Investigation of tail probability using a smooth estimation of a survival function

Serge Thiffeault

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in partial fullfillment of the requirements

for the degree of Master of Science at

Concordia University

Montreal, Quebec, Canada

August, 2004

Abstract

Investigation of tail probability using a smooth estimation of a survival function

Serge Thiffeault

Since the publication of their truncated smooth estimator of a survival function in 1996, Chaubey and Sen continued investigation into its statistical properties under various conditions. The authors had identified a problem of convergence of their estimator when estimating the mean residual life function. To overcome this problem with the convergence, they proposed an untruncated version of their smooth estimator. Nevertheless, in fear of *oversmoothing* with the untruncated version, the authors did not pursue research on this estimator much further.

This thesis deals with the investigation of some statistical properties of the untruncated version of the smooth estimator and studies the problem of *oversmoothing* by using simulations.

Acknowledgments

I wish to express my earnest thanks to my thesis supervisor Prof. Yogendra Chaubey for his advices, patience, constructive criticism and mostly for the trust he has put in me troughtout the preparation of this thesis.

I also wish to thank the Department of Mathematics and Statistics at Concordia University as well as every Professor who taught me during my study. I would like to give a special thank to Ms Anne-Marie Agnew for her kindest and continuous support.

Finally, I wish also to thank my family and friends for their understanding and encouragement during all those years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Preliminaries

1.1 Introduction

In many situations, it may be desirable to estimate the entire cumulative survival function of a random variable T. Most of the time, however, we do not have knowledge about the underlying density. Instead we are given a set of n observations $\{t_1, t_2, \cdots, t_n\}$ of which we assume that they are realizations of independent, identically distributed random variates with the same density as of T. It is our goal to estimate the cumulative survival function on the basis of these observations.

Knowledge of the survival function helps in many aspects. In reliability theory, for example, it is used to determine the probability of success of a unit, in undertaking a mission of a prescribed duration. Pharmaticeutical studies employ the survival function as a basic quantity to describe time-to-event phenomena, the probability of an individual surviving beyond time t (experiencing the event after time x). Interpreting and analyzing the survival function provide a way to obtain many more structural elements such as its density, hazard function, etc.

Let f and F denote the density and distribution function of the random

variable T, respectively, then

$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(x)\,dx, \quad t \geq 0, \text{ defined on } R^{+}. \tag{1.1}$$

The survival function is defined as the complement of the distribution function. That is

$$S(t) = P(T > t) = 1 - F(t) = \int_{t}^{\infty} f(x)\,dx, \quad t \geq 0, \text{ defined on } R^{+}. \tag{1.2}$$

Given a random sample $\{t_1, \ldots, t_n\}$ from the distribution for F, the usual estimator of the unknown probability of the event $\{T \leq t\}$ is given by the observed frequency of its occurrence, known as the empirical distribution function (*edf*)

$$\begin{aligned} F_n(t) &= \frac{\text{Number of } t_i \leq t}{n}, \\ &= n^{-1} \sum_{i=1}^{n} \mathrm{I}[t_i \leq t], \quad t \in R^{+} \end{aligned} \tag{1.3}$$

where $I[\cdot]$ is the indicator function defined as

$$I[\cdot] = \begin{cases} 1 & \text{if } [\cdot] \text{ is true,} \\ 0 & \text{if } [\cdot] \text{ is false.} \end{cases} \tag{1.4}$$

Therefore, the empirical survival function (*esf*) is

$$\begin{aligned} S_n(t) &= P(T > t), \\ &= 1 - P(T \leq t) = 1 - F_n(t), \\ &= n^{-1} \sum_{i=1}^{n} \mathrm{I}[t_i > t], \quad t \in R^{+}. \end{aligned} \tag{1.5}$$

Note that:

- The *esf* is a step function that decreases by $\frac{1}{n}$ just after each observed t if all observations are distinct.
- If we let $t_{n:r}$ denote the $r^{th}(r = 1, 2, \cdots, n)$ order statistic from the random sample $\{t_1, \cdots, t_n\}$, then we have

$$S_n(t) = (n-k)/n, \quad \text{for } t_{n:k} \leq t < t_{n:k+1}, \tag{1.6}$$

  where $k = 0, \cdots, n$, $t_{n:0} = 0$ and $t_{n:n+1} = \infty$.
- $S_n(t) = 1$, for $0 < t \leq t_{n:1}$ and $S_n(t) = 0$, $\forall t > t_{n:n}$.
- $F_n(t)$ and $S_n(t)$ are nonparametric estimators of $F(t)$ and $S(t)$ respectively.

Since $S_n(t)$ is a step function, it is not smooth enough to estimate the corresponding density. Moreover, estimation of a smooth distribution function by a step function may not be particularly attractive. The idea of studying smooth estimator of density and considering problems of accuracy of the estimation were originally posed and studied in the Soviet Union. The results obtained on this subject were due to Glivenko (1934) and Smirnov (1951) who used histograms for estimators. Rosenblatt (1956), Parzen (1962) and Chentsov (1962) made further contributions to the theory of nonparametric estimation of probability densities. In their works, these authors introduced new classes of estimators which generalize histograms. The idea of constructing new nonparametric density estimators is as follows:

Let $x_1, x_2, \cdots, x_n$ be a sequence of independent identically distributed random variables with the distribution law $F(\cdot)$. The empirical distribution $F_n(\cdot)$ constructed from the sample $x_1, x_2, \cdots, x_n$ is a discrete distribution with atoms of weight $1/n$ located at each one of the observed points. Associated with

each observation x_i is a delta-measure $\delta_{x_i}(\cdot)$ concentrated at point x_i and with a sequence of independent observations $x_1, x_2, \cdots, x_n$, the histogram is the arithmetic mean of these measures, *i.e.*

$$f_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(\cdot). \tag{1.7}$$

The above function gives discrete measures concentrated at the sample values $x_1, x_2, \cdots, x_n$. However, if the unknown distribution possesses a smooth density $f(x)$, it is natural to *spread out* each observed measure $\delta_{x_i}(\cdot)$ replacing it by a measure with a certain density concentrated at x in order to estimate it at this point. In practice, we choose this density to be symmetric around the point x and arrive at the following class of estimators:

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right), \tag{1.8}$$

where $k(u)$ is a symmetric density, called the kernel, around 0. For the asymptotic theory to hold for this class of estimators, h must tend to 0 and $nh \to \infty$ as $n \to \infty$.

The choice of the kernel function is not as important as the choice of h. Various methods of determining h have been proposed in the literature along with many kernel functions as given below [see Härdle(1991)].

Kernel	**k(u)**
Uniform	$\frac{1}{2}I(\lvert u\rvert \leq 1)$
Triangle	$(1 - \lvert u\rvert)I(\lvert u\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - u^2)I(\lvert u\rvert \leq 1)$
Quartic	$\frac{15}{16}(1 - u^2)^2 I(\lvert u\rvert \leq 1)$
Triweight	$\frac{35}{32}(1 - u^2)^3 I(\lvert u\rvert \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}u^2\right)$
Cosinus	$\frac{\pi}{4}\cos(\frac{\pi}{2}u)I(\lvert u\rvert \leq 1)$

Among the various approaches studied in the past, their main objectives were to smooth the histogram. Chaubey and Sen (1996), on the other hand

suggested to smooth the empirical survival function directly as given in the next section.

1.2 Smooth Estimator of a Survival Function

Chaubey and Sen (1996) formulated an alternative approach based on the classical Hille's theorem (1948) on uniform smoothing in real analysis and obtained some smooth estimators of $S(t)$ and $f(t)$ which may have some advantages over their counterpart based on the usual kernel method of estimating.

Here, let us illustrate how the classical Hille's theorem (1948) has been incorporated into the new smoothing formulation.

For every $t \in R^+$ and $y \in R^+$, we consider an array $\{w_{nk}(t,y);$ $0 \leq k \leq n;\ n \geq 1\}$ by letting

$$w_{nk}(t,y) = \{(ty)^k/k!\}/\{\sum_{i=0}^{n}(ty)^i/i!\}, \tag{1.9}$$

so that $w_{nk}(t,y)$ is nonnegative and

$$\sum_{k=0}^{n} w_{nk}(t,y) = 1, \quad \forall\, t,y \in R^+. \tag{1.10}$$

Now by letting $\{\lambda_n; n \geq 1\}$ be a sequence of (possibly stochastic) positive numbers, such that as $n \to \infty$, $\lambda_n \to \infty$ almost surely (a.s.), but $n^{-1}\lambda_n \to 0$ (a.s.), enables us to adapt the Hille's theorem (1948), *albeit* in a stochastic setup. The proposed smooth estimator of $S(\cdot)$ is defined as

$$\tilde{S}_n(t) = \sum_{k=0}^{n} w_{nk}(t,\lambda_n)\, S_n\left(\frac{k}{\lambda_n}\right), \quad t \in R^+, \tag{1.11}$$

It has been shown by Chaubey and Sen (1996) that as n increases, for every fixed $t \in R^+$, $w_{nk}(t,y)$ behaves like $e^{-t\lambda_n}(k!)^{-1}(t\lambda_n)^k$, for $k \leq n$, so that for large n, this adaptation is essentially a Poisson mixture of $S_n(\cdot)$ with the

Poisson probabilities adapted to the parameter $t\lambda_n$. In this sense, the adapted parameter $t\lambda_n$ is made to depend on t as well as λ_n, and it monotonically increase as t increases (for a given λ_n or n). The right tiltedness of the Poisson distribution, in its parameter and the monotonicity property of $S_n(\cdot)$, makes $\tilde{S}_n(\cdot)$ monotone as well.

1.2.1 Asymptotic Properties of $\tilde{S}_n(t)$

In this section, we will show that $\tilde{S}_n(t) \to S(t)$ as $\lambda_n \to \infty$ and as $n \to \infty$. We have

$$\left|\tilde{S}_n(t) - S(t)\right| \leq \left|\tilde{S}_n(t) - S^*(t)\right| + \left|S^*(t) - S(t)\right|, \tag{1.12}$$

where $S^*(t) = \int_{-\infty}^{\infty} S(x) dG_{t,h}(x) \to S(t)$, as $n \to \infty$. Also for every t

$$\left|\tilde{S}_n(t) - S^*(t)\right| \leq \max_x \left|S_n(x) - S(x)\right| \int_{-\infty}^{\infty} dG_{t,h}(x). \tag{1.13}$$

It follows that

$$\left|\tilde{S}_n(t) - S^*(t)\right| \leq \left|\tilde{S}_n(t) - S^*(t)\right| \leq \max_x \left|S_n(x) - S(x)\right| \int_{-\infty}^{\infty} dG_{t,h}(x). \tag{1.14}$$

Since by Gilvenko-Cantelli theorem (see Rohatgi (2001), pp. 311)

$$\sup_t \left|\tilde{S}_n(t) - S(t)\right| \to 0 \ a.s. \tag{1.15}$$

the results is confirmed.

The main consistency results are stated in the following theorem:

Theorem 1 *[Chaubey and Sen (1996)] If $S(t)$ is continuous (a.e.), $\lambda_n \to \infty$ and $n^{-1}\lambda_n \to 0$ then*

$$\|\tilde{S}_n - S\| = sup[|\tilde{S}_n(t) - S(t)| : t \in R^+] \ a.s., \qquad as\ n \to \infty. \tag{1.16}$$

Furthermore, because S_n is strongly consistent for S and so is $\tilde{S}_n$, the following theorem depicting their interrelationship was established.

Theorem 2 *[Chaubey and Sen (1996)] Under the hypothesis (on λ_n) of the previous theorem, whenever $f(t)$ is absolutely continuous with a bounded derivative $f'(\cdot)$ a.e. on R^+,*

$$\|\tilde{S}_n - S_n)\| = O\left(n^{\frac{-3}{4}} (\log n)^{1+\delta}\right) \ a.s., \qquad as\ n \to \infty, \tag{1.17}$$

where $\delta(> 0)$ is arbitrary.

We may note that for every $t \in R^+$,

$$\begin{aligned} E\{\tilde{S}_n(t) - S(t)\}^2 = E\{\tilde{S}_n(t) - S_n(t) I(|\tilde{S}_n(t) - S_n(t)| \leq cn^{-3/4})\} \\ + E\{\tilde{S}_n(t) - S_n(t) I(|\tilde{S}_n(t) - S_n(t)| > cn^{-3/4})\}, \end{aligned} \tag{1.18}$$

for a suitable $c : 0 < c < \infty$. Using the basic exponential rates of convergence relating to the Bahadur representation (1971), we have

$$P\{\|\tilde{S}(\cdot)_n - S_n(\cdot)\| > cn^{-3/4}\} = O(n^{-7/4}). \tag{1.19}$$

Therefore by Eq.(1.18) and Eq.(1.19), we have

$$\begin{aligned} E\{(\tilde{S}_n(t) - S_n(t))^2\} &\leq c^2 n^{-3/2} + O(n^{-7/4}) \\ &= O(n^{-3/2}). \end{aligned} \tag{1.20}$$

The kernel method, on the contrary, yields the order of the residual term in Eq.(1.20) as $O(n^{-4/3})$ (*viz.* Azzalini (1981)) so that in this respect, the proposed smooth estimator fares better.

We may define the usual quantile process $Q_n = \{Q_n(t) : 0 \leq t \leq 1\}$ by letting

$$\tilde{Q}_n(t) = \sup\{x : S_n(x) \geq 1 - t\}, \quad 0 \leq t \leq 1. \tag{1.21}$$

Then, for every $t : 0 \leq t \leq 1$, the improved order of approximations in the preceding discussion also pertain to $\tilde{Q}_n(t)$. It may be noted that $\tilde{S}_n(\cdot)$, being differentiable, smooth,and hence in Eq.(1.21), we may as well replace "$\geq$" by "$=$".

In 1999, Chaubey and Sen studied a mean residual life estimator using their smooth estimator. In their study, the authors point out that the smooth estimator is not appropriate to estimate the mean residual life because $\tilde{S}_n(t)$ diverges. By altering the weight function, Chaubey and Sen proposed a modified estimator as following:

$$\tilde{S}_n^0(t) = \sum_{k=0}^{N} w_{nk}^0 (t, \lambda_n) S_n \left(\frac{k}{\lambda_n} \right), \tag{1.22}$$

where

$$w_{nk}^0 (t, \lambda_n) = e^{-t\lambda_n} \frac{(\lambda_n t)^k}{k!} \tag{1.23}$$

and $N \ni S_n \left(\frac{k}{n}\right) = 0$ for $k > N$.

This can be called untruncated version of the Chaubey and Sen's original estimator. The limiting distribution of the weights of truncated version is the Poisson distribution. Therefore, all the asymptotic properties of the truncated version will still apply to this new formulation.

1.3 Overview of the Thesis

This thesis is intended as a qualitative numerical study of the untruncated version of the Chaubey and Sen smooth estimator $\tilde{S}_n^0(t)$. We attempt to analyze some important properties such as bias, standard errors and confidence intervals. We also attempt to investigate the tail behavior in detail.

Chapter 2 presents the statistical distributions that are used to describe survival times. These distributions will be used throughout the rest of the thesis. Chapter 3 deals with the numerical estimation of the smooth estimator $\tilde{S}_n^0(t)$. This subject is extended into Chapter 4 to present some of the statistical properties of $\tilde{S}_n^0(t)$. The confidence intervals and the tail behaviors are assessed and discussed in this chapter.

Chapter 2

Standard Survival Distributions and Their Properties

2.1 Introduction

This chapter introduces some of the statistical distributions that are used in this thesis. They are the Exponential, Gamma, Weibull and Lognormal. These distributions have been chosen because they represent the most frequently used distributions in the reliability engineering and survival analysis fields. Moreover, all distributions except the Exponential attain a wide variety of shapes for various values of their parameters. Thus they can model a great diversity of data and life characteristics.

2.2 Exponential Distribution

Historically, the Exponential distribution was the first widely used lifetime distribution model. This was partly because of the availability of simple statistical methods for it (e.g., Epstein and Sobel (1953)) and partly because of its suitability for representing the lifetimes of many things. The Exponential distribution has been used in areas ranging from studies on the lifetimes of manufactured items (e.g., Davis (1952); Epstein (1958)) to researches in-

volving survival or remission times in chronic diseases (e.g., Feigl and Zelen (1965)).

2.2.1 Properties

The probability density function of the single-parameter Exponential distribution is defined as

$$f(t;\theta) = \frac{1}{\theta}\exp\left(-\frac{t}{\theta}\right), \tag{2.1}$$

where $\theta > 0$ is known as the scale parameter. Its cumulative density function and survival density function are respectively

$$F(t;\theta) = 1 - \exp\left(-\frac{t}{\theta}\right), \tag{2.2}$$

$$S(t;\theta) = \exp\left(-\frac{t}{\theta}\right). \tag{2.3}$$

Some of the specific characteristics of the single-parameter Exponential *pdf* are the following:

Mean: θ,

Standard deviation: θ,

Median: $0.693 \cdot \theta$,

Mode: 0,

Quantiles: $-\theta \log_e(1-p)$,

Moments: For integer $m > 0$, $E\left[(T)^m\right] = m!\theta^m$.

Then $E(t) = \theta$, $Var(T) = \theta^2$.

2.2.2 Motivation for the Exponential Distribution

The motivation for the Exponential distribution is summarized as follows:

- Simplest distribution used in the analysis of survival or reliability data.
- Popular distribution for some kinds of electronic components (e.g., capacitors, high-quality integrated circuits).

- Has the important characteristic; its hazard function is constant (does not depend on time t).

2.3 Weibull Distribution

The Weibull distribution is perhaps the most widely used lifetime distribution model in present. Its application in connection with lifetimes of many types of manufactured items has been widely advocated (e.g., Weibull (1951); Berretoni (1964)). And also the distribution has been used as a model in biomedical applications such as the studies on the time to the occurrence of tumors in human populations (Whittemore and Altschuler (1976)) or in laboratory animals (Pike (1966); Peto et al. (1972)).

2.3.1 Properties

The probability density function of the Weibull distribution is defined as

$$f(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{\left(\frac{t-\gamma}{\eta}\right)^{\beta}}, \tag{2.4}$$

where $f(t) \geq 0$, $t \geq \gamma$, $\beta > 0$, $\eta > 0$, $-\infty < \gamma < \infty$, and β, η are respectively the shape and scale parameters. Its cumulative density function and survival density function are respectively

$$F(t;\beta,\eta,\gamma) = 1 - exp\left[-\left(\frac{t-\gamma}{\eta}\right)^{\beta}\right], \tag{2.5}$$

$$S(t;\beta,\eta,\gamma) = exp\left[-\left(\frac{t-\gamma}{\eta}\right)^{\beta}\right]. \tag{2.6}$$

Some of the specific characteristics of the Weibull *pdf* are the following:

Mean: $\gamma + \eta\Gamma\left(\frac{1}{\beta}+1\right)$,

Standard deviation: $\eta\left[\Gamma\left(\frac{2}{\beta}+1\right) - \left[\Gamma\left(\frac{1}{\beta}+1\right)\right]^2\right]^{1/2}$,

Median: $\gamma + \eta\,(log_e 2)^{1/\beta}$,
Mode: $\gamma + \eta\left(1 - \frac{1}{\beta}\right)^{1/\beta}$,
Quantiles: $\eta\,[\log_e(1-p)]^{1/\beta}$,
Moments: For integer $m > 0$, $E\left(T^m\right) = \eta^m \Gamma\left(1 + m/\beta\right)$. Then
$E(T) = \eta\Gamma\left(1 + \frac{1}{\beta}\right), Var(T) = \eta^2\left[\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)\right]$,
where $\Gamma(\kappa) = \int_0^\infty w^{\kappa-1}\exp(-w)dw$ is the gamma function.

Note: When $\beta = 1$ then $T \sim EXP(\eta)$.

2.3.2 Motivation for the Weibull Distribution

The theory of extreme values shows that the Weibull distribution can be used to model the minimum of a large number of independent positive random variables from a certain class of distributions.

- Failure of the weakest link in a chain with many links with failure mechanisms (e.g. creep or fatigue) in each link acting approximately independent.
- Failure of a system with a large number of components in series and with approximately independent failure mechanisms in each component.

The more common justification for its use is empirical. The Weibull distribution can be used to model failure-time data with a decreasing or increasing hazard function.

2.4 Gamma Distribution

The Gamma distribution is used as a lifetime model (e.g., Gupta and Groll (1961)) although not as widely as the Weibull distribution. This is partly because the survivor and hazard functions of the Gamma distribution are

not expressible in a simple closed form and hence are more difficult to work with than with those of the Weibull distribution. The Gamma distribution, however, does fit a wide variety of lifetime data adequately, and there are failure process models that lead to it (see Buckland (1964), Sec.1.7).

2.4.1 Properties

The general form for *pdf* of the Gamma distribution is defined as

$$f(t) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}(t-r)^{\alpha-1}\exp\left[-\frac{(t-r)}{\beta}\right], \quad \text{for } t > r, \alpha > 0, \beta > 0, \tag{2.7}$$

where α, β and r are shape, scale and location parameter respectively. The standard form of this distribution is obtained by putting $\beta = 1$ and $r = 0$ which is given by

$$f(t) = \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}, \quad \text{for } t \geq 0. \tag{2.8}$$

If $\alpha = 1$, the Gamma distribution reduces to the Exponential distribution. The cumulative distribution function and survival distribution function are respectively

$$F(t;\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\int_0^t e^{-t}t^{\alpha-1}dt, \tag{2.9}$$

$$S(t;\alpha,\beta) = 1 - F(t;\alpha,\beta). \tag{2.10}$$

The Gamma distribution with positive integer α can be derived as the distribution of the waiting time to the αth arrival from a Poisson source with parameter α. It is apparent that the sum of k independent exponential variates with failure rate α has the Gamma distribution with parameters α and k. The continuous random variable t which is distributed according to the probability law,

$$f(t) = \frac{e^{-t}t^{\alpha-1}}{\Gamma(\alpha)} \quad \text{for } \alpha > 0, 0 < t < \infty, \tag{2.11}$$

is known as a Gamma variate with the parameter α, and its distribution is called the Gamma distribution. The mean and variance of this distribution are equal to α, like in a Poisson distribution. The density function which can be seen to be a member of the exponential family is unimodal, positively skewed and Leptokurtic, with its mode at $t = \alpha - 1$ if $\alpha \geq 1$. But distribution Eq.(2.7) has a mode at $t = r + \beta(\alpha - 1)$. If $\alpha < 1$, $f(t)$ tends to infinity as t tends to zero, also if $\alpha = 1$, $\lim_{t \to 0} f(t) = 1$.

The m.g.f. for the Gamma distribution Eq.(2.8) is

$$M_x(t) = (1-t)^{-\alpha}, \quad |t| < 1. \tag{2.12}$$

Thus the cumulant generating function $K(t)$ is given by

$$K_x(t) = lnM_x(t) = ln(1-t)^{-\alpha} = -\alpha \log(1-t), \tag{2.13}$$

$$= \alpha \left[t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots \right], \tag{2.14}$$

from which we can derive

$$\textbf{Mean} = K_1 = \text{ coefficient of } t \text{ in } K_x(t) = \alpha, \tag{2.15}$$

$$\mu_2 = K_2 = \text{ coefficient of } \frac{t^2}{2!} \text{ in } K_x(t) = \alpha, \tag{2.16}$$

$$\mu_3 = K_3 = \text{ coefficient of } \frac{t^3}{3!} \text{ in } K_x(t) = 2\alpha, \tag{2.17}$$

$$K_4 = \text{ coefficient of } \frac{t^4}{4!} \text{ in } K_x(t) = 6\alpha, \tag{2.18}$$

$$\text{Therefore } \mu 4 = K_4 + 3\mu_2^2 = 6\alpha + 3\alpha^2, \tag{2.19}$$

$$\text{Hence, } \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{4}{\alpha}, \tag{2.20}$$

$$\text{and } \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{6}{\alpha}. \tag{2.21}$$

The moments can be found from either the m.g.f., c.f. or directly by integration. From distribution Eq.(2.8), the rth moment about the origin zero is

$$\mu_r' = (\Gamma\alpha)^{-1} \int_0^\infty t^{\alpha+r-1} e^{-t} dt = \frac{\Gamma(\alpha + r)}{\Gamma(\alpha)} \quad \text{for } r = 1, 2, \cdots \tag{2.22}$$

Hence the distribution Eq.(2.8) has a mean = variance = α.

2.4.2 Motivation for the Gamma Distribution

The Gamma distribution gives useful representation of many physical situations. It is used to make realistic adjustments to the Exponential distribution in representing lifetimes in life testing situations. Also the sum of independent exponentially distributed random variables represents a Gamma distribution which leads to the appearance of the theory of random counters and other related topics in association with random process in meteorological precipitation process.

2.5 Lognormal Distribution

The Lognormal distribution, like the Weibull distribution, has been widely used as a lifetime distribution model. It has been used in diverse situations such as the analysis of failure times of electrical insulation (Nelson and Hahn (1972)) and the study of times to the apparency of lung cancer in cigarette smokers (Whittemore and Altschuler (1976)). The distribution is most easily expressed by saying that the lifetime T is log-normally distributed if the logarithm $Y = \log T$ of the lifetime is normally distributed with mean μ and variance σ^2.

2.5.1 Properties

The two-parameter Lognormal distribution *pdf* is given by

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t^* - \mu}{\sigma}\right)^2}, \tag{2.23}$$

where $t^* = log_e t$, $f(t) \geq 0$, $t \geq 0$, $-\infty < t < \infty$, $\sigma > 0$. Its cumulative density

function and survival density function are respectively

$$F(t;\mu,\sigma) = \Phi\left[\frac{log(t)-\mu}{\sigma}\right], \tag{2.24}$$

$$S(t;\mu,\sigma) = 1 - F(t;\mu,\sigma), \tag{2.25}$$

where Φ is the *cdf* for a standardized normal and $\sigma > 0$ is a shape parameter.

Some of the specific characteristics of the Lognormal *pdf* are the following:

Mean: $e^{\mu+\frac{1}{2}\sigma^2}$,

Standard deviation: $\left[\left(e^{2\mu+\sigma^2}\right)\left(e^{\sigma^2}-1\right)\right]^{1/2}$,

Median: e^{μ},

Mode: $e^{\mu-\sigma^2}$,

Quantiles: $exp\left(\mu+\sigma\Phi^{-1}(p)\right)$, where $\Phi^{-1}(p)$ is the p quantile for a standardized normal.

Moments: For integer $m > 0$, $E\left(T^m\right) = exp\left(m\mu + m^2\sigma^2/2\right)$,

$E(T) = exp\left(\mu+\sigma^2/2\right), Var(T) = exp\left(2\mu+\sigma^2\right)\left[exp\left(\sigma^2\right)-1\right]$.

2.5.2 Motivation for the Lognormal Distribution

The motivation for the Lognormal distribution is summarized as follows:

- The Lognormal distribution is a common model for failure times.
- It has been suggested as an appropriate model for failure time caused by a degradation process with combinations of random rates that combine multiplicatively.
- Widely used to describe time to fracture from fatigue crack growth in metals.
- Useful in modeling failure time of a population electronic components with a decreasing hazard function (due to a small proportion of defects in the population).

- Useful for describing the failure-time distribution of certain degradation processes.

Chapter 3

Numerical Investigation of the Smoothing Parameter

3.1 Introduction

This chapter proposes a qualitative numerical analysis of the Chaubey and Sen (1996) smooth estimator $\tilde{S}_n^0(t)$. The goal is to investigate the behavior of the estimator particularly in its tail region with various underlying distributions mentioned in chapter 2. Also we will examine techniques to obtain an optimal value of the smoothing parameter c.

In respect of Hille's theorem (1948), we need to choose λ_n so that $\lambda_n \to \infty$ a.s. as $n \to \infty$. Intuitively we can define

$$\hat{\lambda}_n = \frac{n}{T_{n:n}}, \tag{3.1}$$

where $T_{n:n}$ denotes the largest order statistic corresponding to $\{T_1, \ldots, T_n\}$.

We will show that $\hat{\lambda}_n$ meets the required regularity assumption using the following lemma:

Lemma 1 *Given X has support on $[0, \infty)$. If $E|X| < \infty$ then*

$$n^{-1} \max_{1 \le i \le n} (\, |X_i| \,) \underset{a.s.}{\to} 0 \quad as \quad n \to \infty. \tag{3.2}$$

Proof

Note that $\forall c > 0$ & $Y_i = |X_i|$,

$$P\,[\,\max(Y_1, \cdots, Y_n) > c\,] = 1 - [\,F(c)\,]^n. \tag{3.3}$$

Since $F(c) < 1, \quad \forall c > 0$ then $[\,F(c)\,]^n \to 0$ and

$$P\,[\,\max(Y_1, \cdots, Y_n) > c\,] \to 1 \qquad \forall c > 0. \tag{3.4}$$

Hence we conclude that

$$\max_{1 \le i \le n} |X_i| \underset{a.s.}{\to} \infty \qquad as\ n \to \infty. \tag{3.5}$$

Also, we conclude that $\forall c > 0$ that

$$I\,(T_{n:n} > c) \underset{a.s.}{\to} 1 \quad as\ n \to \infty, \tag{3.6}$$

where $T_{n:n} = \max(Y_1, \cdots, Y_n)$.

Thus

$$n^{-1} T_{n:n} I\,(T_{n:n} > c) \le n^{-1} \sum_{i=1}^{n} T_i I\,(T_i > c)\,. \tag{3.7}$$

But

$$\left| n^{-1} \sum_{i=1}^{n} T_i I\,(T_i > c) \right| \le \left| n^{-1} \sum_{i=1}^{n} T_i I\,(T_i > c) + \int_c^{\infty} t dS(t) \right| + \left| \int_c^{\infty} t dS(t) \right|. \tag{3.8}$$

Taking the limit of both sides as $c \to \infty$ and using the facts that

$$\lim_{c \to \infty} \int_c^{\infty} t dS(t) = 0, \tag{3.9}$$

and

$$n^{-1} \textstyle\sum_{i=1}^{n} T_i I\,(T_i > c) \underset{a.s.}{\to} \int t I(t > c) f(t) = -\int_c^{\infty} t dS(t), \tag{3.10}$$

we conclude from Eqs. (3.7) and (3.8) that

$$\lim_{n \to \infty} n^{-1} T_{n:n} I\,(T_{n:n} > c) = 0\ a.s.. \tag{3.11}$$

Since $I\left(T_{n:n} > c\right) \to 1$ $a.s.$, the above implies that $\lim_{n\to\infty} n^{-1}T_{n:n} = 0$ $a.s.$ Now, what happens if the above requirement is not met? In other words, if one chooses a $\hat{\lambda}_n$ that does not asymptotically converges to ∞. The proposed estimator Eq.(1.22) is defined as

$$\begin{aligned}\tilde{S}_n^0(t) &= \sum_{k=0}^{N} e^{-t\hat{\lambda}_n} \frac{\left(\hat{\lambda}_n t\right)^k}{k!} S_n\left(\frac{k}{\hat{\lambda}_n}\right), \ \mathrm{t} \in R^+, \\ &= w_{n:0}(t, \hat{\lambda}_n) S_n(0) + \cdots + w_{n:\hat{N}}(t, \hat{\lambda}_n) S_n(n/\hat{\lambda}_n).\end{aligned} \tag{3.12}$$

Case 1: Suppose that $\hat{\lambda}_n \to 0$ as $n \to \infty$. It implies that for $\forall t > 0$, we have

$$\begin{aligned}\text{at } k = 0, \ \lim_{\hat{\lambda}_n \to 0} w_{n:0}(t, \hat{\lambda}_n) &= \frac{e^{-\hat{\lambda}_n t}\left(\hat{\lambda}_n t\right)^0}{0!} = 1, \\ \text{at } k \geq 1, \ \lim_{\hat{\lambda}_n \to 0} w_{n:k}(t, \hat{\lambda}_n) &= \frac{e^{-\hat{\lambda}_n t}\left(\hat{\lambda}_n t\right)^k}{k!} = 0.\end{aligned} \tag{3.13}$$

The behavior of the weights $w_{n:k}(t, \hat{\lambda}_n)$ is asymptotically the same as $S_n(0) = 1$. This indicates that $\tilde{S}_n^0(t) \nrightarrow S(t)$ as $n \to \infty$.

Case 2: Suppose that $\hat{\lambda}_n \to \lambda$ as limit$(< \infty)$ as $n \to \infty$ in probability. It implies by the Gilvenko-Cantelli lemma, which says that the empirical survival function of a one dimensional random variable converges uniformly to the true survival function in probability, that

$$\max_{k \leq n} \left| S_n\left(\tfrac{k}{\hat{\lambda}_n}\right) - S\left(\tfrac{k}{\lambda}\right) \right| \to 0, \text{ a.s., as } n \to \infty. \tag{3.14}$$

Hence

$$\tilde{S}_n^0(t) \to \sum_{k=0}^{N} e^{-t\lambda} \frac{(\lambda t)^k}{k!} S\left(\frac{k}{\lambda}\right), \ \mathrm{t} \in R^+. \tag{3.15}$$

Unless t is large, the consistency of $\tilde{S}_n^0(t)$ may not hold since the Poisson mixture on the right side of Eq.(3.15) need not to be close to $S(t)$. Thus by contradiction, we must have $\hat{\lambda}_n \to \infty$ as $n \to \infty$.

Modified $\hat{\lambda}_n$

It is clear from Eq.(1.22) that only the first $N+1$ values of the Poisson probabilities are used in the $\tilde{S}_n^0(t)$ estimations. It implies that some of the weights will be distributed beyond the largest order statistic of the sample data. This results in shortening the right tail, and it may cause *oversmoothing* in some cases.

To prevent or at least minimize this undesired effect, we need to alter the weights' shapes in such way that we bring some of the discarded weights back. This can be accomplished by making the maximum weights of the distribution occur earlier. The adjustment of the weights is possible by considering a smoothing parameter c, namely

$$\hat{\lambda}_n = c\frac{n}{T_{n:n}} \text{ where } c > 0. \tag{3.16}$$

This choice of $\hat{\lambda}_n$ is parallel to the bandwidth selection in kernel method of estimation. By using the smoothing parameter c, we can influence the degree of smoothness of $\tilde{S}_n^0(t)$. The effect of c will be illustrated in the section 3.2 through a numerical study.

Choice of N

The estimated value of N is defined as the smallest value of k so that $S_n\left(\frac{k}{\hat{\lambda}_n}\right) = 0$. That is

$$\hat{N} = inf\{k; S_n\left(\frac{k}{\hat{\lambda}_n}\right) = 0\}. \tag{3.17}$$

Since $S_n\left(\frac{k}{\hat{\lambda}_n}\right) = 0$ when $\frac{k}{\hat{\lambda}_n} = T_{n:n}$, then

$$\begin{aligned} k &= \hat{\lambda}_n T_{n:n}, \\ &= c\frac{n}{T_{n:n}} T_{n:n}, \\ &= cn. \end{aligned} \tag{3.18}$$

Therefore $\hat{N} = cn$ (keep only the integer value).

3.2 Analysis of Weight Distribution of $\tilde{S}_n^0(t)$

The smooth estimator $\tilde{S}_n^0(t)$ is based on the sum of the product of the two main components. First, on the right side of Eq.(3.19) we have $S_n(\cdot)$. It is computed for a set of equally spaced points from 0 to $T_{n:n}$. These points are evaluated at $k/\hat{\lambda}_n$ where k varies from 0 to cn. The estimates of $S_n(\cdot)$ are not dependent on any choice of t.

$$\tilde{S}_n^0(t) = \sum_{k=0}^{cn} w_{nk}^0\left(t, \hat{\lambda}_n\right) S_n\left(\frac{k}{\hat{\lambda}_n}\right), \ \mathrm{t} \in R^+. \tag{3.19}$$

The other main component of $\tilde{S}_n^0(t)$ is the weight function $w_{nk}^0\left(t, \hat{\lambda}_n\right)$. The function is also computed at the same equally spaced points as in $S_n(\cdot)$. However, the weights have been adjusted to surround t. Referring to section 1.2, the weight distributions are based on the Poisson probabilities adapted to the parameter $t\hat{\lambda}_n$. Therefore, the occurrence of the maximum weights always appear around the value of t for $t < T_{n:n}$.

The weight function is analogous to the kernel function in the conventional kernel smoothing method. That is, a weight function $w_{nk}^0\left(t, \hat{\lambda}_n\right)$ is centered around each observation t. In contrast to the kernel, the weight function is not necessary symmetric but still integrates to 1. To illustrate the different forms of the weight distribution, we take a sample of size n equal to 10 and the largest order statistic $T_{n:n}$ "fixed" to 5. Then, we take three values of

$t = \{0.833, 2.5, 5.0\}$ and a fix value of $c = 1.0$. Next, we compute the weight distributions for each t.

Recall that the maximum number of points k used in the sum of Eq.(3.19) is equal to cn where we only keep the integer part of cn. Consequently the number of points is fixed for any given t but varies either by changing n or c.

The effect of t on the weight distributions is displayed in the figure 3.1. Note that the line-graph format is used here only to make the display less cluttered. The weight distribution is not continuous, rather it gives probabilities for discrete value of k. In other words these graphs only "connect-the-dots".

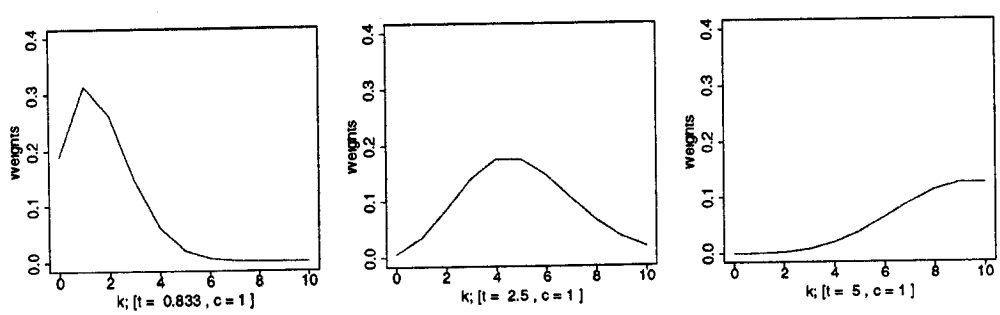

Figure 3.1: Variation of the weight distributions with various t ($c = 1$, $T_{n:n} = 5$, $n = 10$).

It appears that the weights shift from a right-skewed distribution to a symmetric bell shape then to a left-skewed distribution. Also, note that the maximum weight in each panel gradually decreases, and the occurrence of maximum weight moves to the right as the value of t increases. We can see that starting at the left panel at $t = 0.833$, the maximum weight reaches up to 0.314. At $t = 2.5$ the maximum reduces to 0.175, and the value of maximum weight is 0.125 at $t = 5$ as seen in the right panel.

To see this behavior in more detail, table 3.1 shows the weight values for the three values of t. Under our conditions, we have 11 equally spaced points $(k/\hat{\lambda}_n)$ in the estimation. As mentioned previously, the maximum weight at

					$Values of k$					
0	1	2	3	4	5	6	7	8	9	10
					Equally spaced points: $k/\hat{\lambda}_n$					
0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
					Weight function					
					$w_{nk}^0\left(t = 0.833, \hat{\lambda}_n\right)$					
0.189	0.314	0.262	0.145	0.060	0.020	0.005	0.001	2.7e-4	5.1e-5	8.5e-6
					$w_{nk}^0\left(t = 2.5, \hat{\lambda}_n\right)$					
0.006	0.033	0.084	0.140	0.175	0.175	0.146	0.104	0.065	0.036	0.018
					$w_{nk}^0\left(t = 5.0, \hat{\lambda}_n\right)$					
4.5e-5	4.5e-5	0.002	0.007	0.018	0.037	0.063	0.090	0.112	0.125	0.125

Table 3.1: Detailed information about the weight distributions with various t ($c = 1$, $T_{n:n} = 5$, $n = 10$).

$t = 0.833$ is 0.314. This maximum occurs when $k/\hat{\lambda}_n$ is in the neighborhood of 0.833. Similarly, when $t = 2.5$, the maximum weight is achieved when $k/\hat{\lambda}_n$ is in the same magnitude of t.

All values of $S_n(\cdot)$ contribute to the estimation of $\tilde{S}_n^0(t)$ but with different level of importance. The farther the points are from t, the lesser importance is given to $S_n(\cdot)$ at these points.

The behavior of the weights when t increases can possibly create *oversmoothing*. The continuous reduction of the weight as t increases may alienate the contribution of each $S_n(\cdot)$ in the estimation of $\tilde{S}_n^0(t)$ and consequently generate oversmoothed results, particularly at the right tail region.

In fear of *oversmoothing*, we should use the smoothing parameter c to control the level of smoothing. The figure 3.2 shows the effect of c on the weight distributions. The graphs on the horizontal lines represent the effect of t on the distribution while c remains the same. The graphs on the vertical columns display the impact of changing c on the weight distributions while t is fixed.

We observe, for a given t, that the increase of c shifts the distributions to the right and reduces the individual weight values. This behavior is common among all values of t. It also appears that the effects of c generates similar behaviors on the weights as the effect of t.

The figure 3.3 further expands the effect of c on the weight distributions, even beyond the largest order statistic.

On each panel, the systematic reduction of the weight distribution is observed as c increases. Even though this reduction appears less significant in the tail region (t $\geq T_{n:n} = 5$) compared to lower values of t, we are still able to observe the effect of the smoothing parameter.

The variation of the weights in the tail makes the adjustment of the thickness in tail possible. In some situations, it might be quite useful to have a set of estimates corresponding to different c. Those estimates can highlight different aspects in the structure of the data. However, the presentation and interpretation of such curves are quite subjective. Thus, it is necessary to evaluate an approxpriate value of c.

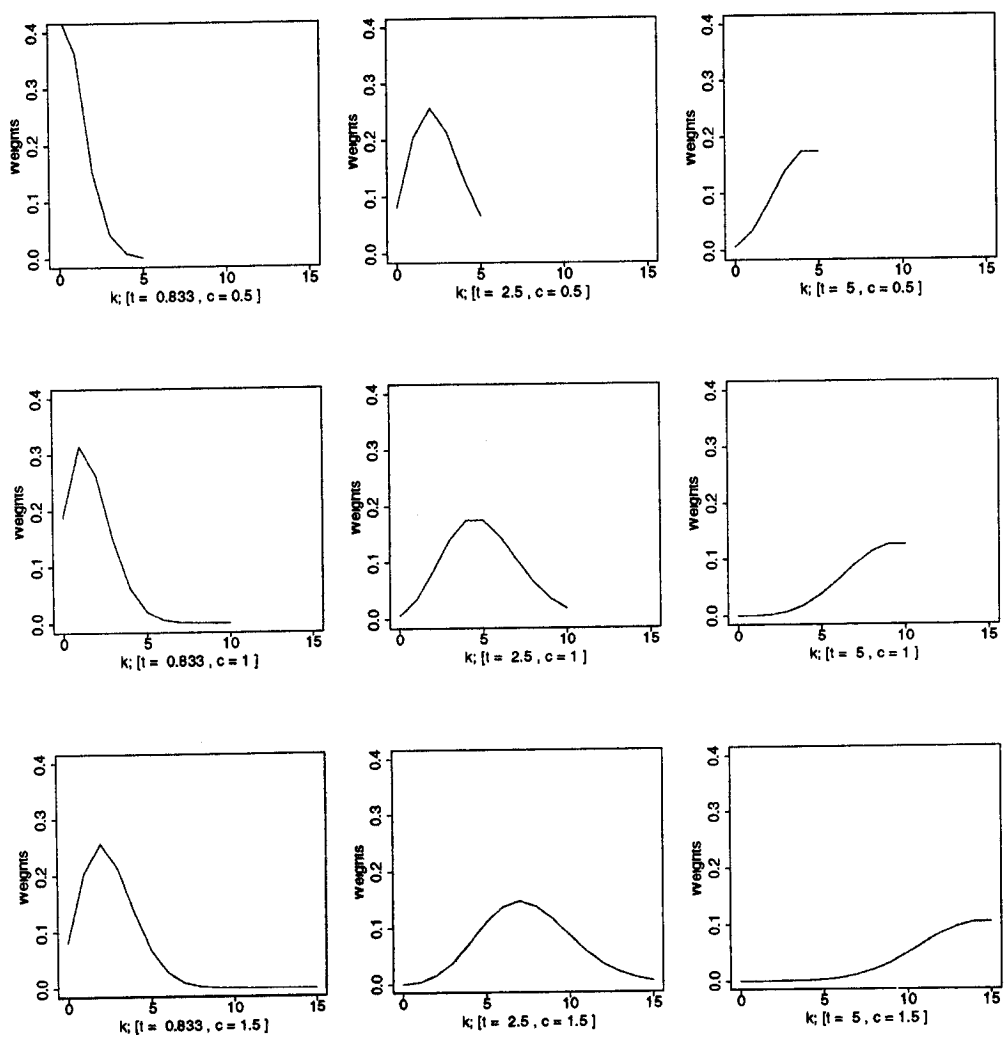

Figure 3.2: Variation of the weight distributions with various t and c ($T_{n:n} = 5$, $n = 10$).

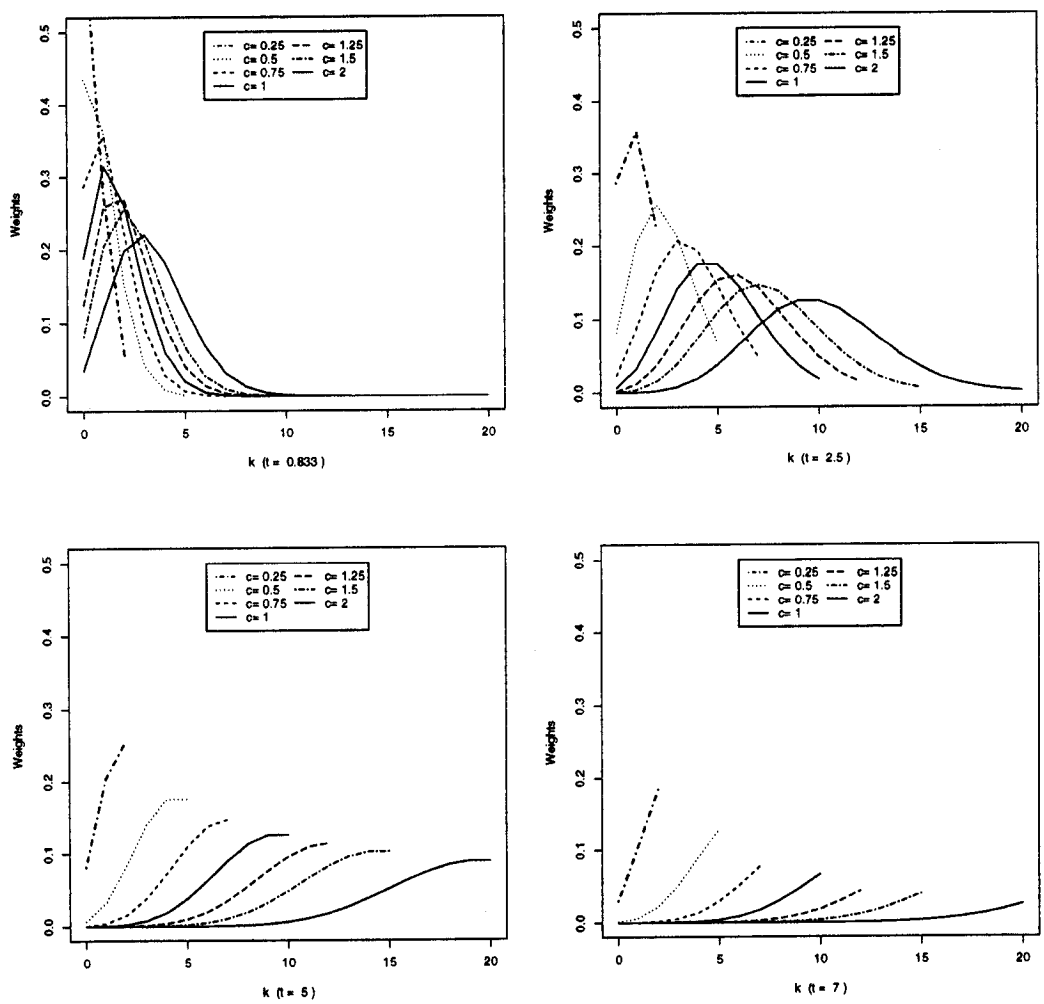

Figure 3.3: Expanded view of the effect of c on the weights distribution with various t.

3.2.1 Effect of c on the Smooth Estimator

To get a feel for the Chaubey and Sen $\tilde{S}_n^0(t)$ estimator, let's consider its estimation with different population distributions. For this study, we choose random samples from the Exponential(1) because of its thin right tail, the Gamma(1,4) and Weibull(1,4) for their moderate tails and the Lognormal(0,1) for its thick tail. Figures 3.4 to 3.7 illustrate the estimated $\tilde{S}_n^0(t)$ with $c = \{0.25, 0, 5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0\}$ and $n = 10$ for these random samples.

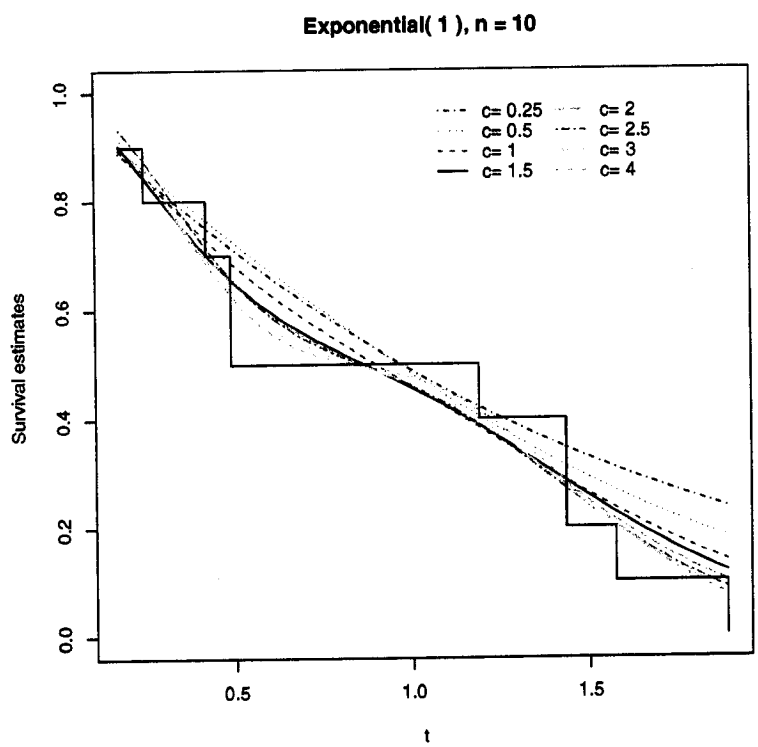

Figure 3.4: Chaubey and Sen smoothed survival function estimations for the Exponential(1) sample.

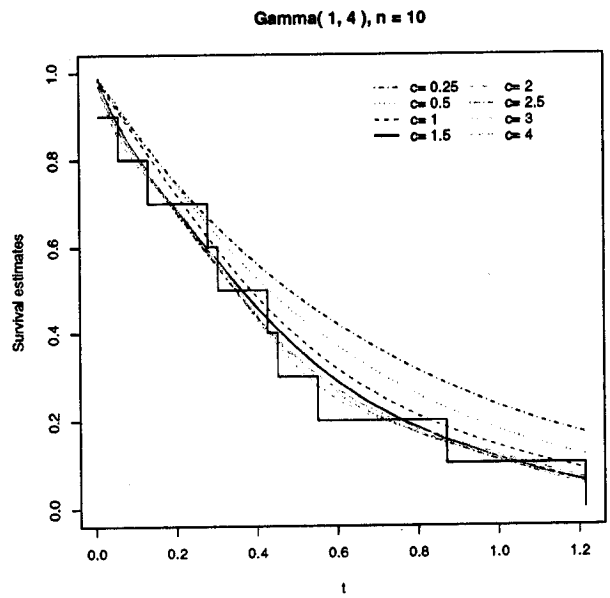

Figure 3.5: Chaubey and Sen smoothed survival function estimations for the Gamma(1,4) sample.

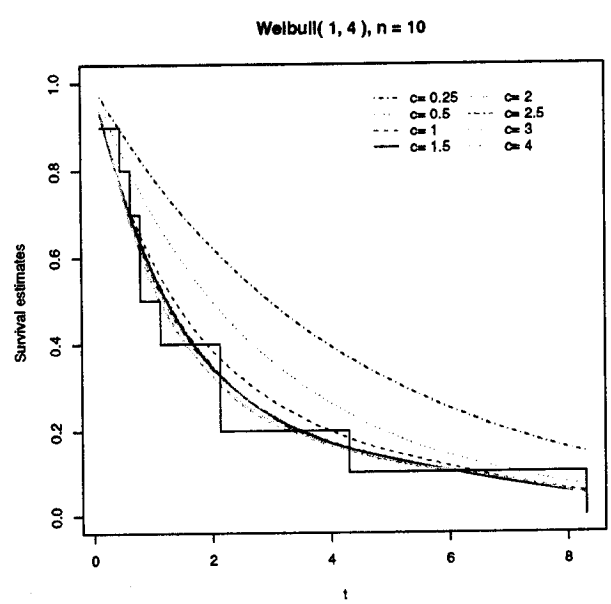

Figure 3.6: Chaubey and Sen smoothed survival function estimations for the Weibull(1,4) sample.

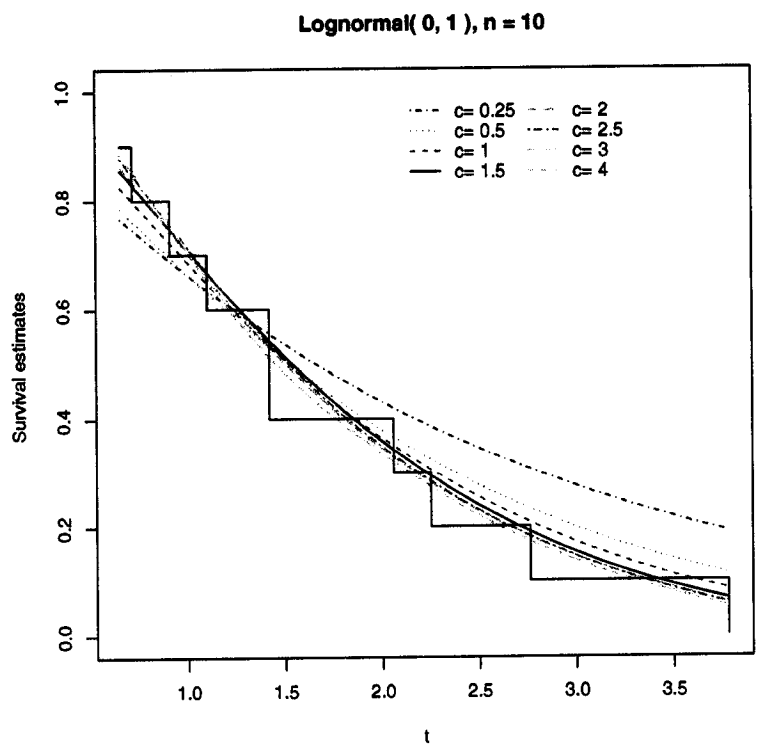

Figure 3.7: Chaubey and Sen smoothed survival function estimations for the Lognormal(0,1) sample.

In all random samples, the thickness of the tail gradually becomes thinner as c increases. The increase in c shifts the weight distributions toward right and reduces the weight values. As a result, only few larger order statistics will contribute to the estimate $\tilde{S}_n^0(t)$. Consequently, the tail will have a propensity to lower. As seen in the previous section, the effect of c diminishes as c becomes larger. In particular, when c is larger than 2.5, the change in the tail is quite small. This may indicates that c may not be so effective beyond certain values.

The range of the estimates derived from the various c at the tail ($t = T_{n:n}$) seems to be wide enough to cover the true tail thickness. This brings the question of what could be the *best* choice of c. The following sections present some techniques for obtaining a reasonable value of c.

3.3 Optimum Value of c

The choice of c is crucial for the smooth estimator $\tilde{S}_n^0(t)$. The qualitative behavior of the estimator described in the last section highlights the need for finding an optimum c in a more systematic approach. Mainly, we have to use a strategy that helps minimizing *oversmoothing* and yet provides a smooth estimation of the empirical survival function.

In this study, we adopt a global measure estimate for finding an optimum value of the smoothing parameter c.

3.3.1 Mean Squared Error

The basic approach considered here is to apply a global measure of closeness between $\tilde{S}_n^0(\cdot)$ and $S_n(\cdot)$. This methodology has the advantage of not focusing on particular region of the function. Rather it emphasizes on the overall appropriateness of the smooth estimates.

We propose to find the value of c which minimizes the mean square difference of the smooth estimator and the empirical survival function.

$$MSE_{emp.}(c) = \frac{\sum_{i=1}^{n} \left(\tilde{S}_n^0(t_i) - S_n(t_i)\right)^2}{n}. \tag{3.20}$$

Clearly, the function $MSE_{emp.}(c)$ is not a smooth function of c since $S_n(\cdot)$ is a step function. Therefore, we must use a numerical search algorithm to obtain the value of c that minimizes $MSE_{emp.}(c)$. To achieve reasonable accuracy, we consider a sequence of 500 sub-intervals of c in $]0,U]$ where U is dependent of the sample size. As seen before, the maximum k is derived from cn. If the sample size and c are large, then k will be very large. Consequently, the portion of the Poisson weights $(\lambda t)^k / k!$ may overflow. For this reason, c should be restricted up to certain value U in practical application.

Using our foregoing samples, we evaluate the smooth function for each of

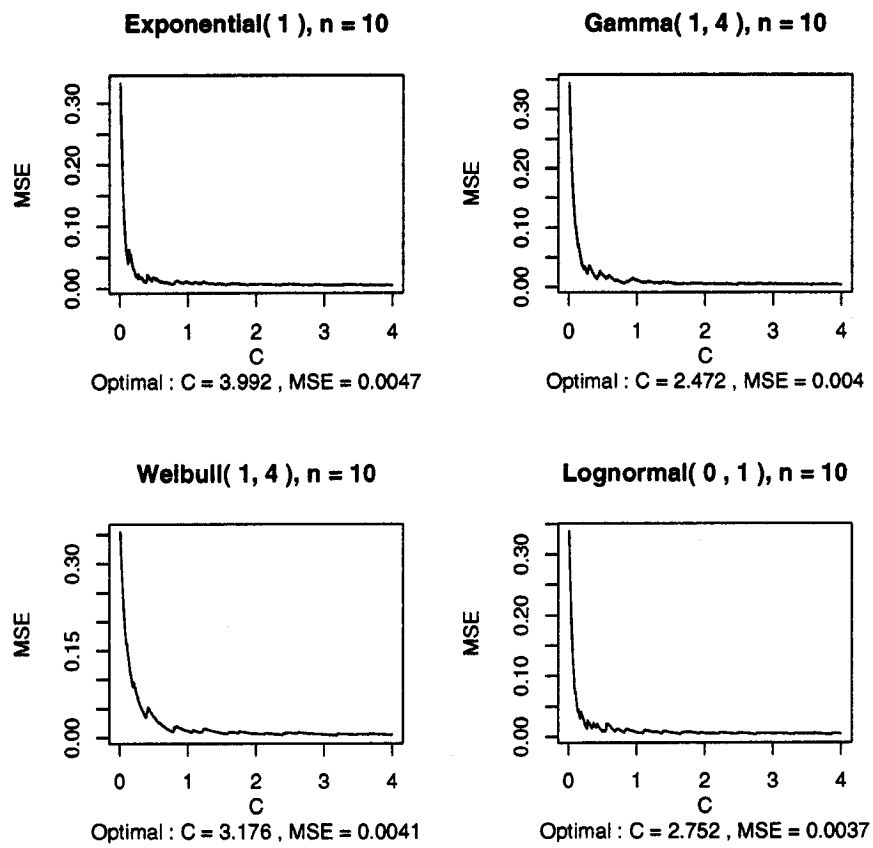

Figure 3.8: Typical optimal search path of the smoothing parameter c. Each panel indicates for the optimal c along with the corresponding $MSE_{emp.}(c)$ of the running samples.

the sub-intervals with c in]0,4]. The value of U was set to 4 to enable the use of the maximum spectrum for c of the distributions. Figures 3.8 shows typical searching paths of c. It is clear that $MSE(c)$ decreases with as c increases and it appears that the minimum MSE values lie in the region with c greater then 1.0 in all cases.

The table 3.2 summarizes the investigation of the optimum c value for different sample sizes. It appears that the Gamma sample in general has the smallest values of optimum c among the distributions. The $MSE(c)$ diminishes significantly as n increases in all samples except the Lognormal in which the

MSE seems to remain constant.

Summary

	Exponential(1)		Gamma(1,4)	
n	**c**	**MSE(c)**	**c**	**MSE(c)**
10	3.992	0.0046	2.472	0.0040
15	3.808	0.0027	3.216	0.0019
20	3.776	0.0012	3.352	0.0013
30	3.776	0.0006	3.824	0.0008

	Weibull(1,4)		Lognormal(0,1)	
n	**c**	**MSE(c)**	**c**	**MSE(c)**
10	3.176	0.0040	2.772	0.0037
15	3.864	0.0023	3.792	0.0016
20	3.688	0.0019	3.944	0.0018
30	3.864	0.0011	3.968	0.0017

Table 3.2: Summary of optimum c values.

Continuing with our random samples, figure 3.9 to 3.12 display the estimator $\tilde{S}_n^0(t)$ at optimum c with sample sizes $n = 10, 15, 20$ and 30. At optimum c, the smooth estimator in general appears to be in accordance with $S_n(\cdot)$; smoothing the bumps and valleys of the $S_n(\cdot)$. Their closeness is even more apparent as n increases.

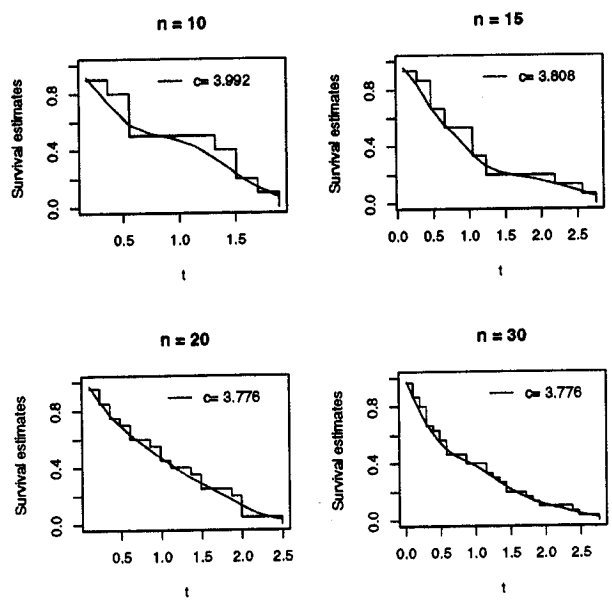

Figure 3.9: Behavior of $\tilde{S}_n^0(t)$ at optimum c as n increases: Exponential(1).

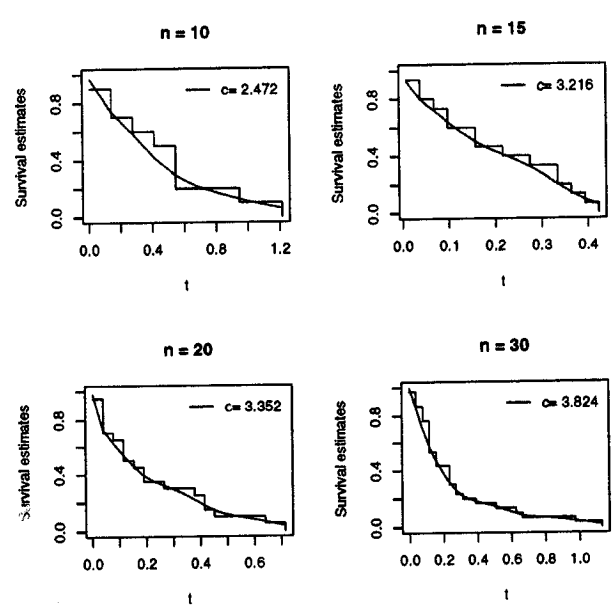

Figure 3.10: Behavior of $\tilde{S}_n^0(t)$ at optimum c as n increases: Gamma(1,4).

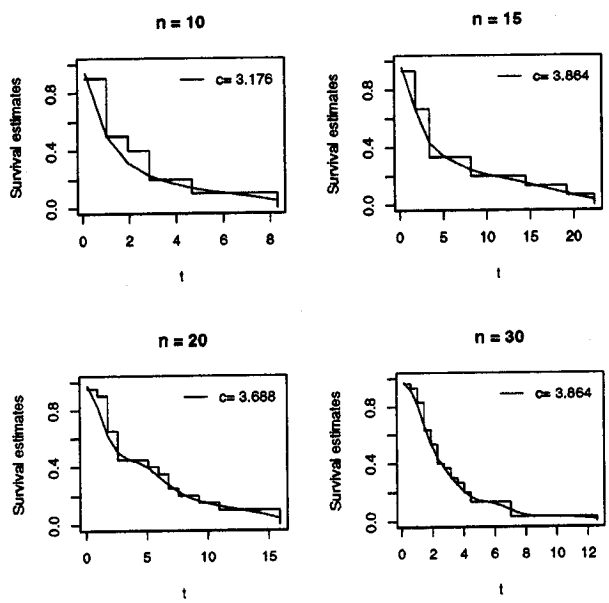

Figure 3.11: Behavior of $\tilde{S}_n^0(t)$ at optimum c as n increases: Weibull(1,4).

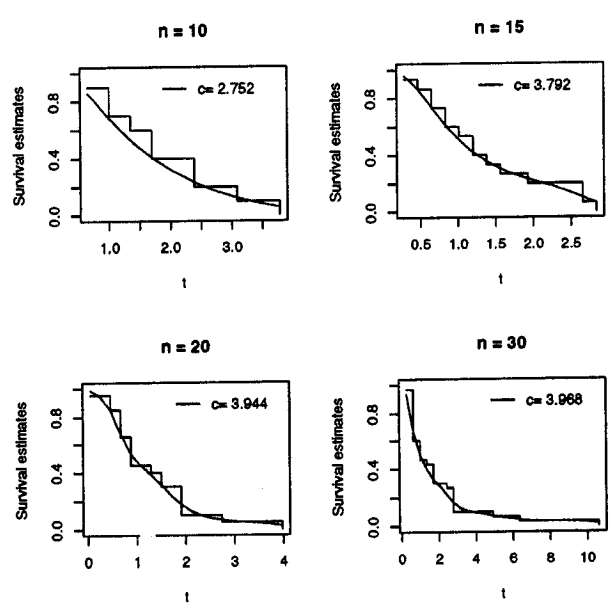

Figure 3.12: Behavior of $\tilde{S}_n^0(t)$ at optimum c as n increases: Lognormal(0,1).

3.3.2 General Model for Optimum c

The computational effort required by the search algorithm of an optimum value of c could be time consuming when the sample size becomes very large. For this reason, we investigate an approximate method to obtain the optimum value of c with less effort.

To proceed, twenty five samples of sizes $n = 10(5)50(10)100$ are considered. For each sample, the mean squared error algorithm is applied. The figure 3.13 illustrates the possible relation between the optimum c values and the sample sizes. The variation in the optimum c is wider when n is small. This is because smaller n allows using higher values of c. It appears that a polynomial model may be suitable. A tentative modeling process has been performed, using polynomial regression in $1/n$ to obtain a general expression of the form of

$$c(n) = B_0 + \frac{B_1}{n} + \frac{B_2}{n^2} + ... + \frac{B_p}{n^p}, \tag{3.21}$$

where p is any positive integer.

Table 3.3 contains the coefficients of such models applied to the four running distributions. It is interesting to observe that the coefficients values among the distribution models are similar in magnitude. In fact, we may postulate the formula as

$$c(n) = -1.6085 + \frac{3.5497}{n} \frac{-8.1581e03}{n^2} + \frac{7.7560e04}{n^3} - \frac{2.6235e05}{n^4}. \tag{3.22}$$

The coefficients were obtained by computing an overall model for all the distributions together. The figure 3.14 gives a visual representation of the model. The above formula seems appropriate for the populations considered in this thesis. However, it provides only crude estimates of optimum smoothing parameter c when time and resource are limited, and it may not be appropriate for other distributions.

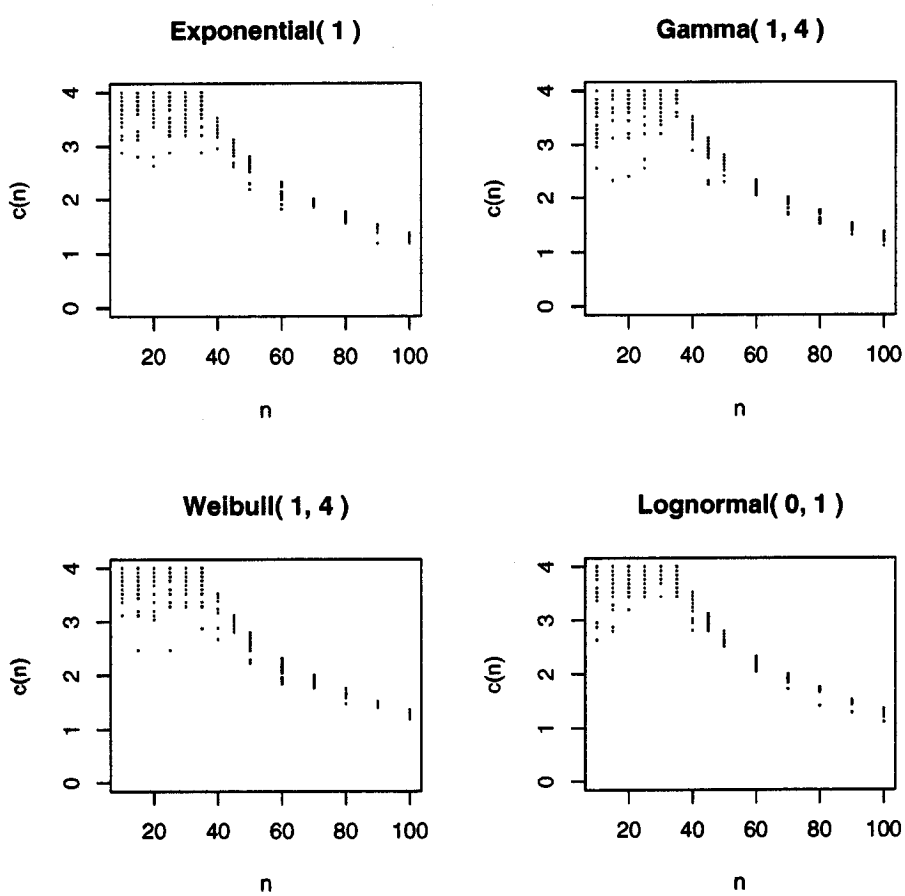

Figure 3.13: Optimum c as a function of n

Coefficient values of the polynomial models

Exponential(1)

B_0	B_1	B_2	B_3	B_4
-1.6609	3.6322e02	-8.6074e03	8.5657e04	-3.0586e05

Gamma(1,4)

B_0	B_1	B_2	B_3	B_4
-1.7124	3.6925e02	-8.7907e03	8.7221e04	-3.0987e05

Weibull(1,4)

B_0	B_1	B_2	B_3	B_4
-1.5211	3.4236e02	-7.5976e03	6.8722e04	-2.1765e05

Lognormal(0,1)

B_0	B_1	B_2	B_3	B_4
-1.5397	3.4502e02	-7.6368e03	6.8639e04	-2.1604e05

Table 3.3: This table contains the coefficient values of the polynomial of degree 4 created for each distributions.

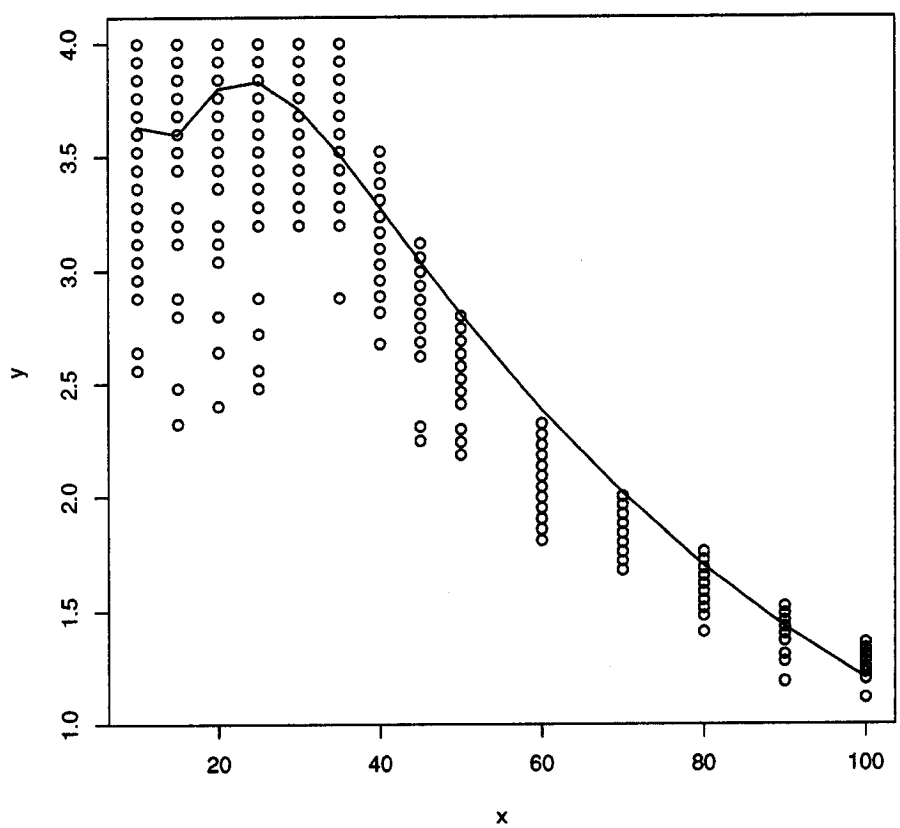

Figure 3.14: Combined model of optimum c as a function of n for all the distributions.

Chapter 4

Statistical Properties of $\tilde{S}_n^0(t)$

4.1 Introduction

In this chapter, we investigate the statistical properties of the untruncated estimator $\tilde{S}_n^0(t)$ using simulations. In particular, we apply Monte Carlo and Bootstrap simulations to obtain estimates of the bias and the variance. We also investigate the construction of confidence intervals for $\tilde{S}_n^0(t)$ estimates. Another important aspect investigated in this chapter is the tail behavior of the smooth estimator.

For simulations, we selected the Exponential, Gamma, Weibull and Lognormal. These families are known to generate a variety of shapes of distributions with non-negative support. Therefore they provide a broad spectrum of choices for the properties to be studied.

Section 4.2 presents a strategy to assess the *oversmoothing* in tails. In section 4.3, we study the appropriateness of several confidence intervals for $\tilde{S}_n^0(t)$.

4.1.1 Monte Carlo Simulation

The sampling distribution of a statistic $\hat{\theta}$ can be considered as the distribution of the values of that statistic calculated from an infinite number of random

samples of size n for a given population. Monte Carlo simulation takes this concept literally, building an estimate of the sampling distribution by drawing a large number of samples of size n randomly from a population, and calculating the statistic for each of these samples.

For our simulation study, we draw R samples each of size n from Exponential(1), Gamma(1,4), Weibull(1,4) and Lognormal(1) distributions. For each of the samples we calculate $\tilde{S}_n^0(t)$ at optimum c. To help the comparison of the results, three t values are chosen to correspond to $S(t) = 0.9$, $S(t) = 0.5$ and $S(t) = 0.1$ for all distributions.

The Monte Carlo $\tilde{S}_n^0(t)$ estimates are calculated for each trial, and the summary statistics are computed as following:

$$mean = \overline{\tilde{S}_n^0(t)} = \sum_{i=1}^{R} \frac{\left[\tilde{S}_n^0(t)\right]_i}{R}, \tag{4.1}$$

$$min = \min\{\left[\tilde{S}_n^0(t)\right]_i\}, \text{ for } i = \{1, 2, ..., R\}, \tag{4.2}$$

$$max = \max\{\left[\tilde{S}_n^0(t)\right]_i\}, \text{ for } i = \{1, 2, ..., R\}, \tag{4.3}$$

$$bias = \sum_{i=1}^{R} \frac{\left[\tilde{S}_n^0(t)\right]_i - [S_n(t)]_i}{R}, \tag{4.4}$$

$$se = \left\{ \sum_{i=1}^{R} \left[\left[\tilde{S}_n^0(t)\right]_i - \sum_{j=1}^{R} \frac{\left[\tilde{S}_n^0(t)\right]_j}{R} \right]^2 /(R-1) \right\}^{1/2} \tag{4.5}$$

We start with $R = 1{,}000$ random samples drawn from our four populations. The sample sizes investigated are $n = 10$, 15, 20, and 30. The summary statistics over the 1,000 estimates are reported in Tables 4.1 and 4.2.

As seen in the tables, the results indicate that the averages of the $\tilde{S}_n^0(t)$ are in general positively biased for the Exponential, Gamma and Weibull samples. The Lognormal samples seem to have negative biases for all n when t is small ($t = 0.2776$).

The standard errors seem to be lower with small values of t and remain relatively constant from the midrange t to the tail. In fact, the errors on average reduce slightly in the tail. Asymptotically, the standard errors of all distributions reduce significantly and present similar estimates, in particular when t increases.

It appears clearly that the Exponential, Gamma and Weibull sample results are comparable in their biases and standard errors. The Lognormal samples show higher biases and standard errors among all distributions. It is a known fact that the Lognormal distribution has a heavy tail. Thus we should expect more variability in its tail than in the other distributions.

We should stress that the values of the biases are significantly large in the tail area, but more importantly when n is small. Although we observe a constant reduction in the bias values as n increases, all values remain positive.

The frequency distributions and normality plots of the figures 4.1 to 4.16 illustrate the variability of the estimates. As it may be seen, the distributions are right skewed when t is small, nearly symmetric in the mid t, and left skewed when t if large. These patterns could be explained partly by the positive biases of $\tilde{S}_n^0(t)$ and the lower and the upper bound of the survival function. In addition, the propensity of the smooth estimator to lift up the tail result in higher values of $\tilde{S}_n^0(t)$ in this area, and thus more estimates are expected to be above the theoretical survival probabilities.

Monte Carlo Simulations

			$\tilde{S}_n^0(t)$				
n	**t**	**S(t)**	**mean**	**bias**	**se**	**min**	**max**
			Exponential(1)				
10	0.1054	0.9	0.9226	0.0270	0.0287	0.7565	0.9882
	0.6931	0.5	0.5955	0.0921	0.1053	0.1663	0.8337
	2.3026	0.1	0.1860	0.0799	0.1006	0.0001	0.5079
15	0.1054	0.9	0.9173	0.0214	0.0265	0.7405	0.9813
	0.6931	0.5	0.5732	0.0743	0.0886	0.1830	0.7983
	2.3026	0.1	0.1644	0.0642	0.0779	0.0001	0.4115
20	0.1054	0.9	0.9142	0.0173	0.0248	0.7952	0.9894
	0.6931	0.5	0.5623	0.0618	0.0781	0.2980	0.7708
	2.3026	0.1	0.1542	0.0536	0.0669	0.0063	0.3979
30	0.1054	0.9	0.9100	0.0109	0.0239	0.8011	0.9709
	0.6931	0.5	0.5470	0.0463	0.0691	0.3155	0.7463
	2.3026	0.1	0.1409	0.0403	0.0545	0.0071	0.3469
			Lognormal(0,1)				
10	0.2776	0.9	0.8729	-0.0227	0.0501	0.6505	0.9832
	1.0000	0.5	0.6154	0.1213	0.1256	0.1855	0.9409
	3.6022	0.1	0.2036	0.1013	0.1426	0.0001	0.8048
15	0.2776	0.9	0.8698	-0.0271	0.0417	0.6925	0.9740
	1.0000	0.5	0.6042	0.1034	0.1013	0.2051	0.9097
	3.6022	0.1	0.1868	0.0835	0.1099	9.7e-5	0.7136
20	0.2776	0.9	0.8671	-0.0308	0.0358	0.7022	0.9702
	1.0000	0.5	0.5942	0.0943	0.0854	0.2882	0.8970
	3.6022	0.1	0.1755	0.0712	0.0907	0.0015	0.6785
30	0.2776	0.9	0.8647	-0.0338	0.0310	0.7623	0.9579
	1.0000	0.5	0.5807	0.0803	0.0710	0.3687	0.8568
	3.6022	0.1	0.1593	0.0556	0.0698	0.0064	0.5757

Table 4.1: Estimates of $\tilde{S}_n^0(t)$ obtained from the Monte Carlo simulations of the Exponential(1) and Lognormal(0,1) random samples (trials = 1,000).

Monte Carlo Simulations

			$\tilde{S}_n^0(t)$				
n	**t**	**S(t)**	**mean**	**bias**	**se**	**min**	**max**
			Gamma(1,4)				
10	0.0263	0.9	0.9207	0.0204	0.0292	0.7156	0.9886
	0.1733	0.5	0.5869	0.0869	0.1072	0.0666	0.8084
	0.5756	0.1	0.1806	0.0797	0.1012	2.5e-6	0.4748
15	0.0263	0.9	0.9169	0.0143	0.0263	0.7816	0.9861
	0.1733	0.5	0.5721	0.0699	0.0897	0.2241	0.8284
	0.5756	0.1	0.1652	0.0623	0.0804	0.0002	0.4247
20	0.0263	0.9	0.9148	0.0148	0.0250	0.8037	0.9813
	0.1733	0.5	0.5633	0.0594	0.0787	0.3023	0.7912
	0.5756	0.1	0.1558	0.0533	0.0678	0.0092	0.3916
30	0.0263	0.9	0.9105	0.0093	0.0237	0.7968	0.9711
	0.1733	0.5	0.5473	0.0452	0.0668	0.3256	0.7300
	0.5756	0.1	0.1428	0.0397	0.0534	0.0041	0.3079
			Weibull(1,4)				
10	0.4214	0.9	0.9232	0.0215	0.0280	0.7487	0.9837
	2.7726	0.5	0.5944	0.0902	0.1054	0.1435	0.8419
	9.2103	0.1	0.1853	0.0893	0.1025	8.8e-5	0.4897
15	0.4214	0.9	0.9179	0.0175	0.0260	0.8143	0.9891
	2.7726	0.5	0.5758	0.0719	0.0890	0.2524	0.8396
	9.2103	0.1	0.1680	0.0612	0.0805	0.0007	0.4263
20	0.4214	0.9	0.9152	0.0153	0.0250	0.8180	0.9781
	2.7726	0.5	0.5636	0.0607	0.0790	0.3112	0.7651
	9.2103	0.1	0.1555	0.0507	0.0667	0.0017	0.3447
30	0.4214	0.9	0.9113	0.0123	0.0232	0.8308	0.9736
	2.7726	0.5	0.5483	0.0461	0.0668	0.3210	0.7425
	9.2103	0.1	0.1417	0.0378	0.0534	0.0068	0.3200

Table 4.2: Estimates of $\tilde{S}_n^0(t)$ obtained from the Monte Carlo simulations of the Gamma(1,4) and Weibull(1,4) random samples (trials = 1,000).

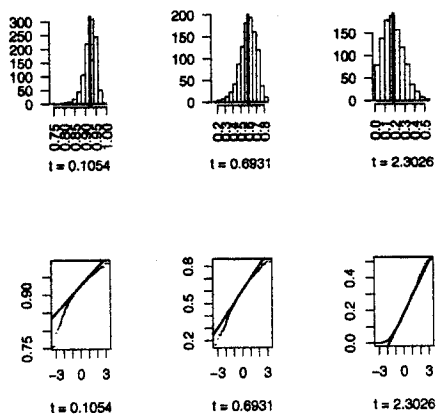

Figure 4.1: Monte Carlo distributions for 1,000 replicates Exponential(1) samples ($n = 10$). The solid lines in each histogram marks the simulation means.

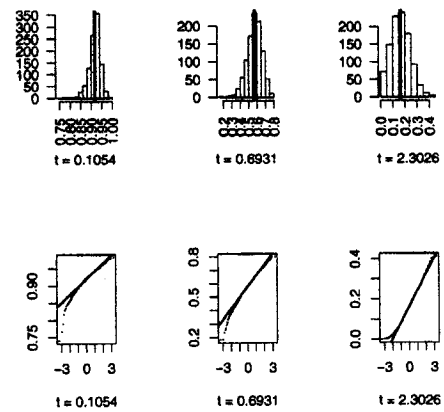

Figure 4.2: Monte Carlo distributions for 1,000 replicates Exponential(1) samples ($n = 15$). The solid lines in each histogram marks the simulation means.

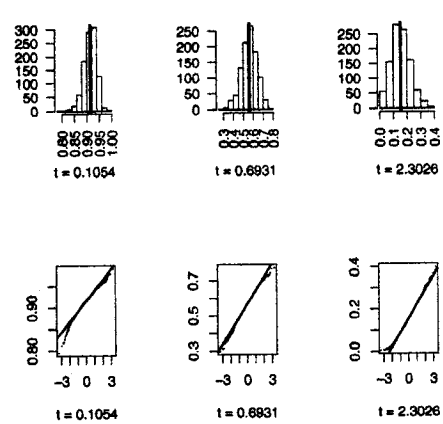

Figure 4.3: Monte Carlo distributions for 1,000 replicates Exponential(1) samples ($n = 20$). The solid lines in each histogram marks the simulation means.

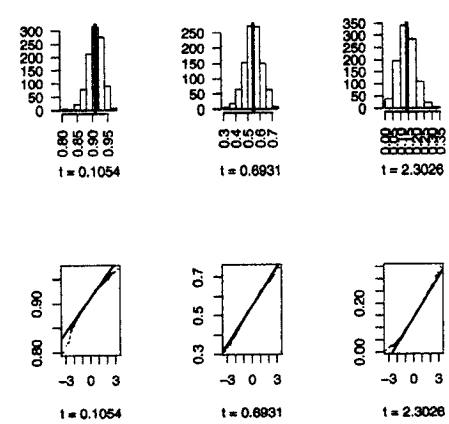

Figure 4.4: Monte Carlo distributions for 1,000 replicates Exponential(1) samples ($n = 30$). The solid lines in each histogram marks the simulation means.

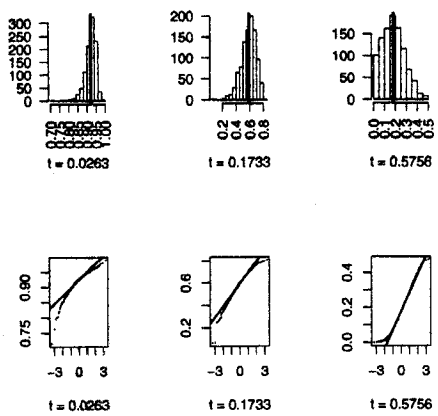

Figure 4.5: Monte Carlo distributions for 1,000 replicates Gamma(1,4) samples (n = 10). The solid lines in each histogram marks the simulation means.

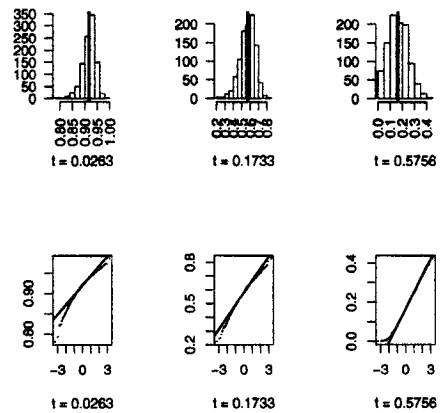

Figure 4.6: Monte Carlo distributions for 1,000 replicates Gamma(1,4) samples (n = 15). The solid lines in each histogram marks the simulation means.

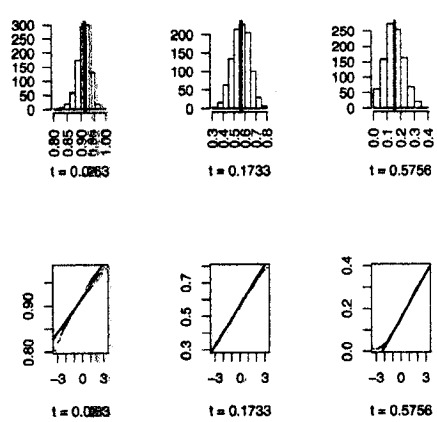

Figure 4.7: Monte Carlo distributions for 1,000 replicates Gamma(1,4) samples (n = 20). The solid lines in each histogram marks the simulation means.

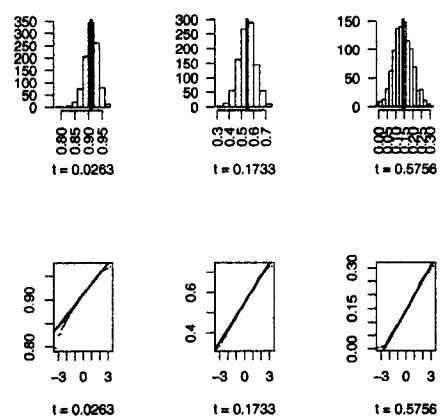

Figure 4.8: Monte Carlo distributions for 1,000 replicates Gamma(1,4) samples (n = 30). The solid lines in each histogram marks the simulation means.

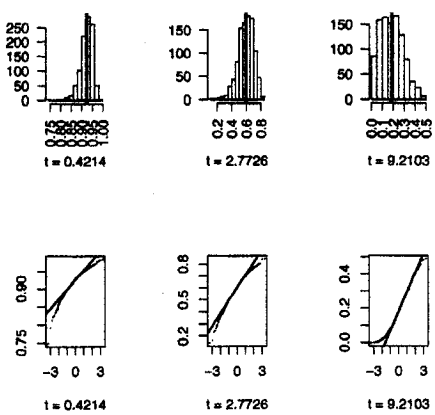

Figure 4.9: Monte Carlo distributions for 1,000 replicates Weibull(1,4) samples ($n = 10$). The solid lines in each histogram marks the simulation means.

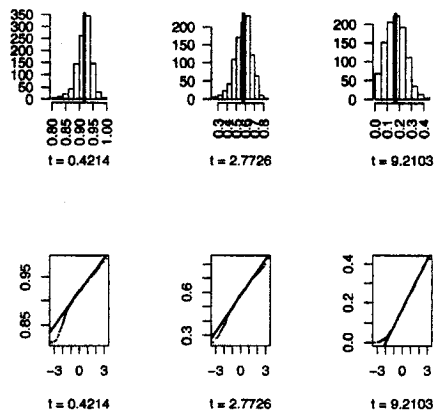

Figure 4.10: Monte Carlo distributions for 1,000 replicates Weibull(1,4) samples ($n = 15$). The solid lines in each histogram marks the simulation means.

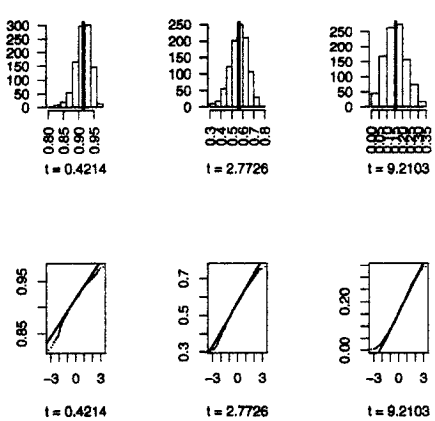

Figure 4.11: Monte Carlo distributions for 1,000 replicates Weibull(1,4) samples ($n = 20$). The solid lines in each histogram marks the simulation means.

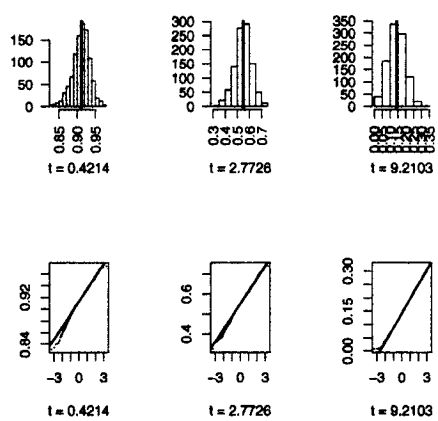

Figure 4.12: Monte Carlo distributions for 1,000 replicates Weibull(1,4) samples ($n = 30$). The solid lines in each histogram marks the simulation means.

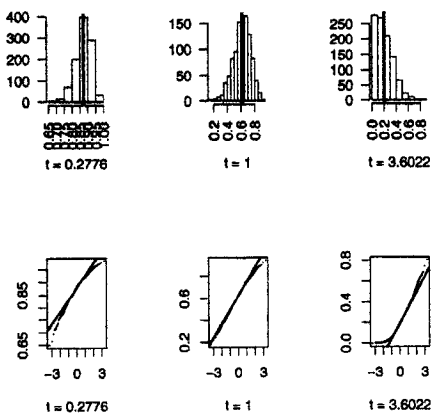

Figure 4.13: Monte Carlo distribution for 1,000 replicates Lognormal(0,1) samples ($n = 10$). The solid lines in each histogram marks the simulation means.

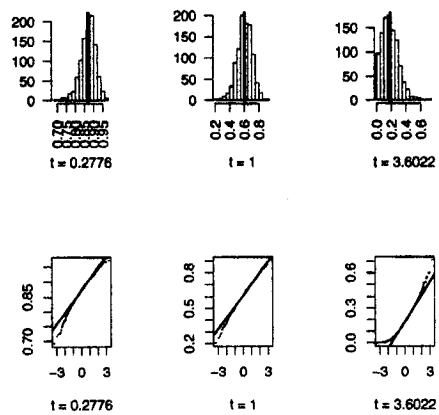

Figure 4.14: Monte Carlo distributions for 1,000 replicates Lognormal(0,1) samples ($n = 15$). The solid lines in each histogram marks the simulation means.

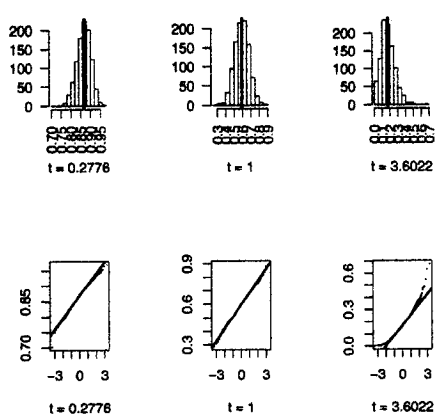

Figure 4.15: Monte Carlo distributions for 1,000 replicates Lognormal(0,1) samples ($n = 20$). The solid lines in each histogram marks the simulation means.

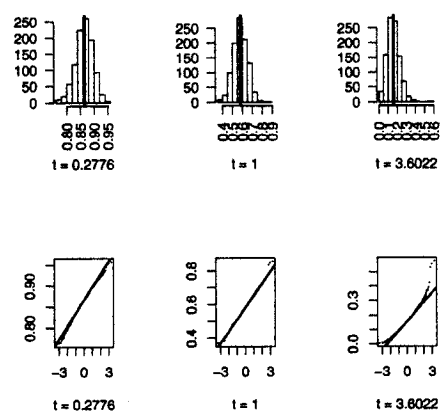

Figure 4.16: Monte Carlo distributions for 1,000 replicates Lognormal(0,1) samples ($n = 30$). The solid lines in each histogram marks the simulation means.

4.1.2 Bootstrap Simulation

The main disadvantage of the Monte Carlo simulation is that it requires the knowledge of the underlying distribution. In practice, one has a sample on hand and does not normally know which distribution the sample comes from. Thus, Monte Carlo simulation can not be performed properly in such a situation. However, in Bootstrap simulation it is possible.

The Bootstrap simulation was introduced by Efron (1979) as a computer-based method for estimating the standard error of a statistic $\hat{\theta}$. It has the advantage of being completely automatic. The Bootstrap estimate of standard error requires no theoretical calculations, and it is available no matter how mathematically complicated the statistic $\hat{\theta}$ may be.

The basic Bootstrap approach is to treat the sample as if it were the population, and then to apply Monte Carlo simulation sampling to generate an empirical estimate of the statistic's sampling distribution. This is accomplished by drawing a large number of resamples of size n from the original sample with random replacement. Although each resample will have the same number of elements as the original sample, through replacement resampling, each resample could have some of the original data points represented more than once, and some not represented at all. Therefore, each of these resamples will likely be slightly and randomly different from the original sample.

The Bootstrap algorithm, described next, is a computational way of obtaining good approximations of the bias and the standard error.

Bootstrap Algorithm:

(i) Evaluate $\tilde{S}_n^0(t)$ for a given sample.

(ii) Select B independent Bootstrap samples, each consisting of n data values drawn with replacement from the original sample.

(iii) Compute the smooth estimator $\tilde{S}_n^0(t)$ for each bootstrap sample.

(iv) Compute the bias and standard error of the smooth estimator of the original sample as

$$\hat{\theta} = \tilde{S}_n^0(t), \quad \text{and} \quad \hat{\theta}_i^* = \left[\tilde{S}_n^0(t)\right]_i. \tag{4.6}$$

The bootstrap bias and standard errors formulas are defined as

$$bias = E\left(\hat{\theta}^*\right) - \hat{\theta} = \frac{\sum_{i=1}^{B} \hat{\theta}_i^*}{B} - S_n(t), \tag{4.7}$$

$$se = \left[\frac{\sum_{i=1}^{B}\left\{\hat{\theta}_i^* - \frac{\sum_{j=1}^{B}\hat{\theta}_j^*}{B}\right\}^2}{(B-1)}\right]^{1/2}. \tag{4.8}$$

Note that when bootstrapping, the original sample should be the empirical survival function and not the smooth function itself.

The practical magnitude of the number of resampling B depends on the tests to be run on the data. Typically, B should be ranging from 50 to 200 to estimate the standard errors of $\hat{\theta}$, and at least 1,000 to estimate the confidence intervals around $\hat{\theta}$ (see Efron & Tibshirani (1986)).

We start by using the 1,000 replicates generated in the Monte Carlo simulation. For each of the sample, we perform a Bootstrap simulation to compute estimates of the smooth estimator $\tilde{S}_n^0(t)$, its bias and its standard error. Each of the Bootstrap simulation uses $B = 999$ as the number of resamples. The Bootstrap smooth estimates $\tilde{S}_n^0(t)$ are evaluated at the optimum c of the original replicates. The frequency distributions of the 999 resamples are the bootstrapped estimat s of the sampling distribution of each $\tilde{S}_n^0(t)$.

The tables 4.3 and 4.4 list the averages of the smooth estimates, biases and standard errors obtained from the Bootstrap simulation over the 1,000 replicates.

Comparison results reveal that the Bootstrap biases for all distributions are in general lower than those of the Monte Carlo simulation. However, the biases in Bootstrap simulation are somewhat higher when t is smallest. We observe also that the Bootstrap standard errors exhibit lower values than the Monte Carlo on all distributions with the exception when the values of t are smaller.

Simulations Summary

			$\tilde{S}_n^0(t)$					
			Monte Carlo			**Bootstrap**		
n	**t**	**S(t)**	$\bar{\tilde{S}}_n^0$	**bias**	**se**	$\bar{\tilde{S}}_n^0$	**bias**	**se**
			Exponential(1)					
10	0.1054	0.9	0.9226	0.0270	0.0287	0.9205	0.0272	0.0297
	0.6931	0.5	0.5955	0.0921	0.1053	0.5825	0.0779	0.0962
	2.3026	0.1	0.1860	0.0799	0.1006	0.1714	0.0657	0.0718
15	0.1054	0.9	0.9173	0.0214	0.0265	0.9153	0.0217	0.0271
	0.6931	0.5	0.5732	0.0743	0.0886	0.5639	0.0674	0.0823
	2.3026	0.1	0.1644	0.0642	0.0779	0.1530	0.0540	0.0615
20	0.1054	0.9	0.9142	0.0173	0.0248	0.9128	0.0175	0.0256
	0.6931	0.5	0.5623	0.0618	0.0781	0.5558	0.0566	0.0753
	2.3026	0.1	0.1542	0.0536	0.0669	0.1463	0.0459	0.0566
30	0.1054	0.9	0.9100	0.0109	0.0239	0.9085	0.0118	0.0244
	0.6931	0.5	0.5470	0.0463	0.0691	0.5408	0.0443	0.0656
	2.3026	0.1	0.1409	0.0403	0.0545	0.1353	0.0349	0.0473
			Lognormal(0,1)					
10	0.2776	0.9	0.8729	-0.0227	0.0501	0.8714	-0.0219	0.0430
	1.0000	0.5	0.6154	0.1213	0.1256	0.5962	0.1010	0.0963
	3.6022	0.1	0.2036	0.1013	0.1426	0.1791	0.0771	0.0827
15	0.2776	0.9	0.8698	-0.0271	0.0417	0.8691	-0.0257	0.0368
	1.0000	0.5	0.6042	0.1034	0.1013	0.5901	0.0916	0.0796
	3.6022	0.1	0.1868	0.0835	0.1099	0.1673	0.0652	0.0720
20	0.2776	0.9	0.8671	-0.0308	0.0358	0.8673	-0.0288	0.0339
	1.0000	0.5	0.5942	0.0943	0.0854	0.5822	0.0836	0.0733
	3.6022	0.1	0.1755	0.0712	0.0907	0.1598	0.0555	0.0670
30	0.2776	0.9	0.8647	-0.0338	0.0310	0.8650	-0.0311	0.0302
	1.0000	0.5	0.5807	0.0803	0.0710	0.5698	0.0736	0.0627
	3.6022	0.1	0.1593	0.0556	0.0698	0.1475	0.0442	0.0552

Table 4.3: Estimates of $\tilde{S}_n^0(t)$ obtained from the Bootstrap simulations of Exponential(1) and Lognormal(0,1) random samples (R = 1,000, B = 999).

Simulations Summary

			$\tilde{S}_n^0(t)$					
			Monte Carlo			**Bootstrap**		
n	**t**	**S(t)**	$\overline{\tilde{S}_n^0}$	**bias**	**se**	$\overline{\tilde{S}_n^0}$	**bias**	**se**
				Gamma(1,4)				
10	0.0263	0.9	0.9207	0.0204	0.0292	0.9187	0.0206	0.0302
	0.1733	0.5	0.5869	0.0869	0.1072	0.5739	0.0727	0.0962
	0.5756	0.1	0.1806	0.0797	0.1012	0.1657	0.0651	0.0713
15	0.0263	0.9	0.9169	0.0143	0.0263	0.9157	0.0150	0.0267
	0.1733	0.5	0.5721	0.0699	0.0897	0.5640	0.0640	0.0810
	0.5756	0.1	0.1652	0.0623	0.0804	0.1545	0.0529	0.0610
20	0.0263	0.9	0.9148	0.0148	0.0250	0.9137	0.0122	0.0255
	0.1733	0.5	0.5633	0.0594	0.0787	0.5576	0.0550	0.0747
	0.5756	0.1	0.1558	0.0533	0.0678	0.1478	0.0454	0.0569
30	0.0263	0.9	0.9105	0.0093	0.0237	0.9089	0.0100	0.0242
	0.1733	0.5	0.5473	0.0452	0.0668	0.5411	0.0432	0.0654
	0.5756	0.1	0.1428	0.0397	0.0534	0.1369	0.0341	0.0480
				Weibull(1,4)				
10	0.4214	0.9	0.9232	0.0215	0.0280	0.9211	0.0216	0.0296
	2.7726	0.5	0.5944	0.0902	0.1054	0.5811	0.0756	0.0959
	9.2103	0.1	0.1853	0.0893	0.1025	0.1705	0.0649	0.0719
15	0.4214	0.9	0.9179	0.0175	0.0260	0.9164	0.0180	0.0266
	2.7726	0.5	0.5758	0.0719	0.0890	0.5674	0.0656	0.0814
	9.2103	0.1	0.1680	0.0612	0.0805	0.1573	0.0517	0.0617
20	0.4214	0.9	0.9152	0.0153	0.0250	0.9139	0.0156	0.0255
	2.7726	0.5	0.5636	0.0607	0.0790	0.5578	0.0562	0.0747
	9.2103	0.1	0.1555	0.0507	0.0667	0.1477	0.0432	0.0568
30	0.4214	0.9	0.9113	0.0123	0.0232	0.9096	0.0130	0.0243
	2.7726	0.5	0.5483	0.0461	0.0668	0.5424	0.0443	0.0654
	9.2103	0.1	0.1417	0.0378	0.0534	0.1361	0.0326	0.0476

Table 4.4: Estimates of $\tilde{S}_n^0(t)$ obtained from the Bootstrap simulations of Gamma(1,4) and Weibull(1,4) random samples (R = 1,000, B = 999).

4.2 Oversmoothing Assessment

The discussion of the previous sections draws attention to the inherent bias of the smooth estimator $\tilde{S}_n^0(t)$ in comparison to the empirical survival function. Although the positive bias could be a good indication that the estimator does not *oversmooth*, we still need to extend the examination of the estimator in this regard.

For a numerical assessment of this problem, we consider comparing the estimator $\tilde{S}_n^0(t)$ against $S_n(t)$ and against the exact s.f. $S(t)$. To proceed, we estimate the probability that $\tilde{S}_n^0(t)$ is lower than $S(t)$. Then, we compute the probability that $S_n(t)$ is lower than $S(t)$. Finally, we evaluate the probability that $\tilde{S}_n^0(t)$ is lower than $S_n(t)$. Here we know that we can compute the exact probability $\mathbf{Pr[S_n(t) < S(t)]}$ by using the binomial distribution, but $\mathbf{Pr[\tilde{S}_n^0(t) < S(t)]}$ can not be computed exactly. The latter probability can be approximated using the central limit theorem, however we have decided to use simulation to compute both probabilities.

We again use the 1,000 replicates of the Monte Carlo simulation to perform our calculations. Tables 4.5 and 4.6 display the resulting probabilities for each of the distribution. Since the *oversmoothing* problem occurs mainly in the tail of a distribution, we emphasize our analysis in this region.

We observe on each distribution that the $\mathbf{Pr[\tilde{S}_n^0(t) < S(t)]}$ is systematically lower than the $\mathbf{Pr[S_n(t) < S(t)]}$ for all values of n. A similar observation can be made in the midrange of the distributions. For instance, with the Exponential distribution ($n = 10$), 21.7% of the samples had their estimate of $\tilde{S}_n^0(t)$ below $S(t)$ at $t = 2.3036$. In comparison, this number is 23.5% for $S_n(t)$. Moreover, this condition seems to accentuate as n increases. With $n = 30$, the probability $\mathbf{Pr[\tilde{S}_n^0(t) < S(t)]}$ is 23.6% versus 40.3% for $S_n(t)$. This behavior is common to all the distributions.

Another point to make is the probability that $\tilde{S}_n^0(t)$ is lower than $S_n(t)$. For all the distribution, the probability of $\tilde{S}_n^0(t)$ being below $S_n(t)$ is considerably low; it is less than 9% throughout.

The biases revealed in the previous sections 4.1.1 and 4.1.2 is the average of the differences between $\tilde{S}_n^0(t)$ and $S_n(t)$. These large positive biases indicate that the smooth estimator is most likely above the empirical survival function. This is confirm by the low probability that $\tilde{S}_n^0(t)$ is lower than $S_n(t)$.

The empirical survival function attains zero value beyond the largest order statistic; however, the smooth estimator assigns some mass in this region so that it will continue to have above zero values. Knowing that the mass of $\tilde{S}_n^0(t)$ remains sufficiently high beyond the largest order statistic of the sample combined with all the above observations allows us to claim that the smooth estimator does not *oversmooth* in the tail, and this is true regardless of the underlying distribution.

Oversmoothing Assessment

n	t	$P\left[\tilde{S}_n^0(t) < S(t)\right]$	$P\left[S_n(t) < S(t)\right]$	$P\left[\tilde{S}_n^0(t) < S_n(t)\right]$
		Exponential(1)		
10	0.1054	0.180	0.289	0.391
10	0.6931	0.174	0.377	0.198
10	**2.3026**	**0.217**	**0.235**	**0.071**
15	0.1054	0.210	0.465	0.435
15	0.6931	0.194	0.500	0.173
15	**2.3026**	**0.218**	**0.552**	**0.067**
20	0.1054	0.267	0.346	0.419
20	0.6931	0.207	0.411	0.161
20	**2.3026**	**0.211**	**0.390**	**0.070**
30	0.1054	0.316	0.370	0.409
30	0.6931	0.239	0.431	0.161
30	**2.3026**	**0.236**	**0.403**	**0.070**
		Lognormal(0,1)		
10	0.2776	0.682	0.275	0.632
10	1.0000	0.186	0.398	0.155
10	**3.6022**	**0.278**	**0.342**	**0.041**
15	0.2776	0.768	0.463	0.667
15	1.0000	0.152	0.506	0.143
15	**3.6022**	**0.236**	**0.527**	**0.045**
20	0.2776	0.822	0.309	0.709
20	1.0000	0.139	0.436	0.117
20	**3.6022**	**0.191**	**0.364**	**0.044**
30	0.2776	0.878	0.363	0.759
30	1.0000	0.130	0.439	0.088
30	**3.6022**	**0.192**	**0.369**	**0.050**

Table 4.5: Comparison between $\tilde{S}_n^0(t)$ and $S_n(t)$ in term of *oversmoothing*.

Oversmoothing Assessment

n	t	$P\left[\tilde{S}_n^0(t) < S(t)\right]$	$P\left[S_n(t) < S(t)\right]$	$P\left[\tilde{S}_n^0(t) < S_n(t)\right]$
		Gamma(1,4)		
10	0.0263	0.198	0.269	0.431
10	0.1733	0.197	0.385	0.195
10	**0.5756**	**0.241**	**0.340**	**0.064**
15	0.0263	0.228	0.440	0.457
15	0.1733	0.201	0.504	0.173
15	**0.5756**	**0.225**	**0.537**	**0.080**
20	0.0263	0.259	0.307	0.465
20	0.1733	0.217	0.419	0.182
20	**0.5756**	**0.218**	**0.383**	**0.074**
30	0.0263	0.305	0.345	0.438
30	0.1733	0.234	0.422	0.153
30	**0.5756**	**0.213**	**0.385**	**0.084**
		Weibull(1,4)		
10	0.4214	0.179	0.271	0.419
10	2.7726	0.171	0.376	0.210
10	**9.2103**	**0.244**	**0.333**	**0.067**
15	0.4214	0.218	0.455	0.456
15	2.7726	0.189	0.489	0.169
15	**9.2103**	**0.219**	**0.508**	**0.083**
20	0.4214	0.246	0.323	0.427
20	2.7726	0.203	0.410	0.156
20	**9.2103**	**0.214**	**0.356**	**0.087**
30	0.4214	0.283	0.362	0.410
30	2.7726	0.220	0.402	0.160
30	**9.2103**	**0.224**	**0.380**	**0.079**

Table 4.6: Comparison between $\tilde{S}_n^0(t)$ and $S_n(t)$ in term of *oversmoothing*.

4.3 Confidence Intervals

The shapes of the simulation distributions approximate the shapes of the sampling distributions. So we can use the simulation distribution in place of the sampling distribution to calculate the confidence intervals. We consider four different types of confidence intervals; Normal, Binomial, Basic and Percentile. The Normal and Binomial intervals were applied to the Monte Carlo simulation, and the Basic and Percentile methods were used in Bootstrap simulation.

4.3.1 Approximate Confidence Intervals

4.3.1.1 Monte Carlo: Normal Confidence Intervals

Let X be the number of $t_i's > t$. It follows that X is distributed as a Binomial random variable

$$X \sim Binom \text{ with } p = S(t). \tag{4.9}$$

The corresponding expectation and variance are given by

$$E[X] = np \qquad \sigma^2[X] = np(1-p), \tag{4.10}$$

subsequently

$$E[S_n(t)] = S(t) \qquad \sigma^2[S_n(t)] = \frac{1}{n}S(t)[1-S(t)]. \tag{4.11}$$

Thus $S_n(t)$ is unbiased and it's variance is of order $O\left(\frac{1}{n}\right)$. Furthermore it can be shown that $S_n(t)$ is a (weakly) consistent estimator of $S(t)$ at each t as

$$\sqrt{n}\,[S_n(t) - S(t)] \xrightarrow{D} N(0, S(t)[1-S(t)]). \tag{4.12}$$

In addition if the difference between $S_n(t)$ and $S(t)$ is considered not only for a fixed t but simultaneously for all t, namely

$$D_n = \sup_t |S_n(t) - S(t)| \xrightarrow{P} 0 \text{ as } n \to \infty, \tag{4.13}$$

a stronger consistency property can be obtained.

Using the above results and referring to Chaubey and Sen (1996), we can establish, for a specified value of t, an approximate 100(1-α)% confidence interval for $\tilde{S}_n^0(t)$ as

$$\left[\tilde{S}_n^0(t) - z_{(1-\alpha/2)} se_{\tilde{S}_n^0(\cdot)}, \tilde{S}_n^0(t) + z_{(1-\alpha/2)} se_{\tilde{S}_n^0(\cdot)}\right], \tag{4.14}$$

where $z_{(1-\alpha/2)}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $se_{\tilde{S}_n^0(\cdot)} = \sqrt{se_{\tilde{S}_n^0(t)}\left[1 - se_{\tilde{S}_n^0(t)}\right]/n}$ is an estimate of the standard error of $\tilde{S}_n^0(t)$.

This is known as the *Standard Normal confidence intervals.*

4.3.1.2 Monte Carlo: Binomial Confidence Intervals

A 100(1-α)% confidence interval for $\tilde{S}_n^0(t)$ based on Binomial sampling can be defined as

$$lcl = 1 - \left[1 + \frac{(1+\tilde{S}_n^0(t))F_{(1-\alpha/2;2n-2n(1-\tilde{S}_n^0(t))+2,2n(1-\tilde{S}_n^0(t)))}}{n(1-\tilde{S}_n^0(t))}^{-1}\right], \tag{4.15}$$

$$ucl = 1 - \left[1 + \frac{n\tilde{S}_n^0(t)}{F_{(1-\alpha/2;2n(1-\tilde{S}_n^0(t))+2,2n-2n(1-\tilde{S}_n^0(t)))}}^{-1}\right]. \tag{4.16}$$

where $F_{(1-\alpha/2;\nu_1,\nu_2)}$ is the 100(1-α/2) quantile of the F distribution with (ν_1, ν_2) degrees of freedom.

The above formulas are called the *exact Binomial-Based confidence interval.* This confidence interval is conservative in the sense that the actual coverage probability is at least equal to $1-\alpha$.

4.3.1.3 Bootstrap: Basic Confidence Intervals

Suppose that T estimates a scalar θ and that we want an interval with left and right tail errors both equal to α. For simplicity we assume that T is

continuous. If the quantiles of $T-\theta$ are denoted by a_p, then

$$Pr(T-\theta \leq a_\alpha) = \alpha = Pr(T-\theta \geq a_{1-\alpha}). \tag{4.17}$$

Rewriting the events $T-\theta \leq a_\alpha$ and $T-\theta \geq a_{1-\alpha}$ as $\theta \geq T-a_\alpha$ and $\theta \leq T-a_{1-\alpha}$ respectively, we see that the $1-2\alpha$ equi-tailed interval has limits

$$\hat{\theta}_\alpha = t - a_{1-\alpha},\ \hat{\theta}_{1-\alpha} = t - a_\alpha. \tag{4.18}$$

However, this ideal solution rarely applies because the distribution of $T-\alpha$ is usually unknown. This leads us to consider approximate methods, most of which are based on approximating the quantiles of $T-\theta$.

Starting at the general confidence interval Eq.(4.18), we can estimate the quantiles a_α and $a_{1-\alpha}$ by the corresponding quantiles of T^*-t. Using the Monte Carlo estimates of quantiles for $T-\theta$, an equi-tailed $(1-2\alpha)$ confidence interval will have limits

$$t - (t^*_{((R+1)(1-\alpha))} - t),\ t - (t^*_{((R+1)\alpha)} - t). \tag{4.19}$$

This is based on the probability implication

$$Pr(a < T-\alpha < b) = 1-2\alpha \ \Rightarrow\ Pr(T-b \leq \theta \leq T-a) = 1-2\alpha. \tag{4.20}$$

The above leads to the *Basic confidence intervals* for θ

$$\hat{\theta}_\alpha = 2t - t^*_{((R+1)(1-\alpha))},\ \hat{\theta}_{1-\alpha} = 2t - t^*_{((R+1)\alpha)}. \tag{4.21}$$

To apply Eq.(4.21) exactly, it is necessary that $(R+1)\alpha$ is an integer. A simple method that works well for approximately normal estimators is linear interpolation on the normal quantile scale. In this case, we define

$$t^*_{((R+1)\alpha)} = t^*_{(k)} + \frac{\Phi^{-1}(\alpha) - \Phi^{-1}\left(\frac{k}{R+1}\right)}{\Phi^{-1}\left(\frac{k+1}{R+1}\right) - \Phi^{-1}\left(\frac{k}{R+1}\right)}(t^*_{(k+1)} - t^*_{(k)}),\ k = [(R+1)\alpha]. \tag{4.22}$$

Clearly such interpolations fail if $k = 0$, R or $R+1$.

4.3.1.4 Bootstrap: Percentile Confidence Intervals

Suppose now that there is some unknown transformation of T, say $U = h(T)$, which has a symmetric distribution. Imagine that we knew h and calculated a $1-2\alpha$ confidence interval for $\phi = h(\theta)$ by applying the basic Bootstrap method Eq.(4.21), except that we first use the symmetry to write $a_\alpha = -a_{1-\alpha}$ in the basic equation Eq.(4.18) as it applies to $U = h(T)$. This would mean that in applying Eq.(4.18) we would take $u - u^*_{((R+1)(1-\alpha))}$ instead of $u^*_{((R+1)\alpha)} - u$ and $u - u^*_{((R+1)\alpha)}$ instead of $u^*_{((R+1)(1-\alpha))} - u$, to estimate the α and $1-\alpha$ quantiles of U. This swap would change the confidence interval limits Eq.(4.21) to

$$u^*_{((R+1)\alpha)},\ u^*_{((R+1)(1-\alpha))}, \tag{4.23}$$

whose transformation back to the θ scale is

$$t^*_{((R+1)\alpha)},\ t^*_{((R+1)(1-\alpha))} \tag{4.24}$$

Remarkably this 1-2α interval for θ does not involve h at all, and so can be computed without knowing h. This interval Eq.(4.24) is known as the *Bootstrap Percentile interval.*

4.3.2 Numerical Study of the Confidence Intervals

We reuse the results of the Monte Carlo and Bootstrap simulations to perform the computation of the confidence intervals. The comparison is made for coverage, average interval width, and interval symmetry. Coverage is measured as the percentage of intervals (out of 1000) that include the true value of $S(t)$. Depending on the distribution, different values of t are used to satisfy $S(t = \{0.1, 0.5, 0.9\})$. The procedure for constructing 95% confidence intervals should produce actual coverage close to 95%.

Average interval width is considered as our main measurement since we are interested in the smallest interval that achieves nominal coverage. Symmetry

is a desirable, but not crucial, property for confidence intervals. We already noticed that our sampling distributions have asymmetric shapes. Therefore, we should expect this to influence our results.

We define symmetry as the degree to which the confidence interval tends to miss the true value of $S(t)$ equally from the right and from the left. Ideally, a 95% confidence interval would not include the true value of $S(t)$ due to underestimation 2.5% (%MR) of the time and, likewise, due to overestimation 2.5% (%ML) of the time. Here, overestimation is taken to indicate that the confidence interval lies entirely below (*i.e.* includes only values less than) the true value of the parameter.

Tables 4.7 to 4.16 display the coverage performance results, and figures 4.17 to 4.28 display the distribution of the interval widths using box plots of the 1,000 confidence intervals of the simulation samples.

We should mention that our interest here is again on the tail region of the survival curves. Therefore, the analysis will be focused on the mid and higher values of t.

At $S(t) = 0.5$, it appears that the Binomial intervals display the largest width values among the intervals and in all distributions with only one exception. When $n = 10$, the Normal interval widths exceed the Binomial intervals. Both Binomial and Normal intervals result in high coverage values, and very low probabilities (<2.5%) in their %ML and %MR . This should not be surprising for the Binomial intervals since it tends to be more conservative in term of coverage probability. The Lognormal distribution for larger sample sizes seems to be the only distribution that display coverages around the nominal value of 95% with the Binomial interval.

The Percentile and Basic intervals display very poor coverage probabilities. The %ML are way too high, and the %MR are too low on all distributions. Also, we notice that these intervals have the smallest widths and share the

same magnitude. The percentage of missed of these intervals can be due to the combination of narrow interval widths and the positive biases of the smooth estimator.

The plots reveal that all distributions have a certain quantity of outliers. which can affect their percentages of missed. They illustrate also that the median of the all intervals widths stay near the middle of the width distributions. This is an indication that the distributions are reasonably symmetric.

At $S(t) = 0.1$, the coverage percentages seem to be around the nominal 95% for the Binomial and Normal intervals on all distributions, but the Lognormal samples have lower coverage percentages for the same intervals. In contrast to $S(t) = 0.5$, the %ML for the Binomial intervals are too high. The %ML is about twice its nominal value while the %MR is always zero. The possible reason for this is the high positive bias values seen in the tail probably which makes the %ML higher than it should.

Once again, the Percentile and Basic intervals continue to display very low coverages and high %ML. Unlike $S(t) = 0.5$, the %MR are very high at $S(t) = 0.1$. The Lognormal distribution shows considerably high %ML and %MR on all intervals. This is probably because of the thickness of the Lognormal tail which is fatter than the other distributions.

The box plots display clearly that the median on all the width distributions and for all samples seem to increase as n increases. The effect of lifting the tail up of the smooth estimator appears to create skewed distributions of $\tilde{S}_n^0(t)$ in the tails.

In summary, we observe that the Percentile and Basic methods tend to have smaller confidence intervals that produce very low coverage probabilities. The Percentile intervals are widely used due to its simplicity but it is not very efficient in non-parametric problems. For the Percentile method to work well, it would be necessary that T is being unbiased on the transformed scale

so that the swap of quantile estimates is correct. As mentioned in Davison (1997), this swap does not usually happen. Also, the method carries the defect of the Basic Bootstrap method. This defect can be observed as the shape of the distribution of T changes as the sampling distribution change from F to $\hat{F}$ even after transformation. In particular, the implied symmetrizing transformation often will not be quite the same as the variance-stabilizing transformation. This may be the cause of the poor performance of the Percentile method.

The Normal and Binomial intervals, on the other hand, seems to produce somewhat better confidence intervals. Although too conservative, both intervals improve asymptotically and produce good intervals in the tail region for the Exponential, Weibull and Gamma distributions.

The Binomial confidence intervals seem to be a *good* choice for approximate confidence intervals for $\tilde{S}_n^0(t)$ at this point. It may give wider limits, but its percentage of misses are relatively low which can be significantly important in practice.

4.4 Conclusion

In this thesis, the appropriateness of the Chaubey and Sen untruncated version smooth estimator has been investigated numerically. The Monte Carlo and Bootstrap simulations reveal similar results in term of biases and standard errors. This establishes a ground for using Bootstrap in practical studies to analyze and investigate some of the statistical properties of the smooth estimator $\tilde{S}_n^0(t)$. It also has been observed that the smooth estimator was suitable for many kinds of population data. Whichever the underlying distribution either Exponential, Gamma, Weibull or Lognormal, each having very distinct shape from the others, the biases and standard errors are in the same magnitude.

We have searched a generalized model for the optimum c which remarkably reduces the calculation effort and time to find suitable estimates.

The *oversmoothing* problem seems to be well handled by the smoothing parameter c. It has been seen that under optimum c, the probability of $\tilde{S}_n^0(t)$ being less than $S(t)$ is quite low. Therefore, it is reasonable to state that there should not be a concern for the fear of *oversmoothing* with the untruncated smooth estimator.

For future research, investigations shall be conducted to establish better confidence intervals with better coverage properties. It is clear from our results that the usual methods do not perform well. The asymmetric shape of the sampling distributions greatly influenced the outcome, particularly in the Bootstrap confidence intervals which resulted in very poor coverages.

One of the major problems of nonparametric Bootstrap is the discreteness of the Bootstrap distribution. For instance, for the sample mean, there are only $\binom{2n-1}{n-1}$ possible values that $\bar{x}^*$ can take on. Fortunately as $n \to \infty$ this number increases quite rapidly so that the sampling distribution becomes like a continuous distribution. However, in some cases, even letting n be large will

not cause a distribution that works well. Therefore, we suggest that other types of intervals such as Bootstrap studentized, studentized (log), BC_a, etc should be included in a simulation study to compare their performance.

Another research area could be initiated to use different set of weights patterns in the estimation of $\tilde{S}_n^0(t)$. For instance, in Chaubey and Sen (2002), Gamma weights were used in the context of the Hille's theorem (1948). It may be possible to explore other types of weight functions.

Exponential(1) Confidence Intervals Coverages					
S(t = 0.1053506) = 0.90, Theoretical coverage = 0.95					
Sample sizes	**Method**	**Actual coverage**	**Average width**	**%ML**	**%MR**
10	Normal theory	99.8	0.2486	0.2	0.0
	Exact binomial	100.0	0.4142	0.0	0.0
	Bootstrap percentile	80.2	0.1152	19.1	0.7
	Basic	66.9	0.1047	32.8	0.3
15	Normal theory	99.7	0.2249	0.3	0.0
	Exact binomial	100.0	0.3349	0.0	0.0
	Bootstrap percentile	85.9	0.1072	13.5	0.6
	Basic	76.7	0.1029	22.7	0.6
20	Normal theory	99.6	0.2101	0.4	0.0
	Exact binomial	100.0	0.2881	0.0	0.0
	Bootstrap percentile	88.1	0.1016	11.0	0.9
	Basic	79.5	0.0998	20.2	0.3
30	Normal theory	99.9	0.1919	0.1	0.0
	Exact binomial	100.0	0.2336	0.0	0.0
	Bootstrap percentile	91.1	0.0969	8.1	0.8
	Basic	83.1	0.0964	16.2	0.7

Table 4.7: Monte Carlo and Bootstrap confidence intervals summary statistics for the Exponential(1) random samples whit $S(t_j) = 0.9$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Exponential(1) Confidence Intervals Coverages					
S($t = 0.6931472$) = 0.50, Theoretical coverage = 0.95					
Sample sizes	**Method**	**Actual coverage**	**Average width**	**%ML**	**%MR**
10	Normal theory	97.8	0.6222	2.0	0.2
	Exact binomial	100.0	0.6054	0.0	0.0
	Bootstrap percentile	79.4	0.3812	18.4	2.2
	Basic	64.5	0.3753	34.6	0.9
15	Normal theory	97.4	0.5100	2.2	0.4
	Exact binomial	99.5	0.5107	0.3	0.2
	Bootstrap percentile	82.7	0.3229	16.0	1.3
	Basic	71.4	0.3224	27.6	1.0
20	Normal theory	97.7	0.4409	2.3	0.0
	Exact binomial	99.7	0.4481	0.3	0.0
	Bootstrap percentile	85.0	0.2974	13.9	1.1
	Basic	76.3	0.2974	23.4	0.3
30	Normal theory	96.9	0.3592	3.0	0.1
	Exact binomial	99.1	0.3693	0.9	0.0
	Bootstrap percentile	89.4	0.2632	9.9	0.7
	Basic	82.0	0.2632	17.7	0.3

Table 4.8: Monte Carlo and Bootstrap confidence intervals summary statistics for the Exponential(1) random samples whit $S(t_j) = 0.5$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Exponential(1) Confidence Intervals Coverages

S(t = 2.3025851) = 0.10, Theoretical coverage = 0.95

Sample sizes	Method	Actual coverage	Average width	%ML	%MR
10	Normal theory	95.8	0.4179	1.5	2.7
	Exact binomial	95.4	0.5035	4.6	0.0
	Bootstrap percentile	81.5	0.2710	10.3	8.2
	Basic	66.1	0.2703	25.7	8.2
15	Normal theory	96.5	0.3407	1.3	2.2
	Exact binomial	96.0	0.4014	4.0	0.0
	Bootstrap percentile	84.7	0.2354	8.0	7.3
	Basic	73.8	0.2346	18.7	7.5
20	Normal theory	96.1	0.2969	1.7	2.2
	Exact binomial	95.5	0.3411	4.5	0.0
	Bootstrap percentile	88.9	0.2237	5.9	5.2
	Basic	79.6	0.2224	14.3	6.1
30	Normal theory	96.9	0.2412	1.7	1.4
	Exact binomial	95.0	0.2691	5.0	0.0
	Bootstrap percentile	87.8	0.1890	7.8	4.4
	Basic	85.1	0.1886	9.7	5.2

Table 4.9: Monte Carlo and Bootstrap confidence intervals summary statistics for the Exponential(1) random samples whit $S(t_j) = 0.1$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Gamma(1,4) Confidence Intervals Coverages					
S(t = 0.02634013) = 0.90, Theoretical coverage = 0.95					
Sample sizes	**Method**	**Actual coverage**	**Average width**	**%ML**	**%MR**
10	Normal theory	99.9	0.2527	0.1	0.0
	Exact binomial	100.0	0.4165	0.0	0.0
	Bootstrap percentile	82.2	0.1175	17.2	0.6
	Basic	68.8	0.1074	30.5	0.7
15	Normal theory	99.9	0.2255	0.1	0.0
	Exact binomial	100.0	0.3353	0.0	0.0
	Bootstrap percentile	83.8	0.1050	15.6	0.6
	Basic	75.1	0.1016	24.5	0.4
20	Normal theory	99.8	0.2091	0.2	0.0
	Exact binomial	100.0	0.2874	0.0	0.0
	Bootstrap percentile	86.8	0.1009	12.5	0.7
	Basic	78.3	0.0987	21.3	0.4
30	Normal theory	99.7	0.1912	0.3	0.0
	Exact binomial	100.0	0.2332	0.0	0.0
	Bootstrap percentile	90.8	0.0963	8.4	0.8
	Basic	85.7	0.0960	14.1	0.2

Table 4.10: Monte Carlo and Bootstrap confidence intervals summary statistics for the Gamma(1,4) random samples whit $S(t_j) = 0.9$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Gamma(1,4) Confidence Intervals Coverages					
S(t = 0.17328680) = 0.50, Theoretical coverage = 0.95					
Sample sizes	**Method**	**Actual coverage**	**Average width**	**%ML**	**%MR**
10	Normal theory	98.1	0.6243	1.7	0.2
	Exact binomial	99.8	0.6066	0.0	0.2
	Bootstrap percentile	81.6	0.3831	16.2	2.2
	Basic	66.1	0.3784	32.7	1.2
15	Normal theory	97.9	0.5100	1.8	0.3
	Exact binomial	99.8	0.5107	0.2	0.0
	Bootstrap percentile	81.9	0.3179	16.1	2.0
	Basic	71.3	0.3178	27.5	1.2
20	Normal theory	97.1	0.4407	2.9	0.0
	Exact binomial	99.3	0.4479	0.7	0.0
	Bootstrap percentile	85.1	0.2952	14.3	0.6
	Basic	74.5	0.2952	25.4	0.1
30	Normal theory	96.6	0.3594	3.3	0.1
	Exact binomial	98.7	0.3695	1.3	0.0
	Bootstrap percentile	90.4	0.2627	8.9	0.7
	Basic	83.7	0.2627	15.9	0.4

Table 4.11: Monte Carlo and Bootstrap confidence intervals summary statistics for the Gamma(1,4) random samples whit $S(t_j) = 0.5$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Gamma(1,4) Confidence Intervals Coverages

$S(t = 0.57564627) = 0.10$, Theoretical coverage $= 0.95$

Sample sizes	Method	Actual coverage	Average width	%ML	%MR
10	Normal theory	95.6	0.4090	1.4	3.0
	Exact binomial	95.3	0.4988	4.7	0.0
	Bootstrap percentile	80.5	0.2678	8.5	11.0
	Basic	65.8	0.2669	23.8	10.4
15	Normal theory	95.2	0.3405	1.5	3.3
	Exact binomial	95.1	0.4013	4.9	0.0
	Bootstrap percentile	83.7	0.2355	8.2	8.1
	Basic	72.4	0.2347	19.4	8.2
20	Normal theory	96.8	0.2983	1.5	1.7
	Exact binomial	94.5	0.3421	5.5	0.0
	Bootstrap percentile	87.6	0.2246	6.5	5.9
	Basic	79.1	0.2233	14.5	6.4
30	Normal theory	86.5	0.2433	1.7	1.8
	Exact binomial	94.5	0.2705	5.5	0.0
	Bootstrap percentile	89.5	0.1923	7.1	3.4
	Basic	87.0	0.1921	8.8	4.2

Table 4.12: Monte Carlo and Bootstrap confidence intervals summary statistics for the Gamma(1,4) random samples whit $S(t_j) = 0.1$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Weibull(1,4) Confidence Intervals Coverages

$\mathbf{S}(t = 0.4214421) = 0.90$, Theoretical coverage $= 0.95$

Sample sizes	Method	Actual coverage	Average width	%ML	%MR
10	Normal theory	99.9	0.2475	0.1	0.0
	Exact binomial	100.0	0.4136	0.0	0.0
	Bootstrap percentile	91.9	0.1150	17.8	0.3
	Basic	65.2	0.1047	34.4	0.4
15	Normal theory	99.5	0.2237	0.5	0.0
	Exact binomial	100.0	0.3342	0.0	0.0
	Bootstrap percentile	96.3	0.1053	13.2	0.5
	Basic	74.8	0.1014	24.8	0.4
20	Normal theory	99.8	0.2085	0.2	0.0
	Exact binomial	100.0	0.2871	0.0	0.0
	Bootstrap percentile	88.1	0.1011	11.2	0.7
	Basic	77.9	0.0991	21.7	0.4
30	Normal theory	99.3	0.1901	0.7	0.0
	Exact binomial	100.0	0.2325	0.0	0.0
	Bootstrap percentile	92.6	0.0966	7.0	0.4
	Basic	85.0	0.0962	15.0	0.0

Table 4.13: Monte Carlo and Bootstrap confidence intervals summary statistics for the Weibull(1,4) random samples whit $S(t_j) = 0.9$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Weibull(1,4) Confidence Intervals Coverages					
S(t = 2.7725887) = 0.50, Theoretical coverage = 0.95					
Sample sizes	**Method**	**Actual coverage**	**Average width**	**%ML**	**%MR**
10	Normal theory	97.8	0.6224	2.0	0.2
	Exact binomial	99.9	0.6056	0.0	0.1
	Bootstrap percentile	79.7	0.3794	18.9	1.4
	Basic	63.4	0.3739	35.3	1.3
15	Normal theory	97.6	0.5095	2.3	0.1
	Exact binomial	99.7	0.5103	0.3	0.0
	Bootstrap percentile	82.9	0.3197	15.7	1.4
	Basic	72.0	0.3194	27.4	0.6
20	Normal theory	97.6	0.4406	2.4	0.0
	Exact binomial	99.4	0.4479	0.6	0.0
	Bootstrap percentile	84.8	0.2953	14.0	1.2
	Basic	75.3	0.2953	24.0	0.7
30	Normal theory	97.1	0.3593	2.8	0.1
	Exact binomial	98.6	0.3695	1.4	0.0
	Bootstrap percentile	91.0	0.2626	8.3	0.7
	Basic	83.5	0.2626	16.2	0.3

Table 4.14: Monte Carlo and Bootstrap confidence intervals summary statistics for the Weibull(1,4) random samples whit $S(t_j) = 0.5$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Weibull(1,4) Confidence Intervals Coverages

S(t = 9.2103404) = 0.10, Theoretical coverage = 0.95

Sample sizes	Method	Actual coverage	Average width	%ML	%MR
10	Normal theory	96.4	0.4159	1.5	2.1
	Exact binomial	94.4	0.5022	5.6	0.0
	Bootstrap percentile	81.2	0.2711	9.7	9.1
	Basic	64.8	0.2702	26.4	8.8
15	Normal theory	96.0	0.3444	1.5	2.5
	Exact binomial	95.0	0.4035	5.0	0.0
	Bootstrap percentile	83.8	0.2377	8.9	7.3
	Basic	71.7	0.2369	20.9	7.4
20	Normal theory	96.9	0.2983	1.2	1.9
	Exact binomial	94.5	0.3421	5.5	0.0
	Bootstrap percentile	88.3	0.2247	6.7	5.0
	Basic	79.7	0.2234	14.6	5.7
30	Normal theory	96.9	0.2420	1.4	1.7
	Exact binomial	95.4	0.2697	4.6	0.0
	Bootstrap percentile	88.7	0.1904	6.9	4.4
	Basic	86.5	0.1900	8.7	4.8

Table 4.15: Monte Carlo and Bootstrap confidence intervals summary statistics for the Weibull(1,4) random samples whit $S(t_j) = 0.1$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Lognormal(0,1) Confidence Intervals Coverages					
S(t = 0.2776062) = 0.90, Theoretical coverage = 0.95					
Sample sizes	**Method**	**Actual coverage**	**Average width**	**%ML**	**%MR**
10	Normal theory	100.0	0.3391	0.0	0.0
	Exact binomial	100.0	0.4651	0.0	0.0
	Bootstrap percentile	88.8	0.1665	1.1	10.1
	Basic	77.6	0.1509	9.8	12.6
15	Normal theory	100.0	0.3029	0.0	0.0
	Exact binomial	99.9	0.3812	0.0	0.1
	Bootstrap percentile	87.5	0.1442	0.9	11.6
	Basic	78.7	0.1378	6.2	15.1
20	Normal theory	100.0	0.2815	0.0	0.0
	Exact binomial	99.9	0.3308	0.0	0.1
	Bootstrap percentile	82.5	0.1324	0.6	16.9
	Basic	80.9	0.1289	3.4	15.7
30	Normal theory	100.0	0.2442	0.0	0.0
	Exact binomial	99.8	0.2690	0.0	0.2
	Bootstrap percentile	79.9	0.1207	0.1	20.0
	Basic	81.5	0.1198	1.2	17.3

Table 4.16: Monte Carlo and Bootstrap confidence intervals summary statistics for the Lognormal(0,1) random samples whit $S(t_j) = 0.9$. The Normal and Binomial C.I. seem to overestimate the coverage percentages while the other two seems to produce underestimated results.

Lognormal(0,1) Confidence Intervals Coverages					
S(t = 1.0000000) = 0.50, Theoretical coverage = 0.95					
Sample sizes	**Method**	**Actual coverage**	**Average width**	**%ML**	**%MR**
10	Normal theory	90.6	0.6024	9.2	0.2
	Exact binomial	98.3	0.5958	1.7	0.0
	Bootstrap percentile	76.9	0.3723	21.0	2.1
	Basic	52.8	0.3477	46.0	1.2
15	Normal theory	90.4	0.5004	9.3	0.3
	Exact binomial	96.7	0.5034	3.2	0.1
	Bootstrap percentile	76.9	0.3092	21.5	1.6
	Basic	53.8	0.3015	45.3	0.9
20	Normal theory	90.5	0.4349	9.4	0.1
	Exact binomial	95.9	0.4430	4.1	0.0
	Bootstrap percentile	79.4	0.2877	20.0	0.6
	Basic	55.7	0.2842	43.9	0.4
30	Normal theory	90.3	0.3557	9.7	0.0
	Exact binomial	95.0	0.3662	5.0	0.0
	Bootstrap percentile	81.1	0.2506	18.4	0.5
	Basic	60.6	0.2493	39.3	0.1

Table 4.17: Monte Carlo and Bootstrap confidence intervals summary statistics for the Lognormal(0,1) random samples whit $S(t_j) = 0.5$. The Normal, Percentile and Basic C.I. seem to underestimate the coverage percentages while the Binomial seems to produce slightly overestimated results.

Lognormal(0,1) Confidence Intervals Coverages					
S(t = 3.6022245) = 0.10, Theoretical coverage = 0.95					
Sample sizes	**Method**	**Actual coverage**	**Average width**	**%ML**	**%MR**
10	Normal theory	87.7	0.4162	7.2	5.1
	Exact binomial	86.1	0.5016	13.9	0.0
	Bootstrap percentile	80.2	0.2953	5.3	14.5
	Basic	52.9	0.2898	32.0	15.1
15	Normal theory	88.2	0.3521	7.8	4.0
	Exact binomial	86.0	0.4090	14.0	0.0
	Bootstrap percentile	85.0	0.2668	5.7	9.3
	Basic	60.3	0.2651	29.6	10.1
20	Normal theory	89.2	0.3101	7.7	3.1
	Exact binomial	86.3	0.3508	13.7	0.0
	Bootstrap percentile	87.6	0.2555	6.4	6.0
	Basic	68.8	0.2532	24.5	6.7
30	Normal theory	90.0	0.2528	8.0	2.0
	Exact binomial	86.5	0.2786	13.5	0.0
	Bootstrap percentile	89.0	0.2152	7.6	3.4
	Basic	75.5	0.2145	20.6	3.9

Table 4.18: Monte Carlo and Bootstrap confidence intervals summary statistics for the Lognormal(0,1) random samples whit $S(t_j) = 0.1$. All confidence intervals seem to underestimate the coverage percentages.

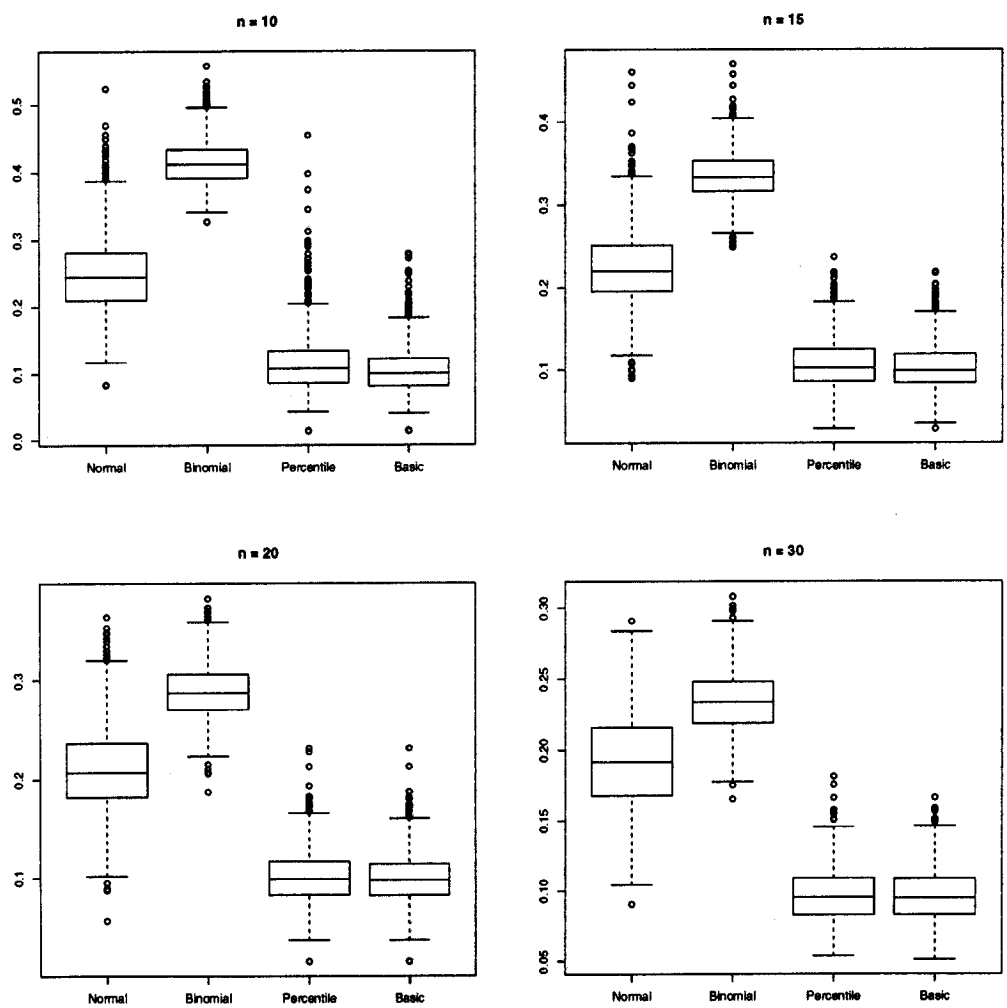

Figure 4.17: Box plots of confidence interval widths of the Exponential(1) samples with $t = 0.1054$ and $S(t) = 0.9$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

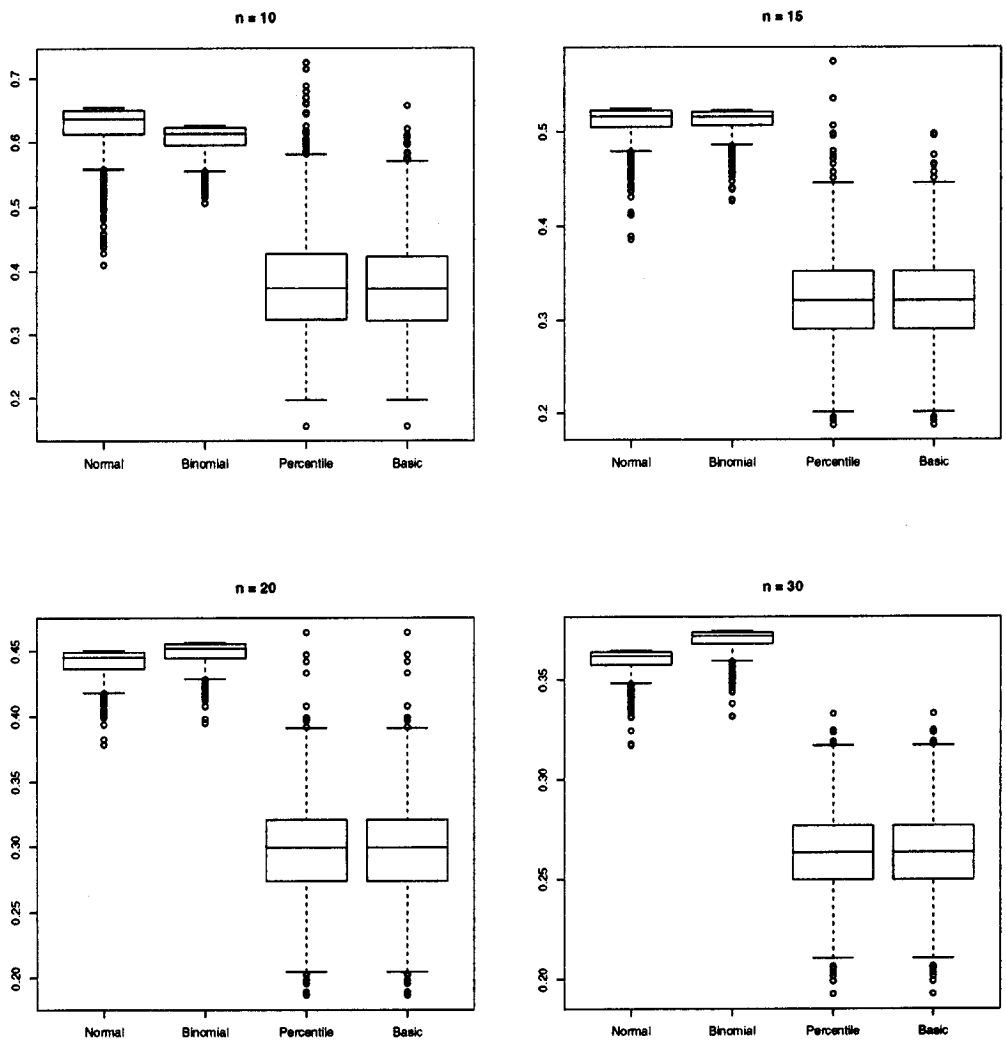

Figure 4.18: Box plots of confidence interval widths of the Exponential(1) samples with $t = 0.6931$ and $S(t) = 0.5$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

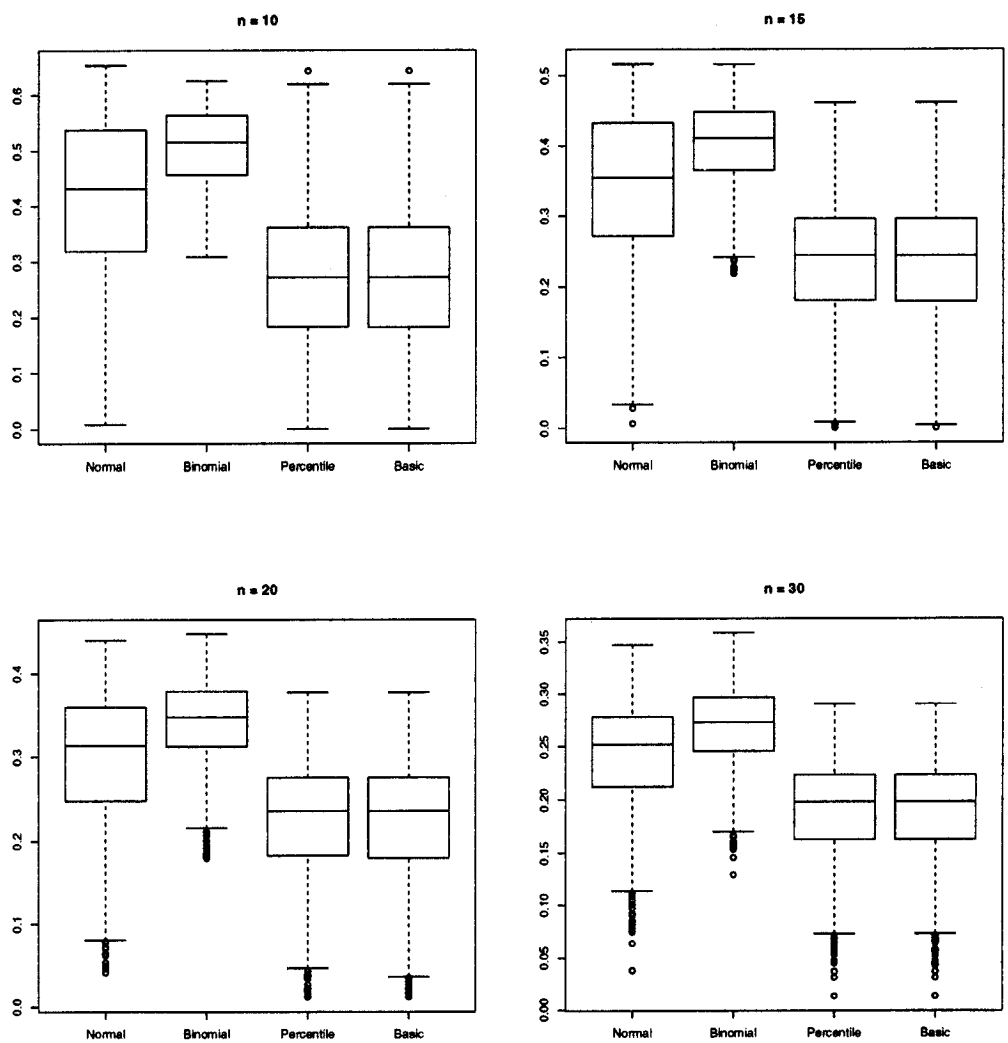

Figure 4.19: Box plots of confidence interval widths of the Exponential(1) samples with $t = 2.3026$ and $S(t) = 0.1$ Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

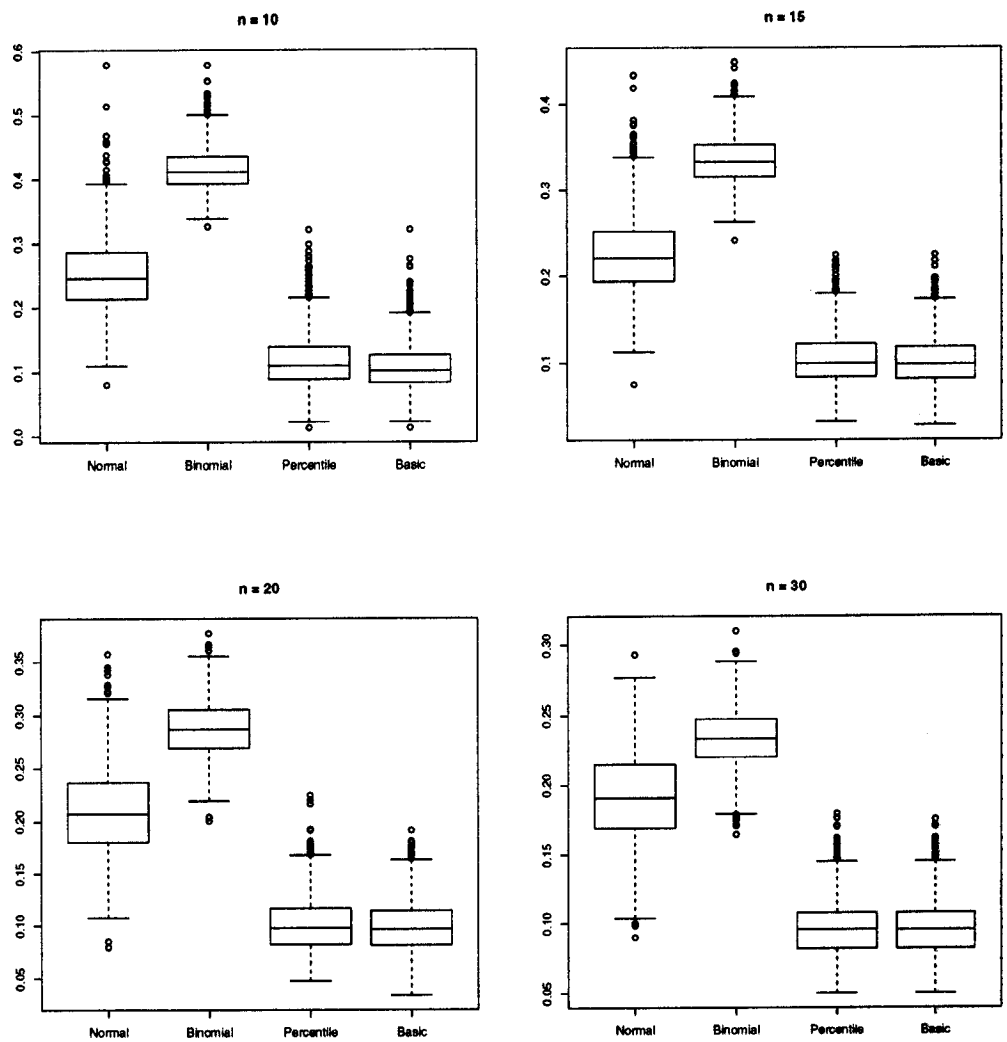

Figure 4.20: Box plots of confidence interval widths of the Gamma(1,4) samples with $t = 0.0263$ and $S(t) = 0.9$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

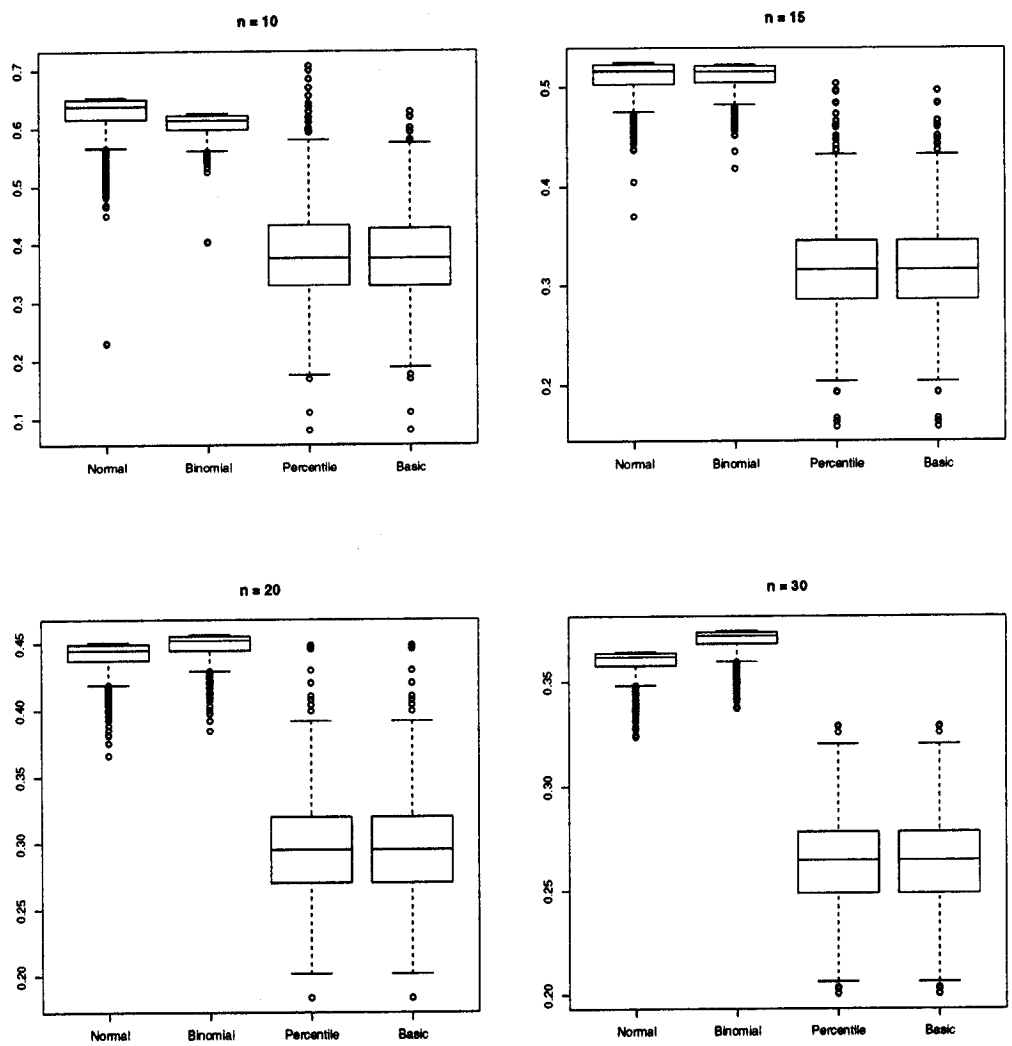

Figure 4.21: Box plots of confidence interval widths of the Gamma(1,4) samples with $t = 0.1733$ and $S(t) = 0.5$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

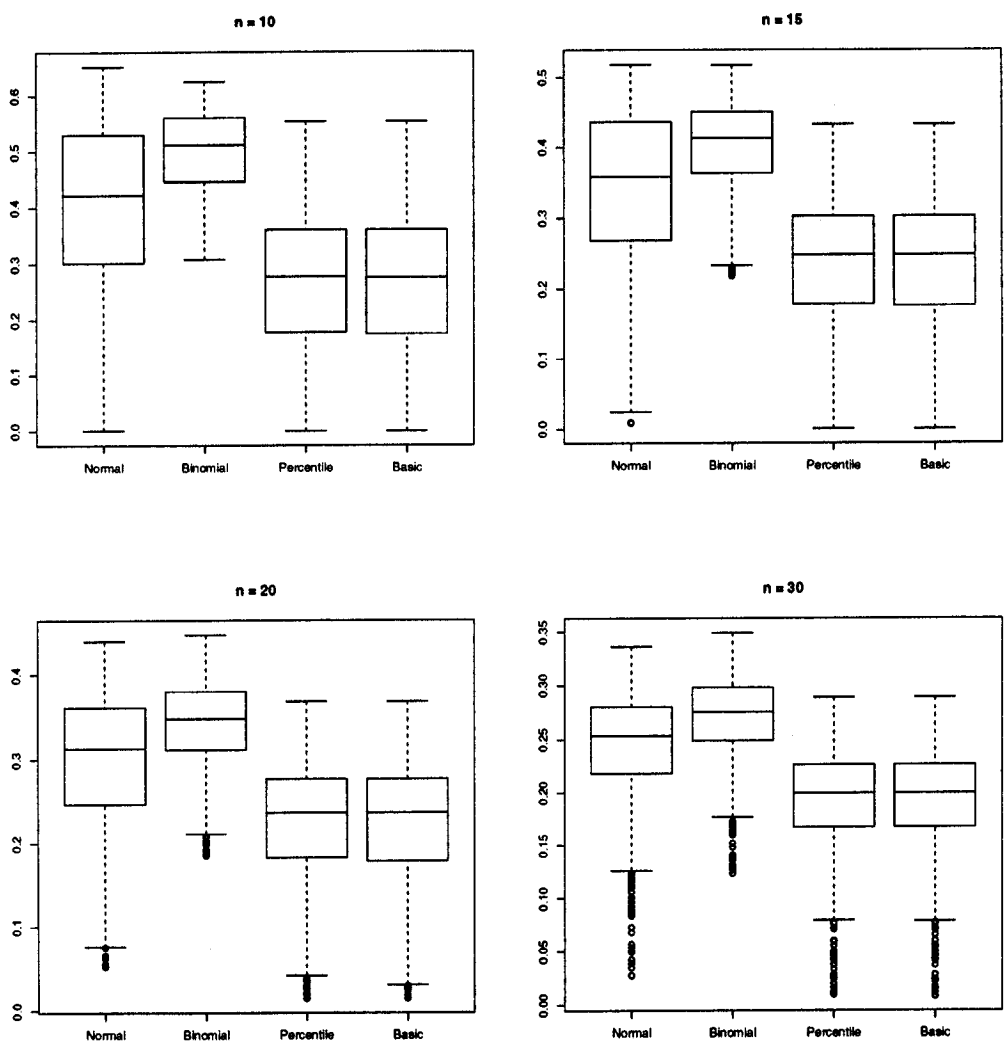

Figure 4.22: Box plots of confidence interval widths of the Gamma(1,4) samples with $t = 0.5756$ and $S(t) = 0.1$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

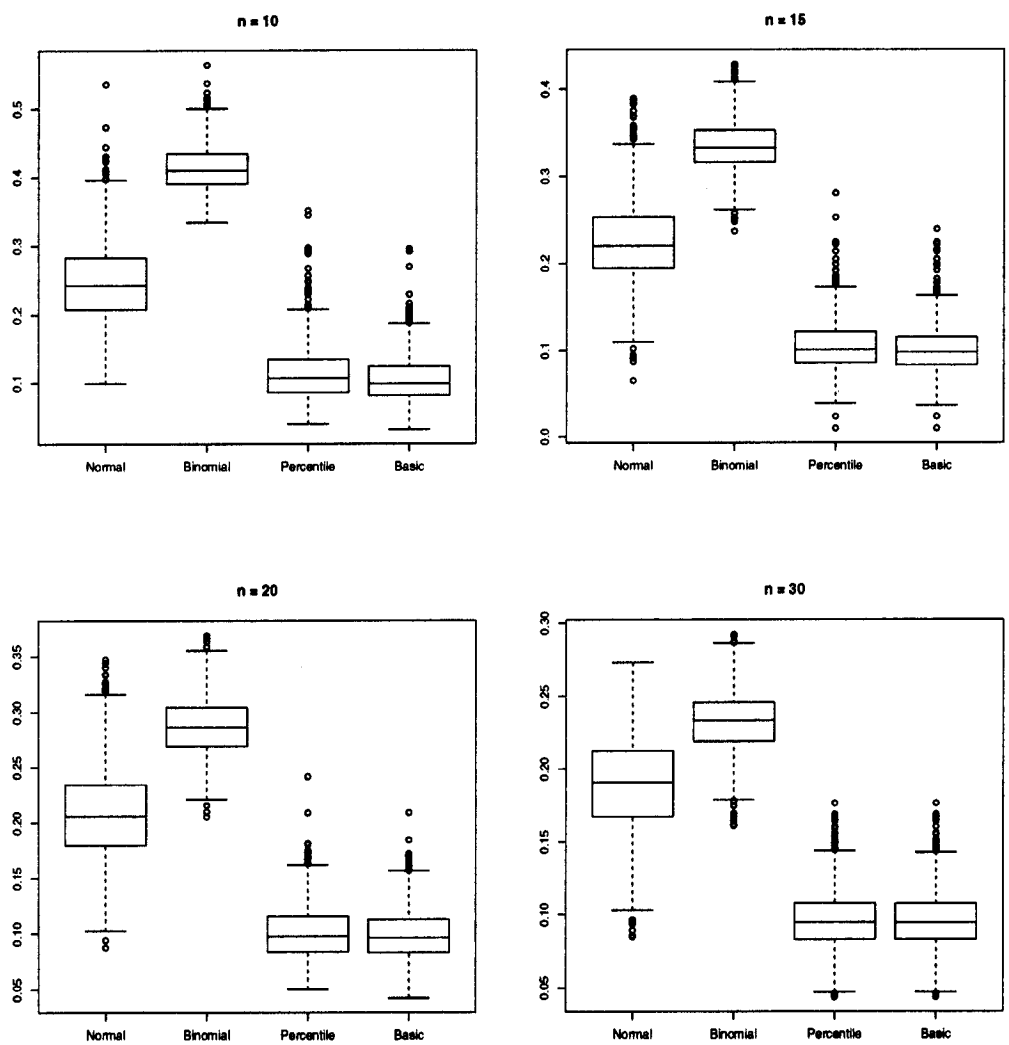

Figure 4.23: Box plots of confidence interval widths of the Weibull(1,4) samples with $t = 0.4214$ and $S(t) = 0.9$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

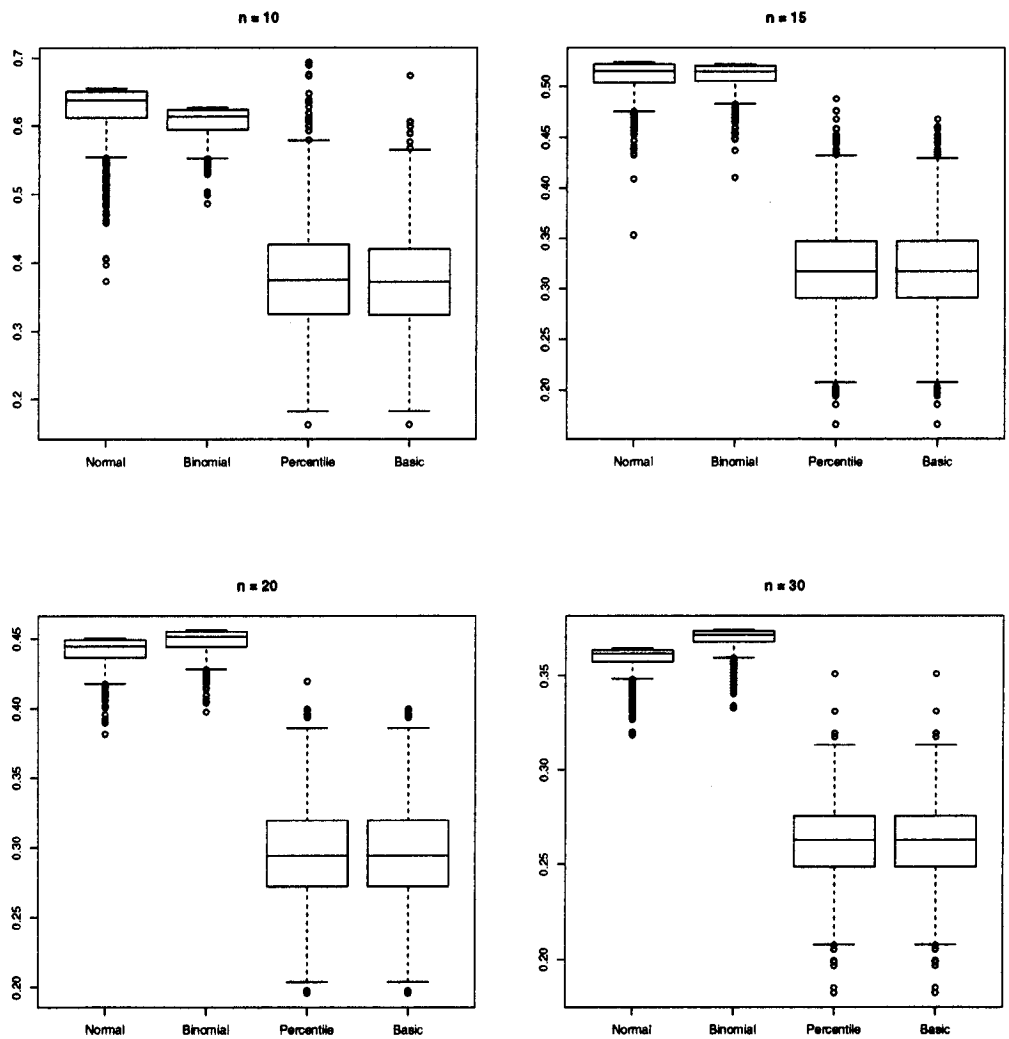

Figure 4.24: Box plots of confidence interval widths of the Weibull(1,4) samples with $t = 2.7725$ and $S(t) = 0.5$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

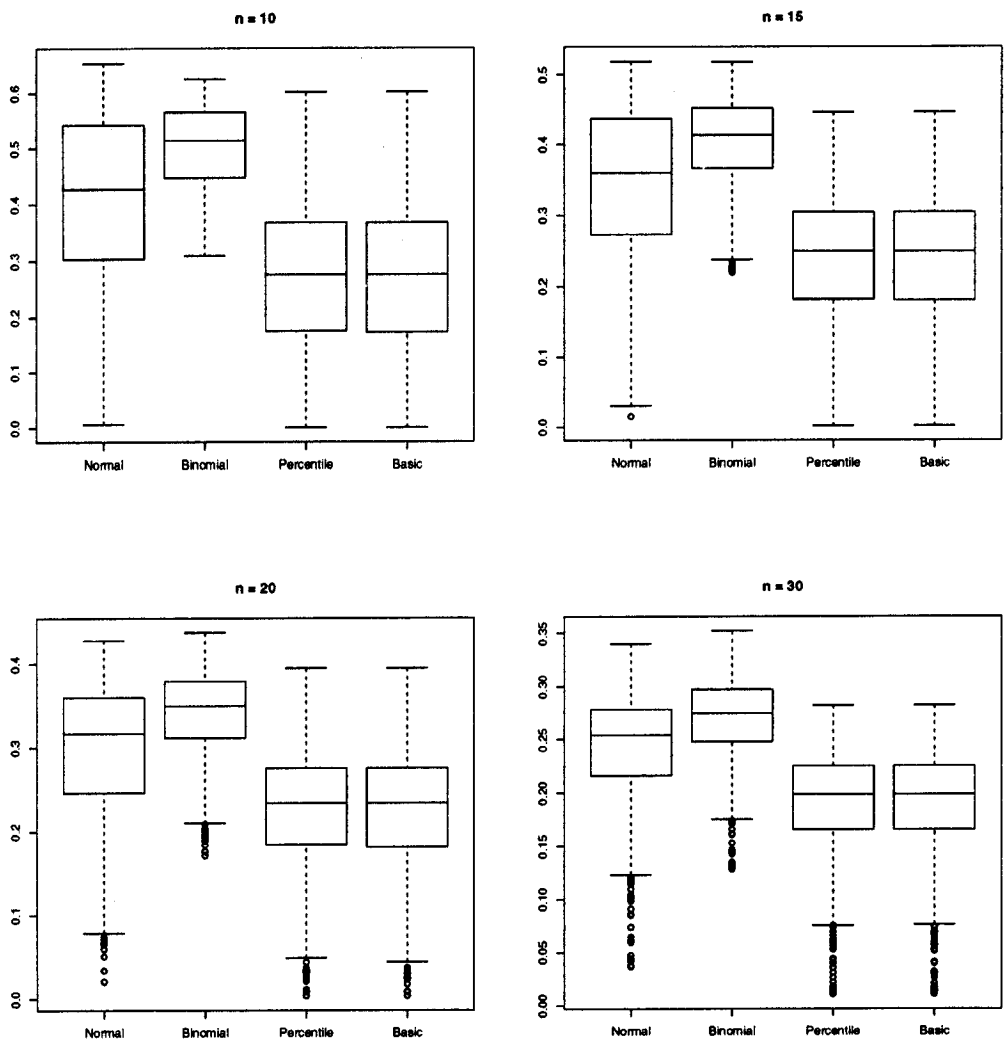

Figure 4.25: Box plots of confidence interval widths of the Weibull(1,4) samples with $t = 9.2103$ and $S(t) = 0.1$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

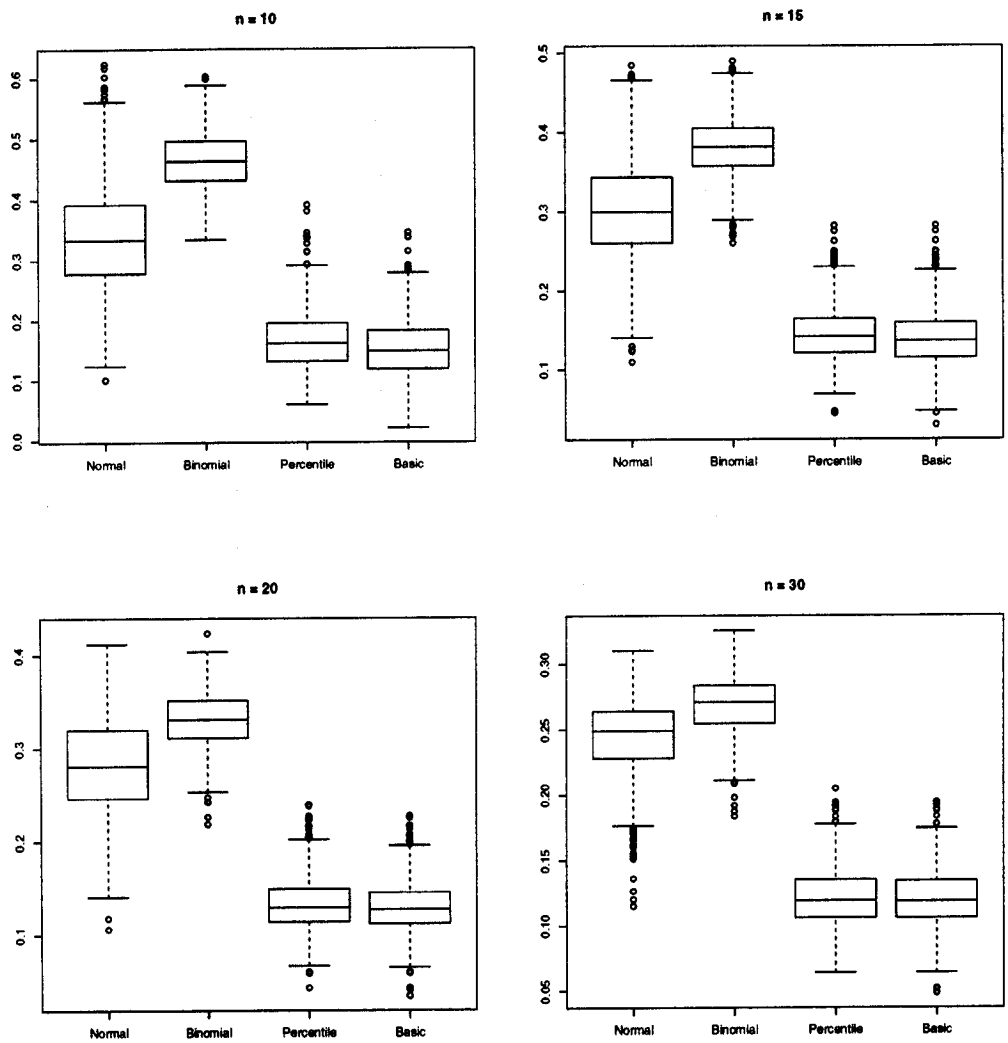

Figure 4.26: Box plots of confidence interval widths of the Lognormal(0,1) samples with $t = 0.2776$ and $S(t) = 0.9$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

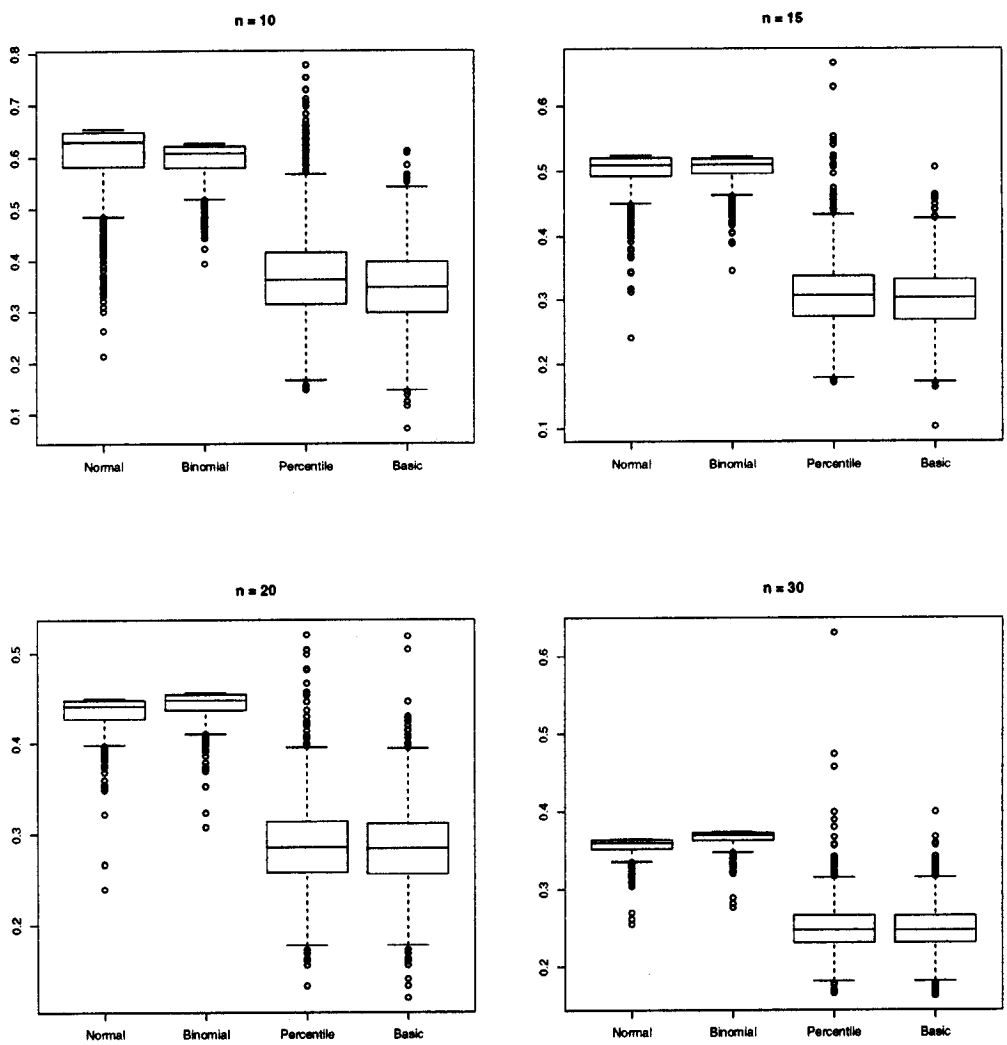

Figure 4.27: Box plots of confidence interval widths of the Lognormal(0,1) samples with $t = 1.000$ and $S(t) = 0.5$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

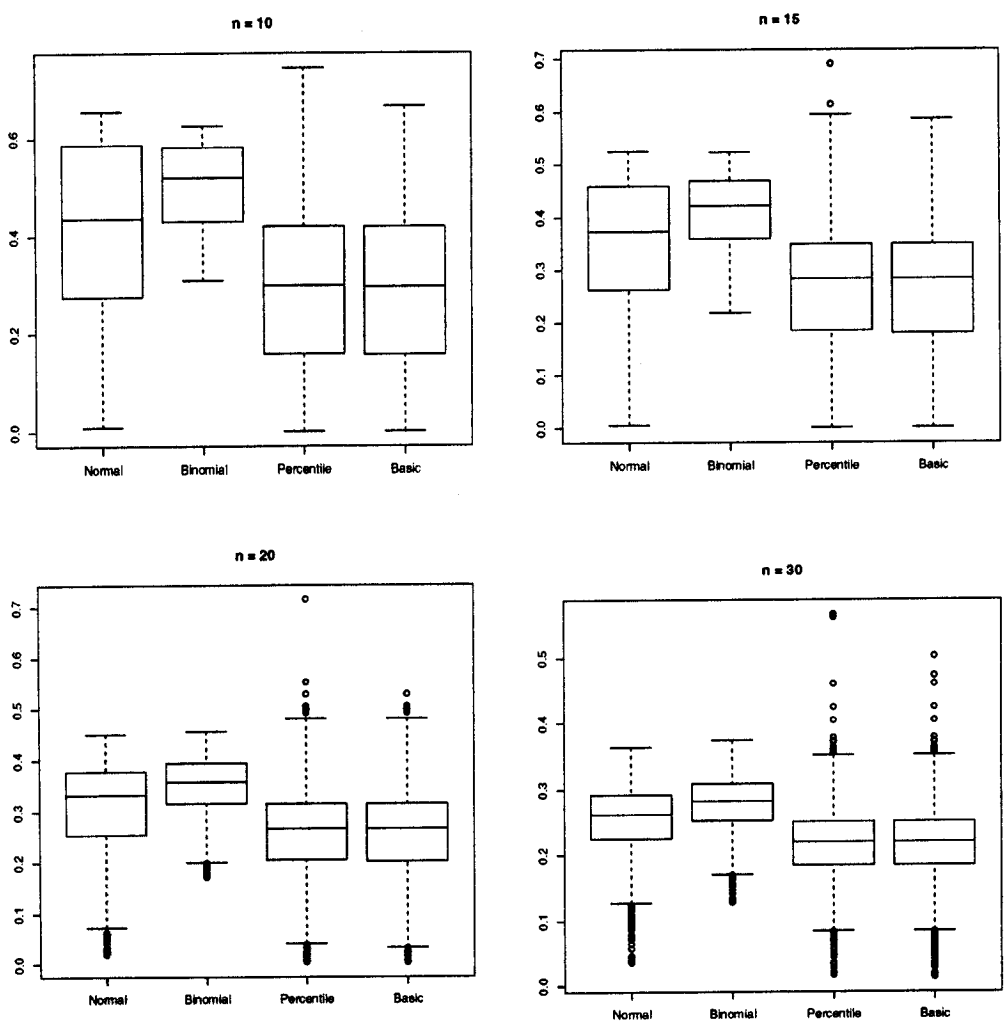

Figure 4.28: Box plots of confidence interval widths of the Lognormal(0,1) samples with $t = 3.6022$ and $S(t) = 0.1$. Normal and Binomial are generated from the Monte Carlo simulations. The Percentile and Basic are generated from the Bootstrap simulations.

Bibliography

[1] Azzalini, A. (1981), "A note on the estimation of a distribution function and quantiles", *Biometrika*, **68**, 326-328.

[2] Bahadur, R.R. (1966), "A Note on Quantiles in Large Sample", *The annals of Mathematical Statistics*, **37**, 577-580.

[3] Berretoni, J.N. (1964), "Practical Applications of the Weibull Distribution", *Industrial Quality Control*, **21**, 71-79.

[4] Bulkland, W.R. (1964), *Statistical Assessment of Life Characteristic - A Bibliographic Guide*, Hafner, New York.

[5] Chambers, J.M. and Hastie, T.J. (1993), *Statistical Models in S*, Chapman and Hall.

[6] Chaubey, Y.P. and Sen, P.K. (1996), "On smooth estimation of survival and density functions", *Statist. Decisions*, **14**, 1-22.

[7] Chaubey, Y.P. and Sen, P.K. (1999), "On smooth estimation of Mean Residual Life", *J. Statis. Plann. Inf*, **75**, 223-236.

[8] Chaubey, Y.P. and Sen, P.K. (1999), "Another look at the kernel density estimator for nonnegative support", *Fourth Biennial international Conference on Statistics, Probability and Related Topics Northern Illinois University, DeKalb, IL, June 14-16, 2002.*

[9] Chaubey, Y.P. and Sen, P.K. (2000), "Tail-behaviour of Survival Functions and Their Smooth Estimators", *Perspectives in statistical sci-*

ences, eds. A.K. Basu et al. Oxford University Press, New Delhi, pp. 87-101

[10] Chencov, N.N. (1962), "Estimation of unknown probability density based on observations", *Dokl. Akad Nauk SSSR*, **147**, 45-48.

[11] Davis, D.J. (1952), "An Analysis of Some Failure Data", *J. Am. Stat. Assoc.*, **47**, 113-150.

[12] Davison, A.C. and Hinkley, D.V. (1997), *Bootstrap Methods and their Application*, Cambridge University Press.

[13] Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall.

[14] Efron, B. and Tibshirani, R.J. (1986), *Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy, Statistical Science* **1**, 54-77.

[15] Efron, B. (1979), "Bootstrap methods: another look at the jackknife", *Annals of Statistics*, **7**, 1-26.

[16] Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for industrial and applied mathematics, Philadelphia

[17] Epstein, B. (1958), "The esponential distribution and its role in life-testing", *Ind. Qual. Control*, **15**, 2-7.

[18] Epstein, B. and Sobel, M. (1953), "Life Testing", *J. Am. Stat. Assoc.*, **48**, 486-502.

[19] Feigl, P. and Zelen, M. (1965), "Estimation of exponential survival probability with concomitant information", *Biometrika*, **21**, 826-838.

[20] Feller, W. (1965), *An Introduction to probability Theory and its Applications, Vol. II*, John Wiley and Sons, New York.

[21] Glivenko, V.I. (1934), *Course in Probability Theory*, Moscow.

[22] Ghosh, J.K. (1971), "A New Proof of the Bahadur Representation of Quantiles and an Applications", *The annals of Mathematical Statistics*, **42**, 1957-1961.

[23] Gupta, S.S. and Groll, P.A. (1961), "Gamma distribution in acceptance sampling based on life tests", *J. Am. Stat. Assoc.*, **56**, 942-970.

[24] Härdle, W. (1991), *Smoothing Techniques with implementation in S*, Springer-Verlag, New York.

[25] Hille, E. and Phillips, R.S. (1957), *Functional Analysis and Semi-groups*, Amer. Math. Soc.

[26] Kececioglu, D. (1993), *Reliability and Life Testing Handbook*, Prentice Hall PTR.

[27] Kiefer, J. (1967), "On Bahadur representation of Sample Quantiles", *The annals of Mathematical Statistics*, **38**, 1323-1342.

[28] Klein, J.P. and Moeschberger, M.L. (1999), *Survival analysis*, Springer-Verlag, New York.

[29] Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

[30] Lehmann, E.L. (1999), *Elements of Large-Sample Theory*, Springer-Verlag New York.

[31] Nelson, W. (1982), *Applied Life Data Analysis*, Wiley, New York.

[32] Nelson, W. and Hahn, G.J. (1972), "Linear Estimation of a Regression Relationship from Censored Data - Part I. Simple Methods and Their Application", *Technometrics*, **14**, 247-269.

[33] Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode", *The annals of Mathematical Statistics*, **33**, 1065-1076.

[34] Peto, M.C. (1966), "A Method of Analysis of a Certain Class of Experiments in Carcinogenesis", *Biometrics*, **22**, 142-161.

[35] Peto, R., Lee, P.N. and Paige, W.S. (1972), "Statistical analysis of the bioassay of continuous carcinogens", *Br. J. Cancer*, **26**, 258-261.

[36] Rohatgi, V.K. and Saleh, A.K.M.E. (2001), *An Introduction to Probability and Statistics*, Wiley, New York.

[37] Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function", *The Annals of Mathematical Statistics*, **27**, 832-837.

[38] Scott, D.W. (1992), *Multivariate Density Estimation*, Wiley.

[39] Sen, D. (1999), *Accelerated Life Testing: Concepts and Models*, Concordia University.

[40] Smirnov, N.V. (1951), "On the approximation of probability densities of random variables", *Scholarly Notes of Moscow State Polytechnical Institute*, **16**, 69-96.

[41] Venables, W.N. and Ripley, B.D. (1997), *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York.

[42] Weibull, W. (1951), "A Statistical Distribution Function of wide Applicability", *Journal of Applied Mechanics*, **18**, 293-297.

[43] Whittermore, A. and Altschuler, B. (1976), "Lung cancer incidence in cigarette smokers: further analysis of Doll and Hill's data for British physicians", *Biometrics*, **32**, 805-816.