On the Equivalence of Tests for Outliers for Pareto and Exponential Distributions

Fang Wang

April 1, 2020

Abstract

This thesis discusses the outlier detection problem with slippage alternative hypothesis for exponential and Pareto distributions. We show that statistics used for testing slippage alternative hypothesis of exponential samples can be easily adapted to test for the Pareto case with the same critical values and power under certain regularity conditions. The exact distributions of some specific statistics are derived, and their performance are examined through a Monte Carlo simulation study. We observed that the classical Dixon statistic has the best performance in terms of power among all test statistics examined.

Acknowledgements

First, I would like to thank my supervisor, Prof. Narayanaswamy Balakrishnan, for his supportive encouragement and patient guidance throughout my time being his student. I really appreciate his crystal clear explanations and insightful comments he gave to me; without those, this thesis could not have been finished.

I would also like to thank all faculty members and staff in our department for their help in my life and study. In particular, I want to thank Prof. Angelo Canty for teaching me many practical skills and useful advices on how to conduct research; my previous work experience with him greatly helped me during this work.

Finally, my thanks go to my family for their love and encouragement throughout the years.

Contents

Chapter 1

Introduction

1.1 Outliers and Outlier Detection Problem

The outliers, in a sample of observations, is a subset of observations that appears to be inconsistent with the rest of the data and the assumption proposed on the dataset [1]. Outliers can arise in many ways, varing from a human recording error to the the natural inherent variability of the data.

The problem of outlier detection is detecting the outliers from a sample of observations and these have found many practical applications such as fraud detection [2] and genetics [3].

Whether an observation is an outlier depends on one's subjective prior knowledge on the data the observation is coming from; hence, the definition of an outlier is subjective. When an implausible observation arises, one may conclude a misrecording happened and discard it, but such an observation may convey some important information about underline population; for example, it may suggests a contamination of the data; that is, some of the observations do not come from the distribution that the data are assumed to have come from, and this motivates the contamination model.

1.1.1 Contamination Model and Slippage Alternative

Suppose we have observations $(x_1 \ldots, x_n)$ which are assumed to be independent and identically distributed (iid) samples of $(X_1, \ldots, X_n)$ and each X_j has distribution F under the null hypothesis H_0. If some of the observations $\{x_1, \ldots, x_r\}$ are unlikely coming from F, then one can conclude that H_0 is false, but there is more than one way to construct the alternative hypothesis $\overline{H}$. We now formally define the null hypothesis as follows:

Definition 1.1 (null hypothesis of contamination model)**.** Let $x_1, \ldots, x_n$ be a sample of n observations. Then under the null hypothesis H_0, $x_1, \ldots, x_n$ are observations of $X_1, \ldots, X_n$, where $X_1, \ldots, X_n$ are independent random variables with common distribution F.

One possible $\overline{H}$ is $\overline{H} : X_j \sim G$ for all X_j, where G is a distribution different from F and this is known as the inherent alternative. One classical example of inherent alternative is the hypothesis testing for one population mean, where F and G are usually same family of distributions but with different expectation.

One can also formalize $\overline{H}$ as the mixture distribution, so that $\overline{H}$ becomes $X_j \sim (1-\lambda)F + \lambda G$ where λ is a small positive real number. One should be cautious about the cyclic argument of the λ being small, since λ being small justifies the outliers are rare but rare outliers also implies λ being small.

In this study, we will consider the alternative hypothesis known as the slippage alliterative constructed as follows: We assume that, expect for a small fixed number r of observations, all arose from F, but theses r observations arose from a modified version of F, say $\overline{F}$. In this study, we will focus on the cases where F and $\overline{F}$ are Pareto and exponential distributions with different parameters, and we will use H_r to denote the slippage alternative hypothesis, where r of the n observations came from $\overline{F}$. We will restrict ourselves to the univariate case, and we can then formally define the slippage alternative hypothesis as follows:

Definition 1.2 (slippage alternative of the contamination model)**.** Let $x_1, \ldots, x_n$ be a sample of n observations with null hypothesis $x_1, \ldots, x_n$ that they are independently from a distribution F. Let $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ be the order

statistics of $x_1,\ldots,x_n$. Then under the slippage alternative H_r, the sample $x_{(1)},\ldots x_{(n-r)}$ are independent observations from distribution F and $x_{(n-r+1)},\ldots,x_{(n)}$ are independent observations from distribution $\overline{F}$ with $F \neq \overline{F}$.

As an illustrative example, we generated six sets of random samples of size 15, where for each random sample, 10 of them came from exponential distribution with mean 1 while other 5 are contaminated coming from exponential distribution with mean 2; the discussion on the sample generation method are postponed to the Section 2.3. The random samples generated are depicted in Figure 1.1, where three sets of samples generated with slippage alternative assumption are plotted in panel A and other three sets of samples without this assumption are given in the panel B.

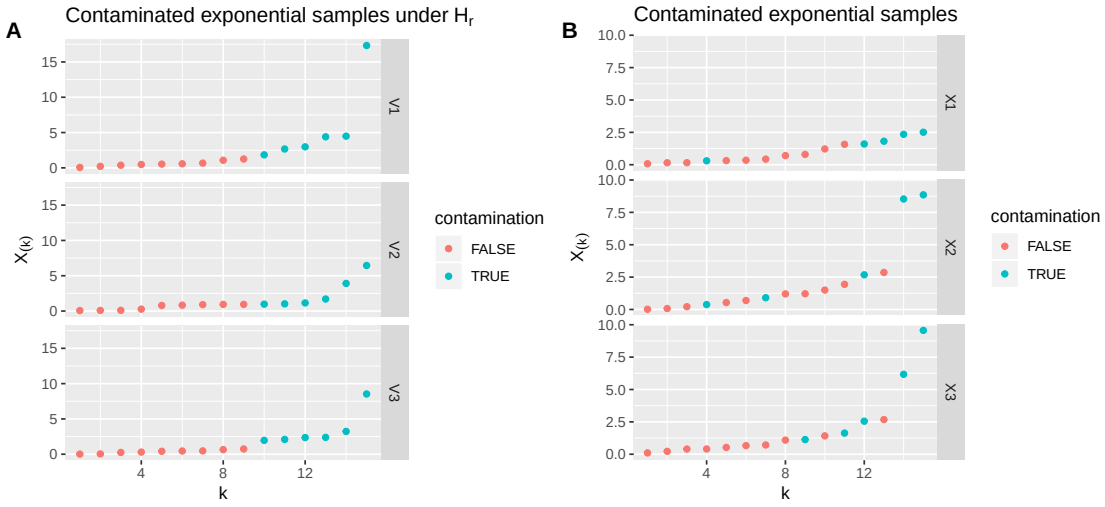

Figure 1.1: Contaminated exponential random samples with $n = 15, r = 5, \mathbb{E}[F] = 1, \mathbb{E}\left[\overline{F}\right] = 2$

A Three random samples of exponential observations under slippage alternative

B Three random samples of exponential observations with contamination but does not have assumption of slippage alternative.

Hereafter, we will use capital letter $X_1,\ldots,X_n$ to denote the random variables, $X_{(1)},\ldots.X_{(n)}$ to denote the order statistics of $X_1,\ldots,X_n$, lower case letter $x_1,\ldots,x_n$ to denote the sample of $X_1,\ldots X_n$ and lower case $\mathbf{x} = (x_1,\ldots,x_n)$ denoted the vector of observations.

1.1.2 Discordancy Test

The hypothesis testing procedures that test for outliers are known as discordancy tests. There are two major tests against H_0 defined in Definition 1.1, and they are sequential tests and block tests [4]. A sequential test for the outlier detection detects the outliers by applying the statistic test repeatedly, starting with the most extreme sample observation; if we reject H_0, then apply the test to the sample with the most extreme sample removed, and so the procedure is stopped once we fail to reject H_0 [5].

The block test is simpler, where the number of possible contamination sample r is specified and the hypothesis testing is done with a single test. The block tests suffer from the problem of swamping and masking, when the number of hypothesized outliers is different from the actual number of outliers, but this is not a problem under slippage alternative. Because block tests are easy to interpret, we restrict ourselves to the case of slippage alternative in our study.

1.2 Organization of the Thesis

The rest of this report is organized as follows: A review of mathematical results that will be used to establish the distribution of test statistics proposed by Zerbet and Nikulin [6] and Jabbari Nooghabi [7] and an introduction of simulation methods used is given in Chapter 2. The exact distributions of test statistics given by Jabbari Nooghabi

and Zerbet and Nikulin, under H_r, are given in Chapter 3. Results of a simulation study and the conclusions drawn are presented in Chapter 4.

Chapter 2

Methods

In this thesis we derive the distribution of Z_r given by Zerbet and Nikulin [6] and $\tilde{R}_r$ given by Jabbari Nooghabi [7] under H_r, forwhich we need some key distributional results that are presented in Section 2.1.

Then, we characterize the distribution of order statistics after monotone transformation assuming some regularity conditions in Section 2.2. In particular, Corollary 2.3 plays a key role in deriving the correct distribution of test statistic $\tilde{Z}_r$ introduced in [7], and justify the equivalence of some test statistics we investigate.

To verify the theoretical results derived and also to observe some other characteristic of the test statistics we study here, we conduct the simulation studies, and the algorithms and some other facts are summarized in Section 2.3.

2.1 Mathematical Results

The follwing lemma will be used in the derivation of the distribution of tests Z_r and R_r to simplify and rewrite the expressions.

Lemma 2.1. *Let $\{a_k\}_{k=1}^m$ be a finite number of distinct real number and $x \in \mathbb{C}$. Then*

$$\prod_{k=1}^{m} \frac{1}{x-a_k} = \sum_{n=1}^{m} \frac{1}{x-a_n} \prod_{i=1, i\neq n}^{m} \frac{1}{a_n - a_i}.$$

Proof. We will prove the result by induction on k. The case for $k=1$ clearly holds. Case for $k=2$ also holds, since

$$\frac{1}{(x-a_1)(x-a_2)} = \frac{1}{(x-a_1)(a_1-a_2)} + \frac{1}{(x-a_2)(a_2-a_1)}.$$

Now assume the case holds for $k = m-1$ and consider the case for $k = m$. Because

$$\begin{aligned}
\prod_{k=1}^{m} \frac{1}{x-a_k} &= \frac{1}{x-a_m} \prod_{k=1}^{m-1} \frac{1}{x-a_k} \\
&= \frac{1}{x-a_m} \sum_{n=1}^{m-1} \frac{1}{x-a_n} \prod_{i=1, i\neq n}^{m-1} \frac{1}{a_n-a_i} && \text{(by induction hypothesis)} \\
&= \sum_{n=1}^{m-1} \left(\frac{1}{a_n-a_m} \frac{1}{x-a_n} + \frac{1}{x-a_m} \frac{1}{a_m-a_n} \right) \prod_{i=1, i\neq n}^{m-1} \frac{1}{a_n-a_i} \\
&= \left[\sum_{n=1}^{m-1} \frac{1}{x-a_n} \prod_{i=1, i\neq n}^{m} \frac{1}{a_n-a_i} \right] + \sum_{n=1}^{m-1} \frac{1}{x-a_m} \frac{1}{a_m-a_n} \prod_{i=1, i\neq n}^{m-1} \frac{1}{a_n-a_i} \\
&= \left[\sum_{n=1}^{m-1} \frac{1}{x-a_n} \prod_{i=1, i\neq n}^{m} \frac{1}{a_n-a_i} \right] + \frac{1}{x-a_m} \sum_{n=1}^{m-1} \frac{1}{a_m-a_n} \prod_{i=1, i\neq n}^{m-1} \frac{1}{a_n-a_i},
\end{aligned}$$

where by induction hypothesis with $x = a_m$ that

$$\sum_{n=1}^{m-1} \frac{1}{a_m - a_n} \prod_{i=1, i\neq n}^{m-1} \frac{1}{a_n - a_i} = \prod_{k=1}^{m-1} \frac{1}{a_m - a_k} = \prod_{i=1, i\neq m}^{m} \frac{1}{a_m - a_i},$$

and therefore

$$\begin{aligned}\prod_{k=1}^{m} \frac{1}{x - a_k} &= \left[\sum_{n=1}^{m-1} \frac{1}{x - a_n} \prod_{i=1, i\neq n}^{m} \frac{1}{a_n - a_i}\right] + \frac{1}{x - a_m} \prod_{i=1, i\neq m}^{m} \frac{1}{a_m - a_i} \\ &= \sum_{n=1}^{m} \frac{1}{x - a_n} \prod_{i=1, i\neq n}^{m} \frac{1}{a_n - a_i}.\end{aligned}$$

This proves the lemma. □

The major tool used for deriving the distribution of Z_r is characteristic function and the following result allows one to obtain the pdf of the distribution of a continuous random variable from its characteristic function.

Lemma 2.2 (Inversion Formula [8])**.** *Let X be a continuous random variable with characteristic function $f(x)$ such that*

$$\int_{-\infty}^{\infty} |f(x)|\,\mathrm{d}x < \infty.$$

Then, the pdf of X, $p(x)$, is given by

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} f(t)\,\mathrm{d}t.$$

Lemma 2.3. *Let b and θ be real numbers and $\theta \neq 0$. Then,*

$$\int_0^{\infty} \frac{e^{-iwz}}{(b/\theta - iz)^r}\,\mathrm{d}z = \frac{2\pi w^{r-1}}{(r-1)!} e^{-wb/\theta}.$$

Proof. Let X be a random variable with Gamma$(r, \theta/b)$ distribution. Then, the pdf $p(x)$ and characteristic function $f(x)$ of X are given by

$$p(x) = \frac{w^{r-1} e^{-w/b}}{\Gamma(r)(\theta/b)^r}, \quad x > 0, \qquad \text{and} \qquad f(x) = (1 - ix\theta/b)^{-r}.$$

Then, it follows from Lemma 2.2 that

$$\begin{aligned}
& & 2\pi p(w) &= \int_{-\infty}^{\infty} e^{-izw} f(t)\,\mathrm{d}z \\
\iff & & 2\pi \frac{w^{r-1} e^{-w\theta/b}}{(\theta/b)^r (r-1)!} &= \int_0^{\infty} \frac{e^{-iwz}}{(1 - iz\theta/b)^r}\,\mathrm{d}z \\
\iff & & 2\pi \frac{w^{r-1} e^{-w\theta/b}}{(r-1)!} &= \int_0^{\infty} \frac{e^{-iwz}}{(b/\theta)^r (1 - iz\theta/b)^r}\,\mathrm{d}z \\
\iff & & 2\pi \frac{w^{r-1} e^{-w\theta/b}}{(r-1)!} &= \int_0^{\infty} \frac{e^{-iwz}}{(b/\theta - iz)^r}\,\mathrm{d}z
\end{aligned}$$

which proves Lemma 2.3. □

The following lemma, was introduced by Chikkagoudar *et al.* [9], allows one to transform the problem of finding the distribution of order statistics into finding the distribution of independent exponential random variables. Since the characteristic function of sum of random variables is the product of the characteristic function of each summand only if the random variables summed up are independent, the following lemma is of critical importance. Prior to introducing the lemma, we first fix the notation and definition of the exponential distribution in the following definition.

Definition 2.1 (Exponential distribution)**.** A random variable X follows exponential distribution with mean parameter $\theta > 0$ if it has pdf of

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0$$

and we denoted it by $X \sim \text{Exp}(\theta)$.

Lemma 2.4. *Let $X_1, \ldots, X_n$ be independent $\text{Exp}(\theta)$ random variables, $X_{(1)}, \ldots X_{(n)}$ be the order statistics of $X_1, \ldots, X_n$, and $Y_j = X_{(j)} - X_{(j-1)}$. Assuming the slippage alternative H_r, we then have*

$$\begin{aligned} &X_{(1)}, X_{(2)}, \ldots, X_{(n-r)} \textit{ derived from } \text{Exp}(\theta); \\ &X_{(n-r+1)}, X_{(n-r+2)}, \ldots, X_{(n)} \textit{ derived from } \text{Exp}(\theta/b), \\ &\quad 0 < b \leqslant 1, \; b \textit{ is unknown.} \end{aligned}$$

Then, $Y_j \sim \text{Exp}(\theta(rb + n - r - j + 1))$ for $1 \leqslant j \leqslant n - r$; $Y_{n-r+j} \sim \text{Exp}((\theta/b)(r - j + 1)^{-1})$ for $1 \leqslant j \leqslant r$.

Proof. Let $\{N_t^k\}_{t \geq 0}$ be independent Poisson counting process with state space $\{0, 1\}$, rate $1/\theta$ for $k \in \{1, \ldots, n - r\}$ and rate θ/b for $k \in \{n - r + 1, \ldots n\}$. Then, each X_k is the sojourn time of N_t^k in the stage 0. Let

$$N_t = \sum_{k=1}^{n} N_t^k,$$

then N_t is a Poisson process with state space $\{0, 1, \ldots, n\}$. Assuming $X_{(0)} = 0$, then $Y_i = X_{(i)} - X_{(i-1)}$ is the sojourn time of N_t in state $i - 1$; that is,

$$\begin{aligned} \mathbb{P}(Y_i > t) &= \mathbb{P}(N_t < i \mid N_t = i - 1) \\ &= \mathbb{P}\left(\sum_{k=i}^{n} N_t^k - \sum_{k=i}^{n} -N_0^k = 0 \right) \\ &= \mathbb{P}\left(\sum_{k=i}^{n} N_t^k - N_0^k = 0 \right) \\ &= \prod_{k=i}^{n} \mathbb{P}\left(N_t^k - N_0^k = 0 \right). \end{aligned}$$

If we assume $i \in \{1, \ldots, n - r\}$, then

$$\begin{aligned} \mathbb{P}(Y_i > t) &= \left[\prod_{k=n-r}^{n} \mathbb{P}\left(N_t^k - N_0^k = 0 \right) \right] \left[\prod_{k=i}^{n-r} \mathbb{P}\left(N_t^k - N_0^k = 0 \right) \right] \\ &= \left[\prod_{k=n-r}^{n} \exp\left\{ -\frac{b}{\theta} t \right\} \right] \left[\prod_{k=i}^{n-r} \exp\left\{ -\frac{t}{\theta} \right\} \right] \\ &= \exp\left\{ -\frac{rb}{\theta} t - \frac{n - r + 1 - i}{\theta} t \right\} \end{aligned}$$

and $Y_i \sim \text{Exp}(\theta(rb + n - r - i + 1))$ as desired. If $i = n - r + j$ for $j \in \{1, \ldots, r\}$ then

$$\begin{aligned} \mathbb{P}\left(Y_{n-r+j} > t \right) &= \prod_{k=n-r+j}^{n} \mathbb{P}\left(N_t^k - N_0^k = 0 \right) \\ &= \exp\left\{ -(r + j + 1) \frac{b}{\theta} t \right\}, \end{aligned}$$

and $Y_j \sim \text{Exp}((\theta/b)(r - j + 1)^{-1})$ for $1 \leqslant j \leqslant r$ as desired. □

2.2 Distribution of Order Statistics Under Monotone Transformation

The goal of this section is to establish the main result that shows certain tests statistics used for testing H_r for the exponential samples can be easily adapted to test H_r for the Pareto samples and the transformed test will have the same power and critical values.

Theorem 2.1. *Let $X_1, X_2, \ldots, X_n$ be continuous random variables with density $f_1, \ldots, f_n$, respectively, where f_i has the same support (a,b) with $-\infty \leqslant a < b \leqslant \infty$. Let $g_1, \ldots, g_n$ be a collection of strictly increasing differentiable functions with domain (a,b) and range $(c,d) \subseteq \mathbb{R}$. Define random variable $Y_i = g_i(X_{(i)})$, for $i = 1, \ldots, n$. Then, the joint pdf of $Y_1, \ldots Y_n$ is given by*

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) = \begin{cases} n! \prod_{i=1}^{n} \left| \frac{\mathrm{d}g_i^{-1}}{\mathrm{d}y} \right| f_i(y_i), & c < y_1 < y_2 < \cdots < y_n < d \\ 0, & \text{elsewhere.} \end{cases}$$

Proof. Since each g_i is monotone increasing and $X_{(1)} < \cdots < X_{(n)}$, evidently the support of of $Y_1, \ldots, Y_n$ is given by $\mathcal{Y} = \{(y_1,\ldots,y_n) : c < y_1 < \cdots < y_n < d\}$, which is the image of the map $T : (x_1,\ldots,x_n) \mapsto (g_1(x_1),\ldots,g_n(x_n))$ from the set $\mathcal{X}_0 = \{a < x_1 < \cdots < x_n < b\}$. In fact, since each g_i is an order preserving bijection, T is a bijection from set $\mathcal{X}_\sigma = \{(x_{\sigma(1)},\ldots,x_{\sigma(n)}) : (x_{\sigma(1)} < \cdots < x_{\sigma(n)}\}$ to $\mathcal{Y}$, where σ is any permutation of the set $\{1,\ldots,n\}$.

Consider the transformation of T restricted on $\mathcal{X}_0$, which is given by $y_1 \mapsto g_1(x_1), \ldots, x_n \mapsto g_n(x_n)$. Then, Jacobian for $T^{-1}_{\mathcal{X}_0}$ is given by

$$|J_0| = \operatorname{abs}\left(\begin{vmatrix} \frac{\mathrm{d}g_1^{-1}}{\mathrm{d}y_1} & 0 & \cdots & 0 \\ 0 & \frac{\mathrm{d}g_2^{-1}}{\mathrm{d}y_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\mathrm{d}g_n^{-1}}{\mathrm{d}y_n} \end{vmatrix} \right),$$

which is equal to $\prod_{i=1}^{n} \left| \frac{\mathrm{d}g_i^{-1}}{\mathrm{d}y} \right|$. Since for any permutation σ, the Jacobian of $T^{-1}|_{\mathcal{X}_\sigma}$ can be obtained by interchanging the row of J_0, it differs from J_0 at most by a sign.

Let S_n denotes all permutations on $\{1,\ldots,n\}$, and J_σ denotes the Jacobian of $T^{-1}_{\mathcal{X}_\sigma}$. Because $\{\mathcal{X}_\sigma\}_{\sigma \in S_n}$ are mutually disjoint and partition the support of $X_1, \ldots, X_n$, the pdf of $(Y_1, \ldots, Y_n)$ is given by

$$\begin{aligned} f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) &= \sum_{\sigma \in S_n} |J_\sigma| f(y_1,\ldots,y_n) \\ &= \begin{cases} n! \prod_{k=1}^{n} \left| \frac{\mathrm{d}g_i^{-1}}{\mathrm{d}y} \right| f_i(y_i), & g(a) < y_1 < y_2 < \cdots < y_n < g(b) \\ 0, & \text{elsewhere,} \end{cases} \end{aligned}$$

where $n!$ comes from S_n having $n!$ elements and the product of f_i follows from the assumption X_i are independent. □

Remark. This result is motivated by the classical proof for the distribution of order statistics. Here, we consider the case when g_i are strictly increasing, one can also show that a similar result holds if g_i are strictly decreasing functions.

By taking g_i to be the identity map and f_i to be the same density f in the previous theorem, we obtain the follwing well known fact.

Corollary 2.1. *Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution of the continuous type having a pdf $f(x)$ that has support (a,b) with $-\infty \leqslant a < b \leqslant \infty$. Let $Y_k = X_{(k)}$ for $k = 1, \ldots, n$, then then the joint pdf of $Y_1, \ldots Y_n$ is given by*

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n), = \begin{cases} n! f(y_1) f(y_2) \cdots f(y_n), & 0 < y_1 < y_2 < \cdots < y_n < b \\ 0, & \text{elsewhere.} \end{cases}$$

Now we shall prove Corollary 2.3, which connects the order statistics of exponential samples and Pareto samples. For this purpose, we first introduce the definition of Pareto distribution.

Definition 2.2 (Pareto distribution)**.** A random variable X follows Pareto(α,θ) distribution if its pdf is given by

$$f(x;\alpha,\theta) = \frac{\alpha\theta^{\alpha}}{x^{\alpha+1}} I_{\{x\geqslant\theta>0\}},$$

where θ and α are both positive parameters.

We note that a Pareto random variable can be transformed into an exponential random variable as shown in the following lemma, and it makes the result given in Corollary 2.3 very intuitive.

Lemma 2.5. *Suppose $X \sim \text{Pareto}(\alpha,\theta)$ and $Y = \log(X/\theta)$. Then, $Y \sim \text{Exp}(\alpha)$.*

Proof. Let f_X and f_Y be the pdf of X and Y, respectively. Then, we have

$$f(x) = \alpha\theta^{\alpha} x^{-(\alpha+1)}$$

for $x \geqslant 0$. Since the function $g : x \mapsto \log(x/\theta)$ is a bijection on the support of X with inverse $g^{-1} : y \mapsto \theta e^{y}$, the pdf of Y is given by

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y))\left|\frac{\mathrm{d}g^{-1}}{\mathrm{d}y}\right| \\ &= \alpha\theta^{\alpha}[\theta e^{y}]^{-(\alpha+1)}|\theta e^{y}| \\ &= \alpha e^{-\alpha y}, \end{aligned}$$

which is the pdf of Exp(α) random variable and the Lemma is thus proved. □

Because the logarithm function is a monotone transformation, one would expect an analogous result of the Lemma 2.5 holds for order statistics from Pareto and exponential distributions, as shown in the Corollary below.

Corollary 2.2. *Let $X_1,\dots X_n$ and $E_1,\dots,E_n$ be independent random variables with $X_k \sim \text{Pareto}(\alpha_k,\theta_k)$ and $E_k \sim \text{Exp}(\alpha_k)$ for $k = 1,\dots,n$. Let $Y_k = E_{(k)}$ and $U_k = \ln\big(X_{(k)}/\theta_k\big)$, for $k = 1,\dots,n$. Then the random vector $\mathbf{U} = (U_1,\dots,U_n)$ has the same distribution as $\mathbf{Y} = (Y_1,\dots,Y_n)$.*

Proof. Because each X_k is independent and the function $g_k(y) = \ln(y/\theta_k)$ is differentiable and strictly increasing on $[0,\infty)$ with range same as its domain, we can apply Theorem 2.1 to obtain the pdf of $\mathbf{U}$. It follows from the definition that the support of $\mathbf{U}$ and $\mathbf{Y}$ are both $\mathcal{U} = \{(u_1,\dots,u_n) : 0 < u_1 < \cdots < u_n < \infty\}$. Let f_k and h_k be the pdfs of X_k and Y_k, respectively. Then for any $(u_1,\dots,u_n) \in \mathcal{U}$, we have

$$\begin{aligned} f_{U_1,\dots,U_n}(u_1,\dots,u_n) &= n!\prod_{k=1}^{n}\left|\frac{\mathrm{d}g_k^{-1}}{\mathrm{d}y}\right| f_k(y_k) \\ &= n!\prod_{k=1}^{n}\alpha_k\theta_k^{\alpha_k}[\theta_k e_k^{u}]^{-(\alpha_k+1)}|\theta_k e^{u_k}| \\ &= n!\prod_{k=1}^{n}\alpha_k e^{-\alpha u_k} \\ &= n!\prod_{k=1}^{n} h_k(u_k). \end{aligned}$$

Then, it follows from Theorem 2.1, with g_i being the identity map, we verify that this is indeed the pdf of $\mathbf{U}$ and the result follows. □

As another application of the previous corollary, we give an analogous result under H_r, which is the main result of this section.

Corollary 2.3. *Let $X_1,\dots,X_n$ be independent random variables with $X_1,\dots,X_{n-r} \sim \text{Pareto}(\alpha_1,\theta)$ and $X_{n-r+1},\dots,X_n \sim \text{Pareto}(\alpha_2,\theta)$. Let $E_1,\dots,E_n$ be independent random variables with $E_1,\dots,E_{n-r} \sim \text{Exp}(\alpha_1)$ and $E_{n-r+1},\dots,E_n \sim \text{Exp}(\alpha_2)$. Let $Y_k = E_{(k)}$ and $U_k = \ln\big(X_{(k)}/\theta\big)$, for $k = 1,\dots,n$. Then the random vector $\mathbf{U} = (U_1,\dots,U_n)$ has the same distribution as $\mathbf{Y} = (Y_1,\dots,Y_n)$ under the slippage alternative H_r.*

Proof. Because $\ln(X_k/\theta)$ has the distribution of $\text{Exp}(\alpha_k)$, it is evident that the random vector $(\ln(X_1/\theta),\dots,\ln(X_n/\theta))$ has the same distribution as $(E_1,\dots,E_n)$. Then, it direct follows from the previous corollary that

$$\begin{aligned}\mathbb{P}\left(U_k \leqslant u \mid H_r\right) &= \mathbb{P}\left(\ln\big(X_{(k)}/\theta\big) \leqslant u \mid \max\{\ln(X_1/\theta),\dots,\ln(X_{n-r}/\theta)\} < \min\{\ln(X_{n-r+1}),\dots,\ln(X_n)\}\right)\\ &= \mathbb{P}\left(Y_k \leqslant u \mid \max\{E_1,\dots,E_{n-r}\} < \min\{E_{n-r+1},\dots,E_n\}\right)\\ &= \mathbb{P}\left(Y_k \leqslant u \mid H_r\right).\end{aligned}$$

□

2.3 Simulation Methods

Monte Carlo methods are experiments that make use of simulated random samples to estimate functionals of statistical densities and we now introduce some methods that will be used.

2.3.1 Slippage Random Sample Generation

Sample Generation Algorithm1

To conduct simulation studies and investigate the performance of statistical tests under slippage alternative H_r defined in Definition 1.2, we need methods that generates random samples under H_r. Our study focuses on the case when F and $\overline{F}$ are exponential or Pareto distributions, but we only need to consider generating samples from exponential case, since it follows from Corollary 2.3 that we can obtain random samples from Pareto distribution under slippage alternative by applying appropriate transformation on the exponential sample obtained.

Algorithm 1 Algorithm for Generating Exponential Random Samples Under H_r

Input:
sample size : n
number of contaminated sample: r
the parameter for contaminated sample : b,θ, where $b \in (0,1), \theta > 0$.
Output: A sample $\mathbf{X} = (x_1,\dots,x_n)$ under H_r.

1: **repeat**
2: generating sample $\mathbf{x} = (x_1,\dots,x_{n-r}) \overset{iid}{\sim} \text{Exp}(\theta)$
3: generating sample $\mathbf{x}^b(x_{n-r+1},\dots,x_n) \overset{iid}{\sim} \text{Exp}(\theta/b)$
4: **until** $\max(\mathbf{x}) < \min(\mathbf{x}^b)$
5: $\mathbf{x} := (x_1,\dots,x_{n-r},x_{n-r+1},\dots,x_n)$
6: **return x**.

The distribution of the samples generated by the above algorithm is indeed for the samples under H_r. Let $\mathbf{X} = (X_1,\dots,X_n)$ with $(X_1,\dots,X_{n-p})$ coming from $\text{Exp}(\theta)$ and $(X_{p+1},\dots,X_n)$ coming from $\text{Exp}(\theta/b)$. Then,

$$\begin{aligned}\mathbb{P}\left(X_{(k)} \leqslant x \mid H_r\right) &= \mathbb{P}\left(X_{(k)} \leqslant x \mid \max\{X_1,\dots,X_{n-p}\} < \min\{X_{n-p+1},\dots,X_n\}\right)\\ &= \mathbb{P}\left(X_{(k)} \leqslant x \mid \text{accept } \mathbf{X}\right).\end{aligned}$$

Efficiency of Algorithm1

Since the algorithm generates the samples by accepting samples that satisfy H_r from samples with contamination but not necessarily satisfy H_r, then the natural question would be, how efficient is Algorithm1; that is, what is the probability that a mixture sample generated will be accepted? To see this, we proceed as follows.

Let $Y = \max\{X_1, \ldots, X_{n-r}\}$ and $Z = \min\{X_{n-r+1}, \ldots, X_n\}$. It then follows that the sample $\mathbf{X}$ will be accepted only if $Y < Z$. It follows from Theorem 5.4.4 of [10] that the pdf for Y and Z is given by

$$f_Y(y) = (n-r)[1-e^{-y/\theta}]^{n-r-1}\frac{e^{-y/\theta}}{\theta}, \quad y > 0,$$

and

$$f_Z(z) = r\left(e^{-bz/\theta}\right)^{r-1}\frac{be^{-bz/\theta}}{\theta}, \quad z > 0.$$

Hence, the acceptance probability is given by

$$\begin{aligned}
\mathbb{P}(Y < Z) &= \int_0^\infty \mathbb{P}(Y < z \mid Z = z) f_Z(z)\,\mathrm{d}z \\
&= \int_0^\infty (1-e^{-z/\theta})^{n-r} r \exp\{-zb/\theta\}^{r-1}\frac{be^{-zb/\theta}}{\theta}\,\mathrm{d}z \\
&= \int_0^\infty \frac{rb}{\theta}[1-\exp\{-z/\theta\}]^{n-r}\exp\{-rbz/\theta\}\,\mathrm{d}z \\
&= \int_0^1 rb[1-w]^{n-r}w^{rb}w^{-1}\,\mathrm{d}w \\
&= rbB(rb, n-r+1),
\end{aligned}$$

where $B(r,s)$ is the complete beta function. For the special case when $b = 1$, the accepting probability becomes

$$\begin{aligned}
\mathbb{P}(Y < Z) &= rB(r, n-r+1) \\
&= \frac{\Gamma(r)\Gamma(n-r+1)}{\Gamma(n+1)} \\
&= \binom{n}{r}^{-1},
\end{aligned}$$

which rapidly decreases to 0 as n increase.

Besides giving an estimation for the efficiency of Algorithm1, the acceptance probability also gives a measure for whether the slippage alternative is appropriate. It could be interpreted as the probability we get a sample $(x_1, \ldots, x_n)$ that satisfies H_r if $x_1, \ldots, x_{n-r}$ arose from $\mathrm{Exp}(\theta)$ distribution and $x_{n-r+1}, \ldots, x_n$ arose from $\mathrm{Exp}(\theta/b)$ distribution.

We plot the efficiency of the algorithm for different choices of b and r in Figure 2.1. As the figure suggests, the slippage assumption is appropriate only when the number of contamination r is small relative to the total sample size n and difference between the contamination distribution $\overline{F}$ and the distribution assumed under the null hypothesis F have a large difference in mean.

2.3.2 Power Estimation

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent samples arose from distribution $F(\theta)$, with θ unknown. Let Θ be the parameter space of θ and we are interested in testing $H_0 : \theta \in \Theta_0$ against $H_a : \theta \in \Theta_a = \Theta \setminus \Theta_0$. Let C be a subset of the sample space such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{X} \in C) = \alpha$$

so that hypothesis testing procedure of H_a against H_0 of rejecting H_0 if $\mathbf{x} \in C$ would have a confidence level of $1-\alpha$. Suppose we want to estimate the power of the test for a given $\theta \in \Theta_a$ which is given by $\gamma_\theta = \mathbb{P}_\theta(\mathbf{x} \in C)$. Consider the indicator function $I_\theta(\mathbf{x}) = \mathbb{1}_{\{\mathbf{x} \in C\}}$, then $I_\theta \sim \mathrm{Ber}(\gamma_\theta)$, and so estimating γ_θ is equivalent to estimating the parameter of I_θ.

It then follows from the law of large numbers that, if $B_1, \ldots, B_N$ are random samples of I_θ, then

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N B_n \overset{\text{a.s}}{\to} \mathbb{E}[I_\theta] = \gamma_\theta.$$

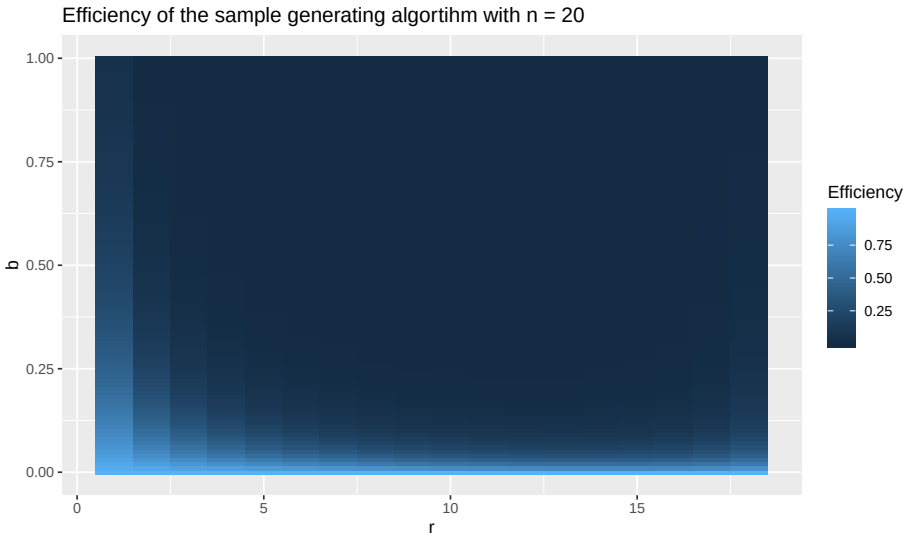

Figure 2.1: The efficiency of Algorithm1 with $n = 12$

The efficiency of the sample generating algorithm for samples under H_r with different contamination number r and contamination rate parameter b.

In particular, if T is a statistic used to test the slippage alternative H_r against H_0 with rejection region C and that we are interested in estimating the power of T. Let $\mathbf{x}_1, \ldots \mathbf{x}_n$ be the random sample generated by Algorithm1. Then, the power of the test can be estimated by

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{T(\mathbf{x}_i) \epsilon C\}}.$$

It then follows from the property of the Bernoulli distribution that the standard error of $\hat{\gamma}_G$ is given by

$$\text{se}(\hat{\gamma}) = \sqrt{\frac{\hat{\gamma}(1-\hat{\gamma})}{n}}.$$

2.3.3 Estimation of Statistical Functions

In this study, we are interested in estimating some unknown distribution $\overline{F}$ which does not have an explicit expression, but we are able to generate random sample of $\overline{F}$. For this purpose, we now introduce some simulation concepts and methods that will be used latter.

Empirical Distribution Functions

With a sample of univariate data points $x_1, \ldots, x_n$ from distribution F, a commonly used tool to summarize the data is using histogram. The shape of the histogram often provides valuable information about the pdf of F, for example, a bell shaped histogram may suggest normality. If we consider the cumulative histogram, then we would expect it has the shape of the CDF of F, and this intuition turns out to be correct as provided in the following definition.

Definition 2.3. Let $X_1, \ldots, X_n$ be a random sample of distribution $F(x)$. Then, we define the *empirical distribution function*(ecdf) $\hat{F}_n$ to be

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

Remark. From this definition, we can see that $\hat{F}_n$ follows direct uniform distribution, as this approach has an assumption that each observed value is equally likely to be observed.

Here, we show that ecdf $\hat{F}_n$ is a pointwise unbiased estimator of F and give a expression for its variance.

Theorem 2.2. *Suppose F is a cumulative distribution function and $\hat{F}_n$ is the empirical cumulative distribution function based on $X_1, \ldots, X_n$. Then, for any point $x \in \mathbb{R}$, $\hat{F}_n(x)$ is an unbiased estimator of $F(x)$ with variance*

$$\mathrm{Var}\left[\hat{F}_n(x)\right] = \frac{F(x)(1-F(x))}{n}.$$

Proof. Let any $x \in \mathbb{R}$ be given. Then,

$$\begin{aligned}
\mathbb{E}\left[\hat{F}_n(x)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant x\}}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\{X_i \leqslant x\}}\right] \\
&= \mathbb{P}\left(X \leqslant x\right) \\
&= F(x)
\end{aligned}$$

and this proves that $\hat{F}_n(x)$ is unbiased. The variance of $\hat{F}_n$ is given by

$$\begin{aligned}
\mathrm{Var}\left[\hat{F}_n(x)\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant x\}}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[\mathbb{1}_{\{X_i \leqslant x\}}\right] \\
&= \frac{F(x)(1-F(x))}{n}
\end{aligned}$$

Since $\mathbb{1}_{\{X_i \leqslant x\}} \sim \mathrm{Ber}(\mathbb{P}\left(X_i \leqslant x\right))$, the proof gets completed. □

Some characteristic ψ of F we are interested in, such as quantile, can often be expressed as a functional of F, say, $\psi = t(F)$. It then follows from the theorem proved above and the plug-in principle that $\hat{\psi} = t(\hat{F})$ will be an estimator of F, which will be used in our study.

Nonparametric Bootstrap Method

Since F is unknown, we often can not find an easy expression for the variance and bias $\hat{\psi}$, and so we wil use bootstrap method to estimate it.

Recall that in this study we are focusing on Pareto and exponential distributions, and so we may assume the unknown distribution F is continuous. Then, it follows from the fact that $F(X) \sim \mathcal{U}(0,1)$ for any continuous random variable X with cdf F; so, we can use the inverse transformation method to simulate $\hat{F}$. More specifically, we can simulate iid samples $U_1, \ldots, U_n$ from $\mathcal{U}(0,1)$ distribution, then the corresponding $X_1, \ldots, X_n$ of $\hat{F}_n$ is generated as

$$X_i = x_{(i)} \quad \text{if} \quad \frac{i-1}{n} < U_i \leqslant \frac{i}{n}, \quad i = 1, \ldots, n,$$

where $x_{(i)}$ is taken to be the i^{th} ordered value of the support of $\hat{F}_n$. It follows from the definition of $\hat{F}_n$ given in Definition 2.3 that simulating sample from $\hat{F}_n$ is equivalent to taking a random sample from the support of $\hat{F}$ with replacement. Hence, we can generate random samples of $\hat{F}_n$ by sampling from the support of $\hat{F}_n$, and we denote this by $\mathbf{X}_1^*, \ldots, \mathbf{X}_B^*$. We can obtain the bootstrap estimates $\hat{\psi}_b = t(\mathbf{X}_b^*)$, for $b = 1, \ldots, B$. Then, we can estimate the bias and variance of $\hat{\psi}$ by

$$\hat{b}(\hat{\psi}) = \frac{1}{B}\sum_{b=1}^{B} (\hat{\psi}_b - \hat{\psi}), \tag{2.1}$$

$$\widehat{\mathrm{Var}}[\hat{\psi}] = \frac{1}{B-1}\sum_{b=1}^{B} \left(\hat{\psi} - \overline{\hat{\psi}_b}\right)^2. \tag{2.2}$$

Chapter 3

Exact Distributions of Discordancy Test Statistics under H_r

In this chapter we derive the exact distributions of the test statistics given by Zerbet and Nikulin [6] and Jabbari Nooghabi [7] under the slippage alternative H_r.

3.1 Discordancy Tests for Exponential Samples

Recalling the contamination model and slippage hypothesis introduced in the Definition 1.2, here we consider the case where the underline distribution is $\text{Exp}(\theta)$ with distribution function $F(x;\theta) = \mathbb{P}(X \leqslant x) = 1 - \exp\{-x/\theta\}$, and the contaminated model is $\overline{F} = \text{Exp}(\theta/b)$ for some unknown constant b, where $0 < b \leqslant 1$.

Zerbet and Nikulin [6] proposed the following statistic to test H_r against H_0:

$$Z_r = \frac{X_{(n-r)} - X_{(1)}}{\sum_{j=n-r+1}^{n}(X_{(j)}) - X_{(1)})}. \tag{3.1}$$

Since we are assuming $b < 1$, under H_r, the test statistic Z_r given in (3.1) has the property that

$$\begin{aligned}\lim_{b\to\infty} \mathbb{E}[Z_r] &= \lim_{b\to\infty} \mathbb{E}\left[\frac{X_{(n-r)} - X_{(1)}}{\sum_{j=n-r+1}^{n}(X_{(j)}) - X_{(1)})}\right] \\ &\leqslant \lim_{b\to\infty} \mathbb{E}\left[\frac{X_{(n-r)}}{X_{(n)} - X_{(1)}}\right] \\ &= 0,\end{aligned}$$

for $X_{(n-r)}$ and $X_{(1)}$ are arising from $\text{Exp}(\theta)$ and $X_{(n)}$ is arising from $\text{Exp}(\theta/b)$ under H_r. Therefore, one should reject H_0 if $Z_r < z_c$, where $z_c(\alpha)$ is the α quantile of Z_r. So, we have corrected the direction of the test claim in [6], where they claimed the test direction is a upper tail test. Here, we derive its distribution, which is the corrected version of Theorem 1 proved in [6].

3.1.1 Distribution of Z_r

Theorem 3.1. *The distribution of the statistic Zr, under H_r, is given by*

$$\begin{aligned}\mathbb{P}(Z_r < Z \mid H_r) = &\frac{b^r \Gamma(rb+n+r)}{\Gamma(rb+1)} \\ &\times \sum_{j=2}^{n-r} \frac{(-1)^{n-j-r}\{b^{-r} - [(rb+n-r-j+1)(z/(1-rz)) + b]^{-r}\}}{(j-2)!(n-j-r)!(rb+n-r-j+1)}, \qquad 0 < x < \frac{1}{r}.\end{aligned}$$

Proof. Suppose H_r holds and consider the random variable U_r given by

$$U_r = \frac{X_{(n-r)} - X_{(1)}}{\sum_{j=n-r+1}^{n}(X_{(j)}) - X_{(n-r)})}.$$

Then

$$\frac{1}{Z_r} - \frac{1}{U_r} = \frac{1}{X_{(n-r)} - X_{(1)}} \sum_{j=1}^{r} X_{(n-r)} - X_{(1)} = r, \tag{3.2}$$

so that $U_r = \frac{Z_r}{1-r}$. We obtain the distribution of Z_r by considering the distribution of U_r. Let

$$V = X_{(n-r)} - X_{(1)} \qquad \text{and} \qquad W = \sum_{j=n-r+1}^{n} X_{(j)} - X_{(n-r)}.$$

Then the test statistic $Z_r = V/W$. Let $Y_j = X_{(j)} - X_{(1)}$. Because

$$\begin{aligned}
\sum_{j=2}^{n-r} Y_j &= \sum_{j=2}^{n-r} X_{(j)} - X_{(j-1)} \\
&= \sum_{j=2}^{n-r} X_{(j)} - \sum_{j=2}^{n-r} X_{(j-1)} \\
&= X_{(n-r)} - X_{(1)} + \sum_{j=1}^{n-r-1} X_{(j)} - \sum_{j=1}^{n-r-1} X_{(j)} \\
&= X_{(n-r)} - X_{(1)}
\end{aligned}$$

and

$$\begin{aligned}
\sum_{j=n-r+1}^{n} (n-j+1)Y_j &= \sum_{j=n-r+1}^{n} X_{(j)} - X_{(j-1)} \\
&= \sum_{j=n-r+1}^{n} X_{(j)} - \sum_{j=n-r}^{n-1} X_{(j-1)} \\
&= \sum_{j=n-r+1}^{n} X_{(j)} - \sum_{j=n-r}^{n-1} X_{(j-1)} \\
&= (n-n+1)X_{(n)} - (n-n+r-1+1)X_{(n-r+1)} + \sum_{j=n-r}^{n-1} (n-j+1)X_{(j)} - (n-j)X_j \\
&= X_{(n)} - rX_{(n-r)} + \sum_{j=n-r}^{n-1} X_{(j)} \\
&= \sum_{j=n-r}^{n} X_{(j)} - X_{(n-r)},
\end{aligned}$$

we have

$$Z_r = \frac{V}{W} = \frac{\sum_{j=2}^{n-r} Y_j}{\sum_{j=n-r+1}^{n}(n-j+1)Y_j}. \tag{3.3}$$

The joint characteristic function (V, W) is

$$
\begin{aligned}
\varphi_{V,W}(t,z) &= \mathbb{E}\left[\exp\{i(Vt+Wz)\}\right] \\
&= \mathbb{E}\left[\exp\left\{i\left(\sum_{j=2}^{n-r} Y_j t + \sum_{j=n-r+1}^{n} (n-j+1)Y_j z\right)\right\}\right] \\
&= \int_{\mathbb{R}^{n-1}} \exp\left\{i\left(\sum_{j=2}^{n-r} Y_j t + \sum_{j=n-r+1}^{n} (n-j+1)Y_j z\right)\right\} \\
&\quad \times f_{(Y_2,\ldots,Y_n)}(y_2,\ldots,y_n)\,\mathrm{d}y_2\cdots\mathrm{d}y_n .
\end{aligned}
$$

It follows from Lemma 2.4 that Y_j, $j=1,\cdots,n-r$, follows $\mathrm{Exp}(\theta(rb+n-r-j+1))$ and Y_{n-r+j} follows $\mathrm{Exp}((\theta/b)(r-j+1)^{-1})$ distribution. Let $a_j=\theta(rb+n-r-j+1)^{-1}$ and $b_j=(\theta/b)(r-j+1)^{-1}$. We can then write $\varphi_{(V,W)}(t,z)$ as

$$
\begin{aligned}
\varphi_{V,W}(t,z) = & \int_{\mathbb{R}^{n-1}} \exp\left\{it(\sum_{j=2}^{n-r} Y_j)\right\}\left[\prod_{k=2}^{n-r}\frac{1}{a_k}e^{-yk/ak}\right] \\
& \times \exp\left\{iz\sum_{j=n-r+1}^{n}(n-j+1)y_j\right\}\left[\prod_{k=1}^{r}\frac{1}{b_k}e^{-y_{n-r+1}/b_k}\right]\mathrm{d}y_2\cdots\mathrm{d}y_n \\
= & \prod_{j=2}^{n-r}\int_0^\infty \frac{1}{a_j}\exp\left\{-\frac{y_j}{a_j}+ity_j\right\}\mathrm{d}y_j \\
& \times \prod_{j=1}^{r}\int_0^\infty \frac{1}{b_j}\exp\left\{-iz(r-j+1)y_{n-r+j}-y_{n-r+j}/b_j\right\}\mathrm{d}y_{n-r+j} \\
= & \prod_{j=2}^{n-r}\int_0^\infty \frac{1}{a_j}\exp\left\{-\frac{y_j}{a_j}+ity_j\right\}\mathrm{d}y_j \\
& \times \prod_{j=1}^{r}\int_0^\infty \frac{1}{b_j}\exp\left\{-y_{n-r+j}\left(\frac{1}{b_j}-iz(r-j+1)\right)\right\}\mathrm{d}y_{n-r+j} \\
= & \left[\prod_{j=2}^{n-r}\frac{1}{a_j}(1/a_j-it)^{-1}\right]\times\left[\prod_{j=1}^{r}\frac{1}{b_j}(1/b_j-iz(r-j+1))^{-1}\right].
\end{aligned}
$$

It then follows from Lemma 2.2 that the joint pdf of (V,W) is given by

$$
\begin{aligned}
f_{(V,W)}(v,w) &= \frac{1}{(2\pi)^2}\int_0^\infty\int_0^\infty \varphi_{(V,W)}(t,z)\exp\{it(tv+wz)\}\,\mathrm{d}t\,\mathrm{d}z \\
&= \frac{1}{(2\pi)^2}\int_0^\infty\left[\prod_{j=2}^{n-r}\frac{1}{a_j}(1/a_j-it)^{-1}\exp\{itv\}\,\mathrm{d}t\right]\times\int_0^\infty\left[\prod_{j=1}^{r}\frac{1}{b_j}(1/b_j-iz(r-j+1))^{-1}e^{-iwz}\,\mathrm{d}z\right]
\end{aligned}
\tag{3.4}
$$

To simplify (3.4), we first use Lemma 2.1 to obtain

$$\begin{aligned}
\prod_{j=2}^{n-r}\frac{1}{1/a_j - it} &= (-1)^{n-r-1}\prod_{j=2}^{n-r}\frac{1}{it-1/a_j} \\
&= (-1)^{n-r-1}\sum_{j=2}^{n-r}\frac{1}{it-1/a_j}\prod_{i\neq 1, i\neq j}^{n-r}\frac{1}{1/a_j-1/a_i} \\
&= (-1)^{n-r-1}\sum_{j=2}^{n-r}\frac{1}{it-1/a_j}\prod_{i\neq 1, i\neq j}^{n-r}\frac{1}{\theta^{-1}(i-j)} \\
&= (-1)^{n-r-1}\sum_{j=2}^{n-r}\frac{\theta^{n-r-2}}{it-1/a_j}\left[\prod_{i=2}^{j-1}\frac{1}{1-j}\right]\left[\prod_{i=j+1}^{n-r}\frac{1}{1-j}\right] \\
&= (-1)^{n-r-1}\sum_{j=2}^{n-r}\frac{\theta^{n-r-2}}{it-1/a_j}\left[\frac{(-1)^{j+2}}{((j-2)!)}\right]\left[\frac{1}{(n-r-j)!}\right] \\
&= \sum_{j=2}^{n-r}\frac{(-1)^{n+j-r-1}\theta^{n-r-2}}{(it-1/a_j)(j-2)!(n-j-r)!},
\end{aligned}$$

$$\begin{aligned}
\prod_{j=1}^{r}(1/b_j - iz(r-j+1))^{-1} &= \prod_{j=1}^{r}\left(\frac{b}{\theta}(r-j-1) - iz(r-j+1)\right)^{-1} \\
&= \prod_{j=1}^{r}((b/\theta - iz)(r-j+1))^{-1} \\
&= (b/\theta - iz)^r\prod_{j=1}^{r}(r-j+1)^{-1} \\
&= \frac{(b/\theta - iz)^r}{r!},
\end{aligned}$$

$$\begin{aligned}
\prod_{j=1}^{r}\frac{1}{a_j} &= \prod_{j=2}^{n-r}\frac{rb+n-r-j+1}{\theta} \\
&= \frac{1}{\theta^{n-r-1}}\frac{\Gamma(rb+n-r)}{\Gamma(rb+1)},
\end{aligned}$$

and

$$\prod_{j=1}^{r}\frac{1}{b_j} = \prod_{j=1}^{r}\frac{b}{\theta}(r-j+1) = \left(\frac{b}{\theta}\right)^r r!. \tag{3.5}$$

We can then express the $f_{V,W}(v,w)$ given in (3.4) as

$$f_{(V,W)}(v,w) = \left[\frac{1}{(2\pi)^2}\sum_{j=2}^{n-r}\frac{\Gamma(rb+n-r)(-1)^{n+j-r-1}}{\Gamma(rb+1)(j-2)!(n-j-r)!}\int_0^\infty \frac{e^{-iv}}{it-1/a_j}\,\mathrm{d}t\right]\times\left[\left(\frac{b}{\theta}\right)^{r+1}\int_0^\infty\frac{e^{-iwz}}{(b/\theta-iz)^r}\,\mathrm{d}z\right].$$

We now claim that

$$\int_0^\infty \frac{e^{-iwz}}{(b/\theta-iz)^r}\,\mathrm{d}z = \frac{2\pi w^{r-1}}{(r-1)!}e^{-wb/\theta}.$$

To show that the above integration identity holds, let X be a random variable having Gamma$(r,\theta/b)$ distribution, with the pdf $p(x)$ and characteristic function $f(x)$ as

$$p(x) = \frac{w^{r-1}e^{-w/b}}{\Gamma(r)\theta^\alpha}, \quad x>0, \qquad \text{and} \qquad f(x) = (1-ix\theta/b)^{-r}.$$

Then it follows from Lemma 2.2 that

$$
\begin{aligned}
& 2\pi p(w) = \int_{-\infty}^{\infty} e^{-izw} f(t)\,\mathrm{d}z \\
\iff \quad & 2\pi \frac{w^{r-1}e^{-w\theta/b}}{(\theta/b)^r (r-1)!} = \int_0^{\infty} \frac{e^{-iwz}}{(1-iz\theta/b)^r}\,\mathrm{d}z \\
\iff \quad & 2\pi \frac{w^{r-1}e^{-w\theta/b}}{(r-1)!} = \int_0^{\infty} \frac{e^{-iwz}}{(b/\theta)^r (1-iz\theta/b)^r}\,\mathrm{d}z \\
\iff \quad & 2\pi \frac{w^{r-1}e^{-w\theta/b}}{(r-1)!} = \int_0^{\infty} \frac{e^{-iwz}}{(b/\theta - iz)^r}\,\mathrm{d}z
\end{aligned}
$$

which proves the claim.

Because $Z_r = V/W$, therefore we can obtain the distribution of Z_r by direct integrate $f_{V,W}(v,w)$ as follows:

$$
\begin{aligned}
\mathbb{P}(U_r < u) &= \mathbb{P}\left(\frac{V}{W} < u\right) \\
&= \mathbb{P}(V < uW) \\
&= \int_0^{\infty}\int_0^{uw} f_{(V,W)}(v,w)\,\mathrm{d}v\,\mathrm{d}w.
\end{aligned}
$$

It follows from the Lemma 2.2 and the characteristic function of $\mathrm{Exp}(1/a_j)$ that

$$
\int_0^{\infty} \frac{e^{-itv}}{it - 1/a_j}\,\mathrm{d}t = -a_j \int_0^{\infty} \frac{e^{-itv}}{1-ita_j}\,\mathrm{d}t = -2\pi \exp\left\{-\frac{1}{a_j}v\right\},
$$

and also by Lemma 2.2, we can simplify $f_{V,W}(v,w)$ to

$$
\begin{aligned}
f_{V,W}(v,w) = \frac{\Gamma(rb+n-r)}{\Gamma(rb+1)}\left(\frac{b}{\theta}\right)^{r+1} \frac{1}{(r-1)!} \\
\times \sum_{j=2}^{n-r} \frac{(-1)^{n+j-r}}{(j-2)!(n-j-r)!} g(v,w),
\end{aligned}
$$

where

$$
g(v,w) = w^{r-1}\exp\left\{-\frac{v}{a_j} - \frac{b}{\theta}w\right\}.
$$

Since

$$
\begin{aligned}
\int_0^{\infty}\int_0^{uw} g(v,w)\,\mathrm{d}v\,\mathrm{d}w &= \int_0^{\infty} w^{r-1}\exp\left\{-\frac{b}{\theta}w\right\}(-a_j)\left[\exp\left\{-\frac{u}{a_j}w\right\} - 1\right]\mathrm{d}w \\
&= -a_j\left[\int_0^{\infty} w^{r-1}\exp\left\{-\left(\frac{b}{\theta} + \frac{1}{a_j}u\right)w\right\}\mathrm{d}w \right. \\
&\quad \left. - \int_0^{\infty} w^{r-1}\exp\left\{-\frac{b}{\theta}w\right\}\mathrm{d}w\right] \\
&= -a_j\left[\Gamma(r)\left(\frac{b}{\theta} + \frac{u}{a_j}\right)^{-r} - \Gamma(r)\left(\frac{b}{\theta}\right)^{-r}\right] \\
&= -a_j(r-1)!\left[\left(\frac{b + u\theta/a_j}{\theta}\right)^{-r} - \left(\frac{b}{\theta}\right)^{-r}\right] \\
&= a_j(r-1)!\theta^r\left[b^{-r} - [u\theta/a_j + b]^{-r}\right],
\end{aligned}
$$

we have

$$\begin{aligned}\int_0^\infty \int_0^{uw} f_{(V,W)}(v,w)\,\mathrm{d}v\,\mathrm{d}w &= \frac{\Gamma(rb+n-r)}{\Gamma(rb+1)}\left(\frac{b}{\theta}\right)^{r+1}\frac{1}{(r-1)!}\times\sum_{j=2}^{n-r}\frac{(-1)^{n+j-r}}{(j-2)!(n-j-r)!}\int_0^\infty\int_0^{uw} g(v,w)\,\mathrm{d}v\,\mathrm{d}w\\ &= \frac{\Gamma(rb+n-r)}{\Gamma(rb+1)}\left(\frac{b}{\theta}\right)^{r+1}\frac{1}{(r-1)!}\sum_{j=2}^{n-r}\frac{(-1)^{n-j-r}}{(j-2)!(n-j-r)!}\\ &\quad\times a_j(r-1)!\theta^r\left[b^{-r}-[u\theta/a_j+b]^{-r}\right]\\ &= \frac{\Gamma(rb+n-r)}{\Gamma(rb+1)}\frac{b^{r+1}}{\theta}\sum_{j=2}^{n-r}\frac{(-1)^{n-j-r}}{(j-2)!(n-j-r)!}a_j\left\{b^{-r}-[u\theta/a_j+b]^{-r}\right\}\\ &= \frac{b^{r+1}\Gamma(rb+n-r)}{\Gamma(rb+1)}\sum_{j=2}^{n-r}\frac{(-1)^{n-j-r}}{(j-2)!(n-j-r)!}\frac{\left[b^{-r}-(u\theta/a_j+b)^{-r}\right]}{\theta/a_j}.\end{aligned}$$

Then, using the fact that $\theta/a_j=(rb+n-r-j+1)$, the above equality becomes

$$\mathbb{P}(U_r<u)=\int_0^\infty\int_0^{uw} f_{(V,W)}(v,w)\,\mathrm{d}v\,\mathrm{d}w=\frac{b\Gamma(rb+n-r)}{\Gamma(rb+1)}\sum_{j=2}^{n-r}\frac{(-1)^{n-j-r}\{b^{-r}-[(rb+n-r-j+1)u+b]^{-1}\}}{(j-2)!(n-j-r)!(rb+n-r-j+1)},$$

and then the required result follows from $U_r=\frac{Z_r}{1-r}$. □

Remark. We remark that the mistake in [6] has been corrected by Mehdi Jabbari Nooghabi, since he proved an analogous result for Pareto distribution in [7], where b corresponds to β and r corresponds to k.

3.2 Discordancy Tests for Pareto Samples

In the previous section, we investigated a statistic test proposed by Zerbet and Nikulin to test the slippage alternative hypothesis for exponential samples. It turns out that the same kind of statistics can be easily adapted to test the slippage alternative for Pareto distribution. This was introduced by Jabbari Nooghabi [7].

To be consistent with the notation used in [7], we will consider the slippage alternative introduced in the Definition 1.2 with $F=\text{Pareto}(\alpha,\theta)$ and $\overline{F}=\text{Pareto}(\alpha\beta,\theta)$ with $\beta<1$.

3.2.1 Test Statistics

Jabbari Nooghabi [7] proposed the statistic $\tilde{Z}_k$ and $\tilde{R}_r$ to test H_r against H_0 for Pareto samples, where

$$\tilde{Z}_r=\frac{\ln\left(X_{(n-r)}\right)-\ln\left(X_{(1)}\right)}{\sum_{j=n-r+1}^{n}\left(\ln\left(X_{(j)}\right)-\ln\left(X_{(1)}\right)\right)}$$

and

$$\tilde{R}_r=\frac{\ln\left(X_{(n-r)}\right)-\ln\left(X_{(1)}\right)}{\ln\left(X_{(n)}\right)-\ln\left(X_{(n-r+1)}\right)}.$$

In [7], the test statistic $\tilde{R}_r$ was first introduced, and was actually used as $1-\tilde{R}_r$. Since $1-R_r$ does not have any theoretical advantage over $\tilde{R}_r$, but $\tilde{R}_r$ has a non-negative support and better interpretability, we will use R_r as the test statistic here.

Theorem 3.2. *Let $X_1,\ldots,X_n$ be a collection of independent Pareto random variable, and under slippage alternative hypothesis H_r. Then the distribution of statistic $\tilde{Z}_r$ is given by*

$$\begin{aligned}\mathbb{P}\left(\tilde{Z}_r<z\,|\,H_r\right)&=\frac{\beta^r\Gamma(r\beta+n-r)}{\Gamma(rb+1)}\sum_{j=2}^{n-r}\sum_{i=2}^{r}\frac{(-1)^{n+i+j-2}}{(n-r-j)!(j-2)!(r-i)!(i-2)!}\\ &\quad\times\frac{[\beta(r-i+1)]^{-1}-[\beta(x-i+1)+z(r\beta+n-r-j+1)]^{-1}}{r\beta+n-r-j+1}.\end{aligned}$$

Proof. We first note that, for any $1 < j, l < n$, we have

$$\ln\big(X_{(j)}/\theta\big) - \ln\big(X_{(l)}/\theta\big) = \ln\big(X_{(j)}\big) - \ln\big(X_{(l)}\big)$$

and so the result follows from the form of Z_r and $\tilde{Z}_r$ and the use of Theorem 3.1 and Corollary 2.3. □

3.2.2 Distribution of $\tilde{R}_r$

In this section, we will derive the distribution of $\tilde{R}_r$ under H_r as given in the following theorem.

Theorem 3.3. *The distribution of the statistic $\tilde{R}_r$ under H_r is given by*

$$\begin{aligned}\mathbb{P}\left(\tilde{R}_r < x \mid H_r\right) &= \frac{\beta^r \Gamma(r\beta + n - r)}{\Gamma(rb+1)} \\ &\quad \times \sum_{j=2}^{n-r} \frac{(-1)^{n-j-r}\{\beta^{-r} - [(r\beta + n - r - j + 1)(x/(1-rx)) + \beta]^{-r}\}}{(j-2)!(n-j-r)!(r\beta + n - r - j + 1)} \qquad 0 < x < \frac{1}{r}.\end{aligned}$$

Proof. We will prove the result using characteristic function. Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics from the sample under $\overline{H}_r$. Then, $\tilde{R}_r$ can be written as

$$\tilde{R}_r = \frac{\sum_{j=2}^{n-k} Y_j}{\sum_{j=n-k+2}^{n} Y_j} = \frac{P}{Q}$$

where $Y_j = \ln\big(X_{(j)}\big) - \ln\big(X_{(j-1)}\big)$, for $j > 1$. It then follows from Lemma 2.4 and Lemma 2.5 that the joint distribution of P and Q is given by

$$\begin{aligned}\phi_{P,Q}(t,s) &= \prod_{j=2}^{n-r} \phi_{Y_j} \prod_{j=2}^{r} \phi_{Y_j} \\ &= \prod_{j=2}^{n-r} \left[\frac{1}{a_j}\left(\frac{1}{a_j} - it\right)\right] \prod_{j=2}^{r} \left[\frac{1}{b_j}\left(\frac{1}{b_j} - is\right)\right],\end{aligned}$$

where $a_j = [\alpha(r\beta + n - r - j + 1)]^{-1}$ and $b_j = [\alpha\beta(r - j + 1)]^{-1}$. Using Lemma 2.2, we have the joint density function of U and V as

$$f_{P,Q}(p,q) = \frac{1}{(2\pi)^2} \int_0^\infty \left[\prod_{j=2}^{n-r} \frac{1}{a_j} (1/a_j - it)^{-1} \exp\{-itp\}\, \mathrm{d}t\right] \times \int_0^\infty \left[\prod_{j=2}^{n-r} \frac{1}{b_j} (1/b_j - it)^{-1} \exp\{-isq\}\, \mathrm{d}t\right].$$

To simplify $f_{P,Q}$, we first use Lemma 2.1 to calculate

$$\begin{aligned}\prod_{j=2}^{k} \frac{1}{1/b_j - is} &= (-1)^r \prod_{j=2}^{r} \frac{1}{it - 1/b_j} \\ &= (-1)^{r-1} \sum_{j=2}^{r} \frac{1}{is - 1/b_j} \prod_{i=2}^{k} \frac{1}{\alpha\beta(k-i)} \\ &= (-1)^{r-1} \frac{1}{(\alpha\beta)^{r-2}} \sum_{j=2}^{k} \frac{1}{is - 1/b_j} \prod_{k=2}^{r} \frac{1}{k-i} \\ &= (-1)^{r-1} \frac{1}{(\alpha\beta)^{r-2}} \sum_{j=2}^{k} \frac{1}{is - 1/b_j} \frac{(-1)^{j-2}}{(j-2)!(k-j)!} \\ &= \sum_{j=2}^{r} \frac{(-1)^{r+j-1}}{(is - 1/b_j)(j-2)(r-j)!(\alpha\beta)^{r-2}}.\end{aligned}$$

It also follows from (3.5) that

$$\prod_{j=2}^{r} \frac{1}{b_j} = (r-1)!(\alpha\beta)^{r-1}.$$

Therefore, with the product identity derived in the proof of Theorem 3.1 we obtained the simplified pdf for $f(p,q)$ is given by

$$f_{P,Q}(p,q) = \frac{\alpha\Gamma(r\beta+n-r)}{\Gamma(r\beta+1)} \sum_{j=2}^{n-r} \frac{(-1)^{n-r+j-1}}{(j-2)(n-r-j)!} \exp\{-n_j p\} \times \alpha\beta(r-1)! \sum_{i=1}^{r} \frac{(-1)^{n-r+j-1}}{(r-1)!(i-2)!} \exp\{-m_i q\},$$

where $n_j = -\alpha(r\beta+n-r-j+1)$ and $m_i = -\alpha\beta(r-i+1)$. Then, the distribution for $\tilde{R}_r$ is given by

$$\begin{aligned}
\mathbb{P}(\tilde{R}_r < x) &= \frac{\alpha^2\beta\Gamma(r\beta+n-r)}{\Gamma(r\beta+1)} \sum_{j=2}^{n-k} \sum_{i=2}^{r} \frac{-1^{n+j+i-2}}{(n-r-j)!(j-2)!(r-1)!(i-2)!} \int_0^{\infty} \int_0^{rq} \exp\{-\alpha n_j p + -\alpha m_i q\} \, \mathrm{d}p \, \mathrm{d}q \\
&= \frac{\alpha^2\beta\Gamma(r\beta+n-r)}{\Gamma(r\beta+1)} \sum_{j=2}^{n-k} \sum_{i=2}^{r} \frac{-1^{n+j+i-2}}{(n-r-j)!(j-2)!(r-1)!(i-2)!} \alpha^{-2} \frac{x}{m_i(n_j x+m_i)} \\
&= \frac{\beta^r\Gamma(r\beta+n-r)}{\Gamma(rb+1)} \sum_{j=2}^{n-r} \sum_{i=2}^{r} \frac{(-1)^{n+i+j-2}}{(n-r-j)!(j-2)!(r-i)!(i-2)!} \times \frac{m_i^{-1} - (m_i+xn_j)^{-1}}{n_i} \\
&= \frac{\beta^r\Gamma(r\beta+n-r)}{\Gamma(rb+1)} \sum_{j=2}^{n-r} \sum_{i=2}^{r} \frac{(-1)^{n+i+j-2}}{(n-r-j)!(j-2)!(r-i)!(i-2)!} \\
&\quad \times \frac{[\beta(r-i+1)]^{-1} - [\beta(x-i+1) + x(r\beta+n-r-j+1)]^{-1}}{r\beta+n-r-j+1}.
\end{aligned}$$

□

Chapter 4

Simulation Study

In this chapter, we will use simulation method introduced earlier in Chapter 4 to numerically verify the exact distributions of the following discordancy test statistics that can be used to test the slippage alternative H_r when we are given random sample $\mathbf{X} = (X_1, \ldots, X_n)$, where each X_i is known to be independent exponential with location parameter. For this simulation study, we have focused on the following three test statistics:

$$D_r(\mathbf{X}) = \frac{X_{(n)} - X_{(n-r)}}{X_{(n)}},$$
$$R_r(\mathbf{X}) = \frac{X_{(n-r)} - X_{(1)}}{X_{(n)} - X_{(n-r+1)}},$$
$$Z_r(\mathbf{X}) = \frac{X_{(n-r)} - X_{(1)}}{\sum_{j=n-r+1}^{n} X_{(j)} - X_{(1)}},$$

where Z_r is first introduced by Zerbet and Nikulin in [6]. Its correct distribution under H_r has been given in Theorem 3.1. The statistic R_r was introduced by Jabbari Nooghabi in [7] as $JZ_r = 1 - \tilde{R}_r$, where JZ_r is the statistic they proposed and $\tilde{R}_r$ is given by

$$\tilde{R}_r(\mathbf{Y}) = \frac{\ln\left(Y_{(n-r)}\right) - \ln\left(X_{(1)}\right)}{\ln\left(Y_{(n)}\right) - \ln\left(Y_{(n-r+1)}\right)}.$$

The corrected distribution of $\tilde{R}_r$ under H_r has been presented in Theorem 3.3. The statistic D_r is the classical Dixon statistic, which we use as a reference for power comparison.

Motivated by Corollary 2.1, we also consider the following test statistics:

$$\tilde{D}_r(\mathbf{Y}) = \frac{\ln\left(Y_{(n)}\right) - \ln\left(Y_{(n-r)}\right)}{\ln\left(Y_{(n)}\right)},$$
$$\tilde{Z}_r(\mathbf{Y}) = \frac{\ln\left(Y_{(n-r)}\right) - \ln\left(Y_{(1)}\right)}{\sum_{j=n-r+1}^{n} \ln\left(Y_{(j)}\right) - \ln\left(Y_{(1)}\right)}$$

Then, it follows from Corollary 2.3 that $\tilde{D}_r$, $\tilde{R}_r$ and $\tilde{Z}_r$ will have the same distribution as D_r, R_r and Z_r if $\mathbf{Y} = (Y_1, \ldots, Y_n)$, where each Y_i is Pareto random variable with scale parameter θ and shape parameter the same as the scale parameter of X_i. Therefore, $\tilde{D}_r$, $\tilde{R}_r$ and $\tilde{Z}_r$ can be used to test H_r for Pareto random samples, and they will have the same critical values and power as D_r, R_r and Z_r respectively.

To avoid confusion, in this chapter, we assume the number of contanminants is r, observation sample size is n for both exponential and Pareto case. When consdering slippage alternative, F and $\overline{F}$ in Definition 1.2 are taken to be $\text{Exp}(\theta)$ and $\text{Exp}(\theta/b)$, respectively, for exponential case; F and $\overline{F}$ are taken to be $\text{Pareto}(\alpha, \theta)$ and $\text{Pareto}(\alpha b, \theta)$,

respectively, for Pareto case. The parameter b is assumed to be a unknown real number in $(0,1)$; α and θ are both fixed but unknown. As done in [6], we taken sample size n to ve 12 and number of contanminants as r to be 3 for the illustrative power comparison.

4.1 Numerical Verification of the Exact Distributions

In Chapter 3, we have derived the exact distribution of $\tilde{R}_r$ in Theorem 3.3 and the exact distribution of Z_r in Theorem 3.1, we will numerically verify them through simulation studies, where simulation sample size is $B = 5000$. The sample generated are depicted in Figure 4.1. As can be seen from the figure, the theoretical distribution agrees with the empirical distribution.

It follows from Theorem 2.2 that the standard error of empirical distribution $\hat{F}(x)$ is estimated by

$$\text{se}(\hat{F}(x)) = \sqrt{\frac{1}{B}\hat{F}(x)(1-\hat{F}(x))},$$

and the 95% confidence intervals are obtained by assuming that the distribution of $F(x)-\hat{F}(x)$ follows normal, and so we have the 95% normal CI for $F(x)$ to be

$$\hat{F}(x) \pm \sqrt{\frac{1}{B}\hat{F}(x)(1-\hat{F}(x))}.$$

Similarly, we generated Pareto samples and compare the theoretical cdf and ecdf and verify the theoretical distribution of $\tilde{R}_r$ given in Theorem 3.3 in Figure 4.2. Since Pareto distribution have a very long tail, we have taken the log of the order statistics in Panel A for illustrataion purpose. One may notice that Panel A of Figure 4.2 and Figure 4.1 are exact the same, and this is not a coincidence. It follows from the Corollary 2.3 that the logarithm distribution of the order statistic is the same as that of exponential distribution with appropriate parameter, and we have used the same seed for all the simulation study.

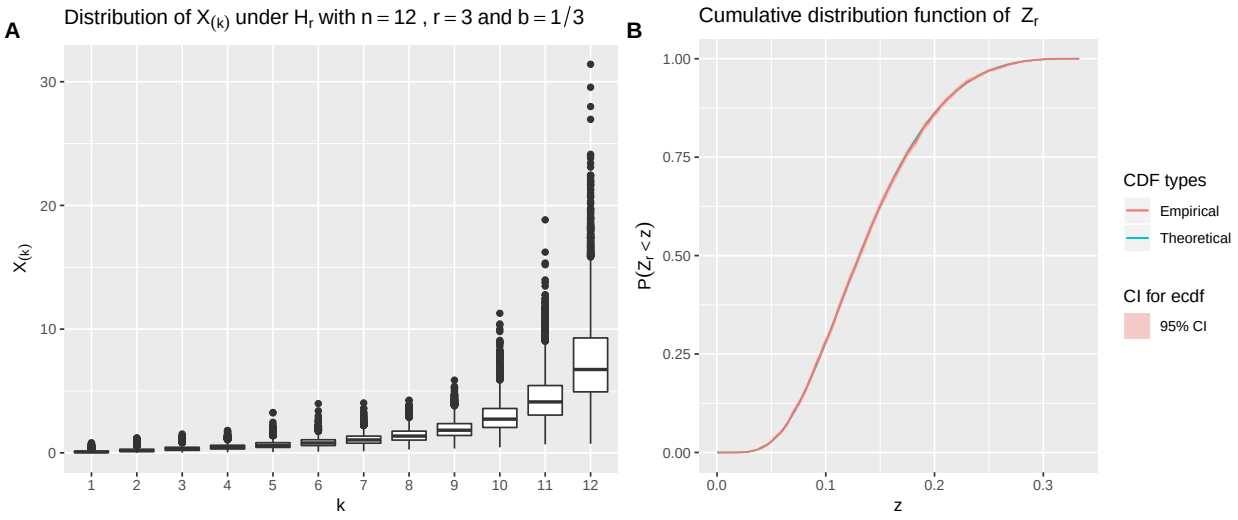

Figure 4.1: Distribution of Z_r under H_r with $n = 12, r = 3, b = 1/3$.

A The box plot of order statistics of the exponential slippage sample generated

B Empirical distribution function and the theoretical distribution function derived in Theorem 3.1

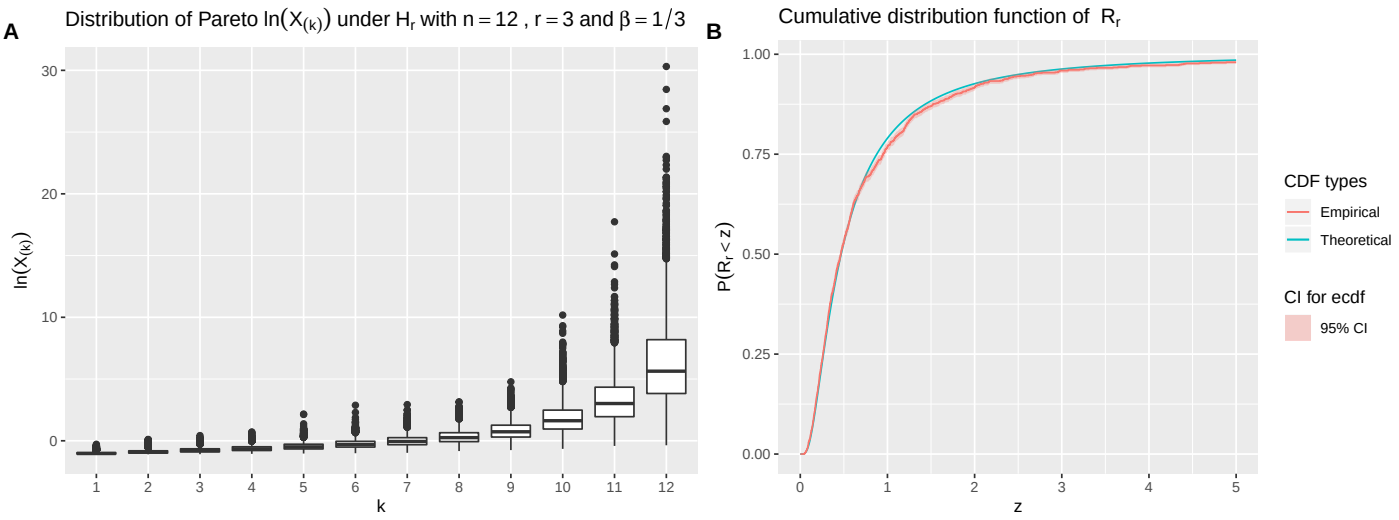

Figure 4.2: Distribution of $\tilde{R}_r$ under H_r with $n = 12, r = 3, \beta = 1/3$.

A The box plot of order statistics of the logarithm of the Pareto samples generated

B Empirical distribution function and the theoretical distribution function derived in Theorem 3.3

4.2 Performance of Test Statistics Under H_0

Here, we examine the performance of the statistics Z_r and $\tilde{R}_r$ under H_0. In particular, we examine the distribution of p-values of the test and verify that the type-I error has been properly controlled. In our study, we compared Z_r with Dixon statistic D_r and $\tilde{D}_r$ for power comparison based on 5000 samples, but we did not obtain the exact distribution of D_r. So, it is not necessary or possible to check the type-I error rate of D_r or $\tilde{D}_r$, since the critical value and distribution of D_r and $\tilde{D}_r$ are estimated with simulation. The distribution of the test statistics and empirical p-values for Z_r are summarized in Figure 4.4. Since we are considering the upper slippage alternative, the p-value is defined to be $\mathbb{P}\left(Z_r < z \,|\, H_0\right)$ for any observed sample statistic z.

As can be seen from Panel B of Figure 4.4, the empirical distribution of p-values is seen to be similar to the random samples of standard uniform distribution, which is the expected result. We fix the type-I error rate to be $\alpha = 0.05$, and use empirical p-values to estimate the actual type-I error rate by

$$\hat{\alpha} = \sum_{i=1}^{B} \frac{1}{B} \mathbb{1}_{\{p_i < 0.05\}}$$

which was found to be $\hat{\alpha} = 0.0522$, which is also acceptable.

We performed the exact the same procedure to examine the distribution of $\tilde{R}_r$ for the slippage samples. Since the distribution of the test statistic $\tilde{R}_r$ is right skewed, the histogram in Panel A is for the sample of $\tilde{R}_r$ after the logarithm transformation. The panel B of the figure shows no contradiction to that empirical p-values follow uniform distribution, which is also acceptable. With the same method, we found the estimated type-I error rate to be $\hat{\alpha} = 0.048$, which is also close.

4.2.1 Critical Values

Since there are mistakes in the derivations of distributions in [6] and [7], here we give tables of the critical values of Z_r and D_r for testing upper outliers with significance level $1 - \alpha$ with $\alpha = 0.05$. The critical values for $\tilde{R}_r$ and Z_r are obtained by finding the root of $F(x) = \alpha$ for x with bisection method, where F is the distribution functions derived in Theorem 3.1 and Theorem 3.3. The critical value for the Dixon statistic was obtained with nonparametric bootstrap method introduced in Section 2.3.3, and the estimation bias has been taken account to by (2.1). A table of standard

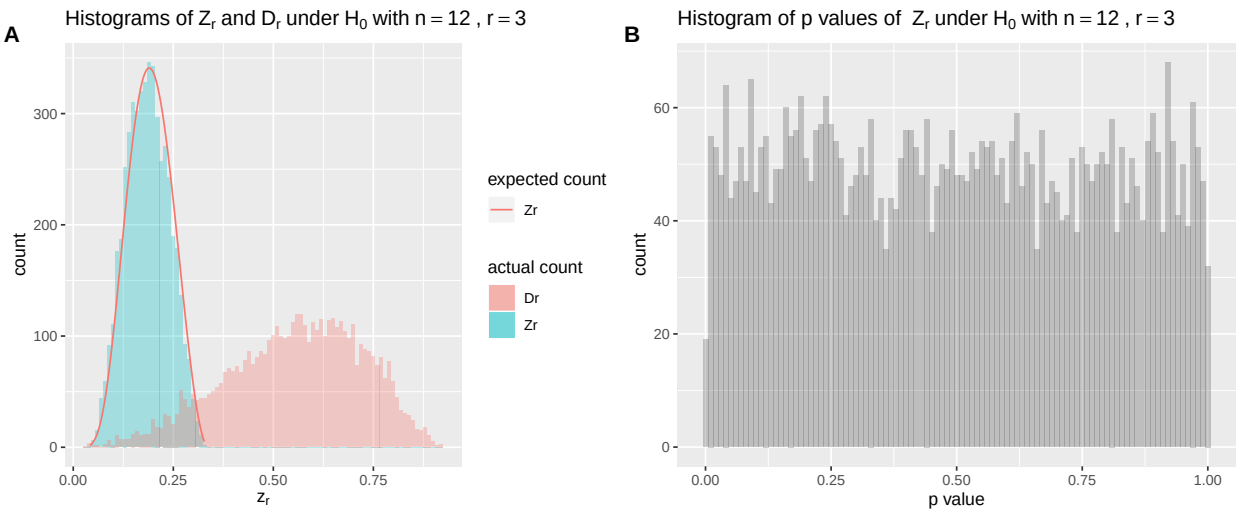

Figure 4.3: Distributions of the test statistics D_r and Z_r under H_r with $n = 12, r = 3, b = 1/3$.

A The histogram of Z_r and D_r; the expected count number of Z_r is estimated with the density of Z_r given in Theorem 3.1

B The distribution of empirical p values, where each empirical value p_i is defined to be $\mathbb{P}\left(Z_r(z_i) < z_{r,\alpha} \mid H_0\right)$, where z_i are the samples generated under H_r.

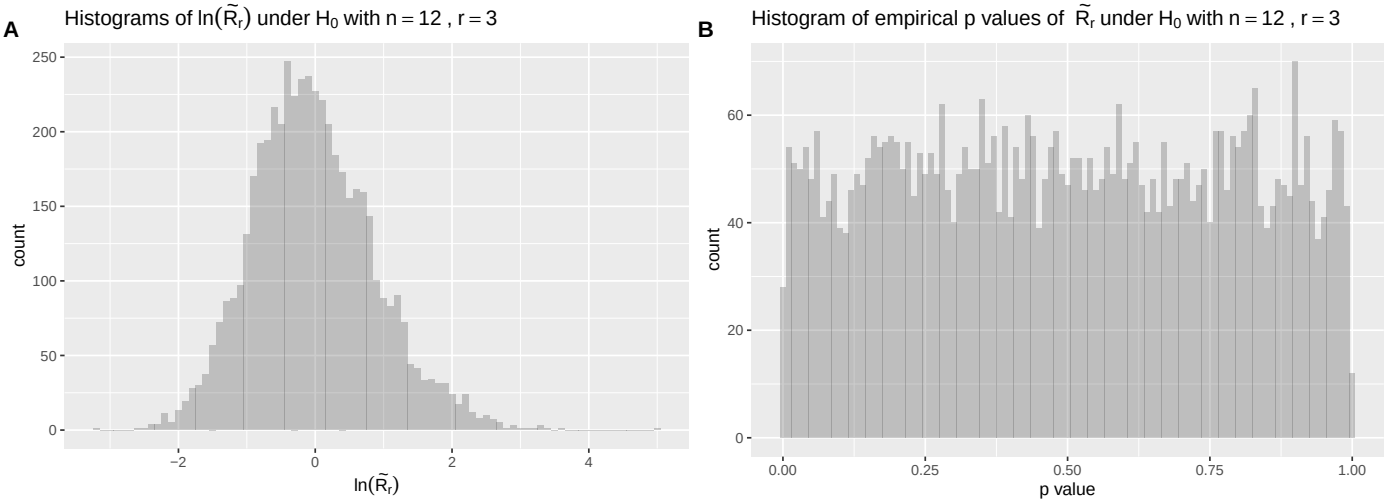

Figure 4.4: Distributions of the test statistic of $\ln(\tilde{R}_r)$ under H_r with $n = 12, r = 3, b = 1/3$.

A The histogram of $\ln(\tilde{R}_r)$

B The distribution of empirical p values, where each empirical value p_i is defined to be $\mathbb{P}\left(\tilde{R}_r(z_i) < r_{k,\alpha} \mid H_0\right)$, where z_i are the samples generated under H_r.

error of the obtained estimates is calculated by (2.2) and presented in Table 4.4. As can be seen from Table4.4, the critical values we obtained are accurate to three decimal places, and this agrees with the result given in [6].

Table 4.1: The critical values of Z_r with $\alpha = 0.05$

n	r					
	1	2	3	4	5	6
6	0.2179255	0.07271396	0.02257252	0.002554801		
7	0.2541362	0.09761256	0.04158413	0.014258365	0.001703935	
8	0.2827005	0.11738195	0.05767611	0.027187769	0.009843320	0.001217544
9	0.3059432	0.13338088	0.07094024	0.038625121	0.019246229	0.007211060
10	0.3253324	0.14660659	0.08194491	0.048345371	0.027852283	0.014371925
11	0.3418340	0.15775129	0.09119986	0.056587249	0.035351931	0.021105592
12	0.3561090	0.16729823	0.09909488	0.063629961	0.041830932	0.027096803

Table 4.2: The estimated critical values of D_r for $\alpha = 0.05$ with Monte Carlo method

n	r					
	1	2	3	4	5	6
6	0.7451293	0.8613298	0.9295339	0.9721648		
7	0.7174043	0.8333060	0.8997864	0.9454283	0.9782023	
8	0.6937633	0.8084582	0.8758362	0.9217053	0.9569222	0.9819938
9	0.6748915	0.7878169	0.8512355	0.9002023	0.9363351	0.9643261
10	0.6572173	0.7696995	0.8354201	0.8819363	0.9175486	0.9458965
11	0.6438796	0.7539956	0.8176284	0.8643931	0.9012357	0.9296735
12	0.6313994	0.7392545	0.8037565	0.8488094	0.8850763	0.9146531

Table 4.3: The estimated critical values of R_r for $\alpha = 0.05$

n	r					
	1	2	3	4	5	6
6	0.1501963	0.04632501	0.00565403			
7	0.2092279	0.08857200	0.03191082	0.004138095		
8	0.2607984	0.12798613	0.06301878	0.024141817	0.003232332	
9	0.3062225	0.16364416	0.09308207	0.048699504	0.019287215	
10	0.3466706	0.19582426	0.12095287	0.073033005	0.039504091	
11	0.3830610	0.22499710	0.14656179	0.096008036	0.059916655	
12	0.4160997	0.25160775	0.17010112	0.117415965	0.079466899	

Table 4.4: The standard errors of the estimate of critical values of D_r for $\alpha = 0.05$

n	r					
	1	2	3	4	5	6
6	0.0003654004	0.0006721107	0.0004517287	0.0002250101		
7	0.0003930877	0.0007751294	0.0005556101	0.0003343549	0.0001549828	
8	0.0004040385	0.0008054345	0.0005574572	0.0003964377	0.0002388118	0.0001299860
9	0.0003597688	0.0008568465	0.0006336613	0.0004514928	0.0003180044	0.0002395700
10	0.0003644383	0.0008356480	0.0006315783	0.0004870168	0.0004525706	0.0002701931
11	0.0003900676	0.0008706266	0.0007761038	0.0005363027	0.0004374640	0.0003245950
12	0.0003680189	0.0008683104	0.0007025402	0.0005738111	0.0004305760	0.0003728830

4.3 Power Estimation

With the critical values obtained, we can now compare the power values of the test statistics. Here we compare the power of Z_r and $\tilde{R}_r$ with the Dixon statistic and the result is given in Figure 4.5. The power of D_r are given by $\gamma_{Dr}(b_0) = \mathbb{P}(D_r \leqslant d_{r,1-\alpha} \mid b = b_0)$, where $d_{r,1-\alpha}$ is the quantile of R_r under H_0, since clearly D_r will tend to 0 as the parameter b tends to 0. Similarly, one can show that R_r and Z_r will tend to 0 as b tends to 0, hence their power are given by $\gamma_{Rr}(b_0) = \mathbb{P}(Z_r > z_{r,1-\alpha} \mid b = b_0)$ and $\gamma_{Z_r}(b_0) = \mathbb{P}(R_r > z_{r,1-\alpha} \mid b = b_0)$.

The power of $\tilde{R}_r$ and Z_r can be obtained directly since we have derived their distributions; the power of Dixon statistic are obtained by Monte Carlo method, where each data point is calculated with 5000 simulated samples. It follows from Corollary 2.3 that the test statistics we have investigated in this study, Z_r, R_r and Dixon statistic, can be used to test the slippage alternative for both exponential and Pareto samples, and we have summarized their power in Figure 4.5. As can be seen from the figure, the Dixon statistic is superior to other two test statistics in terms of power.

Jabbari Nooghabi also suggested the following test statistic [7] to test H_r for Pareto samples:

$$T = \frac{Y_{n-r+1} - Y_{(n-r)}}{\hat{Y}_{(n-r+1)} - Y_{(n-r)}},$$

where $\hat{Y}_{(n-r+1)}$ is the BLU estimator of $Y_{(n-r+1)}$ proposed in [11], and Y has the two-parameter exponential distribution such that $X = e^Y$ will yield the desired Pareto random variables. However, since the sample size n we have considered are relatively small and $\hat{Y}_{(n-r+1)}$ is justified with asymptotic theory, it is not appropriate to consider here in this study.

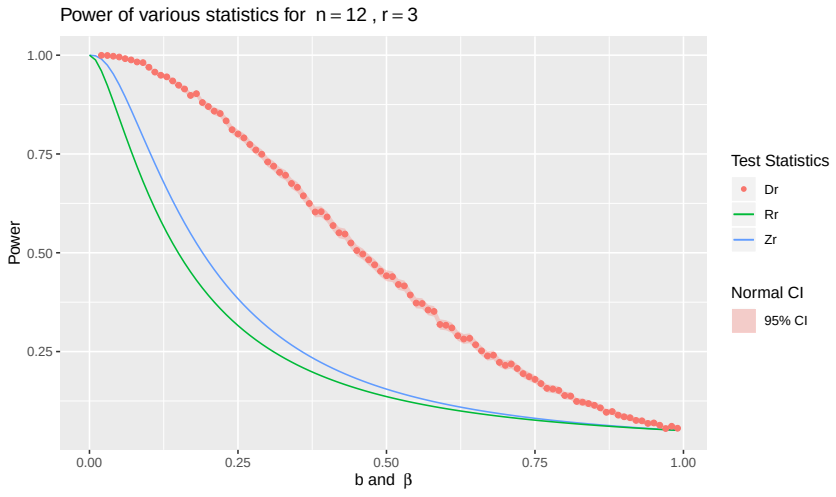

Figure 4.5: Power comparison for various test statistics with $n = 12, r = 3, b = 1/3$.

4.4 Numerical Example

Here we present a illustrative example based on a real-life dataset to evaluate the performance of the tests. We considered Haberman's survival dataset [12], which comes from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. Haberman's survival dataset records the number of positive axillary nodes detected for 305 patients and whether they are still alive 5 years after the surgery, as depicted in Figure 4.6. According to the literature, number of positive axillary nodes are postively correlated to the prognosis of breast cancer [13], so the distribution of detected positive axillary nodes of patients are expected to be different between people who survived and thoese did not. In fact, according to Haberman's survival dataset, the average detected number for survived and passed away patients are found to be 2.80 and 7.46, respectively.

As Figure 4.6 suggests, the distribution of detected positive axillary nodes can be modeled by exponential distribution, with mean different based on the survival status of patients. To examine the performance of the test, we randomly select 12 patients at random. Since 81 out of 305 patient passed away 5 years after the treatment, we can assume 3 out of 12 patient we have drawn are those passed away. So, we should take $n = 12$ and $r = 3$ when construting the slippage alternative hypothesis.

To examine the power of the tests, with each drawn sample $\mathbf{x}_0$, we then calculate the test statistics $R_3(\mathbf{x}_0)$, $D_3(\mathbf{x}_0)$ and $Z_3(\mathbf{x}_0)$. By checking Table 4.2, Table 4.3 and Table 4.1, we will reject H_0 with D_r if $D_3(\mathbf{x}_0) > 0.80375$, with R_r if $R_3(\mathbf{x}_0) < 0.17010$ and with Z_r if $Z_3(\mathbf{x}_0) < 0.09909$.

We have taken 1000 random samples from Haberman's survival dataset and conducted a hypothesis testing with each drawn sample as described. We found that, the hypothesis testing procedures based on D_r, Z_r and R_r have successfully reject H_0 for 475 and 454 and 285 times among 1000 trials conducted, respectively. So, the empirical power comparison based on this real dataset is consistent with the power estimation given in Figure 4.5.

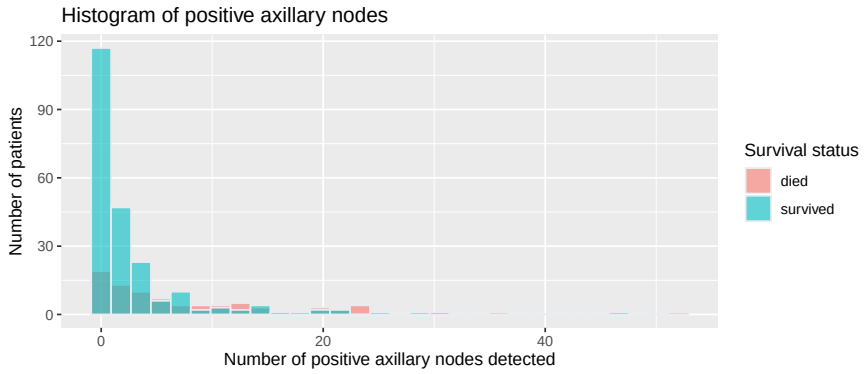

Figure 4.6: Histogram of number positive axillary nodes detected by the survival status of the patients

4.5 Discussions and Conclusions

In this study, we have characterized the distribution of order statistics under smooth monotone transformation in Theorem 2.1. Then we have given methods that transform the test for slippage alternative H_r for exponential samples to the Pareto case. Next, we have investigated some test statistics proposed by Zerbet and Nikulin [6] and Jabbari Nooghabi [7] and derived their distributions under H_r. Besides correcting mistakes in the derivation in [7], the usage of Corollary 2.1 significantly simplified the proof for the distribution of $\tilde{Z}_r$. Finally, we proposed have a modified Dixon statistic $\tilde{D}_r$ that can be used for testing the slippage alternative for the Pareto case.

Through simulation study, we have given a corrected usable table for critical values for all test statistics investigated and have conducted a power comparison for the case with sample size $n = 12$, contamination number $r = 3$ for various choices of parameter b. The simulation study suggests that the Dixon statistics has the best performance in terms of power. However, we did not obtain the theoretical distribution of Dixon statistic. Moreover, the test statistic Z_r and R_r did controlled the type-I error rate under the null hypothesis.

Although there are some mistakes in two papers by Zerbet and Nikulin [6] and Jabbari Nooghabi [7], they do shed some light on the treatment on testing slippage alternative hypothesis. The novelty of the work by Zerbet and Nikulin lies in the fact that they provide a method to derive the distribution of some function of order statistics with the use of characteristic function. In particular, they used Lemma 2.4 to transform the problem of derivation the joint distribution of order statistics into that of finding the distribution of independent exponential random variables, which reduces the complexity of the task. Although not explicitly stating the result as we have done, Jabbari Nooghabi used the relationship between exponential and Pareto distributions to define some test statistics that can be used to test the slippage alternative for the case of Pareto samples, his method motivated us to define the modified Dixon statistic $\tilde{D}_r$.

For future work, one could use the method introduced in [6] to obtain the distribution of D_r. Another challenge is, how to determine whether it is appropriate to assume slippage alternative hypothesis. As we have pointed out in Figure 2.1, the slippage alternative is appropriate only if the number of contanminants is relatively small and mean difference between the contaminated samples and the samples assumed coming from null hypothesis is large enough.

Bibliography

[1] Frank E Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[2] Zakia Ferdousi and Akira Maeda. Unsupervised outlier detection in time series data. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages x121–x121. IEEE, 2006.

[3] Scott S Norton, Jorge Vaquero-Garcia, Nicholas F Lahens, Gregory R Grant, and Yoseph Barash. Outlier detection for improved differential splicing quantification from rna-seq experiments with replicates. *Bioinformatics*, 34(9):1488–1497, 2018.

[4] Vic Barnett and Toby Lewis. *Outliers in statistical data*. Chichester, England, 2 edition, 1993.

[5] Bernard Rosner. On the detection of many outliers. *Technometrics*, 17(2):221–227, 1975.

[6] Aicha Zerbet and Mikhail Nikulin. A new statistic for detecting outliers in exponential case. *Communications in Statistics-Theory and Methods*, 32(3):573–583, 2003.

[7] Mehdi Jabbari Nooghabi. On detecting outliers in the pareto distribution. *Journal of Statistical Computation and Simulation*, 89(8):1466–1481, 2019.

[8] Narayanaswamy Balakrishnan and Valery B Nevzorov. *A primer on statistical distributions*. John Wiley & Sons, Hoboken, New Jersey, 2004.

[9] MS Chikkagoudar and SH Kunchur. Distributions of test statistics for multiple outliers in exponential samples. *Communications in Statistics-Theory and Methods*, 12(18):2127–2142, 1983.

[10] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[11] Kenneth S Kaminsky and Paul I Nelson. Best linear unbiased prediction of order statistics in location and scale families. *Journal of the American Statistical Association*, 70(349):145–150, 1975.

[12] TS Lim. Haberman's survival data set. *UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA*, 1999.

[13] Bernard Fisher, Madeline Bauer, D Lawrence Wickerham, Carol K Redmond, Edwin R Fisher, Anatolio B Cruz, Roger Foster, Bernard Gardner, Harvey Lerner, Richard Margolese, et al. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. an nsabp update. *Cancer*, 52(9):1551–1557, 1983.