1-1-2008

Statistical modeling of skewed data using newly formed parametric distributions

Kahadawala Cooray
University of Nevada, Las Vegas

STATISTICAL MODELING OF SKEWED DATA USING NEWLY FORMED PARAMETRIC DISTRIBUTIONS

by

Kahadawala Cooray

Bachelor of Science
University of Colombo, Sri Lanka
1994

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy Degree in Mathematical Sciences
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
August 2008

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Dissertation Approval

The Graduate College
University of Nevada, Las Vegas

May 12, 2008

The Dissertation prepared by

Kahadawala Cooray

Entitled

Statistical Modeling of Skewed Data Using Newly Formed Parametric Distributions

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mathematical Sciences

Malwane Ananda
Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

1017-52

ABSTRACT

Statistical Modeling of Skewed Data Using Newly Formed Parametric Distributions

by

Kahadawala Cooray

Dr. Malwane M. A. Ananda, Examination Committee Chair
Professor of Statistics
University of Nevada, Las Vegas

Several newly formed continuous parametric distributions are introduced to analyze skewed data. Firstly, a two-parameter smooth continuous lognormal-Pareto composite distribution is introduced for modeling highly positively skewed data. The new density is a lognormal density up to an unknown threshold value and a Pareto density for the remainder. The resulting density is similar in shape to the lognormal density, yet its upper tail is larger than the lognormal density and the tail behavior is quite similar to the Pareto density. Parameter estimation methods and the goodness-of-fit criterion for the new distribution are presented. A large actuarial data set is analyzed to illustrate the better fit and applicability of the new distribution over other leading distributions. Secondly, the Odd Weibull family is introduced for modeling data with a wide variety of hazard functions. This three-parameter family is derived by considering the distributions of the odds of the Weibull and inverse Weibull families. As a result, the Odd Weibull family is not only useful for testing goodness-of-fit of the Weibull and inverse Weibull as submodels, but it is also convenient for modeling

and fitting different data sets, especially in the presence of censoring and truncation. This newly formed family not only possesses all five major hazard shapes: constant, increasing, decreasing, bathtub-shaped and unimodal failure rates, but also has wide variety of density shapes. The model parameters for exact, grouped, censored and truncated data are estimated in two different ways due to the fact that the inverse transformation of the Odd Weibull family does not change its density function. Examples are provided based on survival, reliability, and environmental sciences data to illustrate the variety of density and hazard shapes by analyzing complete and incomplete data. Thirdly, the two-parameter logistic-sinh distribution is introduced for modeling highly negatively skewed data with extreme observations. The resulting family provides not only negatively skewed densities with thick tails, but also variety of monotonic density shapes. The advantages of using the proposed family are demonstrated and compared by illustrating well-known examples. Finally, the folded parametric families are introduced to model the positively skewed data with zero data values.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I wish to express my sincerest gratitude to Professor Malwane M. A. Ananda for his invaluable guidance during the preparation of this dissertation, and for his genuine interest in my professional development and career. I am also thankful for his support and his encouragement during my years in the Department of Mathematical Sciences. I wish to thank the members of my committee, Professors Chih-Hsiang Ho, Hokwon Cho, Sandra Catlin, and Chad Cross for their invaluable contributions to improve my dissertation. I want to thank Professor Dieudonné Phanord for his support and advice on teaching, especially in the first two years of my stay at University of Nevada, Las Vegas.

I am thankful to Mr. Sumith Gunasekera for his kind support, friendship and cooperation to help made my dissertation a success. I would like to thank the students and all the staff in the Mathematics Department and in the University of Nevada, Las Vegas for always being friendly and helpful. The people in the statistics division will always be remembered, as well as for their friendships.

Finally, I want to thank my family members in Sri Lanka for their continuous love, their support, and their encouragement through all these years, and for believing in me and supporting my dreams. I wish to dedicate this dissertation to my parents for their example and their immeasurable love, without it, this work would never have been done.

CHAPTER I

INTRODUCTION

1.1 The background of some continuous univariate distributions

As we know, statistics plays a vital role when making decisions under uncertainty. Once the uncertainty is formalized in terms of probability, then it can be modeled by using probability distributions. Therefore probability distributions play a critical and central role in statistics. Probability distributions which generally involve parameters, are divided into two classes, continuous and discrete. Continuous univariate parametric distributions are widely used in most applications due to their amenability to more elegant mathematical treatment. There exists well over 30 popular continuous univariate parametric distributions. These distributions can be divided into two different classes, regular and nonregular. Here nonregular means the support depends on the parameters of the distribution. Also, they can be divided into two other different classes, lifetime and non-lifetime distributions. For example, the normal distribution is regular non-lifetime distribution, whereas, the lognormal distribution is regular lifetime distribution. Also, the symmetrically truncated Cauchy distribution (Derman 1964) is a nonregular, non-lifetime distribution. Furthermore, the generalized Weibull distribution (Mudholkar *et al.* 1996) or sometimes called embedded Burr (1942) distribution given by the following distribution function is a nonregular lifetime distribution.

$$F(x, \alpha, \theta, \lambda) = 1 - (1 - \lambda(x/\theta)^{\alpha})^{1/\lambda}, \tag{1.1}$$

where $0 \leq x \leq \theta\lambda^{-1/\alpha}$, $0 < \alpha < \infty$, $0 < \theta < \infty$, and $-\infty < \lambda < \infty$.

The probability distributions can be variously specified, in terms of a cumulative distribution function $F(.)$, a density function $f(.)$, or a quantile function $Q(.)$. Some distributions do not have closed-form quantile functions, for example, inverse Gaussian and folded normal distributions. Also, most mixture distributions do not have closed-form expressions for the quantile functions. Furthermore, Tukey's (1960) lambda distribution given by the following quantile function does not have closed-form expression for both density and distribution functions. This distribution is first introduced by Hastings *et al.* (1947).

$$Q(u, \lambda) = [u^{\lambda} - (1 - u)^{\lambda}]/\lambda;\ 0 \leq u \leq 1, -\infty < \lambda < \infty. \tag{1.2}$$

Finally, any good continuous univariate probability distribution should have most of the following:

1. Closed-form expressions for density, distribution, quantile, and moment functions.
2. Least number of parameters in the distribution.
3. Rich density and hazard shapes, including tail shapes.
4. Distribution must be regular.
5. Statistical inference should be technically convenient.
6. Distribution must arise as a plausible physical phenomenon.

1.2 Motivation to construct new distributions

1.2.1 The lognormal-Pareto composite distribution

The Pareto distribution, named after an Italian-born Swiss professor of economics, Vilfredo Pareto (1897) who formulated a model to check how income or wealth was distributed among the individuals in society, is widely used in insurance and actuarial industry. This distribution models the upper portion of most insurance payment data that are commonly encountered in insurance industries. However, insurers are interested to model the entire portion of the payment data, which are frequently distributed in unimodal shape. The Pareto model is not a suitable model for such data due to its non-monotonic density shape, and therefore, the other parametric families such as loglogistic, lognormal, Weibull, inverse Weibull are considered as useful models. However, these parametric models, which possess semi-heavy tails, are inadequate for modeling the heavy tail area of the data distribution. To remedy this situation parametrically, higher order parametric families such as Burr (1942), generalized Pareto models have already been discussed in the literature. In addition, splicing and mixing of the existing distributions have also been discussed. However, since all these remedies are less convenient for modeling and fitting purposes, they have become unattractive to practitioners. The pros and cons of such methodologies are found in Klugman *et al.* (1998) and Everitt and Hand (1981). Further, related works are found in Ramlau-Hansen (1988), Embrechts *et al.* (1999), Beirlant *et al.* (1996), Resnick (1997), Beirlant *et al.* (2004), McNeil (1997), Hogg and Klugman (1984), Hossack *et al.* (1983) and Patrik (1980).

Insurance payment data in actuarial industries are typically highly positively

skewed and distributed with a larger upper tail. Therefore, researchers often tend to use the lognormal distribution or the Pareto distribution to model the data in this field (Klugman *et al.* 1998; Hogg and Klugman 1984). Furthermore, larger loss payments or reinsurance data (Hossack *et al.* 1983; Hogg and Klugman 1984; Beirlant *et al.* 1996) are often modeled by the Pareto distribution. Moreover, to model large claim data, generalized Pareto distribution has been used by several authors such as Resnick (1997) and Beirlant *et al.* (2004).

However, the Pareto distribution, due to the monotonically decreasing shape of the density, does not provide a reasonable fit for many applications when the frequency distribution of the data set is hump-shaped. In these cases, the lognormal distribution is typically used to model these data sets. Even though the lognormal model covers larger data, it fades away to zero more quickly than the Pareto model. In modeling insurance payment data, the lognormal model often fails to provide adequate coverage for higher losses, and thus underestimates payment losses, because the upper tail of the lognormal distribution is much thinner than the Pareto model. Therefore, instead of using the lognormal model for the full data set and ignoring the lower half of the data set, the large insurance payments are typically modeled by the Pareto model (McNeil 1997; Resnick 1997). In fact, the Pareto model covers the behavior of large losses well, but fails to cover the behavior of small losses. Conversely, the lognormal model covers the behavior of small losses well, but fails to cover the behavior of large losses.

The necessity of lognormal and Pareto composition was recognized by several authors through their practical knowledge of loss payment data (Klugman *et al.* 1998;

Patrik 1980). They attempted to address the problem by combining the lognormal model and the Pareto model through splicing method. After partitioning the data into several domains, different probability models were fitted for each domain. Ramlau-Hansen (1988) attempted to handle these types of actuarial data using the loggamma distribution.

Therefore, taking into account the tail behavior of both small and large losses, we were motivated to look for a new avenue to remedy the situation. In order to achieve both of these behaviors in one model, we looked for a desirable composite model, which took the two-parameter lognormal density up to an unknown threshold value and the two-parameter Pareto density for the rest of the model. Differentiability and continuity at the threshold point yield a fine smooth density function called the *lognormal-Pareto composite* (LPC) distribution (Cooray and Ananda 2005) with two unknown parameters. The resulting density given in Chapter II has a larger tail than the lognormal density, as well as a smaller tail than the Pareto density. The shape of the density is similar to the lognormal density, yet its upper tail is larger than the lognormal density, and the tail behavior is quite similar to the Pareto density.

1.2.2 The Odd Weibull distribution

The Weibull distribution, named after the Swedish physicist Waloddi Weibull (1939), having exponential and Rayleigh as submodels, is frequently used for modeling broad variety of lifetime data from reliability, survival, environmental and actuarial sciences. When modeling monotone hazard rates, the Weibull distribution may be an initial choice due to its negatively and positively skewed density shapes.

However, the Weibull distribution does not provide a reasonable parametric fit for some practical applications where the underlying hazard rates may be bathtub or unimodal shapes. In addition, the underlying distribution may be highly negatively or positively skewed with thicker tails. In order to achieve these behaviors from a single distribution, researchers have used different modifications to the Weibull distribution by introducing an additional shape parameter. Such three-parameter extensions for the entire positive real line can be seen from generalized Gamma distribution (Stacy 1962; see Glaser 1980 for all five hazard rates) and exponentiated Weibull family (Mudholkar *et al.* 1995). However, these interesting three-parameter Weibull extensions do not support the modeling of comfortable bathtub-shaped failure rate data, which is often encountered in real life data analyses. In general, middle portions of the hazard curves of such distributions are nearly flat and the corresponding densities have a positive antimode. A variety of distributions for modeling such data and their statistical analyses have appeared in literature. In particular, Rajarshi and Rajarshi (1988), and Shooman (1968) have discussed the issues of such data failure and related applications. Furthermore, to analyze such bathtub-shaped failure data, researchers often use mixture models, which generate a long flat period in the middle portion of the hazard function. However, most mixture models are of less interest to reliability analysts due to several reasons. These reasons include the lack of closed-form expressions in their quantile functions, the complicated nature in parameter estimation techniques, and the necessity of large amount of data in the estimation process. In the case of bathtub-shaped failure distributions, Haupt and Schäbe (1997) have pointed out that the main characteristics, such as moments and quantiles, are not available in

closed-forms. Even the estimations of parameters often resort to extensive iteration procedures.

In particular, distributions with one or two parameters impose strong restrictions on comfortable bathtub-shaped hazard curves, as well as the densities with positive antimodes. In general, at least three parameters are needed to form a flexible bathtub-shaped hazard function. On the other hand, more flexible distributions usually have more than three parameters, and they will become unattractive due to parameter estimation problem.

In the reliability theory, the inverse Weibull distribution has received considerable attention during the past two decades. This inverse Weibull distribution is derived by Keller and Kamath (1982) as a suitable model to describe degradation phenomena of mechanical components such as the dynamic components of diesel engines: pistons, crank shaft, and main bearings. Keller *et al.* (1985) simultaneously used Weibull and inverse Weibull to model the engine parts failure time data of commercial vehicles. Chang (1998) used a mixture of three distributions (Weibull, inverse Weibull, and Gompertz) to analyze the changes in mortality patterns. Gera (1995) proposed a Weibull competing risk model involving a two-parameter Weibull and a two-parameter inverse Weibull distribution. In environmental sciences, Simiu *et al.* (2001) discussed the importance of the inverse Weibull distribution as a reasonable model for analyzing the extreme wind speed data. The other Weibull and inverse Weibull related models, techniques, and applications are found in Murthy, Xie, and Jiang (2004) related to reliability, survival, and environmental disciplines.

A three-parameter generalization of the Weibull distribution is presented in Chap-

ter III to deal with general situations in modeling survival process with *various shapes in the hazard function.* This generalized Weibull family will be referred to as the *Odd Weibull* distribution (Cooray 2006) since it is derived by considering the distributions of the odds of the Weibull and inverse Weibull families. As a result, the Odd Weibull family is not only useful for testing goodness-of-fit of the Weibull and inverse Weibull as submodels, but it is also convenient for modeling and fitting different data sets, especially in the presence of censoring and truncation. In addition, this family accommodates not only all five major hazard shapes: constant, increasing, decreasing, bathtub-shaped and unimodal failure rates, but also has a wide variety of density shapes including the bimodality with one mode at the origin. The model parameters for exact, group, censored and truncated data are estimated in two different ways due to the fact that the inverse transformation of the Odd Weibull family does not change its density function.

1.2.3 The logistic-sinh distribution

Highly negatively skewed data with extreme observations are frequently encountered in reliability and survival analyses. Such data may be incompatible with familiar probability models, and is motivate to explore new models, which are useful to practicing statistician or those who work in the related areas. For the purpose of modeling these data, commonly available parametric families have so far been used by considering such observations as outliers, even though they are true data points. For example, highly negatively skewed distributions such as Gompertz or sinh-normal have been used by ignoring the extreme observations in the right-tail. In addition, distribu-

tions (exponentiated Weibull, generalized gamma, and Weibull), which are flexible to model both negatively and positively skewed data, would not be good choices, since they possess thinner right-tails when they negatively skewed. To remedy the situation parametrically, one can suggest Cauchy or logistic type distributions with appropriate transformations. Moreover, nonparametric and graphical procedures can be used under the poor explanation of the data distribution. As an example to the nonparametric approach, Miller (1983) and Efron (1988) discussed the inefficiencies of the well-developed Kaplan-Meier product limit estimator, which is usually worthless when estimating extreme high quantiles. In engineering sciences, poorly estimated quantiles can lead to serious consequences such as structural failure in buildings and bridges or premature failure in mechanical components. Therefore, parametric modeling is considered as a means of increasing the precision in the estimation of small tail probabilities as noted by Miller (1983). A two-parameter *logistic-sinh* distribution (Cooray 2005) is presented in Chapter IV to model highly negatively skewed data with extreme observations. The resulting family provides not only negatively skewed densities with thick tails, but also variety of monotonic density shapes. Also, the density function has a non-zero density value at the origin.

1.2.4 The folded parametric distributions

In some practical applications, measurements are recorded without their algebraic sign. As a consequence, the underlying distributions of measurements are replaced by distributions of absolute measurements, and the resulting distributions are known as folded distributions. In general, folded distributions are positively skewed and

have non-zero density value at the origin. Therefore, these distributions are useful to analyze the data sets with zero data values. The folded normal distribution and its applications have already been discussed in detail in the statistical literature. In Chapter, we look at some properties and applications of the *folded logistic*, the *folded Cauchy* (Johnson, *et al.* 1994) and the *folded Laplace* distributions.

CHAPTER II

THE LOGNORMAL-PARETO COMPOSITE DISTRIBUTION

2.1 Introduction

The actuarial and insurance industries frequently use the lognormal and the Pareto distributions to model their payment data. These types of payment data are typically very highly positively skewed. Pareto model with a longer and thicker upper tail is used to model the larger loss data values, while larger data values with lower frequencies, as well as smaller data values with higher frequencies, are usually modeled by the lognormal distribution. Even though the lognormal model covers larger data values with lower frequencies, it fades away to zero more quickly than the Pareto model. Furthermore, the Pareto model does not provide a reasonable parametric fit for smaller data values due to the monotonic decreasing shape of its density. Therefore, taking into account the tail behavior of both small and large losses, we were motivated to look for a new avenue to remedy the situation. Here, we introduce a two-parameter smooth continuous *lognormal-Pareto composite* (LPC) density that is a lognormal density up to an unknown threshold value and a Pareto density for the remainder. The resulting two-parameter smooth density is similar in shape to the lognormal density, yet its upper tail is larger than the lognormal density and the tail behavior is quite similar to the Pareto density.

Moment properties such as coefficient of variation, skewness, kurtosis, and limited

expected values of this LPC distribution are derived in Section 2.3. Limited expected values of frequency distributions are widely used in insurance layer analysis. Insurance layers occur due to different deductible and policy limits for each individual. Maximum likelihood parameter estimation technique is presented in Section 2.4 by providing the conditional coverage probabilities (Section 2.5) of those estimators for uncensored samples. Also, least squares parameter estimation method is discussed in Section 2.6, and the critical points for a quantile-quantile (Q-Q) plot correlation coefficient test is provided to assess the assumption of the LPC distribution for a given uncensored data points. Furthermore, Bayesian parameter estimation technique is presented in Section 2.7 by using the Jeffrey's (1961) prior of the LPC distribution. Also related generalized maximum likelihood estimators and their standard errors are obtained by using the joint posterior pdf of shape parameter β (> 0) and scale parameter θ (> 0). Empirical distribution function (EDF) based goodness-of-fit criterion, the Kolmogorov (1933) and Anderson-Darling (1954) test statistics, are discussed in Section 2.8. Simulation studies are carried out to obtain the upper percentage points of these statistics for the LPC distribution. Maximum likelihood parameter estimation technique for right censored data is presented in Section 2.9. Finally, a simulated example and a well-known Danish fire insurance data set with 1492 data points are analyzed and parameters are estimated in Section 2.10 by using the above three methods. Also goodness-of-fit criterion such as chi-squared test statistic and above mentioned EDF based test statistics are used to compare with other leading distributions to show the importance and applicability of this LPC distribution.

2.2 Model derivation

The lognormal-Pareto composite (LPC) density (Cooray and Ananda 2005) can be written as

$$f(x) = \begin{cases} \frac{\beta\theta^{\beta}}{(1+\Phi(k_1))x^{\beta+1}} \exp\left[-0.5\{(\beta/k_1)\ln(x/\theta)\}^2\right] & \text{if} \quad 0 < x \leq \theta \\ \frac{\beta\theta^{\beta}}{(1+\Phi(k_1))x^{\beta+1}} & \text{if} \quad \theta \leq x < \infty \end{cases}, \tag{2.1}$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution and k_1 is a known constant which is given by the positive solution of the equation $\exp(-k_1^2) = 2\pi k_1^2$. This value is $k_1 = 0.372238898$. Also θ (> 0), and β (> 0) are respectively scale and shape parameters of this distribution.

The cumulative distribution function and the quantile function of this distribution can respectively be written as

$$F(x) = \begin{cases} \frac{1}{(1+\Phi(k_1))}\Phi\left((\beta/k_1)\ln(x/\theta) + k_1\right) & \text{if} \quad 0 < x \leq \theta \\ 1 - \frac{1}{(1+\Phi(k_1))}(\theta/x)^{\beta} & \text{if} \quad \theta \leq x < \infty \end{cases}, \tag{2.2}$$

and

$$Q(u) = \begin{cases} \theta\exp\{(k_1/\beta)\left(\Phi^{-1}\left((1+\Phi(k_1))u\right) - k_1\right)\} & \text{if} \quad 0 \leq u \leq u_0 \\ \theta\left\{(1-u)(1+\Phi(k_1))\right\}^{-1/\beta} & \text{if} \quad u_0 \leq u < 1 \end{cases}, \tag{2.3}$$

where $u_0 = \Phi(k_1)/(1+\Phi(k_1))$. Furthermore, the shape of the density surface of the LPC distribution with unit scale $(\theta = 1)$ is given in Figure 2.1.

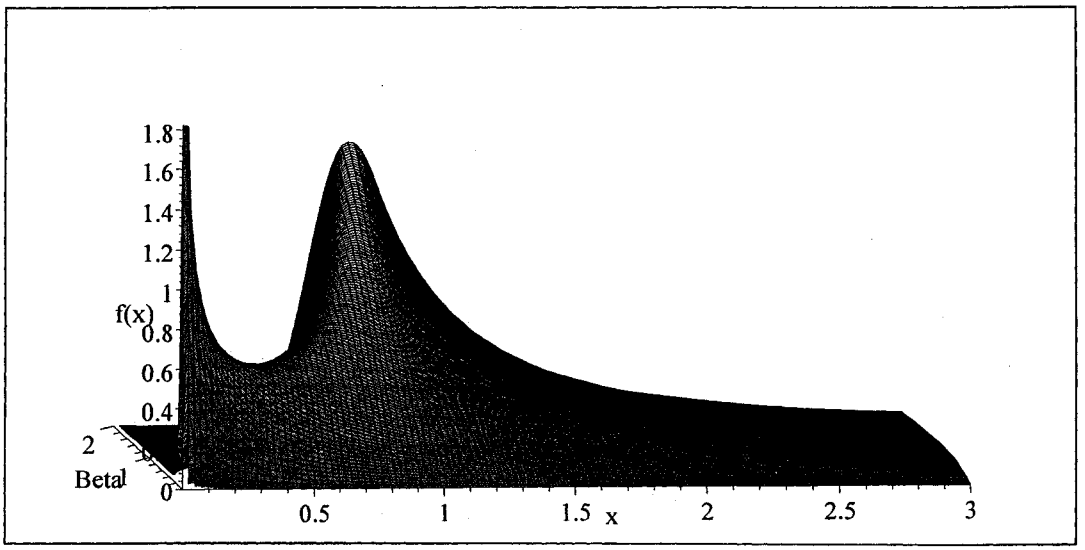

Figure 2.1 Density surface of the LPC distribution for $\theta = 1$.

2.3 Moment properties

The t^{th} raw moment, $E[X^t]$, of the LPC distribution can be obtained from the following equation for $t < \beta$.

$$E\left(X^t\right) = \frac{\theta^t}{1+\Phi\left(k_1\right)}\left\{\Phi\left(k_1 - k_1 t/\beta\right)\exp\left[\frac{1}{2}\left(\frac{k_1}{\beta}\right)^2\left(t^2 - 2\beta t\right)\right] + \frac{\beta}{(\beta - t)}\right\}. \tag{2.4}$$

The coefficient of variation ($CV = \sigma/\mu$), skewness ($\sqrt{\beta_1} = E\left[(X-\mu)^3\right]/\sigma^3$), and kurtosis ($\beta_2 = E\left[(X-\mu)^4\right]/\sigma^4$) for LPC distribution along with some common distributions are plotted in Figure 2.2 and 2.3, where $\mu = E[X]$ and $\sigma = E\left[(X-\mu)^2\right]$. The $\sqrt{\beta_1}$ versus CV graph for some common distributions can be found in Cox and Oakes (1984), and Meeker and Escobar (1998). Note that BISA and GNF stand for Birnbaum-Saunders (Birnbaum and Saunders 1969) and generalized F distributions, respectively. And also some abbreviations are given in the appendix A.

The β_2 versus β_1 graph for some common distributions is available in Pearson and Hartley (1972), and Johnson *et al.* (1994). The reason of creating a (β_1, β_2) plane may be due to having a linear relations between β_1 and β_2 of some Pearsonian family of distributions and as well as inverse Gaussian distributions. For example, the Pearson type III: $2\beta_2 - 3\beta_1 - 6 = 0$, Pearson impossible area of all frequency distributions: $\beta_2 - \beta_1 - 1 = 0$, and the inverse Gaussian: $\beta_2 = 3 + 5\beta_1/3$.

In Figure 2.3, the β_2 versus $\sqrt{\beta_1}$ graph is plotted for some common distributions, due to our interest expanded to the negative skewness regions of those distributions.

The t^{th} limited expected value, $E\left[(X \wedge x)^t\right]$ of the LPC distribution can be obtained from the following equation.

$$E\left[(X \wedge x)^t\right] = \begin{cases} \frac{\theta^t}{1+\Phi(k_1)} \left\{ \begin{array}{l} \Phi(k_1 - \frac{k_1 t}{\beta}) e^{\frac{1}{2}\left(\frac{k_1}{\beta}\right)^2 (t^2 - 2\beta t)} \\ \qquad + \frac{t(x/\theta)^{t-\beta} - \beta}{(\beta - t)} \end{array} \right\}, & \text{for } x > \theta,\ t \neq \beta \\ \frac{\theta^\beta}{1+\Phi(k_1)} \left\{ \frac{1}{2} e^{-\frac{1}{2}k_1^2} + \beta \ln\left(\frac{x}{\theta}\right) + 1 \right\}, & \text{for } x > \theta,\ t = \beta \\ \begin{array}{l} \frac{\theta^t \Phi((\beta/k_1)\ln(x/\theta) + k_1 - k_1 t/\beta)}{1+\Phi(k_1)} e^{\frac{1}{2}\left(\frac{k_1}{\beta}\right)^2 (t^2 - 2\beta t)} \\ \qquad + x^t \left(1 - \frac{\Phi(k_1 + (\beta/k_1)\ln(x/\theta))}{1+\Phi(k_1)}\right) \end{array}, & \text{for } x < \theta. \end{cases} \tag{2.5}$$

The limited expected values are involving in several actuarial quantities (Klugman *et al.* 1998), the mean excess loss, the loss elimination ratio, the expected amount paid per loss, the expected payment per payment, and etc.

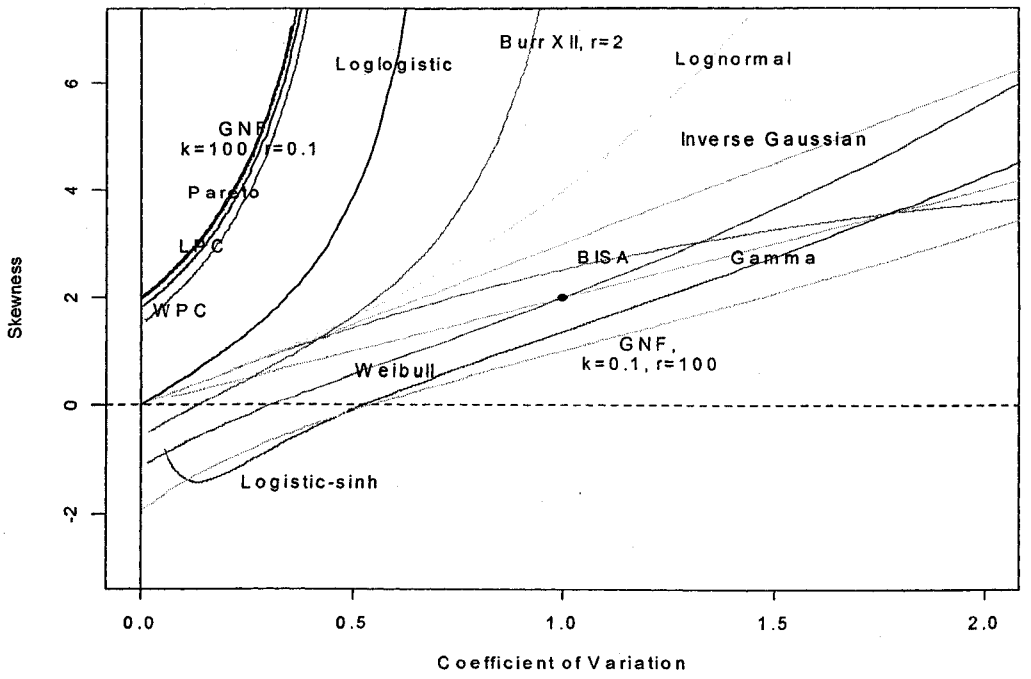

Figure 2.2 $\sqrt{\beta_1}$ versus CV graph for some common distributions.

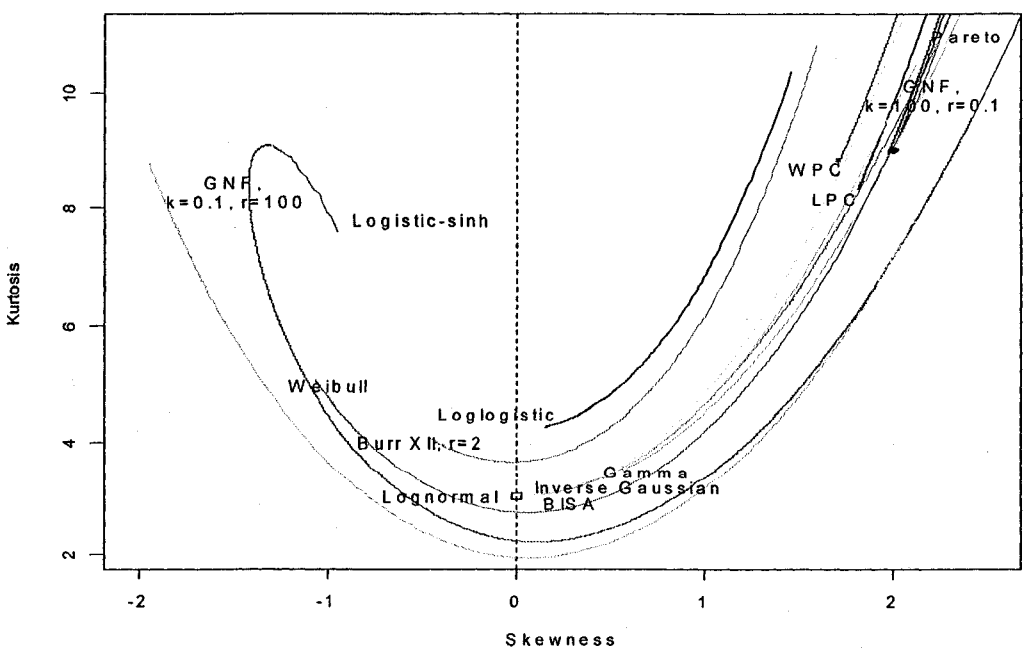

Figure 2.3 β_2 versus $\sqrt{\beta_1}$ graph for some common distributions.

2.4 Maximum likelihood estimators for complete data

Let $X_1, X_2, ..., X_n$ be a random sample from the LPC distribution given in equation (2.1). Suppose the unknown parameter θ is in between the m^{th} observation and $m+1^{\text{th}}$ observation. Therefore, it is reasonable to assume that this is an ordered sample, i.e., $x_1 \leq x_2 \leq x_3 \leq ....x_m \leq \theta \leq x_{m+1} \leq ... \leq x_n$. Then the log-likelihood function is given by

$$\begin{aligned} \ln L(\beta, \theta) &= -n \ln (1 + \Phi(k_1)) + n \ln \beta + n\beta \ln \theta - (1+\beta) \textstyle\sum_{i=1}^{n} \ln x_i \\ &\quad -0.5(\beta/k_1)^2 \textstyle\sum_{i=1}^{m} \ln^2 (x_i/\theta) . \end{aligned} \tag{2.6}$$

An algorithm to evaluate maximum likelihood (ML) estimators

Step 1: for each m ($m = 1, 2, ..., n-1$), calculate $\widehat{\beta}_m$ and $\widehat{\theta}_m$ as follows:

For $m = 1$, $\widehat{\beta}_1 = n \left(\sum_{i=1}^{n} \ln (x_i/x_1)\right)^{-1}$, $\widehat{\theta}_1 = x_1 \prod_{i=1}^{n} (x_i/x_1)^{k_1^2}$.

Otherwise

$$\widehat{\beta}_m = \left(k_1^2 B + \sqrt{k_1^4 B^2 + 4mnk_1^2 A}\right) \Big/ (2A), \tag{2.7}$$

and

$$\widehat{\theta}_m = \left(\exp\left(nk_1^2/\widehat{\beta}_m\right) \prod_i^m x_i\right)^{1/m}, \tag{2.8}$$

where $A = m \sum_{i=1}^{m} (\ln x_i)^2 - \left(\sum_{i=1}^{m} \ln x_i\right)^2$, and $B = n \sum_{i=1}^{m} \ln x_i - m \sum_{i=1}^{n} \ln x_i$.

If $\widehat{\theta}_m$ is in between $x_m \leq \widehat{\theta}_m \leq x_{m+1}$, then the ML estimators of β and θ are

$$\widehat{\beta}_{\text{ML}} = \widehat{\beta}_m, \quad \widehat{\theta}_{\text{ML}} = \widehat{\theta}_m. \tag{2.9}$$

Let us rewrite the equation (2.8) such that

$$\left(nk_1^2/\widehat{\beta}_m\right) + \sum_{i=1}^{m} \ln(x_i/\widehat{\theta}_m) = 0. \tag{2.10}$$

From equation (2.10), there must be at least one x_i value less than $\widehat{\theta}_m$, since n, k_1, and $\widehat{\alpha}_m$ are positive values. Therefore ML estimate of θ cannot occur at x_1.

Step 2: if there is no solution for θ (i.e., $x_n \leq \widehat{\theta}_m$) with the conditions given in Step 1, the ML estimate of β and θ are

$$\widehat{\beta}_{\mathrm{ML}} = nk_1/\sqrt{D}, \tag{2.11}$$

and

$$\widehat{\theta}_{\mathrm{ML}} = \exp\left(k_1^2/\widehat{\beta}_{\mathrm{ML}}\right) \prod_i^n x_i^{1/n}, \tag{2.12}$$

where $D = n\sum_{i=1}^{n} (\ln x_i)^2 - \left(\sum_{i=1}^{n} \ln x_i\right)^2$.

In addition, one can observe that these estimators are such that $\widehat{\theta}_{\mathrm{ML}} = \min_m \left(\widehat{\theta}_m\right)$ and $\widehat{\beta}_{\mathrm{ML}} = \max_m \left(\widehat{\beta}_m\right)$. However, the corresponding log-likelihood value, $l(\widehat{\boldsymbol{\theta}})$, is not equal to $\max_m \ln L(\widehat{\beta}_m, \widehat{\theta}_m)$. But, it is equal to $\ln L(\widehat{\beta}_{\mathrm{ML}}, \widehat{\theta}_{\mathrm{ML}})$.

The associated observed information matrix for estimated parameters can be written as

$$i\left(\widehat{\beta}_{\mathrm{ML}}, \widehat{\theta}_{\mathrm{ML}}\right) = \begin{bmatrix} \frac{2n}{\widehat{\beta}_{\mathrm{ML}}^2} - \frac{1}{\widehat{\beta}_{\mathrm{ML}}} \sum_{i=1}^{n} \ln(x_i/\widehat{\theta}_{\mathrm{ML}}) & \frac{n}{\widehat{\theta}_{\mathrm{ML}}} \\ \frac{n}{\widehat{\theta}_{\mathrm{ML}}} & m\left(\frac{\widehat{\beta}_{\mathrm{ML}}}{k_1 \widehat{\theta}_{\mathrm{ML}}}\right)^2 \end{bmatrix}. \tag{2.13}$$

The asymptotic standard errors of $\widehat{\beta}_{\mathrm{ML}}$ ($SE_{\widehat{\beta}_{\mathrm{ML}}}$) and $\widehat{\theta}_{\mathrm{ML}}$ ($SE_{\widehat{\theta}_{\mathrm{ML}}}$) can be calculated by inverting the observed information matrix (Efron and Hinkley 1978). However, these asymptotic standard errors are conditional on $x_m \leq \widehat{\theta}_{\mathrm{ML}} \leq x_{m+1}$ for a unique

m value. Therefore, the standard errors obtained by using the observed information matrix given in (2.13) are conditional standard errors.

2.5 Approximate conditional coverage probabilities for ML estimators

The coverage probabilities for the ML estimation method with intended confidence levels $\alpha = 0.1$ and $\alpha = 0.05$ are given in Table 2.1. These coverage probabilities are based on 10,000 simulated random samples from the density given in equation (2.1). The random samples are generated by plugging the known values of parameters β and θ (say $\beta = 0.2, \theta = 10$) to the quantile function given in equation (2.3). In addition, n (say $n = 10$), the number of ordered uniform random sample from the uniform distribution, $u \sim U(0,1)$ is required to substitute as u in equation (2.3). In that way, one random sample with size n (say $n = 10$) from the LPC distribution with parameters β and θ (say $\beta = 0.2$, $\theta = 10$) can be generated. In this simulation study, ten thousand such samples are generated to get a single cell value in Table 2.1. The approximate $100\,(1-\alpha)\,\%$ confidence intervals for parameters, β and θ are calculated by using $\left(\widehat{\beta}_{\mathrm{ML}} - Z_{\alpha/2}SE_{\widehat{\beta}_{\mathrm{ML}}},\ \widehat{\beta}_{\mathrm{ML}} + Z_{\alpha/2}SE_{\widehat{\beta}_{\mathrm{ML}}}\right)$ and $\left(\widehat{\theta}_{\mathrm{ML}} - Z_{\alpha/2}SE_{\widehat{\theta}_{\mathrm{ML}}},\ \widehat{\theta}_{\mathrm{ML}} + Z_{\alpha/2}SE_{\widehat{\theta}_{\mathrm{ML}}}\right)$ respectively.

From Table 2.1, one can clearly see that when the sample size increases, the approximate conditional coverage probabilities for the parameters under the maximum likelihood method is getting closer to the intended coverage probabilities. The values in Table 2.1 predict that the parameters are not too overly estimate under the maximum likelihood estimation method. But, one can obtain a desired confidence level by appropriately adjusting the confidence coefficient α.

Table 2.1 Approximate coverage probabilities of LPC

	$n =$	10			20			50		
90% intended	$\theta =$	10	20	50	10	20	50	10	20	50
$\beta = 0.2$	β :	.903	.908	.912	.901	.905	.908	.899	.898	.905
	θ :	.780	.793	.799	.840	.846	.838	.875	.878	.879
$\beta = 0.5$	β :	.910	.910	.908	.903	.904	.906	.900	.898	.904
	θ :	.837	.840	.840	.877	.874	.872	.890	.886	.887
$\beta = 1.0$	β :	.905	.911	.911	.906	.901	.905	.899	.903	.906
	θ :	.837	.845	.843	.872	.877	.874	.891	.895	.890
$\beta = 5.0$	β :	.906	.905	.915	.909	.909	.904	.900	.899	.898
	θ :	.848	.841	.835	.873	.874	.873	.883	.892	.890
95% intended										
$\beta = 0.2$	β :	.956	.960	.961	.953	.955	.957	.953	.949	.954
	θ :	.827	.820	.824	.869	.872	.869	.907	.907	.910
$\beta = 0.5$	β :	.961	.962	.961	.954	.954	.955	.951	.951	.954
	θ :	.878	.881	.878	.916	.917	.912	.934	.933	.935
$\beta = 1.0$	β :	.958	.960	.959	.956	.953	.955	.950	.952	.955
	θ :	.888	.894	.892	.922	.924	.919	.934	.943	.941
$\beta = 5.0$	β :	.964	.958	.961	.956	.955	.954	.952	.949	.951
	θ :	.903	.897	.892	.925	.927	.926	.936	.944	.941

2.6 Least square estimators for complete data

Let $X_1, X_2, ..., X_n$ be a random sample from the LPC distribution given in equation (2.1). Suppose the unknown parameter θ is in between the m^{th} observation and $m+1^{\text{th}}$ observation. As before, it is reasonable to assume that this is an ordered random sample, i.e., $x_1 \leq x_2 \leq x_3 \leq ....x_m \leq \theta \leq x_{m+1} \leq ... \leq x_n$. Then the least square estimators (Gujarati 2002) can be calculated from linear form in equation (2.14) which is obtained by taking the natural logarithm of the quantile function given in equation (2.3) and replacing the cumulative probability with $(i-0.5)/n$. Where n is the total number of data points and $i = 1, 2, 3, ..., n$.

$$Y_i = \begin{cases} aX_{1i} + b & \text{if } Y_i \leq b \\ aX_{2i} + b & \text{if } b \leq Y_i \end{cases}, \tag{2.14}$$

where $X_{1i} = k_1 \left[\Phi^{-1}\left\{(1+\Phi(k_1))\left(\frac{i-0.5}{n}\right)\right\} - k_1\right]$, $X_{2i} = -\ln\left\{(1+\Phi(k_1))\left(\frac{n-i+0.5}{n}\right)\right\}$, $Y_i = \ln x_i$, $a = 1/\beta$, $b = \ln\theta$, $i = 1, 2, 3, ..., n$.

An algorithm to evaluate least square (LS) estimators

Step 1: calculate $\widetilde{\beta}_m$ and $\widetilde{\theta}_m$ from the following equations for $m = 1$ $(m = 1, 2, ...)$

$$\widetilde{\beta}_m = 1/\widetilde{a}_m, \ \widetilde{\theta}_m = \exp(\widetilde{b}_m), \tag{2.15}$$

here $\widetilde{a}_m$ and $\widetilde{b}_m$ are evaluated from

$$\widetilde{a}_m = p/q, \ \widetilde{b}_m = \overline{Y} - \widetilde{a}_m\overline{X}, \tag{2.16}$$

where $p = \sum_{i=1}^{m}\left(X_{1i} - \overline{X}\right)\left(Y_i - \overline{Y}\right) + \sum_{i=m+1}^{n}\left(X_{2i} - \overline{X}\right)\left(Y_i - \overline{Y}\right)$, $\overline{Y} = \sum_{i=1}^{n} Y_i/n$, $\overline{X} = \left(\sum_{i=1}^{m} X_{1i} + \sum_{i=m+1}^{n} X_{2i}\right)/n$, and $q = \sum_{i=1}^{m}\left(X_{1i} - \overline{X}\right)^2 + \sum_{i=m+1}^{n}\left(X_{2i} - \overline{X}\right)^2$.

Also, the variance-covariance matrix of $\widetilde{a}_m$ and $\widetilde{b}_m$ can be written as

$$\begin{bmatrix} \frac{v}{q} & -\frac{v\overline{X}}{q} \\ -\frac{v\overline{X}}{q} & \frac{sv}{nq} \end{bmatrix}, \tag{2.17}$$

where $s = \sum_{i=1}^{m} X_{1i}^2 + \sum_{i=m+1}^{n} X_{2i}^2$, $v = (t - p^2/q)/(n-2)$, and $t = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$.

Step 2: If $\widetilde{\theta}_m \in [x_m,\ x_{m+1}]$, then the least square estimators of β and θ are

$$\widetilde{\beta}_{\mathrm{LS}} = \widetilde{\beta}_m,\ \widetilde{\theta}_{\mathrm{LS}} = \widetilde{\theta}_m. \tag{2.18}$$

Otherwise, repeat the step 1 for the next m value. A unique value for m such that $x_m \leq \widetilde{\theta}_{\mathrm{LS}} \leq x_{m+1}$ can be obtained from this algorithm.

The approximate conditional standard errors of $\widetilde{\beta}_{\mathrm{LS}}$ $(SE_{\widetilde{\beta}_{\mathrm{LS}}})$ and $\widetilde{\theta}_{\mathrm{LS}}$ $(SE_{\widetilde{\theta}_{\mathrm{LS}}})$ can be obtained from the following equations. These equations are derived from the variance-covariance matrix of $\widetilde{a}_m$ and $\widetilde{b}_m$ for a unique m by applying the delta method.

$$SE_{\widetilde{\beta}_{\mathrm{LS}}} = \widetilde{\beta}_{\mathrm{LS}}^2 \sqrt{\frac{v}{q}},\ SE_{\widetilde{\theta}_{\mathrm{LS}}} = \widetilde{\theta}_{\mathrm{LS}} \sqrt{\frac{sv}{nq}}. \tag{2.19}$$

Note that these standard errors are conditioned on $x_m \leq \widetilde{\theta}_{\mathrm{LS}} \leq x_{m+1}$.

Furthermore, one can easily use a quantile-quantile (Q-Q) plot to assess the assumption of LPC distribution for a given uncensored point data set. The Q-Q plot can be obtained by plotting the pairs of ordered observations $(X_{1i},\ Y_i)$ for $i = 1, ..., m$ and $(X_{2i},\ Y_i)$ for $i = m+1, ..., n$. When the points lie very nearly along a straight line, the LPC assumption remains tenable. The straightness of the Q-Q plot can be measured by calculating the correlation coefficient of the points in the plot. The correlation coefficient for the Q-Q plot is defined by

$$r_Q = \frac{p}{\sqrt{qt}}. \tag{2.20}$$

Formally, we reject the hypothesis of the assumption of the LPC distribution at level of significance α if r_Q falls below the appropriate value in Table 2.2. Note that a single entry of this table is obtained by simulating a 100,000 random samples from the LPC distribution with given sample size n.

Table 2.2 Critical points for the correlation coefficient test of LPC

Sample size	Significance levels α				Sample size	Significance levels α			
n	.01	.05	.10	.20	n	.01	.05	.10	.20
5	.774	.848	.885	.914	45	.926	.954	.964	.973
10	.849	.896	.916	.937	50	.929	.957	.967	.975
15	.879	.915	.932	.949	55	.933	.960	.969	.977
20	.895	.928	.942	.957	60	.937	.962	.970	.978
25	.904	.937	.949	.962	75	.943	.967	.974	.981
30	.912	.942	.954	.966	100	.952	.972	.979	.984
35	.917	.947	.959	.969	200	.969	.983	.987	.990
40	.922	.951	.962	.971	300	.976	.987	.990	.993

2.7 Bayesian estimators for complete data

As given in Section 2.4, the likelihood function of the LPC density for the unknown parameters can be written as

$$L(\beta,\theta/x_1,...,x_n)=\frac{\beta^n\theta^{n\beta}}{(1+\Phi(k_1))^n\prod_{i=1}^n x_i^{\beta+1}}e^{-\frac{1}{2}(\frac{\beta}{k_1})^2\sum_{i=1}^m \ln^2(\frac{x_i}{\theta})}. \tag{2.21}$$

We employ the following Jeffrey's (1961) prior (see appendix A for expected information matrix calculation for LPC) for β and θ,

$$\pi(\beta,\theta)=\frac{c}{\theta},\ c>0,\ \theta>0, \tag{2.22}$$

where constant c does not depend on the parameters.

For our simplicity, now let $p_n=(1+\Phi(k_1))^n\prod_{i=1}^n x_i$, $s_n=\sum_{i=1}^n \ln x_i$, $s_m=\sum_{i=1}^m \ln x_i$, $s_{m^2}=\sum_{i=1}^m \ln^2 x_i$, and $d_m=\sqrt{s_{m^2}-s_m^2/m}/k_1$. Then the joint posterior pdf of β and θ can be written as, (see Berger 1985),

$$\begin{aligned}\pi(\beta,\theta/x_1,...,x_n) &= \frac{c\beta^n}{p_n j\theta}e^{\frac{1}{2}\left[\frac{n^2k^2}{m}+\left(\frac{ns_m-ms_n}{md_m}\right)^2\right]}e^{-\frac{1}{2}d_m^2\left[\beta-\left(\frac{ns_m-ms_n}{md_m^2}\right)\right]^2}\\ &\quad e^{-\frac{1}{2}m(\beta/k_1)^2\left[\ln\theta-\left(s_m+nk_1^2/\beta\right)/m\right]^2},\ \beta>0,\ \theta>0,\end{aligned} \tag{2.23}$$

where constant c does not depend on the parameters. Also the value of j can be obtained from the following equations by writing $\left(\frac{ns_m-ms_n}{md_m}\right)=a_0$,

$$j=\frac{2\pi ck_1}{p_n d_m^n\sqrt{m}}e^{\frac{1}{2}\left[n^2k^2/m+a_0^2\right]}I_n, \tag{2.24}$$

where

$$I_n=\int_{-a_0}^{\infty}\frac{1}{\sqrt{2\pi}}(t+a_0)^{n-1}e^{-\frac{1}{2}t^2}dt. \tag{2.25}$$

The value of I_n can be obtained from the following recursive formula with $I_1 = \Phi(a_0)$ and $I_2 = a_0\Phi(a_0) + \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}a_0^2}$.

$$I_{n+2} = nI_n + a_0 I_{n+1}. \tag{2.26}$$

The marginal posterior pdf of parameter β can be written as

$$\pi(\beta/x_1, ..., x_n) = \frac{\sqrt{2\pi}ck_1\beta^{n-1}}{p_n j\sqrt{m}} e^{\frac{1}{2}\left[n^2k^2/m + a_0^2\right]} e^{-\frac{1}{2}d_m^2[\beta - a_0/d_m]^2}, \ \beta > 0. \tag{2.27}$$

The Bayes estimator for the parameter β, $\widehat{\widehat{\beta}}_{\mathrm{B}}$ under the squared error loss function can be obtained from the following equation

$$\widehat{\widehat{\beta}}_{\mathrm{B}} = \begin{cases} \max\limits_m \frac{I_{n+1}}{d_m I_n} & \text{for small } n \\ \max\limits_m \frac{a_0 + \sqrt{a_0^2 + 4n}}{2d_m} & \text{for large } n \end{cases}. \tag{2.28}$$

Suppose in equation (2.28), the maximum value occurred at $m = m_1$.Then the standard error of $\widehat{\widehat{\beta}}_{\mathrm{B}}$ can be obtained as

$$SE_{\widehat{\widehat{\beta}}_{\mathrm{B}}} = \left.\frac{\sqrt{I_n I_{n+2} - I_{n+1}^2}}{d_m I_n}\right|_{m=m_1}. \tag{2.29}$$

The generalized maximum likelihood estimator for parameter β, $\widehat{\widehat{\beta}}_{\mathrm{GML_I}}$ can be obtained from the following equation

$$\widehat{\widehat{\beta}}_{\mathrm{GML_I}} = \left.\left[a_0 + \sqrt{4(n-1) + a_0^2}\right] \Big/ (2d_m)\right|_{m=m_1}. \tag{2.30}$$

The marginal posterior pdf of parameter θ can be written as

$$\pi(\theta/x_1, ..., x_n) = \int_0^\infty \pi(\beta, \theta/x_1, ..., x_n)d\beta, \ \theta > 0. \tag{2.31}$$

As before, the Bayes estimator for the parameter θ, $\widehat{\widehat{\theta}}_{\mathrm{B}}$ under the squared error loss function can be obtained from the following equation

$$\widehat{\widehat{\theta}}_{\mathrm{B}} = \min_{m} \int_0^\infty \theta\pi(\theta/x_1,...,x_n)d\theta = \frac{\sqrt{2\pi}ck_1}{p_n j\sqrt{m}} e^{\left(s_m+\frac{1}{2}n^2k_1^2\right)/m} J_n, \tag{2.32}$$

where

$$J_n = \int_0^\infty \beta^{n-1} \exp\left(\frac{nk_1^2}{m\beta} + \left(\frac{ns_m - ms_n}{m}\right)\beta + \frac{k_1^2}{2m\beta^2} - \frac{d_m^2\beta^2}{2}\right) d\beta \tag{2.33}$$

does not exist. Hence the Bayes estimator, $\widehat{\widehat{\theta}}_{\mathrm{B}}$ does not exist. Similar problem occurred with the Lognormal (Zellner 1971).

Now we consider the conditional posterior pdf of parameter θ (> 0) given β (> 0),

$$\begin{aligned} \pi(\theta/\beta, x_1,...,x_n) &= \frac{\pi(\beta,\theta/x_1,...,x_n)}{\pi(\beta/x_1,...,x_n)} \\ &= \frac{\beta\sqrt{m}}{\sqrt{2\pi}k_1\theta} e^{-\frac{1}{2}m(\beta/k_1)^2\left[\ln\theta-\left(s_m+nk_1^2/\beta\right)/m\right]^2}. \end{aligned} \tag{2.34}$$

Clearly, $\pi(\theta/\beta, x_1,...,x_n)$ is lognormally distributed with $\mu = \left(s_m + nk_1^2/\beta\right)/m$, and $\sigma^2 = \frac{k_1^2}{m\beta^2}$. Hence the conditional Bayes estimator for parameter θ given β is,

$$\widehat{\widehat{\theta}}_{\mathrm{B}}/\left(\beta = \widehat{\widehat{\beta}}_{\mathrm{B}}\right) = \exp\left(\left(s_m + \frac{nk_1^2}{\widehat{\widehat{\beta}}_{\mathrm{B}}} + \frac{k_1^2}{2\widehat{\widehat{\beta}}_{\mathrm{B}}^2}\right)\Big/ m\right)\Bigg|_{m=m_1}. \tag{2.35}$$

The conditional standard error of $\widehat{\widehat{\theta}}_{\mathrm{B}}\Big/\left(\beta = \widehat{\widehat{\beta}}_{\mathrm{B}}\right)$ can be obtained as

$$SE_{\widehat{\widehat{\theta}}_{\mathrm{B}}/\left(\beta=\widehat{\widehat{\beta}}_{\mathrm{B}}\right)} = \exp\left(\left(s_m + \frac{nk_1^2}{\widehat{\widehat{\beta}}_{\mathrm{B}}} + \frac{k_1^2}{2\widehat{\widehat{\beta}}_{\mathrm{B}}^2}\right)\Big/ m\right)\sqrt{e^{\left(\frac{k_1^2}{m\widehat{\widehat{\beta}}_{\mathrm{B}}^2}\right)} - 1}\Bigg|_{m=m_1}. \tag{2.36}$$

The generalized conditional maximum likelihood estimator for θ is,

$$\widehat{\widehat{\theta}}_{\mathrm{GML_I}}\Big/\left(\beta = \widehat{\widehat{\beta}}_{\mathrm{GML_I}}\right) = \exp\left(\left(s_m + \frac{nk_1^2}{\widehat{\widehat{\beta}}_{\mathrm{GML_I}}} - \frac{k_1^2}{\widehat{\widehat{\beta}}_{\mathrm{GML_I}}^2}\right)\Big/ m\right)\Bigg|_{m=m_1}. \tag{2.37}$$

The generalized maximum likelihood estimators for parameters β and θ, i.e. $\widehat{\widehat{\beta}}_{\mathrm{GML_{II}}}$ and $\widehat{\widehat{\theta}}_{\mathrm{GML_{II}}}$ can be obtained from equations (2.38) and (2.39). These equations are derived by differentiating the joint posterior density given in equation (2.23) with respect to the parameters β and θ.

$$\left(ms_{m^2} - s_m^2\right)\beta^4 + \left(ms_n - ns_m\right)k_1^2\beta^3 - mnk_1^2\beta^2 - nk_1^4\beta + k_1^4 = 0\Big|_{m=m_0}.$$

$$\theta = \exp\left(\left(s_m + \frac{nk_1^2}{\beta} - \frac{k_1^2}{\beta^2}\right)\Big/\, m\right)\Big|_{m=m_0}. \qquad \begin{matrix}(2.38)\\(2.39)\end{matrix}$$

Where, at $m = m_0$, the posterior mode has the highest density value.

2.8 Goodness-of-fit tests

In this section we consider the test of fit based on empirical distribution function (EDF). The EDF is a step function calculated from the sample which estimates the population distribution function. EDF statistics are measures of discrepancy between the EDF and the given distribution function, and are used for testing the fit of the sample to the distribution. Here we consider our two-parameter lognormal-Pareto composite distribution. As we know the EDF test statistics are much powerful than the chi-squared test statistic, for example, in order to perform the chi-squared test data must be grouped in which case we may lose some information.

Extensions of EDF statistics to situations involving randomly censored data, the Kaplan-Meier (1958) estimator is generally used for the true distribution. For analog versions of the Kolmogorov-Smirnov, Kuiper, and Cramér-von Mises are found in Koziol (1980), Nair (1981), or Fleming *et al.* (1980). But, these three versions are not computationally convenient and hence we did not used such versions in this dissertation.

The following definitions are taken from D'Agostino and Stephens (1986).

1. The Empirical Distribution Function (EDF)

Suppose a given random sample of size n is $X_1, ..., X_n$ and let $X_{(1)} < ... < X_{(n)}$ be the order statistics: suppose further that the distribution of X is $F(x)$. Here we assume this distribution to be continuous. The EDF is $F_n(x)$, and is defined as

$$F_n(x) = \frac{\text{number of observation } \leq x}{n}; -\infty < x < \infty.$$

More precisely, the definition is

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ \frac{i}{n}, & X_{(1)} \leq x < X_{(i+1)}, i = 1, ..., n-1 \\ 1, & X_{(n)} < x \end{cases} .$$

Thus $F_n(x)$ is a step function, calculated from the data; as x increases it takes step up of height $1/n$ as each sample observation is reached.

2. Kolmogorov - Smirnov Statistics (Supremum Statistics)

The first two EDF statistics, D^+ and D^- are respectively, the largest vertical difference when $F_n(x)$ is greater than $F(x)$, and the largest vertical difference when $F_n(x)$ is smaller than $F(x)$; formerly, $D^+ = \sup_x\{F_n(x) - F(x)\}$ and $D^- = \sup_x\{F(x) - F_n(x)\}$. For calculation purposes, we can rewrite these statistics as, $D^+ = \max_i\{\frac{i}{n} - F(x_{(i)})\}$ and $D^- = \max_i\{F(x_{(i)}) - \frac{i-1}{n}\}$.

The most well known EDF test statistic is D (Kolmogorov 1933), and is defined as

$$D = \max(D^+, D^-).$$

For large n, value of D equals to zero, since $n \to \infty$, $|F_n(x) - F(x)|$ decreases to zero with probability one. Hence $F_n(x)$ is a consistent estimator for $F(x)$.

3. Anderson - Darling Statistic

Anderson - Darling (1954) test statistic is given by

$$\begin{aligned} A^2 &= -n - (1/n)\sum_i^n (2i-1)[\ln F(x_{(i)}) + \ln\{1 - F(x_{(n+1-i)})\}] \\ &= -n - (1/n)\sum_i^n [(2i-1)\ln F(x_{(i)}) + (2n+1-2i)\ln\{1 - F(x_{(i)})\}]. \end{aligned}$$

Anderson-Darling test statistic is much powerful than the Kolmogorov-Smirnov test statistic. Specifically Anderson-Darling test statistic is much sensitive to the tail area of the data distribution whereas the Kolmogorov-Smirnov is more sensitive to the middle portion of the data distribution.

We consider four different cases to calculate the upper percentage points of these statistics for the LPC distribution.

Case 0: both β and θ known;

Case 1: β known, θ unknown;

Case 2: β unknown θ known;

Case 3: both β and θ unknown.

Table 2.3 Upper tail percentage points of D and A^2 statistics, case 0

	Significance level α					
n	.250	.100	.050	.025	.010	.005
	Case 0. Statistic D					
5	.424	.510	.564	.613	.666	.704
10	.306	.369	.409	.446	.488	.517
20	.220	.265	.294	.320	.351	.373
50	.141	.170	.189	.205	.225	.239
100	.100	.120	.134	.146	.161	.170
500	.045	.054	.060	.066	.072	.076
1000	.032	.038	.043	.047	.051	.054
	Case 0. Statistic A^2					
5	1.237	1.950	2.539	3.164	3.992	4.611
10	1.250	1.944	2.510	3.102	3.923	4.569
20	1.243	1.941	2.508	3.097	3.924	4.561
50	1.242	1.922	2.496	3.083	3.911	4.501
100	1.249	1.947	2.504	3.072	3.890	4.486
500	1.249	1.938	2.485	3.055	3.813	4.433
1000	1.249	1.936	2.485	3.067	3.863	4.474

Table 2.4 Upper tail percentage points of D and A^2 statistics, case 1

	Significance level α					
n	.250	.100	.050	.025	.010	.005
	Case 1. Statistic D					
5	.363	.435	.478	.519	.566	.593
10	.262	.315	.349	.379	.415	.439
20	.188	.226	.252	.274	.301	.319
50	.121	.145	.161	.176	.194	.206
100	.086	.103	.115	.125	.138	.146
500	.039	.047	.052	.057	.062	.066
1000	.028	.033	.037	.040	.044	.047
	Case 1. Statistic A^2					
5	.766	1.138	1.432	1.731	2.146	2.543
10	.771	1.154	1.460	1.767	2.234	2.582
20	.781	1.170	1.483	1.825	2.262	2.614
50	.776	1.168	1.486	1.816	2.242	2.581
100	.778	1.171	1.484	1.821	2.289	2.622
500	.780	1.173	1.500	1.838	2.313	2.705
1000	.781	1.174	1.488	1.827	2.287	2.611

Table 2.5 Upper tail percentage points of D and A^2 statistics, case 2

	Significance level α					
n	.250	.100	.050	.025	.010	.005
	Case 2. Statistic D					
5	.416	.500	.562	.618	.693	.740
10	.296	.358	.398	.434	.476	.506
20	.212	.257	.287	.313	.345	.366
50	.136	.164	.183	.200	.220	.235
100	.096	.117	.130	.143	.157	.166
500	.043	.053	.059	.064	.071	.075
1000	.031	.037	.042	.046	.050	.054
	Case 2. Statistic A^2					
5	1.101	1.749	2.404	3.209	4.360	5.822
10	1.100	1.736	2.212	2.688	3.376	3.978
20	1.073	1.740	2.275	2.805	3.468	4.032
50	1.071	1.744	2.307	2.881	3.657	4.221
100	1.069	1.745	2.291	2.871	3.609	4.219
500	1.061	1.723	2.281	2.857	3.682	4.259
1000	1.071	1.754	2.310	2.883	3.648	4.281

Table 2.6 Upper tail percentage points of D and A^2 statistics, case 3

n	Significance level α .250	.100	.050	.025	.010	.005
	Case 3. Statistic D					
5	.292	.336	.364	.386	.407	.419
10	.215	.249	.272	.292	.316	.331
20	.155	.181	.199	.214	.232	.245
50	.100	.117	.128	.138	.150	.158
100	.072	.084	.092	.099	.108	.114
500	.032	.038	.041	.045	.049	.051
1000	.023	.027	.029	.032	.035	.037
	Case 3. Statistic A^2					
5	.477	.615	.715	.803	.911	.996
10	.486	.646	.759	.872	1.022	1.129
20	.487	.654	.779	.897	1.064	1.187
50	.487	.661	.790	.917	1.085	1.209
100	.487	.658	.785	.914	1.084	1.205
500	.490	.663	.791	.918	1.087	1.210
1000	.490	.664	.792	.926	1.098	1.218

2.9 Maximum likelihood estimators for right censored data

In most cases with insurance payments, there is a limit for the maximum amount of payment, i.e., the data are Type I right censored. In this section, we look at the estimation of model parameters of the LPC distribution for such data.

Suppose we have $n+f$ sample values and f of those values are censored at u and as in Section 2.4 the remaining n uncensored ordered values are: $X_1, X_2, ..., X_n$. If the unknown parameter θ is in between the m^{th} observation and $m+1^{\text{th}}$ observation, the log-likelihood function is given by

$$\begin{aligned} \ln L(\beta,\theta) &= -(n+f)\ln\left(1+\Phi\left(k_1\right)\right)+n\ln\beta+(n+f)\beta\ln\theta \\ &\quad -(\beta+1)\textstyle\sum_{i=1}^{n}\ln x_i-0.5(\beta/k_1)^2\sum_{i=1}^{m}\ln^2(x_i/\theta). \end{aligned} \tag{2.40}$$

An algorithm to evaluate maximum likelihood (ML) estimators:

Step 1: for each m $(m=1,2,...,n-1)$, calculate $\widehat{\beta}_m$ and $\widehat{\theta}_m$ as follows:

For $m=1$, $\widehat{\beta}_1 = n\left(f\ln(u/x_1)+\sum_{i=1}^{n}\ln\left(x_i/x_1\right)\right)^{-1}$

$$\widehat{\theta}_1 = x_1(u/x_1)^{(1+f/n)fk^2}\textstyle\prod_{i=1}^{n}(x_i/x_1)^{(1+f/n)k^2}.$$

Otherwise

$$\widehat{\beta}_m = \left(k_1^2C+\sqrt{k_1^4C^2+4mnk_1^2A}\right)\Big/\left(2A\right), \tag{2.41}$$

and

$$\widehat{\theta}_m = \left(\exp\left((n+f)k_1^2/\widehat{\beta}_m\right)\textstyle\prod_i^m x_i\right)^{1/m}, \tag{2.42}$$

where $A = m\sum_{i=1}^{m}\left(\ln x_i\right)^2-\left(\sum_{i=1}^{m}\ln x_i\right)^2$, and $C = mf\ln u-(n+f)\sum_{i=1}^{m}\ln x_i+m\sum_{i=1}^{n}\ln x_i$.

If $\widehat{\theta}_m$ is in between $x_m \le \widehat{\theta}_m \le x_{m+1}$, then the ML estimators of β and θ are

$$\widehat{\beta}_{\mathrm{ML}} = \widehat{\beta}_m, \;\; \widehat{\theta}_{\mathrm{ML}} = \widehat{\theta}_m. \tag{2.43}$$

Let us rewrite the equation (2.42) such that

$$\left((n+f)k_1^2/\widehat{\beta}_m\right) + \textstyle\sum_{i=1}^{m} \ln(x_i/\widehat{\theta}_m) = 0. \tag{2.44}$$

From equation (2.44), there must be at least one x_i value less than $\widehat{\theta}_m$, since n, k_1, and $\widehat{\alpha}_m$ are positive values. Therefore the ML estimate of θ cannot occur at x_1.

Step 2: if there is no solution for θ (i.e., $x_n \leq \widehat{\theta}_m$) with the conditions given in Step 1, the ML estimate of β and θ are

$$\widehat{\beta}_{\mathrm{ML}} = \left(fk_1^2E + \sqrt{f^2k_1^4E^2 + 4n^2k_1^2D}\right) \Big/ (2D), \tag{2.45}$$

and

$$\widehat{\theta}_{\mathrm{ML}} = \left(\exp\left((n+f)k_1^2/\widehat{\beta}_{\mathrm{ML}}\right) \textstyle\prod_i^n x_i\right)^{1/n}, \tag{2.46}$$

where $D = n\sum_{i=1}^{n} (\ln x_i)^2 - \left(\sum_{i=1}^{n} \ln x_i\right)^2$, and $E = \sum_{i=1}^{n} \ln(u/x_i)$.

In this case as in Section 2.4, if $\widehat{\theta}_{\mathrm{ML}}$ is closer to x_1 or x_n, it is easy to show that the LPC distribution is inappropriate and depending on the situation, one need to use the Pareto or the lognormal distributions respectively.

2.10 Illustrative examples

Example 1: Simulated data set.

The following small data set is simulated from the LPC distribution with parameters $\beta = 1$ and $\theta = 10$.

Data set: 7.00, 7.69, 9.19, 9.42, 9.48, 12.21, 15.16, 18.37, 30.07, 6816.38

Table 2.7 provides the estimated parameter values and their standard errors of the fitted LPC distribution using the three different estimation methods. Furthermore, Kolmogorov-Smirnov (D) and Anderson-Darling (A^2) test statistic values are added to this table. Figure 2.4 illustrate the joint posterior pdf surface of β and θ for the simulated data set. Its contours are given in Figure 2.5.

Table 2.7 Estimated parameter, D and A^2 values for simulated data

Estimation Method	Parameter values	D, A^2
Maximum likelihood	$\widehat{\beta}_{\mathrm{ML}} \pm SE_{\widehat{\beta}_{\mathrm{ML}}} = 0.9799 \pm 0.3020$	$D = 0$ .216
	$\widehat{\theta}_{\mathrm{ML}} \pm SE_{\widehat{\theta}_{\mathrm{ML}}} = 11.2685 \pm 2.1517$	$A^2 = 0.770$
Least square	$\widetilde{\beta}_{\mathrm{LS}} \pm SE_{\widetilde{\beta}_{\mathrm{LS}}} = 0.5226 \pm 0.0989$	$D = 0.286$
	$\widetilde{\theta}_{\mathrm{LS}} \pm SE_{\widetilde{\theta}_{\mathrm{LS}}} = 9.6195 \pm 3.4880$	$A^2 = 1.142$
Bayesian	$\widehat{\widehat{\beta}}_{\mathrm{B}} \pm SE_{\widehat{\widehat{\beta}}_{\mathrm{B}}} = 0.8092 \pm 0.2180$	$D = 0.279$
	$\widehat{\widehat{\theta}}_{\mathrm{B}} \pm SE_{\widehat{\widehat{\theta}}_{\mathrm{B}}} = 13.2068 \pm 2.1623$	$A^2 = 1.285$
Generalized mle I	$\widehat{\widehat{\beta}}_{\mathrm{GML_I}} \pm SE_{\widehat{\widehat{\beta}}_{\mathrm{GML_I}}} = 0.7640 \pm 0.2227$	$D = 0.262$
(marginal)	$\widehat{\widehat{\theta}}_{\mathrm{GML_I}} \pm SE_{\widehat{\widehat{\theta}}_{\mathrm{GML_I}}} = 12.8132 \pm 2.1978$	$A^2 = 1.142$
Generalized mle II	$\widehat{\widehat{\beta}}_{\mathrm{GML_{II}}} = 0.8360$	$D = 0.261$
(joint)	$\widehat{\widehat{\theta}}_{\mathrm{GML_{II}}} = 12.6272$	$A^2 = 1.104$

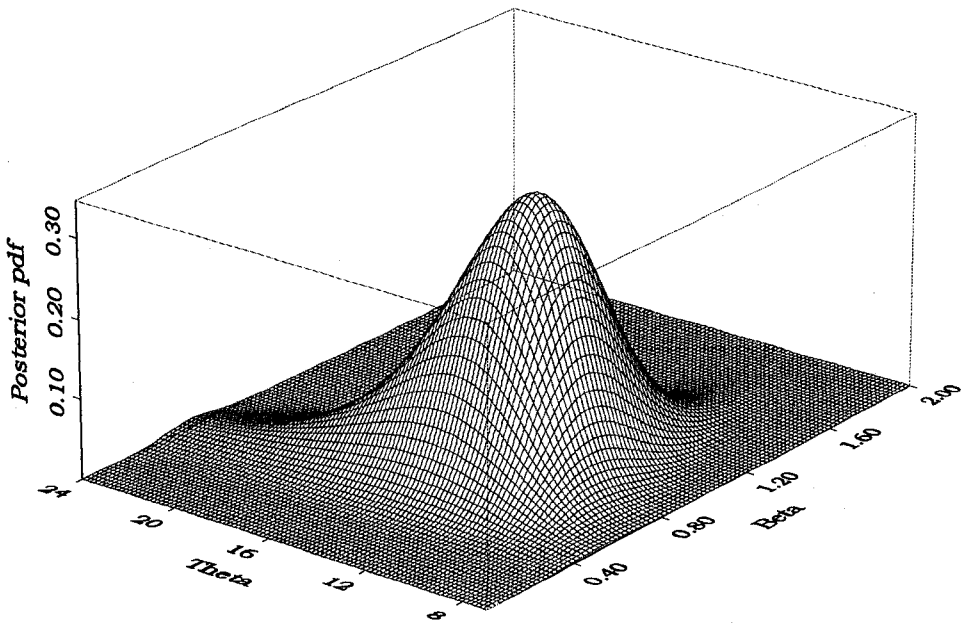

Figure 2.4 Joint posterior pdf surface of β and θ for the simulated data.

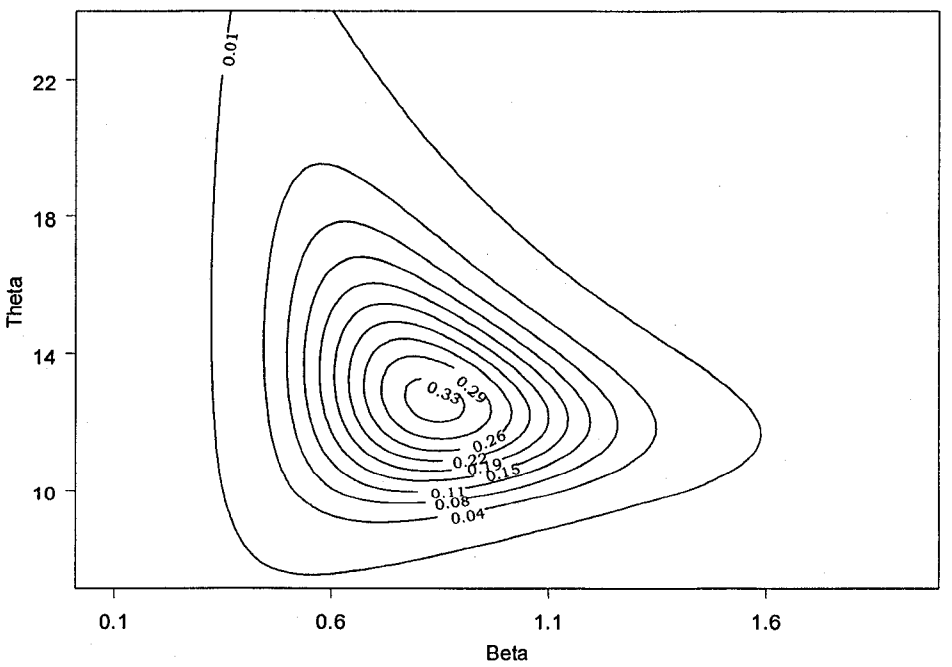

Figure 2.5 Joint posterior pdf contours of β and θ for the simulated data.

Example 2: Danish fire insurance loss data.

This actuarial data set (see appendix B) is taken from Cooray and Ananda (2005). The complete Danish data set consist of 2492 fire insurance losses in Danish Krone (DKK) from the years 1980 to 1990 inclusive. The loss figure is a total loss figure for the events concerned and includes damage to buildings, furniture and personal property as well as loss of profits.

The recorded data have been suitably adjusted to reflect 1985 values. The adjusted loss values in Danish Krone range from (in millions) 0.3134041 to 263.2503660. The figure 2.6 illustrate the histogram of the fire loss data.

As in the previous example, Table 2.8 provides the estimated parameter values and their standard errors of the fitted LPC distribution using the three different estimation methods. Kolmogorov-Smirnov (D) and Anderson-Darling (A^2) test statistic values are added to this table. Figure 2.9 illustrate the Q-Q plot of lognormal, Pareto and LPC distributions to the Danish data under ML estimation method. According to this quantile plot, one can clearly see that the LPC distribution is more reasonable to model the Danish data than other models.

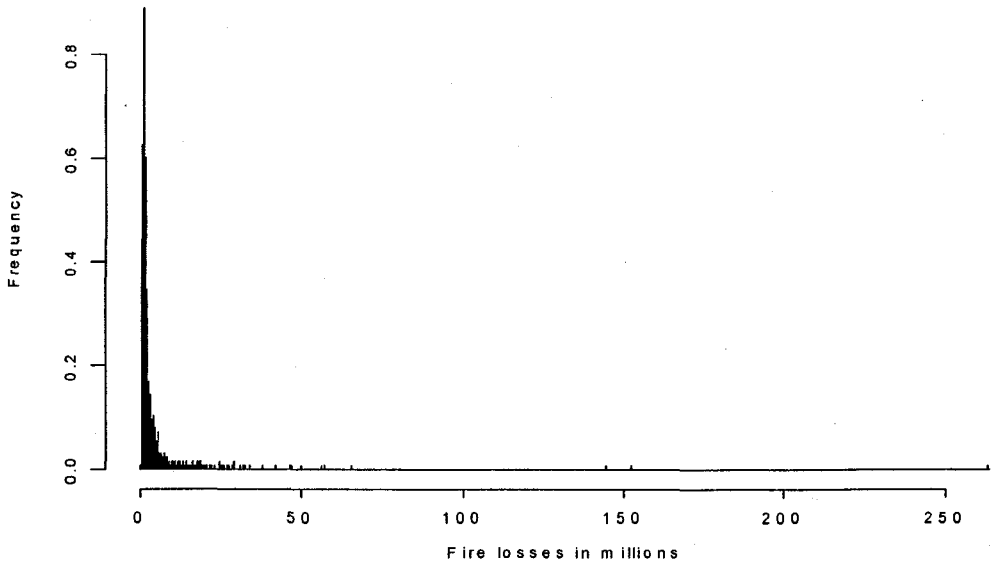

Figure 2.6 Histogram of Danish fire loss data.

Table 2.8 Estimated parameter, D and A^2 values for Danish data

Estimation Method	Parameter values	D, A^2
Maximum likelihood	$\widehat{\beta}_{\mathrm{ML}} \pm SE_{\widehat{\beta}_{\mathrm{ML}}} = 1.4363 \pm 0.0270$	$D = 0.029$
	$\widehat{\theta}_{\mathrm{ML}} \pm SE_{\widehat{\theta}_{\mathrm{ML}}} = 1.3851 \pm 0.0135$	$A^2 = 4.258$
Least square	$\widetilde{\beta}_{\mathrm{LS}} \pm SE_{\widetilde{\beta}_{\mathrm{LS}}} = 1.4126 \pm 0.0017$	$D = 0.036$
	$\widetilde{\theta}_{\mathrm{LS}} \pm SE_{\widetilde{\theta}_{\mathrm{LS}}} = 1.4046 \pm 0.0013$	$A^2 = 4.374$
Bayesian	$\widehat{\widehat{\beta}}_{\mathrm{B}} \pm SE_{\widehat{\widehat{\beta}}_{\mathrm{B}}} = 1.4363 \pm 0.0000$	$D = 0.029$
	$\widehat{\widehat{\theta}}_{\mathrm{B}} \pm SE_{\widehat{\widehat{\theta}}_{\mathrm{B}}} = 1.3852 \pm 0.0117$	$A^2 = 4.258$
Generalized mle I	$\widehat{\widehat{\beta}}_{\mathrm{GML_I}} \pm SE_{\widehat{\widehat{\beta}}_{\mathrm{GML_I}}} = 1.4358 \pm 0.0005$	$D = 0.029$
(marginal)	$\widehat{\widehat{\theta}}_{\mathrm{GML_I}} \pm SE_{\widehat{\widehat{\theta}}_{\mathrm{GML_I}}} = 1.3852 \pm 0.0117$	$A^2 = 4.246$
Generalized mle II	$\widehat{\widehat{\beta}}_{\mathrm{GML_{II}}} = 1.4365$	$D = 0.029$
(joint)	$\widehat{\widehat{\theta}}_{\mathrm{GML_{II}}} = 1.3850$	$A^2 = 4.264$

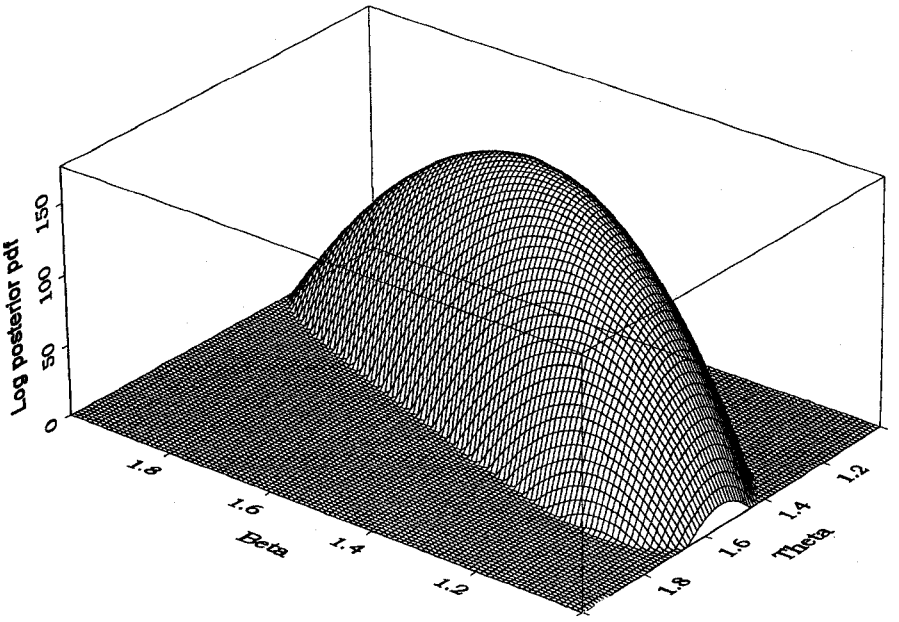

Figure 2.7 Joint log posterior pdf surface of β and θ for the Danish data.

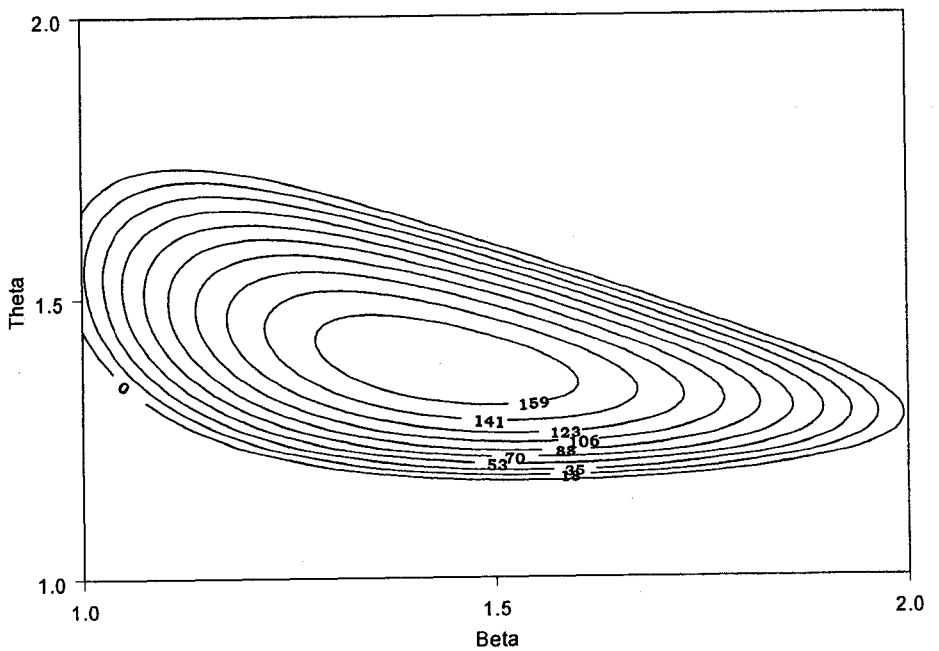

Figure 2.8 Joint log posterior pdf contours of β and θ for the Danish data.

Table 2.9 Estimated goodness-of-fit values for different distributions

Distribution	$l(\widehat{\boldsymbol{\theta}})$	D	A^2
LPC	-3877.84	0.029	4.2585
Lognormal	-4433.89	0.127	85.493
Pareto	-5675.09	0.408	496.64
Loglogistic	-4280.59	0.114	52.502
Inverse Gaussian	-4516.31	0.172	137.48
Gamma	-5243.03	0.201	212.58
Weibull	-5270.47	0.255	219.37

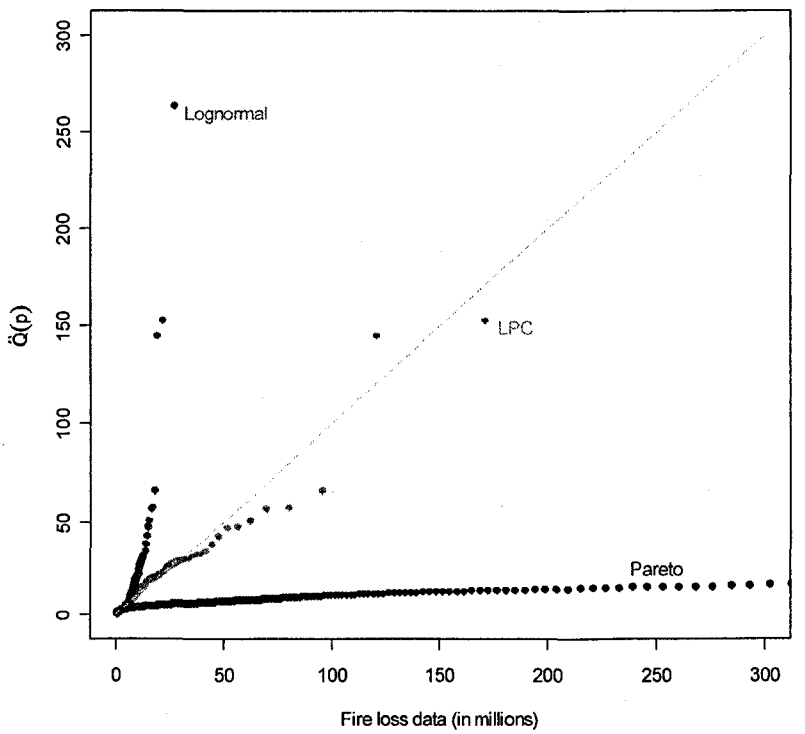

Figure 2.9 Q-Q plot of Danish data for the three distributions.

2.11 The Weibull-Pareto composite distribution

The Weibull distribution, which is frequently used for life data analysis, is composited with the Pareto model to obtain a flexible, reliable long-tailed parametric distribution for modeling unimodal failure rate data. This smooth continuous composition, Weibull-Pareto composite (WPC) family, behaves as a two-parameter Weibull density up to an unknown threshold value, and as a two-parameter Pareto density for the remainder. The two-parameter resulting composite density is similar in shape to the Weibull density, yet the upper tail being larger than the Weibull density, and quite similar in the tail behavior to the Pareto density. The hazard function of the composite family accommodates decreasing and unimodal failure rates, which are separated by the boundary line of the space of shape parameter, gamma, when it equals to a known constant. The maximum likelihood parameter estimation techniques are discussed by providing approximate conditional coverage probabilities for uncensored samples. The advantages of using the new family are demonstrated and compared by illustrating well-known examples: guinea pigs survival time data, head and neck cancer data, and nasopharynx cancer survival data. Another set of authors (Preda and Crumara 2006) is formulated this density to compare with LPC density (Cooray and Ananda 2005) by analyzing a simulated data set. However here our aim is to discuss the flexibility of WPC density and its hazard function for modeling survival data.

2.11.1 Motivation to medical diagnostics

The Weibull distribution, having monotonic increasing, monotonic decreasing, and constant hazard rates, is often used for modeling survival data. However, this density is not an appropriate distribution to model non-monotonic failure rates, in particular bathtub or unimodal shapes. Even though the emphasis has traditionally been placed on models with bathtub-shaped hazard functions, the variety of applications in biostatistical area is appropriately modeled by the densities with unimodal (hump-shaped) hazard functions. This can be illustrated by such examples as survival times of guinea pigs infected with different doses of virulent tubercle bacilli (Bjerkedal 1960), nonresectable gastric carcinoma data (Stablein *et al.* 1981), nasopharynx cancer survival data (West 1987), and head and neck cancer data (Efron 1988).

Since there is an initial increase in risk after successful surgery in biomedical area, the unimodal hazard rate is often used to model survivability. This risk, due to infection, hemorrhage, or other complications after the procedure, is followed by a steady decline as the patients recovers. In another similar example found in epidemiology, patients with tuberculosis have a risk that initially increases and then decreases after the treatment.

The well-known such parametric distributions as loglogistic, lognormal, Birnbaum-Saunders, and inverse Gaussian, which produce unimodal-shaped hazard functions, are desirable for analyzing unimodal failure rate data due to their computational simplicity and popularity among users. However, when these models are inadequate or inappropriate, alternative models or higher order parametric families must be considered for the purpose of modeling such failure rate data. In this regard, nonresectable

gastric carcinoma data (Stablein *et al.* 1981) was analyzed by Ghitany (2001) using a two-parameter compound Rayleigh distribution, which is earlier used by Greenwich (1992) to model uncensored data regarding the survival times of guinea pigs infected with virulent tubercle bacilli (Bjerkedal 1960). This unimodal hazard rate function is particularly useful when the peak time of failure rate is prime interest. For example, if the peak failure time of certain individuals is less than their mean failure time, immediate care must be taken in order to reduce the risk of those individuals.

Furthermore, Glen and Leemis (1997) have used another two-parameter family of lifetime distribution, the arctangent survival distribution, for the purpose of modeling unimodal failure rate data. Efron (1988) used linear (three parameters), cubic (five parameters), and cubic-linear (six parameters) models to analyze the arm A head and neck cancer data. Later, Mudholkar *et al.* (1996) obtained an improved fit for the arm A head and neck cancer data using three-parameter generalized Weibull distributions. A two-parameter composite family of distribution, which is the main topic of this Section, is used to analyze the arm A head and neck cancer data. These analyses reveal that the underlying hazard function for the head and neck cancer data has quite thick upper tails with initial high-risk period. Even though the high-risk period can be modeled by the Weibull type distributions, one can recognize the partial-necessity of Pareto type families for fitting the tail area of such failure data.

In fact, the two-parameter Pareto model supports in modeling longer lifetimes, but fails to cover the behavior of shorter lifetimes. Similarly, the two-parameter Weibull model covers the behavior of shorter lifetimes than it does for the longer lifetimes. Taking into account the tail behavior of both short and long lifetimes, a

natural composition from the Weibull and the Pareto family is found for the purpose of modeling unimodal failure rate data. The two-parameter Weibull density is up to an unknown threshold value and the two-parameter Pareto density for the rest of the model. Differentiability and continuity at the threshold point yield a fine smooth density function called the Weibull-Pareto composite (WPC) distribution with two unknown parameters. The resulting density has the larger right tail than the Weibull density has, and is similar to the Weibull density. Cooray and Ananda (2005) introduced one such two-parameter composition: the lognormal-Pareto composite (LPC) distribution for analyzing highly positively skewed data, which usually arise in insurance industry and actuarial sciences. However, the LPC is not suitable for survival data analyses due to its computational difficulties with the hazard function or survival function. Alternatively, the WPC distribution is a useful lifetime distribution because it has not only closed-form survival and the hazard functions but also a more flexible left tail than the LPC distribution. Finally, the heavy right tail of the WPC distribution is useful for evaluating survivors that fail with less risk once they have survived a certain time threshold.

2.11.2 Model derivation

Let X be a random variable with the pdf

$$f(x) = \begin{cases} cf_1(x) & \text{if } 0 < x \leq \theta \\ cf_2(x) & \text{if } \theta \leq x < \infty \end{cases}, \tag{2.47}$$

where c is the normalizing constant, $f_1(x)$ has the form of the regular Weibull density, and $f_2(x)$ has the form of the two-parameter Pareto density, i.e.,

$$f_1(x) = (\tau/x)(x/\phi)^{\tau} \exp\{-(x/\phi)^{\tau}\},\ x > 0, \tag{2.48}$$

and

$$f_2(x) = (\gamma/x)(\theta/x)^{\gamma},\ x \geq \theta. \tag{2.49}$$

Here $\gamma, \theta, \tau, \phi$ are unknown parameters such that $\gamma > 0, \theta > 0, \tau > 0, \phi > 0$.

Let us impose the continuity and differentiability conditions at θ,

$$f_1(\theta) = f_2(\theta),\ f_1'(\theta) = f_2'(\theta), \tag{2.50}$$

where $f'(\theta)$ is the first derivative of $f(x)$ evaluated at θ. These conditions guarantee that we have a smooth probability density function. These two restrictions reduce the total unknown parameters from four to two. One can show that (see the proof in the end of the section) this composite density can be reparameterized and rewritten as

$$f(x) = \begin{cases} \frac{(k_2+1)^2\gamma}{(2k_2+1)x}\left(\frac{x}{\theta}\right)^{\gamma k_2} \exp\left\{-\left(\frac{k_2+1}{k_2}\right)\left(\frac{x}{\theta}\right)^{\gamma k_2}\right\} & \text{if } 0 < x \leq \theta \\ \left(\frac{k_2+1}{2k_2+1}\right)\left(\frac{\gamma}{x}\right)\left(\frac{\theta}{x}\right)^{\gamma} & \text{if } \theta \leq x < \infty \end{cases}, \tag{2.51}$$

where k_2 is a known constant which is given by the positive solution of the equation $\exp\left(1+\frac{1}{k}\right) = k+1$. This value is $k_2 = 2.857334826$. Here $\tau/\gamma = k_2$ and $c = (k_2+1)/(2k_2+1)$. So this natural composite probability density has only two unknown parameters $\theta > 0$, and $\gamma > 0$. It should be mentioned here that well-known distributions such as Normal, Gamma, inverse Gaussian, Birnbaum-Saunders, etc.,

do not produce a simplified composite distribution with Pareto distribution like the WPC distribution given above.

The cumulative distribution function $(F(x))$, hazard function $(h(x))$, and the quantile function $(Q(u))$ are, respectively, given by

$$F(x) = \begin{cases} \left(\frac{k_2+1}{2k_2+1}\right)\left[1-\exp\left\{-\left(\frac{k_2+1}{k_2}\right)\left(\frac{x}{\theta}\right)^{\gamma k_2}\right\}\right] & \text{if} \quad 0 \le x \le \theta \\ 1-\left(\frac{k_2+1}{2k_2+1}\right)\left(\frac{\theta}{x}\right)^{\gamma} & \text{if} \quad \theta \le x < \infty \end{cases}, \tag{2.52}$$

$$h(x) = \begin{cases} (k_2+1)\left(\frac{\gamma}{x}\right)\left(\frac{x}{\theta}\right)^{\gamma k_2}\left[1+\left(\frac{k_2}{k_2+1}\right)e^{\left(\frac{k_2+1}{k_2}\right)\left(\frac{x}{\theta}\right)^{\gamma k_2}}\right]^{-1} & \text{if} \quad 0 \le x \le \theta \\ \frac{\gamma}{x} & \text{if} \quad \theta \le x < \infty \end{cases}, \tag{2.53}$$

and

$$Q(u) = \begin{cases} \theta\left[\left(\frac{-k_2}{k_2+1}\right)\ln\left\{1-\left(\frac{2k_2+1}{k_2+1}\right)u\right\}\right]^{\frac{1}{\gamma k_2}} & \text{if} \quad 0 \le u \le k_2/(2k_2+1) \\ \theta\left\{(1-u)\left(\frac{2k_2+1}{k_2+1}\right)\right\}^{-1/\gamma} & \text{if} \quad k_2/(2k_2+1) \le u < 1 \end{cases}. \tag{2.54}$$

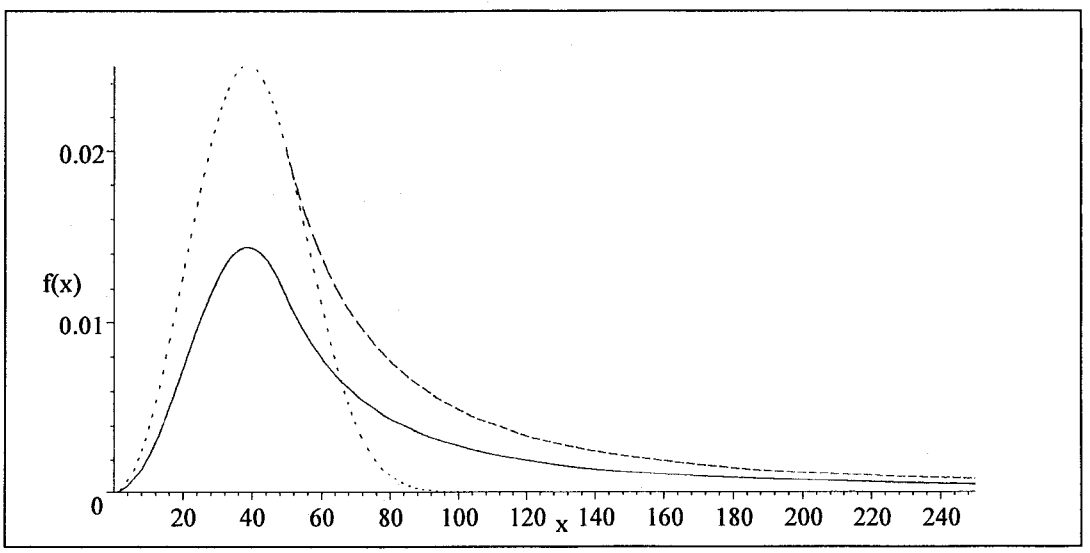

Figure 2.10 Weibull (dotted line), Pareto (dashed line), and WPC (solid line) density curves ($\gamma = 1$, $\theta = 50$).

Figure 2.10 illustrates the shape of the pdf of the WPC distribution, the Weibull distribution, and the Pareto distribution. In this figure, the two densities (Weibull and Pareto) are joining at $\theta = 50$ to make the WPC density. This composite density has a positive mode at $\theta \left\{ \frac{\gamma k_2 - 1}{\gamma(k_2+1)} \right\}^{\frac{1}{\gamma k_2}}$, $\gamma k_2 > 1$, and a thicker tail than the Weibull density.

In Figure 2.10, the dotted line, the dashed line, and the solid line indicate Weibull, Pareto, and WPC distributions respectively. From Figure 2.10, one may clearly see that the WPC density does not fade away to zero too quickly like the Weibull.

The shapes of the hazard function given in equation (2.53) of the WPC distribution is bounded by the parameter space $\gamma > 0$ such that monotone decreasing $(\gamma < 1/k_2)$, and unimodal shape $(\gamma \geq 1/k_2)$. In addition, the peak failure time, t_p can be obtained as a solution of the following equation.

$$(1 - \gamma k_2) \exp\left\{ - \left(\frac{k_2 + 1}{k_2} \right) \left(\frac{t_p}{\theta} \right)^{\gamma k_2} \right\} + \gamma k_2 \left(\frac{t_p}{\theta} \right)^{\gamma k_2} + (1 - \gamma k_2) \left(\frac{k_2}{k_2 + 1} \right) = 0. \tag{2.55}$$

Furthermore, the t^{th} moment, $E(X^t)$ of the composite family for $t < \gamma$ is given by

$$E(X^t) = \theta^t \left(\frac{k_2 + 1}{2k_2 + 1} \right) \left\{ \left(\frac{k_2}{k_2 + 1} \right)^{\frac{t}{\gamma k_2}} \Gamma\left(1 + \frac{t}{\gamma k_2} \right) \Gamma\left(1 + \frac{t}{\gamma k_2}; \frac{k_2 + 1}{k_2} \right) + \frac{\gamma}{\gamma - t} \right\}. \tag{2.56}$$

Here, $\Gamma(.)$ and $\Gamma(.;.)$ are complete and incomplete gamma functions such that, $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt; x > 0$, $\Gamma(x;y) = \frac{1}{\Gamma(x)} \int_0^y e^{-t} t^{x-1} dt$; $x, y > 0$. The t^{th} moment does not exist for $t \geq \gamma$.

The density and hazard surface for scale parameter $\theta = 1$ are respectively given in Figure 2.11 and Figure 2.12 to illustrate the shape variation with respect to the shape parameter γ of the distribution.

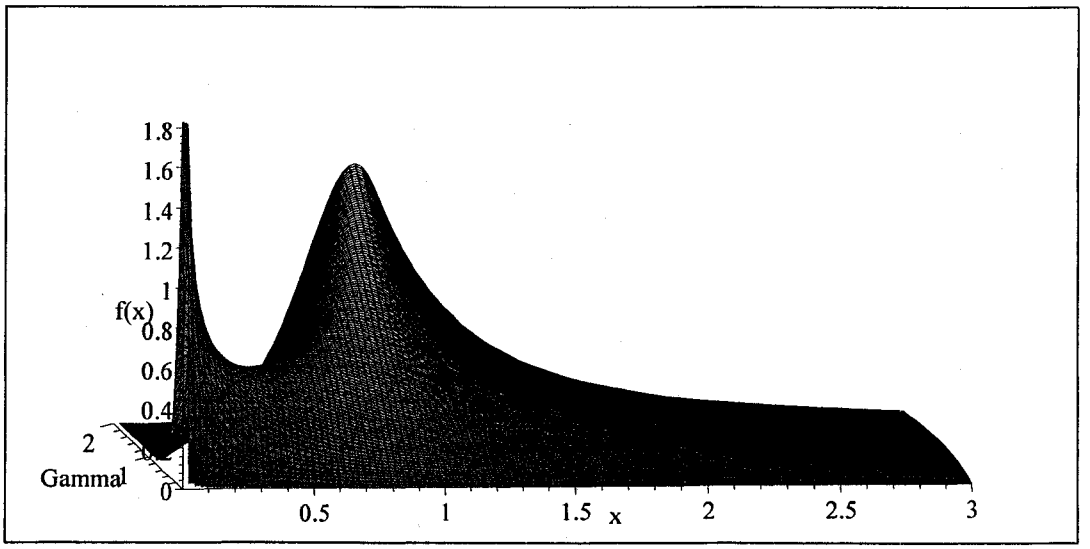

Figure 2.11 Density surface of the WPC distribution for $\theta = 1$.

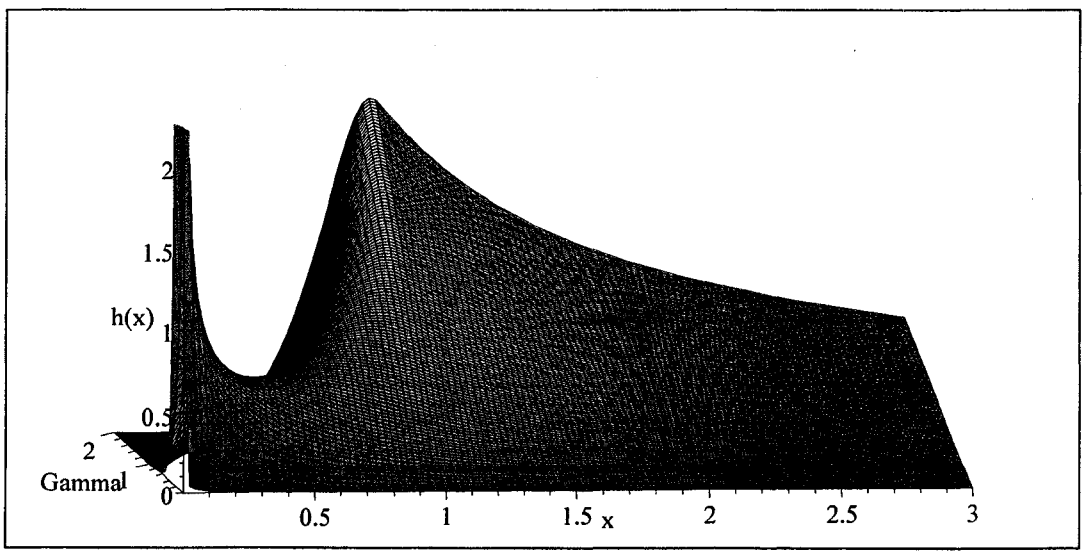

Figure 2.12 Hazard surface of the WPC distribution for $\theta = 1$.

Proof of the composite density given in equation (2.51)*:*

For the density given in equation (2.47), once we impose the continuity and differentiability conditions given in equation (2.50), we get $1+\frac{\tau}{\gamma}=\exp\left(1+\frac{\gamma}{\tau}\right)=\left(\frac{\theta}{\phi}\right)^{\tau}$. Since $\int_0^{\infty} f(x)\,dx=1$, we get $c\left(\int_0^{\theta} f_1(x)\,dx\ +1\right)=1$, here

$$\int_0^{\theta} f_1(x)\,dx=\int_0^{\theta}(\tau/x)(x/\phi)^{\tau}\exp\{-(x/\phi)^{\tau}\}\,dx=1-\exp\left\{-\left(\frac{\theta}{\phi}\right)^{\tau}\right\}=\tau/(\gamma+\tau).$$

Therefore, $c=(\gamma+\tau)/(\gamma+2\tau)$ which yields the equation (2.51).

2.11.3 Parameter estimation under the least square method

Let $X_1, X_2, ..., X_n$ be a random sample from the WPC distribution given in equation (2.51). Suppose the unknown parameter θ is in between the m^{th} observation and $m+1^{\text{th}}$ observation. Therefore, it is reasonable to assume that this is an ordered random sample, i.e., $x_1 \le x_2 \le x_3 \le ....x_m \le \theta \le x_{m+1} \le ... \le x_n$. Then the least square estimators (Gujarati 2002) can be calculated from the following linear form which is obtained by taking the natural logarithm of the quantile function given in equation (2.54) and replacing the cumulative probability with $(i-0.5)/n$. Where n is the total number of data points and $i=1,2,3,...,n$.

$$Y_i=\begin{cases} aX_{1i}+b & \text{if } Y_i \le b \\ aX_{2i}+b & \text{if } b \le Y_i \end{cases}, \tag{2.57}$$

where $X_{1i}=\frac{1}{k_2}\ln\left[\left(\frac{-k_2}{1+k_2}\right)\ln\left\{1-\left(\frac{2k_2+1}{k_2+1}\right)\left(\frac{i-0.5}{n}\right)\right\}\right]$, $X_{2i}=-\ln\left\{\left(\frac{2k_2+1}{k_2+1}\right)\left(\frac{n-i+0.5}{n}\right)\right\}$, $a=1/\gamma$, $b=\ln\theta$, $Y_i=\ln x_i$, $i=1,2,3,...,n$.

An algorithm to evaluate least square (LS) estimators

Step 1: calculate $\widehat{\gamma_m}$ and $\widehat{\theta_m}$ from the following equations for $m=1$ $(m=1,2,...)$

$$\widehat{\gamma_m} = 1/\widehat{a_m},\ \widehat{\theta_m} = \exp(\widehat{b_m}). \tag{2.58}$$

Here $\widehat{a_m}$ and $\widehat{b_m}$ are evaluated from the following equations.

$$\widehat{a_m} = \frac{\sum_{i=1}^{m}\left(X_{1i}-\overline{X}\right)\left(Y_i-\overline{Y}\right)+\sum_{i=m+1}^{n}\left(X_{2i}-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sum_{i=1}^{m}\left(X_{1i}-\overline{X}\right)^2+\sum_{i=m+1}^{n}\left(X_{2i}-\overline{X}\right)^2},\ \widehat{b_m} = \overline{Y}-\widehat{a_m}\overline{X}. \tag{2.59}$$

Where $\overline{X} = \left(\sum_{i=1}^{m} X_{1i} + \sum_{i=m+1}^{n} X_{2i}\right)/n$, and $\overline{Y} = \sum_{i=1}^{n} Y_i$.

Step 2: If $\widehat{\theta}_m \in [x_m,\ x_{m+1}]$, then the least square estimators of γ and θ are

$$\widehat{\gamma}_{\mathrm{LS}} = \widehat{\gamma}_m,\ \widehat{\theta}_{\mathrm{LS}} = \widehat{\theta}_m, \tag{2.60}$$

otherwise, repeat the step 1 for the next m value. A unique value for m such that $x_m \le \widehat{\theta}_{\mathrm{LS}} \le x_{m+1}$ can be obtained from this algorithm.

Approximate conditional standard errors (Se) of $\widehat{\gamma}_{\mathrm{LS}}$ and $\widehat{\theta}_{\mathrm{LS}}$ can be obtained from the following equations which are derived from the variance-covariance matrix of $\widehat{a_m}$ and $\widehat{b_m}$ for a unique m by applying the delta method.

$$Se(\widehat{\theta_m}) = \exp(\widehat{b_m})Se(\widehat{b_m}),\ Se(\widehat{\gamma_m}) = \widehat{a_m}^{-2}Se(\widehat{a_m}). \tag{2.61}$$

Where $Se(\widehat{b_m})$ and $Se(\widehat{a_m})$ are standard errors of the estimators $\widehat{a_m}$ and $\widehat{b_m}$. Note that these standard errors are conditional on $x_m \le \widehat{\theta}_{\mathrm{LS}} \le x_{m+1}$.

One can easily use a quantile-quantile (Q-Q) plot to assess the assumption of WPC for a given uncensored point data set. The Q-Q plot can be obtained by plotting the pairs of ordered observations $(X_{1i},\ Y_i)$ for $i = 1, ..., m$ and $(X_{2i},\ Y_i)$ for

$i = m+1, ..., n$. When the points lie very nearly along a straight line, the WPC assumption remains tenable. The straightness of the Q-Q plot can be measured by calculating the correlation coefficient of the points in the plot. The correlation coefficient for the Q-Q plot is defined by

$$r_Q = \frac{\sum_{i=1}^{m}\left(X_{1i}-\overline{X}\right)\left(Y_i-\overline{Y}\right)+\sum_{i=m+1}^{n}\left(X_{2i}-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{m}\left(X_{1i}-\overline{X}\right)^2+\sum_{i=m+1}^{n}\left(X_{2i}-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}. \tag{2.62}$$

Table 2.10 Critical points for the correlation coefficient test of WPC

Sample size	Significance levels α				Sample size	Significance levels α			
n	.01	.05	.10	.20	n	.01	.05	.10	.20
5	.776	.848	.884	.913	45	.924	.952	.962	.972
10	.838	.890	.912	.934	50	.929	.956	.965	.974
15	.869	.911	.929	.946	55	.933	.958	.967	.975
20	.886	.924	.939	.954	60	.935	.960	.969	.977
25	.897	.932	.946	.959	75	.944	.966	.973	.980
30	.907	.939	.952	.964	100	.952	.972	.978	.987
35	.914	.945	.956	.967	200	.970	.982	.986	.990
40	.920	.949	.960	.970	300	.977	.987	.990	.992

Formally, we reject the hypothesis of the assumption of WPC at level of significance α if r_Q falls below the appropriate value in Table 2.10. Note that a single entry of this table are obtained by simulating a 100,000 random samples from the WPC distribution with given sample size n.

2.11.4 Parameter estimation under the likelihood method

Let $X_1, X_2, ..., X_n$ be a random sample from the WPC distribution given in equation (2.51). As before by assuming the ordered random sample, i.e., $x_1 \leq x_2 \leq x_3 \leq ....x_m \leq \theta \leq x_{m+1} \leq ... \leq x_n$, the log-likelihood function can be written as

$$\begin{aligned} \ln L(\gamma, \theta) &= (m+n)\ln(k_2+1) - (1+\gamma)\textstyle\sum_{i=1}^{n} \ln x_i \\ &\quad + (k_2+1)\gamma \textstyle\sum_{i=1}^{m} \ln x_i + n\ln\gamma - n\ln(2k_2+1) \\ &\quad - \left(\frac{k_2+1}{k_2}\right)\textstyle\sum_{i=1}^{m}\left(\frac{x_i}{\theta}\right)^{\gamma k_2} + \gamma(n - m(k_2+1))\ln\theta. \end{aligned} \quad (2.63)$$

The maximum likelihood (ML) estimators of γ and θ, $\widehat{\gamma}_{\mathrm{ML}}$ and $\widehat{\theta}_{\mathrm{ML}}$ can, respectively, be obtained by maximizing the log-likelihood function ($l(\boldsymbol{\theta}) = \ln L(\gamma, \theta)$). In this case, the log-likelihood function is maximized by solving the score equation $U(\boldsymbol{\theta}) = \frac{\partial l(\theta)}{\partial \theta} = 0$. For large samples, asymptotic normality results hold for the estimated parameters values, i.e., $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} \mathbf{N}_2(\mathbf{0}, i^{-1}(\boldsymbol{\theta}))$, where $\mathbf{N}_2$ denote the bivariate normal distribution and $i(\boldsymbol{\theta})$ is the observed information matrix (Efron and Hinkley 1978) of $\widehat{\boldsymbol{\theta}}$ such that $i(\boldsymbol{\theta}) = -\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}$. The following algorithm provides an easy and straightforward way to compute the maximum likelihood estimators.

An algorithm to evaluate ML estimators

Step 1: calculate $\widehat{\gamma}_m$ from the following nonlinear equation for $m = a+1$ ($m = a+1, a+2, ...$; where $a = \lfloor n/(k_2+1) \rfloor$).

$$(n - m(k_2+1))\left(\frac{\sum_{i=1}^{m} x_i^{\widehat{\gamma}_m k_2} \ln x_i}{\sum_{i=1}^{m} x_i^{\widehat{\gamma}_m k_2}}\right) + (k_2+1)\textstyle\sum_{i=1}^{m}\ln x_i - \sum_{i=1}^{n}\ln x_i + (n/\widehat{\gamma}_m) = 0. \quad (2.64)$$

The corresponding values of θ, $\widehat{\theta}_m$, can be obtained from

$$\widehat{\theta}_m = \left((k_2+1) \sum_{i=1}^{m} x_i^{\widehat{\gamma}_m k_2} / (m(k_2+1) - n) \right)^{1/(\widehat{\gamma}_m k_2)}. \tag{2.65}$$

Step 2: If $\widehat{\theta}_m \in [x_m,\ x_{m+1}]$, then the ML estimators of γ and θ are

$$\widehat{\gamma}_{\mathrm{ML}} = \widehat{\gamma}_m,\ \widehat{\theta}_{\mathrm{ML}} = \widehat{\theta}_m, \tag{2.66}$$

otherwise, repeat the step 1 for the next m value. As before, a unique value for m such that $x_m \leq \widehat{\theta}_{\mathrm{ML}} \leq x_{m+1}$ can be obtained from this algorithm. Furthermore, one can observed these estimators are such that $\widehat{\theta}_{\mathrm{ML}} = \min_m \left(\widehat{\theta}_m \right)$ and $\widehat{\gamma}_{\mathrm{ML}} = \max_m \left(\widehat{\gamma}_m \right)$. However, the corresponding log-likelihood value, $l(\widehat{\boldsymbol{\theta}})$, is not equal to $\max_m \ln L(\widehat{\gamma}_m, \widehat{\theta}_m)$. But, it is equal to $\ln L(\widehat{\gamma}_{\mathrm{ML}}, \widehat{\theta}_{\mathrm{ML}})$. Hence, the associated asymptotic standard errors are conditional on $x_m \leq \widehat{\theta}_{\mathrm{ML}} \leq x_{m+1}$ for a unique m value. Therefore, the calculated asymptotic standard errors given in Section 2.11.7 using the observed information matrix, $i(\boldsymbol{\theta})$, are conditional standard errors.

2.11.5 Approximate coverage probabilities for ML estimators

The coverage probabilities for the maximum likelihood estimation method with intended confidence levels $\alpha = 0.1$ and $\alpha = 0.05$ are given in Table 2.11. These coverage probabilities are based on 10,000 simulated random samples from the density given in equation (2.51). The random samples are generated by plugging the known values of parameters γ and θ (say $\gamma = 0.2$, $\theta = 10$) to the quantile function given in equation (2.54). In addition, n (say $n = 10$) number of ordered uniform random samples from the uniform distribution, $u \sim U(0,1)$ is required to substitute as u in equation (2.54). In that way, one random sample with size n (say $n = 10$) from the WPC distribution with parameters γ and θ (say $\gamma = 0.2$, $\theta = 10$) can be generated.

Table 2.11 Approximate coverage probabilities of WPC

	$n =$	10			20			50		
90% intended	$\theta =$	10	20	50	10	20	50	10	20	50
$\gamma = 0.2$	γ :	.914	.913	.911	.909	.906	.902	.904	.906	.908
	θ :	.775	.782	.780	.829	.828	.829	.870	.872	.867
$\gamma = 0.5$	γ :	.911	.909	.906	.907	.902	.902	.900	.901	.908
	θ :	.828	.821	.824	.866	.860	.869	.892	.886	.887
$\gamma = 1.0$	γ :	.915	.912	.915	.907	.903	.901	.907	.904	.904
	θ :	.823	.831	.827	.871	.863	.865	.887	.890	.888
$\gamma = 2.0$	γ :	.910	.915	.910	.906	.907	.908	.902	.901	.905
	θ :	.823	.827	.832	.866	.859	.868	.882	.883	.887
95% intended										
$\gamma = 0.2$	γ :	.959	.962	.962	.955	.957	.953	.951	.955	.954
	θ :	.804	.807	.811	.855	.857	.856	.899	.901	.897
$\gamma = 0.5$	γ :	.961	.958	.959	.957	.953	.953	.949	.954	.954
	θ :	.872	.863	.869	.910	.903	.910	.935	.931	.931
$\gamma = 1.0$	γ :	.964	.959	.965	.957	.953	.953	.952	.952	.949
	θ :	.879	.880	.879	.919	.917	.913	.936	.940	.939
$\gamma = 2.0$	γ :	.963	.964	.962	.955	.957	.959	.951	.951	.955
	θ :	.880	.884	.885	.920	.915	.922	.936	.934	.939

In this simulation study, ten thousand such samples are generated to get a single cell value in Table 2.11. For this purpose, the subroutine ZBREN in the IMSL (1991) package is used to solve the nonlinear equation given in (2.64). The approximate $100\,(1-\alpha)\,\%$ confidence intervals for parameters, γ and θ are calculated by using $\left(\widehat{\gamma} - Z_{\alpha/2}SE_{\widehat{\gamma}},\ \widehat{\gamma} + Z_{\alpha/2}SE_{\widehat{\gamma}}\right)$ and $\left(\widehat{\theta} - Z_{\alpha/2}SE_{\widehat{\theta}},\ \widehat{\theta} + Z_{\alpha/2}SE_{\widehat{\theta}}\right)$ respectively. Where $SE_{\widehat{\gamma}}$ and $SE_{\widehat{\theta}}$ are asymptotic standard errors of $\widehat{\gamma}$ and $\widehat{\theta}$ respectively.

From Table 2.11, one can clearly see that when the sample size increases, the approximate coverage probabilities for the parameters under the maximum likelihood method are getting closer to the intended coverage probabilities. For small samples, the coverage probabilities of parameter θ are somewhat lower than the intended level. But, one can obtain a desired confidence level by appropriately adjusting the confidence coefficient α. Moreover, the procedure gives a slight over coverage for parameter γ for small samples.

2.11.6 The ML estimation for Type I right censored data

In order to model the Type I right censored data, using the WPC distribution, one can extend the log-likelihood function given in equation (2.63) by introducing an indicator variable δ_i such that

$$\delta_i = \begin{cases} 0 & \text{if } i^{th} \text{ observation is right-censored} \\ 1 & \text{if } i^{th} \text{ observation is not right-censored} \end{cases}, \quad i = 1, 2, ..., n. \tag{2.67}$$

Then the log-likelihood function for the censored sample, $x_1 \leq x_2 \leq x_3 \leq ....x_m \leq \theta \leq x_{m+1} \leq ... \leq x_n$, is

$$\begin{aligned}\ln L(\gamma,\theta) &= \sum_{i=1}^{m}(1-\delta_i)\ln\left[k_2+(k_2+1)\exp\left\{-\left(\frac{k_2+1}{k_2}\right)(x_i/\theta)^{\gamma k_2}\right\}\right] \\ &+\sum_{i=m+1}^{n}\ln\left\{(1+k_2)(\theta/x_i)^{\gamma}\right\}+\sum_{i=1}^{n}\delta_i\ln(\gamma/x_i)-n\ln(2k_2+1) \\ &+\sum_{i=1}^{m}\delta_i\left[\ln\left\{(1+k_2)^2(x_i/\theta)^{\gamma k_2}\right\}-\left(\frac{k_2+1}{k_2}\right)(x_i/\theta)^{\gamma k_2}\right]. \quad (2.68)\end{aligned}$$

The parameter estimation procedure is quite similar to the algorithm given in Section 2.11.4. As before, a unique value for m can be obtained such that $x_m \leq \widehat{\theta}_{\mathrm{ML}} \leq x_{m+1}$. However, due to the censoring of the data set, one cannot observed the estimators are such that $\widehat{\theta}_{\mathrm{ML}} = \min_m\left(\widehat{\theta}_m\right)$ and $\widehat{\gamma}_{\mathrm{ML}} = \max_m\left(\widehat{\gamma}_m\right)$.

In the analysis of numerical examples, especially for censored data, one can easily employ the LE program in BMDP (1992) to solve the nonlinear equations, $U(\boldsymbol{\theta}) = 0$ for the purpose of estimate the parameters. The LE routine also gives the asymptotic conditional standard errors (SE's) of the estimates by inverting the Hessian matrix (Observed information matrix) used in the maximization of the likelihood function.

2.11.7 Illustrative examples

The object of this section, as given in Section 2.10, is to illustrate the use of WPC distribution and to demonstrate its applicability with the aid of real life data. In this regard, three distinctly different examples are presented based on well-known data, which were published in the statistics literature. Specifically, the first example and the rest respectively consider complete and right-censored data. For comparison purposes, the loglogistic (LLG) distribution ($F(x;\varphi,\delta) = 1/(1+(\varphi/x)^{\delta})$; $0 < x$, $0 < \varphi$, $0 < \delta$), which is often used in biomedical area and the embedded Burr (EB)

family (Mudholkar *et al.* 1996), $F(x;\sigma,\beta,\lambda) = 1-1/(1+\lambda(x/\sigma)^{\beta})^{1/\lambda}; 0<\sigma, 0<\beta)$, which is a three-parameter extension of the LLG distribution, are considered. Note that the range of the EB random variable x is $(0,\infty)$ for $\lambda \geqslant 0$ and $(0,\sigma/(-\lambda)^{1/\beta})$ for $\lambda < 0$. Note that other composite parametric families, such as LPC and LLPC (loglogistic-Pareto composite) gives very poor fit to the following data sets, hence we disregard to present such analysis to reduce the length of this section.

Example 1. Guinea pigs survival time data

This example (see the Appendix B for the data set) is abstracted from Bjerkedal (1960) represents the survival times in days of guinea pigs after infected with virulent tubercle bacilli.

In this example, guinea pigs survival time data is analyzed by using the three models; LLG, WPC, and EB distributions. The expected deaths, E_j, $j = 1,2,...,11$, the estimated parameter values, the log-likelihood ($l(\widehat{\boldsymbol{\theta}})$) values, the chi-squared values, and the corresponding p-values are given in Table 2.12. Estimated values given in this table are obtained by using the maximum likelihood method. Note that the chi-squared test has been performed by dividing the survival times into 11 classes with upper limits 65, 75, 85, 95, 105, 115, 130, 150, 200, 300, ∞. For our convenience, the number of classes is obtained by the formula (D'Agostino and Stephens 1986), $M \approx 2n^{2/5}$, where M and n are, respectively, the number of classes and the sample size.

Table 2.12 Estimated values of three models for guinea pigs data

Time interval (in days)		Observed deaths O_j	Expected deaths E_j		
			LLG	WPC	EB
0	65	7	10.1202	5.8134	6.7870
65	75	4	4.9156	5.5120	6.0610
75	85	10	5.5793	7.6829	7.3846
85	95	6	5.8731	8.5915	7.4753
95	105	11	5.8081	7.4037	6.7267
105	115	5	5.4629	5.6706	5.6858
115	130	5	7.1917	6.2924	6.6793
130	150	6	7.4949	5.7645	6.2649
150	200	8	10.5776	7.8834	8.4318
200	300	4	6.4282	5.9618	5.9639
300	up	6	2.5484	5.4238	4.5398
Parameters $\pm$ SE			$\widehat{\delta} = 3.3450$	$\widehat{\gamma} = 1.8289$	$\widehat{\beta} = 5.8142$
			$SE_{\widehat{\delta}} = 0.3348$	$SE_{\widehat{\gamma}} = 0.2091$	$SE_{\widehat{\beta}} = 1.2229$
			$\widehat{\varphi} = 111.69$	$\widehat{\theta} = 98.789$	$\widehat{\lambda} = 2.8055$
			$SE_{\widehat{\varphi}} = 6.7602$	$SE_{\widehat{\theta}} = 5.6042$	$SE_{\widehat{\lambda}} = 0.8884$
					$\widehat{\sigma} = 94.414$
					$SE_{\widehat{\sigma}} = 6.9813$
$l(\widehat{\boldsymbol{\theta}})$			-401.575	-396.937	-397.260
χ^2_{df}			$\chi^2_6 = 8.5266$	$\chi^2_8 = 4.9472$	$\chi^2_6 = 5.2021$
p-value			0.2020	0.7632	0.5182

The estimated values indicate that the WPC distribution gives a very good fit to the guinea pigs survival time data. Furthermore, the adequacy of the fit is further strengthened by illustrating the survival function of the WPC distribution along with the Kaplan-Meier curve (see Figure 2.13). In order to compare, the fitted survival curves of the LLG and EB distributions are included in Figure 2.13. In addition, the least square estimators for this data set using the equations (2.58) and (2.59) are $\widehat{\gamma}_{\text{LS}} \pm SE_{\widehat{\gamma}_{\text{LS}}} = 1.9176 \pm 0.0337$, $\widehat{\theta}_{\text{LS}} \pm SE_{\widehat{\theta}_{\text{LS}}} = 97.3436 \pm 0.9963$. Also, the estimated correlation coefficient using the equation (2.60), $\widehat{r_Q} = 0.9894$, is well above the corresponding table value (see Table 2.10) for 20% significance level at sample size 75. Therefore the associated Q-Q plot, which have not been plotted to reduce the length of the chapter, can be assume as nearly a straight line. Hence the WPC distribution gives a better fit to the guinea pigs survival time data.

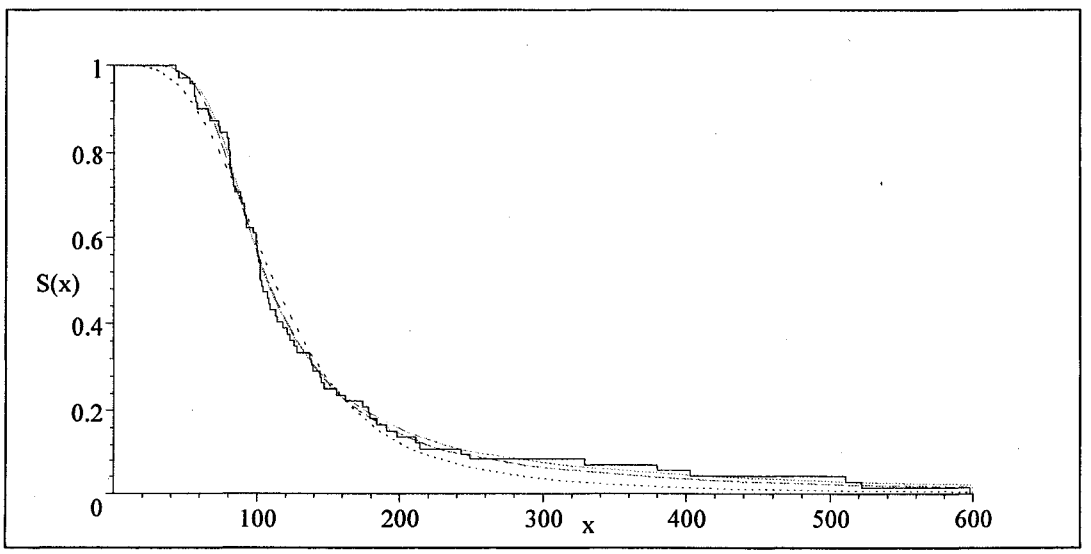

Figure 2.13. Fitted survival curves for guinea pigs data. Kaplan-Meier survival curve (step function), LLG (dotted line), WPC (solid line), EB(dashed line).

Example 2. Arm A head and neck cancer data

This example (see the Appendix B for the data set) represents the survival times in days of head and neck cancer patients after a treatment considered earlier by Efron (1988). This clinical trial data consist of 51 patients with radiation therapy alone denoted by arm A. Nine patients were lost to follow-up and were regarded as right censored.

This data set is analyzed by Mudholkar *et al.* (1996) using the EB distribution. In this example, the WPC distribution is used to reanalyze the arm A clinical trial data to demonstrate and illustrate its flexibility towards modeling the unimodal failure rate data.

Using the estimation procedure describe above, it is easy to fit the WPC distribution to arm A data. The expected deaths, the estimated parameter values, the log-likelihood ($l(\widehat{\boldsymbol{\theta}})$) value, the chi-squared value, and the corresponding p-value are given in Table 2.13. For comparison purposes, the estimated values of the LLG and EB models are also included in Table 2.13. Note that these estimated values are calculated before converting to the months of survival times.

The p-value for this right censored data is based on discretized method introduced by Efron (1988). In order to calculate the p-value, same discretization given by Efron (1988) and Mudholkar *et al.* (1996) is used in Table 2.13. The data given in this table includes the signed deviance residuals, R_j given by the formula (McCullagh and Nelder 1998),

$$R_j \equiv \sqrt{2} sign(S_j - E_j) \left\{ S_j \ln \frac{S_j}{E_j} + (N_j - S_j) \ln \frac{N_j - S_j}{N_j - E_j} \right\}^{1/2}, \tag{2.69}$$

to the fitted hazard functions of three models for arm A clinical trial data. Where, N_j- total no. of patients at risk at the beginning of each interval j, $j = 1, ..., 13$ for arm A data, S_j- observed death at the end of each interval, E_j- expected death at the end of each interval, for details/notations behind the procedure see Efron (1988).

The p-values given in Table 2.13 indicate that arm A clinical trial data fits better with the WPC distribution. As in the previous example, the fitness is further strengthened by illustrating the survival curve of the WPC distribution along with the Kaplan-Meier curve (see Figure 2.14). In order to compare the fitness by graphically, the fitted survival curves of the LLG and EB families are also included in Figure 2.14.

Note that, commonly available unimodal failure rate parametric distributions such as lognormal, loglogistic, Birnbaum-Saunders, inverse Gaussian, Pareto, Weibull, etc., are inappropriate for head and neck cancer high-risk (hump-shaped) failure data.

Table 2.13 Estimated values of three models for Arm A clinical data

Time interval	At risk	Dead	Expected deaths E_j		
(in Months)	N_j	S_j	LLG	WPC	EB
0-1	51	1	2.1579	0.8205	1.1041
1-2	50	2	3.7516	2.7763	3.3181
2-3	48	5	4.3510	4.5631	4.7337
3-4	42	2	4.1379	5.2975	4.8086
4-6	72	15	7.3402	10.0243	8.4241
6-8	49	3	4.8815	5.3995	5.1926
8-11	56	4	5.1725	4.5899	5.0092
11-14	45	3	3.6909	2.7672	3.2540
14-18	45	2	3.2224	2.1675	2.6454
18-24	46	2	2.7618	1.7016	2.1276
24-31	49	0	2.3790	1.3704	1.7393
31-38	47	2	1.8940	1.0494	1.3413
38-47	28	1	0.9489	0.5127	0.6578

(continued)

Table 2.13 (continued)

			$\widehat{\beta} = 2.1285$
	$\widehat{\delta} = 1.5265$	$\widehat{\gamma} = 0.7650$	$SE_{\widehat{\beta}} = 0.5273$
	$SE_{\widehat{\delta}} = 0.1977$	$SE_{\widehat{\gamma}} = 0.1127$	$\widehat{\lambda} = 2.1522$
	$\widehat{\varphi} = 238.34$	$\widehat{\theta} = 178.75$	$SE_{\widehat{\lambda}} = 0.9365$
	$SE_{\widehat{\varphi}} = 38.453$	$SE_{\widehat{\theta}} = 21.204$	$\widehat{\sigma} = 182.26$
			$SE_{\widehat{\sigma}} = 38.886$
$l(\widehat{\boldsymbol{\theta}})$	-292.380	-290.383	-291.263
$\sum R_j^2$, df	17.603, 11	11.295, 11	13.493, 10
p-value	0.0913	0.4189	0.1974

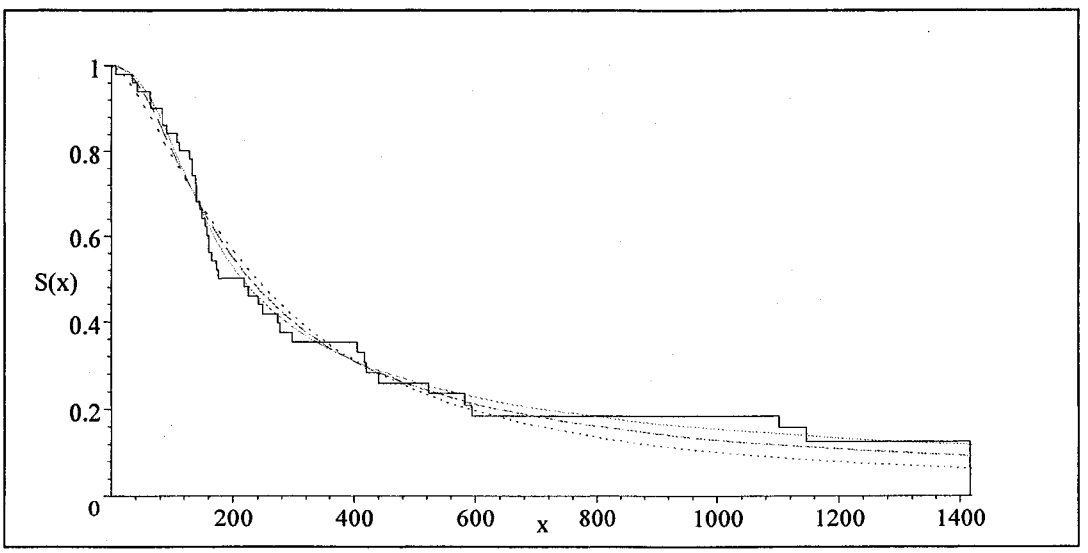

Figure 2.14. Fitted survival curves for Arm A cancer data. Kaplan-Meier survival curve (step function), LLG (dotted line), WPC (solid line), EB (dashed line).

Example 3. Nasopharynx cancer survival data

The data set (data are given in the appendix B) of this example is taken from McKeague (2000) and given by West (1987, 1992) who studied the data on 181 nasopharynx cancer patients. Their cancer careers, culminating in either death (127 cases) or censoring (54 cases), are recorded to the nearest month, ranging from 1 to 177 months. Our analysis is restricted to these two variables, even though the data set contains several covariates. The WPC distribution is used to analyze the nasopharynx data set for further strengthen its flexibility in the analysis of unimodal failure rate data. As before, the estimated values of the three models are given in Table 2.15. Furthermore, fitted survival curves are illustrated in Figure 2.15 along with the Kaplan-Meier curve.

Table 2.14 Estimated values of three models for cancer data

Time interval	At risk	Dead	Expected deaths E_j		
(in Years)	N_j	S_j	LLG	WPC	EB
0-0.5	181	19	24.8687	18.9620	19.9235
0.5-1	160	35	27.5625	33.6312	34.4229
1-1.5	112	17	19.0157	24.9446	22.3195
1.5-2	93	16	14.8340	16.3005	15.3688
2-2.5	72	11	10.6288	9.7891	9.8764
2.5-3	60	9	8.1670	6.6652	6.9629
3-3.5	50	1	6.2805	4.6961	5.0052
3.5-4	48	4	5.5787	3.9053	4.2147
4-4.5	41	4	4.4242	2.9423	3.2021
4.5-5	37	4	3.7203	2.3752	2.5999
5-5.5	32	3	3.0088	1.8583	2.0426
5.5-6	23	0	2.0292	1.2194	1.3445
6-6.5	22	0	1.8271	1.0729	1.1859
6.5-7	21	0	1.6466	0.9482	1.0500
7-7.5	21	0	1.5588	0.8828	0.9790
7.5-8	21	1	1.4795	0.8258	0.9169
8-8.5	19	1	1.2735	0.7018	0.7800
8.5-9	18	0	1.1502	0.6269	0.6973
9-9.5	16	0	0.9767	0.5271	0.5867

(continued)

Table 2.14 (continued)

9.5-10	15	0	0.8764	0.4688	0.5221
10-10.5	13	0	0.7282	0.3865	0.4306
10.5-11	13	0	0.6993	0.3685	0.4108
11-11.5	12	1	0.6208	0.3250	0.3624
11.5-12	11	0	0.5480	0.2852	0.3182
12-12.5	8	0	0.3843	0.1990	0.2220
12.5-13	6	0	0.2783	0.1434	0.1600
13-13.5	6	0	0.2690	0.1380	0.1540
13.5-14	6	1	0.2603	0.1330	0.1485
14-14.5	4	0	0.1681	0.0855	0.0955
14.5-15	1	0	0.0407	0.0207	0.0231
					$\widehat{\beta} = 2.0913$
			$\widehat{\delta} = 1.3012$	$\widehat{\gamma} = 0.6093$	$SE_{\widehat{\beta}} = 0.3177$
			$SE_{\widehat{\delta}} = .0957$	$SE_{\widehat{\gamma}} = 0.0513$	$\widehat{\lambda} = 3.0664$
			$\widehat{\varphi} = 26.147$	$\widehat{\theta} = 18.499$	$SE_{\widehat{\lambda}} = 0.8079$
			$SE_{\widehat{\varphi}} = 2.7275$	$SE_{\widehat{\theta}} = 1.8698$	$\widehat{\sigma} = 15.862$
					$SE_{\widehat{\sigma}} = 2.5183$
$l(\widehat{\boldsymbol{\theta}})$			−607.318	−600.781	−600.883
$\sum R_j^2$, df			41.708, 28	29.812, 28	28.860, 27
p-value			0.0462	0.3722	0.3677

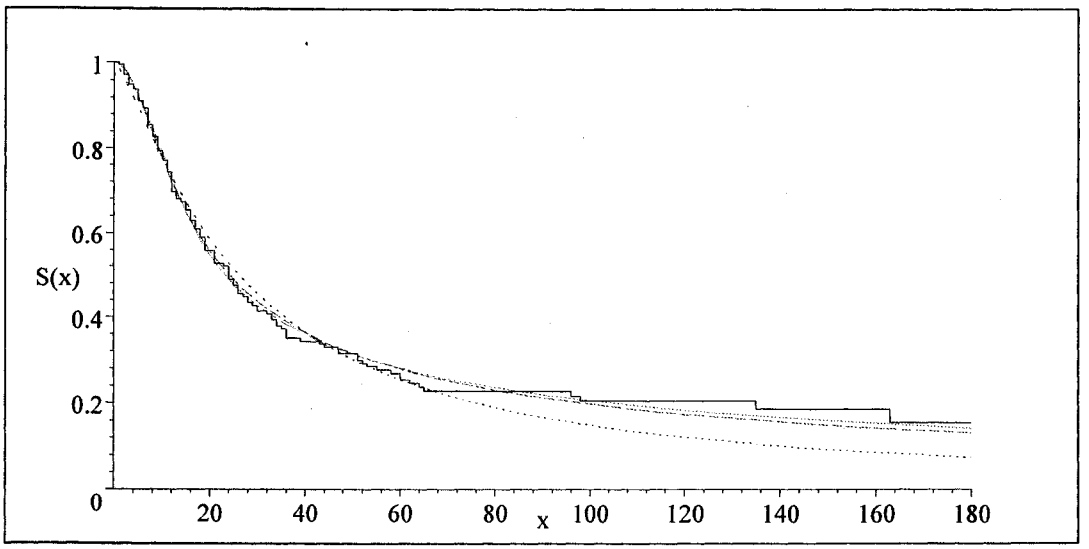

Figure 2.15. Fitted survival curves for nasopharynx data. Kaplan-Meier survival curve (step function), LLG (dotted line), WPC (solid line), EB (dashed line).

Once again, commonly available parametric families are not appropriate to analyze the nasopharynx cancer survival time data.

Example 4. Stimulus-response time data (an additional example)

The data in this example (data set is given in the appendix B) represents the reaction time of one subject in 180 trials of a psychological experiment (Whitmore 1986). In each trial the subject was asked to decide whether the distance between two dots displayed on a monitor placed 10ft away was long or short. The dots remained visible until the subject made a response. The reaction time for each trial is the length of time from stimulus to response in milliseconds.

This data was analyzed by Whitmore (1986) using inverse Gaussian and normal-gamma mixture to obtain a proper fit. The sample log-likelihood values for the inverse Gaussian, normal-gamma mixture, and truncated normal-gamma mixture are

-1186.8, -1179.5, and -1176.3 respectively.

In this example, stimulus-response time data is reanalyzed by the WPC distribution. Using the iterative procedure and the likelihood method given above, it is easy to fit the WPC distribution. The estimated parameter values and the log-likelihood ($l(\widehat{\boldsymbol{\theta}})$) value are given in Table 2.15.

Table 2.15 Estimated values of WPC for stimulus-response time data

Time	0-450	450-550	550-650	650-750	750-850
O_i	4	31	70	31	11
E_i	4.9925	32.8102	62.9321	33.7033	17.4904
Time	850-950	950-1050	1050-1150	1150-1250	
O_i	10	7	2	3	
E_i	9.8178	5.8611	3.6772	2.4034	
Time	1250-1350	1350-1450	1450-1550	1550-up	
O_i	2	0	4	5	
E_i	1.6256	1.1321	0.8084	2.7458	

$\widehat{\alpha} \pm SE_{\widehat{\alpha}}$	$\widehat{\theta} \pm SE_{\widehat{\theta}}$	$l(\widehat{\boldsymbol{\theta}})$	χ^2_6	p-value
3.8695 ± 0.27713	606.8435 ± 8.6987	-1169.5	7.6151	0.2677

The stimulus-response time data are grouped with observed frequencies (O_i), $i = 1, ..., 13$ for the purpose of obtaining the goodness-of-fit chi-squared value. Expected frequencies (E_i) in Table 2.15 show that E_1, E_8, E_9, E_{10}, E_{11}, E_{12}, and E_{13} are smaller than 5. The expected frequency, E_1 is close to 5, hence kept it as it is. The other

expected values are pooled such that (E_8, E_9), and (E_{10}, E_{11}, E_{12}, E_{13}) to get the chi-squared with 6 df, $\varkappa_6^2 = 7.6151$, with a p-value of 0.2677. This indicates that the WPC distribution gives an improved fit to the stimulus-response time data. The Figure 2.16 illustrates the fitted density curve of the WPC distribution to the response time data with its histogram.

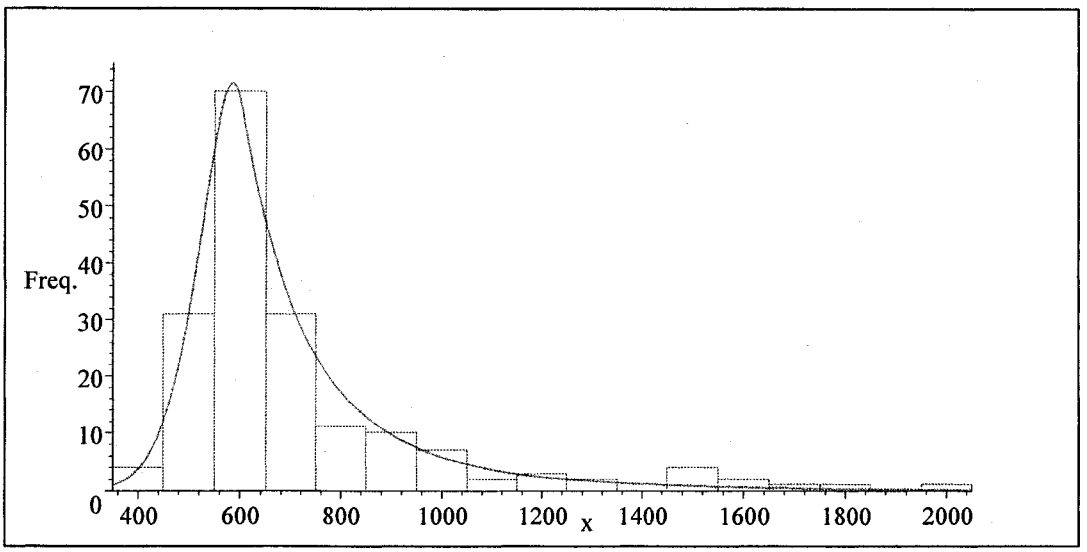

Figure 2.16. Histogram and WPC density for stimulus-response time data.

2.12 The loglogistic-Pareto composite distribution

Let X be a random variable with the pdf

$$f(x) = \begin{cases} \frac{1}{c_0} f_1(x) & \text{if } 0 < x \leq \omega \\ \frac{1}{c_0} f_2(x) & \text{if } \omega \leq x < \infty \end{cases}, \tag{2.70}$$

where c_0 is the normalizing constant, $f_1(x)$ has the form of the regular loglogistic density, and $f_2(x)$ has the form of the two-parameter Pareto density, i.e.,

$$f_1(x) = (\tau/x)(x/\lambda)^{\tau}\left[1 + (x/\lambda)^{\tau}\right]^{-2}, \; x > 0, \tag{2.71}$$

and

$$f_2(x) = (\delta/x)(\omega/x)^{\delta}, \; x \geq \omega. \tag{2.72}$$

Here δ, τ, λ, ω are unknown parameters such that $\delta > 0$, $\tau > 0$, $\lambda > 0$, $\omega > 0$.

Let us impose the continuity and differentiability conditions at ω, where it yields,

$$f_1(\omega) = f_2(\omega), \text{ and } f_1'(\omega) = f_2'(\omega). \tag{2.73}$$

Where $f'(\omega)$ is the first derivative of $f(x)$ evaluated at ω. These conditions guarantee a smooth probability density function. These two restrictions reduce the total unknown parameters from four to two. One can show that (the proof is omitted due to similarity with the LPC distribution) the loglogistic-Pareto composite (LLPC) density can be reparameterized and rewritten as

$$f(x) = \begin{cases} \frac{1}{k_3}(k_3+1)^2 \left(\frac{\delta}{x}\right)\left(\frac{x}{\omega}\right)^{\frac{1}{k_3}(k_3+1)^2\delta}\left[1+k_3\left(\frac{x}{\omega}\right)^{\frac{1}{k_3}(k_3+1)^2\delta}\right]^{-2} & \text{if} \quad 0 < x \leq \omega \\ \frac{1}{k_3}\left(\frac{\delta}{x}\right)\left(\frac{\omega}{x}\right)^{\delta} & \text{if} \quad \omega \leq x < \infty \end{cases}, \tag{2.74}$$

where $k_3 = c_0 = \left(\sqrt{5}+1\right)/2 \approx 1.618034$, which is called the golden ratio (ϕ), also known as the divine proportion, or golden mean. So this natural composite probability density has only two unknown parameters $\delta > 0$, and $\omega > 0$.

The cumulative distribution function, $F(x)$, and the quantile function, $Q(u)$, are, respectively, given by

$$F(x) = \begin{cases} \left(\frac{x}{\omega}\right)^{\frac{1}{k_3}(k_3+1)^2\delta}\left[1+k_3\left(\frac{x}{\omega}\right)^{\frac{1}{k_3}(k_3+1)^2\delta}\right]^{-1} & \text{if} \quad 0 \leq x \leq \omega \\ 1-\frac{1}{k_3}\left(\frac{\omega}{x}\right)^{\delta} & \text{if} \quad \omega \leq x < \infty \end{cases}, \tag{2.75}$$

and

$$Q(u) = \begin{cases} \omega\left(\frac{u}{1-k_3 u}\right)^{\frac{k_3}{(1+k_3)^2\delta}} & \text{if} \quad 0 \le u \le 1/k_3^2 \\ \omega\left(\frac{1}{k_3(1-u)}\right)^{1/\delta} & \text{if} \quad 1/k_3^2 \le u < 1 \end{cases} . \tag{2.76}$$

The following three curves in Figure 2.17 demonstrate the shape of the pdf of the LLPC distribution, the loglogistic distribution, and the Pareto distribution. Here, the two densities, loglogistic and Pareto, are joined at $\omega = 50$ to make the LLPC density. The variation of the LLPC density with parameter δ and parameter ω is, respectively, given in Figure 2.18 and Figure 2.19.

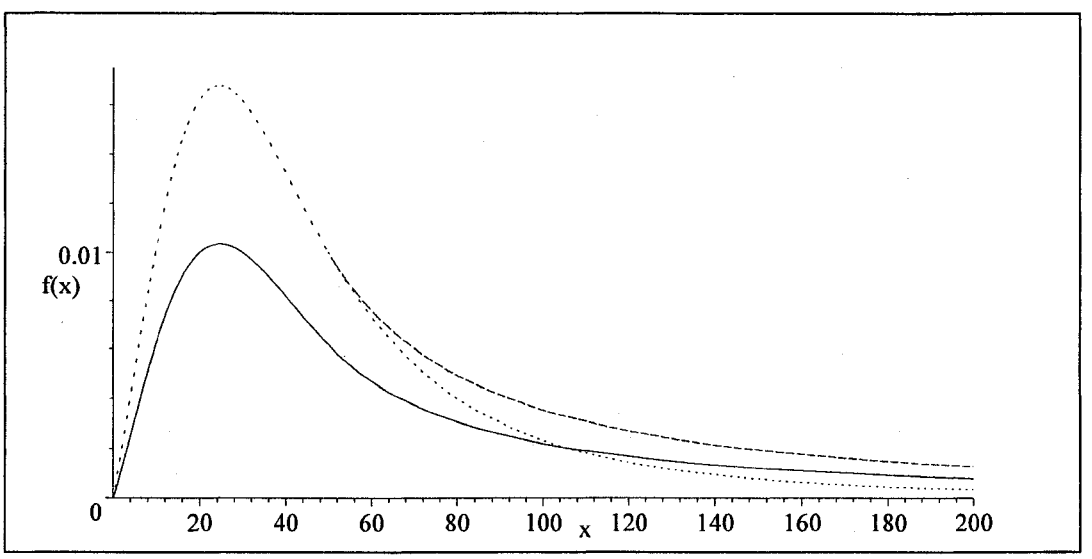

Figure 2.17 Loglogistic (dotted line), Pareto (dashed line), and LLPC (solid line) density curves ($\delta = 0.5$, $\omega = 50$).

In Figure 2.17, the dotted line, the dashed line, and the solid line indicate loglogistic, Pareto, and LLPC distribution, respectively, and it is clear that the right tail

of LLPC density graph does not approach to zero quicker than that of the loglogistic does.

In Figure 2.18, the dotted line, the dashed line, the dot-dashed line, and the solid line, respectively, represent the density curves of the LLPC family for the parameter values, $\delta = 0.5$, $\delta = 1.0$, $\delta = 1.5$, and $\delta = 2.0$.

In Figure 2.19, the dotted line, the dashed line, the dot-dashed line, and the solid line, respectively, represent the density curves of the LLPC distribution for the parameter values, $\omega = 100$, $\omega = 75$, $\omega = 50$, and $\omega = 25$.

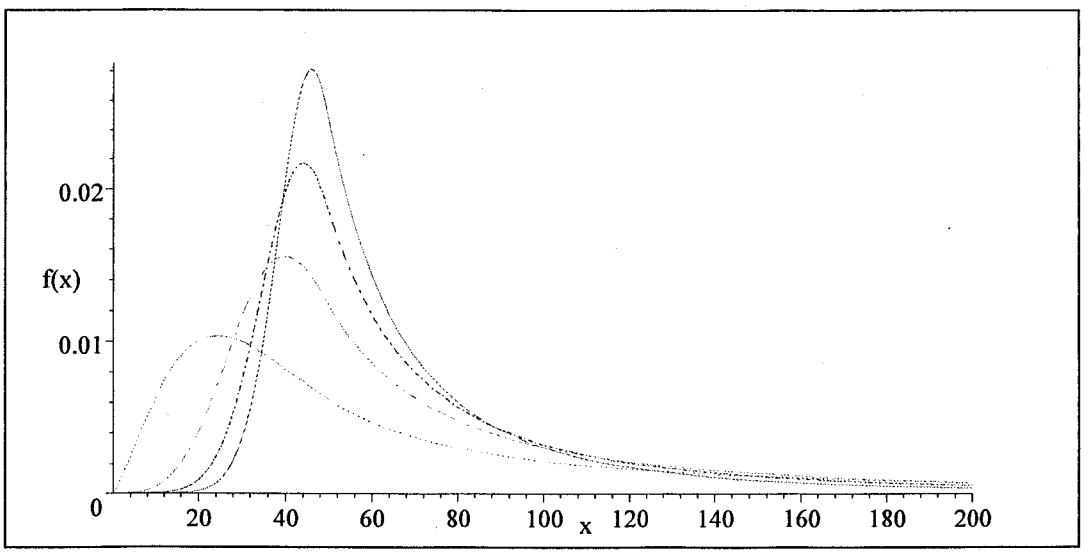

Figure 2.18 LLPC density curves with $\omega = 50$. For $\delta = 0.5$ (dotted line), $\delta = 1.0$ (dashed line), $\delta = 1.5$ (dot-dashed line), and $\delta = 2.0$ (solid line).

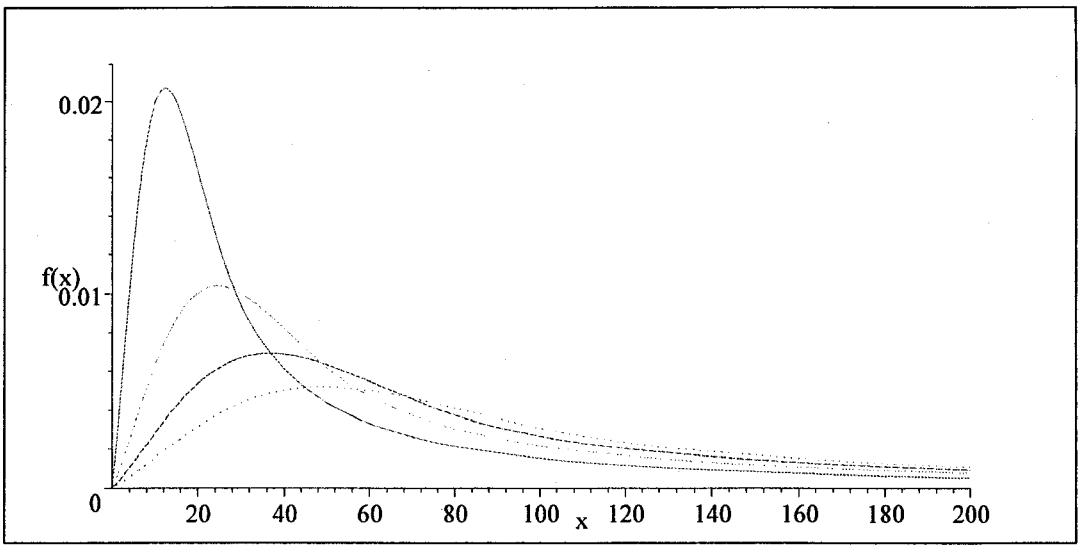

Figure 2.19 LLPC density curves with $\delta = 0.5$. For $\omega = 100$ (dotted line), $\omega = 75$ (dashed line), $\omega = 50$ (dot-dashed line), and $\omega = 25$ (solid line).

2.13 The inverse Weibull-Pareto composite distribution

Let Y be a random variable, like earlier, with the pdf

$$g(y) = \begin{cases} \kappa g_1(y) & \text{if } 0 < y \leq \varphi \\ \kappa g_2(y) & \text{if } \varphi \leq y < \infty \end{cases}. \tag{2.77}$$

Where κ is the normalizing constant, $g_1(y)$ has the form of the regular inverse Weibull density, and $g_2(y)$ has the form of the two-parameter Pareto density, i.e.,

$$g_1(y) = (\xi/y)(\mu/y)^{\xi} \exp\left[-(\mu/y)^{\xi}\right], \; y > 0, \tag{2.78}$$

and

$$g_2(y) = (\gamma/y)(\varphi/y)^{\gamma}, \; y \geq \varphi. \tag{2.79}$$

Here γ, μ, ξ, φ are unknown parameters such that $\gamma > 0$, $\mu > 0$, $\xi > 0$, $\varphi > 0$.

The continuity and differentiability conditions at φ yield

$$g_1(\varphi) = g_2(\varphi), \text{ and } g_1'(\varphi) = g_2'(\varphi). \tag{2.80}$$

Where $g'(\varphi)$ is the first derivative of $g(y)$ evaluated at φ. These conditions guarantee a smooth probability density function. These two restrictions reduce the total unknown parameters from four to two. One can show that (the proof is omitted) the inverse Weibull-Pareto composite (IWPC) density can be reparameterized and rewritten as

$$f(y) = \begin{cases} \frac{k_4^2}{1-k_4}\left(\frac{\gamma}{y}\right)\left(\frac{\varphi}{y}\right)^{\frac{\gamma}{1-k_4}} \exp\left[-k_4\left(\frac{\varphi}{y}\right)^{\frac{\gamma}{1-k_4}}\right] & \text{if} \quad 0 < y \leq \varphi \\ k_4\left(\frac{\gamma}{y}\right)\left(\frac{\varphi}{y}\right)^{\gamma} & \text{if} \quad \varphi \leq y < \infty \end{cases}, \tag{2.81}$$

where $k_4 = \kappa$ is a known constants, which is given by the positive solution of the equation $\exp(-k) = (1-k)/k$. The approximate value is, $k_4 = 0.659046068445$. As before, this natural composite probability density has only two unknown parameters $\gamma > 0$, and $\varphi > 0$.

The cumulative distribution function, $F(y)$, and the quantile function, $Q(p)$, are, respectively, given by

$$F(y) = \begin{cases} k_4 \exp\left[-k_4\left(\frac{\varphi}{y}\right)^{\frac{\gamma}{1-k_4}}\right] & \text{if} \quad 0 \leq y \leq \varphi \\ 1 - k_4\left(\frac{\varphi}{y}\right)^{\gamma} & \text{if} \quad \varphi \leq y < \infty \end{cases}, \tag{2.82}$$

and

$$Q(p) = \begin{cases} \varphi\left(\frac{1}{k_4}\ln\left(\frac{k_4}{p}\right)\right)^{-\frac{1-k_4}{\gamma}} & \text{if} \quad 0 \leq p \leq 1-k_4 \\ \varphi\left(\frac{1-p}{k_4}\right)^{-1/\gamma} & \text{if} \quad 1-k_4 \leq p < 1 \end{cases}. \tag{2.83}$$

As previously done, the graph of pdf of the IWPC distribution and its construction through the inverse Weibull density and the Pareto density are given in Figure 2.20. In this figure, the two densities, inverse Weibull and Pareto, are joined at $\varphi = 50$ to make the IWPC density, and one can easily graph the variation of the IWPC density for a given values of parameters γ and φ.

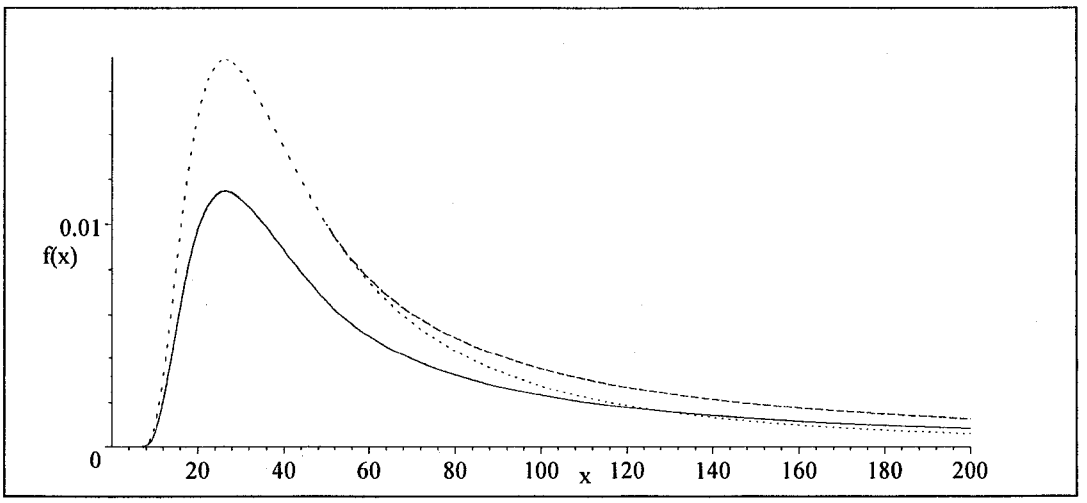

Figure 2.20 Inverse Weibull (dotted line), Pareto (dashed line), and IWPC (solid line) density curves ($\gamma = 0.5$, $\varphi = 50$).

In Figure 2.20, dotted line, dashed line, and solid line indicate inverse Weibull, Pareto, and IWPC distributions, respectively. It is clear that the tail of the LLPC distribution does not approach to zero too faster than that of the inverse Weibull does.

2.14 Grouped likelihood procedure for Pareto composite distributions

For the purpose of analyzing the grouped data, first we consider the LLPC distribution by estimating its parameters using the maximum likelihood method.

Suppose that the data consists of r intervals and the j^{th} interval, i.e. (c_{j-1}, c_j), has n_j observations for $j = 1, 2, 3, ..., r$; $c_0 = 0$. The r^{th} open interval, i.e. (c_{r-1}, ∞), contains n_r observations, and the total number of observations, n, can be written as $\sum_{j=1}^{r} n_j$. Now suppose the unknown parameter ω is in the i^{th} interval such that $c_{i-1} \leqslant \phi \leqslant c_i$, $1 \leqslant i \leqslant r$. Therefore the log-likelihood function of the LLPC distribution for grouped data can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{i-1} n_j \ln\left[F_1(c_j;\delta,\omega) - F_1(c_{j-1};\delta,\omega)\right] + n_i \ln\left[F_2(c_i;\delta,\omega) - F_1(c_{i-1};\delta,\omega)\right] \\ &\quad + \sum_{j=i+1}^{r} n_j \ln\left[F_2(c_j;\delta,\omega) - F_2(c_{j-1};\delta,\omega)\right]. \end{aligned} \tag{2.84}$$

Where $F_1(.;\delta,\omega)$ and $F_2(.;\delta,\omega)$ are such that

$$F(c_j) = \begin{cases} F_1(c_j;\delta,\omega) = \frac{1}{k_3}\left(\frac{c_j}{\omega}\right)^{\frac{1}{k_3}(k_3+1)^2\delta}\left[\frac{1}{k_3} + \left(\frac{c_j}{\omega}\right)^{\frac{1}{k_3}(k_3+1)^2\delta}\right]^{-1} & \text{if } 0 \leq c_j \leq \omega \\ F_2(c_j;\delta,\omega) = 1 - \frac{1}{k_3}\left(\frac{\omega}{c_j}\right)^{\delta} & \text{if } \omega \leq c_j < \infty \end{cases} . \tag{2.85}$$

In this case, one can find the values of δ and ω by changing ω over the interval $(0,\ \infty)$, and then maximizing the $l(\boldsymbol{\theta})$ by solving the score equation $U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$. There is a unique positive value for $\widehat{\omega}$, for fixed i such that $c_{i-1} \leqslant \widehat{\omega} \leqslant c_i$ can be obtained from this maximization. Note that one may need to check only $(r-1)$ such intervals. This procedure can be applied to the other Pareto composite distributions.

2.14.1 Grouped data example

Analysis of grouped data from Danish fire-insurance losses

The analyses are provided in Table 2.16. The estimated log-likelihood values, the chi-squared values, and the p-values indicate that the LPC distribution as the better-fit to the grouped data from Danish fire insurance losses.

Table 2.16 Estimated values of four composite models for Danish data

Loss (10^6 DKK)		Number of fires	Expected frequencies for LPC	LLPC	IWPC	WPC
0	1	325	325.51	326.25	323.83	336.15
1	2	1263	1259.91	1256.65	1276.72	1226.75
2	3	371	385.23	387.87	370.53	411.06
3	4	169	169.26	169.97	165.11	175.78
4	5	110	92.42	92.63	91.08	94.10
5	6	68	57.19	57.23	56.82	57.33
6	7	29	38.41	38.39	38.41	38.01
7	8	26	27.33	27.29	27.49	26.75
8	10	22	35.89	35.79	36.35	34.66
10	12	24	22.21	22.11	22.68	21.12
12	15	25	20.64	20.52	21.24	19.33
15	20	24	18.83	18.68	19.57	17.30
20	∞	36	39.16	38.61	42.18	33.68
Parameter values			$\widehat{\beta} = 1.364508$	$\widehat{\delta} = 1.371934$	$\widehat{\gamma} = 1.325037$	$\widehat{\gamma} = 1.440740$
			$\widehat{\theta} = 1.372916$	$\widehat{\omega} = 1.361907$	$\widehat{\varphi} = 1.261128$	$\widehat{\theta} = 1.481544$
Log-likelihood value			-4120.68	-4120.74	-4121.07	-4123.27
Pearson χ^2_{10} value			16.41	16.56	17.08	21.89
p-value			0.089	0.085	0.073	0.016

CHAPTER III

THE ODD WEIBULL FAMILY

3.1 Introduction

A generalization of the Weibull family is derived by considering the distributions of the odds of the Weibull and inverse Weibull families. This generalized Weibull distribution is henceforth referred to as the *Odd Weibull* family (Cooray 2006). The name "Odd Weibull" originates from the idea of evaluating the odds of death of a Weibull random variable. This generalization accommodates not only all five major hazard shapes; constant, increasing, decreasing, bathtub-shaped and unimodal failure rates, but also has a wide variety of density shapes including the bimodality with one mode at the origin. This bimodality corresponds to comfortable bathtub-shaped hazard curves (e.g. see figure by Shooman 1968). Furthermore, the maximum likelihood large-sample procedure, which is often used in life data analyses, can easily be implemented for this model, and is computationally convenient for censored data. The new family is suitable for discriminating between Weibull and inverse Weibull models, and is adopted for testing goodness-of-fit of Weibull and inverse Weibull as submodels. The inverse (reciprocal) transformation of the new family is the same as the original distribution, and is less common among the distributions with rich hazard rate functions, i.e., hazard rate functions that can take all five major hazard rate shapes. For example, inverse transformations of the lognormal, loglogistic, and

Birnbaum-Saunders (1969) distribution are the same as their original distributions. This property can be used to estimate the Odd Weibull parameters for a given data set in two different ways. The first method parameters are estimated by considering them to be initially positive. In the other method, same parameters are estimated by taking the inverse sample of the data set considering parameters to be initially negative. Also, this two way estimation method is extended to analyze the group, right censored, left censored, interval censored, right truncated and left truncated data that frequently arise in survival analyses.

Moreover, the total time on test (TTT) transform procedure is used as a tool to identify the hazard behavior of the proposed distribution. To measure the discrepancy between empirical and fitted TTT transforms, a previously proposed test statistic (Aarset 1987) is used. Simulation studies are carried out to obtain the upper percentage points of this statistic for the Odd Weibull family. To emphasize the flexibility and better-fitness of this family over the other leading parametric distributions, we provided an analysis of three different examples to illustrate: increasing, bathtub, and unimodal failure rates. Specifically, the first example (Section 3.7.1) contains 208 data points, which represent the ages at death in weeks for male mice exposed to 240r of gamma radiation (Kimball 1960). The second example (Section 3.7.2) represents time to failure of 50 devices put on a life test at time zero (Aarset 1987). The third example (Section 3.7.3) is a twin data set, consisting of alluvial diamonds from the Bougban (683 stones) and Damaya (444 stones) deposits in Guinea of West Africa (Beirlant *et al.* 1996). In each of these examples, the scaled fitted TTT graph is plotted along with the scaled empirical TTT graph by providing an approximate

pointwise confidence band to the scaled fitted TTT graph. Other than the likelihood estimation method, as a traditional small sample estimation procedure, the percentile matching technique is carried out.

To advance the applicability of this family; group, right censored and truncated-interval censored lifetime data is analyzed and compared with associated Kaplan-Meier (1958) curve and other leading lifetime distributions. Specifically, the fourth example (Section 3.7.4) is a positively skewed density shape and large grouped data set regarding the frequency distribution of hospital stays (in days) of 2311 schizophrenic patients (Whitmore 1986). The fifth example (Section 3.7.5) is a bimodal density shape interval censored data set regarding the drug resistance (time in months to resistance to Zidovudine) of 31 AIDS patients. Finally, the sixth example (Section 3.7.6) is negatively skewed density shapes left truncated and interval censored large twin data set (Pan and Chappell 1998, 2002) regarding the loss of functional independence of people of age 65 years or older. This twin data set consists of 421 non-poor male group and 609 non-poor female group.

We use the Odd Weibull distribution to compare and assess the accuracy for testing the exponentiality (Section 3.8) of Kolmogorov-Smirnov (1933), Anderson-Darling (1954), and Cramér von Mises (1937) test statistics. Furthermore, the empirical method to find the Odd Weibull aliases (Section 3.9) of some common distributions is presented by calculating the Moor's Kurtosis (1988) and Galton's skewness (1883). This method is useful for finding the aliases which do not have finite moments.

Moreover, the exponential transformation of the Odd Weibull distribution leads to a parent distribution for both smallest and largest extreme value distributions (Section

3.10). A large data set (Coles 2001), consisting of two oceanographic variables - wave and surge height (2894 data points for each variable), is analyzed and compared with other leading distributions (Section 3.10.1). Lastly, the logarithmic transformation of the Odd Weibull family leads to a parent distribution for both Power and Pareto distribution (Section 3.11).

3.2 Model derivation

The Odd Weibull family is considered as a suitable answer to the following two questions found in survival analysis:

1. What are the odds that an individual will die prior to time X, if X follows a certain life distribution W?

2. If these odds follow some other life distribution L, then what is the corrected distribution of X?

Obviously, the answer to the first question is very straightforward and depends on the distribution of W. However, the answer to the second question will vary due to the choice of both L and W. Let's answer the first question by representing odds that an individual will die prior to time X in terms of its survival function $S_X(x)$ as $(1 - S_X(x))/S_X(x)$, where $S_X(x) = \Pr(X \geq x)$, $x \in (0, \infty)$. Here one can denote this ratio, the odds of death, by y ($y \in (0, \infty)$), and it can be considered a random variable. Suppose that we are interested in modeling the randomness of the "odds of death" using an appropriate parametric distribution, say, $F_Y(y)$. Then, we can write

$$\Pr(Y \leq y) = F_Y(y) = F_Y\left(\frac{1 - S_X(x)}{S_X(x)}\right). \tag{3.1}$$

Let us consider the loglogistic distribution to model this randomness with its cdf

given by $F_Y(y) = 1 - (1 + y^{\gamma})^{-1}$; $0 < \gamma < \infty$. Perhaps, it is a desirable candidate to model this randomness, since the analog power transformation exists as follows

$$\left(\frac{1 - S_X(x)}{S_X(x)}\right) = \left(\frac{1 - S_Y(y)}{S_Y(y)}\right)^{1/\gamma}. \tag{3.2}$$

Where $S_Y(y) = 1 - F_Y(y)$, and γ can be considered as a correction parameter of the W distribution.

Now suppose the lifetime random variable X follows the Weibull distribution with its survival function $S_X(x) = e^{-(\frac{x}{\theta})^{\alpha}}$; $0 < x < \infty$, $0 < \alpha$, $0 < \theta$. Then the cdf of the corrected distribution of X is

$$F_X(x) = 1 - \left(1 + \left(e^{(\frac{x}{\theta})^{\alpha}} - 1\right)^{\gamma}\right)^{-1};\ 0 < x < \infty, 0 < \alpha,\ 0 < \gamma, 0 < \theta. \tag{3.3}$$

If the lifetime random variable X follows the inverse Weibull distribution with its survival function $S_X(x) = 1 - e^{-(\frac{x}{\theta})^{\alpha}}$; $0 < x < \infty$, $\alpha < 0$, $0 < \theta$, then the corrected distribution of X as

$$F_X(x) = 1 - \left(1 + \left(e^{(\frac{x}{\theta})^{\alpha}} - 1\right)^{-\gamma}\right)^{-1};\ 0 < x < \infty, \alpha < 0,\ 0 < \gamma, 0 < \theta. \tag{3.4}$$

One can easily combine equation (3.3) and equation (3.4) by writing the correction parameter $\beta = \pm\gamma$, $\gamma > 0$, to obtain the cdf of the Odd Weibull family as

$$F(x; \alpha, \beta, \theta) = 1 - \left(1 + \left(e^{(\frac{x}{\theta})^{\alpha}} - 1\right)^{\beta}\right)^{-1};\ 0 < x < \infty,\ 0 < \theta,\ 0 < \alpha\beta. \tag{3.5}$$

Hence the parameter β is the log odd ratio between the Odd Weibull and the Weibull distribution. The corresponding pdf, hazard function, and the quantile function are, respectively,

$$f(x;\alpha,\beta,\theta) = (\frac{\alpha\beta}{x})(\frac{x}{\theta})^{\alpha} e^{(\frac{x}{\theta})^{\alpha}} \left(e^{(\frac{x}{\theta})^{\alpha}} - 1\right)^{\beta-1} \left(1 + \left(e^{(\frac{x}{\theta})^{\alpha}} - 1\right)^{\beta}\right)^{-2}, \tag{3.6}$$

$$h(x;\alpha,\beta,\theta) = (\frac{\alpha\beta}{x})(\frac{x}{\theta})^{\alpha} e^{(\frac{x}{\theta})^{\alpha}} \left(e^{(\frac{x}{\theta})^{\alpha}} - 1\right)^{\beta-1} \left(1 + \left(e^{(\frac{x}{\theta})^{\alpha}} - 1\right)^{\beta}\right)^{-1}, \tag{3.7}$$

and

$$Q(u) = F^{-1}(u) = \theta \ln^{1/\alpha}\left(1 + \left(\frac{u}{1-u}\right)^{1/\beta}\right);\ 0 < u < 1. \tag{3.8}$$

It should be noted that there may be some other ways to derive this distribution by variable transformations and parameter addition (removal) of existing distributions, due to its simple form as a life distribution. It is clear that the $1/x$ transformation of the Odd Weibull family does not change its density form. Furthermore, the Odd Weibull gives the Weibull density when $\beta = 1$, and gives the inverse Weibull density when $\beta = -1$. The Odd Weibull family is asymptotically equivalent to the loglogistic distribution for larger values of θ.

Figure 3.1 and 3.2, respectively, show the Odd Weibull hazard and corresponding density curves for different parameter values. When both shape parameters, α and β, of the Odd Weibull family are negative, the hazard function given in equation (3.7) is unimodal. When parameters α and β are positive, the major shapes of the hazard function are separated by the boundary line of the space of shape parameters, $\alpha = 1$ and $\alpha\beta = 1$. Specifically, when $(\alpha > 1,\ \alpha\beta > 1)$, $(\alpha < 1,\ \alpha\beta < 1)$, $(\alpha > 1,\ \alpha\beta \leq 1)$, and $(\alpha < 1,\ \alpha\beta \geq 1)$ the shapes of the hazard function are, respectively, increasing, decreasing, bathtub, and unimodal. In addition, when α and β are positive, some

other shapes of the hazard function may appear in the two regions ($\alpha > 1,\ \alpha\beta > 1$) and ($\alpha < 1,\ \alpha\beta < 1$). However, due to complexities of the derivatives of the hazard function, the boundary lines of the parameter space are obtained numerically. Table 3.1 provides the various shapes of the hazard function for different Odd Weibull shape parameters.

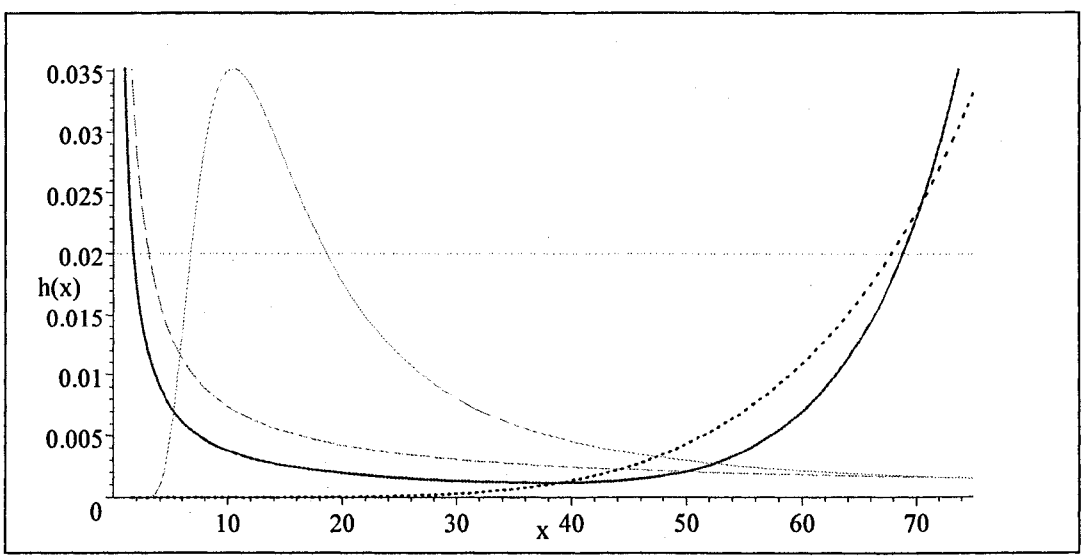

Figure 3.1 Odd Weibull hazard curves. The dark dotted line ($\alpha = 9,\ \beta = 0.7,\ \theta = 85$), the dashed line ($\alpha = 0.5,\ \beta = 0.3,\ \theta = 100$), the dot dashed line ($\alpha = 1,\ \beta = 1,\ \theta = 50$), the dark solid line ($\alpha = 8,\ \beta = 0.01,\ \theta = 45$), and the solid line ($\alpha = -1.5,\ \beta = -0.1,\ \theta = 75$), respectively, represent increasing, decreasing, constant, bathtub, and unimodal failure rates.

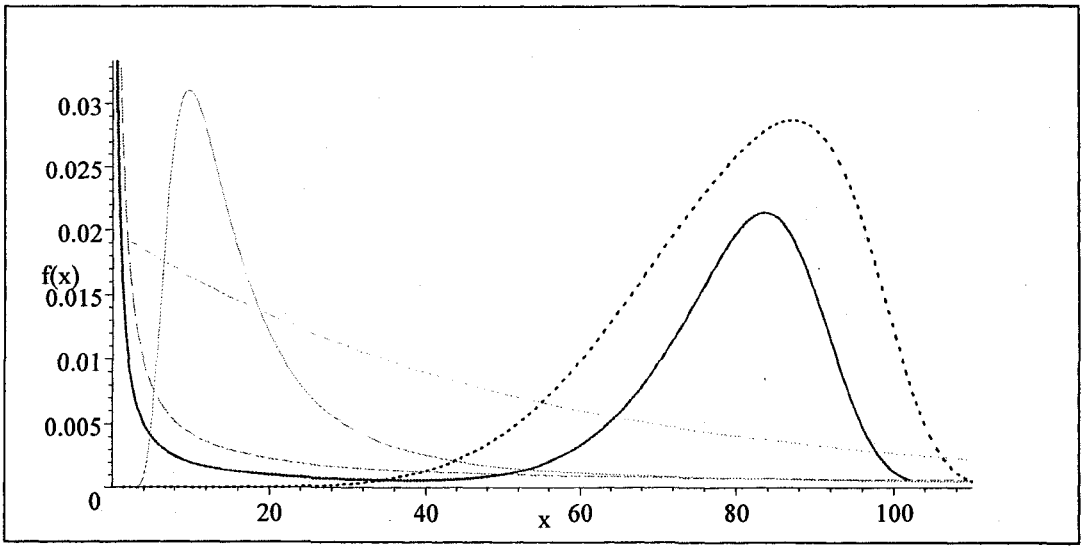

Figure 3.2 Odd Weibull density curves. These curves correspond to the hazard curves given in Figure 3.1.

Table 3.1 Hazard behavior of the Odd Weibull family

Parameter Space	Failure Rate Behavior
$\alpha = 1, \beta = 1$	constant (exponential)
$\beta = 1$	monotone (Weibull)
$\alpha = -1, \beta = -1$	unimodal (inverse exponential)
$\beta = -1$	unimodal (inverse Weibull)
$\alpha, \beta < 0$	unimodal
$\alpha, \alpha\beta > 1$	increasing ($\alpha\beta \approx 1 \rightarrow$ inverse S-shape)
$0 < \alpha, \alpha\beta < 1$	decreasing ($\alpha\beta \approx 1 \rightarrow$ S-shape)
$\alpha > 1, 0 < \alpha\beta \leq 1$	bathtub
$0 < \alpha < 1, \alpha\beta \geq 1$	unimodal

3.3 Parameter estimation under percentile matching technique

As an initial estimation method, the percentile matching technique can quickly implemented for the Odd Weibull distribution to estimate the parameters. These estimators are unique, since the Odd Weibull distribution function is strictly increasing. This is a very historical estimation method and is available in most preliminary statistics books. To apply this method, first order the data set and find the median $(x_{0.5})$ and then first $(x_{0.25})$ and third $(x_{0.75})$ quartiles. Hence one can easily formulate the following three equations to estimate the Odd Weibull parameters.

$$\alpha = \frac{\ln\left(\frac{\ln\left[1+(1/3)^{1/\beta}\right]}{\ln 2}\right)}{\ln\left(x_{0.25}/x_{0.5}\right)} = \frac{\ln\left(\frac{\ln\left[1+(3)^{1/\beta}\right]}{\ln 2}\right)}{\ln\left(x_{0.75}/x_{0.5}\right)}, \quad \theta = x_{0.5}\ln^{-1/\alpha} 2. \tag{3.9}$$

3.4 Two way parameter estimation under the likelihood method

In the following five subsections, we presented the likelihood method of estimating the Odd Weibull parameters in two different ways, specifically applications to complete, grouped, randomly right & left censored, randomly right & left truncated, and interval censored data. The likelihood function may be constructed by concerning all these events together (Klein and Moeschberger 1997, Lawless 2003).

$$\begin{aligned} L(\boldsymbol{\theta}) \;=\; & \left(\prod_{j\epsilon E} f\left(x_j;\boldsymbol{\theta}\right)\right)\left(\prod_{j\epsilon G}\left[S\left(c_{j-1};\boldsymbol{\theta}\right)-S\left(c_j;\boldsymbol{\theta}\right)\right]^{n_j}\right)\left(\prod_{j\epsilon RC} S\left(rc_j;\boldsymbol{\theta}\right)\right) \\ & \left(\prod_{j\epsilon LC}\left[1-S\left(lc_j;\boldsymbol{\theta}\right)\right]\right)\left(\prod_{j\epsilon RT}\frac{f\left(rt_j;\boldsymbol{\theta}\right)}{\left[1-S\left(rt_j;\boldsymbol{\theta}\right)\right]}\right)\left(\prod_{j\epsilon LT}\frac{f\left(lt_j;\boldsymbol{\theta}\right)}{S\left(lt_j;\boldsymbol{\theta}\right)}\right) \\ & \left(\prod_{j\epsilon IC}\left[S\left(lic_j;\boldsymbol{\theta}\right)-S\left(ric_j;\boldsymbol{\theta}\right)\right]\right). \end{aligned} \tag{3.10}$$

Where,

$L(\boldsymbol{\theta})-$ likelihood function, $\boldsymbol{\theta}-$ unknown parameter vector to be estimated, $f-$ density function, $S-$ survival function.

Also $f(x_j;\boldsymbol{\theta})$ – for exact or complete data points, x_j- exact data points, $E-$ set of exact data points.

$[S(c_{j-1};\boldsymbol{\theta}) - S(c_j;\boldsymbol{\theta})]^{n_j}$ – for grouped data points, n_j- number of data points in the j^{th} group, i.e., $n_j \epsilon [c_{j-1}, c_j)$, c_j- upper limit of the j^{th} group, $G-$ set of grouped data points.

$S(rc_j;\boldsymbol{\theta})$ – for right censored data points, rc_j- right censored data points, $RC-$ set of right censored data points.

$[1 - S(lc_j;\boldsymbol{\theta})]$ – for left censored data points, lc_j- left censored data points, $LC-$ set of left censored data points.

$f(rt_j;\boldsymbol{\theta}) / [1 - S(rt_j;\boldsymbol{\theta})]$ – for right truncated data points, rt_j- right truncated data points, $RT-$ set of right truncated data points.

$f(lt_j;\boldsymbol{\theta}) / S(lt_j;\boldsymbol{\theta})$ – for left truncated data points, lt_j- left truncated data points, $LT-$ set of left truncated data points.

$[S(lic_j;\boldsymbol{\theta}) - S(ric_j;\boldsymbol{\theta})]$ – for interval censored data points, lic_j- left limit of j^{th} interval censored data points, ric_j- right limit of j^{th} interval censored data points, $IC-$ set of interval censored data points.

The score function and the observed information matrix for $\boldsymbol{\theta}$ are, respectively,

$$\frac{\partial l}{\partial \boldsymbol{\theta}} \text{ and } i(\boldsymbol{\theta}) = \frac{-\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \tag{3.11}$$

where $l = \log L(\boldsymbol{\theta})$.

The likelihood procedures are available in Cox and Oakes (1984), Kalbfleisch and Prentice (2002), Rao (1973), and Lawless (2003). In our numerical examples given in Section 3.7, the asymptotic standard errors (SE's) of the estimates are obtained by inverting the observed information matrix of the log-likelihood function (Efron and Hinkley 1978).

3.4.1 Applications to the complete data

Large number of complete data examples are available in literature. Some examples are given in the appendix B. The standard maximum likelihood procedure can be used to estimate the Odd Weibull parameters, for such given data, by maximizing the log-likelihood function ($l(\boldsymbol{\theta}) = \ln L(\alpha, \beta, \theta)$). Lets assume $x_1, x_2, ..., x_n$ are an ordered random sample (i.e. $x_1 \leq x_2 \leq ... \leq x_n$) from the Odd Weibull family.

By using equation (3.10) the log-likelihood function of the Odd Weibull family can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{n} \left\{ \ln\left(\alpha\beta\theta^{-\alpha}\right) + (\alpha-1)\ln x_j + \left(\frac{x_j}{\theta}\right)^{\alpha} + (\beta-1)\ln\left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right) \right\} \\ &\quad -2\sum_{j=1}^{n} \ln\left(1 + \left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right)^{\beta}\right). \end{aligned} \tag{3.12}$$

Let $X_1, X_2, ..., X_n$ be a random sample from the positive region (i.e., $\alpha > 0$, $\beta > 0$) of the Odd Weibull family. Without loss of generality, we can assume that this is an ordered sample, i.e., $x_1 \leq x_2 \leq ... \leq x_n$. Let us denote the corresponding log-likelihood function $l_+^e(\boldsymbol{\theta}) = \ln L(\alpha_+, \beta_+, \theta_+)$, which can be obtained from equation (3.12), where $\alpha_+ = \alpha$, $\beta_+ = \beta$, $\theta_+ = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

Now suppose the inverse random sample of $X_1, X_2, ..., X_n$, i.e., $1/X_n, 1/X_{n-1}, ...,$ $1/X_1$ are coming from the negative region (i.e., $\alpha < 0$, $\beta < 0$) of the Odd Weibull family. As before, we can assume that this is an ordered sample, i.e., $1/x_n \leq 1/x_{n-1} \leq ... \leq 1/x_1$. Let us denote the corresponding log-likelihood function $l_-^e(\boldsymbol{\theta}) = \ln L(\alpha_-, \beta_-, \theta_-)$, which can be obtained from equation (3.12), where $\alpha_- = \alpha$, $\beta_- = \beta$, $\theta_- = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

The following four relations then hold.

$$\alpha_+ = -\alpha_-,\ \beta_+ = -\beta_-,\ \theta_+ = 1/\theta_-,\ l_+^e(\boldsymbol{\theta}) = l_-^e(\boldsymbol{\theta}) - 2\sum_{j=1}^{n} \ln x_j. \tag{3.13}$$

Note that without direct calculation, by using the delta method, the estimated standard errors of the negative parameter values can be obtained from observed information matrix of the positive parameter values. Clearly following standard error relations hold.

$$SE_{\widehat{\alpha_-}} = SE_{\widehat{\alpha_+}},\ SE_{\widehat{\beta_-}} = SE_{\widehat{\beta_+}},\ SE_{\widehat{\theta_-}} \approx \frac{1}{\widehat{\theta_+^2}} SE_{\widehat{\theta_+}}. \tag{3.14}$$

According to these relations, the Odd Weibull family allows us to estimate its parameters in two different ways. Moreover, one can easily prove that this density is strictly unimodal and positively skewed when $\alpha < 0$, $\beta < 0$. Therefore, this two way estimation method is useful to avoid some computational issues involved in the likelihood procedure, especially when the Odd Weibull densities are non-unimodal.

3.4.2 Applications to the grouped data

Grouped data are usually large samples and available in many fields such as actuarial sciences, demographic studies, reliability analysis, and etc. This type of data is usually arise due to grouping the original point data by data collectors. Some examples are given in the appendix B. Suppose that the data consists of r intervals and the j^{th} interval, i.e. (c_{j-1}, c_j), has n_j observations for $j = 1, 2, 3, ..., r$; $c_0 = 0$. The r^{th} open interval, i.e. (c_{r-1}, ∞), contains n_r observations, and the total number of observations, n, can be written as $\sum_{j=1}^{r} n_j$. Therefore using equation (3.10) the log-likelihood function of the Odd Weibull family for grouped data can be written as

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{r} n_j \ln \left[\left(1 + \left(e^{(\frac{c_{j-1}}{\theta})^{\alpha}} - 1 \right)^{\beta} \right)^{-1} - \left(1 + \left(e^{(\frac{c_j}{\theta})^{\alpha}} - 1 \right)^{\beta} \right)^{-1} \right]. \tag{3.15}$$

Let's consider the following grouped data example. Table 3.2a represents the original grouped data. Table 3.2b is created by inverting the values in Table 3.2a. Let the data are in Table 3.2a coming from the positive region (i.e., $\alpha > 0$, $\beta > 0$) of the Odd Weibull family. As in the previous subsection we can denote the corresponding log-likelihood function $l_+^g(\boldsymbol{\theta}) = \ln L(\alpha_+, \beta_+, \theta_+)$, which can be obtained from equation (3.15), where $\alpha_+ = \alpha$, $\beta_+ = \beta$, $\theta_+ = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

Now suppose the inverted data in Table 3.2b are coming from the negative region (i.e., $\alpha < 0$, $\beta < 0$) of the Odd Weibull family. As before, we can denote the corresponding log-likelihood function $l_-^g(\boldsymbol{\theta}) = \ln L(\alpha_-, \beta_-, \theta_-)$, which can be obtained from equation (3.15), where $\alpha_- = \alpha$, $\beta_- = \beta$, $\theta_- = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

Table 3.2 Grouped data and its inverted structure

Table 3.2a		Table 3.2b	
Data interval	Frequency	Data interval	Frequency
$0 < x < c_1$	n_1	$0 < 1/x \leq 1/c_{r-1}$	n_r
$c_1 \leq x < c_2$	n_2	$1/c_{r-1} < 1/x \leq 1/c_{r-2}$	n_{r-1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c_{r-1} \leq x < \infty$	n_r	$1/c_1 < 1/x < \infty$	n_1

The following four relations then hold.

$$\alpha_+ = -\alpha_-,\ \beta_+ = -\beta_-,\ \theta_+ = 1/\theta_-,\ l_+^g(\boldsymbol{\theta}) = l_-^g(\boldsymbol{\theta}). \tag{3.16}$$

Once again, without direct calculation, by using the delta method, the estimated standard errors of the negative parameter values can be obtained from observed information matrix of the positive parameter values. Therefore the same relations given in equation (3.14) are valid.

3.4.3 Applications to the randomly right censored data

Randomly right censored data mostly appear in reliability and survival analysis. These data are categorize into several parts due to the censoring nature of the data sets. Appendix B provides the data of Type I censoring, Type I progressive censoring, generalized Type I censoring, and random censoring. The likelihood function given in equation (3.17) is the general form to analyze all of these censoring data.

Suppose we want to analyze the randomly right censored data. Note that, as before, $x_1, x_2, ..., x_n$ are an ordered random sample (i.e. $x_1 \leq x_2 \leq ... \leq x_n$) from the

Odd Weibull family with some values are right censored. Then, by using equation (3.10), the log-likelihood function of the Odd Weibull family can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{n} \delta_j \left\{ \ln\left(\alpha\beta\theta^{-\alpha}\right) + (\alpha-1)\ln x_j + \left(\frac{x_j}{\theta}\right)^{\alpha} + (\beta-1)\ln\left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right) \right\} \\ &\quad - \sum_{j=1}^{n} (1+\delta_j) \ln\left(1 + \left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right)^{\beta}\right), \end{aligned} \tag{3.17}$$

where

$$\delta_j = \begin{cases} 0 & \text{if } j^{th} \text{ observation is right censored, } j = 1, 2, ..., n \\ 1 & \text{else.} \end{cases}$$

Similarly, by using equation (3.10), one can write the following log-likelihood function of the Odd Weibull family to analyze the left censored data.

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{n} \delta_j \left\{ \ln\left(\alpha\beta\theta^{-\alpha}\right) + (\alpha-1)\ln x_j + \left(\frac{x_j}{\theta}\right)^{\alpha} - \ln\left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right) \right\} \\ &\quad + \beta \sum_{j=1}^{n} \ln\left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right) - \sum_{j=1}^{n} (1+\delta_j) \ln\left(1 + \left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right)^{\beta}\right) \end{aligned} \tag{3.18}$$

where $x_1, x_2, ..., x_n$ are an ordered random sample (i.e. $x_1 \leq x_2 \leq ... \leq x_n$) from the Odd Weibull family with some values are left censored, and

$$\delta_j = \begin{cases} 0 & \text{if } j^{th} \text{ observation is left censored, } j = 1, 2, ..., n \\ 1 & \text{else.} \end{cases}$$

Now let $X_1, X_2, ..., X_n$ be an original random sample from the positive region (i.e., $\alpha > 0, \beta > 0$) of the Odd Weibull family with some values are right censored. Without loss of generality, we can assume that this is an ordered sample, i.e., $x_1 \leq x_2 \leq ... \leq x_n$. Let us denote the corresponding log-likelihood function $l_{+}^{rc}(\boldsymbol{\theta}) = \ln L(\alpha_+, \beta_+, \theta_+)$, which can be obtained from equation (3.17), where $\alpha_+ = \alpha$, $\beta_+ = \beta$, $\theta_+ = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

Now suppose the inverse random sample of $X_1, X_2, ..., X_n$, i.e., $1/X_n, 1/X_{n-1}, ...,$ $1/X_1$ are coming from the negative region (i.e., $\alpha < 0$, $\beta < 0$) of the Odd Weibull family. As before, we can assume that this is an ordered sample, i.e., $1/x_n \leq 1/x_{n-1} \leq$ $... \leq 1/x_1$. Let us treat this reciprocal data set as left censored and denote the corresponding log-likelihood function $l_-^{lc}(\boldsymbol{\theta}) = \ln L(\alpha_-, \beta_-, \theta_-)$, which can be obtained from equation (3.18), where $\alpha_- = \alpha$, $\beta_- = \beta$, $\theta_- = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

The following four relations then hold.

$$\alpha_+ = -\alpha_-,\ \beta_+ = -\beta_-,\ \theta_+ = 1/\theta_-,\ l_+^{rc}(\boldsymbol{\theta}) = l_-^{lc}(\boldsymbol{\theta}) - 2\sum_{j=1}^{n} \delta_j \ln x_j. \tag{3.19}$$

Once again, without direct calculation, by using the delta method, the estimated standard errors of the negative parameter values can be obtained from observed information matrix of the positive parameter values. Therefore the same relations given in equation (3.14) are valid.

3.4.4 Applications to the randomly truncated data

Randomly truncated data mostly appear in survival analysis. These data are usually divided into two categories, namely left truncated and right truncated data. Left truncation usually arises due to delayed entry of individuals to a event of interest. Right truncation occurs when only individuals who have experienced the event are included in the sample, and any individual who has yet to experience the event is not observed (Klein and Moeschberger 1997). Examples are given in the appendix B. Suppose we want to analyze the randomly left truncated data. Note that, as

before, $x_1, x_2, ..., x_n$ are an ordered random sample (i.e. $x_1 \leq x_2 \leq ... \leq x_n$) from the Odd Weibull family with some values are left truncated at lt_j, where $lt_j \leq x_j$ for $j = 1, 2, ..., n$. If $lt_j = 0$ then the corresponding x_j value is not left truncated. Then, by using equation (3.10), the log-likelihood function of the Odd Weibull family can be written as

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{n} \left\{ \ln\left(\alpha\beta\theta^{-\alpha}\right) + (\alpha-1)\ln x_j + \left(\frac{x_j}{\theta}\right)^{\alpha} + (\beta-1)\ln\left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right) \right\} \\ &\quad -2\sum_{j=1}^{n} \ln\left(1 + \left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right)^{\beta}\right) + \sum_{j=1}^{n} \ln\left(1 + \left(e^{(\frac{lt_j}{\theta})^{\alpha}} - 1\right)^{\beta}\right). \end{aligned} \quad (3.20)$$

Similarly, by using equation (3.10), one can easily write the following log-likelihood function of the Odd Weibull family to analyze the right truncated data.

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{n} \left\{ \ln\left(\alpha\beta\theta^{-\alpha}\right) + (\alpha-1)\ln x_j + \left(\frac{x_j}{\theta}\right)^{\alpha} + (\beta-1)\ln\left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right) \right\} \\ &\quad -2\sum_{j=1}^{n} \ln\left(1 + \left(e^{(\frac{x_j}{\theta})^{\alpha}} - 1\right)^{\beta}\right) \\ &\quad +\sum_{j=1}^{n} \ln\left(1 - 1 \Big/ \left(1 + \left(e^{(\frac{rt_j}{\theta})^{\alpha}} - 1\right)^{\beta}\right)\right), \end{aligned} \quad (3.21)$$

where $x_1, x_2, ..., x_n$ are an ordered random sample (i.e. $x_1 \leq x_2 \leq ... \leq x_n$) from the Odd Weibull family with some values are right truncated at rt_j, where $rt_j \leq x_j$ for $j = 1, 2, ..., n$. If $rt_j = \infty$ then the corresponding x_j value is not right truncated.

Now let $X_1, X_2, ..., X_n$ be an original random sample from the positive region (i.e., $\alpha > 0$, $\beta > 0$) of the Odd Weibull family with some values are left truncated at lt_j, where $lt_j \leq x_j$ for $j = 1, 2, ..., n$. If $lt_j = 0$ then the corresponding x_j value is not left truncated. Without loss of generality, we can assume that this is an ordered sample, i.e., $x_1 \leq x_2 \leq ... \leq x_n$. Let us denote the corresponding log-likelihood function $l_{+}^{lt}(\boldsymbol{\theta}) = \ln L(\alpha_+, \beta_+, \theta_+)$, which can be obtained from equation (3.20), where $\alpha_+ = \alpha$,

$\beta_+ = \beta$, $\theta_+ = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

Now suppose the inverse random sample of $X_1, X_2, ..., X_n$, i.e., $1/X_n, 1/X_{n-1}, ...,$ $1/X_1$ are coming from the negative region (i.e., $\alpha < 0$, $\beta < 0$) of the Odd Weibull family. As before, we can assume that this is an ordered sample, i.e., $1/x_n \leq 1/x_{n-1} \leq$ $... \leq 1/x_1$. These values are now right truncated at $1/lt_j$, where $1/x_j \leq 1/lt_j$ for $j = 1, 2, ..., n$. If $1/lt_j = \infty$ then the corresponding x_j value is not right truncated. Therefore, for this reciprocal data set we can denote the corresponding log-likelihood function $l_-^{rt}(\boldsymbol{\theta}) = \ln L(\alpha_-, \beta_-, \theta_-)$, which can be obtained from equation (3.21), where $\alpha_- = \alpha$, $\beta_- = \beta$, $\theta_- = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

The following four relations then hold.

$$\alpha_+ = -\alpha_-,\ \beta_+ = -\beta_-,\ \theta_+ = 1/\theta_-,\ l_+^{lt}(\boldsymbol{\theta}) = l_-^{rt}(\boldsymbol{\theta}) - 2\sum_{j=1}^{n} \ln x_j. \tag{3.22}$$

Once again, without direct calculation, by using the delta method, the estimated standard errors of the negative parameter values can be obtained from observed information matrix of the positive parameter values. Therefore the same relations given in equation (3.14) are valid.

3.4.5 Applications to the interval censored data

Suppose we want to analyze interval censored data (see examples in the appendix B). Let us assume the actual data points lie between lic_j and ric_j, for $j = 1, 2, ..., n$, where lic_j and ric_j are, respectively, left and right limit of j^{th} data point. Note that

$lic_1, lic_2, ..., lic_n$ or $ric_1, ric_2, ..., ric_n$ is not necessarily an ordered limit set. Therefore using equation (3.10) the log-likelihood function of the Odd Weibull family for interval censored data can be written as

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln \left[\left(1 + \left(e^{(\frac{lic_j}{\theta})^{\alpha}} - 1 \right)^{\beta} \right)^{-1} - \left(1 + \left(e^{(\frac{ric_j}{\theta})^{\alpha}} - 1 \right)^{\beta} \right)^{-1} \right]. \tag{3.23}$$

Let's consider the following interval censored data example. Table 3.3a represents the original interval censored data. Table 3.3b is created by inverting the values in Table 3.3a.

Table 3.3 Interval censored data and its inverted structure

Table 3.3a		Table 3.3b	
Censoring point		Censoring point	
left limit	right limit	left limit	right limit
lic_1	ric_1	$1/ric_1$	$1/lic_1$
lic_2	ric_2	$1/ric_2$	$1/lic_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
lic_n	ric_n	$1/ric_n$	$1/lic_n$

Let the data are in Table 3.3a coming from the positive region (i.e., $\alpha > 0$, $\beta > 0$) of the Odd Weibull family. As in the previous subsection we can denote the corresponding log-likelihood function $l_+^{ic}(\boldsymbol{\theta}) = \ln L(\alpha_+, \beta_+, \theta_+)$, which can be obtained from equation (3.23), where $\alpha_+ = \alpha$, $\beta_+ = \beta$, $\theta_+ = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

Now suppose the inverted data in Table 3.3b are coming from the negative region (i.e., $\alpha < 0$, $\beta < 0$) of the Odd Weibull family. As before, we can denote the corresponding log-likelihood function $l_{-}^{ic}(\boldsymbol{\theta}) = \ln L(\alpha_{-}, \beta_{-}, \theta_{-})$, which can be obtained from equation (3.23), where $\alpha_{-} = \alpha$, $\beta_{-} = \beta$, $\theta_{-} = \theta$. Also these parameters and their standard errors can be estimated by using equation (3.11).

The following four relations then hold.

$$\alpha_{+} = -\alpha_{-},\ \beta_{+} = -\beta_{-},\ \theta_{+} = 1/\theta_{-},\ l_{+}^{ic}(\boldsymbol{\theta}) = l_{-}^{ic}(\boldsymbol{\theta}). \tag{3.24}$$

Once again, without direct calculation, by using the delta method, the estimated standard errors of the negative parameter values can be obtained from observed information matrix of the positive parameter values. Therefore the same relations given in equation (3.14) are valid.

3.5 Goodness-of-fit

Comparing the goodness-of-fit of a Weibull model is complicated by the large magnitude of the class of alternatives. By restricting the alternatives to the Odd Weibull family, we can use the usual likelihood ratio statistics for testing the adequacy of the Weibull and inverse Weibull submodels. The null hypotheses, $H_{011} : \beta = 1$, $H_{012} : (\alpha = 1, \beta = 1)$, $H_{021} : \beta = -1$, and $H_{022} : (\alpha = -1, \beta = -1)$ respectively correspond to the Weibull, exponential, inverse Weibull, and inverse exponential submodels of the Odd Weibull family. The likelihood ratio statistics for $H_{0ij}\,(i = 1, 2; j = 1, 2)$ are:

$$\Lambda = \sup_{R_{0ij}} L(\alpha, \beta, \theta) \,/\, \sup_{UR} L(\alpha, \beta, \theta),\ i = 1, 2;\ j = 1, 2. \tag{3.25}$$

Where, R_{0ij} is the restricted parametric space corresponding to H_{0ij}, $i = 1, 2; j = 1, 2$. UR is the unrestricted parameter space.

In terms of the ML estimates, the likelihood ratio statistics reduce to:

$$\begin{aligned} \Lambda_{11} &= L(\alpha_w, \beta = 1, \theta_w)/L(\alpha, \beta, \theta); \ \Lambda_{12} = L(\alpha = 1, \beta = 1, \theta_e)/L(\alpha, \beta, \theta); \\ \Lambda_{21} &= L(\alpha_{iw}, \beta = -1, \theta_{iw})/L(\alpha, \beta, \theta); \ \Lambda_{22} = L(\alpha = -1, \beta = -1, \theta_{ie})/L(\alpha, \beta, \theta). \end{aligned} \tag{3.26}$$

Under the null hypothesis, $-2 \cdot \ln(\Lambda_{11})$, $-2 \cdot \ln(\Lambda_{12})$, $-2 \cdot \ln(\Lambda_{21})$, and $-2 \cdot \ln(\Lambda_{22})$ respectively follow χ^2 distribution with degrees of freedom 1, 2, 1, and 2. The use of the Odd Weibull family for modeling and testing goodness-of-fit hypotheses is given in Section 3.8.

3.6 Total time on test transforms

3.6.1 TTT transforms of the Odd Weibull family

In the analysis of lifetime data by the Odd Weibull family, predetermination of the sign and range of α and β can be obtained by using the empirical TTT procedure. For this purpose, the empirical TTT transform $(H_N^{-1}(r/n))$ can be used as a tool to identify the hazard shape for the given data set.

The scaled empirical TTT is given by

$$\phi_N(r/n) = H_N^{-1}(r/n)/H_N^{-1}(1) = \left(\sum_{i=1}^{r} X_{i:n} + (n-r) X_{r:n}\right) / \sum_{i=1}^{n} X_i. \tag{3.27}$$

Where $r = 1, ..., n$ and $X_{i:n}$, $i = 1, ..., n$ represent the order statistics of the sample.

If the empirical TTT transform is convex, concave, convex then concave, and concave then convex, the shape of the corresponding hazard function for such failure data is decreasing, increasing, bathtub, and unimodal respectively (Barlow and Campo 1975; Aarset 1987; Mudholkar *et al.* 1996).

The scaled TTT transform for the Odd Weibull family can be defined as

$$\phi_F(u) = H_F^{-1}(u)/H_F^{-1}(1);\ 0 < u < 1, \tag{3.28}$$

where $H_F^{-1}(u) = \int_0^{F^{-1}(u)} (1 - F(x))\, dx$. The $F^{-1}(u)$ is given in equation (3.8).

Adequacy of the model for the given uncensored data can be illustrated by using the plot of $\phi_F(u)$. The $\phi_F(u)$ of the Odd Weibull family is indeed a function of parameters α and β for a given value u. The typical shapes of $\phi_F(u)$ of the Odd Weibull family given by the following formula for $0 < u < 1$, are illustrated in Figure 3.3.

$$\begin{aligned}\phi_F(u) &= \left[\int_0^u \ln^{1/\alpha}\left(1+\left(\frac{t}{1-t}\right)^{1/\beta}\right) dt + (1-u)\ln^{1/\alpha}\left(1+\left(\frac{u}{1-u}\right)^{1/\beta}\right)\right] \\ &\quad \Big/ \left(\int_0^1 \ln^{1/\alpha}\left(1+\left(\frac{t}{1-t}\right)^{1/\beta}\right) dt\right).\end{aligned} \tag{3.29}$$

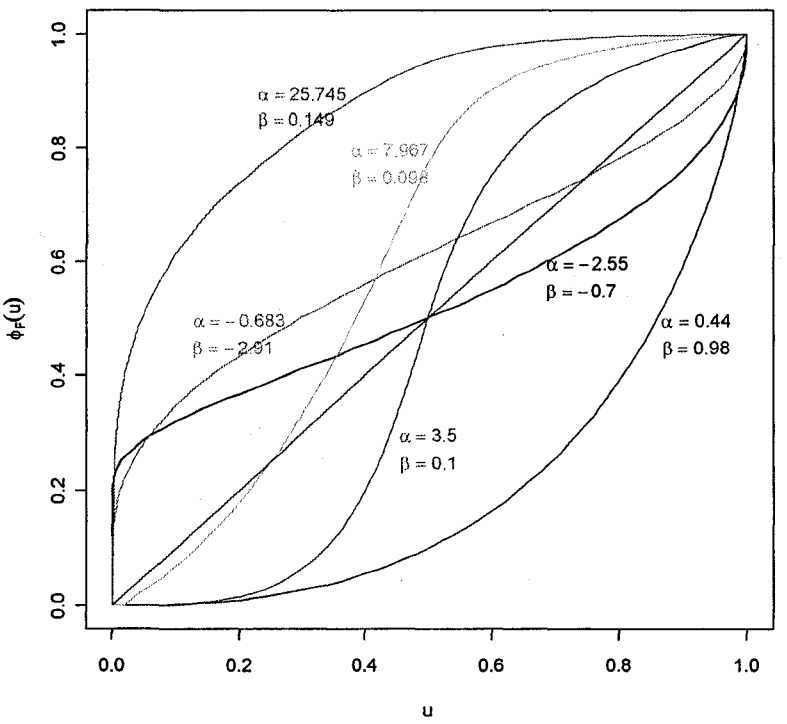

Figure 3.3 Typical shapes of $\phi_F(u)$ of the Odd Weibull family. The above line ($\alpha = 25.745$, $\beta = 0.149$), the below line ($\alpha = 0.44$, $\beta = 0.98$), the lines (above: $\alpha = 7.967$, $\beta = 0.098$, below: $\alpha = 3.5$, $\beta = 0.1$), and the lines (above: $\alpha = -0.683$, $\beta = -2.91$, below: $\alpha = -2.55$, $\beta = -0.7$) respectively represent increasing, decreasing, bathtub, and unimodal failure rates.

Note that, the ML estimates were used to plot the fitted $\phi_F(u)$ curves for the examples given in Section 3.7. In these cases, we graphically investigate the effect of the substitution of ML estimates for the unknown parameters by constructing an approximate pointwise confidence band for $\phi_F(u)$. See Figure 3.4, 3.5, 3.6a, and 3.6b. This is done by calculating the asymptotic variance of $\phi_F(u)$ for given u $(0 < u < 1)$, using the delta method.

For the present purpose, to test the goodness-of-fit based on the TTT plot, one can check the values of the following test statistic (Aarset 1987).

$$R_n = \int_0^1 T_n^2(u)du, \tag{3.30}$$

where $T_n(u) = \sqrt{n}\{\phi_N(r/n) - \phi_F(u)\}$, and the asymptotic distribution of R_n under exponentiality is obtained. The null hypothesis is rejected when R_n is large, and we can use it to measure the discrepancy between an empirical and a fitted TTT plot.

3.6.2 Simulation studies

To gain some insight into the adequacy of the above test statistic (3.30) for the Odd Weibull family, simulation studies are conducted as follows: 100,000 random samples are generated from the Odd Weibull family and R_n values are calculated. 90%, 95%, and 99% upper percentage points of R_n are recorded. This procedure is repeated for different parameter values and sample sizes. The results are given in Table 3.4. Parameters of the Odd Weibull family are chosen in a way to produce different types of failure rates.

From this table, one can see that the R_n values decrease with sample size, except when the parameters of the Odd Weibull family represent unimodal failure rates.

Table 3.4 Upper percentage points of R_n for the Odd Weibull family

		90%			95%			99%		
		Sample size								
α	β	10	20	50	10	20	50	10	20	50
-3.0	-0.5	1.532	1.900	2.597	1.917	2.405	3.379	2.672	3.445	5.162
-2.0	-2.0	0.141	0.152	0.161	0.198	0.217	0.230	0.363	0.419	0.447
-2.0	-1.0	0.712	0.865	1.089	0.929	1.169	1.532	1.441	2.105	3.277
-1.0	-2.0	0.612	0.730	0.915	0.815	0.999	1.309	1.305	1.923	3.061
-1.0	-1.0	3.598	5.090	8.014	4.099	5.879	9.460	4.999	7.359	12.04
-0.5	-3.0	1.054	1.315	1.813	1.321	1.693	2.413	1.863	2.536	4.406
0.5	0.5	1.020	0.844	0.733	1.378	1.199	1.047	2.105	2.043	1.831
0.5	1.0	1.011	0.895	0.814	1.382	1.271	1.161	2.190	2.169	2.018
0.5	3.0	0.425	0.431	0.448	0.588	0.602	0.633	0.963	1.011	1.085
0.5	5.0	0.200	0.201	0.205	0.281	0.285	0.290	0.484	0.486	0.506
1.0	0.5	0.423	0.327	0.282	0.660	0.497	0.414	1.284	0.999	0.773
1.0	1.0	0.365	0.321	0.300	0.546	0.472	0.430	1.025	0.880	0.773
1.0	5.0	0.054	0.053	0.052	0.078	0.075	0.074	0.140	0.132	0.127
5.0	0.1	0.058	0.041	0.034	0.101	0.067	0.051	0.265	0.156	0.103
5.0	0.5	0.058	0.050	0.046	0.097	0.078	0.067	0.211	0.163	0.129
5.0	1.0	0.030	0.028	0.027	0.048	0.041	0.039	0.099	0.083	0.071
8.0	0.1	0.050	0.036	0.029	0.091	0.062	0.045	0.251	0.150	0.098
8.0	1.0	0.014	0.013	0.012	0.022	0.019	0.018	0.047	0.039	0.033
8.0	5.0	0.001	0.001	0.001	0.002	0.001	0.001	0.003	0.003	0.002

3.7 Illustrative examples

3.7.1 Increasing failure rate and uncensored data

This example contains 208 data points (data set is given in the appendix B), which represent the ages at death in weeks for male mice exposed to 240r of gamma radiation (Kimball 1960). The empirical TTT plot for this data set indicates an increasing hazard rate (see dark dotted line in Figure 3.4). The Odd Weibull family becomes versatile because it expands the Weibull family into a larger family due to an additional shape parameter. Therefore, it is important to identify the parameter space with increasing failure rate. From Section 3.2, when $\alpha > 1$, $\alpha\beta > 1$, the hazard function of the Odd Weibull family is increasing. Therefore, one can make an initial guess to estimate the parameters. Also, our quick parameter estimation method, the percentile estimation method also called crude estimation method given in Section 3.3, provides the estimated parameter values of the Odd Weibull family as $\widetilde{\alpha} = 10.5369$, $\widetilde{\beta} = 0.4299$, $\widetilde{\theta} = 130.4599$.

Table 3.5a provides the estimated parameter values and the log-likelihood values of the fitted Odd Weibull family to the original and inverse sample of mice data. Specifically, the inverse sample of mice data is analyzed by using the negative region (Section 3.4.1) of the Odd Weibull family.

Furthermore, the fitness of the Odd Weibull family to the original mice data is illustrated by the scaled fitted TTT graph (see dark solid line in Figure 3.4) with its 5% confidence band (solid lines). The calculated R_n value of the original sample of mice data is 0.0107, which is smaller than 90% upper percentage point of R_n, 0.0249. Therefore, we do not reject the Odd Weibull fit for the mice data.

Table 3.5a Estimated values of the Odd Weibull family for mice data

Sample	$\widehat{\alpha} \pm SE_{\widehat{\alpha}}$	$\widehat{\beta} \pm SE_{\widehat{\beta}}$	$\widehat{\theta} \pm SE_{\widehat{\theta}}$	$l(\widehat{\boldsymbol{\theta}})$
Original	6.2278 ± 0.8326	0.7495 ± 0.1221	131.45 ± 1.9535	-988.89
Inverse	-6.2278 ± 0.8328	-0.7495 ± 0.1220	0.0076 ± 0.0001	993.25

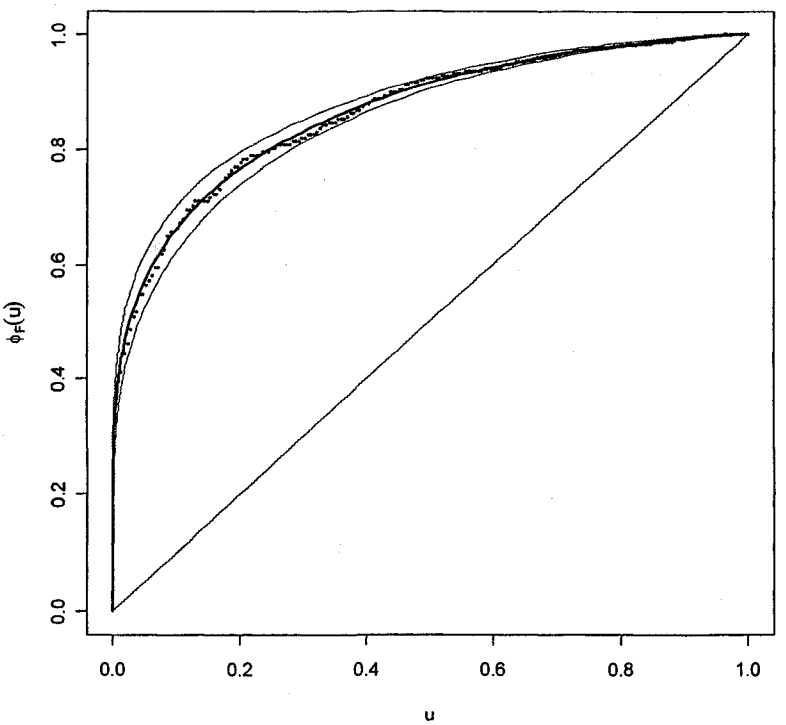

Figure 3.4 Total time on test transforms of the mice data. Scaled empirical (dark dotted line), scaled Odd Weibull fit (dark solid line), 5% confidence band (light solid lines).

Table 3.5b provides the likelihood ratio statistics of the Odd Weibull model for the two null hypotheses $H_{011} : \beta = 1$ to the Weibull submodel (Λ_{11}) and $H_{021} : \beta = -1$

to the inverse Weibull submodel (Λ_{21}). These values indicate an adequate fit to the Weibull model, whereas the inverse Weibull fit is inappropriate.

Table 3.5b Likelihood ratio tests of subhypotheses for mice data

Null hypothesis		χ^2_1	p-value
H_{011}	$-2 \cdot \ln(\Lambda_{11}) =$	2.62	0.11
H_{021}	$-2 \cdot \ln(\Lambda_{21}) =$	150.34	0.00

3.7.2 Bathtub-shaped failure rate and uncensored data

This example (data set is given in the appendix B) represents time to failure of 50 devices put on a life test at time zero (Aarset 1987). The empirical TTT procedure indicates a bathtub hazard shape for this data set (see dark dotted line in Figure 3.5). As before, when modeling bathtub-shaped failure rate data it is important to identify the parameter space of the Odd Weibull family, which actually produces a bathtub hazard shape. It could be observed that the major shape of the hazard function of the Odd Weibull family is bathtub-shaped when $1 < \alpha < \infty$, $0 < \alpha\beta \leq 1$. Also, our quick parameter estimation method, the percentile estimation method given in Section 3.3, provides the estimated parameter values of the Odd Weibull family as $\widetilde{\alpha} = 4.1694$, $\widetilde{\beta} = 0.1775$, $\widetilde{\theta} = 52.9565$.

As in the previous example, the estimated values are given in Table 3.6a.

Table 3.6a Estimated values of the Odd Weibull family for device data

Sample	$\widehat{\alpha} \pm SE_{\widehat{\alpha}}$	$\widehat{\beta} \pm SE_{\widehat{\beta}}$	$\widehat{\theta} \pm SE_{\widehat{\theta}}$	$l(\widehat{\boldsymbol{\theta}})$
Original	6.9657 ± 0.5903	0.0921 ± 0.0171	53.509 ± 2.4733	-215.88
Inverse	-0.0921 ± 0.0140	-0.0921 ± 0.0140	0.0187 ± 0.0006	92.02

In addition, the adequacy of the fit is strengthened by illustrating the scaled fitted TTT graph (see dark solid lines in Figure 3.5) with its 5% confidence band (solid lines). This indicates that the data is better-fit with the Odd Weibull family. The calculated R_n value of the device failure data is 0.0482, which is slightly smaller than the 95% upper percentage point of R_n, 0.0485. Therefore, it is hard to reject the Odd Weibull fit for the device failure data.

Mudholkar *et al.* (1996) used this example to illustrate the flexibility of their generalized Weibull family under the nonregular case. For comparison purposes, their generalized Weibull fit is illustrated in Figure 3.5 (see light dashed line).

Meanwhile, the likelihood ratio statistics given in Table 3.6b of the Odd Weibull family for the null hypotheses $H_{011} : \beta = 1$ and $H_{021} : \beta = -1$ corresponding to the Weibull and inverse Weibull submodels indicate their inappropriateness.

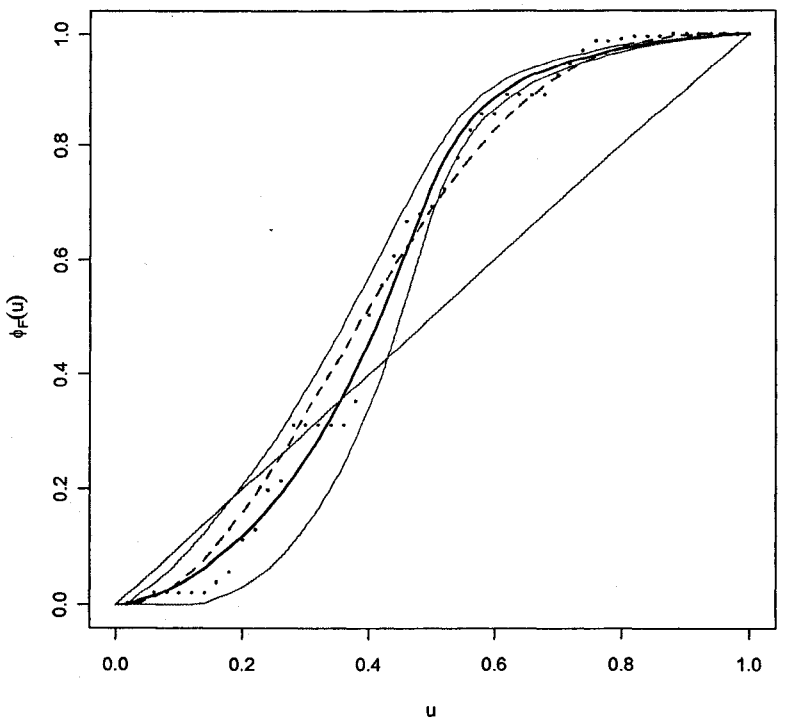

Figure 3.5 Total time on test transforms of the device data. Scaled empirical (dark dotted line), scaled Odd Weibull fit (dark solid line), scaled Mudholkar's generalized Weibull fit (light dashed line), 5% confidence band for the Odd Weibull fit (light solid lines).

Table 3.6b Likelihood ratio tests of subhypotheses for device data

Null hypothesis		χ^2_1	p-value
H_{011}	$-2 \cdot \ln(\Lambda_{11}) =$	50.65	0.00
H_{021}	$-2 \cdot \ln(\Lambda_{21}) =$	98.28	0.00

3.7.3 Unimodal failure rate and uncensored data

The twin data set (data sets are given in the appendix B) consists of alluvial diamonds from the Bougban and Damaya deposits in Guinea of West Africa (Beirlant *et al.* 1996). The sampling program on Bougban recovered 683 stones, whereas the Damaya sampling recorded 444 stones. In fact, the empirical TTT procedure gives unimodal hazard shapes for the twin data set (see dark dotted lines in Figures 3.6a and 3.6b). From Section 3.2, we observed that the hazard function of the Odd Weibull family is unimodal when its shape parameters, α and β, are both negative or $\alpha < 1$ and $\alpha\beta \geq 1$. Therefore, to estimate its parameters for unimodal failure rate data, one can pick the values from one of these ranges as an initial guess. The best-fitting parameters for the Odd Weibull family can be obtained by choosing the maximum value out of the following two likelihood functions, one from the range of both shape parameters $\alpha < 0, \beta < 0$ and the other function is from $\alpha < 1, \alpha\beta \geq 1$. The percentile estimation method given in Section 3.3, provides the estimated parameter values of the Odd Weibull family for Bougban as $\widetilde{\alpha} = -2.4026$, $\widetilde{\beta} = -0.7473$, $\widetilde{\theta} = 0.2232$ and for Damaya as $\widetilde{\alpha} = -2.0561$, $\widetilde{\beta} = -0.8561$, $\widetilde{\theta} = 0.3933$.

The estimated values are given in Table 3.7a. Specifically, the inverse samples of diamond data are analyzed by using the positive region (Section 3.4.1) of the Odd Weibull family.

The Odd Weibull fit to the original diamond data using scaled fitted TTT graphs (dark solid lines) with their 5% confidence bands (light solid lines) are illustrated in Figure 3.6a (Bougban) and Figure 3.6b (Damaya). The calculated R_n values of the Bougban and Damaya data are, respectively, 0.1111 and 0.0464, which are quite

smaller than their 90% upper percentage points of 4.9550 and 2.6762.

Table 3.7b gives the likelihood ratio statistics of the Odd Weibull family for the twin data set. These values indicate that the Weibull and inverse Weibull submodels are inappropriate for the diamond data.

Table 3.7a Estimated values of the Odd Weibull family for diamond data

	Original sample of		Inverse sample of	
	Bougban	Damaya	Bougban	Damaya
$\widehat{\alpha}$	-1.3431	-0.6107	1.3430	0.6107
$\pm SE_{\widehat{\alpha}}$	$\pm$ 0.1372	$\pm$ 0.1811	$\pm$ 0.1372	$\pm$ 0.1811
$\widehat{\beta}$	-1.2884	-2.9853	1.2885	2.9853
$\pm SE_{\widehat{\beta}}$	$\pm$ 0.1501	$\pm$ 0.9173	$\pm$ 0.1502	$\pm$ 0.9173
$\widehat{\theta}$	0.2008	0.2642	4.9807	3.7844
$\pm SE_{\widehat{\theta}}$	$\pm$ 0.0064	$\pm$ 0.0460	$\pm$ 0.1594	$\pm$ 0.6585
$l(\widehat{\boldsymbol{\theta}})$	113.49	-167.06	-1564.67	-773.02

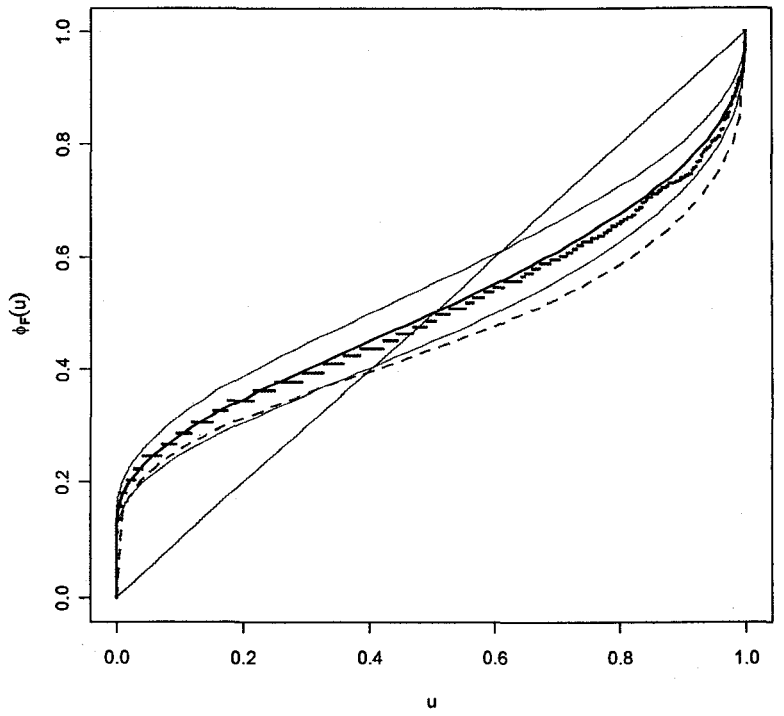

Figure 3.6a Total time on test transforms of the Bougban data. Scaled empirical (dark dotted line), scaled Odd Weibull fit (dark solid line), scaled Mudholkar's generalized Weibull fit (light dashed line), 5% confidence band (light solid lines).

Table 3.7b Likelihood ratio tests of subhypotheses for diamond data

Null	Bougban			Damaya		
hypothesis		χ^2_1	p-value		χ^2_1	p-value
H_{011}	$-2 \cdot \ln(\Lambda_{11}) =$	483.19	0.00	$-2 \cdot \ln(\Lambda_{11}) =$	230.90	0.00
H_{021}	$-2 \cdot \ln(\Lambda_{21}) =$	5.90	0.02	$-2 \cdot \ln(\Lambda_{21}) =$	46.79	0.00

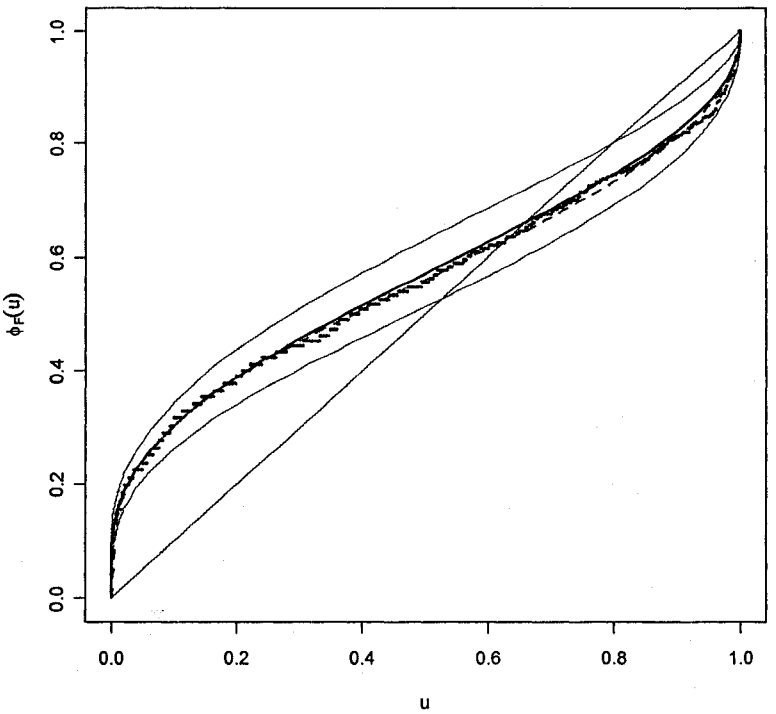

Figure 3.6b Total time on test transforms of the Damaya data. Scaled empirical (dark dotted line), scaled Odd Weibull fit (dark solid line), scaled Mudholkar's generalized Weibull fit (light dashed line), 5% confidence band (light solid lines).

3.7.4 Positively skewed density shape and grouped data

This example presents in Table 3.8 is a hospital-stay frequency distribution for 2311 schizophrenic patients taken from the Maryland Psychiatric Case Register. This data set was earlier analyzed by Eaton & Whitmore (1977) to discuss the appropriateness of the inverse Gaussian distribution as a model for the hospital stay pattern. Later, Whitmore (1986) noted that any simple model is inappropriate to explain the hospital stay pattern. Therefore, he formulated the normal-gamma mixture model to provide a clear improvement in fit relative to the unmixed inverse Gaussian model.

Table 3.8 The estimated values of the Odd Weibull family for hospital data

Stay		Observed	Fitted	Stay		Observed	Fitted
0	10	113	112.4490	160	170	24	33.8400
10	20	188	189.9422	170	180	41	30.9542
20	30	190	184.6042	180	190	33	28.4141
30	40	163	164.3855	190	200	20	26.1680
40	50	125	142.8408	200	300	164	179.6849
50	60	127	123.4291	300	400	102	99.5764
60	70	122	106.8266	400	500	78	63.1077
70	80	83	92.8782	500	600	48	43.5385
80	90	90	81.2119	600	700	33	31.8453
90	100	67	71.4399	700	800	23	24.3075
100	110	76	63.2195	800	900	14	19.1666
110	120	52	56.2659	900	1100	21	28.3075
120	130	51	50.3484	1100	1300	20	19.9163
130	140	38	45.2815	1300	1500	15	14.7866
140	150	43	40.9168	1500	2000	14	24.3998
150	160	44	37.1349	2000	∞	89	79.8125

$\widehat{\alpha} \pm SE_{\widehat{\alpha}} = -0.3623 \pm 0.0633$ $\quad l(\widehat{\boldsymbol{\theta}}) = -7448.9394$

$\widehat{\beta} \pm SE_{\widehat{\beta}} = -2.3778 \pm 0.4280$ $\quad \chi^2_{28} = 35.0182$

$\widehat{\theta} \pm SE_{\widehat{\theta}} = 30.7417 \pm 5.2278$ $\quad$ p-value $= 0.1693$

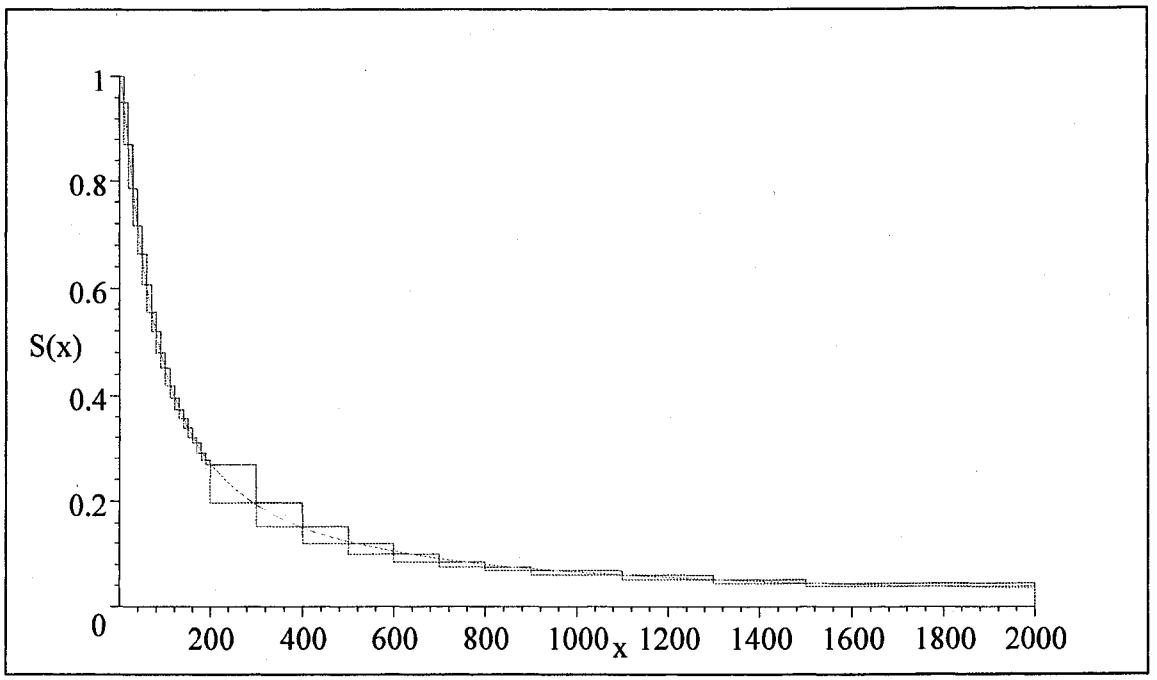

Figure 3.7 Fitted survival curves for the hospital-stay pattern data. The upper step function, the lower step function, and the solid line are, respectively, represent the Kaplan-Meier upper bound, Kaplan-Meier lower bound, and the fitted Odd Weibull curve.

Estimated log-likelihood values of unmixed inverse Gaussian and normal-gamma mixture model are, respectively, -7474.9 and -7452.9. He added an extra parameter for both models to provide an exact fit in class 0-10 days. Then the fitted chi-squared values and p-values of unmixed inverse Gaussian and normal-gamma mixture model are, respectively, $\chi^2_{28} = 82.7$, p-value $= 0.0000$ and $\chi^2_{26} = 44.2$, p-value $= 0.0144$. This is a large data set and hence this fit is much better than the inverse Gaussian fit. We used our Odd Weibull model to analyze this hospital-stay pattern and the estimated values of the fitted Odd Weibull family are given in Table 3.8. The fitted chi-squared value ($\chi^2_{28} = 35.0182$) and the p-value ($= 0.1693$) of the Odd Weibull

model indicate that it is a better model to analyze the hospital-stay pattern data. Furthermore, Figure 3.7 illustrate the fitted Odd Weibull survival curve along with the Kaplan-Meier survival band. But, we do not include the fitted survival curves for inverse Gaussian and normal-gamma mixture model, since they do not have closed-form survival functions.

3.7.5 Bimodal density shape and interval censored data

This interval censored data (data set is given in the appendix B) is taken from Ryan and Lindsey (1998) and is originally analyzed by Richman *et al.* (1990) regarding the drug resistance (time in months to resistance to Zidovudine) of 31 AIDS patients. To analyze this type of data Kaplan-Meier related nonparametric techniques have been developed, see for instance Peto (1973) and Turnbull (1976). Due to computational simplicity, Turnbull's (1976) method (self consistency iterative algorithm; R-codes for this algorithm are given in the appendix B) have been used by many authors. The Kaplan-Meier survival curves usually well bracket the Turnbull's (1976) survival curve. However, due to heavy censoring or may be specific configuration of the drug resistance data set, the Kaplan-Meier survival curves does not bracket the Turnbull's (1976) survival curve. Therefore Ryan and Lindsey (1998) recommended parametric models to analyze the drug resistance data set. They used Weibull, piecewise exponential and logspline models which we were not plotted in Figure 3.8, to track the survival curves of the drug resistance data set. Note that other parametric families may be better choices to estimate the true survival curve of the drug resistance data set.

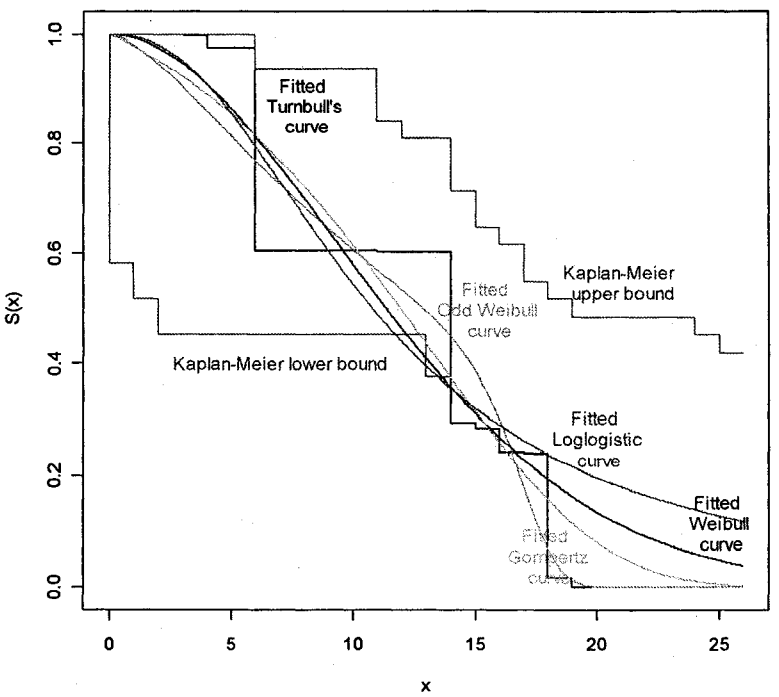

Figure 3.8 Fitted survival curves for interval censored resistance data.

In this regard, we used the drug resistance data set to illustrate the flexibility and applicability of the Odd Weibull family for modeling interval censored data. The estimated parameter and log-likelihood values of the Odd Weibull family for the drug resistance data are, respectively, $\widehat{\alpha} \pm SE_{\widehat{\alpha}} = 9.4445 \pm 14.0762$, $\widehat{\beta} \pm SE_{\widehat{\beta}} = 0.1569 \pm 0.2613$, $\widehat{\theta} \pm SE_{\widehat{\theta}} = 13.3759 \pm 3.0019$, $l(\widehat{\boldsymbol{\theta}}) = -19.0405$. Figure 3.8 illustrate the Odd Weibull survival curve along with the Kaplan-Meier survival curves. Also for comparison purposes, we plotted the fitted Gompertz and loglogistic survival curves in Figure 3.8. The Kaplan-Meier survival curves, as expected, well bracket the fitted Odd Weibull survival curve than the fitted Weibull survival curve. Also the right tail area of the Odd Weibull survival curve well matches with the Turnbull's (1976)

survival curve. Therefore, the Odd Weibull family may be useful to make further analysis of the interval censored drug resistance data set.

3.7.6 Negatively skewed density shape and left truncated interval censored data

This example is a left truncated and interval censored increasing failure rate twin data set (Pan and Chappell 1998, 2002), regarding the loss of functional independence of people of age 65 years or older. This twin data set (data are in the appendix B) consists of 421 non-poor male group and 609 non-poor female group. In this example, we do not motivate to provide nonparametric graphical comparison with Odd Weibull family since left truncated and interval censored data are a very complicated form of incomplete data arising in survival data analysis. Here, our aim is to show the flexibility of Odd Weibull family for modeling complicated form of incomplete data.

The estimated parameter and likelihood values of the Odd Weibull family for the functional independence data are;

Non-poor female: $\widehat{\alpha} \pm SE_{\widehat{\alpha}} = 10.0451 \pm 1.3515$, $\widehat{\beta} \pm SE_{\widehat{\beta}} = 0.9808 \pm 0.2452$, $\widehat{\theta} \pm SE_{\widehat{\theta}} = 81.4178 \pm 1.2653$, $l(\widehat{\boldsymbol{\theta}}) = -629.9040$.

Non-poor male: $\widehat{\alpha} \pm SE_{\widehat{\alpha}} = 6.9247 \pm 1.1200$, $\widehat{\beta} \pm SE_{\widehat{\beta}} = 0.6583 \pm 0.1903$, $\widehat{\theta} \pm SE_{\widehat{\theta}} = 71.4860 \pm 5.0491$, $l(\widehat{\boldsymbol{\theta}}) = -473.2271$.

The fitted Odd Weibull survival curves, for the non-poor female (dotted line) and for the non-poor male (solid line), are given in Figure 3.9

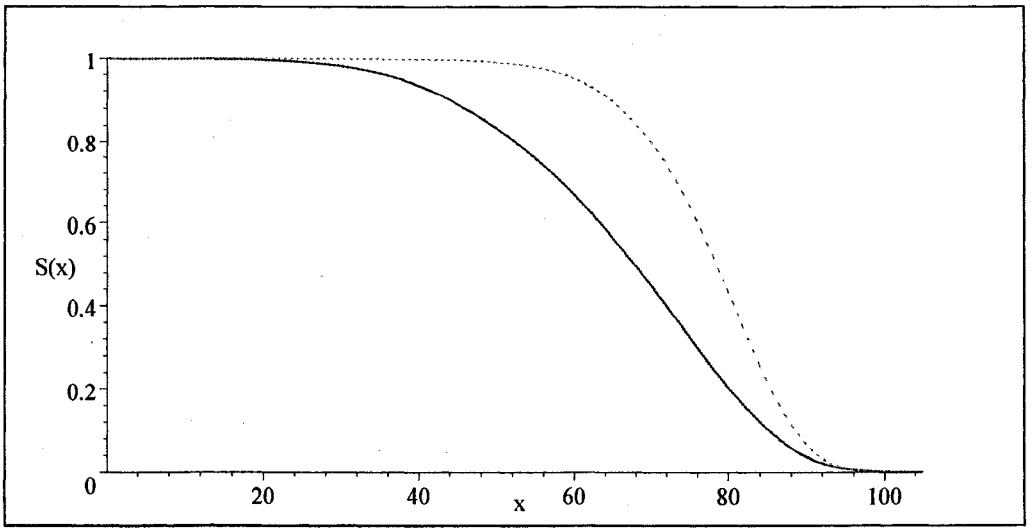

Figure 3.9 Fitted survival curves for functional independence data. Non-poor female (dotted line) and non-poor male (solid line).

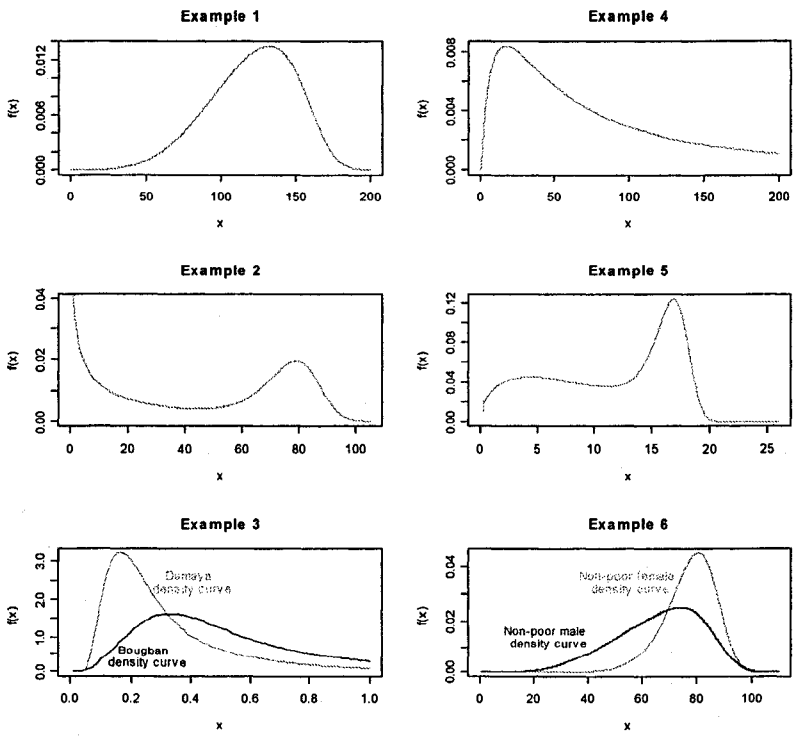

Figure 3.10 Fitted Odd Weibull density curves for examples 1 through 6.

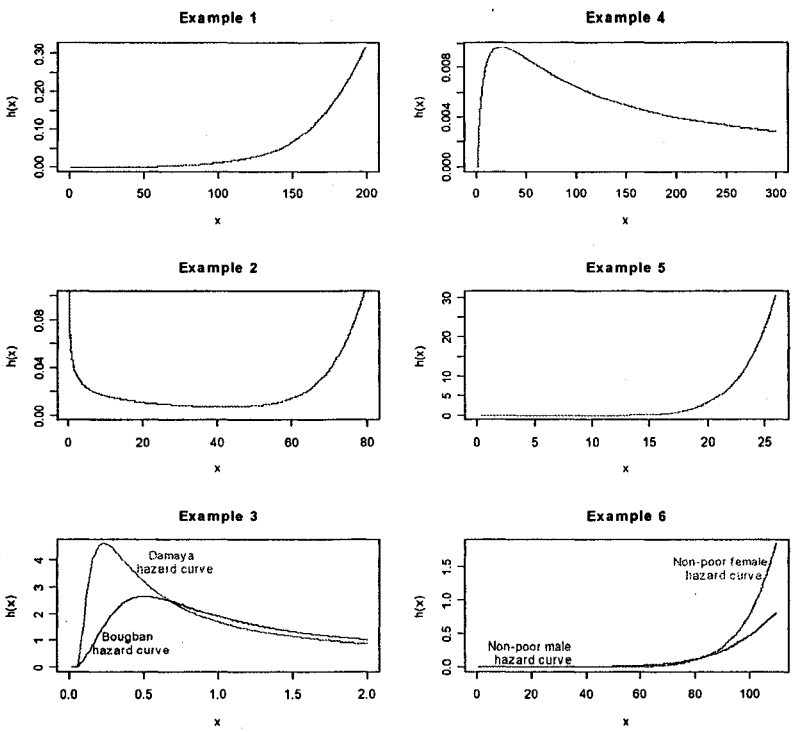

Figure 3.11 Fitted Odd Weibull hazard curves for examples 1 through 6.

3.8 Exponentiality test

A graphical device called *isotones* (equal tensions or strengths) is first formulated by Lin (1977) and later used by Mudholkar *et al.* (1991) for testing normality of the Shapiro-Wilk $W-$ test, Vasicek's entropy test, and Lin and Mudholkar's Z_p- test by the generalized Tukey lambda family. Also Kollia (1989) used this graphical device to construct isotones of four tests namely; χ^2_{CSY} test (Csörgő *et al.* 1975), entropy test $K_{3,20}$, the two-sided bivariate F-test (Lin and Mudholkar 1980), and Gnedenko's (Gnedenko *et al.* 1969) $Q(r)$ test for testing the exponentiality by the generalized Weibull family (Mudholkar *et al.* 1996).

In this section, we used isotones to compare and assess the accuracy for testing the exponentiality of Kolmogorov-Smirnov ($T(D)$, 1933), Anderson-Darling ($T(A)$,

1954), and Cramér von Mises ($T(U)$, 1937) test statistics by the Odd Weibull family. $T(D)$, $T(A)$, and $T(U)$ are modified forms for testing exponentiality given by the following equations.

$$T(D) = (D - 0.2/n)(\sqrt{n} + 0.26 + 0.5/\sqrt{n}).$$

$$T(A) = A^2(1.0 + 0.6/n).$$

$$T(U) = U^2(1.0 + 0.16/n).$$

Where D and A^2 are given in Chapter II Section 2.8, and

$$U^2 = \sum_{i=1}^{n} \left\{F(x_{(i)}) - (2i - 1)/(2n)\right\}^2 + 1/(12n).$$

The isotones are based on *ideal samples* called *profiles* from the members of the two-shape parameter ($\lambda = \alpha$, $\mu = \beta/\alpha$, $\theta = 1$) Odd Weibull family. The sensitivity surface of the values of the goodness-of-fit test statistic calculated from these profiles on the (λ, μ) plane for the three statistics are, respectively, illustrate in Figure 3.12a, 3.12b, and 3.12c. The isotones are the contours of these surfaces and are, respectively, illustrate in Figure 3.13a, 3.13b, and 3.13c. The contours visualize the exponentiality departure of the statistics on the (λ, μ) plane starting from $\lambda = 1$, $\mu = 1$. A sample size of 50 (see Mudholkar *et al.* 1991) is used to construct these surfaces and their contours.

Figure 3.14a, 3.14b, and 3.14c are, respectively, illustrate the upper tail 90% and 95% probability contours of $T(D)$, $T(A)$, and $T(U)$ for testing exponentiality with sample size 50 by the Odd Weibull family. Also, figure 3.15a, 3.15b, and 3.15c are, respectively, illustrate the upper tail 90% probability contours with sample size 20 and

50 of $T(D)$, $T(A)$, and $T(U)$ for testing exponentiality by the Odd Weibull family. These upper percentage values are well tabulated by Pearson and Hartley (1972). From the two set of figures (3.14 and 3.15), one can clearly see that the isotones shrink closer to the exponential point as rejecting probability level (α) or sample size (n) increase. This is analogous to the increase in power as α or n increases.

Figure 3.16a and 3.16b are, respectively, illustrate the 90% and 95% probability contours of superimposition of the three statistics with sample size 50 for testing exponentiality on the (λ, μ) plane. From these two figures, we can clearly see that $T(A)$ and $T(U)$ have relatively higher strength for testing exponentiality than the $T(D)$ test statistic. As we expected, the square tests ($T(A)$ and $T(U)$) are more powerful than the $T(D)$. Also, EDF based test is more powerful than the chi-squared tests.

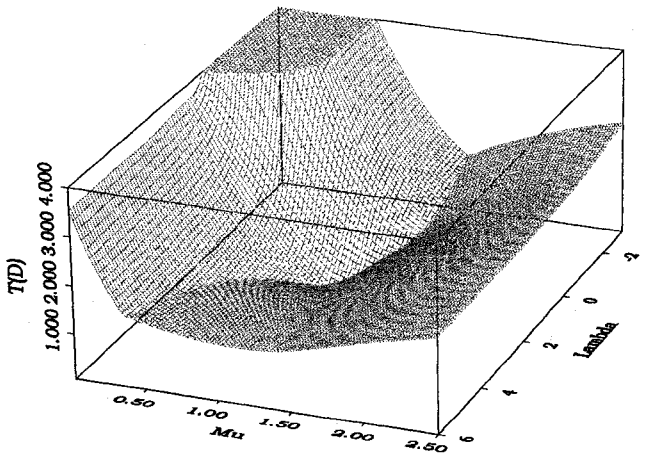

Figure 3.12a Kolmogorov-Smirnov sensitivity surface.

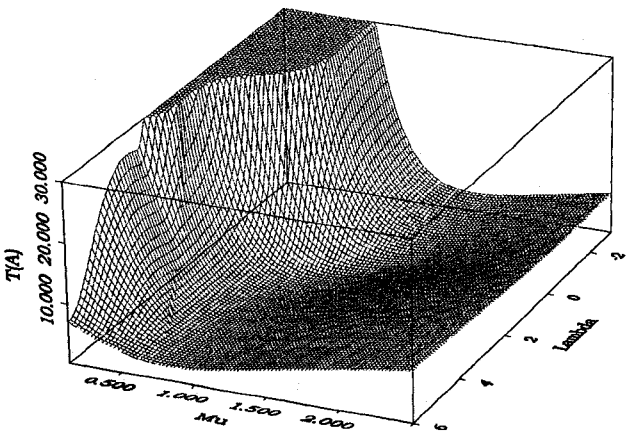

Figure 3.12b Anderson-Darling sensitivity surface.

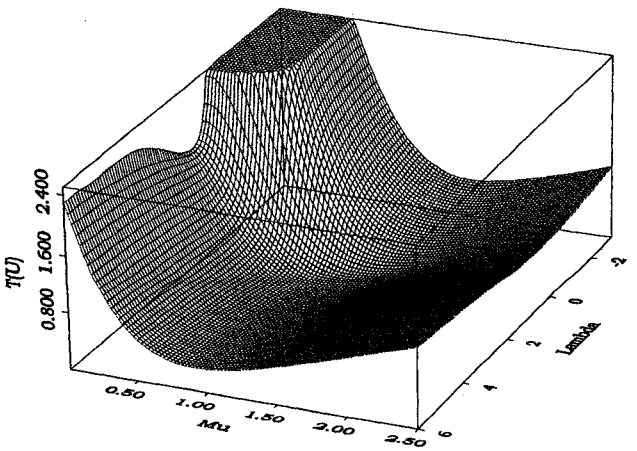

Figure 3.12c Cramér von Mises sensitivity surface.

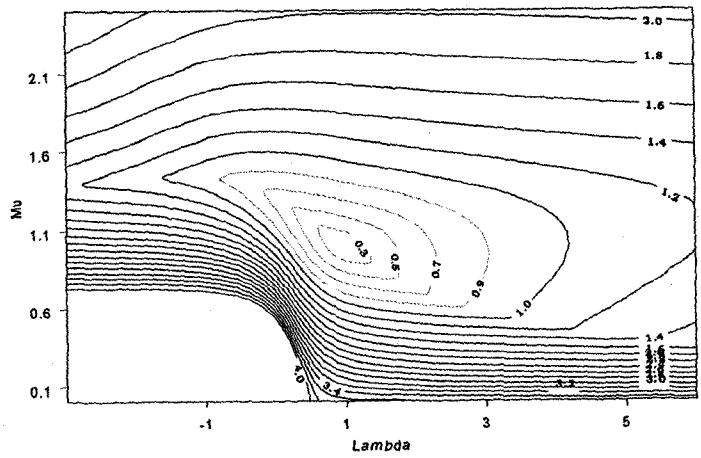

Figure 3.13a Kolmogorov-Smirnov contour plot.

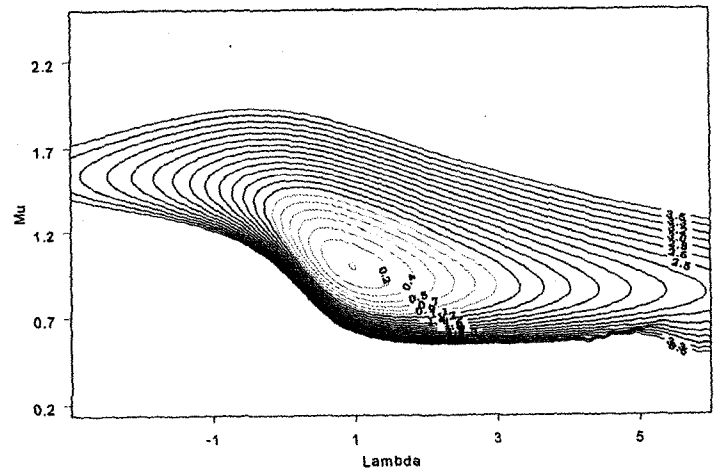

Figure 3.13b Anderson-Darling contour plot.

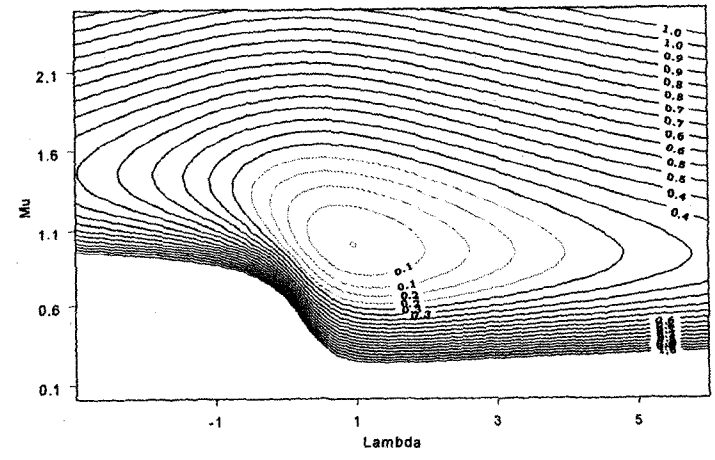

Figure 3.13c Cramér von Mises contour plot.

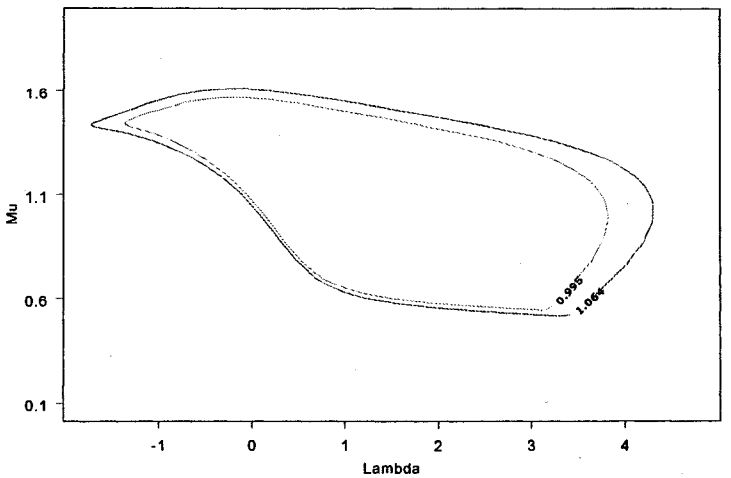

Figure 3.14a The $\alpha = 0.05$ and 0.10 of $T(D)$.

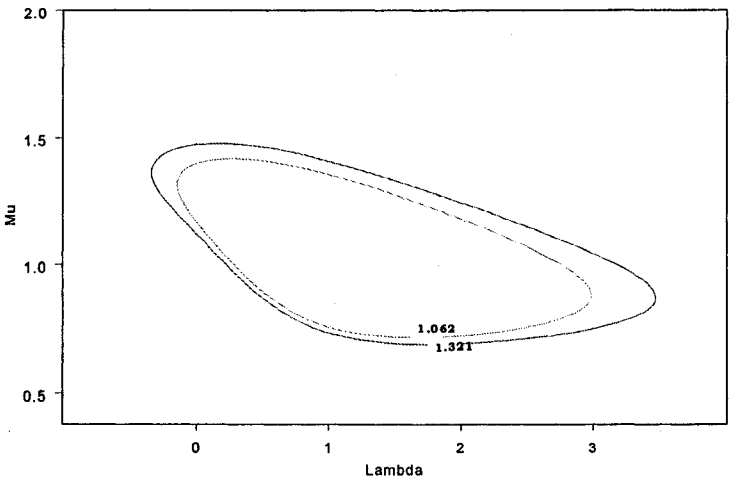

Figure 3.14b The $\alpha = 0.05$ and 0.10 of $T(A)$.

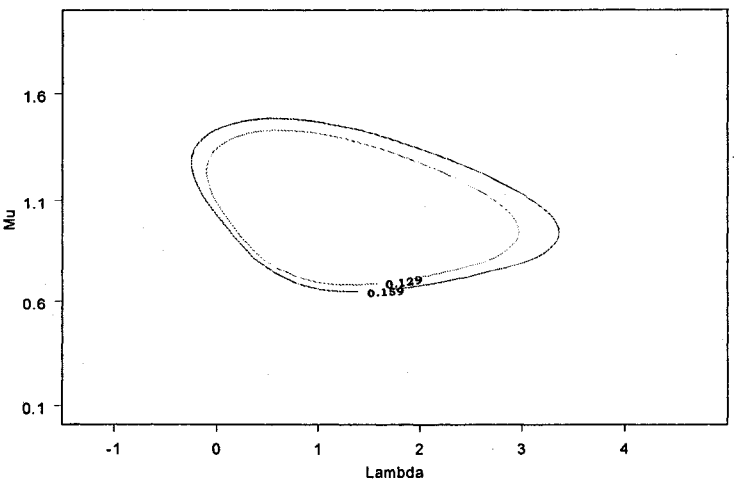

Figure 3.14c The $\alpha = 0.05$ and 0.10 of $T(U)$.

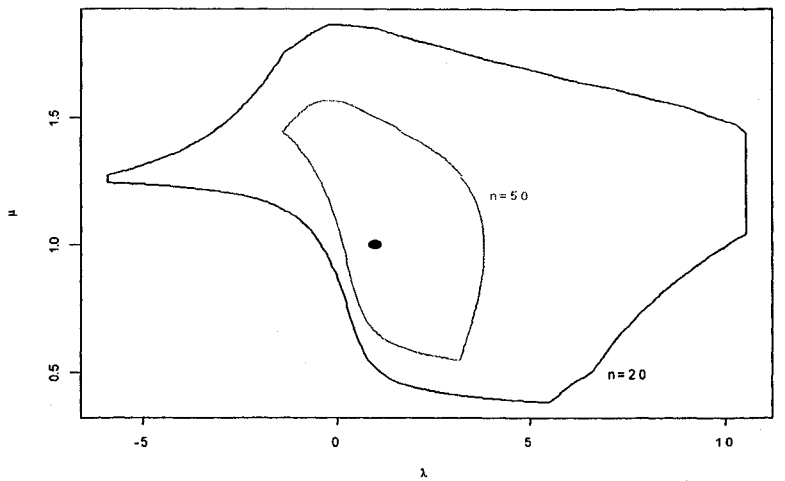

Figure 3.15a The $\alpha = 0.10$ of $T(D)$ for $n = 20$, and $n = 50$.

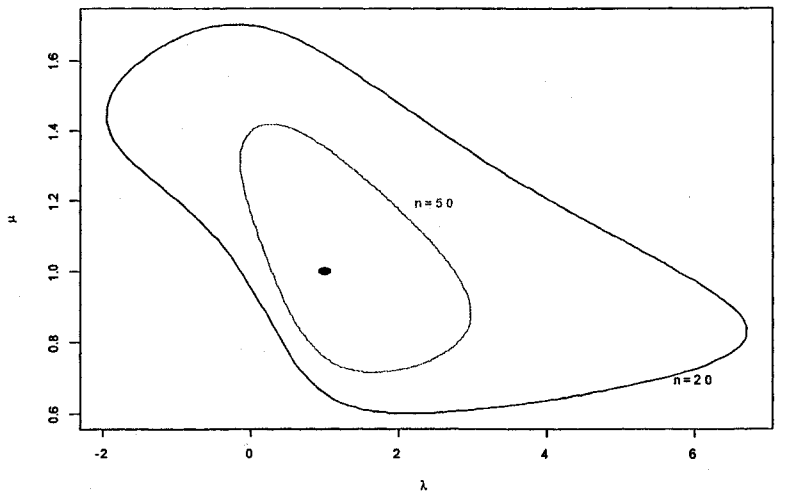

Figure 3.15b The $\alpha = 0.10$ of $T(A)$ for $n = 20$, and $n = 50$.

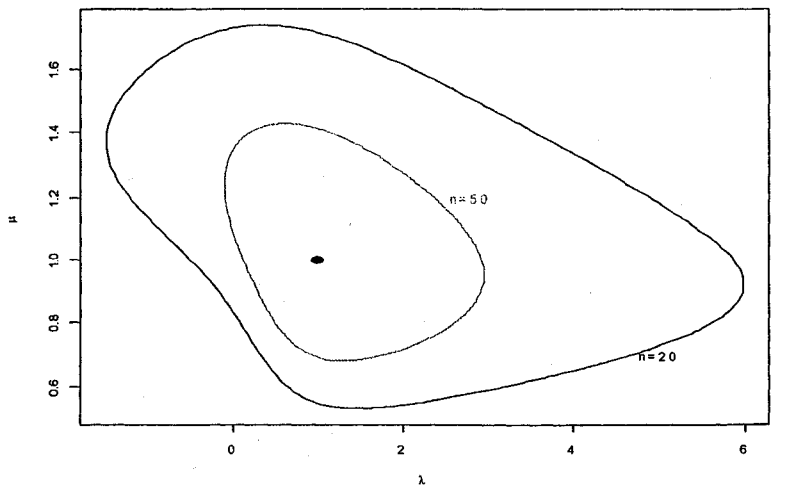

Figure 3.15c The $\alpha = 0.10$ of $T(U)$ for $n = 20$, and $n = 50$.

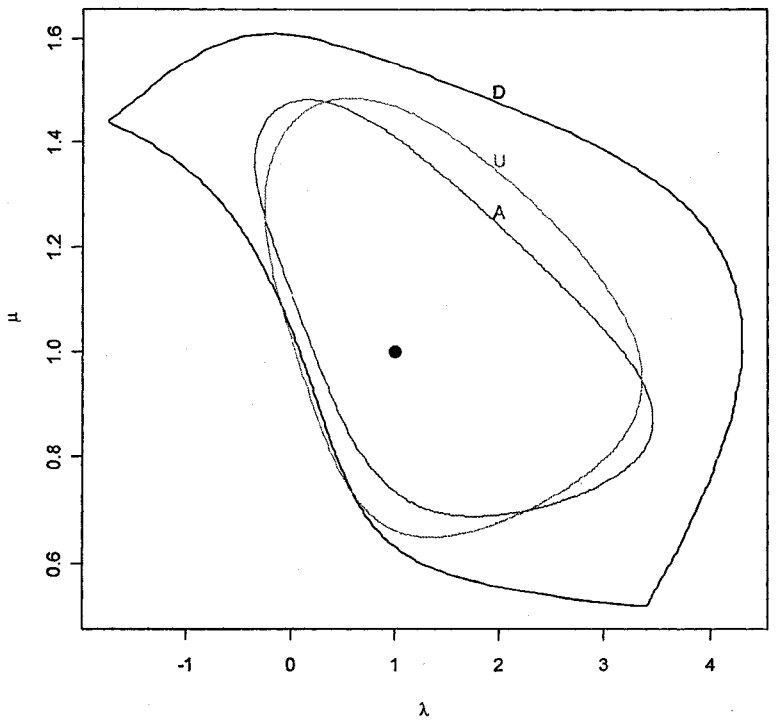

Figure 3.16a The $\alpha = 0.05$ of $T(D)$, $T(A)$, and $T(U)$.

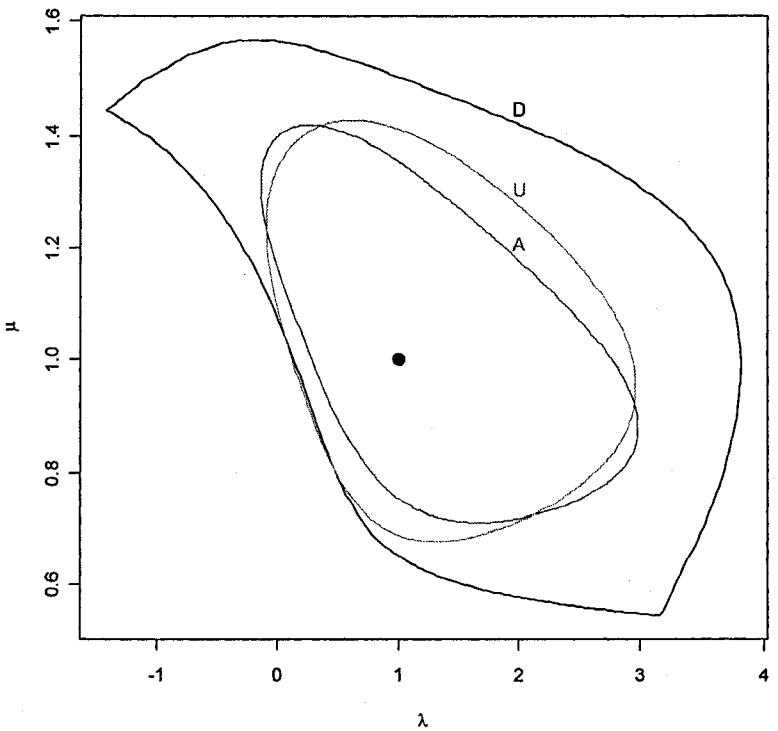

Figure 3.16b The $\alpha = 0.10$ of $T(D)$, $T(A)$, and $T(U)$.

3.9 Odd Weibull aliases of some common distributions via Galton's skewness and Moor's kurtosis

To enlighten the structure of the Odd Weibull family, one can compare it with commonly available parametric families in the coefficient of skewness ($\sqrt{\beta_1}$) and kurtosis (β_2) plane. Similar analysis has done by Johnson *et al.* (1982) and Mudholkar *et al.* (1994, 1996) using the (β_1, β_2) plane. However, these analysis are limited to the distributions with finite moments. For example, Cauchy related distributions such as half Cauchy, folded Cauchy, truncated Cauchy, and log Cauchy which are use in lifetime data analysis, cannot be compared with other commonly available parametric families in the (β_1, β_2) plane. Furthermore, some two-parameter common distributions such as loglogistic, Pareto, inverse Pareto, inverse Weibull, inverse Gamma, paralogistic, inverse paralogistic, and etc., do not have a finite moment of some values of their shape parameter space. Also, some Odd Weibull aliases do not have a finite moments and hence creating a (β_1, β_2) or ($\sqrt{\beta_1}$, β_2) plane for the Odd Weibull family is less important.

It is well known, the moment base kurtosis value is 6 for the Laplace and infinite for the Cauchy. Balanda (1987) has pointed out that the moment base comparison is inadequate for the Laplace and the Cauchy distributions. Since it does not recognize the dominant features such as the Laplace's dramatic peak and the Cauchy's long tail. Furthermore, Horn (1983) identified that the Laplace has more peaked than that of the Cauchy, while Rosenberger and Gasko (1983) identified the heavier tail of the Cauchy than the Laplace. The skewness and kurtosis comparison of the two models may support this argument.

As an answer to the above discussion we can use quantile function based skewness (S) and kurtosis (K) plane to compare almost all of the commonly available parametric families. This quantile based idea has been mentioned by Moor (1988), but a continuation of such work does not appear during the last two decades. However, our aim here is not just to create a quantile based (S, K) plane, but also to find the Odd Weibull aliases in the (S, K) plane.

Galton's (1883) measure of skewness ($-1 \leq S \leq 1$) is defined by

$$S = \frac{(\text{Upper quartile} - \text{Median}) + (\text{Lower quartile} - \text{Median})}{\text{Interquartile distance}}. \tag{3.31}$$

Moor's (1988) measure of kurtosis ($K > 0$) is defined by

$$K = \frac{(\text{Seventh octile} - \text{Fifth octile}) + (\text{Third octile} - \text{First octile})}{\text{Interquartile distance}}. \tag{3.32}$$

One can easily obtain the S and K values for the Odd Weibull distribution ($\lambda = \alpha$, $\mu = \beta/\alpha$, $0 < \mu < \infty$, $-\infty < \lambda < \infty$) as

$$S = \frac{\ln^{1/\lambda}\left(1+3^{\lambda/\mu}\right) + \ln^{1/\lambda}\left(1+3^{-\lambda/\mu}\right) - 2\ln^{1/\lambda} 2}{\ln^{1/\lambda}\left(1+3^{\lambda/\mu}\right) - \ln^{1/\lambda}\left(1+3^{-\lambda/\mu}\right)}, \tag{3.33}$$

and

$$K = \frac{\ln^{1/\lambda}\left(1+7^{\lambda/\mu}\right) - \ln^{1/\lambda}\left(1+7^{-\lambda/\mu}\right) - \ln^{1/\lambda}\left(1+(5/3)^{\lambda/\mu}\right) + \ln^{1/\lambda}\left(1+(5/3)^{-\lambda/\mu}\right)}{\ln^{1/\lambda}\left(1+3^{\lambda/\mu}\right) - \ln^{1/\lambda}\left(1+3^{-\lambda/\mu}\right)}. \tag{3.34}$$

Note that skewness and kurtosis values does not depend on the location and scale parameters of a distribution. The Quantile, Galton's skewness and Moor's kurtosis

functions for various continuous univariate distributions are given in the appendix A. In order to identify such distributions the following abbreviations are used.

Uniform (U), Normal (N), Logistic (L), Laplace (LA), Cauchy (C), Sinh-normal (SN), Sinh-logistic (SL), Sinh-Cauchy (SC), Smallest extreme (SEV), Largest extreme (Gumbel) (LEV), Half-normal (HN), Half-logistic (HL), Half Laplace (Exponential) (EXP), Half-Cauchy (HC), Folded-logistic (FL), Folded Laplace (FLA), Folded-Cauchy (FC), Lognormal (LN), Loglogistic (LL), LogCauchy (LC), Power distribution (PO), Pareto distribution (PA), Weibull Distribution (W), Gamma distribution (GA), Gompertz distribution (G), Logistic-sinh distribution (LS), Gompertz-sinh distribution (GS), Birnbaum-Saunders distribution (BS), Inverse exponential distribution (IEXP), Inverse Weibull distribution (IW), Lognormal-Pareto composite distribution (LPC), Weibull-Pareto composite distribution (WPC), Weibull-inverse Weibull composite distribution (WIW). Note that due to the computational difficulties of calculating the (S, K) values for folded normal and inverse Gaussian distribution, we will ignore them in our current study.

Figure 3.17a represents some common distributions (maximum one shape parameter) in the (S, K) plane, whereas its magnified graph is given in Figure 3.17b. The location-scale parametric families gives a single point on the (S, K) plane, whereas a single curve represents one shape parameter family of distributions. From these two graphs, one can see that most of the commonly available distributions are crowded in a specific region in the (S, K) plane, and this may be the reason most natural data sets arise from that region.

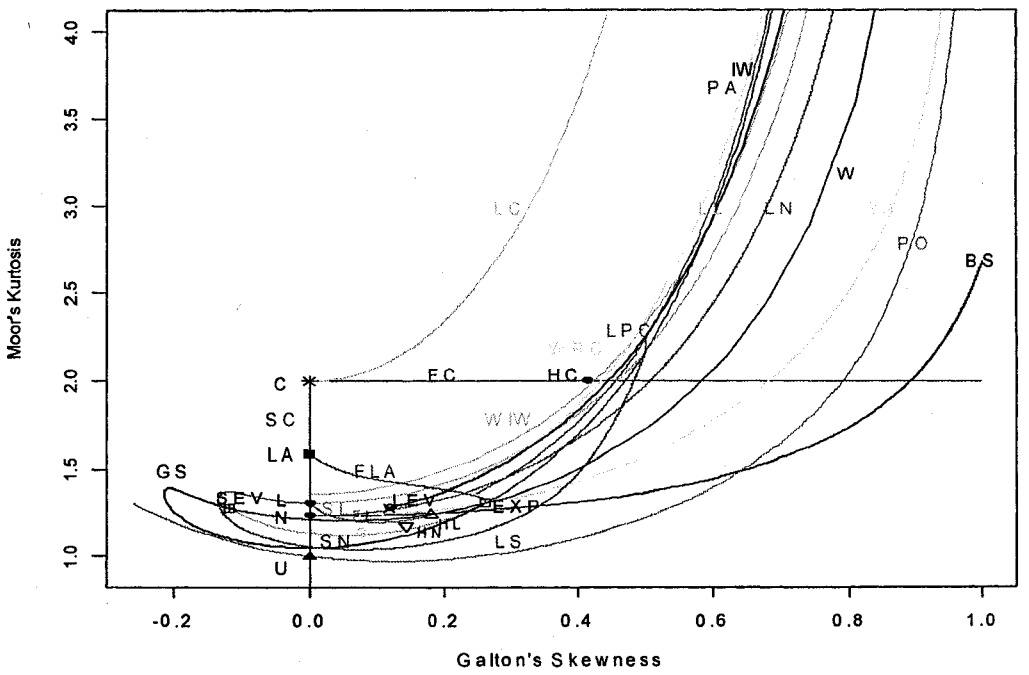

Figure 3.17a Some common distributions in the (S, K) plane.

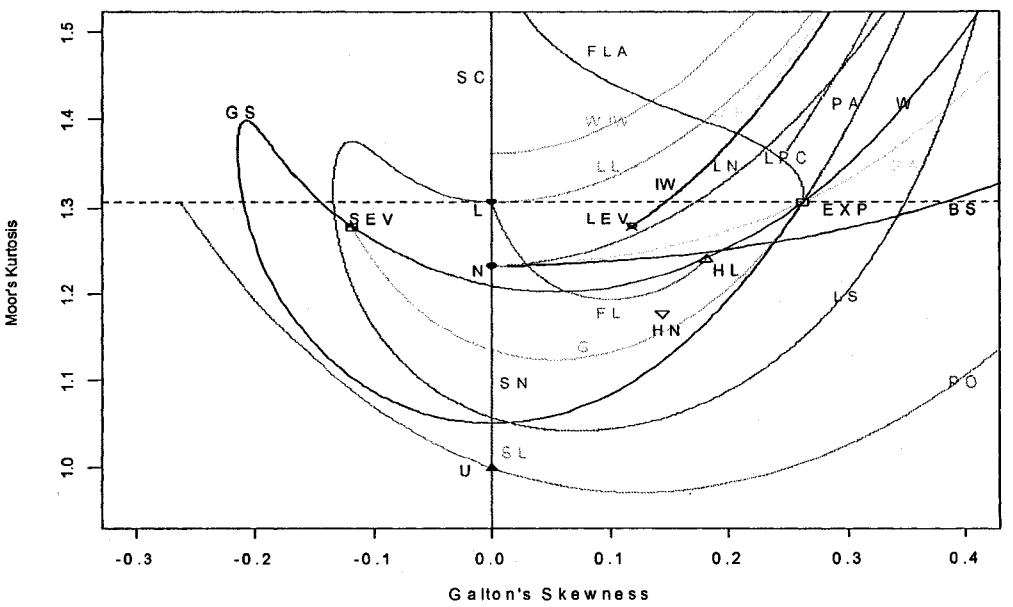

Figure 3.17b Some common distributions in the (S, K) plane (magnified).

Figure 3.18a and 3.18b, respectively, represent the Galton's skewness surface and contours of the Odd Weibull family on the (λ, μ) plane. Figure 3.18c and 3.18d, respectively, represent the Moor's kurtosis surface and contours of the Odd Weibull family on the (λ, μ) plane. Figure 3.19 shows the Odd Weibull aliases of some common distribution on the (λ, μ) plane.

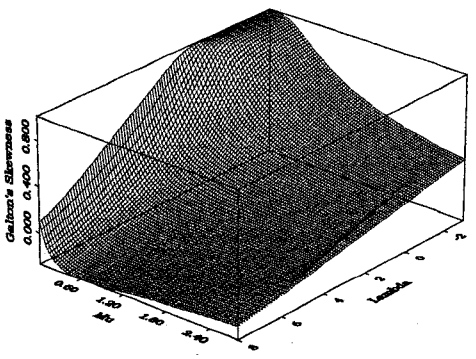

Figure 3.18a Galton's skewness surface.

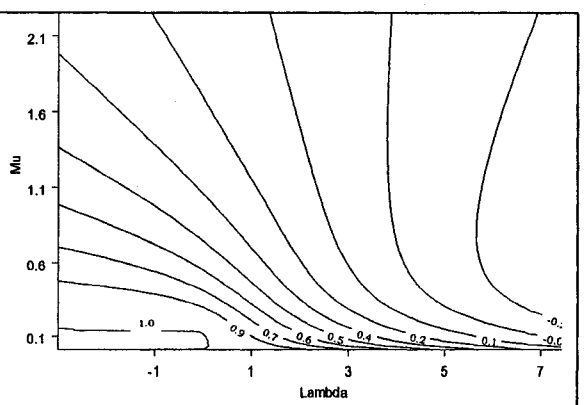

Figure 3.18b Galton's skewness contours.

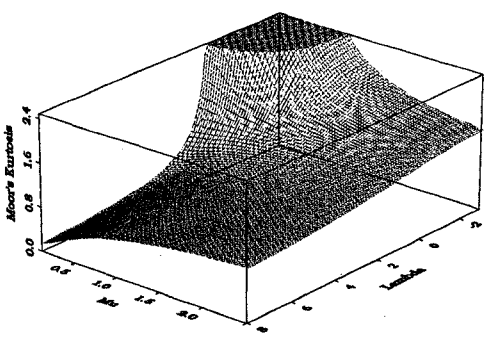

Figure 3.18c Moor's kurtosis surface.

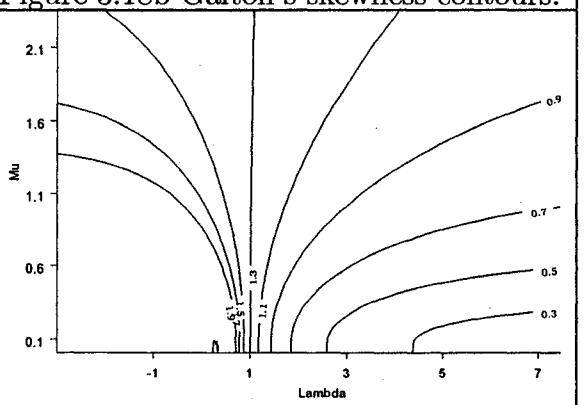

Figure 3.18d Moor's kurtosis contours.

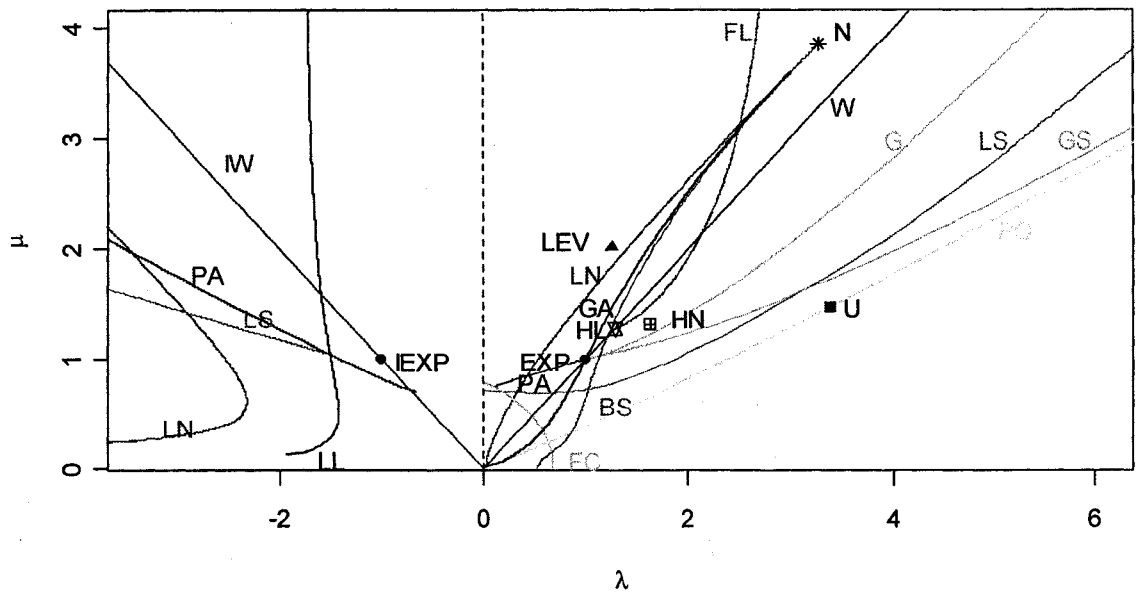

Figure 3.19 Odd Weibull aliases of some common distributions.

3.10 The exponential Odd Weibull family

An exponential transformation of the Odd Weibull family leads to the following distribution.

$$F(x;\beta,\mu,\sigma) = 1 - \left(1 + \left(e^{e^{\left(\frac{x-\mu}{\sigma}\right)}} - 1\right)^{\beta}\right)^{-1}, \tag{3.35}$$

where $-\infty < x < \infty$, $-\infty < \mu < \infty$, and $0 < \beta\sigma < \infty$.

When $\beta = 1$ this exponential Odd Weibull (EOW) model reduces to the smallest extreme value (SEV) distribution, whereas, when $\beta = -1$ this exponential Odd Weibull (EOW) model reduces to the largest extreme value (LEV, Gumbel) distribution.

The density and the quantile function of the distribution are, respectively,

$$f(x;\beta,\mu,\sigma)=\frac{\beta}{\sigma}e^{\left(\frac{x-\mu}{\sigma}\right)}e^{e^{\left(\frac{x-\mu}{\sigma}\right)}}\left(e^{e^{\left(\frac{x-\mu}{\sigma}\right)}}-1\right)^{\beta-1}\left(1+\left(e^{e^{\left(\frac{x-\mu}{\sigma}\right)}}-1\right)^{\beta}\right)^{-2}, \tag{3.36}$$

and

$$Q(u)=\mu+\sigma\ln\ln\left\{1+\left(\frac{u}{1-u}\right)^{1/\beta}\right\}. \tag{3.37}$$

Where $-\infty<x<\infty$, $-\infty<\mu<\infty$, $0<\beta\sigma<\infty$, and $0\leq u\leq 1$.

The Galton's skewness and Moor's kurtosis functions for the EOW family can respectively be written as

$$G=\frac{\ln\left\{\ln\left(1+3^{1/\beta}\right)\ln\left(1+3^{-1/\beta}\right)/\left(\ln 2\right)^{2}\right\}}{\ln\left\{\ln\left(1+3^{1/\beta}\right)/\ln\left(1+3^{-1/\beta}\right)\right\}}, \tag{3.38}$$

and

$$K=\frac{\ln\left\{\frac{\ln\left(1+7^{1/\beta}\right)\ln\left(1+(5/3)^{-1/\beta}\right)}{\ln\left(1+7^{-1/\beta}\right)\ln\left(1+(5/3)^{1/\beta}\right)}\right\}}{\ln\left\{\ln\left(1+3^{1/\beta}\right)/\ln\left(1+3^{-1/\beta}\right)\right\}}, \tag{3.39}$$

where $-\infty<\beta<\infty$.

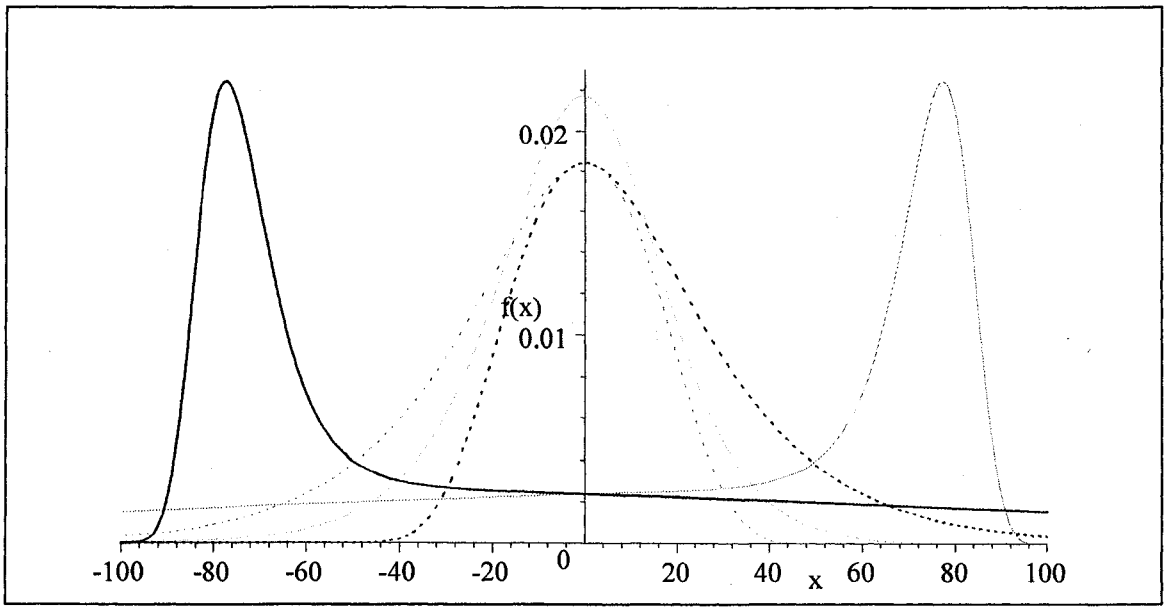

Figure 3.20 Exponential Odd Weibull density curves. The dark solid line ($\mu = -50$, $\sigma = -10$, $\beta = -0.1$), the light solid line ($\mu = 50$, $\sigma = 10$, $\beta = 0.1$), the dark dotted line ($\mu = 0$, $\sigma = -20$, $\beta = -1$), the light dotted line ($\mu = 0$, $\sigma = 20$, $\beta = 1$), and the dashed line ($\mu = 28$, $\sigma = 80$, $\beta = 5$).

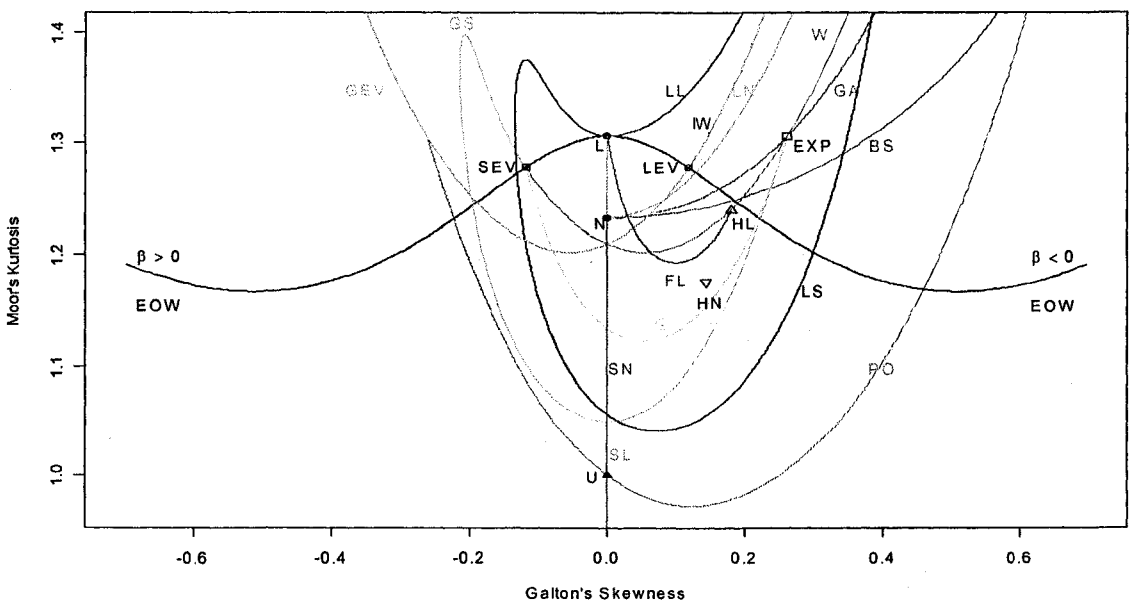

Figure 3.21 Shape of EOW and GEV distributions in the (S, K) plane

3.10.1 Analysis of wave-surge data

This example is a concurrent measurements of two oceanographic variables - wave and surge height at a single location off south-west England (Coles 2001). This is a large data set with 2894 data points (see appendix B) for each variables - wave and surge heights measured in meters. As noted by Coles (2001), the scatter plot of wave and surge data suggests a tendency for extremes of one variable to coincide with extremes of the other. The importance of this example is to identify such phenomenon, as the impact of simultaneous extremes may be much greater than if extremes of either component occur in isolation. Therefore to calculate the probability of simultaneously extreme events, multivariate extreme value models can be used. For this purpose, Coles (2001) used three different bivariate models: logistic (one parameter), bilogistic (two parameters) and Dirichlet (two parameters) for wave-surge data under the point process analysis. Maximized log-likelihood values of three bivariate models indicate that the Dirichlet model is a better fit for wave-surge data. To apply these bivariate models, each variables, wave and surge heights are individually modeled by the generalized extreme value (GEV) distribution and the parameters are estimated under the likelihood method, and then the data are transformed according to the standard Fréchet scale.

Alternative to the GEV distribution, one can used better extreme value models to transform the data according to the standard Fréchet scale to apply the above mentioned bivariate models. Therefore, we used the exponential Odd Weibull (EOW) distribution for individual modeling of these variables and the fitness is compared with Gumbel and GEV distributions. Estimated values of these three models are

given in Table 3.9 and Table 3.10. These estimated values indicate that the EOW model gives a better fit than the Gumbel or GEV models for both wave and surge data. Alternatively, the other known parametric families gave a very poor fit than the EOW fit given on Table 3.9 and 3.10.

The chi-squared tests given in Tables 3.9 and 3.10 are performed according to the computer generated (arbitrarily chosen) class intervals for wave and surge height data. It would be more appropriate if we use equal probability class intervals. The number of classes can be chosen by the formula (D'Agostino and Stephens 1986), $M \approx 2n^{2/5}$, where M and n are, respectively, the number of classes and the sample size. Then $M \approx 48$ for both wave and surge data. Hence, fitted chi-squared values and p-values for wave and surge data are given in Table 3.11. The pros and cons of chi-squared test and related methodologies are found in Greenwood and Nikulin (1996).

Table 3.9 Estimated values of three models for wave height data

Height interval (in meters)		Observed frequency	Expected frequencies for Gumbel	GEV	EOW
$-\infty$	0.625	18	64.1174	32.1974	21.5226
0.625	0.875	78	70.8973	65.8626	64.1641
0.875	1.125	131	110.7791	122.0659	137.3998
1.125	1.375	211	152.2238	178.8105	205.4916
1.375	1.625	217	188.5474	221.9692	241.2622
1.625	1.875	239	214.7636	245.1004	245.7542
1.875	2.125	254	228.6314	249.0507	232.8437
2.125	2.375	216	230.4802	238.5196	213.5211
2.375	2.625	205	222.3554	219.0020	193.1621
2.625	2.875	188	207.0552	195.1452	173.6982
2.875	3.125	143	187.3871	170.2461	155.5782
3.125	3.375	126	165.7373	146.3468	138.7694
3.375	3.625	118	143.9034	124.5391	123.1561
3.625	3.875	109	123.0999	105.2771	108.6621
3.875	4.125	97	104.0497	88.6271	95.2638
4.125	4.375	76	87.1037	74.4413	82.9692
4.375	4.625	63	72.3539	62.4707	71.7927
4.625	4.875	63	59.7276	52.4321	61.7373
4.875	5.125	49	49.0576	44.0456	52.7850

continued

Table 3.9 (continued)

Height interval (in meters)		Observed frequency	Expected frequencies for Gumbel	GEV	EOW
5.125	5.375	48	40.1311	37.0538	44.8948
5.375	5.625	35	32.7222	31.2291	38.0055
5.625	5.875	41	26.6114	26.3759	32.0407
5.875	6.125	30	21.5962	22.3286	26.9149
6.125	6.375	20	17.4965	18.9488	22.5389
6.375	6.625	22	14.1557	16.1214	18.8242
6.625	6.875	19	11.4403	13.7513	15.6860
6.875	7.125	18	9.2376	11.7602	13.0459
7.125	7.375	10	7.4537	10.0835	10.8326
7.375	7.625	10	6.0109	8.6682	8.9826
7.625	7.875	9	4.8451	7.4705	7.4400
7.875	8.125	11	3.9040	6.4544	6.1564
8.125	8.375	8	3.1447	5.5902	5.0903
8.375	∞	12	12.9797	42.0148	24.0136
$\widehat{\mu}\pm\ SE_{\widehat{\mu}}$			2.1625 ± 0.0224	2.0739 ± 0.0232	2.2106 ± 0.0224
$\widehat{\sigma}\pm\ SE_{\widehat{\sigma}}$			1.1495 ± 0.0174	1.0744 ± 0.0182	$-0.7400\pm\ 0.0240$
$\widehat{\beta}\pm\ SE_{\widehat{\beta}}$				0.1476 ± 0.0173	$-0.5743\pm\ 0.0230$
$l(\widehat{\boldsymbol{\theta}})$			-5068.6830	-5026.0151	-4996.7272
χ^2_{df}			$\chi^2_{28} = 161.8114$	$\chi^2_{29} = 77.9198$	$\chi^2_{29} = 33.5904$
p-value			0.0000	0.0000	0.2546

Table 3.10 Estimated values of three models for surge height data

Height interval (in meters)		Observed frequency	Expected frequencies for Gumbel	GEV	EOW
$-\infty$	-0.2875	7	0.5621	3.1221	10.4720
-0.2875	-0.2625	8	1.9241	4.8425	7.9928
-0.2625	-0.2375	10	6.0040	10.1775	12.8678
-0.2375	-0.2125	39	14.9255	19.1956	20.0009
-0.2125	-0.1875	29	30.7155	32.8439	30.0720
-0.1875	-0.1625	54	54.0305	51.4877	43.7736
-0.1625	-0.1375	50	83.4402	74.6222	61.6716
-0.1375	-0.1125	87	115.6717	100.8121	83.9835
-0.1125	-0.0875	108	146.6319	127.8993	110.2860
-0.0875	-0.0625	128	172.6052	153.4106	139.2261
-0.0625	-0.0375	158	191.0922	175.0273	168.3851
-0.0375	-0.0125	202	201.0899	190.9772	194.4808
-0.0125	0.0125	209	202.9098	200.2607	214.0069
0.0125	0.0375	219	197.7585	202.6890	224.1724
0.0375	0.0625	235	187.2826	198.7707	223.7496
0.0625	0.0875	230	173.2011	189.5101	213.4033
0.0875	0.1125	179	157.0661	176.1805	195.3470
0.1125	0.1375	184	140.1428	160.1210	172.5551
0.1375	0.1625	132	123.3783	142.5828	147.9411

continued

Table 3.10 (continued)

Height interval (in meters)		Observed frequency	Expected frequencies for Gumbel	GEV	EOW
0.1625	0.1875	124	107.4229	124.6337	123.8115
0.1875	0.2125	103	92.6789	107.1129	101.6713
0.2125	0.2375	83	79.3562	90.6260	82.2910
0.2375	0.2625	73	67.5254	75.5652	65.8937
0.2625	0.2875	51	57.1627	62.1435	52.3572
0.2875	0.3125	36	48.1843	50.4340	41.3783
0.3125	0.3375	38	40.4734	40.4072	32.5850
0.3375	0.3625	22	33.8976	31.9645	25.6037
0.3625	0.3875	24	28.3220	24.9648	20.0941
0.3875	0.4125	17	23.6166	19.2456	15.7629
0.4125	0.4375	13	19.6607	14.6380	12.3661
0.4375	0.4625	13	16.3452	10.9773	9.7056
0.4625	0.4875	10	13.5736	8.1094	7.6228
0.4875	∞	19	65.3486	18.6454	28.4693
$\widehat{\mu}\pm SE_{\widehat{\mu}}$			-0.0069 ± 0.0026	0.0026 ± 0.0027	-0.0484 ± 0.0100
$\widehat{\sigma}\pm SE_{\widehat{\sigma}}$			0.1308 ± 0.0018	0.1323 ± 0.0018	-0.2726 ± 0.0281
$\widehat{\beta}\pm SE_{\widehat{\beta}}$				-0.1341 ± 0.0083	-2.4329 ± 0.2637
$l(\widehat{\boldsymbol{\theta}})$			1465.4623	1540.0158	1561.9937
χ^2_{df}			$\chi^2_{28} = 221.2781$	$\chi^2_{28} = 79.7487$	$\chi^2_{29} = 41.2333$
p-value			0.0000	0.0000	0.0657

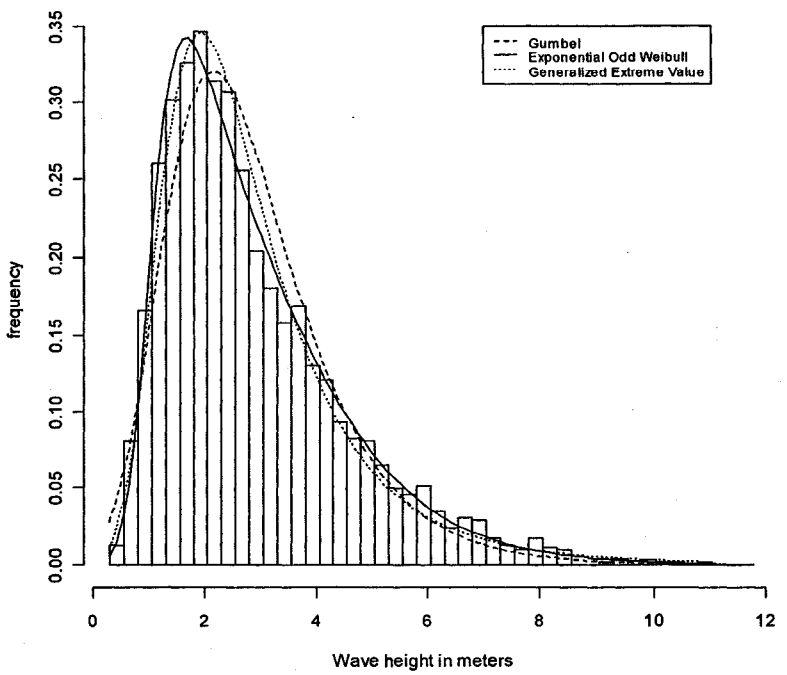

Figure 3.22 Fitted EOW, GEV, and Gumbel densities for Wave data.

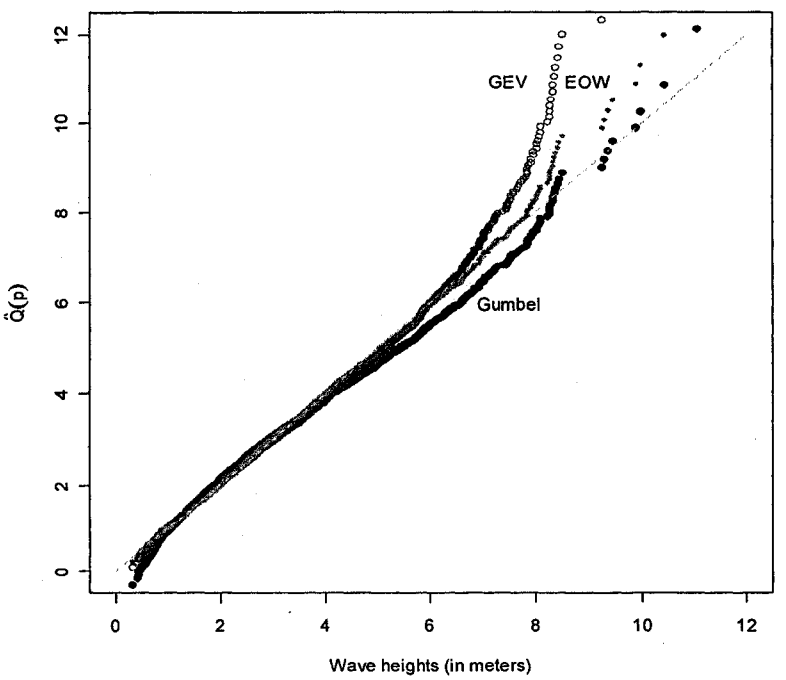

Figure 3.23 Fitted EOW, GEV, and Gumbel Q-Q plots for Wave data.

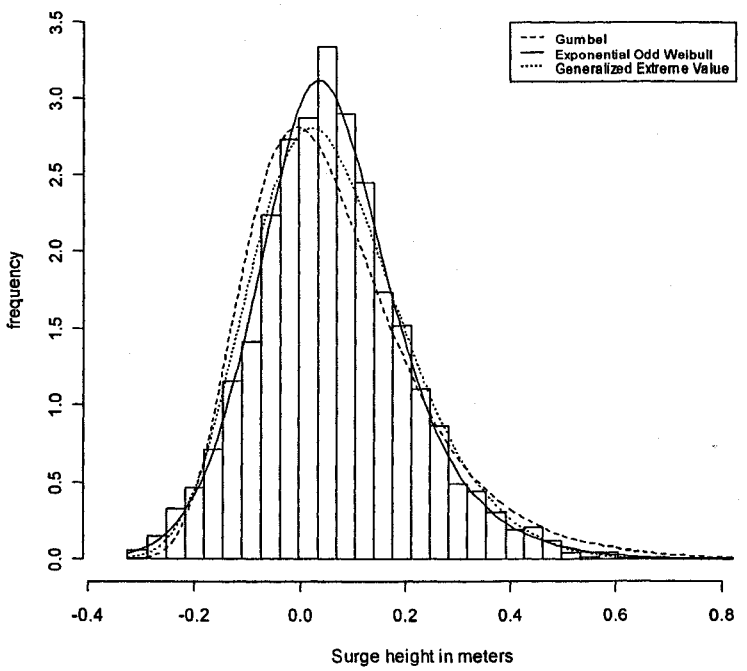

Figure 3.24 Fitted EOW, GEV, and Gumbel densities for Surge data.

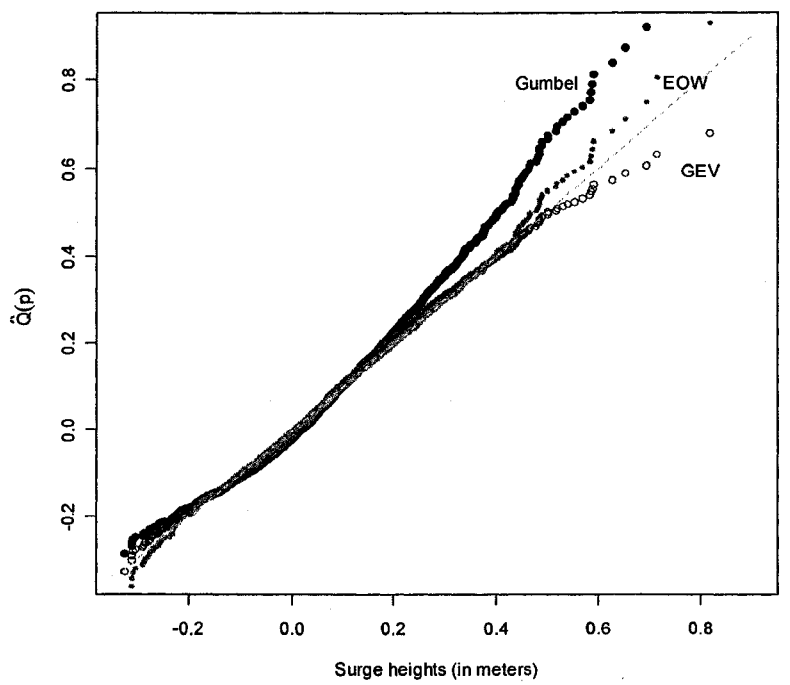

Figure 3.25 Fitted EOW, GEV, and Gumbel Q-Q plots for Surge data.

Table 3.11 Chi-squared and p-values under equal probability classes

	Gumbel	GEV	EOW
		Wave data	
χ^2_{df}	$\chi^2_{45} = 173.1569$	$\chi^2_{44} = 91.6199$	$\chi^2_{44} = 49.4250$
p-value	0.0000	0.0000	0.2655
		Surge data	
χ^2_{df}	$\chi^2_{45} = 201.9834$	$\chi^2_{44} = 81.4029$	$\chi^2_{44} = 40.5681$
p-value	0.0000	0.0005	0.6195

3.11 The log Odd Weibull family

The following Power-Pareto model (PPG) is presented by Gilchrist (2000) by multiplying the quantile functions of Power and Pareto distribution. The quantile function of the PPG distribution is given by

$$Q(u) = \theta u^{\gamma}/(1-u)^{\delta};\ 0 < \gamma,\ \delta,\ \theta < \infty,\ 0 \leq u \leq 1. \tag{3.40}$$

Note that the PPG distribution does not provide closed-form expressions for either density or distribution even though Pareto and Power distributions are submodels.

Alternatively here we presented the better Power-Pareto distribution by logarithmic transformation to the Odd Weibull distribution. It's cumulative distribution function, density function, hazard function and quantile function are, respectively, given by

$$F(x) = 1 - 1/[1 + \{(x/\theta)^{\alpha} - 1\}^{\beta}], \tag{3.41}$$

$$f(x) = (\alpha\beta/x)(x/\theta)^{\alpha}\{(x/\theta)^{\alpha} - 1\}^{\beta-1}/[1 + \{(x/\theta)^{\alpha} - 1\}^{\beta}]^2, \tag{3.42}$$

$$h(x) = (\alpha\beta/x)(x/\theta)^{\alpha}\{(x/\theta)^{\alpha} - 1\}^{\beta-1}/[1 + \{(x/\theta)^{\alpha} - 1\}^{\beta}], \tag{3.43}$$

and

$$Q(u) = \theta\left\{1 + \left(\frac{u}{1-u}\right)^{1/\beta}\right\}^{1/\alpha}. \tag{3.44}$$

Where, if $\alpha < 0$, and $\beta < 0$, then $0 \leq x \leq \theta$, otherwise $\theta \leq x < \infty$, and $0 \leq u \leq 1$. Note that when $\beta = 1$ and $\beta = -1$ this density represents the Power and Pareto distributions respectively. When $\theta = 1$, α, $\beta < 0$, the shape of the arising distribution is exactly as the beta distribution. Also note that when $\alpha = -1$ this Power-Pareto Model reduces to a submodel of the following distribution given by Balakrishnan (1992). It's distribution function is

$$F(x) = \left\{1 + e^{-\gamma}\left(\frac{1}{y} - 1\right)^{\delta}\right\}^{-1}, \tag{3.45}$$

where $0 < y < 1$, $0 < \gamma < \infty$, and $0 < \delta < \infty$.

Finally, the study of the log Odd Weibull family is an open question to interested readers.

CHAPTER IV

THE LOGISTIC-SINH DISTRIBUTION

4.1 Introduction

A two-parameter family of distributions is derived to model *highly negatively skewed data with extreme observations.* This distribution is referred to as the *logistic-sinh* (LS) distribution, since it is derived from the logistic distribution by appropriately replacing an exponential term with a hyperbolic sine term. The resulting family provides not only negatively skewed densities with thick tails but also variety of monotonic density shapes. The space of shape parameter, lambda greater than zero, is divided by the boundary line of lambda equals one into two regions over which the hazard function is, respectively, increasing and bathtub-shaped. The maximum likelihood parameter estimation techniques are discussed by providing approximate coverage probabilities for uncensored samples. The advantages of using this distribution are demonstrated and compared through well-known examples. In addition, the proposed family permits proportional hazard modeling. If a baseline hazard function of a distribution is $h(t)$ (or equivalently survival function $S(t)$), then for any $\xi > 0$, the new hazard function, $\xi h(t)$, (or equivalently the new survival function, $[S(t)]^{\xi}$), is also its member. This closure property of the proposed family is desirable, interesting, and convenient in studies involving multi-sample occurring in repair-reuse type reliability situations.

In the literature, two-parameter bathtub-shaped failure rate distributions are presented by several authors. Smith and Bain (1975) introduced the exponential power-life-testing distribution, and they used it to analyze the fifth bus motor failure data given by Davis (1952). Later, Dhillon (1981) used it to analyze some field failure data. Chen (2000) introduced a two-parameter bathtub-shaped failure rate lifetime distribution. However, this model is not convenient for parametric inferences due to the lack of scale parameter of the family. For the purpose of modeling life data, an extension of this two-parameter model is presented by Murthy *et al.* (2004).

Inspired by the work done by several authors, the LS family was derived to model highly negatively skewed data with extreme observations. After introducing the logistic distribution (Verhulst 1838, 1845) as a tool to study the population data, many other authors used it to solve some bioassay problems, used it as an income distributions and a growth model, etc., (see Balakrishnan 1992). For the first time, Plackett (1959) used it to analyze the survival data. Later, Balakrishnan and Cohen (1990) illustrated single-parameter half-logistic as a useful lifetime distribution. Burr (1942) proposed the replacement of random variable of type I logistic function with hyperbolic sine function; see also Johnson *et al.* (1994). Smith and Naylor (1987) analyzed the glass fiber strength data by using three-parameter Weibull distribution for the purpose of modeling unusually shaped likelihoods. Finally, using the exponentiated Weibull distribution, Mudholkar *et al.* (1995) advanced the analysis of the classical bus motor failure data while Lindsey (1997) gave an alternative analysis to the bus motor failure data using parametric multiplicative intensity models.

4.2 The model and its properties

Consider the half-logistic distribution with its cdf $F(x) = 1 - (1 + .5\,(e^x - 1))^{-1}$; $0 \leq x < \infty$. The logistic-sinh (LS) distribution is obtained by replacing the term $(e^x - 1)$ of the half-logistic distribution by $\sinh(e^x - 1)$ to obtain a negatively skewed density function. Alternative functional forms are discussed at the end of this section.

After suitably setting the parameters $\lambda(> 0)$ and $\theta(> 0)$, the cumulative distribution function, probability density function, hazard function, and the quantile function of the LS distribution can respectively be written as

$$F(x;\lambda,\theta) = 1 - (1 + \lambda \sinh(\exp(x/\theta) - 1))^{-1}, \tag{4.1}$$

$$f(x;\lambda,\theta) = (\lambda/\theta)\exp(x/\theta)\cosh(\exp(x/\theta) - 1)(1 + \lambda \sinh(\exp(x/\theta) - 1))^{-2}, \tag{4.2}$$

$$h(x;\lambda,\theta) = (\lambda/\theta)\exp(x/\theta)\cosh(\exp(x/\theta) - 1)(1 + \lambda \sinh(\exp(x/\theta) - 1))^{-1}, \tag{4.3}$$

and

$$Q(u) = F^{-1}(u) = \theta \ln\left(1 + \operatorname{arc\,sinh}\left(\frac{u}{\lambda(1-u)}\right)\right). \tag{4.4}$$

Where $0 < x < \infty$, $0 < \lambda < \infty$, $0 < \theta < \infty$, and $0 < u < 1$.

The kth moment of the LS distribution,

$$E(X^k) = \int_0^1 Q^k(u)du, \tag{4.5}$$

with $Q(u)$ given by (4.4), does not have a closed-form expression and must be evaluated numerically.

The quantiles of the distributions in the family, which are readily available from (4.4), can be used to construct quantile analogs of moment-based descriptive measures. In addition, the quantiles permit us to generate random data that describes the density given in (4.2). See Gilchrist (2000) for a comprehensive account of statistical modeling with quantile functions.

The LS distribution can be useful in the analysis of human survival time data, since it can be highly negatively skewed and non-zero density at the origin. For larger values of θ, and $\lambda = 2$, the LS distribution converges to an exponential distribution.

Figure 4.1 shows the LS density curves for different parameter values. The solid line, the dotted line, the dashed line, and the dot-dashed line indicate the parameter values $(\lambda = 0.025, \theta = 40)$, $(\lambda = 0.7, \theta = 40)$, $(\lambda = 1, \theta = 30)$, and $(\lambda = 3, \theta = 100)$, respectively.

Figure 4.2 illustrate the hazard curves of the LS distribution for different parameter values. The solid line $(\lambda = 0.5, \theta = 40)$, the dotted line $(\lambda = 1, \theta = 20)$, the dashed line $(\lambda = 7, \theta = 15)$, and the dot-dashed line $(\lambda = 10, \theta = 30)$.

Table 4.1 gives the mean $(E[X/\theta])$ and the standard deviation $(\sqrt{V(X/\theta)})$ for given parameter θ, and the coefficient of variation (CV) for different values of parameter λ of the LS distribution. This table would be useful to the reliability analysts in order to get starting points for the iterative method employed in Section 4.4.

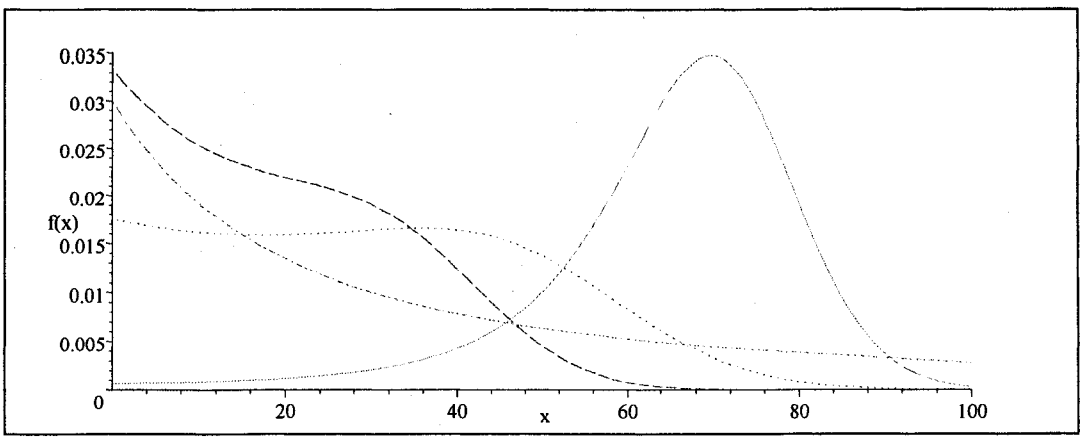

Figure 4.1 Typical density curves of the logistic-sinh distribution. The solid line ($\lambda = 0.025, \theta = 40$), the dotted line ($\lambda = 0.7, \theta = 40$), the dashed line ($\lambda = 1, \theta = 30$), and the dot-dashed line ($\lambda = 3, \theta = 100$).

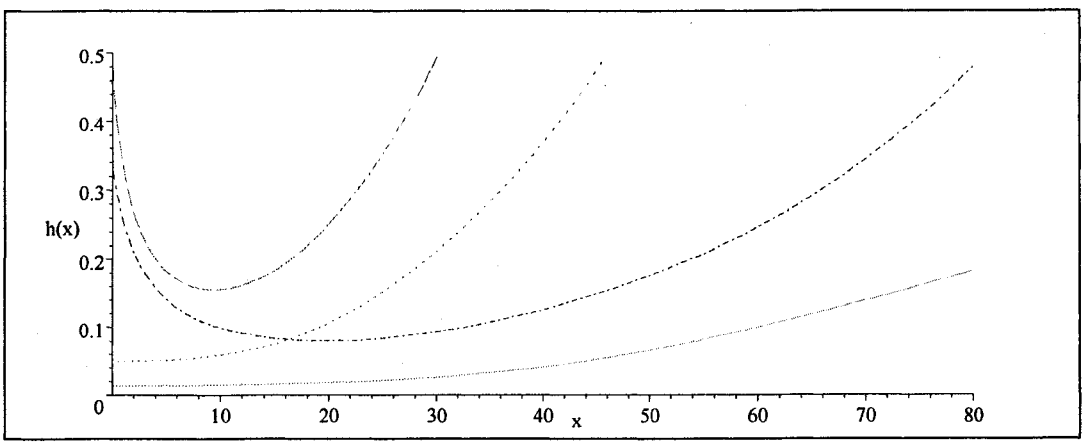

Figure 4.2 Typical hazard curves of the logistic-sinh distribution. The solid line ($\lambda = 0.5, \theta = 40$), the dotted line ($\lambda = 1, \theta = 20$), the dashed line ($\lambda = 7, \theta = 15$), and the dot-dashed line ($\lambda = 10, \theta = 30$).

For example, when $\theta = 15$ and $\lambda = 0.2$, the mean and the standard deviation can be calculated from this table such that, the mean $= 15 \times 1.1357 = 17.0355$, and the

standard deviation $= 15 \times 0.4763 = 7.1445$. The corresponding coefficient of variation, CV can directly be read from the table such that $CV = 0.4194$.

Figure 4.3 illustrates the shape variation (Y) of $E[X/\theta]$ (dotted line), $\sqrt{V(X/\theta)}$ (dashed line), and CV (solid line) against the parameter λ of the distribution.

Table 4.1 Mean, standard deviation, and CV of the LS distribution

λ	$E[X/\theta]$	$\sqrt{V(X/\theta)}$	CV	λ	$E[X/\theta]$	$\sqrt{V(X/\theta)}$	CV
0.01	1.7952	0.3250	0.1810	1	0.6902	0.4732	0.6857
0.02	1.6696	0.3614	0.2165	2	0.5134	0.4350	0.8472
0.05	1.4794	0.4124	0.2787	5	0.3210	0.3592	1.1190
0.10	1.3157	0.4485	0.3409	10	0.2129	0.2941	1.3813
0.20	1.1357	0.4763	0.4194	20	0.1355	0.2314	1.7077
0.50	0.8817	0.4889	0.5545	50	0.0708	0.1609	2.2726

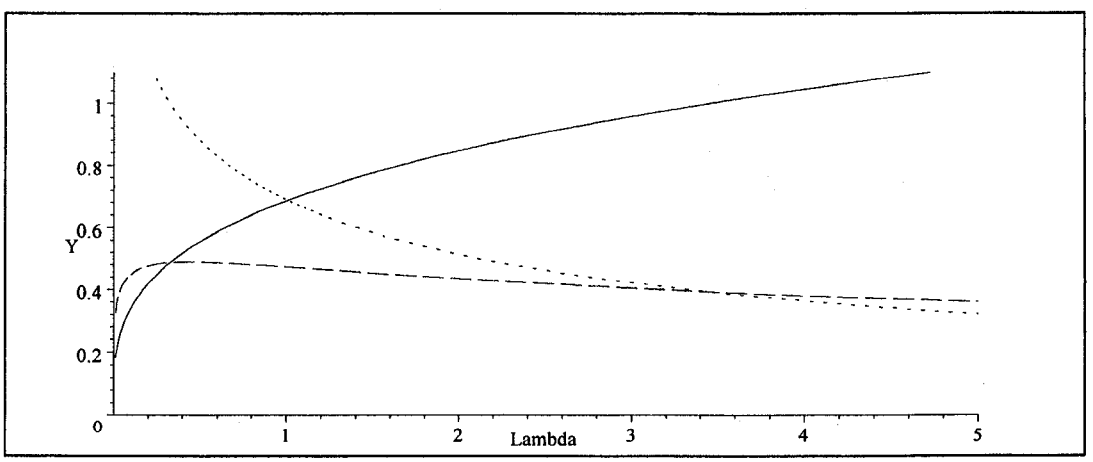

Figure 4.3 Mean, standard, and CV shape variations of LS. For given θ $E[X/\theta]$ (dotted line), $\sqrt{V(X/\theta)}$ (dashed line), and CV (solid line) with respect to parameter λ.

Furthermore, the CV does not depend on parameter θ. Consequently, the parameter λ is directly related to the CV of the distribution. The CV is important in reliability analysis. Barlow and Proschan (1981) showed that any life distribution with increasing (decreasing) hazard function has a $CV \leq 1$ ($CV > 1$). The theorem 4.1 (given below) shows that the logistic-sinh hazard function is increasing when $\lambda \leq 1$, i.e. $CV \leq 0.6857$. In addition, the numerical investigation shows that $CV = 1$ when $\lambda = 3.45$. This is due to mixture of decreasing and increasing shapes (bathtub-shaped) of the hazard function of the LS distribution, when $\lambda \geq 1$.

The proof of Theorem 4.1 can be verified using elementary calculus, and the shapes of the hazard function given by the theorem are illustrated in Figure 4.4. The solid line, and the dotted line indicate the parameter values ($\lambda = 20, \theta = 40$), and ($\lambda = 0.5, \theta = 45$) respectively.

Theorem 4.1

The shapes of the LS hazard function (4.3) are bounded by the parameter space $\lambda(> 0)$ such that monotone increasing ($\lambda \leq 1$) and bathtub-shaped ($\lambda > 1$).

Proof

Consider the following form of the hazard function given in (4.3).

$$h(x) = (\lambda/\theta) \exp(x/\theta) / (\operatorname{sec h}(\exp(x/\theta) - 1) + \lambda \tanh(\exp(x/\theta) - 1)),$$

where $0 \leq x < \infty$, $0 < \lambda < \infty$, and $0 < \theta < \infty$. One can show that $\lim_{x \to \infty} h(x) \to \infty$, and $h(0) = \lambda/\theta$.

The first derivative of $h(x)$ is,

$$h'(x) = \left(\frac{\lambda e^{x/\theta}}{\theta^2}\right) \left(\frac{(e^{x/\theta} + \lambda/2) \sinh(e^{x/\theta} - 1) - \lambda e^{x/\theta} + \cosh(e^{x/\theta} - 1)}{(1 + \lambda \sinh(e^{x/\theta} - 1))^2} \right).$$

Now $h'(0) = (\frac{\lambda}{\theta^2})(1-\lambda)$, hence $h'(0) \gtrless 0$, if $\lambda \lessgtr 1$. If $h(x)$ has a local maximum or minimum at c, then $h'(c) = 0$. It follows that, $\lambda e^{c/\theta} = .5\lambda \sinh\left(2(e^{c/\theta}-1)\right) + \cosh\left(e^{c/\theta}-1\right) + e^{c/\theta}\sinh\left(e^{c/\theta}-1\right)$.

The second derivative at c is,

$$h''(c) = (\frac{\lambda e^{2c/\theta}}{\theta^3})\left(\frac{\lambda\left(\cosh\left(2(e^{c/\theta}-1)\right)-1\right)+2\sinh\left(e^{c/\theta}-1\right)+e^{c/\theta}\cosh\left(e^{c/\theta}-1\right)}{\left(1+\lambda\sinh\left(e^{c/\theta}-1\right)\right)^2}\right).$$

Clearly, $h''(c) > 0$, if $c > 0$. Therefore, $x = c$ is a local minimum of $h(x)$. i.e., the hazard function is bathtub-shaped. This is possible only if $\lambda > 1$ due to the first derivative result at $x = 0$, $h'(0) < 0$. Furthermore, when $\lambda \leq 1$, the hazard function is increasing.

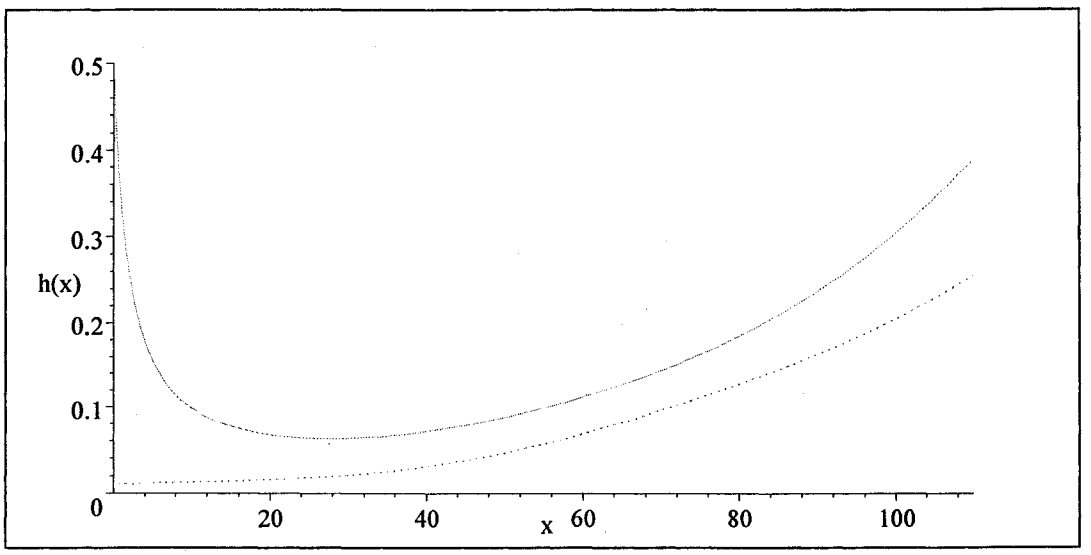

Figure 4.4 The hazard curves describe by the theorem 4.1 of LS. The solid line $(\lambda = 20, \theta = 40)$, the dotted line $(\lambda = 0.5, \theta = 45)$.

Alternative functional forms for the LS distribution

Consider the following function,

$$F(x;\lambda,\theta) = 1 - (1 + \lambda g(x;\theta))^{-1},\ 0 < \lambda < \infty,\ 0 < \theta < \infty,$$

where $g(.)$ is any increasing function in $\mathbb{R}^+$ such that $g(0) = 0$ and $g(\infty) = \infty$, is also a distribution function over $0 \leq x < \infty$. In order to obtain a negatively skewed density function over $0 \leq x < \infty$, one can select $g(.)$ such that $\lim_{x\to\infty}(g(x)/\sinh(x)) \to \infty$. For example, following functional forms would be possible for $g(.)$:

$g_1(x;\theta) = \sinh(\sinh(x/\theta))$, $g_2(x;\theta) = \exp(\exp(x/\theta) - 1) - 1$,

$g_3(x;\theta) = \exp(\sinh(x/\theta)) - 1$.

The hazard function of $g_1(.)$ selection gives only bathtub-shaped failure rates. The hazard function corresponding to $g_2(.)$ selection would be interesting due to sharp bounds of the hazard rate such that $\lambda > 2$ bathtub-shaped, and $\lambda \leq 2$ increasing. Although, this selection does not provide rich density shapes as the LS distribution. For example, the second bus motor failure data (Davis 1952) given in Example 2, Section 4.7, does not give a good fit with $g_2(.)$ selection. The $g_3(.)$ selection also gives sharp bounds of the hazard rate such that $\lambda > 1$ bathtub-shape, and $\lambda \leq 1$ increasing. This selection provides better density shapes than $g_2(.)$ selection. However, the $g_3(.)$ selection does not provide rich monotonic density shapes like the LS distribution. In addition, the density graph of the $g_3(.)$ selection shows a light positive antimode, which is not very apparent like in the LS density graph. This may lead to getting lower p-values in the analysis of data sets like the bus motor failure data (Davis 1952).

4.3 Parametric inference

Typically, parametric inference of distributions like the LS to the given data are based on likelihood methods and their asymptotic theory (Cox and Oakes 1984; Lawless 2003; Rao 1973). Estimates of the parameters are obtained by maximizing the log-likelihood function ($l(\boldsymbol{\theta}) = \ln L(x_1, x_2, ..., x_n; \lambda, \theta)$).

The log-likelihood function of the LS distribution is given by

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{n} \delta_j \left\{ \ln(\lambda/\theta) + \frac{x_j}{\theta} + \ln\left(\cosh\left(e^{x_j/\theta} - 1\right)\right) \right\} \\ &\quad - \sum_{j=1}^{n} (1+\delta_j) \ln\left(1 + \lambda \sinh\left(e^{x_j/\theta} - 1\right)\right), \end{aligned} \tag{4.6}$$

where δ_j is such that

$$\delta_j = \begin{cases} 0 & \text{if } j^{th} \text{ observation is right censored} \\ 1 & \text{if } j^{th} \text{ observation is not right censored} \end{cases} \qquad j = 1, 2, ..., n.$$

For the purpose of analyzing the grouped data (see example 2, Section 4.7) by estimating the parameters of the LS distribution under the ML method, one can assume that the data consists of r intervals and the jth interval, i.e. (c_{j-1}, c_j), has n_j observations for j = 1, 2, 3, ..., r ; $c_0 = 0$. The rth open interval, i.e. (c_{r-1}, ∞), contains n_r observations, and the total number of observations, n can be written as $\sum_{j=1}^{r} n_j$. Therefore the log-likelihood function of the LS distribution for grouped data can be written as

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{r} n_j \ln\left[F(c_j; \lambda, \theta) - F(c_{j-1}; \lambda, \theta)\right], \tag{4.7}$$

where $F(.; \lambda, \theta)$ is the cumulative distribution function of the LS distribution.

In this case, the log-likelihood function is maximized by solving the score equation $U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$. For large samples, asymptotic normality results hold for estimated parameters due to the convergence in distribution, i.e., $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow \mathbf{N}_2(\mathbf{0}, I^{-1}(\boldsymbol{\theta}))$, where $\mathbf{N}_2$ denotes the bivariate normal distribution and $I(\boldsymbol{\theta})$ is the expected Fisher information matrix of $\widehat{\boldsymbol{\theta}}$ such that $I(\boldsymbol{\theta}) = -\mathbf{E}\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$. Because the limiting distribution $\mathbf{N}_2$ is continuous, the convergence is actually uniform. The variance-covariance matrix, $I(\boldsymbol{\theta})$ is useful to construct approximate confidence intervals for individual parameters and functions of such parameters (Rao 1973). Moreover, whether the data are complete or right censored, inference procedures based on maximum likelihood large-sample theory can be applied in a straightforward way (Lawless 2003).

The fitting of the LS distribution by solving the score equation involving in the log-likelihood function can be facilitated using computer programs such as DNEQNF in IMSL (1991) and LE in BMDP (1992). The LE program in BMDP (1992), which uses Newton-Raphson type algorithm to maximize the likelihood function. It can be employed to estimate unknown parameters, whether the data is in complete, grouped, truncated, or right censored form. The LE routine also gives the asymptotic standard errors (SE's) of the estimates by inverting the Hessian matrix used in the maximization of the likelihood function, unless the information matrix is ill-conditioned. The information matrix may be ill-conditioned due to singularity or near singularity of the Hessian matrix. In this situation, the LE routine will set the asymptotic standard error of a parameter to zero by warning the following: Linear dependence among the parameters; The parameter is fixed; The parameter is on the boundary. For example, in the analysis of Sample 2 in Example 1, Section 4.7 using the three-parameter

Weibull distribution, one can see that the LE in BMDP (1992) produces invalid results.

4.4 Approximate coverage probabilities

The coverage probabilities for the maximum likelihood estimation method with intended confidence levels $\alpha = 0.1$ and $\alpha = 0.05$ are given in Table 4.2. These coverage probabilities are based on 10,000 simulated random samples from the density given in equation (4.2). The random samples are generated by plugging the known values of parameters λ and θ (say $\lambda = 0.01$, $\theta = 10$) to the quantile function given in equation (4.4). In addition, n (say $n = 10$) number of ordered uniform random sample from the uniform distribution, $u \sim U(0,1)$ is required to substitute as u in equation (4.4). In that way, one random sample of size n (say $n = 10$) from the LS distribution with parameters λ and θ (say $\lambda = 0.01$, $\theta = 10$) can be generated. In this simulation study, ten thousand such samples are generated to get a single cell value in Table 4.2. For this purpose, the subroutine DNEQNF in the IMSL (1991) package is used to solve the nonlinear equations involving in the likelihood procedure. The approximate $100\,(1-\alpha)\,\%$ confidence intervals for parameters, λ and θ are calculated by using $\left(\widehat{\lambda} - Z_{\alpha/2}SE_{\widehat{\lambda}},\ \widehat{\lambda} + Z_{\alpha/2}SE_{\widehat{\lambda}}\right)$ and $\left(\widehat{\theta} - Z_{\alpha/2}SE_{\widehat{\theta}},\ \widehat{\theta} + Z_{\alpha/2}SE_{\widehat{\theta}}\right)$ respectively. Where $SE_{\widehat{\lambda}}$ and $SE_{\widehat{\theta}}$ are respectively asymptotic standard errors of $\widehat{\lambda}$ and $\widehat{\theta}$, which are taken from the observed information matrix (Efron and Hinkley 1978).

Table 4.2 Approximate coverage probabilities of the LS

	$n =$	10			20			50		
90% intended	$\theta =$	10	20	50	10	20	50	10	20	50
$\lambda = 0.01$	λ :	.815	.802	.767	.812	.807	.791	.849	.849	.847
	θ :	.914	.916	.883	.894	.891	.873	.894	.888	.893
$\lambda = 0.10$	λ :	.758	.766	.758	.842	.854	.838	.872	.871	.869
	θ :	.841	.836	.824	.885	.885	.879	.887	.890	.890
$\lambda = 1.0$	λ :	.782	.782	.753	.835	.837	.842	.881	.874	.877
	θ :	.775	.766	.752	.832	.836	.839	.879	.872	.874
$\lambda = 10.0$	λ :	.532	.501	.414	.676	.670	.680	.793	.791	.800
	θ :	.498	.485	.410	.639	.628	.630	.770	.762	.767
95% intended										
$\lambda = 0.01$	λ :	.839	.829	.797	.842	.840	.821	.887	.882	.886
	θ :	.955	.956	.929	.939	.941	.922	.942	.936	.942
$\lambda = 0.10$	λ :	.792	.800	.790	.878	.888	.875	.911	.913	.903
	θ :	.888	.884	.876	.928	.934	.926	.936	.938	.936
$\lambda = 1.0$	λ :	.820	.818	.798	.873	.873	.879	.920	.917	.916
	θ :	.815	.811	.796	.876	.878	.877	.923	.918	.921
$\lambda = 10.0$	λ :	.573	.538	.451	.723	.717	.724	.836	.834	.841
	θ :	.532	.522	.443	.678	.668	.667	.810	.802	.806

From Table 4.2, one can see that when the sample size increases, the approximate coverage probabilities for the parameters under the maximum likelihood method is getting closer to the intended coverage probabilities. Although, the coverage probabilities decrease, when the shape parameter λ increases from one. But, one can obtain a desired confidence level by appropriately adjusting the confidence coefficient α. Moreover, the procedure gives an over coverage for parameter θ for small samples (e.g. when $n = 10$) and small values of parameters ($\theta = 10, 20$; $\lambda = 0.01$). Except for this case, the values in Table 4.2 predict that the parameters are not overly estimate under the maximum likelihood estimation method.

4.5 Actual coverage probabilities

The coverage probabilities for the maximum likelihood estimation method with intended confidence levels $\alpha = 0.1$ and $\alpha = 0.05$ are given in Table 4.3. These coverage probabilities are based on 10,000 simulated random samples.

In this simulation studies, the subroutine DNEQNF in the IMSL (1991) package is used to solve the nonlinear equations involving in the likelihood procedure. The approximate $100\,(1-\alpha)\,\%$ confidence intervals for parameters λ and θ are calculated by using $\left(\widehat{\lambda} - Z_{\alpha/2}SE_{\widehat{\lambda}},\ \widehat{\lambda} + Z_{\alpha/2}SE_{\widehat{\lambda}}\right)$ and $\left(\widehat{\theta} - Z_{\alpha/2}SE_{\widehat{\theta}},\ \widehat{\theta} + Z_{\alpha/2}SE_{\widehat{\theta}}\right)$ respectively. Where $SE_{\widehat{\lambda}}$ and $SE_{\widehat{\theta}}$ are respectively asymptotic standard errors of $\widehat{\lambda}$ and $\widehat{\theta}$, which were taken from the expected information matrix. In order to find the expected values involving in the information matrix, the Monte Carlo simulation method is used. For this purpose, 100,000 samples were used from the estimated LS density.

Table 4.3 Actual coverage probabilities of the LS

n	$\theta =$ / $\lambda =$	50 .100	50 .500	50 1.00	50 2.00	50 5.00	100 .100	100 .500	100 1.00	100 2.00	100 5.00
		90% intended									
10	λ:	.760	.770	.767	.745	.563	.749	.750	.765	.734	.552
	θ:	.833	.781	.757	.714	.572	.823	.775	.758	.716	.579
20	λ:	.811	.843	.835	.825	.774	.819	.832	.837	.823	.771
	θ:	.862	.851	.839	.813	.740	.848	.851	.837	.806	.741
30	λ:	.826	.861	.862	.850	.812	.826	.855	.862	.851	.803
	θ:	.867	.868	.860	.829	.787	.866	.868	.855	.836	.782
50	λ:	.855	.875	.875	.872	.847	.854	.879	.875	.873	.848
	θ:	.882	.879	.874	.863	.821	.876	.878	.876	.863	.820
		95% intended									
10	λ:	.794	.811	.805	.785	.613	.782	.792	.806	.779	.604
	θ:	.884	.828	.804	.759	.612	.871	.823	.800	.759	.618
20	λ:	.851	.878	.872	.863	.813	.849	.869	.872	.861	.812
	θ:	.908	.897	.879	.854	.780	.909	.893	.878	.847	.781
30	λ:	.865	.898	.898	.889	.852	.868	.891	.898	.887	.846
	θ:	.921	.917	.902	.875	.824	.919	.915	.898	.876	.819
50	λ:	.900	.917	.913	.911	.885	.896	.921	.918	.912	.886
	θ:	.932	.926	.921	.903	.862	.929	.931	.919	.906	.862

From Table 4.3, one can clearly see that as the sample size increases, the actual coverage probabilities for the parameters under the maximum likelihood method is getting closer to the intended coverage probabilities. On the other hand, the coverage probabilities decreases, when the shape parameter λ increases from one. However, one can obtain a desired confidence level by appropriately adjusting the confidence coefficient α. Moreover these values predict that the parameters are not overly estimate under the maximum likelihood estimation method.

4.6 Illustrative examples

The object of this section is to illustrate the use of proposed LS distribution and to demonstrate its applicability and better fit with the aid of real life data. In this regard, three distinctly different examples are presented based on well-known data, which were published in statistics literature. Specifically, first, second, and third examples respectively consider complete, grouped, and right censored data.

Example 1. Glass fiber strength data

The following glass fiber data (see data in the appendix B) are experimental strength values of two lengths, 1.5cm, and 15cm, from the National Physical Laboratory in England (Smith and Naylor 1987). Preliminary inspection of the data reveals possible outliers in the lower end point of the sample, the smallest observation in Sample 1 (strength value = 0.55) and the smallest two in Sample 2 (strength values = 0.37, 0.40). The authors used three-parameter Weibull distribution to model the two data sets and concluded that the Bayesian techniques appear to be better choice

for handling unusually shaped likelihoods than the maximum likelihood techniques. Furthermore, they attempted to fit alternative models, but they do not match with the end points of the data distribution, even though such models reduce the discrepancy between maximum likelihood and Bayesian analysis. In general, the tested glass material's experimental strength values have a wide statistical spread due to, in part, grip-induced breakage. Therefore, longer tail distributions would be appropriate to model such data. Hence, we can use these data sets to show the flexibility and applicability of the LS distribution in the presence of some extreme observations.

Furthermore, in order to select the best fitting model out of the three-parameter Weibull distribution and the LS distribution, the Akaikie information criterion (AIC) can be used. This criterion is based on the log-likelihood value $l(\widehat{\boldsymbol{\theta}})$, and the number of parameters in the distribution (p). The AIC attempts to balance the need for a model which fits the data vary well to that of having a simple model with few parameters. It is defined as $r = l(\widehat{\boldsymbol{\theta}}) - 2p$. The distribution with the largest r value is the distribution that fits the data the best.

The estimated parameter values, mean failure time ($\widehat{e}$) with their asymptotic standard errors, and the r values for the LS distribution are given below.

Sample 1: Strengths 1.5cm Fiber ($\widehat{\lambda} \pm SE_{\widehat{\lambda}} = 0.01820 \pm 0.008429$, $\widehat{\theta} \pm SE_{\widehat{\theta}} = 0.89622 \pm 0.035912$, $\widehat{e} \pm SE_{\widehat{e}} = 1.512434 \pm 0.355602$; $r = -16.371$).

Sample 2: Strengths 15cm Fiber ($\widehat{\lambda} \pm SE_{\widehat{\lambda}} = 0.03495 \pm 0.016598$, $\widehat{\theta} \pm SE_{\widehat{\theta}} = 0.72346 \pm 0.034692$, $\widehat{e} \pm SE_{\widehat{e}} = 1.126601 \pm 0.289001$; $r = -6.647$).

Estimated parameter values and the r values for the three-parameter Weibull

distribution ($F(x;\beta,\phi,\mu) = 1-\exp\left(-\left(\frac{x-\mu}{\phi}\right)^{\beta}\right)$, $\mu \leq x < \infty$, $-\infty < \mu < \infty$, $0 < \beta$, $\phi < \infty$) are as follows. Note that for Sample 2, the LE program in BMDP (1992) does not produce correct values for asymptotic standard errors due to the Hessian matrix is nearly singular. Therefore, the calculations are done using the IMSL (1991).

Sample 1: Strengths 1.5cm Fiber ($\widehat{\beta} \pm SE_{\widehat{\beta}} = 11.855837 \pm 9.832166$, $\widehat{\phi} \pm SE_{\widehat{\phi}} = 3.235026 \pm 2.624502$, $\widehat{\mu} \pm SE_{\widehat{\mu}} = -1.593400 \pm 2.614103$; $r = -20.285$).

Sample 2: Strengths 15cm Fiber ($\widehat{\beta} \pm SE_{\widehat{\beta}} = 21.311912 \pm 49.031591$, $\widehat{\phi} \pm SE_{\widehat{\phi}} = 4.720435 \pm 10.677731$, $\widehat{\mu} \pm SE_{\widehat{\mu}} = -3.471851 \pm 10.664725$; $r = -8.082$).

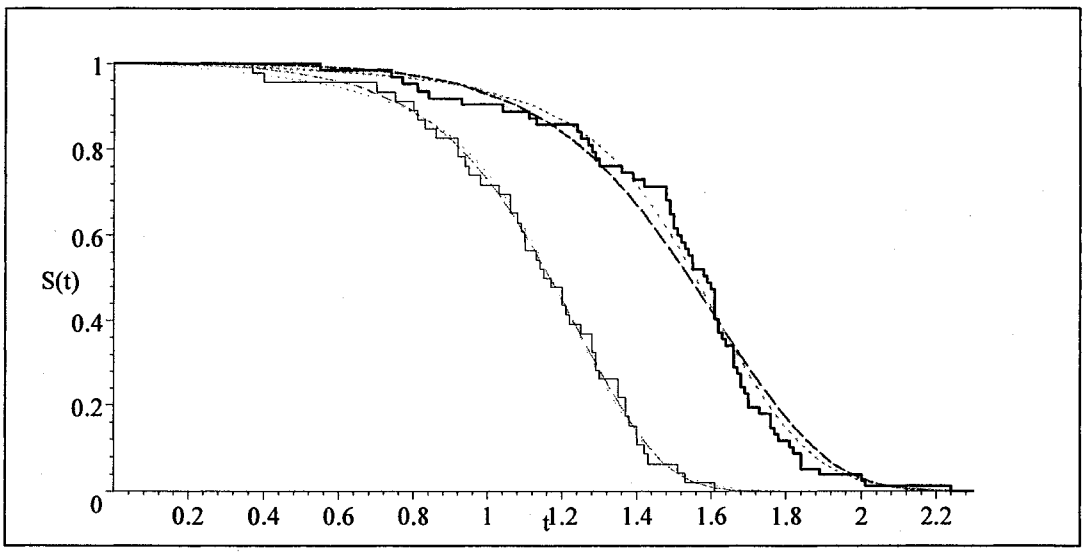

Figure 4.5 Fitted LS and three-parameter Weibull curves for fiber data. Dark lines for sample 1, light lines for sample 2. Kaplan-Meier survival curve (solid lines), logistic-sinh (dotted lines), three-parameter Weibull (dashed lines).

The r values of AIC given under the Sample 1 and 2 indicate better fit to the LS distribution than the three-parameter Weibull distribution. Furthermore, the

adequacy of the fits for the two samples (dark lines for Sample 1, light lines for Sample 2 in Figure 4.5) is further strengthened by illustrating the LS survival curve (dotted lines) along with the Kaplan-Meier curve (solid lines). For comparison purposes, the three-parameter Weibull survival curves for the two samples are also included in Figure 4.5 (dashed lines).

It should be mentioned here that negatively skewed distributions such as Gompertz, Weibull, exponential power-life-testing distribution, etc., provide a good fit to Sample 2 as well. However, the well-known distributions such as lognormal, gamma, Pareto, Weibull, exponential power-life-testing distribution, etc., perform very poorly with the data set given in Sample 1, and even the Gompertz distribution gives a very poor fit with r value of AIC, $r = -18.808$. This is because the final data point in Sample 1 (strength value = 2.24) is an outlier to the Gompertz distribution.

Example 2. Bus motor failure data.

The classical five bus motor failure data (see data in the appendix B) are firstly considered and analyzed by Davis (1952). The results take into account the time to the first and succeeding major motor failures for 191 buses operated by a large city bus company, with time being the number of thousand miles driven. Failure was either abrupt, in which some part broke or the motor would not run; or, by definition, when the maximum power was produced, as measured by a dynamometer, motor rate fell below a fixed percentage of the normal rated value. Failures of motor accessories, which could be easily replaced, were not included in these data.

Davis used the truncated normal distribution to analyze the first two motor failure data and the exponential distribution for the second and succeeding failures. In

the analysis, in terms of the chi-squared goodness-of-fit, he found that both models are poorly fit to the second bus motor failure data. Bain (1974, 1978) adapted a three-parameter quadratic hazard model for the purpose of obtaining a good fit to the second bus motor failure data. Later, Mudholkar *et al.* (1995) used three-parameter exponentiated Weibull model to analyze the five motor failure data. Lindsey (1997) gave an alternative analysis to the bus motor failure data using parametric multiplicative intensity models. However, he considers data that are grouped more coarsely than the data given by Davis (1952).

In this example, the two-parameter LS distribution is used to reanalyze the five classical bus motor failure data. The reanalysis is based on the data given in the appendix B, which can be compared (see Table 4.4) with the three-parameter exponentiated Weibull model ($F(y;\gamma,\delta,\varphi) = (1-\exp(-(y/\varphi)^{\gamma}))^{\delta}$; $0 < y < \infty$, $\gamma > 0$, $\delta > 0$, $\varphi > 0$, Mudholkar *et al.* 1995). Table 4.4 gives the estimated parameter values, log-likelihood values ($l(\widehat{\boldsymbol{\theta}})$), chi-squared values ($\chi^2_{df}$), and p-values for the bus motor failure data. The estimated mean failure, $\widehat{e}$ and its asymptotic standard error, $SE_{\widehat{e}}$, are also included. In this estimation process, equation (4.7) is used as a log-likelihood function, $l(\boldsymbol{\theta})$. The asymptotic standard errors are calculated by inverting the Hessian matrix. The Hessian matrix is obtained by partially differentiating equation (4.7). The $\widehat{e}$ value is calculated by substituting the estimated parameter values, $\widehat{\lambda}$ and $\widehat{\theta}$ into equation (4.5), with $k = 1$. Its standard error is calculated from the inverse of the Hessian matrix (observed information matrix) and a direct application of the delta method. For this purpose, partial differentiation with respect to the parameter λ and θ of equation (4.5) with $k = 1$, is used.

Table 4.4 Estimated values of LS for the five bus motor failure data

	1st	2nd	3rd	4th	5th
	The two-parameter logistic-sinh distribution				
$\widehat{\lambda}$	0.155413	0.644780	1.328801	1.487961	1.484886
$\pm\, SE_{\widehat{\lambda}}$	$\pm$ 0.03023	$\pm$ 0.16110	$\pm$ 0.36493	$\pm$ 0.41183	$\pm$ 0.41032
$\widehat{\theta}$	79.94892	84.66868	88.42888	68.32537	59.90192
$\pm\, SE_{\widehat{\theta}}$	$\pm$ 2.4516	$\pm$ 5.7378	$\pm$ 10.2375	$\pm$ 7.7739	$\pm$ 6.69700
$\widehat{e}$	96.16751	68.62946	54.40017	40.05453	35.14708
$\pm\, SE_{\widehat{e}}$	$\pm$ 11.7328	$\pm$ 11.9867	$\pm$ 9.84241	$\pm$ 7.56824	$\pm$ 6.78529
$l(\widehat{\boldsymbol{\theta}})$	-380.541	-202.447	-177.167	-147.709	-125.207
χ^2_{df}	$\chi^2_5 = 1.1738$	$\chi^2_4 = 3.6084$	$\chi^2_3 = 0.9779$	$\chi^2_2 = 0.8239$	$\chi^2_2 = 2.9557$
p-value	0.9474	0.4616	0.8066	0.6623	0.2281
	The three-parameter exponentiated Weibull distribution				
$\widehat{\gamma}$	7.234813	18.8858	3.9365	16.9445	0.9584
$\pm\, SE_{\widehat{\gamma}}$	$\pm$ 1.82478	$\pm$ 7.1188	$\pm$ 4.3156	$\pm$ 2.4063	$\pm$ 0.7917
$\widehat{\delta}$	0.277454	0.0506	0.1909	0.0385	1.8311
$\pm\, SE_{\widehat{\delta}}$	$\pm$ 0.0878	$\pm$ 0.0203	$\pm$ 0.2344	$\pm$ 0.00	$\pm$ 3.3503
$\widehat{\varphi}$	138.99086	134.7123	118.0331	98.0984	24.8345
$\pm\, SE_{\widehat{\varphi}}$	$\pm$ 6.00212	$\pm$ 5.4195	$\pm$ 23.1313	$\pm$ 6.5577	$\pm$ 37.8164
$l(\widehat{\boldsymbol{\theta}})$	-381.811	-201.707	-176.987	-147.372	-123.826
χ^2_{df}	$\chi^2_3 = 2.6700$	$\chi^2_3 = 1.9485$	$\chi^2_2 = 0.6438$	$\chi^2_1 = 0.1343$	$\chi^2_1 = 0.0868$
p-value	0.4454	0.5832	0.7248	0.7140	0.7683

These values indicate excellent to moderate fit to the two-parameter LS distribution. In view of Theorem 4.1, the first two motor failures indicate increasing hazard shapes ($\widehat{\lambda} \leq 1$), whereas the other three motor failures indicate bathtub hazard shapes ($\widehat{\lambda} > 1$). However, the slope changes for the 3rd motor failure given in Figure 4.6 are subtle and hard to see graphically. A similar problem was pointed out by Mudholkar *et al.* (1995) in the analysis of second motor failure data using the exponentiated Weibull distribution.

Figure 4.6 shows the fitted LS hazard curves for the five motor failure data. The dark solid line, the dark dashed line, the dotted line, the dashed line, and the solid line represent first, second, third, fourth, and fifth motor failures respectively.

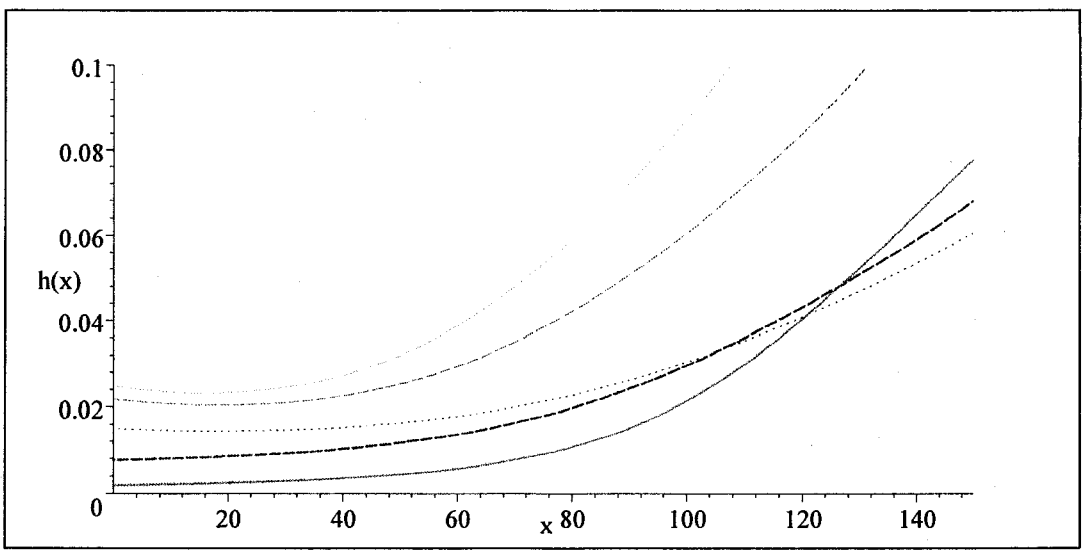

Figure 4.6 Fitted LS hazard curves for five motor failure data. First (dark solid line), second (dark dashed line), third (dotted line), fourth (dashed line), and fifth (solid line).

Example 3. Oklahoma diabetic data

The following survival times (in years) represent the *first 40 male patients* enrolled in a mortality study (see data in the appendix B) of Oklahoma diabetic Indians (Lee and Wang 2003). This example is a part of larger sample of 1012 Oklahoma Indians with non-insulin-dependent diabetes mellitus (NIDDM)), and the data were examined in 1972-1980.

This example is analyzed to convince the applicability and better-fitness of the LS distribution for right censored data. The estimated parameter values, mean failure time ($\widehat{e}$) with their asymptotic standard errors, and the log-likelihood value for the LS distribution are ($\widehat{\lambda} \pm SE_{\widehat{\lambda}} = 0.169642 \pm 0.100923$, $\widehat{\theta} \pm SE_{\widehat{\theta}} = 11.42646 \pm 1.641168$, $\widehat{e} \pm SE_{\widehat{e}} = 13.4802 \pm 4.42253$; $l(\widehat{\boldsymbol{\theta}}) = -74.362$). The equation (4.6) is used as a log-likelihood function, $l(\boldsymbol{\theta})$ to estimate the parameter values. The asymptotic standard errors are obtained by inverting the Hessian matrix of equation (4.6). As in the previous example, the $\widehat{e}$ value is calculated by substituting the estimated parameter values, $\widehat{\lambda}$ and $\widehat{\theta}$ into the equation (4.5), with $k = 1$. In order to find its standard errors by using the delta method, the partial derivatives of equation (4.5) with respect to the parameter λ and θ have to be evaluated.

The p-values for this right censored data are based on discretized method introduced by Efron (1988). This data was discretized as in Table 4.5, which includes the signed deviance residuals, R_j given by formula (2.69) (McCullagh and Nelder 1998); to the fitted hazard function of the LS distribution. Where, N_j is the total number of patients at risk at the beginning of each interval j, $j = 1, ..., 14$. S_j and E_j are respectively observed and expected deaths at the end of each interval j. The

p-value in Table 4.5, indicates that the LS distribution provides a good fit for the diabetic data, and the fitness is illustrated in Figure 4.7 (the dotted line) along with the Kaplan-Meier curve (solid line). Once again, well-known distributions, such as lognormal, gamma, Weibull, Pareto, etc., are not appropriate to analyze this data set due to their poor fits.

Table 4.5 Residual analysis of LS hazard for the diabetic data

j	0-1	1-2	2-3	3-4	4-5	5-6	6-7
N_j	40	39	39	38	36	35	33
S_j	1	0	1	2	1	2	2
E_j	0.62	0.65	0.71	0.77	0.82	0.91	0.99
R_j	0.45	-1.15	0.32	1.18	0.20	1.01	0.91
j	7-8	8-9	9-10	10-11	11-12	12-13	13-14
N_j	31	31	31	30	29	28	29
S_j	0	0	1	0	1	3	5
E_j	1.10	1.32	1.62	1.94	2.36	2.88	4.23
R_j	-1.50	-1.64	-0.54	-2.00	-1.03	0.08	0.40
$\sum R_j^2 = 15.37$			d.f. $= 12$		p-value $= 0.2216$		

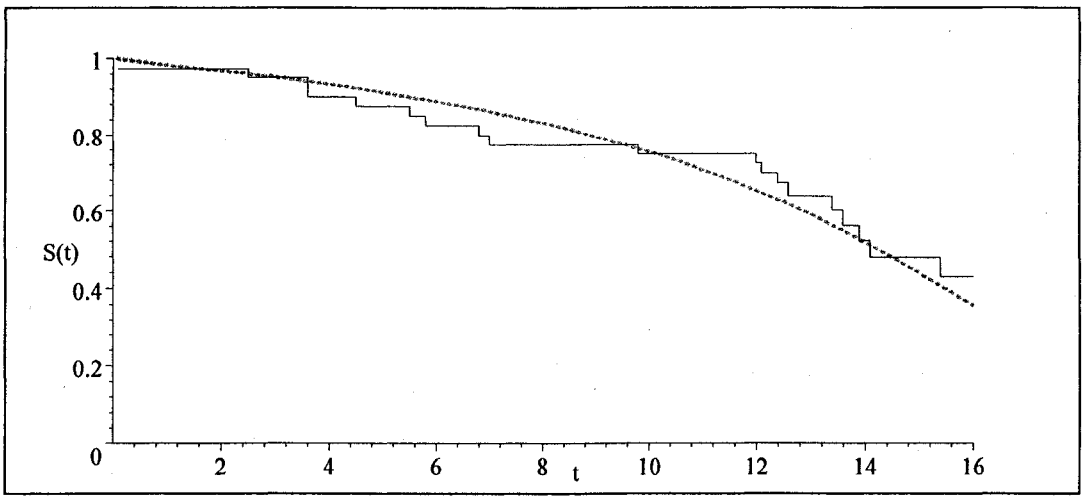

Figure 4.7 Fitted LS survival curves for diabetic data. Kaplan-Meier survival curve (solid line), logistic-sinh (dotted line).

4.7 Reanalyzing the bus motor failure data using the ELS

In this section, a third parameter $\beta(>0)$ to the LS distribution (4.2) is introduced to reanalyze the bus motor failure data given in the appendix B. The cumulative distribution function, the probability density function, and the hazard function of this extended logistic-sinh model (ELS) can respectively be written as

$$F(x;\lambda,\beta,\theta) = 1 - \left(1 + \lambda \sinh^{\beta}\left(\exp\left(x/\theta\right) - 1\right)\right)^{-1}, \tag{4.8}$$

$$f(x;\lambda,\beta,\theta) = \left(\frac{\lambda\beta}{\theta}\right)\frac{\exp\left(x/\theta\right)\cosh\left(\exp\left(x/\theta\right) - 1\right)\sinh^{\beta-1}\left(\exp\left(x/\theta\right) - 1\right)}{\left(1 + \lambda \sinh^{\beta}\left(\exp\left(x/\theta\right) - 1\right)\right)^{2}}, \tag{4.9}$$

and

$$h(x;\lambda,\beta,\theta) = \left(\frac{\lambda\beta}{\theta}\right)\frac{\exp\left(x/\theta\right)\cosh\left(\exp\left(x/\theta\right) - 1\right)\sinh^{\beta-1}\left(\exp\left(x/\theta\right) - 1\right)}{1 + \lambda \sinh^{\beta}\left(\exp\left(x/\theta\right) - 1\right)}, \tag{4.10}$$

where $0 < x < \infty, 0 < \lambda < \infty, 0 < \beta < \infty$, and $0 < \theta < \infty$.

The likelihood procedure given in Section 4.4 can easily be extended to estimate the parameters of the ELS model given in (4.9).

Table 4.6 gives the estimated parameter values, log-likelihood values ($l(\widehat{\boldsymbol{\theta}})$), and chi-squared values (χ^2_{df}) with respective p-values of the ELS model for the bus motor failure data given in the appendix B. The fitted ELS density functions, and hazard functions are given in Figure 4.8, and 4.9, respectively.

Table 4.6 Estimated values of ELS for the five bus motor failure data

	1st	2nd	3rd	4th	5th
$\widehat{\lambda}$	0.16951	0.47710	1.13889	1.11165	5.15389
$SE_{\widehat{\lambda}}$	0.04039	0.12686	0.50098	0.45011	7.03415
$\widehat{\beta}$	1.18860	0.69817	0.91209	0.81119	1.55681
$SE_{\widehat{\beta}}$	0.23276	0.17728	0.21683	0.23771	0.39121
$\widehat{\theta}$	86.8134	68.9998	81.6755	58.3025	97.7103
$SE_{\widehat{\theta}}$	8.88948	9.72197	17.8659	12.7621	49.3878
$l(\widehat{\boldsymbol{\theta}})$	-380.181	-201.242	-177.089	-147.428	-123.803
χ^2_{df}	$\chi^2_4 = 0.4038$	$\chi^2_3 = 1.0225$	$\chi^2_2 = 0.8594$	$\chi^2_1 = 0.2409$	$\chi^2_1 = 0.0406$
p-value	0.9822	0.7958	0.6507	0.6236	0.8403

Unlike the other parametric families, which were given in literature, the associated p-values of the chi-squared goodness-of-fit test to the ELS model for the five motor

failure data is over 60%. Furthermore, using a simple calculator, the log-likelihood functions of the ELS model can be maximized for the five motor failure data.

Figure 4.8 shows the fitted ELS density curves for the five motor failure data given in the appendix B. The dark solid line, dark dashed line, dotted line, dashed line and solid line represent first, second, third, fourth, and fifth motor failures respectively.

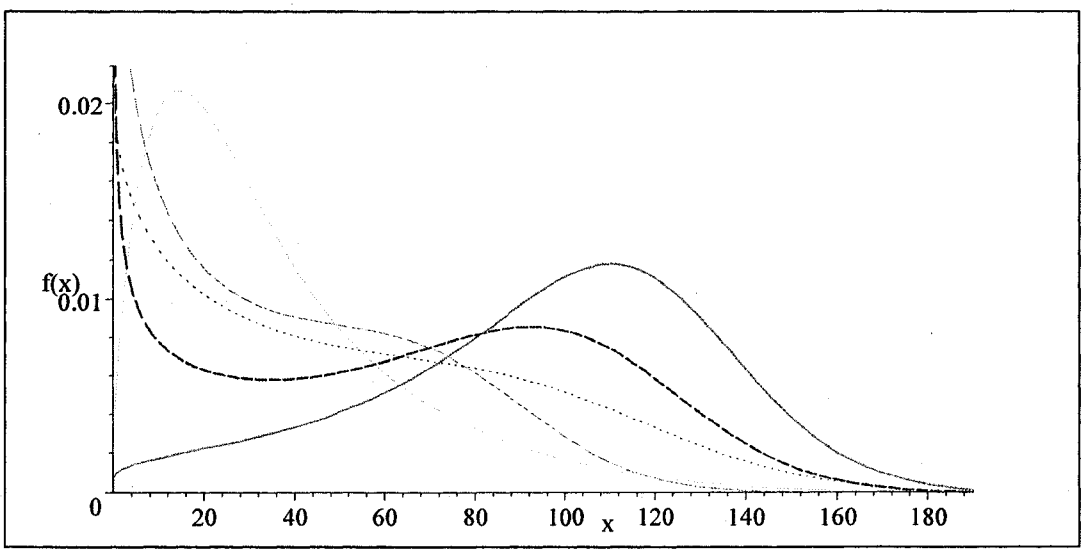

Figure 4.8 Fitted ELS density curves for five motor failure data. The dark solid line, the dark dashed line, the dotted line, the dashed line, and the solid line, respectively, represent the first, the second, the third, the fourth, and the fifth motor failure data.

Figure 4.9 shows the fitted ELS hazard curves for the five motor failure data given in the appendix B. The dark solid line, dark dashed line, dotted line, dashed line and solid line represent the first, second, third, fourth, and fifth motor failures respectively. This figure indicate three different hazard shapes for the five different bus motor failure data. Specifically, the first motor failure indicates an increasing

failure rate; and the second, third, and fourth motor failures indicate bathtub shape failure rates; the failure rate of the fifth motor indicates initially increasing and again increasing.

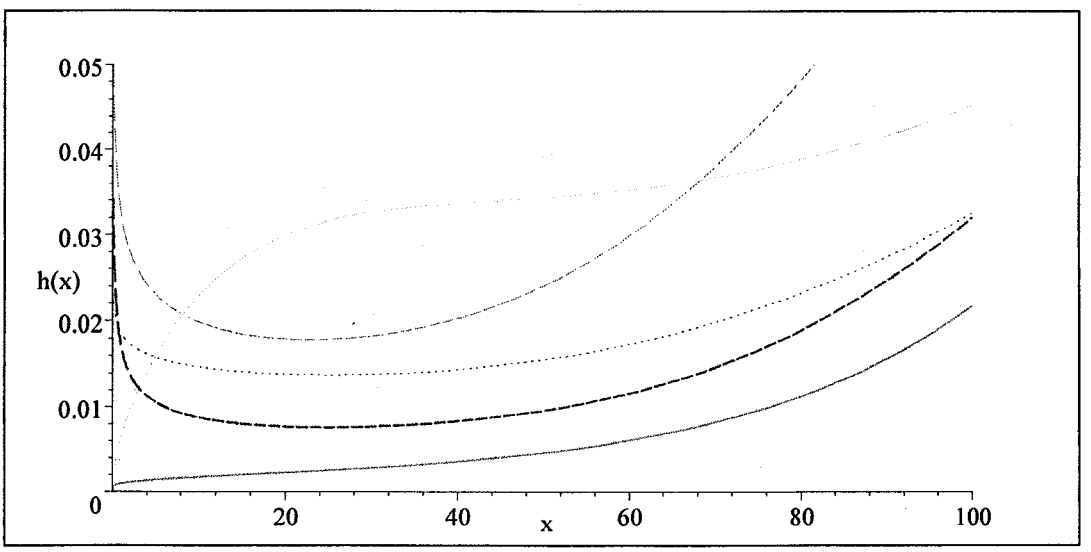

Figure 4.9 Fitted ELS hazard curves for five motor failure data. The dark solid line, the dark dashed line, the dotted line, the dashed line, and the solid line, respectively, represent the first, the second, the third, the fourth, and the fifth motor failure data.

4.8 The Gompertz-sinh distribution

In this section, we explore two- and three-parameter families of distributions in order to model highly negative-skewed data which arise frequently in survival analysis. Since the two-parameter distribution is derived from the Gompertz distribution by appropriately replacing the index of the exponential term with a hyperbolic sine term, it is henceforth referred to as *Gompertz-sinh* (GS) distribution. The resulting distribution possesses a lighter right tail than that of the Gompertz distribution,

which is often used to model highly negative-skewed data. Moreover, we generalize the Gompertz-sinh distribution by simply adding a second shape parameter as an exponent to its distribution function to accommodate a variety of density shapes as well as non-decreasing hazard shapes. This generalization is referred to as *exponentiated Gompertz-sinh* (EGS) distribution. The maximum likelihood parameter estimation techniques are discussed by providing approximate coverage probabilities for uncensored samples. Furthermore, the applicability and flexibility are demonstrated and illustrated by citing real data examples.

4.8.1 Motivation to analyze the aging process

The Gompertz distribution (Gompertz 1825), developed from the mortality law, is often used to model highly negative-skewed data in survival analysis. However, the Gompertz distribution does not provide a reasonable fit for highly negative-skewed data found in some practical applications in which the underlying distribution possesses a thinner and shorter right tails. Such situations have so far been remedied with nonparametric and graphical procedures although they poorly analyze the data, see Miller (1983) and Efron (1988). Moreover, some researchers use higher order parametric models or combine the existing distributions even though they need large amount of data to estimate the parameters. One such two-parameter composite model was introduced by Cooray and Ananda (2005) to model highly positively skewed data which usually arise in insurance industry and actuarial sciences.

Many researchers in actuarial sciences, demographic studies and statistics have so far used different modifications closely related to the Gompertz or exponential dis-

tribution, for example, the two-parameter exponential power life-testing distribution (Smith and Bain 1975), a quadratic hazard function (Bain 1974), the two-parameter lifetime distribution (Chen 2000), and the well known three-parameter Gompertz-Makeham distribution (Makeham 1860) to model some failure data or to construct life tables.

Burr (1942) first suggested the hyperbolic sine transformations to the logistic distribution function, see also Johnson *et al.* (1994). Interestingly, Barndorff-Nielsen (1978) introduced a family of generalized hyperbolic distributions by analyzing directly the density functions of the exponential family. Later Rieck and Nedelman (1991) used hyperbolic sine transformation in the standard normal distribution by establishing the relationship between the sinh-normal and the Birnbaum-Saunders distributions (1969). In addition, Cooray (2005) proposed the two-parameter logistic-sinh family which possesses bathtub-shaped and increasing failure rates to model lifetime data.

4.8.2 The model and its properties

The Gompertz distribution is widely used as a parametric model to identify the natural death behavior of a population of humans or animals. For example, deaths caused by chronic disease conditions (e.g. diabetes) occur more frequently than the natural deaths. To model such death data, the Gompertz distribution,

$$F(x;\gamma,\phi) = 1 - \exp\left(-\gamma(e^{x/\phi} - 1)\right);\ 0 \leq x,\ 0 < \gamma,\ 0 < \phi \qquad (4.11)$$

is modified by replacing the term $(e^{y/\phi} - 1)$ with $\sinh(e^{x/\theta} - 1)$ giving

$$F(x;\mu,\theta) = 1 - \exp\left(-\mu \sinh(e^{x/\theta} - 1)\right);\ 0 \leq x < \infty,\ 0 < \mu < \infty,\ 0 < \theta < \infty, \tag{4.12}$$

which is now called as *Gompertz-sinh* (GS) distribution. This transformation assigned more probabilities to the left tail of the distribution than that of the Gompertz distribution. Its probability density function is given by,

$$f(x;\mu,\theta) = (\mu/\theta)\, e^{x/\theta} \cosh(e^{x/\theta} - 1) \exp\left(-\mu \sinh(e^{x/\theta} - 1)\right), \tag{4.13}$$

where $0 \leq x < \infty$, $0 < \mu < \infty$, and $0 < \theta < \infty$. Furthermore, the quantile function and the hazard function are, respectively, given by

$$Q(u) = F^{-1}(u) = \theta \ln\left(1 - \operatorname{arc\,sinh}\left(\mu^{-1} \ln(1-u)\right)\right), \tag{4.14}$$

and

$$h(x;\mu,\theta) = (\mu/\theta)\, e^{x/\theta} \cosh(e^{x/\theta} - 1);\ 0 \leq x < \infty, 0 < \mu < \infty, 0 < \theta < \infty. \tag{4.15}$$

Where $0 < u < 1$, $0 \leq x < \infty$, $0 < \mu < \infty$, and $0 < \theta < \infty$.

The GS family can be useful in the analysis of human survival time data, since this highly negative-skewed distribution has a non-zero density at the origin. For large values of θ, the GS family converges to an exponential distribution with a mean θ/μ, i.e., the GS density curve moves to the left hand side by keeping its unimodality. On the other hand, $\lim_{x\to\infty} F_{GS}(x;\mu,\theta)/F_G(x;\gamma,\phi) \to 0$ will show that the GS density has a thinner right tail than that of the Gompertz density. Here, $F_{GS}(x;\mu,\theta)$ and $F_G(x;\gamma,\phi)$ are distribution functions of GS and Gompertz distributions, respectively.

Figure 4.10a and Figure 4.10b, respectively, represent the GS density and its hazard curves for parameter values $\mu = 0.01$ (solid line), $\mu = 0.1$ (dashed line), $\mu = 0.5$ (dotted line) for $\theta = 50$.

Furthermore, in order to visualize the different tail behavior of the Gompertz and the GS densities, we matched the first and the third quartiles, $Q1$ and $Q3$, of the two models. They are illustrated in Figure 4.11, the solid lines and the dashed lines, respectively, represent the GS and the Gompertz distributions. The matched quartile values in the figure from left to right are, respectively, ($Q1 = 25$, $Q3 = 55$), ($Q1 = 50$, $Q3 = 70$), and ($Q1 = 80$, $Q3 = 90$). These density curves indicate that the GS density function possesses a thicker left tail and thinner right tail than that of the Gompertz density function.

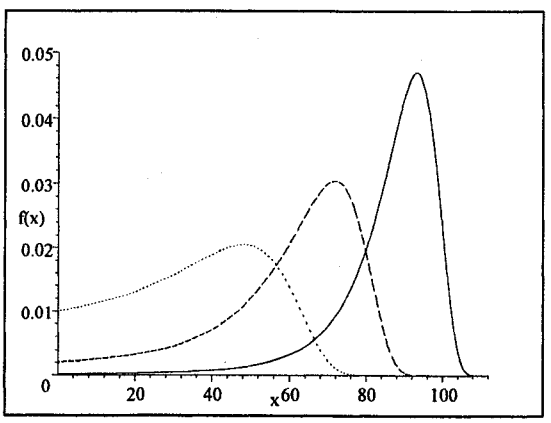

Figure 4.10a GS density curves.

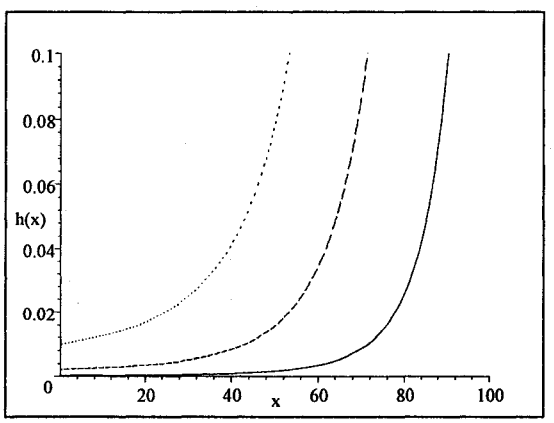

Figure 4.10b GS hazard curves.

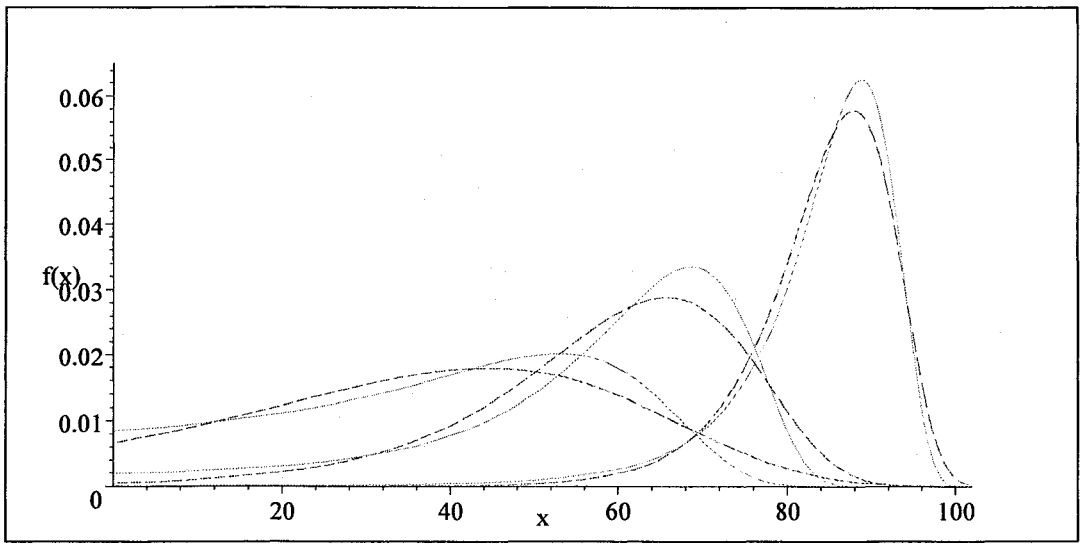

Figure 4.11 Matched first and third quartiles of Gompertz (dashed lines) and GS (solid lines).

Figure 4.12a and Figure 4.12b, respectively, represent the skewness (S $=\sqrt{\beta_1}$) and kurtosis (K $= \beta_2$) variations against the shape parameters of the two models. The solid lines and the dashed lines, respectively, represent the GS distribution with $\mu = X$ and the Gompertz model with $\gamma = X$. These quantities, the skewness and the kurtosis are not depend on the scale parameters of the two models. From figure 4.12a, one can clearly see that the GS distribution is more highly negatively skewed than the Gompertz distribution. For example, when $\gamma = 0.001$, the Gompertz distribution gives the skewness value, -1.0076 with its kurtosis value, 4.5018. The minimum attainable kurtosis for the Gompertz family is 2.2717 and it occurs when $\gamma = 0.181$ with skewness value, 0.0527. Similarly, when $\mu = 0.001$, the GS distribution gives the skewness value, -2.4592 with its kurtosis value, 15.2929. The minimum attainable kurtosis for the GS distribution is 1.9945 and it occurs when $\mu = 0.696$ with skewness

value, -0.0133.

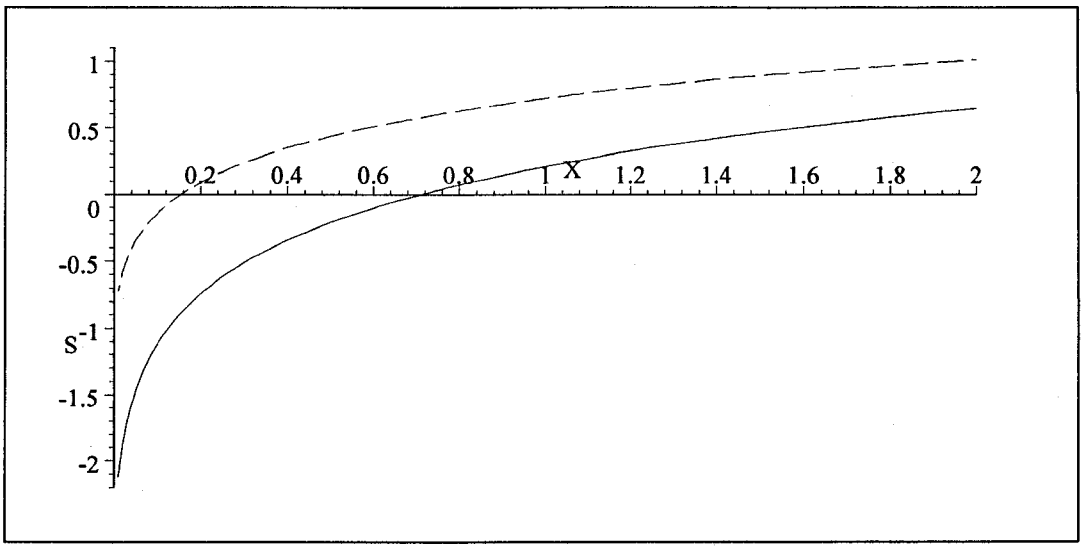

Figure 4.12a Skewness (S) variation with shape parameter (X) of GS (solid line) and G (dashed line).

Overall, the GS distribution is more highly negatively skewed than the Gompertz distribution and the upper tail of the GS distribution is thinner than that of the Gompertz distribution. Therefore, in the presence of highly negatively skewed data, the new GS distribution provides a better fit than the Gompertz distribution. The details of this assertion are exemplified by the data given in the examples in Section 4.8.6.

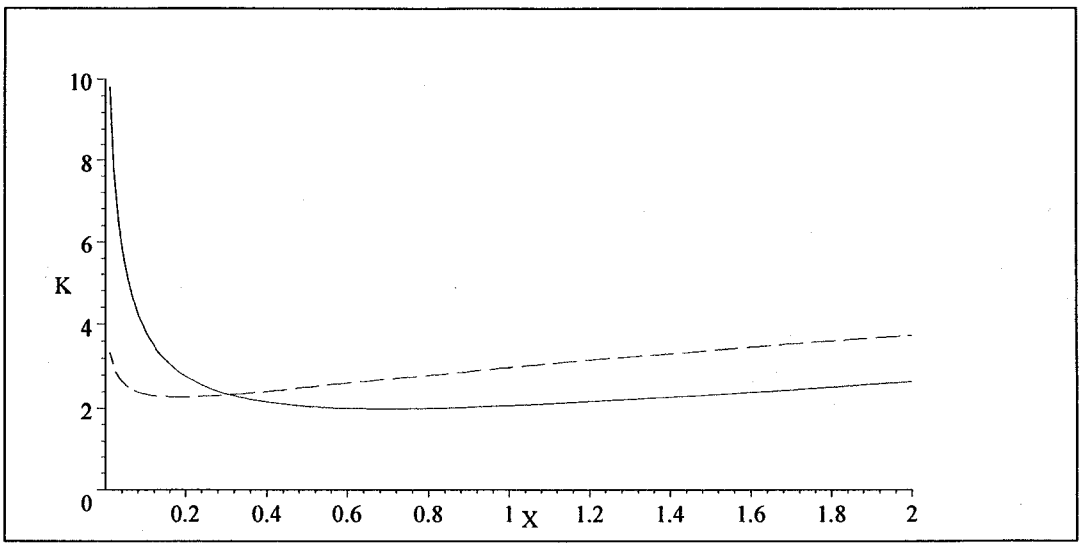

Figure 4.12b Kurtosis (K) variation with shape parameter (X) of GS (solid line) and G (dashed line).

Alternative functional forms for the GS distribution

Consider the following function,

$$F(x;\mu,\theta) = 1 - \exp\left(-\mu g\left(x;\theta\right)\right),$$

where $0 < x < \infty, 0 < \mu < \infty,\ 0 < \theta < \infty$, and $g\left(.\right)$ is any increasing function in $\mathbb{R}^{+}$ such that $g\left(0\right) = 0$ and $g\left(\infty\right) = \infty$. Then $F(x;\mu,\theta)$ is also a distribution function over $0 \leq x < \infty$. In order to obtain a negatively skewed density function over $0 \leq x < \infty$, one can select $g\left(.\right)$ such that $\lim_{x\to\infty}(g(x)/x^{3.6}) \to \infty$. The value 3.6 is obtained as a minimum value of the shape parameter for negative skewness of the Weibull distribution. For example, the following functional forms would be possible for $g(.)$ to model highly negatively skewed data.

$g_1\left(x;\theta\right) = \sinh\left(\sinh\left(x/\theta\right)\right),\ g_2\left(x;\theta\right) = \exp\left(\exp\left(x/\theta\right) - 1\right) - 1,$

$g_3\left(x;\theta\right) = \exp\left(\sinh\left(x/\theta\right)\right) - 1.$

The density function corresponding to $g_1(.)$ selection gives an antimode for lower μ values, even though it gives higher skewness than the GS density. The distribution of $g_2(.)$ selection skewed negatively much less than the GS distribution and hence it may not be much different from the Gompertz distribution in the analysis of highly negatively skewed data. The density function corresponding to $g_3(.)$ selection is interesting, since it shows much closer relation to the GS distribution in terms of density shapes, hazard shapes, skewness, and kurtosis. Although, in the analysis of data sets like the Badenscallie burial data given in Example 1 in Section 4.8.6, gives a lower p-values than that of the GS distribution. Therefore, in this section we are not interested in providing a detail analysis of the $g_3(.)$ selection.

4.8.3 The exponentiated Gompertz-sinh family

In survival analysis, increasing failure rates and bathtub-shaped failure rates or *the curve of deaths* (Bowers *et al.* 1986) which represent the shape of human mortality (see the solid line in Figure 4.13) are commonly encountered. Distributions with one- or two-parameters impose strong restrictions on bathtub hazard shapes or the shapes of the curve of deaths. In general, at least three parameters are needed to form flexible bathtub-shaped hazard functions. On the other hand, more flexible distributions usually have more than three parameters and they will become unattractive due to problems related to parameter estimation. These arguments, together with consideration of computational simplicity, led us to search for a useful generalization with a minimum number of parameters, with the capacity to describe increasing and bathtub-shaped hazard as well as the curve of deaths. In this section, we generalized

the GS distribution by exponentiating its distribution function with an additional shape parameter. Therefore, this proposed family of distributions is referred to as the exponentiated Gompertz-sinh distribution (EGS).

The distribution function and the probability density function of the EGS distribution are, respectively, given by

$$F(x) = \left(1 - \exp\left(-\mu \sinh(e^{x/\theta} - 1)\right)\right)^{\beta}, \tag{4.16}$$

and

$$f(x) = \left(\frac{\beta\mu}{\theta}\right) e^{x/\theta} \cosh(e^{x/\theta} - 1) e^{-\mu \sinh(e^{x/\theta}-1)} \left(1 - e^{-\mu \sinh(e^{x/\theta}-1)}\right)^{\beta-1}. \tag{4.17}$$

Where $0 < x < \infty$, $0 < \mu < \infty$, $0 < \beta < \infty$, and $0 < \theta < \infty$.

Figure 4.13 shows the EGS density curves for different parameter values. Solid line, dashed line, dotted line, and dot dashed line indicate, respectively, the parameter values $(\mu = 0.01, \beta = 0.4, \theta = 45)$, $(\mu = 1, \beta = 0.75, \theta = 70)$, $(\mu = 7, \beta = 3, \theta = 120)$, and $(\mu = 0.2, \beta = 2, \theta = 60)$. The graphical analysis of the density function shows that the curve of deaths of this family will appear whenever $\mu < \beta < 1$. Moreover, when ($\mu < 1$ and $\beta \geqslant 1$) or ($\mu \geqslant 1$ and $\beta > 1$) the density is unimodal-shaped.

The Figure 4.14 shows the EGS hazard curves for different parameter values. Solid line (which illustrates the so called curve of deaths), dotted line, dashed line, and dot dashed line indicate, respectively, the parameter values $(\mu = 3, \beta = 0.1, \theta = 100)$, $(\mu = 0.8, \beta = 2, \theta = 80)$, $(\mu = 100, \beta = 4, \theta = 600)$, and $(\mu = 10000, \beta = 1, \theta = 100000)$. Except for larger values of θ, whenever $\beta \lesssim 1$, the EGS hazard function is, respectively, bathtub-shaped or increasing.

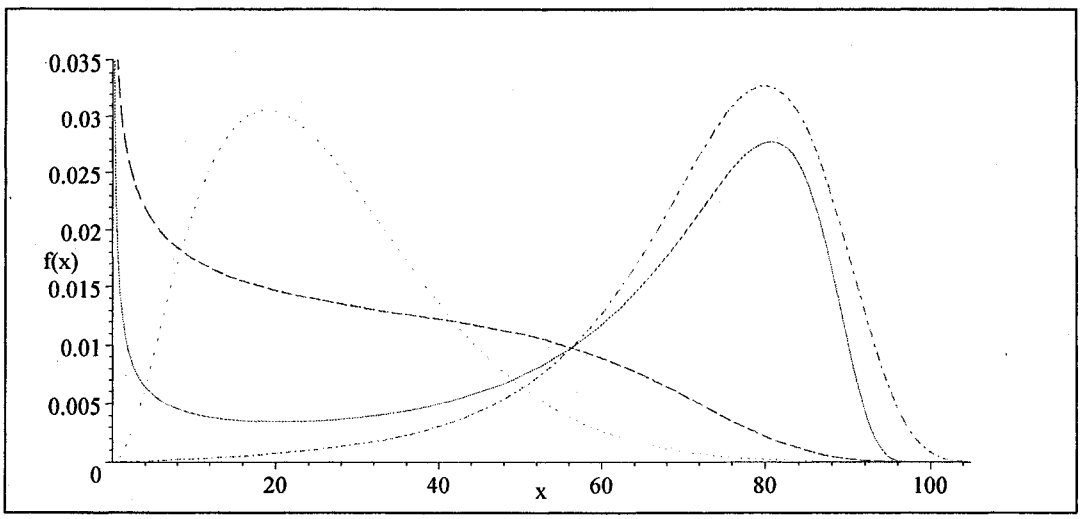

Figure 4.13 EGS density curves. Solid line ($\mu = 0.01, \beta = 0.4, \theta = 45$), dashed line ($\mu = 1, \beta = 0.75, \theta = 70$), dotted line ($\mu = 7, \beta = 3, \theta = 120$), and dot dashed line ($\mu = 0.2, \beta = 2, \theta = 60$).

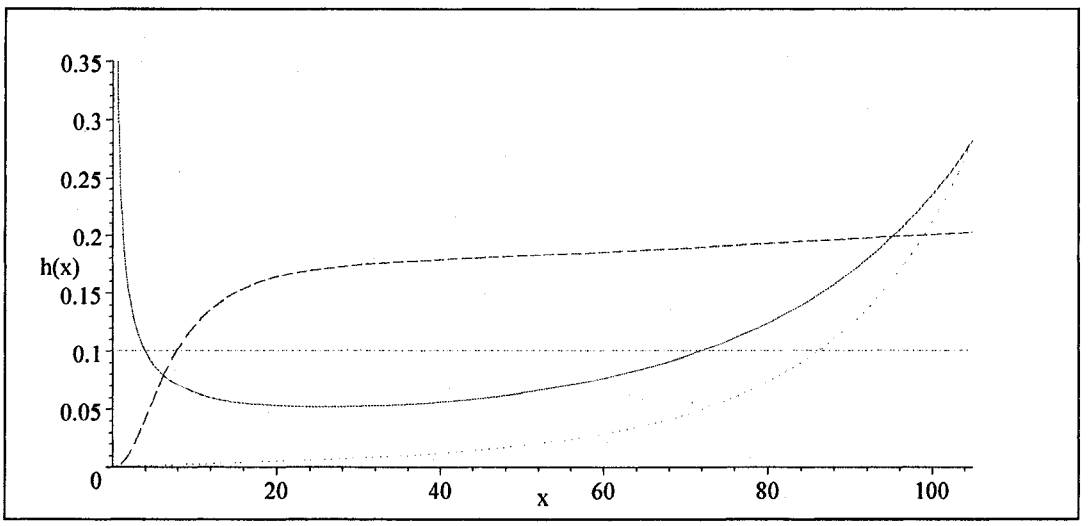

Figure 4.14 EGS hazard curves. Solid line ($\mu = 3, \beta = 0.1, \theta = 100$), dotted line ($\mu = 0.8, \beta = 2, \theta = 80$), dashed line ($\mu = 100, \beta = 4, \theta = 600$), and dot-dashed line ($\mu = 10000, \beta = 1, \theta = 100000$).

4.8.4 Parametric inference

The parametric inference of distributions like the GS and EGS family for the given data are typically based on likelihood methods and their asymptotic theory (Cox and Oakes 1984; Lawless 2003; Rao 1973). For our simplicity, we provide the estimation procedure for the EGS model. One can obtain the required results for the GS model by setting the parameter $\beta = 1$. The estimates of the parameters are obtained by maximizing the log-likelihood function ($l(\boldsymbol{\theta}) = \ln L(x_1, x_2, ..., x_n; \mu, \beta, \theta)$).

The log-likelihood function of the EGS family is given by

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{n} \delta_j \left\{ \ln\left(\cosh\left(e^{x_j/\theta} - 1\right)\right) - \mu \sinh\left(e^{x_j/\theta} - 1\right) \right\} \\ &\quad + \sum_{j=1}^{n} \delta_j \left\{ \ln\left(\mu\beta/\theta\right) + \frac{x_j}{\theta} + (\beta - 1)\ln(1 - \exp(-\mu \sinh\left(e^{x_j/\theta} - 1\right))) \right\} \\ &\quad + \sum_{j=1}^{n} (1 - \delta_j) \ln\left(1 - (1 - \exp(-\mu \sinh\left(e^{x_j/\theta} - 1\right)))^{\beta}\right), \end{aligned} \tag{4.18}$$

where δ_j is such that

$$\delta_j = \begin{cases} 0 & \text{if } j^{th} \text{ observation is right-censored} \\ 1 & \text{if } j^{th} \text{ observation is not right-censored} \end{cases} \qquad j = 1, 2, ..., n.$$

For the purpose of analyzing the grouped data by estimating the parameters of the EGS model under the ML method, one can assume that the data consists of r intervals and the jth interval, i.e. (c_{j-1}, c_j), has n_j observations for $j = 1, 2, 3, ..., r$; $c_0 = 0$. The rth open interval, i.e. (c_{r-1}, ∞), contains n_r observations, and the total number of observations, n can be written as $\sum_{j=1}^{r} n_j$. Therefore the log-likelihood function of the EGS model for grouped data can be written as

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{r} n_j \ln\left[F(c_j; \mu, \beta, \theta) - F(c_{j-1}; \mu, \beta, \theta)\right], \tag{4.19}$$

where $F(.;\mu,\beta,\theta)$ is the distribution function of the EGS family.

In this case, the log-likelihood function is maximized by solving the score equation $U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$. For large samples, asymptotic normality results hold for estimated parameters values, i.e., $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} \mathbf{N}_3(\mathbf{0},\ I^{-1}(\boldsymbol{\theta}))$, where $\mathbf{N}_3$ denote the trivariate normal distribution and $I(\boldsymbol{\theta})$ is the expected Fisher information matrix of $\widehat{\boldsymbol{\theta}}$ such that $I(\boldsymbol{\theta}) = -\mathbf{E}\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$. The variance-covariance matrix, $I(\boldsymbol{\theta})$ is useful to construct approximate confidence intervals for individual parameters and functions of such parameters (Rao 1973). Moreover, whether the data are complete or right censored, inference procedures based on maximum likelihood large-sample theory can be applied in a straightforward way (Lawless 2003).

The fitting of the EGS family by solving the score equation involving in the log-likelihood function can be facilitated using computer programs such as DNEQNF in IMSL (1991) and LE in BMDP (1992). The LE program in BMDP (1992), which uses Newton-Raphson type algorithm to maximize the likelihood function. It can easily be employed to estimate unknown parameters, whether the data is in complete, grouped, truncated or censored form. The LE routine also gives the asymptotic standard errors (SE's) of the estimates by inverting the Hessian matrix used in the maximization of the likelihood function, unless the information matrix is ill-conditioned. The information matrix may be ill conditioned due to singularity or near singularity of the Hessian matrix.

4.8.5 Approximate coverage probabilities of the GS distribution

The coverage probabilities of the GS distribution for the maximum likelihood estimation method with intended confidence levels $\alpha = 0.1$ and $\alpha = 0.05$ are given in Table 4.7. These coverage probabilities are based on 100,000 simulated random samples from the density given in (4.13). The random samples are generated by plugging the known values of parameters μ and θ (say $\mu = 0.01$, $\theta = 10$) to the quantile function given in equation (4.14). In addition, n (say $n = 10$) number of ordered uniform random sample from the uniform distribution, $u \sim U(0,1)$ is required to substitute as u in equation (4.14). In that way, one random sample with size n (say $n = 10$) from the GS distribution with parameters μ and θ (say $\mu = 0.01$, $\theta = 10$) can be generated. In this simulation study, ten thousand such samples are generated to get a single cell value in Table 4.7. For this purpose, the subroutine ZBREN in the IMSL (1991) package is used to solve the following nonlinear equation of $\widehat{\theta}$, the ML estimator of θ,

$$\widehat{\mu}\sum_{i=1}^{n} x_i e^{x_i/\widehat{\theta}} \cosh(e^{x_i/\widehat{\theta}} - 1) - \sum_{i=1}^{n} x_i e^{x_i/\widehat{\theta}} \tanh(e^{x_i/\widehat{\theta}} - 1) - \sum_{i=1}^{n} x_i - n\widehat{\theta} = 0, \tag{4.20}$$

where $\widehat{\mu}$, the ML estimator of μ, is given by,

$$\widehat{\mu} = \left(\frac{1}{n}\sum_{i=1}^{n} \sinh(e^{x_i/\widehat{\theta}} - 1)\right)^{-1}. \tag{4.21}$$

Approximate $100\,(1-\alpha)\,\%$ confidence intervals for, μ and θ are, respectively, calculated by using $\left(\widehat{\mu} - Z_{\alpha/2}SE_{\widehat{\mu}},\ \widehat{\mu} + Z_{\alpha/2}SE_{\widehat{\mu}}\right)$ and $\left(\widehat{\theta} - Z_{\alpha/2}SE_{\widehat{\theta}},\ \widehat{\theta} + Z_{\alpha/2}SE_{\widehat{\theta}}\right)$. Where $SE_{\widehat{\mu}}$ and $SE_{\widehat{\theta}}$ are, respectively, asymptotic standard errors of $\widehat{\mu}$ and $\widehat{\theta}$, which are taken from the observed information matrix (Efron and Hinkley 1978).

From Table 4.7, one can see that when the sample size increases, the approximate coverage probabilities for the parameters under the maximum likelihood method are getting closer to the intended coverage probabilities. In addition, the approximate coverage probabilities decrease when the shape parameter μ decreases from 0.5. The values in Table 4.7 predict that the parameters do not overly estimate under the maximum likelihood estimation method.

Table 4.7 Approximate coverage probabilities of the GS

	$n =$	10			20			50		
90% intended	$\theta =$	10	20	50	10	20	50	10	20	50
$\mu = .005$	μ:	.690	.690	.688	.767	.769	.769	.839	.840	.840
	θ:	.822	.824	.823	.860	.861	.859	.884	.886	.884
$\mu = .010$	μ:	.704	.702	.704	.779	.782	.784	.846	.845	.845
	θ:	.822	.823	.823	.859	.862	.864	.884	.883	.883
$\mu = .100$	μ:	.756	.757	.759	.821	.823	.824	.867	.867	.868
	θ:	.813	.812	.816	.859	.859	.858	.882	.884	.882
$\mu = .500$	μ:	.771	.770	.770	.832	.832	.831	.877	.875	.875
	θ:	.762	.763	.762	.827	.827	.825	.875	.875	.872
95% intended										
$\mu = .005$	μ:	.716	.718	.716	.798	.801	.800	.874	.874	.874
	θ:	.874	.875	.875	.911	.912	.912	.934	.936	.935
$\mu = .010$	μ:	.733	.731	.733	.812	.814	.817	.882	.881	.881
	θ:	.874	.874	.874	.910	.913	.913	.935	.934	.933
$\mu = .100$	μ:	.790	.791	.794	.857	.858	.859	.907	.908	.909
	θ:	.861	.860	.864	.905	.906	.906	.931	.932	.932
$\mu = .500$	μ:	.811	.810	.809	.872	.871	.870	.917	.916	.916
	θ:	.810	.812	.810	.873	.872	.871	.919	.918	.917

4.8.6 Illustrative examples

In this section, two different examples are provided to illustrate the applicability and flexibility of the GS and the EGS models over the Gompertz distribution.

Example 1. Badenscallie burial data

This example (see data in the appendix B) is provided by Sprent and Smeeton (2000) regarding the death times of male members of Scottish clan. The authors provided several uncensored data sets, and they pointed out that the McAlpha clan data set possesses reasonably approximate pattern of death ages. Furthermore, the sample sizes of the other data sets are quite small and therefore we analyze only the McAlpha clan data set using Gompertz, GS, and EGS family.

The ordered data given below are the age of death of 59 male members of Scottish McAlpha clan in the burial ground at Badenscallie in the Coigach district of Wester Ross, Scotland. Ages are given for complete years, e.g. 0 means before first birthday, and 79 means on or after 79th but before 80th birthday, according to the information on the tombstone.

This example was analyzed using Gompertz, GS, and EGS models, by treating the observations X_i, $i = 1, 2, ..., 59$ as $X_i = \lfloor \widetilde{X}_i \rfloor$, where $\widetilde{X}_1, \widetilde{X}_2, ..., \widetilde{X}_{59}$ have one of the distributions specified previously.

The EGS family is used by replacing the X_i values with $\max(X_i, \delta)$ for some small $\delta > 0$ and to pretend that the latter values come from the EGS model. Then the results depend strongly on the threshold δ. Table 4.8a provides the estimated values by varying δ for the EGS model. It also includes the estimated values for the Gompertz and GS model with $\delta = 0$. It should be mentioned here that among the

other well known distributions, the Gompertz distribution is the closest and natural competitor for the GS distribution. In addition, the well-known distributions such as, lognormal, gamma, Weibull, Pareto, etc., perform very poorly with this data set, and even the Gompertz distribution gives a very poor fit with p-value 0.02.

The Pearson's goodness-of-fit chi-squared test has been performed by using a 10-year class width by treating the data as grouped data (see table 4.8b). For this purpose we use the log-likelihood function given in equation (4.18), Section 4.8.4 to estimate the parameters of the GS and EGS model.

From Table 4.8a and 4.8b, one can see that the values for measures of log-likelihood, Kolmogorov-Smirnov (D) statistics and especially the higher p-values have emphasized that the EGS gives very good fit to the death data for McAlpha clan, whereas GS provides an acceptable fit. Note that the fitted log-likelihood values of the EGS model for two threshold values should not be taken too seriously, since the EGS densities are unbounded at zero if $\beta < 1$. The fitness is further strengthened by illustrating the fitted survival function of the EGS model with threshold parameter $\delta = 0.1$ along with the Kaplan-Meier curve. For comparison purposes the fitted survival function of GS and Gompertz models are also included. Figure 4.15, the Kaplan-Meier curve, the EGS survival fit with $\delta = 0.1$, the GS survival fit, and the Gompertz survival fit are, respectively, represented by the step function, the dark solid line, the dashed line, and the dotted line. Note that one can obtain different chi-squared values by considering the minimum expected frequencies as 5.

Table 4.8a Estimated values of three models for the burial data

Model	δ	Parameters $\pm$ SE	$l(\widehat{\boldsymbol{\theta}})$	D
Gompertz	0	$\widehat{\gamma} = 0.02626 \pm 0.014869$ $\widehat{\phi} = 19.9801 \pm 2.689128$	-267.85	0.17
GS	0	$\widehat{\mu} = 0.14440 \pm 0.049334$ $\widehat{\theta} = 58.7310 \pm 3.211787$	-259.75	0.11
EGS	0.0001	$\widehat{\mu} = 0.00292 \pm 0.003789$ $\widehat{\beta} = 0.29550 \pm 0.070728$ $\widehat{\theta} = 43.1740 \pm 2.995459$	-242.23	0.07
EGS	0.1	$\widehat{\mu} = 0.00925 \pm 0.010309$ $\widehat{\beta} = 0.39100 \pm 0.098251$ $\widehat{\theta} = 46.0917 \pm 3.413815$	-251.37	0.07

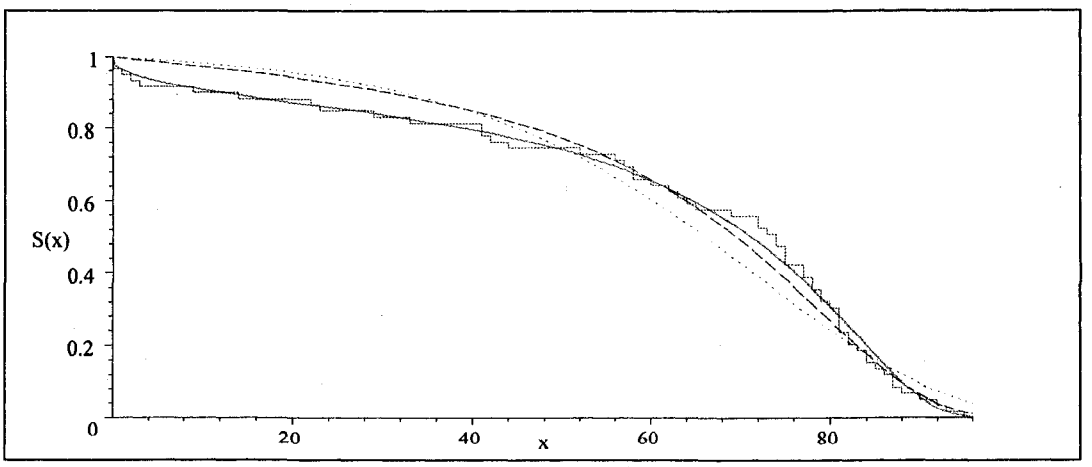

Figure 4.15 Fitted G, GS, and EGS survival curves for burial data. Kaplan-Meier curve (step function), Gompertz (dotted line), GS (dashed line), and EGS with $\delta = 0.1$ (dark solid line).

Table 4.8b Estimated values of three models for the grouped burial data

Age interval (Years)		Observed frequency	Expected frequency		
			Gompertz	GS	EGS
0	10	6	1.1020	1.6556	5.8041
10	20	1	1.7183	1.9638	1.9771
20	30	3	2.6405	2.4332	1.8631
30	40	1	3.9641	3.1852	2.1943
40	50	4	5.7341	4.4185	3.0079
50	60	5	7.8181	6.4113	4.5862
60	70	6	9.7035	9.3444	7.5013
70	80	14	10.3798	12.4262	12.4324
80	90	15	8.7871	11.9957	15.7652
90	up	4	7.1523	5.1659	3.8685
Parameters $\pm$ SE			$\widehat{\gamma} = 0.03154$ ± 0.01828 $\widehat{\phi} = 21.3360$ $\pm$ 3.13861	$\widehat{\alpha} = 0.15732$ $\pm$ 0.05623 $\widehat{\theta} = 60.4329$ ± 3.79220	$\widehat{\alpha} = 0.00452$ ± 0.00894 $\widehat{\beta} = 0.34298$ ± 0.16552 $\widehat{\theta} = 44.2816$ ± 5.28530
$l(\widehat{\boldsymbol{\theta}})$			-131.03	-125.29	-120.52
χ^2_{df}			$\chi^2_7 = 34.33$	$\chi^2_7 = 16.27$	$\chi^2_6 = 2.74$
p-value			0.00	0.02	0.84

Example 2. Diabetic data

In this example (see data in the appendix B), we analyze the following right censored data set related to the survival times (in years) of 149 diabetic patients who were followed for 17 years (Lee and Wang 2003):

For this data set, we are not providing Kolmogorov-Smirnov statistic values due to computational complexities for censored samples. However, one such method to handle right censored data is given in Fleming *et al.* (1980). The p-values for this right censored data are based on discretized method introduced by Efron (1988). The data were discretized as of Table 4.9b in which includes the signed deviance residuals, R_j, given by the formula (2.69) (see McCullagh and Nelder 1998); where, N_j is total number patients at risk at the beginning of each interval j, $j = 1, ..., 11$, S_j is observed death at the end of each interval, E_j is expected death at the end of each interval, for the three hazard models (Efron 1988).

The results from three models are given in Tables 4.9a and 4.9b. Once again, both likelihood values and p-values indicate that the EGS, GS distributions fit better than the Gompertz distribution. Note that the parameters of the EGS family are estimated by replacing the data, X_i, values with $\max(X_i, \delta)$ for some small $\delta > 0$ and to pretend that the latter values come from the EGS model. Then the results depend strongly on the threshold δ. However, not like the previous example, the effect of δ is not too considerable, since the data set consists a single zero data value. Therefore, we analyze this example using $\delta = 0.05$. Once again, the well-known distributions such as, lognormal, gamma, Weibull, Pareto, etc., are not appropriate to analyze this example due their poor fits.

Table 4.9a Estimated values of the three model for the diabetic data

Model	Parameters $\pm$ SE	$l(\widehat{\boldsymbol{\theta}})$	p-value
Gompertz	$\widehat{\gamma} = 0.01307 \pm 0.0051$ $\widehat{\phi} = 2.97980 \pm 0.2443$	-343.84	0.01
GS	$\widehat{\alpha} = 0.12229 \pm 0.0271$ $\widehat{\theta} = 9.95490 \pm 0.3396$	-335.97	0.30
EGS	$\widehat{\alpha} = 0.07245 \pm 0.0394$ $\widehat{\beta} = 0.80240 \pm 0.1546$ $\widehat{\theta} = 9.41168 \pm 0.5448$	-335.29	0.37

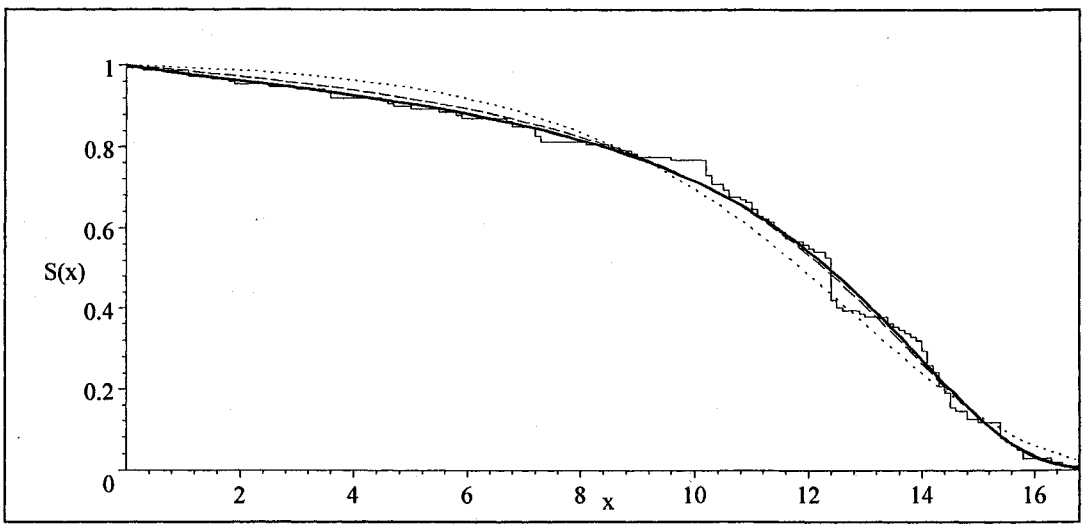

Figure 4.16 Fitted G, GS, and EGS survival curves for diabetic data. Kaplan-Meier (step function), G (dotted line), GS (dashed line), and EGS (dark solid line).

As before, The fitness is strengthened by illustrating the fitted survival functions of the EGS and GS models along with the Kaplan-Meier curve. For comparison

purposes the fitted survival function of the Gompertz model is also included. In Figure 4.16, the Kaplan-Meier curve, the EGS survival fit, the GS survival fit, and the Gompertz survival fit are, respectively, represented by the step function, dark solid line, dashed line, and the dotted line.

Table 4.9b Residual analysis of the three hazard models for the diabetic data

Class interval		N_j	S_j	Expected death (E_j)			Deviance residual (R_j)		
				G*	GS	EGS	G*	GS	EGS
0	1	149	2	0.78	1.93	3.16	1.16	0.05	-0.71
1	2	147	5	1.07	2.13	2.63	2.77	1.69	1.32
2	3	141	2	1.44	2.32	2.60	0.44	-0.22	-0.39
3	4	138	3	1.97	2.63	2.77	0.67	0.22	0.14
4	5	133	4	2.65	3.01	3.04	0.78	0.55	0.53
5	6	128	3	3.57	3.52	3.45	-0.32	-0.29	-0.25
6	7	123	3	4.80	4.24	4.07	-0.90	-0.65	-0.57
7	8	117	5	6.39	5.22	4.94	-0.59	-0.10	0.03
8	9	109	5	8.32	6.50	6.11	-1.29	-0.63	-0.48
9	11	205	17	26.20	20.52	19.23	-2.04	-0.84	-0.54
11	17	246	75	88.75	93.12	91.85	-1.85	-2.42	-2.25

$*$ where G stands for Gompertz distribution.

CHAPTER V

FOLDED PARAMETRIC FAMILIES

5.1 Introduction

Physical measurements like dimensions including time and angles in scientific experiments are frequently recorded without their algebraic sign. The directions of those physical quantities measured with respect to a frame of reference in most practical applications are considered to be unimportant and ignored. As a consequence, the underlying distribution of measurements is replaced by a distribution of absolute measurements. When the underlying distribution is normal, logistic, Laplace, and Cauchy the resulting distribution is, respectively, called the "folded normal", "folded logistic", "folded Laplace", and "folded Cauchy" distribution.

For example, whenever a difference or deviation is measured, or measurements like length, distance, or angle is taken on either side of a line of reference, and when the algebraic sign is unknown, disregarded, or lost, the resulting distribution of these absolute measurements can range in shape from thinnest right tail (half normal; Daniel 1959) via medium right tail (half logistic; Balakrishnan 1992) to thickest right tail (half Cauchy; Johnson *et al.* 1994) or concerning the more variability thinnest right tail (folded normal; Leone *et al.* 1961) via median right tail (folded logistic; Cooray *et al.* 2006) to thickest right tail (folded Cauchy; Johnson *et al.* 1994). The effect of dropping the sign adds the otherwise negative values to the positive values. Geomet-

rically, this amounts to the folding of the negative side of the distribution onto the positive side. Traditionally, for the convenience of writing properties of distributions, researchers have been folding distributions at the mean. The half normal (Daniel 1959), half logistic (Balakrishnan 1992), and two-fold t distribution (half t distribution) (Psarakis and Panaretos 1990) are some examples for such half distributions.

But, Leone *et al.* (1961) folded the normal distribution at a general point from the mean and, using first and second moments, gave the method of moment estimators. Later, Elandt (1961) proposed an alternative method of moment estimators for these parameters using the second and the fourth moments. Sundberg (1974) gives statistical inference procedures for the folded normal distribution, and for some other related work see Nelson (1980) and Risvi (1971).

In this section, we consider the above mentioned four folded family of distributions, which are derivations of the original four pdf folded at a general point rather than at their mean. Estimation procedures are discussed through real data examples. Note that all these folded distributions are positively skewed and have non zero density value at the origin and, therefore these models are useful to analyze the data sets with zero data values.

5.2 The folded normal distribution

The density function and the distribution function of the folded normal distribution (Leone *et al.* 1961) are, respectively, given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\left[e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} + e^{-\frac{1}{2}(\frac{x+\mu}{\sigma})^2}\right], \tag{5.1}$$

and

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) + \Phi\left(\frac{x+\mu}{\sigma}\right) - 1. \tag{5.2}$$

Where $0 \leq x < \infty$, $-\infty < \mu < \infty$, and $0 < \sigma < \infty$.

Median (m) and positive mode (m_0) of the distribution can, respectively, be calculated from

$$\Phi\left(\frac{m-\mu}{\sigma}\right) + \Phi\left(\frac{m+\mu}{\sigma}\right) - 1.5 = 0, \tag{5.3}$$

and

$$(m_0 - \mu)e^{2m_0\mu/\sigma^2} + m_0 + \mu = 0. \tag{5.4}$$

We will not discuss any analysis of the folded normal distribution, since one can find related discussion in Leone *et al.* (1961), Elandt (1961), Risvi (1971), Sundberg (1974), and Nelson (1980).

5.3 The folded logistic distribution

The density function, the distribution and the quantile function of the folded logistic distribution (Cooray *et al.* 2006) are, respectively, given by

$$\begin{aligned} f(x) &= \frac{e^{(\frac{x-\mu}{\sigma})}}{\sigma(1+e^{(\frac{x-\mu}{\sigma})})^2} + \frac{e^{(\frac{x+\mu}{\sigma})}}{\sigma(1+e^{(\frac{x+\mu}{\sigma})})^2} = \frac{e^{-(\frac{x-\mu}{\sigma})}}{\sigma(1+e^{-(\frac{x-\mu}{\sigma})})^2} + \frac{e^{-(\frac{x+\mu}{\sigma})}}{\sigma(1+e^{-(\frac{x+\mu}{\sigma})})^2} \\ &= \frac{1}{4\sigma}\text{sech}^2(\frac{x-\mu}{2\sigma}) + \frac{1}{4\sigma}\text{sech}^2(\frac{x+\mu}{2\sigma}), \end{aligned} \tag{5.5}$$

$$\begin{aligned} F(x) &= 1 - \frac{1}{1+e^{(\frac{x-\mu}{\sigma})}} - \frac{1}{1+e^{(\frac{x+\mu}{\sigma})}} = \frac{1}{1+e^{-(\frac{x-\mu}{\sigma})}} + \frac{1}{1+e^{-(\frac{x+\mu}{\sigma})}} - 1 \\ &= \frac{1}{2}\tanh(\frac{x-\mu}{2\sigma}) + \frac{1}{2}\tanh(\frac{x+\mu}{2\sigma}), \end{aligned} \tag{5.6}$$

and

$$Q(u) = \sigma \ln[\frac{(1+k^2)u + \{4k^2 + (1-k^2)^2u^2\}^{1/2}}{2k(1-u)}]. \tag{5.7}$$

Where $0 \leq x < \infty$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$, $0 \leq u \leq 1$, and $k = e^{\mu/\sigma}$.

Figure 5.1 shows some characteristic shapes of folded logistic distribution varying with μ and σ.

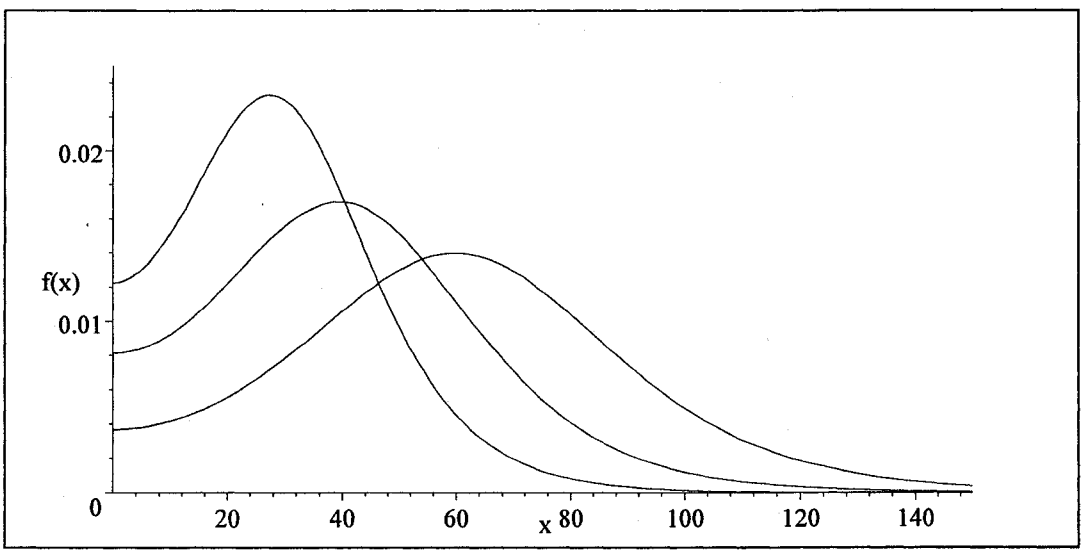

Figure 5.1 Folded logistic density curves varying with μ and σ.

Further, moments of the folded logistic distribution can be obtained from the following equations.

$$\begin{aligned} E[X^{2r-1}] &= 2(2r-1)\mu^{2r-1}\{\frac{\sigma}{\mu}\ln(1+e^{-\frac{\mu}{\sigma}}) + \sum_{k=1}^{\infty}\binom{2r-2}{k}(-\frac{\sigma}{\mu})^{k+1}I_k(\frac{\mu}{\sigma}) \\ &\quad + \sum_{k=2}^{\infty}\binom{2r-2}{2k-2}(\frac{\sigma}{\mu})^{2k-1}(1-2^{2-2k})\Gamma(2k-1)\zeta(2k-1)\} + \mu^{2r-1} \end{aligned} \tag{5.8}$$

$$E[X^{2r}] = 4r\mu^{2r}\sum_{k=1}^{\infty}\binom{2r-1}{2k-1}(\frac{\sigma}{\mu})^{2k}(1-2^{1-2k})\Gamma(2k)\zeta(2k) + \mu^{2r}. \tag{5.9}$$

Where $r = 1, 2, 3, ...$ and $I_k(\frac{\mu}{\sigma}) = \int_0^{\mu/\sigma} t^k (1+e^t)^{-1} dt$, $k = 1, 2, 3, ...$ The integral $I_k(a)$ can be expressed as a summation of incomplete, $\Gamma(a,\ b)$, and complete, $\Gamma(a)$, gamma functions as follows.

$$I_k(a) = \sum_{i=0}^{\infty} \frac{(-1)^i \Gamma(k+1)\Gamma(k+1, (1+i)a)}{(1+i)^{(1+k)}}. \tag{5.10}$$

Also $\zeta(.)$ is a Riemann zeta function (Abramowitz and Stegun 1972; Chaudhry and Zubair 2001) such that

$$\zeta(2n) = (n+1/2)^{-1} \sum_{k=1}^{n-1} \zeta(2k)\,\zeta(2n-2k), \quad n = 2, 3, 4, ... \tag{5.11}$$

and

$$\begin{aligned} \zeta(2n+1) &= (-1)^n \frac{(2\pi)^{2n}}{n(2^{2n+1}-1)} [\sum_{k=1}^{n-1} \frac{(-1)^{k-1} k}{(2n-2k)!} \frac{\zeta(2k)}{\pi^{2k}} \\ &\quad + \sum_{k=0}^{\infty} \frac{(2k)!}{(2n+2k)!} \frac{\zeta(2k)}{2^{2k}}], \quad n = 1, 2, 3, ... \end{aligned} \tag{5.12}$$

with $\zeta(2) = \frac{\pi^2}{6}$.

The first two moments, median and positive mode of the distribution can respectively be obtained as

$$E(X) = \mu + 2\sigma \ln(1 + e^{-\mu/\sigma}), \tag{5.13}$$

$$E(X^2) = \mu^2 + \pi^2\sigma^2/3, \tag{5.14}$$

$$\text{Median} = \mu + \sigma \ln\{\frac{1}{2}[1 + q^2 + \sqrt{16q^2 + (1-q^2)^2}]\}, \tag{5.15}$$

and

$$\text{Mode} = \mu + \sigma \ln \left[\frac{\left\{ 1 - 6q^2 + q^4 + \sqrt{1 - 16q^2 + 30q^4 - 16q^6 + q^8} \right\}}{2\left(1 + q^2\right)} \right], \tag{5.16}$$

where $q = e^{-\mu/\sigma}$.

Maximum likelihood estimates of μ and σ are obtained in the usual manner by differentiating the log-likelihood function

$$\ln L(\mu, \sigma) = -n \ln 4\sigma + \sum_{i=1}^{n} \ln\{\operatorname{sech}^2(\frac{x_i - \mu}{2\sigma}) + \operatorname{sech}^2(\frac{x_i + \mu}{2\sigma})\}, \tag{5.17}$$

with respect to μ and σ.

The maximum likelihood estimators of μ and σ can be found by numerically solving the following two equations.

$$\sum_{i=1}^{n} \frac{\operatorname{sech}^2(\frac{x_i - \widehat{\mu}}{2\widehat{\sigma}}) \tanh(\frac{x_i - \widehat{\mu}}{2\widehat{\sigma}}) - \operatorname{sech}^2(\frac{x_i + \widehat{\mu}}{2\widehat{\sigma}}) \tanh(\frac{x_i + \widehat{\mu}}{2\widehat{\sigma}})}{\operatorname{sech}^2(\frac{x_i - \widehat{\mu}}{2\widehat{\sigma}}) + \operatorname{sech}^2(\frac{x_i + \widehat{\mu}}{2\widehat{\sigma}})} = 0. \tag{5.18}$$

$$\sum_{i=1}^{n} x_i \frac{\operatorname{sech}^2(\frac{x_i - \widehat{\mu}}{2\widehat{\sigma}}) \tanh(\frac{x_i - \widehat{\mu}}{2\widehat{\sigma}}) + \operatorname{sech}^2(\frac{x_i + \widehat{\mu}}{2\widehat{\sigma}}) \tanh(\frac{x_i + \widehat{\mu}}{2\widehat{\sigma}})}{\operatorname{sech}^2(\frac{x_i - \widehat{\mu}}{2\widehat{\sigma}}) + \operatorname{sech}^2(\frac{x_i + \widehat{\mu}}{2\widehat{\sigma}})} - n\widehat{\sigma} = 0. \tag{5.19}$$

Because of the complexity of calculation we further suggest a moment method of estimation for μ and σ. We have used method of moment estimation procedure similar to the method for folded normal distribution proposed by Leone *et al.* (1961) and Elandt (1961). The estimators of method of moment using first and second, and second and fourth would respectively be given by

$$\widetilde{\mu} = \widetilde{k}\widetilde{\sigma}, \text{ and } \widetilde{\sigma} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\widetilde{k}^2 + (\pi^2/3)}}, \tag{5.20}$$

where $\widetilde{k}$ is a solution of the following equation

$$\{k + 2\ln(1+e^{-k})\}^2 \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \{k^2 + (\pi^2/3)\}\{\frac{1}{n}\sum_{i=1}^{n} x_i\}^2 = 0, \tag{5.21}$$

and

$$\widetilde{\widetilde{\mu}} = \frac{1}{2}\sqrt{30(\frac{1}{n}\sum_{i=1}^{n} x_i^2)^2 - 5(\frac{1}{n}\sum_{i=1}^{n} x_i^4)} - \frac{3}{2}(\frac{1}{n}\sum_{i=1}^{n} x_i^2), \tag{5.22}$$

$$\widetilde{\widetilde{\sigma}}^2 = \frac{3}{\pi^2}\{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \widetilde{\widetilde{\mu}}^2\}. \tag{5.23}$$

5.4 The folded Laplace distribution

The density function, the distribution and the quantile function of the folded Laplace distribution are, respectively, given by

$$\begin{aligned} f(x) &= \begin{cases} \frac{1}{2\sigma}e^{\left(\frac{x-\mu}{\sigma}\right)} + \frac{1}{2\sigma}e^{-\left(\frac{x+\mu}{\sigma}\right)} & \text{if} \quad 0 \le x \le \mu \\ \frac{1}{2\sigma}e^{-\left(\frac{x-\mu}{\sigma}\right)} + \frac{1}{2\sigma}e^{-\left(\frac{x+\mu}{\sigma}\right)} & \text{if} \quad \mu \le x < \infty \end{cases} \\ &= \begin{cases} \frac{1}{k\sigma}\cosh(x/\sigma) & \text{if} \quad 0 \le x \le \mu \\ \left(\frac{1+k^2}{2k\sigma}\right)e^{-x/\sigma} & \text{if} \quad \mu \le x < \infty \end{cases}, \end{aligned} \tag{5.24}$$

$$\begin{aligned} F(x) &= \begin{cases} \frac{1}{2}e^{\left(\frac{x-\mu}{\sigma}\right)} - \frac{1}{2}e^{-\left(\frac{x+\mu}{\sigma}\right)} & \text{if} \quad 0 \le x \le \mu \\ 1 - \frac{1}{2}e^{-\left(\frac{x-\mu}{\sigma}\right)} - \frac{1}{2}e^{-\left(\frac{x+\mu}{\sigma}\right)} & \text{if} \quad \mu \le x < \infty \end{cases} \\ &= \begin{cases} \frac{1}{k}\sinh(x/\sigma) & \text{if} \quad 0 \le x \le \mu \\ 1 - \left(\frac{1+k^2}{2k}\right)e^{-x/\sigma} & \text{if} \quad \mu \le x < \infty \end{cases}, \end{aligned} \tag{5.25}$$

and

$$Q(u) = \begin{cases} \sigma \sinh^{-1}(ku) & \text{if} \quad u \leq (1 - 1/k^2)/2 \\ \sigma \ln\{\frac{1+k^2}{2k(1-u)}\} & \text{if} \quad u \geq (1 - 1/k^2)/2 \end{cases} . \tag{5.26}$$

Where $k = e^{\mu/\sigma}$, $0 < \mu < \infty$, $0 < \sigma < \infty$, and $0 \leq u \leq 1$.

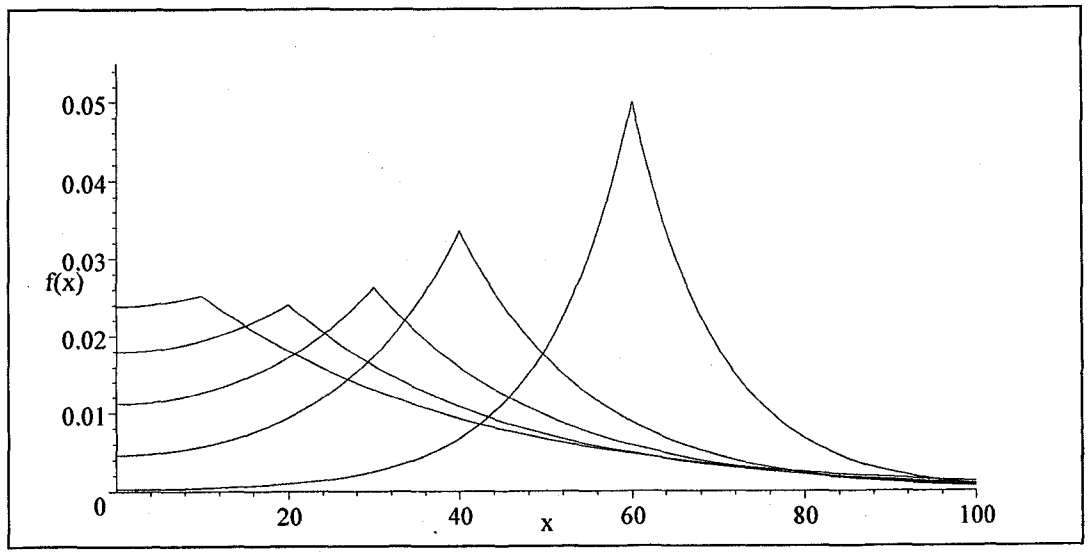

Figure 5.2 Folded Laplace density curves varying with μ and σ.

The moment generating function is

$$M_X(t) = \frac{\sigma t + k^{1+\sigma t}}{k\,(1 - \sigma^2 t^2)}; \quad |t| < 1/\sigma. \tag{5.27}$$

The mean, median, and mode of the distribution are, respectively,

$$\text{Mean} = \sigma\,(\ln k + 1/k) = \mu + \sigma e^{-\mu/\sigma}, \tag{5.28}$$

$$\text{Median} = \sigma \ln\,(k + 1/k) = 2\mu + \sigma \ln(1 + e^{-2\mu/\sigma}), \tag{5.29}$$

and

$$\text{Mode} = \mu. \tag{5.30}$$

More Laplace related distributions and their generalizations with applications can be found in Kotz *et al.* (2001).

5.5 The folded Cauchy distribution

The density function of the folded Cauchy distribution (Johnson *et al.* 1994) is given by

$$f(x) = \frac{1}{\pi\sigma}\left[\frac{1}{1+\left(\frac{x-\mu}{\sigma}\right)^2} + \frac{1}{1+\left(\frac{x+\mu}{\sigma}\right)^2}\right], \tag{5.31}$$

where $0 \leq x < \infty$, $-\infty < \mu < \infty$, and $0 < \sigma < \infty$.

This density can be reparameterize and rewritten as

$$f(x) = \frac{\gamma}{\pi\theta}\frac{1+(\theta/x)^2}{1+\gamma^2\left(\frac{x}{\theta}-\frac{\theta}{x}\right)^2}, \tag{5.32}$$

where $0 \leq x < \infty$, $0 < \gamma < \infty$, $0 < \theta < \infty$, $\theta^2 = \mu^2 + \sigma^2$, and $\gamma = \theta/(2\sigma)$. Note that this form is a harmonic transformation of the Cauchy distribution given by its distribution function

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\left\{\gamma\left(\frac{x}{\theta}-\frac{\theta}{x}\right)\right\}, \tag{5.33}$$

where $0 \leq x < \infty$, $0 < \gamma < \infty$, $0 < \theta < \infty$.

$$\text{Mode} = \begin{cases} 0 & \text{if} \quad \gamma \leq 1/\sqrt{3} \\ \theta\left[2\left\{1-1/\left(2\gamma\right)^2\right\}^{1/2}-1\right]^{1/2} & \text{if} \quad 1/\sqrt{3} < \gamma \end{cases}. \tag{5.34}$$

The parameter θ is the median of the distribution, also it is the geometric mean of the lower and upper α^{th} percentiles of the distribution. The parameter γ is the ratio between median and the difference of upper and lower α^{th} percentiles of the distribution. Hence the MacGillivray's (1992) skewness function $(\gamma_X(u))$ and Galton's skewness function (G) are same for this distribution. The MacGillivray's (1992) skewness function is

$$\gamma_X(u) = \frac{F^{-1}(u) + F^{-1}(1-u) - 2F^{-1}(1/2)}{F^{-1}(u) - F^{-1}(1-u)}, \ 1/2 \leq u \leq 1. \tag{5.35}$$

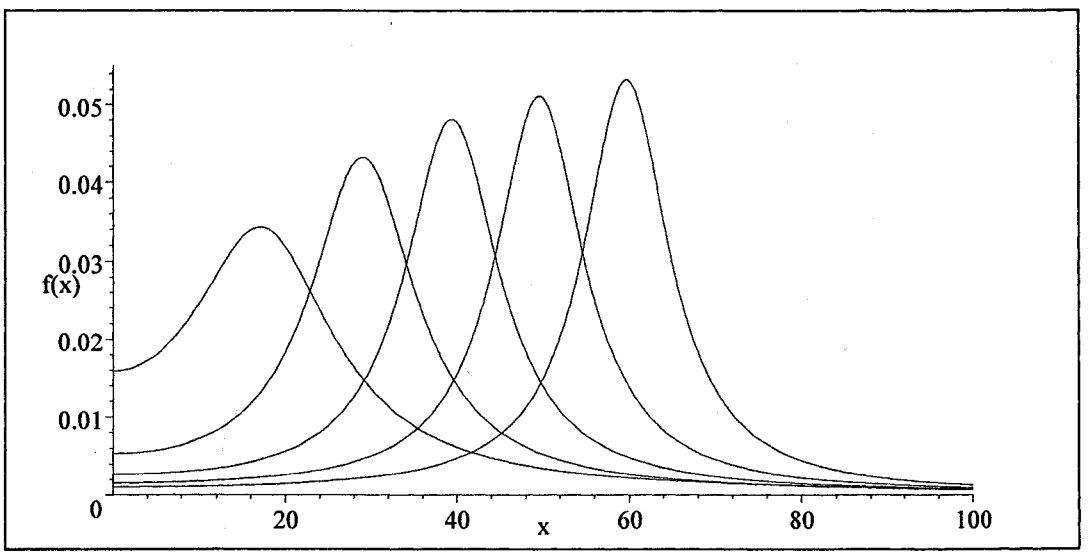

Figure 5.3 Folded Cauchy density curves varying with γ and θ.

The coverage probabilities for the maximum likelihood estimation method with intended confidence levels $\alpha = 0.1$ and $\alpha = 0.05$ are given in Table 5.1. These coverage probabilities are based on 10,000 simulated random samples from the density given in equation (5.32). This analysis is very similar to the coverage probabilities given in Section 4.4.

Table 5.1 Approximate coverage probabilities of the folded Cauchy

	$n =$	12			24			48		
90% intended	$\theta =$	10	20	50	10	20	50	10	20	50
$\gamma = 0.1$	γ :	.847	.824	.817	.933	.925	.924	.957	.962	.942
	θ :	.665	.656	.671	.830	.801	.792	.851	.851	.855
$\gamma = 0.5$	γ :	.922	.913	.922	.942	.939	.939	.960	.963	.961
	θ :	.758	.754	.761	.822	.830	.822	.871	.859	.859
$\gamma = 1.0$	γ :	.931	.936	.932	.957	.957	.955	.969	.970	.970
	θ :	.791	.780	.795	.853	.844	.848	.876	.878	.871
$\gamma = 10.0$	γ :	.785	.788	.781	.845	.848	.849	.900	.907	.901
	θ :	.836	.840	.836	.870	.870	.871	.885	.877	.877
95% intended										
$\gamma = 0.1$	γ :	.875	.850	.842	.951	.943	.951	.976	.977	.962
	θ :	.705	.694	.718	.884	.848	.854	.896	.907	.907
$\gamma = 0.5$	γ :	.943	.941	.943	.962	.961	.960	.975	.980	.979
	θ :	.818	.805	.811	.876	.884	.879	.917	.915	.913
$\gamma = 1.0$	γ :	.953	.956	.956	.975	.975	.973	.985	.986	.987
	θ :	.853	.854	.848	.903	.902	.902	.928	.931	.926
$\gamma = 10.0$	γ :	.840	.844	.835	.897	.895	.897	.935	.941	.939
	θ :	.889	.894	.889	.921	.921	.920	.941	.934	.934

From Table 5.1, one can see that when the sample size increases, the approximate coverage probabilities for the parameters under the maximum likelihood method is getting closer to the intended coverage probabilities.

The Figure 5.4 shows the above four folded families with same mode (30) and same median (31). Equal mode is used to identify the peakendess of the four folded families, whereas the equal median is used to identify the upper tail variations of the four folded families.

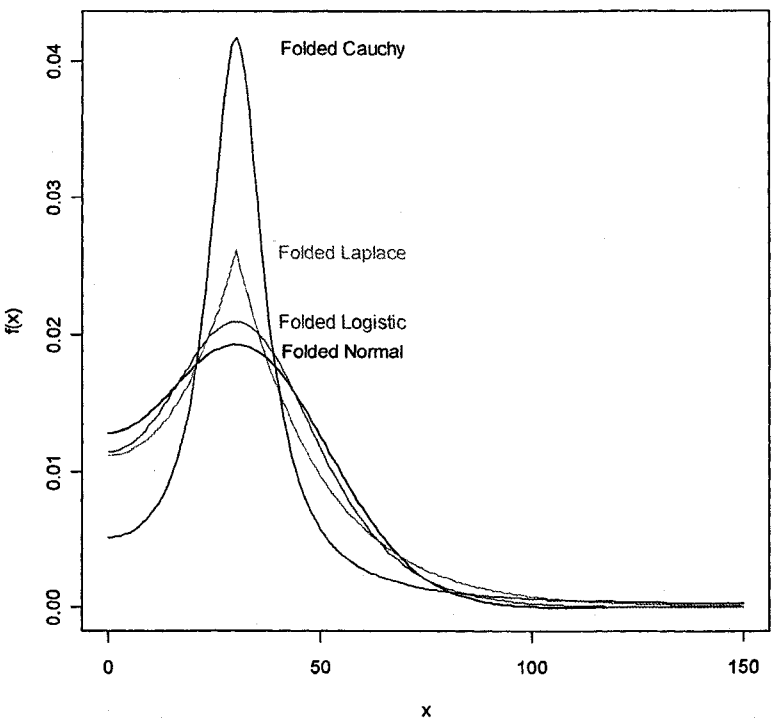

Figure 5.4 Four folded families with same mode (30) and same median (31).

5.6 Illustrative example

Example 1. This complete data set (see appendix B for the data set under the C8 column) concerns the urinary excretion rates (mg/24 hr) of the Tetra hydrocor-

ticosteron steroid metabolite for 86 patients with Cushing's syndrome (Aitchison *et al.* 2005). The normal range, based on the 37 normal healthy adults, of the urinary excretion rates of the Tetra hydrocorticosteron steroid metabolite is 0.02-.22. In this example, urinary excretion rates of 86 patients are analyzed using the four folded parametric distributions to examine how much they differ from the normal range. For this specific data set, the first three data points are 0,0,0. Also, this data set is distributed as a unimodal density and unimodal hazard shape. Therefore distributions like, gamma, loglogistic, lognormal, Weibull, inverse Gaussian, etc., cannot be use to model this data. A positively skewed non zero density would be a better choice to model these types of data.

Table 5.2 provides the estimated values of the four folded distributions for the urinary excretion rates data (cush data). Note that the chi-squared test has been performed by grouping the data into 11 classes with upper values of the non-open intervals, 0.04, 0.08, 0.12, 0.16, 0.2, 0.28, 0.36, 0.48, 0.68, 1.08.

Table 5.2 Estimated values of four folded distributions for cush data

Distribution	Parameter	$l(\boldsymbol{\theta})$	χ^2_8	p-value
Folded normal	$*\widehat{\mu} = 0.0000$, $\widehat{\sigma} = 1.5073$	-97.71	267.4	0.0000
Folded logistic	$*\widehat{\mu} = 0.0000$, $\widehat{\sigma} = 0.4644$	-59.21	97.18	0.0000
Folded Laplace	$*\widehat{\mu} = 0.0000$, $\widehat{\sigma} = 0.5867$	-40.15	53.31	0.0000
Folded Cauchy	$\widehat{\theta} = 0.1758$, $\widehat{\gamma} = 0.4980$	-12.37	11.53	0.1734

* Due to specific configuration of the data set, folded normal, folded logistic, and folded Laplace converge, respectively, to half normal, half logistic, and half Laplace

(exponential) distributions when estimating their parameters under the likelihood method.

The estimated values given in Table 5.2 indicate that the folded Cauchy distribution provides better fit to the urinary excretion rates data. Furthermore, the adequacy of the fit is further strengthened by illustrating the survival function of the folded Cauchy distribution (see solid line in Figure 5.5) along with the Kaplan-Meier curve. In order to compare, the fitted survival curves of the folded normal (dashed line), folded logistic (dotted line), and folded Laplace (dark dashed line) distributions are included in Figure 5.5.

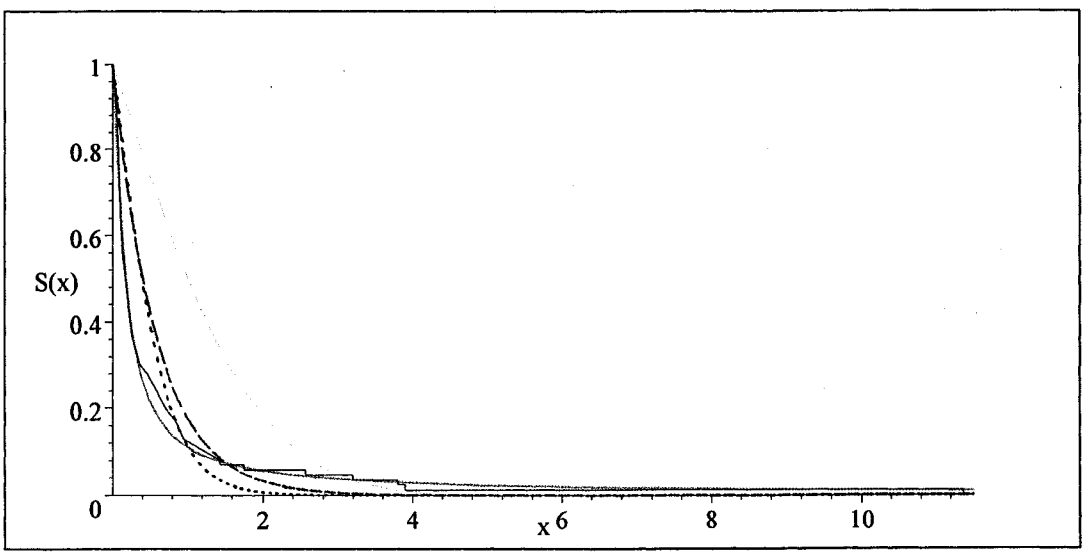

Figure 5.5 Fitted survival curves of the four folded family for cush data. Solid line, dark solid line, dark dashed line, dark dotted line and dashed line are, respectively, fitted Kaplan-Meier, folded Cauchy, folded Laplace, folded logistic, and folded normal curves.

CHAPTER VI

OVERVIEW, SUMMARY, AND FUTURE WORKS

6.1 Overview

This dissertation is about newly formed two- and three-parameter distributions. With the computational technology increases, these distributions are developed to benefit the modest modeling advantage. In this regard, for the most part of this dissertation, the model flexibility and applicability are thoroughly concerned with the aid of real data to establish the practical advantage of newly developed distributions. In addition, the model simplicities are simultaneously concerned in terms of the number of parameters involving in the new distributions. Also, regularity of the distributions and closed-form solutions for the density, the distribution, and the quantile functions are considered. Furthermore, some parts of this dissertation look at constructing the new distributions via plausible physical phenomenon. Moreover, a two way parameter estimation method is introduced with the construction of a new distribution. To advance further the applicability and flexibility of these distributions; complete, grouped, censored, and truncated data found in survival, reliability, and actuarial sciences are analyzed, illustrated, and compared with other leading distributions. Overall these newly developed distributions and their inherent properties can properly be distinguished from the other leading distributions. Hence, one can add these new distributions to the existing inventory of continuous univariate parametric

distributions.

6.2 Summary

In the first part of Chapter II, we review and suggest the remedy for the problem of handling data from two different models, the lognormal and the Pareto, by using one composite model, the composite lognormal-Pareto model. This new development, which has a promising approach for data modeling in the actuarial and the insurance industries, may be very useful for practitioners who have been handling lognormal and Pareto data separately for their research work. Actuaries who encounter smaller data values with higher frequencies, as well as occasional larger data values with lower frequencies are now exposed to a new avenue of this composite model which has a longer right tail than most of the non-monotonic positively skewed two-parameter density functions. The newly introduced composite lognormal-Pareto density is similar in shape to the lognormal density, and its upper tail is larger than the lognormal density. The new model can easily tackle the situation when the lognormal model underestimates the tail probability.

A two-parameter family of distribution, which is a natural composition of Weibull and Pareto family, is presented in the second part of Chapter II as an alternative to several well-known distributions such as lognormal, loglogistic, inverse Gaussian, etc., to model the unimodal failure rate data. Even though, a large number of unimodal failure rate life distributions are available, the Weibull-Pareto composite family is useful to the survival analyst due to its flexible left tail from Weibull model and thick and longer upper tail from Pareto model as well as closed-form survival and hazard

functions. Its flexibility, reliability, and applicability to survival data are demonstrated and emphasized using well-known examples. Specifically, the arm A head and neck cancer data given in example two, and nasopharynx cancer survival time data given in example three show a better fit to the Weibull-Pareto composite family than to the other parametric families. Furthermore, whether the data is complete or right censored, maximum likelihood parameter estimation techniques can be easily implemented for this model, and the related algorithms are quite simple. Finally, two more Pareto composite families are introduced and their flexibility is compared by analyzing a grouped data example found in actuarial sciences.

In Chapter III, a generalization of the Weibull distribution, the Odd Weibull family, is presented for modeling different types of failure rate data. Its applicability for modeling various failure rate data such as increasing, comfortable bathtub, unimodal, and etc., is demonstrated and emphasized using well-known examples by illustrating the scaled empirical and scaled fitted TTT plots. The upper percentage points of a test statistic, which measure the goodness-of-fit based on TTT plot, is tabulated for different parameter values of the Odd Weibull family. Furthermore, permissibility of testing the goodness-of-fit of the Weibull and inverse Weibull as submodels of the Odd Weibull family was also demonstrated in the examples. The inverse transformation of the Odd Weibull family is the same as the original distribution and is uncommon among the distributions having bathtub-shaped failure rates. Using this property and the maximum likelihood procedure, parameters of the Odd Weibull family is estimated in two different ways for complete, grouped, censored and truncated samples. This is actually useful to avoid some computational difficulties involved in

the likelihood function, especially when the densities of the Odd Weibull family are non-unimodal.

In the first part of Chapter IV, a two-parameter family of distributions, which is an alternative to such several well-known distributions as Gompertz, Weibull, exponential power-life-testing, etc., is presented to handle highly negatively skewed data with extreme observations. Its flexibility and applicability for lifetime data is demonstrated and emphasized using well-known examples. Especially, the parametric fit for the glass fiber strength data given in the first example, and the five-motor failure data given in the second example are improved by the logistic-sinh distribution. Furthermore, whether the data are complete or censored, maximum likelihood parameter estimation techniques can easily be implemented for this model, and the related computation procedures are quite simple. In addition, the logistic-sinh family is closed under proportional hazard modeling. It can be used to study the multi-population studies and is also amenable to simpler method of analyses and inferences. Because of this flexible nature of the logistic-sinh family, it can easily fit the five-motor failure data given in the second example with higher probability. Finally, an extension of the logistic-sinh family is introduced by reanalyzing the bus motor failure data.

In the second part of Chapter IV, we have presented two- and three-parameter families of distributions to model lifetime data. The two-parameter family can be used as an alternative to the well-known Gompertz distribution or other negatively skewed distribution, to model highly negatively skewed data that usually arise in life testing and survival analysis. Its flexibility, reliability, and applicability to lifetime data have been demonstrated and emphasized using well-known examples. Especially, when the

data are highly negatively skewed, the Gompertz-sinh distribution often provides a better model than the Gompertz distribution. In addition, the Gompertz-sinh family is more highly negatively skewed than the Gompertz family. The three-parameter exponentiated Gompertz-sinh family accommodates a wide variety of density shapes and non-decreasing hazard shapes, and can especially be useful for modeling human or animal survival time data due to its thick lower tail of the density function. This three-parameter model gives better-fit for both examples that we analyzed in the example section. Specifically, the parametric fit of the burial data given in the first example is improved by the exponentiated Gompertz-sinh distribution. Furthermore, whether the data are complete or censored, maximum likelihood parameter estimation techniques can easily be implemented for this model, and the related computation procedures are quite simple. Finally, the two models that we present here are worthwhile to survival and reliability analyst due to their flexibility and simplicity towards the data modeling specifically when the underlying distributions are negatively skewed.

Chapter V presented some folded parametric families: folded normal, folded logistic, folded Laplace, and folded Cauchy. The folded normal distribution has previously been studied. Therefore, we discussed some properties of the folded logistic, folded Laplace, and folded Cauchy distributions. The folded distributions are positively skewed and have non-zero density value at the origin. Therefore, these distributions are useful to analyze the data sets with zero data values.

Finally, following pointwise specific features give the importance of the distributions presented in subsequent chapters.

1. The composite Pareto family is useful for modeling highly positively skewed,

unimodal density shape, and thick upper tail data.

2. The logistic-sinh distribution is useful for modeling highly negatively skewed, unimodal non-zero density shape, and thick tails data.

3. The folded family is useful for modeling positively skewed unimodal non-zero density shape data.

4. The Odd Weibull family is useful since:

a. The Weibull and the inverse Weibull are submodels,

b. It has all five major hazard shapes,

c. It has longer useful lifetime when it is exhibiting a bathtub hazard shape,

d. The reciprocal transformation does not change the density function and hence the Odd Weibull parameters can be estimated two ways when analyzing exact, grouped, censored, and truncated data.

6.3 Future works

One of the major weaknesses of the composite distributions is; when estimating the parameters, the associated standard errors are not the required marginal standard errors, and they are conditional standard errors. Therefore, calculating the marginal standard errors of these composite models is an open problem.

Some disadvantages of the Odd Weibull distribution are difficult to conduct moment based statistical analysis, for example, calculating the generalized p-values. Also, the Odd Weibull distribution is computationally inconvenience to extend to analyze multivariate data. In addition, small sample analysis such as Bayesian technique is not performed using the Odd Weibull distribution. Furthermore, the study

of the log Odd Weibull family is an open question to interested readers.

Again, using the logistic-sinh or the Gompertz-sinh distrbution, small sample analysis such as Bayesian technique is not performed in this dissertation.

APPENDIX A

DERIVATIVES AND FORMULAS

1. Derivation of the Fisher information matrix for the lognormal-Pareto composite distribution

Let $X_1, X_2, ..., X_n$ be a random sample from the composite lognormal-Pareto model given in equation (2.1). Suppose the unknown parameter θ is in between the m^{th} observation and $m+1^{\text{th}}$ observation. Therefore, it is reasonable to assume that this is an ordered random sample, i.e., $x_1 \leq x_2 \leq x_3 \leq ....x_m \leq \theta \leq x_{m+1} \leq ... \leq x_n$. Then one can write the following equation by using the fact that the area under the density curve equals to 1,

$$\sum_{i=1}^{n} \int_{-\infty}^{\infty} f(x_i) dx_i = \int_{-\infty}^{\infty} \left(\sum_{i=1}^{n} f(x_i) \right) dx_i = n. \tag{A.1}$$

For the lognormal-Pareto composite density,

$$\sum_{i=1}^{n} \left(\int_{0}^{\theta} f_1(x_i) dx_i + \int_{\theta}^{\infty} f_2(x_i) dx_i \right) = n, \tag{A.2}$$

where $\ln f_1(x) \propto \ln \beta - 0.5 \left((\beta / k_1) \ln (x/\theta) + k_1 \right)^2$ and $\ln f_2(x) \propto \beta \ln \theta + \ln \beta - \beta \ln x$. Then

$$\sum_{i=1}^{m} \int_{0}^{\theta} f_1(x_i) dx_i + \sum_{i=m+1}^{n} \int_{\theta}^{\infty} f_2(x_i) dx_i = n. \tag{A.3}$$

$$\sum_{i=1}^{m} \int_{0}^{\theta} f_1(x_i) dx_i + \sum_{i=m+1}^{n} \int_{\theta}^{\infty} f_2(x_i) dx_i = n. \tag{A.4}$$

Now differentiating w.r.t. θ, we get

$$\sum_{i=1}^{m}\int_0^{\theta}\frac{\partial f_1(x_i)}{\partial\theta}dx_i + mf_1(\theta) + \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\frac{\partial f_2(x_i)}{\partial\theta}dx_i - (n-m)\,f_2(\theta) = 0.$$

$$\sum_{i=1}^{m}\int_0^{\theta}\frac{\frac{\partial f_1(x_i)}{\partial\theta}}{f_1(x_i)}f_1(x_i)dx_i + \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\frac{\frac{\partial f_2(x_i)}{\partial\theta}}{f_2(x_i)}f_2(x_i)dx_i + \frac{(2m-n)\,\beta}{(1+\Phi(k_1))\,\theta} = 0. \tag{A.5}$$

$$\sum_{i=1}^{m}\int_0^{\theta}\frac{\partial \ln f_1(x_i)}{\partial\theta}f_1(x_i)dx_i + \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\frac{\partial \ln f_2(x_i)}{\partial\theta}f_2(x_i)dx_i + \frac{(2m-n)\,\beta}{(1+\Phi(k_1))\,\theta} = 0. \tag{A.6}$$

(A.7)

Again differentiating w.r.t. θ, we get

$$\begin{aligned}
0 &= \sum_{i=1}^{m}\int_0^{\theta}\frac{\partial^2 \ln f_1(x_i)}{\partial\theta^2}f_1(x_i)dx_i + \sum_{i=1}^{m}\int_0^{\theta}\left[\frac{\partial \ln f_1(x_i)}{\partial\theta}\right]^2 f_1(x_i)dx_i \\
&\quad + \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\frac{\partial^2 \ln f_2(x_i)}{\partial\theta^2}f_2(x_i)dx_i + \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\left[\frac{\partial \ln f_2(x_i)}{\partial\theta}\right]^2 f_2(x_i)dx_i \\
&\quad + \frac{m\beta}{\theta}f_1(\theta) - (n-m)\,\frac{\beta}{\theta}f_2(\theta) - \frac{(2m-n)\,\beta}{(1+\Phi(k_1))\,\theta^2}.
\end{aligned} \tag{A.8}$$

$$\begin{aligned}
I[\theta\theta] &= \sum_{i=1}^{m}\int_0^{\theta}\left[\frac{\partial \ln f_1(x_i)}{\partial\theta}\right]^2 f_1(x_i)dx_i + \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\left[\frac{\partial \ln f_2(x_i)}{\partial\theta}\right]^2 f_2(x_i)dx_i \\
&= -\sum_{i=1}^{m}\int_0^{\theta}\frac{\partial^2 \ln f_1(x_i)}{\partial\theta^2}f_1(x_i)dx_i - \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\frac{\partial^2 \ln f_2(x_i)}{\partial\theta^2}f_2(x_i)dx_i \\
&\quad + \frac{\{2m-n+(n-m)\,\beta-m\beta\}\,\beta}{(1+\Phi(k_1))\,\theta^2} \\
&= \left\{\frac{n-2m+\frac{m\Phi(k_1)}{k_1^2}}{1+\Phi(k_1)}\right\}(\beta/\theta)^2.
\end{aligned} \tag{A.9}$$

Similarly,

$$\begin{aligned}
I[\theta\beta] &= \sum_{i=1}^{m}\int_0^{\theta}\left[\frac{\partial \ln f_1(x_i)}{\partial\theta}\right]\left[\frac{\partial \ln f_1(x_i)}{\partial\beta}\right] f_1(x_i)dx_i \\
&\quad + \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\left[\frac{\partial \ln f_2(x_i)}{\partial\theta}\right]\left[\frac{\partial \ln f_2(x_i)}{\partial\beta}\right] f_2(x_i)dx_i \\
&= -\sum_{i=1}^{m}\int_0^{\theta}\frac{\partial^2 \ln f_1(x_i)}{\partial\beta\partial\theta}f_1(x_i)dx_i - \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\frac{\partial^2 \ln f_2(x_i)}{\partial\beta\partial\theta}f_2(x_i)dx_i \\
&\quad + \frac{n-2m}{(1+\Phi(k_1))\,\theta} \\
&= m/\theta.
\end{aligned} \tag{A.10}$$

$$
\begin{aligned}
I[\beta\beta] &= \sum_{i=1}^{m}\int_0^{\theta}\left[\frac{\partial \ln f_1(x_i)}{\partial\beta}\right]^2 f_1(x_i)dx_i + \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\left[\frac{\partial \ln f_2(x_i)}{\partial\beta}\right]^2 f_2(x_i)dx_i \\
&= -\sum_{i=1}^{m}\int_0^{\theta}\frac{\partial^2 \ln f_1(x_i)}{\partial\beta^2} f_1(x_i)dx_i - \sum_{i=m+1}^{n}\int_{\theta}^{\infty}\frac{\partial^2 \ln f_2(x_i)}{\partial\beta^2} f_2(x_i)dx_i \\
&= \frac{n-m+mk_1^2+mk_1^2\Phi(k_1)+2m\Phi(k_1)}{(1+\Phi(k_1))\beta^2}. \qquad \text{(A.11)}
\end{aligned}
$$

Now let $p = \frac{\Phi(k_1)}{k_1} - 2$, $q = k_1^2(1+\Phi(k_1)) + 2\Phi(k_1) - 1$, then one can write the Fisher information matrix as

$$
I[\theta,\beta] = \begin{bmatrix} \left(\frac{n+mp}{1+\Phi(k_1)}\right)\left(\frac{\beta}{\theta}\right)^2 & \frac{m}{\theta} \\ \frac{m}{\theta} & \left(\frac{n+mq}{1+\Phi(k_1)}\right)\left(\frac{1}{\beta}\right)^2 \end{bmatrix}. \qquad \text{(A.12)}
$$

2. The Quantile and related functions

The Quantile ($Q(u)$), Galton's skewness (G), and Moor's kurtosis (K) functions for various continuous univariate distributions are given below.

(a). Symmetrical Unimodal distributions ($G = 0$)

(i). Uniform (U)

$Q(u) = a + (b-a)u$; $0 \le a < b < \infty$, $0 \le u \le 1$; $K = 1$.

(ii). Normal (N)

$Q(u) = \mu + \sigma\Phi^{-1}(u)$; $-\infty < \mu < \infty, 0 < \sigma < \infty$, $0 \le u \le 1$; $K = 1.233095115$.

(iii). Logistic (L)

$Q(u) = \mu + \sigma \ln(\frac{u}{1-u})$; $-\infty < \mu < \infty, 0 < \sigma < \infty$, $0 \le u \le 1$; $K = 1.306270228$.

(iv). Laplace (LA)

$$Q(u) = \begin{cases} \mu + \sigma \ln(2u) & \text{if} \quad u \leq 1/2 \\ \mu - \sigma \ln(2(1-u)) & \text{if} \quad u \geq 1/2 \end{cases} ; -\infty < \mu < \infty,\ 0 < \sigma < \infty,\ 0 \leq u \leq 1;\ K = 1.584962501.$$

(v). Cauchy (C)

$Q(u) = \mu + \sigma \tan[\pi(u - 0.5)]; -\infty < \mu < \infty, 0 < \sigma < \infty,\ 0 \leq u \leq 1;\ K = 2.$

(b). Symmetrical Bimodal distributions ($G = 0$)

(i). Sinh-normal (SN)

$Q(u) = \theta \sinh^{-1}[\frac{1}{\alpha}\Phi^{-1}(u)]; 0 < \alpha < \infty, 0 < \theta < \infty,\ 0 \leq u \leq 1;\ K\epsilon[0, 1.233095115].$

(ii). Sinh-logistic (SL)

$Q(u) = \theta \sinh^{-1}[\frac{1}{\alpha} \ln(\frac{u}{1-u})]; 0 < \alpha < \infty, 0 < \theta < \infty,\ 0 \leq u \leq 1;\ K\epsilon[0, 1.306270228].$

(iii). Sinh-Cauchy (SC)

$Q(u) = \theta \sinh^{-1}[\frac{1}{\alpha} \tan\{\pi(u - 0.5)\}]; 0 < \alpha < \infty, 0 < \theta < \infty,\ 0 \leq u \leq 1;\ K\epsilon[0, 2].$

(c). Extreme value distributions

(i). Smallest extreme value (SEV)

$Q(u) = \mu + \sigma \ln \ln(\frac{1}{1-u}); -\infty < \mu < \infty, 0 < \sigma < \infty,\ 0 \leq u \leq 1;\ G = -0.118432588,\ K = 1.278103155.$

(ii). Largest extreme value (Gumbel) (LEV)

$Q(u) = \mu - \sigma \ln \ln(\frac{1}{u}); -\infty < \mu < \infty, 0 < \sigma < \infty,\ 0 \leq u \leq 1;\ G = 0.118432588,\ K = 1.278103155.$

(d). Half distributions

(i). Half-normal (HN)

$Q(u) = \mu + \sigma\Phi^{-1}(\frac{1+u}{2}); -\infty < \mu < \infty, 0 < \sigma < \infty,\ 0 \leq u \leq 1;\ G = 0.144292171,\ K = 1.176419296.$

(ii). Half-logistic (HL)

$Q(u) = \mu + \sigma \ln(\frac{1+u}{1-u}); -\infty < \mu < \infty, 0 < \sigma < \infty,\ 0 \leq u \leq 1;\ G = 0.180833387,$

$K = 1.239547938.$

(iii). Half Laplace or Exponential (EXP)

$Q(u) = \mu - \sigma \ln(1-u); -\infty < \mu < \infty, 0 < \sigma < \infty,\ 0 \leq u \leq 1;\ G = 0.261859507,$

$K = 1.306270228.$

(iv). Half-Cauchy (HC)

$Q(u) = \mu + \sigma \tan[\pi u/2]; -\infty < \mu < \infty, 0 < \sigma < \infty,\ 0 \leq u \leq 1;\ G = 0.414213562,$

$K = 2.$

(e). Folded distributions

(i). Folded-logistic (FL)

$Q(u) = \sigma \ln[\frac{(1+k^2)u+\{4k^2+(1-k^2)^2u^2\}^{1/2}}{2k(1-u)}]; k = e^{\mu/\sigma}, -\infty < \mu < \infty, 0 < \sigma < \infty,$

$0 \leq u \leq 1$; There is no short form formula for G, and K.

(ii). Folded Laplace (FLA)

$$Q(u) = \begin{cases} \sigma \sinh^{-1}(ku) & \text{if} \quad u \leq (1-1/k^2)/2 \\ \sigma \ln\{\frac{1+k^2}{2k(1-u)}\} & \text{if} \quad u \geq (1-1/k^2)/2 \end{cases}; k = e^{\mu/\sigma},\ 0 < \mu < \infty,\ 0 < \sigma <$$

$\infty,\ 0 \leq u \leq 1$; There is no short form formula for G, and K.

(iii). Folded-Cauchy (FC)

$Q(u) = \frac{\theta}{2\alpha}[\tan\{\pi(u-0.5)\} + \sqrt{\tan^2\{\pi(u-0.5)\} + 4\alpha^2}]; 0 < \alpha < \infty, 0 < \theta < \infty,$

$0 \leq u \leq 1;\ G = \sqrt{1+4\alpha^2} - 2\alpha,\ K = 2.$

(f). Logarithmic transformed distributions

(i). Lognormal (LN)

$Q(u) = e^{\mu+\sigma\Phi^{-1}(u)}; -\infty < \mu < \infty, 0 < \sigma < \infty, 0 \leq u \leq 1; G = \tanh\{0.5\sigma\Phi^{-1}(3/4)\},$

$K = [\sinh\{\sigma\Phi^{-1}(7/8)\} - \sinh\{\sigma\Phi^{-1}(5/8)\}]/\sinh\{\sigma\Phi^{-1}(3/4)\}$.

(ii). Loglogistic (LL)

$Q(u) = \theta(\frac{u}{1-u})^{1/\alpha}; 0 < \alpha < \infty, 0 < \theta < \infty,\ 0 \leq u \leq 1;\ G = \tanh\{\ln(3)/(2\alpha)\}$, $K = [\sinh\{\ln(7)/\alpha\} - \sinh\{\ln(5/3)/\alpha\}]/\sinh\{\ln(3)/\alpha\}$.

(iv). LogCauchy (LC)

$Q(u) = \theta e^{\frac{1}{\alpha}\tan\{\pi(u-0.5)\}}; 0 < \alpha < \infty, 0 < \theta < \infty,\ 0 \leq u \leq 1;\ G = \tanh(0.5/\alpha)$, $K = 2\cosh(\sqrt{2}/\alpha)$.

(g). Power distribution (PO)

$Q(u) = bu^{1/c};\ 0 < b < \infty,\ 0 < c < \infty,\ 0 \leq u \leq 1,\ G = (3^{1/c} - 2^{1+1/c} + 1)/(3^{1/c} - 1)$, $K = (7^{1/c} - 5^{1/c} + 3^{1/c} - 1)/(6^{1/c} - 2^{1/c})$.

(h). Pareto distribution (PA)

$Q(u) = b(1-u)^{-1/c};\ 0 < b < \infty,\ 0 < c < \infty,\ 0 \leq u \leq 1,\ G = (3^{-1/c} - 2^{1-1/c} + 1)/(1 - 3^{-1/c})$, $K = (1 - 3^{-1/c} + 5^{-1/c} - 7^{-1/c})/(2^{-1/c} - 6^{-1/c})$.

(i). Weibull Distribution (W)

$Q(u) = \theta \ln^{1/\alpha}\left(\frac{1}{1-u}\right); 0 < \alpha < \infty, 0 < \theta < \infty,\ 0 \leq u \leq 1$; There is no short form formula for G, and K.

(j). Gamma distribution (GA)

There is no closed-form quantile function.

(k). Gompertz distribution (G)

$Q(u) = \theta \ln\left(1 - \alpha^{-1}\ln(1-u)\right);\ 0 < u < 1, 0 < \alpha < \infty, 0 < \theta < \infty$; There is no short form formula for G, and K.

(l). Logistic-sinh distribution (LS)

$Q(u) = \theta \ln\left(1 + \operatorname{arc\,sinh}\left(\frac{u}{\lambda(1-u)}\right)\right)$; $0 < u < 1, 0 < \lambda < \infty, 0 < \theta < \infty$; There is no short form formula for G, and K.

(m). Gompertz-sinh distribution (GS)

$Q(u) = \theta \ln\left(1 - \operatorname{arc\,sinh}\left(\mu^{-1} \ln(1-u)\right)\right)$; $0 < u < 1, 0 < \mu < \infty, 0 < \theta < \infty$; There is no short form formula for G, and K.

(n). Birnbaum-Saunders distribution (BS)

$Q(u) = \left(\frac{\sigma\Phi^{-1}(u) + \sqrt{4\mu + (\sigma\Phi^{-1}(u))^2}}{2}\right)^2$; $-\infty < \mu < \infty, 0 < \sigma < \infty$, $0 \leq u \leq 1$; There is no short form formula for G, and K.

(o). Inverse exponential distribution (IEXP)

$Q(u) = \theta \ln^{-1}\left(\frac{1}{u}\right)$; $0 < \theta < \infty, 0 \leq u \leq 1$; $G = 0.476280986$, $K = 2.141741023$.

(p). Inverse Weibull distribution (IW)

$Q(u) = \theta \ln^{-1/\alpha}\left(\frac{1}{u}\right)$; $0 < \alpha < \infty, 0 < \theta < \infty$, $0 \leq u \leq 1$; There is no short form formula for G, and K.

(q). Lognormal-Pareto composite distribution (LPC)

$$Q(u) = \begin{cases} \theta \exp\left\{(k_1/\beta)\left(\Phi^{-1}\left((1 + \Phi(k_1))u\right) - k_1\right)\right\} & \text{if} \quad 0 \leq u \leq u_0 \\ \theta\left\{(1-u)(1 + \Phi(k_1))\right\}^{-1/\beta} & \text{if} \quad u_0 \leq u < 1. \end{cases}$$

Where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution, $u_0 = \Phi(k_1) / (1 + \Phi(k_1))$, and $k_1 = 0.372238898$. Also θ (> 0), and β (> 0) are, respectively, scale and shape parameters of this distribution.

There is no short form formula for G, and K.

(r). Weibull-Pareto composite distribution (WPC)

$$Q(u) = \begin{cases} \theta\left[\left(\frac{-k}{k+1}\right)\ln\left\{1 - \left(\frac{2k+1}{k+1}\right)u\right\}\right]^{\frac{1}{\gamma k}} & \text{if} \quad 0 \leq u \leq k/(2k+1) \\ \theta\left\{(1-u)\left(\frac{2k+1}{k+1}\right)\right\}^{-1/\gamma} & \text{if} \quad k/(2k+1) \leq u < 1 \end{cases}$$

Where $k = 2.857334826$. Also θ (> 0), and γ (> 0) are, respectively, scale and shape parameters of this distribution.

(s). Weibull-inverse Weibull composite distribution (WIW)

$$Q(u) = \begin{cases} \theta \ln^{1/\alpha} \left[1/\left\{1 - 2\left(1 - e^{-1}\right) u\right\}\right] & \text{if} \quad 0 \leq u \leq 1/2 \\ \theta \ln^{-1/\alpha} \left[1/\left\{2\left(1 - e^{-1}\right) u + 2e^{-1} - 1\right\}\right] & \text{if} \quad 1/2 \leq u < 1 \end{cases}$$

Where θ (> 0), and α (> 0) are, respectively, scale and shape parameters of this distribution.

APPENDIX B

DATA AND CODES

Appendix B provides data sets that are related to the reliability, medical, environmental, and actuarial sciences. We briefly discuss their failure rates and basic shape of the data distribution where necessary although some of the data sets have already been analyzed in previous chapters. For the most part, the original data source are provided for reader interest. Some of these data sets have also been analyzed by other researchers and their source of such analyses are provided as well.

The data in the appendix B appear in seven different format:

1. Complete data.
2. Grouped data.
3. Right censored data.
4. Interval censored data.
5. Right truncated data.
6. Left truncated and right censored data.
7. Left truncated and interval censored data.

For each of these format BMDP (1992) codes are provided by using the Odd Weibull model. One can replace this Odd Weibull model with different parametric models to enable these codes where necessary. The data are given in dat.xls file in the CD-ROM. Also some useful R-codes are given at the end.

1. Complete data

1.1 Juran and Gryna's electronic ground support equipment failure data

This 105 data points represents the time in hours to failure for a unit of electronic ground support equipment, which was taken from Juran and Gryna (1970), and also can found in Kolb and Ross (1980). Later Elsayed (1996) used these data to determine the nonparametric renewal function to estimate the expected number of failures under the discrete time approach. The hazard shape of this data represents the bathtub-shaped failure rate.

1.2 Aarset's device failure data

The data represents the times to failure of 50 devices put on a life test at time zero (Aarset 1987). Later, Mudholkar *et al.* (1996) used this data set to illustrate the flexibility of generalized Weibull family. The hazard shape of this data represents the bathtub-shaped failure rate. Reanalysis of Aarset's data is given in Chapter III Section 3.7.

1.3 Single exposure gamma irradiated mice mortality data

The 208 data points represent the ages at death in weeks for male mice exposed to 240r of gamma radiation (Fürth *et al.* 1959; Kimball 1960). This data is also available in Elandt-Johnson and Johnson (1980), and Lawless (2003). The hazard shape of this data represents the increasing failure rate. Reanalysis of this data set is given in Chapter III Section 3.7.

1.4 Multiple exposure gamma irradiated mice mortality data

This 47 survival times, in units of 2 months, of continuous whole-body gamma-radiation at an intensity of 2.2 standard units of radiation daily. The data are abstracted from Sampford (1952) and initially studied by Lorentz *et al.* (1947). The hazard shape of these data represent the increasing failure rate.

1.5 Glass fiber strength data

The glass fiber data are experimental strength values of two lengths, 1.5cm (63 data points), and 15cm (46 data points), from the National Physical Laboratory in England (Smith and Naylor 1987). The authors used three-parameter Weibull distribution to model the two data sets. The hazard shape of this data represents the increasing failure rate. Our analysis of these data sets are given in Chapter IV Section 4.7.

1.6 Guinea pigs survival time data

This data set (study M, regimen 5.5), which is abstracted from Bjerkedal (1960) represents the survival times of guinea pigs after infected with virulent tubercle bacilli. The hazard shape of this data represents the unimodal-shaped failure rate. Our analysis for this data set is given in Chapter II Section 2.11.7.

1.7 Stimulus-response time data

This data represents the reaction time of one subject in 180 trials of a psychological experiment (Whitmore 1986). In each trial the subject was asked to decide whether the distance between two dots displayed on a monitor placed 10ft away was long or short. The dots remained visible until the subject made a response. The reaction

time for each trial is the length of time from stimulus to response in milliseconds. The hazard shape of this data represents the unimodal-shaped failure rate. Our analysis for this data set is given in Chapter II Section 2.11.7.

1.8 Diamond data

This twin data set consists of alluvial diamonds from the Bougban and Damaya deposits in Guinea of West Africa (Beirlant *et al.* 1996). The deposits have been well explored by a systematic 100 × 50 meter sampling grid of unit samples of 8.85 square meters. The sampling program on Bougban recovered 683 stones, whereas the Damaya sampling yielded 444 stones. The data represent the unimodal-shaped failure rates. The analysis of these data using the odd Weibull family are given in Chapter III Section 3.7.

1.9 Urinary excretion rates data

This complete data set concerns the urinary excretion rates of the tetra hydro-corticosteron steroid metabolite for 86 patients with Cushing's syndrome (Aitchison *et al.* 2005). The hazard shape of this data represents the unimodal-shaped failure rate. The urinary excretion rates data are analyzed by using the folded distributions given in Chapter V Section 5.6.

1.10 Danish fire insurance data

This complete Danish data set (McNeil 1997, Resnick 1997) consist of 2492 fire insurance losses in Danish Krone (DKK) from the years 1980 to 1990 inclusive. The loss figure is a total loss figure for the events concerned and includes damage to buildings, furniture and personal property as well as loss of profits. The recorded data

have been suitably adjusted to reflect 1985 values. The adjusted loss values in Danish Krone range from (in millions) 0.3134041 to 263.2503660. McNeil (1997) analyzed the upper portion of this data, which consist of 2156 losses over one million Danish Krone, as an example to the use of extreme value theory by estimating the tails of loss severity distributions. For the upper portion of the data, he used two-parameter shifted Pareto model as a parametric model and concluded that the two-parameter shifted Pareto model is a useful model for estimating the tails of loss severity distributions. Resnick (1997) analyzed the full Danish data set to demonstrate several alternative statistical techniques and plotting devices that can be used for assessing the appropriateness of heavy tailed models, and justified McNeil's decision to drop losses below one million DKK to use the Pareto model. Embrechts *et al.* (1999) used the data set (upper portion) in their book to discuss the Pareto model as a useful loss severity distribution. Our analysis for this data set is given in Chapter II Section 2.10 and Section 2.14.1.

1.11 Badenscallie burial data

The following data set is the age of death of male members of Scottish McAlpha clan in the burial ground at Badenscallie in the Coigach district of Wester Ross, Scotland (Sprent and Smeeton 2000). Ages are given for complete years, e.g. 0 means before first birthday and 79 means on or after 79th but before 80th birthday, according to the information on the tombstone. The data were collected in June 1987. The authors pointed out that the McAlpha clan data set possesses reasonably approximate pattern of death ages for the all four clan. The hazard shape of this data represents the bathtub-shaped failure rate. Furthermore, the analysis of the McAlpha

clan data set is given in Chapter IV Section 4.8.6.

1.12 Wave and surge height data

This example is a concurrent measurements of two oceanographic variables - wave and surge height at a single location off south-west England (Coles 2001). This is a large data set with 2894 data points for each variables - wave and surge heights measured in meters. As noted by Coles (2001), the scatter plot of wave and surge data suggests a tendency for extremes of one variable to coincide with extremes of the other. Our analysis for this twin data set is given in Chapter III Section 3.10.1.

1.13 BMDP code using the Odd Weibull model for Juran and Gryna's electronic ground support equipment failure data

```
/INPUT VARIABLES=1.
FORMAT=FREE.
/VARIABLE NAMES=time.
/ESTIMATE PARAMETERS=3.
/PARAMETER NAMES=a,b,c.
INITIAL=5,.1, 80.
/DENSITY F=(a*b/time)*((time/c)**a)*EXP((time/c)**a)
*((EXP((time/c)**a)-1)**(b-1))
/((1+(EXP((time/c)**a)-1)**b)**2).
/END
/END
```

Note: The point data column should be insert between the last two /END command.

2. Grouped data

2.1 Bus motor failure data

The classical five bus motor failure data are firstly considered and analyzed by Davis (1952). The results take into account the time to the first and succeeding major motor failures for 191 buses operated by a large city bus company, with time being the number of thousand miles driven. Failure was either abrupt, in which some part broke or the motor would not run. Failures of motor accessories, which could be easily replaced, were not included in these data.

Davis used the truncated normal distribution to analyze the first two motor failure data and the exponential distribution for the second and succeeding failures. In the analysis, in terms of chi-squared goodness-of-fit, he found that both models are poorly fit to the second bus motor failure data. Bain (1974) adapted a three-parameter quadratic hazard model for the purpose of obtaining a good fit to the second bus motor failure data. Later, Mudholkar *et al.* (1995) used three-parameter exponentiated Weibull model to analyze the five motor failure data. Lindsey (1997) gave an alternative analysis to the bus motor failure data using parametric multiplicative intensity models. However, he considers data that are grouped more coarsely than the data given by Davis (1952). Our analysis for this data set is given in Chapter IV Section 4.7 and 4.8.

2.2 Hospital stay length data

This example represents hospital-stay frequency distribution for 2311 schizophrenic patients taken from the Maryland Psychiatric Case Register. This data set was earlier analyzed by Eaton & Whitmore (1977) to discuss the appropriateness of the inverse Gaussian distribution as a model for the hospital stay pattern. Later, Whitmore (1986) noted that any simple model is inappropriate to explain the hospital stay pattern. Therefore, he formulated the normal-gamma mixture model to provide a clear improvement in fit relative to the unmixed inverse Gaussian model. The hazard shape of this data represents the unimodal-shaped failure rate. Our analysis for this data set is given in Chapter III Section 3.7.

2.3 BMDP code using the Odd Weibull model for the second bus motor failure data

```
/INPUT VARIABLES=1.
FORMAT=FREE.
/VARIABLE NAME=count.
/ESTIMATE PARAMETER=3.
/PARAMETER NAME=a,b,c.
INITIAL=4,0.25,75.
/DENSITY U1=(1+(EXP((20/c)**a)-1)**b)**(-1).
U2=(1+(EXP((40/c)**a)-1)**b)**(-1).
U3=(1+(EXP((60/c)**a)-1)**b)**(-1).
U4=(1+(EXP((80/c)**a)-1)**b)**(-1).
```

U5=(1+(EXP((100/c)**a)-1)**b)**(-1).

U6=(1+(EXP((120/c)**a)-1)**b)**(-1).

IF (KASE EQ 1) THEN LNF=count*LN(1-U1).

IF (KASE EQ 2) THEN LNF=count*LN(U1-U2).

IF (KASE EQ 3) THEN LNF=count*LN(U2-U3).

IF (KASE EQ 4) THEN LNF=count*LN(U3-U4).

IF (KASE EQ 5) THEN LNF=count*LN(U4-U5).

IF (KASE EQ 6) THEN LNF=count*LN(U5-U6).

IF (KASE EQ 7) THEN LNF=count*LN(U6).

/END

/END

Note: The point data column should be inserted between the last two /END command.

3. Right censored data

3.1 Head and neck cancer data

The following data represents the survival times in days of head and neck cancer patients after two different treatments considered earlier by Efron (1988) from a two-arm clinical trial. This clinical trial data consists of 51 patients with radiation therapy alone denoted by arm A and 45 patients with radiation plus chemotherapy denoted by arm B. Nine and fourteen patients were lost to follow-up respectively in arm A and arm B and were regarded as right censored. Mudholkar *et al.* (1995) used arm A clinical trial data to demonstrate the flexibility of exponentiated Weibull distribution

to unimodal-shaped failure rate data. Meanwhile, the generalized Weibull model (Mudholkar *et al.* 1996) gives considerably improved fit for the two-arm clinical trial data. Our analysis for the arm A data set is given in Chapter II Section 2.11.7.

3.2 Nasopharynx cancer survival data

The data set of this example is taken from McKeague (2000) and given by West (1987, 1992) who studied the data on 181 nasopharynx cancer patients. Their cancer careers, culminating in either death (127 cases) or censoring (54 cases), are recorded to the nearest month, ranging from 1 to 177 months. Our analysis for this data set is given in Chapter II Section 2.11.7.

3.3 Oklahoma diabetic data

The survival times (in years) represent the *first 40 male patients* enrolled in a mortality study of Oklahoma diabetic Indians (Lee and Wang 2003). This example is a part of larger sample of 1012 Oklahoma Indians with non-insulin-dependent diabetes mellitus (NIDDM)), and the data were examined in 1972-1980. The hazard shape of this data represents the increasing failure rate. Analysis of this data set is given in Chapter IV Section 4.7.

3.4 Diabetic data

This example represents the survival times (in years) of 149 diabetic patients who were followed for 17 years (Lee and Wang 2003). The hazard shape of this data represents the increasing failure rate. Analysis of this data set is given in Chapter IV Section 4.8.6.

3.5 BMDP code using the Odd Weibull model for the arm A head and neck cancer data

```
/INPUT VARIABLES=2.
FORMAT=FREE.
/VARIABLE NAMES=time,cen.
/ESTIMATE PARAMETERS=3.
/PARAMETER NAMES=a,b,c.
INITIAL=-0.25,-4.0,50.
/DENSITY EX1=(a*b/time)*((time/c)**a)*EXP((time/c)**a)
*((EXP((time/c)**a)-1)**(b-1))/((1+(EXP((time/c)**a)-1)**b)**2).
EX2=(1+(EXP((time/c)**a)-1)**b)**(-1).
IF(cen==1) THEN F=EX1.
IF(cen==0) THEN F=EX2.
/END
/END
```

Note: The point data values in the first column and the censoring indicator in the second column (1 = dead, 0 = censored) should be inserted between the last two /END command.

4. Interval censored data

4.1 Breast cancer data

Beadle *et al.* (1984a and b) report a retrospective study carried out to compare the cosmetic effects of radiotherapy alone versus radiotherapy and adjuvant chemotherapy

on women with early breast cancer. This twin data set is discussed by Klein and Moeschberger (1997), Finkelstein and Wolfe (1985), and Ryan and Lindsey (1998).

4.2 Drug resistance of AIDS patients data

This interval censored data set is taken from Ryan and Lindsey (1998) and is originally analyzed by Richman *et al.* (1990) regarding the drug resistance (time in months to resistance to Zidovudine) of 31 AIDS patients. Our analysis for this data set is given in Chapter III Section 3.7.

4.3 BMDP code using the Odd Weibull model for the radiotherapy and adjuvant chemotherapy data

```
/INPUT VARIABLES=2.
FORMAT=FREE.
/VARIABLE NAMES=left,right.
/ESTIMATE PARAMETERS=3.
/PARAMETER NAMES=a,b,c.
INITIAL=1,1,50.
/DENSITY U=1/(1+(EXP((left/c)**a)-1)**b).
V=1/(1+(EXP((right/c)**a)-1)**b).
IF (right NE 61) THEN LNF=LN(U-V).
IF (right EQ 61) THEN LNF=LN(U).
/END
/END
```

Note: The right and left limit point data values are in the two columns which

should be inserted between the last two /END command. 61 mean right open interval.

5. Right truncated data

5.1 AIDS blood-transfusion data 1

This example is taken from Wang (1989) and is initially analyzed by Kalbfleisch and Lawless (1989). In acquired immune deficiency syndrome (AIDS) studies survival time is usually defined as the time from the human immunodeficiency virus (HIV) infection to the diagnosis of AIDS. Only individuals who have developed AIDS prior to the end of the study period are included in the study. Infected individuals who have yet to develop AIDS are not included and hence the data set is right truncated. The data are presented through two variables: time in months from the transfusion to the diagnosis of AIDS, truncation in months from transfusion to the end of the study period (July 1989). Also data are categorized into three age groups: "children" aged 1-4, "adults" aged 5-59, and "elderly patients" aged 60 and older.

5.2 AIDS blood-transfusion data 2

This example is initially analyzed by Lagakos *et al.* (1988) and is also available in Klein and Moeschberger (1997). As in the previous example, important measurement is the induction period between infection with the AIDS virus and the onset of clinical AIDS. This time is sometimes referred to as the latency period or incubation period. The data are presented through two variables: time in months from the transfusion to the diagnosis of AIDS, truncation in months from transfusion to the end of the study period (June 30, 1986). Also data are categorized into two age groups: children and adults. There are 37 children and 258 adults included in this study.

5.3 BMDP code using the Odd Weibull model for AIDS transfusion data 1 for children aged 1-4

```
/INPUT VARIABLES=2.
FORMAT=FREE.
/VARIABLE NAMES=time, trunc.
/ESTIMATE PARAMETERS=3.
/PARAMETER NAMES=a,b,c.
INITIAL=3.5,0.5,25.
/DENSITY U=(1+(EXP((trunc/c)**a)-1)**(-b)).
F=U*(a*b/time)*((time/c)**a)*EXP((time/c)**a)
*((EXP((time/c)**a)-1)**(b-1))/((1+(EXP((time/c)**a)-1)**b)**2).
/END
/END
```

Note: The time data values in the first column and the right truncated values in the second column should be inserted between the last two /END command.

6. Left truncated and right censored data

6.1 Channing House data

Channing House is a retirement center located in Palo Alto, California. Data consist of ages at death of 462 individuals (97 males and 365 females), who were in residence during the period of January 1964 to July 1975. This data has been reported by Hyde (1980) and also available in Klein and Moeschberger (1997). Data reports the age in months when members of the community died or left the center

and the ages when individuals entered the community. Also individuals must survive to a sufficient age to enter the retirement community. Therefore, the lifetimes of this data set are left truncated and right censored.

6.2 BMDP code using the Odd Weibull model for Channing House data for 97 males

```
/INPUT VARIABLES=3.
FORMAT=FREE.
/VARIABLE NAMES=trunc,time,cen.
/ESTIMATE PARAMETERS=3.
/PARAMETER NAMES=a,b,c.
INITIAL=6,1,1000.
/DENSITY U=(1+(EXP((trunc/c)**a)-1)**b).
EX1=(a*b/time)*((time/c)**a)*EXP((time/c)**a)
*((EXP((time/c)**a)-1)**(b-1))/((1+(EXP((time/c)**a)-1)**b)**2).
EX2=(1+(EXP((time/c)**a)-1)**b)**(-1).
IF(cen==1) THEN F=U*EX1.
IF(cen==0) THEN F=U*EX2.
/END
/END
```

Note: The truncated data values in the first column, the death time in the second column, and the right censoring indicator in the third column (1 = dead, 0 = censored) should be inserted between the last two /END command.

7. Left truncated and interval censored data

7.1 Functional independence data

This example is a left truncated and interval censored increasing failure rate twin data set (Pan and Chappell 1998, 2002), regarding the loss of functional independence of people of age 65 years or older. This twin data set consists of 421 non-poor male group and 609 non-poor female group. Our analysis for this data set is given in Chapter III Section 3.7.

7.2 BMDP code using the Odd Weibull model for functional independence data for non-poor female

```
/INPUT VARIABLES=3.
FORMAT=FREE.
/VARIABLE NAMES=truc,left,right.
/ESTIMATE PARAMETERS=3.
/PARAMETER NAMES=a,b,c.
INITIAL=8,1,100.
/DENSITY U=(1+(EXP((truc/c)**a)-1)**b)/(1+(EXP((left/c)**a)-1)**b).
V=(1+(EXP((truc/c)**a)-1)**b)/(1+(EXP((right/c)**a)-1)**b).
IF (right NE 9999) THEN LNF=LN(U-V).
IF (right EQ 9999) THEN LNF=LN(U).
/END
/END
```

Note: The truncated data values in the first column, the left limit in the second

column, and the right limit in the third column should be inserted between the last two /END command.

R-codes

1. for Turnbull curve given in Section 3.7.5

```
require(survival)
cria.tau <- function(data){
l <- data$left
r <- data$right
tau <- sort(unique(c(l,r[is.finite(r)])))
return(tau)
}
S.ini <- function(tau){
m<-length(tau)
ekm<-survfit(Surv(tau[1:m-1],rep(1,m-1)))
So<-c(1,ekm$surv)
p <- -diff(So)
return(p)
}
cria.A <- function(data,tau){
tau12 <- cbind(tau[-length(tau)],tau[-1])
interv <- function(x,inf,sup) ifelse(x[1]>=inf & x[2]<=sup,1,0)
A <- apply(tau12,1,interv,inf=data$left,sup=data$right)
```

```
id.lin.zero <- which(apply(A==0, 1, all))
if(length(id.lin.zero)>0) A <- A[-id.lin.zero, ]
return(A)
}
Turnbull <- function(p, A, data, eps=1e-3, iter.max=200, verbose=FALSE){
n<-nrow(A)
m<-ncol(A)
Q<-matrix(1,m)
iter <- 0
repeat {
iter <- iter + 1
diff<- (Q-p)
maxdiff<-max(abs(as.vector(diff)))
if (verbose)
print(maxdiff)
if (maxdiff<eps | iter>=iter.max)
break
Q<-p
C<-A%*%p
p<-p*((t(A)%*%(1/C))/n)
}
cat("Iterations = ", iter,"\n")
cat("Max difference = ", maxdiff,"\n")
```

```
cat("Convergence criteria: Max difference < 1e-3","\n")
dimnames(p)<-list(NULL,c("P Estimate"))
surv<-round(c(1,1-cumsum(p)),digits=5)
right <- data$right
if(any(!(is.finite(right)))){
t <- max(right[is.finite(right)])
return(list(time=tau[tau<t],surv=surv[tau<t]))
}
else
return(list(time=tau,surv=surv))
}
dat <- read.table("C:/Documents and Settings/Desktop/aids.txt",header=T)
dat$right[is.na(dat$right)] <- Inf
tau <- cria.tau(dat)
p <- S.ini(tau=tau)
A <- cria.A(data=dat,tau=tau)
tb <- Turnbull(p,A,dat)
tb
plot(tb$time,tb$surv,lty=1, col = 1,type="s",ylim=c(0,1),xlim=range(c(0,26)),
xlab="x",ylab="S(x)")
text(8,0.87,"Fitted \n Turnbull's \n curve",col=1)
```

REFERENCES

Aarset MV (1987) How to identify a bathtub hazard rate. *IEEE Transactions on Reliability* **R-36**, 106-8.

Abramowitz M and Stegun IA (1972) *Handbook of Mathematical Functions.* Dover Publications.

Aitchison J, Kay JW, and Lauder IJ (2005) *Statistical Concepts and Applications in Clinical Medicine.* Chapman & Hall.

Anderson TW and Darling DA (1954) A test of goodness of fit. *Journal of the American Statistical Association* **49**, 765-9.

Bain LJ (1978) *Statistical Analysis of Reliability and Life-Testing Models.* Volume 24. Statistics: Textbooks And Monographs.

Bain LJ (1974) Analysis for the linear failure rate life testing distribution. *Technometrics* **16**, 551-9.

Balakrishnan N (1992) *Handbook of the Logistic Distribution.* Volume 123. Statistics: text books and monographs.

Balakrishnan N and Cohen AC (1990) *Order Statistics and Inference: Estimation Methods.* Academic Press, Boston.

Balanda (1987) Kurtosis comparisons of the Cauchy and double exponential distributions. *Communications in Statistics-Theory and Methods* **16**, 579-92.

Barlow RE and Campo R (1975) Total time on test processes and applications

to failure data analysis. In Barlow RE, Fussell JB and Singpurwalla ND, editors, *Reliability and Fault Tree Analysis.* Society for Industrial and Applied Mathematics, 451-81.

Barlow RE and Proschan F (1981) *Statistical Theory of Reliability and Life Testing.* To Begin With.

Barndorff-Nielsen O (1978) Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics* **5**, 151-7.

Beadle GF, Come S, Hendeson C, Silver B, and Hellman SAH (1984 a) The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. *International journal of Radiation Oncology, Biology and Physics* **10**, 2131-7.

Beadle GF, Harris JR, Silver B, Botnick L and Hellman SAH (1984 b) Cosmetic results following primary radiation therapy for early breast cancer. *Cancer* **54**, 2911-8.

Beirlant J, Teugels JL and Vynckier P (1996) *Practical Analysis of Extreme Values.* Leuven University Press.

Beirlant J, Joossens E and Segers J (2004) Generalized Pareto fit to the Society of Actuaries' large claims database. *North American Actuarial Journal* **8**, 108-11.

Berger JO (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag.

Birnbaum ZW and Saunders SC (1969) A new family of life distributions. *Journal of Applied Probability* **6**, 319-27.

Bjerkedal T (1960) Acquisition of resistance in guinea pigs infected with different

doses of virulent tubercle bacilli. *American Journal of Hygiene* **72**, 130-48.

BMDP (1992) *Statistical Software Manual.* University of California Press.

Bowers N, Gerber H, Hickman J, Jones D and Nesbitt C (1986) *Actuarial Mathematics.* Society of Actuaries.

Burr IW (1942) Cumulative frequency functions. *Annals of Mathematical Statistics* **13**, 215-32.

Chang SC (1998) Using parametric statistical models to estimate mortality structure: the case of Taiwan. *Journal of Actuarial Practice* **6** (1 & 2).

Chaudhry MA and Zubair SM (2001) *On A Class of Incomplete Gamma Functions with Applications.* Chapman & Hall.

Chen Z (2000) A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics and Probability Letters* **49**, 155-61.

Coles S (2001) *An Introduction to Statistical Modeling of Extreme Values.* Springer.

Cooray K (2005) Analyzing lifetime data with long-tailed skewed distribution: the logistic-sinh family. *Statistical Modelling* **5**, 343-58.

Cooray K (2006) Generalization of the Weibull distribution: the odd Weibull family. *Statistical Modelling* **6**, 265-77.

Cooray K (2005) The Weibull-Pareto composite family with applications to the analysis of unimodal failure data. *Submitted for publication.*

Cooray K and Ananda MA (2008) A generalization of the half-normal distribution with applications to lifetime data. *Communications in Statistics-Theory and Methods* **37**, 1323-37.

Cooray K and Ananda MA (2005) Modeling actuarial data with a composite

lognormal-Pareto model. *Scandinavian Actuarial Journal* **105**, 321-34.

Cooray K, Gunasekera S and Ananda MA (2006) The folded logistic distribution. *Communications in Statistics-Theory and Methods* **35**, 385-93.

Cox DR and Oakes D (1984) *Analysis of Survival Data.* Chapman & Hall.

Csörgő M, Seshadri V, and Yalovsky M (1975) Application of characterizations in the area of goodness-of-fit. In: Patil GP, Kotz S, and Ord JK (eds.), *Statistical distribution in Scientific Work*, **2** Reidel, Boston, 79-90.

D'Agostino RB and Stephens MA (1986) *Godness-of-Fit Techniques.* Marcel Dekker.

Daniel C (1959) Use of half-normal plots in interpreting two level of experiments. *Technometrics* **1**, 311-41.

Davis DJ (1952) An analysis of some failure data. *Journal of the American Statistical Association* **47**, 113-50.

Dhillon BS (1981) Life distributions. *IEEE Transactions on Reliability* **R-30**, 457-60.

Derman, C. (1964) Some notes on the Cauchy distribution. *National Bureau of Standards Technical note*, Nos. 3-6, Washington, DC.

Eaton WW and Whitmore GA (1977) Length of stay as a stochastic process: a general approach and application to hospitalization for schizophrenia. *Journal of Mathematical Sociology* **5**, 273-92.

Efron B (1988) Logistic regression, survival analysis and the Kaplan-Meier curve. *Journal of the American Statistical Association* **83**, 414-25.

Efron B and Hinkley DV (1978) Assessing the accuracy of the maximum likelihood

estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457-87.

Elandt RC (1961) The folded normal distribution: two methods of estimating parameters from moments. *Technometrics* **3**, 551-62.

Elandt-Johnson RC and Johnson NL (1980) *Survival Models and Data Analysis.* John Wiley & Sons.

Elsayed EA (1996) *Reliability Engineering.* Addison Wesley Longman, Inc.

Embrechts P, Klüppelberg C and Mikosch T (1999) *Modelling Extremal Events for Insurance and Finance.* Springer-Verlag.

Everitt BS and Hand DJ (1981) *Finite Mixture Distributions.* Chapman & Hall.

Finkelstein DM and Wolfe RA (1985) A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933-45.

Fleming TR, O'Fallon JR, O'Brien PC and Harrington DP (1980) Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36**, 607-25.

Fürth J, Upton AC and Kimball AW (1959) Late pathologic effects of atomic detonation and their pathogenesis. *Radiation Research* **1**, 243-64.

Gera AE (1995) Lifetime modelling with aid of a modified complementary Weibull distribution. *Quality and Reliability Engineering International* **11**, 379-88.

Ghitany ME (2001) A compound Rayleigh survival model and its application to randomly censored data. *Statistical Papers* **33**, 187-202.

Gilchrist WG (2000) *Statistical Modelling with Quantile Functions.* Chapman & Hall.

Glaser RE (1980) Bathtub and related failure rate characterizations. *Journal of*

the American Statistical Association **75**, 667-72.

Glen AG and Leemis LM (1997) The arctangent survival distribution. *Journal of Quality Technology* **29**, 205-10.

Gnedenko BV, Belyayev YK, and Solvyev AD (1969) *Mathematical Methods of Reliability Theory.* Academic Press, New York.

Gompertz B (1825) On the nature of the function expressive of the law of human mortality and on a new mode of determining the value of life contingencies. *Philosophical Transactions, Royal Society of London* **115**, 513-85.

Greenwich MA (1992) Unimodal hazard rate function and its failure distribution. *Statistical Papers* **33**, 187-202.

Greenwood PE and Nikulin MS (1996) *A guide to Chi-squared Testing.* John Wiley & Sons.

Gujarati ND (2002) *Basic Econometrics.* McGraw-Hill.

Hastings C, Mosteller F, Tukey JW, and Winsor CP (1947) Low moments for small samples: a comparative study of statistics, *Annals of Mathematical Statistics* **18**, 413-26

Haupt E and Schäbe H (1997) The TTT transformation and a new bathtub distribution model. *Journal of Statistical planning and Inference* **60**, 229-40.

Hogg RV and Klugman SA (1984) *Loss Distributions.* John Wiley & Sons.

Horn PS (1983) A measure of peakedness. *The American Statistician* **37**, 55-6.

Hossack IB, Pollard JH and Zehnwirth B (1983) *Introductory Statistics with Applications in General Insurance.* Cambridge University Press.

Hyde (1980) Survival Analysis with Incomplete Observations. In *Biostatistic*

Casebook, Miller RG, Efron B, Brown BW, and Moses LE, eds. John Wiley & Sons, 31-46.

IMSL (1991) *International Mathematical and Statistical Libraries.* User's Manual.

Jeffreys H (1961) *Theory of Probability.* Oxford University Press, London.

Johnson NL, Kotz S and Balakrishnan N (1994) *Continuous Univariate Distributions.* John Wiley & Sons.

Juran JM and Gryna FM (1970) *Quality Planning and Analysis.* McGraw-Hill.

Kalbfleisch JD and Lawless JF (1989) Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association* **84**, 360-72.

Kalbfleisch JD and Prentice RL (2002) *The Statistical Analysis of Failure Time Data.* John Wiley & Sons.

Keller AZ, Giblin MT and Farnworth NR (1985) Reliability analysis of commercial vehicle engines. *Reliability Engineering* **10**, 15-25.

Keller AZ and Kamath ARR (1982) Alternative reliability models for mechanical systems. *Proceeding of the Third International Conference on Reliability and Maintainability*, 411-15, Toulouse, France.

Kimball AW (1960) Estimation of mortality intensities in animal experiments. *Biometrics* **16**, 505-21.

Klein JP and Moeschberger ML (1997) *Survival Analysis.* Springer-Verlag.

Klugman SA, Panjer HH and Willmot GE (1998) *Loss Models From Data to Decisions.* John Wiley & Sons.

Kolb J and Ross SS (1980) *Product Safety and Liability.* McGraw-Hill.

Kollia GD (1989) A study of some quantile function families: Isotones and other applications. Unpublished PhD Thesis, The University of Rochester.

Kolmogorov AN (1933) Sulla determinizione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari.* **4**, pp. 83–91.

Kotz S, Kozubowski TJ, and Podgórski K (2001) *The Laplace Distribution and Generalizations.* Birkhäuser.

Koziol JA (1980) Goodness-of-fit tests for randomly censored data. *Biometrika* **67**, 693-96

Lagakos SW and Mosteller F (1981) A case study of statistics in the regulatory process: The FD&C red no. 40 experiments. *Journal of the National Cancer Institute* **66**, 197-212.

Lagakos SW, Barraj LM, and Gruttola VD (1988). Nonparametric analysis of truncated survival data. *Biometrika* **75**, 515-23.

Lawless JF (2003) *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons.

Lee ET and Wang JW (2003) *Statistical Methods for Survival Data Analysis.* John Wiley & Sons.

Leone FC, Nelson LS and Nottingham RB (1961) The folded normal distribution. *Technometrics* **3**, 543-50.

Lin CT (1977) Construction and evaluation of tests of fit–applications of characterizations and profiles of distributions. Unpublished PhD Thesis, The University of Rochester.

Lin CT and Mudholkar GS (1980) A test of exponentiality based on the bivariate

F distribution. *Technometrics* **22**, 79-82.

Lindsey JK (1997) Parametric multiplicative intensities models fitted to bus motor failure data. *Applied Statistics* **46**, 245-52.

MacGillivray HL (1992) Shape properties of the g- and h- and Johnson families. *Communications in Statistics-Theory and Methods* **21**, 1233-50.

Makeham WM (1860) On the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries* **6**, 301-10.

McCullagh P and Nelder J (1998) *Generalized Linear Models.* Chapman & Hall.

McKeague IW, Tighiouart M (2000) Bayesian estimators for conditional hazard functions. *Biometrics* **56**, 1007-15.

McNeil A (1997) Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin* **27**, 117-37.

Meeker WQ and Escobar LA (1998) *Statistical Methods for Reliability Data.* John Wiley & Sons.

Miller RG (1983) What price Kaplan-Meier?. *Biometrics* **39**, 1077-81.

Moors JJ (1988) A quantile alternative for kurtosis. *The Statistician* **37**, 25-32.

Mudholkar GS and Kollia GD (1994) Generalized Weibull family: a structural analysis. *Communications in Statistics-Theory and Methods* **23**, 1149–71.

Mudholkar GS, Kollia GD, Lin CT, and Patel KR (1991) A graphical procedure for comparing goodness-of-fit tests. *Journal of the Royal Statistical Society* **53**, 221-32.

Mudholkar GS, Srivastava DK and Freimer M (1995) The exponentiated Weibull family: a reanalysis of the bus-motor-failure data. *Technometrics* **37**, 436-45.

Mudholkar GS, Srivastava DK and Kollia GD (1996) A generalization of the

Weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association* **91**, 1575-83.

Murthy DNP, Xie M and Jiang R (2004) *Weibull Models.* John Wiley & Sons.

Nair VN (1981) Plots and tests for goodness of fit with randomly censored data. *Biometrika* **68**, 99-103.

Nelson LS (1980) The folded normal distribution. *Journal of Quality Technology* **12**, 236-38.

Pan W and Chappell R (1998) Estimating survival curves with left-truncated and interval-censored data under monotone hazards. *Biometrics* **54**, 1053-60.

Pan W and Chappell R (2002) Estimation in the Cox proportional hazards model with left-truncated and interval-censored data. *Biometrics* **58**, 64-70.

Pareto V (1897) *Curs d'Economie Politique*, Paris: Rouge et Cie.

Patrik G (1980) Estimating casualty insurance loss amount distributions. *Proceedings of the Casualty Actuarial Society* **LXVII**, 57-109.

Peto R (1973) Experimental survival curves for interval-censored data. *Applied Statistics* **22**, 86-91.

Pearson ES and Hartley HO (1972) *Biometrika Tables for Statisticians.* Cambridge University Press.

Plackett RL (1959) The analysis of life test data. *Technometrics* **1**, 9-19.

Preda V and Ciumara R (2006) On composite models: Weibull-Pareto and lognormal-Pareto. - A comparative study-. *Journal for Economic Forecasting* **3**, 32-46.

Psarakis S and Panaretos J (1990) The folded t distribution. *Communications in Statistics-Theory and Methods* **19**, 2717-34.

Rajarshi S and Rajarshi MB (1988) Bathtub distributions: a review. *Communications in Statistics-Theory and Methods* **17**, 2597-621.

Ramlau-Hansen H (1988) A solvency study in non-life insurance. part 1. Analyses of fire, windstrom and glass claims. *Scandinavian Actuarial Journal* **1-2**, 3-34.

Rao CR (1973) *Linear Statistical Inference and Its Applications.* John Wiley & Sons.

Resnick SI (1997) Discussion of the Danish data on large fire insurance losses. *ASTIN Bulletin* **27**, 139-51.

Richman DD, Grimes JM, and Lagakos SW (1990). Effect of stage of disease and drug dose on zidovudine susceptibilities of isolates of human immunodeficiency virus. *Journal of AIDS* **3**, 743-6.

Rieck JR and Nedelman JR (1991) A log-linear model for the Birnbaum-Saunders distribution. *Technometrics* **33**, 51-60.

Risvi MH (1971) Some selection problems involving folded normal distribution. *Technometrics* **13**, 355-69.

Rosenberger JL and Gasko M (1983) Comparing location estimators: Trimmed means, medians and trimean, *In Understanding Robust and Exploratory Data Analysis*, (eds., Hoaglin DC, Mosteller F, and Tukey JW), John Wiley, 297-338.

Ryan LM and Lindsey JC (1998) Tutorials in Biostatistics; Methods for interval-censored data. *Statistics in Medicine* **17**, 219-38.

Sampford MR (1952) The estimation of response-time distributions II : Multi-stimulus distributions. *Biometrics* **8**, 307-69.

Shooman ML (1968) *Probabilistic Reliability: An Engineering Approach.* McGraw-

Hill Book Company.

Simiu E, Heckert NA, Filliben JJ and Johnson SK (2001) Extreme wind load estimates based on the Gumbel distribution of dynamic pressures: an assessment. *Structural Safety* **23**, 221-29.

Smith RM and Bain LJ (1975) An exponential power life-testing distribution. *Communications in Statistics-Theory and Methods* **4**, 469-81.

Smith RL and Naylor JC (1987) A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Applied Statistics* **36**, 358-69.

Sprent P and Smeeton NC (2000) *Applied Nonparametric Statistical Methods.* Chapman & Hall.

Stablein DM, Carter WH, Novak JW (1981) Analysis of survival data with non-proportional hazard functions. *Controlled Clinical Trials* **2**, 149-59.

Stacy EW (1962) A generalization of the gamma distribution. *Annals of Mathematical Statistics* **33**, 1187-92.

Stacy EW and Mihram GA (1965) Parameter estimation for a generalized gamma distribution. *Technometrics* **7**, 349-58.

Sundberg R (1974) On estimation and testing for the folded normal distribution. *Communications in Statistics-Theory and Methods* **3**, 55-72.

Tukey JW (1960) The practical relationship between the common transformations of percentages of counts and of amounts, *Technical Report* **36**, Statistical Techniques Research Group, Princeton University.

Turnbull BW (1976) The empirical distribution function from arbitrarily grouped, censored and truncated data. *journal of Royal Statistical Society* B **38**, 290-5.

Verhulst PJ (1838) Notice sur la lois que la population suit dans sons accroissement. *Correspondance Mathématique et Physique* **10**, 113-21.

Verhulst PJ (1845) Recherches mathématiques sur la loi d'accroissement de la population. *Académie de Bruxelles* **84**, 742-8.

Wang MC (1989) A semiparametric models for randomly truncated data. *Journal of the American Statistical Association* **65**, 1601-9.

Weibull W (1939) Statistical theory of the strength of materials. *Ingenioor Vetenskps Akademiens Handlingar* **151**, 1-45.

West M (1987) Analysis of nasopharynx cancer survival data using dynamic Bayesian models. *Warwick Research Report* 109 and *Technical Report* 7-1987, Department of Mathematics, II, University of Rome.

West M (1992) Modelling time-varying hazards and covariate effects. *In Survival Analysis: State of the Art.* Klein JP and Goel PK, editors, Kluwer, 47-62.

Whitmore GA (1986) Normal-gamma mixtures of inverse Gaussian distributions. *Scandinavian Journal of Statistics* **13**, 211-20.

Zellner A (1971) Bayesian and non-Bayesian analysis of the log-normal distribution and log-normal regression. *Journal of the American Statistical Association* **66**, 327-30.

VITA

Graduate College
University of Nevada, Las Vegas

Kahadawala Cooray

Home Address:
1601 East University Avenue, 205
Las Vegas, NV 89119

Degrees:
Bachelor of Science, Chemistry and Mathematics, 1994
University of Colombo, Sri Lanka

Special Honors and Awards:
Received the Wolzinger Family Research Scholarship for the 2007-2008 academic year at the University of Nevada, Las Vegas.

Nominated in the Marquis Who's Who in America, 63rd Edition.

Publications:
Cooray, K. (2005). Analyzing lifetime data with long-tailed skewed distribution: the logistic-sinh family. Statistical Modelling 5, 343-358.

Cooray, K. (2006). Generalization of the Weibull distribution: the Odd Weibull family. Statistical Modelling 6, 265-277.

Cooray, K. and Ananda, M. M. A. (2005). Modeling actuarial data with a composite lognormal-Pareto model. Scandinavian Actuarial Journal 105, 321-334.

Cooray, K., Gunasekera, S., and Ananda, M. M. A. (2006). The folded logistic distribution. Communications in Statistics-Theory and Methods 35, 385-393.

Cooray, K. and Ananda, M. M. A. (2008). A generalization of the half-normal distribution with applications to lifetime data. Communications in Statistics-Theory and Methods 37, 1323-1337.

Dissertation Title: Statistical Modeling of Skewed Data Using Newly Formed Parametric Distributions

Dissertation Examination Committee:

Chairperson, Prof. Malwane M. A. Ananda, Ph. D.
Committee Member, Prof. Chih-Hsiang Ho, Ph. D.
Committee Member, Prof. Hokwon Cho, Ph. D.
Committee Member, Prof. Sandra Catlin, Ph. D.
Graduate Faculty Representative, Prof. Chad Cross, Ph. D.