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

Fang Wang

McMaster University

April 9, 2020

Outline

* Introduction
* Discordancy tests
* Distribution of order statistics under monotone transformation
* Power comparision and real dataset application

Introduction: Outliers

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.

Introduction: H_0

Definition (null hypothesis of contamination model)

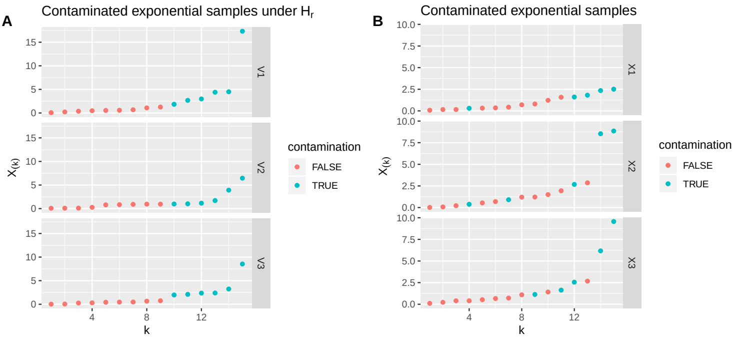
Let $x_1, \ldots, x_n$ be a sample of n observations. Then under the null hypothesis H_0, $x_1, \ldots, x_n$ are observations of $X_1, \ldots, X_n$, where $X_1, \ldots, X_n$ are independent random variables with common distribution F.

Introduction: H_r

Definition (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 $F \neq \overline{F}$.

Contamination Model



A: Contaminated samples with H_r

B: Contaminated samples without H_r

Exponential Distribution

Definition (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 $X \sim \mathrm{Exp}(\theta)$.

Poission Process

* Let $N(t)$ be a Poisson process with rate parameter $1/\theta$, and stage space $\{0,1\}$, then the sojourn time of N_t follows $\mathrm{Exp}(\theta)$ distribution.
* If $Y_1, \ldots, Y_k$ are iid exponential samples, they are also sojourn times of some Poisson processes $N_1, \ldots, N_k$.
* Let $N = N_1 + \ldots + N_k$, then $Y_{(i)} - Y_{(i-1)}$ are sojourn time of N in stage i_1

Pareto Distribution

Definition (Pareto distribution)

A random variable X follows $\mathrm{Pareto}(\alpha, \theta)$ distribution if its pdf is given by

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

where θ and α are both positive parameters.

Remark

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

H_r for Exponential and Pareto Distributions

Let α and θ be two positive real numbers and $b \in (0,1)$.

Exponential Case F is $\text{Exp}(\theta)$ and $\overline{F}$ is $\text{Exp}(\theta/b)$

Pareto Case F is $\text{Pareto}(\alpha, \theta)$, $\overline{F}$ is $\text{Pareto}(\alpha b, \theta)$

Simulate Exponential Sample Under H_r

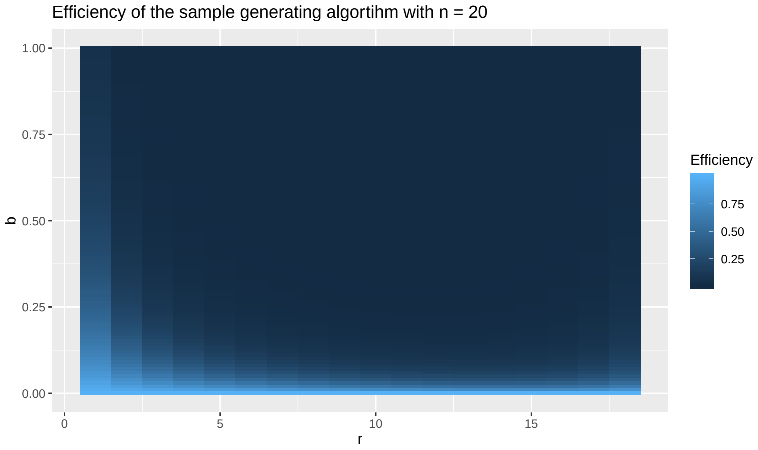
1. Generate $n - r$ observations from F and r observation from $\overline{F}$.
2. Combined total n observations into a single observation $\mathbf{x}$
3. Accept $\mathbf{x}$ if it satisfies H_r.

If F is $\mathrm{Exp}(\theta)$ and $\overline{F}$ is $\mathrm{Exp}(\theta/b)$,the acceptance probability is

$$\begin{aligned}\mathbb{P}\left(\text{Acceptance}\right) &= \mathbb{P}\left(\max\{X_1, \ldots, X_{n-r}\} < \min\{X_{n-r+1}, \ldots, X_n\}\right) \\ &= rbB(rb, n-r+1),\end{aligned}$$

where $B(r, s)$ is the complete beta function.

Acceptance Probability for Various Parameters



Discordancy Test Statistics for Exponential H_r

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

Discordancy Test Statistics for Pareto H_r

$$\tilde{D}_r(\mathbf{Y}) = \frac{\ln\big(Y_{(n)}\big) - \ln\big(Y_{(n-r)}\big)}{\ln\big(Y_{(n)}\big)},$$

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

$$\tilde{Z}_r(\mathbf{Y}) = \frac{\ln\big(Y_{(n-r)}\big) - \ln\big(Y_{(1)}\big)}{\sum_{j=n-r+1}^{n} \ln\big(Y_{(j)}\big) - \ln\big(Y_{(1)}\big)}.$$

Distributional Result

Theorem

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

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

Corollary

Let $X_1, \dots X_n$ and $E_1, \dots, E_n$ be independent random variables with $X_k \sim \mathrm{Pareto}(\alpha_k, \theta_k)$ and $E_k \sim \mathrm{Exp}(\alpha_k)$ 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 $\mathbf{Y} = (Y_1, \dots, Y_n)$.

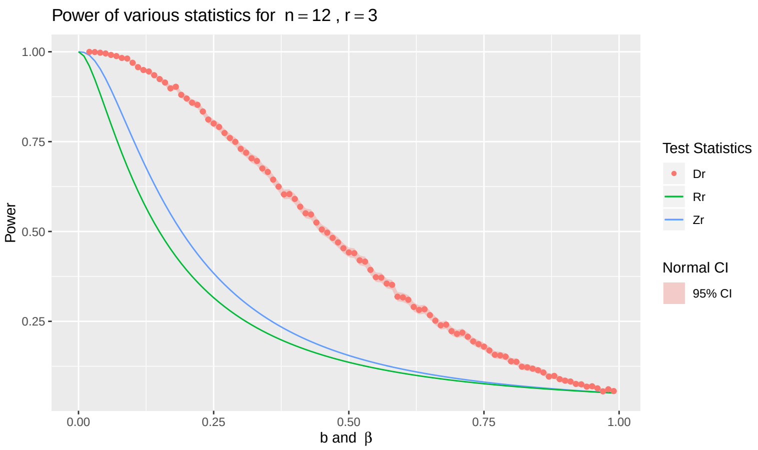
Remark

The conclusion holds under H_r.

Equivalence of Tests

* $D_r(\mathbf{Y}) \stackrel{\text{D}}{=} \tilde{D}_r(\mathbf{X})$, $Z_r(\mathbf{Y}) \stackrel{\text{D}}{=} \tilde{Z}_r(\mathbf{X})$ and $R_r(\mathbf{Y}) \stackrel{\text{D}}{=} \tilde{R}_r(\mathbf{X})$ under H_r.
* Statistics tests based on D_r, R_r and Z_r would have the same power and critical values as $\tilde{D}_r$, $\tilde{R}_r$ and $\tilde{Z}_r$, respectively.
* statistics used for testing slippage alternative hypothesis of exponential samples can be easily adapted to test for the Pareto case

Power of Tests



Real Dataset Application

* Haberman's survival dataset
* Mean of parameter of two distribution: $\theta_1 = 2.80, \theta_2 = 7.46,$
* Sample size: $N_1 = 85, N_2 = 244$
* Estimated power:
  $\gamma(\hat{D}_r) = 0.475, \gamma(\hat{Z}_r) = 0.455, \gamma(\hat{R}_r) = 0.28.$