Chapter 1

Introduction

1.1 Introduction

An inlier in a set of data is an observation or subset of observations not necessarily all zeros, which appears to be inconsistent with the remaining data set. For example: consider the following example as a natural occurrence of a physical phenomenon: 0, 0, 0, 0, 0.01, 0.05, 0.06, 0.71, 1.91, 1.2, 1.76, 2.54, 2.72, 3.07, 3.91 and 3.99. Here the first four observations are instantaneous failures, next three observations may be treated as early failures (by specifying delta δ=0.06 or 0.08) and others may be treated as coming from any positive distribution *F*. The observations which are identified as instantaneous and early failures together are called inliers, introduced first time by Muralidharan and Kale (2002). In outlier's concept, they may be termed as spurious observations, but unlike outlier concept, we don't discard such observations from analysis and inferences. An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. In many cases outliers exist in the form of errors of observation or mis-recording due to human errors. Outliers are the surprisingly extreme values occurring on both sides of the distribution whereas inliers occur on left hand side of the distribution. Inliers are integral part of the data and cannot be neglected. For an exhaustive survey and theory of outliers one may refer to Barnett and Lewis (1994) and Rousseeuw and Leroy (1987) the references contained therein. Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks. Most of the earliest univariate methods for outlier detection rely on the assumption of an underlying known distribution of the data, which is assumed to be identically and independently distributed (i.i.d.). Moreover, many discordance tests for detecting univariate outliers further assume that the distribution parameters and the type of expected outliers are also known (Barnett and Lewis, 1994). In real world for data-mining applications these assumptions are often violated. In some of the examples discussed above, the inlier observation also becomes a part of outlier observations.

In literature some authors have defined inliers as those observations which are not outliers (Barnett and Lewis, 1994). One can refer Akkaya and Tiku (2005) for this.

Some specific real life situations, where inlier observations are natural occurrences can be described by the following examples:

- In auditing some population elements contain no errors, whereas other population elements contain errors of varying amounts. The distribution of errors can, therefore, be viewed as a mixture of two distinguishable distributions, one with a discrete probability mass at zero and the other a continuous distribution of non-zero positive and/or negative error amounts. The main statistical objective in this auditing problem is to provide a statistical bound for the total error amount in the population.

- In the mass production of technological components of hardware, intended to function over a period of time, some components may fail on installation and therefore have zero life lengths. A component that does not fail on installation will have a life length that is a positive random variable whose distribution may take different forms. Thus, the overall distribution of lifetimes, which includes the duds, is a nonstandard mixture.

- In the study of tumor characteristics, two variates may be recorded. The first is the absence (0) or presence (1) of a tumor and the second is tumor size measured on a continuous scale. In this problem, it is sometimes of interest to consider a marginal tumor measurement that is 0 with nonzero probability and the other a continuous distribution.

- In studies of genetic birth defects, children can be characterized by two variates, a discrete or categorical variable to indicate if one is not affected, affected and born dead, or affected and born alive and a continuous variable measuring the survival time of affected children born alive. The conditional distribution of survival time given, this first variable is undefined for children who are not affected and born dead, and nontrivial for children who are born alive. In some cases it may be necessary to consider the conditional survival time distribution for affected children as a mixture of a mass point at 0 and a nontrivial distribution.

- In measurements of physical performances scores of patients with a debilitating disease such as multiple sclerosis, there will be frequent zero measurements from those giving no performance and many observations with graded positive performance.

- In studies of methods for removing certain behaviors (e.g. predatory behavior or salt consumption), the amount of the behavior which is exhibited at a certain point in time may be measured. In this context, complete absence of the target behavior may represent a different result than would a reduction from a baseline level of the behavior. Thus, one would model the distribution of activity levels as a mixture of a discrete value of zero and a continuous random level.

- Time until remission is of interest in studies of drug effectiveness for treatment of certain diseases. Some patients respond and some do not. The distribution is a mixture of a mass point at 0 and a nontrivial continuous distribution of positive remission times. The problem can be considered for instantaneous failure.

- In a quite different context, important problems exist in time-series analysis in which there are mixed spectra containing both discrete and continuous components.

- The data recorded for a rainy season can be seen as a combination of zeros (no rainfall) and positive observations (days having nominal or marginal rain reported) etc.

From the above examples, it is seen that the values including zeroes and close to zeroes are important as well as significant in most of the cases. Thus inliers are more natural than the outliers, where most of the time after the detection of outlier(s), the observation(s) may not be considered for further analysis. As a consequence, the modeling of inliers distribution is more important than the detection. Below we discuss some possible models treated in this thesis for detection, estimation and testing.

1.2 Models

Various inlier prone models and their statistical significances are studied in this thesis`. We have considered the following models in various chapters. They are used for analysis of mixture distribution of inliers and target observations, and for estimating the parameters of mixture distribution. Comparison of models are also been done to know which model fits well to the data.

1.2.1 Instantaneous failure model

Consider a model $\Im = \{F(x,\theta), x \geq 0, \theta > 0\}$ where $F(x,\theta)$ is a continuous failure time distribution function (df) with $F(0)=0$. To accommodate a real life situation, where instantaneous failures are observed at the origin, the model $\Im$ is modified to $\mathcal{G} = \{G(x,\theta,\alpha), x \geq 0, \theta \in \Omega, 0 < \alpha < 1\}$ by using a mixture in the proportion $1\text{-}\alpha$ and α respectively of a singular random variable Z at zero and with a random variable X with distribution function $F \in \Im$. Thus the modified failure time distribution has the pdf

$$g(x,\theta,\alpha) = \begin{cases} 1-\alpha, & x=0 \\ \alpha f(x,\theta), & x>0 \end{cases} \tag{1.2.1}$$

This model has been studied by many authors. The problem of inference about (α,θ) has received considerable attention particularly when X is exponential with mean θ. Some of the early references are Aitchison (1955), Kleyle and Dahiya (1975), Jayade and Prasad (1990), Vannman (1991), Muralidharan (1999, 2000), Kale and Muralidharan (2000) and references contained therein.

Aitchson (1955) stated the problem of determining efficient estimates of the mean and variance of a distribution specified by (i) non zero probability that the variables assumes a zero value and (ii) a conditional distribution for the positive values of the variable. The estimation problem was analyzed and its implications for the Pearson type-III, exponential, lognormal and Poisson series conditional distribution were investigated. Kleyle and Dahiya (1975) have considered estimation of parameters of mixture distribution of binomial and exponential population. The exact bias and mean square error (MSE) of the estimator is derived and computed for different values of parameters. They had also shown the exact MSE approaches to asymptotic MSE as n increases. Jayade and Prasad (1990) studied the problem of estimation of parameters of a mixture of degenerate and exponential distribution. A new sampling scheme was proposed and the exact bias and MSE of the MLE of the parameters was derived. Moment estimators and their approximate bias and MSE were also obtained. Muralidharan (1999), (2000) obtained tests for the mixing proportion in the mixture of a degenerate rate and exponential distribution. The UMVUE and Bayes estimator of the reliability for some selective prior when the mixing proportion is known and unknown are derived. Muralidharan and Kale (2002) considered the case where F is a two parameter Gamma distribution with shape parameter β and scale parameter θ and obtained confidence interval for $\phi=\alpha\beta\theta$ assuming α known and unknown respectively. Singh (2008) obtained UMVUE for mixture of instantaneous and positive observation from exponential families.

1.2.2 Early failure model

To accommodate early failures, the family $\Im$ is modified to new distribution $\mathcal{G}_1 = \{G_1(x,\theta,\alpha), x\geq 0, \theta\in\Omega, 0<\alpha<1\}$ where the d.f. corresponding to $g_1 \in \mathcal{G}_1$ is given by

$$G_1(x,\theta,\alpha) = (1-\alpha)H(x)+\alpha F(x,\theta)$$

where $H(x)$ is a d.f. with $H(\delta)$=1 for δ sufficiently small and assumed known and specified in advance. The corresponding pdf is then given by

$$g_1(x,\alpha,\theta)=\begin{cases}0, & x<\delta\\ 1-\alpha+\alpha F(\delta,\theta), & x=\delta\\ \alpha f(x,\theta), & x>\delta\end{cases} \tag{1.2.2}$$

Some of the references which treat early failure analysis with exponential distributions are Kale and Muralidharan (2000), Kale (2001) and Muralidharan (2002), wherein they have treated early failures as inliers using the sample configurations. Muralidharan (2005) has presented in his paper, estimation of parameters in presence of early failures. Kale and Muralidharan (2007) obtained MLE for parameter θ of the target distribution F and parameter ϕ of the contaminating population $\mathcal{G}$ assuming number of inliers is known. Muralidharan and Lathika (2008) studied analysis of instantaneous and early failures in Weibull distribution.

Kale and Muralidharan (2008) studied inlier detection using Schwarz information criterion. The estimation of mixture density of inliers and target observation can be viewed as special case of mixture distribution.

1.2.3 Nearly instantaneous failure model

As seen in the data set discussed above, if the observations are closed to zeroes, they can be termed as nearly instantaneous failures. Although the model described in (1.2.2) incorporates inliers for a specified value of δ, there are some drawbacks for the model (1.2.2). This is rectified in the following model as a complete mixture of two distributions. Thus, the nearly instantaneous gives the modification, the density function is given by:

$$f(x)=(1-p)f_1(x)+pf_2(x) \tag{1.2.3}$$

where

$$f_1(x)=\delta_d(x-x_0)=\begin{cases}\frac{1}{d}, & x_0\le x\le x_0+d\\ 0, & otherwise\end{cases}$$

and $f_2(x)$ can be considered as any other lifetime distribution of target population. A mixture distribution involving two-parameter Weibull distributions has been thoroughly studied by Lai, Khoo, Muralidharan and Xie (2007). The importance of the model is that we can obtain the reliability function and hazard function in closed form. The characteristics of the model, such as survival rate, hazard rate and mean residual life, are studied for various distributions in various chapters for particular cases of $f_2(x)$.

1.2.4 M_k inliers Models and L_k inliers Models

Suppose that n units are put on test and n_0 units fail instantaneously and $(n\text{-}n_0)$ failure time are available. Out of these positive observations we have to determine which are inliers or early failures. Before the start of the experiment we are unaware of which unit fail instantaneously or will produce early failures. These experimental conditions are to be modeled in M_k inlier model for given k. Let us denote failure times of these (n-n_0) unit as $\left(X_1,X_2,......X_{n-n_0}\right)$. Then in M_k inlier model, (n- n_0 - k) are considered from target population with pdf $f\in\Im$ and k observations are from inlier population $g\in\mathcal{G}$. Thus the joint pdf of $\left(X_1,X_2,......X_{n-n_0}\right)$ can be written as

$$L\left(x_1,x_2,...x_{n-n_0}\mid f,g,v\right)=\left\{\prod_{i\in v}g(x_i)\right\}\left\{\prod_{i\notin v}f(x_i,\theta)\right\},f\in\Im,v\in V\text{ and }g\in\mathcal{G}\quad(1.2.4)$$

where v is the new parameter representing set of inliers and ranges over V, the set of integers $\left(i_1,i_2,......i_k\right)$ chosen out of $\left(1,2,....[n-n_0]\right)$ and therefore with cardinality $\binom{n-n_0}{k}$. This is so far similar to the model M_k for k outlier. The main difference in M_k inlier model is that $\Psi(x)=\dfrac{\partial G}{\partial F}=\dfrac{g(x)}{f(x)}$ is assumed to be strictly decreasing function of X. The theorem stated below is used to write the likelihood function under M_k and L_k reproduced from Muralidharan (2010), for continuity.

Theorem 1.2.1: Let $\left(X_{(1)}<X_{(2)}<......<X_{(n-n_0)}\right)$ be the order statistics observations and $\left(R_1,R_2,....R_{n-n_0}\right)$ be the corresponding rank order statistics then $Max\,\varphi\left(r_1,r_2,....r_k\right)=\varphi\left(1,2,......k\right)$ and $x_{(1)},x_{(2)}......x_{(k)}$ have the maximum probability of being inliers.

Here we give only the important outline of the theorem. Assume that model contains n-n_0 positive observations. Then for one inlier model considered is

Proof: Consider M_1 and $P\left[R_1=r_1,X_{(r_1)}=X_{ir_1}\mid x_{ir_1},g\right]=\varphi\left(r_1\right)$. Then

$$\varphi\left(r_1\right)=\left(\frac{n-n_0-1}{r_1-1}\right)\int\left[F(x)\right]^{r_1-1}\left[1-F(x)\right]^{n-n_0-r_1}dG(x)$$

now

$$\Psi(x)=\frac{\partial G}{\partial F}=\frac{g(x)}{f(x)},$$

therefore,

$$\varphi(r_1)=\left(\frac{n-n_0-1}{r_1-1}\right)\int_0^\infty y^{r_1-1}[1-y]^{n-n_0-r_1}\Psi\left[F^{-1}(y)\right]dy$$

$$=\frac{1}{n-n_0}E\left[\Psi_1(y_r)\right] \tag{1.2.5}$$

Now y_r is a beta random variable with parameters $(r_1,n-n_0-r_1+1)$. Note that, $r_1=1,2,\ldots n-n_0$ is stochastically ordered sequence, since h is such that $\frac{h(y_r+1)}{h(y_r)}\,\alpha\,\frac{y}{1-y}$ which is strictly increasing function of y over (0,1). $\Psi\left[F^{-1}(y)\right]$ is decreasing function of y by as per our assumption. Therefore, from the result of Lehman (1959) it follows that $\varphi(1)>\varphi(2)>\ldots..>\varphi(n)$ and $X_{(1)}$ has maximum probability of being an inlier. Let $\varphi(r_1,r_2)=$Probability that $X_{(r_1)}$ and $X_{(r_2)}$ are inliers for $1\le r_1\le r_2\le n.$ for model M$_2$, where

$$\varphi(r_1,r_2)=\frac{(n-n_0-2)!2!}{(r_1-1)!(r_2-r_1-1)!(n-n_0-r_2)!}\int\int_{0<x<y<\infty}\left[F(x)\right]^{r_1-1}\left[F(y)-F(x)\right]^{r_2-r_1-1}$$
$$\left[1-F(y)\right]^{n-n_0-r_2}dG(x)dG(y)$$

$$=\frac{(n-n_0-2)!2!}{(n-n_0-r_2)!}\int_0^1\left[\int_0^v\frac{u^{r_1-1}(v-u)^{r_2-r_1-1}\Psi\left[F^{-1}(u)\right]}{(r_1-1)!(r_2-r_1-1)!}du\right][1-v]^{n-n_0-r_2}\Psi\left[F^{-1}(v)\right]dv$$

Then one can show $Max_{r_1<r_2}\varphi(r_1,r_2)=\varphi(1,2)$ and $x_{(1)}$ and $x_{(2)}$ have maximum probability of being inliers.

Generalizing the above result we can show that

$Max_{r_1\le r_2\le\ldots\ldots\le r_k\le n-n_0}\varphi(r_1,r_2,\ldots.r_k)=\varphi(1,2,\ldots\ldots k)$ and hence $x_{(1)},x_{(2)}\ldots\ldots x_{(k)}$

have the maximum probability of being inliers.

For other detailed proof of the theorem, one may refer to the paper by Muralidharan (2010). Thus the generalized form of $\varphi(\bullet)$ with k inliers is

$$\varphi(r_1,r_2,....r_k)=\frac{(n-n_0-k)!k!}{(r_1-1)!(r_2-r_1-1)!.....(n-n_0-k)}\int_{0<w_1<w_2<....w_k<1}\{w_1^{r_1-1}[w_2-w_2]^{r_2-r_1-1}......$$

$$....[1-w_k]^{n-n_0-n_k}\Psi\left[F^{-1}(w_1)\right]......\Psi\left[F^{-1}(w_k)\right]\}dw_1dw_2.....dw_k$$

Now fixing $(r_2,r_3,...,r_k)$ and $(w_2,w_3,...,w_k)$ we can show that $\phi(r_1,r_2,...,r_k)$ as decreasing function of r_1 for $1\le r_1\le r_2$.

Thus the model for M_k inlier is

$$L(x\,|\,g,f,\hat{v})=\prod_{i=1}^{k}g(x_{(i)})\prod_{i=k+1}^{n-n_0}f(x_{(i)}),\ f\in\mathfrak{I},g\in \mathrm{G}, \qquad (1.2.6)$$

But $L(x\,|\,g,f,\hat{v})$ is likelihood and not joint pdf of $x_{(1)},x_{(2)}......x_{(n-n_0)}$.

The model for L_k inlier is therefore

$$L(\underline{x}\,|\,g,f)=\frac{(n-n_0)!k!}{\phi_k(F,G)}\prod_{i=1}^{k}g(x_{(i)})\prod_{i=k+1}^{n-n_0}f(x_{(i)}),\ f\in\mathfrak{I},g\in \mathcal{G}, \qquad (1.2.7)$$

where $\varphi_k(F,G)=\varphi(r_1,r_2,...,r_k)$ is the normalizing constant to make L_k a pdf. The model is called as labeled slippage model and it can be derived as model from M_k with $(Y_1,Y_2,...,Y_k)$ are i.i.d.as $\mathcal{G}$, and $(V_1,V_2,...,V_{n-n_0})$as i.i.d from $\mathfrak{I}$ and with the additional condition $Max(Y_1,Y_2,...,Y_k)\le Min(V_1,V_2,...,V_{n-n_0})$.

1.3 Information criteria for inliers

The most important use of information criterion is, that it helps us in model selection, from the set of different models which all fit the data. These criterion are suitable when the underlying distribution and inlier distribution are available. It is an exploratory data analysis approach as no formal statistical inference is performed. The Akaike information criterion is a measure of the relative goodness of fit_of a statistical model. It was developed by Hirotsugu Akaike, under the name of "an information criterion" (*AIC*), and was first published by Akaike (1974). It is grounded in the concept of information entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality. It can be said to describe the tradeoff between bias and variance in model construction, or loosely speaking between accuracy and complexity of the model. In statistics, the Bayesian information criterion (*BIC*) or Schwarz criterion (also *SBC, SBIC*) is a criterion for model selection among a class of parametric models with different numbers of parameters. Choosing a model to optimize *BIC* is a form of regularization. When estimating model parameters using maximum likelihood estimation, it is possible to increase the likelihood by adding parameters, which may result in overfitting. The *BIC* resolves this problem by introducing a penalty term for the number of parameters in the model. This penalty is larger in the BIC than in the related *AIC*. The *BIC* was developed by Gideon E. Schwarz (1978), who gave a Bayesian argument for adopting it. It is closely related to the Akaike information criterion (*AIC*). In fact, Akaike was so impressed with Schwarz's Bayesian formalism that he developed his own Bayesian formalism, now often referred to as the *ABIC* for "a Bayesian Information Criterion" or more casually "Akaike's Bayesian Information Criterion". The BIC is an asymptotic result derived under the assumptions that the data distribution is in the exponential family.

Given any two estimated models, the model with the lower value of *BIC* is the one to be preferred. The *BIC* is an increasing function of σ_e^2 (variance) and an increasing function of *p*, where *p* is number of parameters of population under study. That is, unexplained variation in the dependent variable and the number of explanatory variables increase the value of *BIC*. Hence, lower BIC implies either fewer explanatory variables, better fit, or both. The *BIC* generally penalizes free parameters more strongly than does the Akaike information criterion, though it depends on the size of *n* and relative magnitude of *n* and *k*. It is important to keep in mind that the *BIC* can be used to compare estimated models only when the numerical values of the dependent variable are identical for all estimates being compared. The models being compared need not be nested, unlike the case when models are being compared using an *F* or likelihood ratio test.

The following information criteria are used in all the chapters. The Schwarz's Information criterion as given by $SIC = -2\ln L(\Theta) + p\ln n,$ Schwarz's Bayesian Information criterion as obtained by $BIC = -\ln L(\Theta) + \frac{0.5(p\ln n)}{n}$ and Hannan-Quinn criterion as given by $HQ = -\ln L(\Theta) + p\ln(\ln(n))$ to detect the inliers, where L(Θ) the maximum likelihood function and *p* is the number of free parameters that need to be estimated under the model.

Before discussing the tests of hypothesis we provide another theorem, again reproduced from Kale and Muralidharan (2007), which will help to understand the inlier distribution from among the other distributions. If *F* and *G* are respectively given by

$$F(x,\theta) = 1 - \exp(-x\theta), \quad x > 0, \theta > 0,$$

and

$$G(x,\phi) = 1 - \exp(-x\phi), \quad x > 0, \phi > 0 \text{ where } \phi = \lambda\theta,\ \lambda > 0$$

Using theorem (1.2.1), the labeled slippage alternative of $r \geq 1$ are discordant observation H_r, the joint distribution of the ordered statistics is given by

$$f\left(x_{(1)}, x_{(2)} \ldots . x_{(n)} \mid H_r\right) = \frac{(n-r)!r!\lambda^r}{\varphi(1,2,\ldots r)} \exp\left\{-\lambda\sum_{i=1}^{r} x_{(i)} - \sum_{i=k+1}^{n} x_{(i)}\right\} \qquad (1.3.1)$$

where the normalizing factor is given by

$$\varphi(1,2,\ldots r) = \frac{n-r}{\lambda} B\left[r+1, \frac{n-r}{\lambda}\right], \lambda > 0, r \geq 1.$$

Then we have the following theorem

Theorem 1.3.1: Under the labeled slippage alternative, $H_{r,}$ $\frac{x_{(1)}}{x_{(i)}} \xrightarrow{P} 0$ as $\lambda \to \infty$, for $i = r+1, r+2, \ldots..n$.

Proof: From (1.3.1) the joint density of $X_{(1)}$ and $X_{(r+1)}$ can be obtained as

$$f\left(x_{(1)}, x_{(r+1)}\right) = \frac{(n-r)r\lambda}{\varphi(1,2,\ldots..r)} e^{-\lambda x_{(1)}} \left[e^{-\lambda x_{(1)}} - e^{-\lambda x_{(r+1)}}\right]^{k-1} e^{-\left(n-rx_{(r+1)}\right)} \qquad (1.3.2)$$

where $0 \le x_{(1)} \le x_{(2)}$

$$= \frac{(n-r)re^{-\left(n-rx_{(r+1)}\right)}}{k\varphi(1,2,\ldots..r)} \left(e^{-\lambda x_{(1)}} - e^{-\lambda x_{(r+1)}}\right)^{k}$$

$$f\left(x_{(1)} \mid x_{(r+1)}\right) = \frac{\lambda e^{-\lambda x_{(1)}}}{k\varphi(1,2,\ldots..r)} \frac{\left(e^{-\lambda x_{(1)}} - e^{-\lambda x_{(r+1)}}\right)^{k-1}}{\left(e^{-\lambda x_{(2)}} - e^{-\lambda x_{(r+1)}}\right)^{k}}$$

hence for all $a \in (0,1)$, we get

$$P\left(\frac{X_{(1)}}{X_{(r+1)}} < a \mid H_r\right) =$$

$$\frac{(n-r)r\lambda}{\varphi(1,2,\ldots.r)} \sum_{i=0}^{r-1} \frac{(-1)^i \binom{r-1}{i}}{\left[\lambda(r-1-i)+(n-1)\right]\left\{\frac{1}{a}\left[\lambda(r-1-i)+(n-r)\right]+\lambda(i+1)\right\}}$$

. One gets $\frac{x_{(1)}}{x_{(r+1)}} \to 0$ as $\lambda \to \infty$. which proves the theorem given the condition $0 \le \frac{x_{(1)}}{x_{(i)}} \le \frac{x_{(1)}}{x_{(r+1)}} \Rightarrow X_{(r+1)} < X_{(i)}, i = r+2, \ldots.n$.

1.4 Testing of hypothesis

The main objective of this thesis is to detect (estimate) number of inliers in a given data. After detecting the number of inliers, using some model, we subjected the finding to test whether our results are true in light of a random sample. For which we have used various test to do this. Some Traditionally used tests are discussed below: In most of the test procedure, our main objective is to test the hypothesis:

$$H_0 : X_{(1)}, X_{(2)} \ldots X_{(n)} \text{ are from } F(x, \theta) \text{ and}$$

$$H_1 : X_{(1)}, X_{(2)} \ldots X_{(r)} \text{ are from } G(x, \phi) \text{ and } X_{(r+1)}, X_{(r+2)} \ldots X_{(n)} \text{ are from } F(x, \theta), \quad (1.4.1)$$

For a hypothesis of the form in equation (1.4.1) one can construct likelihood ratio test for testing inliers in the usual way. For example the underlying density is exponential, then the likelihood ratio test for one inlier by Kale and Muralidharan (2007) is obtained as

Reject H_0 if

$$\frac{T}{x_{(1)}} > c \, , \qquad (1.4.2)$$

where $T = \sum_{i=1}^{n} X_{(i)}$. And the value of $c = \dfrac{n}{1-(1-\alpha)^{\frac{1}{n-1}}} - 1.$

Also the power of the test for one inlier is given by

$$P_1(\lambda) = 1 - \left(\frac{c-n+1}{c+\lambda}\right) \quad \text{where} \quad \lambda = \frac{\theta}{\phi} \qquad (1.4.3)$$

Specifically if $X_1, X_2, \ldots\ldots X_n$ are independent and identically distributed r.v's having mixture distribution with likelihood is

$$L(x, \phi, \theta, p) = \prod \{(1-p) g(x_i) + p f(x_i)\} \qquad (1.4.4)$$

then the objective is to test

$$H_0 : p = 1 \quad \text{against} \quad H_1 : p < 1$$

for which we can have the following tests:

1.4.1 Locally most powerful test

The LMP test critical region for equation (1.4.4) is given by

$$\left[\underline{x} \mid \frac{\partial L(x,\phi,\theta,p)}{\partial p} \mid H_0 \right] \le C \tag{1.4.5}$$

where C is such that

$$P\left\{\left[\underline{x} \mid \frac{\partial L(x,\phi,\theta,p)}{\partial p} \mid H_0 \right] \le C\right\} = \alpha, \text{ the size of test.}$$

1.4.2 A Large sample test

A large sample test for the hypothesis (1.4.4) can be constructed using the asymptotic binomial distribution of the parameter of p: The large sample test for the hypothesis

$H_0 : p \ge p_0$ against $H_1 : p < p_0$, p_0 specified.

The test statistics is given by

$$Z_{cal} = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0 q_0}}, \quad q_0 = 1 - p_0 \tag{1.4.6}$$

and we reject H_0 if $Z_{cal} < Z_\alpha$ where α is level of significance. p denotes proportion of observations from target population.

1.5 Inlier estimation through Sequential Probability Ratio Test (SPRT)

Here first we want to test the hypothesis whether an observation belongs to inliers population with pdf $g(x,\phi)$ against hypothesis that it belongs to target population with pdf $f(x,\theta)$.

That is if $L_1 = \prod_{i=1}^{r} f(x_i,\theta)$ and $L_0 = \prod_{i=1}^{r} g(x_i,\phi)$ denote likelihood function under target and inlier population respectively, then the SPRT is the likelihood ratio λ_r is given by

$$\lambda_r = \frac{L_1}{L_0}$$

or equivalently

$$\ln \lambda_r = \sum_{i=1}^{r} \ln \frac{f\left(x_{(i)},\theta\right)}{g\left(x_{(i)},\phi\right)} = \sum_{i=1}^{r} z_{(i)} \qquad r = 1,2,....n \tag{1.5.1}$$

For deciding number of inliers r we continue to take additional observations till we reject H$_0$. That is

if $\sum_{i=1}^{r} z_{(i)} \leq \ln B$ accept H_0 and take the next observation.

and

if $\sum_{i=1}^{r} z_{(i)} \geq \ln A$ reject H_0 and stop.

The corresponding r represents the first observation from $f\left(x_{(i)},\theta\right)$ and the previous (r-1) observations from $g\left(x_{(i)},\phi\right)$. Thus the number of inliers will be r - 1.

1.6 Most powerful test for detection of inliers when underlying parameters are specified

If we are interested in testing $H_0 : g\left(x_{(i)},\phi\right)$ against $H_1 : f\left(x_{(i)},\theta\right)$ (i.e. whether data is from inlier population against data is from target population) a MPT can be constructed for ξ, the common parameter of interest, then the hypothesis can be equivalently written as

$$H_0 : \xi = \phi \text{ against } H_1 : \xi = \theta. \tag{1.6.1}$$

In the above frame, both H_0 and H_1 are simple and hence the most powerful test according to NP lemma is

$$\psi(x)=\begin{cases}1, & \dfrac{P_1(x)}{P_0(x)}>C\alpha\\ 0, & \dfrac{P_1(x)}{P_0(x)}<C\alpha\end{cases} \tag{1.6.2}$$

where the constant $C\alpha$ can be obtained using the size condition. For specific distributional model, the value of $C\alpha$ can be numerically computed.

In chapters to follow, we have studied many other test procedures and interesting properties of the models. For situation specific, we have changed the notation and theoretical development to establish proper continuity. We now provide the chapter wise summary of the thesis, in brief.

Chapter 1 gives a detailed introduction of the study and its need. An exhaustive literature survey on study of inliers is discussed. The utility and applicability of inlier distributions are also discussed in length and breadth. Various real life examples and their application areas are discussed in this chapter.

Chapter 2 discusses Pareto distribution as a inlier model for file sizes on the internet, insurance losses, and financial behavior of the stock market and in telecommunication systems. The proposed study is a further look at suitability of Pareto distributions in the context of life testing experiments where data involves instantaneous and early failures. We provide the inferences on parameters of modified forms of Pareto type distributions involving one and two parameters. The methods are illustrated on simulated data sets and on a real life data. We have discussed different criteria for detection of inliers and studied the sensitivity of various distributions with respect to different hypothesis. Through many other characteristics, we have shown that the Pareto distribution is much better than that of the Weibull distribution, in identifying inliers, and inlier models

In **Chapter 3,** we study the estimation of inliers in Normal distribution. The masking effect problem for correctly identifying the inliers is discussed with respect to various test procedures. Test for detecting a single inlier, H_o against H_1 is based on symmetric functions of observations or on functions of order statistics. In the *k*-inlier model, the joint distribution of order statistics $X_{(1)},X_{(2)},....X_{(n)}$ is same as that under the exchangeable model introduced by Kale (1998) where it is assumed that any set $X_{i_1},X_{i_2},....X_{i_k}$ has priori equal probability of being independent and identically distributed as G_λ and the remaining (*n-k*) observation are distributed as *F*, the distribution function of target population.

The study of inliers in Weibull models is the content of **chapter 4**. Apart from the regular estimation of inliers, we have also discussed the model specific estimation when the total realizations are assumed to be from either Model-1 or Model-2. If we assume, the data $\underline{X}$=(x_1, x_2,...... x_n) whose joint distribution is unknown, and if we have two competing models with parametric density $f_j(x;\theta_j,\alpha)$, $\theta_j \in \Theta_j$, where Θ_j is the Parametric space. Model-1 is selected with inliers and target population both having Weibull distribution with same shape parameter whereas other model-2 has Weibull with same scale parameter. We have also used predictive approach to model selection using exponential model. The SPRT test is conducted to detect number of inliers in both the models. Conditional test and Predictive method are also incorporated to detect inliers in exponential models.

In **chapter 5** we study the usefulness of mixture distribution and modified distribution for inlier study. Mixture distributions have been extensively used in a wide variety of important practical situations where data can be viewed as arising from two or more populations mixed in varying proportions. Mixture of distribution refers to the situation in which ith distribution out of k underlying distribution is chosen with probability p_i ,i=1,2,....*k*. Mixture distribution having *k*=2 components are extensively studied in literature. For example a probability model for the life of a electronic product can be described as the mixture of two unimodel distribution, one representing the life of inliers and other for target observations. We have listed down the methods which will be useful in detecting inliers present in the sample data. The graphs representing mixture of inliers and target populations, for exponential families are also plotted.

The inlier detection in generalized distribution is included in **chapter 6.** A generalized treatment for estimation and detection of inliers is discussed in this chapter. We also studied estimation of parameters of mixture distribution for particular cases. Apart from this we have derived the test for one inlier in the data set.

At the end we have given an exhaustive and extensive bibliography. As an output of the thesis, two articles have been published and couple of papers is on the way for publication. About three papers are ready for communication.