USE OF THE LOGNORMAL DISTRIBTUON FOR SURVIVAL DATA: INFERENCE AND ROBUSTNESS

by

Stephen Overduin

B.Sc., University College of the Fraser Valley, 2002

PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the Department
of
Statistics and Actuarial Science

Simon Fraser University
Fall 2004

APPROVAL

Name: Stephen Overduin

Degree: Master of Science

Title of project: Use of the Lognormal Distribution for Survival Data: Inference and Robustness

Examining Committee: Dr. Tim Swartz
Chair

Dr. Michael A. Stephens
Senior Supervisor
Simon Fraser University

Dr. Richard A. Lockhart
Simon Fraser University

Dr. Rachel Altman
External Examiner
Simon Fraser University

Date Approved: Dec 09/04

SIMON FRASER UNIVERSITY

PARTIAL COPYRIGHT LICENCE

Abstract

Two data sets are presented and various distributions, including the lognormal, are fitted to the data. A method is given to calculate exact confidence intervals for the quantiles of the lognormal distribution. The coverage probability of the confidence intervals is examined when the lognormal distribution is the correct model, and for various departures from lognormality. In addition, the connection between the coverage probability and the p-value from a goodness-of-fit test is explored.

Acknowledgements

I would like to thank the whole Statistics department and the graduate students, who all were a great help to me in completing this degree.

In particular, I would like to thank my supervisor, Dr. Michael Stephens. It is his patience that stands out most to me, as he has guided me in the ideas and writing of this project. He has taught me many things and I am indebted to him for passing on his wealth of knowledge, for supporting me in my work, and for helping me persevere.

Thanks also goes to Dr. Richard Lockhart for his willingness to give expert advice and necessary clarification.

The project was motivated by two talks given by Dr. Judy-Anne Chapman and Dr. Patricia Tai at the Statistical Society of Canada's 2004 annual meeting. Dr. Judy-Anne Chapman gave an overview of the use of the lognormal model in breast cancer investigations. Dr. Patricia Tai supplied an example with real data where the lognormal distribution was used. I am grateful to both of them for providing the motivation and for being willing to supply the data sets for this project.

A final thanks goes to Leanne and Leah as well as my extended family, who supported me in all things.

Commit your way to the LORD,
Trust also in Him,
And He shall bring it to pass.
Psalm 37:5

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many data analyses begin by fitting a distribution or model to data and then make subsequent inference; the resulting inference depends on which model is fitted. The focus in this project is on fitting the lognormal distribution to survival data and then going on to calculate confidence intervals for the quantiles. In practice, other models may provide a more adequate fit to the data, and this motivates the following questions concerning robustness:

1. Are confidence intervals for quantiles, based on the lognormal assumption, sensitive to departures from lognormality?

2. Is the outcome of a goodness-of-fit (GoF) test for lognormality connected to the performance of the confidence interval?

These questions are first examined in the context of two data sets and then more generally through simulation studies. In this first chapter we present the data and suggest several distributions to fit the data. In the second chapter we describe and apply GoF procedures. In the third chapter we examine confidence intervals for the quantiles, focusing especially on the lognormal distribution. In the fourth chapter we present the simulation studies and in the final chapter we conclude with summary remarks.

1.1 Survival Data

Survival data are very common in the medical field. For example, consider a group of patients who suffer from a particular cancer. A treatment is applied to each patient and they are followed until the study period ends. During the study period, a patient may die, be declared in remission, relapse, or be removed from the study due to certain circumstances. All these factors are considered in determining the survival time for a patient. For example, for those patients who die of the disease during the study, their exact survival time is known. Patients who are alive at the end of the study but not in remission are censored; that is, their survival time is known up to a lower bound. At the end of the study these survival times are examined to assess the treatment effect and to estimate survival rates.

A method for analyzing such data is given by Boag (1948). The first step in the analysis is to test the fit of the lognormal distribution to the group of patients who died of the disease during the study period. If the lognormal distribution gives an adequate fit, the analysis proceeds to assess treatment effects and calculate survival rates using the information from the full data set. This method is still implemented today by Dr. Patricia Tai (2003), an oncologist at the Saskatchewan Cancer Agency.

The two data sets below give survival times in months for patients who died from a particular cancer. Both sets were kindly supplied by Tai. Table 1.1 gives survival times for 184 patients who had limited stage small-cell lung cancer (LC). Table 1.2 gives survival times for 38 patients who died of cervical cancer (CC). The complete CC data set is currently on-line at www.ssc.ca/documents/case_studies/2002/cervical_e.html.

Histograms for the LC and CC data are given in Figures 1.1 and 1.2 respectively. Both histograms show distributions which are skewed to the right. There is one distinct mode in the histogram for the LC data, while the histogram for the CC data suggests the possibility of two modes.

If the lognormal assumption is reasonable for the data, transforming the data using

The Lung Cancer Data Set							
4.04	4.70	5.82	6.15	7.07	7.36	7.56	7.76
7.82	7.86	7.86	7.89	8.15	8.19	8.84	9.04
9.17	9.24	9.47	9.67	10.03	10.06	10.13	10.26
10.32	10.36	10.42	10.42	10.45	10.52	10.52	10.72
10.75	10.75	11.15	11.18	11.28	11.34	11.47	11.77
11.80	11.93	12.03	12.30	12.39	12.53	12.53	12.53
12.56	12.56	12.82	12.95	13.05	13.12	13.15	13.18
13.28	13.32	13.74	13.91	14.04	14.17	14.33	14.33
14.93	14.93	14.99	15.02	15.02	15.12	15.35	15.52
15.58	15.88	15.95	15.95	16.01	16.11	16.14	16.27
16.41	16.41	16.60	16.67	16.77	17.13	17.16	17.23
17.52	17.79	17.82	17.98	18.02	18.02	18.48	18.61
18.81	18.81	19.13	19.17	19.20	19.20	19.30	19.46
19.53	19.63	19.73	19.82	19.86	19.89	20.05	20.12
20.19	20.22	20.28	20.32	20.65	20.65	20.68	20.68
20.78	20.81	20.84	21.11	21.14	21.47	21.50	21.70
21.80	21.90	22.45	22.62	23.31	23.54	23.57	23.64
23.70	23.70	23.70	23.84	24.03	24.16	24.20	24.46
24.46	24.69	24.72	24.79	25.18	25.35	25.45	25.97
25.97	27.12	27.16	27.48	27.65	28.04	28.27	28.64
29.10	29.98	30.02	30.05	30.97	31.27	32.55	32.61
33.83	34.88	35.38	36.62	38.37	42.38	43.00	44.42
44.65	47.28	47.64	53.82	55.69	57.50	58.82	64.64

Table 1.1: The survival times in months for lung cancer patients

The Cervical Cancer Data Set							
5.26	6.64	8.38	9.80	11.08	11.18	12.56	12.66
13.45	14.14	17.46	17.52	20.91	21.67	23.18	25.74
25.78	32.55	34.13	37.55	38.07	38.70	39.85	41.88
50.83	51.16	53.98	55.96	57.11	62.50	66.08	67.82
67.86	70.55	78.05	82.78	96.13	100.67		

Table 1.2: The survival times in months for cervical cancer patients

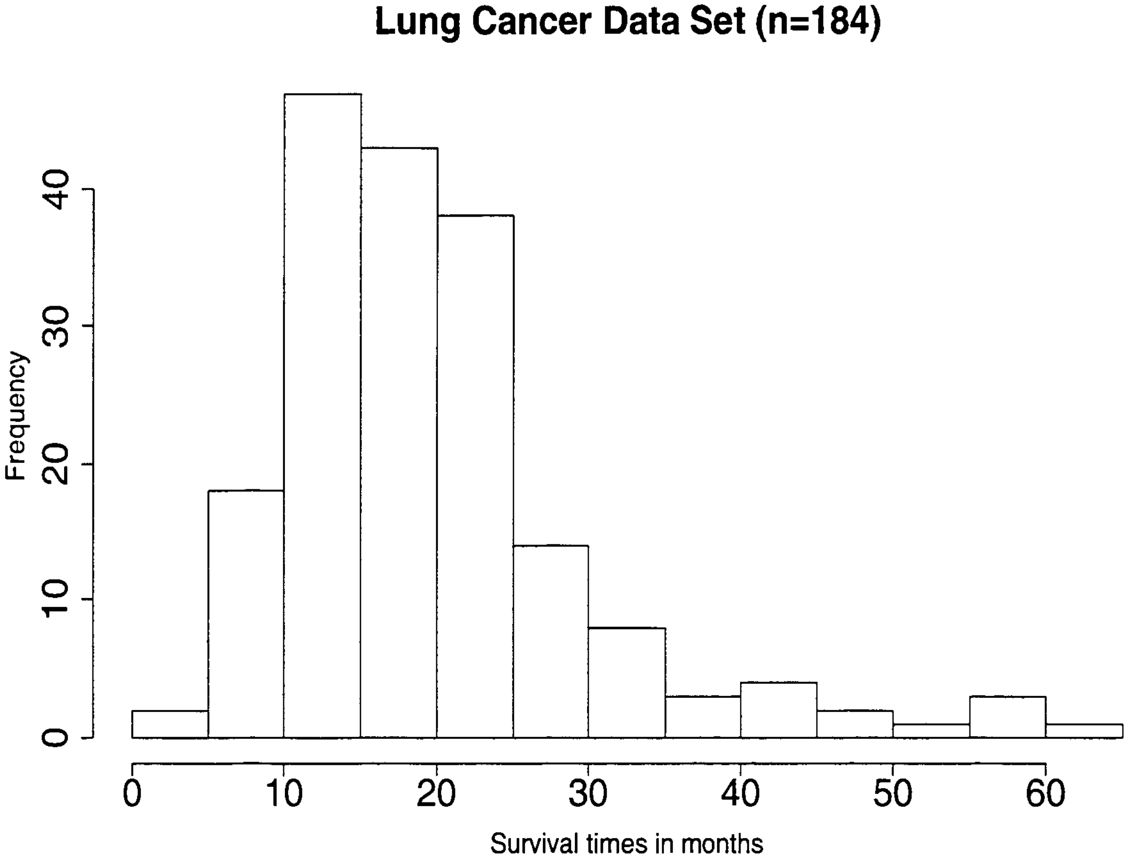

Figure 1.1: A histogram of the LC data

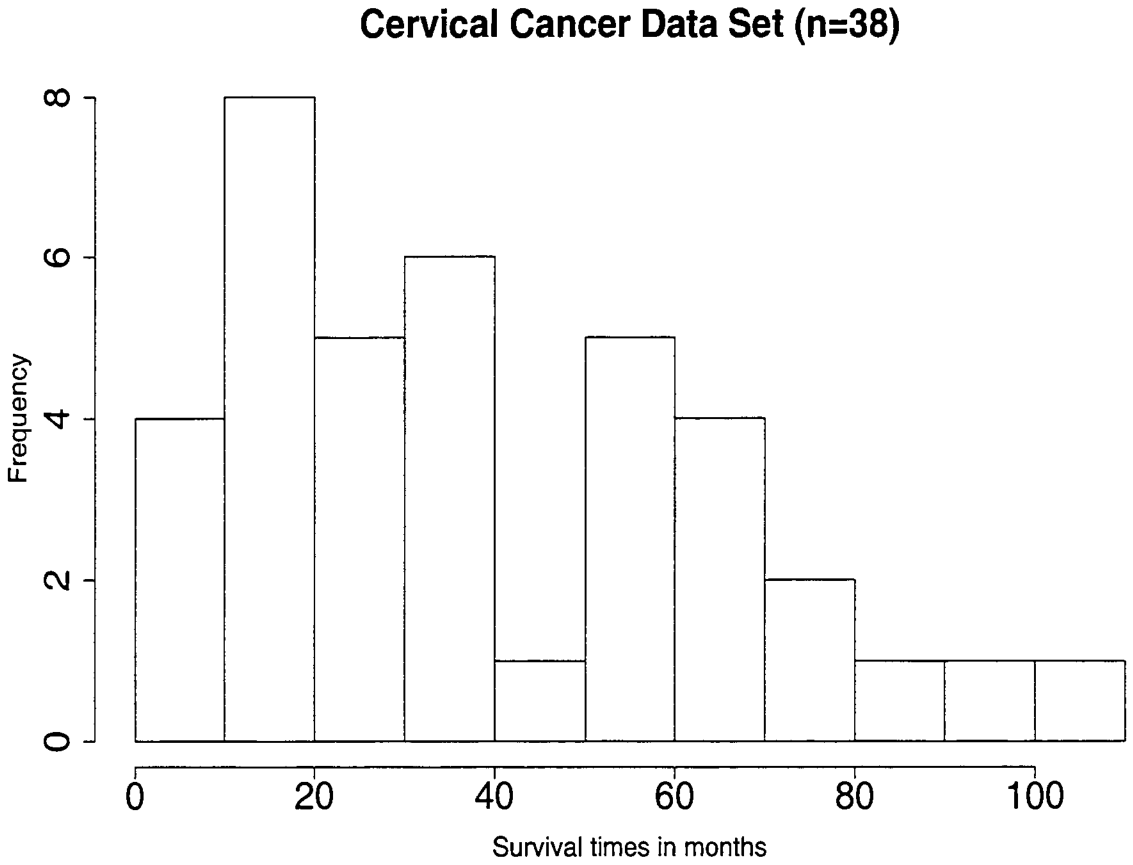

Figure 1.2: A histogram of the CC data

the log transformation will produce a bell-shaped histogram. Histograms for the log survival times are given in Figures 1.3 and 1.4. The histogram for the transformed LC data does appear bell-shaped, but the histogram for the transformed CC data is not as convincing.

There are other distributions used in survival data analysis; see Lawless (2003). Four of these distributions, including the lognormal, are described in the next section.

1.2 Survival Distributions

Some general notation is given first. Let X be a random variable with density function $f(x;\theta)$ and distribution function $F(x;\theta) = \int_{-\infty}^{x} f(t;\theta)dt$. Here, θ is a vector of parameters which is usually unknown and estimated from the data; the estimate is denoted by $\hat{\theta}$.

We refer to $X_1, X_2, \ldots, X_n$ as a sample of size n from $F(x;\theta)$, and denote the order statistics as $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. The sample mean for the X-set is denoted by $\overline{X} = \frac{1}{n}\sum X_i$ and the sample variance by $s_X^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$. All sums run from 1 to n. In some cases, the transformation $Y_i = \log X_i$ is made, and a Y-set is obtained from the X-set. The sample mean and variance of the Y-set are defined in a similar fashion and denoted $\overline{Y}$ and s_Y^2 respectively.

The p^{th} quantile of X is the value x_p such that $F(x_p;\theta) = p$. Solving this expression for x_p, we obtain the inverse cumulative distribution function, $F^{-1}(p;\theta)$, termed the quantile function.

For each distribution, $f(x;\theta)$, $F(x;\theta)$, and $F^{-1}(p;\theta)$ are given, explicitly if possible. Using the technique of maximum likelihood (ML) estimation, we give the ML estimates or the ML equations that need to be solved. For the latter, existing algorithms like the Newton-Raphson method can be implemented to obtain a solution.

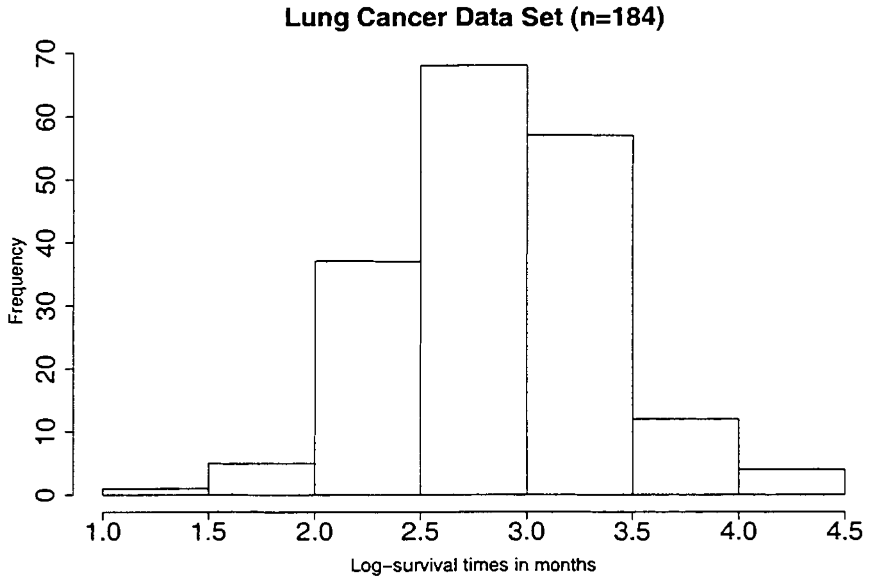

Figure 1.3: A histogram of the transformed LC data

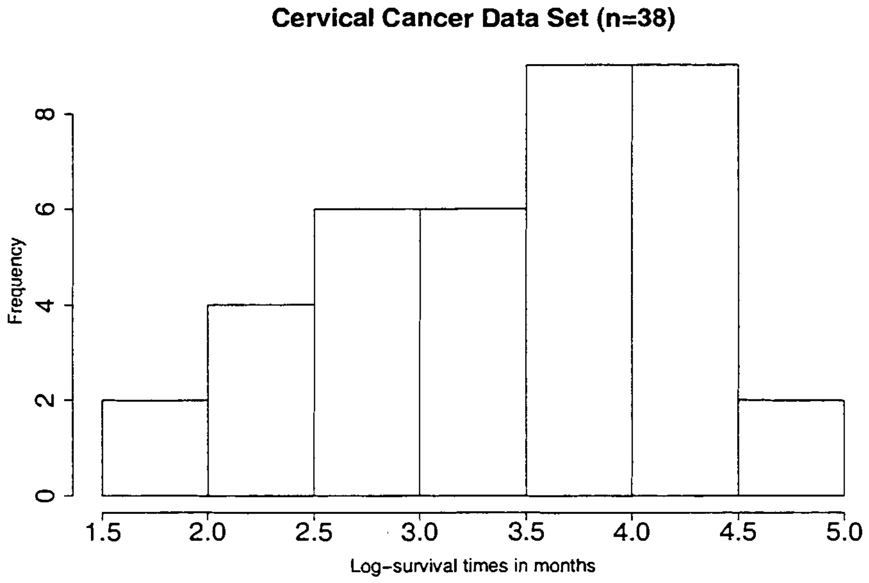

Figure 1.4: A histogram of the transformed CC data

The Exponential: EXP(β)

The exponential distribution was one of the first distributions used in survival analysis and is the simplest of the four distributions. The parameter β is a positive scale parameter.

Density Function:	$f(x;\beta) = \frac{1}{\beta}\exp(-x/\beta)$ for $x \geq 0$
Distribution Function:	$F(x;\beta) = 1 - \exp(-x/\beta)$ for $x \geq 0$
Quantile Function:	$F^{-1}(p;\beta) = x_p = -\beta\log(1-p)$ for $0 < p < 1$
ML Estimate:	$\hat{\beta} = \overline{X}$

The Weibull: WB(α, β)

The Weibull distribution is widely used today in many survival data applications. In addition to its scale parameter (β), it has a shape parameter (α), making it more flexible than the exponential distribution. Both parameters are positive. The Weibull distribution is equivalent to the exponential distribution when $\alpha = 1$.

Density Function: $f(x;\alpha,\beta) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp(-(x/\beta)^{\alpha})$ for $x \geq 0$

Distribution Function: $F(x;\alpha,\beta) = 1 - \exp(-(x/\beta)^{\alpha})$ for $x \geq 0$

Quantile Function: $F^{-1}(p;\alpha,\beta) = x_p = (-\beta^{\alpha}\log(1-p))^{1/\alpha}$ for $0 < p < 1$

ML Equations: $\alpha = \left(\frac{\sum X_i^{\alpha}\log X_i}{\sum X_i^{\alpha}} - \frac{1}{n}\sum \log X_i\right)^{-1}$

$\beta = (\frac{1}{n}\sum X_i^{\alpha})^{1/\alpha}$

Finding estimates for the Weibull distribution can be done indirectly by transforming the data. The transformation $Y = \log X$ gives the extreme value distribution in terms of a location (η) and scale (γ) parameter and estimation and inference for a location-scale distribution is easier to apply in practice. The parameters are transformed as $\eta = \log\beta$ and $\gamma = \frac{1}{\alpha}$ and give the following functions and ML equations for the extreme value distribution.

Density Function: $f(y;\eta,\gamma) = \frac{1}{\gamma}\exp(\frac{y-\eta}{\gamma} - \exp(\frac{y-\eta}{\gamma}))$ for $-\infty < y < \infty$

Distribution Function: $F(y;\eta,\gamma) = 1 - \exp(-\exp(\frac{y-\eta}{\gamma}))$ for $-\infty < y < \infty$

Quantile Function: $F^{-1}(p;\eta,\gamma) = y_p = \eta + \gamma\log(-\log(1-p))$ for $0 < p < 1$

ML Equations: $\eta = -\gamma\log\left[\frac{n}{\sum\exp(Y_i/\gamma)}\right]$

$$\gamma = -\overline{Y} + \frac{\sum Y_i\exp(Y_i/\gamma)}{\sum\exp(Y_i/\gamma)}$$

Using the estimates for the extreme value distribution, estimates for the Weibull distribution are then $\hat{\alpha} = \frac{1}{\hat{\gamma}}$ and $\hat{\beta} = \exp(\hat{\eta})$.

The Gamma: $\mathrm{G}(\alpha,\beta)$

The gamma distribution is similar to the Weibull in terms of flexibility and use, but in this case no transformation to a location-scale distribution is available. Both its shape (α) and scale (β) parameters are positive, and like the Weibull, the gamma reduces to the exponential when $\alpha = 1$.

Density Function: $f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}\exp(-x/\beta)$ for $x \geq 0$

Distribution Function: $F(x;\alpha,\beta) = \int_0^x f(u;\alpha,\beta)du$ for $x \geq 0$

Quantile Function: $F^{-1}(p;\alpha,\beta) = x_p$ for $0 < p < 1$

ML Equations: $\alpha = \exp\left\{\Psi(\alpha) + \log\overline{X} - \frac{1}{n}\sum\log X_i\right\}$

$\beta = \frac{\overline{X}}{\alpha}$

where $\Psi(\alpha)$ is the digamma function.

The Lognormal: LN(μ, σ^2)

The lognormal distribution has been used to fit a wide variety of cancer survival data (see Tai 2003). With two parameters, it also is very flexible. The range of μ is $-\infty$ to ∞, and $\sigma > 0$. The functions and estimates are as follows.

Density Function: $f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma x}\exp(\frac{-(\log x-\mu)^2}{2\sigma^2})$ for $x \geq 0$

Distribution Function: $F(x;\mu,\sigma^2)=\int_0^x f(u;\mu,\sigma^2)du$ for $x \geq 0$

Quantile Function: $F^{-1}(p;\mu,\sigma^2)=x_p$ for $0<p<1$

ML Estimates: $\hat{\mu}=\frac{1}{n}\sum \log X_i$

$\hat{\sigma}^2=\frac{1}{n-1}\sum(\log X_i-\hat{\mu})^2$

If a random variable X has a lognormal distribution, the random variable $Y=\log X$ is normally distributed, denoted N(μ,σ^2). This allows the more well-known analysis techniques for the normal distribution to be applied to lognormal data through transformation. The functions and estimates for the normal distribution are given below.

Density Function: $f(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp(\frac{-(x-\mu)^2}{2\sigma^2})$ for $-\infty<y<\infty$

Distribution Function: $F(y;\mu,\sigma^2)=\int_{-\infty}^y f(u;\mu,\sigma^2)du$ for $-\infty<y<\infty$

Quantile Function: $F^{-1}(p;\mu,\sigma^2)=y_p$ for $0<p<1$

ML Estimates: $\hat{\mu}=\overline{Y}$

$\hat{\sigma}^2=s_Y^2$

For the lognormal and normal distributions the correct ML estimate of σ^2 is $\frac{(n-1)\hat{\sigma}^2}{n}$, but as $\hat{\sigma}^2$ is the more common estimate, and in large samples the difference vanishes, it

	EXP(β)	WB(α,β)		G(α,β)		LN(μ,σ^2)	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\mu}$	$\hat{\sigma}^2$
LC Data	19.746	1.995	22.386	4.203	4.699	2.859	0.246
CC Data	38.990	1.520	43.368	1.980	19.690	3.390	0.645

Table 1.3: Parameter estimates for the fitted distributions for each data set

will be treated as the ML estimate for σ^2.

The estimated parameters for the LC and CC data sets are given in Table 1.3. Using these estimates, the density functions for the fitted distributions are plotted in Figures 1.5 and 1.6 for each data set, overlaying a relative histogram. These plots indicate to what extent the fitted distributions follow the shape of the data.

In both plots, the Weibull, gamma, and lognormal follow the shape of the data reasonably well and appear closest to one another in the upper tail of the distribution. In both cases, the exponential model is unable to capture the shape and is reasonably close to the others only in the upper tail. One way to look more closely at the differences among the distributions is to examine the quantiles of the fitted distributions. Estimates for various quantiles are given in Table 1.4 for both data sets.

The values in the table highlight that while differences exist, they may be small. For example, in the LC data set, quantile estimates for the Weibull, gamma, and lognormal fits give very similar estimates for nearly every quantile. There is less similarity however in the CC data set, especially for the upper quantiles. If estimating the upper quantiles is important in this situation, model choice becomes an issue, and GoF testing becomes a very important step in the data analysis. Which distribution gives the best fit for these two data sets? This question is answered in Chapter 2.

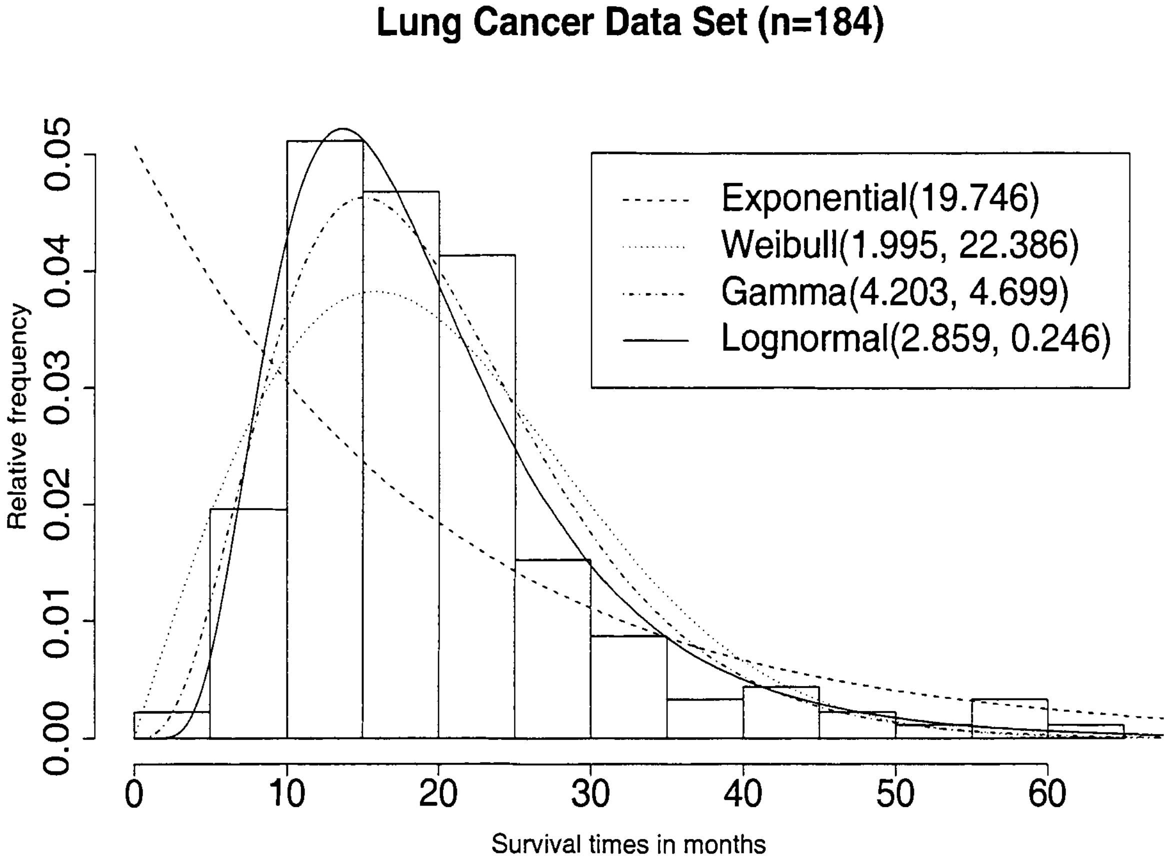

Figure 1.5: A relative histogram of the LC data with fitted density curves

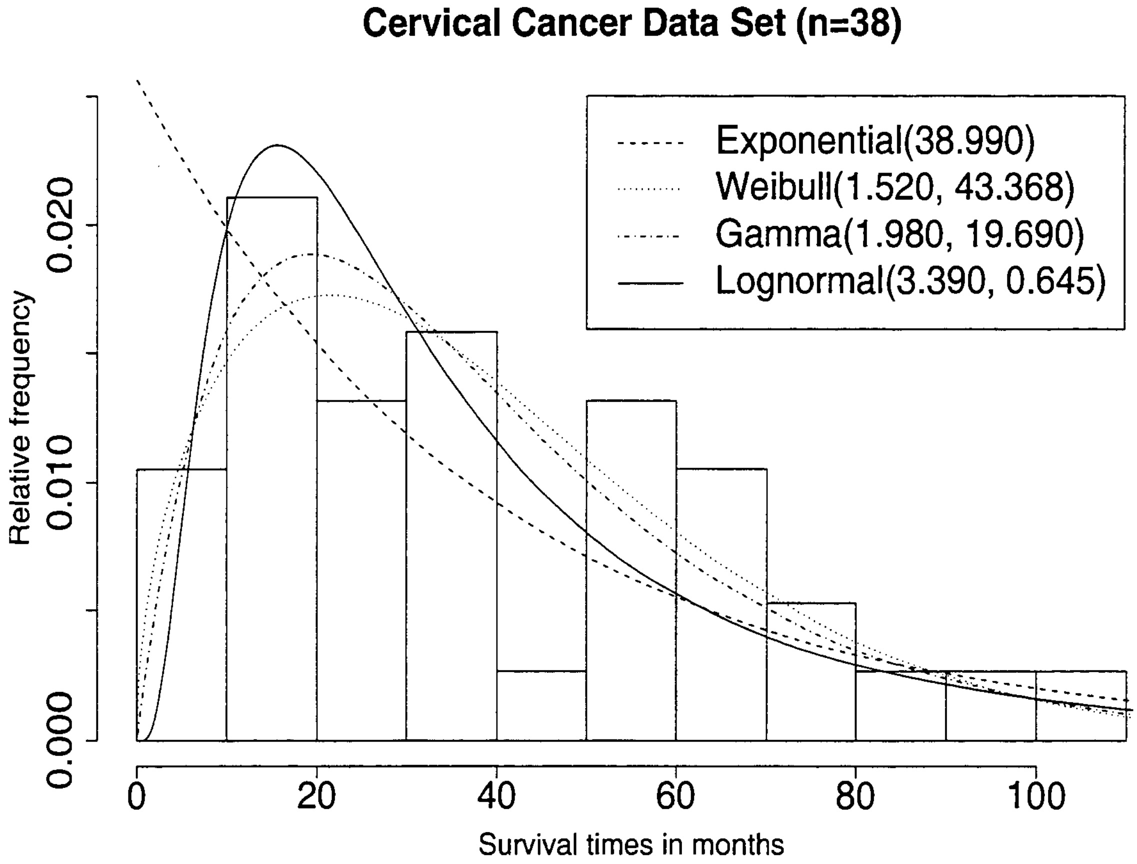

Figure 1.6: A relative histogram of the CC data with fitted density curves

	Value of p for p^{th} quantile x_p						
	0.010	0.050	0.100	0.500	0.900	0.950	0.990
LC Data Set ($n = 184$)							
EXP(19.746)	0.198	1.013	2.080	13.687	45.467	59.154	90.934
WB(1.995, 22.386)	2.232	5.052	7.247	18.629	34.004	38.798	48.129
G(4.203, 4.699)	4.279	6.977	8.838	18.204	32.654	37.782	48.707
LN(2.859, 0.246)	5.501	7.714	9.238	17.449	32.957	39.468	55.349
CC Data Set ($n = 38$)							
EXP(38.990)	0.392	2.000	4.108	27.026	89.778	116.804	179.556
WB(1.520, 43.368)	2.105	6.149	9.872	34.078	75.057	89.239	118.405
G(1.980, 19.690)	2.839	6.841	10.270	32.659	76.008	92.774	129.983
LN(3.390, 0.645)	4.578	7.914	10.596	29.667	83.060	111.209	192.267

Table 1.4: Estimates for several quantiles based on both the LC and CC data sets

Chapter 2

Goodness-of-Fit

Testing GoF is an important step in validating assumptions and strengthening the credibility of further analysis. For example, Tai (2003) has two phases in her analysis: the first phase is a GoF test for lognormality and the second is drawing inference under the lognormal assumption. In this chapter, a GoF procedure is presented and applied to the two data sets.

2.1 Testing Goodness-of-Fit

The null hypothesis for a GoF test of the distribution of a random sample of size n is:

$$H_0 : \text{ the distribution of the sample is } F(x; \theta).$$

Many statistics have been proposed for testing H_0. A well-known group of statistics is based on the empirical distribution function (EDF). The EDF is denoted by $F_n(x)$ and is defined as follows:

$$F_n(x) = \begin{cases} 0 & \text{if } x < X_{(1)}; \\ \frac{i}{n} & \text{if } X_{(i)} \leq x < X_{(i+1)}, \text{ for } i = 1, \ldots, n-1; \\ 1 & \text{if } x \geq X_{(n)}; \end{cases}$$

The plot of $F_n(x)$ against x is a step function giving the proportion of observations less than or equal to x. If H_0 is true, the EDF should mirror the null distribution, $F(x;\theta)$ and EDF statistics are based on $F_n(x) - F(x;\theta)$. Two of the more well-known and powerful statistics are the Cramér von Mises statistic, W^2, and the Anderson-Darling statistic, A^2, defined as

$$W^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x;\theta))^2 dF(x;\theta)$$
$$A^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x;\theta))^2 \psi(x) dF(x;\theta)$$

where $\psi(x) = [F(x;\theta)(1 - F(x;\theta))]^{-1}$. The weight function in A^2 is the variance of the EDF function and gives more prominence to the tails of the distributions. In general, A^2 is a more powerful statistic than W^2.

In practice, the Probability Integral Transform (PIT), $Z_i = F(X_i;\theta)$ is performed, which, if θ is known, produces a Z-set which is uniformly distributed on $[0,1]$. Performing the PIT does not cause the values of the statistics to change. The statistics are expressed in terms of $F_n(z) - z$ as

$$W^2 = n \int_0^1 (F_n(z) - z)^2 dz$$
$$A^2 = n \int_0^1 \frac{(F_n(z) - z)^2}{z(1-z)} dz.$$

Computing formulas based on the Z-set are given below.

$$W^2 = \sum \left(Z_{(i)} - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$$

$$A^2 = -n - \frac{1}{n} \sum (2i-1) \left(\log Z_{(i)} + \log(1 - Z_{(n+1-i)}) \right)$$

Usually, θ is unknown and the Z-set is calculated using $F(x;\hat{\theta})$ instead. Now the Z-set is no longer uniform, but its EDF is still plotted and compared against the uniform distribution and the computing formulas above are also used.

Finding the p-value for the Test

The quantiles or percentage points of the distributions of the statistics under the null hypothesis are required in order to obtain a p-value for the GoF test. Asymptotic percentage points of the distributions of W^2 and A^2 are given, for example by Stephens (1986), for a number of common null distributions. In many cases, however, these tables are limited to upper tail percentage points. One exception is the test of normality. Then, the distribution of W^2 or A^2 depends only on the sample size. Using this fact, a table has been given by Stephens (1986) to calculate p-values anywhere along the $[0,1]$ interval. This table proves useful in testing for normality when a fast and efficient method of computing an exact p-value is required and is used in in Chapter 4 for the simulation studies.

An alternative method of obtaining a p-value is by the parametric bootstrap. By the parametric bootstrap, we refer to the following procedure:

1. Generate an X'-set of size n from $F(x;\hat{\theta})$.

2. Calculate $\hat{\theta}'$ based on the X'-set.

3. Calculate $Z_i' = F(X_i';\hat{\theta}')$, $i = 1 \ldots n$.

4. Calculate the test statistics W^2 and A^2, using the Z'-set.

5. Repeat steps 1-4 M times where M is large.

After step 5 is complete, the M values of W^2 and M values of A^2 give suitable approximations to their respective distributions under H_0. For example, the p-value of A^2 is obtained by calculating the proportion of simulated A^2 values which fall beyond the observed A^2 value.

Two errors are present in this procedure which prevent the p-value from being exact. The first error is the error due to simulation because M is finite. The second error in this procedure is that the X'-sets are generated from the wrong distribution, that is $F(x;\hat{\theta})$ and not $F(x;\theta)$. However, the error has implications for the p-value only when the distribution of the test statistic depends on the true parameters. Therefore, no error is made in testing the fit of the normal distribution, for example. In contrast, an error is made in testing fit to the gamma distribution as the distributions of W^2 and A^2 depend on the true value of the shape parameter. Despite these errors, the parametric bootstrap, being easy to implement on the computer, remains a common procedure and is an improvement on the asymptotic tables.

Once the p-value has been obtained, the fitted distributions are ranked by their p-value and the model with the highest p-value is considered to give the best fit.

2.2 Application to the Data Sets

In Figures 2.1 and 2.2, the EDF is plotted along with the fitted distributions for the two data sets. As in Figures 1.5 and 1.6, these plots indicate how well the EDF follows the fitted distributions. It remains difficult, however, to determine which of the Weibull, gamma, or lognormal distributions best describe the data in each case. For the CC data set (Figure 2.2), the upper tail of the EDF follows more closely the fitted Weibull and gamma models, which was not as evident in Figure 1.6.

In Figures 2.3 and 2.4, the four fitted distributions are separately compared with the uniform distribution function after doing the PIT. From Figure 2.3 it is clear that the

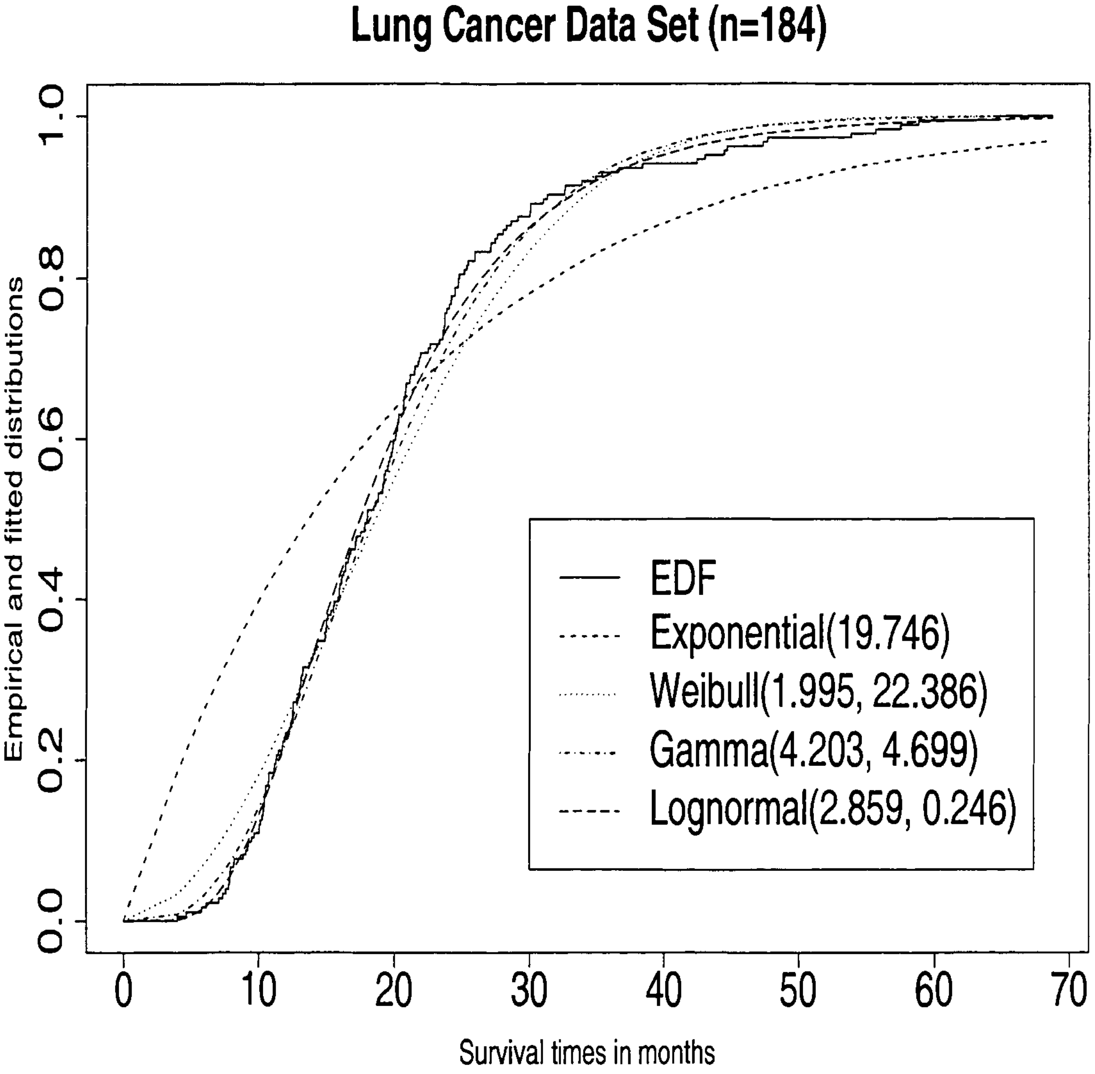

Figure 2.1: EDF of the LC data set and four fitted distributions

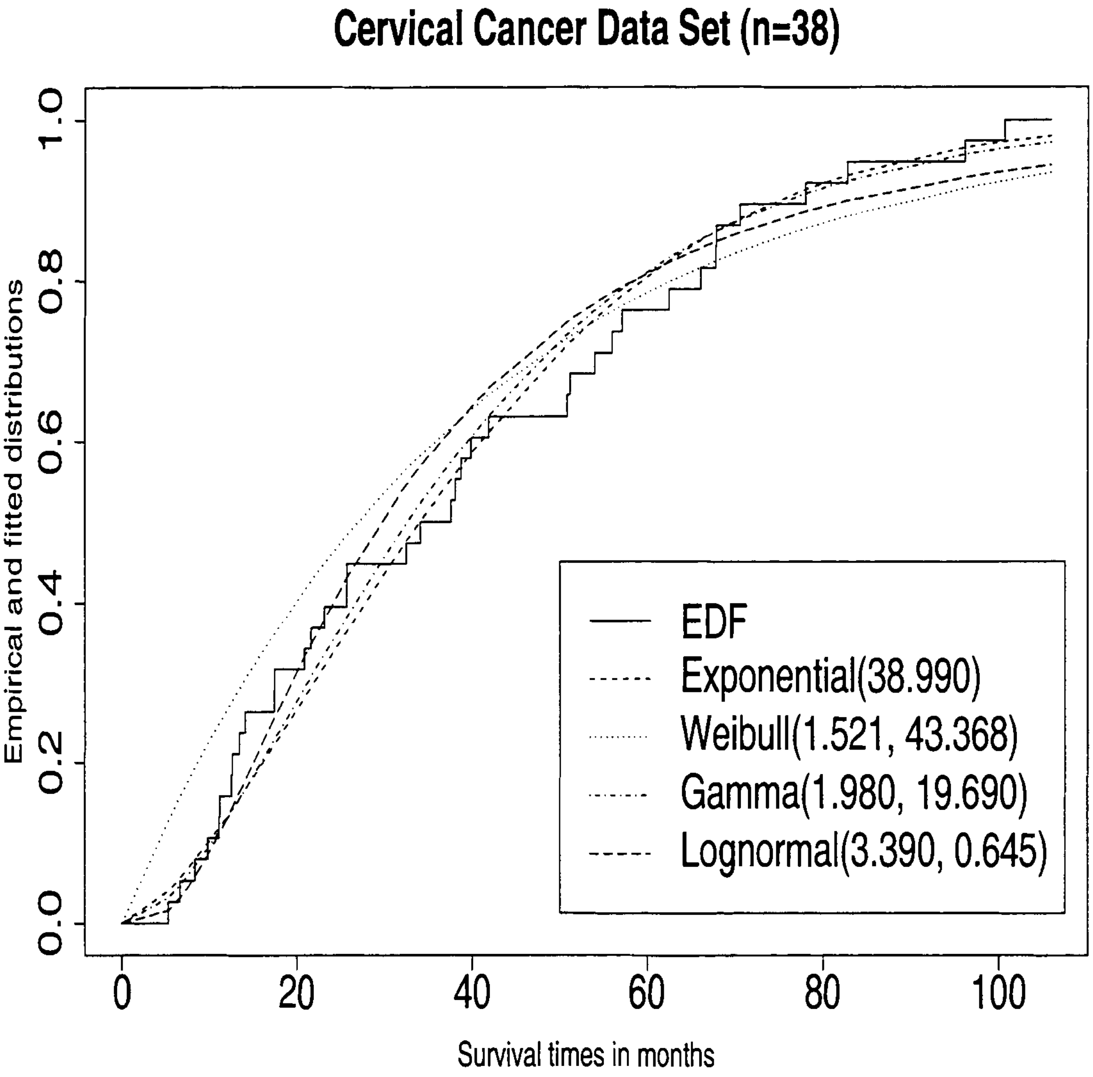

Figure 2.2: EDF of the CC data set and four fitted distributions

		W^2	p-value	A^2	p-value
LC Data	EXP(19.746)	4.6642	0.0000	24.092	0.0000
($n = 184$)	WB(1.995, 22.386)	0.4547	0.0000	3.2546	0.0000
	G(4.203, 4.699)	0.1391	0.0383	1.0808	0.0095
	LN(2.859, 0.246)	0.0595	0.3820	0.3947	0.3698
CC Data	EXP(38.990)	0.2261	0.0457	1.5113	0.0279
($n = 38$)	WB(1.520, 43.368)	0.0686	0.2823	0.4302	0.3107
	G(1.980, 19.690)	0.0741	0.2693	0.4554	0.2821
	LN(3.390, 0.645)	0.0978	0.1120	0.5876	0.1160

Table 2.1: Test statistics and p-values for various distributions and both data sets

Weibull, gamma, and lognormal fits to the LC data set lie close to the straight line. The closest are the gamma and lognormal fits, while the exponential clearly does not fit well. For the CC data set in Figure 2.4, it is more difficult to determine which distribution lies closest to the straight line, as the distributions appear very similar in general.

It is tempting to think that the 'best fitting' distribution is the one which lies closest to the straight line; in other words, the one which gives the smallest value for the statistic. However, the ranking of the fitted distributions based on the value of the statistic may not be the same as the ranking based on the p-values, as the distribution of the statistic depends on the null distribution being tested. In Table 2.1, values of W^2 and A^2 and their p-values are given. The p-values were obtained using the parametric bootstrap method with $M = 10000$.

Consider the LC data set first. Despite the similarity in fits for the Weibull, gamma, and lognormal distributions evident in Figure 2.3, the p-value for the lognormal fit is much larger than for the gamma or Weibull. In fact, the Weibull gives a p-value of 0.0000 for both W^2 and A^2. Using the typical significance levels of 0.05 or 0.10, tests for a Weibull or gamma fit would reject H_0, while that for the lognormal would not reject H_0 by a relatively wide margin. In this case, where sample size, n, is large, the power of the test is relatively high greatly increased and therefore allows a greater ability to

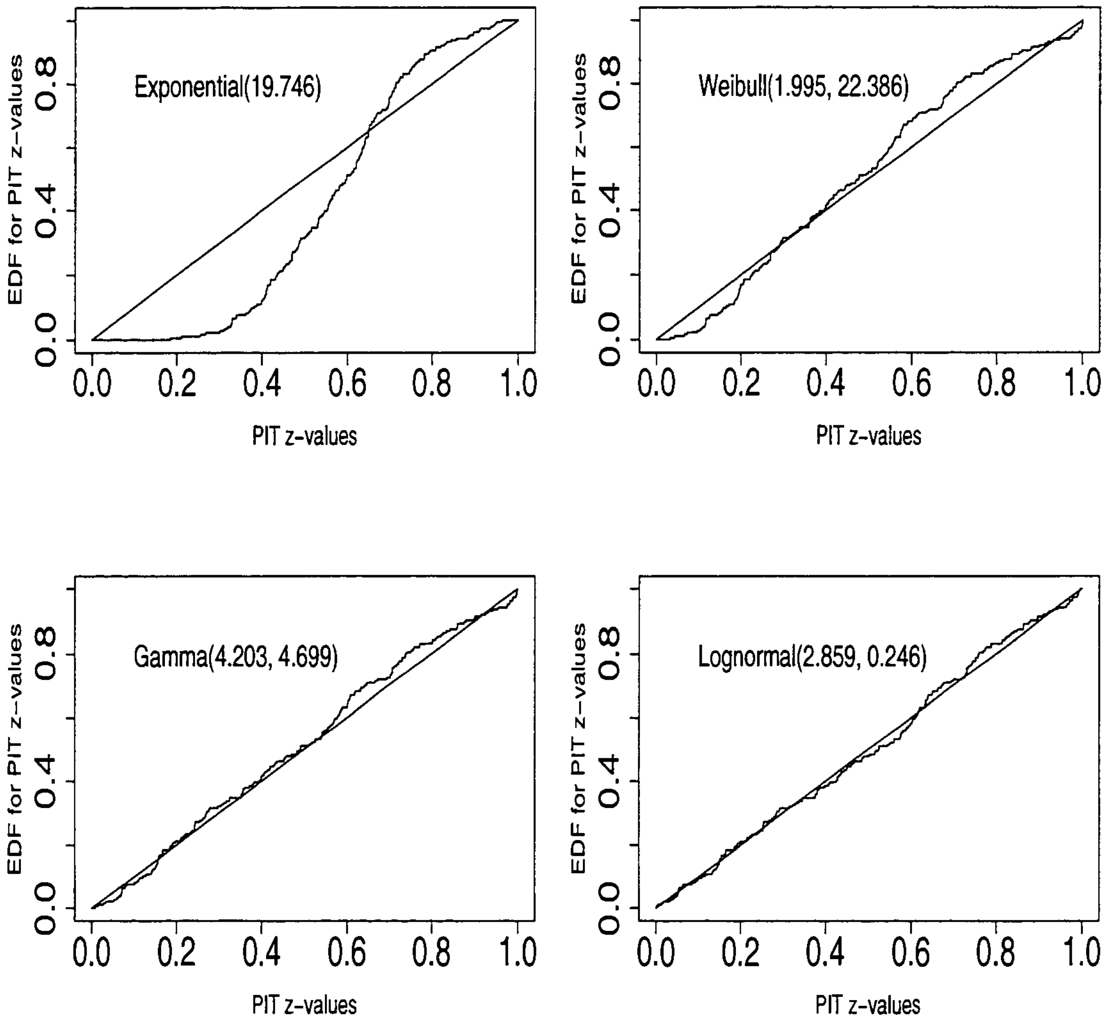

Figure 2.3: EDF plots for the PIT z-values from the four distributions fitted to the LC data set

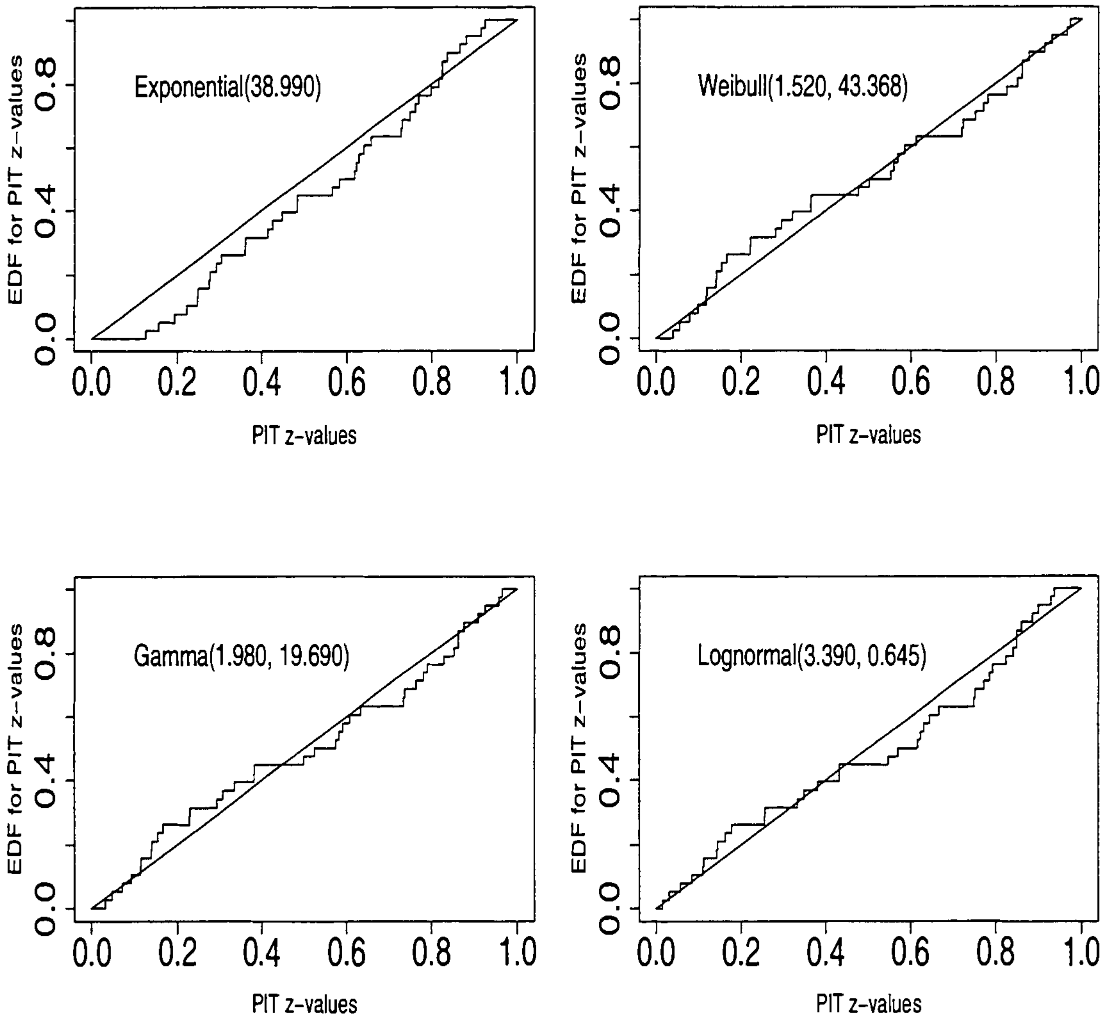

Figure 2.4: EDF plots for the PIT z-values from the four distributions fitted to the CC data set

distinguish among various models.

The CC data set is somewhat more interesting. The p-values confirm that the Weibull, gamma, and lognormal provide adequate fits, though the p-value for the lognormal sits very near the 0.10 significance level. Even the exponential distribution, though it is rejected at the 0.05 significance level, has a p-value very close to 0.05 using W^2.

Based on the ranks of the p-values, the best fitting distribution to the LC data is the lognormal, while the Weibull distribution gives the best fit to the CC data. The decision seems clear for the LC data but less so for the CC data. In fact, the fits for the CC data give an example of the situation described in the opening paragraphs of the first chapter. Even though the lognormal model produces an adequate fit, the Weibull and gamma fits both have a higher p-value, giving a better fit to the data. What is the impact on subsequent inference in this situation if the lognormal model is chosen instead of the better fitting Weibull or gamma models? In particular, how do confidence intervals for the quantiles compare among the various fitted distributions? This question is examined in Chapter 3.

Chapter 3

Confidence Intervals

Point estimates for various quantiles of the fitted distributions are given in Chapter 1. Standard errors or confidence intervals should be included to provide an idea of the accuracy of the estimates. Procedures are given in this chapter to calculate confidence intervals for the quantile point estimates, with a special emphasis on the method for the lognormal quantiles. The procedures are then applied to the two data sets.

To construct a confidence interval (CI) for the p^{th} quantile, x_p, we need to find numbers L_p and U_p that satisfy

$$P(L_p < x_p < U_p) = 1 - \alpha, \tag{3.1}$$

and we say that (L_p, U_p) is a $100(1-\alpha)\%$ CI for x_p. One method of finding the numbers L_p and U_p is to construct a pivotal quantity, or pivot. In general, a random variable $G(X_1, X_2, \ldots, X_n, \theta)$ is a pivot for θ if the distribution of $G(X_1, X_2, \ldots, X_n, \theta)$ does not depend on θ (Casella and Berger, 2002). For example, given an X-set from a $N(\mu, 1)$ distribution, the random variable $G(X_1, X_2, \ldots, X_n, \mu) = \sqrt{n}(\overline{X} - \mu)$ follows a $N(0, 1)$ distribution and therefore can be used as a pivot for μ. For quantiles, a random variable is needed which includes x_p in its construction, but whose distribution does not depend on x_p. The problem is then reduced to finding the percentage points of the pivotal distribution.

In general, methods which satisfy (3.1) exactly are difficult to obtain and may be computationally intensive. As a results, approximate methods based on asymptotic theory or simulations have been suggested, and are adequate given a reasonably large sample size. Exact methods are first discussed for the exponential and lognormal quantiles, followed by some approximate methods applicable to all four distributions.

3.1 Exact Methods

Exact methods are available (see Lawless 2003) for any location-scale distribution, even when there are censored data. We consider exact methods for the exponential and lognormal distributions for a complete (uncensored) sample.

3.1.1 The Exponential Distribution

The procedure for the exponential distribution is based on the following well-known result: given a random sample $X_1, X_2, \ldots, X_n$ from an exponential distribution, the distribution of $2n\frac{\overline{X}}{\beta}$ is χ^2_{2n}.

The quantile function for the exponential distribution is $x_p = -\beta \log(1-p)$, and the ML estimate is $\hat{x}_p = -\overline{X}\log(1-p)$. As $\frac{\hat{x}_p}{x_p} = \frac{\overline{X}}{\beta}$, the pivotal quantity $2n\frac{\hat{x}_p}{x_p}$ follows a χ^2_{2n} distribution and can be used to give an exact CI for the p^{th} quantile of the exponential distribution.

Let C be a χ^2_{2n} random variable and let c_α be the value which satisfies, $P(C < c_\alpha) = \alpha$. Then

$$P(c_{\frac{\alpha}{2}} < 2n\frac{\hat{x}_p}{x_p} < c_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

Pivoting on x_p, we obtain

$$P\left(\frac{2n\hat{x}_p}{c_{1-\frac{\alpha}{2}}} < x_p < \frac{2n\hat{x}_p}{c_{\frac{\alpha}{2}}}\right) = 1 - \alpha$$

and the $100(1-\alpha)\%$ CI is given by

$$\left(\frac{2n\hat{x}_p}{c_{1-\frac{\alpha}{2}}}, \frac{2n\hat{x}_p}{c_{\frac{\alpha}{2}}}\right).$$

The percentage points for the χ^2_{2n} distribution can be easily obtained in tables or using standard statistical software.

3.1.2 The Lognormal Distribution

A CI for the lognormal quantile x_p is obtained by transforming the CI for the normal quantile y_p. If the interval (L_p, U_p) gives a CI for y_p, then $(\exp(L_p), \exp(U_p))$ gives a CI for x_p. The exact CI procedure for y_p is based on a pivot which follows the non-central t-distribution and is given by Lawless (2003) and Johnson, Kotz and Balakrishnan (1995). The details are outlined below.

If X follows a $LN(\mu, \sigma^2)$ distribution, then $Y = \log X$ follows a $N(\mu, \sigma^2)$ distribution and $y_p = \mu + z_p\sigma$ where z_p is the value of the p^{th} quantile for the standard normal distribution, $N(0,1)$. Given a random X-set, the log transformation produces a Y-set with ML estimates $\hat{\mu} = \overline{Y}$ and $\hat{\sigma}^2 = S^2_Y$. The pivotal quantity $Q_p = \frac{\sqrt{n}(\overline{Y}-y_p)}{S_Y}$ follows a non-central t-distribution with $n-1$ degrees of freedom and non-centrality parameter, $-\sqrt{n}z_p$.

Let $t_v(\delta)$ follow a non-central t-distribution with v degrees of freedom and non-centrality parameter δ and let $t'_{v,\alpha}(\delta)$ be the value such that $P(t_v(\delta) < t'_{v,\alpha}(\delta)) = \alpha$. Using the fact that $t'_{v,\alpha}(-\delta) = -t'_{v,1-\alpha}(\delta)$, (see Johnson et al. 1995), we can write

$$P\left(t'_{n-1,\frac{\alpha}{2}}(-\sqrt{n}z_p) < \frac{\sqrt{n}(\overline{Y}-y_p)}{S_Y} < -t'_{n-1,\frac{\alpha}{2}}(\sqrt{n}z_p)\right) = 1 - \alpha.$$

Pivoting on y_p, the $100(1-\alpha)\%$ CI for y_p is

$$\left(\overline{Y} + t'_{n-1,\frac{\alpha}{2}}(\sqrt{n}z_p)\frac{S_Y}{\sqrt{n}}, \overline{Y} - t'_{n-1,\frac{\alpha}{2}}(-\sqrt{n}z_p)\frac{S_Y}{\sqrt{n}}\right).$$

The difficulty in applying this method is finding the percentage points for the non-central t-distribution. The percentage points can be obtained through tables in the literature (Owen 1962), SAS, or through IMSLIB (available in FORTRAN). If access to tables or statistical software is limited, however, it becomes difficult to calculate an exact CI.

We define a new pivotal quantity in what follows, and give a method to calculate its percentage points.

The New Pivotal

Let $\hat{y}_p = \overline{Y} + z_p S_Y$ be an estimate for y_p and consider the pivotal quantity

$$W_p = \frac{\sqrt{n}(y_p - \hat{y}_p)}{S_Y}. \tag{3.2}$$

Note that

$$W_p = -(Q_p + \sqrt{n}z_p). \tag{3.3}$$

Let $w_{p,\alpha}$ be the value such that $P(W_p < w_{p,\alpha}) = \alpha$. A $100(1-\alpha)\%$CI for y_p is given by

$$\left(\hat{y}_p + w_{p,\frac{\alpha}{2}}\frac{S_Y}{\sqrt{n}}, \hat{y}_p + w_{p,1-\frac{\alpha}{2}}\frac{S_Y}{\sqrt{n}}\right). \tag{3.4}$$

Since W_p includes $\hat{y}_p$ explicitly, it is intuitively more attractive than Q_p, and also has two advantages over Q_p:

1. As $n \to \infty$, W_p is asymptotically normal. Q_p, on the other hand, goes off to $\pm\infty$ depending on the quantile of interest. To see this, we write $Q_p = \frac{\sqrt{n}(\overline{Y}-\mu)}{S_Y} - \frac{\sqrt{n}z_p\sigma}{S_Y}$. Now $\frac{\sqrt{n}(\overline{Y}-\mu)}{S_Y} \to N(0,1)$ (see Casella and Berger, 2002) and $\frac{\sqrt{n}z_p\sigma}{S_Y} \to \pm\infty$ depending on the sign of z_p.

2. Due to its limiting normal distribution, the percentage points of W_p for a particular p and α will produce simple curves when plotted against $1/\sqrt{n}$, anchored at an exact intercept when $n = \infty$.

Asymptotic Distribution of W_p

The pivot $W_p = U + V_p$, where $U = \frac{\sqrt{n}(\mu - \overline{Y})}{S_Y}$ and $V_p = \sqrt{n} z_p (\frac{\sigma}{S_Y} - 1)$. Now $-U$ follows the central t-distribution with $n - 1$ degrees of freedom, and converges to a $N(0,1)$ distribution (Casella and Berger 2002) and therefore so does U.

The second component, V_p, converges to a $N(0, \frac{z_p^2}{2})$ distribution. To see this, we use a first-order Taylor series expansion of the function $g(S_Y) = \frac{\sigma}{S_Y}$ about σ. That is, for some σ^* between S_y and σ

$$g(S_Y) = g(\sigma) + g'(\sigma)(S_Y - \sigma) + g''(\sigma^*)\frac{(S_Y - \sigma)^2}{2},$$

where $g''(\sigma^*)\frac{(S_Y - \sigma)^2}{2}$ goes to 0 as S_Y converges to σ in probability. Therefore we write

$$\frac{\sigma}{S_y} \approx 1 - \frac{S_Y - \sigma}{\sigma}$$

from which we obtain

$$V_p \approx -z_p\sqrt{n}\left(\frac{S_Y - \sigma}{\sigma}\right).$$

We multiply the top and bottom of the right-hand side of the above equation by $(S_Y + \sigma)$ and use the fact that S_y converges to σ in probability to give approximately $-z_p \frac{\sqrt{n}(S_Y^2 - \sigma^2)}{2\sigma^2} = -\frac{z_p}{2}\sqrt{n}\left(\frac{S_Y^2}{\sigma^2} - 1\right)$.

The term, $\frac{S_Y^2}{\sigma^2}$, follows a $\frac{\chi_{n-1}^2}{n-1}$ distribution where $n-1$ is the degrees of freedom. Since the limiting distribution of $\sqrt{v}\left(\frac{\chi_v^2}{v} - 1\right)$ is $N(0,2)$ for large v, the limiting distribution of $\sqrt{n}\left(\frac{S_Y^2}{\sigma^2}-1\right)$ is $N(0,2)$. Therefore the limiting distribution of V_p is $N(0,\frac{z_p^2}{2})$.

If the two components of W_p were independent, the joint distribution of U and V_p would converge to the joint distribution of two independent normals. This would be sufficient to say that W_p converges to the sum of the two limiting normal distributions. However, though the covariance of U and V_p is 0, they are not independent since S_Y is in both terms.

Therefore, we write $(U,V_p)=(U^*,V_p)+(\sqrt{n}(\mu-\overline{Y})\left(\frac{1}{S_Y}-\frac{1}{\sigma}\right),0)$, where $U^*=\frac{\sqrt{n}(\mu-\overline{Y})}{\sigma}$. Then (U^*,V_p) are independent. In the second term, because S_Y converges to σ in probability and $\sqrt{n}(\mu-\overline{Y})$ converges to $N(0,\sigma^2)$, Slutsky's Theorem states that $(\sqrt{n}(\mu-\overline{Y})\left(\frac{1}{S_Y}-\frac{1}{\sigma}\right),0)$ converges to $(0,0)$ in distribution. It follows that (U,V_p) and (U^*,V_p) converge to the same asymptotic distribution.

The asymptotic distribution of W_p is therefore $N(0,A_p)$, where $A_p=1+\frac{z_p^2}{2}$. The moments of W_p give an indication of the speed of convergence.

The Moments of W_p

We denote the k^{th} moment of W_p about the origin as $\mu'_{p,k}=E(W_p^k)$.

Calculating $\mu'_{p,k}$ requires finding $E\left(\frac{1}{S_Y^k}\right)$. Using the fact that $\frac{(n-1)S_Y^2}{\sigma^2}$ follows a $\chi^2_{(n-1)}$ distribution, we find:

$$E\left(\frac{1}{S_Y^k}\right)=\left(\frac{n-1}{2\sigma^2}\right)^{k/2}\frac{\Gamma(\frac{n-1-k}{2})}{\Gamma(\frac{n-1}{2})}.$$

Setting $C_{n,k}=\left(\frac{n-1}{2}\right)^{k/2}\frac{\Gamma(\frac{n-1-k}{2})}{\Gamma(\frac{n-1}{2})}$, the moments of W_p are given below:

	$p = 0.5$		$p = 0.6$		$p = 0.7$		$p = 0.8$		$p = 0.9$	
n	$\sqrt{\beta_1}$	β_2	$\sqrt{\beta_1}$	β_2	$\sqrt{\beta_1}$	β_2	$\sqrt{\beta_1}$	β_2	$\sqrt{\beta_1}$	β_2
10	0.000	4.200	0.376	4.427	0.716	5.032	1.002	5.848	1.234	6.742
20	0.000	3.400	0.204	3.460	0.394	3.623	0.558	3.854	0.699	4.123
50	0.000	3.133	0.114	3.151	0.220	3.200	0.315	3.271	0.397	3.357
100	0.000	3.063	0.077	3.071	0.150	3.094	0.215	3.127	0.272	3.167
250	0.000	3.024	0.048	3.028	0.093	3.036	0.133	3.049	0.169	3.064

Table 3.1: Values of $\sqrt{\beta_1}$ and β_2 for W_p at various sample sizes

$$\mu'_{p,1} = \sqrt{n} z_p \left[C_{n,1} - 1\right],$$

$$\mu'_{p,2} = n z_p^2 \left[C_{n,2}\left(\frac{1}{n z_p^2} + 1\right) - 2C_{n,1} + 1\right],$$

$$\mu'_{p,3} = n^{3/2} z_p^3 \left[C_{n,3}\left(\frac{3}{n z_p^2} + 1\right) - 3C_{n,2}\left(\frac{1}{n z_p^2} + 1\right) + 3C_{n,1} - 1\right],$$

$$\mu'_{p,4} = n^2 z_p^4 \left[C_{n,4}\left(\frac{3}{n^2 z_p^4} + \frac{6}{n z_p^2} + 1\right) - 4C_{n,3}\left(\frac{3}{n z_p^2} + 1\right) + 6C_{n,2}\left(\frac{1}{n z_p^2} + 1\right) - 4C_{n,1} + 1\right].$$

The central moments are defined as $\mu_{p,k} = E((W_p - \mu'_{p,1})^k)$, and the skew and kurtosis are defined as $\sqrt{\beta_1} = \sqrt{\frac{\mu_{p,3}^2}{\mu_{p,2}^3}}$ and $\beta_2 = \frac{\mu_{p,4}}{\mu_{p,2}^2}$ respectively. For a normal random variable, the values for $\sqrt{\beta_1}$ and β_2 are 0 and 3 respectively. In Table 3.1 the values of $\sqrt{\beta_1}$ and β_2 for W_p are given for various values of n and p, illustrating the speed of convergence to the values 0 and 3.

Percentage Points of W_p

The approximate percentage points of W_p can be found by fitting Pearson curves (see Solomon and Stephens 1978) or Cornish-Fisher expansions (see Kendall and Stuart 1977). These methods use the moments given above, and give very good results when the distribution is close to normal.

The points for W_p can be found exactly, however, using (3.3) which gives

$$w_{p,\alpha} = -\left(t'_{n-1,1-\alpha}(-\sqrt{n}z_p) + \sqrt{n}z_p\right). \tag{3.5}$$

The points for the non-central t-distribution are available in SAS or through IMSLIB, and the points for W_p are therefore easily obtained. Some caution is needed in making use of the algorithms in SAS and IMSLIB, since for large sample sizes, the algorithm can fail to give correct output. SAS was found to be accurate for a greater sample size than IMSLIB, and we therefore used SAS to first obtain the percentage points of Q_p from which we calculated the percentage points of W_p using (3.5).

Plots of the Percentage Points for Q_p and W_p

The percentage points for both Q_p and W_p are plotted to illustrate the advantage of the limiting normal distribution for W_p.

In Figure 3.1, the percentage points of Q_p are plotted against $1/\sqrt{n}$ for $p = 0.9$ and various α values. As $1/\sqrt{n} \to 0$, the percentage points decrease to $-\infty$

In Figure 3.2, the percentage points of W_p are plotted against $1/\sqrt{n}$ for the same values of p and α. As $1/\sqrt{n} \to 0$, each curve approaches an intercept or anchor point given by $z_\alpha\sqrt{A_p}$. The smoothness of the curves and the anchor points invite us to approximate these curves by a function of $1/\sqrt{n}$.

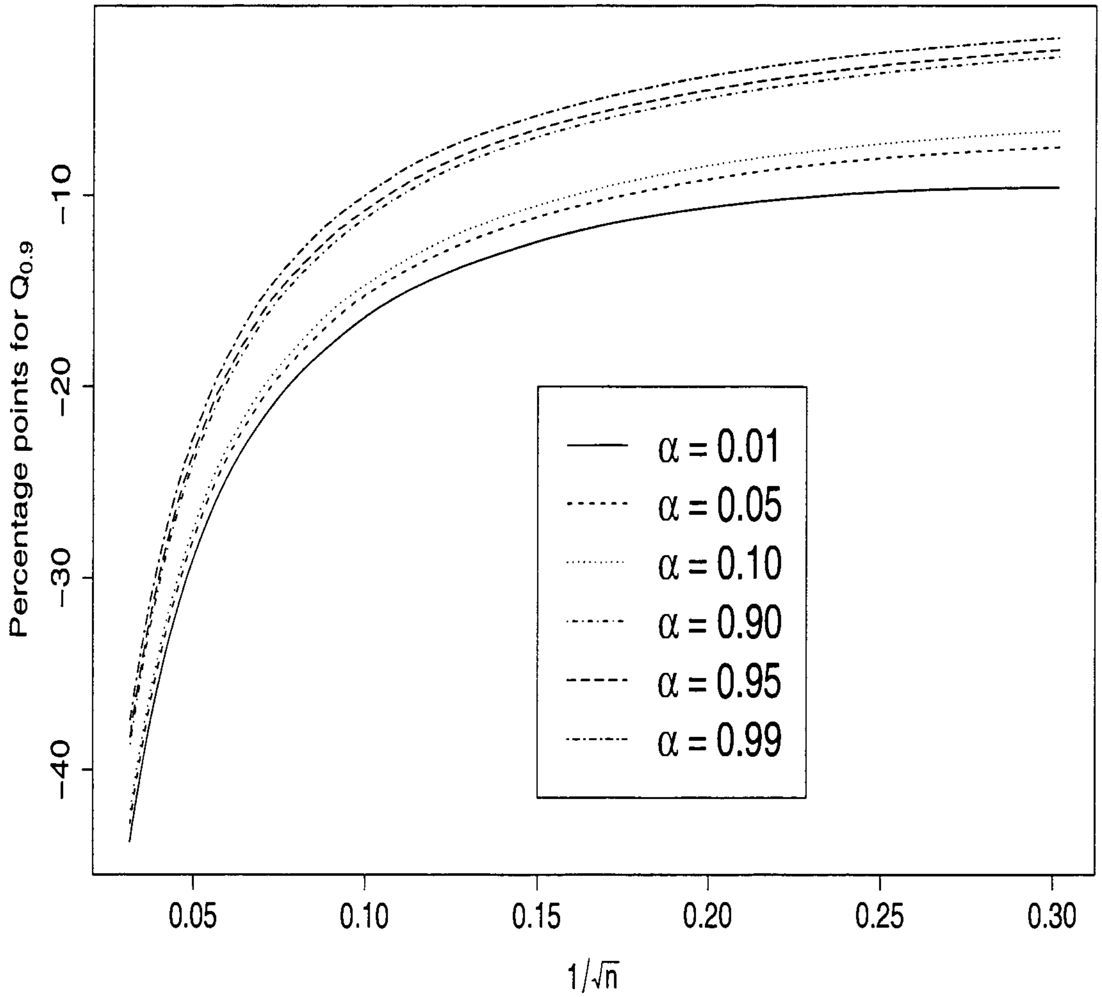

Figure 3.1: The percentage points of $Q_{0.90}$ plotted against $1/\sqrt{n}$ at various levels of α.

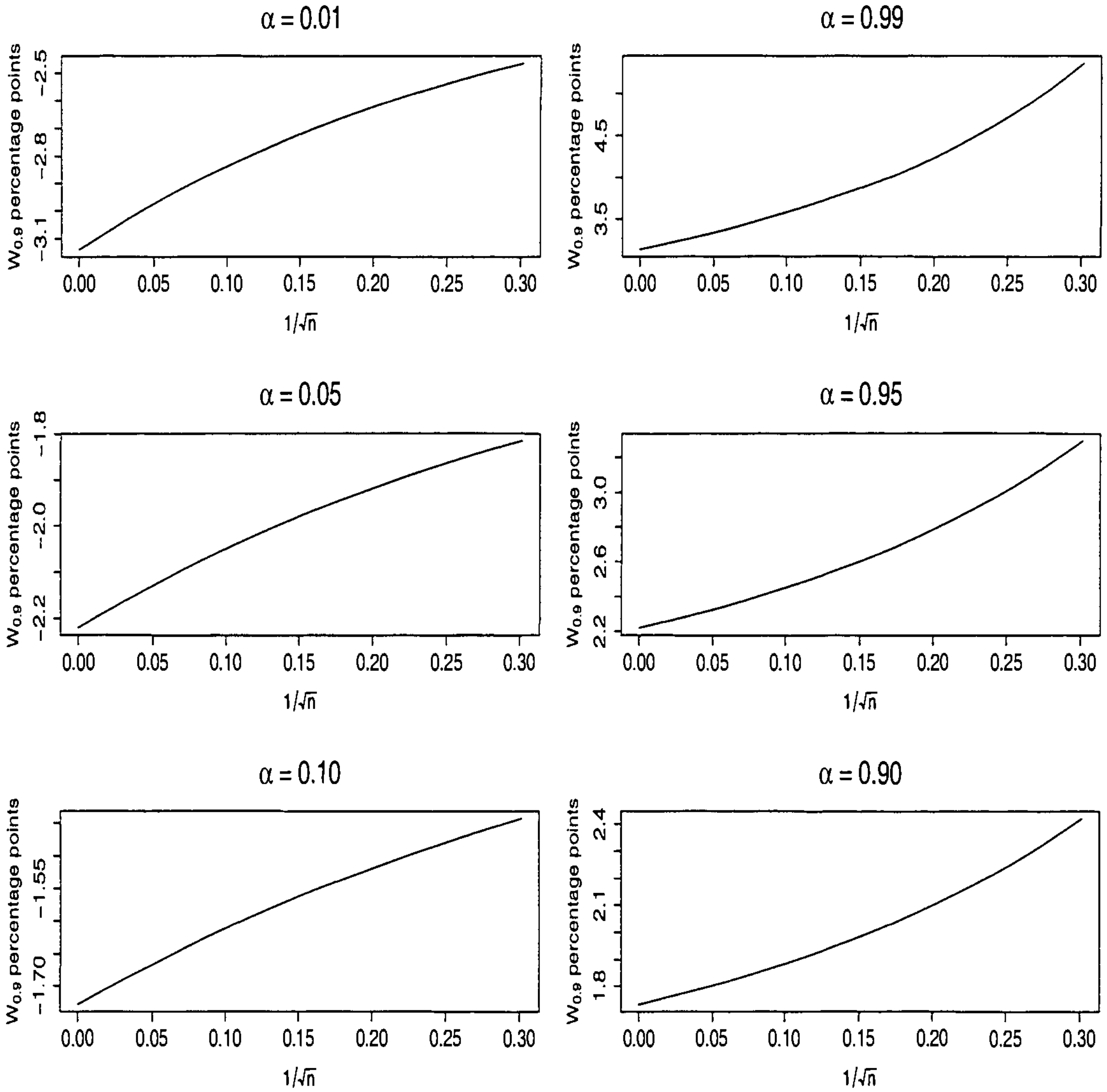

Figure 3.2: The percentage points of $W_{0.90}$ plotted against $1/\sqrt{n}$ at various levels of α.

	α					
Approximations	0.01	0.05	0.10	0.90	0.95	0.99
$a\left(1+\frac{b}{\sqrt{n}}\right)$	0.512216	0.114393	0.044139	0.254238	0.714793	3.953214
$a\left(1+\frac{b}{\sqrt{n}}+\frac{c}{n}\right)$	0.003480	0.000756	0.000295	0.004173	0.013536	0.101660
$a\left(1+\frac{b}{\sqrt{n}}+\frac{c}{n}+\frac{d}{n^2}\right)$	0.000223	0.000041	0.000016	0.000006	0.000008	0.000048

Table 3.2: Residual sums of squares for various approximations to the points of $W_{0.90}$ at six α levels.

Approximating the Curves of W_p

The following approximating function is considered:

$$w_{p,\alpha} \approx a\left(1+\frac{b}{\sqrt{n}}+\frac{c}{n}+\frac{d}{n^2}\right). \tag{3.6}$$

The value of a in this approximation is the intercept, $z_\alpha\sqrt{A_p}$, and the (b, c, d) co-efficients are obtained in the following way. Subtracting a from the percentage points, $w_{p,\alpha}$, we have

$$w_{p,\alpha} - a \approx \frac{b'}{\sqrt{n}}+\frac{c'}{n}+\frac{d'}{n^2}$$

The least-squares method is applied to obtain values for b', c', and, d'. These values are subsequently divided by a to obtain final values for the co-efficients as given in (3.6). Using four coefficients in the approximation was an empirical choice, motivated by first fitting approximations using two and three coefficients. In order to give some measure of the difference between the different expansions, the sum of squared residuals is given in Table 3.2.

The set of co-efficients (a, b, c, d) are obtained for a number of (p, α) combinations and are given in Table 3.3. The table gives values of p from 0.50 to 0.99, giving the co-efficients needed for the CI of an upper quantile, but the same table can be used to obtain a CI for a lower quantile. In particular, since $t'_{n-1,\alpha}(-\delta) = -t'_{n-1,1-\alpha}(\delta)$ and $z_p = -z_{1-p}$, it follows that $w_{p,\alpha} = -w_{1-p,1-\alpha}$.

Using Table 3.3, the procedure to calculate a CI for the quantile is straight-forward. For example, if a 95% CI for the 90^{th} quantile is required, we use the table and calculate $w_{0.90,0.025}$ and $w_{0.90,0.975}$ based on (3.6). The confidence limits are found using (3.4), where values for $\hat{y}_p$ and S_Y are computed from the data.

The approximations using (a, b, c, d) do very well, which we demonstrate in Figure 3.3 for $p = 0.90$ and various α-levels. On the plots, the fitted curves are indistinguishable from the true curves, deviating slightly only when the sample size is small. In Figure 3.4, the fitted expansions using only two (a, b) and three (a, b, c) coefficients are plotted along with the percentage points, again for $p = 0.90$ and various α-levels. Moving from two to three coefficients shows a definite improvement However, it is arguable whether moving from three to four coefficients is in fact necessary.

Large Samples

For very large sample sizes, $w_{p,\alpha}$ approaches $z_\alpha\sqrt{1+\frac{z_p^2}{2}}$, and in this case, the CI is approximately

$$\left(\hat{y}_p + z_{\frac{\alpha}{2}}\sqrt{1+\frac{z_p^2}{2}}\frac{S_Y}{\sqrt{n}}, \hat{y}_p + z_{1-\frac{\alpha}{2}}\sqrt{1+\frac{z_p^2}{2}}\frac{S_Y}{\sqrt{n}}\right).$$

Using the fact that $z_{1-\frac{\alpha}{2}} = -z_{\frac{\alpha}{2}}$, it can be written as

$$\left(\hat{y}_p + z_{\frac{\alpha}{2}}\sqrt{1+\frac{z_p^2}{2}}\frac{S_Y}{\sqrt{n}}, \hat{y}_p - z_{\frac{\alpha}{2}}\sqrt{1+\frac{z_p^2}{2}}\frac{S_Y}{\sqrt{n}}\right).$$

This large sample case gives the same result as the approximate Wald CI, described in next section.

p		α									
		0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
0.50	*a*	-2.576	-2.326	-1.960	-1.645	-1.282	1.282	1.645	1.960	2.326	2.576
	b	0.003	0.002	0.001	0.001	0.000	0.000	0.001	0.001	0.002	0.003
	c	1.844	1.559	1.186	0.912	0.653	0.653	0.912	1.186	1.559	1.844
	d	7.289	5.401	3.400	2.232	1.340	1.340	2.232	3.400	5.401	7.289
0.55	*a*	-2.586	-2.336	-1.968	-1.651	-1.287	1.287	1.651	1.968	2.336	2.586
	b	-0.168	-0.155	-0.136	-0.120	-0.103	0.104	0.122	0.139	0.160	0.175
	c	1.792	1.516	1.153	0.888	0.636	0.683	0.950	1.231	1.615	1.908
	d	5.420	3.913	2.347	1.454	0.791	1.928	3.058	4.505	6.947	9.230
0.60	*a*	-2.617	-2.363	-1.991	- 1.671	-1.302	1.302	1.671	1.991	2.363	2.617
	b	-0.337	-0.310	-0.271	-0.240	-0.206	0.207	0.241	0.274	0.315	0.344
	c	1.755	1.486	1.135	0.877	0.633	0.727	1.000	1.290	1.684	1.985
	d	3.652	2.518	1.361	0.732	0.289	2.551	3.917	5.648	8.540	11.214
0.65	*a*	-2.670	-2.411	-2.031	-1.705	-1.328	1.328	1.705	2.031	2.411	2.670
	b	-0.502	-0.461	-0.404	-0.356	-0.307	0.308	0.358	0.406	0.466	0.509
	c	1.732	1.472	1.131	0.880	0.643	0.782	1.063	1.360	1.765	2.075
	d	2.006	1.217	0.458	0.076	-0.164	3.201	4.804	6.816	10.153	13.208
0.70	*a*	-2.747	-2.481	-2.090	-1.754	-1.367	1.367	1.754	2.090	2.481	2.747
	b	-0.660	-0.607	-0.531	-0.469	-0.405	0.406	0.471	0.534	0.612	0.667
	c	1.726	1.473	1.140	0.897	0.665	0.849	1.137	1.442	1.857	2.175
	d	0.503	0.041	-0.350	-0.508	-0.563	3.870	5.705	7.994	11.772	15.232
0.75	*a*	-2.854	-2.577	-2.171	-1.822	-1.420	1.420	1.822	2.171	2.577	2.854
	b	-0.811	-0.746	-0.653	-0.578	-0.500	0.501	0.580	0.656	0.751	0.819
	c	1.737	1.490	1.166	0.927	0.700	0.927	1.222	1.535	1.961	2.284
	d	-0.827	-0.996	-1.065	-1.020	-0.918	4.550	6.611	9.182	13.396	17.271
0.80	*a*	-2.997	-2.707	-2.281	-1.914	-1.491	1.491	1.914	2.281	2.707	2.997
	b	-0.953	-0.878	-0.770	-0.682	-0.592	0.593	0.684	0.773	0.884	0.962
	c	1.767	1.523	1.205	0.970	0.747	1.015	1.318	1.639	2.077	2.410
	d	-1.982	-1.890	-1.677	-1.461	-1.226	5.250	7.531	10.371	15.001	19.234
0.85	*a*	-3.194	-2.884	-2.430	-2.039	-1.589	1.589	2.039	2.430	2.884	3.194
	b	-1.089	-1.003	-0.881	-0.782	-0.681	0.682	0.784	0.885	1.010	1.098
	c	1.817	1.575	1.260	1.027	0.806	1.114	1.425	1.755	2.205	2.552
	d	-2.954	-2.644	-2.197	-1.839	-1.495	5.965	8.463	11.568	16.629	21.176
0.90	*a*	-3.476	-3.139	-2.645	-2.220	-1.729	1.729	2.220	2.645	3.139	3.476
	b	-1.220	-1.125	-0.989	-0.880	-0.769	0.771	0.883	0.993	1.132	1.230
	c	1.889	1.648	1.333	1.100	0.878	1.227	1.548	1.888	2.352	2.706
	d	-3.752	-3.272	-2.635	-2.168	-1.739	6.718	9.430	12.795	18.274	23.247
0.95	*a*	-3.951	-3.568	-3.006	-2.523	-1.966	1.966	2.523	3.006	3.568	3.951
	b	-1.354	-1.250	-1.101	-0.982	-0.863	0.865	0.985	1.106	1.257	1.365
	c	1.995	1.752	1.433	1.198	0.972	1.365	1.697	2.050	2.532	2.900
	d	-4.415	-3.803	-3.022	-2.472	-1.979	7.554	10.502	14.156	20.086	25.439
0.99	*a*	-4.959	-4.478	-3.773	-3.166	-2.467	2.467	3.166	3.773	4.478	4.959
	b	-1.488	-1.375	-1.216	-1.088	-0.963	0.966	1.093	1.223	1.386	1.504
	c	2.106	1.863	1.544	1.308	1.080	1.547	1.895	2.266	2.767	3.145
	d	-4.602	-3.956	-3.140	-2.571	-2.070	8.442	11.646	15.608	22.057	27.921

Table 3.3: Coefficients (a, b, c, d) required to obtain percentage points $w_{p,\alpha}$ for W_p.

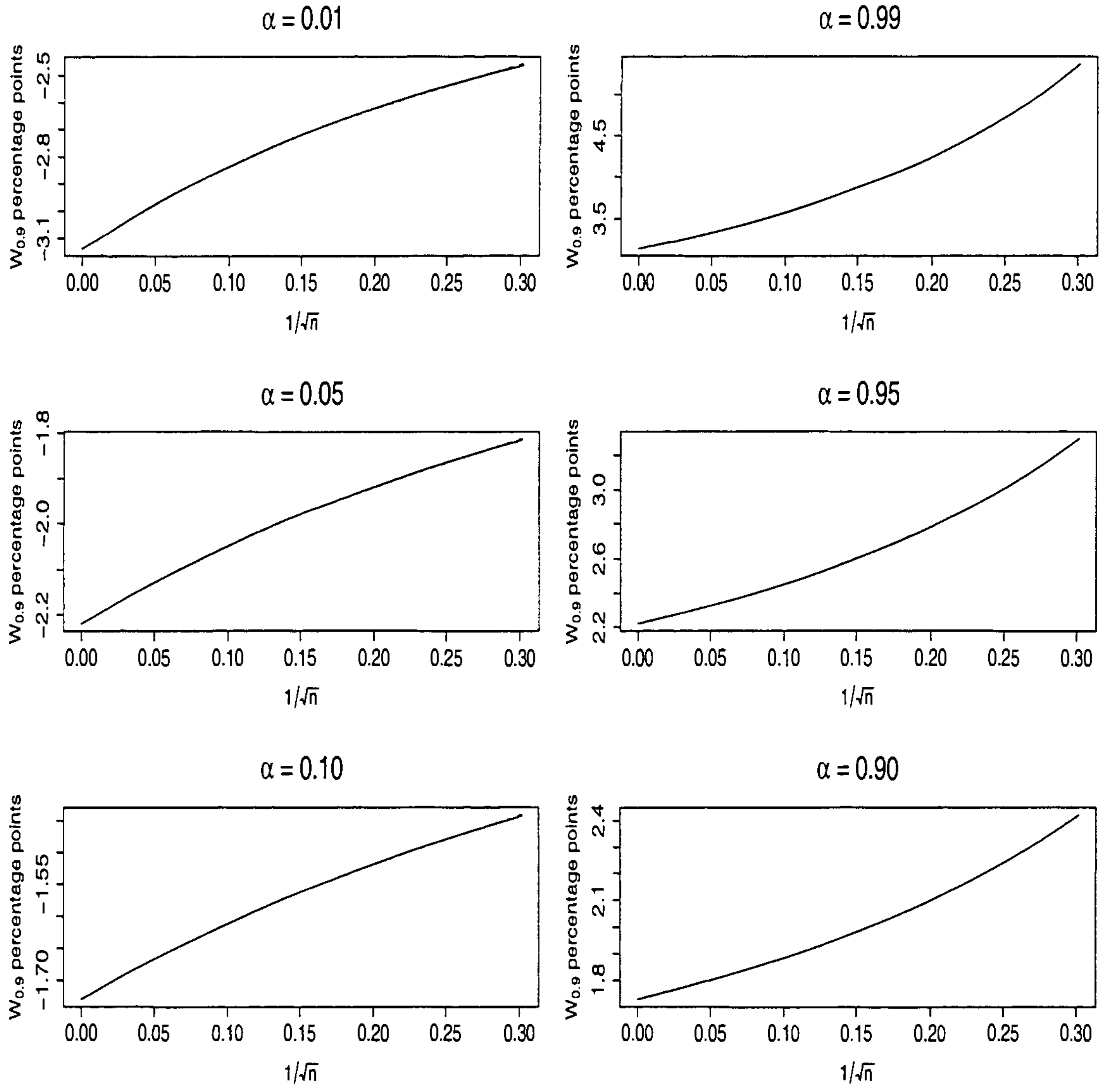

Figure 3.3: The percentage points for $W_{0.90}$, at various levels of α, plotted against $1/\sqrt{n}$ along with the fitted curves using the approximation given in (3.6). The solid line represents the percentage points, and the dashed line represents the fitted curve.

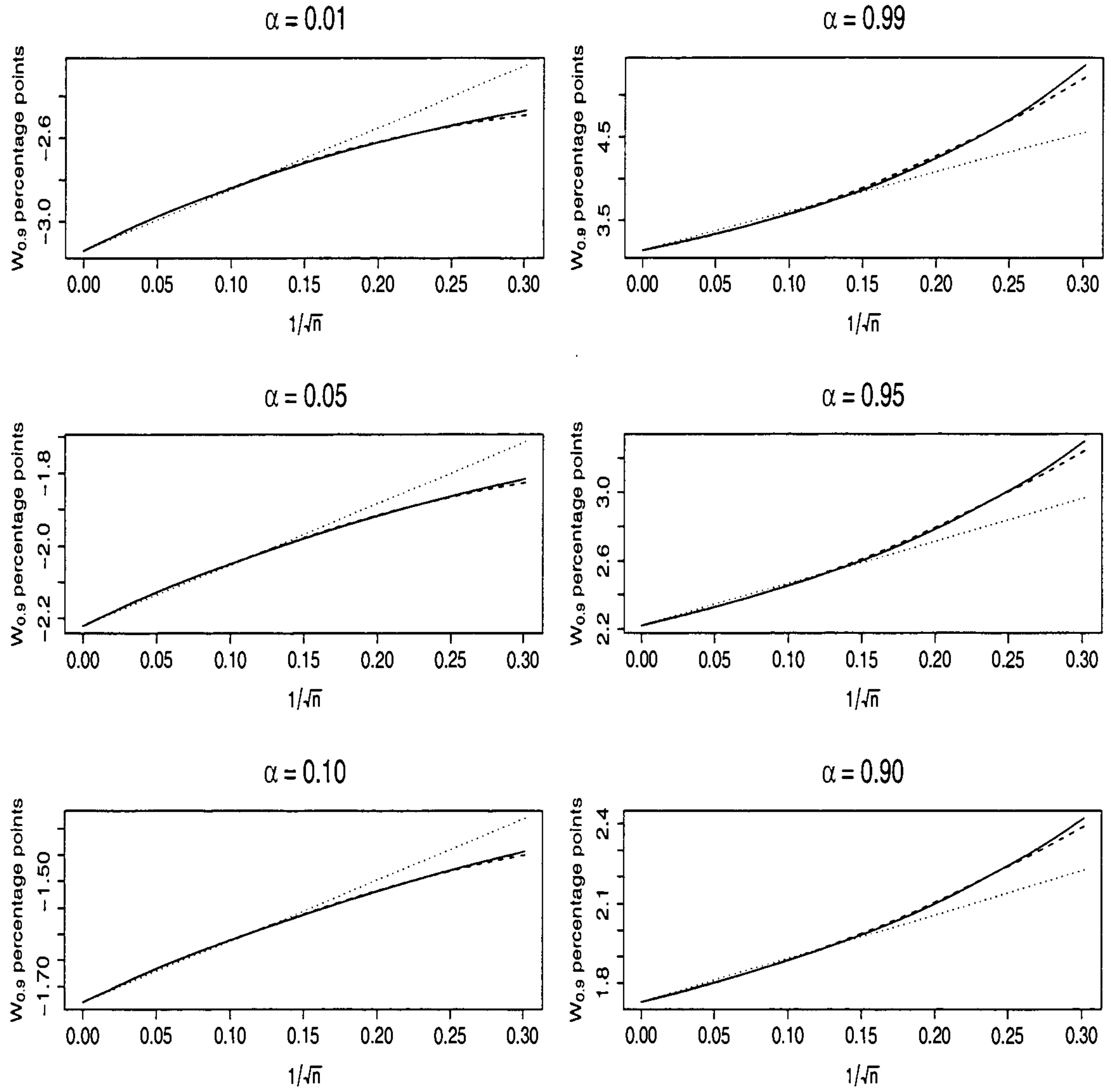

Figure 3.4: The percentage points for $W_{0.90}$ at various levels of α, plotted against $1/\sqrt{n}$ along with the fitted curves. The solid line represents the percentage points, the dashed line represents the fitted curve using three coefficients, and the dotted line represents the fitted curve using two coefficients.

3.2 Approximate Methods

The Wald Method

The Wald method is a well-known likelihood-based procedure for calculating an approximate CI and, under mild conditions, performs well in large samples. For a location-scale distribution or for a distribution which can be transformed to a location-scale distribution, the Wald CI is easy to compute for quantiles, and is therefore suitable for the exponential, Weibull, and lognormal distributions. Below we give the details, based on Lawless' (2003) general procedure for a location-scale distribution.

Let x_p be the quantile of a location-scale distribution with parameters u and b respectively, and let ω_p be the quantile of the same distribution, but with $u = 0$ and $b = 1$. Then $x_p = u + \omega_p b$, which we can estimate by $\hat{x}_p = \hat{u} + \omega_p \hat{b}$ using ML estimates $\hat{u}$ and $\hat{b}$. The pivotal quantity is

$$Z_p = \frac{\hat{x}_p - x_p}{se(\hat{x}_p)},$$

where

$$se(\hat{x}_p) = (v\hat{a}r(\hat{u}) + \omega_p^2 v\hat{a}r(\hat{b}) + 2\omega_p c\hat{o}v(\hat{u}, \hat{b}))^{1/2}. \tag{3.7}$$

Based on assumption of the asymptotic normality of ML estimates, Z_p is approximately $N(0,1)$. Let Z be a $N(0,1)$ random variable and z_α be the value such that $P(Z < z_\alpha) = \alpha$. A Wald $100(1-\alpha)\%$ CI for x_p is given by

$$\left(\hat{x}_p + z_{\frac{\alpha}{2}} se(\hat{x}_p), \hat{x}_p - z_{\frac{\alpha}{2}} se(\hat{x}_p)\right)$$

Let $\theta = (\theta_1, \theta_2)' = (u, b)'$. The variance and covariance terms in (3.7) come from the asymptotic covariance matrix for $\hat{\theta}$, which is the inverse of Fisher's observed information matrix evaluated at $\hat{\theta}$. Let $l(\theta)$ denote the log-likelihood function and $I(\theta)$ denote Fisher's observed information matrix. The $(i,j)^{th}$ entry in $I(\theta)$, is $-\frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j}$ $i, j = 1, 2$.

The inverse of $I(\hat{\theta})$ gives the variance and covariance terms: the diagonal elements of $I(\hat{\theta})^{-1}$ give the variances and the off-diagonal elements give the covariances.

For the exponential distribution with only a scale parameter, β, $I(\hat{\beta}) = \frac{n}{\hat{\beta}^2}$, and therefore $se(\hat{\beta}) = \frac{\hat{\beta}}{\sqrt{n}}$. Since $\omega_p = -\log(1-p)$ for the exponential, the $se(\hat{x}_p) = se(\omega_p\hat{\beta}) = \frac{\hat{x}_p}{\sqrt{n}}$.

For the lognormal and Weibull distributions, the standard errors are expressed in terms of their location-scale counterparts, the normal and extreme value distributions.

For the normal distribution,

$$I(\hat{\mu}, \hat{\sigma}) = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{2n}{\hat{\sigma}^2} \end{pmatrix}.$$

Therefore, $se(\hat{y}_p) = \sqrt{\frac{\hat{\sigma}^2}{n}\left(1 + \frac{\omega_p^2}{2}\right)}$ where ω_p is the p^{th} quantile for the standard normal distribution.

The standard error for the quantiles of the extreme value distribution cannot be written explicitly. Fisher's observed information matrix, evaluated at $\hat{\gamma}$ and $\hat{\eta}$, is

$$I(\hat{\gamma}, \hat{\eta}) = \begin{pmatrix} \frac{n}{\hat{\gamma}^2} & \frac{\sum(Z_i \exp(Z_i))}{\hat{\gamma}^2} \\ \frac{\sum(Z_i \exp(Z_i))}{\hat{\gamma}^2} & \frac{n}{\hat{\gamma}^2} + \frac{\sum(Z_i^2 \exp(Z_i))}{\hat{\gamma}^2} \end{pmatrix},$$

where $Z_i = \frac{Y_i - \hat{\gamma}}{\hat{\eta}}$. Using $\omega_p = \log(-\log(1-p))$ for the extreme value distribution and the variance-covariance terms from $I(\hat{\gamma}, \hat{\eta})^{-1}$, $se(\hat{y}_p)$ can be calculated as in (3.7)

The Bootstrap Method

Another method of approximating the distribution of $\hat{x}_p$ is the bootstrap. There are two ways to perform the bootstrap: the parametric and the non-parametric (Lawless 2003). An example of the parametric bootstrap has already been described in Chapter 2, where we generated p-values. We give the steps again, here in the context of a CI for the p^{th} quantile.

1. Generate a X'-set of size n from $F(x;\hat{\theta})$.
2. Calculate $\hat{\theta}'$ based on the X'-set.
3. Calculate $\hat{x}_p'$.
4. Repeat steps 1-3 M times where M is large.

At the end of step 4, M estimates of x_p will provide an approximation to its distribution, and the limits of the CI can easily be obtained. For example, if $M = 10000$, a 95% CI for x_p is obtained by first putting the 10000 estimates of x_p into ascending order and then taking the 250^{th} and 9750^{th} values from the ordered set of $\hat{x}_p'$ values.

For the non-parametric bootstrap, the procedure is the same, except for the first step. Instead of generating an X'-set from $F(x;\hat{\theta})$, we resample with replacement from the X-set to obtain an X'-set. In other words, instead of sampling from $F(x;\hat{\theta})$, we sample from the EDF.

3.3 Application to the Data Sets

We now apply these methods to the two data sets. In Tables 3.4 and 3.5, 95% confidence intervals are given for the median and three upper quantiles for each of the four fitted distributions. The methods used are given in brackets. The two bootstrap methods are differentiated by P (parametric) and NP (non-parametric).

In the LC data set, comparing exact and approximate methods, there is almost exact agreement, indicating that the approximations do very well when the sample size is large. In the CC data set, though the exact and approximate intervals match well at the median, the discrepancy increases in the upper quantiles.

The degree of similarity in the point estimates (see Table 1.4) is also reflected here. The confidence intervals for the LC data set give very similar results for the Weibull, gamma, and lognormal distributions. For the CC data set, though there is a large degree of overlap, the similarity in confidence limits is not as strong.

Connection Between the CI and the GoF Test

It is interesting to compare the resulting confidence intervals in light of the GoF test results in Chapter 2.

Consider first the results in Table 3.4, based on the LC data set. The GoF results in Table 2.1 report only the lognormal distribution as an adequate fit (p-value > 0.05). Yet the confidence intervals under both the Weibull and gamma assumptions are very similar to the results based on the lognormal assumption. For example, all three models give roughly 30 to 36 months as the CI for the 90^{th} quantile. The choice of the Weibull, gamma, or lognormal distribution does not seem to matter, at least in the case of a CI for the 90^{th} quantile.

For the CC data set, a very different picture arises. The GoF results in Table 2.1 report a p-value above 0.05 for the Weibull, gamma, and lognormal distributions. However, in this case, the degree of overlap in the intervals for the upper quantiles is much lower (see Table 3.5). For example, the Weibull and gamma models estimate the 90^{th} quantile to be between 59 and 92 months and 68 and 95 months respectively, while the lognormal model estimates it roughly between 60 and 125 months. This discrepancy increases for the 95^{th} and 99^{th} quantiles. The conclusions from confidence intervals for the upper quantiles is therefore dependent on which model is chosen.

Consider the scenario where the statistician is given survival times (with no censoring) and is asked to see if the lognormal distribution gives a good fit to the data. If a p-value > 0.05 is obtained, the statisician goes on to make inference. What happens when the data are in fact from another distribution? Does the assumption of lognormality produce misleading results? In particular, how wrong are the conclusions from a

	Value of p for p^{th} quantile, x_p			
Distribution	0.50	0.90	0.95	0.99
EXP(Exact)	(11.91, 15.90)	(39.55, 52.82)	(51.46, 68.73)	(79.10, 105.65)
EXP(Wald)	(11.71, 15.67)	(38.90, 52.04)	(50.61, 67.70)	(77.79, 104.07)
EXP(P Bootstrap)	(11.76, 15.77)	(39.14, 52.13)	(51.03, 67.85)	(78.17, 104.69)
EXP(NP Bootstrap)	(12.67, 14.78)	(42.05, 49.04)	(54.76, 64.01)	(84.11, 98.23)
WB(Wald)	(17.11, 20.28)	(31.57, 36.63)	(35.87, 41.97)	(44.01, 52.63)
WB(P Bootstrap)	(17.09, 20.26)	(31.39, 36.57)	(35.57, 41.93)	(43.54, 52.74)
WB(NP Bootstrap)	(17.30, 20.06)	(30.42, 37.46)	(34.32, 43.12)	(41.61, 54.50)
G(P Bootstrap)	(16.94, 19.54)	(30.07, 35.36)	(34.48, 41.11)	(43.87, 53.65)
G(NP Bootstrap)	(16.93, 19.56)	(29.44, 36.00)	(33.70, 41.99)	(42.58, 54.95)
LN(Exact)	(16.23, 18.75)	(30.11, 36.60)	(35.64, 44.51)	(48.76, 64.45)
LN(Wald)	(16.24, 18.75)	(29.92, 36.31)	(35.36, 44.06)	(48.21, 63.54)
LN(P Bootstrap)	(16.24, 18.73)	(29.92, 36.24)	(35.34, 43.99)	(48.09, 63.51)
LN(NP Bootstrap)	(16.25, 18.73)	(29.78, 36.31)	(35.11, 44.08)	(47.72, 63.59)

Table 3.4: 95% confidence intervals, using various methods, for the middle and upper quantiles of the fitted distributions to the LC data set

CI for a quantile when the data are assumed to be lognormal but are in fact distributed as either Weibull or gamma? The issue is one of robustness, and this is examined for the lognormal distribution in Chapter 4.

	Value of p for p^{th} quantile, x_p			
Distribution	0.50	0.90	0.95	0.99
EXP(Exact)	(20.14, 38.19)	(66.89, 126.87)	(87.03, 165.06)	(133.79, 253.73)
EXP(Wald)	(18.43, 35.62)	(61.23, 118.32)	(79.67, 153.94)	(122.47, 236.65)
EXP(P Bootstrap)	(19.19, 36.00)	(63.86, 120.73)	(82.44, 156.42)	(127.61, 240.68)
EXP(NP Bootstrap)	(21.42, 32.85)	(71.37, 109.93)	(92.80, 141.16)	(142.21, 218.11)
WB(Wald)	(26.61, 43.63)	(60.23, 93.53)	(70.38, 113.15)	(89.61, 156.45)
WB(P Bootstrap)	(26.51, 43.00)	(58.69, 91.13)	(68.55, 110.54)	(86.53, 150.99)
WB(NP Bootstrap)	(26.06, 43.23)	(60.32, 87.88)	(71.04, 104.75)	(92.83, 140.56)
G(P Bootstrap)	(25.46, 41.33)	(57.82, 95.21)	(69.88, 117.65)	(93.78, 168.99)
G(NP Bootstrap)	(25.46, 41.05)	(59.57, 90.95)	(72.32, 111.06)	(100.39, 156.93)
LN(Exact)	(22.78, 38.63)	(61.48, 126.46)	(79.49, 180.99)	(127.14, 358.81)
LN(Wald)	(22.98, 38.30)	(58.84, 117.24)	(75.16, 164.55)	(117.59, 314.38)
LN(P Bootstrap)	(22.93, 38.07)	(58.71, 117.38)	(74.78, 164.94)	(117.10, 313.59)
LN(NP Bootstrap)	(22.81, 37.85)	(62.26, 103.05)	(81.35, 140.32)	(132.19, 254.68)

Table 3.5: 95% confidence intervals, using various methods, for the middle and upper quantiles of the fitted distributions to the CC data set

Chapter 4

Model Robustness

A desirable feature of a statistical model is robustness. That is, under the assumed model, subsequent inference procedures perform well, even when there are departures from the assumed model. In this chapter, we examine the robustness of the lognormal assumption in connection with confidence intervals for the quantiles.

Some work in this area is as follows. Lefante Jr. and Shah (2002) examine various CI methods for the mean of a lognormal distribution. The coverage probabilities and interval widths for the various methods are compared for several lognormal distributions and also for some gamma alternatives. Modarres, Nayak, and Gastwirth (2002) examine the performance of upper quantile estimation for several distributions. The distributions include the lognormal, the log-logistic, and the log-double exponential. Confidence intervals and coverage probability (CP) are not examined.

We investigate the performance of the CP for the quantile confidence intervals and, in addition, we relate the p-value from a GoF test to the CP. The performance is first examined when the distribution or parent population is indeed lognormal. We then look at the performance of the CP calculated under the lognormal assumption when the parent population is either Weibull or gamma. Therefore, three simulation studies are conducted.

Outline of the Simulations

The general outline for each simulation is as follows. An x-set is generated from the parent distribution and is assumed to be lognormal. The y-set ($y_i = \log x_i$) is calculated, and a GoF test for normality is performed for the y-set using the Anderson-Darling statistic A^2. A p-value is obtained using the tables by Stephens (1986, Tables 4.7 and 4.9). A 95% CI for the quantile of interest is calculated by fitting the lognormal distribution and using the coefficients in Table 3.3. An indicator variable records whether or not the CI covers the quantile of the true parent distribution.

The two factors that are varied are the quantile and sample size. Two quantiles are considered: the 50^{th} and 90^{th}, denoted $x_{0.5}$ and $x_{0.9}$ respectively. Fifteen sample sizes are considered: 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 50, 100, 200, and 500. For each combination of factors, 10000 runs were made.

For a location-sacle distribution, variation in location and/or scale parameters does not change the CP of the CI procedure, calculated under the assumption the data are normally distributed.. This is shown below for a location-scale distribution and since the Weibull and lognormal distributions are transformed to location-scale distributions in order to analyze the data, there is no need to vary the parameters for these or the exponential distribution. For the gamma distribution, changes in the shape parameter, but not the scale, will affect the CP. Therefore, in addition to varying the quantile and sample size, the shape parameter is also a factor in the simulation when the parent population is gamma.

Invariant Coverage Probabilities

Suppose Y follows a location-scale distribution with location parameter u and scale parameter b; then $Z = \frac{Y-u}{b}$ follows the standard distribution with $u = 0$ and $b = 1$. Let y_p and z_p denote the respective p^{th} quantiles, related by $y_p = u + bz_p$.

Given a z-set, a y-set is calculated using $y_i = u + bz_i$ for any u and b. The y-set is a now a sample from the distribution with parameters u and b. Note that $\overline{y} = u + b\overline{z}$ and $s_y = bs_z$. Under the normal assumption, the z-set is $N(0, 1)$ and the y-set is $N(u, b^2)$. A CI for the p^{th} quantile is a function of the sample mean and standard deviation. That is, the CI for z_p can be written as $(\overline{z}+Ls_z, \overline{z}+Us_z)$ for appropriate values of L and U which do not depend on the parameters. Similarly, a CI for y_p is given by $(\overline{y} + Ls_y, \overline{y} + Us_y)$ for the same values of L and U.

Notice that $\overline{z}+Ls_z < z_p < \overline{z}+Us_z$ if and only if $u+b(\overline{z}+Ls_z) < u+bz_p < u+b(\overline{z}+Us_z)$ if and only if $\overline{y} + Ls_y < y_p < \overline{y} + Us_y$. Therefore, z_p is inside the CI calculated from the z-set if and only if y_p is inside the CI calculated from the y-set.

Hence, coverage probabilities under the normal assumption are invariant to location and/or scale changes. A similar argument can be made for the scale parameter of the gamma distribution.

4.1 Parent Population: Lognormal

Since the coverage probabilities are invariant, without loss of generality, we generate samples from the lognormal distribution with $\mu = 0$ and $\sigma^2 = 1$. Confidence intervals are calculated for both $x_{0.5}$ and $x_{0.9}$.

The overall CP is plotted against $1/n$ for both quantiles in Figure 4.1. The 95% target CP is maintained regardless of sample size as expected.

The CP for each quantile is now examined in the light of the p-values resulting from the GoF test. The p-values have been divided into three categories: p-value < 0.05, $0.05 <$ p-value < 0.10, and p-value > 0.10.

The coverage probabilities for the confidence intervals are plotted against $1/n$ for each quantile in Figures 4.2 and 4.3 respectively, separated by the p-value categories. In both cases, the results appear more variable if the p-value is below 0.10. However, this increased variability is due to the simulation error and not to the p-value itself. In

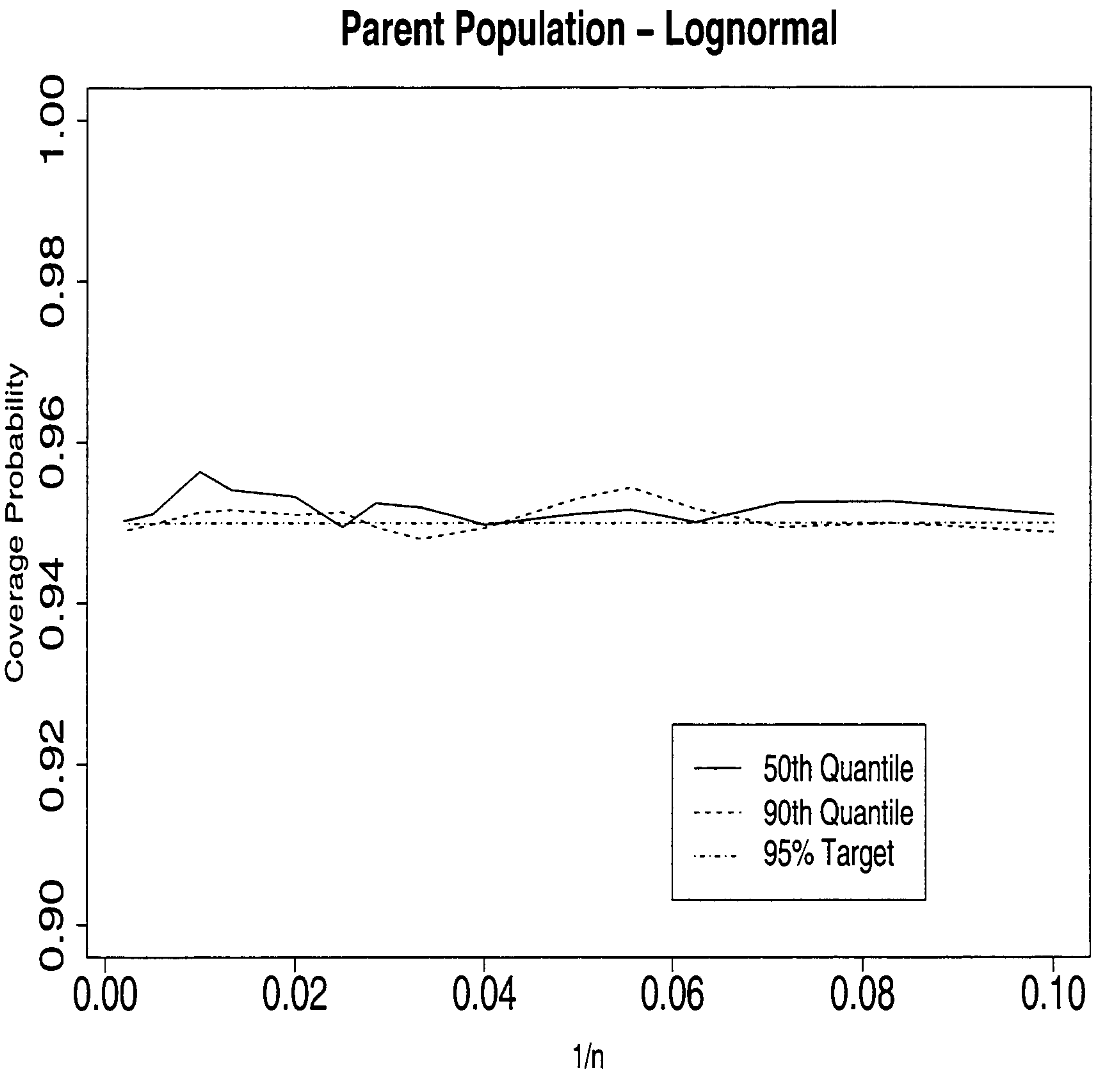

Figure 4.1: The true coverage probability of a 95% CI for $x_{0.5}$ and $x_{0.9}$ calculated under the lognormal assumption when the parent population is lognormal.

both plots, there is no visible connection between the CP and the p-value, as for each category the CP remains very close to the 95% target CP.

The results provide reassurance that, under the correct model, the CI gives an accurate coverage probability as expected. In addition, it indicates that there is no connection between the p-value and the performance of the CI procedure. This lack of connection can be explained by Basu's theorem, which states that if a complete and minimal sufficient statistic exists, then it is independent of any other ancillary statistic (Casella and Berger 2002). In this case, the minimal sufficient statistic is $(\overline{y}, s_y^2)$ and the ancillary statistic is A^2. Since the confidence limits are functions of the minimal sufficient statistic, and the p-value is a function of A^2, the CI, and therefore the CP, are independent of the p-value.

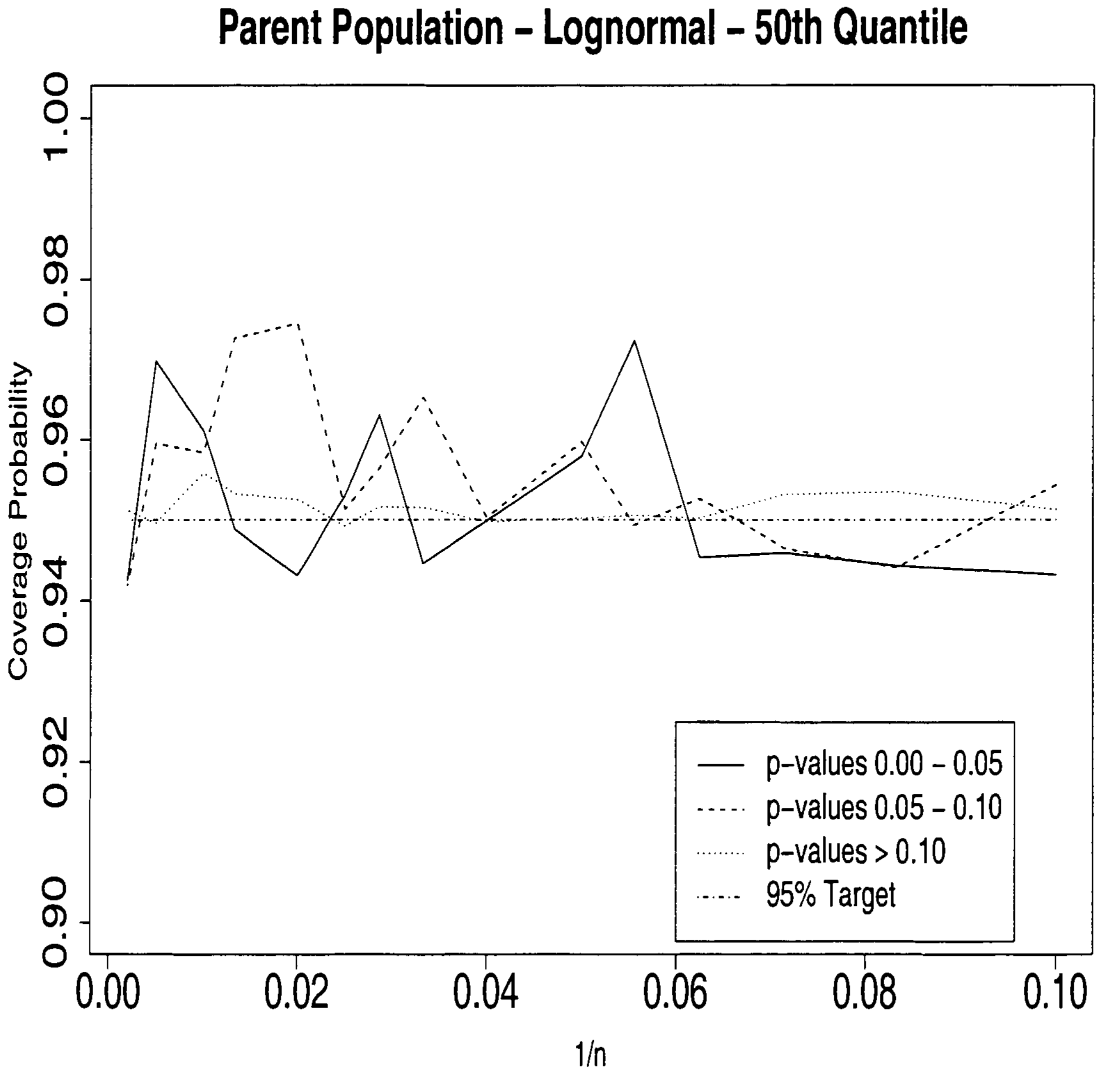

Figure 4.2: The true coverage probability of a 95% CI for $x_{0.5}$ calculated under the lognormal assumption when the parent population is lognormal. The CP is separated by the p-value from a GoF test for lognormality.

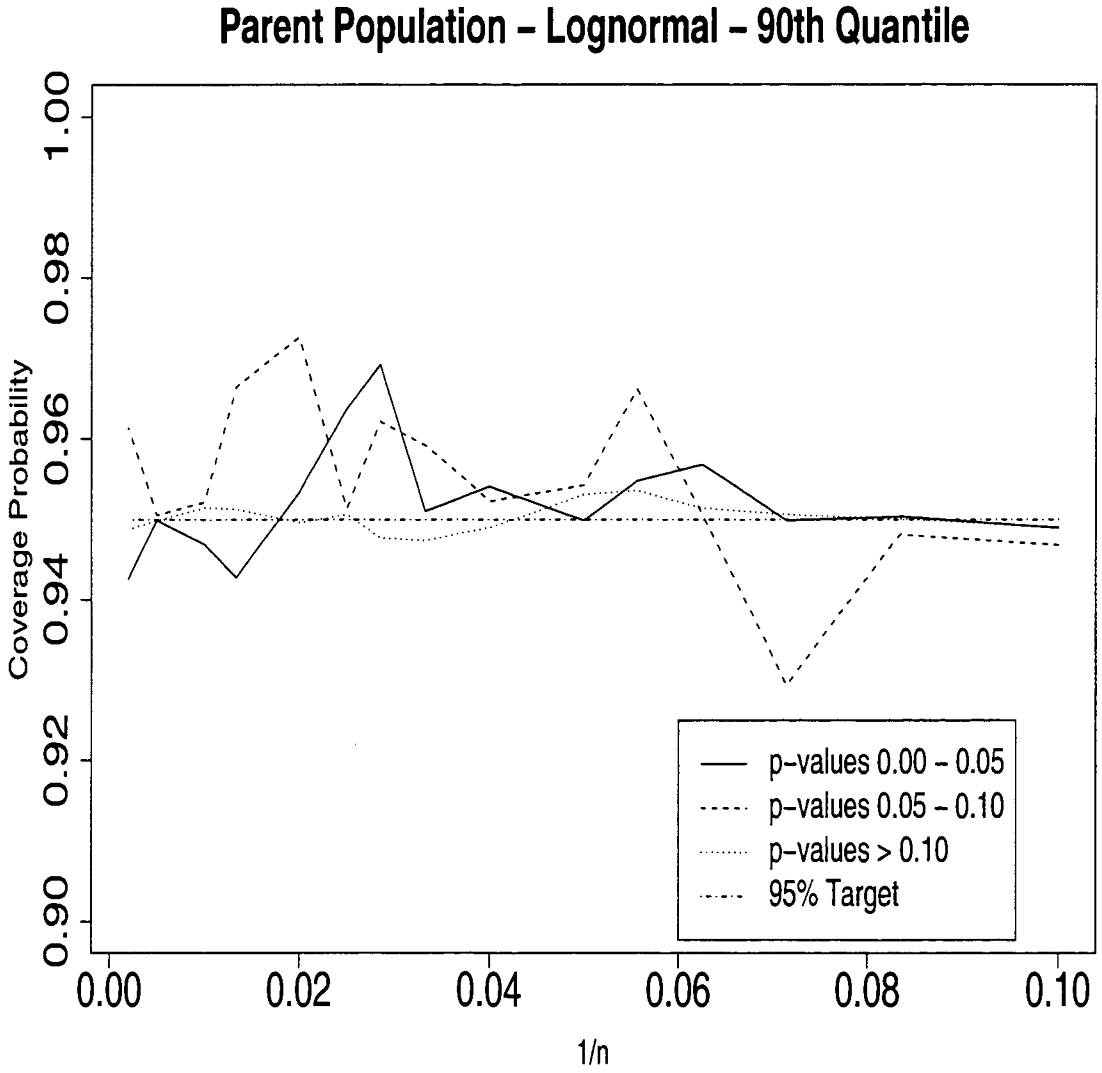

Figure 4.3: The true coverage probability of a 95% CI for $x_{0.9}$ calculated under the lognormal assumption when the parent population is lognormal. The CP is separated by the p-value from a GoF test for lognormality.

4.2 Parent Population: Weibull

Since the coverage probabilities are invariant, without loss of generality, we generate samples from the Weibull distribution with $\alpha = \beta = 1$. Under the lognormal assumption, 95% confidence intervals are calculated for $x_{0.5}$ and $x_{0.9}$.

Coverage Probability

In Figure 4.4 the CP for each quantile is plotted against $1/n$ and the target coverage of 95% is also provided. The plot shows that for $n < 50$, the procedure maintains at least 80% coverage for both quantiles, and for $n < 20$, the procedure gives close to 95% coverage. In fact, the coverage for $x_{0.9}$ exceeds 95% for $n < 20$ and is higher than the coverage for $x_{0.5}$ for all sample sizes. The high CP in small samples can be attributed to the larger width of the interval. Insistence on fitting the lognormal distribution in small samples under the Weibull alternative promises a good CP, but the width of the interval may be too wide to be useful. Larger samples yield narrower intervals but at the expense of decreasing the CP.

Coverage Probability and p-values

The CP of the CI for $x_{0.5}$ is plotted against $1/n$ in Figure 4.5, now separated by the p-value. There is little variation among the separate curves, and they follow the pattern for the quantile shown in Figure 4.4. In general, there appears to be no obvious connection between the p-value and the CP.

Interestingly, for small sample sizes, when the fit is bad (p-value < 0.05), the CP is slightly higher than when the fit is good. At the same time, for small sample sizes, the power of the GoF test is low, and therefore most p-values will be greater than 0.05. This is demonstrated in Figure 4.6 where the percentage of p-values below 0.05 and below 0.10 is plotted against $1/n$. In other words, it gives the power of the GoF test for

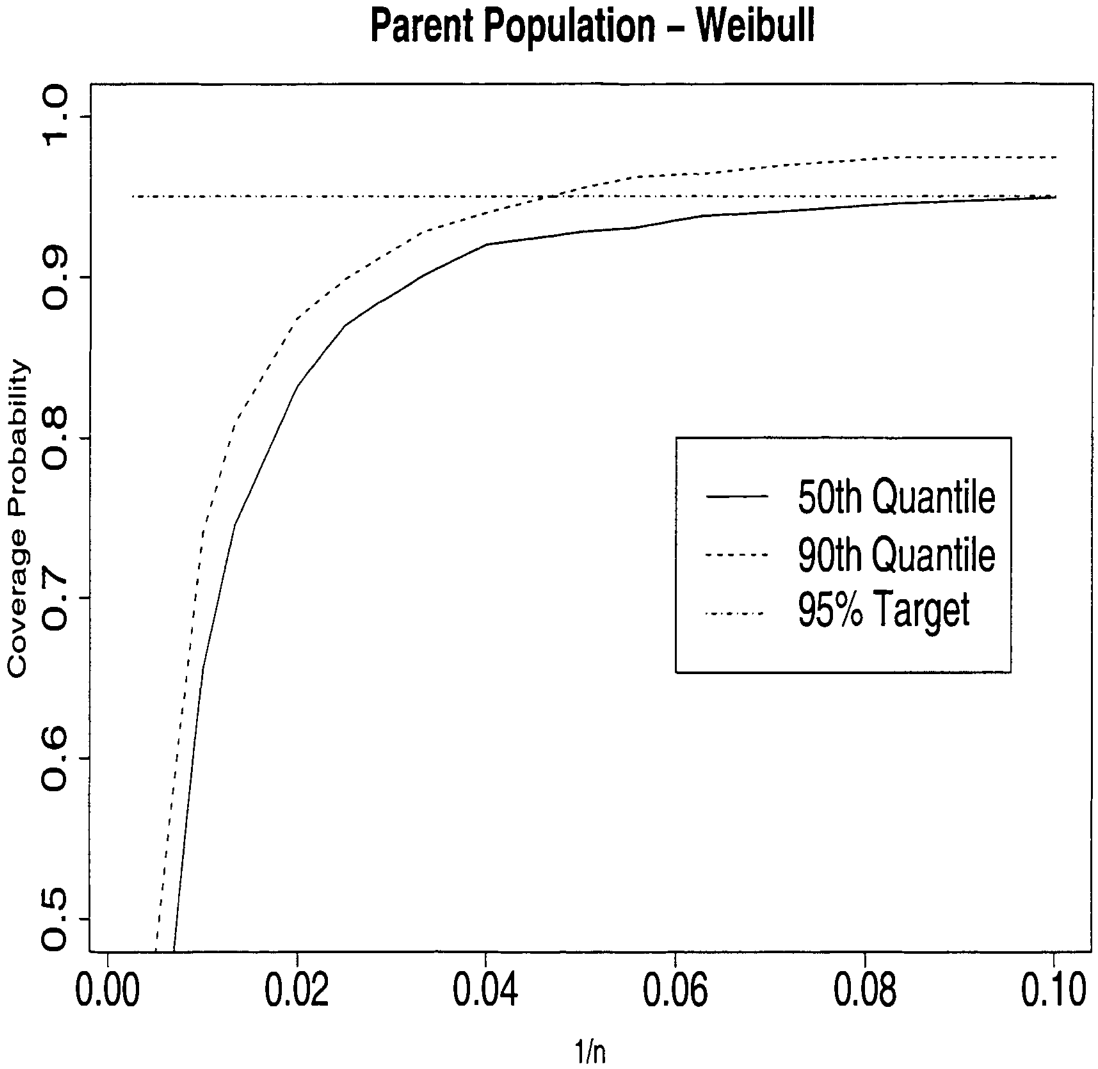

Figure 4.4: The true coverage probability of a 95% CI for $x_{0.5}$ and $x_{0.9}$ calculated under the lognormal assumption when the parent population is Weibull.

lognormality against the Weibull alternative at the two levels of significance. It will be the same regardless of which quantile is being considered. For example, with a sample size of 20 the GoF test will reject H_0 about 30% of the time at the 5% significance level and about 40% at the 10% significance level.

The CP for the CI for $x_{0.9}$ is plotted against $1/n$ in Figure 4.7, but here the results are much different. In this case, the CP is very good for p-values > 0.05, regardless of sample size, but deteriorates as sample size increases for p-values less than 0.05. In other words, if the GoF test yields an acceptable p-value for a lognormal fit even though the true distribution is Weibull, the CP for the 95% CI for $x_{0.9}$ still gives roughly 95% coverage.

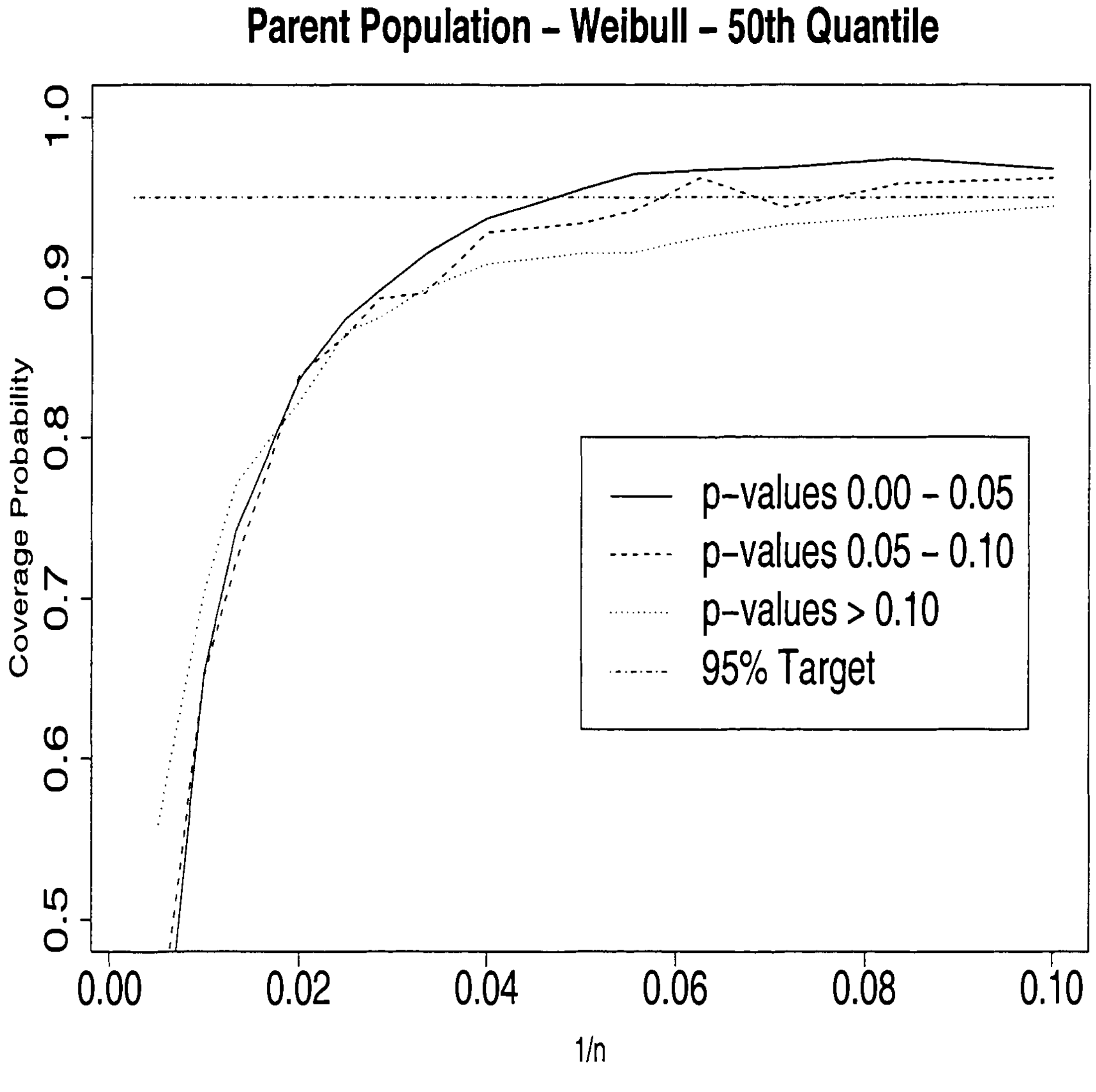

Figure 4.5: The true coverage probability of a 95% CI for $x_{0.5}$ calculated under the lognormal assumption when the parent population is Weibull. The CP is separated by the p-value from a GoF test for lognormality.

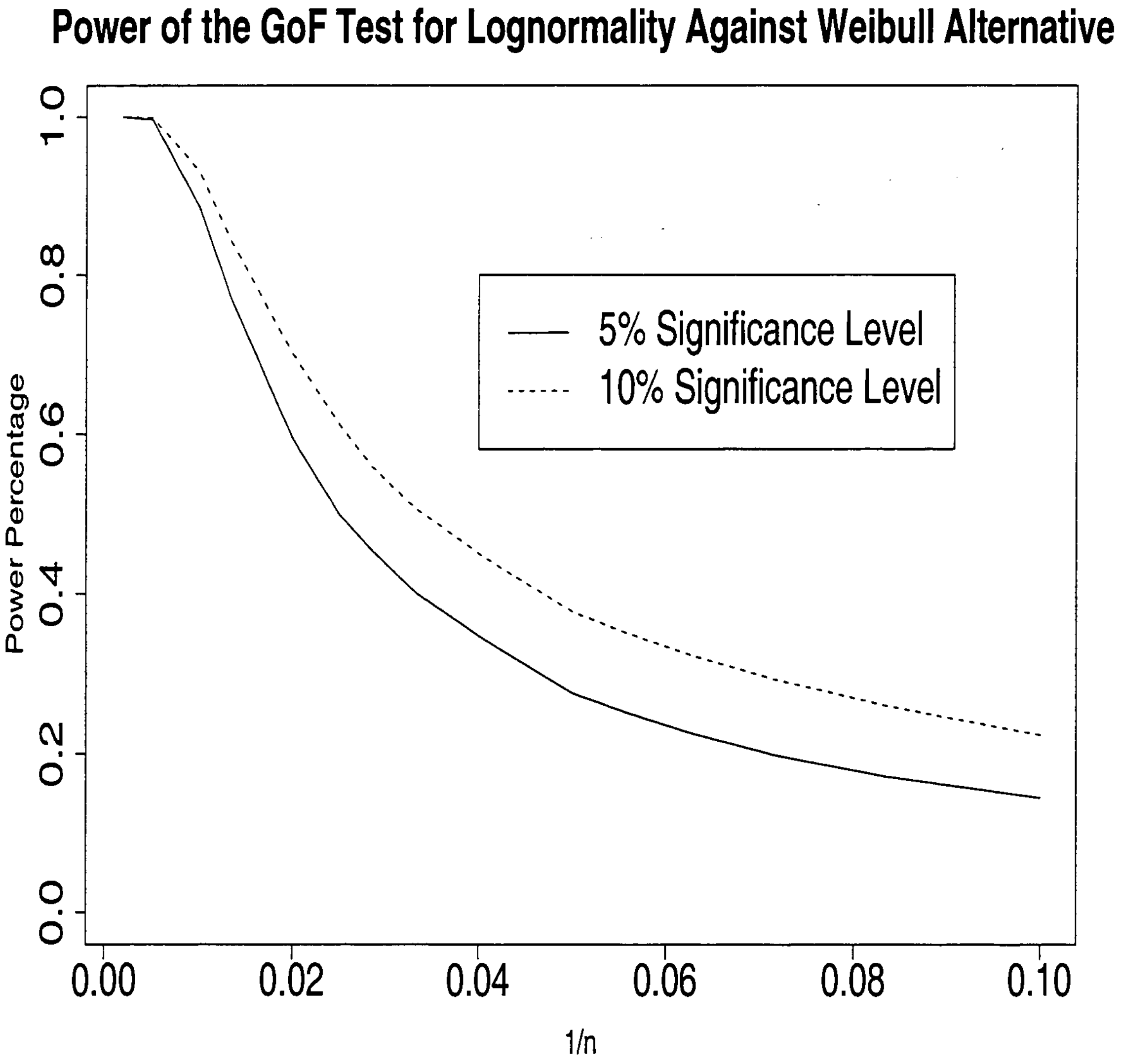

Figure 4.6: Percentage of p-values which fall below either 0.05 or 0.10 for a specific sample size.

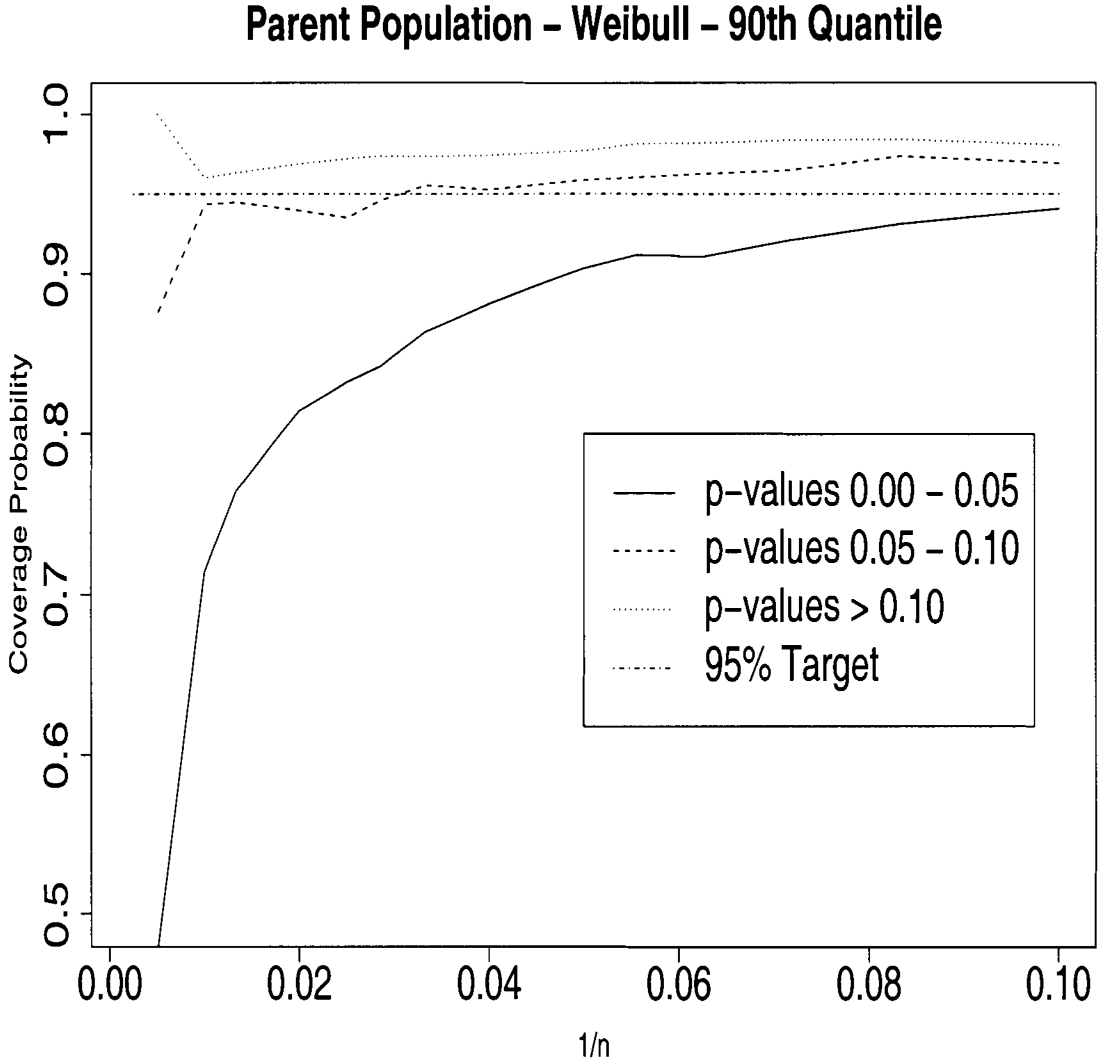

Figure 4.7: The true coverage probability of a 95% CI for $x_{0.9}$ calculated under the lognormal assumption when the parent population is Weibull. The CP is separated by the p-value from a GoF test for lognormality.

4.3 Parent Population: Gamma

The shape parameter for the gamma alternative is also varied, in addition to the two quantiles and fifteen sample sizes. Samples are generated from the gamma with $\beta = 1$ and $\alpha = 0.10, 0.25, 0.50, 0.75, 1, 2, 5,$ and 10.

Coverage Probability

The overall CP is plotted in Figures 4.8 and 4.9 for $x_{0.5}$ and $x_{0.9}$, respectively. The solid line gives the coverage probability when the shape parameter $\alpha = 1$. The dashed lines to the right of the solid line correspond to decreasing values of α. The CP becomes successively worse as α decreases. The dotted lines to the left of the solid line correspond to increasing values of α. The CP becomes successively better as α increases. The CP for the two quantiles gives similar results, though the plots show the confidence interval for $x_{0.9}$ has a slightly higher CP.

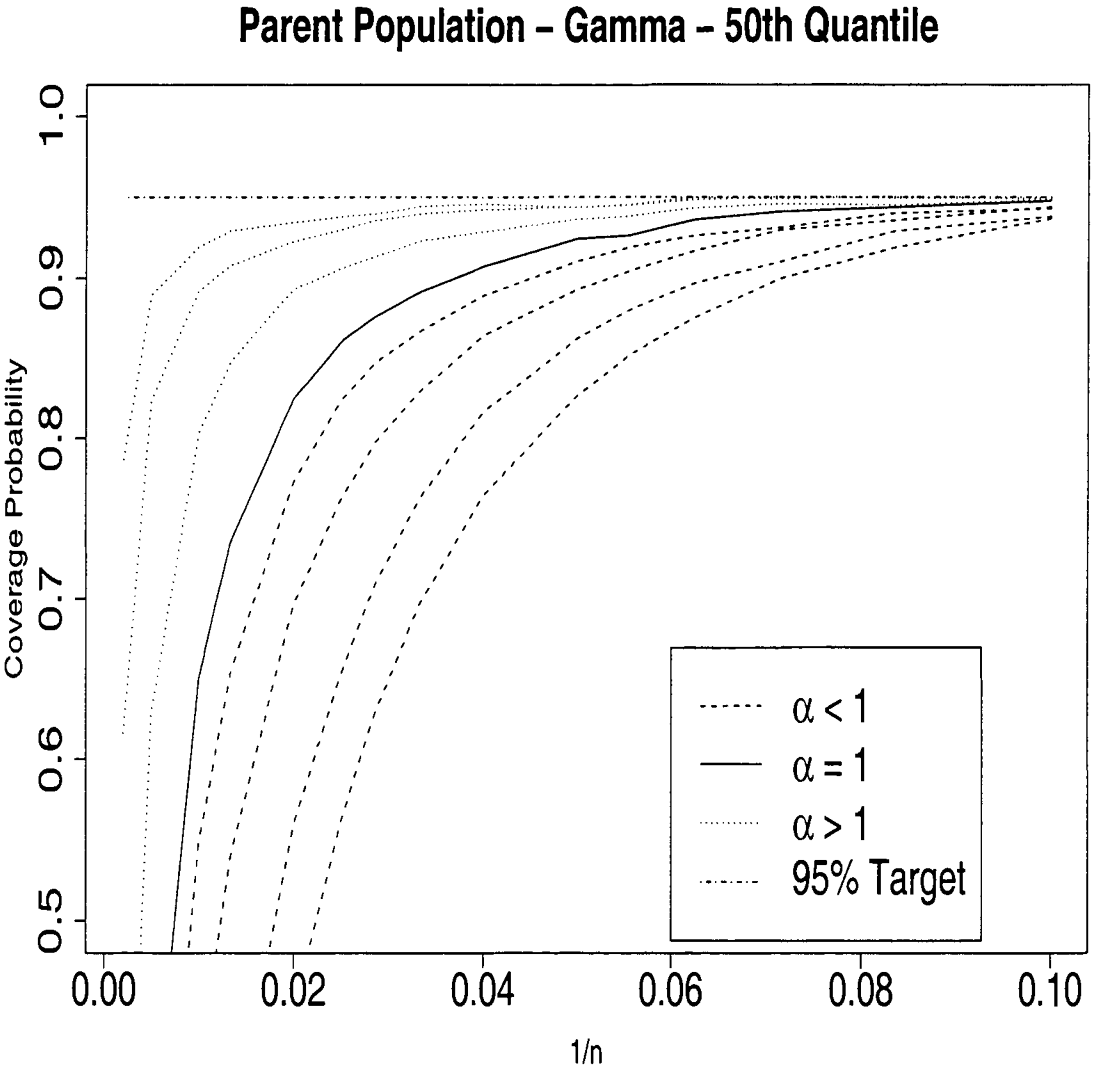

Figure 4.8: The true coverage probability of a 95% CI for $x_{0.5}$ calculated under the lognormal assumption when the parent population is gamma. The CP is separated by the shape parameter α.

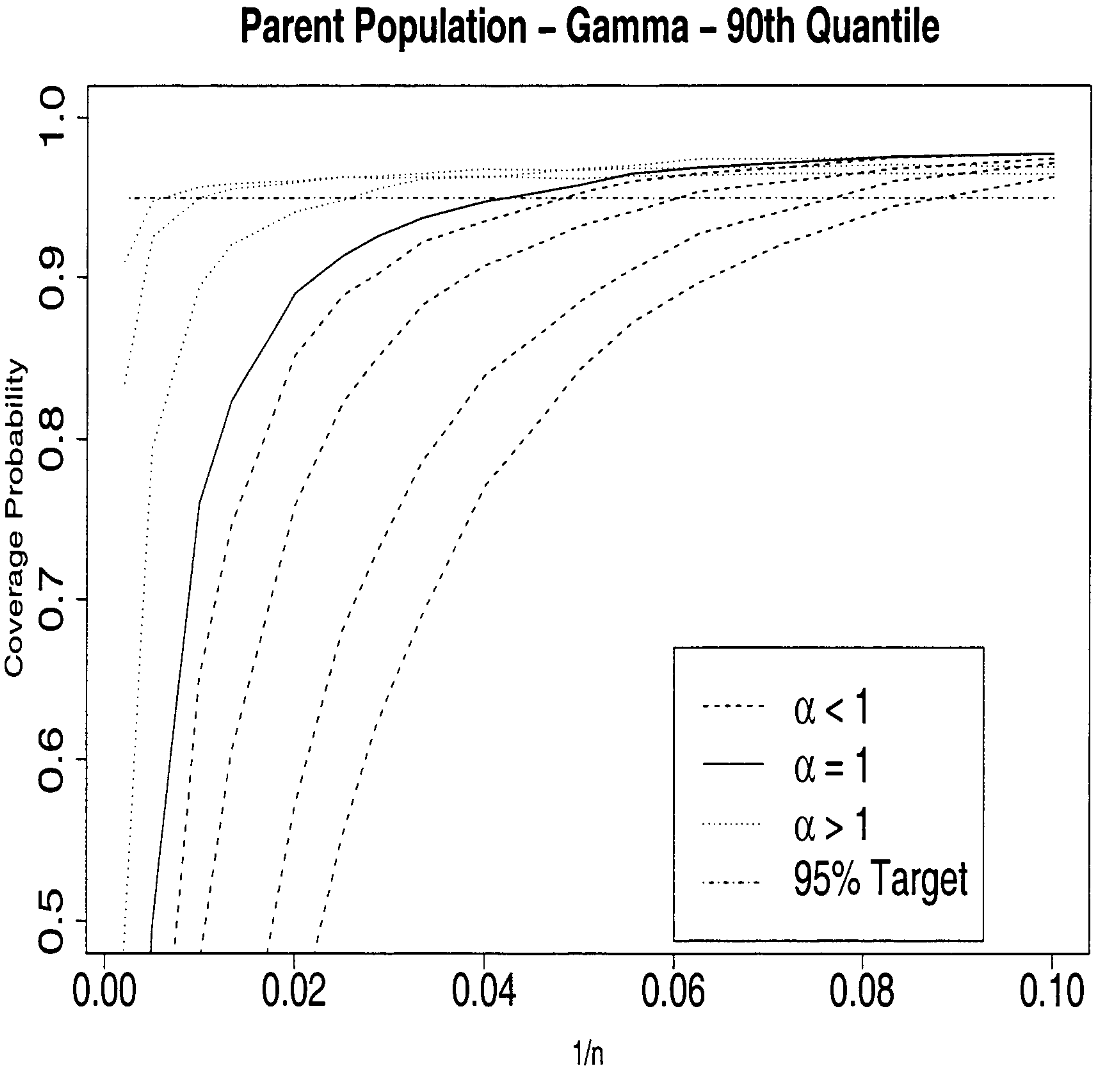

Figure 4.9: The true coverage probability of a 95% CI for $x_{0.9}$ calculated under the lognormal assumption when the parent population is gamma. The CP is separated by the shape parameter α.

Coverage Probability and p-values

The CP for each quantile is examined in light of the p-values. Figures 4.10-4.11 give the CP of the 95% CI for $x_{0.50}$ at the different values of α. There appears to be no connection between the p-value and the CP, as the three lines on each plot roughly follow each other, regardless of the p-value. In addition, as α increases, the CP for all p-values approaches the 95% target level.

Figures 4.12-4.13 give similar plots for $x_{0.9}$. Here there is a connection between the p-value and the CP, especially when $\alpha < 1$. As long as the p-value is above 0.05, a very good CP is maintained even for larger samples, while low p-values correspond to a poor CP, which worsens as the sample size increases. As α increases, however, this connection fades: the CP is above 90% for large values of α (say, $\alpha = 5$ or 10) regardless of the p-value.

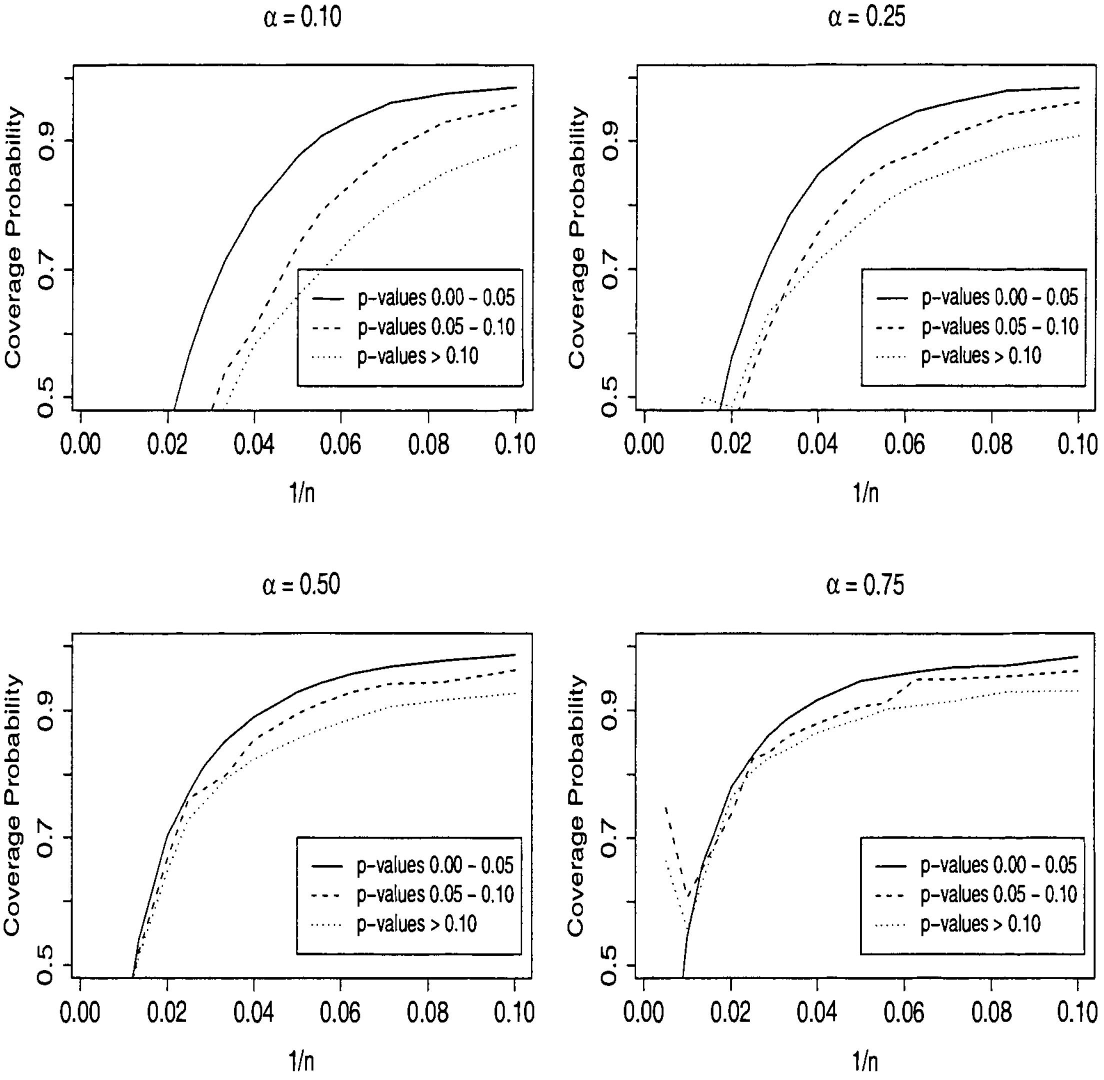

Figure 4.10: The true coverage probability of a 95% CI for $x_{0.5}$ calculated under the lognormal assumption when the parent population is gamma. The CP is separated by the p-value from a GoF test for lognormality. Shape parameter values are 0.10, 0.25, 0.5, and 0.75.

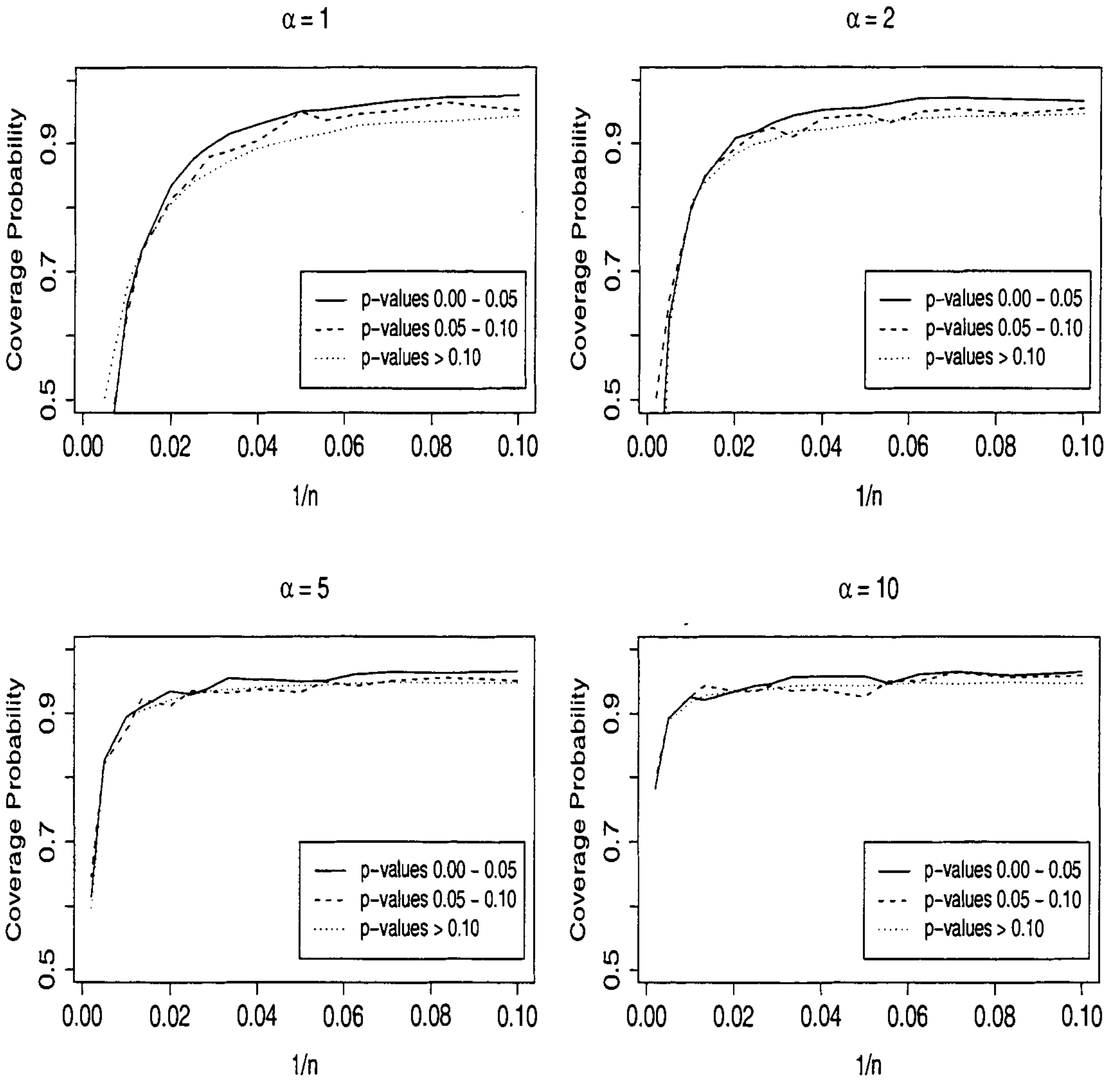

Figure 4.11: The true coverage probability of a 95% CI for $x_{0.5}$ calculated under the lognormal assumption when the parent population is gamma. The CP is separated by the p-value from a GoF test for lognormality. Shape parameter values are 1, 2, 5, and 10.

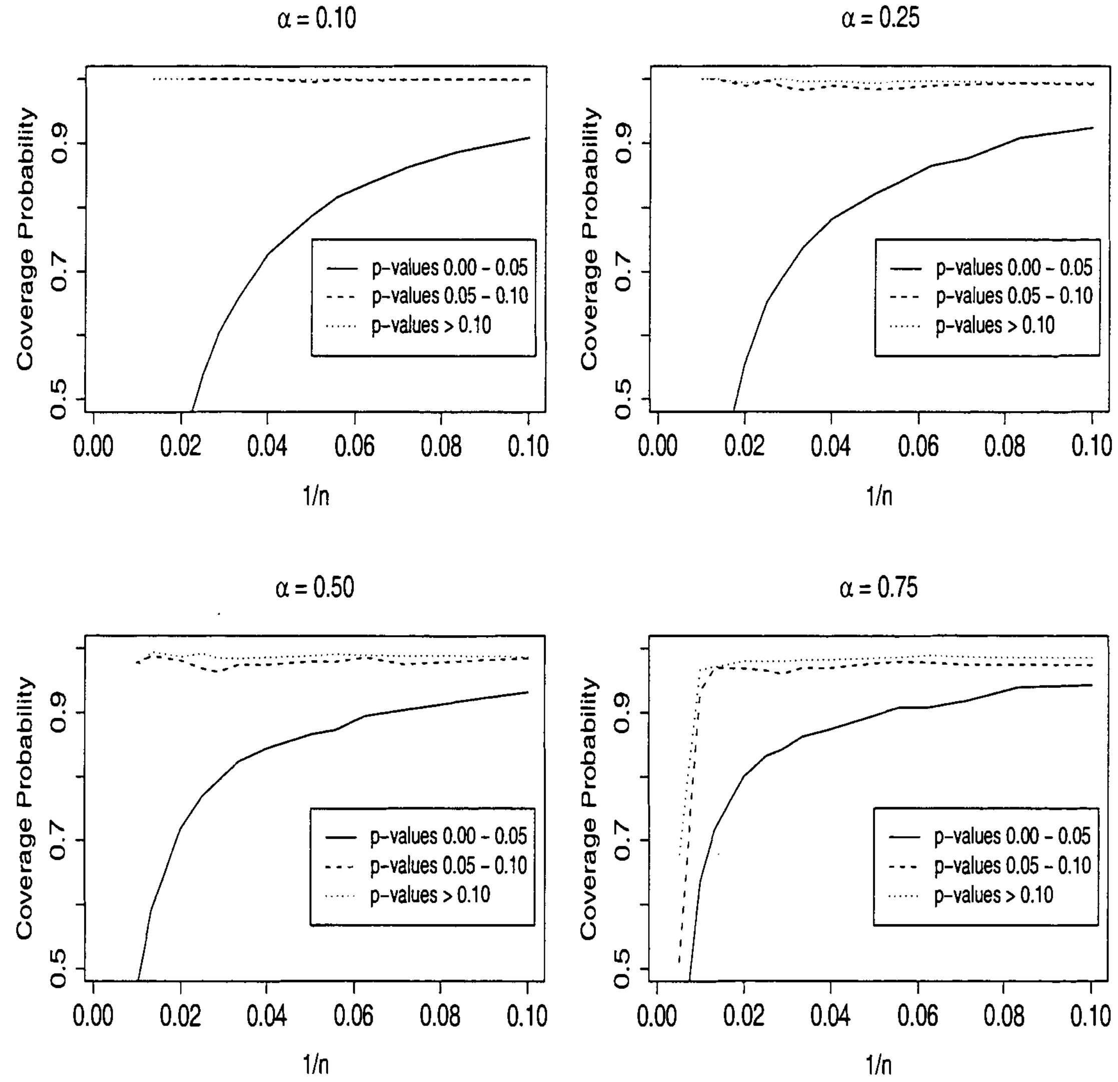

Figure 4.12: The true coverage probability of a 95% CI for $x_{0.9}$ calculated under the lognormal assumption when the parent population is gamma. The CP is separated by the p-value from a GoF test for lognormality. Shape parameter values are 0.10, 0.25, 0.5, and 0.75.

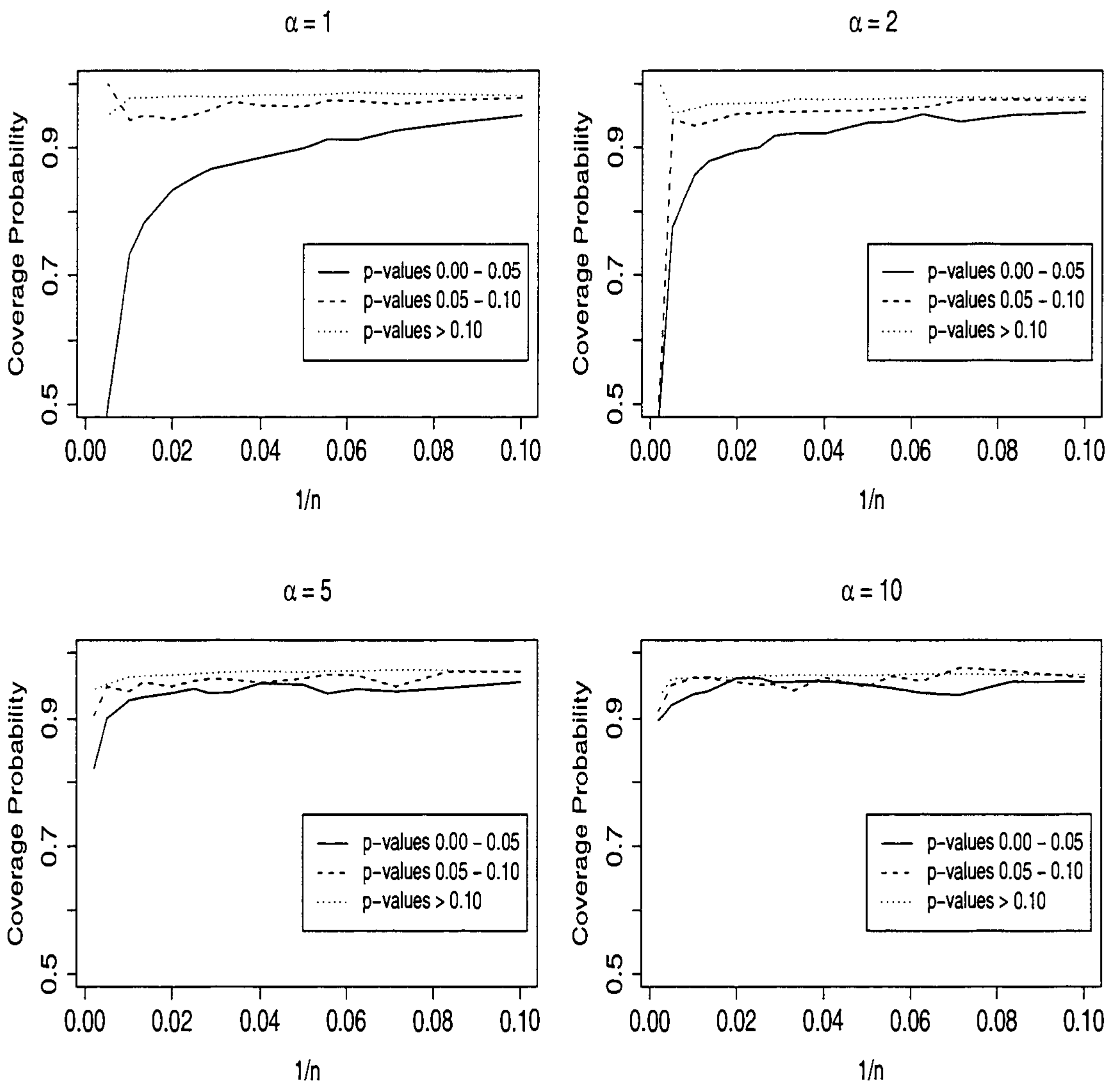

Figure 4.13: The true coverage probability of a 95% CI for $x_{0.9}$ calculated under the lognormal assumption when the parent population is gamma. The CP is separated by the p-value from a GoF test for lognormality. Shape parameter values are 1, 2, 5, and 10.

4.4 Return to the Data

In the LC data set, the Weibull and gamma fits are very poor, giving p-values of < 0.0001 and 0.0095 respectively for A^2. The lack of fit shows up strongly because the sample size is large (n=184). However, the lognormal fit is very good with a p-value of 0.3698. The results in Table 3.4 give reasonably narrow confidence intervals, as one would expect from such a large sample, and our simulation studies (see Figures 4.1-4.3) suggest that the coverage probability will be 95% as expected.

For the CC data set, with such a small sample (n=38), the lognormal fit is not as good as the Weibull and gamma fits, with a p-value of 0.1160 for the lognormal fit compared to 0.3107 for the Weibull fit and 0.2821 for the gamma fit.

Our simulation results show (see Figures 4.5 and 4.7) that if the data were in fact Weibull, with a sample size of 38 and a p-value of 0.1160, the lognormal assumption would give a CP of a 95% CI of approximately 85% for the 50^{th} quantile, and at least 95% for the 90^{th} quantile.

For the gamma fit to the CC data, the shape parameter was 1.98. If the data were indeed gamma with a shape parameter near 2, but the lognormal distribution were fitted, Figures 4.11 and 4.13 suggest that with a sample size of 38 and a p-value of 0.1160, the CP of a 95% CI is about 90% for the 50^{th} quantile and 95% for the 90^{th} quantile.

These high coverage probabilities are reassuring in terms of fitting the lognormal distribution to this particular data set, but again the price of a high CP is larger intervals which may in fact be too large to be useful. Note that in Table 3.5, the confidence intervals for the Weibull and gamma fits are narrower.

Chapter 5

Final Remarks

The purpose of this project was to explore the use of the lognormal distribution in the context of survival data. What follows is a summary and comments on future work.

Summary

1. Two data sets were provided, and four distributions, including the lognormal, were used to fit the data. GoF testing was done to determine which model gave the best fit to the data.

2. Confidence intervals for the quantiles were considered. The focus was on calculating confidence intervals for the quantiles of the lognormal distribution. A new pivotal quantity was proposed and its percentage points were well-approximated by an expression which depended only on the sample size. This led to a simple procedure to calculate the CI for the quantile.

3. The performance of the 95% CI was examined by looking at the CP when the parent population was indeed lognormal. The robustness of the lognormal assumption was examined by looking at the CP of the confidence intervals for Weibull and gamma alternatives. The CP was studied for a middle (50^{th}) and upper (90^{th})

quantile for a number of sample sizes. The following remarks summarize the results of the simulation studies.

- Under the lognormal distribution, the CP was 95% for both quantiles regardless of the sample size and p-value. This can be explained using theoretical results.
- For Weibull and gamma alternatives, 95% confidence intervals calculated under the lognormal assumption yielded good CP for small samples, but as sample size increased, the coverage probability decreased, as expected.
- The connection between the CP and the p-value was examined. For both the Weibull and gamma alternatives, a low p-value (below 0.05) corresponded to a lower coverage probability in the case of the 90^{th} quantile, but not the 50^{th} quantile.
- Since the gamma distribution with a large shape parameter is similar to the lognormal distribution, results for the gamma distribution with a large shape parameter are similar to the lognormal results.

Future Work

The following are suggestions for future work:

- The exact CI method for the lognormal quantile applies only to complete samples. In practice, censored observations are common, in which case the table given in Chapter 3 cannot be used. We want to examine the behaviour of W_p with censored data.
- The focus of the simulation studies was on coverage probability. We want to extend this work to examine the widths of the intervals.

Bibliography

Boag, J. (1948). The presentation and analysis of the results of radiotherapy. Part I. Introduction. *British Journal of Radiology* **21** 128-138.

Casella, G. and Berger, R. (2002). *Statistical Inference.* Duxbury, California.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995) *Continuous Univariate Distributions, Volume 2.* Wiley, New York.

Kendall, M. G. and Stuart, A. (1977). *The Advanced Theory of Statistics, Vol. I, 4th ed.* MacMillan, New York.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data.* Wiley, New York.

Lefante Jr, J. and Shah, A. (2002). Robustness properties of lognormal confidence intervals for lognormal and gamma distributed data. *Communications in Statistics: Theory and Methods*, **31**(11) 1939-1957.

Modarres, R., Nayak, T., and Gastwirth, J. (2002). Estimation of upper quantiles under model and parameter uncertainty. *Computational Statistics and Data Analysis* **39** 529-554.

Owen, D. (1962). *Handbook of Statistical Tables.* Addison-Wesley Publishing, Massachusetts.

Solomon, H. and Stephens, M.A. (1978). Approximations to density functions using pearson curves. *Journal of the American Statistical Association*, **73** 153-160.

Stephens, Michael A. (1986). Tests based on EDF statistics, Chapter 4 in *Goodness of Fit Techniques*, (D'Agostino, R.B. and Stephens, M.A., eds.), Marcel Dekker, New York.

Tai, P. (2003). Twenty-year follow-up study of long-term survival of limited-stage small-cell lung cancer and overview of prognostic and treatment factors. *Int. Journal of Radiation Oncology Biol. Phys*, **56**(3) 626-633.