A Simulation Study on Mann Whitney Statistic

Jishu Adhikary
Mainack Paul
Sreejit Roy

April 15, 2022

Contents

1 Introduction **3**

2 Formulation of the Two Sample Problem **4**

3 Empirical Verification **6**
3.1 U is distribution free under null 6
3.2 Expectation of U under null 15
3.3 Variance of U under null 16

4 Results on Size and Power **20**
4.1 Empirical Size 20
4.2 Empirical Power 25

5 Large Sample Properties **32**

6 A Case Study **45**

7 References and Bibliography **49**

8 Acknowledgements **49**

1 Introduction

Suppose there are two populations with corresponding distributions F and G. A very common question regarding the populations is: *are these populations exactly same?* I.e., in terms of distribution functions, is $F(x) = G(x)\ \forall x$?

At this stage, it is important to point out that further proceedings are difficult without any additional assumptions. If we are willing to make parametric model assumptions about the forms of the underlying populations and assume that differences between the two populations occur only with respect to some parameters, for example, say, the means or the variances, it is sometimes possible to derive the *best test* in Neyman-Pearson framework[1]. For example, if we assume that the underlying populations are normally distributed, it is well known that the two-sample Student's t test for equality of means and the F test for equality of variances are respectively the *best tests*. The performances of these two tests are also well known. However, as we would note in section 2, these and other classical tests are sensitive to violations of the fundamental model assumptions inherent in the derivation and construction of these tests. Therefore, any conclusions arrived at using such tests will be valid as long as the underlying assumptions are valid. If there is any reason to suspect a violation of any of the assumptions, or if sufficient information to judge their validity is unavailable, or if a completely general test of equality for unspecified distributions is desired, nonparametric procedures would be much more fitting.

Even though we are willing to use a nonparametric procedure, some assumptions need to be made. A common assumption is that of a *location model*, defined at the beginning of the next section. Based on that we would introduce the Mann Whitney statistic, based on which we would try to answer the question posed at the beginning of this section. We would also try to verify some common properties and arrive at some remarks.

[1] any text on parametric testing of hypothesis serves as a good introduction to this topic.

2 Formulation of the Two Sample Problem

First, consider the following definition:

Definition 2.1. *(Location Model)* Consider two random variables X and Y with corresponding distribution functions F and G. If

$$G(x) = Pr(Y \leqslant x) = Pr(X \leqslant x - \theta) = F(x - \theta)\ \forall x \text{ and } \theta \neq 0, \tag{1}$$

then we say that X and Y are from the same location model.

Remark 2.1. Definition 2.1 means that $X + \theta$ and Y have the same distribution. In other words, X is distributed as $Y - \theta$.

Remark 2.2. If $\theta = 0$, both the populations considered in definition 2.1 are same. If $\theta > 0 (< 0)$, the population of Y is shifted to the right (left).

Remark 2.3. Under the location model, both the populations have the same shape and same variance. In fact, the amount of shift θ is equal to the difference between any two respective location parameters or quantiles of the same order.

The two sample problem, in terms of hypotheses is as follows:

$$H_0 : F(x) = G(x)\ \forall x \tag{2}$$

against

$$H_A : not\ H_0 \tag{3}$$

In particular, H_A can be any of the following:

$$H_1 : F(x) \geqslant G(x) \forall x;\ F(x) > G(x)\ for\ atleast\ one\ x \tag{4}$$

$$H_2 : F(x) \leqslant G(x) \forall x;\ F(x) < G(x)\ for\ atleast\ one\ x \tag{5}$$

$$H_A : F(x) \neq G(x)\ for\ some\ x \tag{6}$$

Note that under the location model, (2), (4) - (6) simplifies to the following:

$$H_0 : \theta = 0;\ H_1 : \theta > 0;\ H_2 : \theta < 0;\ H_3 : \theta \neq 0 \tag{7}$$

For any testing problem, a prerequisite is to draw a random sample. Here, for the two sample problem considered under the location model, we draw two independent samples of size n and m respectively from F and G. I.e.,

$$X_i \overset{iid}{\sim} F(.),\ i = 1(1)n \tag{8}$$

and

$$Y_i \overset{iid}{\sim} G(.),\ i = 1(1)m \tag{9}$$

Finally, define $N = n + m$.

In the following subsection, we define the Mann Whitney statistic.

Mann Whitney U Statistic

Mann Whitney U test[1] is based on the idea that when the combined sample is arranged in an increasing order of magnitude, the pattern exhibited provides information about the relationship between their populations. Mann Whitney U statistic is based on the magnitude of Y's in relation to the X's. I.e., the position of Y's in the combined ordered sequence. A sample pattern of arrangement where most of the Y's are greater than most of the X's, or vice versa, or both, would be evidence against a random mixing, i.e., against H_0.

Definition 2.2. *(Mann Whitney U Statistic)*

$$U \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{m}\phi_{ij}, \tag{10}$$

where $\phi_{ij} \coloneqq \begin{cases} 1, & \text{if } X_i > Y_j \\ 0, & \text{otherwise.} \end{cases}$

Remark 2.4. U is distribution free under H_0.

Remark 2.5. $E_{H_0}(U) = nm/2$; $var_{H_0}(U) = nm(N+1)/12$.

Remark 2.6. Reject H_0 in favour of H_1 for small values of U.

Remark 2.7. Reject H_0 in favour of H_2 for large values of U.

Remark 2.8. Reject H_0 in favour of H_3 for values of U that are too small or too large.

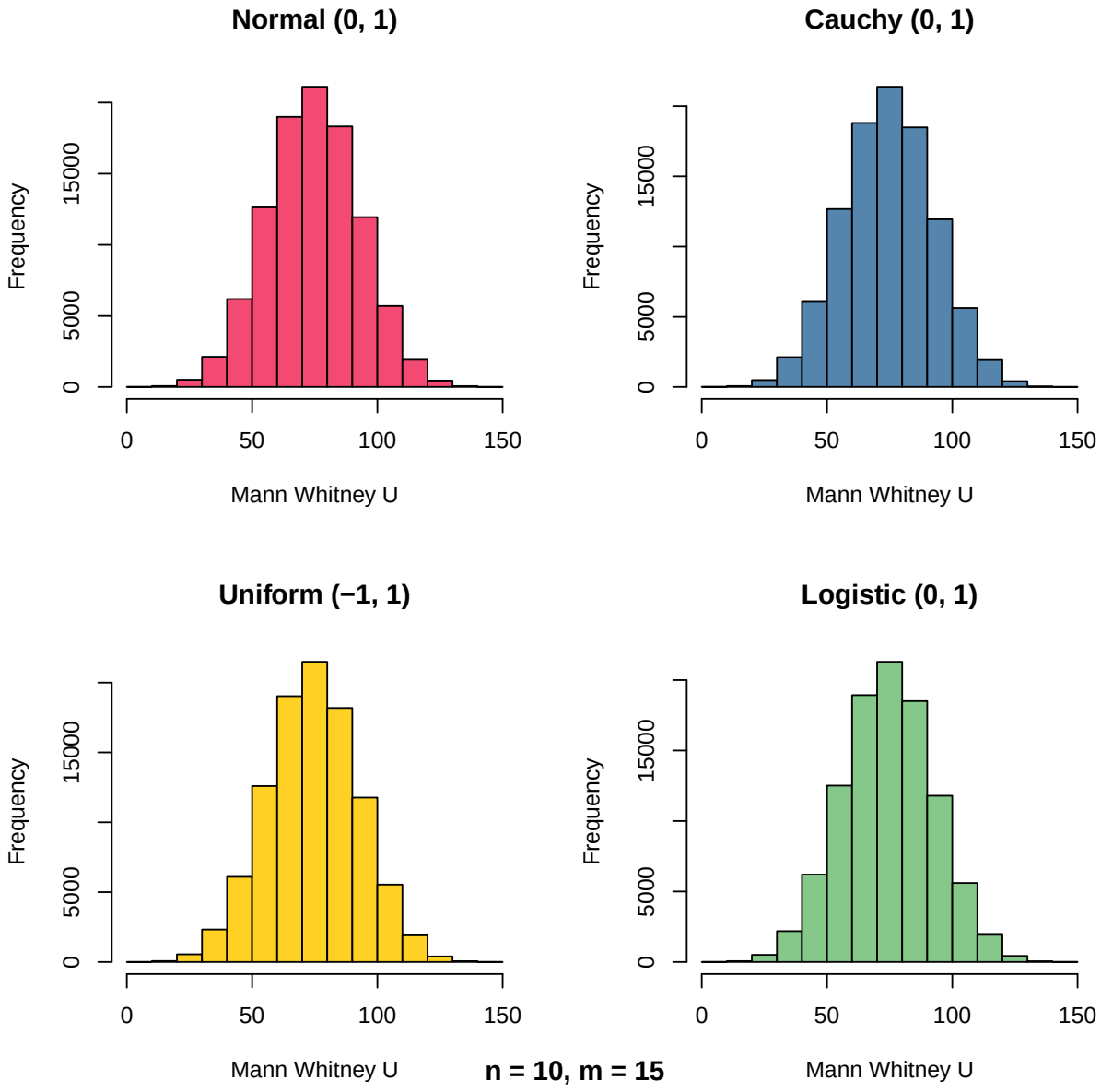
We empirically verify remarks 4 and 5 in the next section.

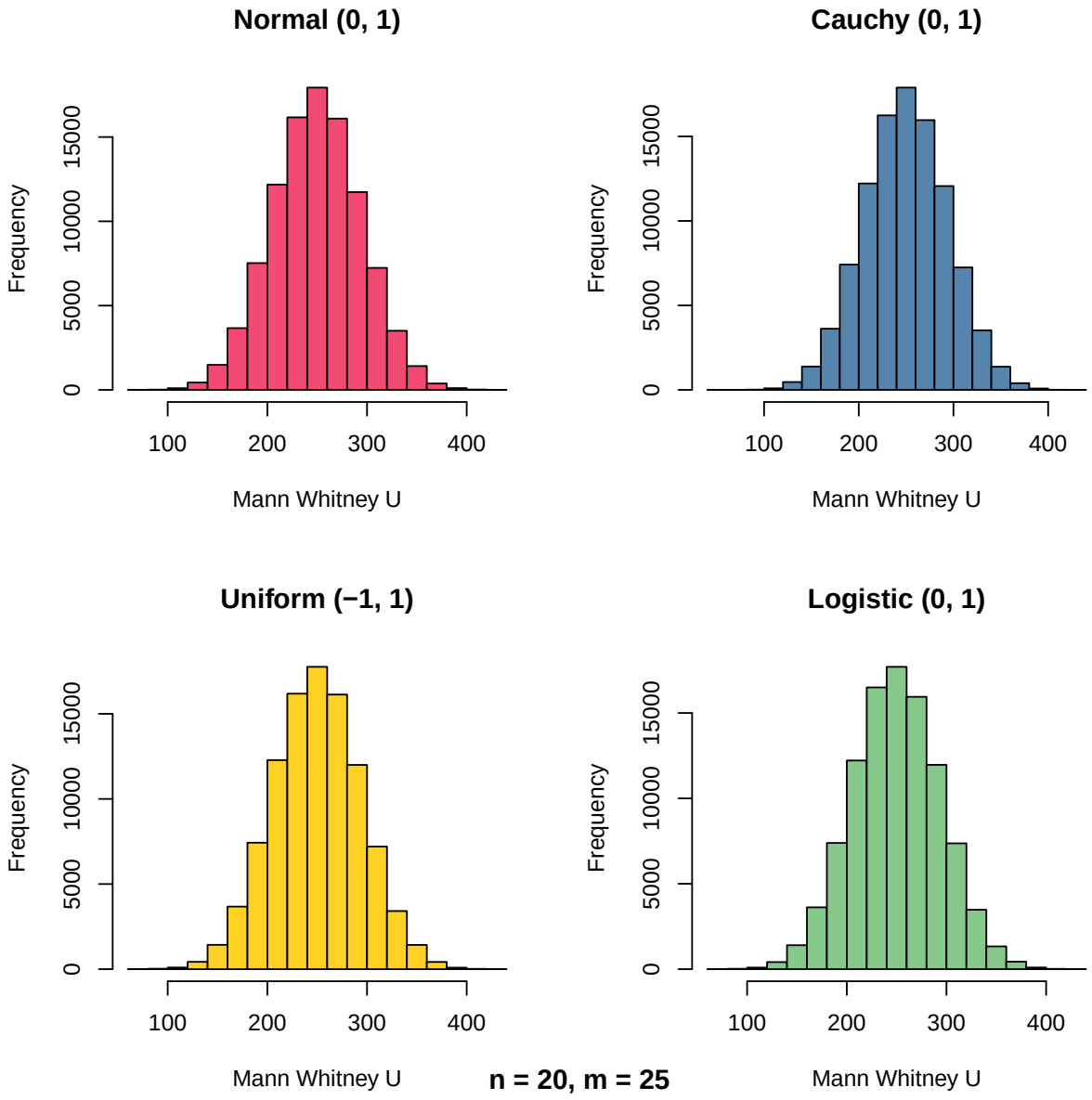
[1] Mann, Henry B.; Whitney, Donald R. (1947). *On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other.* Annals of Mathematical Statistics. 18 (1): 50–60.

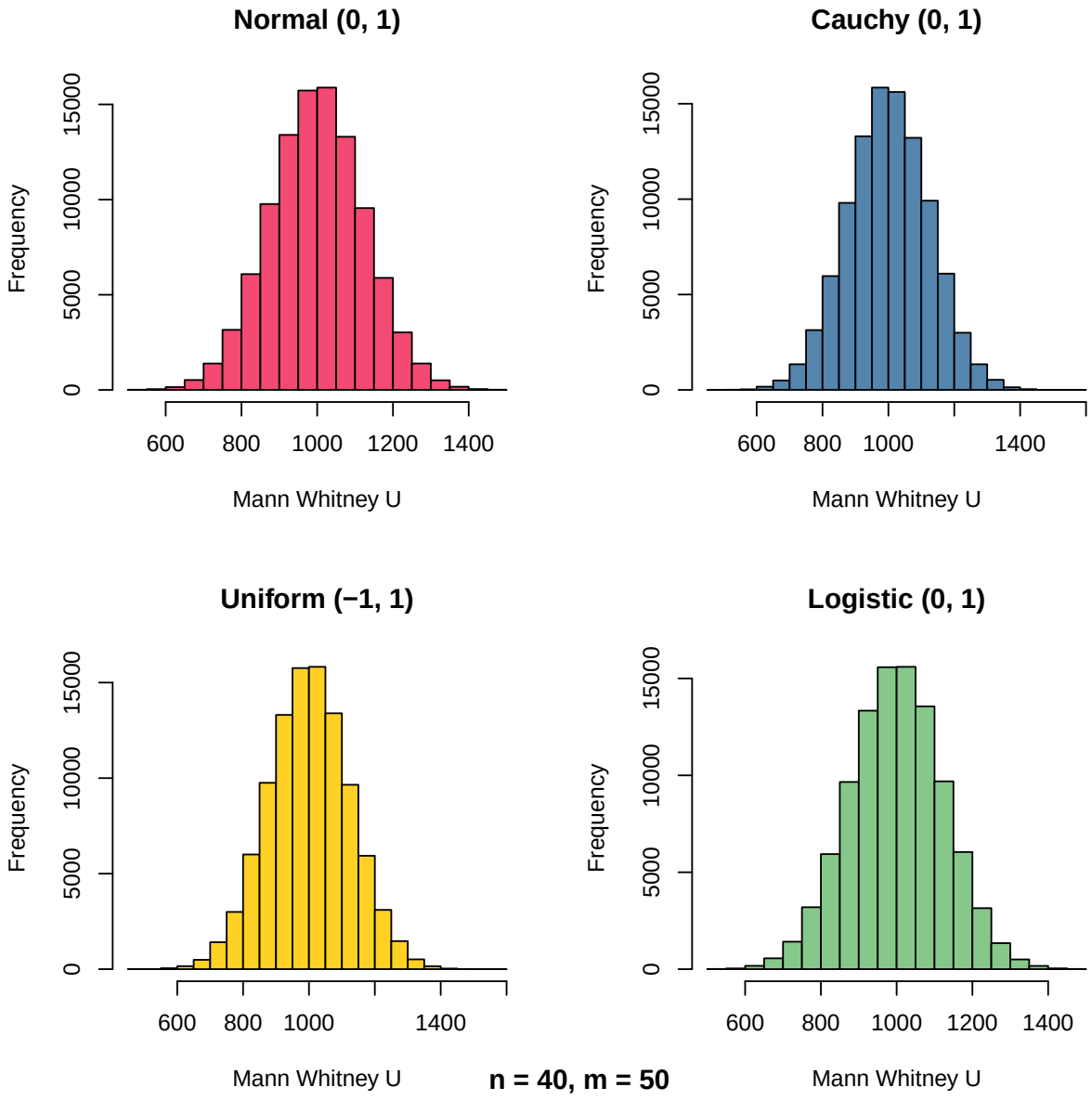
3 Empirical Verification

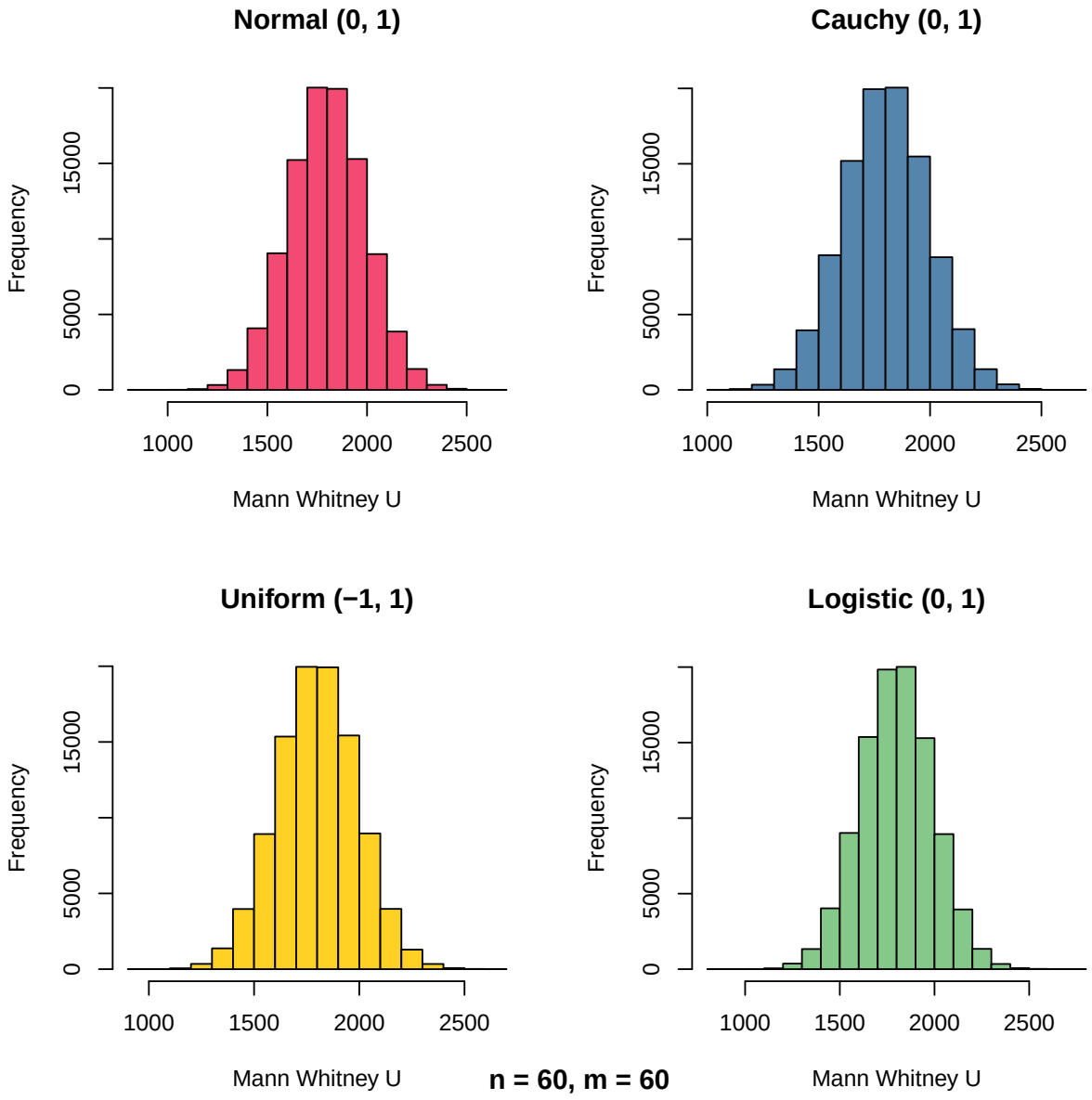
3.1 U is distribution free under null

The methodology of the verificaton is simple - we consider different continuous distributions from location model. We then try to show that for a given pair $(n, m) \in \mathbb{N}^2$, the structure of the distribution of U is same irrespective of the underlying distribution of the population.




Normal (0, 1)
Cauchy (0, 1)
Uniform (−1, 1)
Logistic (0, 1)
Frequency
0
5000
10000
15000
100
200
300
400
Mann Whitney U
n = 20, m = 25


Normal (0, 1)
Frequency
0
5000
10000
15000
600
800
1000
1200
1400
Mann Whitney U
Cauchy (0, 1)
Frequency
0
5000
10000
15000
600
800
1000
1400
Mann Whitney U
Uniform (−1, 1)
Frequency
0
5000
10000
15000
600
800
1000
1400
Mann Whitney U
Logistic (0, 1)
Frequency
0
5000
10000
15000
600
800
1000
1200
1400
Mann Whitney U
n = 40, m = 50



Remark 3.1. Observe from the figures that irrespective of the underlying distribution, the form of the distribution of of the Mann Whitney statistic is mostly the same for a given n and m.[1] This conforms to remark 2.4.

A remark about the advantage over parametric methods

As we have noted in section 3.1, U statistic is distribution free under the null hypothesis. Now suppose we decided not to use the Mann Whitney test, rather to opt for the *classical two sample t test*. For that, we would calculate the *t statistic*, instead of the Mann Whitney U statistic. The

[1] An interesting question: is the distribution characterised by the pair (n, m), or simply N? I.e., for (n_1, m_1), $(n_2, m_2) \ni n_1 + m_1 = n_2 + m_2$, the distributions same?

following is the definition of *two sample t statistic*:

Definition 3.1. *(Two Sample t Statistic)* For two samples $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$, the statistic is defined as follows:

$$t := \frac{\bar{X} - \bar{Y}}{s(\frac{1}{n} + \frac{1}{m})^{1/2}}, \tag{11}$$

where, $s^2 := \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2}$.

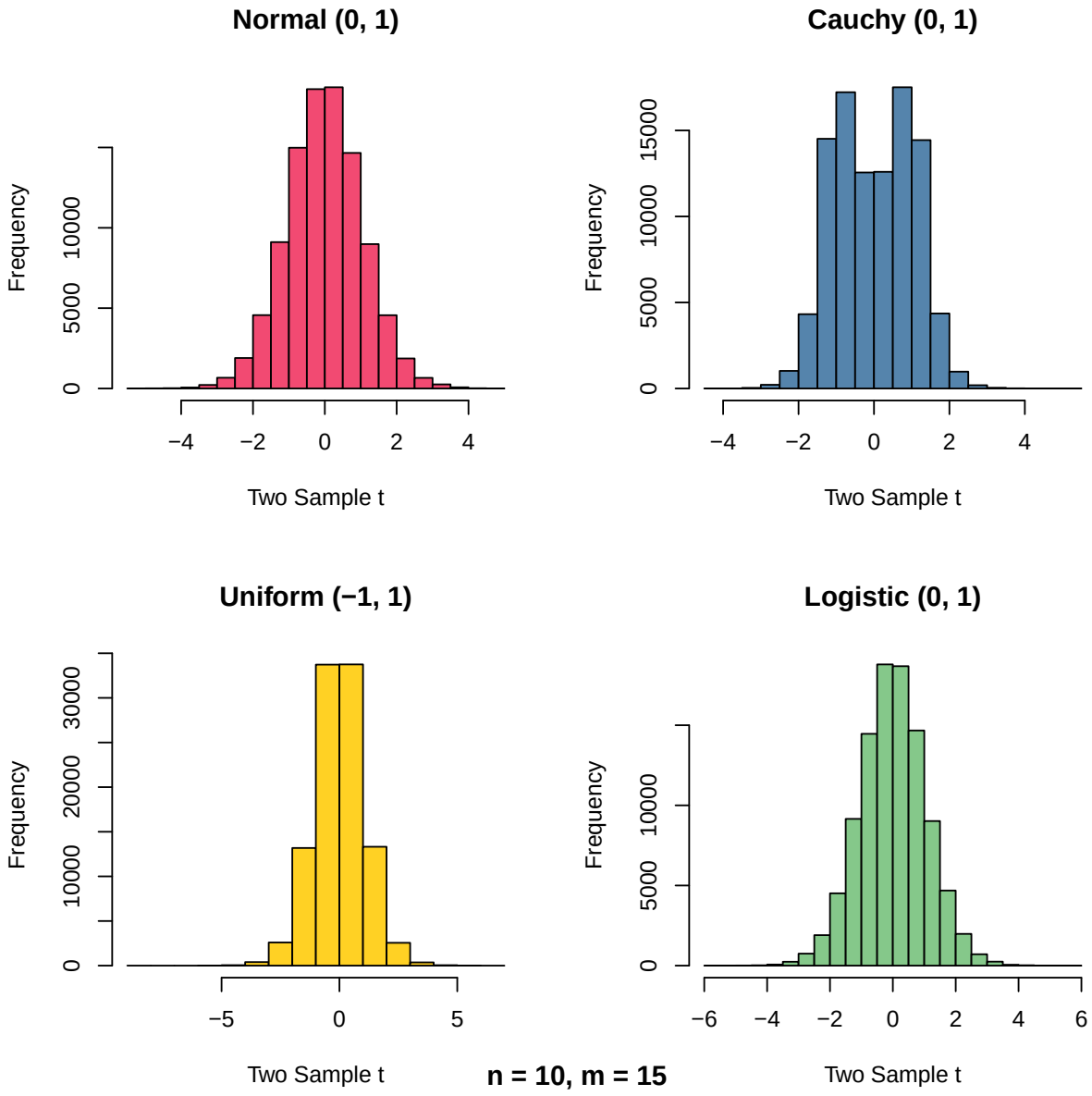
Remark 3.2. $t \overset{H_0}{\sim} t_{m+n-2}$.

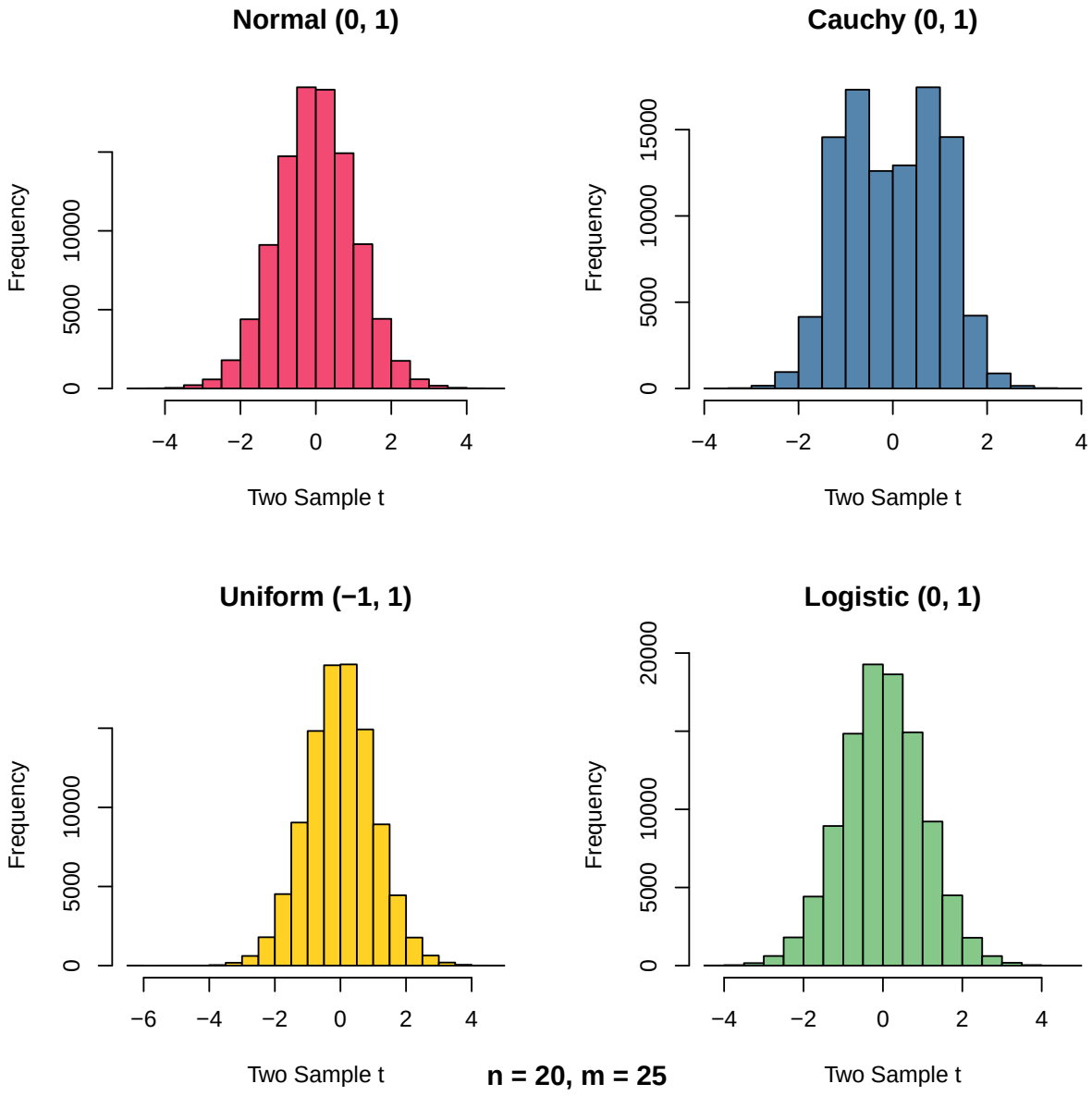
Remark 3.3. Reject H_0 in favour of H_1 for large values of t.

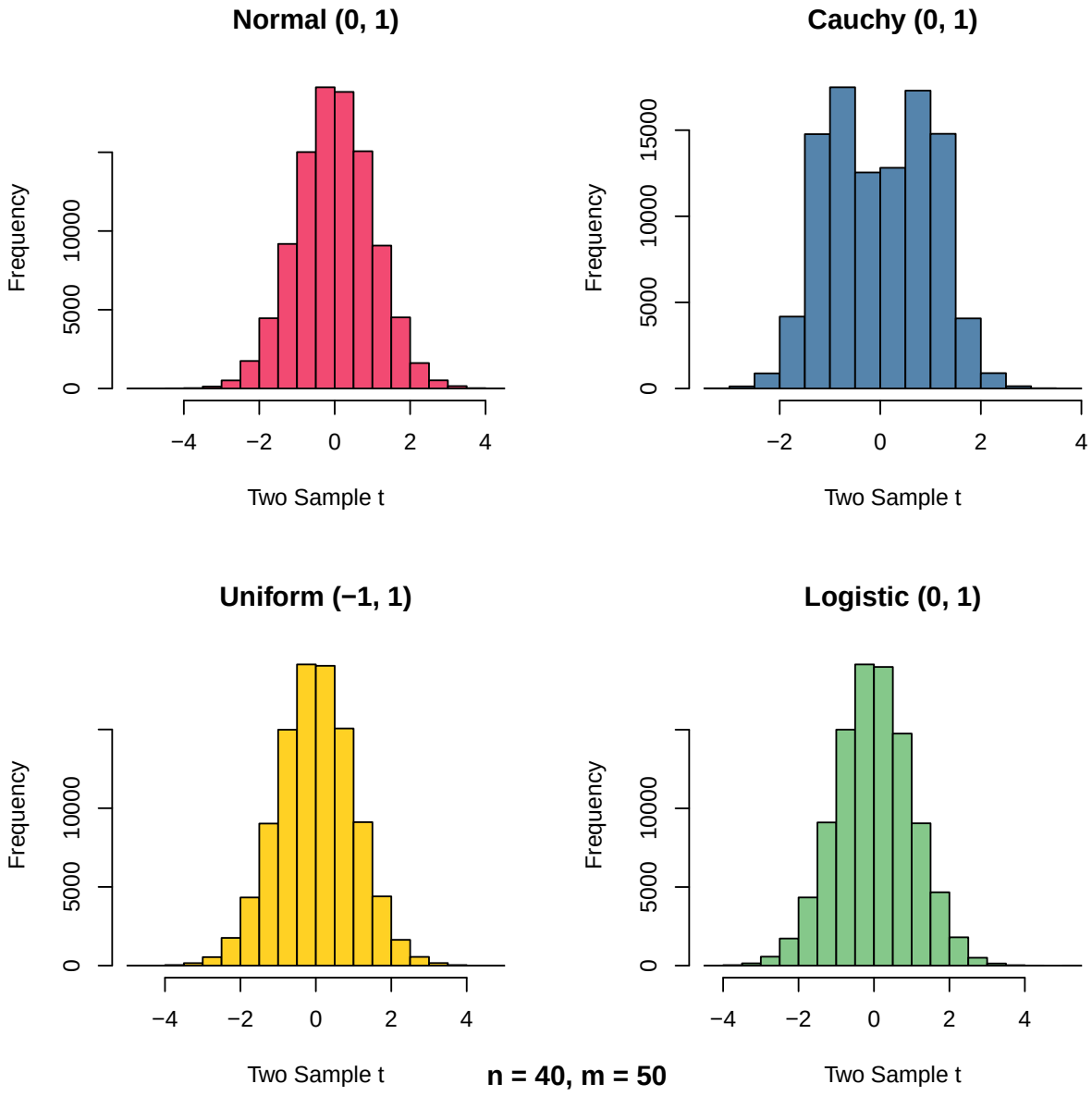
Remark 3.4. Reject H_0 in favour of H_2 for small values of t.

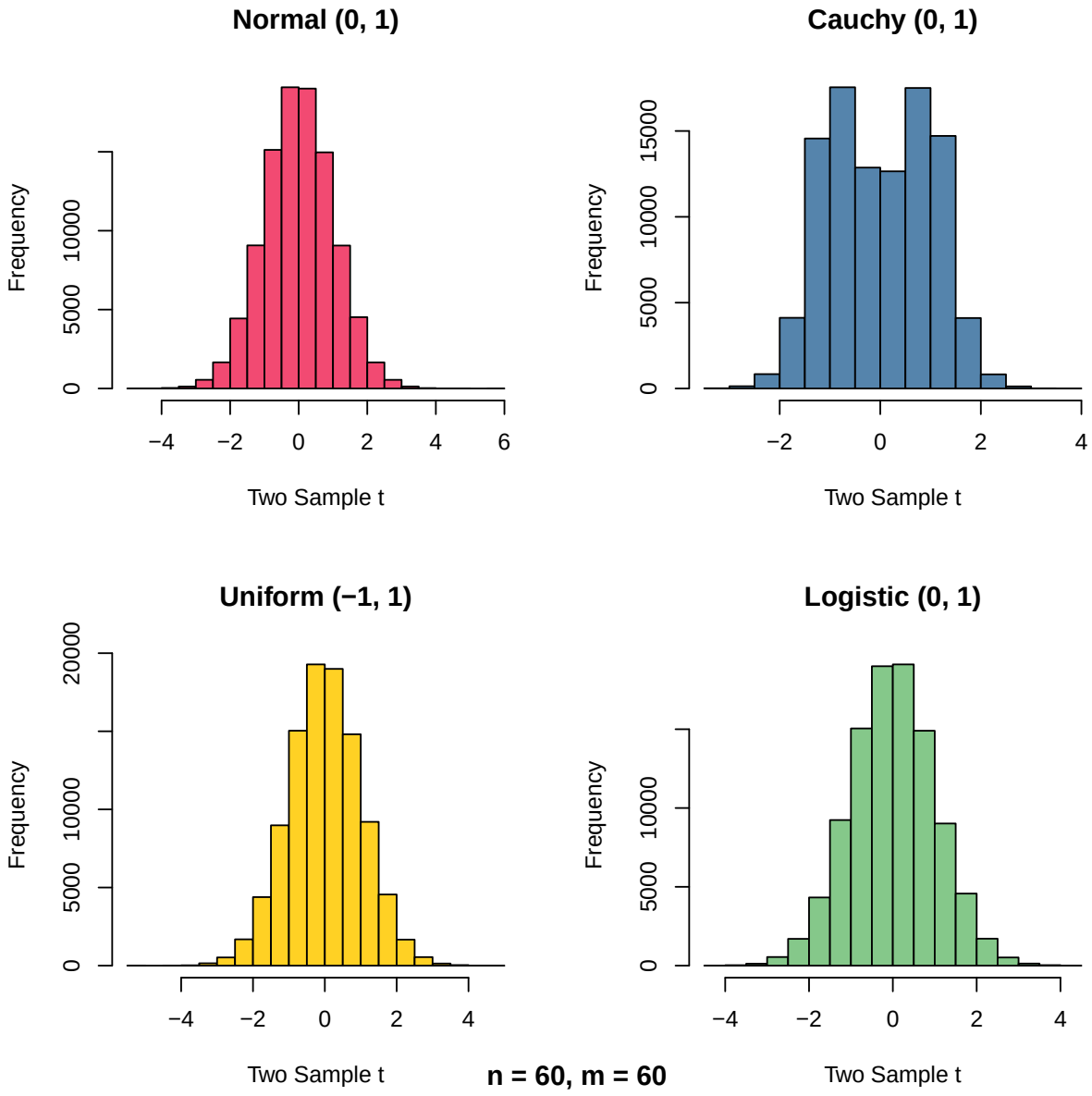
Remark 3.5. Reject H_0 in favour of H_3 for values of t that are too small or too large.

An obvious but interesting question that arises is that given the two sample problem, should we prefer *t test* over *Mann Whitney test* or the vice versa? To answer that, let us first look at the following figures:


Normal (0, 1)
Frequency
0
5000
10000
−4
−2
0
2
4
Two Sample t
Cauchy (0, 1)
Frequency
0
5000
10000
15000
−4
−2
0
2
4
Two Sample t
Uniform (−1, 1)
Frequency
0
10000
20000
30000
−5
0
5
Two Sample t
Logistic (0, 1)
Frequency
0
5000
10000
−6
−4
−2
0
2
4
6
Two Sample t
n = 10, m = 15


Normal (0, 1)
Frequency
0
5000
10000
−4
−2
0
2
4
Two Sample t
Cauchy (0, 1)
Frequency
0
5000
10000
15000
−4
−2
0
2
4
Two Sample t
Uniform (−1, 1)
Frequency
0
5000
10000
−6
−4
−2
0
2
4
Two Sample t
Logistic (0, 1)
Frequency
0
5000
10000
20000
−4
−2
0
2
4
Two Sample t
n = 20, m = 25


Normal (0, 1)
Frequency
0
5000
10000
−4
−2
0
2
4
Two Sample t
Cauchy (0, 1)
Frequency
0
5000
10000
15000
−2
0
2
4
Two Sample t
Uniform (−1, 1)
Frequency
0
5000
10000
−4
−2
0
2
4
Two Sample t
Logistic (0, 1)
Frequency
0
5000
10000
−4
−2
0
2
4
Two Sample t
n = 40, m = 50

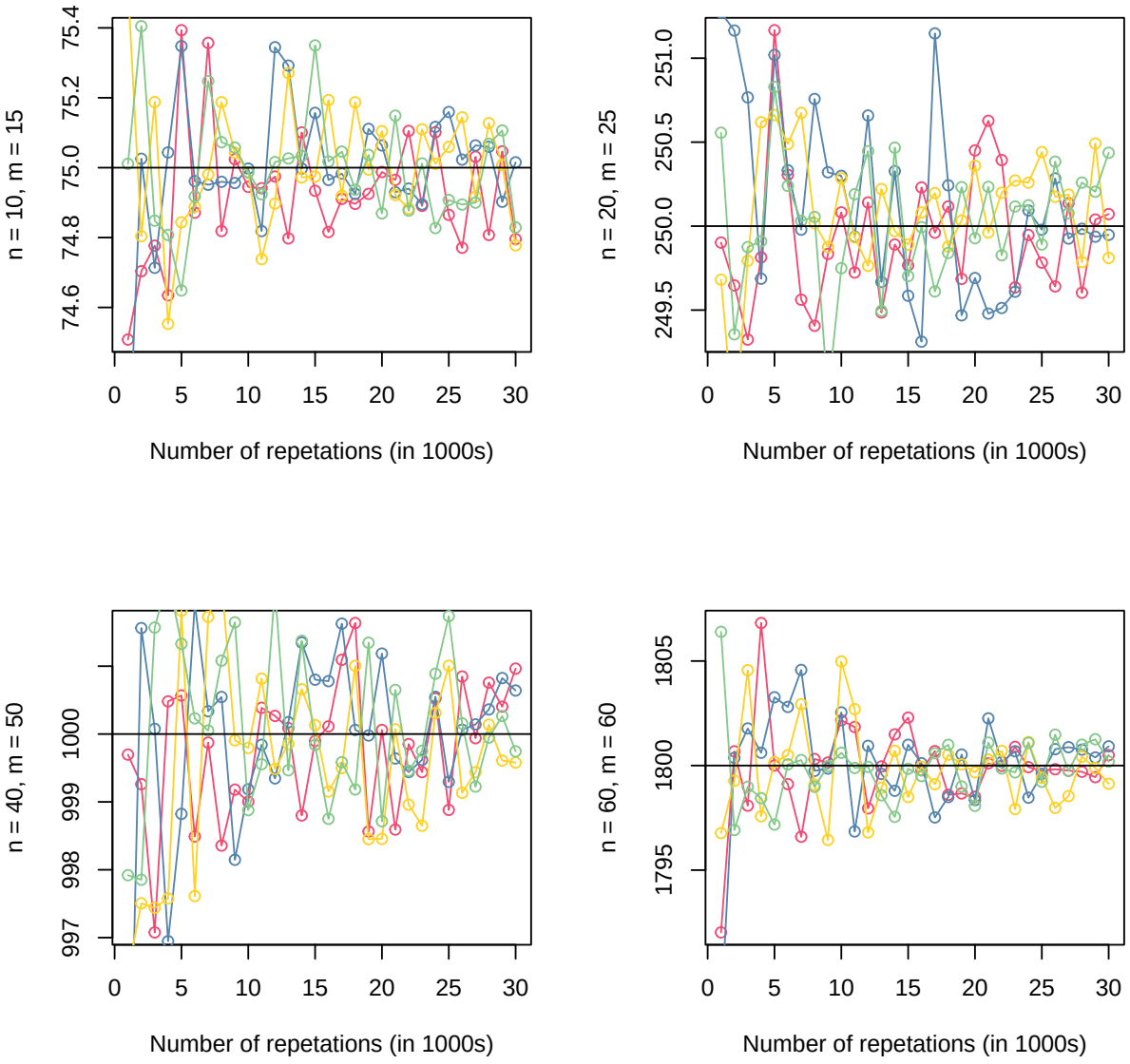


Observe that the curves depend on the underlying distribution, but we have noted in remark 3.2 that under the null, the two sample t statistic follows a t distribution. This indicates that something must be wrong - either the definition or the curves! Actually though, none is false. To understand what is actually happening, we need to delve deeper into the genesis of the distribution of t under the null. Actually, the central limit theorem plays a key role for the distribution to hold, which, in turn, depends on some regularity conditions. One such condition is that of the existance of the first order raw moment of the underlying distribution. Now look at the figures again. The curves for Cauchy distribution are significantly different from the rest for each of the panels. Recall that for Cauchy distribution, first order (and thus any higher order) moment is non-existent. This explains the seemingly wild observation. This also answers the question posed in the preceeding paragraph; as Mann Whitney test is invariant of the underlying distribution, it should always be preferred over

two sample t statistic whenever nothing concrete can be said about the data generating mechanism.

3.2 Expectation of U under null

Consider the following figures:[1]



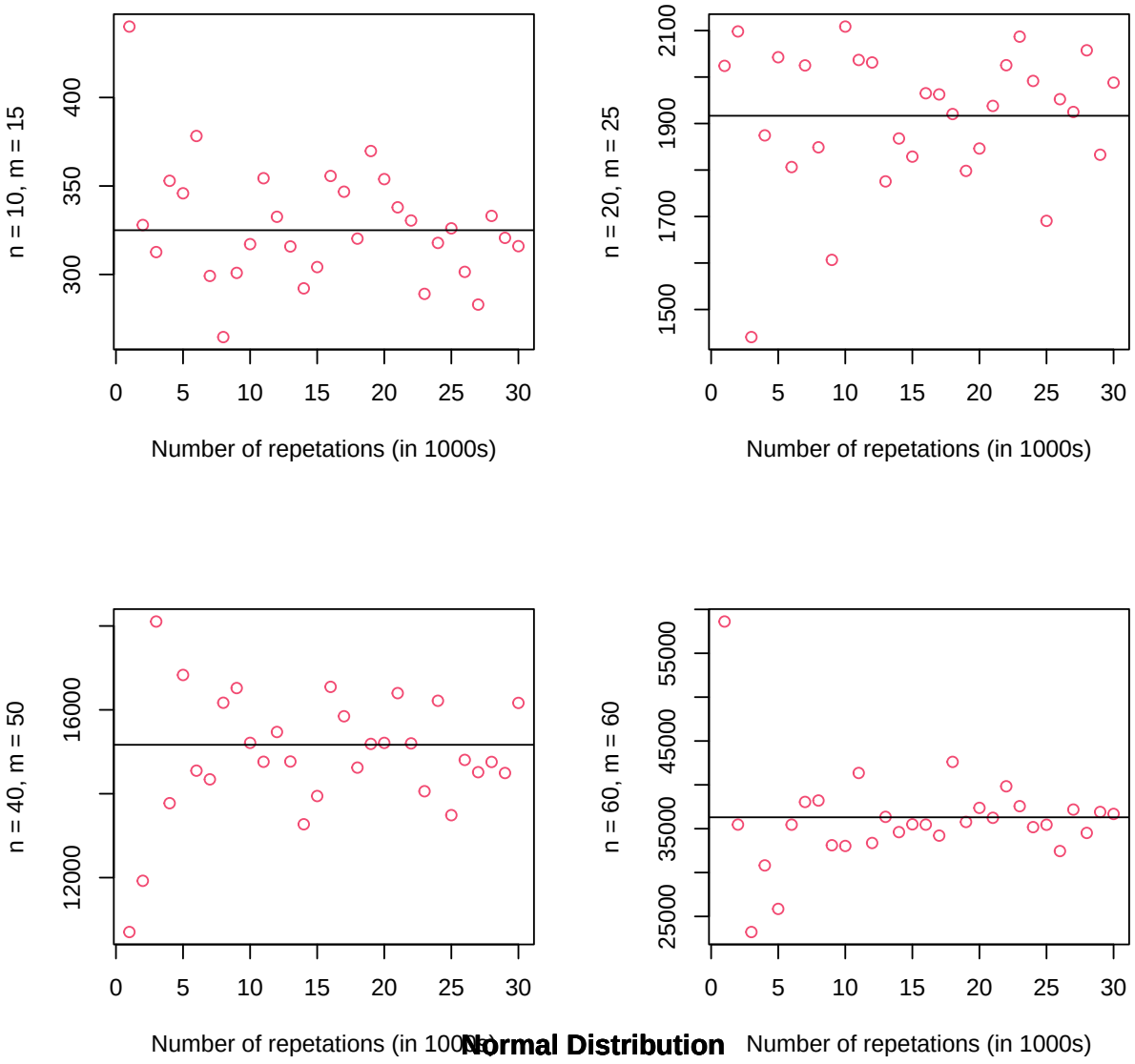
Remark 3.6. For a fixed pair (n, m), as number of repetations increase, the sample mean tends to the population mean under the null.

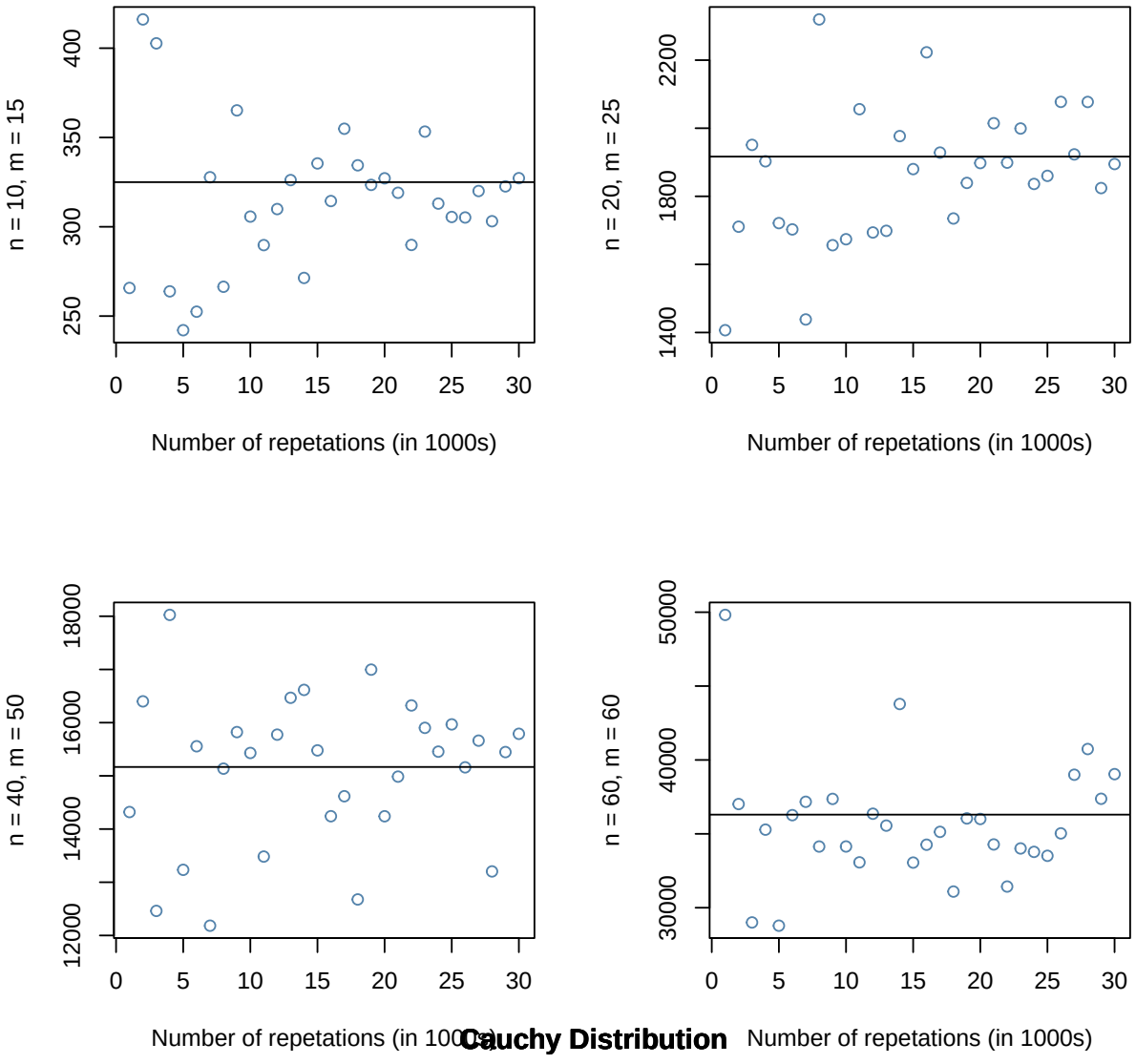
Remark 3.7. As both n and m increases, the sample mean tends to the population mean under the null faster in the sense that it tends to the desired for even smaller number of repetations.

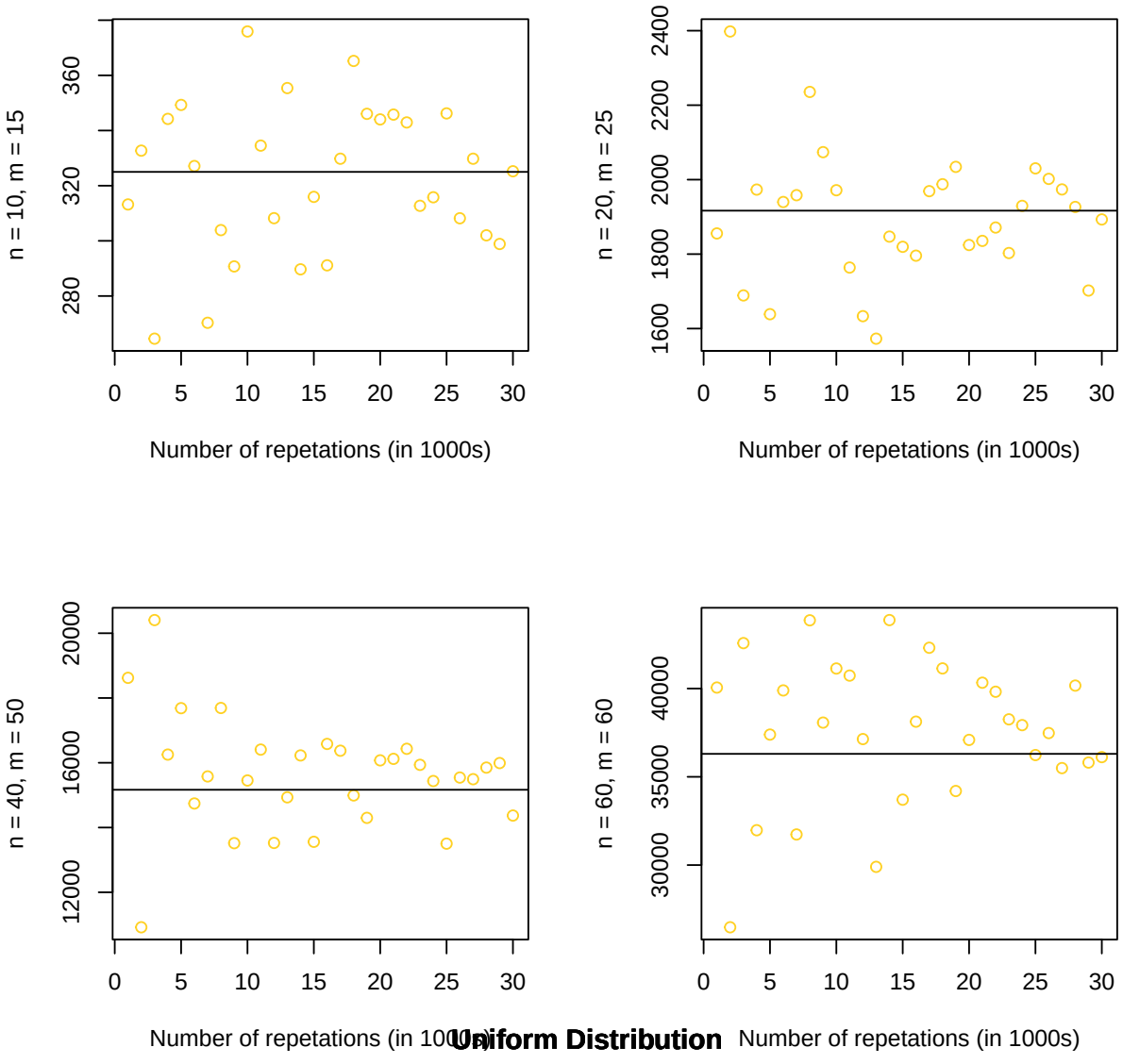
[1] Red: Normal, Blue: Cauchy, Yellow: Uniform, Green: Logistic

3.3 Variance of U under null

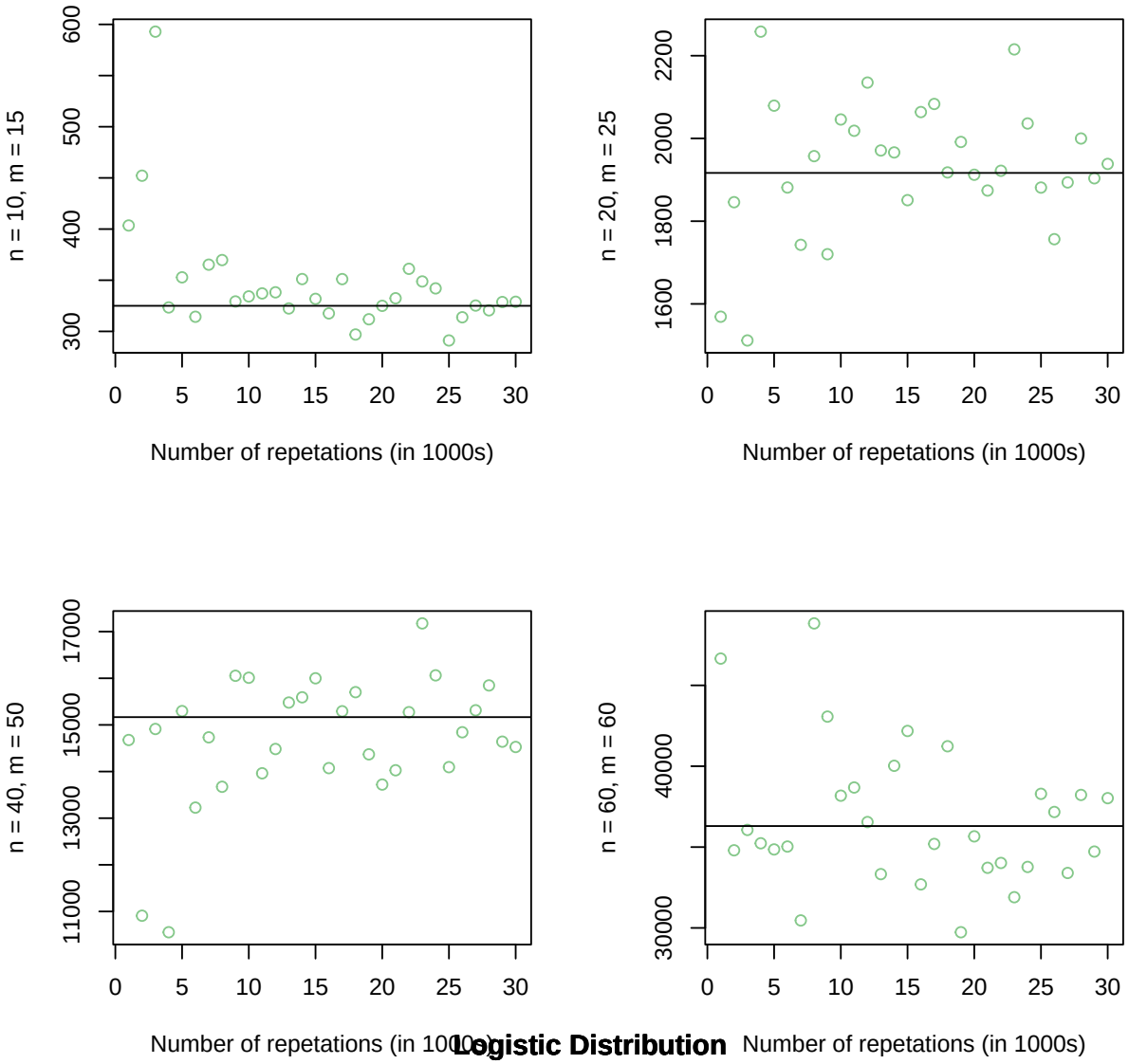
Consider the following figures:




n = 10, m = 15
250
300
350
400
0
5
10
15
20
25
30
Number of repetations (in 1000s)
n = 20, m = 25
1400
1800
2200
Number of repetations (in 1000s)
n = 40, m = 50
12000
14000
16000
18000
Number of repetations (in 1000s)
n = 60, m = 60
30000
40000
50000
Number of repetations (in 1000s)
Cauchy Distribution



Uniform Distribution



Remark 3.8. For a fixed pair (n, m), as number of repetations increase, the sample variance tends to the population variance under the null.

Remark 3.9. As both n and m increases, the sample variance tends to the population variance under the null faster in the sense that it tends to the desired for even smaller number of repetations.

4 Results on Size and Power

We have already seen that Mann Whitney test is indeed a good test in terms of its behaviour under the null hypothesis. But there are still things that we need to observe for concluding anything strong about the use of the statistic. For example, we need to observe the size of the test for different underlying distributions, as well as different values of the parameter. Observing how the power curve changes in terms of the underlying distribution and parameter values is also important. In this section, we do this, as well as compare the obtained results to those corresponding to the classical two sample t test.

4.1 Empirical Size

Definition 4.1. *(Empirical Size)* For a given test with rejection region S; under H_0, empirical size s_E is defined to be:

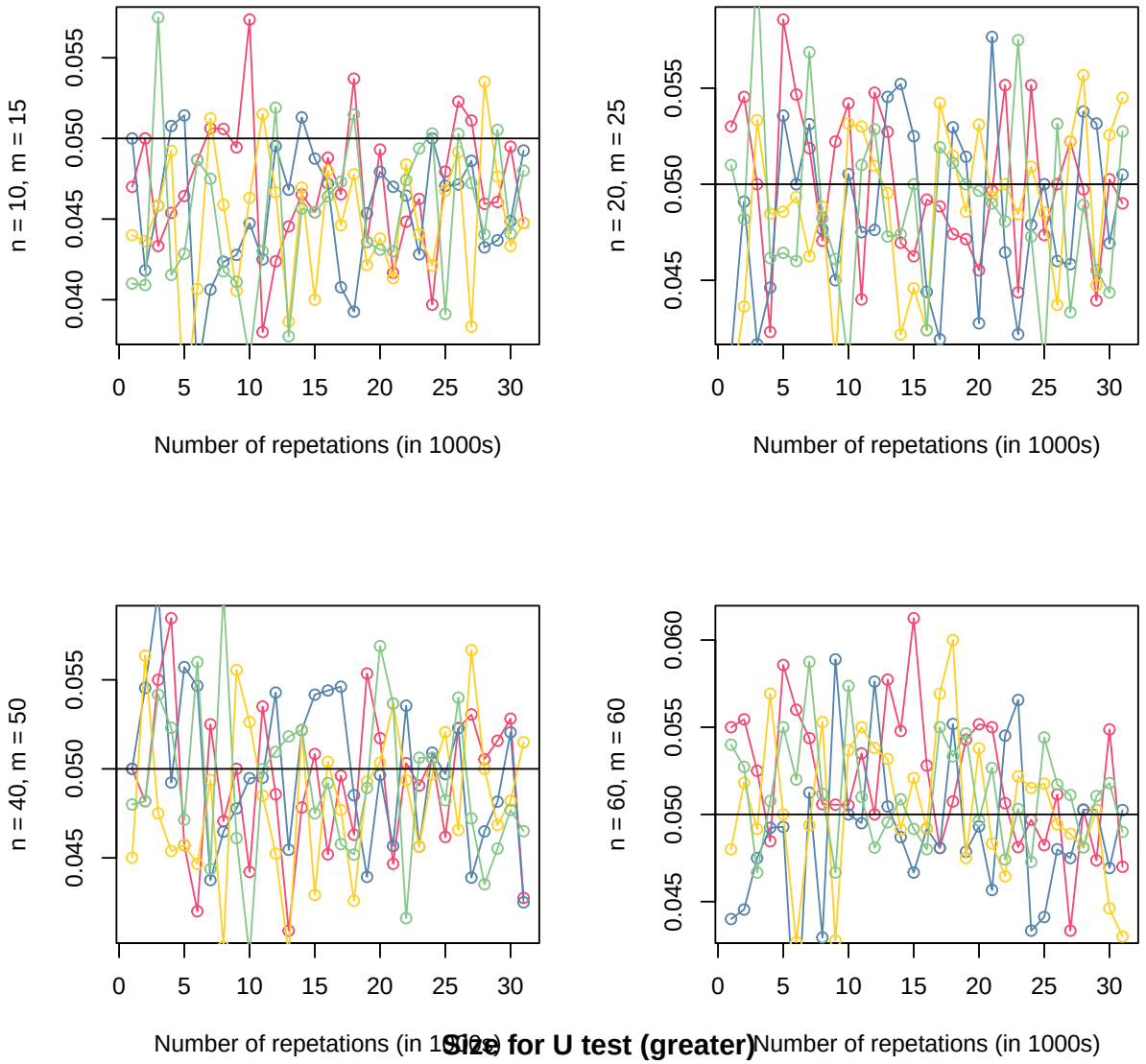
$$s_E := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in S), \tag{12}$$

where, $X_1,\ X_2,\ ..., X_n$ is a drawn random sample from the distribution of the test statistic.

We use definition 4.1 for exploring the properties of size of the Mann Whitney test.

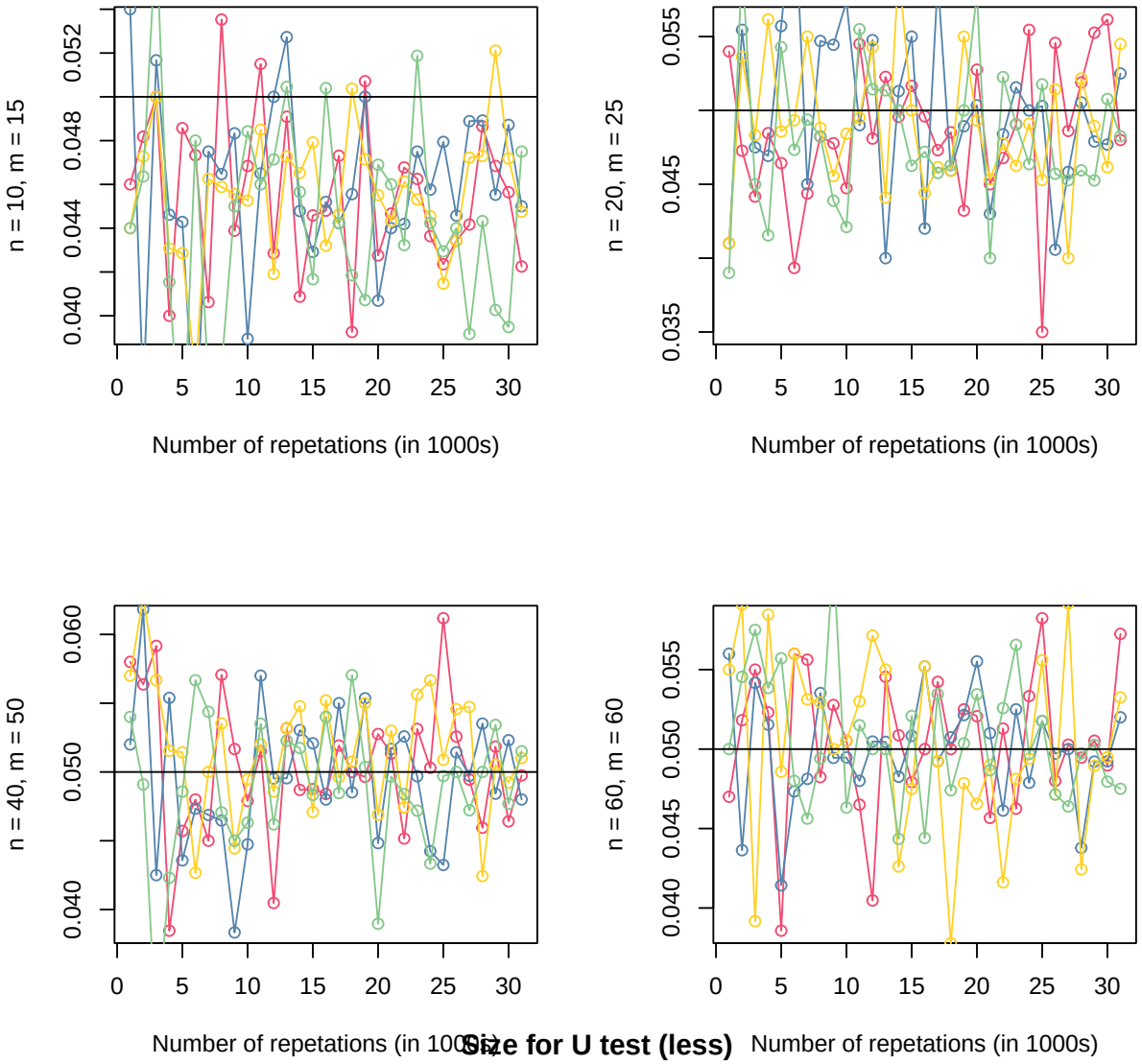
Let us first consider the alternative H_1.

```
## Warning in wilcox.test.default(x, y, type):  cannot compute exact p-value with ties
```

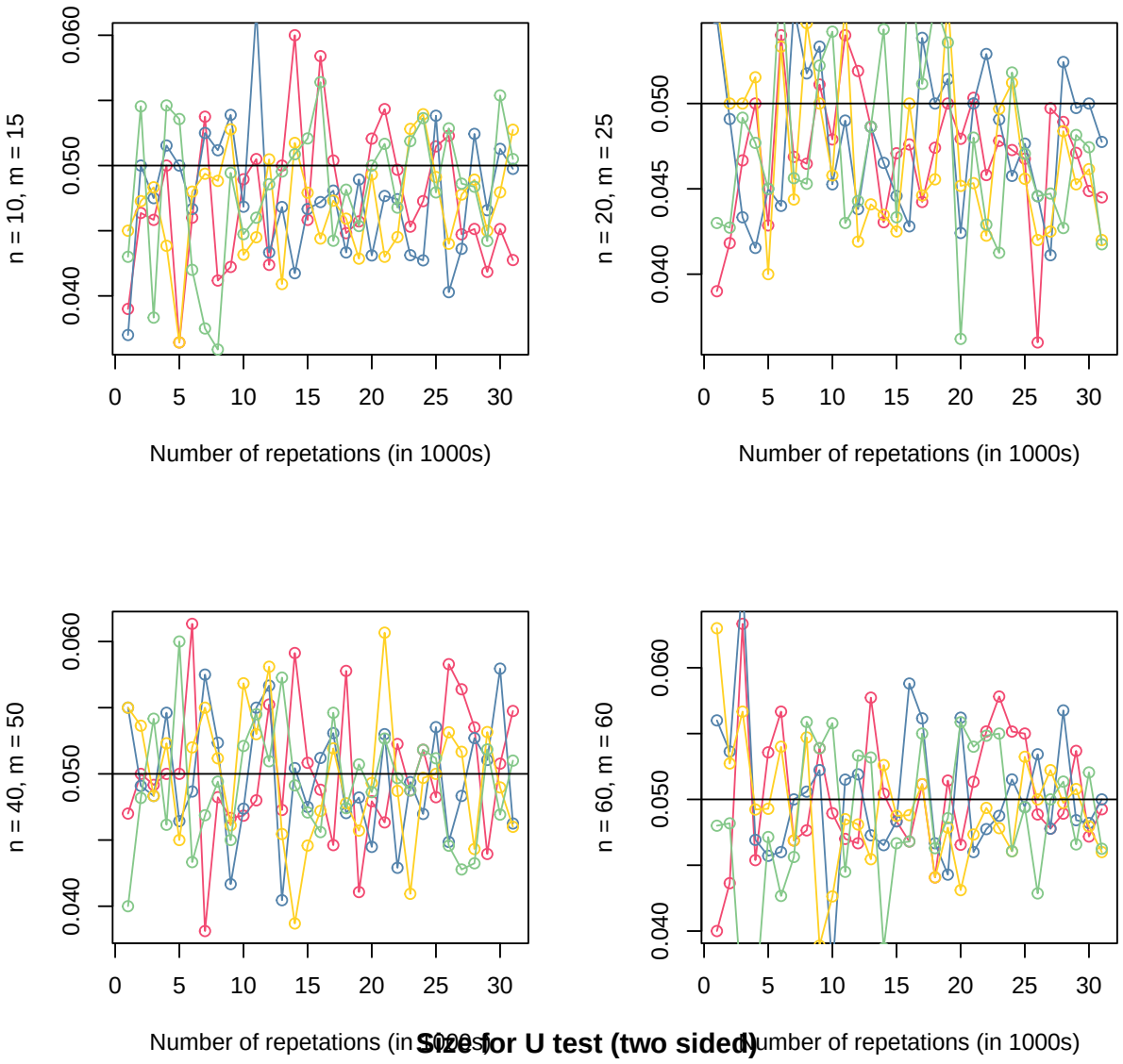


Remark 4.1. Observe that irrespective of the underlying distribution, we are able to obtain level α tests for each (n, m). Also, as the number of repetations increase, the empirical size gets closer to the given α.

Now consider the other two alternatives respectively.



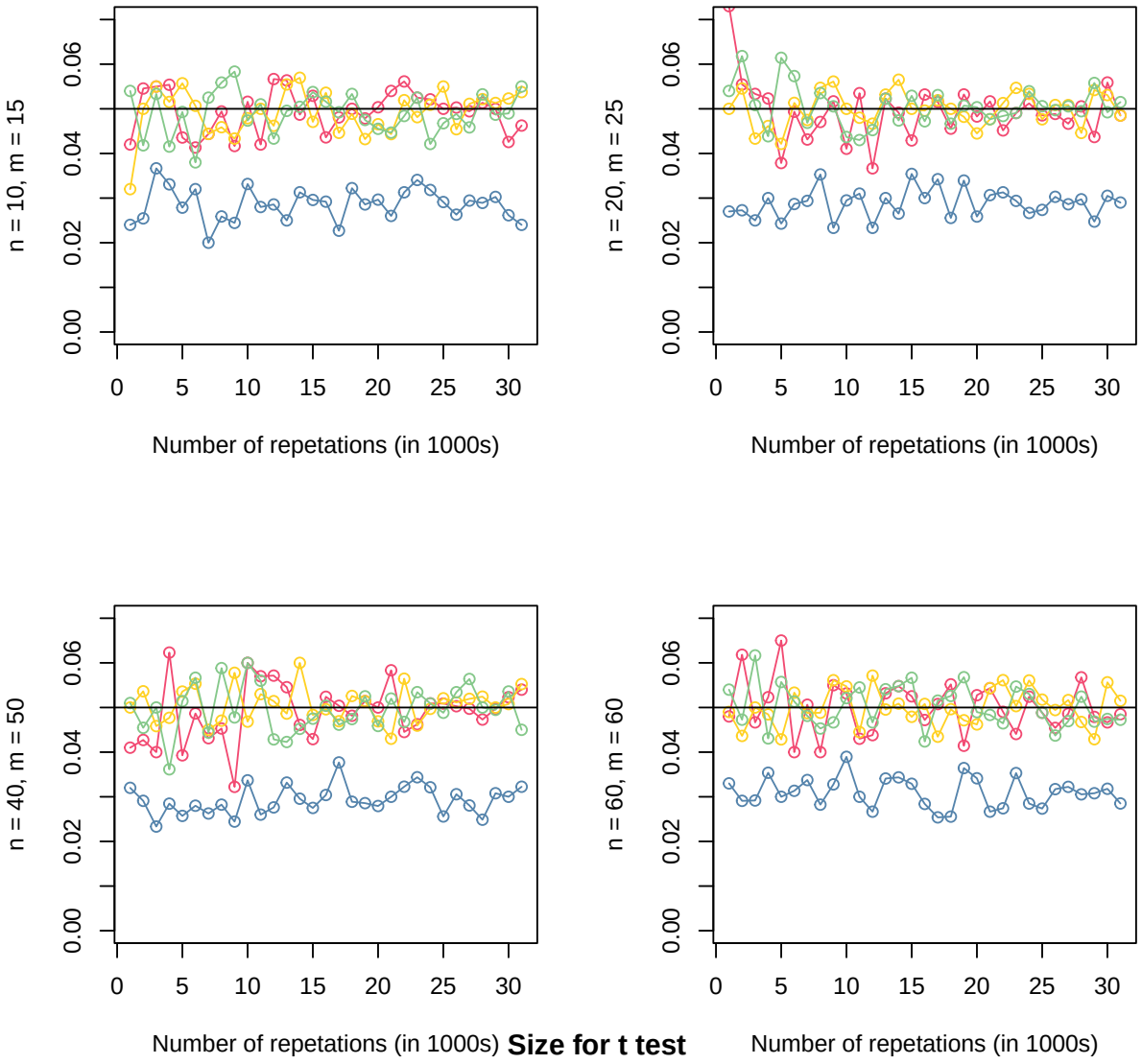
Size for U test (less)



Remark 4.2. Remark 4.1 holds true for the other two alternatives too.

A comparison with the parametric counterpart

Let us take a look at the size curves for the two sample t test: Consider the alternative H_1 only.



Remark 4.3. For all the distributions considered sans Cauchy, as repetations increase, empirical size goes to the desired level.

Remark 4.4. For Cauchy distribution, the obtained empirical sizes are well below the desired level. This validates the fact that the *two sample t statistic* is not distribution free. Therefore, for unknown data generating mechanisms, Mann Whitney statistic (and hence the test) is more well suited than its parametric counterpart.

4.2 Empirical Power

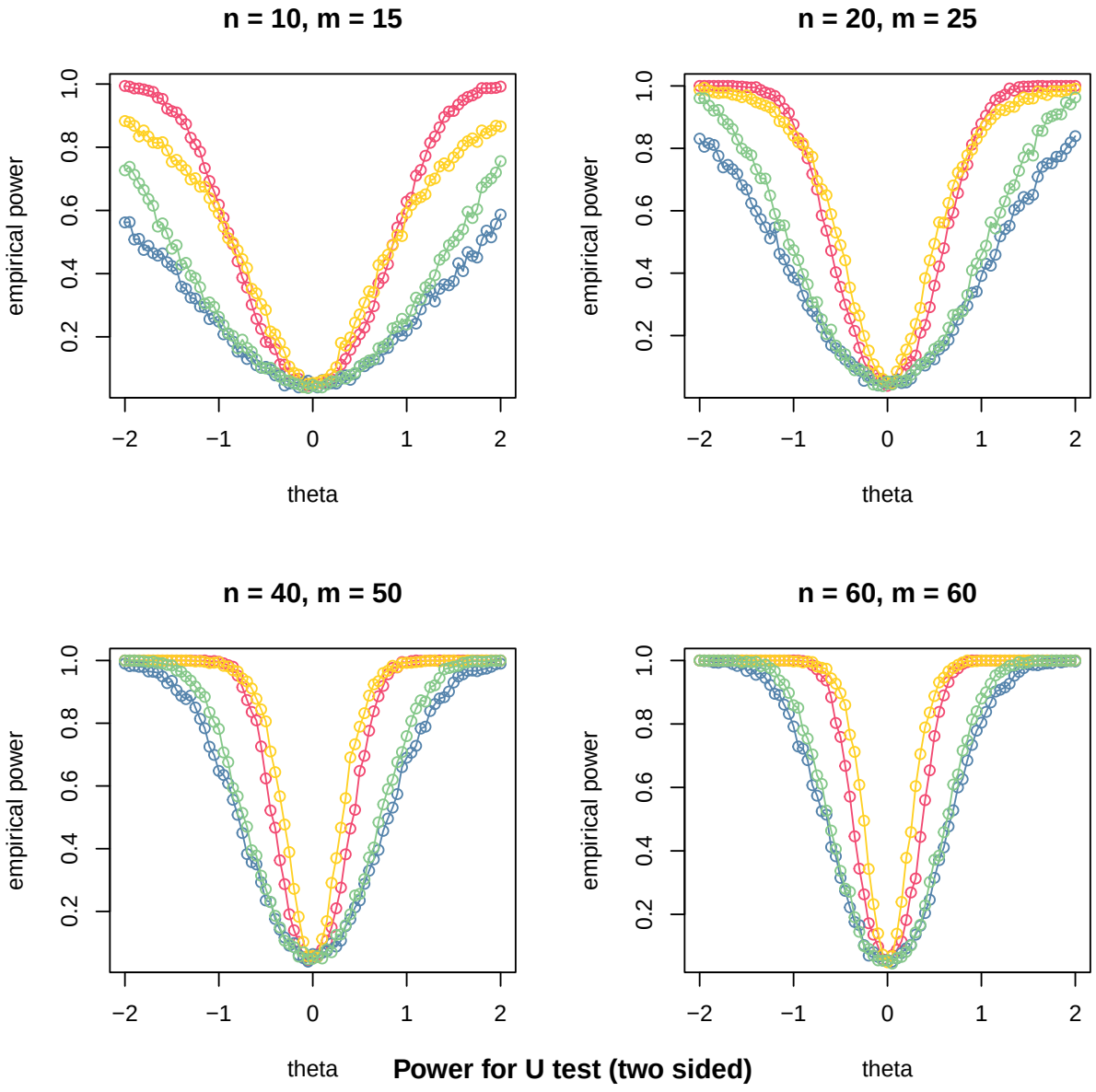
Definition 4.2. *(Empirical Power)* For a given test with rejection region S; under H_1, empirical power β_E is defined to be:

$$\beta_E := \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \in S), \tag{13}$$

where, $X_1,\ X_2,\ ..., X_n$ is a drawn random sample from the distribution of the test statistic.

We use definition 4.2 for exploring the power curve of the Mann Whitney test.

Let us first consider the two sided alternative.



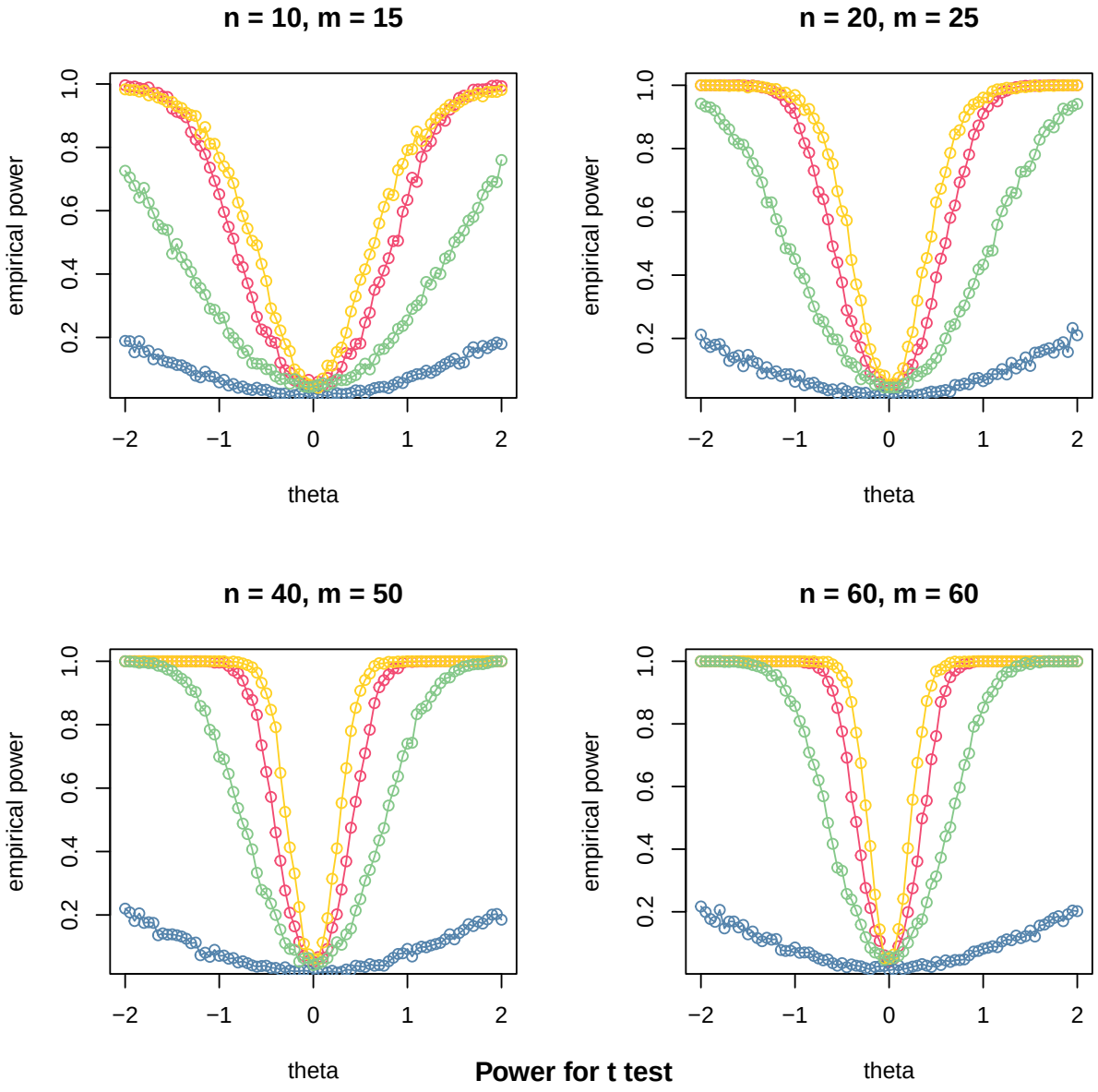
Remark 4.5. As (n, m) increases, the test becomes more powerful in even closer neighbourhoods of $\theta = 0$.

Remark 4.6. It is easy to observe that the power function is not distribution free. As a matter of fact, the figures indicate that the power function of normal and uniform performs significantly better than that corresponding to Cauchy or logistic.

Remark 4.7. Similarly we can consider the other two alternatives respectively.

A comparison with the parametric counterpart

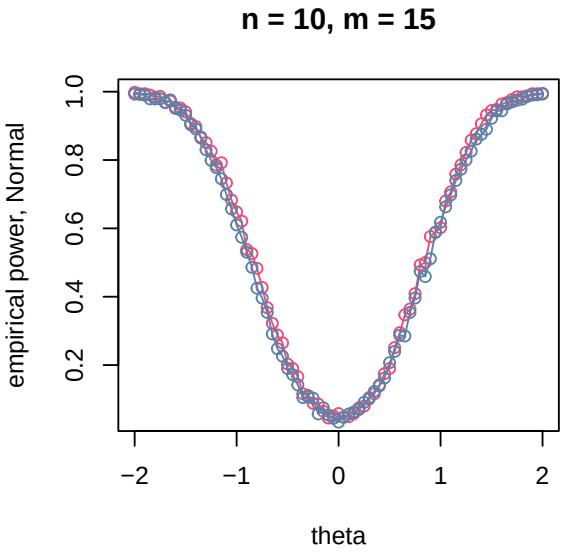
Let us take a look at the power curves for the two sample t test: Consider the two sided alternative only.

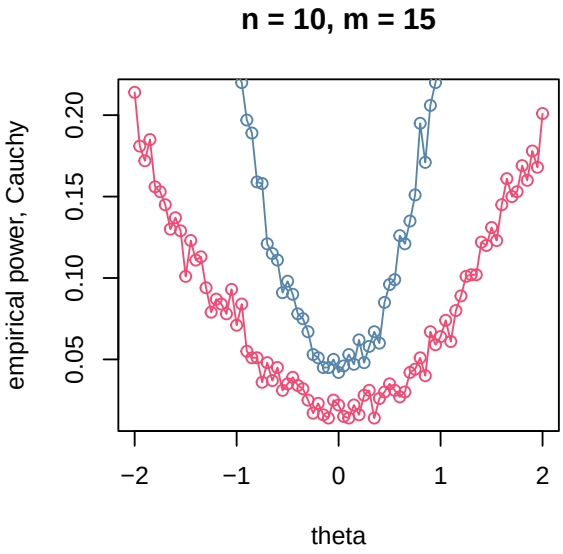


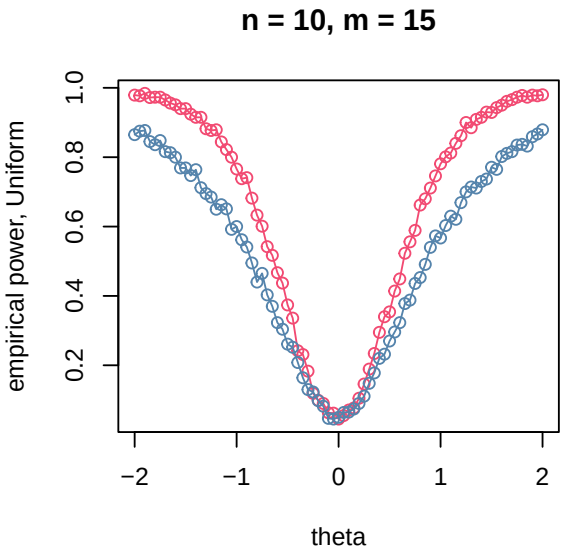
Remark 4.8. As (n, m) increases, the test becomes more powerful in even closer neighbourhoods of $\theta = 0$.

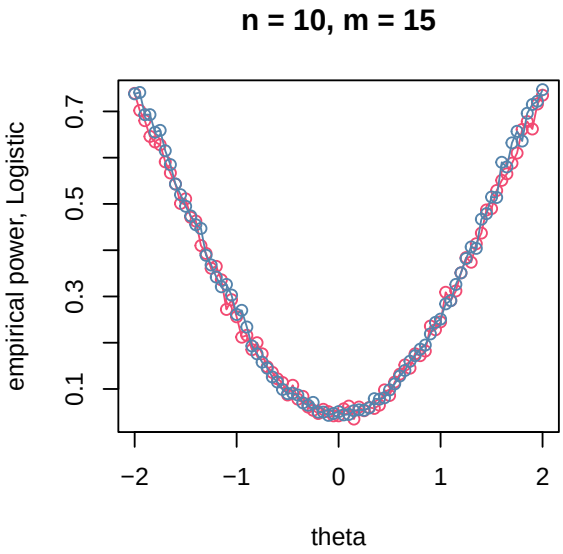
Remark 4.9. For Cauchy distribution, the obtained empirical power curve is very poor. This corresponds to our previous remarks about t test for Cauchy data generating mechanism.

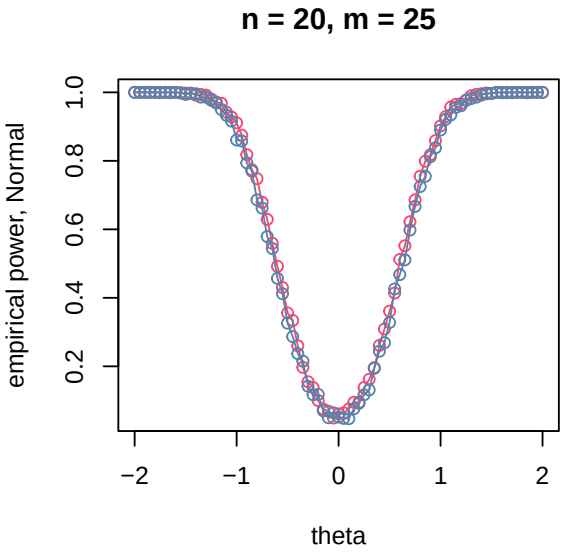
We now compare the power curves under the two sided alternative for t test and Mann Whitney test for each of the distributions separately.

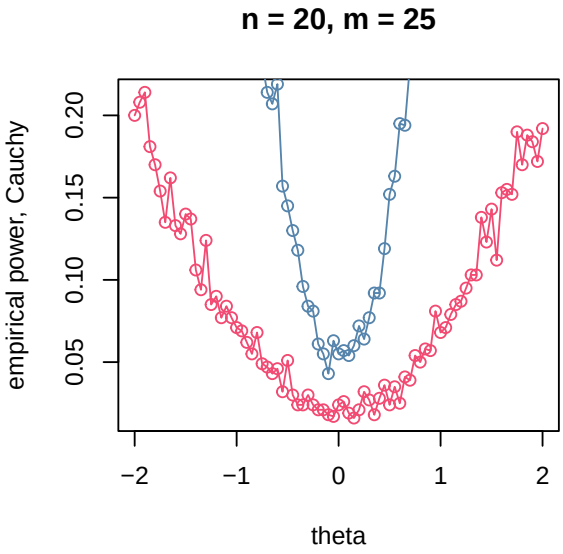

n = 10, m = 15
empirical power, Normal
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
theta

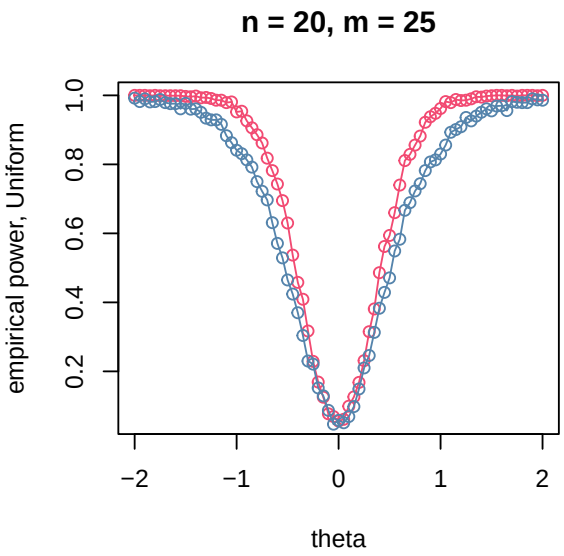

n = 10, m = 15
empirical power, Cauchy
0.05
0.10
0.15
0.20
−2
−1
0
1
2
theta

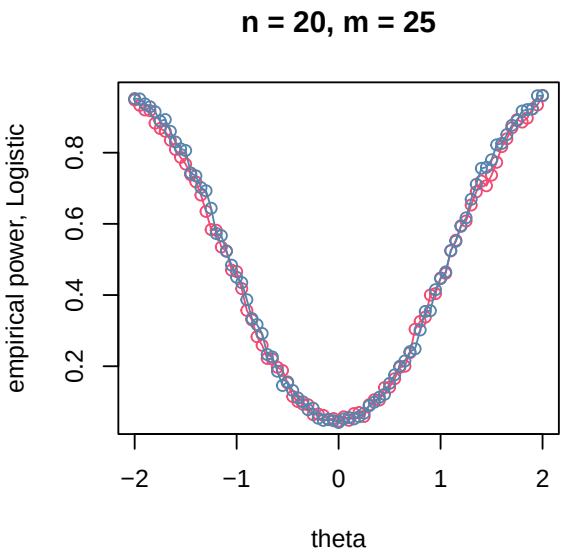

n = 10, m = 15
empirical power, Uniform
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
theta

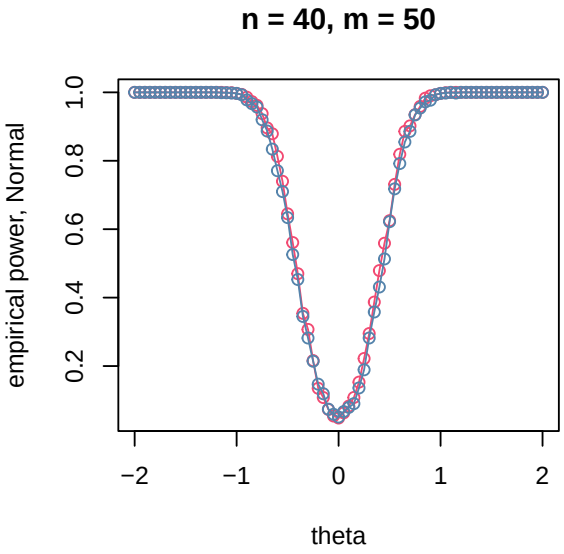

n = 10, m = 15
empirical power, Logistic
0.1
0.3
0.5
0.7
−2
−1
0
1
2
theta

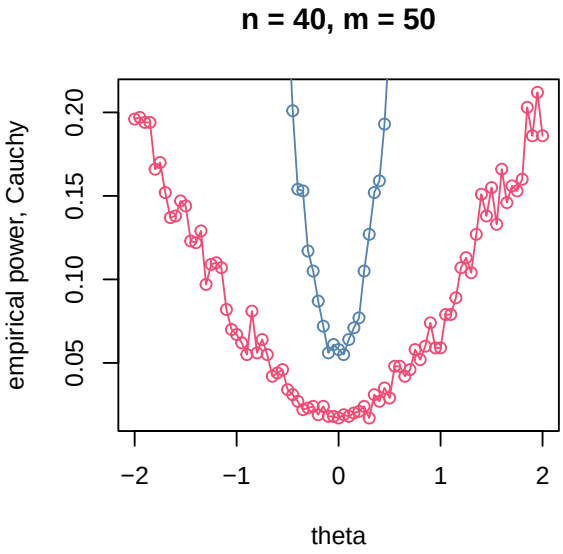

n = 20, m = 25
empirical power, Normal
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
theta

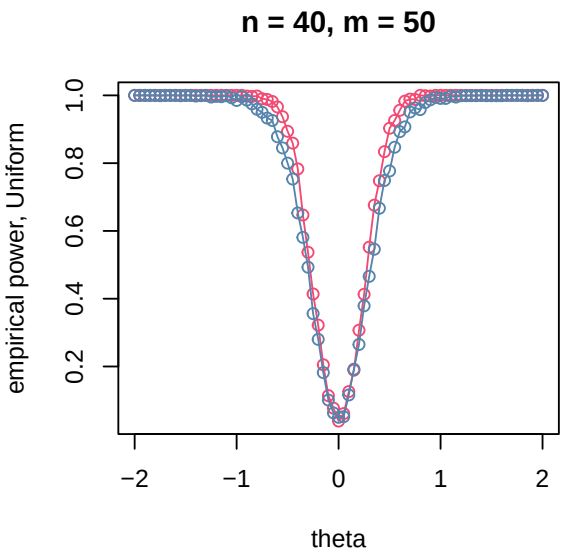

n = 20, m = 25
empirical power, Cauchy
0.05
0.10
0.15
0.20
−2
−1
0
1
2
theta

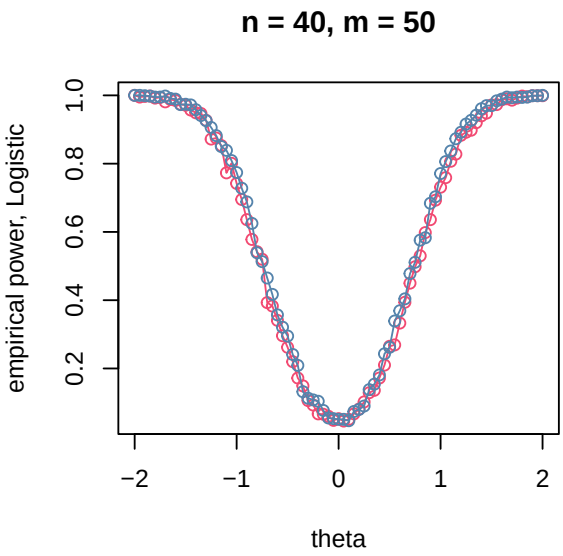

n = 20, m = 25
empirical power, Uniform
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
theta

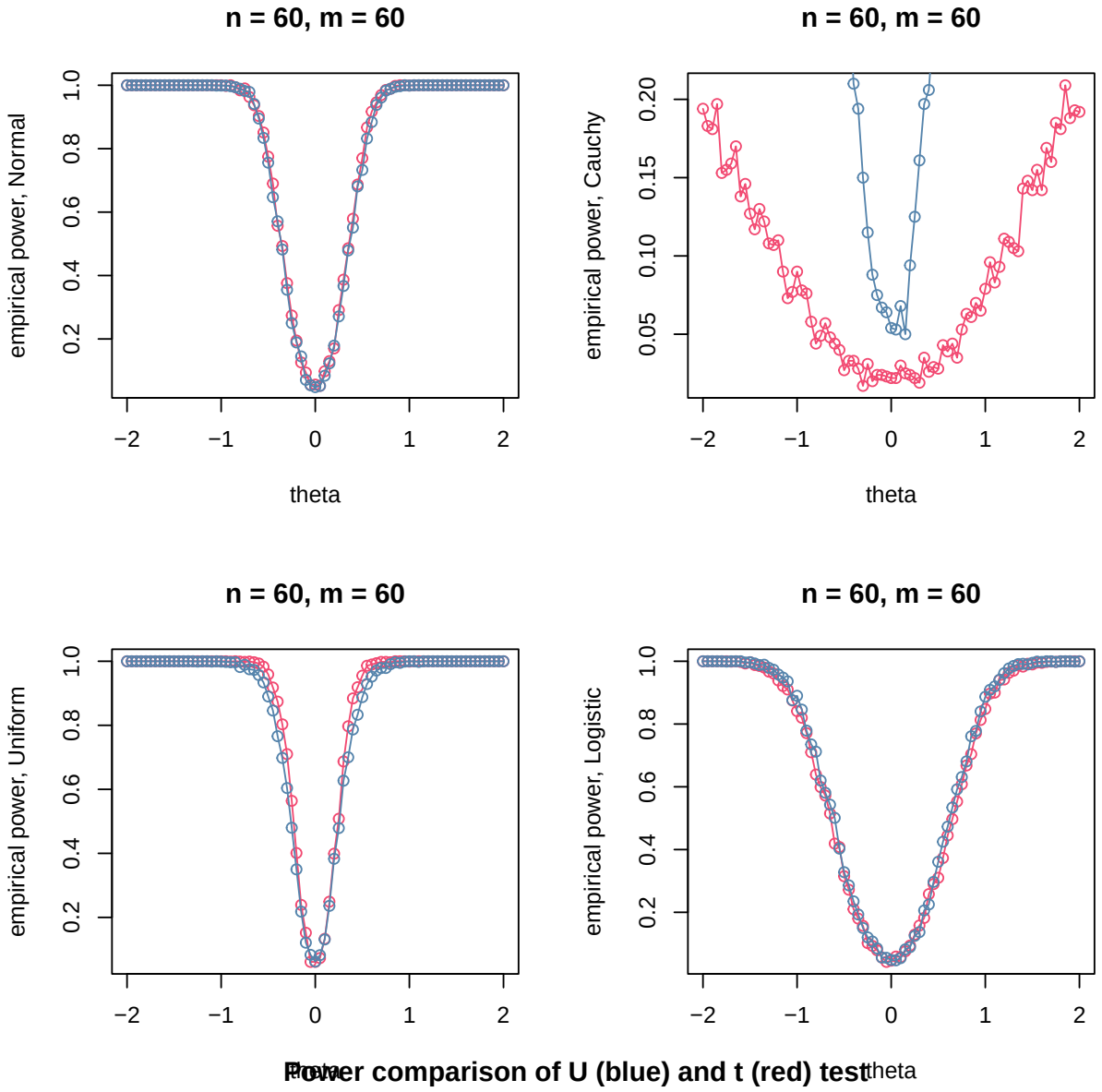

n = 20, m = 25
empirical power, Logistic
0.2
0.4
0.6
0.8
−2
−1
0
1
2
theta


n = 40, m = 50
empirical power, Normal
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
theta


n = 40, m = 50
empirical power, Cauchy
0.05
0.10
0.15
0.20
−2
−1
0
1
2
theta


n = 40, m = 50
empirical power, Uniform
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
theta


n = 40, m = 50
empirical power, Logistic
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
theta



Remark 4.10. For normal distribution, t test has more power than the U test. This is obvious, as t test is the most powerful test for a normal sample.

Remark 4.11. For a Cauchy sample, as noted earlier, U test is far better in terms of power curve.

Remark 4.12. For an uniform sample, t test is significantly better than U for small N, and marginally better for larger N, in terms of the power curve.

Remark 4.13. For logistic sample, both t and U tests exhibit more or less the same power curve.

5 Large Sample Properties

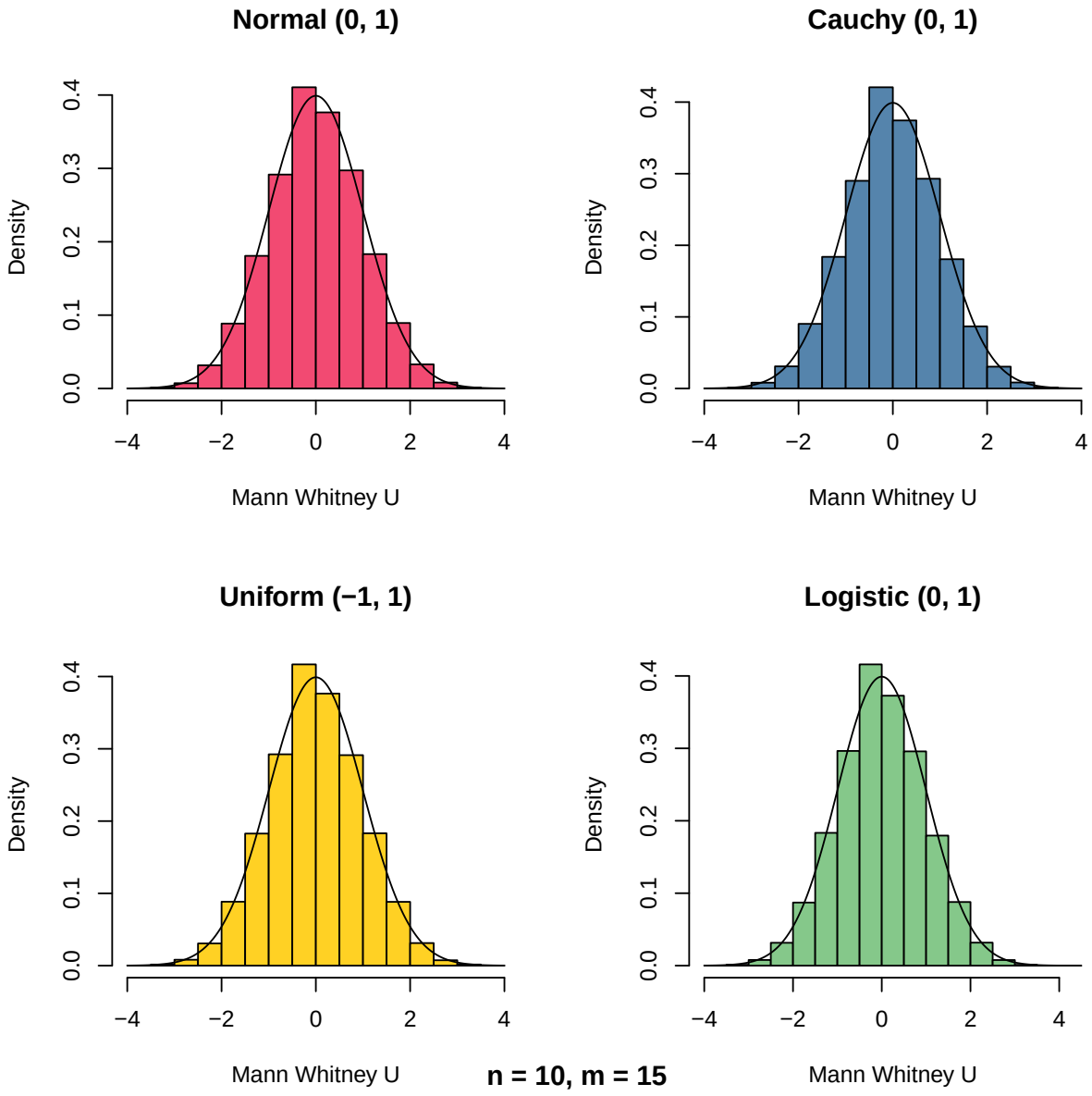
Theorem 5.1. For large N,

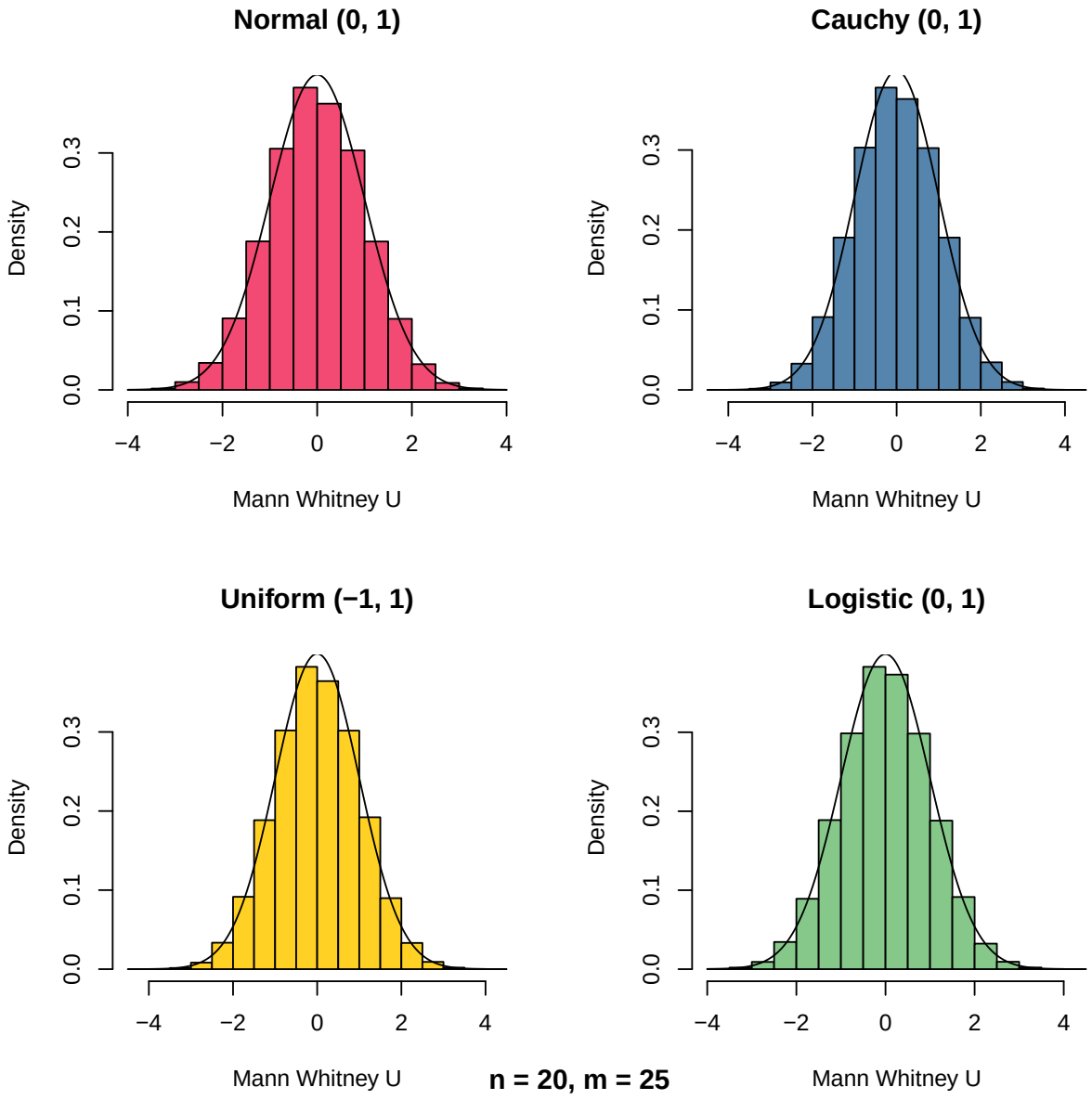
$$Z_U \coloneqq \frac{U - mn/2}{(mn(N+1)/12)^{1/2}} \overset{H_0}{\sim} N(0,1) \tag{14}$$

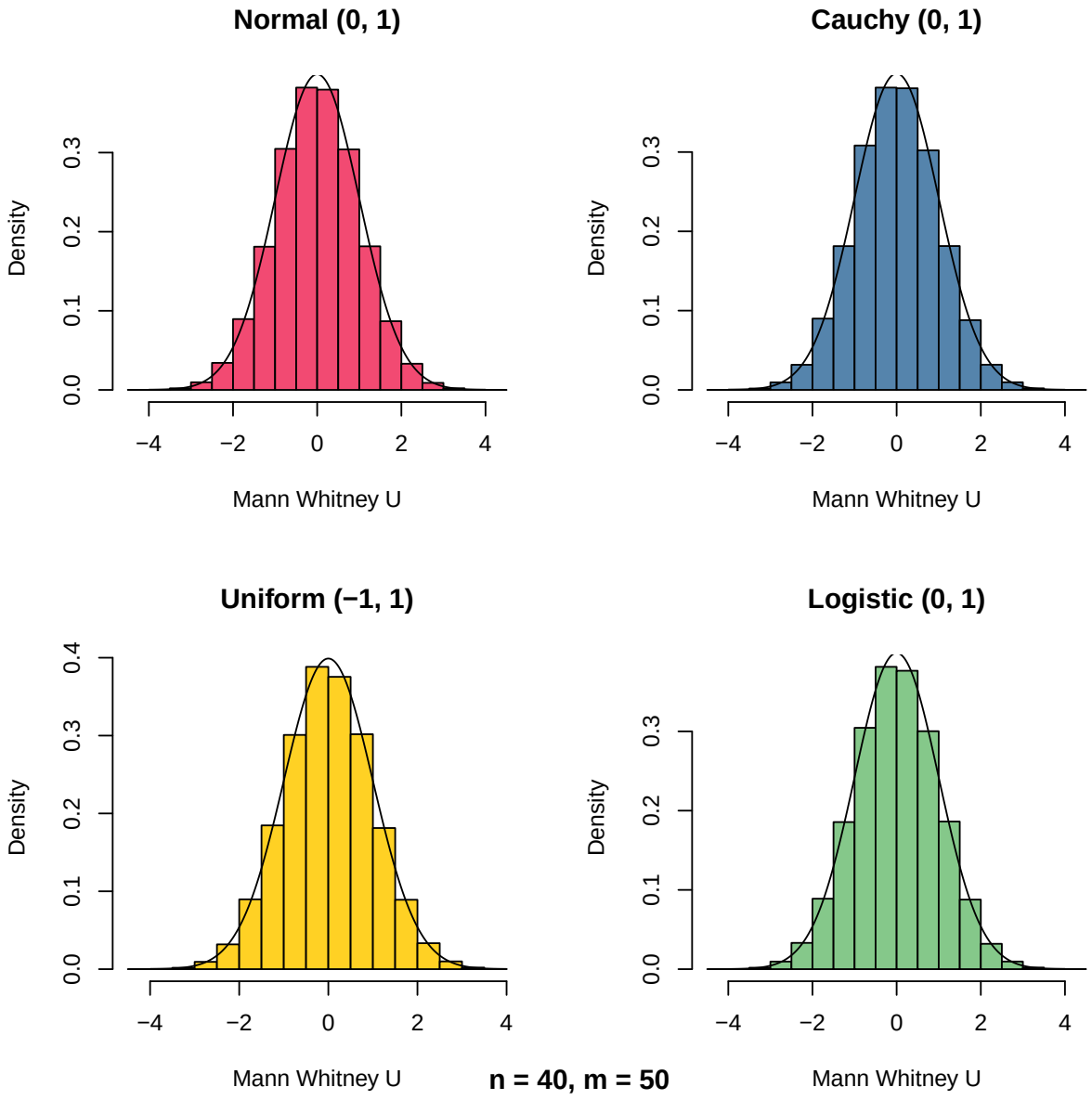
Theorem 5.2. For large N,

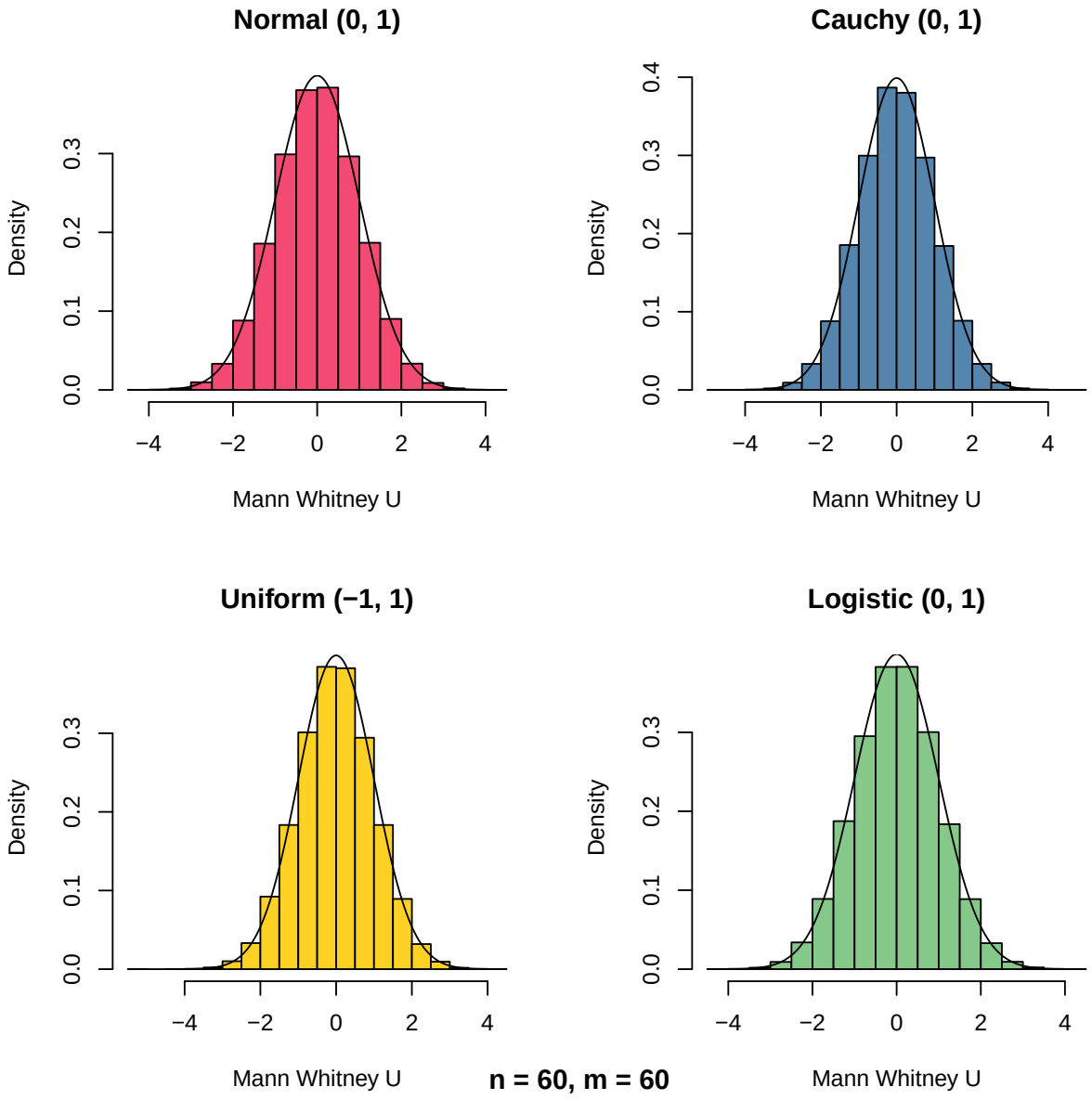
$$Z_t \coloneqq \frac{t}{((N-2)/(N-4))^{1/2}} \overset{H_0}{\sim} N(0,1) \tag{15}$$

We now try to figure out the attainment of the aforementioned theorems, for different underlying distributions, different values of N, etc.

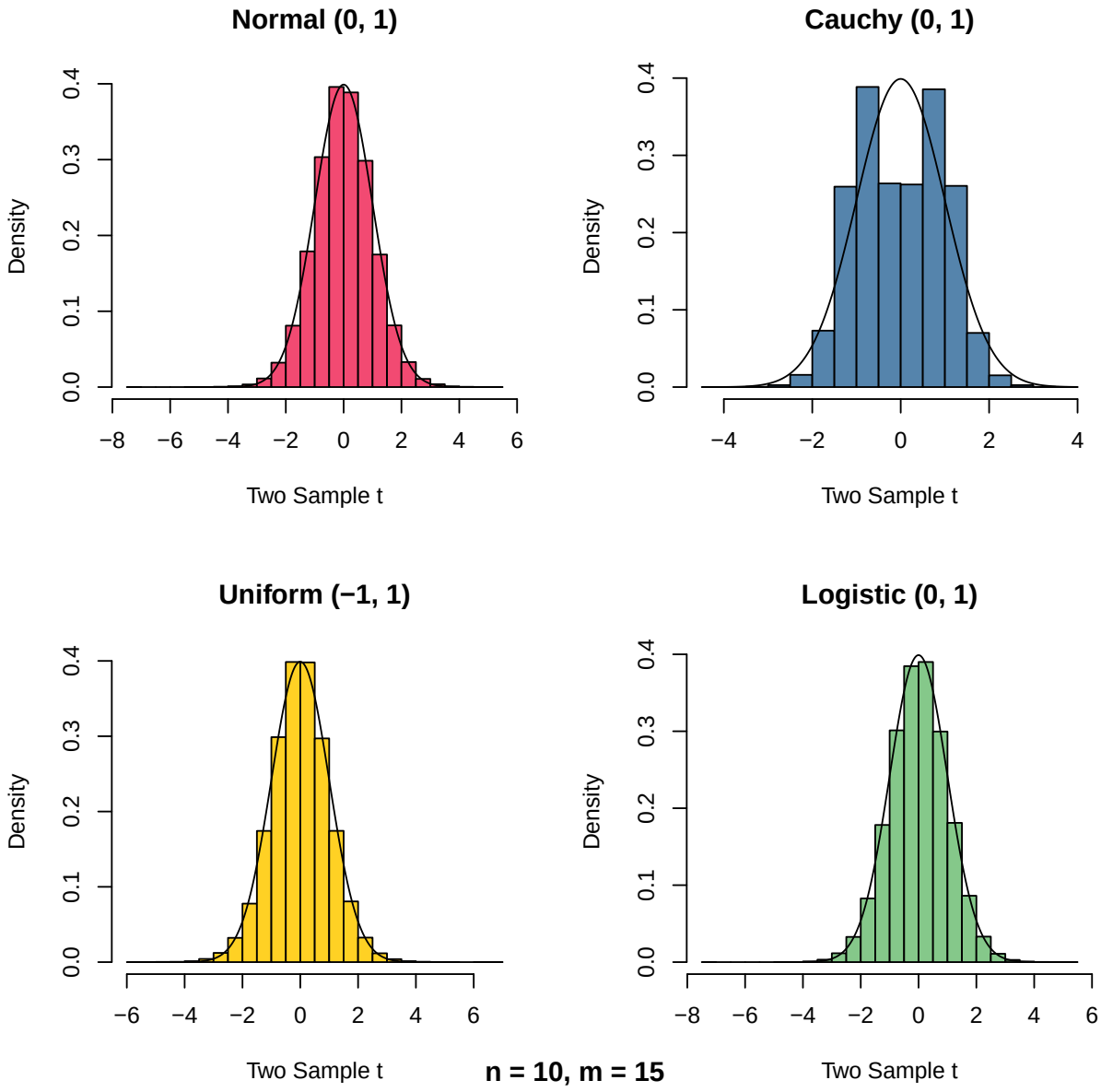

Normal (0, 1)
Cauchy (0, 1)
Uniform (−1, 1)
Logistic (0, 1)
Density
Mann Whitney U
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
n = 10, m = 15

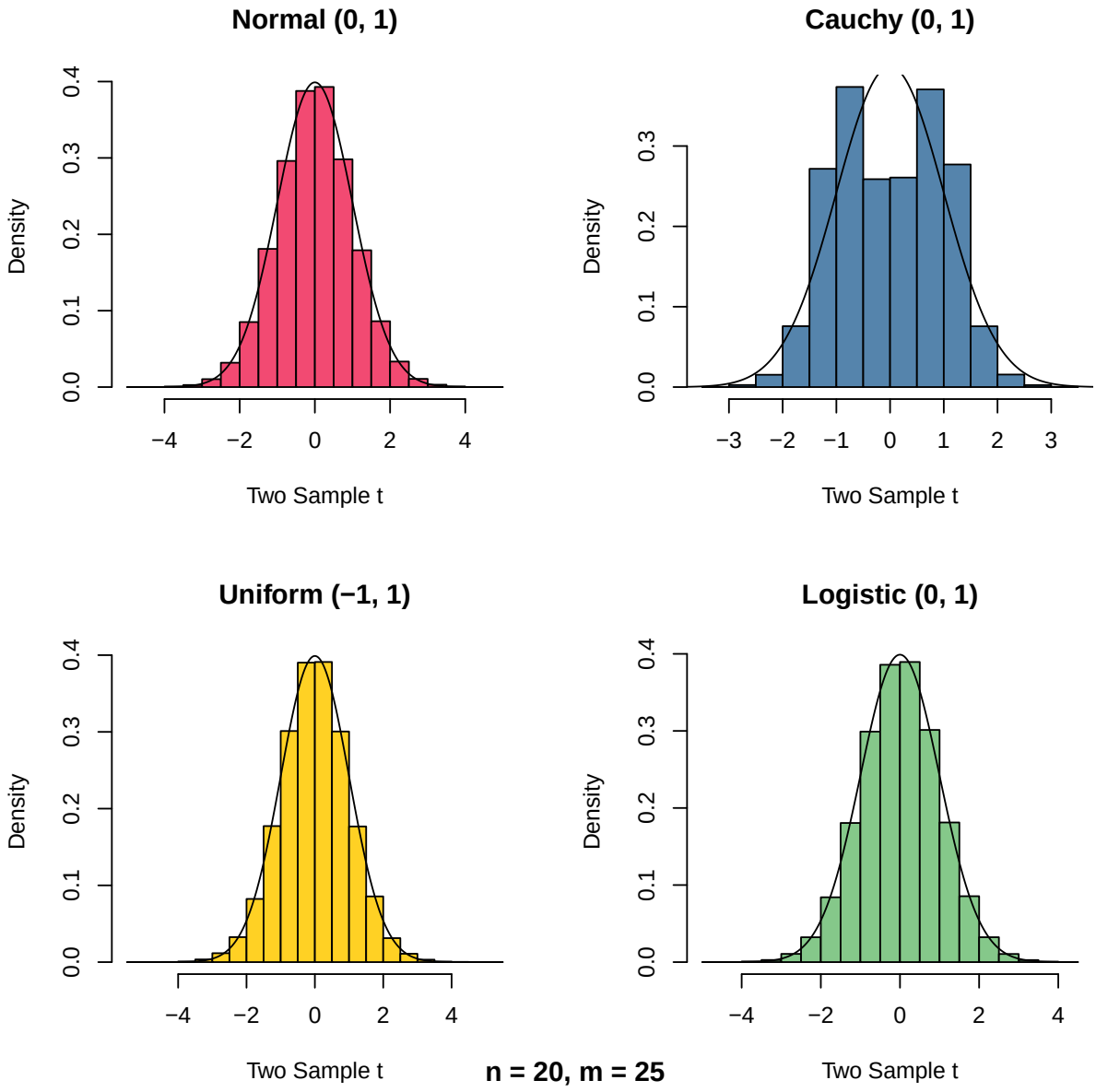

Normal (0, 1)
Cauchy (0, 1)
Uniform (−1, 1)
Logistic (0, 1)
Density
0.0
0.1
0.2
0.3
−4
−2
0
2
4
Mann Whitney U
n = 20, m = 25

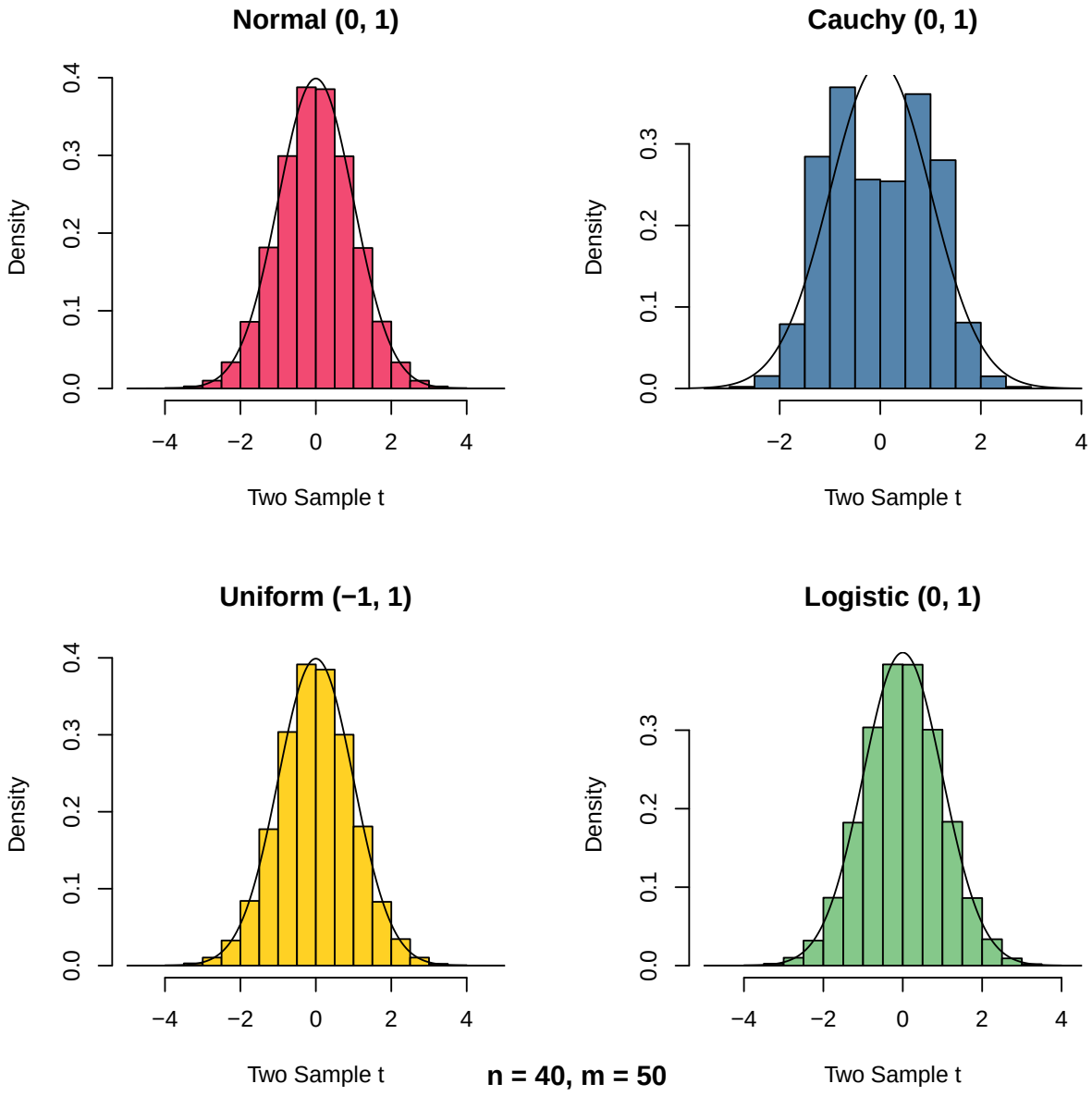

Normal (0, 1)
Cauchy (0, 1)
Uniform (−1, 1)
Logistic (0, 1)
Density
Mann Whitney U
n = 40, m = 50

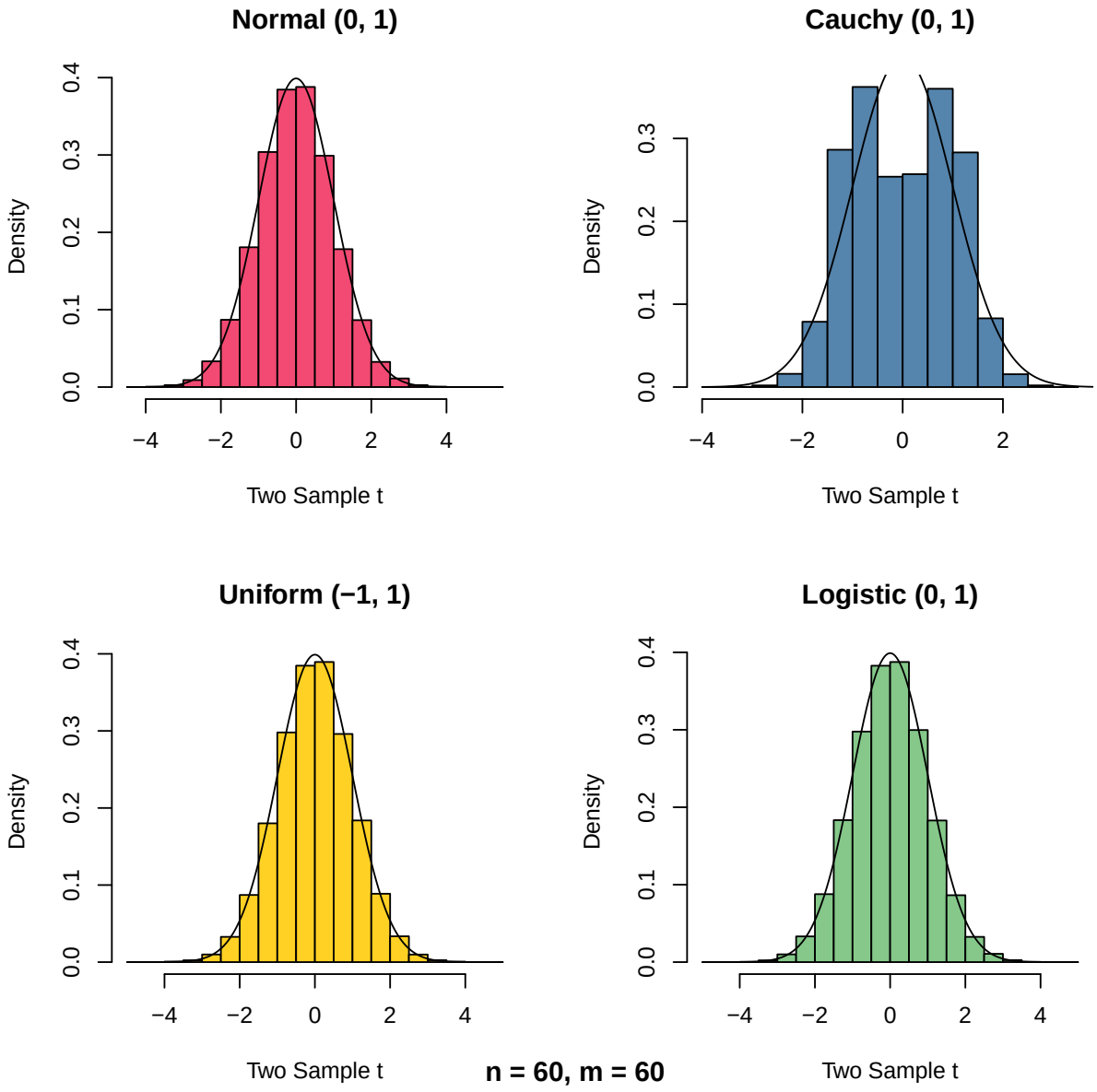


Remark 5.1. As N increases, the distribution of U statistic gets more and more closer to the standard normal distribution, irrespective of the underlying distribution of the sample.


Normal (0, 1)
Cauchy (0, 1)
Uniform (−1, 1)
Logistic (0, 1)
Density
Two Sample t
n = 10, m = 15

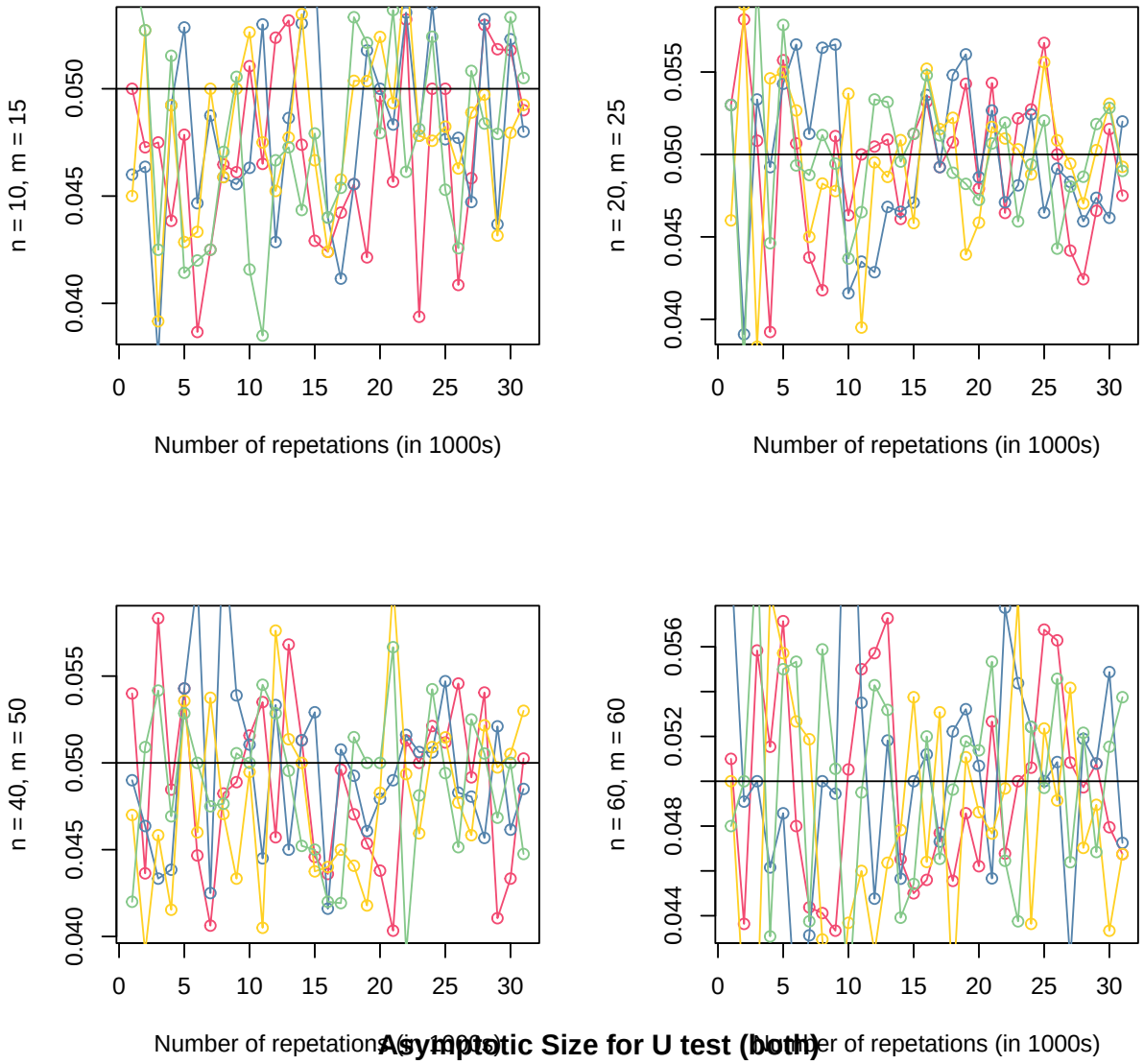

Normal (0, 1)
Cauchy (0, 1)
Uniform (−1, 1)
Logistic (0, 1)
Density
Two Sample t
n = 20, m = 25


Normal (0, 1)
Cauchy (0, 1)
Uniform (−1, 1)
Logistic (0, 1)
Density
Two Sample t
n = 40, m = 50

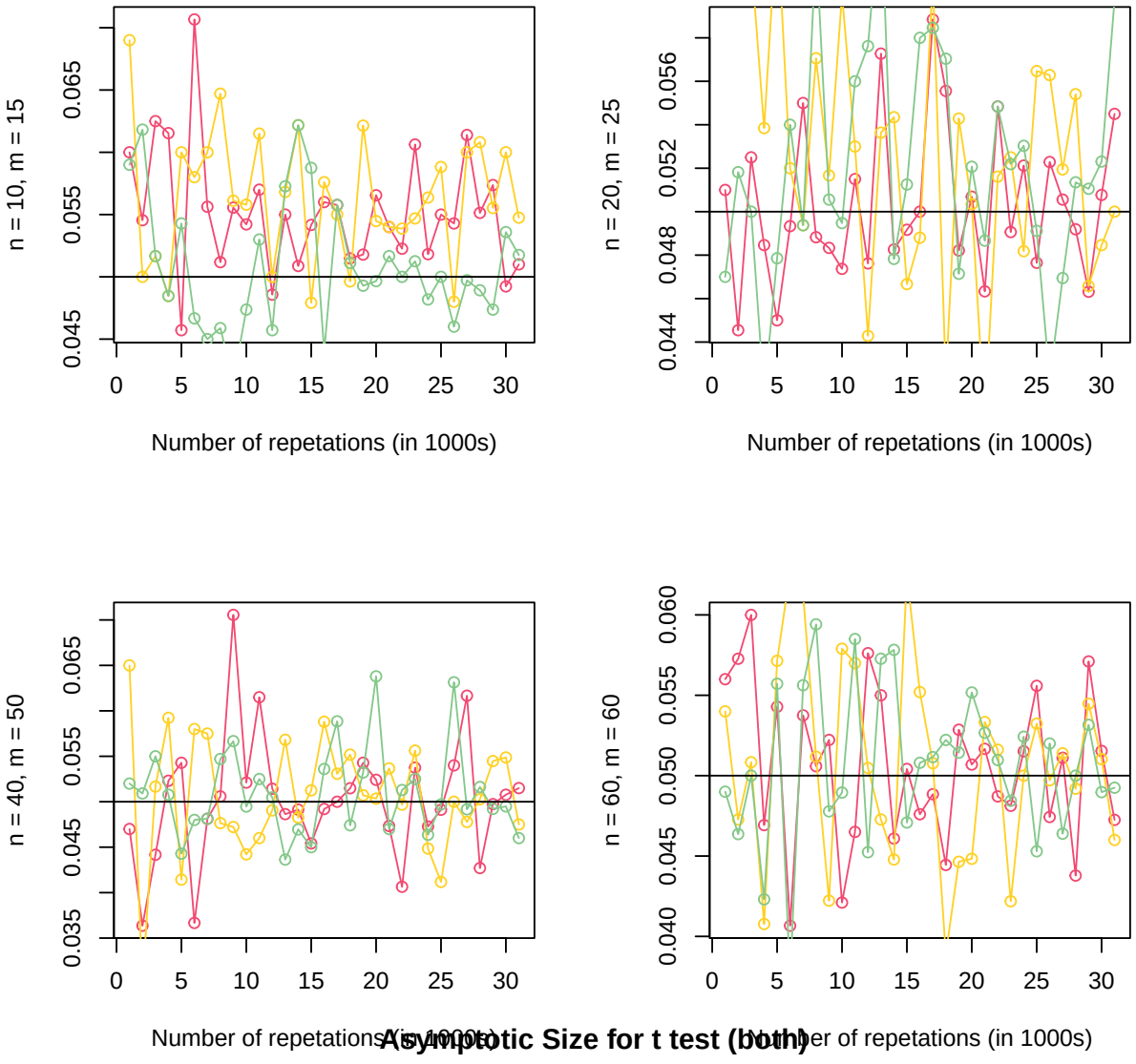


Remark 5.2. Although all the other distributions tend to the standard normal distribution, t statistic for Cauchy remains bimodal, which is expected. This, once again, exhibits the usefulness of Mann Whitney over two sample t for unknown data generating mechanism.

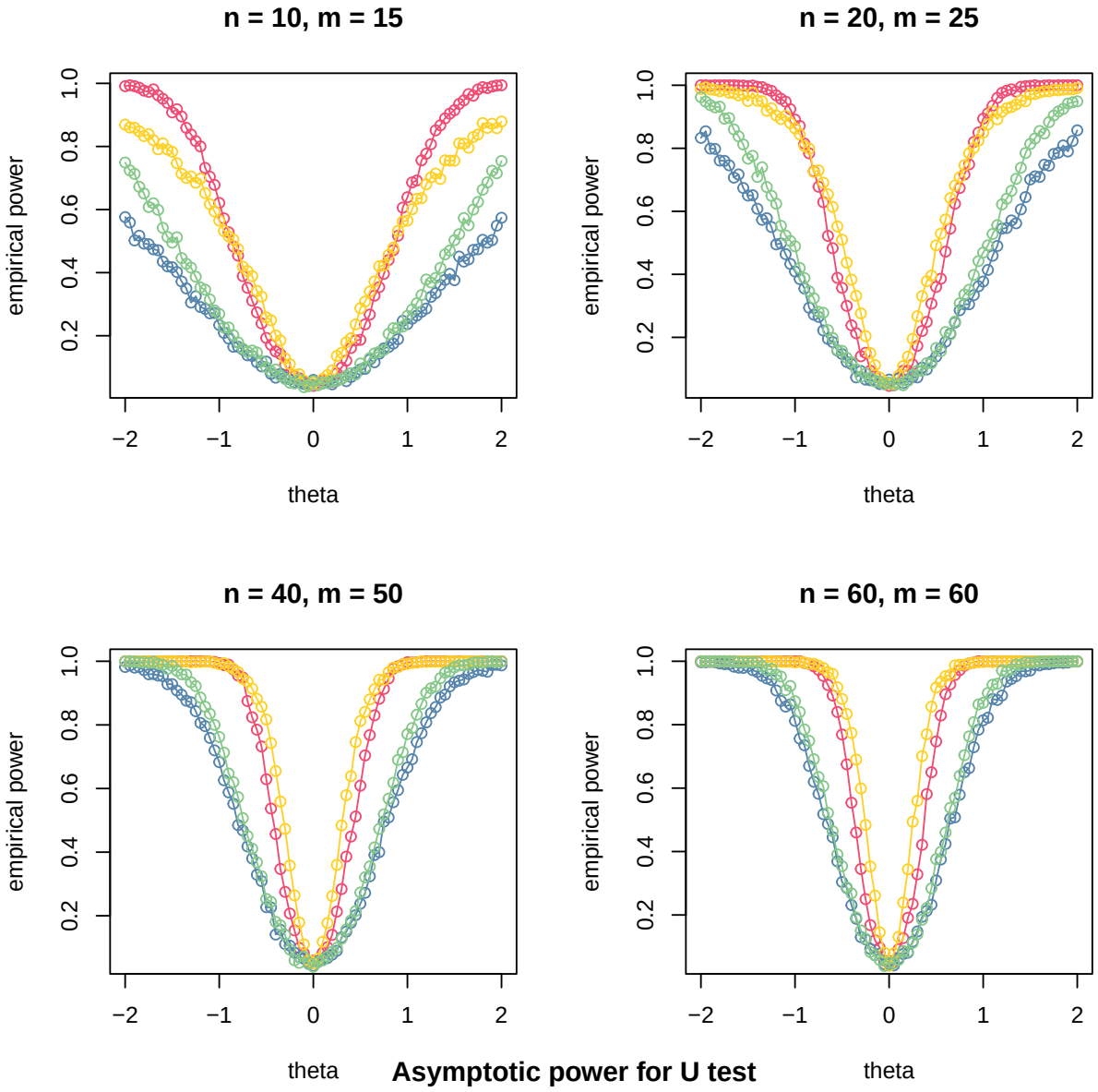
Now consider the asymptotic size and power curves:

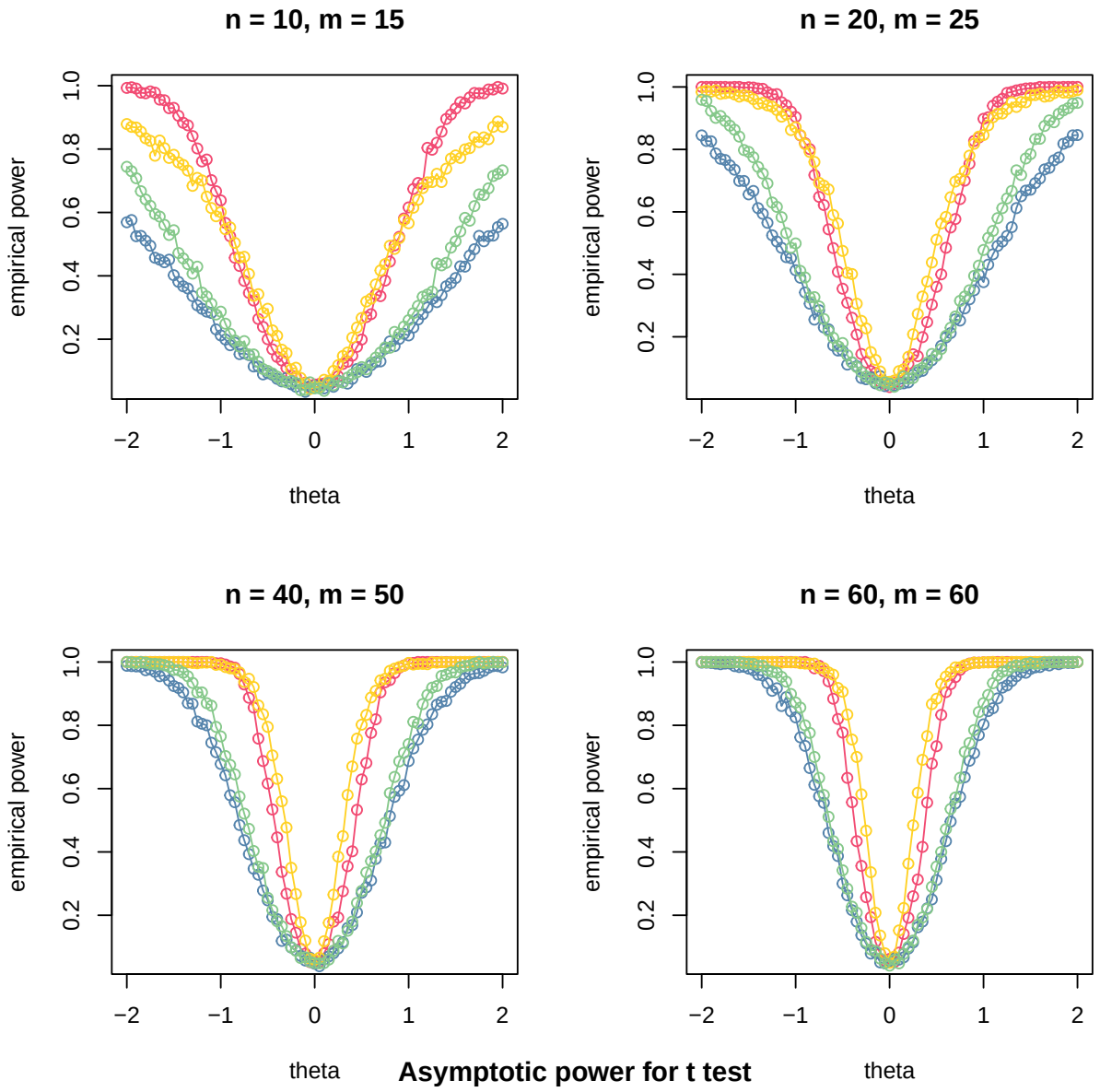


Asymptotic Size for U test (both)



Asymptotic Size for t test (both)


n = 10, m = 15
n = 20, m = 25
n = 40, m = 50
n = 60, m = 60
empirical power
theta
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
Asymptotic power for U test



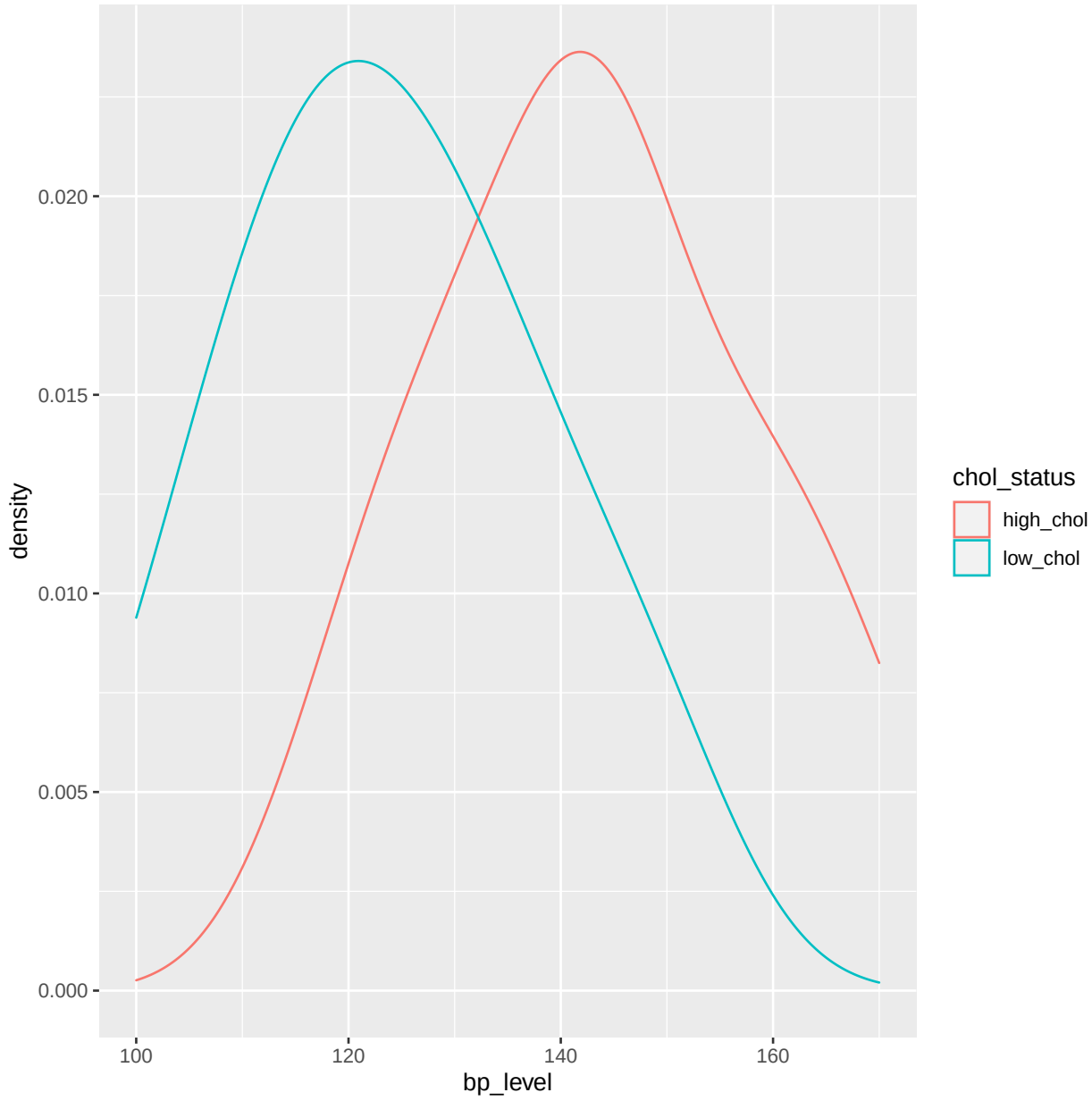
Remark 5.3. We observe from these asymptotic plots that as N increases, asymptotic properties start to hold, as expected.

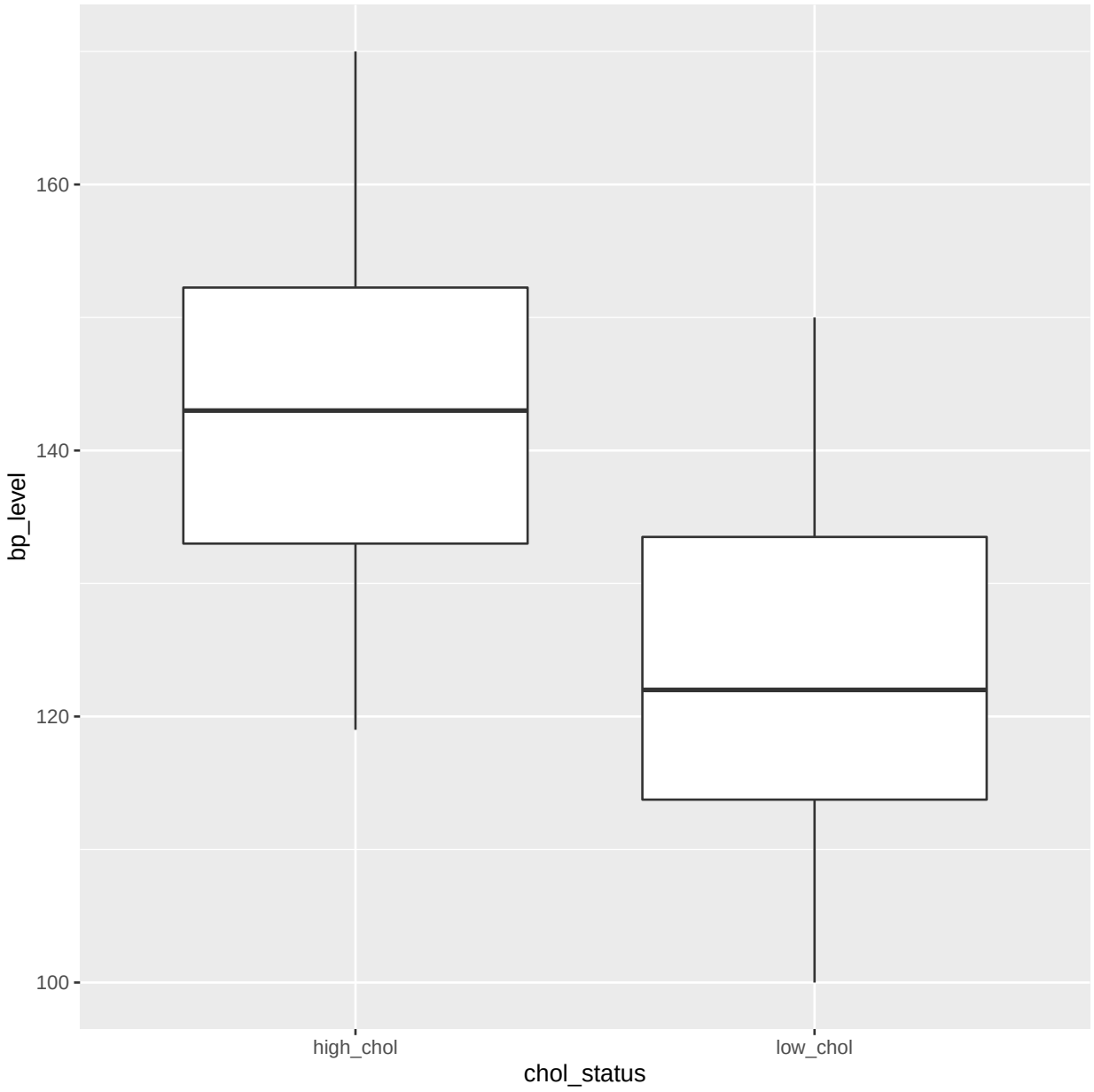
6 A Case Study

Consider the following dataset:

bp for low cholesterol	bp for high cholesterol
120	119
110	130
115	145
143	141
129	150
133	161
135	159
150	125
109	134
100	145
123	140
121	170

This data set is regarding the blood pressure levels of two groups of patients having lower and higher cholesterol levels. As we know, high cholesterol level causes higher blood pressure. The two samples of patients are drawn independently of each other. We will here test if there is any significant difference between the distributions of two class of patients.


density
0.020
0.015
0.010
0.005
0.000
100
120
140
160
bp_level
chol_status
high_chol
low_chol


160
140
120
100
bp_level
high_chol
low_chol
chol_status

Looking at the box plot and the empirical density plots we get that the median of patients having higher cholesterol is higher than that of patients having lower cholesterol and the distributions are more or less the same. So, if we assume that the underlying distribution of bp of patients having lower cholesterol to be $F(x)$ and that of having higher cholesterol to be $F(x-\theta)$. Then our hypothesis of interest is $H_0 : \theta = 0$ against $H_1 : \theta > 0$ (i.e., distribution of bp of patients having higher cholesterol is stochastically larger than the the distribution of bp of patients having lower cholesterol).

For this, we calculate the Mann Whitney statistic and calculate the corresponding p-value.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  bp_2 and bp_1
## W = 117.5, p-value = 0.004672
## alternative hypothesis: true location shift is greater than 0
```

Remark 6.1. Based on this statistic we calculate the p value to be 0.004672. So we reject the null at 0.01 or 0.05 level of significance. We therefore conclude that the distribution of bp of patients having higher cholesterol is stochastically larger than the distribution of bp of patients having lower cholesterol.

7 References and Bibliography

1. Gibbons, J.D., Chakraborti, S., *Nonparametric Statistical Inference, Fourth Edition*, Marcel Dekker, Inc.

2. https://www.rdocumentation.org

8 Acknowledgements

We would like to express our deeply felt gratitude and regards towards **Prof. Dr. Isha Dewan ma'am**. Her continuous support, suggestions, and guidance helped us in solving our doubts and figuring out methodologies and procedures. Her help and vision made it possible for us to finish the project.