ and $\beta_2 = 3.1006$ for the sampling distribution of δ suggest a Pearsonian Type I curve, which for the given empirical frequency

distribution is

$$f(x) = (29.331)\left(1 + \frac{x}{55.875}\right)^{17.062}\left(1 - \frac{x}{265.236}\right)^{80.992} \tag{5.11}$$

where x = δ - 52.939.

Simpson's rule was used to generate expected frequency counts for comparison with the observed sampling distribution. This produced a chi-square statistic of $\chi^2 = 8.983$ which carried four degrees of freedom. This compares favourably with $\chi^2(4, .05) = 9.488$. Hence the Type I Pearsonian curve fits the empirical sampling distribution of the predictive interval length (δ). The shape of the curve (5.11) is illustrated in Figure 5.6, reflecting a positively skewed distribution. To examine the effect of prior information on the resultant distribution of the predictive interval length, several other similar runs were performed varying the b_i parameters from zero to fifty in steps of ten units. The general shape of the empirical frequency distribution remained the same as depicted in Figure 5.6 with the sole difference being the extent of the upper tail. As the b_i parameters increase, the upper tail elongates due to the effects on the shape of the predictive distribution seen in the previous section. It should be noted that the effect of the prior information on the predictive interval length's distribution is not surprising due to the fact that the sample size used was quite small (n=25).

Further simulations were run with varying levels of censoring. Results were as one might expect in that the width of the predictive interval increased proportionately with the amount of sample information lost through censoring.

Figure 5.6

PREDICTIVE INTERVAL LENGTH: PEARSONIAN CURVE
TYPE I CURVE FITTED TO SAMPLING DISTRIBUTION

PDF

PREDICTIVE INTERVAL LENGTH
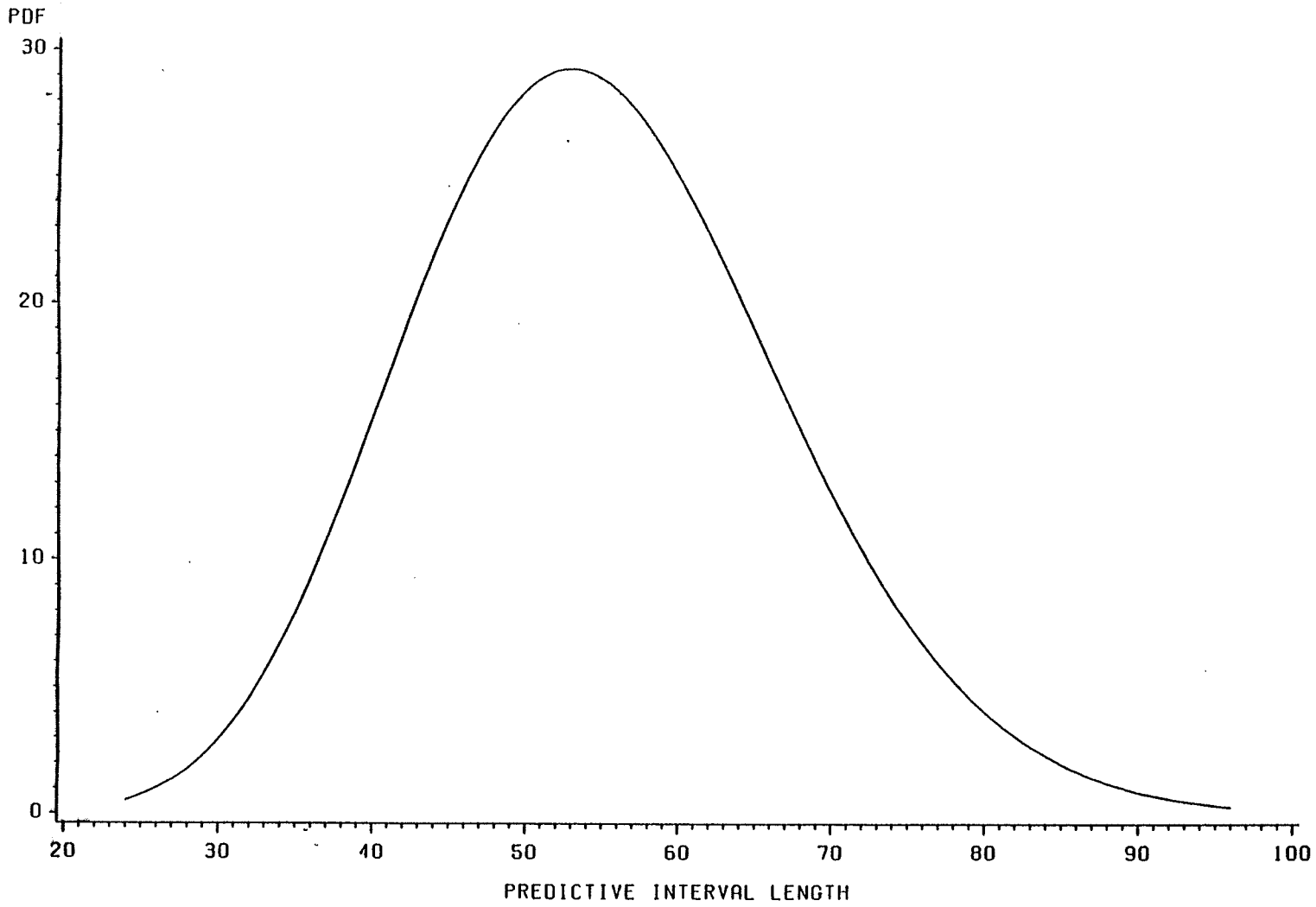
5.6 Complete Samples

In some situations complete samples become available because no censoring occurs. Many of the findings relevant to the censored samples case already presented are directly applicable to the complete samples case. The algebra simplifies dramatically when censoring is absent however and so it is of interest to consider this situation, if nothing more than a special case of a censored environment. The formulae of this section could be obtained by setting $T=\infty$, in the previous section, but the resulting equations are not obvious. Further insight into the environment is gained by looking at the problem from a different perspective. With complete sample information, it is interesting to investigate results for the previously used samples if a noninformative prior is used. To avoid redundancy, only cursory algebraic work is presented here. Detailed algebra is found in Sloan and Sinha (1988).

Algebraic work becomes much simpler due to the absence of any combinatoric terms which were necessary in the censored case. The likelihood is now

$$L(p, \theta_1, \theta_2 \mid t_{11}, \ldots, t_{1n_1}; \; t_{21}, \ldots, t_{2n_2}) \propto \frac{p^{n_1} q^{n_2}}{\theta_1^{n_1} \theta_2^{n_2}} \exp\left[-\frac{n_1 \bar{t}_1}{\theta_1} - \frac{n_2 \bar{t}_2}{\theta_2}\right]$$

where $q=1-p$.

Since complete sample information is now available and the impact of prior information was demonstrated in the previous section, a noninformative prior distribution will be used here to compare the relative importance of prior information versus moderate censoring. This is the primary reason for examining the complete samples case.

Jeffreys' invariant prior (Sinha, 1983) is

$$g(p, \theta_1, \theta_2) \propto \frac{1}{\theta_1 \theta_2}$$

so that the joint posterior distribution is

$$\Pi(p, \theta_1, \theta_2 | \underline{t}) = \frac{\left(n_1 \bar{t}_1\right)^{n_1} \left(n_2 \bar{t}_2\right)^{n_2} p^{n_1} q^{n_2}}{B(n_1+1, n_2+1) \Gamma(n_1) \Gamma(n_2) \theta_1^{n_1+1} \theta_2^{n_2+1}} \exp\left[-\frac{n_1 \bar{t}_1}{\theta_1} - \frac{n_2 \bar{t}_2}{\theta_2}\right]$$

for $0 \leq p \leq 1$ and $\theta_1, \theta_2 > 0$.

Following the triple integral form for the predictive distribution (5.7), it is found that the combinatorics are replaced by beta functions all of which cancel to produce, after considerable algebra

$$h(y|\underline{t}) = \frac{1}{n_1+n_2+2} \left[\frac{n_1(n_1+1)(n_1\bar{t}_1)^{n_1}}{(n_1\bar{t}_1+y)^{n_1+1}} + \frac{n_2(n_2+1)(n_2\bar{t}_2)^{n_2}}{(n_2\bar{t}_2+y)^{n_2+1}}\right] \qquad (5.12)$$

for $0 < y < \infty$.

Proceeding as with the censored samples case, equation (5.9) is used to set up the system of equations that will produce the $100(1-\alpha)\%$ predictive limits (L,U)

$$\frac{1}{n_1+n_2+2}\left[(n_1+1)\left[\frac{n_1\bar{t}_1}{n_1\bar{t}_1+L}\right]^{n_1} + (n_2+1)\left[\frac{n_2\bar{t}_2}{n_2\bar{t}_2+L}\right]^{n_2}\right] = 1-\alpha/2$$

and

$$(5.13)$$

$$\frac{1}{n_1+n_2+2}\left[(n_1+1)\left[\frac{n_1\bar{t}_1}{n_1\bar{t}_1+U}\right]^{n_1} + (n_2+1)\left[\frac{n_2\bar{t}_2}{n_2\bar{t}_2+U}\right]^{n_2}\right] = \alpha/2$$

which can once again be solved by iterative search techniques.

We reuse the generated samples from the censored case for the previous section and construct 95% predictive intervals. The n=100 and n=25 samples produce intervals of (1.04, 164.97) and (1.10, 110.09) with interval widths of 163.93 and 108.99 respectively. These intervals are almost as wide as the censored samples using the noninformative prior parameter settings. This would indicate even more strongly than before the importance of the prior information, even for a moderate sample size. Moderate censoring, then, does not impact as heavily on the predictive distribution as the amount of prior information available.

Monte Carlo simulation was subsequently undertaken for the n=100 sample size and the 1000 samples from the previous section were used. The predictive intervals averaged $\delta=183.21$ in width with a standard deviation of 18.45 units. The typical (median) interval was (1.18, 184.39). The large interval width maintained throughout, reinforcing the concept that the noninformative prior is more damaging in terms of lack of information than is moderate censoring.

The momental constants for the sampling distribution of the predictive interval width were $\beta_1=.0377$ and $\beta_2=3.0957$, which are indicative of a normal distribution. The Shapiro-Wilk test for normality indicated that a normal curve does adequately represent the sampling distribution (p=.7102). Sampling distributions of both the lower and upper endpoints follow a normal curve (p-values of .5236 and .3256 respectively). The complete sample information does at least induce normality to the predictive interval, reducing the possibility of anomalously large intervals. As was seen, however, ample prior information is necessary to bring the predictive interval's average length down to a reasonable level.

5.7 Summary

The advantage of prior information is clearly demonstrated in predicting future observations. In a state of ignorance, using noninformative prior settings, we see that the predictive intervals become very wide regardless of the amount of sample information. In such a situation the predictive intervals are likely of little practical value. With appropriately specified prior information however, the precision with which one can predict future observations increases dramatically.

In the censored case, the sampling distribution for the length of predictive intervals is positively skewed and nonnormal. If one of the subsamples is very small or severely censored, the result is an unduly large interval. This causes the tail of the sampling distribution to be elongated. When complete sample information is available the interval length distribution tends toward normality, with the extremely large intervals of the upper tail from the censored case being brought back into line. As one might expect, the predictive intervals become shorter when complete sample information is available. The greater the degree of censoring, the wider the interval becomes. The potential for dominance by the prior information is still present, however.

The author is indebted to a referee for suggesting the investigation of a conjugate prior distribution. Although the general shape of the predictive distribution and the sampling distribution for the length of the predictive interval were comparable using a noninformative prior, the effect of the prior distribution, as seen above, was impressive.

Algorithms used to produce predictive intervals in both the censored and uncensored case were included, under the name PREDSIM, in Press' (1989) review of statistical software available for Bayesian analysis.

CHAPTER 6: PREDICTIVE INTERVALS FOR A MIXTURE OF WEIBULL FAILURE-TIME DISTRIBUTIONS

6.1 Introduction

In this chapter, predictive intervals of a future observation for a mixture of Weibull distributions for complete and time-censored samples are studied assuming vague prior information. The sampling environment once again follows the definition of the previous chapter from Mendenhall and Hader (1958). Effects of the sample size and degree of censoring on the predictive interval are explored via Monte Carlo simulations. A portion of this work has been published in the South African Journal of Statistics by Sinha and Sloan (1989).

As described in the previous chapter, one often encounters situations where the underlying distribution is not homogeneous, but may consist of two or more subpopulations mixed in unknown proportions. Classical examples typically involve electrical components manufactured under different processing conditions or specifications. After having been sampled, each item of the population is attributable to the appropriate subpopulation. The mixture proportion p is an unknown entity. Titterington, Smith and Makov (1988) provide a detailed exposition of the theory and application of such mixture distributions.

The subject distribution of this chapter is a mixture of two Weibull subpopulations. The components under study are assumed to have lifetimes t, which follow a composite distribution such that

$$F(t) = pF_1(t) + (1-p)F_2(t)$$

where p is an unknown mixture proportion parameter and

$$F_i(t) = 1 - \exp\left[-\frac{t^{\alpha_i}}{\theta_i}\right]$$

which produces the corresponding mixture density function

$$f(t \mid p, \alpha_1, \alpha_2, \theta_1, \theta_2) = \frac{p\alpha_1}{\theta_1} t^{\alpha_1 - 1} \exp\left[-\frac{t^{\alpha_1}}{\theta_1}\right] + \frac{(1-p)\alpha_2}{\theta_2} t^{\alpha_2 - 1} \exp\left[-\frac{t^{\alpha_2 - 1}}{\theta_2}\right]$$

$$t, \alpha_1, \alpha_2, \theta_1, \theta_2 > 0 \qquad 0 < p < 1 \qquad\qquad (6.1)$$

A considerable amount of work has been done involving this model. Sinha (1987b) provides numerical algorithms for estimating the parameters involved $(p, \alpha_1, \alpha_2, \theta_1, \theta_2,)$ under a Bayesian framework. Lawless (1982) gives maximum likelihood estimators of the parameters as well as tests of hypothesis for investigating equivalence of the scale parameters for the two subpopulations. Cheng and Fu (1982) give a weighted least-squares alternative to maximum likelihood estimation of the parameters for the mixture of two Weibull distributions. McCool (1975, 1979) studied Weibull distributions and the effect of censoring.

Recall that the Bayesian predictive distribution of a future observation (typically a component lifetime) Y is defined as the posterior expectation of the density function

$$h(y \mid \underline{t}) \propto \int\int\int\int\int f(y \mid p, \alpha_1, \alpha_2, \theta_1, \theta_2) \Pi(p, \alpha_1, \alpha_2, \theta_1, \theta_2 \mid \underline{t}) \, dp \, d\alpha_1 \, d\alpha_2 \, d\theta_1 \, d\theta_2$$

where Π is the joint posterior distribution of the parameters.

Predictive distributions under industrial settings have been extensively studied. Englehardt and Bain (1979) provide maximum likelihood based prediction limits for the single sample Weibull distribution. Predictive distributions and intervals have been

derived for related lifetesting distributions such as the inverse Gaussian (Chhikara and Guttman, 1982) and the two-parameter exponential distribution (Lawless, 1977).

In this chapter n units are assumed to be sampled from the mixture of Weibull distributions model (6.1) by being subjected to some life testing experiment and letting $\underline{x} = (x_1, x_2, \ldots, x_n)$ be the failure times of these units. The complications of the Mendenhall and Hader (1958) sampling environment are also assumed to be present. Further assume that an additional set of n components of the same kind are to be put into future use and let $\underline{y} = (y_1, y_2, \ldots, y_n)$ be the future failure times of these components. The predictive distribution and the corresponding predictive interval of the future failure time Y of a single component on the basis of the previously observed life test data $\underline{x}$ will be derived using a Bayesian approach. Although the example of two subpopulations will be used, these results are directly generalisable to any arbitrary number of subpopulations.

6.2 Complete Samples Case

Let t_{ij} denote the failure time of the jth unit belonging to the ith subpopulation so that the observed sample drawn from the mixture of Weibull's distribution (6.1) may be represented by $\underline{t} = \{t_{ij} | j = 1, 2, \ldots, n_i; i = 1, 2\}$ and be thought of in terms of two subsamples that are identifiable a posteriori as being comprised of n_1 and n_2 components respectively.

As in the previous chapter, the attributability of the failed components to the appropriate subpopulation alters the likelihood function so that it is more than merely a product of the mixture density function 6.1. For further details the reader is referred to the discussion in Chapter five.

Due to the complex sampling situation, the likelihood function is more complicated than merely taking the product of the mixture density 6.1. The process of deriving the likelihood function described in chapter five is directly applicable here since the sampling environment is the same. Hence, given this complicated sampling environment, the likelihood function for this sample is

$$L(p, \alpha_1, \alpha_2, \theta_1, \theta_2 | \underline{t}, T) \propto \frac{p^{n_1}(1-p)^{n_2}\alpha^{n_1}\alpha^{n_2}}{\theta_1^{n_1}\theta_2^{n_2}}\lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2-1}\exp\left[-\frac{\sum_{j=1}^{n_1} t_{1j}^{\alpha_1}}{\theta_1}\right]\exp\left[-\frac{\sum_{j=1}^{n_2} t_{2j}^{\alpha_2}}{\theta_2}\right]$$

$$(6.2)$$

where

$$\lambda_i = \prod_{j=1}^{n_i} t_{ij}$$

As usual with Bayes theory, a decision must be made about the form of the prior distribution of the parameters. Following Jeffreys (1983), a state of vague knowledge will assume to exist about the five parameters. Appropriate choices under this assumption for the mixture parameter p and the shape parameters α_1 and α_2 are uniform prior distributions. Specifically, the marginal priors are

$$p{\sim}U(0,1), \quad \alpha_i{\sim}U(0,\infty) \quad for \ i=1,2.$$

Further, Jeffreys sets the scale parameter (θ_1, θ_2) priors as

$$g_i(\theta_i) \propto \frac{1}{\theta_i} \quad for \ i=1,2.$$

Assuming prior independence of knowledge regarding the five systemic parameters, the joint prior distribution is simply the product of the marginals such that

$$g(p, \alpha_1, \alpha_2, \theta_1, \theta_2) \propto \frac{1}{\theta_1 \theta_2} \quad . \tag{6.3}$$

Combining the likelihood (6.2) and the prior distribution (6.3), the joint posterior distribution is

$$\Pi(p, \alpha_1, \alpha_2, \theta_1, \theta_2 | \underline{t}) \propto \frac{p^{n_1}(1-p)^{n_2} \alpha_1^{n_1} \alpha_2^{n_2} \lambda_1^{\alpha_1-1} \lambda_2^{\alpha_2-1}}{\theta_1^{n_1+1} \theta_2^{n_2+1}} \exp\left[-\frac{\sum_{j=1}^{n_1} t_{1j}^{\alpha_1}}{\theta_1} - \frac{\sum_{j=1}^{n_2} t_{2j}^{\alpha_2}}{\theta_2} \right] . \tag{6.4}$$

The predictive density formula was given in the previous chapter as the integrated product of the posterior distribution and the parent distribution. Substituting the appropriate forms (6.4) and (6.1) respectively, the predictive density function of a future observation y may be found, after some algebra, to be $h(y|\underline{t})$ such that

$$h(y|\underline{t}) = \mathbb{C}\left[n_1(n_1+1) \int_0^\infty \frac{\lambda_1^{\alpha_1-1} \alpha_1^{n_1+1} y^{\alpha_1-1}}{\left[\sum_{j=1}^{n_1} t_{1j}^{\alpha_1} + y^{\alpha_1}\right]^{n_1+1}} d\alpha_1 \int_0^\infty \frac{\lambda_2^{\alpha_2-1} \alpha_2^{n_2}}{\left[\sum_{j=1}^{n_2} t_{2j}^{\alpha_2}\right]^{n_2}} d\alpha_2 \right.$$

$$\left. + n_2(n_2+1) \int_0^\infty \frac{\lambda_1^{\alpha_1-1} \alpha_1^{n_1}}{\left[\sum_{j=1}^{n_1} t_{1j}^{\alpha_1}\right]^{n_1}} d\alpha_1 \int_0^\infty \frac{\lambda_2^{\alpha_2-1} \alpha_2^{n_2+1} y^{\alpha_2-1}}{\left[\sum_{j=1}^{n_2} t_{2j}^{\alpha_2} + y^{\alpha_2}\right]^{n_2+1}} d\alpha_2 \right] \tag{6.5}$$

where

$$C^{-1} = (n+2) \left[\int_0^\infty \frac{\lambda_1^{\alpha_1-1}\alpha_1^{n_1}}{\left[\sum_{j=1}^{n_1} t_{1j}^{\alpha_1}\right]^{n_1}} d\alpha_1 \int_0^\infty \frac{\lambda_2^{\alpha_2-1}\alpha_2^{n_2}}{\left[\sum_{j=1}^{n_2} t_{2j}^{\alpha_2}\right]^{n_2}} d\alpha2 \right]$$

Unfortunately, equation (6.5) leaves six integrals which do not lead to closed forms. Numerical integration routines must hence be employed to evaluate the predictive density function. Producing the predictive distribution thus becomes a computer intensive exercise as it is necessary to evaluate six integrals at each point along the distribution.

The next step is to produce an equal-tail $100(1-\alpha)\%$ prediction interval [L,U]. This adds one more numerical integration step to the process as L and U are solutions to the equation

$$\frac{\alpha}{2} = \int_0^L h(y|\underline{t}) dy = \int_U^\infty h(y|\underline{t}) dy \qquad (6.6)$$

To produce the predictive interval a numerical integration of an already numerically integrated function must be performed.

6.3 Censored Samples Case

In most practical life testing applications it is impractical to perform the experimentation until the nth failure time is observed. Time censoring typically is performed so that a censored sample of r component lifetimes is observed. As before, each failure time can be attributed to the appropriate subpopulation after the unit has

failed so that the r failure times can be dichotomized into r_1 and r_2 observations from the two subpopulations respectively. Information regarding n_1 and n_2 is now unavailable and all that is known is that n-r components survived to time T. These n-r items are not attibutable to either subpopulation and as such must be represented only by the knowledge that they are drawn from the mixture population density 6.1. The sample can now be described by $\underline{t} = \{t_{11}, t_{12}, \ldots, t_{1r_1}; \ t_{21}, t_{22}, \ldots, t_{2r_2}\}, \quad t_{ij} \leq T.$

Incorporating the Mendenhall and Hader (1958) approach to this sampling environment once again, the likelihood function is obtained as

$$L(p, \alpha_1, \alpha_2, \theta_1, \theta_2 \mid \underline{t}, T) \propto \frac{p^{r_1}(1-p)^{r_2}\alpha^{r_1}\alpha^{r_2}}{\theta_1^{r_1}\theta_2^{r_2}}\lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2-1}\exp\left[-\frac{\sum_{j=1}^{r_1}t_{1j}^{\alpha_1}}{\theta_1} - \frac{\sum_{j=1}^{r_2}t_{2j}^{\alpha_2}}{\theta_2}\right]$$
$$\cdot \left[p\cdot\exp\left(-\frac{T^{\alpha_1}}{\theta_1}\right) + (1-p)\exp\left(-\frac{T^{\alpha_2}}{\theta_2}\right)\right]^{n-r}$$

with λ_i defined as above, except now the product represents a product over the non-censored observations.

The censoring has no impact on the prior information, so the posterior distribution is found in the same manner as for the complete samples case, but produces a more complicated result of

$$\Pi(p, \alpha_1, \alpha_2, \theta_1, \theta_2 \mid \underline{t}, T) \propto \frac{\alpha_1^{r_1}\alpha_2^{r_2}\lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2-1}}{\theta_1^{r_1+1}\theta_2^{r_2+1}}\left[\sum_{k=0}^{n-r}\binom{n-r}{k}p^{n-k-r_2}(1-p)^{r_2+k}\right]$$
$$\cdot \exp\left[-\frac{\sum_{j=1}^{r_1}t_{1j}^{\alpha_1}+(n-r-k)T^{\alpha_1}}{\theta_1} - \frac{\sum_{j=1}^{r_2}t_{2j}^{\alpha_2}+kT^{\alpha_2}}{\theta_2}\right]$$

Again, combining the posterior and parent distributions and performing the tractable

integrations involving p, θ_1 and θ_2 the predictive distribution of a future component lifetime is given by

$$h(y \mid \underline{t}, T) = C \sum_{k=0}^{n-r} \binom{n-r}{k} B(n-k-r_2+1, r_2+k+1)$$

$$\cdot \left[r_1(n-k-r_2+1) \int_0^\infty \frac{\lambda_1^{\alpha_1-1} \alpha_1^{r_1+1} y^{\alpha_1-1}}{\left[\sum_{j=1}^{r_1} t_{1j}^{\alpha_1} + (n-r-k) T^{\alpha_1} + y^{\alpha_1} \right]^{r_1+1}} d\alpha_1 \int_0^\infty \frac{\lambda_2^{\alpha_2-1} \alpha_2^{r_2}}{\left[\sum_{j=1}^{r_2} t_{2j}^{\alpha_2} + kT^{\alpha_2} \right]^{r_2}} d\alpha_2 \right.$$

$$\left. + r_2(r_2+k+1) \int_0^\infty \frac{\lambda_1^{\alpha_1-1} \alpha_1^{r_1}}{\left[\sum_{j=1}^{r_1} t_{1j}^{\alpha_1} + (n-r-k) T^{\alpha_1} \right]^{r_1}} d\alpha_1 \int_0^\infty \frac{\lambda_2^{\alpha_2-1} \alpha_2^{r_2+1} y^{\alpha_2-1}}{\left[\sum_{j=1}^{r_2} t_{2j}^{\alpha_2} + kt^{\alpha_2} + y^{\alpha_2} \right]^{r_2+1}} d\alpha_2 \right]$$

where

$$C^{-1} = (n+2) \sum_{k=0}^{n-r} \binom{n-r}{k} B(n-k-r_2+1, r_2+k+1)$$

$$\cdot \left[\int_0^\infty \frac{\lambda_1^{\alpha_1-1} \alpha_1^{r_1}}{\left[\sum_{j=1}^{r_1} t_{1j}^{\alpha_1} + (n-r-k) t^{\alpha_1} \right]^{r_1}} d\alpha_1 \int_0^\infty \frac{\lambda_2^{\alpha_2-1} \alpha_2^{r_2}}{\left[\sum_{j=1}^{n_2} t_{2j}^{\alpha_2} + kT^{\alpha_2} \right]^{r_2}} d\alpha 2 \right]$$

The prediction interval is again intractable by any means but numerical integration and is the solution to the system of equations (6.6). Naturally, the censored case involves a much more complicated expression than the complete samples case. To evaluate any point on the predictive density function six integrals must be numerically evaluated for each term of the summation running from zero to the number of observed failure times.

6.4 Example

Lawless (1982) contains an example dataset for the mixture of Weibull's model.

Failure times for twenty specimens subjected to accelerated life testing were observed, being comprised of ten specimens from each of two types of polyethylene cable. An underlying Weibull process was a reasonable assumption for the lifetime distribution.

Type I: 5.1, 9.2, 9.3, 11.8, 17.7, 19.4, 22.1, 26.7, 37.3, 60.0

Type II: 11.0, 15.1, 18.3, 24.0, 29.1, 38.6, 44.2, 45.1, 50.9, 70.0

It would not be reasonable to assume the distribution is a singular Weibull process. Clearly, the Type II dataset has a longer average lifetime, substantiated by an average sample lifetime of 34.6 in comparison to an average of 21.86 for the Type I insulation.

To assess the assumption of Weibull parent populations, empirical goodness of fit tests (D'Agostino and Stephens) were used. Tests were performed on each subsample to confirm that each was likely to have arisen from a Weibull process (p-values of 0.47 and 0.92 respectively). An omnibus test of the overall mixture of Weibull's model (6.1) was also done to verify that the model was appropriate for these data (p-value of 0.88). These tests were developed and implemented in the Microsoft BASIC language under the program name EDFIT and are available on request. Parametric values were not available for the given samples so maximum likelihood estimates as given by Lawless (1982) were used in the application of the empirical goodness of fit test procedures.

Construction of the predictive distribution is made difficult by the fact that the integration involving the systemic parameters was intractable. Since in most applied situations, the true value of the parameters will be unknown, it is necessary to estimate these entities to be able to construct the predictive distribution. As with the empirical goodness of fit tests, maximum likelihood estimates of the parameters are substituted for parametric values.

Another practical consideration arose in the computer implementation of the previous algebra. There are several parameterizations of the Weibull model, all of which are equally valid mathematically, differing only in the way in which the parameters are defined. For certain algebraic conveniences or intuitive applications, for example, it might be useful to define a scale parameter θ as $\beta=\theta^{-1}$. This minor algebraic consideration can have implications when computers enter the process and accuracy becomes a concern. For this model, parameterization is critical to the accuracy of the results due to the necessity of numerically integrating a large number of functions for each point on the predictive distribution.

In implementing the parameterization seen in previous sections, it was found to be numerically inconvenient, leading to consistent overflow and underflow conditions, even for small sample sizes. Although correctable, whenever such a situation occurs in programming, some degree of accuracy is lost.

A parameterization due to Lawless (1982) turned out to be much better suited to numerical work. The difference lies merely in the definition of the θ scale parameters. The Lawless form defines the scale parameter in terms of the algebra of the previous section as θ^{α}. This minor change does not alter the algebraic results, but does make for more practical computer implementation because it reduces the amount of exponentiation work required by the numerical integration routines.

As an illustration, for the given sample the mle's under the Lawless parameterization are

$$\hat{\alpha}_{1L}=1.51, \quad \hat{\alpha}_{2L}=2.11, \quad \hat{\theta}_{1L}=24.40, \quad \hat{\theta}_{2L}=39.25$$

whereas under the original parameterization the scale parameters are much larger.

Specifically, for the given sample the original constants would have been

$$\hat{\theta}_1 = (\hat{\theta}_{1L})^{\alpha_1} = (24.40)^{1.51} = 124.44$$

$$and$$

$$\hat{\theta}_2 = (\hat{\theta}_{2L})^{\alpha_2} = (39.25)^{2.11} = 2306.76$$

Even with the improved parameterization, the normalizing constants involved in this small sample are of the order 1×10^{23}. It should be noted that with careful implementation and corrective action for the overflow conditions the results found with the Lawless parameterization can be duplicated using the original algebra. The algorithms under the original parameterization take more than triple the CPU time, however.

Using the complete samples from the insulation data above, the predictive distribution and 95% predictive interval were constructed. The predictive density function is of a shape similar to those which will be presented in the next section and so is omitted here. The 95% predictive interval was found to be (2.184,62.500) which covers all of the given data except for one point. Although the interval is quite wide, it is reasonable given the fact that a noninformative a priori state is assumed. As was seen with the mixture of exponentials model in the previous chapter, incorporating accurate prior information to the process would reduce the width of the interval.

The censored samples case was then considered by censoring at time T=55, which reduces each subsample by one observation. Because for the given sample this amounts to excising obvious tail observations, the change in the predictive distribution was expected to be considerable. By censoring 10% of the dataset, the 95% predictive interval is now (2.023, 73.456), an increase of 18% in width. Movement of the interval is more marked in the upper tail due to the positive skewness of the parent Weibull

distributions.

6.5 Monte Carlo Simulation

Simulated samples allow for more informed analysis of the mixture of Weibull's model as the parametric values are within our control and can be used instead of the substituted maximum likelihood estimates as was done with the previous given sample. Both methods were investigated however, and the results were comparable.

Sample generation followed the method of Marsaglia (1961), which involves a two-stage randomization process. First the IMSL uniform random sample generation subroutine GGUBS was used to identify from which subpopulation an observation was to be drawn. Subsequently, the exponential variate generator REXP from IMSl was used along with appropriate transformation to produce the desired Weibull observation. As with the given sample, all generated samples underwent empirical goodness of fit testing procedures. Each subsample was tested as well as the adequacy of the overall mixture of Weibull's model. Any sample that failed to pass all three tests was rejected and another sample generated in its place.

The algorithms for the predictive distribution and interval was implemented on an Amdahl 470 mainframe using the PL/I programming language. The censored sample algorithm was much more expensive than the complete samples counterpoint, as might be expected. CPU time for the censored sample case ranged from double to five times that of the complete samples case, depending upon the sample size and degree of censoring.

To examine the relationship between sample size, degree of censoring and the

resultant predictive distribution, a series of simulations were run. Parametric settings of $p=0.75$, $\alpha_1=2.0$, $\alpha_2=3.0$, $\theta_1=5.0$, $\theta_2=6.0$ were used. Total sample size ran from very small ($n=10$) to moderate ($n=50$). The amount of censoring was controlled so that the percentage of sample information available that was used in calculations (PIU) ranged form 40% to the complete sampling case. Widths for the resultant predictive intervals are given in Table 6.1.

Table 6.1: Width Of 95% Predictive Intervals For Given N And PIU

	Percentage of Sample Information Used				
n	40%	50%	60%	70%	100%
10	6.44	5.22	3.75	3.24	2.23
15	6.40	4.02	3.75	2.87	2.58
20	6.18	3.18	2.82	2.69	2.34
30	5.04	3.96	2.88	2.56	2.57
40	2.52	2.87	2.77	2.88	2.88
50	2.25	2.35	2.84	3.83	3.05

The results of the simulation are not uniform across either the amount of sample information used or the sample size. This is due largely to the degree of censoring in the tail of the distribution. One outlying observation left in or censored out caused great shifts in the predictive interval, much in the way that was seen for the given sample in the previous section. The degree to which these outlying observations had an impact on the resultant predictive distribution did relate to the amount of sample information and sample size, analogous to the way that a sample mean is sensitive to outliers. If the sample size is sufficiently large, the predictive interval was affected to a lesser degree by the inclusion or exclusion of an observation from the upper tail of the parent

population than if the sample size was very small. It is remarkable, however, that because of these outlying observations, which arise due to the skewness of the parent Weibull populations, there are some situations where the predictive interval is actually narrower for the censored case than it is for the complete samples case.

This phenomenon notwithstanding, some general inferences can be drawn from the simulation results. The effect of censoring is much more pronounced for small samples (n < 30) due to the lack of remaining information. The predictive intervals for small censored samples are noticeably wider than those for the complete samples case which is what one would expect since we only have partial information with censored samples. As n increases, however, the two distributions trend to convergence.

For small fixed n (moving across any particular row of Table 6.1), as the percentage of sample information used increases the width of the predictive interval decreases and ultimately tends to the uncensored prediction interval. For larger n, more information remaining after censoring produces more stable interval estimates.

The predictive distributions for the complete samples case and 60% PIU censored samples case are given for the six sample sizes of Table 6.1 in Figures 6.1 through 6.6 respectively. In Figure 6.1, the impact of censoring on a small sample size is dramatic. The two distributions have markedly different upper tails although both demonstrate considerable positive skewness.

As n increases, flipping through the figures reveals a general, albeit inconsistent, convergence. This is a graphic illustration of the concept noted above that implied the secondary importance of censoring relative to the size of the sample. By the time n = 50 is reached in Figure 6.6, the complete and censored samples case's predictive

distributions are almost superimposed. The interim distribution in Figure 6.5 actually displays a situation where the censored distribution has a shorter tail than that of the complete samples case. This is indicative of the situation mentioned above.

6.6 Summary

In this chapter, the Mendenhall and Hader (1958) sampling environment involving the mixture of Weibull's model was seen to be much more algebraically complex than the mixture of exponentials model in Chapter five. The intractability of the integrals in the predictive distribution forms posed a major problem. Although not mathematically profound, the steps involved for the numerical work were trecherous and fraught with potential for measurement error. For the more complex censored samples case with even moderate sample size, five minutes of CPU time were required to complete the many numerical integrations.

These results hold practical implications for the use of censoring in order to predict future observations. Findings would suggest that if the sample size is sufficiently large, the exact moment that sampling is stopped is not a critical issue. In terms of predicting future observations, once the bulk of the distributional information is represented through sufficient sample size, the amount of censoring that has taken place is of secondary importance. This would seem to indicate that the adjustment made to the predictive distribution by the incorporation of the censoring time T adequately estimates the remaining distributional information. For large sample experimentation, then, if the experiment is costly to keep running, it would be equally as effective to subject a large number of components to testing for a shorter perior of time. The increased sample size

PREDICTIVE DENSITY FUNCTION
MIXTURE OF WEIBULL DISTRIBUTIONS
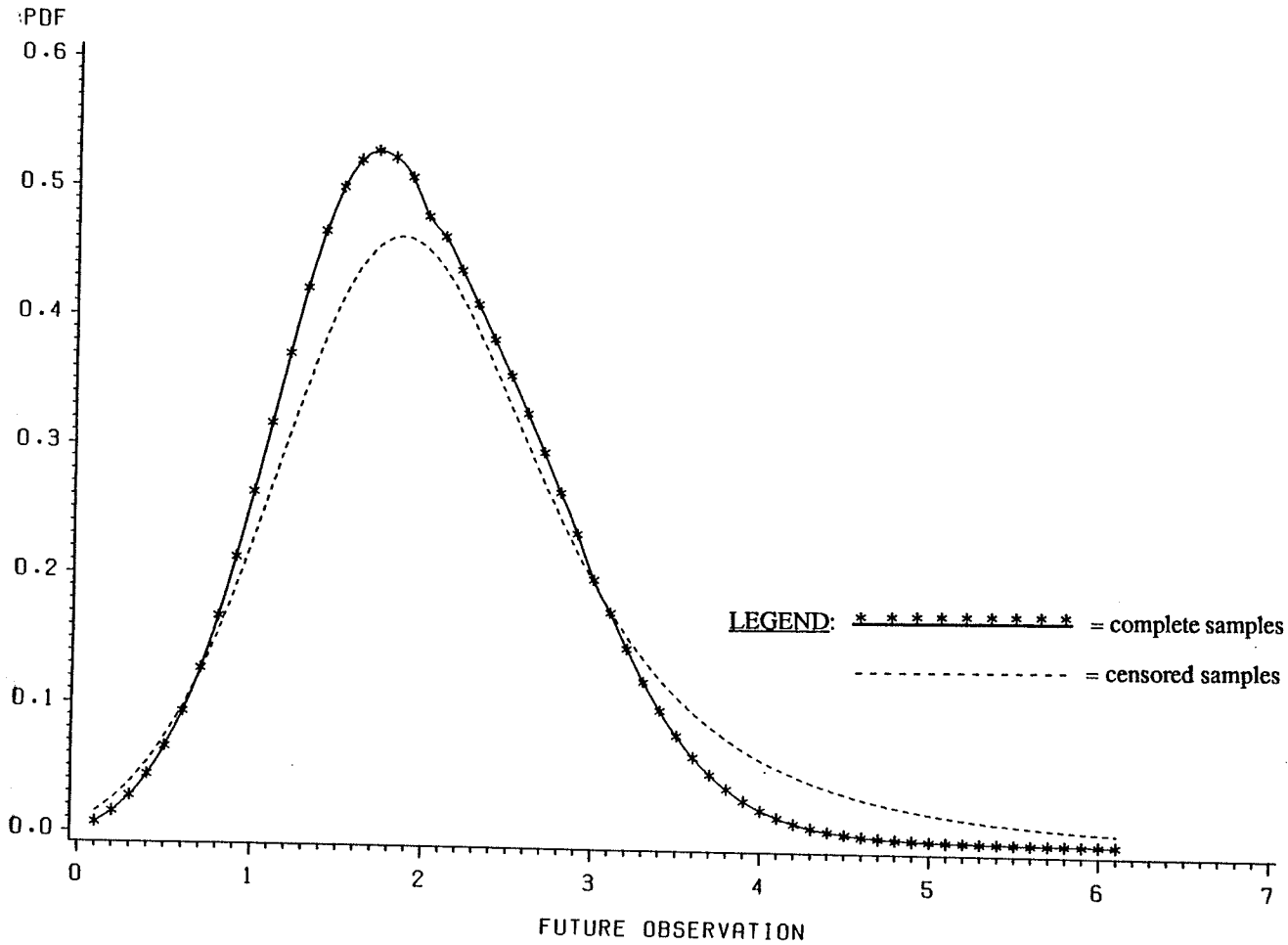## COMPLETE AND CENSORED SAMPLES CASE

Figure 6.1

LEGEND: $*$ $*$ $*$ $*$ $*$ $*$ $*$ $*$ = complete samples

- - - - - - - - - - = censored samples

FUTURE OBSERVATION
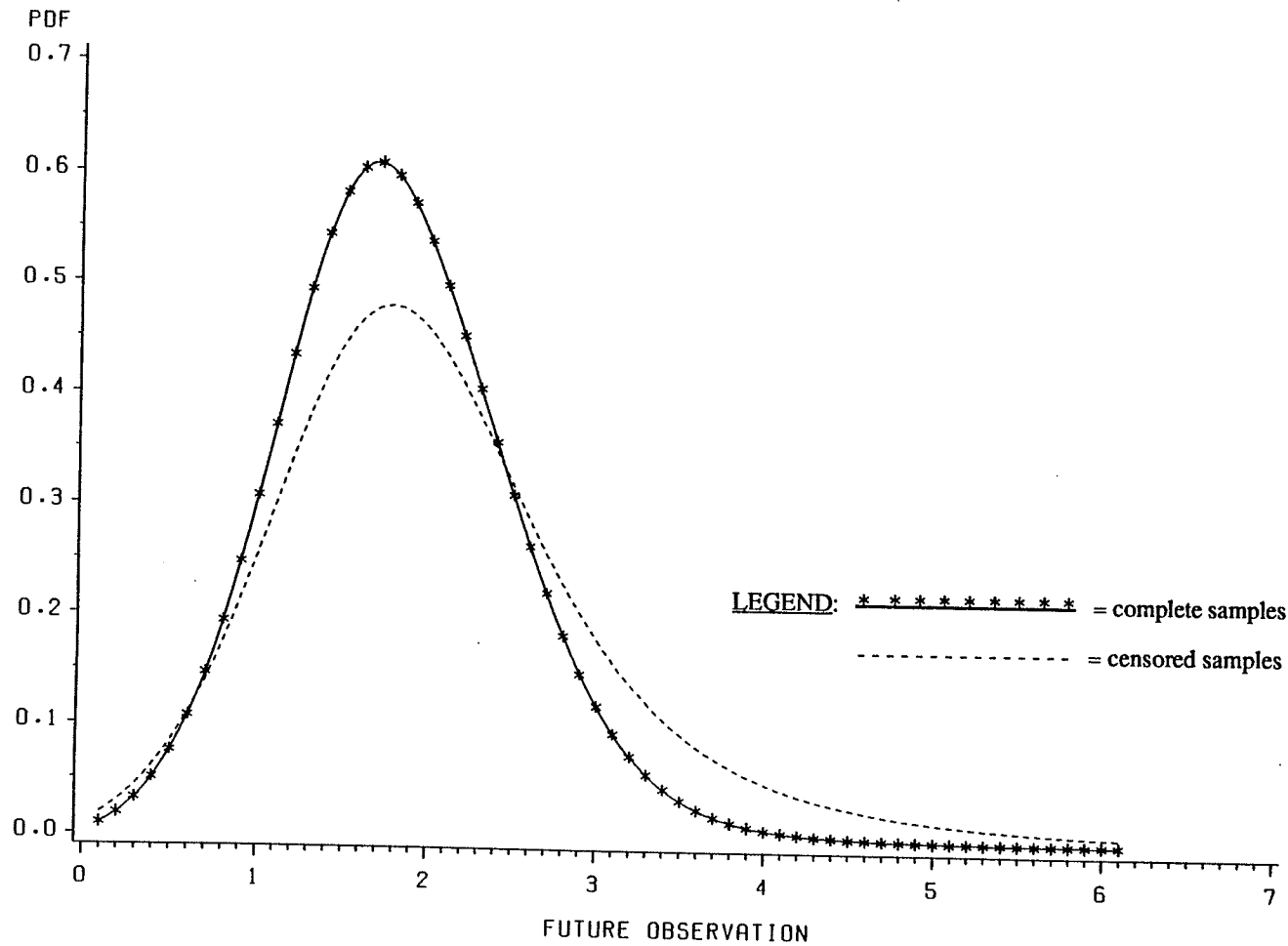
PARAMETERS: $P_1 = 2, P_2 = 3, \theta_1 = 5, \theta_2 = 6, n = 10$

PREDICTIVE DENSITY FUNCTION
MIXTURE OF WEIBULL DISTRIBUTIONS
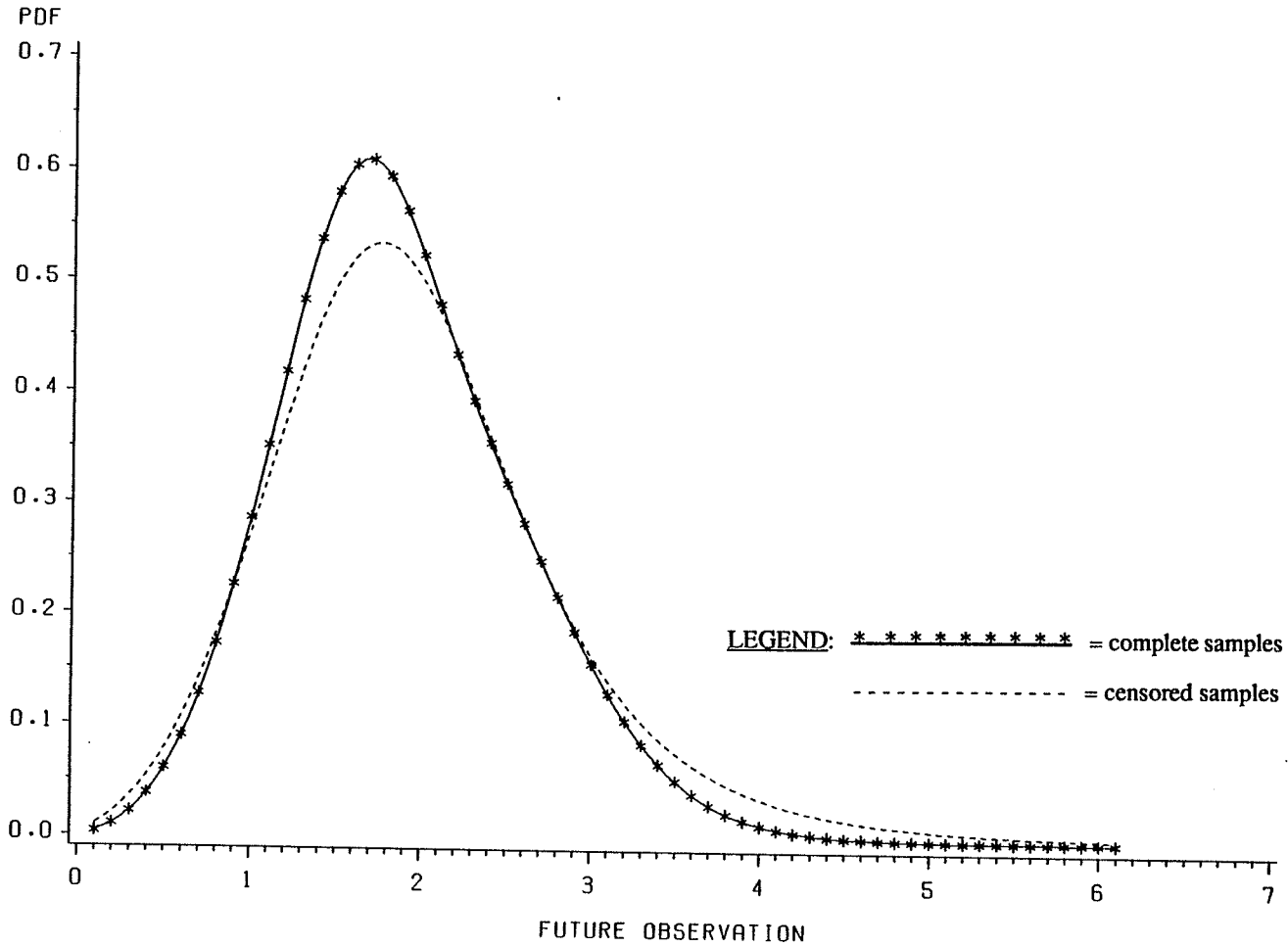## COMPLETE AND CENSORED SAMPLES CASE

Figure 6.2

LEGEND: $* * * * * * * * * *$ = complete samples

$- - - - - - - - - - - -$ = censored samples

FUTURE OBSERVATION

PARAMETERS: $P_1 = 2, P_2 = 3, \theta_1 = 5, \theta_2 = 6, n = 15$

PREDICTIVE DENSITY FUNCTION

Figure 6.3

MIXTURE OF WEIBULL DISTRIBUTIONS
COMPLETE AND CENSORED SAMPLES CASE

LEGEND: * * * * * * * * * = complete samples

- - - - - - - - - - - = censored samples
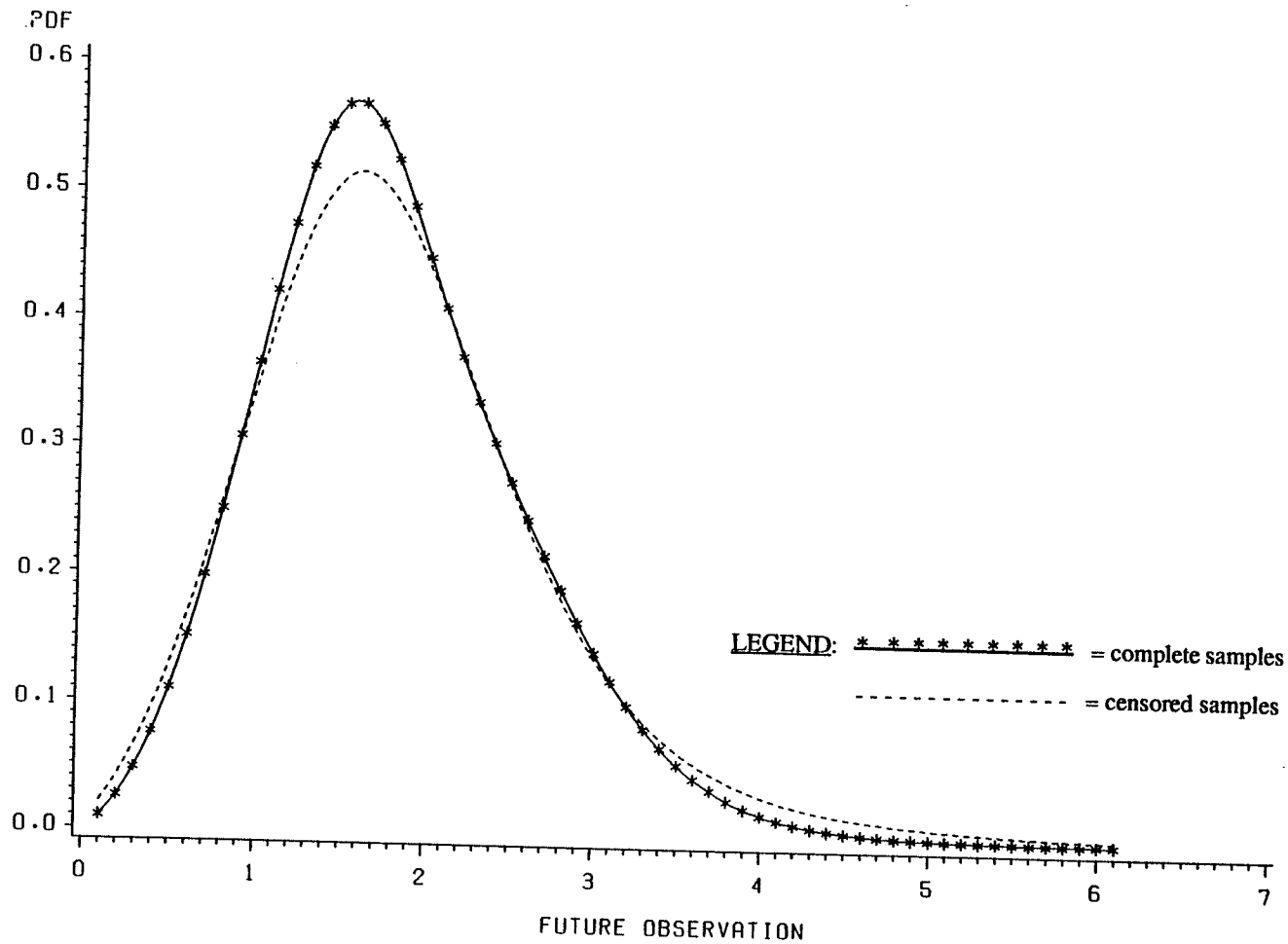
FUTURE OBSERVATION

PARAMETERS: $P_1 = 2, P_2 = 3, \theta_1 = 5, \theta_2 = 6, n = 20$

PREDICTIVE DENSITY FUNCTION

Figure 6.4

MIXTURE OF WEIBULL DISTRIBUTIONS
COMPLETE AND CENSORED SAMPLES CASE

FUTURE OBSERVATION

PARAMETERS: $P_1 = 2, P_2 = 3, \theta_1 = 5, \theta_2 = 6, n = 30$

LEGEND: ＊ ＊ ＊ ＊ ＊ ＊ ＊ ＊ ＊ = complete samples

－ － － － － － － － － － － = censored samples

PREDICTIVE DENSITY FUNCTION
MIXTURE OF WEIBULL DISTRIBUTIONS
COMPLETE AND CENSORED SAMPLES CASE

Figure 6.5

LEGEND: * * * * * * * * * = complete samples

- - - - - - - - - - - - - - = censored samples

FUTURE OBSERVATION

PARAMETERS: $P_1 = 2, P_2 = 3, \theta_1 = 5, \theta_2 = 6, n = 40$

Figure 6.6

PREDICTIVE DENSITY FUNCTION
MIXTURE OF WEIBULL DISTRIBUTIONS
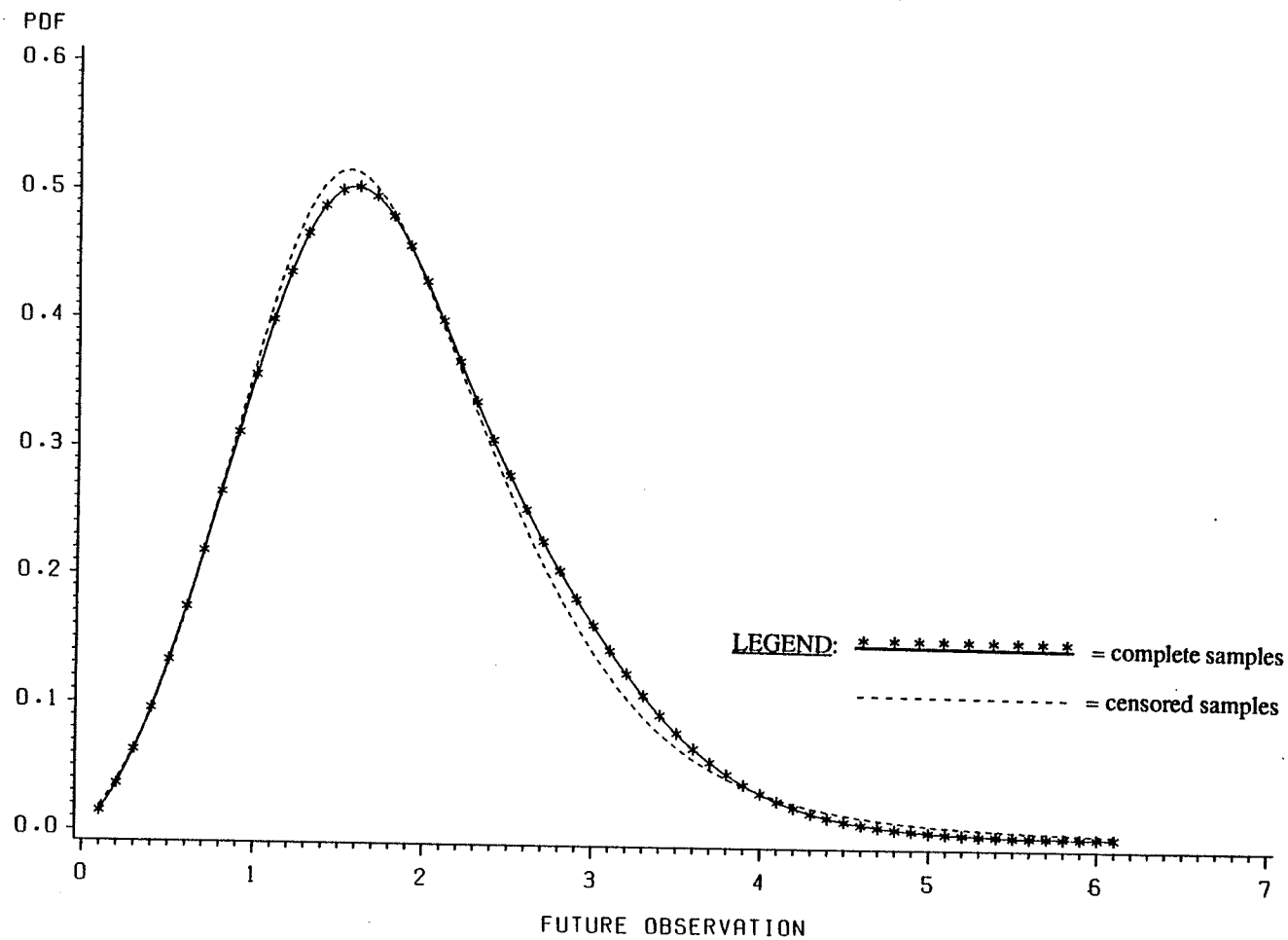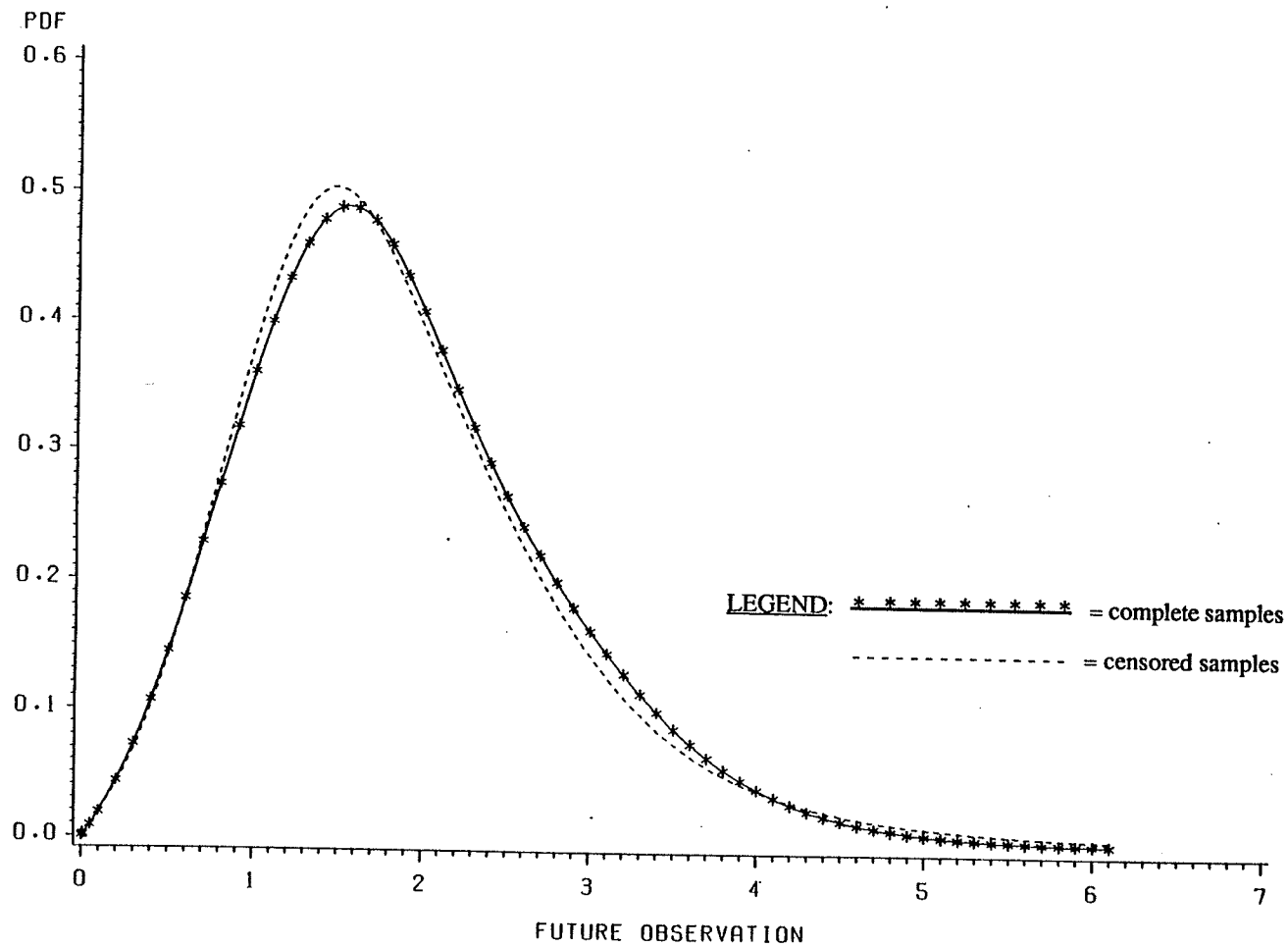### COMPLETE AND CENSORED SAMPLES CASE

PARAMETERS: $P_1 = 2, P_2 = 3, \theta_1 = 5, \theta_2 = 6, n = 50$

would offset the lost information due to censoring. The importance of the loss of sample information through censoring was seen to have differing degrees of importance, depending upon the base sample size.

CHAPTER 7: BAYESIAN p-CHARTS FOR PROCESS CONTROL

7.1 Introduction

First detailed by Dodge and Roming (1959), attribute sampling plans have been studied extensively under various conditions. Duncan (1974) contains an extensive list of references. They are an integral part of basic quality control methodology (see Montgomery, 1985, for example). They apply to production environments where the items under study are declared to be of acceptable quality or not based on the number of items found to be nonconforming to accepted standards. As first proposed by Shewhart(1931), control charts were derived from the classical, or frequentist perspective. Attribute sampling plans (MIL-STD-105D, 1963 for example) were developed under the simplistic assumption of the normal approximation to the binomial distribution, an assumption which several authors have demonstrated to have potentially misleading ramifications. Furthermore, the classical approach assumes that the probability of a defect is constant. In an assembly line situation, this assumption often fails in practice.

Approaching quality control from a Bayesian perspective has been considered by several authors (Hald, 1981 provides an extensive reference list). Attribute sampling under a Bayesian framework has been detailed by the American Society for Quality Control in Calvin (1984). The primary difference between this approach and the classical method is to produce an estimator that will more readily incorporate new knowledge on the process variation. The way in which the intervals are used as process watchdogs is no different from the usage of the classical interval formulation. It is expected, however, that the intrinsic nature of the Bayesian approach will produce more

accurate interval estimators and hence a more accurately measured process.

In this chapter attribute control charts are further examined (also referred to as p-charts) under the Bayesian framework. The traditional Bayesian approach merely substitutes the Bayes estimator of process variation for its classical counterpart in the p-chart formula (Calvin, 1984). The use of predictive intervals, highest posterior density(HPD) intervals and predictive p-charts as alternatives to the classical p-charts and traditional Bayesian p-charts is investigated.

7.2 Alternatives for p-chart Control Limits

7.2.1 Model Preliminaries

Consider an attribute sampling environment where a sample of n items is tested, producing x defects. The pdf of x is the binomial B(n,p) where p is the probability of an item being defective.

Shewhart proposed a general formulation for a control chart of a quality characteristic. If S is a statistic that estimates such a characteristic, then knowing the mean μ_S and the standard deviation σ_S for the statistic S allows us to construct a k-sigma Shewhart control chart as follows

$$(LCL, UCL) = (\mu_S - k\sigma_S, \mu_S + k\sigma_S)$$

Typically, k=3.

If the characteristic of interest is the proportion p of nonconforming (or defective) units in the population, then the classical p-chart is

$$(LCL, UCL) = \hat{p} \pm 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (7.1)$$

where $\hat{p}$ is the usual sample average proportion of nonconforming units.

In testing r samples (running r experiments), the likelihood function is

$$L(p|\underline{x}) = \prod_{i=1}^{r} \begin{bmatrix} n \\ x_i \end{bmatrix} p^{\sum_{i=1}^{r} x_i}(1-p)^{nr-\sum_{i=1}^{r} x_i}$$

Let $t = \sum_{i=1}^{r} x_i$ be the total number of defective units found in the r samples tested. Then

the sample proportion defectives for the classical p-chart is

$$\hat{p} = \frac{t}{nr}$$

and the likelihood function can be rewritten as

$$L(p|x) \propto p^t(1-p)^{nr-t}$$

Under a Bayesian framework, p is considered to be a random variable with conjugate prior distribution

$$g(p) \propto p^{t-1}(1-p)^{m-1}$$

so that p has a beta prior.

Combining prior and sample information, the posterior distribution of p is

$$\Pi(p|\underline{x}) \propto p^{t+\ell-1}(1-p)^{nr+m-t-1} \qquad 0 \le p \le 1$$

which is clearly a beta distribution with parameters $t+\ell$ and $nr+m-t$.

Assuming squared error loss, the Bayes estimator p^* of p is the mean of the posterior distribution which is easily found to be

$$p^* = \frac{t+\ell}{\ell+m+nr} \ .$$

Under the Bayesian framework, the classic confidence intervals are referred to as credible intervals. The interpretation of a credibility interval differs markedly from that of a confidence interval in that one refers precisely to the probability that the parameter under consideration falls within a specified interval, conditional on the given observations. One such technique of finding credible intervals is the so-called highest posterior density (HPD) interval. An interval (LCL,UCL) is sought which satisfies two conditions:

1) F(UCL) - F(LCL) = 1-α

2) (LCL,UCL) is the shortest among all intervals satisfying (1).

For unimodal distributions, the second condition is replaced by

2) the posterior density at the interval endpoints is identical

i.e., $\Pi(p=LCL|\underline{x})=\Pi(p=UCL|\underline{x})$.

where Π represents the posterior density.

Since the posterior distribution is beta with parameters strictly greater than one, the posterior distribution must be unimodal with mode at

$$mode[\Pi(p)] \;=\; \frac{t+\ell-1}{nr+\ell+m-2} \quad .$$

This will provide a starting point for the numerical search routine.

7.2.2 The Predictive Interval

Given the prior and sample information, the predictive distribution for the number of nonconforming items to be found in a future lot of n items is given by

$$h(y|\underline{x}) \;=\; \int_0^\infty f(y)\pi(p|\underline{x})dp$$

where f(y) is the parent binomial distribution of the process. After collecting terms this becomes

$$h(y|\underline{x}) \propto \binom{n}{y}\int_0^1 p^{t+\ell+y-1}(1-p)^{nr+m+n-y-t-1}dp$$

$$\propto \binom{n}{y}B(t+\ell+y,nr+m+n-y-t)$$

After removing some constants and collecting like terms

$$h(y|\underline{x}) \propto \binom{t+\ell+y-1}{y}\binom{nr+m+n-t-y-1}{n-y}$$

$$\propto \left[\begin{matrix} t+\ell+y-1 \\ t+\ell-1 \end{matrix}\right]\left[\begin{matrix} nr+m+n-t-y-1 \\ nr+m-t-1 \end{matrix}\right]$$

$$h(y|\underline{x}) = C\left[\begin{matrix} \nu_1+y-1 \\ \nu_1-1 \end{matrix}\right]\left[\begin{matrix} \nu_2+n-y-1 \\ \nu_2-1 \end{matrix}\right]$$

where

$$C^{-1} = \sum_{y=0}^{n}\left[\begin{matrix} \nu_1+y-1 \\ \nu_1-1 \end{matrix}\right]\left[\begin{matrix} \nu_2+n-y-1 \\ \nu_2-1 \end{matrix}\right], \ \ \nu_1=t+\ell, \ \nu_2=nr+m-t$$

To evaluate this function, consider the following expansions

$$(1-z)^{-\nu_1} = \sum_{j=0}^{\infty}\left[\begin{matrix} \nu_1+j-1 \\ j \end{matrix}\right]z^j$$

and

$$(1-z)^{-\nu_2} = \sum_{k=0}^{\infty}\left[\begin{matrix} \nu_2+k-1 \\ k \end{matrix}\right]z^k$$

Multiply the two series together and collect the coefficients of z^n.

$$\left[\begin{matrix} \nu_1+\nu_2+n-1 \\ n \end{matrix}\right] = \sum_{j=0}^{n}\left[\begin{matrix} \nu_1+j-1 \\ j \end{matrix}\right]\left[\begin{matrix} \nu_2+n-j-1 \\ n-j \end{matrix}\right]$$

$$\left[\begin{matrix} \nu_1+\nu_2+n-1 \\ \nu_1+\nu_2-1 \end{matrix}\right] = \sum_{j=0}^{n}\left[\begin{matrix} \nu_1+j-1 \\ \nu_1-1 \end{matrix}\right]\left[\begin{matrix} \nu_2+n-j-1 \\ \nu_2-1 \end{matrix}\right]$$

which is the inverse of our normalizing constant C.

Incorporating this information, the predictive distribution for the number of

defectives in a future lot of n items is found to be

$$h(y|\underline{x}) = \frac{\begin{bmatrix} \nu_1+y-1 \\ \nu_1-1 \end{bmatrix}\begin{bmatrix} \nu_2+n-y-1 \\ \nu_2-1 \end{bmatrix}}{\begin{bmatrix} \nu_1+\nu_2+n-1 \\ \nu_1+\nu_2-1 \end{bmatrix}} \qquad (7.2)$$

for y=0,1,2...,n and hence y is beta-binomial b(y,n,ν_2,ν_1).

The mean of the predictive distribution is

$$E(y|\underline{x}) = \frac{t+\ell}{\ell+m+n+r}$$

and is a Bayes estimator of y under squared error loss.

A 100(1-α)% predictive interval for the number of defects in a future lot of n items is composed of the endpoints (L,U) such that

$$\sum_{y=0}^{L} h(y|\underline{x}) = \sum_{y=U}^{n} h(y|\underline{x}) = \alpha$$

Calvin (1984) discusses the predictive distribution (7.2) as being a Polya distribution. He arrives at the so-called Polya distribution by assuming that the process is not stable. The process is assumed to vary about its average in an individual manner with the probability of an item being nonconforming varying over time.

Calvin uses this distribution to replace σ_p in the p-chart limit formula by

$$\sigma_p^2 = \frac{\hat{p}(1-\hat{p})}{n}\begin{bmatrix} n+l+m \\ 1+l+m \end{bmatrix} \qquad (7.3)$$

This variance form is interesting in that it will always produce wider control limits than its classical counterpart. In this manner, however, Calvin states that the Polya distribution will give a better assessment of the true sampling risks than the binomial

model of the frequentist approach. Hald (1981) states that this variance formulation is significantly larger than that of the classical binomial model.

All four of the alternative approaches to a p-chart have been derived. Before exploring the relative merits of each approach, a more basic problem must be addressed, namely, estimating the prior distribution parameters upon which all the Bayes methods rely.

7.2.3 Prior Estimation

In quality control applications, it is very likely that little knowledge will be available about the prior parameters ℓ and m. Two approaches are open to the researcher under such circumstances.

One may assume a state of ignorance and make use of a noninformative prior. In this case one could use the uniform prior $g(p)=1$. A second approach was to use an asymptotically linear invariant (ALI) prior as set out in Hartigan (1964) by putting $\ell=m=0$. Another alternative by Jeffreys was to set $\ell=m=0.5$ (Hartigan, 1983). As has been demonstrated under several applications, however, such an assumption typically produces predictive intervals that are so wide as to be practically useless.

A second and more informative approach is to use the Empirical Bayes methods due to Robbins (1964) to estimate the prior parameters ℓ and m from the sample data.

Consider the marginal distribution of x, the number of defects in the lot, from our model by integrating over p. This produces

$$f(x) \propto \binom{n}{x} \int_0^1 p^{x+\ell-1}(1-p)^{n+m-x-1}dp$$

$$\propto \binom{n}{x} B(x+\ell, n+m-x)$$

and following the algebra in the previous section,

$$f(x) \propto \binom{m+n-x-1}{m-1} \binom{x+\ell-1}{\ell-1}$$

Restoring the normalizing constant, the marginal distribution is then obtained

$$f(x) = \frac{\binom{m+n-x-1}{n-1} \binom{x+\ell-1}{\ell-1}}{\binom{m+n+\ell-1}{m+\ell-1}} \quad .$$

Hence, the marginal distribution of x is also beta-binomial b(x,n,m,l).

The forms for the mean and variance of the beta-binomial distribution are

$$E(X) = \frac{n\ell}{\ell+m}$$

$$Var(X) = \frac{n\ell m(\ell+m+n)}{(\ell+m)^2(\ell+m+1)} \quad .$$

Define the two first sample moments as follows

$$a = \bar{X}_r = \frac{\sum_{i=1}^{r} x_i}{r} = \frac{t}{r}$$

$$v = s_x^2 = \frac{1}{r-1}\sum_{i=1}^{r}(x_i - a)^2 \quad .$$

Equating the true moments to their sample counterparts produces the system of equations

$$a = \frac{n\ell}{\ell + m}$$

$$v = \frac{n\ell m(\ell + m + n)}{(\ell + m)^2(\ell + m + 1)} \quad .$$

From the first equation of the system derive

$$\ell + m = \frac{n\ell}{a}$$

and substituting this into the second equation gives

$$v = \frac{n\left[\frac{a}{n}\right]\left[1 - \frac{a}{n}\right]\left[\frac{n\ell}{a} + n\right]}{\left[\frac{n\ell}{a} + 1\right]} \quad .$$

Finally, solve for ℓ and m to produce estimates for the prior parameters

$$l^* = \left[\frac{a}{n}\right]\left[\frac{a(n-a) - v}{v - a\left[1 - \frac{a}{n}\right]}\right]$$

$$m^* = \frac{\ell^*(n-a)}{a} \quad .$$

These preliminary estimators for the beta prior parameters allow for much greater precision in the posterior and predictive information as will be seen later. A problem arises in their use with the possibility that the estimators can be negative. When one is negative the other must be also. In such a case it is sensible to use the noninformative priors mentioned above. Since little or no information about the true prior parameters is assumed under each of the noninformative approaches, a uniform prior ($\ell = m = 1$) will

be used when the empirical prior estimates are unacceptable.

Negative values for ℓ^* and m^* occur under special circumstances. It should be noted that only two cases exist. Either both prior estimates are positive or both are negative. Negative values are more likely to occur when the true proportion defectives in the population is small. From a sample perspective it is also true that if the observed variability in the number of defectives across samples is large relative to the average number of defectives, then v is overestimated. This causes ℓ^* to be negative.

7.2.4 Effect Of Prior Parameters

This section gives some insight to the impact on the posterior of using our ℓ^* and m^* priors over the noninformative priors. As previously stated, the noninformative priors give little or no information about the true value of p, the population proportion of nonconforming units. Using the empirical Bayes estimates for the beta prior parameters ℓ and m, it will be seen that a prior is produced that contains a considerable amount of information about p. Although this will be further detailed in the discussion of the HPD and predictive interval methodology, it is useful to put in context at this point the value of the empirical quantities.

Figures 7.1 and 7.2 display the resultant prior distributions observed from generated data. In a practical setting this sort of exercise typically takes the form of a pre-production run. It is readily apparent from the two graphs that these priors give a good deal of accurate information about the prior distribution. Clearly, including such information in the analysis should produce more accurate results than merely using a flat uniform distribution for the prior.

Figure 7.1

ESTIMATED BETA PRIOR DISTRIBUTION
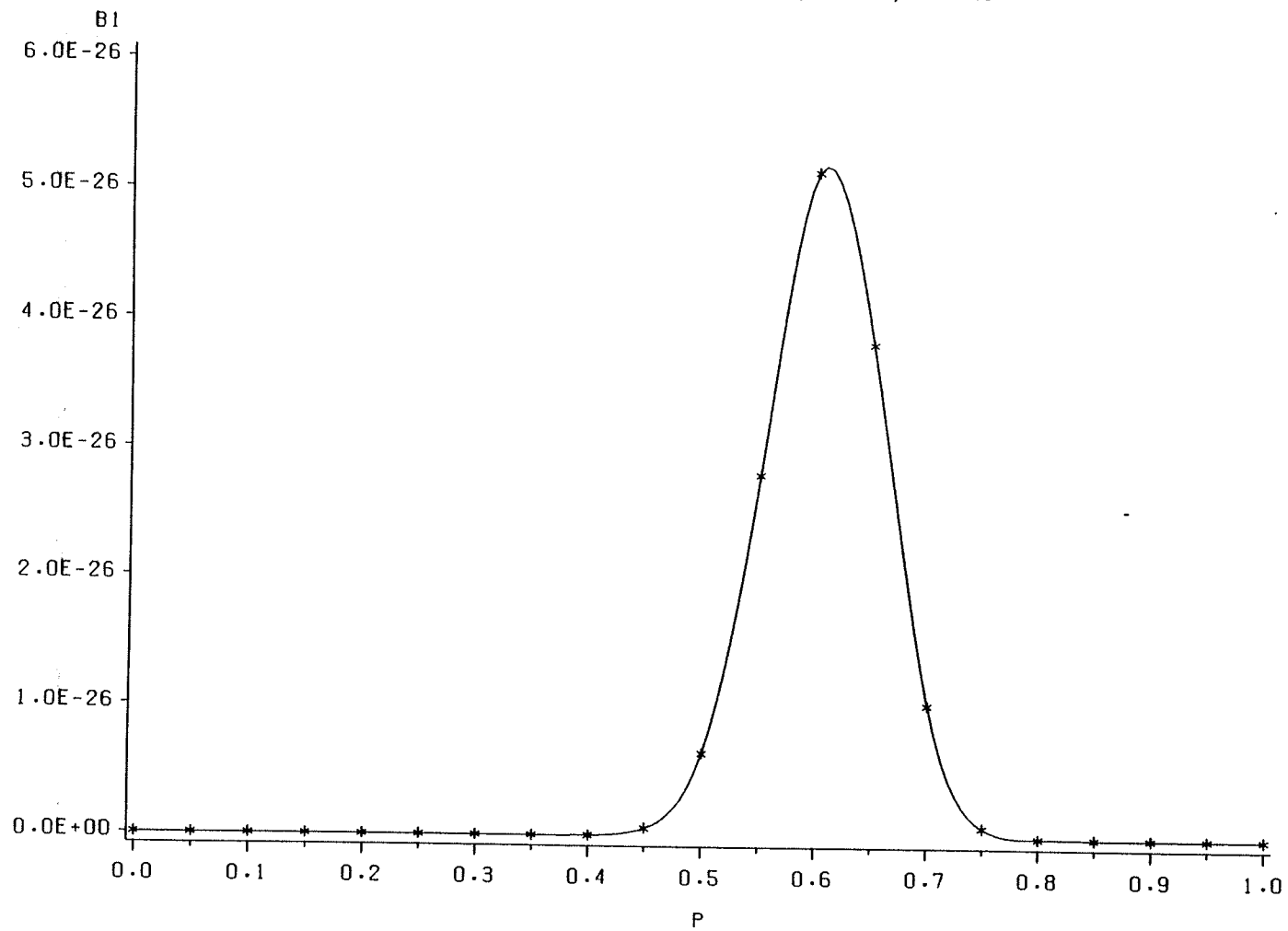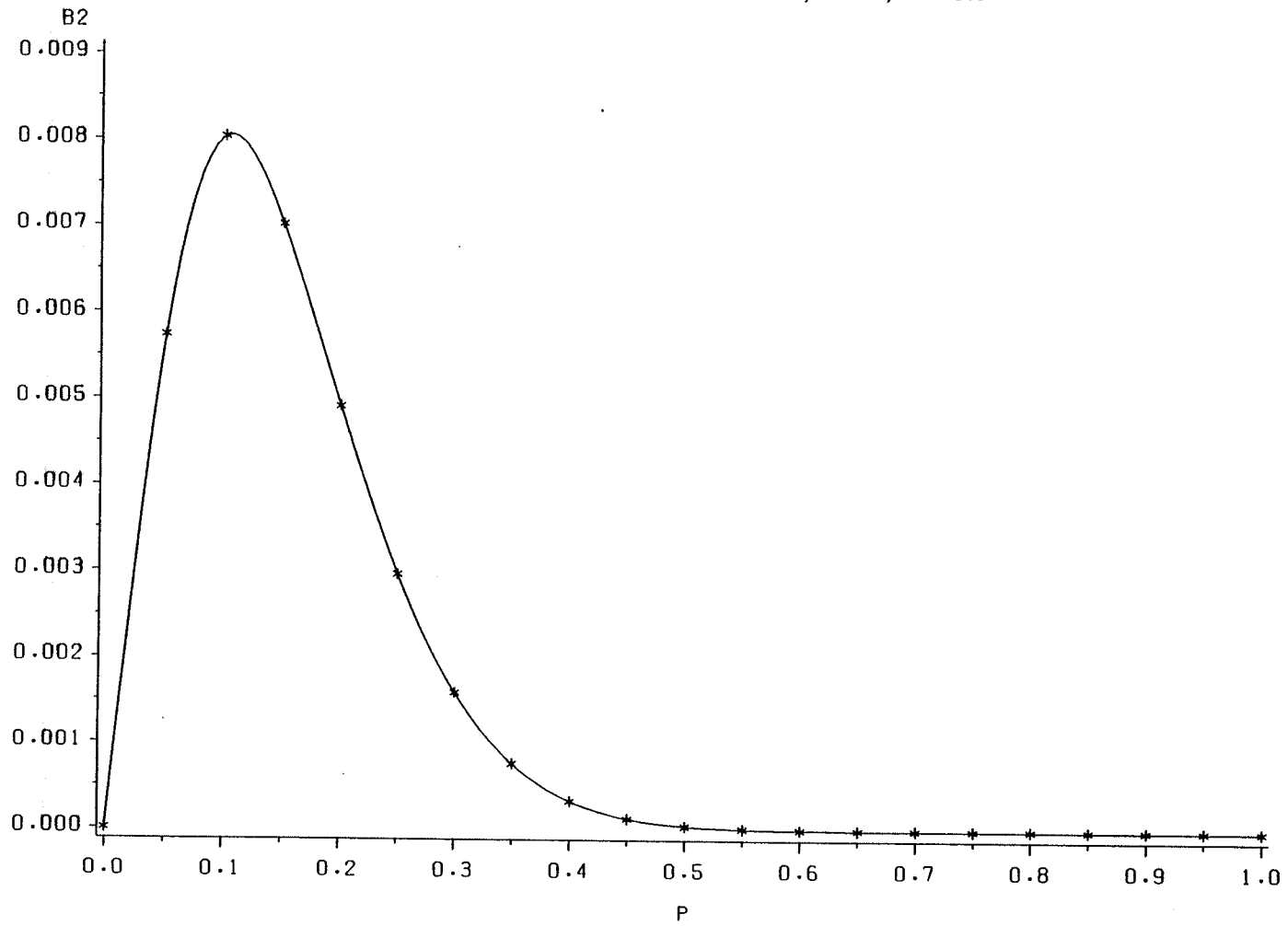TRUE PARAMETERS: P=0.6, L=53.0, M=35.0

Figure 7.2

ESTIMATED BETA PRIOR DISTRIBUTION
TRUE PARAMETERS: P=0.1, L=1.5, M=13.0

7.2.5 Comparisons Of The Four Methods

Now that the theoretical foundation has been built, comparisons may be made among the four alternative approaches. The classical p-chart will be compared against:

1) the semi-Bayesian p-chart which uses p^* and σ_p from (7.3)

2) the predictive interval

3) the HPD interval.

Results are produced for selected sample sizes and sample results. To begin, consider example datasets drawn from Montgomery (1985) on samples of $n=50$ orange concentrate cans. Three different samples are presented in total, referred to as the M1, M2 and M3 samples. Specifications for the three samples are given in Table 7.1.

Table 7.1: Specifications For Montgomery Samples

Set	r	t	$\hat{p}$	p^*	(l^*,m^*)	Classical	Semi-Bayesian
M1	30	347	.2313	.2313	(5.6,18.6)	(.0524,.4102) $\delta=.3578$	(-.0757,.5384) $\delta=.6141$
M2	24	131	.1092	.1098	(1.0,1.0)	(-.0231,.2415) $\delta=.2646$	(-.4410,.6607) $\delta=1.102$
M3	40	218	.1090	.1090	(118.6,970)	(-.0232,.2412) $\delta=.2644$	(-.0262,.2442) $\delta=.2704$

Montgomery describes the first sample as a "warm-up" for the process and so the results are markedly different from the other two samples. The usual estimate ($\hat{p}$) and Bayes estimate (p^*) are comparable for all three samples. Very different values are observed for our empirical prior parameter estimates. M2 produces negative values for the estimates so the uniform noninformative settings are used. M3 produces extraordinarily large values.

The standard p-chart 3-sigma limits are given along with the semi-Bayes

alternative and the respective interval length δ. The Bayes limits are much wider, as noticed by Calvin. The negative lower limits are problematic. Montgomery, as other authors, suggest that a rescreening of the data for outliers would produce an interval with positive endpoints, but it is clear that if the true proportion of nonconforming items is small ($<.25$), then the control chart lower limit will be negative. The standard fixup in practical settings is to use zero as the lower limit and ignore the problem.

In comparing the four methods one is actually comparing three distributions: the predictive, binomial and posterior distribution which form the basis of the four methods. Figure 7.3 depicts the three distributions for the Montgomery M1 sample. The distributions are depicted as continuous for comparison sake, but it should be noted that the predictive is definitely discrete whereas the binomial graph could be thought of as the normal approximation. Most striking is the marked kurtosis of the posterior distribution. This is mainly due to good information, both from the empirical prior and the moderate sample size. Obviously any interval based on the posterior distribution will be smaller than its counterparts from the other methods. The predictive distribution shows variability similar to that of the binomial, but shares the same measure of centrality that the posterior does. As with any predictive distribution, it by definition must be wider than the parent distribution upon which it is based.

To compare the four methods it is necessary to set a common k-sigma level. This is accomplished by first noting that of the four methods, only the predictive interval is restricted to working with a discrete distribution. One cannot, therefore set an arbitrary significance level and compare the four methods. Instead, an arbitrary level must be set and the procedure must come as close as possible to that level for the discrete predictive

Figure 7.3

COMPARISON OF P DISTRIBUTIONS
PREDICTIVE VS BINOMIAL VS POSTERIOR
RESULTS FOR GENERATED SAMPLE M3
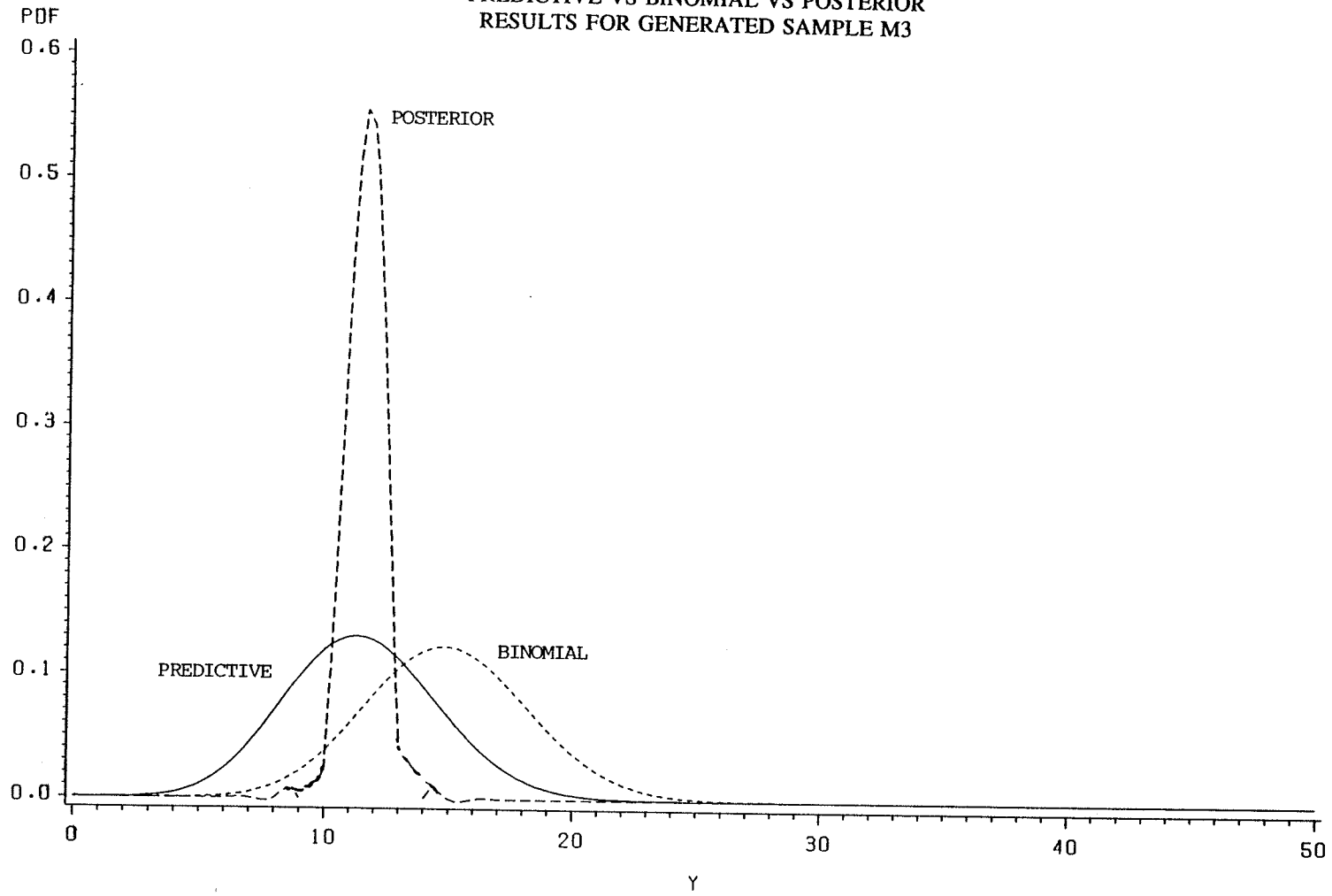
distribution. The observed significance level is then taken for the predictive interval and intervals are constructed using the same level for the other three methods. For the classical and semi-Bayes intervals this means merely changing the percentile of the standard normal variate. For the HPD interval, it merely requires a different target value for the numerical search routine. In all cases, intervals are constructed covering 95% of the distribution. In Table 7.2 these comparable intervals are presented.

Table 7.2: Comparable Intervals For Montgomery Samples

Set	Classical	Semi-Bayesian	Predictive	HPD
M1	(.1024,.3603) δ=.2579	(.0100,.4527) δ=.4427	(.1200,.3600) δ=.2400	(.2071,.2561) δ=.0490
M2	(.0201,.1982) δ=.1781	(-.2610,.4806) δ=.7416	(.0400,.2000) δ=.1600	(.0908,.1296) δ=.0388
M3	(.0189,.1991) δ=.1802	(.0169,.2011) δ=.1842	(.0400,.2000) δ=.1600	(.0976,.1206) δ=.0230

The semi-Bayes intervals are clearly the poorest estimates. The predictive intervals, although suffering from their discrete nature in that the same interval is produced for M2 and M3 are smaller than their classical counterparts. As expected, however, the HPD intervals zero in on p with much greater accuracy than any of the other methods.

The impact of the empirical prior parameter estimates can be seen by comparing the M2 and M3 HPD intervals. The point estimates are almost identical and both samples contain considerable information to the extent that one would expect it to dominate prior information. M2 suffers, however, relative to M3 because by using the empirical estimates instead of the noninformative values, the M3 interval is noticeably smaller, as desired.

The predictive interval and HPD interval approaches do not suffer from the

negative lower limit problem. Although the predictive lower limit can become zero due to the discreteness of the underlying distribution, it is still a valid zero and not an arbitrary bandaid to cover up for the technique's failing, such as is the case with the classical interval. The HPD interval cannot attain a zero lower limit as it uses numerical integration across a continuous distribution in p.

The results for these three samples are convincing. The Bayes methodology was proposed, however, to examine the impact on small samples. Because it is well known that the normal approximation to the binomial becomes poor as n decreases, it was expected that a Bayes exact approach would produce more accurate results. As has been seen above, the domination of the empirical prior estimates is impressive even at moderate sample sizes.

A number of samples were generated from a parent binomial process and the four methods applied. Table 7.3 below presents the specifications of four such samples which will form the basis for our discussion.

Table 7.3: Specifications For Generated Samples

Set	r	t	$\hat{p}$	p^*	(l^*, m^*)	Classical	Semi-Bayesian
G1	5	19	.3800	.3846	(-4.1,-6.7)	(-.0805,.8405) $\delta=.9210$	(-.536,1.315) $\delta=1.8513$
G2	10	47	.3133	.3133	(41.4,90.8)	(-.0460,.6726) $\delta=.7186$	(-.064,.6910) $\delta=.7554$
G3	6	6	.0667	.0667	(3.2,44.8)	(-.1266,.2599) $\delta=.3865$	(-.152,.2858) $\delta=.4378$
G4	5	17	.3400	.3400	(6.2,12.0)	(-.2050,.8850) $\delta=1.090$	(-.560,1.060) $\delta=1.6203$

G1 and G4 have n=10 while G2 and G3 have n=15. G1 is interesting for comparison

with the other three samples due to the negative prior empirical estimates. Note that once again the noninformative uniform prior is substituted for the negative values. The comments regarding the estimators of p and the usual 3-sigma limits given for the Montgomery samples above remain the same.

Figures 7.4 and 7.5 depict the three distributions mentioned previously for the G1 and G3 samples respectively. This comparison is interesting because G1 uses the noninformative prior whereas G3 uses the empirical estimates. It is notable that the posterior curve is still much narrower in the noninformative case, though it is not as impressive as in the G3 case. Figures for G2 and G4 provide similar information and are therefore omitted.

Table 7.4 presents the comparable intervals of the four techniques, again using a level of 95%. As before, the actual observed level of significance is not exactly 95% because of discreteness, but the four intervals are all constructed at the observed level so that they are directly comparable.

Table 7.4: Comparable Intervals For Generated Samples

Set	Classical	Semi-Bayesian	Predictive	HPD
G1	(.0430,.7170) $\delta=.6740$	(-.2895,1.059) $\delta=1.3485$	(.1000,.7000) $\delta=.6000$	(.2421,.5325) $\delta=.2904$
G2	(.0493,.5773) $\delta=.5280$	(.0358,.5909) $\delta=.5551$	(.0667,.5333) $\delta=.4666$	(.2535,.3748) $\delta=.1213$
G3	(-.0825,.2159) $\delta=.2984$	(-.1025,.2358) $\delta=.3383$	(.0000,.2000) $\delta=.2000$	(.0241,.1194) $\delta=.0953$
G4	(.0106,.6094) $\delta=.5988$	(-.0594,.7394) $\delta=.7988$	(.1000,.7000) $\delta=.6000$	(.2187,.4670) $\delta=.2483$

Again it is seen that the prior information in G4 produces a smaller interval than in G1. The G3 results indicate that the most reasonable interval is likely the HPD as it is the

THERE IS NO PAGE 276

Figure 7.4

COMPARISON OF P DISTRIBUTIONS
PREDICTIVE VS BINOMIAL VS POSTERIOR
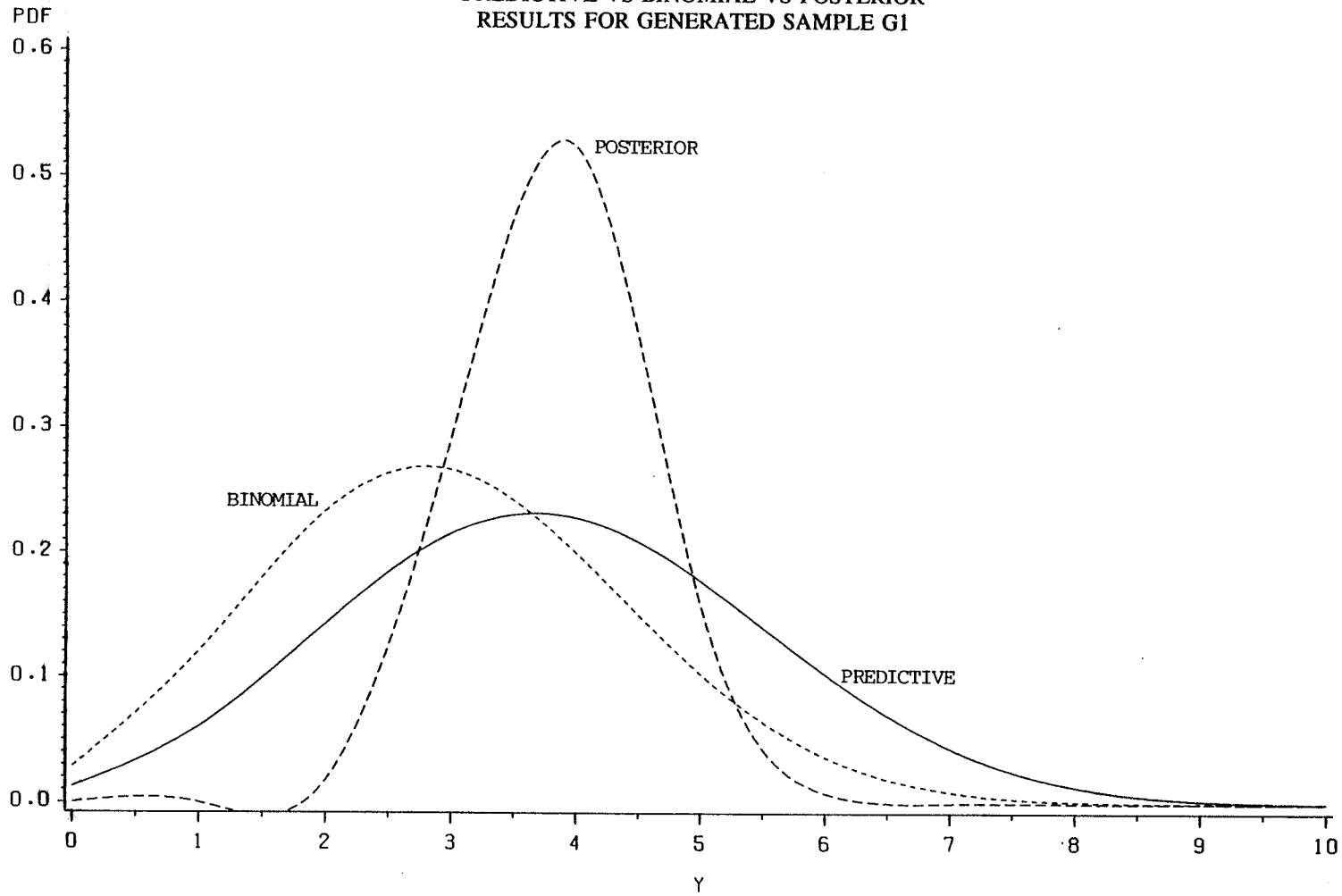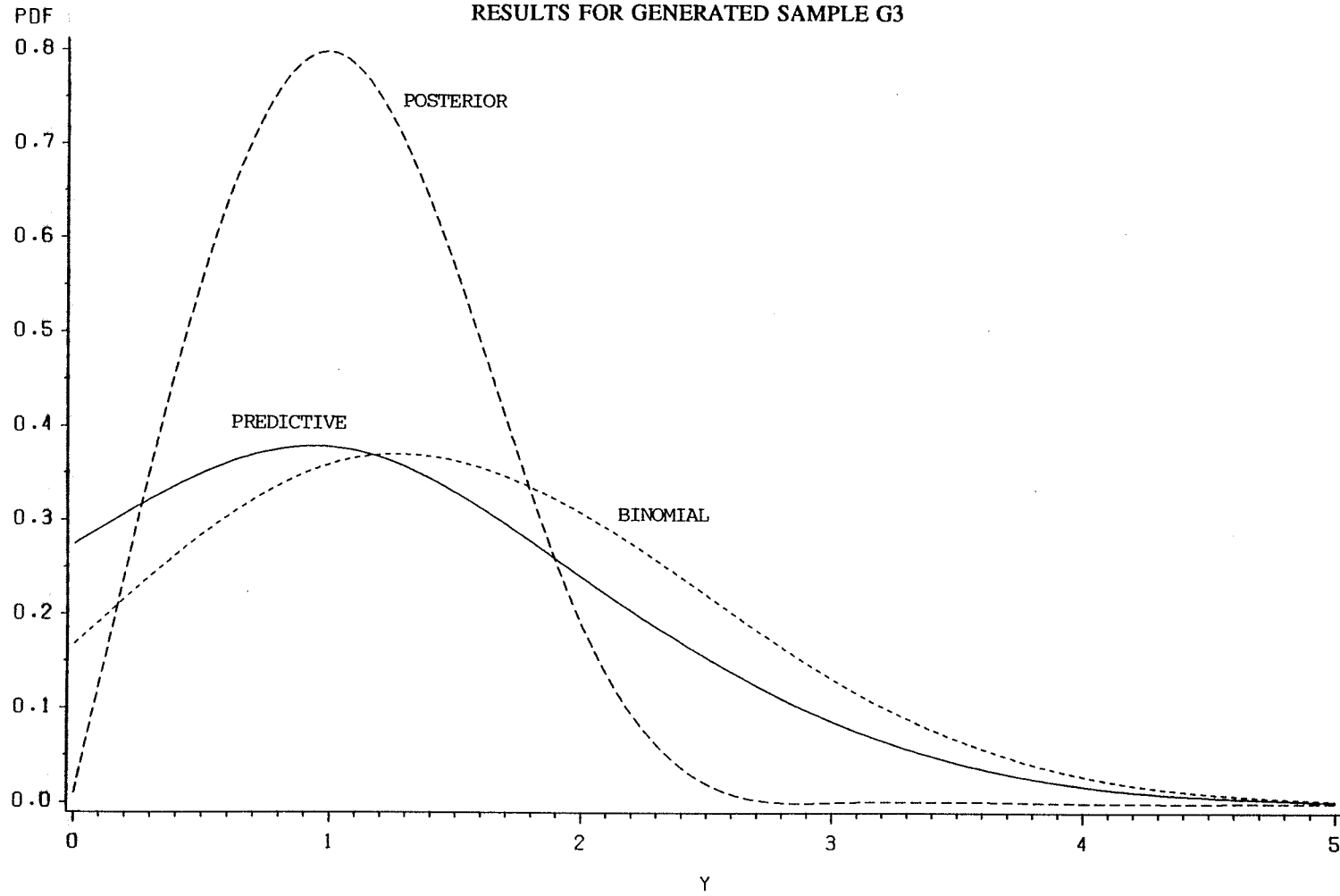### RESULTS FOR GENERATED SAMPLE G1

Figure 7.5

COMPARISON OF P DISTRIBUTIONS
PREDICTIVE VS BINOMIAL VS POSTERIOR
RESULTS FOR GENERATED SAMPLE G3

only one that does not have zero as a member. However, the predictive interval does have the practical advantage for the discreteness of the model situation.

To gain further insight into the comparative performance of the four intervals, 1000 sampling runs consisting of ten lots drawn from a B(10,0.3) population were generated. Intervals for each of the four methods were then constructed as before and summary statistics created for the run of 1000 samplings. The average and standard deviation for the length of the resultant intervals are given in Table 7.5.

Table 7.5: Interval Comparison For 1000 Generated Samples

Interval	Avg length	s.d. length	max upper bound
Classical	.689	.051	.837
Semi-Bayes	1.125	.307	1.2454
Predictive	.574	.044	.800
HPD	.197	.038	.561

The order of performance remains as it was for the single sample results. The HPD interval is by far the shortest, while the predictive interval is consistently shorter than its classical counterpart. The semi-Bayes interval performs terribly in comparison. In terms of coverage of the true proportion defective, the classical, semi-Bayes and predictive intervals covered the true value of the proportion nonconforming units for all of the 1000 sampling cycles. The HPD intervals covered the true p 95% of the time, which is to be expected.

Several points of interest were uncovered via the simulation. As mentioned before, when the lot variance of defectives is high relative to the average lot defectives, the estimators for the prior parameters l^* and m^* become negative, causing the traditional noninformative prior to be used. This will happen with increasing frequency as n and

p decrease. In our simulation 55.9% of the samples produced negative prior parameter estimates. The average length of the intervals where positive prior estimates were obtained produced an average interval length of 0.570, noticeably smaller than the full 1000 sample result of 0.574. The positive prior parameter estimates were as large as $(\ell^*, m^*) = (3544, 7531)$, although more than half of the samples with positive prior estimates produced $(\ell^*, m^*) < (10, 10)$. The higher the prior parameter estimates, the shorter the resultant predictive and HPD intervals. Negative lower bounds were evidenced in 34.5% of the classical intervals and 87.6% of the semi-Bayes intervals respectively.

7.2.6 Summary

Two new approaches were presented to attribute sampling control limits that are definite improvements on the classical approach. The predictive interval approach maintains the discreteness of the model situation while producing smaller interval estimates than the classical p-chart at the same coverage level. Using the posterior distribution to produce an HPD interval incorporates all sources of information to produce markedly smaller intervals.

The estimation of prior parameters by empirical methods is a definite asset. Even with the possibility of negative estimates, the approach does no worse than simply using a state of ignorance about the parameters. Even in the face of considerable information from the data, the empirical prior estimates have an impact on the resultant interval estimators for p.

7.3 Predictive p-Charts

The intrinsic advantage that Bayesian methods have over classical techniques is the constant updating of model parameters through conditional probability. This allowed for the construction of the predictive distribution for a future lot proportion defective. This information would be useful to incorporate into the traditional p-chart as a supplementary indicator of process control. In this section such a chart is proposed, which will be referred to as a predictive p-chart.

7.3.1 Construction of the Predictive p-Chart

As with the previous section's work, this method will be most easily applied to situations in which a presample is observed on a process to give foundation figures for the construction of control limits. Once these limits have been established, it is proposed to monitor the probability that the next lot will produce an observed proportion of defectives that falls outside the limits. This predictive probability is updated after each sampling and superimposed on the traditional p-chart. In this way, a trend for an upcoming problem in the process control can be spotted more readily than through the use of a p-chart alone. This holds important implications for practitioners in that the sooner a problem is detected, the lower the cost of the problem.

The method of construction is straightforward:

1) Construct the traditional p-chart limits as usual.

2) Use the beta-binomial predictive distribution of the previous section to calculate the probability that the next sample will produce a proportion of defectives outside the p-chart limits.

3) Repeat step 2 before each sampling and plot the probability on a separate y-axis superimposed on the traditional p-chart.

Note that although the traditional p-chart 3-σ limits are used for demonstration purposes, any of the alternative limits proposed in the previous section could be used without loss of generality. Furthermore, 2-σ limits could be used as well if desired. The technique is robust with respect to the limits used.

Figure 7.6 provides an example of the predictive p-chart. Data was generated for an artificial B(10,0.1) process in control (i.e. stable). The figure can be read as a typical p-chart if the left y-axis is used and the asterisk points are ignored. The right-hand axis and asterisk points represent the scale and observed values for the probability that the next sample from the process will produce an observed level of defectives outside the traditional 3-σ process control limits. The limits were established by a presample of 15 lots to be (LCL,UCL)=(-.1826,.3693). By convention, since the lower control limit is negative and therefore meaningless, the zero horizontal is used in its place.

Some mention of the right-hand scaling technicalities must be made. The reader will notice that the scale is not uniform. The method of scaling is such that the right hand axis values may be ignored below the upper control limit. The remainder of the scale is adjusted so that the range of the observed predictive probability falls above the corresponding horizontal of the left-hand upper control limit. To make the graph readable, the right-hand y-axis is scaled so that the range of the predictive probabilities lie outside the graph space occupied for the traditional p-chart. As can be seen in Figure 7.6, the observed predictive probabilities fall above 0.01. This value is used as a starting point for plotting of the predictive probabilities so that all the predictive probability

Figure 7.6   PREDICTIVE P-CHARTS
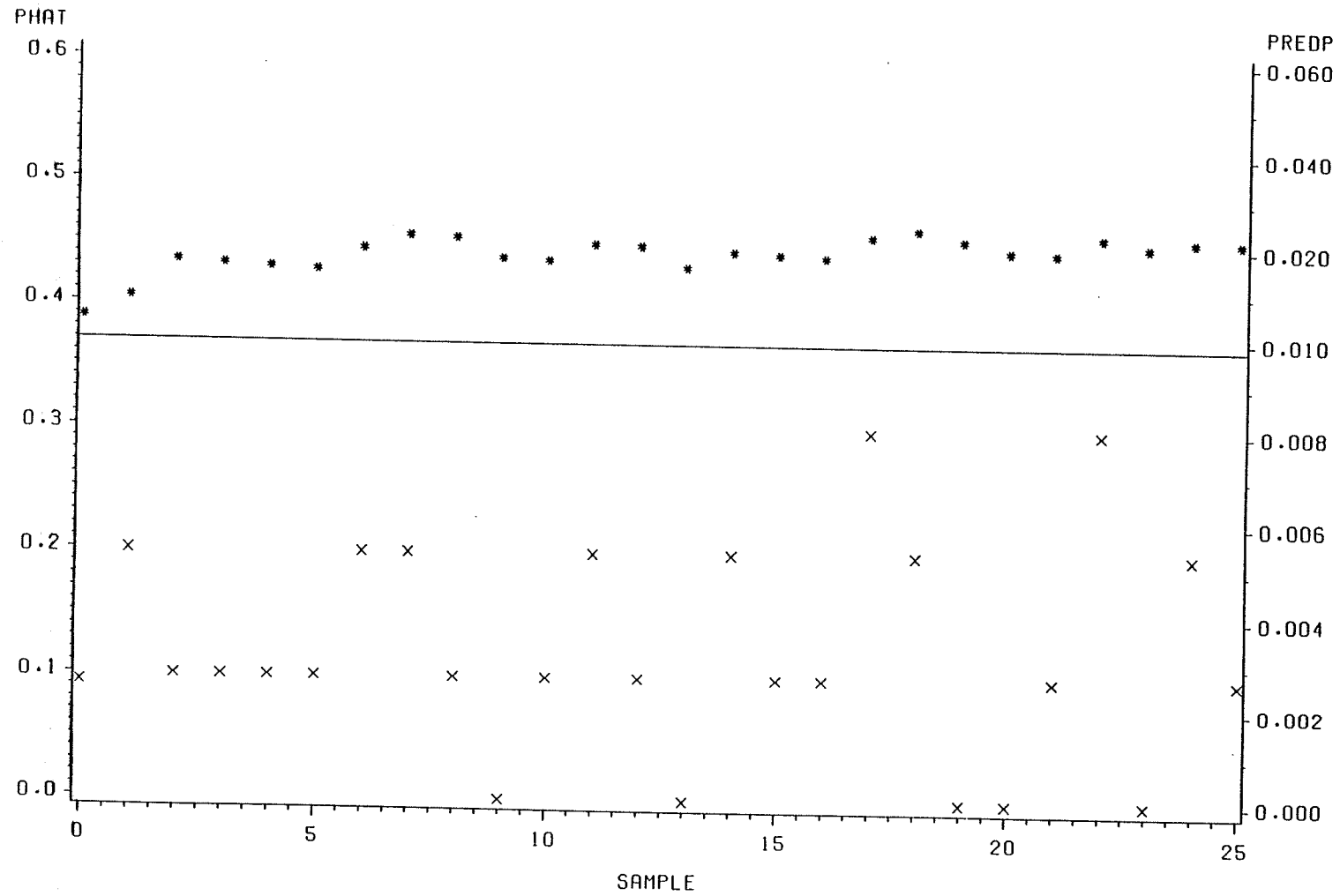
LEGEND: X - PROPORTION DEFECTIVE
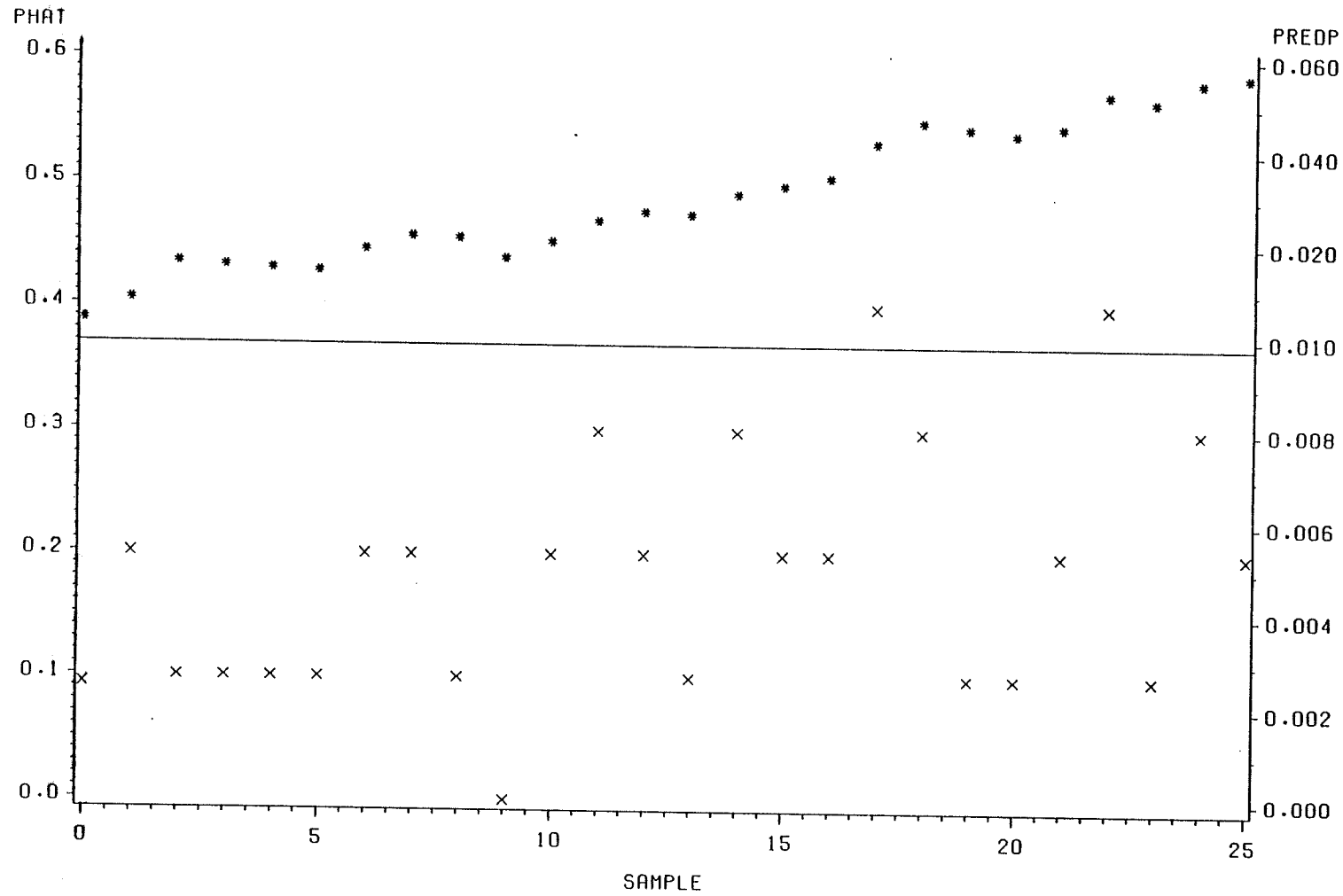   * - PR(NEXT SAMPLE OUT OF CONTROL)

points lie above the traditional p-chart's plotting area. Traditional p-chart points that fall outside the control limits are made more noticeable as they will fall into the space of the predictive p-chart. The predictive p-chart y-axis for Figure 7.6 is wider than the observed data for this example only so that comparison against the results in Figure 7.7 can be made. This comparison is interesting because Figure 7.6 represents a process in control whereas Figure 7.7 portrays a process that runs out of control after the tenth sample.

It may be argued that the inclusion of a right-hand scale and second set of points may be confusing to some practitioners. For such situations, it would be just as easy to produce a separate plot of the predictive probabilities. It is preferable, however, that the information contained in the predictive distribution be weighed in conjunction with that of the traditional p-chart in the same way it is preferable to examine location and dispersion of a dataset. Combining both types of information into a parsimonious display, therefore, is desirable. Furthermore, the plot can easily be viewed as two separate entities. If one ignores the right-hand y-axis and predictive probability points(the *'s), what remains is the traditional p-chart. Looking only at the plot above the upper control limit and reading the right-hand y-axis produces a picture of the predictive p-chart.

Following is a detailed description for the first two lots sampled. At time zero, information is available from the presample of fifteen lots of size ten that produced the traditional 3-σ limits. The prior parameters of the beta prior are estimated as in the previous section and produce the predictive distribution. Using this distribution the Pr(next sample falls outside 3-σ limits)=Pr(number of defectives in the next sample is

Figure 7.7   PREDICTIVE P-CHARTS

LEGEND: X - PROPORTION DEFECTIVE
          * - PR(NEXT SAMPLE OUT OF CONTROL)

less than 0 or greater than .3693) is calculated to be 0.0117. This value is plotted against the right-hand y-axis. The presample process average defectives is also plotted, represented by the X at time zero. A supplementary sample of ten items is then taken at time one and 20% defectives are observed. Using this information, the prior and predictive distributions are updated as per the previous section. The process is then repeated by calculating the probability of the next sample (now referring to time two) falling outside the 3-σ limits. This value turns out to be 0.0140, up from the time zero value due to the fact that the present percent lot defectives was above the process average. It is this intrinsic updating of information that is gained from the predictive p-chart. At time two, a further lot is taken and 10% observed defectives are found. Although this information indicates that the process is in control, the previous runs are not ignored by the predictive distribution and the predictive probability of the sample at time three falling outside the interval is 0.180. Only in subsequent lots when the lot quality level is maintained does the carryover effect of the present lot cause the predictive probability decrease. Hence the predictive p-chart is more than a mere echo of the present lot performance.

7.3.2 Examples

Figure 7.6 is a good exemplary plot for the technique, but not terribly exciting as far as process control is concerned. By both the traditional p-chart (the X's) and the predictive p-chart (the *'s) it is obvious that the process is in control, with random fluctuation being the only differentiating component. Although the predictive probability shows jumps and dives, the overall appearance is that of a white noise process.

Now that the basic construction of the predictive p-chart has been described, some illustrative examples are presented. Figure 7.7 takes the process seen in Figure 7.6 and adds a disruptive influence at time T=10. Again the predictive p-chart y-axes in Figures 7.6 and 7.7 are the same for easy comparison. This disruption quietly adds a 10% additional rate of defectives to the process. This causes the points at time T=17 and T=22 to fall outside the control limits, indicative of a process out of control. Evidence that the process has suffered shortly after time T=10 is much more obvious from the predictive p-chart values as represented by an almost monotone increasing plot. The plot is much more dynamic than the traditional p-chart and a practitioner would be more likely to spot the problem, which was introduced at time T=10, before the process produces a lot outside the control limits at time T=17.

For a real application, the Montgomery M1 dataset discussed in the previous section is presented in Figure 7.8. The first ten lots are used as the presample in this situation and the latter twenty lots are considered to be supplementary samples. The process clearly has control problems and in this case does not necessarily demand the use of the p-chart to see that a problem exists at time T=5. The predictive p-chart does however give a better overview of the carryover effect inherent in the process under the assumptions of the Bayes model. As stated above, the predictive p-chart is meant to supplement, not replace, the traditional p-chart.

To further drive home the point that the predictive p-chart can spot a trend before examination of the traditional p-chart, a specially "designed" observed process was constructed and is presented in Figure 7.9. A preliminary sample from a B(10,0.1) of fifteen lots was used to produce control limits of (LCL,UCL)=(-.1815,.3615). The run

Figure 7.8

PREDICTIVE P-CHARTS
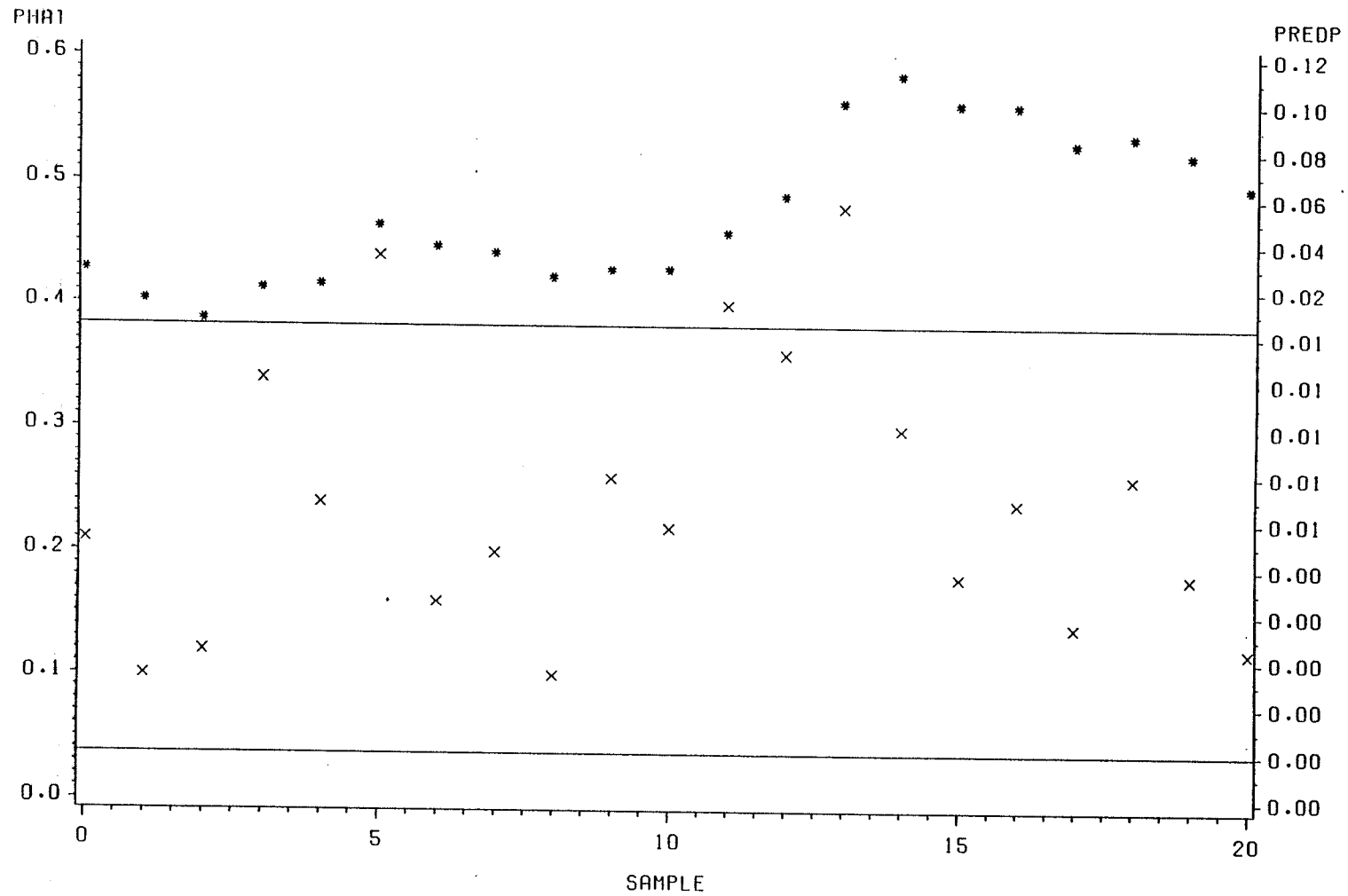
LEGEND: X – PROPORTION DEFECTIVE
         * – PR(NEXT SAMPLE OUT OF CONTROL)

of supplementary samples was then designed to remain in control by all traditional measures and to demonstrate an out-of-control point late in the process. The traditional p-chart of Figure 7.9 looks to be in control until time $T=18$ when a lot falls outside the 3-σ limits. The predictive p-chart, however, demonstrates that a problem has been brewing since time $T=0$ and would have caused the practitioner to raise the alarm much earlier.

7.3.3 Discussion

The reader will notice that no mention of precise stopping rules have been presented with regard to the predictive p-chart. This is due to the fact that other considerations such as cost factors and the importance/consequences of stopping the process must be incorporated into the setting of any such arbitrary alarm condition. Just as some practitioners use 2-σ limits instead of 3-σ limits, so will the actual point for stopping a process due to evidence provided by the predictive p-chart be a largely arbitrary and situation-specific decision.

The level of probability expected deserves some discussion. In all our examples, the probability of the next sample falling outside the 3-σ control limits is below 0.2. This is not surprising when one considers that the 3-σ limits are used. Under normal distribution theory, if the process is under control, the probability that an observation falls outside 3-σ limits is less than 0.003. Predictive probabilities are observed to be as large as 40 times this level in order of magnitude. It is for this reason that a great deal of forethought must be given to the exact criteria for stopping the process. Using a rule such as stopping the process if the probability doubles from one sample to the next may

PREDICTIVE P-CHARTS

Figure 7.9

PHAT

PREDP

LEGEND: X – PROPORTION DEFECTIVE
         * – PR(NEXT SAMPLE OUT OF CONTROL)

SAMPLE

290
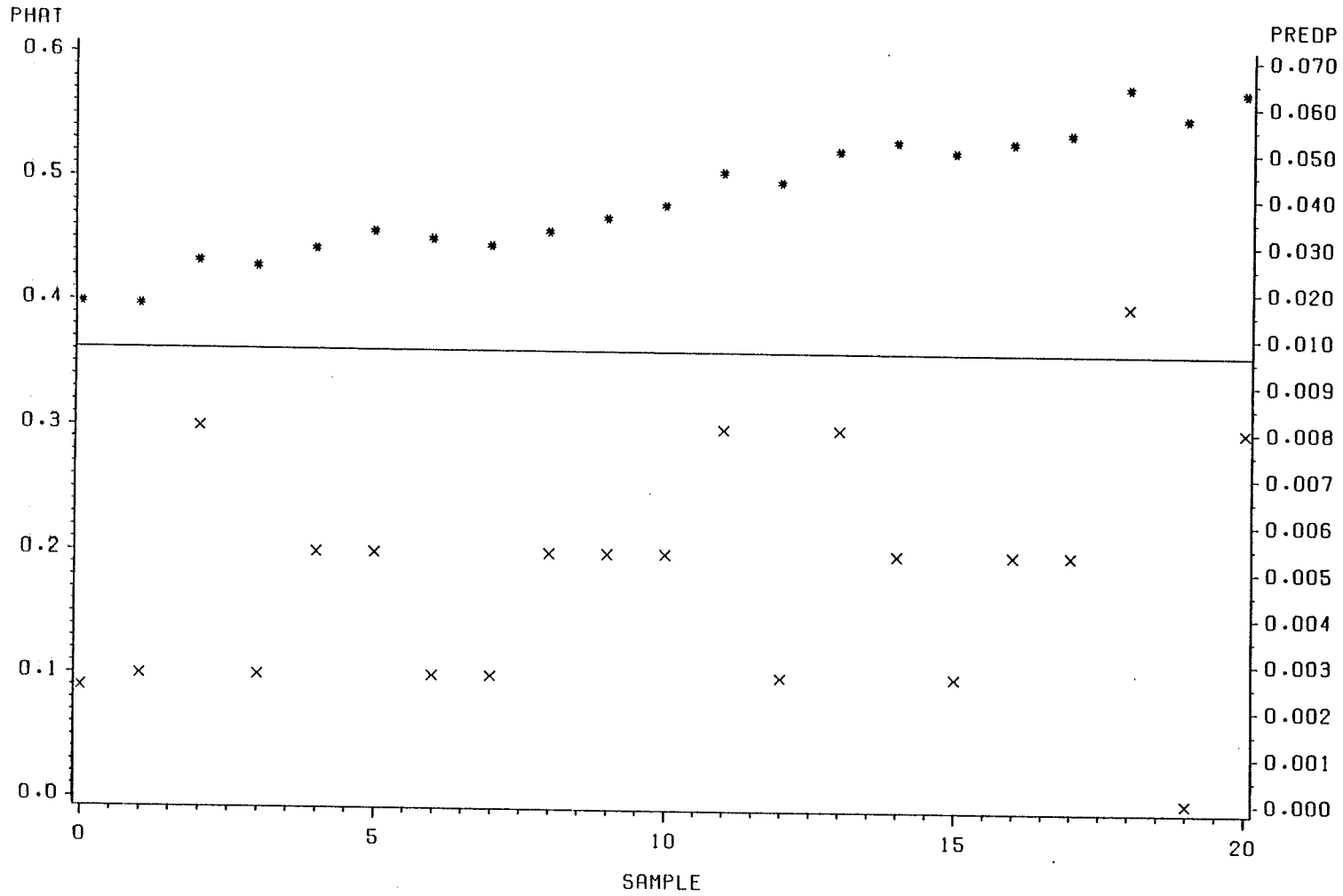
be too lenient. Demanding a stoppage after a 5% rise in the predictive probability may be too strict. Such decisions are certainly relevant and will no doubt be the source of great discussion. Because this discussion is best situated in the hands of the practitioner rather than the mathematician, speculation on such stopping rules is beyond the purview of the statistician. The important contribution of the technique is a parsimonious presentation of supplementary information to the practitioner which may allow for an earlier detection of a trend or a process going out of control. The exact definition of what will constitute an alarm situation is best left to quality control engineers.

It should also be pointed out that this approach has some relatives in the literature. The CUSUM chart of Shewhart(1931) for example includes the carryover effect inherent in the Bayesian underpinnings of the predictive p-chart. It does not however give probabilistic information directly. Hunter's(1986) exponentially weighted moving average (EWMA) control chart also incorporates the concept of updating process performance information from a time series approach. The EWMA chart, however, is more difficult to read than the predictive p-chart in that it overlaps two process results.

The required computation work to produce a predictive p-chart, while more involved than the traditional p-chart, is easily accomplished. The algorithms were coded in FORTRAN-77 and the plot produced via SAS/GRAPH on an AMDAHL 470 mainframe. Algorithms are available from the author upon request.

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. (1965), Handbook of Mathematical Functions, Dover Publications Inc., New York.

Adatia, A. and Chan, L. K. (1985), "Robust estimators of the three-parameter Weibull distribution", IEEE Transactions on Reliability 34, 39-43.

Anderson, T. W. (1958), An Introduction to Multivariate Statistical Analysis, Wiley, New York.

Anderson, T. W. and Darling, D. (1954), "A test of goodness of fit", Journal of the American Statistical Association 49, 765-769.

Aitchison, J. and Brown, J. (1957), The Lognormal Distribution, Cambridge University Press.

Amin, N. A. (1981), The Application of a New Method of Parameter Estimation to the Three-parameter Lognormal and Three-parameter Weibull distributions, unpublished Ph. D. thesis, University of Wales.

Archer, N. P. (1980), "A computational technique for maximum likelihood estimation with Weibull models", IEEE Transactions on Reliability 29, 57-62.

Bagchi, P. and Kadane, J. B. (1991), "Laplace approximations to posterior moments and marginal distributions on circles, spheres and cylinders", Canadian Journal of Statistics 19, 67-77.

Bain, L. J. (1978), Statistical Analysis of Reliability and Life-testing Models, Marcel Dekker, New York.

Basu, D. (1975), "Statistical information and likelihood", Sankhya Series A 37, 1-71.

Bayes, Rev. T. (1763), "An essay towards solving a problem in the doctrine of chances", Philosophical Transactions of the Royal Society 53, 370-418. Reprinted in Biometrika 45, 296-315.

Bannerjee, A. K. and Bhattacharyya, G. K. (1979), "Bayesian results for the inverse Gaussian distribution with an application", Technometrics 21, 247-251.

Berger, J. O. and Bernardo, J. M. (1989), "Estimating a product of means: Bayesian analysis with reference priors", Journal of the American Statistical Association 84, 200-207.

Berger, J. (1990), Comment on "The present position of Bayesian statistics", by D. V. Lindley, Statistical Science 5, 71-75.

Bhattacharya, S. K. (1967), "Bayesian approach to life testing and reliability estimation", Journal of the American Statistical Association 62, 48-62.

Blischke, W. R. (1974), "On non-regular estimation II: Estimation of the location parameter of the Gamma and Weibull distributions", Communications in Statistics 12, 1109-1129.

Bowman, K. O. and Shenton, L. R. (1988), Properties of Estimators for the Gamma Distribution, Marcel Dekker, New York.

Box, G. E. P. and Tiao, G. (1973), Bayesian Inference in Statistical Analysis, Addison-Wesley, New York.

Calitz, F. (1973), "Maximum likelihood estimation of the parameters of the three-parameter lognormal distribution - a reconsideration", Australian Journal of Statistics 15, 185-190.

Calvin, T. W. (1984), How and When to Perform Bayesian Acceptance Sampling, American Society for Quality Control, Milwaukee, Wisconsin.

Cheng, S. W. and Fu, J. C. (1982), "Estimation of mixed Weibull parameters in life testing", IEEE Transactions on Reliability 31, 377-381.

Cheng, S. W., Fu, J. C. and Sinha, S. K. (1985), "An empirical procedure for estimating the parameters of a mixed exponential life testing model", IEEE Transactions on Reliability 34, 60-64.

Chhikara, R. S. and Guttman, I. (1982), "Prediction limits for the inverse Gaussian distribution", Technometrics 24, 319-324.

Cohen, A. C. (1951), "Estimating parameters of logarithmic-normal distributions by maximum likelihood", Journal of the American Statistical Association 46, 206-212.

Cohen, A. C. (1965), "Maximum likelihood estimation in the Weibull distribution based on complete and censored samples", Technometrics 5, 579-588.

Cohen, A. C. and Whitten, B. (1980), "Estimation in the three-parameter Lognormal distribution", Journal of the American Statistical Association 75, 399-404.

Cohen, A. C. and Whitten, B. (1982a), "Modified maximum likelihood and modified moment estimators for the three-parameter Weibull distribution", Communications

in Statistics 11, 2631-2656.

Cohen, A. C. and Whitten, B. (1982b), "Modified maximum likelihood and modified moment estimators for the three-parameter Gamma distribution", Communication in Statistics 11, 197-216.

Cohen, A. C., Whitten, B. and Ding, Y. (1984), "Modified moment estimation for the three-parameter Weibull distribution", Journal of Quality Technology 16, 159-167.

Cohen, A. C. and Whitten, B. J. (1988), Parameter Estimation in Reliability and Life-Span Models, Marcel Dekker, New York.

Cramer, H. (1946), Mathematical Methods of Statistics, Princeton University Press.

Crow, E. L. and Shimizu, K. (1988), Lognormal Distributions: Theory and Applications, Marcel Dekker Inc., New York.

D'Agostino, R. B. and Stephens, M. B. (ed.), (1986), Goodness of Fit Techniques, Marcel Dekker, New York.

Dawid, A., Stone, M. and Zidek, J. (1973), "Marginalization paradoxes in Bayesian and structured inference", Journal of the Royal Statistical Society series B 35, 189-233.

DeBruijn, N. (1961), Asymptotic Methods in Analysis, Dover Publications Inc.

Dodge, H. F. and Roming, H. G. (1959), Sampling Inspection Tables, John Wiley, New York.

Duncan, A. J. (1974), Quality Control and Industrial Statistics, 4th ed., Irwin, Homewood, Illinois.

Dubey, S. D. (1967), "Some percentile estimators for Weibull parameters", Technometrics 9, 119-129.

Dumonceaux, R. and Antle, C. E. (1973), "Discrimination between the lognormal and the Weibull distributions", Technometrics 15, 923-926.

Elderton, W. and Johnson, N. (1969), Systems of Frequency Curves, Cambridge University Press.

Efron, B. et al (1986), "Why isn't everyone a Bayesian", American Statistician 40, 1-11.

Englehardt, M. and Bain, L. J. (1979), "Prediction limits and two-sample problems with complete or censored Weibull data", Technometrics 21,

233-237.

Englehardt, M. and Lee, J. (1979), "Prediction limits and two sample problems with complete or censored Weibull data", Technometrics 21, 233-237.

Finney, D. J. (1941), "On the distribution of a variate whose logarithm is normally distributed", Journal of the Royal Statistical Society, Series B 7,   155-161.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling based approaches to calculating marginal densities", Journal of the American Statistical Association 85, 398-409.

Giesbrecht, F. and Kempthorne, O. (1976), "Maximum likelihood estimation with the three-parameter Lognormal", Journal of the Royal Statistical Society, Series B 38, 257-264.

Haan, C. T. and Beer, C. E. (1967), "Determination of maximum likelihood estimation for the three-parameter Weibull distribution", Iowa State Journal of Science 42, 37-42.

Hager, H. and Bain, L. (1970), "Inferential procedures for the generalized Gamma distribution", Journal of the American Statistical Society 65, 1601-1609.

Hald, A. (1981), Statistical Theory of Sampling Inspection by Attributes, Academic Press, New York.

Harter, H. L. (1967), "Maximum likelihood estimation of the four-parameter generalized Gamma population from complete and censored samples", Technometrics 9, 159-165.

Harter, H. L. (1969), Order Statistics and Their Use in Testing and Estimation, Aerospace Research Laboratories.

Harter, H. L. (1971), "Some optimization problems in parameter estimation", in Optimizing Methods in Statistics, Academic Press.

Harter, H. L. and Moore, A. (1965), "Maximum likelihood estimation of the parameters of Gamma and Weibull populations from complete and censored samples", Technometrics 7, 639-643.

Harter, H. L. and Moore, A. (1966), "Local maximum likelihood estimation of the parameters of the three-parameter Lognormal population from complete and censored samples", Journal of the American Statistical Association 61, 842-851.

Hartigan, S. A. (1983), Bayes Theory, Springer-Verlag, New York.

Hill, B. M. (1963), "The three-parameter Lognormal distribution and Bayesian analysis of a point-source epidemic", Journal of the American Statistical Association 58, 72-84.

Howlader, H. (1982), "On the study of the posteriors of parameters of some well known distributions under proper and improper priors", unpublished Ph. D. thesis, University of Manitoba.

Howlader, H. and Weiss, G. (1987a), "On Bayesian estimation of Cauchy parameters", to appear in Sankhya series A.

Howlader, H. and Weiss, G. (1987b), "Bayesian reliability estimation of a two-parameter Cauchy distribution", to appear in Biometrical Journal of the Institut fur Mathematik der Akademie der Wissenschaften der DDR.

Howlader, H. and Weiss, G. (1987c), "Considerations on the approximation of Bayesian ratios of integrals", unpublished manuscript.

Hunter, J. S. (1986), "The exponentially weighted moving average", Journal of Quality Technology 18, 203-210.

IMSL (1975), International Mathematical and Statistical Library Manual, IMSL Inc., Houston.

Jeffreys, H. (1983), Theory of Probability, Clarendon Press, Oxford.

Johnson, N. and Kotz, S. (1970), Continuous Univariate Distributions, Houghton Mifflin Company, Boston.

Johnson, R. and Haskell, J. (1983), "Sampling properties of estimators of a Weibull distribution of use in the lumber industry", Journal of the Statistical Society of Canada 11, 155-169.

Johnson, R. A. (1967), "An asymptotic expansion for posterior distributions", Annals of Mathematical Statistics 38, 1899-1907.

Johnson, R. A. (1970), "Asymptotic expansions associated with posterior distributions", Annals of Mathematical Statistics 41, 851-864.

Kappenman, R. F. (1985), "Estimation for the three-parameter Weibull, lognormal and gamma distributions", Computational Statistics and Data Analysis 3, 11-23.

Kass, R., Tierney, L. and Kadane, J. B. (1988), "Asymptotics in Bayesian computation", In Bayesian Statistics 3 (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 261-278, Oxford University

Press.

Kass, R., Tierney, L. and Kadane, J. B. (1990), "The validity of posterior expansions based on Laplace's method", Essays in Honor of George Barnard (S. Geisser, J. Hodges, J. Press and A. Zellner, eds.), North Holland 473-488.

Kendall, M. G. and Stuart, A. (1973), The Advanced Theory of Statistics, Volume 2, Hafner Publishing, New York.

Kennedy, J. and Gentle, G. (1980), Statistical Computing, Marcel Dekker, New York.

Laplace, P. S. (1776), "Recherches sur l'integration des equations differentielles aux differences finies, et sur leur usage dans la theorie des hasards", Memoires de mathematique et de physique presentes a l'Academie royale des sciences, par divers savants, et lus dans ses assemblees, 37-162. Reprinted in Laplace's Oeuvres completes 8, 69-197. Translated by S. Stigler, (1986), Statistical Science 1, 364-378.

Lawless, J. F. (1977), "Prediction intervals for the two-parameter exponential distribution", Technometrics 16, 241-244.

Lawless, J. F. (1982), Statistical Models and Methods for Lifetime Data, John Wiley, New York.

LeCam, L. (1970), "On the assumptions used to prove asymptotic normality of maximum likelihood estimates", Annals of Mathematical Statistics 41, 802-828.

Lemon, G. H. (1974), "Maximum likelihood estimation for the three-parameter Weibull distribution based on censored samples", Technical Report of the Convair Aerospace Division of General Dynamics.

Leonard, T., Hsu, J. S. J. and Tsui, K. (1989), "Bayesian marginal inference", Journal of the American Statistical Association 84, 1051-1058.

Lindley, D. V. (1965), Introduction to Probability and Statistics from a Bayesian Viewpoint: Volume 2, Inference, Cambridge University Press.

Lindley, D. V. (1975), "The future of Statistics - A Bayesian 21st century", Supplement to Advances in Applied Probability 7, 106-115.

Lindley, D. V. (1980), "Approximate bayesian methods (with discussants)", Trabajos de Estadistica y de Investigacion Operativa 31, 232-245.

Lindley, D. V. (1990), "The 1988 Wald memorial lectures: The present position

in Bayesian statistics (with discussants)", Statistical Science 1, 44-89.

Lye, L. M., Sinha, S. K. and Booy, C. (1988), "Bayesian analysis of the T-year events for flood data fitted by a three-parameter lognormal distribution", Civil Engineering Systems 5, 81-86.

Marsaglia, G. (1961), "Expressing a random variable in terms of uniform random variables", Annals of Mathematical Statistics 32, 894-898.

McCool, J. I. (1975), Inferential Techniques for Weibull Populations II, Wright Patterson Air Force Base, Ohio. ARL Technical Report ARL-75-0233.

McCool, J. I. (1979), "Analysis of single classification experiments based on censored samples from the two-parameter Weibull distribution", Journal of Statistical Planning and Inference 3, 39-68.

Mendenhall, W. and Hader, R. J. (1958), "Estimation of the parameters of mixed exponentially distributed failure time distributions from censored life test data", Biometrika 45, 504-519.

MIL-STD-105D. (1963), Sampling Procedures and Tables for Inspection by Attributes, Government Printing Office, Washington, D.C.

Montgomery, D. C. (1985), Introduction to Statistical Quality Control, John Wiley, New York.

Naylor, J. C. (1982a), Some Numerical Aspects of Bayesian Inference, unpublished PhD thesis, University of Nottingham.

Naylor, J. C. and Smith, A. (1982b), "Applications of a method for the efficient computation of posterior distributions", Applied Statistics 31, 214-225.

Naylor, J. C. and Smith, A. (1983), "A contamination model in clinical chemistry: an illustration of a method for the efficient computation of posterior distributions", The Statistician 32, 82-87.

Padget, W. J. (1988) Chapter Six: Bayesian Estimation in Crow, E. L. and Shimizu, K. (1988), Lognormal Distributions: Theory and Applications, Marcel Dekker Inc., New York.

Parr, V. and Webster, J. (1965), "A method for disriminating between failure density functions used in reliability predictions", Technometrics 7, 1-10.

Pearson, E. S. and Hartley, H. O. (1966), Biometrika Tables for Statisticians, Volume I, Cambridge University Press.

Pope, J., Lehrer, B. and Stevens, J.P. (1980), "A multiphasic reading screening procedure", Journal of Learning Disabilities 13, 98-102.

Press, S. J. (1989), Bayesian Statistics: Principles, Models and Applications, John Wiley, New York.

Protter, M. H. and Morrey, C. B. (1964), Modern Mathematical Analysis, Addison-Wesley, Don Mills.

Raiffa, H. and Schlaifer, R. (1961), Applied Statistical Decision Theory, Harvard University Press.

Reilly, P. (1976), "The numerical computation of posterior distributions in Bayesian statistical inference", Applied Statistics 25, 201-209.

Rice, J. R. (1983), Numerical Methods, Software and Analysis, McGraw-Hill Inc.

Richards, F. S. (1961), "A method of maximum likelihood estimation", Journal of the Royal Statistical Society series B 23, 469-475.

Robbins, H. (1964), "An empirical Bayes approach to statistical problems", Annals of Mathematical Statistics 35, 1-35.

Rockette, H., Antle, C. and Klimko, L. (1974), "Maximum likelihood estimation with the Weibull model", Journal of the American Statistical Association 69, 246-249.

Shapiro, S. S. and Wilk, M. B. (1965), "An analysis of variance test for normality (complete samples)", Biometrika 52, 591-611.

Shewhart, W. A. (1931), Economic Control of Quality, D. Van Nostrand Co., New York.

Sinha, S. K. (1983), "Bayesian estimates ofthe parameters and reliability function from mixed exponentially distributed time-censored life test data", Communications in Statistics 12, 141-151.

Sinha, S. K. (1986), "Bayesian estimation of the reliability function of the inverse Gaussian distribution", Statistics and Probability Letters 4, 319-323.

Sinha, S. K. (1986b), Reliability and Life Testing, Wiley Eastern Ltd, New York.

Sinha, S. K. (1987), "Efficiency of Bayes estimators based on Lindley's approximation", unpublished manuscript.

Sinha, S. K. (1987b), "Bayesian estimation of the parameters and reliability function of a mixture of Weibull life distributions", Journal of Statistical

Planning and Inference 16, 377-387.

Sinha, S. K. and Sloan, J. A. (1988), "Bayes estimation of the parameters and reliability function of the three-parameter Weibull distribution", IEEE Transactions on Reliability 37, 364-369.

Sinha, S. K. and Sloan, J. A. (1989), "Prediction intervals for a mixture of Weibull failure-time distributions - a Bayesian approach", South African Journal of Statistics 23, 119-130.

Sloan, J. A. and Sinha, S. K. (1991), "Bayesian prediction intervals for a mixture of exponential failure-time distributions for complete and censored samples", Statistics and Probability Letters vol. 2 #6 June(to appear).

Smith, A. F. M. (1986), "Some Bayesian thoughts on modelling and model choice", The Statistician 35, 97-102.

Smith, A. F. M. et al (1985), "The implementation of the Bayesian paradigm", Communications in Statistics 14, 1079-1102.

Smith, A. F. M. et al (1987), "Progress with numerical and graphical methods for practical Bayesian statistics", The Statistician 36, 75-82.

Smith, R. (1985), "Maximum likelihood estimation in a class of nonregular cases", Biometrika 72, 67-90.

Stacey, E. and Mihram, G. (1965), "Parameter estimation for a generalized gamma distribution", Technometrics 7, 349-357.

Steen, P. J. and Stickler, D. J. (1976), "A sewage pollution study of beaches from Cardiff to Ogmore", Department of Applied Biology Internal Report, University of Wales, Cardiff.

Stigler, S. M. (1986), "Laplace's memoir on inverse probability", Statistical Science 1, 359-362.

Thoman, D. R., Bain, L. J. and Antle, C. E. (1970), "Maximum likelihood estimation, exact confidence intervals for reliability, and tolerance limits in the Weibull distribution", Technometrics 12, 363-371.

Tierney, L. and Kadane, J. B. (1986), "Accurate approximations for posterior moments and marginals", Journal of the American Statistical Association 81, 82-86.

Tierney, L., Kass, R. and Kadane, J. B. (1989), "Fully exponential Laplace approximation of expectations and variances of non-positive functions", Journal of the American Statistical Association 84, 710-716.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1988), Statistical Analysis of Finite Mixture Distributions, Wiley, New York.

Van Dijk, H. K., Hop, J. P. and Louter, A. S. (1987), "An algorithm for the computation of posterior moments and densities using simple importance sampling", The Statistician 36, 83-90.

Walker, A. (1969), "On the asymptotic behavior of posterior distributions", Journal of the Royal Statistical Society series B 31, 80-88.

Weibull, W. (1939), "A statistical theory of the strength of material", Ingeniors Vetenskaps Akademiens Handligar 151, Stockholm.

Weiss, G. (1988), "Invariant Bayes estimators", Proceedings of the Interface of Computers and Statistics 17, 232-235.

Wingo, D. R. (1972), "Maximum likelihood estimation of the parameters of the Weibull distribution by modified quasilinearization", IEEE Transactions on Reliability 21, 89-93.

Zanakis, S. (1977), "Computational experience with some nonlinear optimization algorithms in deriving maximum likelihood estimates for the three-parameter Weibull distribution", Algorithmic Methods in Probability 7, 63-77.

Zanakis, S. (1979a), "A simulation study of some simple estimators of the three-parameter Weibull distribution", Statistics and Computer Simulation 9, 101-116.

Zanakis, S. (1979b), "Extended pattern search with transformations for the three-parameter Weibull problem", Management Science 25, 1149-1161.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions", in Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, P. Goel and A. Zellner, eds., North-Holland, New York.