Two Sample Location Problem: Mann-Whitney Test

Kaulik Poddar Saheli Datta Tiyasa Dutta

2024-01-05

INTRODUCTION:

- Let's get introduced to Non-Parametric testing through an example.
- Suppose we have n students of class 10 from school A and m students from school B. Suppose we want to test whether the average marks obtained by the students of School A is less than that of School B.
- We can formulate this as a testing problem , and the test we can perform here is Mann-Whitney U Test, proposed by Henry Mann and his student Donald Whitney around 1947.
- This is also known as Wilcoxon Rank Sum Test.

ASSUMPTION:

- Consider two population with distribution function F(x) and G(x), where G(x)=F(x-θ).
- We assume that F and G are continuous.
- We draw two sets of random sample (x_1,x_2,.....x_n) from F and (y_1,y_2,.....y_m) from G independently.

OBJECTIVE:

- We have two independent populations from F and G respectively where we assume F and G to be continuous and G(x)=F(x-θ).
- Through our presentation we want to show when the distributions are completely unknown how to test the whether two samples are likely to derive from the same population.
- The first step is to check whether the test statistic is distribution free under H_0 and we are also interested in finding the asymptotic distribution of the test statistic under H_0 and H_1.
- We are also interested in checking what happens if the distribution is not continuous , that is if it is discrete whether the test statistic is distribution free under H_0 etc.
- We will estimate size and power from different distributions and make a power comparison.
- The parametric counterpart is also to be explored using the parametric test for location parameter.
- Apart from all these we will try to throw some light into the estimation part of the location parameter, what happens if outliers are present in the data, comparison of power curve for parametric and non parametric set up for different distributions.

Set up:

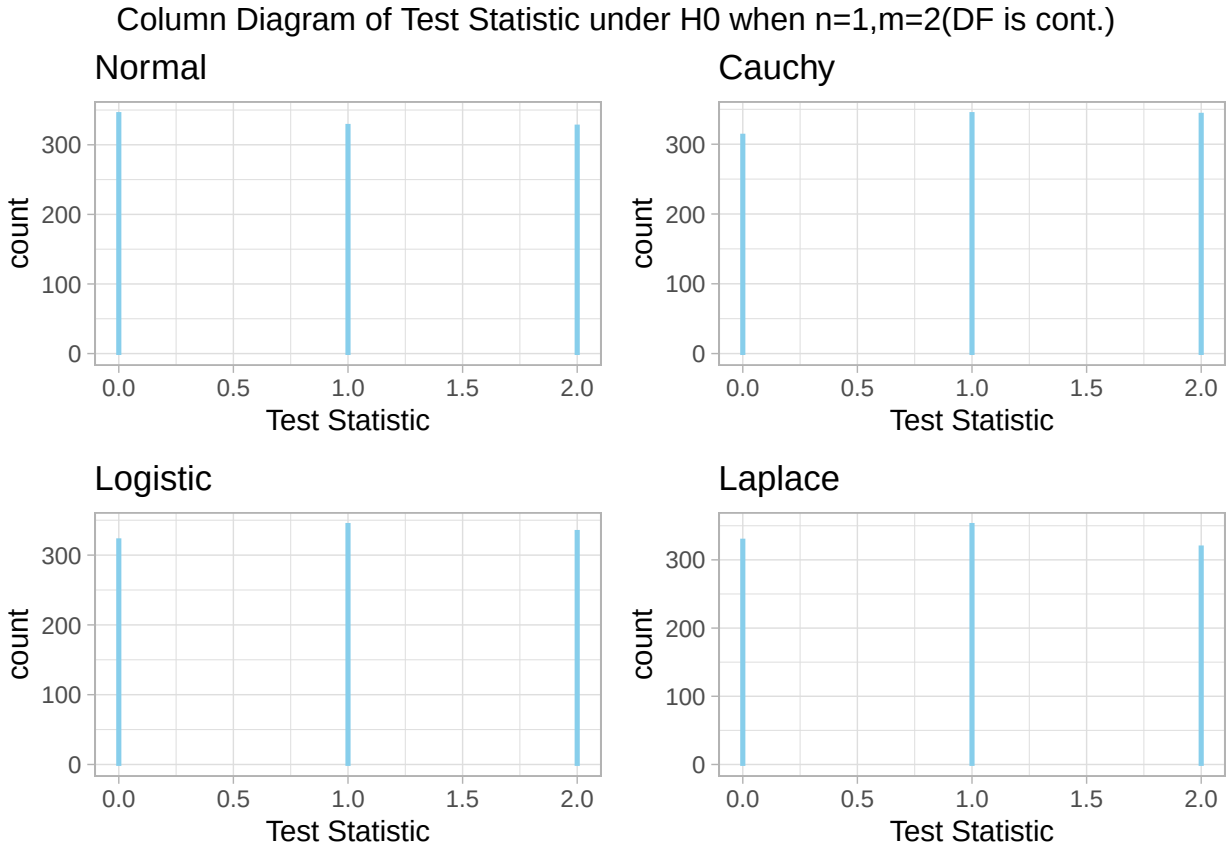
Want to test:

- We are interested in testing H_0: $\theta = 0$ vs H_1: $\theta > 0$.
- (Similary we can do the test for the alternatives H_2: $\theta < 0$ and H_3: $\theta \neq 0$).

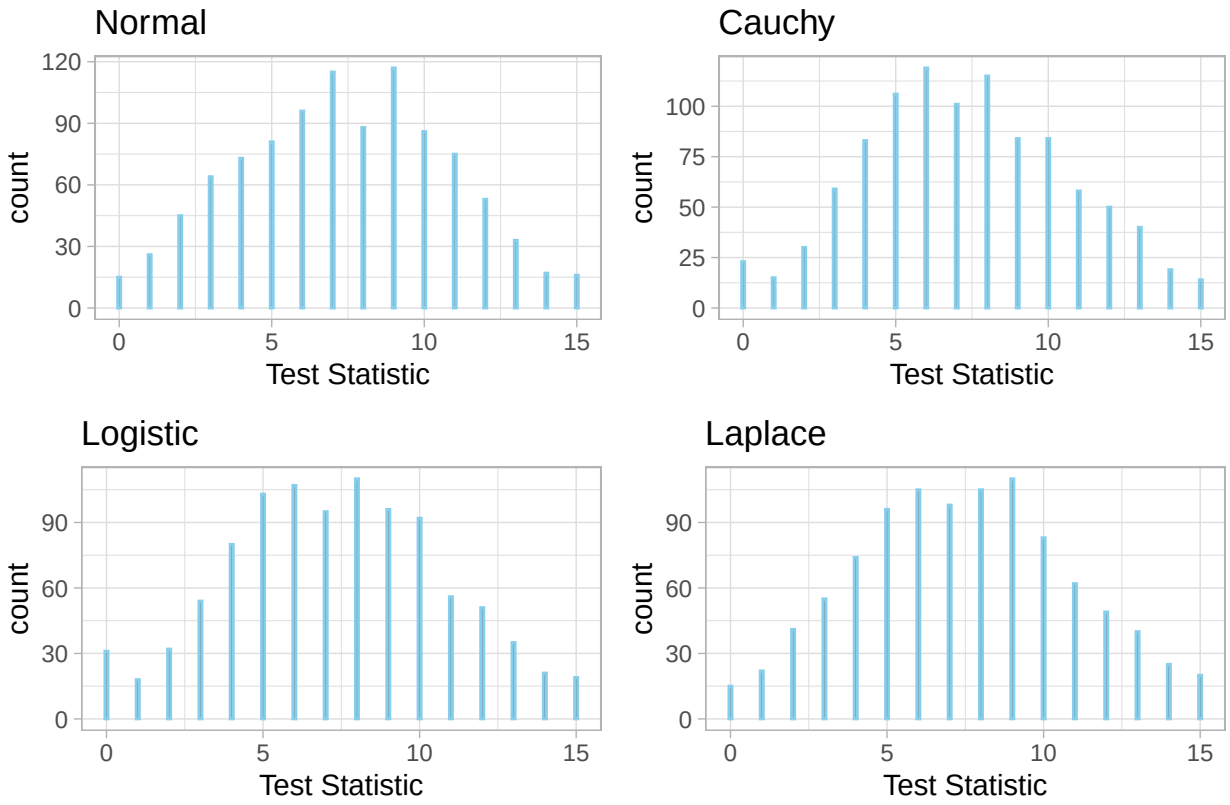
Defining Test Statistic:

- First we define, $\phi(X_i,Y_j)= I(X_i>Y_j)$
- $U = \sum_{i=1}^{n} \sum_{j=1}^{m} \phi(X_i,Y_j)$
- We will see later that U is distribution free under H_0.
- Under H_0, $E[U]=\frac{mn}{2}$ and $Var(U)=\frac{mn(m+n+1)}{12}$
- We will reject H_0 in favour of H_1 for small values of U.

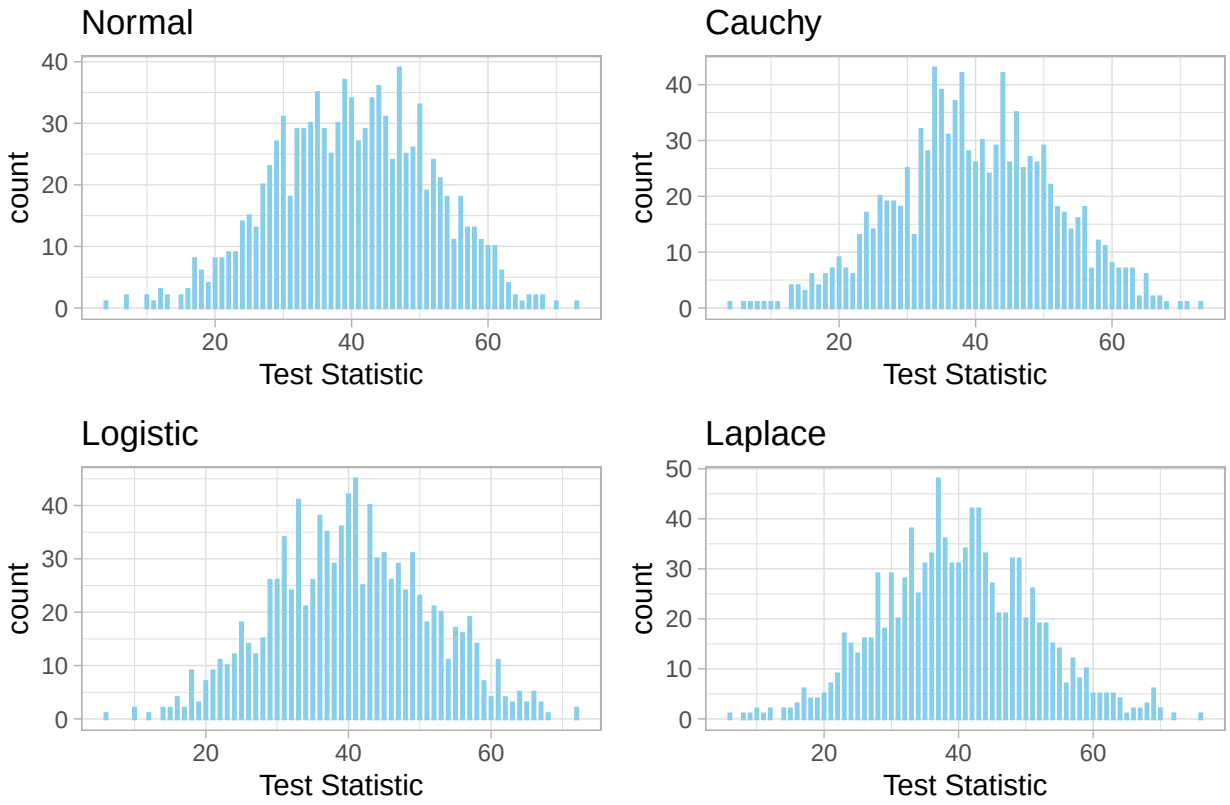
Checking for Distribution free Of Mann Whitney Statistic Under H_0 when the Distribution Function is Continuous:

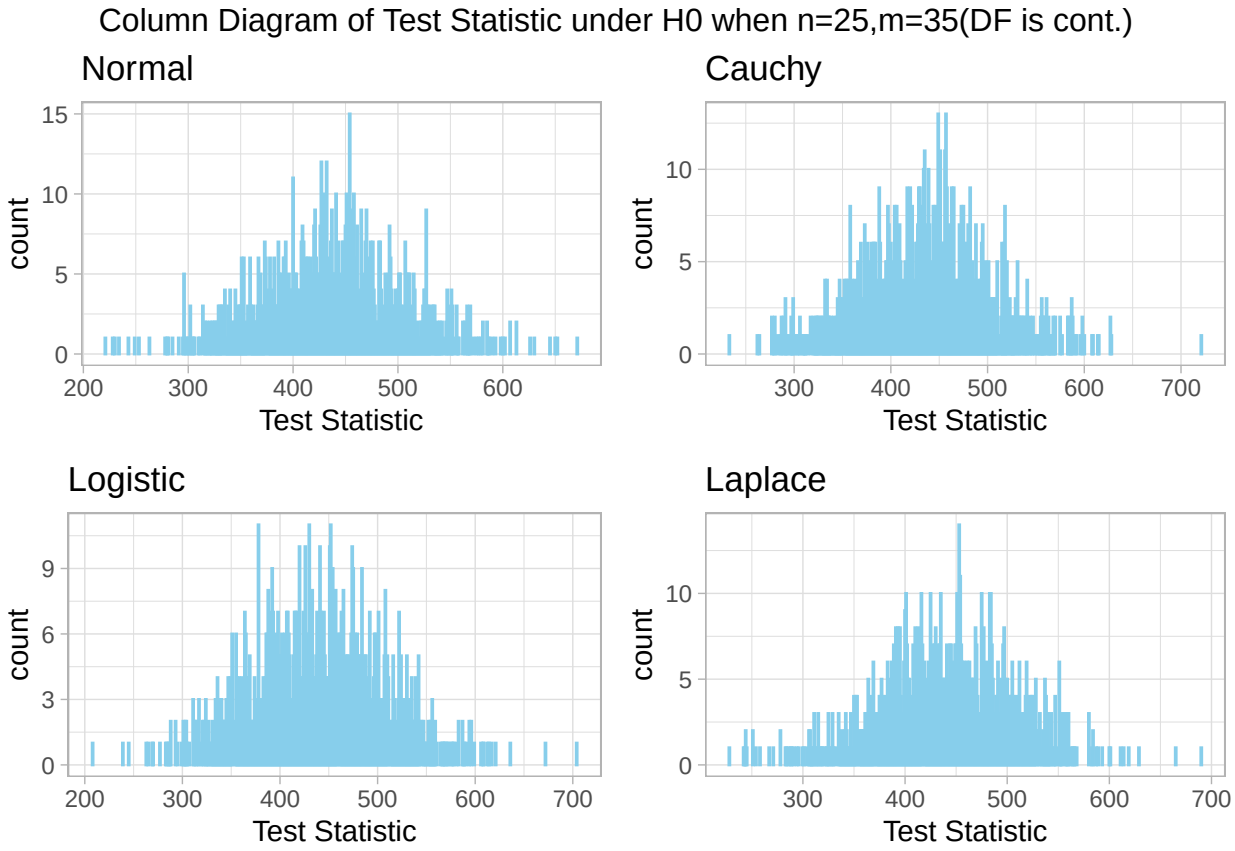


Column Diagram of Test Statistic under H0 when n=3,m=5(DF is cont.)



Column Diagram of Test Statistic under H0 when n=10,m=8(DF is cont,)



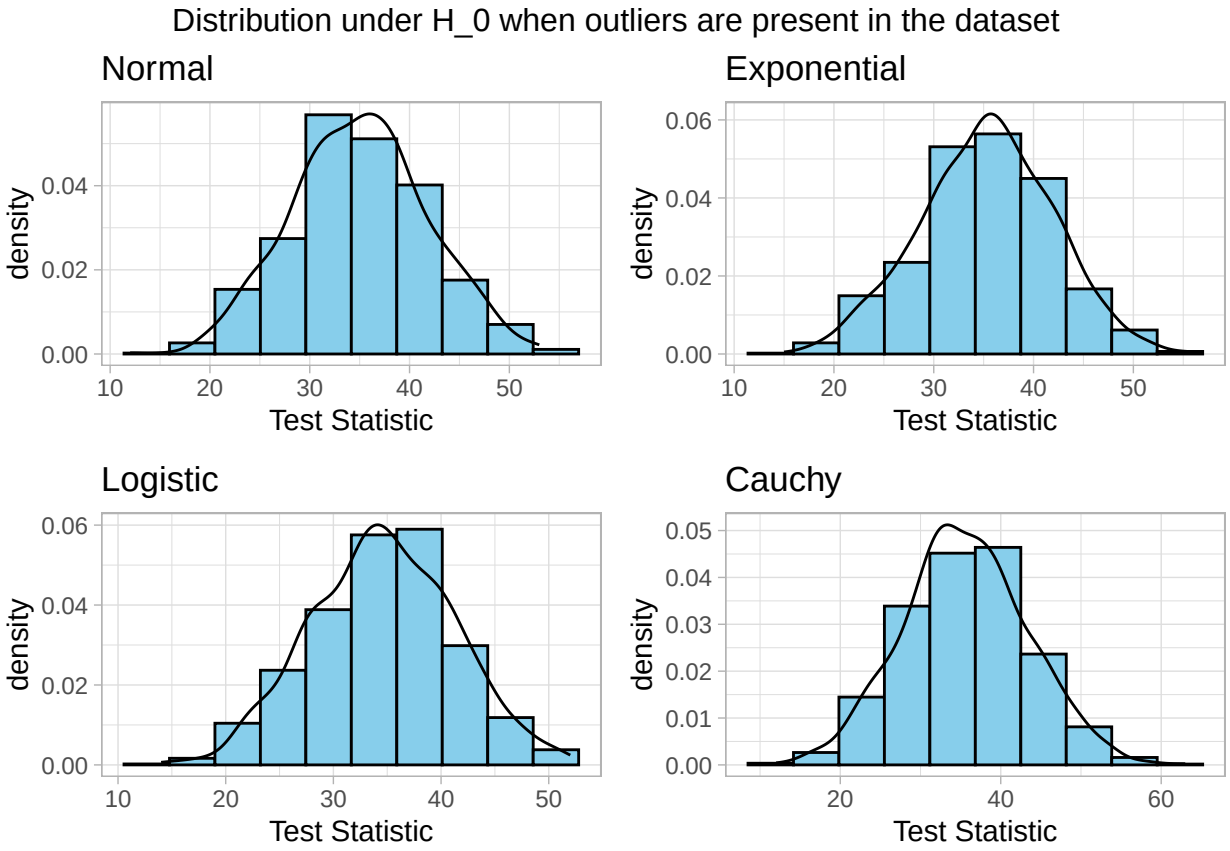


COMMENT:

- In the light of the given data we can conclude that Mann Whitney Statistic is distribution free under H_0 when F,G are continuous.

Robustness of Mann Whitney U Statistics

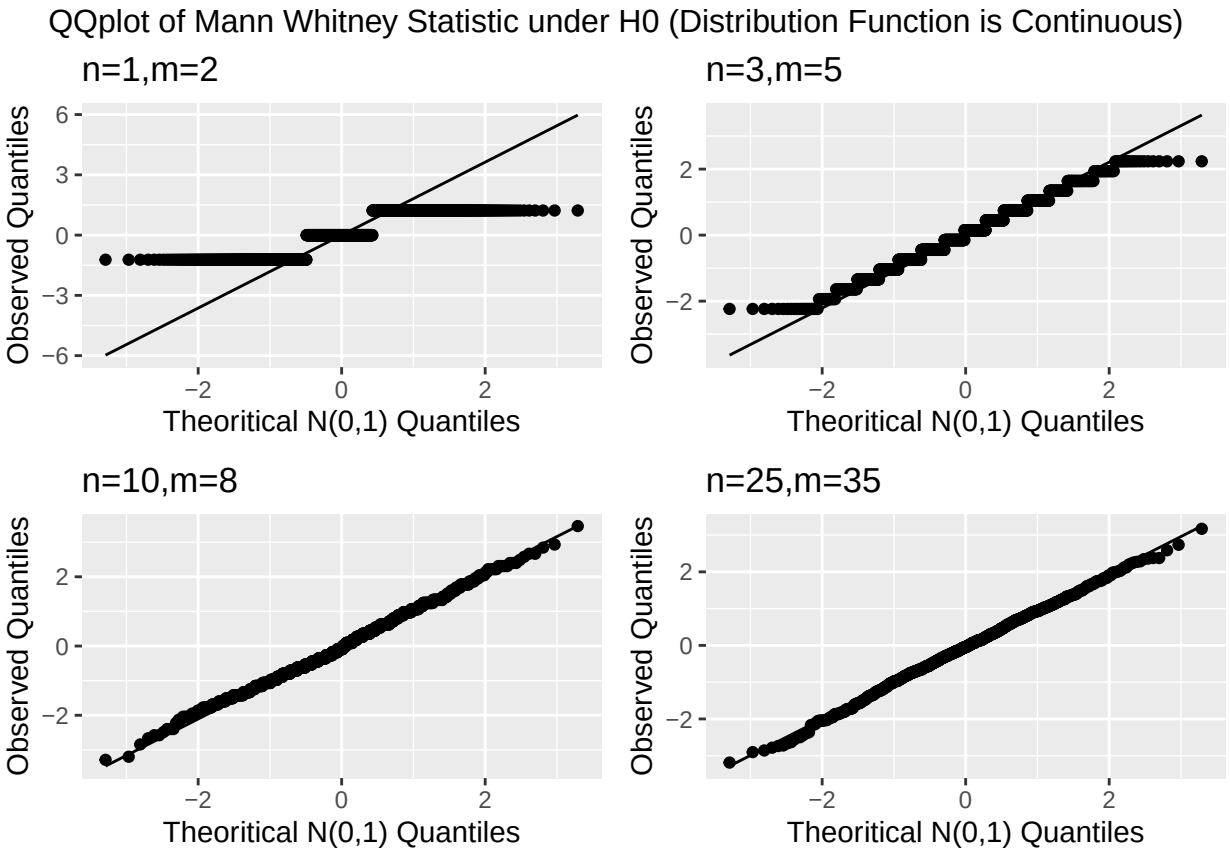
- Here we generate data from Normal(0,1), Exponential(0,1), Logistic(0,1), Cauchy(0,1) with added noise from uniform(10,20) and see how the distribution of U statistics behave when the data contains outliers.



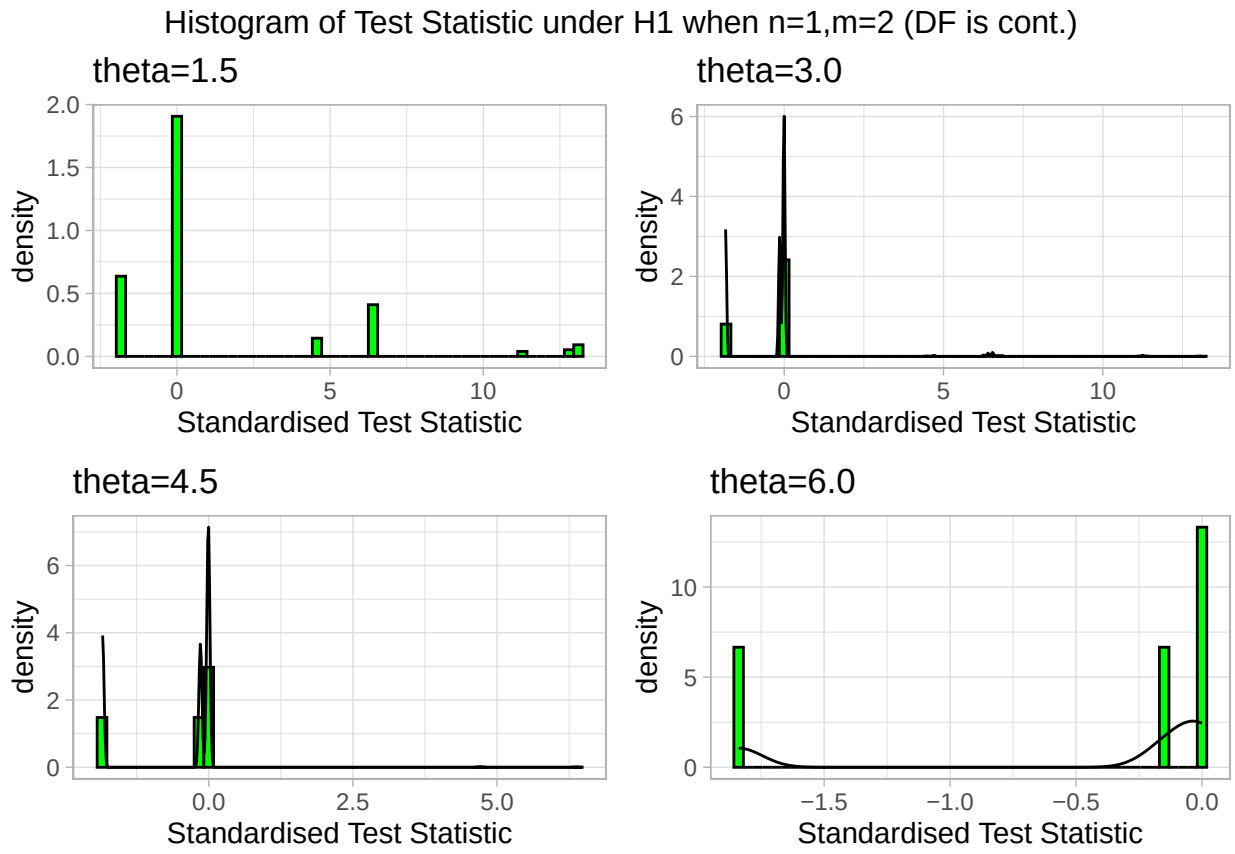
Asymptotic Distribution of Test Statistic when the Distribution Functions are Continuous:

- Here, we want to find the asymptotic distribution of the test statistic under H_0 and H_1.
- Here, we choose F as the CDF of N(0,1) and G as the CDF of N(θ,1).

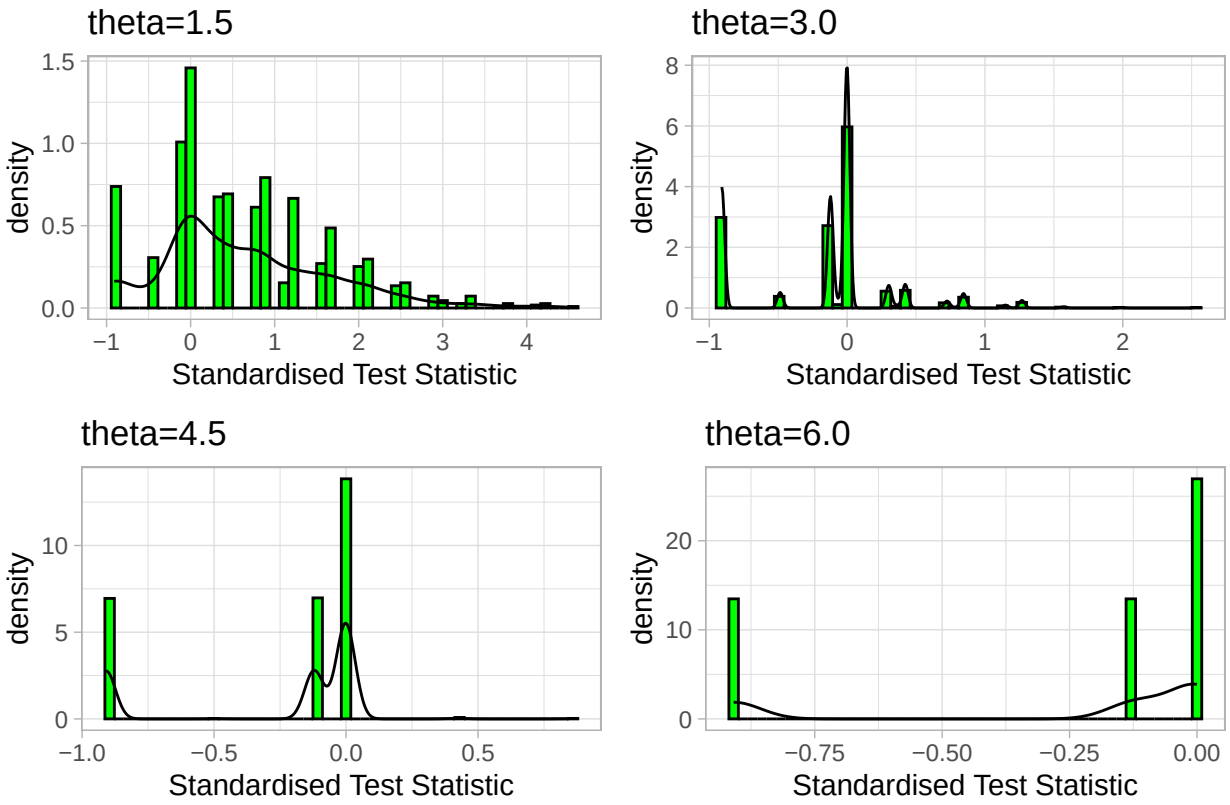
Under H_0:


QQplot of Mann Whitney Statistic under H0 (Distribution Function is Continuous)
n=1,m=2
n=3,m=5
n=10,m=8
n=25,m=35
Observed Quantiles
Theoritical N(0,1) Quantiles

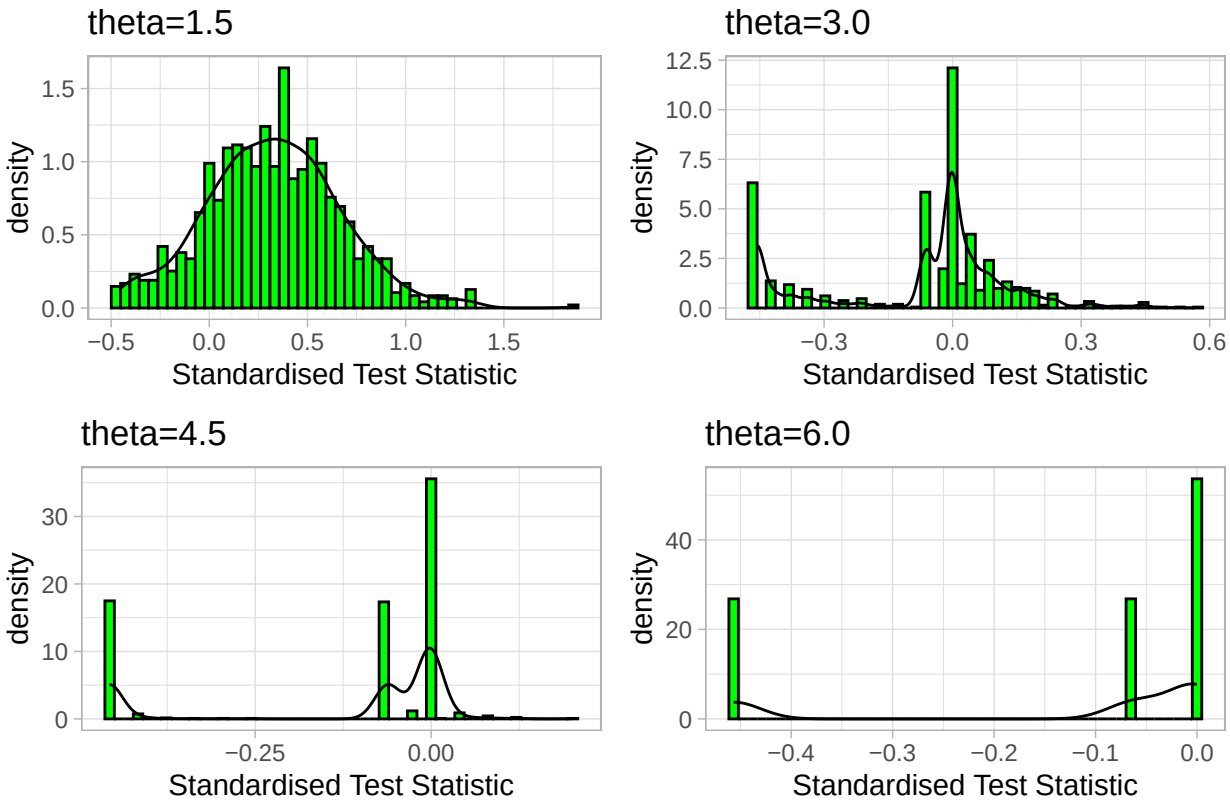
Under H1:

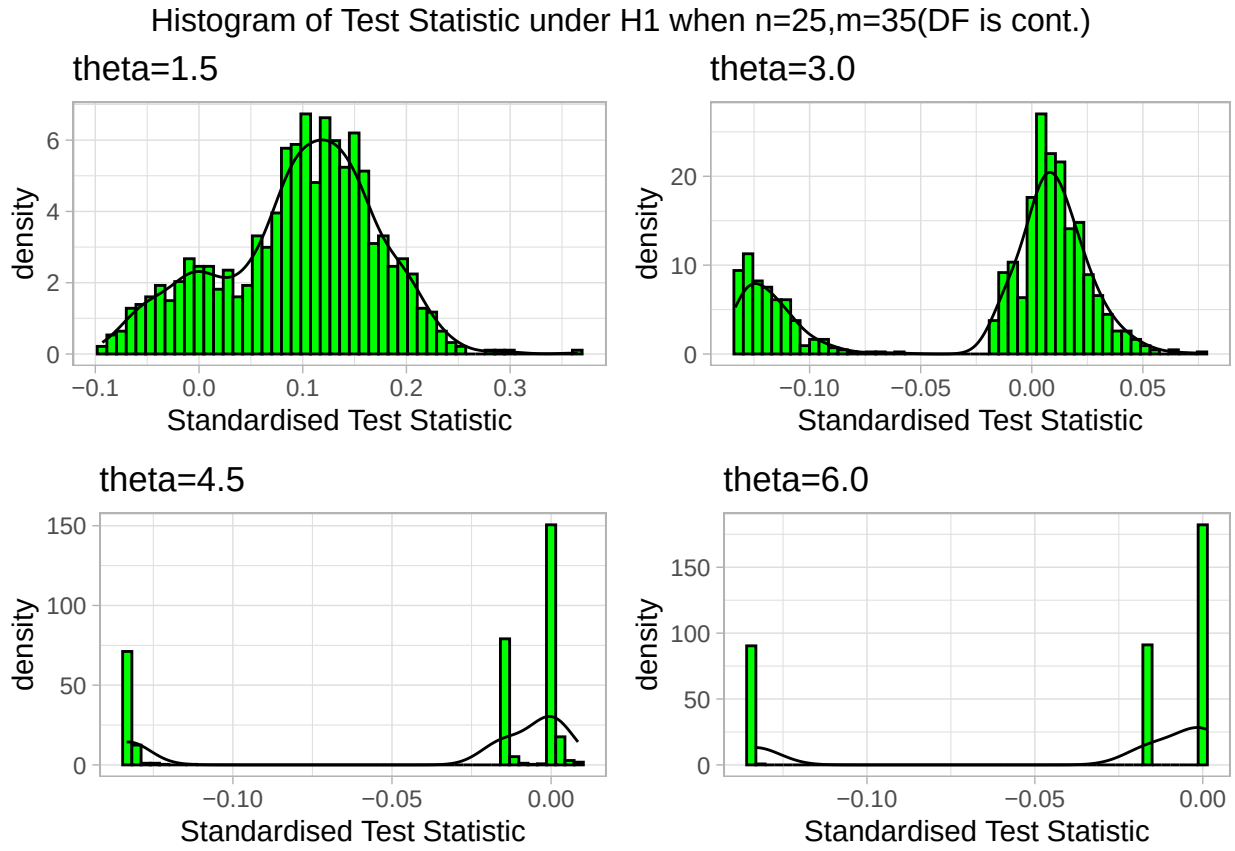

Histogram of Test Statistic under H1 when n=1,m=2 (DF is cont.)
theta=1.5
theta=3.0
theta=4.5
theta=6.0
density
Standardised Test Statistic

Histogram of Test Statistic under H1 when n=3,m=5(DF is cont.)



Histogram of Test Statistic under H1 when n=10,m=8(DF is cont.)

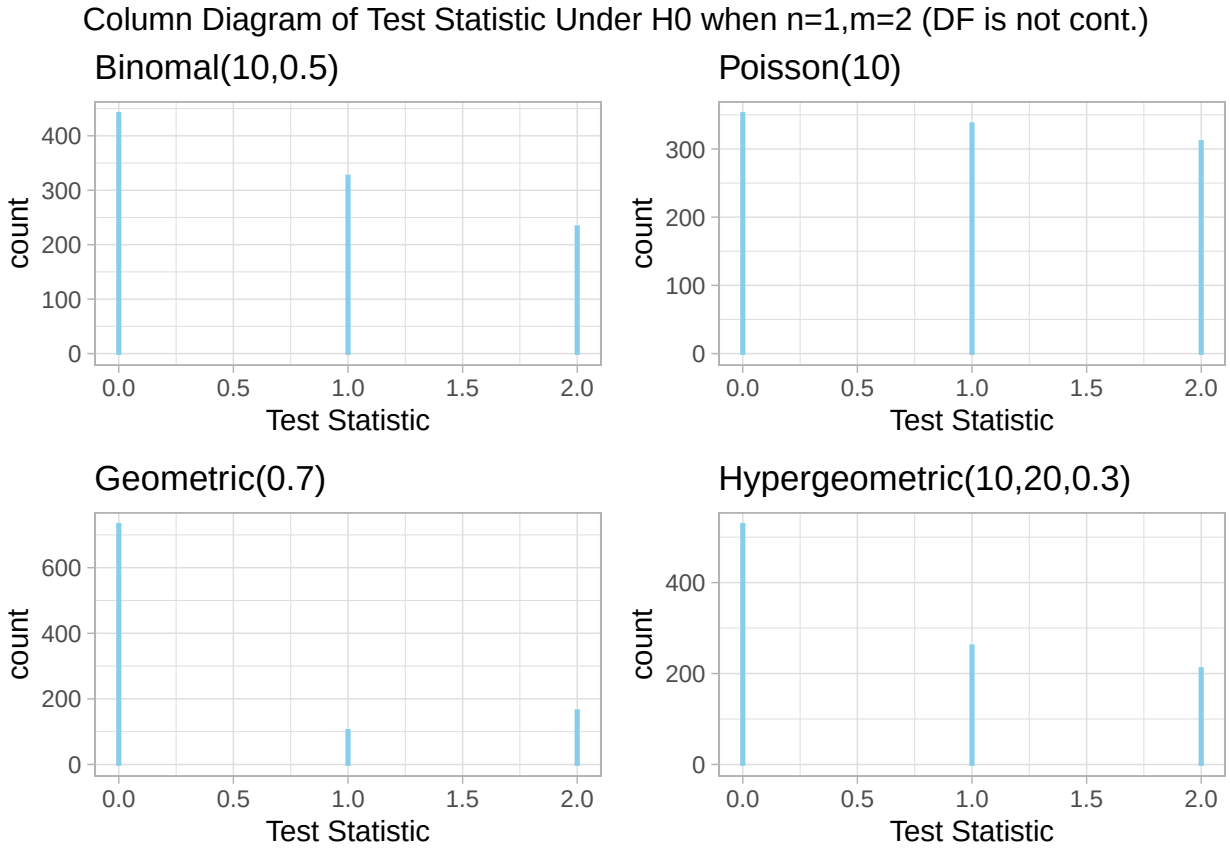




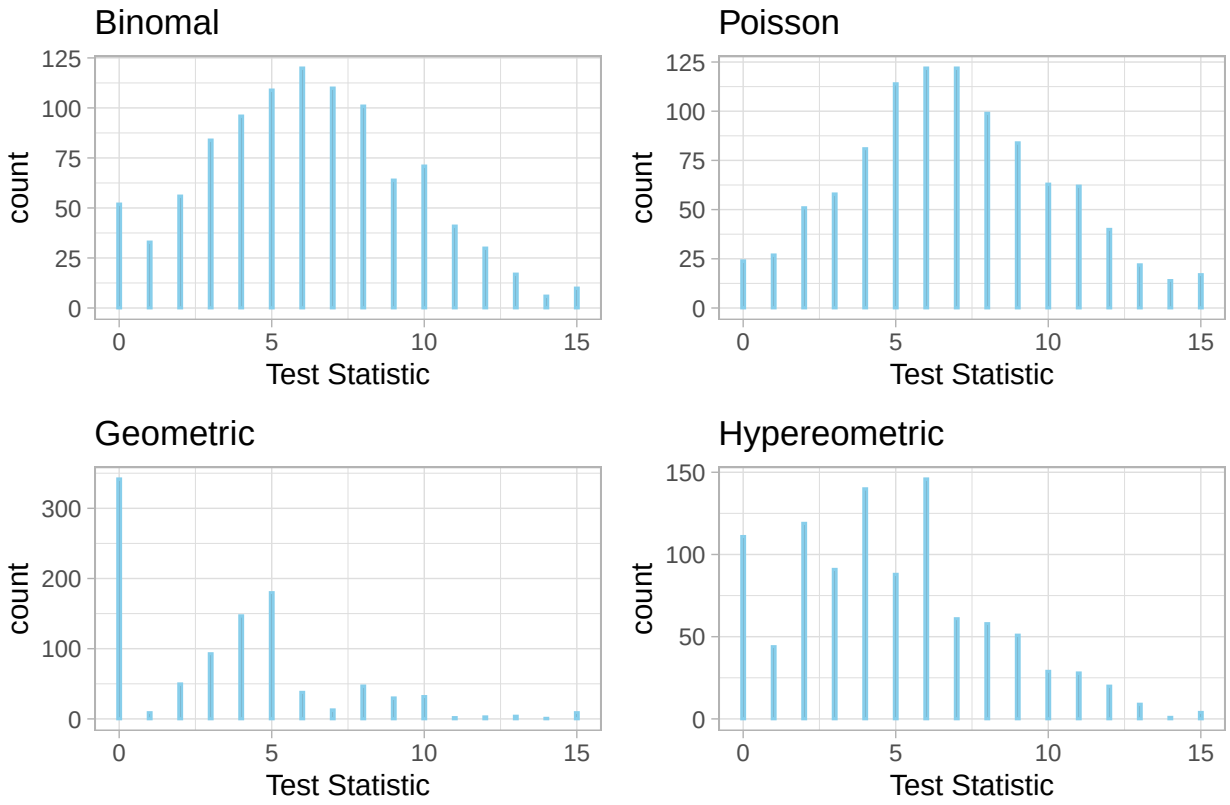
COMMENT:

- Thus when F,G are continuous under H0, we can see that the distribution of standardized Mann Whitney Statistic tends to N(0,1) as the value of (m+n) increases.
- But when F,G are continuous under H1, we can see that there is no specific asymptotic distribution of Mann Whitney Statistic.

Checking for Distribution free Of Mann Whitney Statistic when the Distribution Function is not Continuous:

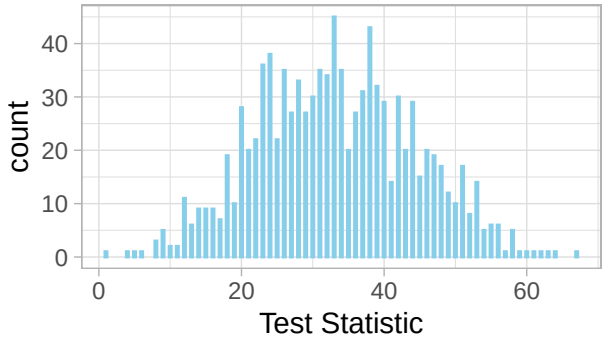


Column Diagram of Test Statistic under H0 when n=3,m=5(DF is not cont.)

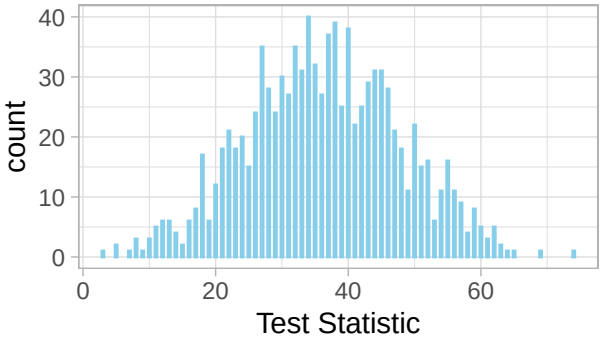


Column Diagram of Test Statistic under H0 when n=10,m=8(DF is not cont.)

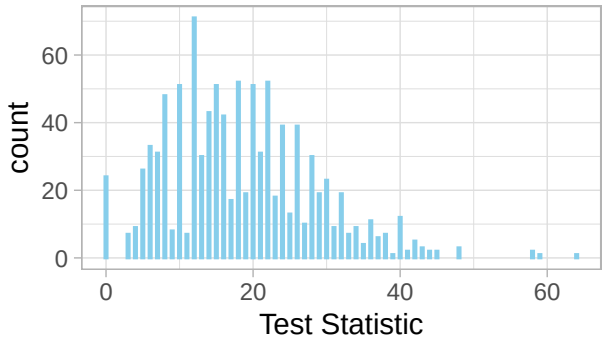
Binomial



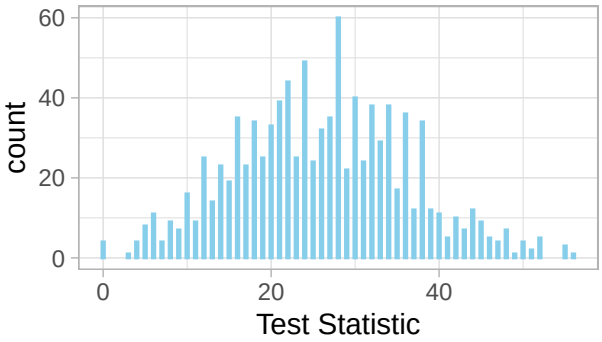
Poisson

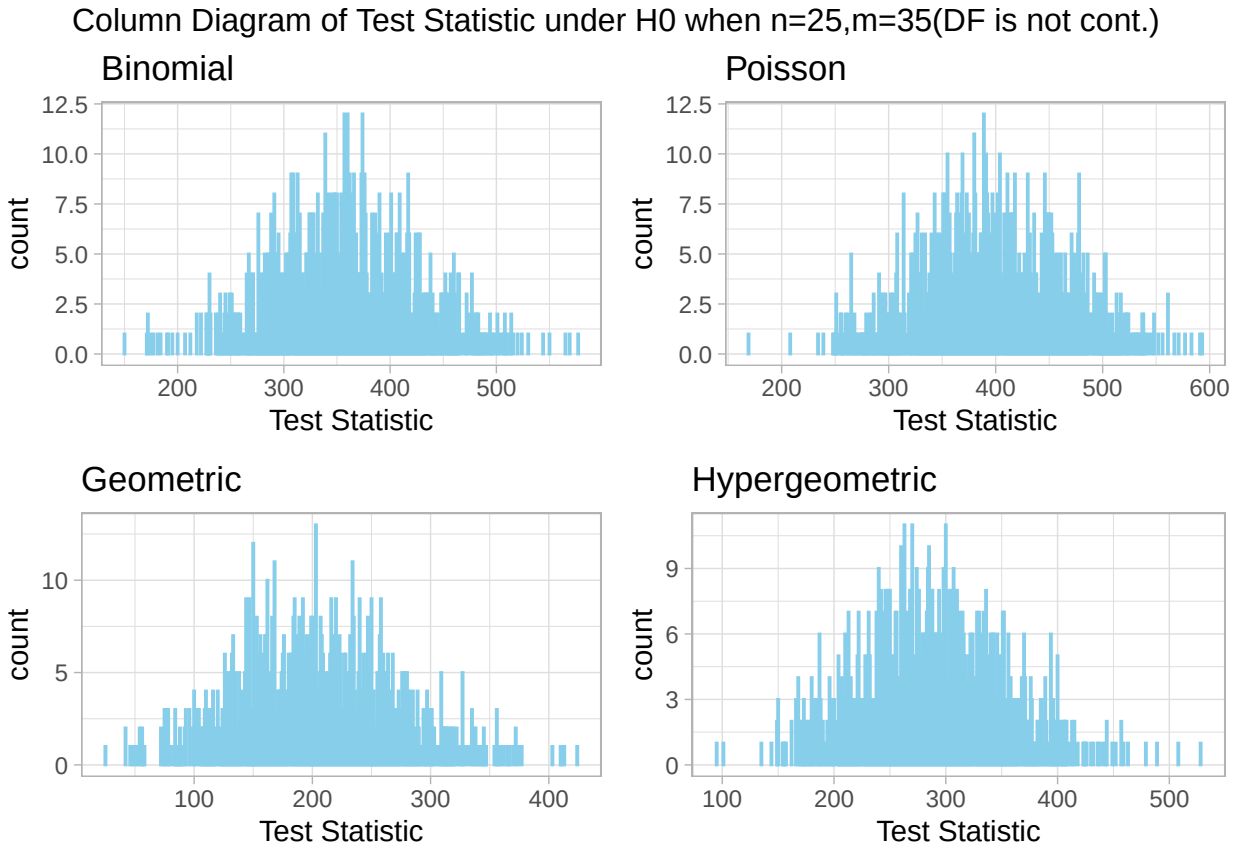


Geometric



Hypergeometric





COMMENT:

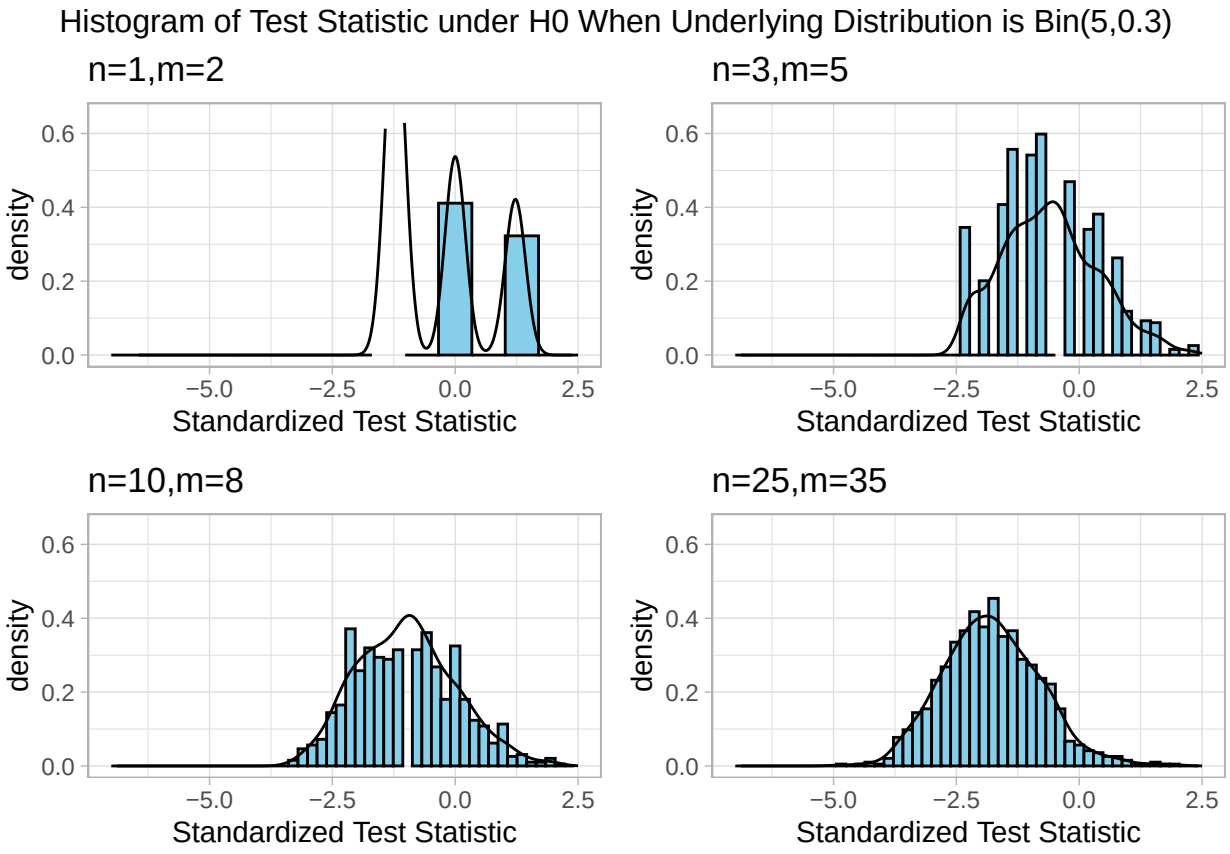
- Thus when the distribution functions are not continuous the Mann Whitney Statistic is not distribution free under H_0.

Asymptotic Distribution of Test Statistic when the Distribution Functions are not continuous:

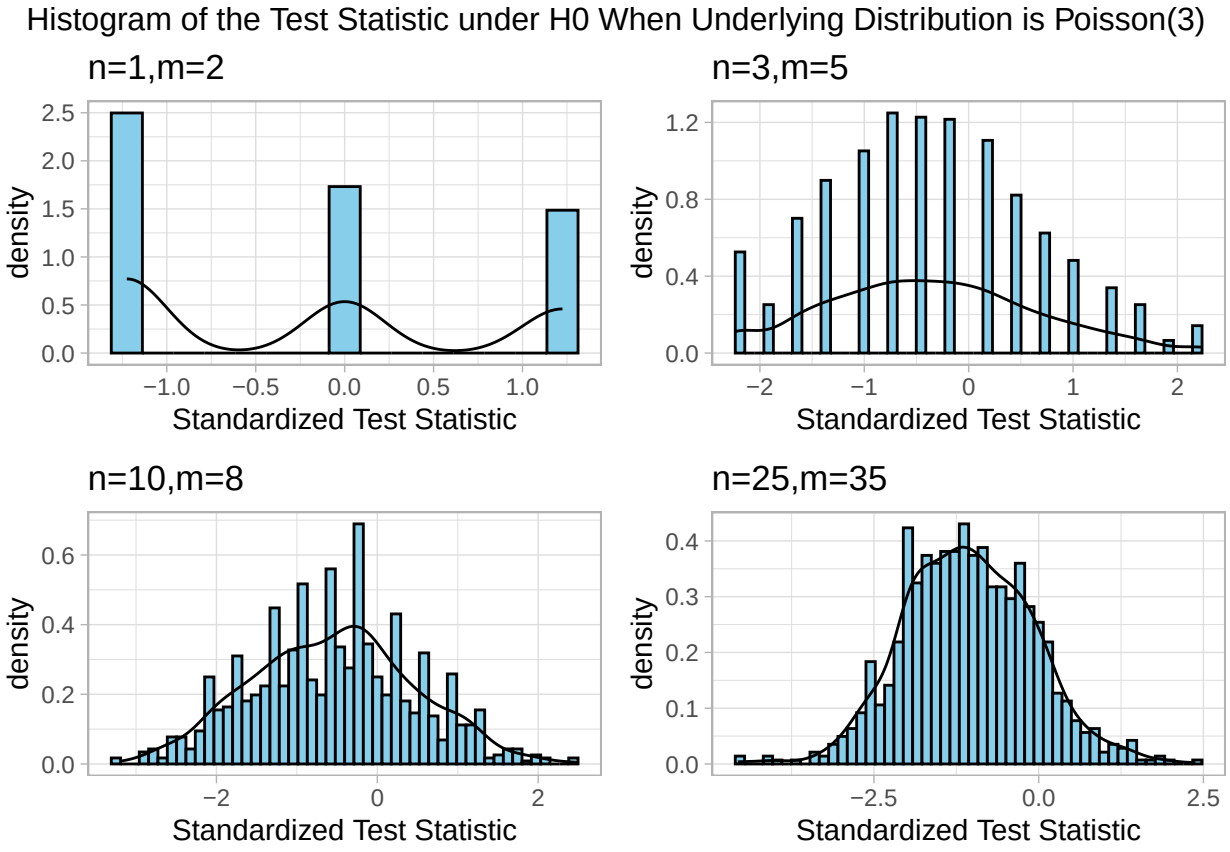
- Here, we want to find the asymptotic distribution of the test statistic under H_0 and H_1.
- Here, we generate the observations from Bin(5,0.3) and Poisson(3) under H_0.
- We generate the observations from Bin(5,p) under H_1.

Under H_0:

The observations are coming from Bin(5,0.3):

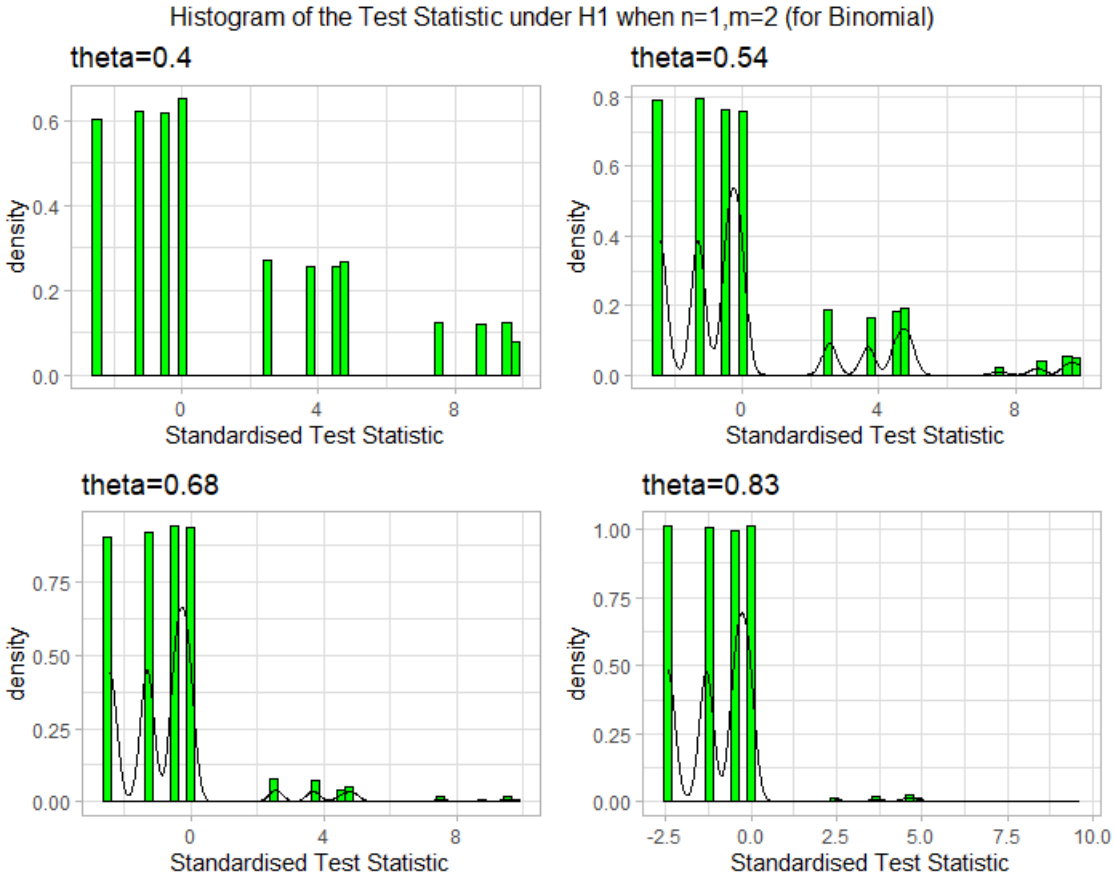


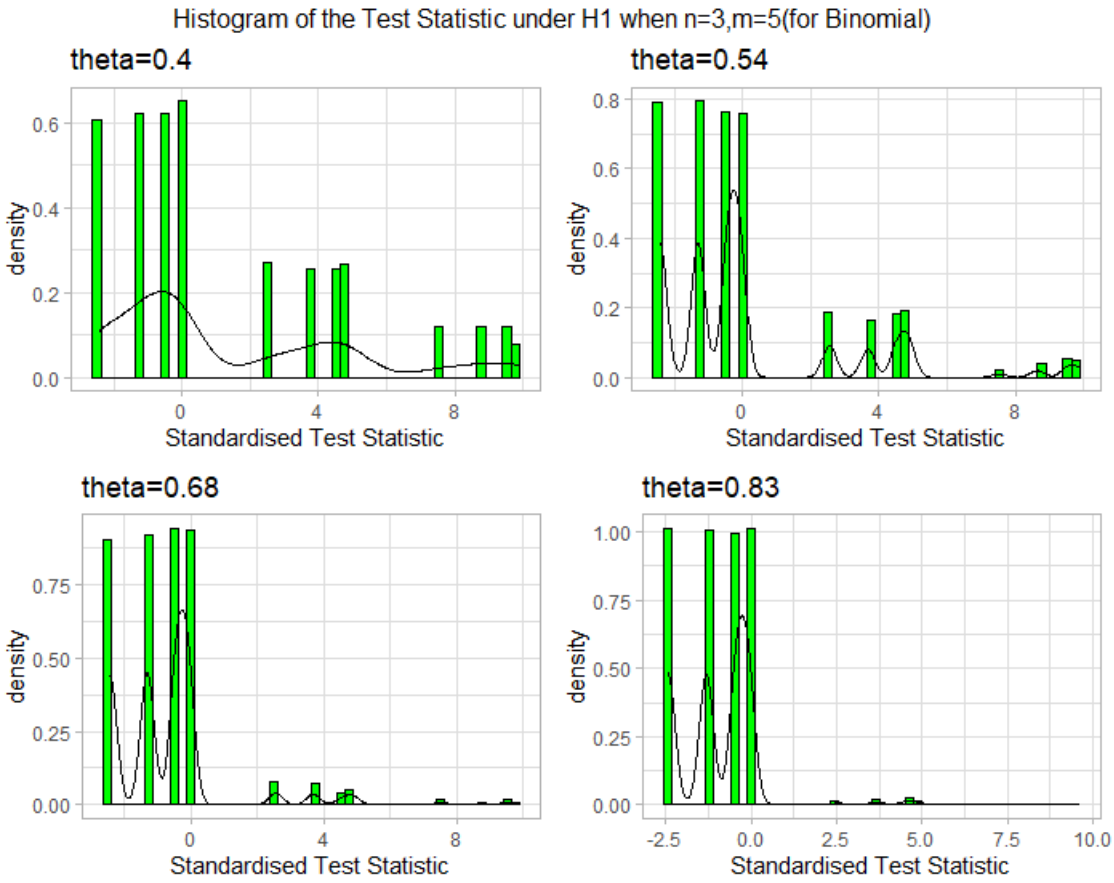
The observations are drawn from Poisson(3):

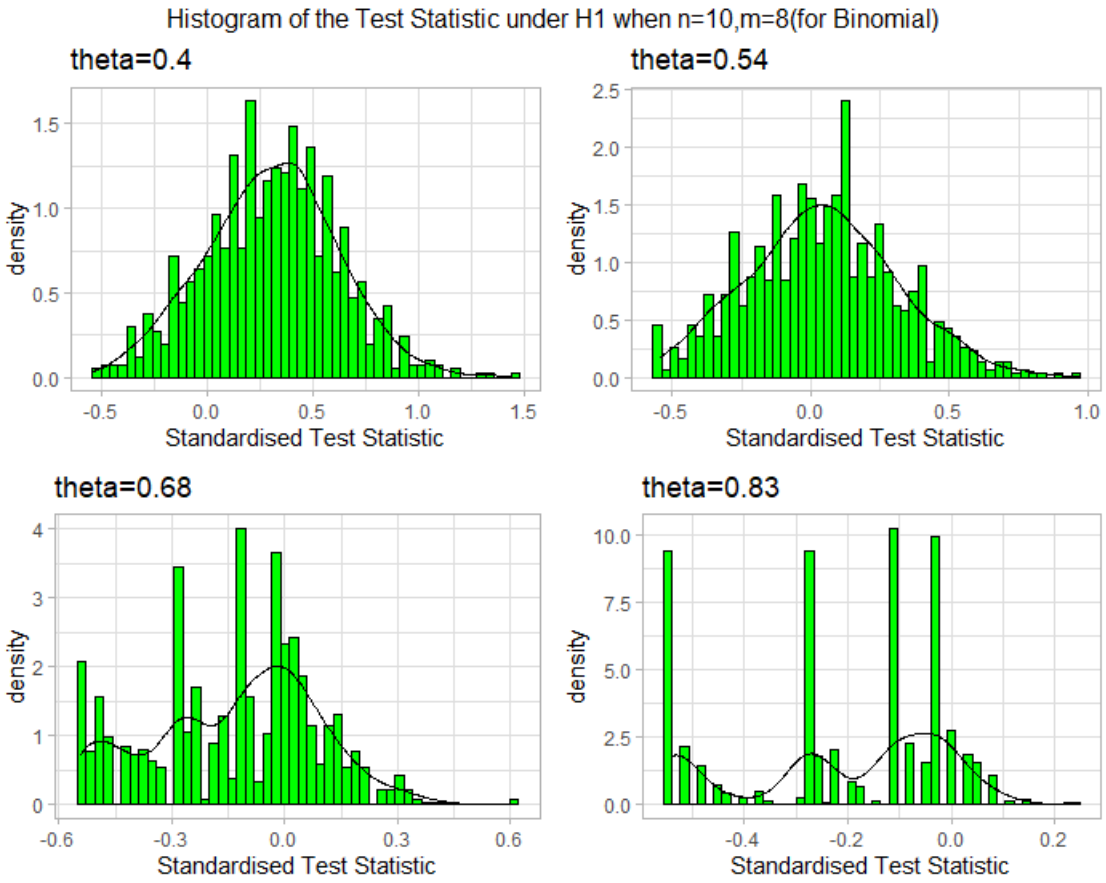


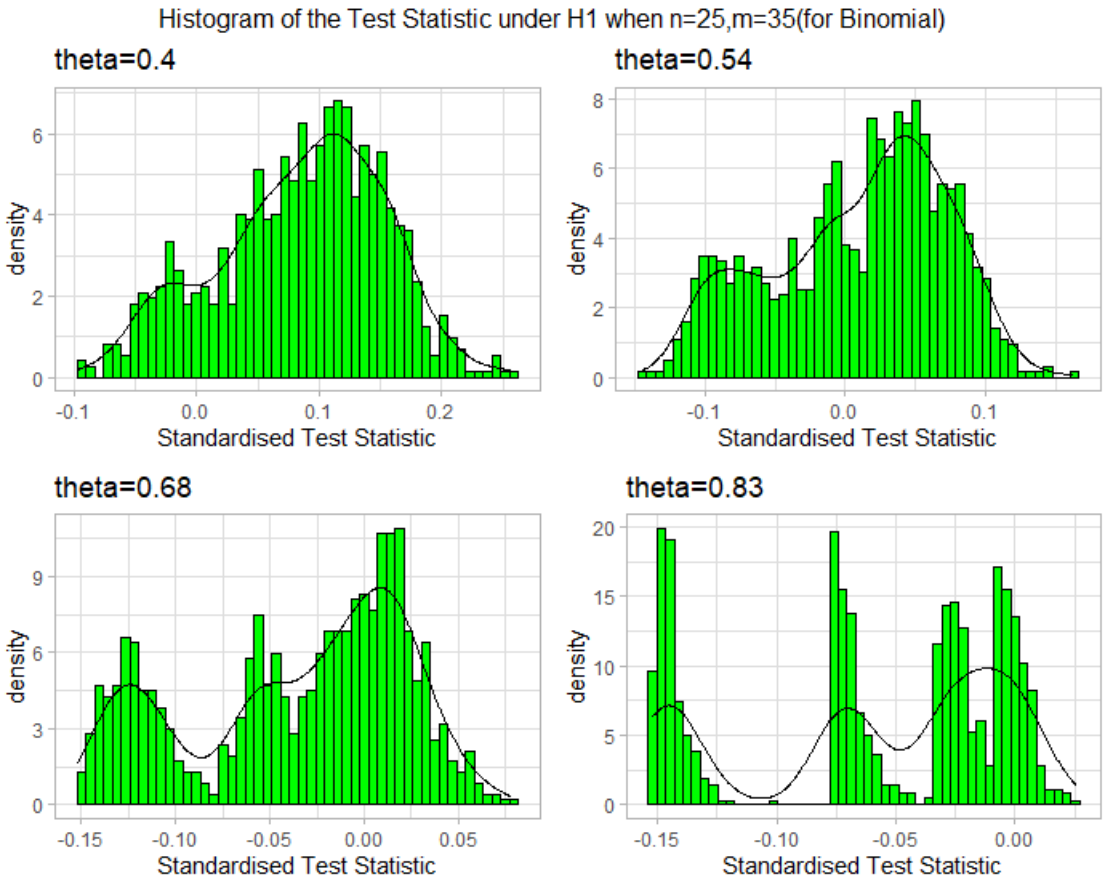
Under H_1:

The observations are coming from Bin(5,p):




Histogram of the Test Statistic under H1 when n=3,m=5(for Binomial)
theta=0.4
theta=0.54
theta=0.68
theta=0.83
density
Standardised Test Statistic


Histogram of the Test Statistic under H1 when n=10,m=8(for Binomial)
theta=0.4
theta=0.54
theta=0.68
theta=0.83
density
Standardised Test Statistic



COMMENT:

- Thus we can see that in this case the asymptotic distribution of Mann Whitney Statistic is a symmetric distribution but we can't say its normal.
- But it is quite difficult to find out which distribution since it changes as we change the underlying distribution.

Power and Size of the test H_0 vs H_1

- Here we take 4 different choices of level of significance which are .001, .01, .05, .1. The critical values corresponding to these levels are obtained from the table of exact null distribution of Mann- Whitney U statistic. (Source:)
- Sizes are simulated for small as well as for large sample sizes

Simulated size in case of small sample size(n=5, m=8)

- Given below the table that shows the simulated size of the test when the observations are generated from Normal(0,1) and Cauchy(0,1)

level of significance	size(Normal)	size(Cauchy)
0.001	0.000	0.000
0.010	0.008	0.006
0.050	0.048	0.053
0.100	0.086	0.091

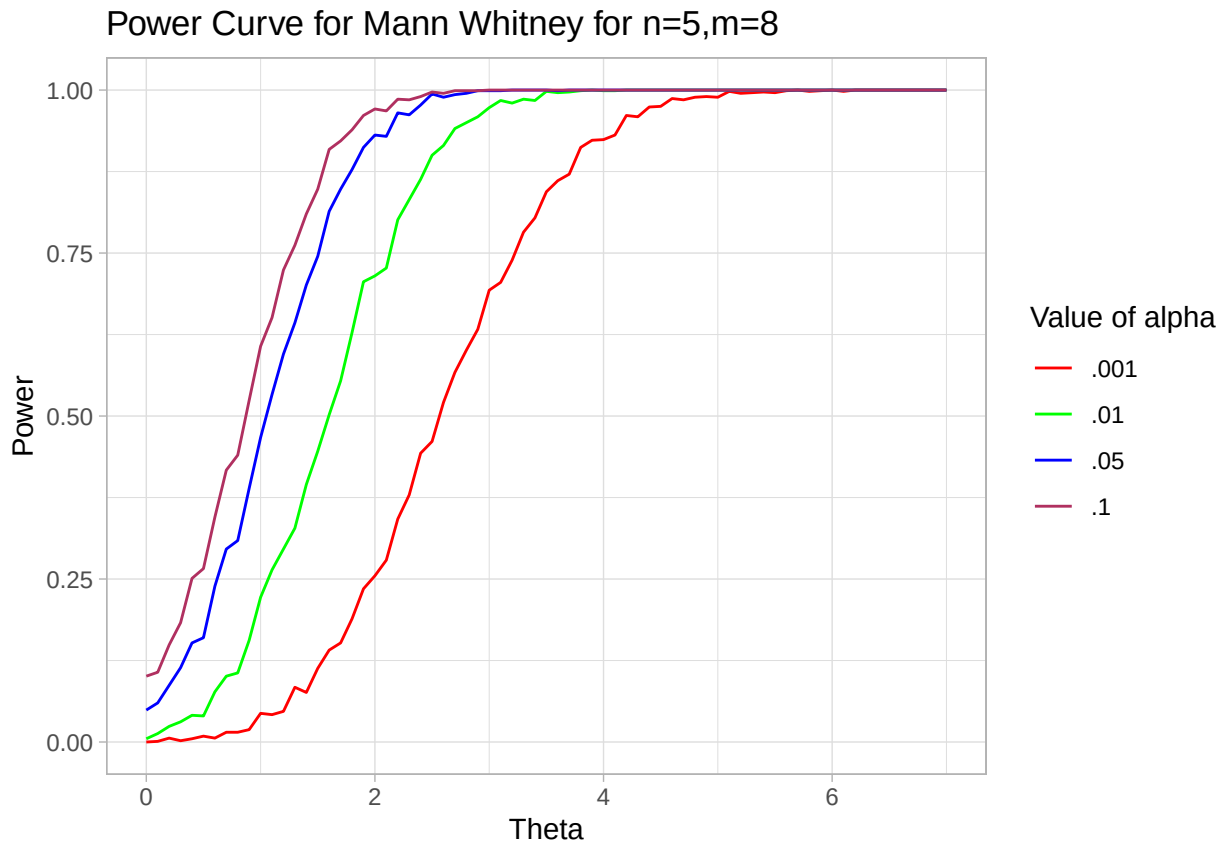
level of significance	size(Normal)	size(Cauchy)

Simulated size in case of large sample size(n=100, m=200)

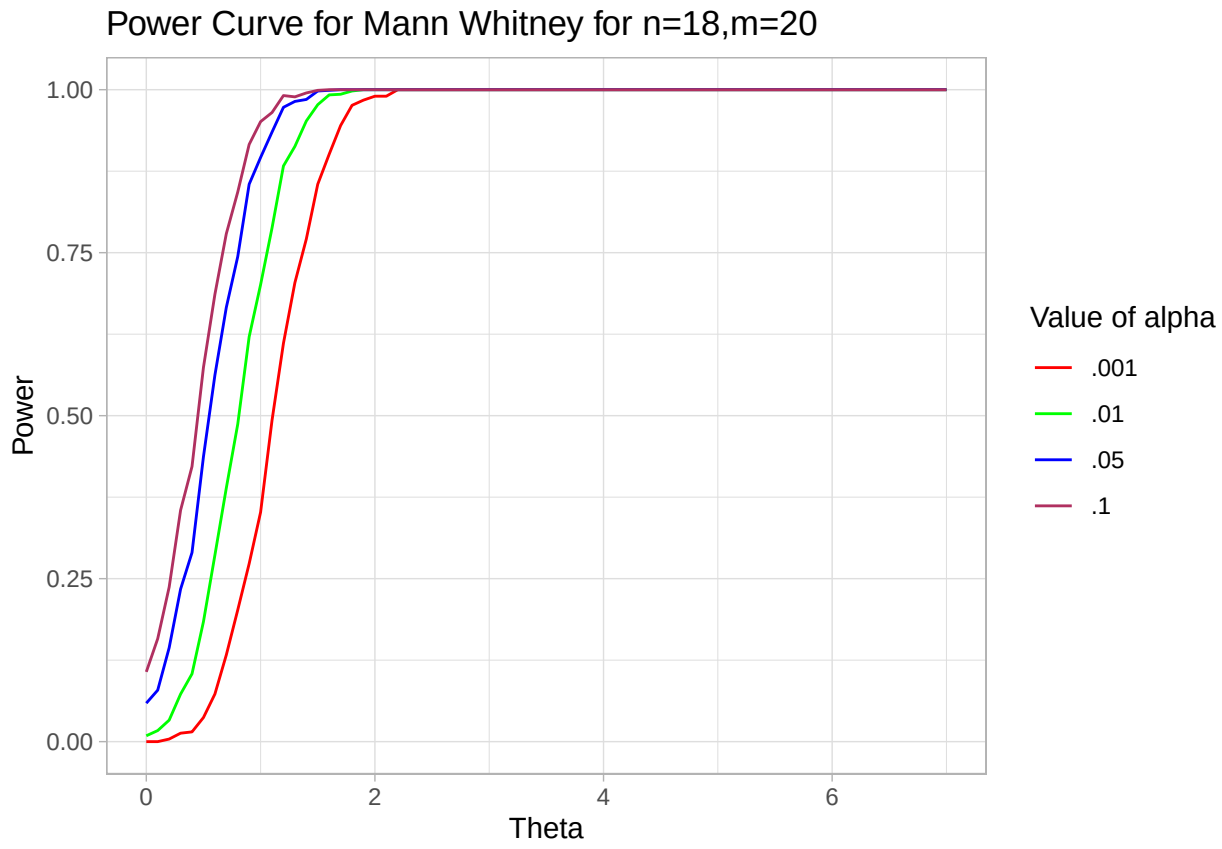
- Since we know that under our null hypothesis, the mann whitney statistics after standardization follows Normal(0,1), the critical values here are obtained from the normal quantiles itself.

level of significance	size
0.001	0.002
0.010	0.007
0.050	0.045
0.100	0.104

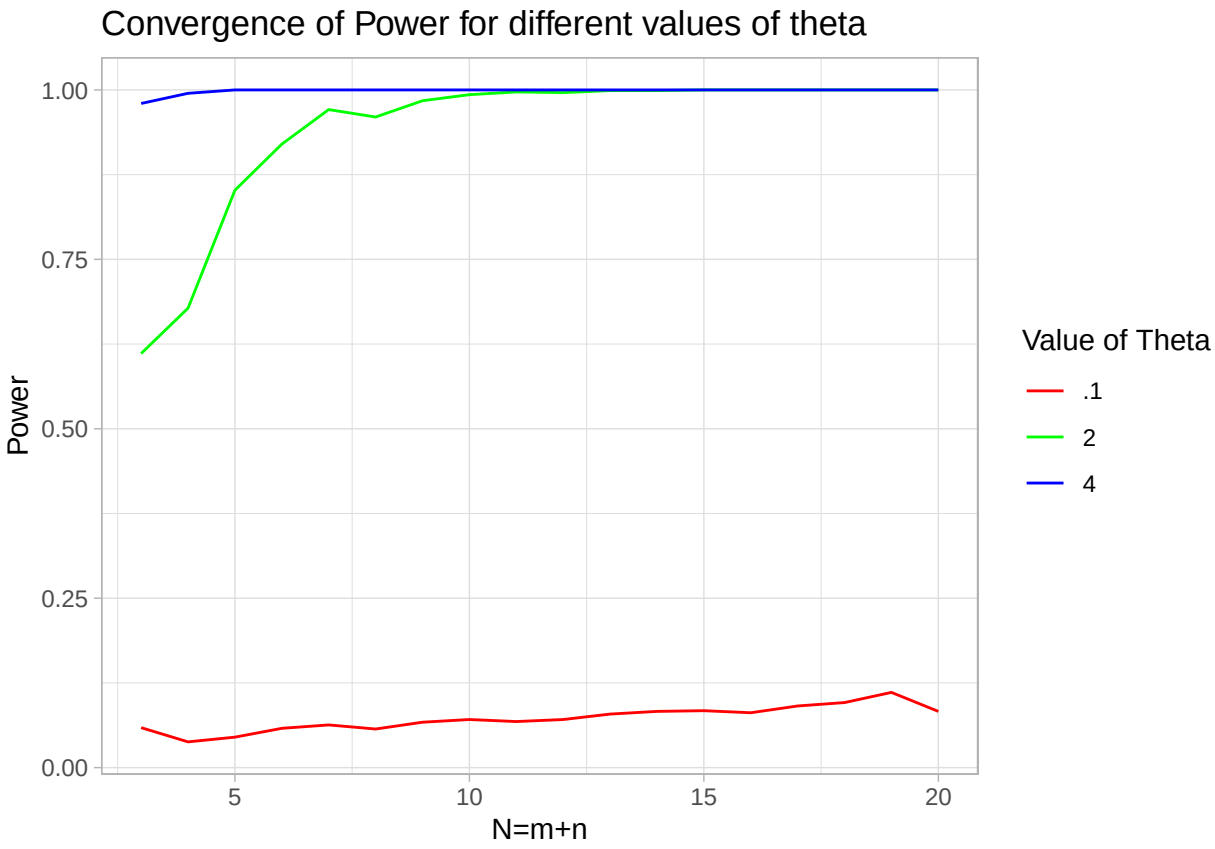
Simulated power curve in case of small sample size(n=5, m=8)



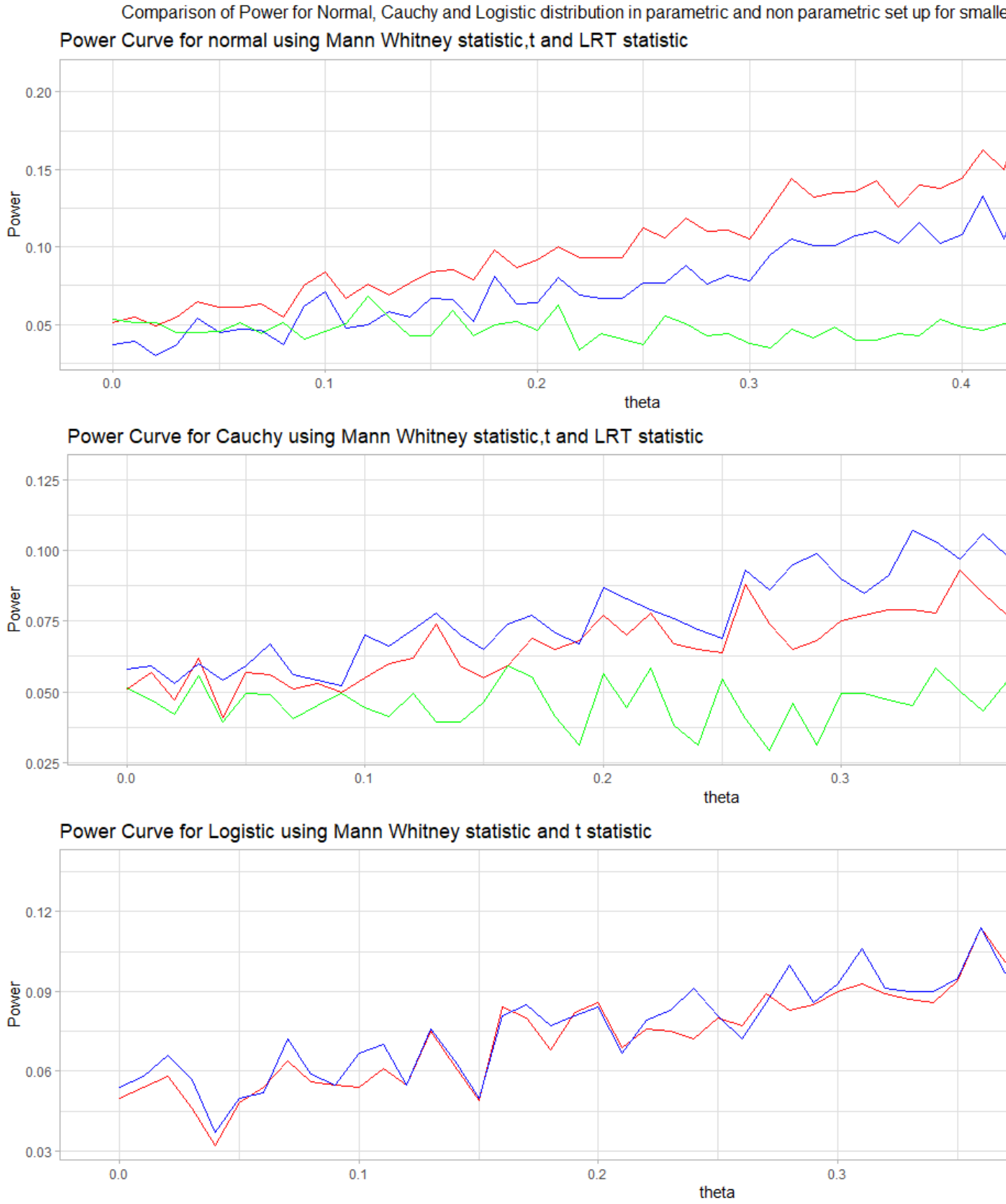
Simulated power curve in case of large sample size(n=18, m=20)



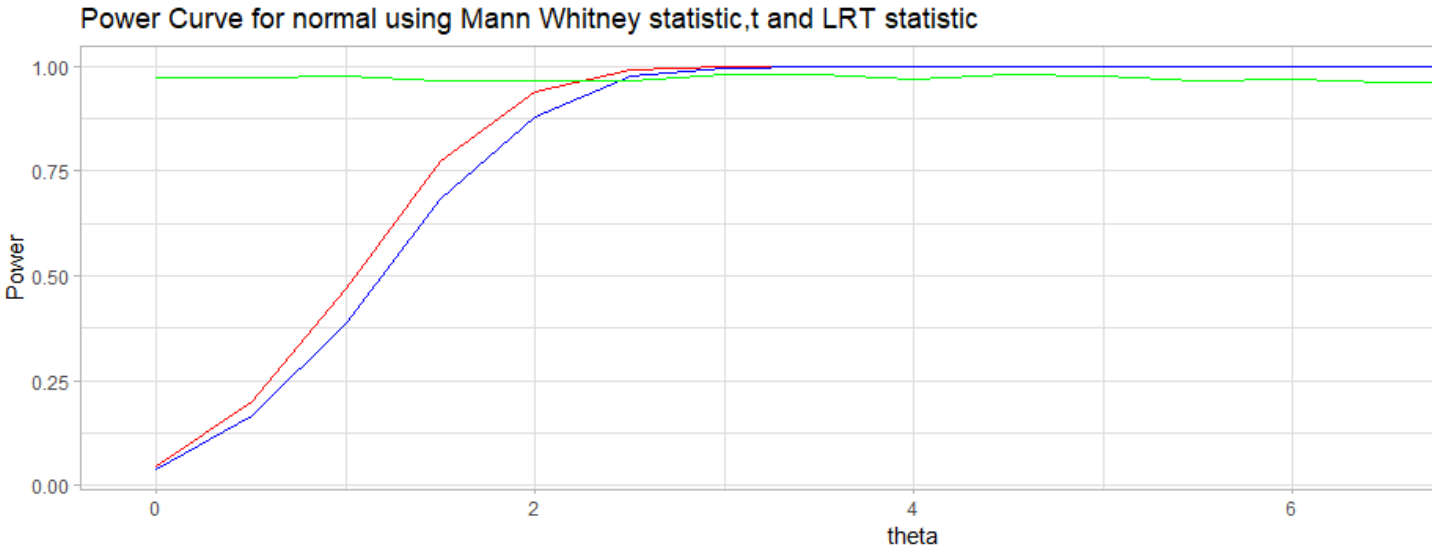
Asymptotic Consistency

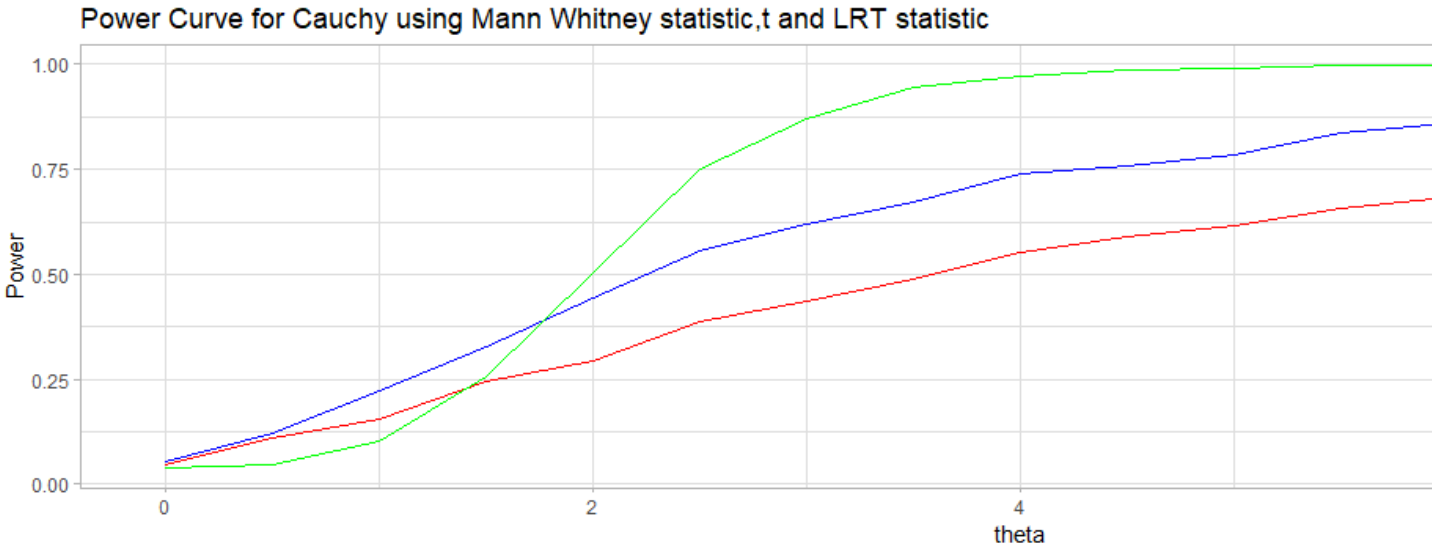


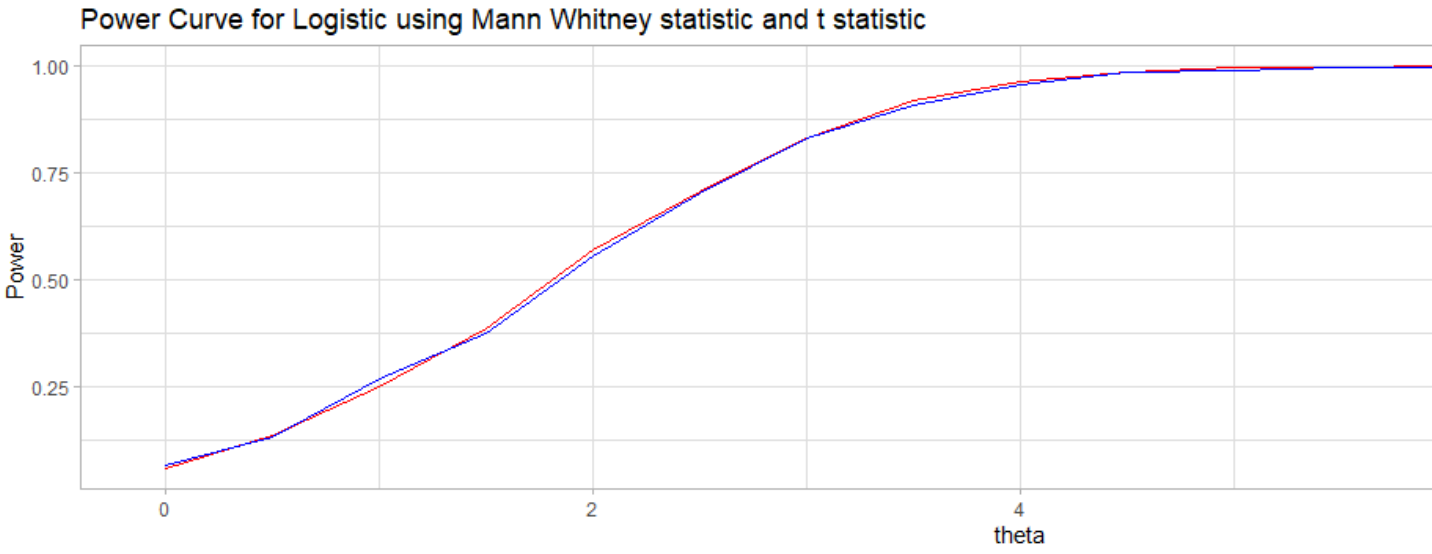
PARAMETRIC COUNTERPART:



Comparison of Power for Normal, Cauchy and Logistic distribution in parametric and non parametric set up for high values



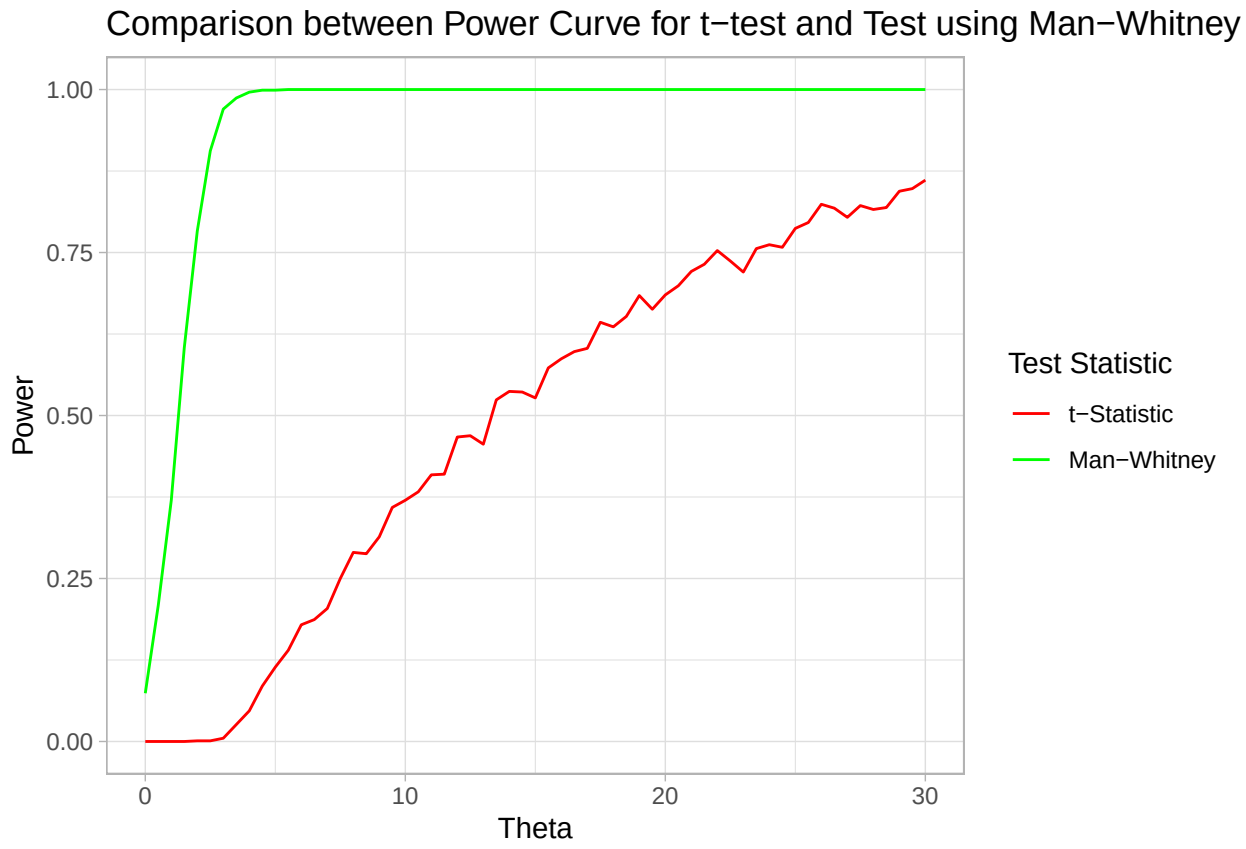




COMMENT :

- Firstly we observe three distributions(normal,cauchy and logistic) for sample sizes n=5 and m=7 respectively.
- We observe that if we take a very small neighbourhood of theta around 0 for normal distribution the power of the t-test is maximum which is quite obvious but the power of the non-parametric test Mann-Whitney is showing more power than likelihood ratio test.
- For Cauchy and Logistic distribution the Mann-Whitney shows more power than t-test and LRT, for Logistic its quite justified since for Logistic distribution Mann-Whitney is the Locally most powerful test.
- When we consider a more wider neighbourhood of theta we see that for normal distribution the LRT always shows power 1 which is quite counter intuitive but our hunch is this is happening due to the inefficiency of the optim function.
- For Cauchy distribution as the value of theta increases the power of the LRT increases than that of Mann-Whitney and t-test which is justified as LRT is a parametric it should have power more than that of a non-parametric test.
- For Logistic still Mann-Whitney shows more power but the plot seems if we take more wider neighbourhood of theta the power of the t-test will exceed that of the Mann-Whitney power.

Robustness: t Test and Test based on Mann Whitney U statistic

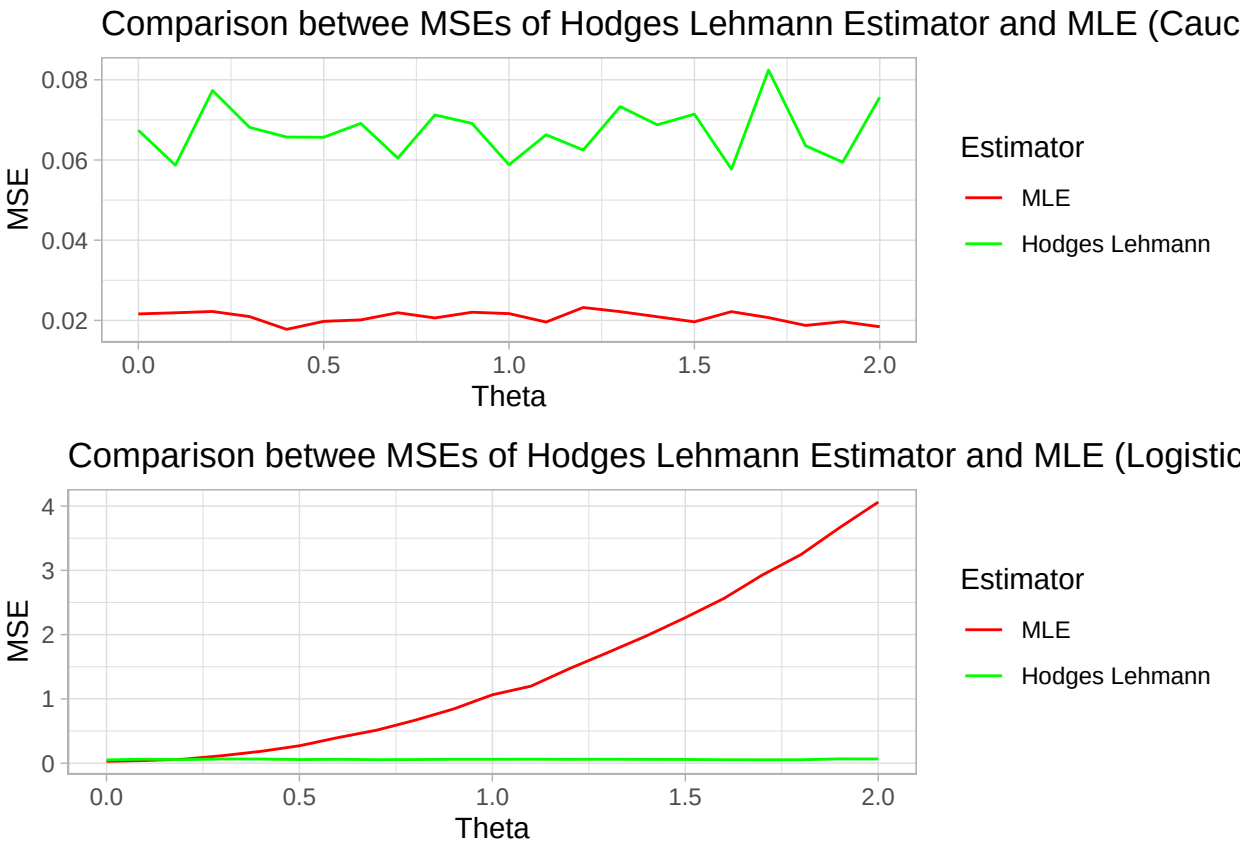


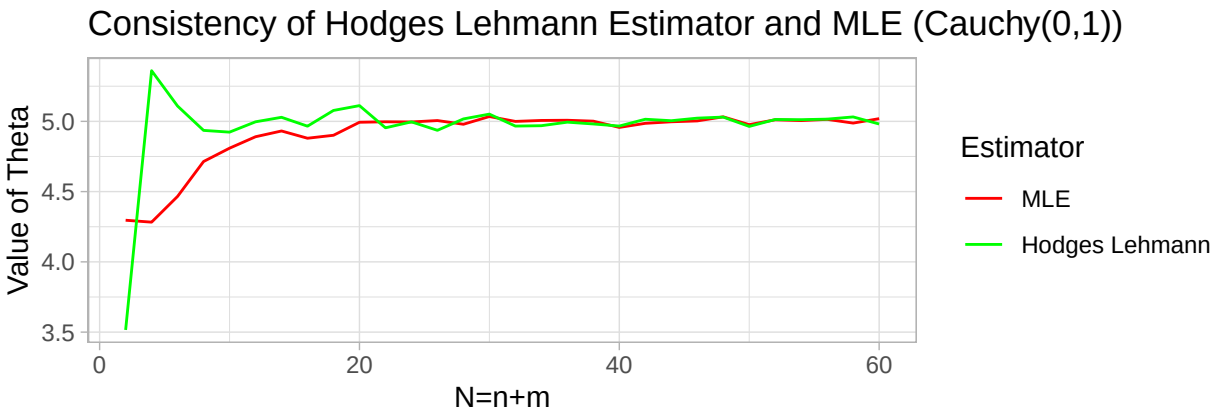
Two Sample Estimation: Hodges Lehmann Estimator

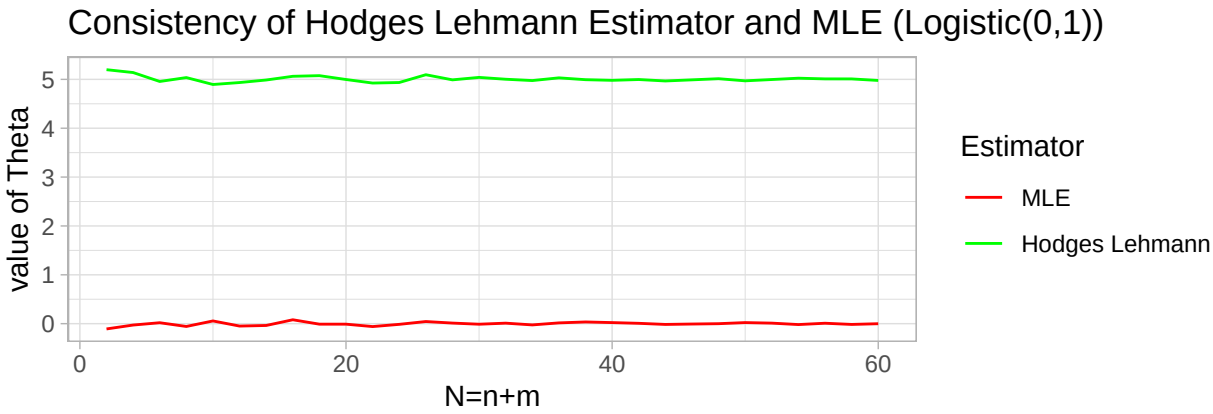
- Under the current setup we can directly derive the Hodges Lehmann Estimator for estimating the location shift.

Since the MLEs of both Logistic and Cauchy distribution do not have a closed form, here we tried to compare

whether the Hodges Lehmann estimator performs better than the Maximum Liklihood Estimator. The motivation of this comparison is if we find the Hodges Lehmann Estimator performs better than the MLE then we may use Hodges Lehmann estimator instead of MLE as Hodges Lehmann estimator has a closed form and we can calculate it easily.







COMMENT

- In case of Cauchy distribution we can definitely see that MLE has less Mean squared Error. Also, it can be noticed that MLE and Hodges Lehmann both are consistent for the location parameter θ.
- Due to inefficiency of the 'optim', 'nleqslv' functions in R, it seems that the comparison of MSEs of MLE and Hodges Lehmann Estimator of Logistic distribution gives us absurd result. We tried to do the simulation by solving the score equations but the results are similar. So, in this case we could not draw a conclusion regarding this coparison