Chapter 1
INTRODUCTION AND PRELIMINARIES

1.1 Introduction

Skewed distributions play an important role in the analysis of sample data originating from life span, reaction time, reliability, survivor, and related studies. In the field of engineering, skewed distributions are employed as model for life tests and for the distribution of characteristics such as dimension, strength, and hardness of materials and products. They are also useful as model for distributions of voltage, amperage, capacitance, resistance, and other characteristics of interest in electric and electronic devices. In medicine and biology, skewed distributions play a role in survival studies and in distributions of characteristics such as blood pressure, other measures of vital functions, cholesterol level, and various other measures of body chemistry. In economics, they serve as model for distributions of income, sales volume, tax collection, insurance premiums and claims, and for other items of interest in economics and financial studies. Accordingly, these distributions are important in business, engineering, quality control, medical and biological sciences, and in all areas

of physical sciences. They are of importance in the area of research and development.

Statistical modeling for skewed lifetime data is done through well known distributions like Weibull, exponential, gamma, lognormal etc. In practice statisticians can come across number of situations wherein the existing well known models fail to model observed data. So in such cases, it is necessary to provide suitable model for such data sets. In recent years new models have been proposed by modifying existing well known models. Exponentiated family of distributions is one of them and is proposed by introducing additional parameter to existing distribution. i.e. Suppose $G(x,\theta)$ is the cumulative distribution function (c.d.f.) of a continuous base line distribution with parameter θ then c.d.f of corresponding exponentiated family of distribution is given by

$$F(x, \alpha,\theta)= (G(x,\theta))^{\alpha} \quad ,\alpha>0 , \theta>0, x \in S, \qquad (1.1.1)$$

where α is the additional parameter, θ may be a vector and S is the support of x independent of parameters .

One can also define $F(x, \alpha,\theta)$ in terms of survival function of the existing distribution, $s(x,\theta)$ as

$$F(x, \alpha,\theta) = 1- (1-G(x,\theta))^{\alpha} = 1- (s(x,\theta))^{\alpha} \quad ,\alpha>0 , \theta>0, x \in S. \quad (1.1.2)$$

In this chapter, Section 1.2 consists of brief review of the literature on exponentiated family of distributions. The motivation for the present work has been provided in Section 1.3. Section 1.4 deals with design of thesis. Some pre-requisite definitions and basic results are reported in Section 1.5 and chapterwise summary is given in Section 1.6.

1.2 Review of Literature

Initially, Mudholkar and Srivastava (1993) proposed a three parameter distribution (one scale and two shape parameters), the exponentiated Weibull distribution as extension of the Weibull family, which contains distributions with bathtub shaped and unimodal failure rates besides a broader class of distributions with monotone failure rates. The cumulative distribution function (c.d.f.) of an exponentiated Weibull random variable X is given by

$$F(x)=(1-\exp(-(\lambda x)^{\theta}))^{\alpha} \quad x>0,\ \alpha>0,\ \lambda>0,\ \theta>0, \qquad (1.2.1)$$

where α, θ are two shape parameters and λ is a scale parameter. The corresponding probability density function (p.d.f.) is

$$f(x)=\theta\,\alpha\,\lambda^{\theta}\,x^{\theta-1}\exp(-(\lambda x)^{\theta})\ (1-\exp(-(\lambda x)^{\theta}))^{\alpha-1},\ x>0,\ \alpha,\ \lambda,\ \theta>0. \qquad (1.2.2)$$

Mudholkar, Srivastava, Freimer (1995), illustrated the usefulness and flexibility of the exponentiated Weibull family by reanalyzing five classical

data sets on bus-motor failures from Davis (1952) and Efron's (1988) clinical trial data pertaining to a head and neck cancer. It is observed that the exponentiated Weibull has a better fit than the two parameter Weibull or one parameter exponential distribution, which are special cases of the exponentiated Weibull distribution.

Gupta, Gupta and Gupta (1998), proposed to model failure time data by $F(t)=(G(t))^{\theta}$, where G(t) is the baseline distribution function and θ is a positive real number. They studied monotonicity of the failure rates in general and some order relations are examined. In literature, this model has been called as Lehman alternatives, when θ is in fact a positive integer. Lehman (1953) has studied such alternatives to define various nonparametric hypotheses and has computed the approximate power of certain rank tests using large sample theory.

Gupta et al. (1998) also introduced exponentiated exponential (EE) distribution, which is a particular member of exponentiated Weibull distribution. The c.d.f. and p.d.f. of EE distribution is given by equations (1.2.1) and (1.2.2) respectively, where $\theta=1$. Gupta and Kundu (2001a) observed that two parameter EE distribution or generalized exponential (GE) distribution can be used quite effectively to analyze positive lifetime data, particularly, in place of two parameter gamma or Weibull distributions. The

genesis of EE distribution, several properties, different estimation procedure and their properties, closeness of this distribution with gamma distribution, estimation of stress-strength parameter are discussed in series of papers Gupta and Kundu (2001a, 2001b, 2002, 2003, 2005).

Kundu, Gupta and Manglick (2005) proposed a very convenient method to generate a normal random variable using EE distribution. The new method is compared with the other existing methods and it is observed that the proposed method is quite competitive with most of the existing methods in terms of the Kolmogorv-Smirnov distances and the corresponding p-values.

Recently Surles and Padgett (2001) introduced a two parameter scaled Burr Type X distribution and named correctly as the generalized Rayleigh (GR) distribution, which is also a particular member of exponentiated Weibull distribution. The c.d.f. and p.d.f. of GR distribution is given by (1.2.1) and (1.2.2) respectively, where $\theta=2$. The GR distribution is a positively skewed unimodal distribution useful in modeling strength data.

Raqab and Kundu (2005) considered the estimation of R = P(Y<X), where Y and X are two independent scaled Burr Type X (GR) distribution having the same scale parameters. Kundu and Gupta (2005) considered the estimation of R = P(Y<X), where Y and X are two independent GE

distribution having the same scale parameters. Both the papers deal with the maximum likelihood estimator of R and its asymptotic distribution is used to construct an asymptotic confidence interval of R. Different point estimators like maximum likelihood estimator, uniformly minimum variance unbiased estimator, and Bayesian estimator for known scale parameter have discussed and compared their performances through simulations.

Jeevanand and Nair (1994) discussed the problem of estimating R = P(Y<X), where X and Y are independent exponential random variables and the sample from each population contains one spurious observation. Kim and Chung (2006) considered the same problem by taking Burr-type X distribution instead of exponential distribution. In both the papers, the Bayes estimates are derived for exchangeable and identifiable cases.

Nadarajah (2005) introduced exponentiated Gumbel distribution that generalized the standard Gumbel distribution. The c.d.f of the standard Gumbel distribution is

$$F(x)=\left(\exp\left(-\exp\left(\frac{-(x-\mu)}{\sigma}\right)\right)\right),\ \sigma>0\ ,-\infty<\mu<\infty,\ -\infty<x<\infty. \qquad (1.2.3)$$

The c.d.f. of exponentiated Gumbel distribution, in terms of survival function, is defined as

$$F(x)=1-\left(1-\exp\left(-\exp\left(\frac{-(x-\mu)}{\sigma}\right)\right)\right)^{\alpha}, \alpha,\sigma>0, -\infty<\mu<\infty, -\infty<x<\infty. \quad (1.2.4)$$

The corresponding p.d.f. of (1.2.4) is given by

$$f(x)=\frac{\alpha}{\sigma}\left(1-\exp\left(-\exp\left(\frac{-(x-\mu)}{\sigma}\right)\right)\right)^{\alpha-1}\exp\left(\frac{-(x-\mu)}{\sigma}\right)\exp\left(-\exp\left(\frac{-(x-\mu)}{\sigma}\right)\right)$$

Nadarajah (2006) provides a comprehensive treatment of the mathematical properties of this new distribution and illustrates its applicability for modeling rainfall data from Orlando, Florida.

Shirke, Kumbhar and Kundu (2005) introduced exponentiated scale family of distributions and provided an asymptotic upper β-expectation and β-content γ-level tolerance intervals. Expected coverage of a proposed β-expectation tolerance interval has been obtained. They have also obtained bootstrap based tolerance limit for Lawless (1982) data assuming EE distribution suitable for the data.

1.3 Motivation of Present Work

Literature survey revealed that there is a scope to study exponentiated family of distributions in general and its members with regard to distributional properties, properties of estimation procedures like consistency, asymptotic normality, point and interval estimation, testing of

hypothesis, estimation of stress-strength parameters, tolerance intervals and statistical modeling for the real life data sets. We have considered the members of exponentiated family of distributions that are positively skewed distributions, i.e. distributions are skewed to the right. They have applications in theory of reliability and analysis of skewed data occurs in all areasof business, engineering and medical, biological and physical sciences.

Salient features of the work reported in the thesis are as follows.

(i) We study the distributional properties and some reliability measures such as hazard rate, reverse hazard rate and stochastic orderings, of exponentiated scale family of distributions and exponentiated scale and location family of distributions and provide graphical illustrations.

(ii) We study inference about parameters involved in the distributions, point estimation, interval estimation and testing of hypothesis.

(iii) Some new exponentiated type distributions, with base line distributions like gamma, Gumbel, normal and lognormal, are defined and studied as mentioned in (i) and (ii) above. These distributions are used to model real life data.

(iv) Modified exponentiated scale family of distributions, a mixture of singular distribution at zero and two parameter exponentiated scale family of distributions, has been introduced and is used to accommodate instantaneous failures.

(v) Bayesian and non-Bayesian inference of $R = P(Y<X)$ are provided for exponentiated scale family of distributions. Simulation study for testing of R has been reported to distributions like exponentiated gamma distribution and exponentiated Gumbel distribution.

(vi) Bayes estimator of R has been obtained for exponentiated scale family of distributions containing one spurious observation. An application to the EE distribution is provided.

(vii) Based on grouped data, inference for exponentiated scale and location family of distributions has been provided. Point and interval estimation has been provided using maximum likelihood method. Tolerance intervals based on grouped data are provided for exponentiated scale family of distributions. An application to EE distribution with real life

data set has been provided with simulation study.

1.4 Pre-requisite Definitions

In this section, we have discussed some important definitions and results on statistical inference and calculus useful in the subsequent part of the thesis.

Maximum likelihood estimator (MLE)

Let a random sample $\underline{x}=(x_1,x_2,.....x_n)$ is from a distribution having p.d.f. $f(x,\underline{\theta})$, $\underline{\theta}=(\theta_1,\theta_2,...,\theta_k)$, the vector of parameters belongs to a set $\underline{\theta}\in\Theta\subset R^k$. The likelihood function L of θ given the sample observations is defined to be

$$L(\underline{\theta}/\underline{x})\propto f(\underline{x};\underline{\theta}),\quad \underline{\theta}\in\Theta.$$

Suppose $\hat{\theta}=T(x_1,x_2,.....x_n)$ is a nontrivial function of $x_1,x_2,.....x_n$ such that $L(\hat{\theta}/\underline{x})=\sup_{\theta\in\Theta} L(\underline{\theta}/\underline{x})$ then $\hat{\theta}$ is called as maximum likelihood estimator of θ.

Invariance and asymptotic properties of MLE

(1): Let $\hat{\theta}$ be the maximum likelihood estimator of θ, where θ is assumed to be a scalar. If $\Psi(\theta)$ is a function of θ then the MLE of $\Psi(\theta)$ is $\Psi(\hat{\theta})$.

(2): Under regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is a consistent asymptotically normal (CAN) estimator of θ.

i.e. $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0,1/I(\theta))$, where $I(\theta)$ is known as Fisher information and is given by $I(\theta) = E_\theta\left(\frac{d \ln f(\underline{x},\theta)}{d\theta}\right)^2 = E_\theta\left(\frac{-d^2 \ln f(\underline{x},\theta)}{d\theta^2}\right)$.

The likelihood ratio test

Let $\underline{\theta}$ be a vector of parameters with set of possible values Θ. One can test the hypothesis $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$,

where Θ_0 and Θ_1 are two disjoint subsets of Θ and $\Theta_1 = \Theta - \Theta_0$.

A ratio of likelihoods $\Lambda = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$ where the denominator is maximized likelihood function with respect to parameter $\underline{\theta}$ and the numerator is maximized only after some or all of the parameters have been restricted by H_0. The likelihood ratio statistic is $\Lambda = -2\log\lambda$. The asymptotic distribution of Λ under H_0 has chi-square distribution with degrees of freedom given by the number of parameters which are estimated under H_1 but fixed under H_0.

Newton-Raphson method

Let r be the root of a real valued function $f(x)$, that is, $f(r) = 0$. Assume that $f'(x) \neq 0$.

Let x_0 be an approximate initial solution of $f(x)=0$ then an improved solution to $f(x)=0$ is obtained by iterative formula $x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$, for i=1,2,..... . This process will generate a sequence of numbers $\{x_n\}$ which approximates r. This technique of successive approximations of roots using above iterative formula is called Newton-Raphson method.

Leibniz rule for differentiating an integral

Let $I(t) = \int_{g(t)}^{h(t)} f(x;t)\,dx$, where $f(.;.), g(.)$ and $h(.)$ are assumed differentiable then $\frac{dI}{dt} = \int_{g(t)}^{h(t)} \frac{df}{dt}\,dx + f(h(t);t)\frac{dh}{dt} - f(g(t);t)\frac{dg}{dt}$.

AIC and BIC

AIC stands for Akaikie Information Criterion. BIC stands for Bayesian Information Criterion. Both criteria are used to select the best fitting model among various models. These two criteria are based on log-likelihood value (L) and number of parameters in the distribution (k) and defined as AIC = L- 2k and BIC = L-(k/2) log(n) where n is sample size. The distribution with the largest AIC (BIC) value is the distribution that fits the data the best.

Tolerance intervals

Let U be a statistic based on data observed from a distribution with density function f(x, $\underline{\theta}$) where $\underline{\theta}$ represents a vector of unknown parameters then the interval (-∞, U) is a β -expectation tolerance interval if

$$E\left[\int_{-\infty}^{U} f(x,\theta)dx\right]=\beta \text{ , for every } \underline{\theta}\in\Theta$$

and interval (-∞, U) is upper β content γ level tolerance interval if

$$P\left[\int_{-\infty}^{U} f(x,\theta)dx \geq \beta\right]=\gamma \text{ , for given } \beta \text{ , } \gamma \in (0,1).$$

C is called the coverage of tolerance interval (-∞, U), if $C=\int_{-\infty}^{U} f(x,\theta)dx$.

1.5 Chapterwise Summary

In addition to the first introductory chapter, the thesis contains four more chapters. The chapterwise summary is presented in brief as follows.

In Chapter 2, we discuss exponentiated scale family of distributions. The c.d.f. of two parameter exponentiated scale family of distributions is defined as $F(x;\alpha,\theta)=\left(G\left(\frac{x}{\theta}\right)\right)^{\alpha}$, α>0, θ>0, x ∈ S, (1.5.1)

where G(.) is the c.d.f. of baseline distribution with scale parameter θ.

Some members of the exponentiated scale (ES) family of distributions are exponentiated exponential (EE) distribution, exponentiated gamma (EG) distribution, exponentiated Frechet (EF) distribution, and exponentiated Gumbel (EGum) distribution. There are situations wherein members of this family can be used to model real life skewed data.

We study the distributional properties and the estimation of parameters of this family in general and for EG and EGum distribution in particular. We obtain the MLE by using iterative procedure viz. Newton-Raphson method, since the associated likelihood equations do not lead to close form solution. We discuss asymptotic properties of MLE and obtain expressions for Fisher information matrix which is used to obtain asymptotic confidence intervals. Also, we discuss Efron's (1982) bootstrap percentile method to obtain asymptotic bootstrap confidence interval. Testing of parameters based on asymptotic distribution using likelihood ratio test has been discussed. Part of this work (the one related to EGum) is accepted for publication in TRAJECTORY journal (Kakade and Shirke (2007a)).

When the scale parameter is known, we obtain the exact distribution of an estimator of shape parameter and is used to obtain exact confidence interval for shape parameter. Testing of shape parameter based on exact and asymptotic distributions has been discussed.

We apply EG and an EGum distributions to real life data sets. It is observed that EGum distribution is more suitable model than two parameter gamma, Weibull and EE distributions. Similarly, we illustrate that the EG distribution can be used as a possible alternative to gamma, Weibull and EE distributions for analyzing above mentioned real life data. Part of this work is accepted for publication in International Journal of Agricultural and Statistical Science (Kakade and Shirke (2007b)).

Modified exponentiated scale family of distributions is provided as a mixture of singular distribution at zero and two parameter ES family of distributions to analyze instantaneous failures in ES family of distributions. We consider two real life data sets namely Vanmann data (1991) and rainfall data of Jalgaon for the year 1961. We apply the modified EE model for the above data sets and obtain the estimates of the parameters based on instantaneous failures and approximate 95 % confidence intervals.

Chapter 3 deals with exponentiated scale and location family of distributions. Three parameter exponentiated scale and location family is defined by c.d.f., $F(x;\alpha,\mu,\theta)=\left(G\left(\frac{x-\mu}{\theta}\right)\right)^{\alpha}$, $\alpha>0$, $\theta>0$, $\mu \in R$, $x \in S$, (1.5.2)

where G(.) is the c.d.f. of baseline distribution with parameters μ & θ.

We study the distributional properties and estimation of parameters based on maximum likelihood method. An asymptotic property of MLE is discussed and it provides Fisher information matrix. Testing of a shape parameter based on likelihood ratio test has been discussed. We apply this procedure to members of exponentiated scale and location family of distributions namely exponentiated normal (EN) and exponentiated lognormal (ELN) distributions.

We use EN distribution to model Smith and Naylor (1987) data set and ELN distribution to model Lawless (1982) data set and Airplane polished window data set. We observe that an EN and ELN distributions fit quite good to the above data sets. ELN distribution can be used to analyze above data sets in place of gamma, Weibull, lognormal and EE distributions. Part of this work is published in International Journal of Agricultural and Statistical Science (Kakade and Shirke (2006)).

Chapter 4 presents the inference of $R = P(Y<X)$, when X and Y are independent but not identically distributed random variables from exponentiated scale family of distributions. We obtain MLE of R and its asymptotic distribution. When scale parameter is known, the exact distribution of MLE of R has been obtained. Exact and asymptotic confidence intervals for R are provided. Uniformly minimum variance

unbiased estimator and Bayesian estimator for R have been discussed. Performances of the estimators are studied through simulations.

Testing of reliability R based on exact and asymptotic distributions of the MLE are discussed along with simulation study. Comparison of parametric test with usual nonparametric Wilcoxon-Mann-Whitney test is also considered.

The part of work related to exponentiated scale family of distributions has been submitted for publication (Shirke and Kakade (2007c)).

As a particular case, inference about R is considered, when X and Y are independent but not identically distributed i) EG and ii) EGum variables.

When the common scale parameter is unknown, the bootstrap confidence interval, based on the MLE of R, works well even when the sample size is small. The performance of the MLE is quite satisfactory in terms of bias and mean squared error (MSE). It is observed that when sample size increases, the MSE decreases, supporting the consistency property of the MLE of R.

Inference of R for EGum distribution has been submitted for publication (Kakade and Shirke (2007d)).

The problem of estimating $R = P(Y<X)$ when X and Y are independent exponentiated scale family distributed random variables and the

sample from each population contains a discordant observation is also discussed in the Chapter 4. A discordant observation is defined as an observation that appears surprising or discrepant under exchangeable model and identifiable model. The exchangeable model assumes that discordant observation is not identifiable and any observation in sample is as likely to be discordant as any other but the identifiable model assumes that a discordant observation is identifiable and we treat the largest observation in the sample as a discordant observation.

Bayes estimates of R based on random sample of size n are derived for exchangeable and identifiable models. We apply this procedure to exponentiated exponential distribution. Part of the work has been submitted for possible publication (Kakade and Shirke (2007e)).

Many times in a life testing problem, due to several reasons, it is not possible to record exact time of the failure of components. Hence, it is more economical to observe number of failures of components in predefined time intervals, which form grouped data. Chapter 5 is devoted to inferences for exponentiated scale and location family of distributions based on grouped data. Based on grouped data, point estimation using maximum likelihood method and asymptotic distribution of MLE have been discussed. Asymptotic confidence interval, β-expectation tolerance interval and

β-content γ-level tolerance interval based on MLE's of parameters have been discussed. Third bus motor failure data (Davis,1952) has been modeled by EE distribution. Simulation study has been carried out for providing MLE, asymptotic confidence interval, percentile bootstrap confidence interval, β-expectation tolerance interval and β-content γ-level tolerance interval.

The thesis ends with the discussion of the scope for further research, the MATLAB programmes which are used for simulations have provided in Appendices A,B and C. The list of references is presented in Bibliography.

Throughout this thesis we labeled Theorem, Lemma and Table numbers by the notation (a.b), where 'a' represents Chapter number and 'b' represents that number. Similarly equations are labeled by (a.b.c), where 'a' represents Chapter number and 'b' represents section number and 'c' represents equation number.

The next chapter deals with exponentiated scale family of distributions in which distributional properties and parametric estimation are carried out for EG and EGum distributions.