Brno University of Technology

University of L'Aquila

Double-Degree Master's Programme - InterMaths
Applied and Interdisciplinary Mathematics

Master of Science **Mathematical Engineering**	**Master of Science** **Mathematical Engineering**
Brno University of Technology (BUT)	University of L'Aquila (UAQ)

Master's Thesis

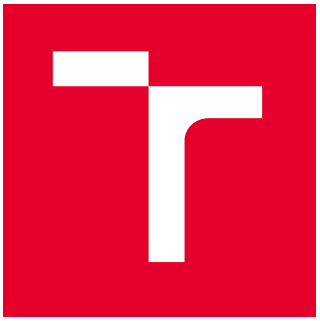
Properties of Cauchy Distribution and their Applications

Supervisor
doc. Mgr Zuzana Hübnerová,

Candidate
Samson Quaye

Student ID (UAQ): 274464
Student ID (BUT): 243843

Academic Year 2021/2022



BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF MECHANICAL ENGINEERING

FAKULTA STROJNÍHO INŽENÝRSTVÍ

INSTITUTE OF MATHEMATICS

ÚSTAV MATEMATIKY

PROPERTIES OF CAUCHY DISTRIBUTION AND THEIR APPLICATIONS

VLASTNOSTI CAUCHYHO ROZDĚLENÍ A JEJICH UŽITÍ

MASTER'S THESIS

DIPLOMOVÁ PRÁCE

AUTHOR — **Samson Quaye**

AUTOR PRÁCE

SUPERVISOR — **doc. Mgr. Zuzana Hübnerová, Ph.D.**

VEDOUCÍ PRÁCE

BRNO 2022

BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Mechanical Engineering

MASTER'S THESIS

Brno, 2022 Samson Quaye

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF MECHANICAL ENGINEERING

Assignment Master's Thesis

Institut:	Institute of Mathematics
Student:	**Samson Quaye**
Degree programm:	Mathematical Engineering
Branch:	
Supervisor:	**doc. Mgr. Zuzana Hübnerová, Ph.D.**
Academic year:	2021/22

As provided for by the Act No. 111/98 Coll. on higher education institutions and the BUT Study and Examination Regulations, the director of the Institute hereby assigns the following topic of Master's Thesis:

Properties of Cauchy Distribution and their Applications

Brief Description:

Cauchy distribution has specific properties, which imply the unsuitability of some statistical methods of parameter estimates and inference on them. However, this distribution is applied in technical areas.

Master's Thesis goals:

Study of the properties of Cauchy distribution.
Application in real or simulated data analysis.

Recommended bibliography:

CASELLA, G., BERGER, R. L. Statistical Inference. Belmont, CA: Duxbury, 2001. ISBN 0-534-24312-6.

Deadline for submission Master's Thesis is given by the Schedule of the Academic year 2021/22

In Brno,

L. S.

prof. RNDr. Josef Šlapal, CSc. Director of the Institute	doc. Ing. Jaroslav Katolický, Ph.D. FME dean

Abstract

The Cauchy distribution plays a major role in biology, mathematics, physics and many related disciplines. As a consequence, a parameter estimation methodology for data which is distributed according to a Cauchy distribution is of importance. Nevertheless, the Cauchy distribution is well known for causing difficulties with classical approaches to parameter estimation. This diploma thesis is concerned with the study of the properties of Cauchy distribution. Several robust estimations of its location parameter are presented. A simulation study programmed in Python allows us to compare the performance of these parameter estimates in Bi-Cauchy ROC curve estimation. Moreover, some theoretical properties of the Bi-Cauchy ROC curve are studied.

Keywords

Cauchy distribution, Parameter estimation, Huber M-estimator, Hodges Lehmann estimator, Trimmed Mean, Sens Mean, ROC curves.

Quaye. S.: *Properties of Cauchy Distribution and their Applications*, Brno University of Technology, Faculty of Mechanical Engineering, 2022. 53 s. Supervisor: doc.Mgr. Zuzana Hübnerová, Ph.D.

I declare that I have worked on this thesis independently under a supervision of doc. Mgr. Zuzana Hübnerová, Ph.D. and using the sources listed in the bibliography.

Samson Quaye

I would like to express my gratitude to my supervisor doc. Mgr. Zuzana Hübnerová, Ph.D. for being a window through which I peeped into the world of statistics, for her outstanding education, suggestion and patience to me. Without her patience and sacrifice of time I would not have completed this thesis. To my Mom and Dad, words will fail to convey my gratitude for your ever present help and love! thank you so much for your relentless prayers‘ and support. I want to thank my classmates, Joshua Adeleke, Asante Michael, Moshood Olarewaju, Tunmbi Okediran for their friendship and encouragement. To almighty God for his grace and blessings, I’m utmost grateful.

Samson Quaye

Contents

1 INTRODUCTION

Cauchy distribution is a distribution which has a bell-shaped density function with heavier tails which are symmetric and unimodal. The distribution was named after a French mathematician, Augustin-Louis Cauchy (1789–1857). In recent times, Cauchy distribution has been of interest and many controversies has risen from it due to some properties it possesses making it difficult to estimate its parameters, hence mathematical approaches to it is quite hard. It appears that the distribution may look like the normal distribution but its tails does not recede quickly as those of the normal distribution. There is a great difference between the Cauchy distribution and the normal distribution indeed, as the ratio of two independent standard normal random variable gives a Cauchy random variable (details of this can be found in [1]). Also, the standard Cauchy random variable has a student's t-distribution with one degree of freedom. The shape of the density function is sometimes called the witch of Agnesi [2]. In many fields of study fat tailed distributions like the Cauchy distribution never arise but in other fields they do arise in diverse ways. For example, in biology, the distribution that seeds end up from the parent plant may have a "fat-tailed" distribution if most fall by plant are dispersed by wind, water or animals. It can be shown in quantum mechanically that for a state which decomposes exponentially with time, the spread of energy throughout the state is expressed by Cauchy distribution [3]. [4] showed another fascinating application of the distribution in the sense that, he described the distribution of hypo-centers on focal spheres of earthquakes. Cauchy distribution can again be used to analyse polar and non-polar liquids in porous glasses [5].

From the study of the probability distribution, a distribution function is completely determined by its moment generating function in case it exists. The most common moments are the mean $\mu = E(X)$ and the variance $E(X^2) - \mu^2$ which shows how spread out the distribution is, and other higher moments like the skewness and kurtosis. In the case of Cauchy distribution there exist no mean, variance or moment generating function. In addition, the characteristic function(cf) and the cumulative distribution function(cdf) exist. The approach of finding the characteristic function and the cumulative distribution function is by using its probability density function(pdf). The central moments are undefined, this is due to the long tails of the density function. The pdf of $C(\mu, \lambda)$ is

$$f_X(x) = \frac{\lambda}{\pi\left(\lambda^2 + (x-\mu)^2\right)} \quad -\infty < x < \infty \tag{1.1}$$

where $\mu \in (-\infty, \infty)$ and $\lambda > 0$.

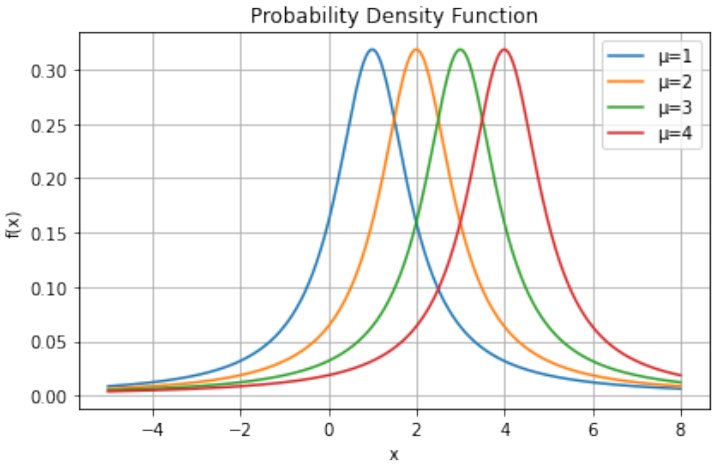


Figure 1: PDF of Cauchy Distribution with different location parameters and a constant scale parameter $\lambda = 1$

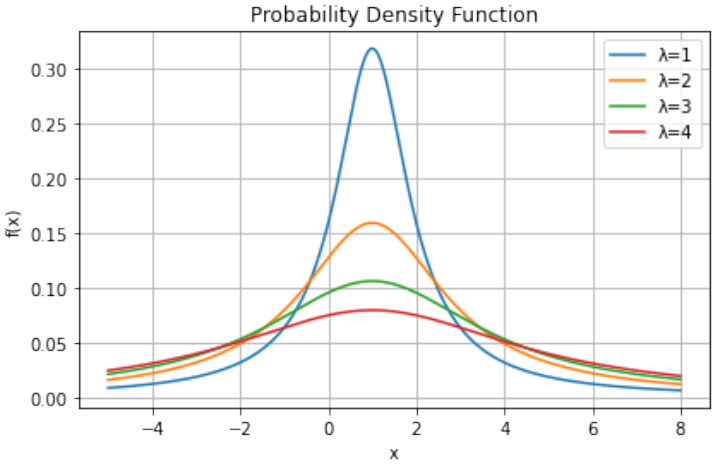


Figure 2: PDF of Cauchy Distribution with different scale parameter and constant location parameter $\mu = 1$

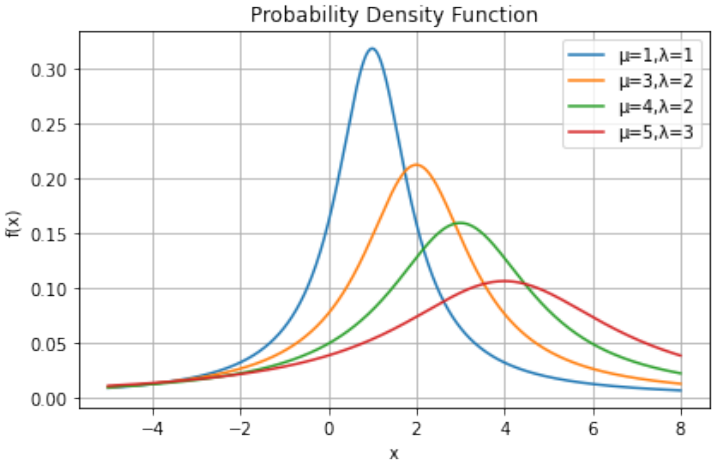


Figure 3: PDF of Cauchy Distribution with different and location and scale parameter

1.1 Motivation

The first objective of the thesis is to study the properties of Cauchy distribution. The second goal is to apply the distribution in real life or simulated data analysis. Owing to the fact that the distribution possesses some properties which makes estimation of its location parameter a difficult task, we shall study new estimation approaches which are based on order statistics called the robust estimators. They are Huber M-estimate, Hodges Lehmann estimate, Trimmed Mean, Winsorized Mean and the Sen's Mean.

We will then apply it in receiver operating characteristic (ROC) curves where we will build a parametric ROC model for variables with Cauchy distribution and estimate its parameters using the robust estimators. Conclusions will be made after different set of simulations are made to know which of the estimators provides the most suitable estimate of the true ROC.

1.2 Structure of thesis

The thesis work is organised as follows: Theoritical part and practical part.

- The second section is allocated to studying some properties of the distribution and justifying them with proofs.
- The third section is allocated to parameter estimation where we study the different robust estimators and their theoretical approaches. This section is mainly based on the book [6].
- The fourth section is also allocated to the receiver operating characteristic (ROC) curves where we discuss its properties and operation and derive the ROC model based on Cauchy distribution. Some theoretical part and assumptions made in this section was taken from the book [7] and [8].
- The fifth section is allocated to simulation of data in Python using the derived ROC model in section four. The parameters of the model are estimated by the different types of the robust estimators.
- The last section is devoted to main conclusions of the work.

2 PROPERTIES OF CAUCHY DISTRIBUTION

2.1 Expected Value

Expected value or mean of a random variable $g(X)$, denoted by $Eg(X)$ is

$$Eg(X) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \chi} g(x) f_X(x), & \text{if } X \text{ is discrete} \end{cases}$$

as long as the integral or sum exists. On condition that $E|g(X)| = \infty$, we say that Eg(X) does not exist [1].

Proposition 2.1. *The standard Cauchy random variable is a random variable with location parameter $\mu = 0$ and scale parameter $\lambda = 1$, denoted by $C(0, 1)$. It is the type whose expected value does not exist, that is, one with probability density function*

$$f_X(x) = \frac{1}{\pi(1 + x^2)} \qquad -\infty < x < \infty \tag{2.1}$$

It is trivial to check that $\int_{-\infty}^{\infty} f_X(x)dx = 1$, but $E|X|$ does not exist.

Proof.

$$E(X) = \int_{-\infty}^{\infty} \frac{x}{\pi(1 + x^2)} dx = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x}{(1 + x^2)} dx,$$

but this integral does not converge. The support of a Cauchy random variable is $-\infty$ to ∞, so the integral representing the $E(X)$ is an improper integral, i.e., the limit of the definite integral of $xf(x)$ as the limits of integration approach minus infinity and plus infinity, the integral is undefined. Thus,

$$\mathbb{E}[X] = \lim_{L \to -\infty} \int_L^0 x f_X(x) dx + \lim_{U \to \infty} \int_0^U x f_X(x) dx \tag{2.2}$$

$$\mathbb{E}[X] = \lim_{L \to -\infty} \int_L^0 x \frac{1}{\pi(1 + x^2)} dx + \lim_{U \to \infty} \int_0^U x \frac{1}{\pi(1 + x^2)} dx \tag{2.3}$$

$$\mathbb{E}[X] = \lim_{L \to -\infty} \left[\frac{1}{2\pi} \ln(1 + x^2) \right]_L^0 + \lim_{U \to \infty} \left[\frac{1}{2\pi} \ln(1 + x^2) \right]_0^U \tag{2.4}$$

$$\mathbb{E}[X] = \lim_{L \to -\infty} -\frac{1}{2\pi} \ln(1 + L^2) + \lim_{U \to \infty} \frac{1}{2\pi} \ln(1 + U^2) \tag{2.5}$$

$$\mathbb{E}[X] = -\infty + \infty \tag{2.6}$$

From (2.6) we can see that $E|X|$ does not exist, this is because both integrals are not finite. In other words $-\infty + \infty$ is not a number. Note that, the "tails" of the pdf decay as $(\frac{1}{x^2})$, and hence very large outcomes are possible. Since $E|X|$, does not exist it follows that no moments of the Cauchy distribution exist or, in other words, all absolute moments equals ∞. Particularly, the moment generating function does not exist. □

2.2 Characteristic Function

Definition 2.2. The characteristic function of a random variable X is

$$\begin{aligned}\phi(t) &= E[e^{itX}] \\ &= E[\cos(tX) + i\sin(tX)]\end{aligned} \tag{2.7}$$

$\forall t \in \mathbb{R}$ and $i = \sqrt{-1}$. Since $|\cos(tX)|, |\sin(tX)| \leq 1$, the characteristic function is defined for all t for any random variable.

The characteristic function is defined in the discrete or continuous case where is defined as the Fourier transform of the probability function using the Fourier transform parameters. It can be written as

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itX} dP_x$$

If X is a continuous random variable with density function $f_X(x)$, then

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itX} f_X(x) dx$$

The advantage with the characteristic function is that it always exists, unlike the moment generating function, which can be infinite everywhere except $s = 0$.

Proposition 2.3. *Suppose X is a standard Cauchy variable, then its pdf is given by (2.1). The characteristic function can be written in closed form as*

$$\phi(t) = e^{-|t|} \quad t \in \mathbb{R}.$$

The above expression is not entirely trivial to obtain. On the other hand, we find here that the characteristic function for the Cauchy random variable exists everywhere. This is essentially because the integral defining the characteristic function converges absolutely, and hence uniformly, for all $t \in \mathbb{R}$. Characteristic functions are thus particularly useful in handling heavy-tailed random variables, for which the corresponding moment generating functions do not exist.

Proof. We show this by the Cauchy's residue theorem by evaluation of the contour integrals. We consider three cases of t thus, when $t = 0$, $t > 0$ and $t < 0$. Indeed, it requires considering two separate contour integrals for $t > 0$ and $t < 0$. The integral defining the characteristic function is given by

$$\begin{aligned}\phi(t) &= \int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)} dx \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{itx}}{(1+x^2)} dx.\end{aligned}$$

For the case when $t = 0$

$$\begin{aligned}\phi(0) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{i(0)x}}{(1+x^2)} dx \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{(1+x^2)} dx \\ &= \frac{1}{\pi} \left[\arctan(x)\right]_{-\infty}^{\infty} \\ &= 1.\end{aligned}$$

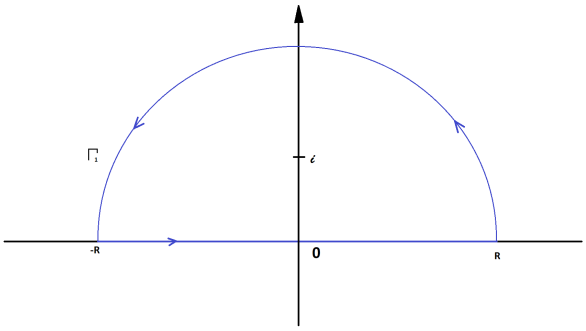


Figure 4: C_1, for $t > 0$

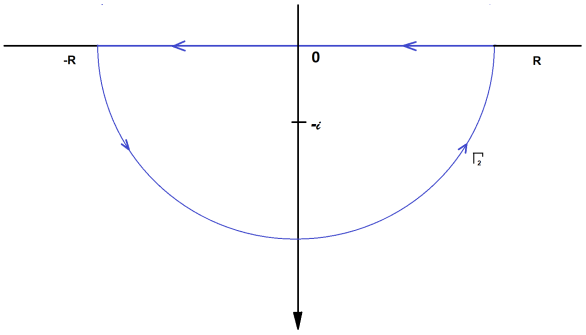


Figure 5: C_2, for $t < 0$

Let

$$f(z) = \frac{e^{itz}}{1+z^2}$$

the poles of $f(z)$ are solutions of $(z+i)(z-i) = 0$ where $z \in \mathbb{C}$

For $t > 0$, we consider Figure 4

$$\int_{C_1} f(z)dz = \int_{-R}^{R} f(z)dz + \int_{\Gamma_1} f(z)dz$$

The left hand side $\int_{C_1} f(z)dz$ we use the residue theorem, there's a pole inside C_1 at $z = i$ so we calculate the residue of the function $f(z)$ at $z = i$

$$\begin{aligned}
\int_{C_1} f(z)dz &= 2\pi i Res_{z=i} f(z) \\
&= 2\pi i \lim_{z\to i}(z-i)\frac{e^{itz}}{(z^2+1)} \\
&= 2\pi i \lim_{z\to i}(z-i)\frac{e^{itz}}{(z+i)(z-i)} \\
&= \pi e^{-t}
\end{aligned}$$

For $\int_{-R}^{R} f(z)dz$ we do not need to worry more about it because as $R \to \infty$ it approaches the original integral of the Cauchy distribution. Then for $\int_{\Gamma_1} f(z)dz$ we have

$$\left|\int_\Gamma f(z)dz\right| \leq \text{length}(\Gamma) \times \max_{z\in\Gamma} |f(z)|$$

But Γ is a half circle $\implies$ $Length(\Gamma) = \pi R$. Considering the $\max_{z\in\Gamma} |f(z)|$ we can write $|f(z)| = \left|\frac{e^{itz}}{z^2+1}\right|$, then by the triangle inequality $|z^2+1| \geq |z^2| - 1 = |z|^2 - 1$. But on the curve $|z| = R$ because we are tracing out a semicircle with radius R , where $R > 0$. Hence, $|z|^2 - 1 = R^2 - 1 \implies \frac{1}{|z^2+1|} \leq \frac{1}{R^2-1}$ which takes care of the denominator. Since z is a complex number then the curve Γ_1 can be parameterized by $z = Re^{it}$ for $t \in (0, \pi)$. By this then the numerator can be written as

$$\begin{aligned}
|e^{itz}| &= |e^{itRe^{it}}| \\
&= |e^{itR(\cos t + i\sin t)}| \\
&= |e^{itR\cos t + i^2 tR\sin t}| \\
&= |e^{itR\cos t}| \cdot |e^{i^2 tR\sin t}|
\end{aligned}$$

Since t and R are real numbers, it implies that $\cos t$ is also a real number, hence $|e^{itR\cos t}| = 1$. Then $|e^{i^2 tR\sin t}| = e^{-tRsint}$.

$$\left|\int_{\Gamma_1} f(z)dz\right| \implies \left|\int_{\Gamma_1} f(z)dz\right| \leq \pi R \times \max_{z\in\Gamma} \frac{e^{-tR\sin t}}{R^2-1} \tag{2.10}$$

but if we take the limit as $R \to \infty$ then (2.10) goes to 0. Therefore $\int_\Gamma f(z)dz = 0$. All in all, for the case $t > 0$ we have

$$\phi(t) = \int_\infty^\infty \frac{e^{itx}}{1+x^2}dx = \pi e^{-t}. \tag{2.11}$$

Finally, for the case $t < 0$ we set new integrals because we consider the contour in Figure 5. Thus

$$\int_{C_2} f(z)dz = \int_R^{-R} f(z)dz + \int_{\Gamma_2} f(z)dz.$$

Just as before, to deal with the integral on the left $\int_{C_2} f(z)dz$ we have to use the residual theorem and looking at the contour in Figure (5) is a simple pole, a pole with multiplicity of 1 which happens at $-i$. We just need to calculate the residue at $z = -i$.

$$\begin{aligned}
\int_{C_2} f(z)dz &= 2\pi i Res_{z=-i} f(z) \\
&= 2\pi i \lim_{z\to -i}(z+i)\frac{e^{itz}}{(z^2+1)} \\
&= 2\pi i \lim_{z\to i}(z+i)\frac{e^{itz}}{(z+i)(z-i)} \\
&= -\pi e^t
\end{aligned}$$

Also, $\int_{\Gamma_2} f(z)dz$ is going to be done the same way using the same type of argument. That is

$$\left|\int_{\Gamma_2} f(z)dz\right| \leq \text{length}(\Gamma_2) \times \max_{z\in\Gamma_2} |f(z)|. \tag{2.13}$$

It follows the same approach as before but t is defined in the range $t \in [-\pi, 0]$ and the length$(\Gamma) = \pi R$ but since our $t < 0$ it implies that $\sin t \leq 0$ which shows that $t \sin t \geq 0$, then we can clearly conclude (2.13) goes to 0 as $R \to \infty$. Hence for $t < 0$, we have

$$\int_{\infty}^{-\infty} \frac{e^{itx}}{1+x^2}dx = -\pi e^t$$

$$\implies \phi(t) = \int_{-\infty}^{\infty} \frac{e^{itx}}{1+x^2}dx = \pi e^t \qquad . \tag{2.14}$$

So we conclude that for the cases $t = 0$, $t > 0$ and $t < 0$

$$\phi(t) = \int_{-\infty}^{\infty} \frac{e^{itx}}{(1+x^2)}dx = \begin{cases} \pi e^{-t} & \text{for} \quad t > 0 \\ \pi & \text{for} \quad t = 0 \\ \pi e^{t} & \text{for} \quad t < 0 \end{cases}$$

So one best way of summarizing this function is $\phi(t) = \pi e^{-|t|}$, $\forall t \in \mathbb{R}$.
Then

$$\phi(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{e^{itx}}{(1+x^2)}dx = \frac{1}{\pi}(\pi e^{-|t|}) = e^{-|t|}.$$

□

Remark 1. Assume $X \sim C(0,1)$, then we will say that $Y \sim C(\mu,\lambda)$ if $Y = \mu + X\lambda$, where μ is the location parameter and λ is the scale parameter, for $\lambda > 0$. We know $\Phi_{X(t)} = E[e^i itX] = e^{-|t|}$, to derive the characteristic function,

$$\begin{aligned}
\Phi_Y(t) = \Phi_{\mu+\lambda X} &= E\left[e^{it(\mu+\lambda X)}\right] \\
&= E\left[e^{it\mu} \cdot e^{it\lambda X}\right] \\
&= e^{it\mu} \cdot E\left[e^{it\lambda X}\right] \\
&= e^{it\mu} \cdot \phi_X(t\lambda) \\
&= e^{it\mu} \cdot e^{-|\lambda t|} \\
&= e^{it\mu} \cdot e^{-\lambda|t|} \\
&= e^{it\mu-\lambda|t|} \qquad \text{for } \lambda > 0
\end{aligned}$$

hence the characteristic function for Cauchy variable Y with two parameters μ and λ is given by $\Phi_Y(t) = E[e^{itY}] = e^{it\mu-\lambda|t|}$

Proposition 2.4. *Suppose $Y_1, Y_2, ..., Y_n$ are independent Cauchy variables such that each Y_i is Cauchy variable with parameters (μ_i, λ_i). We find the distribution of $\sum_{j=1}^n Y_i$.*

Proof. Since all $Y_i's$ are Cauchy variable, the characteristic function of Y_j,

$$\Phi_{Y_j}(t) = E[e^{it\mu_j-\lambda_j|t|}] \qquad \forall j = 1, ..n.$$

To obtain the characteristic function of $\sum_{j=1}^n Y_i$ that is,

$$\begin{aligned}
\Phi_{\sum Y_j}(t) &= E[e^{it\sum Y_j}] \\
&= \prod_{j=1}^{n} Ee^{itY_j} \\
&= \prod_{j=1}^{n} \phi_{Y_j(t)} \\
&= \prod_{j=1}^{n} e^{it\mu_j-\lambda_j|t|} \\
&= e^{it(\sum \mu_j)-(\sum \lambda_j)|t|}
\end{aligned}$$

and as we know this is the characteristic function of Cauchy distribution with parameters $\sum \mu_j$ and $\sum \lambda_j$ and by the inversion theorem we say that the distribution of $\sum Y_j$ is also Cauchy with parameters $\sum \mu_j$ and $\sum \lambda_j$. □

Proposition 2.5. *If $X_1, X_2, ..., X_n$ are independent standard Cauchy variable. Then the distribution of its average, $\bar{X} = \frac{1}{n}\sum X_j$ is also standard Cauchy.*

Proof. The general characteristic function of a probability distribution is defined as $\phi_X(t) = E(e^{itX}), \forall t \in \mathbb{R}$. The characteristic function of the average $\bar{X} = \frac{1}{n}\sum X_j$ can be written as

$$\begin{aligned}\phi_{\frac{1}{n}\sum X_i}(t) &= E\left(e^{it\frac{1}{n}X_1} \cdot e^{it\frac{1}{n}X_2} \dots e^{it\frac{1}{n}X_n}\right) \\ &= E\left(e^{it\frac{1}{n}X_1}\right) \cdot E\left(e^{it\frac{1}{n}X_2}\right) \dots E\left(e^{it\frac{1}{n}X_n}\right). \\ &= \phi_{X_1}\left(\frac{t}{n}\right) \cdot \phi_{X_2}\left(\frac{t}{n}\right) \dots \phi_{X_n}\left(\frac{t}{n}\right).\end{aligned}$$

If each X_i is Cauchy distributed, we know that, a random variables has the standard Cauchy distribution, $X \sim C(0,1)$ if and only if its characteristic function is $\Phi_X(t) = e^{-|t|}$ from Proposition (2.1). Then

$$\begin{aligned}\phi_{X_1}\left(\frac{t}{n}\right) \cdot \phi_{X_2}\left(\frac{t}{n}\right) \dots \phi_{X_n}\left(\frac{t}{n}\right) &= e^{\left(-\frac{1}{n}|t|\right)} \cdot e^{\left(-\frac{1}{n}|t|\right)} \dots e^{\left(-\frac{1}{n}|t|\right)} \\ &= e^{\left(-\frac{1}{n}\sum_{i=1}^{n}|t|\right)} \\ &= e^{\left(-\frac{1}{n}\cdot n|t|\right)} \\ &= e^{-|t|}\end{aligned}$$

Hence, if random variables X_i are independent standard Cauchy variable then the distribution of its mean is also standard Cauchy. □

Remark 2. Proposition 2.5 can be explained in a more general sense thus, $C(\mu, \lambda)$. Suppose $Y_1, Y_2, ..., Y_n$ are independent Cauchy variable with two parameters (μ,λ), then the distribution of its average $\bar{Y} = \frac{1}{n}\sum Y_j$ is also Cauchy distributed.
Following the same approach in (2.4) we have

$$\begin{aligned}\phi_{\frac{1}{n}\sum Y_i}(t) &= E\left(e^{it\frac{1}{n}\sum_{i=1}^{n}Y_i}\right) \\ &= E\left(e^{it\frac{1}{n}Y_1}\right) \cdot E\left(e^{it\frac{1}{n}Y_2}\right) \dots E\left(e^{it\frac{1}{n}Y_n}\right) \\ &= \phi_{Y_1}\left(\frac{t}{n}\right) \cdot \phi_{Y_2}\left(\frac{t}{n}\right) \dots \phi_{Y_n}\left(\frac{t}{n}\right).\end{aligned}$$

Then from Remark1 we know the characteristic function of Cauchy variable with two parameters $C(\mu, \lambda)$, is given by $\phi_Y(t) = E[e^{itY}] = e^{it\mu - \lambda|t|}$ and since each Y_i is Cauchy distributed. It follows that

$$\begin{aligned}\phi_{Y_1}\left(\frac{t}{n}\right) \cdot \phi_{Y_2}\left(\frac{t}{n}\right) \dots \phi_{Y_n}\left(\frac{t}{n}\right) &= e^{\left(\frac{1}{n}(it\mu - \lambda|t|)\right)} \cdot e^{\left(\frac{1}{n}(it\mu - \lambda|t|)\right)} \dots e^{\left(\frac{1}{n}(it\mu - \lambda|t|)\right)} \\ &= e^{\left(\frac{1}{n}\sum_{k=1}^{n} it\mu - \lambda|t|\right)} \\ &= e^{\left(\frac{1}{n}(nit\mu - n\lambda|t|)\right)} \\ &= e^{(}it\mu - \lambda|t|)\end{aligned} \tag{2.16}$$

Therefore, if random variables Y_i are independent Cauchy variable with two parameters, $Y \sim C(\mu, \lambda)$, then the distribution of its mean is also $Cauchy(\mu, \lambda)$.

Proposition 2.6. *Let $X_1, X_2, ...$ be independent identically Cauchy distributed random variables, then $\frac{1}{n}\sum_{i=1}^{n} X_i$ is again Cauchy of the same distribution, which is readily shown in Proposition 2.5. The weak law of large numbers cannot be applied to Cauchy distribution (because the first moment is infinite).*

Proof. We can proof this by showing that, if $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in distribution to the Cauchy distribution as $n \to \infty$ and compare this with the conclusion of the weak law of large numbers. As shown previously, a random variable X obeying the Cauchy distribution with the location parameter $x_0 \in \mathbb{R}$ and the scale parameter $\lambda \in \mathbb{R}_+$,

- Its moment generating functions does not exist(specifically its mean and variance are undefined)
- Its characteristic function is $e^{(\mu it - \lambda|t|)}$

By the property of the characteristic function of Cauchy shown in the previous remark 2 that

$$\begin{aligned}
\phi_{nA_n}(t) &= \phi_{X_1}(t)\phi_{X_2}(t)\dots\phi_{X_n}(t) \\
&= e^{(\sum_{i=1}^{n} x_0 it - \lambda|t|)} \\
&= e^{(nx_0 it - n\lambda|t|)} \\
\phi_{A_n}(t) &= \phi_{nA_n}\left(\frac{t}{n}\right) \\
&= e^{\left(nx_0 i\frac{t}{n} - n\lambda\left|\frac{t}{n}\right|\right)} \\
&= e^{(x_0 it - \lambda|t|)}
\end{aligned}$$

If we take the limit as $n \to \infty$

$$\lim_{n\to\infty} \phi_{A_n}(t) = e^{(x_0 it - \lambda|t|)} \tag{2.17}$$

Then by the uniqueness theorem, the left hand side of (2.17) is the characteristic function of the Cauchy distribution with the location x_0 and scale parameter λ. If we denote A obeying this distribution. Then, as $n \to \infty$, A_n converges in distribution to A, which also obeys the Cauchy distribution. From [9] the following theorem of the weak law of large numbers holds

Theorem 2.7. ***(Weak Law of large numbers)*** *Let $X_1, X_2, ...$ be a sequence of identically distributed uncorrelated random variables. Let mean, $\mu = E\{X_1\}$ and variance, $\Sigma = E\{(X_1 - \mu)(X_1 - \mu)'\}$ be finite. Then the averages*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu, \quad \textit{in} \quad \textit{probability.}$$

It might be intriguing to apply to this results the weak law of large numbers due to the form of A_n, taking x_0 as μ. However, the mean and variance are undefined as stated earlier and shown in Proposition 2.1. Therefore the weak law of large numbers does not apply to this results. □

2.3 Cumulative Distribution Function

The cumulative distribution function (CDF) calculates the cumulative probability for a given x-value. It is used to determine the probability that a random observation that is taken from the population will be less than or equal to a certain value. You can also use

this information to determine the probability that an observation will be greater than a certain value, or between two values.

Definition 2.8. The cumulative distribution function or cdf of a continuous random variable X, denoted by $F_X(x)$ is defined as

$$F_X(x) = P_X(X \leq x) = \int_{-\infty}^{x} f_X(t)dt, \quad \text{for all } x \in \mathbb{R}$$

where $f_X(t)$ is the probability density function.

Proposition 2.9. *Suppose X is a standard Cauchy variable with pdf (2.1) then it's cumulative distribution function is*

$$F_X(x) = P_X(X \leq x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x) \qquad -\infty < x < \infty \tag{2.18}$$

Proof. We can show this by integrating (2.1) over the interval $-\infty < t < x$

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f_X(t)dt \\ &= \int_{-\infty}^{x} \frac{1}{\pi\,(1+t^2)}dt \\ &= \frac{1}{\pi}\int_{-\infty}^{x} \frac{1}{(1+t^2)}dt \\ &= \frac{1}{\pi}[\arctan(t)]_{-\infty}^{x} \\ &= \frac{1}{\pi}[\arctan(x) - \arctan(-\infty)] \\ &= \frac{1}{\pi}\left[\arctan(x) + \frac{\pi}{2}\right] \\ &= \frac{1}{2} + \frac{1}{\pi}\arctan(x), \end{aligned} \tag{2.19}$$

hence, $F_X(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$ is the cumulative distribution function of standard Cauchy distribution. □

Remark 3. In the case of $\mathrm{C}(\mu, \lambda)$, Proposition 2.9 can be shown in a more detailed way thus, integrating the pdf over the same interval.
If

$$F(x) = \int_{-\infty}^{x} \frac{\lambda}{\pi} \cdot \frac{1}{\lambda^2 + (t-\mu)^2}dt$$

Put

$$\frac{t-\mu}{\lambda} = y \quad \Rightarrow \quad t - \mu = \lambda y$$

and

$$dt = \lambda dy$$

which means the interval can be written as

$$-\infty < y < \frac{x-\mu}{\lambda}$$

then

$$
\begin{aligned}
F(x) &= \frac{\lambda}{\pi}\int_{-\infty}^{\frac{x-\mu}{\lambda}} \frac{1}{\lambda^2+(\lambda y)^2}\lambda dy \\
&= \frac{\lambda^2}{\pi\lambda^2}\int_{-\infty}^{\frac{x-u}{x}} \frac{1}{1+y^2}dy \\
&= \frac{1}{\pi}[\arctan(y)]_{-\infty}^{\frac{x-\mu}{\lambda}} \\
&= \frac{1}{\pi}\left[\arctan\left(\frac{x-\mu}{\lambda}\right) - \arctan(-\infty)\right] \\
&= \frac{1}{\pi}\left[\arctan\left(\frac{x-\mu}{\lambda}\right) + \frac{\pi}{2}\right] \\
&= \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\lambda}\right) \qquad \text{for all } \lambda > 0,
\end{aligned}
\tag{2.20}
$$

therefore, $F(x) = \frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{x-\mu}{\lambda}\right)$ is the cumulative distribution function of Cauchy(μ, λ).

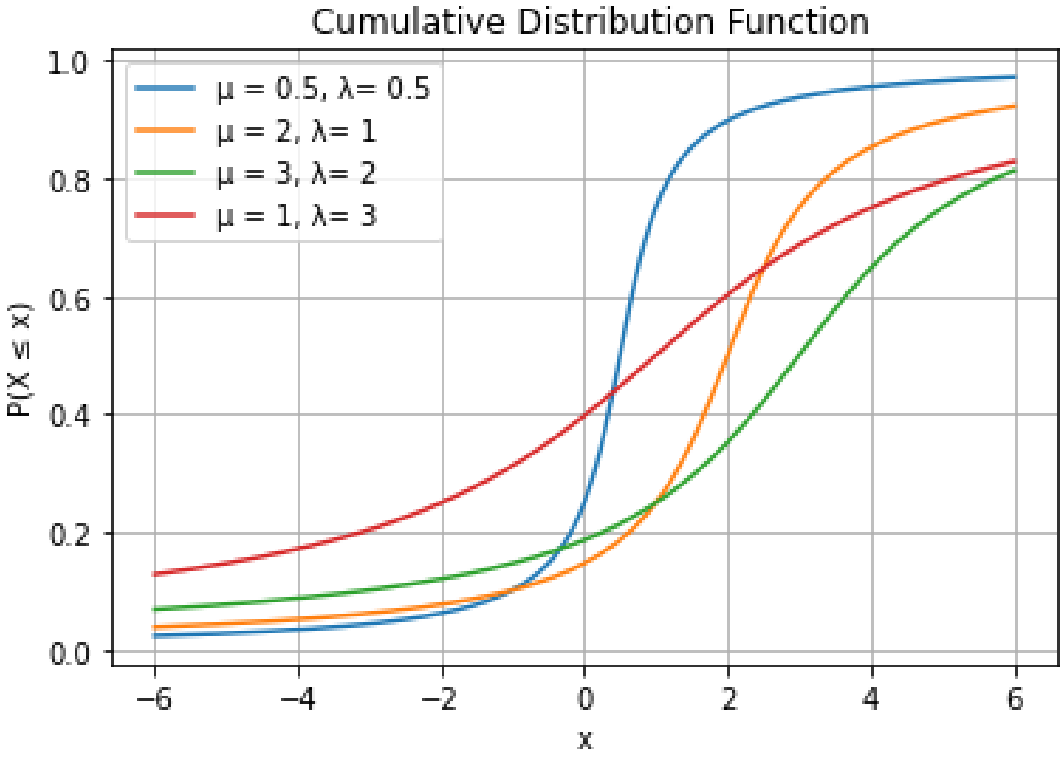


Figure 6: CDF of Cauchy Distribution with different location and scale parameter

3 PARAMETER ESTIMATION

This chapter is mainly from the book [6].

Parameter estimation is the refers to using data to estimate the parameters of a distribution. The main goal of estimation of parameters is to estimate the unknown parameters of a fixed model in order to fit some noisy measurements. For parameter estimation, the basic step is to choose a parametric model of a function which is a family of functions that can be described with a finite set of parameters. It can be denoted by $F : \mathbb{R}^n \to \mathbb{R}^n$. Data that are used to estimate the parameters of a model are normally contaminated with noise. Sometimes, it is not pertinent to assume that, this noise are normally distributed. This is due to the fact that there can be evident errors in the data which is not bound to happen if we only consider normally-distributed noise. The false measurements that are seen in the data are called outliers. Outliers are data points that are nowhere near other data points. Particularly, outliers are just the unusual values in a data set. They are problematic for analysis in statistics due to the fact that they can cause tests to either miss important findings or deform real results. An estimator is called robust when these outliers do not affect the estimation much. Robust statistical methods were developed to compliment the classical procedures when the data violates classical assumptions. Owing to the fact that, Cauchy is peculiar with regards to its heavy tails and difficulty in estimation of its location parameter. Then analysis of its data needs to take some technical approach to be possible. Different robust estimators were adopted to estimate the location parameter which can further aid in its application. We shall mainly deal with three broad classes of robust estimators of the real parameter thus, M-estimators, L-estimators, and R-estimators. Cauchy distribution can often be cited as an example of the computational failure of the maximum likelihood method of estimation. The substantial reason summarized by [10] are: (1) the likelihood equation have $2n - 1$ roots which are real and complex; (2) not either numerical solution or analytical solution of the likelihood equation can easily be obtained; (3) it is only one of the real roots (the global maximum) of the likelihood that does not tend to ∞ or $-\infty$ as $n \to \infty$, all the others does. The method of moment estimation fails since the mean and variance of the Cauchy distribution does not exist. [11] proposed a new unbiased L-estimator based on order statistics, which is not only asymptotically efficient but outperforms existing L-estimators in terms of finite sample efficiency. [12] discussed the properties of various estimators of the central position of the Cauchy distribution which is notorious for the divergent nature of its first and higher moments and evaluated the results of using different kinds of estimators by simulation of different kinds of experiment using Monte Carlo simulations. [13] defined robust estimators for the parameters of Cauchy distribution based on the probability integral method where, the estimators were simple, robust and consistent, but asymptotically less efficient than the maximum likelihood estimators which are not robust. [14] showed that the likelihood equations obey a certain structure that appears in other robust parameter estimation problems.

Definition 3.1. Let $X_1, X_2, ..., X_n \sim iid, X_1 \sim F$. An M-estimator T_n is defined as a solution of the minimization problem

$$\sum_{i=1}^{n} \rho(X_i, t) := min \tag{3.1}$$

where $\rho(.,.)$ is a properly chosen function. If ρ is differentiable in t with a continuous derivative $\psi(X,t) = \frac{\partial}{\partial t}\rho(X,t)$, then T_n is a root or one of the roots of the equation

$$\sum_{i=1}^{n} \psi(X_i, t) = 0 \tag{3.2}$$

M-estimator is a robust method used as an alternative to be the least square method when the data under study has outliers, extreme observations, or does not follow a normal distribution. The fundamentals of the M-estimators consist of taking distributions that have a probability density function with "heavy tails". A distribution having heavy tails means that, large errors are less unlikely than it would be with normal distribution.Different types of M-estimators based on several noise distributions have been suggested. In this study the M-estimator we would consider is;

3.1 Huber M-estimate

Huber(1964) introduced M-estimators and their asymptotic properties and they were an important part of the modern robust statistics.In the strive of defining a robust criterion, Huber(1964) considered an agreement between the mean and the median. The mean criterion is a square and in the "tails" the square gives too much weight to big observations. In contrast, it is well known robust location estimator.
The M-estimator for the location parameter μ is defined as a solution of

$$\sum_{i=1}^{n} \psi(\frac{x_i - t}{\sigma}) = 0. \tag{3.3}$$

Taken $\sigma = 1$ the solution is

$$\sum_{i=1}^{n} \psi(x_i - t) = 0.$$

The function ψ is linear in a bounded segment $[-k, k]$ and constant outside this segment. The Huber M-estimate is defined by the function ψ where

$$\psi_k(x) = \begin{cases} k, & x \geq k \\ x & -k \leq x \leq k \\ -k & x \leq -k, \end{cases} \tag{3.4}$$

where $k > 0$ is a fixed constant.The constant k, which can also be called a *turning parameter*, controls the mix, with small values of k yielding a more "median-like" estimator. Constant k depends of the number of outliers in the set. As Huber proved in (1964), an estimator, produced by the function (3.4) is mini-maximally robust for a contaminated normal distribution, while the value of k depends on the contamination ratio. A more practical and natural question is whether there exists a distribution F such that the Huber M-estimator is the maximal likelihood estimator μ for $F(x - \mu)$, that is, such that ψ is the likelihood function of F. A distribution of such property really exists, and its density is normal in internal $[-k, k]$ and exponential outside [6].

3.2 Hodges-Lehmann Estimate (R-estimate)

We define rank R_i as

$$R_i = \sum I_{[X_j \leq X_i]}, \qquad i = 1, ..., n \tag{3.5}$$

which is widely used for location parameter tests(Wilcoxon test) and that is $R_i = n\hat{F}_n(X_i)$, $i = 1, ..., n$, where $\hat{F}_n$ is empirical distribution function of $X_1, ..., X_n$. The ranks remains unchanged with respect to the class of monotone transformations of observations. The pros of the rank test include; [6],

- The estimators are position and scale equivariant.
- The most important one is that, the distribution of the test criterion under the hypothesis of randomness $i.e.$,if $X_1, ..., X_n$ are independent and identically distributed with a continuous distribution function) is independent of the distribution of observations.

A common complaint about non-parametric statistics is that they allow you to assess statistical significance but they do not allow you to measure effect size. Fortunately, there are techniques for measuring the size of an effect that are insensitive to distributional unpleasantness (like skewness, kurtosis or outliers) just like non-parametric statistical tests are. One of such methods is the Hodges-Lehmann estimator. Hodges and Lehmann (1963) proposed that estimators could be obtained by inverting rank tests, and the class of such estimators is termed as R-estimators. R-estimate can be defined for a lot of models, reasonably for all where the rank tests make sense and the test criterion is symmetric about a known center or has other appropriate property under the null hypothesis.
Hodges–Lehmann estimator plays a fundamental role when applying standard rank-based methods (in particular, the Wilcoxon signed rank test).The Walsh averages of n observations refers to all pairwise averages,$(X_i + X_j)/2$ for all $i \leq j$.The Hodges-Lehmann estimator is the median of all Walsh averages, specifically,

$$T_{nH} = med_{i \leq j} \frac{X_i + X_j}{2}. \tag{3.6}$$

3.3 Sen's Estimate (L-estimator)

L-estimators are based on order statistics $X_{n:1} \leq ... \leq X_{n:n}$ of random sample $X_1, ..., X_n$. The general L-estimator can be given in the form

$$T_n = \sum_{i=1}^{n} c_{ni} h(X_{n:i}) + \sum_{j=1}^{k} a_j h^*(X_{n:[np_j]+1}) \tag{3.7}$$

where $c_{n1}, ..., c_{nn}$ and $a_1, ..., a_k$ are given coefficients, $0 < p_1 < ... < p_k < 1$ and $h(.)$ and $h^*(.)$ are given functions. Also, the coefficient $c_{ni}, 1 \leq i \leq n$ are generated by a bounded weight function $J : [0, 1] \to \mathbb{R}$ in the following way:

$$c_{ni} = \int_{\frac{i-1}{n}}^{\frac{i}{n}} J(s)ds \qquad i = 1, ..., n \tag{3.8}$$

approximately

$$c_{ni} = \frac{1}{n} J(\frac{i}{n+1}) \qquad i = 1, ..., n \tag{3.9}$$

The first component of the L-estimator (3.7) generally involves all order statistics, while the second component is a linear combination of several (finitely many) sample quantiles.

The major advantage of the L-estimator is that it has a closed form and is computationally easy [11]. The simplest examples suggested by Jurečková and Picek [6] of an L-estimator of the distribution location are the sample median and the midrange, in which the later is defined as

$$T_n = \frac{X_{1:n} + X_{n:n}}{2} \tag{3.10}$$

A simple L-estimator of distribution scale is the sample range or

$$R_n = X_{n:n} - X_{1:n}. \tag{3.11}$$

An interesting L-estimator that have immediate connection to the L-moments is the Sen weighted mean. A special location statistics, which is based on the order statistics, is the Sen weighted mean (Sen, 1964) or the quantity $\mathcal{T}_{n,k}$. The $\mathcal{T}_{n,k}$ is a robust estimator [6] of the mean of a distribution and is defined as

$$\mathcal{T}_{n,k} = \binom{n}{2k+1}^{-1} \sum_{i=1}^{n} \binom{i-1}{k} \binom{n-i}{k} X_{n:i},$$

where $X_{n:i}$ are the sample order statistics and k is a weighting or trimming parameter. Note that $\mathcal{T}_{n,0} = \mu = \overline{X}_n$ or the arithmetic mean and $\mathcal{T}_{n,k}$ is the sample median if either n is even and $k = (n/2) - 1$ or n is odd and $k = (n-1)/2$.

3.4 Trimmed Mean(L-etsimator)

The trimmed mean is another type of unbiased estimator.To compute a trimmed mean, we remove a predetermined amount of observations on each side of a distribution, and average the remaining observations. $\alpha - trimmed$ mean $0 < \alpha < 0.5$is the average of the central quantiles. It is given by

$$\bar{X}_{n\alpha} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{n:i}$$

where $X_{n:i}$ are the sample order statistics [6].

3.5 Winsorized Mean(L-estimator)

Winsorization is a method of reducing the influence of outliers in your data set.The Winsorized mean is similar to the trimmed mean, except that rather than deleting the extreme values, they are set equal to the largest (or smaller) values.
The average of the Winsorized mean is a typical example of the general form (3.7) which possess two components

$$\begin{aligned} &\bar{W}_{n\alpha} = T(F_n) \\ &= \tfrac{1}{n}\left\{[n\alpha]X_{n:[n\alpha]+1} + \textstyle\sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{n:i} + [n\alpha]X_{n:n-[n\alpha]}\right\} \\ &= \textstyle\sum_{i=1}^{n} c_{ni} X_{n:i} + \frac{[n\alpha]+1}{n} X_{n:[n\alpha]+1} + \frac{[n\alpha]+1}{n} X_{n:n-[n\alpha]} \end{aligned}$$

where

$$c_{ni} = \begin{cases} \frac{1}{n} & \dots \quad 1 + [n\alpha] < i < n - [n\alpha] \\ 0 & \dots \quad \text{otherwise} \end{cases}$$

The extreme quantiles are not trimmed but substituted with quantiles $X_{n:[n\alpha]+1}$ and $X_{n:n-[n\alpha]}$ sequentially [6].

4 RECIEVER OPERATING CHARACTERISTIC CURVE (ROC) CURVE

4.1 Summary of ROC curves

The ROC curve assesses the performance classification methods used to identify observations by type. For instance, the ROC curve may indicate the performance of (1) a software filter which categorizes emails as ham or spam, (2) a medical diagnostic test which categorizes patients as healthy or sick, or (3) an astrophysics approach which classifies galaxies as elliptical or spiral. The receiver operating characteristic (ROC) curves have been in use for a long time, having risen in the context of signal detection theory which was developed in the $1950s$ as [15] discussed. To produce ROC curve, a sample of observations with known classes must be available. Often, the true ROC function may be a continuous curve which remains unknown.

4.2 Classification methods, scores, and error rates

The ROC curve portrays a classification method's ability to distinguish classes and depicts all attainable error rates across thresholds. Furthermore, the ROC curve's standard axes of true positive and false positive rates helps to compare various classification methods. Assume that we have a classification method that accesses observations to establish to which one of the two classes they belong. For consistency with established terms of the ROC curves, label one class positive and the other class negative. Considering a given observation, the classification method gives back a number called a score. On top of multiple observations, the classification method produces two distributions of scores, one for the negative class and the other for positive class. Suppose that high scores implies the positive class and that low scores also implies the negative. We want to measure the method how the method's performance is selective between the two classes [8].
Following, for the purposes of testing this performance, let us assume we have an observation whose true sources are known. We apply the classification method to this test data to obtain scores for all observations. For a threshold c on the scores, if the score for a given observation is greater than c, we assume the observation came from the positive class. Otherwise, the negative class is been considered as the source of the observation. Supposing we support the wrong assumption, we commit an error. Also, when we favor the negative class when in truth an observation comes from the positive class, the error is termed as false negative(FN). Likewise, when we favor the positive class when in truth an observation comes from the negative class, the error is termed as a false positive(FP). With multiple observations we can compute the rates of these errors. We will only highlight on true positive rate (TPR) and false positive rate(FPR), which define the axes of the ROC plot. Another summary index that has been proposed is the point on the ROC curve where sensitivity is equal to specificity, $TPR = 1 - FPR$ [8]. The ROC curve can be constructed as a plot of sensitivity(TPR) versus 1-specificity(FPR). Particularly, the ROC plots the FPR on the x-axis and the TPR on the y-axis.

$$TPR(Sensitivity) = \frac{TP}{TP + FN}$$

and

$$FPR(1 - Specificity) = \frac{FP}{TN + FP}$$

Definition 4.1 (ROC)**.** Using a threshold c, define a binary test result Y as

$$\begin{aligned} \text{positive if} \quad & Y \geq c, \\ \text{negative if} \quad & Y < c. \end{aligned}$$

Let the corresponding true and false positive fractions at the threshold c be TPR(c) and FPR(c),respectively where

$$\begin{aligned} TPF(c) &= P[Y \geq c|D = 1], \\ FPF(c) &= P[Y \geq c|D = 0]. \end{aligned} \tag{4.1}$$

The ROC curve is the whole set of possible true and false positive rates attainable by dividing Y with different thresholds. That is, the ROC curve is

$$ROC(.) = \{(FPR(c), TPR(c)), c \in (-\infty, \infty)\} \tag{4.2}$$

We will discover that, as the threshold c increases, both $FPR(c)$ and $TPR(c)$ decrease. At one extreme, $c = \infty$, we have $\lim_{c\to\infty} TPR(c) = 0$ and $\lim_{c\to\infty} FPR(c) = 0$. At the other, $c = -\infty$, we have $\lim_{c\to-\infty}$ TPR(c) $= 1$ and $\lim_{c\to-\infty}$ FPR(c) $= 1$ which implies that, the ROC is a monotone increasing function in the positive quadrant as illustrated in Figure(3).We can also write the ROC curve as

$$ROC(.) = \{(t, ROC(t)), t \in (0, 1)\} \tag{4.3}$$

where the ROC function maps t to TPC(c), and c is the threshold corresponding to FPF(c)=t.

Mathematically, the ROC curve can be described as a monotone increasing function in the unit square tied down at the boundary points $(0, 0)$ and $(1, 1)$. A perfect classifier completely separates diseased and non-diseased subjects thus, $(FPR = 0, TPR = 1)$. Its ROC curve is along the left and upper borders of the positive unit quadrant,whereas, an uninformative ROC curve for a continuous test result Y is a diagonal of 45°. Alternatively, an uninformative test is one such that $TPR(c) = FPR(c)$ for every threshold t and this situation is represented by ROC curve $ROC(t) = t$, which is a line with unit slope. Most ROC curves lie between those of the perfect and useless tests. Better tests have ROC curves closer to the upper left corner.

The primary numerical index used to describe the behavior of the ROC curve is the area under the ROC curve (AUC), defined by

$$AUC = \int_0^1 ROC(t)dt \tag{4.4}$$

4.3 Empirical ROC Curve

The empirical ROC curve is a more general type of the ROC curve which is non-parametric. It does not require any distribution pattern of test values. It uses the trapezoidal rule in the sense that it calculates the area by joining the points at each interval of the observed values of continuous test. One advantage of this method is that no structural hypothesis is made on the form of the plot, and the underlying distributions of the results of the two groups are not specified. The precision of statistical conclusions based on the empirical ROC curve is reduced compared to a model-based estimator when the genuine ROC curve is a smooth function (at least when the model is correctly specified). Nevertheless, the estimator has some drawbacks, and it may suffer from large variability, particularly for small sample sizes [16].

Definition 4.2. The empirical estimator of the ROC curve applies the definition of the ROC curve to the observed data. That is, for every possible threshold c, the empirical true and false positive ratio are calculated as

$$\hat{TPR}(c) = \sum_{i=1}^{n_D} I\left(Y_{D_i} \geq c\right)/n_D,$$
$$\hat{FPR}(c) = \sum_{j=1}^{n_{\bar{D}}} I\left(Y_{\bar{D}_j} \geq c\right)/n_{\bar{D}}$$

where D and $\bar{D}$ represent the positive and negative responses, n_Y and $n_{\bar{Y}}$ are the total number of positive and the negative responses, Y_D and $Y_{\bar{D}}$ are the random diagnostic variables in the positive and the negative responses.

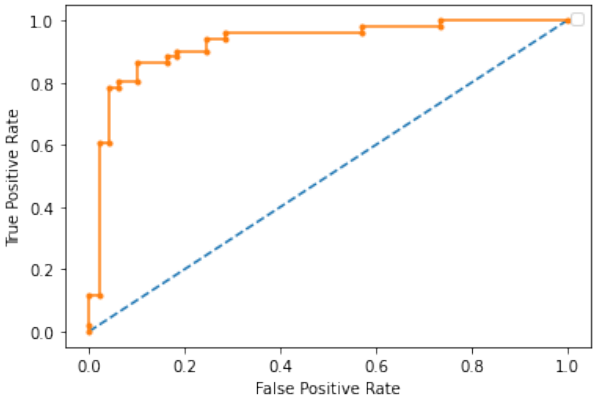


Figure 7: An example of an empirical ROC curve denoted by n_D and $n_{\bar{D}}$

Particularly, each point on the empirical ROC curve represents an individual threshold value. The points are connected to form the curve. Threshold values that result in low false positive rates tend to result in low true positive rate as well. The empirical ROC curve is a plot of the true positive rate $\hat{TPR}(c)$ versus the false positive rate $\hat{FPR}$ foe all $c \in (-\infty, \infty)$. Also, as the true positive increases, the false positive also increases. By [7] the AUC of the empirical estimator is defined as

$$AUC = \frac{1}{n_D n_{\bar{D}}} \sum_{j=1}^{n_{\bar{D}}} \sum_{i=1}^{n_D} \left(I\left(Y_{D_i} > Y_{\bar{D}_j}\right) + \frac{1}{2} I\left(Y_{D_i} = Y_{\bar{D}_j}\right) \right) \tag{4.5}$$

4.4 Binormal ROC Curve

The Binormal model for ROC assumes that under some unspecified, monotonic increasing transformation, the negative distribution of the diagonal variables in positive classes follows normal distributions [7]. To evaluate the binormal ROC curve, the sample mean and sample variances are estimated from the known positive group, and again for the known negative group. These sample means and sample variances are used to identify two normal distributions. The ROC curve is later generated from the two normal distributions. In a case the two normal distributions overlap, Binormal ROC curve is closer to the 45-degree diagonal line. When the two overlap only in the tails, the Binormal ROC curve has a much greater distance from the 45-degree diagonal line. The choice of the binormal estimator to fit the ROC curve is often justified by theoretical considerations, mathematical tractability, familiarity with the normal model, or simply a favorable agreement or convenience. [17] presents a table outlining the most common arguments in favor of using this estimator. But some authors also argue that the binormal estimator is robust.It is used in the sense of robust statistics, meaning that the presence of a certain amount of observations coming from a non-normal distribution will yield reliable results.

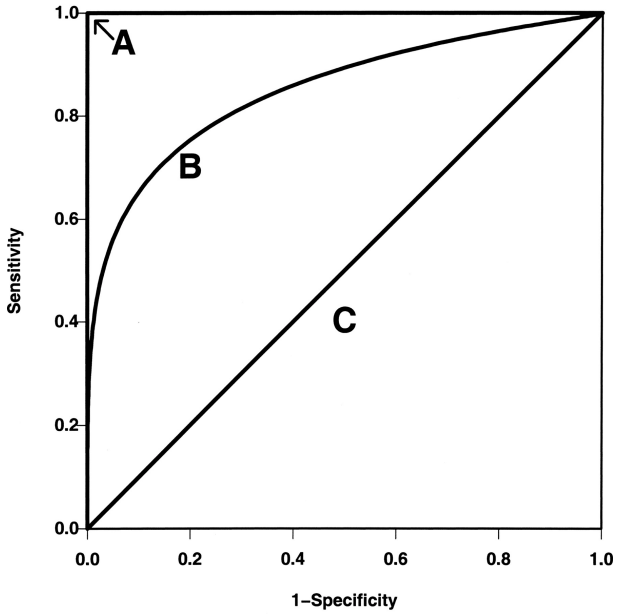


Figure 8: Three hypothetical ROC curves representing the accuracy of an ideal test (line A) on the upper and left axes in the unit square, a typical ROC curve (curve B), and a diagonal line corresponding to an uninformative test (line C). As test accuracy improves, the ROC curve moves toward line A

.

In figure (8), line A represents a perfect test with $AUC = 1$, curve B represents a typical ROC curve (for example $AUC = 0.85$), and a diagonal line also known as the chance line (line C) corresponding to uninformative test with $AUC = 0.5$. As test accuracy improves, the ROC curve moves toward A, and the AUC approaches 1. Clearly, if two tests are ordered with test A uniformly better than test B in the sense that [7]

$$ROC_A(t) \geq ROC_B(t) \qquad \forall t \in (0,1) \tag{4.6}$$

which can clearly be seen in (fig.4), then their AUC statistics are ordered;

$$AUC_A \geq AUC_B.$$

Functional Form of the Binormal ROC Curve
To derive the functional form of the ROC curve, we can assume that test results are normally distributed in the diseased (D) and non-diseased($\tilde{D}$) populations.
If

$$Y_D \sim N(\mu_D, \sigma_D^2) \qquad Y_{\tilde{D}} \sim N(\mu_{\tilde{D}}, \sigma_{\tilde{D}}^2)$$

then

$$ROC(t) = \Phi(a + b\Phi^{-1}(t))$$

where

$$a = \frac{\mu_D - \mu_{\tilde{D}}}{\sigma_D} \qquad b = \frac{\sigma_{\tilde{D}}}{\sigma_D}$$

and Φ denotes the standard normal cumulative distribution function [7].

Proof. The Binormal ROC curve with a given threshold c is commonly assessed using the probabilities that correctly classify outcomes thus, the $FPR(c)$ and $TPR(c)$ can be define respectively as

$$\mathrm{FPR}(c) = \mathrm{P}\left[Y_{\bar{D}} > c\right] = \Phi\left(\frac{\mu_{\bar{D}} - c}{\sigma_{\bar{D}}}\right)$$
$$\mathrm{TPR}(c) = \mathrm{P}\left[Y_D > c\right] = \Phi\left(\frac{\mu_D - c}{\sigma_D}\right)$$

For a false positive fraction t, we see that $c = \mu_{\bar{D}} - \sigma_{\bar{D}}\Phi^{-1}(t)$ is the corresponding threshold for the test positivity criterion. Hence.

$$\begin{aligned}\mathrm{ROC}(t) = \mathrm{TPR}(c) &= \Phi\left(\frac{\mu_D - c}{\sigma_D}\right) \\ &= \Phi\left(\frac{\mu_D - \mu_{\bar{D}} + \sigma_{\bar{D}}\Phi^{-1}(t)}{\sigma_D}\right) \\ &= \Phi\left(a + b\Phi^{-1}(t)\right).\end{aligned}$$

□

Definition 4.3. The binormal ROC is defined by the form

$$ROC(t) = \Phi(a + b\Phi^{-1}(t)) \tag{4.7}$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. We call a the intercept and b the slope for binormal ROC curve.As per our convention that largest test result are more indicative of disease, we have $\mu_D > \mu_{\bar{D}}$ so $a > 0$.The slope b is positive by definition.

The binormal ROC provides a good approximation to a wide range of ROC curves that occur in practice. Note that if the slope $b = 1$, then the binormal ROC curve is concave everywhere. To see this, observe that the slope of the ROC curve at t is likelihood ratio at the corresponding threshold c, which can be given as

$$\frac{f_D(C)}{f_{\bar{D}}(C)} = \left(\frac{\sigma_{\bar{D}}}{\sigma_D}\right)\exp\left\{\frac{-(c-\mu_D)^2}{2\sigma_D^2} + \frac{(c-\mu_{\tilde{D}})^2}{2\sigma_{\tilde{D}}^2}\right\},$$

when $b = 1$, we have $\sigma_D = \sigma_{\bar{D}}$ and the common value is denoted by σ.The likelihood ratio then reduces to $\exp\{c(\mu_D - \mu_{\bar{D}})/\sigma^2 - ({\mu_D}^2 - {\mu_{\bar{D}}}^2)/2\sigma^2\}$. As c decreases (t increases), we seee that the slope decreases because $\mu_D > \mu_{\bar{D}}$. Hence when $b = 1$ the slope of the ROC curve is monotone decreasing in t. Again, if $b \neq 1$ the monotonicity criterion fails. For $b > 1$, the likelihood ratio decreases and then increases. Conversely, for $b < 1$ the likelihood ratio increase and then decrease as t range from 0 to 1. This produces the irregularities in the ROC curve where it falls below the uninformative test ROC curve, $ROC(t) = t$. The AUC has a simple analytic form when the ROC curve is binormal. The AUC for binormal ROC curve is [7]

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right).$$

4.5 Bi-Cauchy ROC

Here, we derive the functional form of ROC model for Cauchy similar to the binormal ROC curve. Let the distribution in the true positive class be Cauchy with location parameter μ_D and scale parameter σ_D^2 and false positive class be Cauchy with location parameter $\mu_{\bar{D}}$ and scale parameter $\sigma_{\bar{D}}^2$ under this transformation

$$Y_D \sim C(\mu_D, \sigma_D^2) \qquad Y_{\bar{D}} \sim C(\mu_{\bar{D}}, \sigma_{\bar{D}}^2).$$

Since the ROC curve with a given threshold c is commonly assessed using the probabilities that correctly defined outcomes as stated in the case of the binormal ROC we can define the $TPR(c)$ as

$$\begin{aligned} TPR(c) = P(Y_D > c) &= 1 - P(Y_D \leq c) \\ &= 1 - F\left(\frac{c-\mu_D}{\sigma_D}\right) \\ &= F\left(-\frac{c-\mu_D}{\sigma_D}\right) \\ &= F\left(\frac{\mu_D - c}{\sigma_D}\right) \end{aligned} \tag{4.8}$$

From above, $P(Y_D \leq c) = F\left(\frac{c-\mu_D}{\sigma_D}\right)$ by standardization. The false positive rate $FPR(c)$ follows the same approach. Hence

$$\begin{aligned} FPR(c) = P(Y_{\bar{D}} > c) &= 1 - P(Y_{\bar{D}} \leq c) \\ &= F\left(\frac{\mu_{\bar{D}} - c}{\sigma_{\bar{D}}}\right). \end{aligned} \tag{4.9}$$

For the false positive fraction t, if we rearrange the $FPR(c)$ equation,that is, taking its inverse transformation we have

$$F^{-1}(FPR(c)) = \frac{\mu_{\bar{D}}}{\sigma_{\bar{D}}} - \frac{c}{\sigma_{\bar{D}}} \tag{4.10}$$

we see that

$$c = \mu_{\bar{D}} - \sigma_{\bar{D}} F^{-1}(t), \tag{4.11}$$

which can be substituted into (4.9) to get the Bi-Cauchy ROC

$$\begin{aligned}\text{ROC}(t) = TPR(c) &= F\left(\frac{u_D - c}{\sigma_D}\right) \\ &= F\left(\frac{u_D - u_D + \sigma_D F^{-1}(t)}{\sigma_D}\right) \\ &= F\left(\frac{u_D - u_{\bar{D}}}{\sigma_D} + \frac{\sigma_{\tilde{D}} F^{-1}(t)}{\sigma_D}\right)\end{aligned}$$

introducing

$$\alpha = \frac{u_D - u_{\bar{D}}}{\sigma_D} \quad \text{and} \quad \beta = \frac{\sigma_{\bar{D}}}{\sigma_D}$$

we have

$$ROC(t) = F\left(\alpha + \beta F^{-1}(t)\right), \tag{4.12}$$

where (4.12) is the Bi-Cauchy ROC model. We see that, the model has two parameters, α which is the intercept, β is the slope, withholding the fact stated by [7] that larger test results are more indicative of disease thus $\mu_D > \mu_{\bar{D}}$, then we can conclude that $\alpha > 0$ and also the slope β is positive by definition. F is the cumulative distribution function of standard Cauchy distribution as shown in Proposition (2.9) and F^{-1} is the inverse of the cumulative distribution function. Since all terms are known the Bi-Cauchy ROC model in (4.12) is

$$ROC(t) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\alpha + \beta\tan\left(\pi t - \frac{\pi}{2}\right)\right) \quad \forall t \in (0,1) \tag{4.13}$$

The nature and behavior of the curve is dependent on the slope α and the intercept β. From [7] when $\beta = 1$, the Binormal ROC curve is concave everywhere. This assertion holds for the Bi-Cauchy ROC in the sense that, for a constant β the curves are concave but for different values of β there are anomalies in the nature of the curves which can be seen in Figure 10. Proper ROC curves are concave and symmetric and never cross each other, this makes it reliable in comparison between different subjects, Figure 9 shows a proper Bi-Cauchy ROC curves.

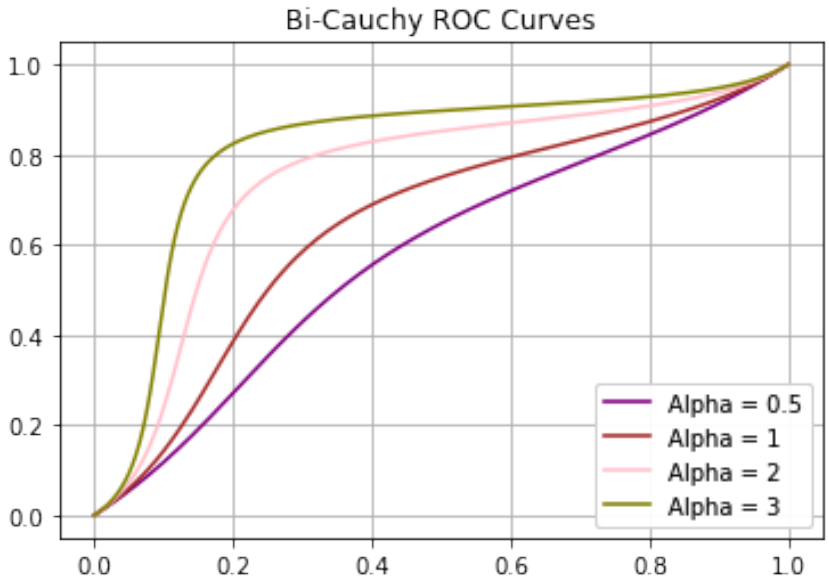


Figure 9: Bi-Cauchy ROC curves with a constant slope ($\beta = 1$)

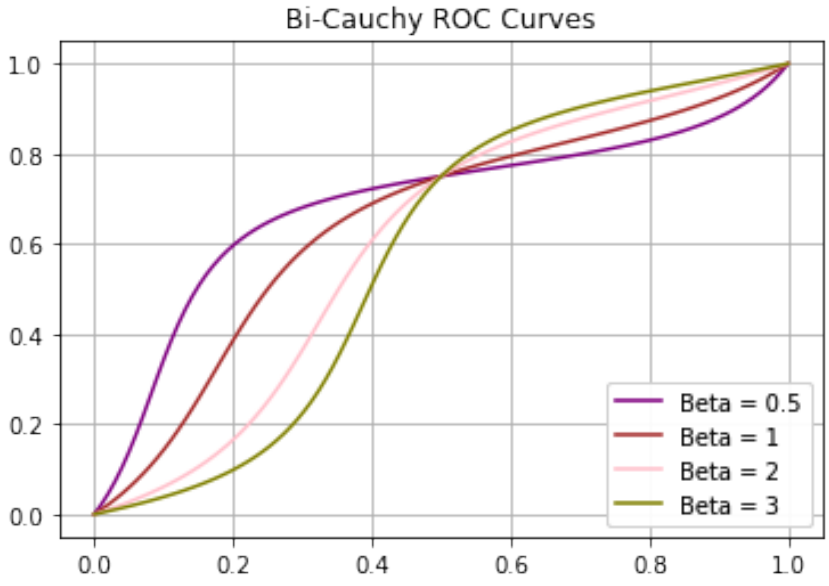


Figure 10: Bi-Cauchy ROC curves with a constant Intercept ($\alpha = 1$)

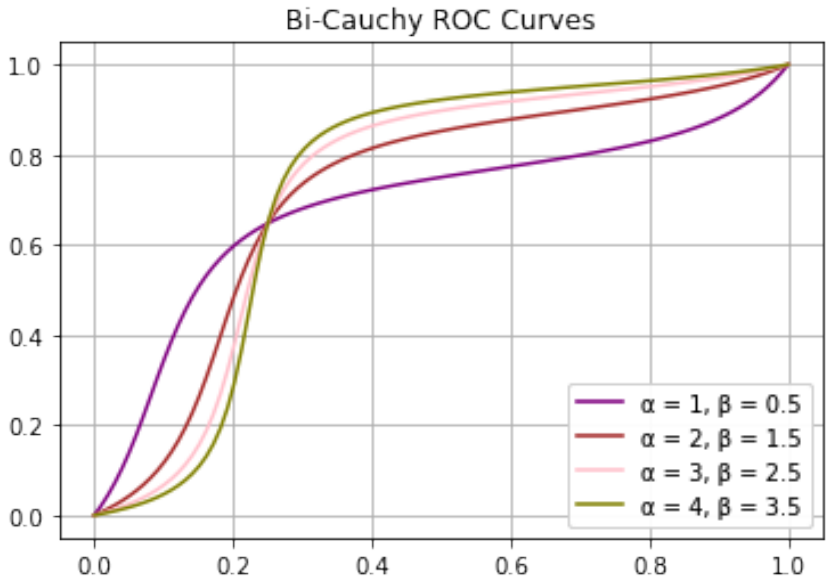


Figure 11: Bi-Cauchy ROC curves with different α and β values

AUC of the Bi-Cauchy ROC curve
From [7] the area under the curve is

$$AUC = P[Y_D > Y_{\bar{D}}], \tag{4.14}$$

where Y_D and $Y_{\bar{D}}$ are independent randomly chosen test results from the diseased and non-diseased population, respectively. [18] base on the same assumption in (4.14) in its estimates in clinical trials. It can be calculated as the average of TPR accross all FPR. Then basing on the same results (4.14) which was used to derived the AUC for the Binormal ROC curve by [7], we can derive the Bi-Cauchy ROC in the sense that

$$AUC = P[Y_D > Y_{\bar{D}}] = P[Y_D - Y_{\bar{D}} > 0]$$

Let

$$K = P[Y_D - Y_{\bar{D}}].$$

Then using the properties of the characteristic function it can be shown that

$$K \sim C(\mu_D - \mu_{\bar{D}}, \sigma_D^2 + \sigma_{\bar{D}}^2)$$

which implies that

$$\begin{aligned} P(K>0) &= 1 - P(K \leq 0) \\ &= 1 - F\left(\frac{-\mu_D + \mu_{\bar{D}}}{\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}}\right) \\ &= F\left(\frac{\mu_D - \mu_{\bar{D}}}{\sqrt{\sigma_D^2 + \sigma_{\bar{D}}^2}}\right) \\ &= F\left(\frac{\mu_D - \mu_{\bar{D}}}{\sqrt{\sigma_D^2 \times \left(1 + \frac{\sigma_{\bar{D}}^2}{\sigma_D^2}\right)}}\right) \\ &= F\left(\frac{\mu_D - \mu_{\bar{D}}}{\sigma_D} \div \sqrt{1 + \frac{\sigma_{\bar{D}}^2}{\sigma_D^2}}\right) \end{aligned} \tag{4.15}$$

if we denote

$$\alpha = \frac{u_D - u_{\bar{D}}}{\sigma_D} \quad \text{and} \quad \beta = \frac{\sigma_{\bar{D}}}{\sigma_D}.$$

From the assertion above, the AUC of the Bi-Cauchy ROC can be written in the following convenient closed form

$$AUC = F\left(\frac{\alpha}{\sqrt{1+\beta^2}}\right), \tag{4.16}$$

g since F is a known function, thus the cumulative distribution function of Cauchy distribution as show in Proposition 2.9, it follows that the AUC of the Bi-Cauchy ROC model is

$$AUC = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{\alpha}{\sqrt{1+\beta^2}}\right). \tag{4.17}$$

4.5.1 Point of Inflection of the Bi-Cauchy ROC Model

A point of inflection is established where the graph of a function changes concavity. From basic calculus results concerning concave functions it follows that the an inflection is where a curve changes from concave upwards to concave downwards. The first derivative of the function gives the slope and the second derivative tells us whether the slope increases or decreases. In the case the second derivative is positive the curve is concave upwards and vice-versa is concave downwards. When $\beta \neq 1$, Bi-Cauchy ROC curve is not "proper" as is not concave throughout the whole ROC square as seen in Figure(10). We seek to find a point in the model (4.13) where the curve inflects. This is possible by finding the first and second derivative. To locate the inflection point, we set the second derivative to zero, and solve the equation.
For the model,

$$\mathrm{ROC}(t) = \frac{1}{2} + \frac{1}{\pi}\arctan(\alpha + \beta\tan(\pi t - \pi/2)) \quad \forall t \in (0,1).$$

The first derivative

$$\begin{aligned}\frac{\partial ROC(t)}{\partial t} &= \frac{\partial}{\partial t}\left(\frac{1}{2}+\frac{1}{\pi}\arctan(\alpha+\beta\tan(\pi t-\pi/2))\right)\\ &= \frac{\partial}{\partial t}\left(\frac{1}{2}\right)+\frac{\partial}{\partial t}\left(\frac{1}{\pi}\arctan(\alpha+\beta\tan(\pi t-\pi/2))\right)\\ &= \frac{1}{\pi}\left(\frac{1}{1+(\alpha+\beta\tan(\pi t-\pi/2))^2}\times\frac{\beta\pi}{\cos^2(\pi t-\pi/2)}\right)\\ &= \left(\frac{\beta}{(1+(\alpha+\beta\tan(\pi t-\pi/2))^2)\cos^2(\pi t-\pi/2)}\right).\end{aligned}$$

For the second derivative we differentiate the first derivative by the approach of the product rule and we have

$$\begin{aligned}\frac{\partial^2 ROC(t)}{\partial t^2} &= \Big[\beta(1+\left(\alpha+\beta\tan(\pi t-\pi/2)^2\right)\Big)^{-1}\left(-\left(\cos^2(\pi t-\pi/2)\right)^{-2}\right)(2\cos(\pi t-\pi/2))\\ &\quad(-\sin(\pi t-\pi/2))\pi\Big]+\Big[\cos^2(\pi t-\pi/2)\left(-\left(1+\left(\alpha+\beta\tan(\pi t-\pi/2)^2\right)^{-2}\right.\right.\\ &\quad(2\alpha+2\beta\tan(\pi t-\pi/2))\left(\frac{\beta\pi}{\cos^2(\pi t-\pi/2)}\right)\Big],\end{aligned}$$

by simplification of the second derivative and equating it to zero, it follows that,

$$\frac{2\sin(\pi t-\pi/2)\times\pi}{(1+(\alpha+\beta\tan(\pi t-\pi/2))^2)\,(\cos(\pi t-\pi/2))^3}-\frac{2(\alpha+\beta\tan(\pi t-\pi/2))\times\beta\pi}{(1+(\alpha+\beta\tan(\pi t-\pi/2))^2)^2}=0$$

$$\sin(\pi t-\pi/2)\,(1+(\alpha+\beta\tan(\pi t-\pi/2)))=\beta(\alpha+\beta\tan(\pi t-\pi/2))(\cos(\pi t-\pi/2))^3$$

$$\implies \tan(\pi t-\pi/2)(1+(\alpha+\beta\tan(\pi t-\pi/2))=\beta(\alpha+\beta\tan(\pi t-\pi/2))(\cos(\pi t-\pi/2))^2.$$

Let

$$D=\tan(\pi t-\pi/2)$$

Then we have the expression

$$D\,(1+\alpha+\beta D)=(\alpha+\beta D)\times\beta\left(\cos^2(\pi t-\pi/2)\right) \tag{4.18}$$

But

$$\cos^2(\pi t-\pi/2)=\frac{1}{1+\tan^2(\pi t-\pi/2)} \tag{4.19}$$

Putting (4.19) into (4.18) we have

$$\begin{aligned}D+\alpha D+\beta D^2 &= (\alpha+\beta D)\times\frac{\beta}{1+D^2}\\ \Rightarrow D+\alpha D+\beta D^2 &= \frac{\alpha\beta}{1+D^2}+\frac{\beta^2 D}{1+D^2}.\end{aligned}$$

By simplifying the above expression we have

$$\beta D^4+\alpha D^3+D^3+\beta D^2+\alpha D-\beta^2 D+D-\alpha\beta=0, \tag{4.20}$$

If we assume $\alpha=1$ and $\beta=1$ then we have

$$D^4+2D^3+D^2+D-1=0. \tag{4.21}$$

It is not trivial to find the roots of the polynomial (4.21), we find the roots numerically and the real roots are $D_1 = 0.48402$ and $D_2 = -1.8972$. Since

$$D = \tan(\pi t - \pi/2),$$

then for $D_1 = 0.48402$ we have

$$\begin{aligned}
&\tan(\pi t - \pi/2) = 0.48402 \\
&(\pi t - \pi/2) = \arctan(048402) \\
&\pi t = \text{arctann}(048402) + \pi/2. \\
&t = \frac{1}{\pi}\arctan(048402) + 2. \\
&\Rightarrow t_1 = 10.221,
\end{aligned}$$

also for $D_2 = -1.8972$ we have

$$\begin{aligned}
\tan(\pi t - \pi/2) &= -1.8972 \\
\pi t - \pi/2 &= \arctan(-1.8972) \\
\pi t &= \arctan(-1.8972) + \pi/2 \\
t &= \frac{1}{\pi}\arctan(-1.8972) + 2 \\
\Rightarrow t_2 &= -17.801.
\end{aligned}$$

From above we can infer that there are many roots of the second derivative and finding the point by which the Bi-Cauchy ROC curve inflects will be complex. This is because the points t_1 and t_2 above were obtained when we assumed the constants $\alpha = 1$ and $\beta = 1$. How about the cases when there are different values of α and β. We find the roots numerically to see the behavior for different values of the constants by using the expression (4.20). Since the value of D is known then the second derivative for $t \in (0, 1)$ is

$$\begin{aligned}
\beta\tan^4(\pi t - \pi/2) + \alpha\tan^3(\pi t - \pi/2) + \tan^3(\pi t - \pi/2) + \beta\tan^2(\pi t - \pi/2) + \\
\alpha\tan(\pi t - \pi/2) - \beta^2\tan(\pi t - \pi/2) + \tan(\pi t - \pi/2) - \alpha\beta = 0.
\end{aligned} \tag{4.22}$$

We simulate using the function (4.22) to see the behavior of the function and this can be seen in the figure below. From the figure we see for different values of α, β and $t \in (0, 1)$ all the curves remains constant at the same point in the middle before the switch which shows the behavior of the points of inflection of the Bi-Cauchy ROC curve.

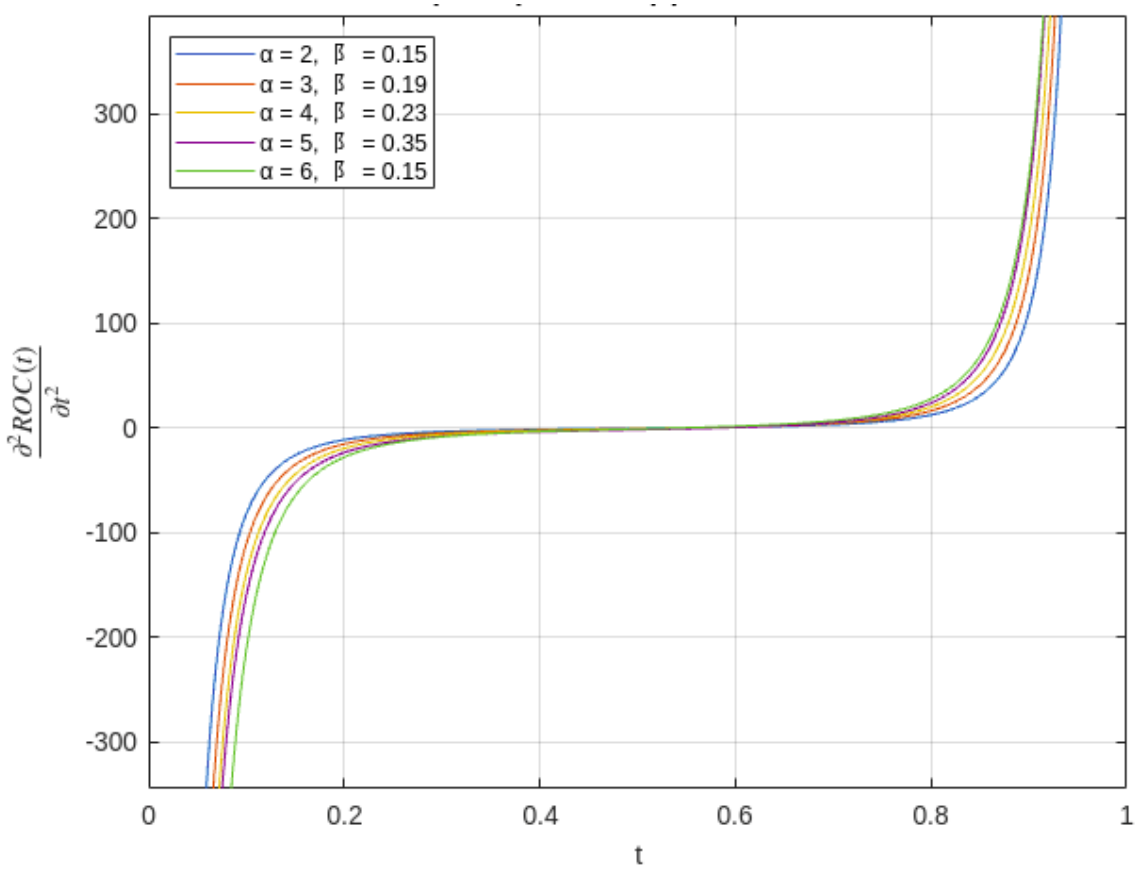


Figure 12: Point of Inflection of the Bi-Cauchy ROC curve for different values of α and β

CASE STUDY

Suppose that the test results are Cauchy distributed in the diseased and non-diseased populations with location parameter $(\mu_D, \mu_{\bar{D}})$ and scale parameter $(1, 1)$ thus the case

$$Y_D \sim C(\mu_D, 1) \qquad Y_{\bar{D}} \sim C(\mu_{\bar{D}}, 1).$$

Then for any threshold c, we can express the FPR and TPR using the cdf of standard Cauchy distribution as

$$\begin{aligned} \text{FPR}(c) &= \text{P}\left[Y_{\bar{D}} > c\right] = F\left(\frac{\mu_{\bar{D}} - c}{1}\right) \\ \text{TPR}(c) &= \text{P}\left[Y_D > c\right] = F\left(\frac{\mu_D - c}{1}\right), \end{aligned}$$

to arrive at a simple expression for ROC(t) for the Cauchy distributed test, we take the inverse function of the first term $FPR(c)$, we can see that $c = \mu_{\bar{D}} - F^{-1}(t)$ which will be the corresponding threshold for the test positivity criterion. Therefore,

$$\begin{aligned} \text{ROC}(t) = \text{TPR}(c) &= F\left(\frac{\mu_D - c}{1}\right) \\ &= F\left(\mu_D - (\mu_{\bar{D}} - F^{-1}(t))\right) \\ &= F\left(\mu_D - \mu_{\bar{D}} + F^{-1}(t)\right), \end{aligned} \tag{4.23}$$

hence for Cauchy distributed diagnosing random variable with distribution parameters $C(\mu_D, 1)$ and $C(\mu_{\bar{D}}, 1)$ the Bi-Cauchy model can be written as

$$ROC(t) = F(\mu_D - \mu_{\bar{D}} + F^{-1}(t)). \tag{4.24}$$

Since the functions F and $F^{-1}(t)$ are known the Bi-Cauchy ROC model is

$$ROC(t) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\mu_D - \mu_{\bar{D}} + \tan\left(\pi t - \frac{\pi}{2}\right)\right) \quad \forall t \in (0,1) \tag{4.25}$$

5 RESULTS OF SIMULATIONS

5.1 Results 1

In this section, I performed simulations based on the previously described model in pyhton software version 3.6.0]. We have our Bi-Cauchy ROC model

$$ROC(t) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\mu_D - \mu_{\bar{D}} + \tan\left(\pi t - \frac{\pi}{2}\right)\right) \quad \forall t \in (0,1)$$

we simulate for different values of mu_D and $\mu_{\bar{D}}$ and compared it with the empirical ROC curve as discussed earlier in (4.3). I chose a random set of data for $n = 100$ for both diseased and non-diseased population. The data was chosen from [cardiovascular-disease-dataset]. The discussed estimators in section 3 thus, Huber M-estimate, Hodges Lehmann estimate(R-estimate), Sen's mean(L-estimator), Trimmed mean(L-estimator) and the Winsorized mean(L-estimator) were used to estimate the the chosen data set for both diseased and non-diseased population.The data were chosen just get a values for each estimates in order to apply in the Bi-Cauchy model.

The estimator values were then used to generate a random set of observations which are Cauchy distributed for both the diseased and non-diseased population The different values of the estimates were set as the location parameter μ_D and $\mu_{\bar{D}}$ to perform these simulations.

The Cauchy random variables generated were used to plot different empirical ROC curves and compared with Bi-Cauchy ROC. The results of the estimator values are shown in the table below

Estimator	Diseased Population	Non-Diseased Population
Mean($\bar{x}$)	26.0831	25.0013
Huber M-estimate(M-estimator)	25.7658	25.0784
Hodges Lehmann Estimate(R-estimator)	25.7850	25.1625
Sen' Estimate(L-estimator)	26.0831	25.0013
5% Trimmed Mean(L-estimator)	25.7658	25.0698
10% Trimmed Mean(L-estimator)	25.7748	25.25175
5% Winsorized Mean(L-estimator)	25.7467	24.9033
10% Winsorized Mean(L-estimator)	25.7408	24.9234

Table 1: Table Showing different estimator values

The results for the empirical and Bi-Cauchy ROC curves for the different estimates with their AUCs are shown as follows

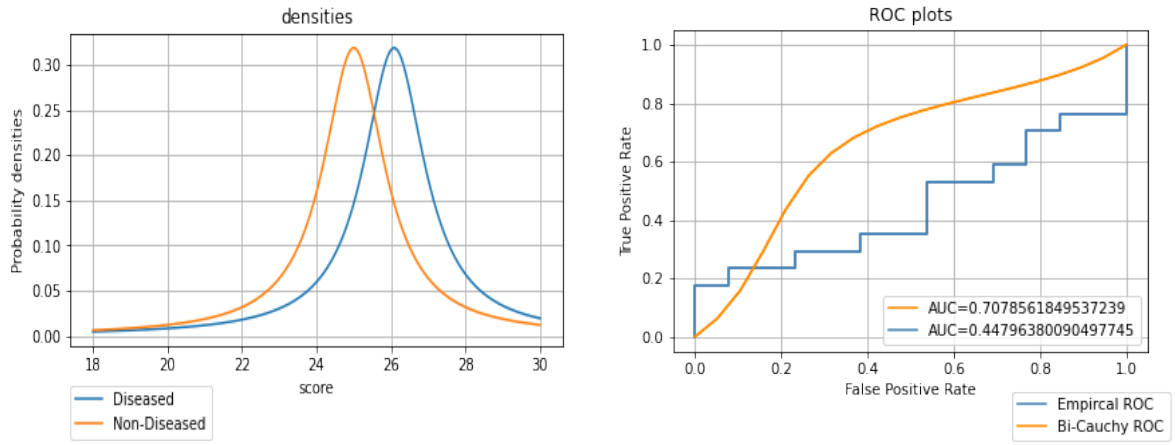


Figure 13: ROC curve when the diseased and non-diseased class follows a Cauchy distribution with the Mean ($\mu_D = 26.0831$ and $\mu_{\bar{D}} = 25.0013$ estimates in Table 1)

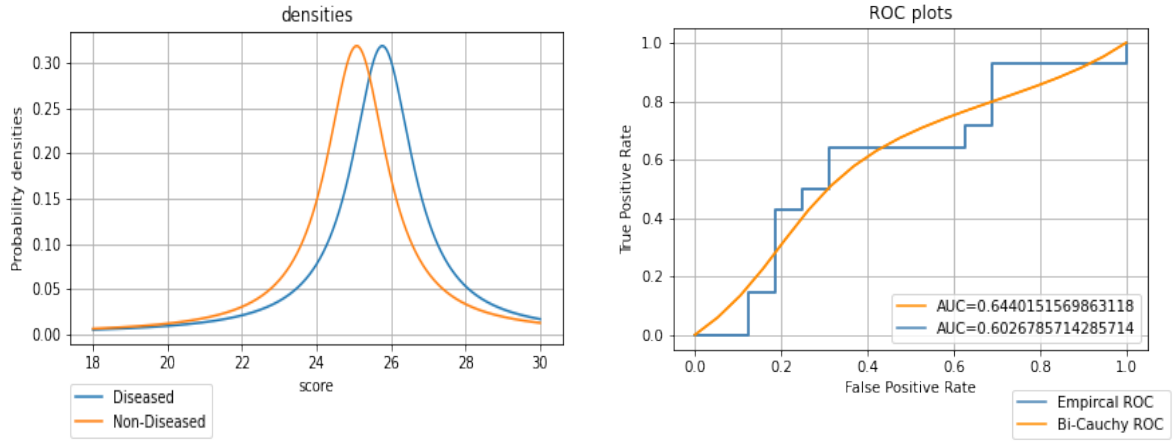


Figure 14: ROC curve when the diseased and non-diseased class follows a Cauchy distribution with the Huber M-estimate ($\mu_D = 25.7658$ and $\mu_{\bar{D}} = 25.0784$ estimates in Table 1)

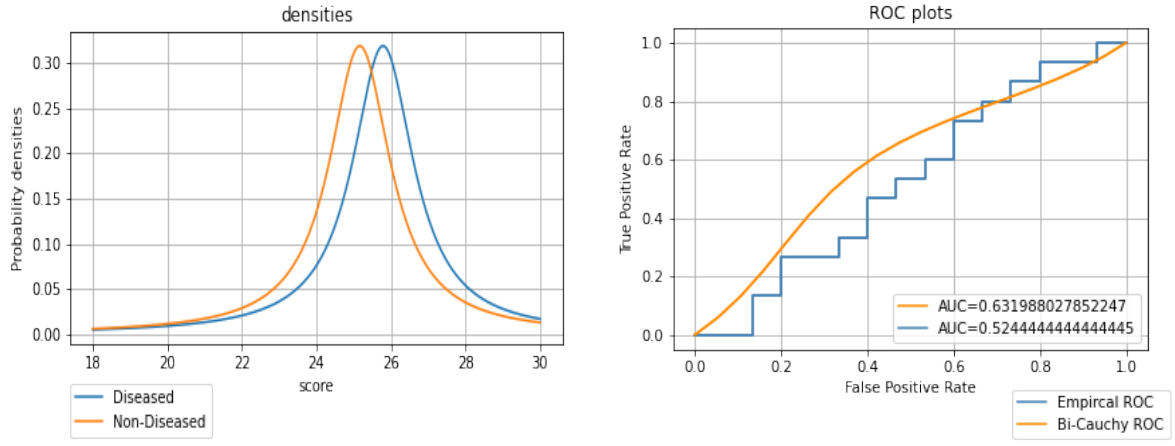


Figure 15: ROC curve when the diseased and non-diseased class follows a Cauchy distribution with the Hodges Lehmann Estimate ($\mu_D = 25.7850$ and $\mu_{\bar{D}} = 25.1625$ estimates in Table 1)

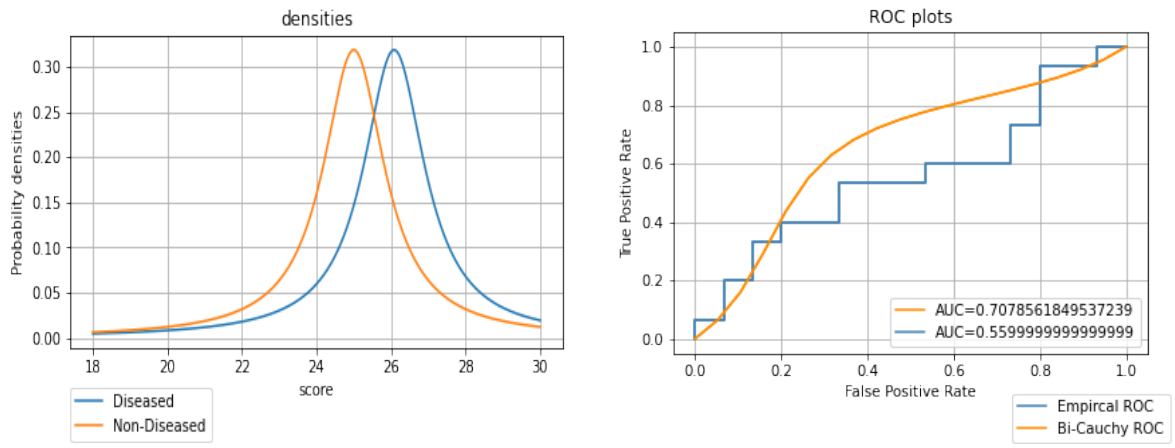


Figure 16: ROC curve when the diseased and non-diseased class follows a Cauchy distribution with the Sen Weighted Mean ($\mu_D = 26.0831$ and $\mu_{\bar{D}} = 25.0013$ estimates in Table 1)

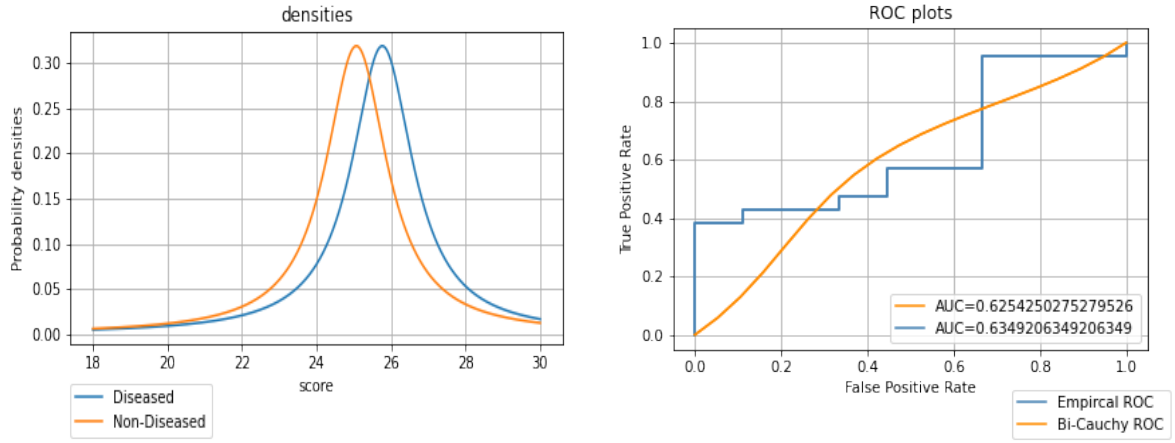


Figure 17: ROC curve when the diseased and non-diseased class follows a Cauchy distribution with 5%-Trimmed Mean ($\mu_D = 25.7658$ and $\mu_{\bar{D}} = 25.0698$ estimates in Table 1)

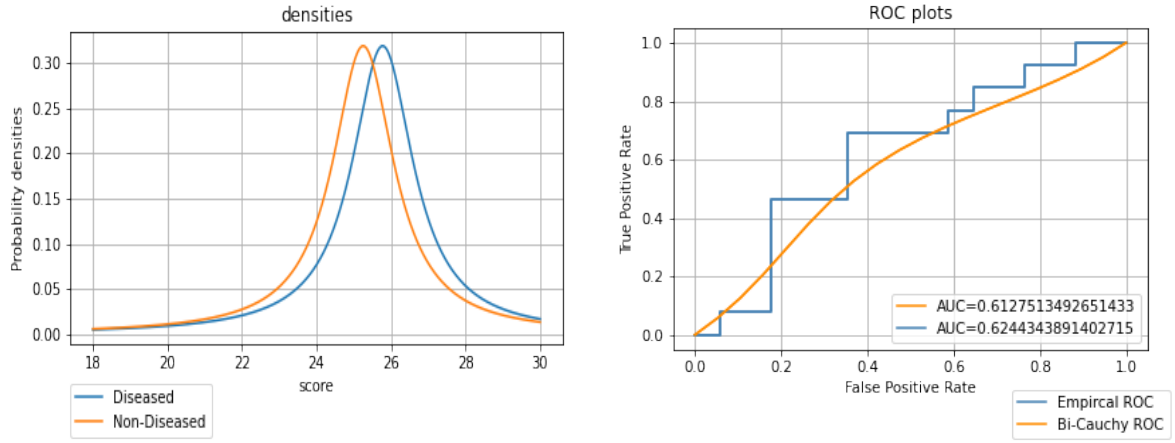


Figure 18: ROC curve when the diseased and non-diseased class follows a Cauchy distribution with 10%-Trimmed Mean ($\mu_D = 25.7748$ and $\mu_{\bar{D}} = 25.25175$ estimates in Table 1)

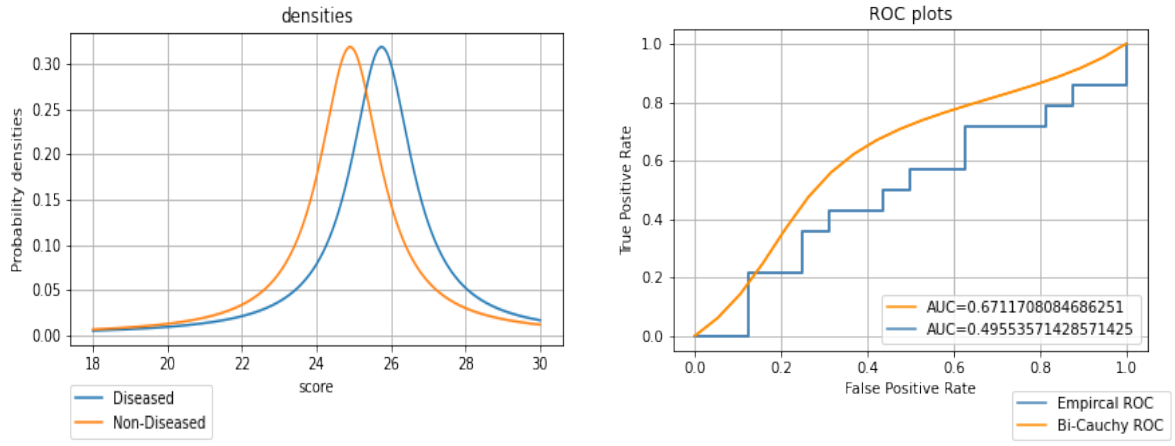


Figure 19: ROC curve when the diseased and non-diseased class follows a Cauchy distribution with 5%- Winsorized Mean ($\mu_D = 25.7467$ and $\mu_{\bar{D}} = 25.9033$ estimates in Table 1)

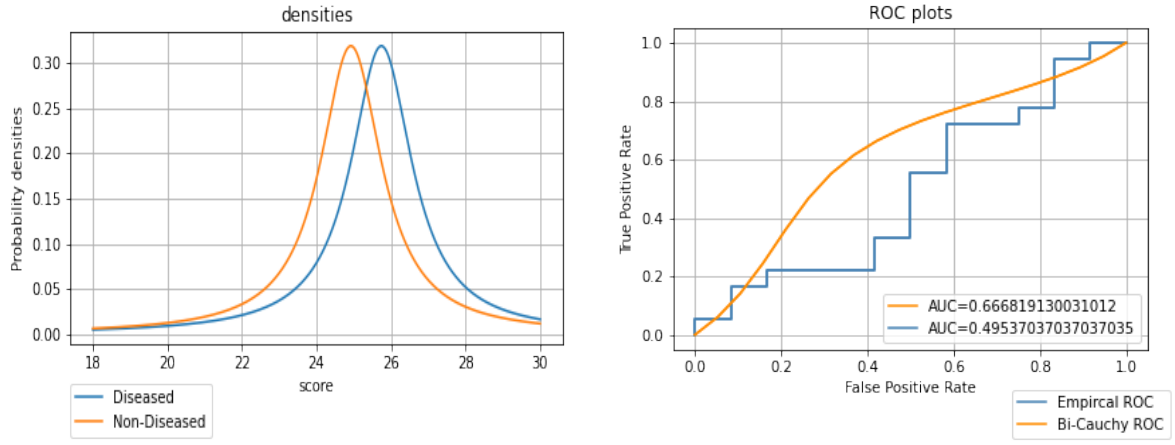


Figure 20: ROC curve when the diseased and non-diseased class follows a Cauchy distribution with 10%- Winsorized Mean ($\mu_D = 25.7408$ and $\mu_{\bar{D}} = 24.9234$ estimates in Table 1)

From the above ROC plots, it is evident that, the estimate with the average mean Figure 13 shows a vast difference between the empirical ROC and the true(Bi-Cauchy) ROC curve as compared to the other estimates. Also, comparing the AUCs, the AUC for the empirical is far lower than that of the Bi-Cauchy ROC curve as compared to the difference between the AUCs of other estimates.This also shows why the mean deviates for Cauchy distribution and why the weak law of large numbers does not hold for Cauchy distribution as explained in Proposition 2.6.

Also, the same values of the estimators are considered and the empirical is been simulated by a combination of all the data sets from both the diseased and non-diseased population from each of the estimators. The result of the simulation is seen in the Figure 21. From the figure, we can see that estimation with the mean moves farther from the empirical as compared to the other estimators. In the order of suitability the closest one to the empirical is the 10%-Trimmed mean followed by the Hodges Lehmann estimate, Huber M-estimator, 5%-Trimmed mean, 10%-Winsorized mean, 5%-Winsorized mean, Sen's mean and the last was the average mean.

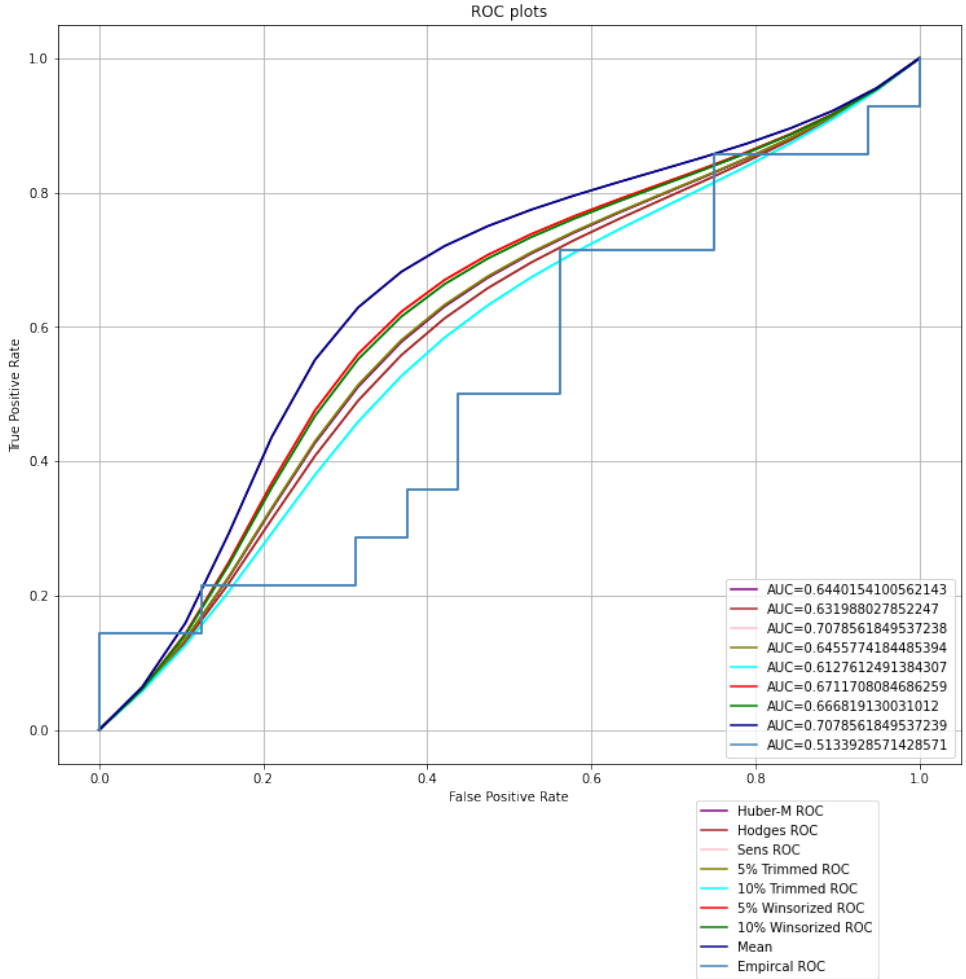


Figure 21: ROC Plots comparing the empirical ROC and all Bi-Cauchy ROC of different estimates

5.2 Results 2

A new set of simulations were performed to check how the various estimators can be applied to Cauchy distributed data using the same ROC approach. I choose two constant values ($\mu_D = 28$) and ($\mu_{\bar{D}} = 26.5$) as the location parameter for both diseased and non-diseased subjects with the same model

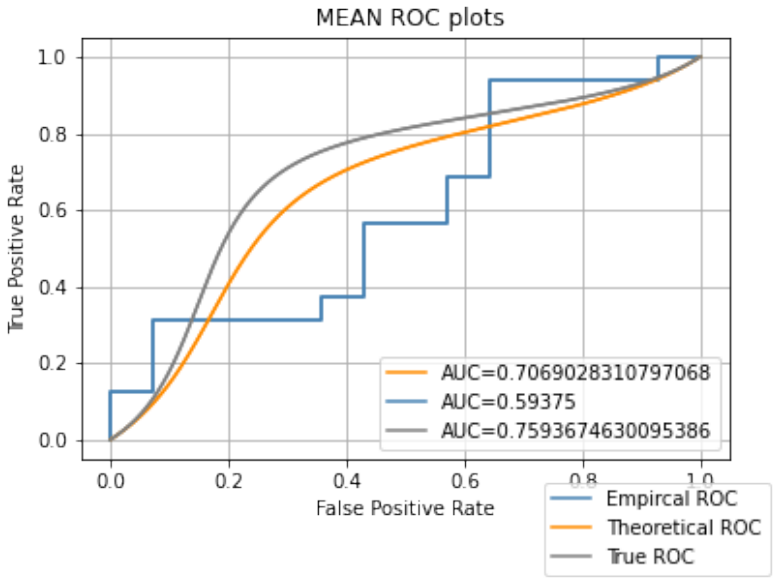
$$ROC(t) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\mu_D - \mu_{\bar{D}} + \tan\left(\pi t - \frac{\pi}{2}\right)\right) \quad \forall t \in (0,1)$$

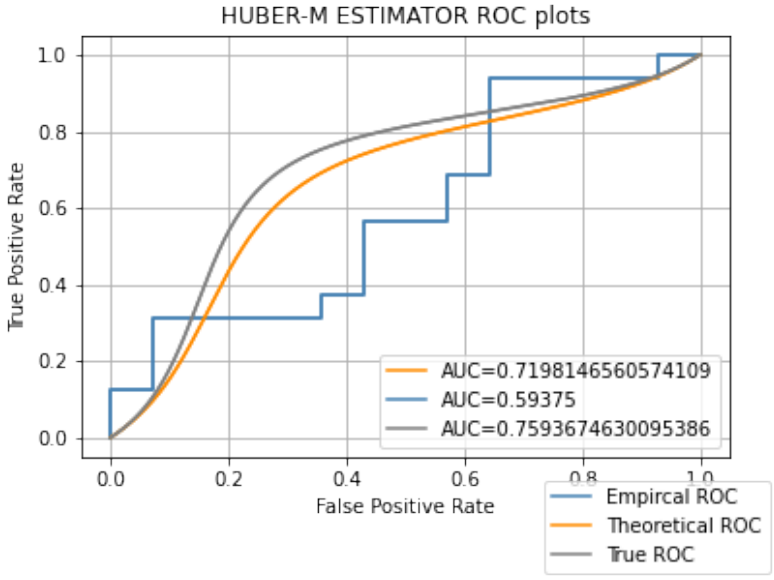
The constant values ($\mu_D = 28$) and ($\mu_{\bar{D}} = 26.5$) were used to generate Cauchy distributed random variables, 100 for each value. The various estimators were then used to estimate each of the set of variables for both the diseased and non-diseased populations and applied in ROC plots. The constant values were used to plot the true Bi-Cauchy ROC curves and the estimated values were used for theoretical Bi-Cauchy ROC curves and plotted with the empirical ROC. The empirical ROC was generated from the same 100 set of randomly generated variable for both populations. Results for the estimator values are shown in the table below

Estimator	Diseased Population	Non-Diseased Population
Mean($\bar{x}$)	27.879129	26.804028
Huber M-estimate(M-estimator)	27.86196	26.69340
Hodges Lehmann Estimate(R-estimator)	27.85858	26.64251
Sen's Estimate(L-estimator)	27.87913	26.80403
5% Trimmed Mean(L-estimator)	27.90481	26.72807
10% Trimmed Mean(L-estimator)	27.86493	26.69325
5% Winsorized Mean(L-estimator)	27.91092	26.775005
10% Winsorized Mean(L-estimator)	27.90945	26.70798

Table 2: Table Showing different estimator values for

The ROC plots comparing the Empirical ROC curve, True ROC curve and the Theoretical ROC curve for two set of observations. Small sample size $n = 100$ and large sample size $n = 500$ are shown as follows





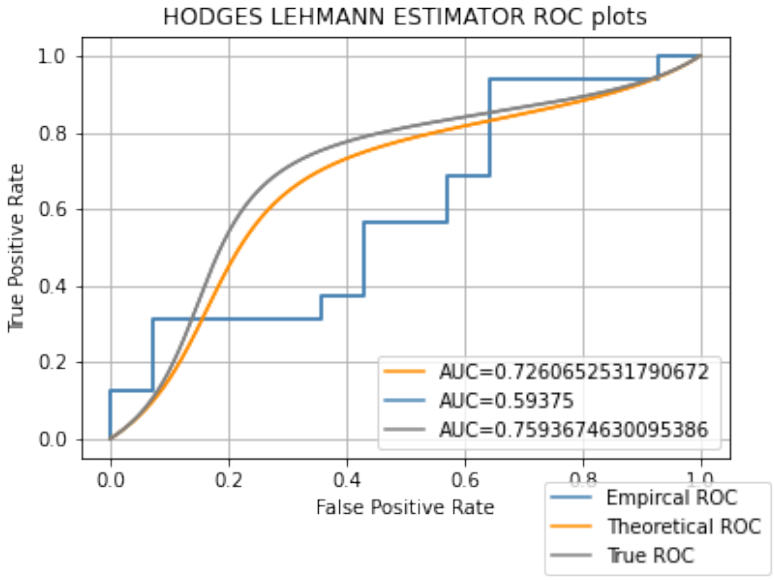


Figure 22: Simulation of small sample size n=100

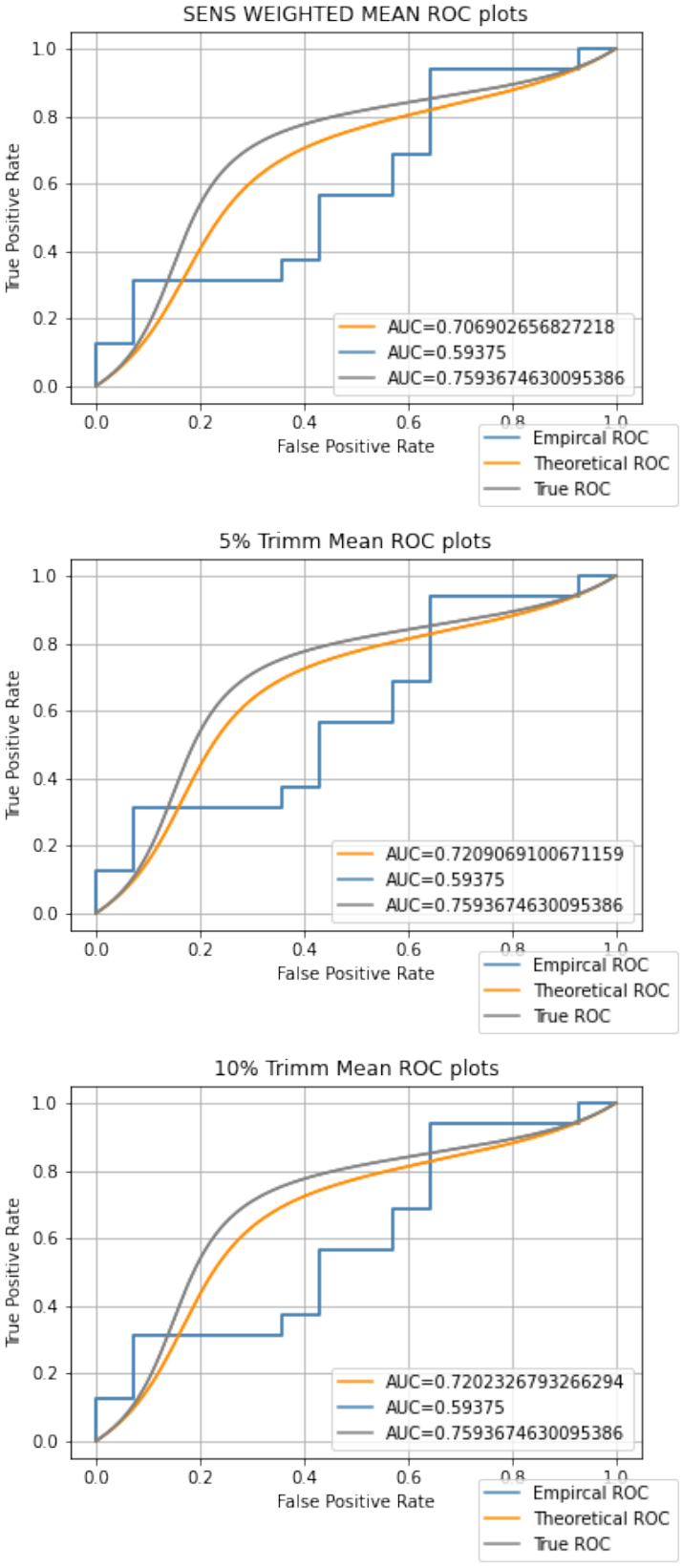


Figure 23: Simulation of small sample size n=100

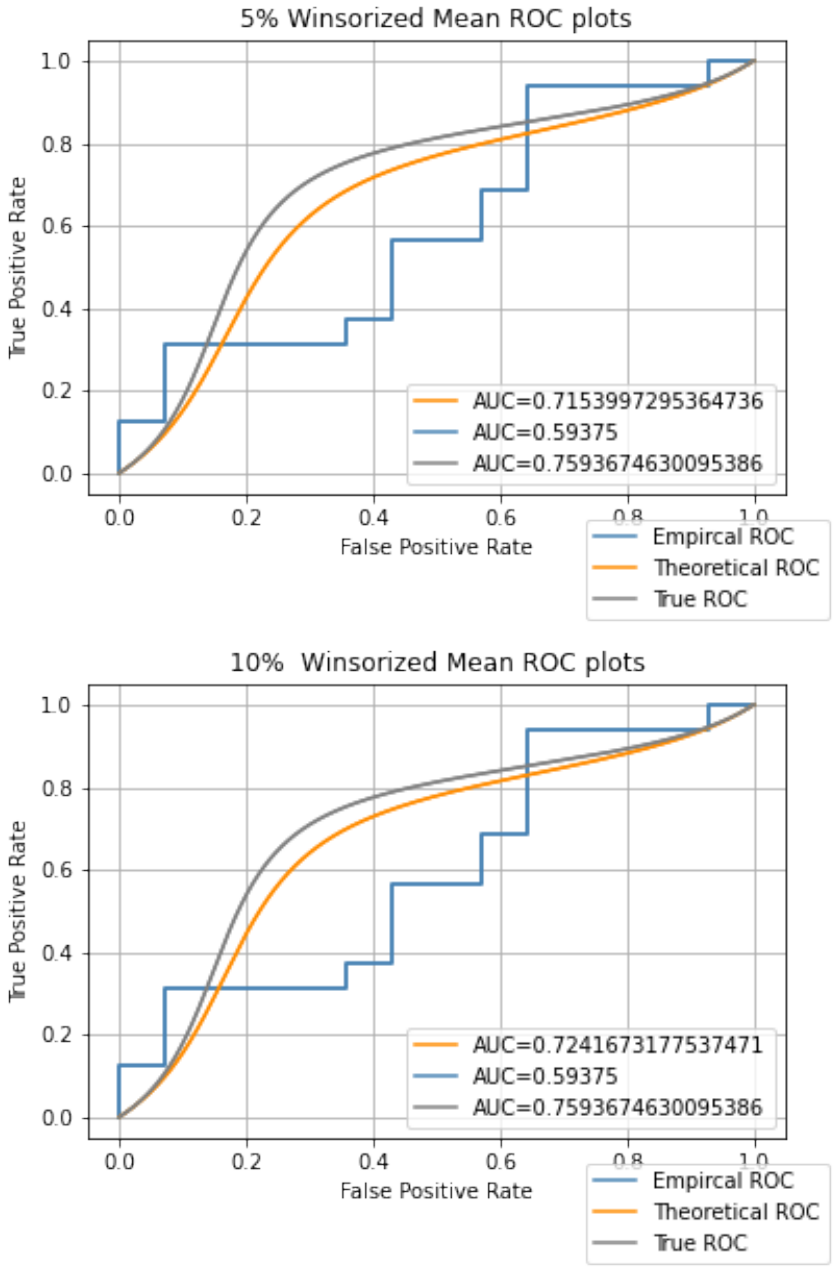


Figure 24: Simulation of small sample size n=100

With reference to the ROC plots above, we can see by the AUC,s of the theoretical Bi-Cauchy ROC curve which is very close to the true Bi-Cauchy ROC curve is the Hodges Lehmann estimator followed by 10%-Winsorized mean, 10%-Trimmed mean, 5%-Trimmed mean, Huber M-estimator, 5%-Winsorized mean, Sens mean and the average mean. The Sens mean and the average mean are the most farther from the true Bi-Cauchy ROC curve. Figure (6) below shows a comparison of the various true Bi-Cauchy ROC, empirical roc and theoretical ROC plots of the estimates.

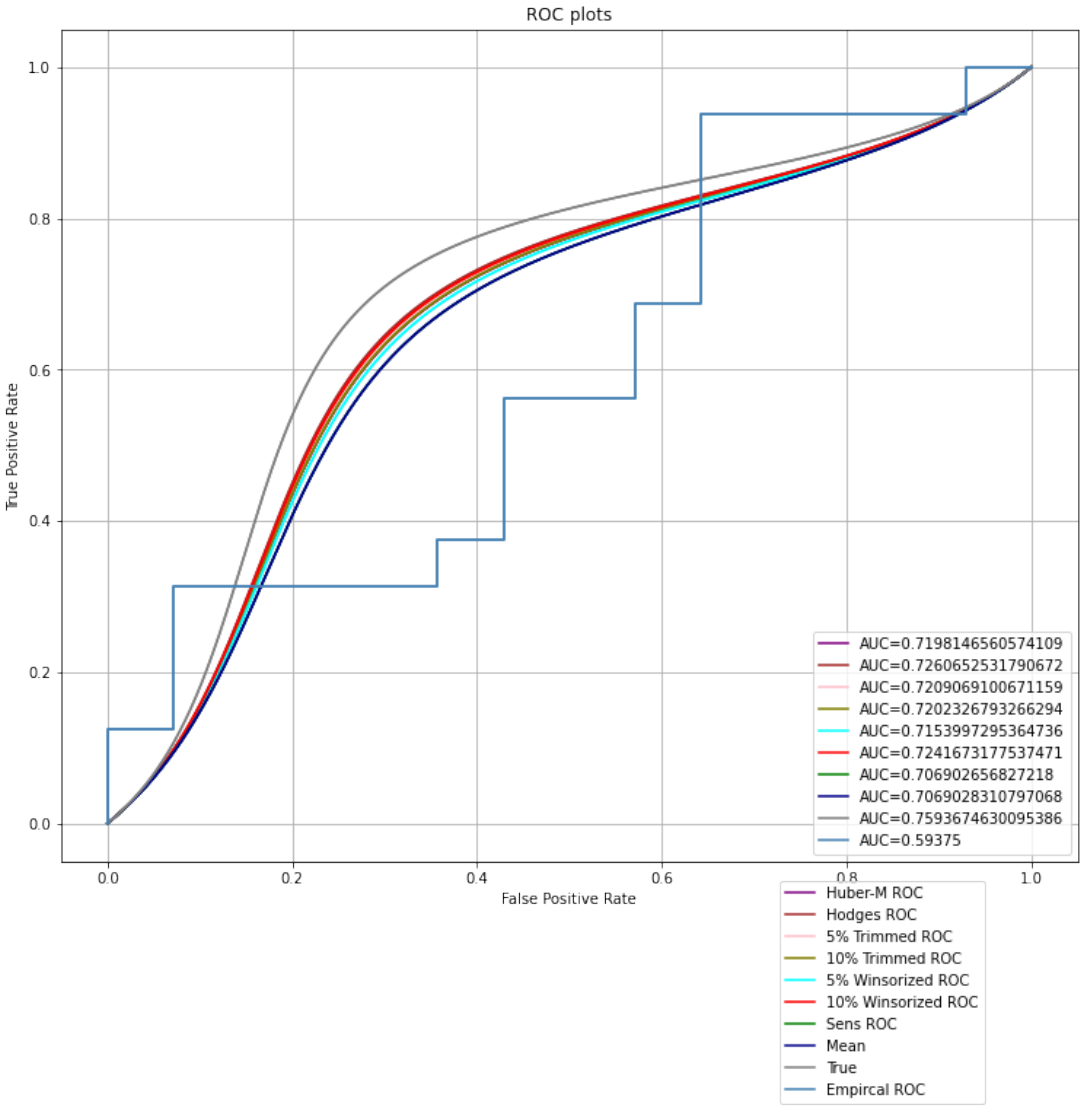


Figure 25: Simulation of small sample size $n = 100$ of the True Bi-Cauchy ROC, Theoretical Bi-Cauchy ROC and the empirical ROC of each of the estimators estimators.

We simulated for a large sample size, for $n = 500$. Large sample size increases the estimator values and this increases the AUCs of the Bi-Cauchy ROC curves because the higher the difference between μ_D and $\mu_{\bar{D}}$ the more concave the Bi-Cauchy ROC curves. But this reduces the AUC of the empirical because the empirical is dependent on the number of observation n as studied in section 4.3.

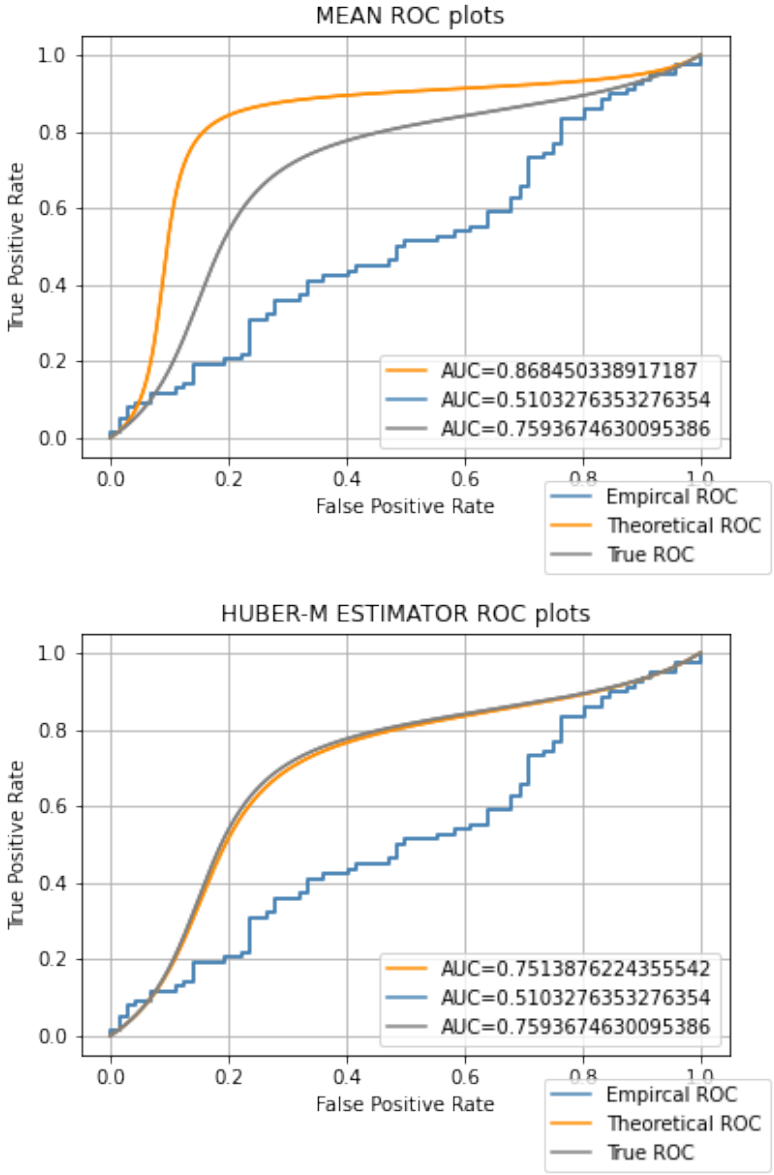


Figure 26: Simulations of Large sample size $n = 500$

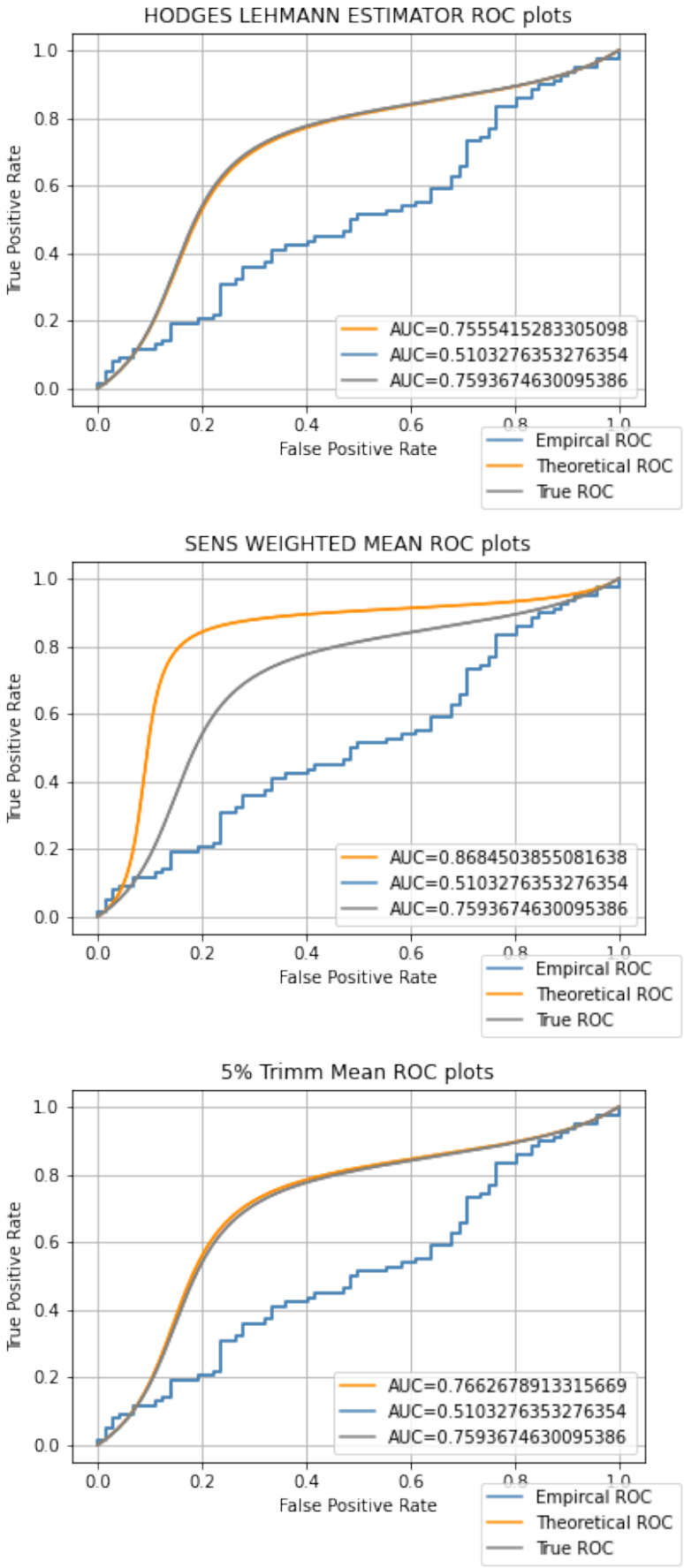


Figure 27: Simulations of Large sample size $n = 500$

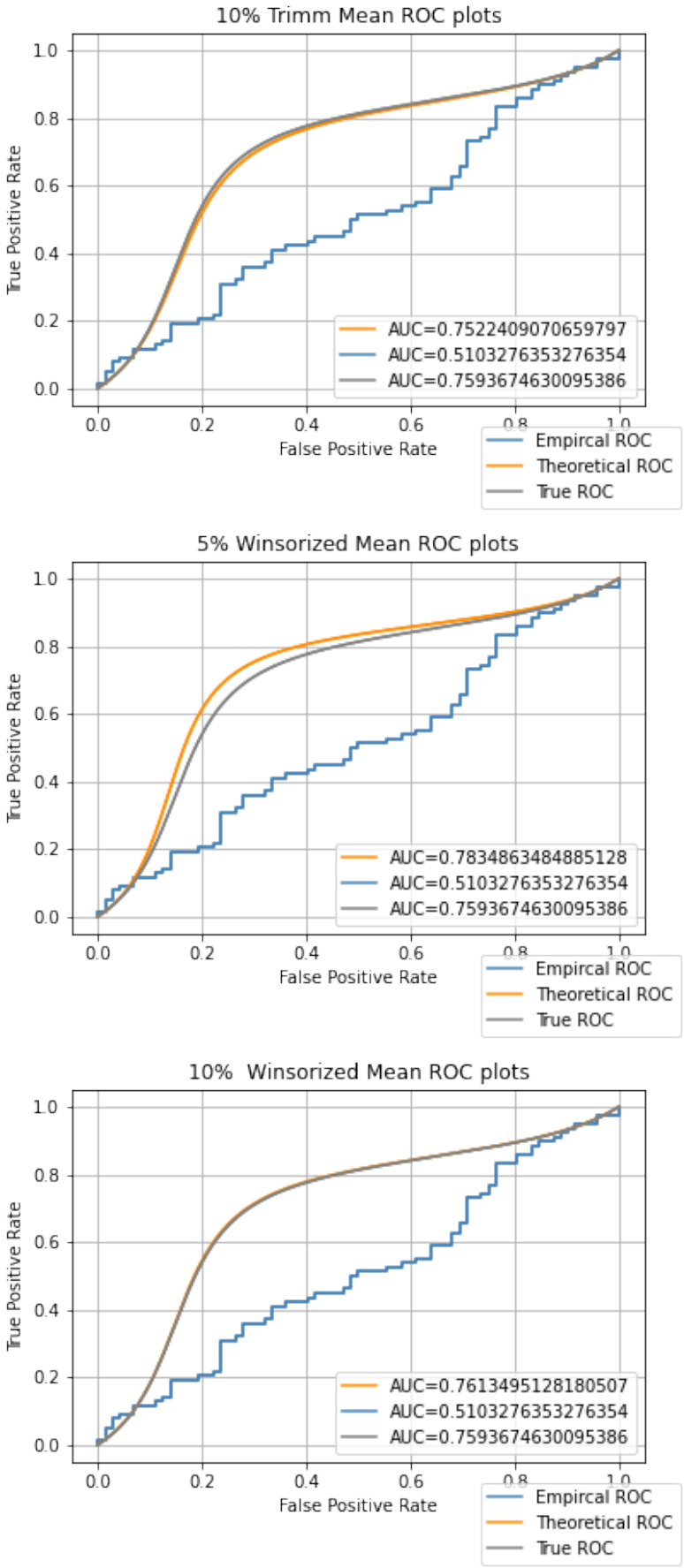


Figure 28: Simulations of Large sample size $n = 500$

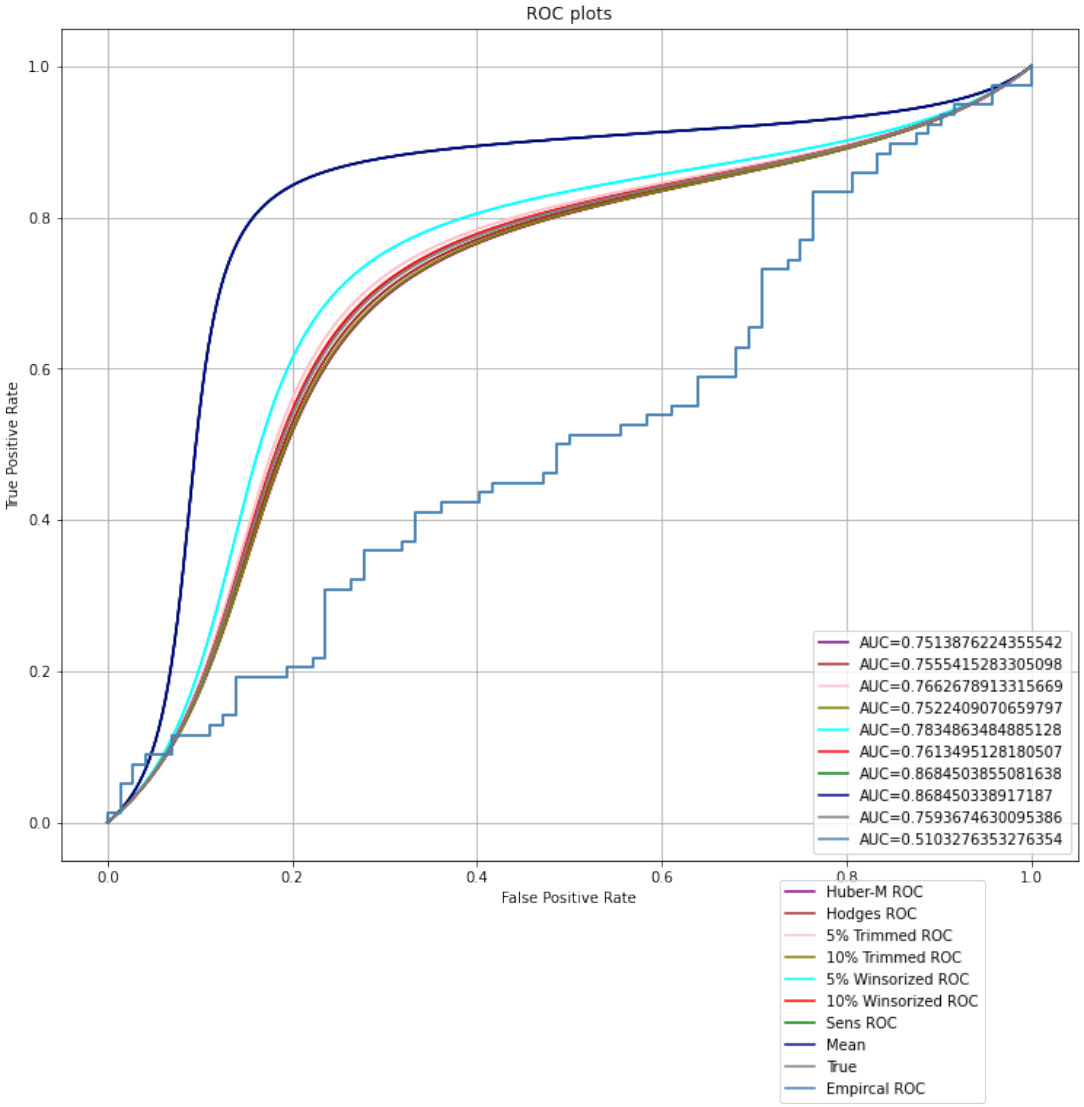


Figure 29: Simulations of Large sample size $n = 500$ of all the estimates

The values for both the diseased and non-diseased population were set in a such a way that the difference will be a more bit wider. We set the diseased population to $\mu_D = 28$ and non-diseased to $\mu_{\bar{D}} = 22$. Simulation done for both small sample $n = 100$ size and large sample size $n = 500$ to see how the concavity of the ROC improves. Results is seen below

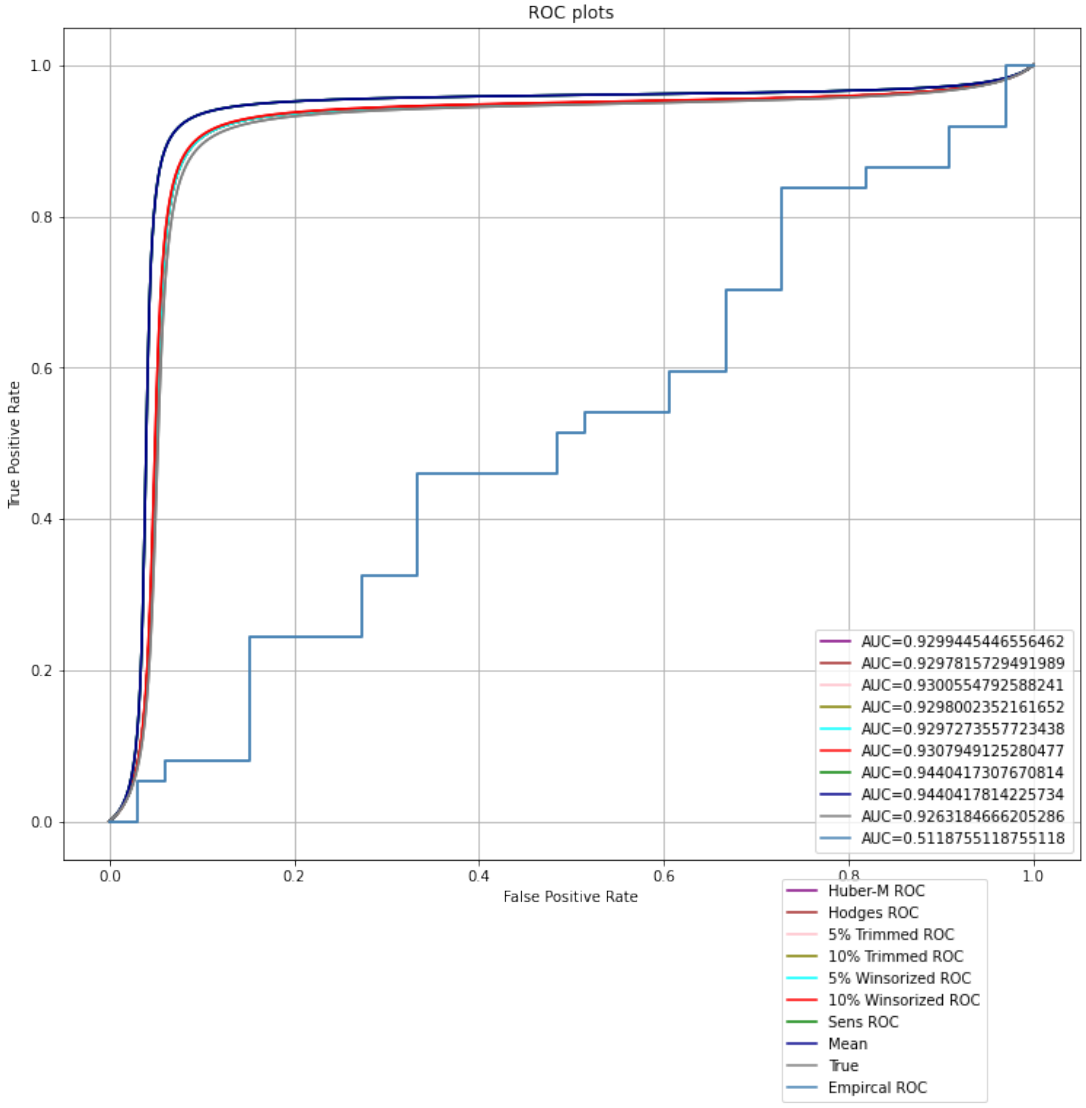


Figure 30: Simulations of sample size $n = 100$ of all the estimates with $\mu_D = 28$ and $\mu_{\bar{D}} = 22$

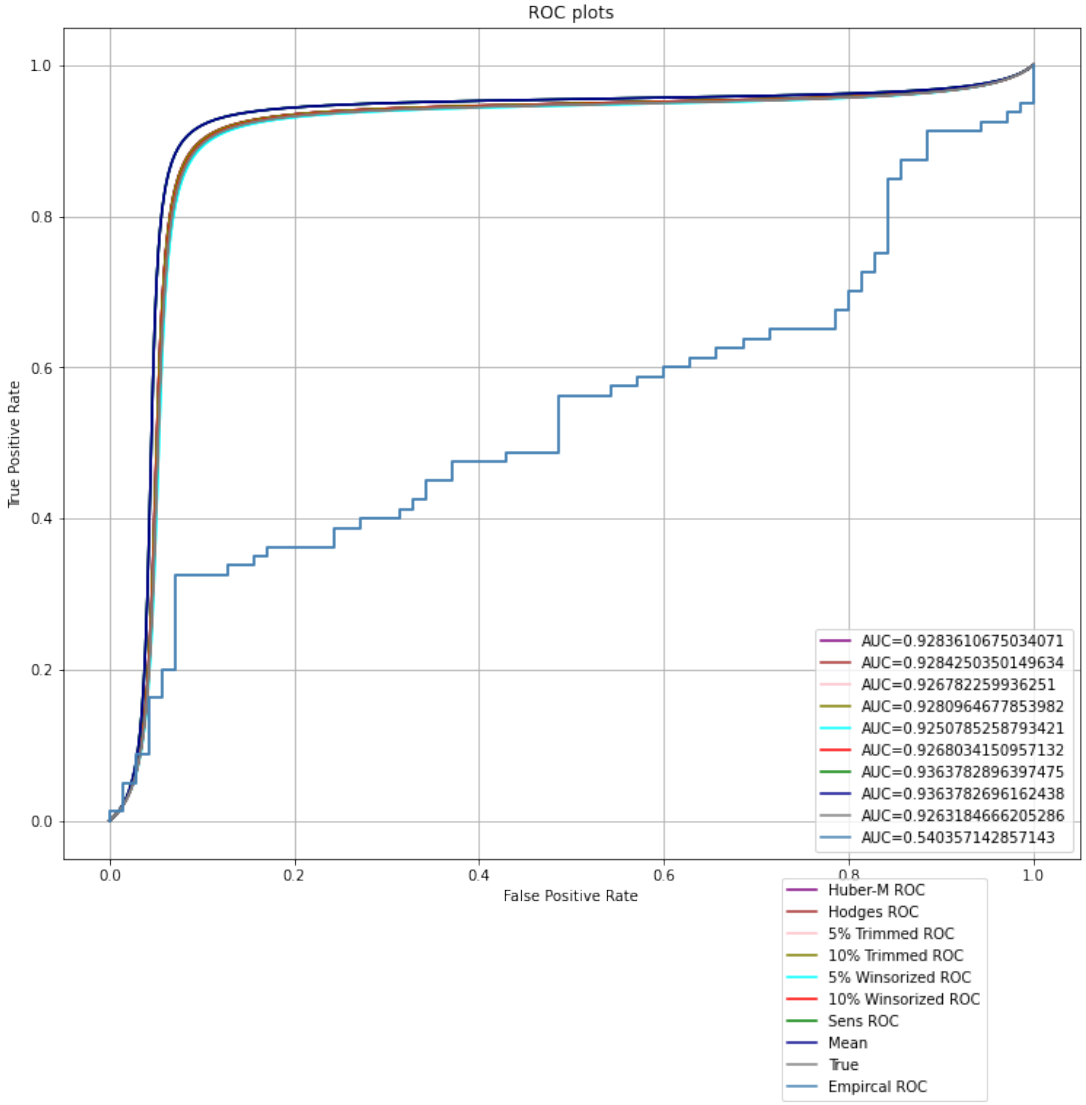


Figure 31: Simulations of sample size $n = 500$ of all the estimates with $\mu_D = 28$ and $\mu_{\bar{D}} = 22$

5.3 Theoritical Comments

- Comparing both simulations in section 5.1 and 5.2, in both cases the ROC estimate based on the average estimate of μ_D and $\mu_{\bar{D}}$ deviates from all the other estimates in its estimation which assert to the fact that the conditions for the weak law of large numbers are not fulfilled and thus the average does not converge to the location parameter in probability. This can be seen from Figure 21, 25 and 29.
- Furthermore, we were considering the order of their performance in a model with Cauchy distribution. To compare from the two results, estimates with the Hodges Lehmann estimator, Huber M-estimate, both 5% and 10% Trimmed mean, 5% Winsorized mean and the 10% Winsorized mean show to be more suitable in their estimation. The average mean and the Sens mean were not suitable. In summary, Hodges Lehmann estimator, Huber M-estimate, both 5% and 10% Trimmed mean, 5% Winsorized and the 10% Winsorized are more suitable in the estimation of the location parameter of Cauchy distribution.
- Owing to the fact that, ROC analysis is a method for evaluating the accuracy or performance of medical diagnostic test, then for any real life data that is Cauchy distributed, this technical approach can also be used to give out the best diagnosis to a patient using the robust estimators aforementioned.
- This can be possible by the help of the AUC since it is a summary measure that gives a very meaningful interpretation. In the case of a perfect classification the Bi-Cauchy ROC curve will reach the point of highest theoretical accuracy *i.e* the sensitivity and specificity will both be 100% and the AUC will tend to one (the highest possible value).
- The AUC values was between the range of $0.5 - 0.8$ in the simulations, in analysis of ROC 0.5 gives a non-informative diagnostic curve but all the data used in the simulation were randomly generated not a real life data or accurate.
- In Figure 30 and Figure 31 where we considered a wider difference $\mu_D = 28$ and $\mu_{\bar{D}} = 22$. The AUC values increased and the Bi-Cauchy ROC curves moved a bit higher to the top left corner. [19] argued to the fact that concavity is a characteristic of the parametric ROC curves because it guarantees that the ROC will never cross the main diagonal line. Since the Bi-Cauchy ROC is parametric, then this characteristic is of importance to its functionality.
- The technical approach suggested in this work would also be more important in solving real life problems that requires analysis of data. All that is required is to perform a goodness of fit test on the data set to wrap Cauchy distribution then apply this approach to give results. An example of such statistics which is mentioned in literature for fitting real life data to fit Cauchy distribution is the Watson's U-squared statistics. [20] also gave an approach for performing a goodness of fit test for Cauchy distribution based on empirical characteristic function.

6 CONCLUSION

In this thesis, a new approach of estimation of the location parameter of Cauchy distribution was studied. The main controversies of Cauchy distribution was the difficulties in estimating its parameters. Numerous approaches have already been used in literature both numerical and analytical approach to estimate the parameters. A set of robust estimators were suggested in this work and were later applied in the receiver operating characteristic curves where a number of simulations were performed based on a model derived in Cauchy sense called the Bi-Cauchy ROC curve. This thesis is divided into six chapters. In chapter 2, we studied some properties of Cauchy distribution, in particular the expected value, the characteristic function and the cumulative distribution function.

Furthermore, in chapter 3 we studied the theoretical operation of the different types of robust estimators which are the Huber M-estimator, Hodges Lehmann estimator, Sens mean, Trimmed mean and the winsorized mean. In Chapter 4, we studied the receiver operating Characteristic (ROC) curve where we derived Bi-Cauchy ROC model and applied it in the simulation of data where we considered mainly diseased and non-diseased population.

Chapter 5 was mainly devoted to simulations and results. The simulations were performed in Python software in Google Colab [version 3.6.0]. Two approaches were used to perform the simulations and some theoretical comments were given. The results showed that although estimation of the location parameters Cauchy distribution is difficult but the suggested robust estimators are suitable in its operation.

BIBLIOGRAPHY

[1] G. Casella and R. L. Berger, *Statistical inference*. Cengage Learning, 2021.

[2] S. M. Stigler, "Studies in the history of probability and statistics. cauchy and the witch of agnesi: An historical note on the cauchy distribution," *Biometrika*, pp. 375–380, 1974.

[3] B. P. Roe, *Probability and statistics in experimental physics*. Springer Science & Business Media, 2012.

[4] Y. Kagan, "Correlations of earthquake focal mechanisms," *Geophysical Journal International*, vol. 110, no. 2, pp. 305–320, 1992.

[5] S. Stapf, R. Kimmich, and R.-O. Seitter, "Proton and deuteron field-cycling nmr relaxometry of liquids in porous glasses: evidence for lévy-walk statistics," *Physical review letters*, vol. 75, no. 15, p. 2855, 1995.

[6] M. J.Jureckova, J.Picek, *Robust statistical methods with R*. Chapman and Hall/CRC, 2nd ed. ed., 2019.

[7] P. Qiu *et al.*, "The statistical evaluation of medical tests for classification and prediction,margaret '.sullivan pepe," *Journal of the American Statistical Association*, vol. 100, pp. 705–705, 2005.

[8] R. B. Patterson, *Goodness-of-fit tests and function estimators for receiver operating characteristic (ROC) curves: Inference from perpendicular distances*. George Mason University, 2012.

[9] S. Zacks, *Examples and Problems in Mathematical Statistics*. John Wiley & Sons, 2013.

[10] Z. Bai and J. Fu, "On the maximum-likelihood estimator for the location parameter of a cauchy distribution," *Canadian Journal of Statistics*, vol. 15, no. 2, pp. 137–146, 1987.

[11] J. Zhang, "A highly efficient l-estimator for the location parameter of the cauchy distribution," *Computational statistics*, vol. 25, no. 1, pp. 97–105, 2010.

[12] K. Hanson and D. Wolf, "Estimators for the cauchy distribution," in *Maximum Entropy and Bayesian Methods*, pp. 255–263, Springer, 1996.

[13] S. Fegyverneki, "A simple robust estimation for parameters of cauchy distribution," *Miskolc Math. Notes*, vol. 14, no. 3, pp. 887–892, 2013.

[14] S. Schuster, "Parameter estimation for the cauchy distribution," in *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 350–353, IEEE, 2012.

[15] J. A. Swets, R. M. Dawes, and J. Monahan, "Better decisions through science," *Scientific American*, vol. 283, no. 4, pp. 82–87, 2000.

[16] A. Jokiel-Rokita and M. Pulit, "Nonparametric estimation of the roc curve based on smoothed empirical distribution functions," *Statistics and Computing*, vol. 23, no. 6, pp. 703–712, 2013.

[17] J. A. Hanley, "The robustness of the" binormal" assumptions used in fitting roc curves," *Medical decision making*, vol. 8, no. 3, pp. 197–203, 1988.

[18] B. Rosner and D. Grove, "Use of the mann–whitney u-test for clustered data," *Statistics in medicine*, vol. 18, no. 11, pp. 1387–1400, 1999.

[19] D. D. Dorfman, K. S. Berbaum, C. E. Metz, R. V. Lenth, J. A. Hanley, and H. A. Dagga, "Proper receiver operating characteristic analysis: the bigamma model," *Academic Radiology*, vol. 4, no. 2, pp. 138–149, 1997.

[20] N. Gürtler and N. Henze, "Goodness-of-fit tests for the cauchy distribution based on the empirical characteristic function," *Annals of the Institute of Statistical Mathematics*, vol. 52, no. 2, pp. 267–286, 2000.

Appendices

A Bi-Cauchy ROC curves

```
# -*- coding: utf-8 -*-
"""Bi-Cauchy Curves

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1Y774PEYLf9xxQtX7-AibYv9wWXxeJjAo
"""

#import packages
import math
import numpy as np
import matplotlib.pyplot as plt
#Bi-Cauchy ROC curves
pi = math.pi
tan = np.tan
arctan = np.arctan
plt.title("Bi-Cauchy ROC Curves")
x = np.linspace(0.00001, 1, num=200)
b = 1
colors = ['purple','brown','pink', 'olive','cyan','red','green','indigo','darkblue','darkcyan']
col_count = 0
seq = [0.5,1,2,3]
h = []
for i in seq:
  a = i
  y= (1/2) + (1/pi)*arctan(a + b*tan(pi*x +  (pi/2)))
  globals()[f"h_{col_count}"], = plt.plot(x,y,color = colors[col_count], label="Alpha = "+str(i) )
  h.append(globals()[f"h_{col_count}"],)
  col_count = col_count + 1
leg = plt.legend(handles=h, loc='lower right')
ax = plt.gca().add_artist(leg)

fig = plt.gcf()
plt.grid()
plt.show()

import math
import numpy as np
import matplotlib.pyplot as plt
plt.title("Bi-Cauchy ROC Curves")
pi = math.pi
tan = np.tan
arctan = np.arctan
x = np.linspace(0.00001, 1, num=100)
a = 1
colors = ['purple','brown','pink', 'olive','cyan','red','green','indigo','darkblue','darkcyan']
```

```
col_count = 0
seq = [0.5,1,2,3]
h = []
for i in seq:
  b = i
  y= (1/2) + (1/pi)*arctan(a + b*tan(pi*x +  (pi/2)))
  globals()[f"h_{col_count}"], = plt.plot(x,y,color = colors[col_count],
    label="\u03B2 = "+str(i) )
  h.append(globals()[f"h_{col_count}"],)
  col_count = col_count + 1
leg = plt.legend(handles=h, loc='lower right')
ax = plt.gca().add_artist(leg)
fig = plt.gcf()
plt.grid()
plt.show()

import math
import numpy as np
import matplotlib.pyplot as plt
plt.title("Bi-Cauchy ROC Curves")
pi = math.pi
tan = np.tan
arctan = np.arctan
x = np.linspace(0.00001, 1, num=100)
b = [0.5,1.5,2.5,3.5]
a = [1,2,3,4]
colors = ['purple','brown','pink', 'olive','cyan','red','green','indigo','
    darkblue','darkcyan']
col_count = 0
h = []
for i in range(len(a)):
  y= (1/2) + (1/pi)*arctan(a[i] + b[i]*tan(pi*x +  (pi/2)))
  globals()[f"h_{col_count}"], = plt.plot(x,y,color = colors[col_count],
    label="\u03B1 = "+str(a[i])+", \u03B2 = " + str(b[i]))
  h.append(globals()[f"h_{col_count}"],)
  col_count = col_count + 1
leg = plt.legend(handles=h, loc='lower right')
ax = plt.gca().add_artist(leg)

fig = plt.gcf()
plt.grid()
plt.show()
```

B Simulation Results 2

```
# -*- coding: utf-8 -*-
"""RESULTS1

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1
    gVKpiAz8cALqJot0vZLFX7yJ6gSwUfY0
"""

#import PACKAGES
import numpy as np
```

```
import pandas as pd
import pandas.util.testing as tm
from statsmodels.robust.scale import huber
import array as arr
from scipy.stats import cauchy
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve,auc
from scipy.stats import wilcoxon
from scipy import stats
#import statistics as stats

#Defining terms for the empirical ROC
def plot_roc(y_test,probs):
    fpr,tpr,threshold=roc_curve(y_test,probs)
    roc_auc=auc(fpr,tpr)
    print('ROC AUC=%0.2f'%roc_auc)
    plt.plot(fpr,tpr,label='AUC=%0.2f'%roc_auc,color='darkorange')
    plt.legend(loc='lower right')
    plt.plot([0,1],[0,1],'b—')
    plt.xlim([0,1])
    plt.ylim([0,1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.show()

#packages to read data and split
#Logistic regression is used for binary classification problem
#It help split the data into training and testing
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import matplotlib.pyplot as plt

#Read data for all the estimates
data = pd.read_csv('Together2.csv')
h = list(data.columns.values)
h[:-1]

X = data[h[:-1]]
y = data['Outcome']

#split the dataset into training (70%) and testing (30%) sets
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,
    random_state=0)

#instantiate the model
log_regression = LogisticRegression()

#fit the model using the training data
log_regression.fit(X_train,y_train)

#define metrics
y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)

```

```
#create ROC curve
plt.plot(fpr,tpr)
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.grid()
plt.show()

#define metrics
import math
y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

#Bi-Cauchy constants after subtracting the not-diseased from the diseased
c1 = 0.6874
c2 = 0.6225
c3 = 1.0818
c4 = 0.696
c5 = 0.52305
c6 = 0.8434
c7 = 0.8174
c8 = 1.0818000000000012
estimates = ['Huber-M ROC', 'Hodges ROC', 'Sens ROC', '5% Trimmed ROC', '
    10% Trimmed ROC', '5% Winsorized ROC', '10% Winsorized ROC','Mean']
const = [c1,c2,c3,c4,c5,c6,c7,c8]
colors = ['purple','brown','pink', 'olive','cyan','red','green','darkblue',
    'darkcyan','darkcyan']
col_count = 0
t = np.linspace(0,1,20)
ct = []
ct2 = []
x = np.pi*t - np.pi/2
for i in const:
  y = i + np.tan(x)
  z = 1/2 + 1/np.pi*np.arctan(y)

  auc2 = 1/2 + 1/np.pi*np.arctan(i/math.sqrt(2))

  globals()[f"h1_{col_count}"], = plt.plot(t,z,color = colors[col_count],
    label="AUC="+str(auc2))
  globals()[f"h2_{col_count}"], = plt.plot(t,z,color = colors[col_count],
    label=estimates[col_count])
  ct.append(globals()[f"h1_{col_count}"],)
  ct2.append(globals()[f"h2_{col_count}"],)
  col_count = col_count + 1




#create ROC curve
h32, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
h3, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
plt.title("ROC plots")
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
```

```
#plt.plot(tpr, np.tan(tpr))

ct.append(h32)
ct2.append(h3)
leg = plt.legend(handles=ct, loc='lower right')
ax = plt.gca().add_artist(leg)
leg = plt.legend(handles=ct2, loc=3, bbox_to_anchor=(0.7,-0.3))
fig = plt.gcf()
fig.set_size_inches(12.5, 10.5)
fig.savefig('test2png.png', dpi=100)
plt.grid()
plt.show()
```

C Simulation Results 1

```
# -*- coding: utf-8 -*-
"""RESULT2

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1wihcrpGAczJ_6hliI3aV-D7eY-MN88Rc
"""

#import all packages
import numpy as np
import pandas as pd
import pandas.util.testing as tm
from statsmodels.robust.scale import huber
import array as arr
from scipy.stats import cauchy
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve,auc
from scipy.stats import wilcoxon
from scipy import stats

from scipy.stats.mstats import winsorize

# Commented out IPython magic to ensure Python compatibility.
#This is a very important package for running Sens Mean
# %load_ext rpy2.ipython

#Random Cauchy variables for diseased population
rdiseased = cauchy.rvs(loc = 28, scale =1, size=100)
rdiseased

#Random Cauchy variables for notdiseased population
rnotdiseased = cauchy.rvs(loc = 26.5, scale =1, size=100)
rnotdiseased

#This reads the data file that will be upladed
df = pd.read_csv('Result2.csv')

#Defining metric for Empirical ROC
def plot_roc(y_test,probs):
    fpr,tpr,threshold=roc_curve(y_test,probs)
```

```
    roc_auc=auc(fpr,tpr)
    print('ROC AUC=%0.2f'%roc_auc)
    plt.plot(fpr,tpr,label='AUC=%0.2f'%roc_auc,color='darkorange')
    plt.legend(loc='lower right')
    plt.plot([0,1],[0,1],'b—')
    plt.xlim([0,1])
    plt.ylim([0,1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.show()

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import matplotlib.pyplot as plt

data= df
#define the predictor variables and the response variable
X = data[['Diseased', 'Not Diseased']]
y = data['Outcome']

#split the dataset into training (70%) and testing (30%) sets
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,
    random_state=0)

#instantiate the model
log_regression = LogisticRegression()

#fit the model using the training data
log_regression.fit(X_train,y_train)

#define metrics
y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)

#create ROC curve
plt.plot(fpr,tpr)
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.grid()
plt.show()

"""**HUBER ESTMATE**"""

#HUBER ESTIMATE COMPUTATION
import math
Huberdisease = huber(df['Diseased'])
Hubernotdisease = huber(df['Not Diseased'])
print(Huberdisease)

diff1 = Huberdisease[0] - Hubernotdisease[0]
print(diff1)

#define metrics

```

```
y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

const = diff1;
t = np.linspace(0,1,1000)
x = np.pi*t - np.pi/2
y = const + np.tan(x)
z = 1/2 + 1/np.pi*np.arctan(y)
#auc2 = np.trapz(z,t)
auc2 = 1/2 + 1/np.pi*np.arctan(const/math.sqrt(2))


y2 =  (28-26.5) + np.tan(x)
const2 = 1.5
z2 =  1/2 + 1/np.pi*np.arctan(y2)
#auc3 = np.trapz(z2,t)
auc3 = 1/2 + 1/np.pi*np.arctan(const2/math.sqrt(2))

#create ROC curve
h22, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
h2, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
plt.title("HUBER-M ESTIMATOR ROC plots")
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
#plt.plot(tpr, np.tan(tpr))
h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
h1, = plt.plot(t,z,color = 'darkorange', label="Theoretical ROC")
h33, = plt.plot(t,z2,color = 'grey',label="AUC="+str(auc3))
h3, = plt.plot(t,z2,color = 'grey', label="True ROC")

leg = plt.legend(handles=[h11,h22,h33], loc='lower right')
ax = plt.gca().add_artist(leg)
leg = plt.legend(handles=[h2,h1,h3], loc=3, bbox_to_anchor=(0.7,-0.3))
plt.grid()
plt.show()

"""**HODGES**

"""

#HODGES COMPUTATION
import statistics as stats
hL = [];
ard = ['Diseased','Not Diseased' ]
for i in range(2):
  l_input = df[ard[i]];
  l_avgs = [];

  i = 0;
  j = 0;

  while i < len(l_input):
    while j < len(l_input):
      l_avgs.append(stats.mean([l_input[i], l_input[j]]))
      j = j + 1
    i = i + 1
```

```
    j = i

  hl_est = stats.median(l_avgs)
  hL.append(hl_est)

diff2 = hL[0] - hL[1]
print(diff2)



#Define Metrics
y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

const = diff2;
t = np.linspace(0,1,1000)
x = np.pi*t - np.pi/2
y = const + np.tan(x)
z = 1/2 + 1/np.pi*np.arctan(y)
#auc2 = np.trapz(z,t)
auc2 = 1/2 + 1/np.pi*np.arctan(const/math.sqrt(2))

y2 =  (28-26.5) + np.tan(x)
const2 = 1.5
z2 =  1/2 + 1/np.pi*np.arctan(y2)
#auc3 = np.trapz(z2,t)
auc3 = 1/2 + 1/np.pi*np.arctan(const2/math.sqrt(2))

#create ROC curve
h22, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
h2, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
plt.title("HODGES LEHMANN ESTIMATOR ROC plots")
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
#plt.plot(tpr, np.tan(tpr))
h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
h1, = plt.plot(t,z,color = 'darkorange', label="Theoretical ROC")
h33, = plt.plot(t,z2,color = 'grey',label="AUC="+str(auc3))
h3, = plt.plot(t,z2,color = 'grey', label="True ROC")

leg = plt.legend(handles=[h11,h22,h33], loc='lower right')
ax = plt.gca().add_artist(leg)
leg = plt.legend(handles=[h2,h1,h3], loc=3, bbox_to_anchor=(0.7,-0.3))
plt.grid()
plt.show()

"""**5% Trimm Mean**"""

#Trimm Mean Computation
from scipy import stats
pernt = 0.05
y = arr.array('d', [ stats.trim_mean(df['Diseased'], pernt),stats.trim_mean
    (df['Not Diseased'], pernt)])
difftrim5 = y[0]-y[1]
print(difftrim5)
diff3 = difftrim5
```

```
211
212
213
214 #Define Metrics
215 y_pred_proba = log_regression.predict_proba(X_test)[::,1]
216 fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
217 auc = metrics.roc_auc_score(y_test, y_pred_proba)
218
219 const = difftrim5;
220 t = np.linspace(0,1,1000)
221 x = np.pi*t - np.pi/2
222 y = const + np.tan(x)
223 z = 1/2 + 1/np.pi*np.arctan(y)
224 #auc2 = np.trapz(z,t)
225 auc2 = 1/2 + 1/np.pi*np.arctan(const/math.sqrt(2))
226
227 y2 =  (28-26.5) + np.tan(x)
228 conts2 = 1.5
229 z2 =  1/2 + 1/np.pi*np.arctan(y2)
230 #auc3 = np.trapz(z2,t)
231 auc3 = 1/2 + 1/np.pi*np.arctan(const2/math.sqrt(2))
232
233 #create ROC curve
234 h22, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
235 h2, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
236 plt.title("5% Trimm Mean ROC plots")
237 plt.ylabel('True Positive Rate')
238 plt.xlabel('False Positive Rate')
239 #plt.plot(tpr, np.tan(tpr))
240 h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
241 h1, = plt.plot(t,z,color = 'darkorange', label="Theoretical ROC")
242 h33, = plt.plot(t,z2,color = 'grey',label="AUC="+str(auc3))
243 h3, = plt.plot(t,z2,color = 'grey', label="True ROC")
244
245 leg = plt.legend(handles=[h11,h22,h33], loc='lower right')
246 ax = plt.gca().add_artist(leg)
247 leg = plt.legend(handles=[h2,h1,h3], loc=3, bbox_to_anchor=(0.7,-0.3))
248 plt.grid()
249 plt.show()
250
251 """**10% Trimm Mean**"""
252
253 #Trimm Mean Computation
254 pernt = 0.10
255 y = arr.array('d', [ stats.trim_mean(df['Diseased'], pernt),stats.trim_mean
      (df['Not Diseased'], pernt)])
256 difftrim10 = y[0]-y[1]
257 print(difftrim10)
258 diff4 = difftrim10
259 print(y[1])
260
261 #Define Metrics
262 y_pred_proba = log_regression.predict_proba(X_test)[::,1]
263 fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
264 auc = metrics.roc_auc_score(y_test, y_pred_proba)
265
266 const = difftrim10;
```

```
t = np.linspace(0,1,1000)
x = np.pi*t - np.pi/2
y = const + np.tan(x)
z = 1/2 + 1/np.pi*np.arctan(y)
#auc2 = np.trapz(z,t)
auc2 = 1/2 + 1/np.pi*np.arctan(const/math.sqrt(2))


y2 =  (28-26.5) + np.tan(x)
const2 = 1.5
z2 =  1/2 + 1/np.pi*np.arctan(y2)
#auc3 = np.trapz(z2,t)
auc3 = 1/2 + 1/np.pi*np.arctan(const2/math.sqrt(2))


#create ROC curve
h22, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
h2, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
plt.title("10% Trimm Mean ROC plots")
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
#plt.plot(tpr, np.tan(tpr))
h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
h1, = plt.plot(t,z,color = 'darkorange', label="Theoretical ROC")
h33, = plt.plot(t,z2,color = 'grey',label="AUC="+str(auc3))
h3, = plt.plot(t,z2,color = 'grey', label="True ROC")

leg = plt.legend(handles=[h11,h22,h33], loc='lower right')
ax = plt.gca().add_artist(leg)
leg = plt.legend(handles=[h2,h1,h3], loc=3, bbox_to_anchor=(0.7,-0.3))
plt.grid()
plt.show()

"""**5% Winsorized Mean**

"""

#Winssorised mean computation
winsorizedarray_adep = winsorize(df['Not Diseased'], limits=[0.05, 0.05])
WinsorizedMean_adep = np.mean(winsorizedarray_adep)
winsorizedarray_obes = winsorize(df['Diseased'], limits=[0.05, 0.05])
WinsorizedMean_obes = np.mean(winsorizedarray_obes)
y1 = arr.array('d', [WinsorizedMean_obes,WinsorizedMean_adep])
Winn5 = y1[0]-y1[1]
print(Winn5)
diff5 = Winn5



#Define Metrics
y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

const = Winn5;
t = np.linspace(0,1,1000)
x = np.pi*t - np.pi/2
```

```
y = const + np.tan(x)
z = 1/2 + 1/np.pi*np.arctan(y)
#auc2 = np.trapz(z,t)
auc2 = 1/2 + 1/np.pi*np.arctan(const/math.sqrt(2))


y2 =  (28-26.5) + np.tan(x)
const2 = 1.5
z2 =  1/2 + 1/np.pi*np.arctan(y2)
#auc3 = np.trapz(z2,t)
auc3 = 1/2 + 1/np.pi*np.arctan(const2/math.sqrt(2))


#create ROC curve
h22, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
h2, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
plt.title("5% Winsorized Mean ROC plots")
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
#plt.plot(tpr, np.tan(tpr))
h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
h1, = plt.plot(t,z,color = 'darkorange', label="Theoretical ROC")
h33, = plt.plot(t,z2,color = 'grey',label="AUC="+str(auc3))
h3, = plt.plot(t,z2,color = 'grey', label="True ROC")

leg = plt.legend(handles=[h11,h22,h33], loc='lower right')
ax = plt.gca().add_artist(leg)
leg = plt.legend(handles=[h2,h1,h3], loc=3, bbox_to_anchor=(0.7,-0.3))
plt.grid()
plt.show()

"""10% Winsorized Mean"""

#Winssorised mean computation
winsorizedarray_adep = winsorize(df['Not Diseased'], limits=[0.1, 0.1])
WinsorizedMean_adep = np.mean(winsorizedarray_adep)
winsorizedarray_obes = winsorize(df['Diseased'], limits=[0.1, 0.1])
WinsorizedMean_obes = np.mean(winsorizedarray_obes)
y1 = arr.array('d', [WinsorizedMean_obes,WinsorizedMean_adep])
Winn10 = y1[0]-y1[1]
print(Winn10)
diff6 = Winn10
print(y1[1])


#Define Metrics
y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

const = Winn10;
t = np.linspace(0,1,1000)
x = np.pi*t - np.pi/2
y = const + np.tan(x)
z = 1/2 + 1/np.pi*np.arctan(y)
#auc2 = np.trapz(z,t)
auc2 = 1/2 + 1/np.pi*np.arctan(const/math.sqrt(2))
```

```
381
382
383 y2 =  (28-26.5) + np.tan(x)
384 const2 = 1.5
385 z2 =  1/2 + 1/np.pi*np.arctan(y2)
386 #auc3 = np.trapz(z2,t)
387 auc3 = 1/2 + 1/np.pi*np.arctan(const2/math.sqrt(2))
388
389
390 #create ROC curve
391 h22, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
392 h2, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
393 plt.title("10%  Winsorized Mean ROC plots")
394 plt.ylabel('True Positive Rate')
395 plt.xlabel('False Positive Rate')
396 #plt.plot(tpr, np.tan(tpr))
397 h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
398 h1, = plt.plot(t,z,color = 'darkorange', label="Theoretical ROC")
399 h33, = plt.plot(t,z2,color = 'grey',label="AUC="+str(auc3))
400 h3, = plt.plot(t,z2,color = 'grey', label="True ROC")
401
402 leg = plt.legend(handles=[h11,h22,h33], loc='lower right')
403 ax = plt.gca().add_artist(leg)
404 leg = plt.legend(handles=[h2,h1,h3], loc=3, bbox_to_anchor=(0.7,-0.3))
405 plt.grid()
406 plt.show()
407
408 """**SENS WEIGHTED MEAN**"""
409
410 # Commented out IPython magic to ensure Python compatibility.
411 # %%R
412 # install.packages("lmomco")
413 # library("lmomco")
414 #
415 # df <- read.csv(file = 'Both.csv')
416 #
417 # sen1 <- sen.mean(df[1]);
418 # sen2 <- sen.mean(df[2]);
419 #
420 # #   list1 <- append(list1 , sen)
421 # x <- c(sen1,sen2)
422 #
423 # print(x)
424 #
425 #
426
427 diffsen = 27.87913- 26.80403
428 print(diffsen)
429 diff7 = diffsen
430 #Define Metrics
431 y_pred_proba = log_regression.predict_proba(X_test)[::,1]
432 fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
433 auc = metrics.roc_auc_score(y_test, y_pred_proba)
434
435 const = diffsen;
436 t = np.linspace(0,1,1000)
437 x = np.pi*t - np.pi/2
```

```
y = const + np.tan(x)
z = 1/2 + 1/np.pi*np.arctan(y)
#auc2 = np.trapz(z,t)
auc2 = 1/2 + 1/np.pi*np.arctan(const/math.sqrt(2))

y2 =  (28-26.5) + np.tan(x)
const2 = 1.5
z2 =  1/2 + 1/np.pi*np.arctan(y2)
#auc3 = np.trapz(z2,t)
auc3 = 1/2 + 1/np.pi*np.arctan(const2/math.sqrt(2))

#create ROC curve
h22, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
h2, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
plt.title("SENS WEIGHTED MEAN ROC plots")
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
#plt.plot(tpr, np.tan(tpr))
h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
h1, = plt.plot(t,z,color = 'darkorange', label="Theoretical ROC")
h33, = plt.plot(t,z2,color = 'grey',label="AUC="+str(auc3))
h3, = plt.plot(t,z2,color = 'grey', label="True ROC")

leg = plt.legend(handles=[h11,h22,h33], loc='lower right')
ax = plt.gca().add_artist(leg)
leg = plt.legend(handles=[h2,h1,h3], loc=3, bbox_to_anchor=(0.7,-0.3))
plt.grid()
plt.show()

"""**MEAN**"""

# Computation of mean
mean_1 = np.mean(df['Not Diseased'])
mean_2 = np.mean(df['Diseased'])
mean = [mean_1,mean_2]
mean

diffmean = mean[1]-mean[0]
print(diffmean)
diff8 = diffmean
print( mean)

#Define Metrics
y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

const = diffmean;
t = np.linspace(0,1,1000)
x = np.pi*t - np.pi/2
y = const + np.tan(x)
z = 1/2 + 1/np.pi*np.arctan(y)
#auc2 = np.trapz(z,t)
auc2 = 1/2 + 1/np.pi*np.arctan(const/math.sqrt(2))

y2 =  (28-26.5) + np.tan(x)
const2 = 1.5
```

```
z2 =  1/2 + 1/np.pi*np.arctan(y2)
#auc3 = np.trapz(z2,t)
auc3 = 1/2 + 1/np.pi*np.arctan(const2/math.sqrt(2))

#create ROC curve
h22, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
h2, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
plt.title(" MEAN ROC plots")
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
#plt.plot(tpr, np.tan(tpr))
h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
h1, = plt.plot(t,z,color = 'darkorange', label="Theoretical ROC")
h33, = plt.plot(t,z2,color = 'grey',label="AUC="+str(auc3))
h3, = plt.plot(t,z2,color = 'grey', label="True ROC")

leg = plt.legend(handles=[h11,h22,h33], loc='lower right')
ax = plt.gca().add_artist(leg)
leg = plt.legend(handles=[h2,h1,h3], loc=3, bbox_to_anchor=(0.7,-0.3))
plt.grid()
plt.show()

"""**PLOTS OF ALL ESTIMATORS**"""

const = [diff1,diff2,diff3,diff4,diff5,diff6,diff7,diff8,const2]
estimates = ['Huber-M ROC', 'Hodges ROC', '5% Trimmed ROC', '10% Trimmed
    ROC','5% Winsorized ROC','10% Winsorized ROC', 'Sens ROC', 'Mean',"True"
    ]
colors = ['purple','brown','pink', 'olive','cyan','red','green','darkblue',
    'grey','darkcyan']
col_count = 0
t = np.linspace(0,1,1000)
ct = []
ct2 = []

x = np.pi*t - np.pi/2
for i in const:
  y = i + np.tan(x)
  z = 1/2 + 1/np.pi*np.arctan(y)
  auc2 = 1/2 + 1/np.pi*np.arctan(i/math.sqrt(2))
  globals()[f"h1_{col_count}"], = plt.plot(t,z,color = colors[col_count],
    label="AUC="+str(auc2))
  globals()[f"h2_{col_count}"], = plt.plot(t,z,color = colors[col_count],
    label=estimates[col_count])
  ct.append(globals()[f"h1_{col_count}"],)
  ct2.append(globals()[f"h2_{col_count}"],)
  col_count = col_count + 1




#h11, = plt.plot(t,z,color = 'darkorange',label="AUC="+str(auc2))
#h1, = plt.plot(t,z,color = 'darkorange', label="Bi-Cauchy ROC")


#create ROC curve
h32, = plt.plot(fpr,tpr,color = 'steelblue', label="AUC="+str(auc))
```

```
547 h3, =  plt.plot(fpr,tpr,color = 'steelblue', label="Empircal ROC")
548 plt.title("ROC plots")
549 plt.ylabel('True Positive Rate')
550 plt.xlabel('False Positive Rate')
551 #plt.plot(tpr, np.tan(tpr))
552
553 ct.append(h32)
554 ct2.append(h3)
555 leg = plt.legend(handles=ct, loc='lower right')
556 ax = plt.gca().add_artist(leg)
557 leg = plt.legend(handles=ct2, loc=3, bbox_to_anchor=(0.7,-0.3))
558 fig = plt.gcf()
559 fig.set_size_inches(12.5, 10.5)
560 fig.savefig('test2png.png', dpi=100)
561 plt.grid()
562 plt.show()
```