CHAPTER I

Introduction

1.1. Population data and outliers

Population data are a vital resource for many programmes of government and private sectors like public health, development of educational institutions etc. Trends in population estimates are important to facilitate analysis of particular disease and risk factors over time. Population projections are used by government departments in a wide range of ways, including assisting long-term fiscal and economic planning, to forecast future demands for services and to help devise strategies to deal with changing demographics. These data may misrepresent the facts if some of the entries were wrongly entered. Such entries may be called as outlying observations.

In this thesis, we intend to consider the problem of outliers in population data. An outlier is an observation that deviates in some sense from the rest of the observations. Outliers have strong influence on estimates of the parameters of a model that is being fitted to the data. Several literature on outliers like Grubbs (1969), Gumbel (1960), Kendall and Buckland (1957) and Gentleman and Wilk (1975) are available. Barnett and Lewis (1993), Kale (1979), Hawkins (1980), David and Nagaraja (2003) and Beckman and Cook (1983) have done excellent survey work in this field. It is clear that the inclusion of outlying observations in any analysis may yield quite erroneous conclusions. Moreover, at times, these observations may themselves be of interest (Joshi, 1969).

1.2 Effect of presence of outlying observations in the estimate of the parameters

In chapter II, effects of presence of outlying observations in the estimate of the parameters were considered. For this, the data used was that of Census 2011 of India. For making any inference about outliers in a sample, it is necessary to know the distribution of the population from which the observations were drawn. For identifying their distribution, the software ‘Easy Fit’ was used. In most of the cases of population data, it was found that Johnson S_B distribution to be the best fitting distribution, in some cases Pareto and in some other cases, Burr XII distributions was found to be the best fit. Hence, in this chapter the effect of presence of one outlying observation was studied for Johnson S_B distribution only.
The p.d.f of Johnson S_B distribution is given by

\[ f(x; \xi, \lambda, \gamma, \delta) = \frac{\delta}{\sqrt{2\pi}} \left( \frac{\lambda}{(x-\xi)} \right)^{\frac{\lambda}{2}} \exp \left\{ -\frac{1}{2} \left( y + \delta \ln \left( \frac{x-\xi}{\lambda-(x-\xi)} \right) \right)^2 \right\}. \]

\[ \xi \leq x \leq \xi + \lambda, \delta > 0, -\infty < \gamma < \infty, \lambda > 0, -\infty < \xi < \infty. \]

The estimates of the parameters of this distribution as given by George & Ramachandran (2011) using maximum likelihood least square method was considered are as follows.

\[ \hat{\gamma} = -\frac{\delta \sum_{i=1}^{n-1} \theta \left( \frac{x_i-\xi}{\lambda} \right)}{n-1} = -\delta \bar{g}. \]

\[ \delta^2 = \frac{\sum_{i=1}^{n-1} \left[ \theta \left( \frac{x_i-\xi}{\lambda} \right) \right]^2 - \frac{1}{n-1} \left[ \sum_{i=1}^{n-1} \theta \left( \frac{x_i-\xi}{\lambda} \right) \right]^2}{\text{var}(g)} = \frac{1}{n-1} \text{var}(\theta). \]

where, \( g \left( \frac{x-\xi}{\lambda} \right) = \log \left( \frac{x-\xi}{\lambda-(x-\xi)} \right) \), \( \bar{g} \) is the mean and \( \text{var}(g) \) is the variance of the values of \( g \) defined here.

The estimates of \( \lambda \) and \( \xi \) were as follows:

\[ \hat{\lambda} = \frac{(n-1) \sum_{i=1}^{n-1} x_i f^{-1} \left( \frac{z_i-\gamma}{\delta} \right) - \sum_{i=1}^{n-1} f^{-1} \left( \frac{z_i-\gamma}{\delta} \right) \sum_{i=1}^{n-1} x_i}{(n-1) \sum_{i=1}^{n-1} f^{-1} \left( \frac{z_i-\gamma}{\delta} \right)^2 - \left[ \sum_{i=1}^{n-1} f^{-1} \left( \frac{z_i-\gamma}{\delta} \right) \right]^2}. \]

\[ \hat{\xi} = \bar{x} - \lambda \cdot \text{mean} \left[ f^{-1} \left( \frac{z-\gamma}{\delta} \right) \right], \]

where \( z = \frac{x-\xi}{\lambda} \) is a standard normal variate. Hence, the quantiles of \( x \) and the corresponding quantiles of \( z \) can be considered as paired observations. When there were 100 or more \( x \) values, the percentiles 1 through 99 were considered, while for \( k \) number of data points of \( x \), where \( k \) is less than 100, \( k - 1 \) quantiles of \( x \) were considered. These \( k - 1 \) quantiles of \( x \) and the corresponding \( k - 1 \) quantiles of \( z \) were considered as paired observations.

Then the largest observation was shifted to the upper side with a shifted location parameter by an amount \( a \) and its effect on estimates of location, scale and on both the shape parameters were studied. It was found that there was no effect of shift in the estimate of the location parameter. Also the percentage variations were obtained for the estimates of scale and both the shape parameters due to a shift \( 'a' \) in the largest observation. These are given by
\[
\lambda_{LP} = \frac{[\bar{x} - \bar{z}]}{\lambda} \times 100, \gamma_{LP} = \frac{|\bar{y} - \bar{p}|}{\bar{p}} \times 100 \text{ and } \delta_{LP} = \frac{|\bar{\delta} - \bar{\delta}|}{\bar{\delta}} \times 100 \text{ respectively.}
\]

Later, the smallest observation was shifted to the lower side with a shifted location parameter by an amount \(b\) and its effects on estimates of location, scale and on both the shape parameters were studied. It was found that there was no effect of shift in the estimate of the location parameter. Also the percentage variations were obtained for the estimates of scale and both the shape parameters due to a shift ‘\(b\)’ in the largest observation. These are given by
\[
\lambda_{SP} = \frac{[\bar{x} - \bar{z}]}{\lambda} \times 100, \gamma_{SP} = \frac{|\bar{y} - \bar{p}|}{\bar{p}} \times 100 \text{ and } \delta_{SP} = \frac{|\bar{\delta} - \bar{\delta}|}{\bar{\delta}} \times 100 \text{ respectively.}
\]

Similarly, the effect of shift in the scale parameter of the largest observation by an amount \(c\) was studied. Again it was found that there is no effect of shift in the estimate of the location parameter. The percentage variations so obtained are given by
\[
\tilde{\lambda}_{LP} = \frac{[\bar{x} - \bar{z}]}{\lambda} \times 100, \tilde{\gamma}_{LP} = \frac{|\bar{y} - \bar{p}|}{\bar{p}} \times 100 \text{ and } \tilde{\delta}_{LP} = \frac{|\bar{\delta} - \bar{\delta}|}{\bar{\delta}} \times 100 \text{ respectively.}
\]

Thereafter the effects of the shift in the scale of the smallest observation by amount \(d\) were observed. Again it was found that there is no effect of shift in the estimate of the location parameter. The percentage variations so obtained are
\[
\tilde{\lambda}_{SP} = \frac{[\bar{x} - \bar{z}]}{\lambda} \times 100, \tilde{\gamma}_{SP} = \frac{|\bar{y} - \bar{p}|}{\bar{p}} \times 100 \text{ and } \tilde{\delta}_{SP} = \frac{|\bar{\delta} - \bar{\delta}|}{\bar{\delta}} \times 100 \text{ respectively.}
\]

Examples were discussed for showing the working procedure of the formulae so obtained.

It was concluded that even a large variations in sample size does not have much effect in the variation of the estimates of any of the parameters of Johnson SB distribution, while any shift in the parameters certainly lead to variation in the estimate of the parameters. It was also observed that the percentage variation of the estimate of \(\lambda\) is almost a constant for any variation in any of the parameters.

1.3 Outlier detection procedures for a sample from a Johnson SB distribution when parameters are known

In chapter III, the detection of an upper, lower, a pair and several outlying observations, were studied with a sample from Johnson SB distribution with known parameters.

For derivation of a test statistic for detection of an outlier, Johnson SB distribution is firstly transformed into Standard normal distribution using some suitable transformation. Then some
test statistics for a sample from a Normal distribution as given in Barnett and Lewis (1994) were used. From these statistics, the corresponding statistics for identifying the outlying observations in a sample from Johnson $S_B$ distribution were derived. The following test statistics,

\[
W_U = \left( \frac{x(n) - \xi}{x(n-1) - \xi} \right) \left( \frac{\lambda - (x(n-1) - \xi)}{\lambda - (x(n) - \xi)} \right)^\delta,
\]

\[
W_L = \left( \frac{x(2) - \xi}{x(1) - \xi} \right) \left( \frac{\lambda - (x(1) - \xi)}{\lambda - (x(2) - \xi)} \right)^\delta,
\]

\[
W_{LU} = \left( \frac{x(n) - \xi}{x(1) - \xi} \right) \left( \frac{\lambda - (x(1) - \xi)}{\lambda - (x(n) - \xi)} \right)^\delta,
\]

\[
W_P = \left( \frac{x(n-k+1) - \xi}{x(k) - \xi} \right) \left( \frac{\lambda - (x(k) - \xi)}{\lambda - (x(n-k+1) - \xi)} \right)^\delta,
\]

were proposed for the detection of an upper, lower, a lower and upper pair and for $k$-lower and $k$-upper pairs of outlying observations respectively.

These statistics were applied to a real life data with a planted outlying observation and it was found that all the proposed statistics were efficient enough in identifying the planted observation. Performance studies of all the statistics were also discussed for the slippage alternatives of all the parameters.

It was concluded that the test statistics suggested for detection of an upper, a lower, a lower and upper pair and for $k$-lower and $k$-upper pairs outlying observations in a sample from Johnson $S_B$ distribution with known parameters, was performing very efficiently.

### 1.4 Outlier detection procedures for a sample from a Johnson $S_B$ distribution when parameters are unknown

In chapter IV, outlier detection procedures of an upper, lower, a pair and several outlying observations in a sample from a Johnson $S_B$ distribution with unknown parameters were considered. Four test statistics $W_U^\prime$, $W_L^\prime$, $W_{LU}^\prime$ and $W_P^\prime$ were suggested for detection of an upper, lower, a lower and upper pair and for $k$-lower and $k$-upper pairs outlying observations respectively. Hence, for estimation of the unknown parameters, the MLE least square estimates suggested by George & Ramachandran (2011) were used with a trimmed sample obtained after removing the suspected outlying observation(s).
The test statistics for the case of an upper, lower, lower and upper pair and for \( k\)- lower and \( k\)-upper pair are obtained respectively are as follows.

\[
W_U' = \left\{ \frac{x_{(n)} - \xi}{x_{(n-1)} - \xi} \right\} \left\{ \frac{\lambda - (x_{(n-1)} - \xi)}{\lambda - x_{(n)} - \xi} \right\}^\delta,
\]

\[
W_L' = \left\{ \frac{x_{(2)} - \xi}{x_{(1)} - \xi} \right\} \left\{ \frac{\lambda - (x_{(1)} - \xi)}{\lambda - (x_{(2)} - \xi)} \right\}^\delta,
\]

\[
W_{LU}' = \left\{ \frac{x_{(n)} - \xi}{x_{(1)} - \xi} \right\} \left\{ \frac{\lambda - (x_{(1)} - \xi)}{\lambda - (x_{(n)} - \xi)} \right\}^\delta,
\]

\[
W_P' = \left\{ \frac{x_{(n-k+1)} - \xi}{x_{(k)} - \xi} \right\} \left\{ \frac{\lambda - (x_{(k)} - \xi)}{\lambda - (x_{(n-k+1)} - \xi)} \right\}^\delta,
\]

where \( X_{(k)} \) is the \( k^{th} \) order statistic, \( k = 1, 2, \cdots, n \) of a sample of size \( n \). Simulation technique was used here for obtaining the critical values of these test statistics, as the derivation of the density function of these statistics were very complicated. On comparison of these critical values with that of the one obtained for the case of a sample from a Johnson SB distribution with known parameters, it can be seen that the values are very close to each other. These statistics were applied to a census data with a planted contaminant observation and it was found that all the proposed statistics were identifying the planted observation. Performance studies of all the statistics were also made using simulation technique for the slippage alternatives of all the parameters.

It was concluded that the suggested test statistic for an upper, a lower and for a lower and upper pair are performing very well for higher values of the shift, whereas for the lower values of the shift, performance of the statistic are not satisfactory. While for \( k\)-lower and \( k\)-upper pairs the performance of the test statistic is well for all the values of shift. Thus the suggested test statistic for \( k\)-lower and \( k\)-upper pairs can be used effectively for the detection of outlying observations for a sample from Johnson SB distribution with unknown parameters.

### 1.5 Detection of an outlier in a sample from Pareto and Burr XII distributions with known parameters

In chapter V, detection procedures for a single outlier, either in the upper side or in the lower side of a sample from a Pareto distribution and Burr XII distribution have been discussed.
In this chapter, two test statistics were developed for detection of an upper outlier as well as of a lower ones using a result that states that the distribution of the ratio of two consecutive order statistics, \( r^{th} \) and \( (r+1)^{th} \) of a sample of size \( n \) from a Pareto distribution with scale parameter \( \mu \) and shape parameter \( \gamma \) will also follow a Pareto distribution with scale parameter 1 and shape parameter \( (n-r)\gamma \).

For detection of an upper outlier, the test statistic \( Z_U = T_U^\gamma \), is proposed, where \( T_U = \left( \frac{y_n}{y_{(n-1)}} \right) \).

The distribution of \( Z_U \) was obtained and its critical values are obtained.

In the same way, the test statistic \( Z_L = T_L^{(n-1)\gamma} \), is proposed, where \( T_L = \left( \frac{y_{(1)}}{y_{(2)}} \right) \) for detection of a lower outlier. The critical values of \( Z_L \) turned out to be same as that of \( Z_U \).

For highlighting the proposed utility of the procedure in both the cases, the statistics were applied to certain real life data. Performance study was carried out in both the cases using simulation technique.

It was concluded that the test statistics suggested for detection of an upper outlier in a sample from Pareto distribution with known scale and shape parameters, were performing efficiently. In the case of detection of lower outlier, shifting the scale parameter does not have much effect on the observation, due to which this test statistic may result in masking effect.

For developing a test statistic for detection of an outlier in a sample from a Burr XII distribution, transformation of Burr XII density to Pareto density with a suitable transformation was considered. Then using the test statistic proposed for the Pareto distribution, a statistic for identifying the outlying observations in a sample from Burr XII distribution was derived by transforming the test statistic used in Pareto distribution to that of Burr XII variable. Thus the following test statistics,

\[
Z_U^* = \left\{ \frac{\beta + (x_{(n)})^\alpha}{\beta + (x_{(n-1)})^\alpha} \right\}^\gamma .
\]

\[
Z_L^* = \left\{ \frac{\beta + (x_{(2)})^\alpha}{\beta + (x_{(1)})^\alpha} \right\}^{(n-1)\gamma} .
\]

were proposed for the detection of an upper and a lower outlying observations respectively, in a sample from Burr XII distribution with known parameters. The critical values of \( Z_U^* \) and
$Z_L^*$ turned out to be same as that of $Z_U$, the proposed statistic in the case of a sample from a Pareto distribution with known scale and shape parameters.

A sample from Census 2011 of India was considered for highlighting the utility of the proposed results. The performance study was also studied using simulation technique for the slippage alternative of scale and for both the shape parameters.

It was concluded that, the performance of the suggested test statistic for detection of an upper outlying observation was quite good, while the performance of the test statistic suggested for detection of a lower outlying observation was very poor.

1.6 Detection of an outlier in a sample from Pareto and Burr XII distributions with unknown parameters

In chapter VI the Pareto and Burr XII distribution with unknown parameters were considered. The detection of an upper and a lower outlying observation were discussed. For the case of Pareto distribution two test statistics $Z_U'$ and $Z_L'$ have been suggested to detect an upper and a lower outlying observations respectively to test the null hypothesis. Hence, for the estimation of the unknown parameters, the moment estimates suggested by Quandt (1966) were used, using a trimmed sample obtained after removing the suspected outlying observation(s). The estimator of shape parameter $\gamma$ is as follows.

$$\hat{\gamma}^* = \frac{n \bar{y} - y_1'}{n \bar{y} - \bar{y}'}$$

where, $\bar{y}$ is the sample mean and $y_1'$ is the smallest sample value.

The test statistics for the case of detection of an upper and a lower outlier were obtained respectively are as follows.

$$Z_U' = T_U^{\hat{\gamma}^*}, \text{ where } T_U = \left( \frac{y(n)}{Y(n-1)} \right);$$

$$Z_L' = T_L^{\hat{\gamma}^*}, \text{ where } T_L = \left( \frac{Y(2)}{Y(1)} \right)$$

where, $Y(k)$ is the $k^{th}$ order statistic, $k = 1, 2, \cdots, n$ of a sample of size $n$.

The simulation technique was used here for obtaining the critical values, as the derivation of the density function was very complicated. On comparison of these critical values with that of the one with known parameters, it can be seen that the values are not close to each other,
because the test statistic as well as critical values both depends upon sample size $n$ in this case.

Examples are also given for highlighting the utility of the proposed results. Also the performances of all the cases were studied using simulation technique.

It was concluded that the test statistics suggested for detection of an upper outlier in a sample from Pareto distribution with unknown parameters, are performing efficiently. While in the case of detection of lower outlier, performance of the test statistic is not good enough.

For the case of Burr XII distribution two test statistics $Z_{U}^{**}$ and $Z_{L}^{**}$ have been suggested to detect an upper and a lower outlying observations respectively to test the null hypothesis. In this case, both the shape parameters are assumed to be unknown and scale parameter to be known. Hence, for the estimation of the unknown parameters, maximum likelihood estimators of censored and uncensored data of Wang, Keats and Zimmer (1995) were used, using a trimmed sample obtained after removing the suspected outlying observation(s). For the case of upper outlier the estimators of all the parameters are as follows

\[
\hat{\gamma}_U' = \frac{n-1}{\log \sum_{i=1}^{n-1} \left( \frac{x(i)}{\rho} \right)^{\alpha} + \log (x(n-1)^{a} + 1)}
\]

The likelihood equations for estimating $\alpha$ and $\beta$ were obtained as follow.

\[
\frac{n-1}{\alpha} + \log \sum_{i=1}^{n-1} \left( \frac{x(i)}{\beta} \right)^{\gamma} - (\gamma + 1) \sum_{i=1}^{n-1} \left( \frac{x(i)}{\beta} \right)^{\gamma} + 1 - \gamma \left( \frac{x(n-1)}{\beta} \right)^{\gamma} + 1 = 0.
\]

\[
\Rightarrow - (\alpha - 1) + (\gamma + 1) \alpha \frac{\sum_{i=1}^{n-1} x(i)^{a}}{\sum_{i=1}^{n-1} \left[ (x(i)^{a} + \beta a \right]} + \gamma \alpha \frac{x(n-1)^{a}}{(x(n-1)^{a} + \beta a} = 0.
\]

In order to solve the above equations, Newton Raphson iterative procedure was employed by assuming $\beta$ to be known and solving for the unknown parameter $\alpha$, which give the estimates of the shape parameter $\hat{\alpha}_U'$ of $\alpha$.

The test statistic for the case of detection of an upper outlier is given as

\[
Z_{U}^{**} = \left\{ \frac{\hat{\beta}_U' \hat{\alpha}_U' + (x(n)) \hat{\alpha}_U'}{\hat{\beta}_U' \hat{\alpha}_U' + (x(n-1)) \hat{\alpha}_U'} \right\},
\]

where, $x(n-1)$ and $x(n)$ denotes the largest two observations of the sample.
For the case of lower outlier, the estimators of all the parameters are as follows

\[ \hat{\gamma}_L = \frac{n-1}{\log \sum_{i=1}^{n} \left( \frac{x_i}{\beta} \right)^{\alpha} + 1} \]

The likelihood equations for estimating \( \alpha \) and \( \beta \) were obtained as follow.

\[ \frac{n-1}{\alpha} + \log \sum_{i=2}^{n} \left( \frac{x_i}{\beta} \right)^{\alpha} - (\gamma + 1) \sum_{i=2}^{n} \left( \frac{x_i}{\beta} \right)^{\alpha} + 1 + \gamma \left( \frac{x_2}{\beta} \right)^{\alpha} \log \left( \frac{x_2}{\beta} \right)^{\alpha} + 1 = 0. \]

\[ \Rightarrow -(\alpha - 1) + (\gamma + 1) \alpha \frac{\sum_{i=2}^{n} x_i^{\alpha}}{\left( \frac{x_1}{\beta} \right)^{\alpha} + 1} - \gamma \alpha \frac{x_2^{\alpha}}{\left( \frac{x_2}{\beta} \right)^{\alpha} + 1} = 0. \]

In order to solve the above equations, Newton Raphson iterative procedure was employed by assuming scale parameter \( \beta \) to be known and solving for the other unknown parameter \( \alpha \), which give the estimates of the shape parameter \( \hat{\alpha}_U \) of \( \alpha \).

The test statistic for the case of detection of a lower outlier is given as

\[ Z_L^{**} = \left\{ \frac{\beta_L \hat{\alpha}_L + x_2^{\alpha}}{\beta_L \hat{\alpha}_L + x_1^{\alpha}} \right\} \hat{\gamma}_L \]

where, \( x_1 \) and \( x_2 \) denotes the smallest two observations of the sample.

The simulation technique was used here for obtaining the critical values, as the derivation of the density function was very tedious. It can be observed from the critical value tables that, the variation between different critical values exist only in the second decimal place onwards.

Also it can be seen that the critical values increase as the sample size increases.

A sample from Census 2011 of India was considered for highlighting the utility of the proposed results. The performance study was also studied using simulation technique.

It can be concluded from the above discussions, that the two test statistics suggested for the detection of an upper and for a lower outlier respectively, were performing efficiently. Thus the use of the test statistic \( Z_U^{**} \) and \( Z_L^{**} \) for detection of an upper and lower outlying observation respectively, can be recommended for the Burr XII distribution with unknown shape parameters and known scale parameter.