Replication of Results from “*Parametric Versus Non-Parametric Statistics In The Analysis Of Randomized Trials With Non-Normally Distributed Data*”

By: Spencer Hunt

1 Introduction

The application of parametric methods to data with non-Gaussian distributions is often viewed as dubious at best, and outright wrong in other cases. This begs the question, just how inaccurate are parametric tests, such as the t-test, in cases where the assumption of normality is violated? Is there a clear improvement in accuracy when using a non-parametric approach, such as the Mann-Whitney test? If so, what is the extent of the difference between parametric and non-parametric tests when applied to the same datasets? These are the questions that the paper ”Parametric Versus Non-Parametric Statistics In The Analysis Of Randomized Trials With Non-Normally Distributed Data” by Andrew J Vickers sought to answer. In this paper, I will reproduce some of the tests in the aforementioned paper, and recreate table 3, and compare our results with the originals.

2 Summary of Paper

2.1 Background

It is widely accepted that in the case of non-normally distributed data, using a non-parametric test is the most logical approach. This is a result of the parametric tests often assuming normality, which is an assumption that is obviously violated with non-normal datasets. Often the power of the parametric tests can be quite low when the data violates the normality assumptions. For a 2 sample case, a Wilcoxon-Mann-Whitney(WMW) test is a common non-parametric choice, while the parametric test of choice is often a t-test or ANCOVA.

2.2 Summary

In the paper ”Parametric Versus Non-Parametric Statistics In The Analysis Of Randomized Trials With Non-Normally Distributed Data”, there were several tests and subsequent comparisons made. These tests were the result of generating random samples emulating existing empirical distributions from various sources, such as headache or shoulder pain levels before and after a treatment effect. They introduced a treatment effect to one of two groups from the empirical distribution and then compared them using both a WMW test and ANCOVA. The power of those tests to detect the treatment effect were compared using a ratio, and the results tabulated. This was repeated for multiple sample sizes and correlation coefficients. The result was a table that showed both the relative efficacy of one test vs another in certain scenarios, as well as the degree to which one outperformed the other under a variety of conditions. The non-parametric WMW test was used over other non-parametric tests because it is the most powerful test for 2 sample cases. ANCOVA was used because it has been shown to be more powerful than the t-test. It also adjusts for any chance baseline imbalances; it can be extended to incorporate randomization strata as co-variates, which has been shown to increase power; it can also be extended to incorporate time effects where measures are repeated. In their analysis, they found that ANCOVA is generally superior to Mann-Whitney. Smaller sample sizes and correlations near the extremes reduce the advantage of ANCOVA in favor of the WMW test. ANCOVA outperformed WMW for most distributions under most circumstances. ANCOVA has a major advantage over any non-parametric method: it provides an estimate for the size of the difference between group, that is, an effect size. Clinicians and patients generally want to know not just whether a treatment helps, but how much it helps, so they can determine whether it is worth the time, effort, risks and expense.

2.3 Objectives

In some cases, a parametric test may outperform a non-parametric test, despite the data not being normally distributed, or vice versa. I seek to compare the relative power between the t-test and WMW tests across multiple non-normally distributed datasets with different values of ρ and sample sizes using R to generate the simulated data. In our report, I aim to recreate Table 3 from the original report and compare the two tables for similarities and differences. It is worth noting that in the original paper, the software used for simulations was Stata as opposed to R. It is possible that this could lead to discrepancies in results, but such an outcome is unlikely and not consistently detectable, so for the purposes of this paper, it was ignored. In the article, they compared the WMW test to an ANCOVA test, but instead, I compared the WMW test to the parametric t-test. For the purpose of our paper, the additional properties of ANCOVA over the t-test are not needed. I also add to the original paper by analyzing the power of the Siegel-Tukey test for the same simulated datasets with a treatment effect of multiplying the variable by 0.8, rather than the adding 2 to the variable.

3 Statistical Analysis

3.1 Data Simulation

I started our analysis by simulating the data as they did in the article. I generated data X_1 and X_2 from a bivariate normal distribution of size N and specified ρ, by creating a mean vector of 0's, and a variance matrix with the diagonals equal to 1, and top right and bottom left corners equal to our specified ρ. By doing so, I can create two vectors of simulated data with the same properties and no deviations between the variables to be able to be detected by the test, while also being able to control the covariance between the observations. This allows us to be able to conduct our tests with control over the linear relationships between the data and find the power of the tests as covariance increases. I would expect that as the covariance approaches 1, I would have an easier time detecting a difference between observations since the variables are closely related to each other. I then applied the polynomial, $Y_i = 14.8 + 16.5X_i + 7.5X_i^2 - 1.15X_i^3$ where $i = 1, 2$, to transform the data into a slightly skewed dataset that represented a random sample of pain scores from an empirical distribution that was non-normally distributed. I then applied a treatment effect to one of the variables as $Z_1 = Y_1$, and $Z_2 = Y_2 + 2$, so that there is a difference that can be detected when I run the tests. Now that I have our data, I can run our tests and calculate the power for the different combinations of ρ and N.

3.2 Running the tests

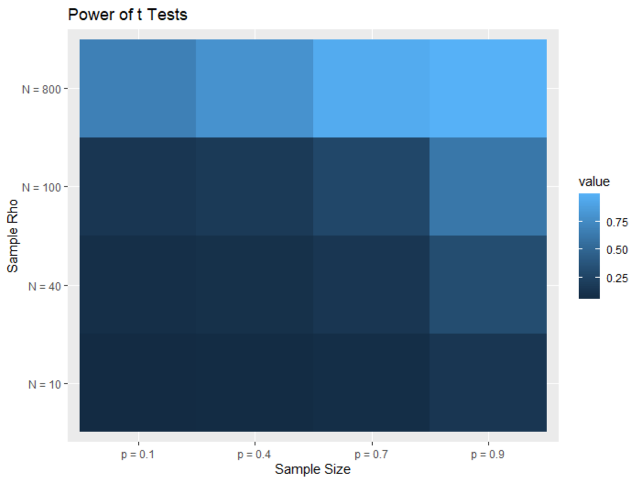
In the article, they used combinations of $\rho = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.70, 0.8, 0.9$ and $N = 10, 20, 30, 40, 60, 100, 200, 400, 800$, but for simplicity and ease of analysis, I only used $\rho = 0.1, 0.4, 0.7, 0.9$, and $N = 10, 40, 100, 800$. I believe I can still come to the same conclusions with fewer combinations. To calculate the power, I ran 1000 iterations in a for loop of simulating a data set as described above, running the test on the dataset, calculating the power of the individual test, then storing the power of the specific test into an empty vector. After I obtained 1000 power observations for one specified combination of ρ and N, I took the average of these power values to indicate the overall power of the test for specified values of ρ, and N. I did this same process for the t-test and the WMW test. The results are listed below. Table 1 consists of the power calculations of the parametric t-test, and table 2 consists of the power calculations of the non-parametric WMW test. The last plots are a heatmap of the power values to have a better visual understanding of how the power changes.

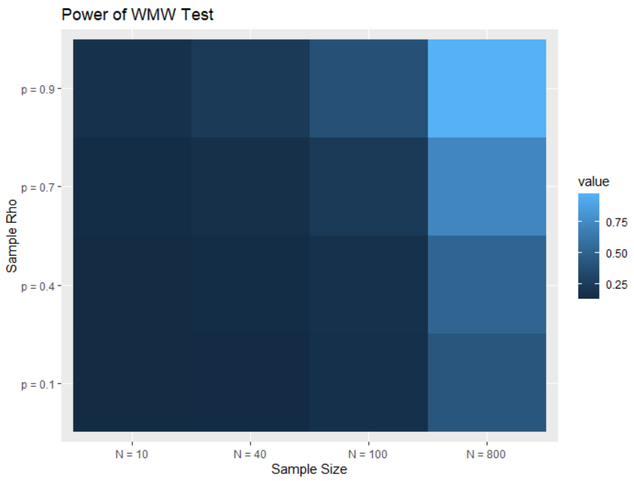
	N = 10	N = 40	N = 100	N = 800
ρ = 0.1	0.05198	0.08168	0.13799	0.67885
ρ = 0.4	0.05773	0.09572	0.17245	0.80054
ρ = 0.7	0.07708	0.140984	0.27359	0.96305
ρ = 0.9	0.142775	0.317389	0.614301	0.999987

Table 1: Power of paired t-tests

	N = 10	N = 40	N = 100	N = 800
ρ = 0.1	0.130165	0.1265	0.1566	0.43094
ρ = 0.4	0.12507	0.13318	0.161595	0.507353
ρ = 0.7	0.13580	0.15662	0.21712	0.717960
ρ = 0.9	0.167484	0.22985	0.371973	0.979230

Table 2: Power of WMW tests





3.3 Ratio of the Power Results

Now that I have obtained the different powers of the parametric t-test and the non-parametric WMW test, I can calculate the power ratio between the t-test and WMW test for different combinations of ρ and N. This ratio allows us to better understand how much more powerful one test is than the other. The ratio is calculated by dividing the power of the WMW test by the power of the t-test. Values greater than 1 indicate the WMW test had a higher power for the specified ρ and N, and values less than 1 indicate the t-test had a higher power for the specified ρ and N. The ratio is listed in table 3 below. Observations close to 1 indicate equal power between the tests. I can now compare our recreated table to the table they created in the original report.

	ρ = 0.1	ρ = 0.4	ρ = 0.7	ρ = 0.9
N = 10	2.50414	2.16646	1.76181	1.17306
N = 40	1.54873	1.3913	1.1109	0.72334
N = 100	1.13486	0.93705	0.79359	0.60552
N = 800	0.63481	0.63376	0.74551	0.97924

Table 3: Ratio between t-test and WMW test

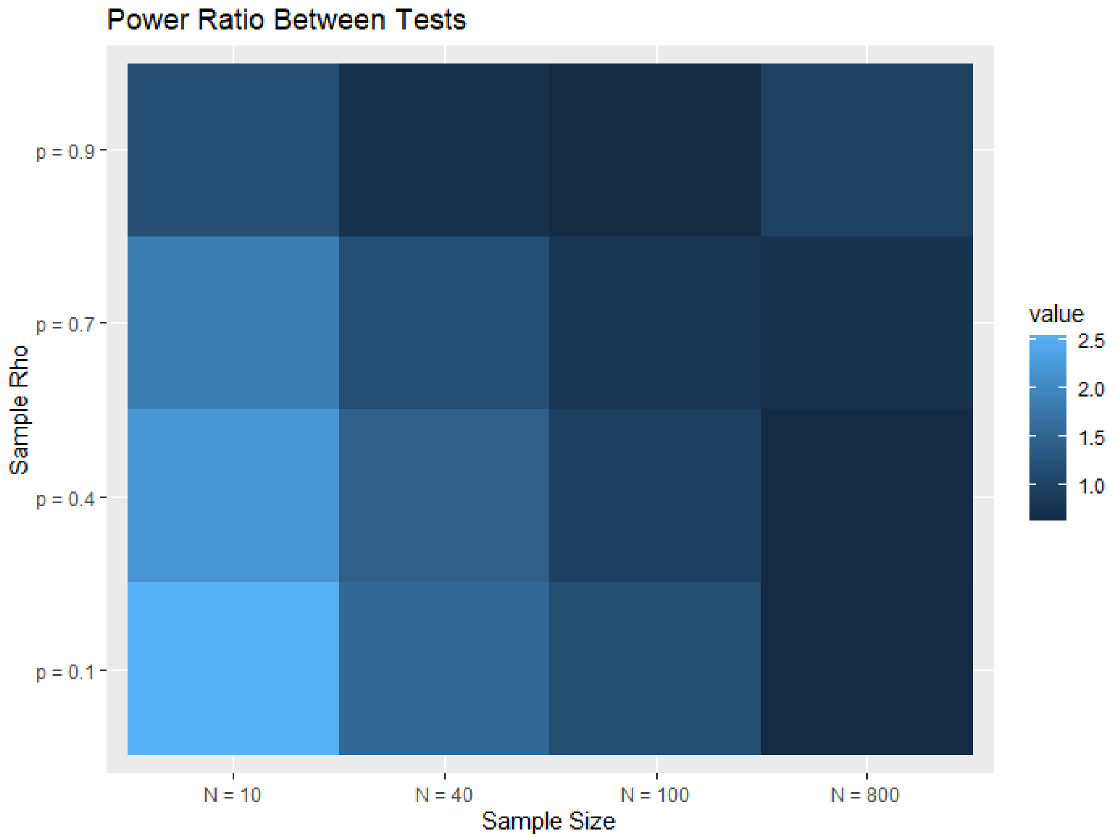
4 Discussion

Our results are consistent with those in the paper, despite our divergence from its exact methodology. Our use of a t-test as an alternate parametric test as opposed to ANCOVA seems to reinforce the conclusions drawn in the original paper; that the WMW test outperforms its parametric counterparts in cases of low sample size or high correlation, whereas the t-test only pulls ahead in cases of low correlation and high sample size. I can better visualize these results from the heatmap listed below. The power of the WMW test is significantly higher when N and p are low, and gradually gets worse as ρ increases. I can also see that when N is large, the WMW test does start to gain on the t-test as ρ increases. It never overtakes the t-test, but the power values do seem to converge in power. These results are somewhat surprising given the common sentiment discussed previously around the use of non-parametric tests as opposed to parametric options, as our findings suggest that it would be preferable to use a non-parametric approach in some cases. If our understanding of ANCOVA testing were to improve, re-running the exact process from the original paper would be interesting, as I could easily compare our results directly to those in the paper.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.8125	0.9773	0.601	0.9403	0.9722	0.8834	1.2129	1.0851	1.1320
20	1.0160	1.1093	0.8172	0.9348	0.7742	0.9905	0.8124	1.0058	1.0231
30	1.0000	0.9167	0.7785	0.8617	0.8027	0.7756	0.8689	0.8760	1.0096
40	0.8866	0.8596	0.8120	0.7332	0.7365	0.7556	0.9172	0.9067	0.929
60	0.8925	0.8996	0.7752	0.7632	0.7418	0.8277	0.8728	0.8841	0.9892
100	0.8822	0.8594	0.7816	0.7259	0.7071	0.8277	0.8639	0.8702	0.9259
200	0.8484	0.8030	0.7611	0.6920	0.6979	0.7591	0.8793	0.8888	-
400	0.8512	0.8292	0.7392	0.7113	0.6707	0.8029	0.8336	-	-
800	0.8781	0.9087	-	-	-	-	-	-	-

Table 4: This is Table 3 from the original paper

The table above (Table 4 from the original paper), displays the ratio of the powers of the two tests with given set of values for ρ and N. A value greater than 1 represents WMW having a higher power, and a value less than 1 represents ANCOVA showing a higher power. The blank cells indicate the power of one or both tests was equal to 1. In the plot below, I have provided a heatmap of the power values from our analysis. The lighter the color is, the better the WMW test did compared to the t-test.



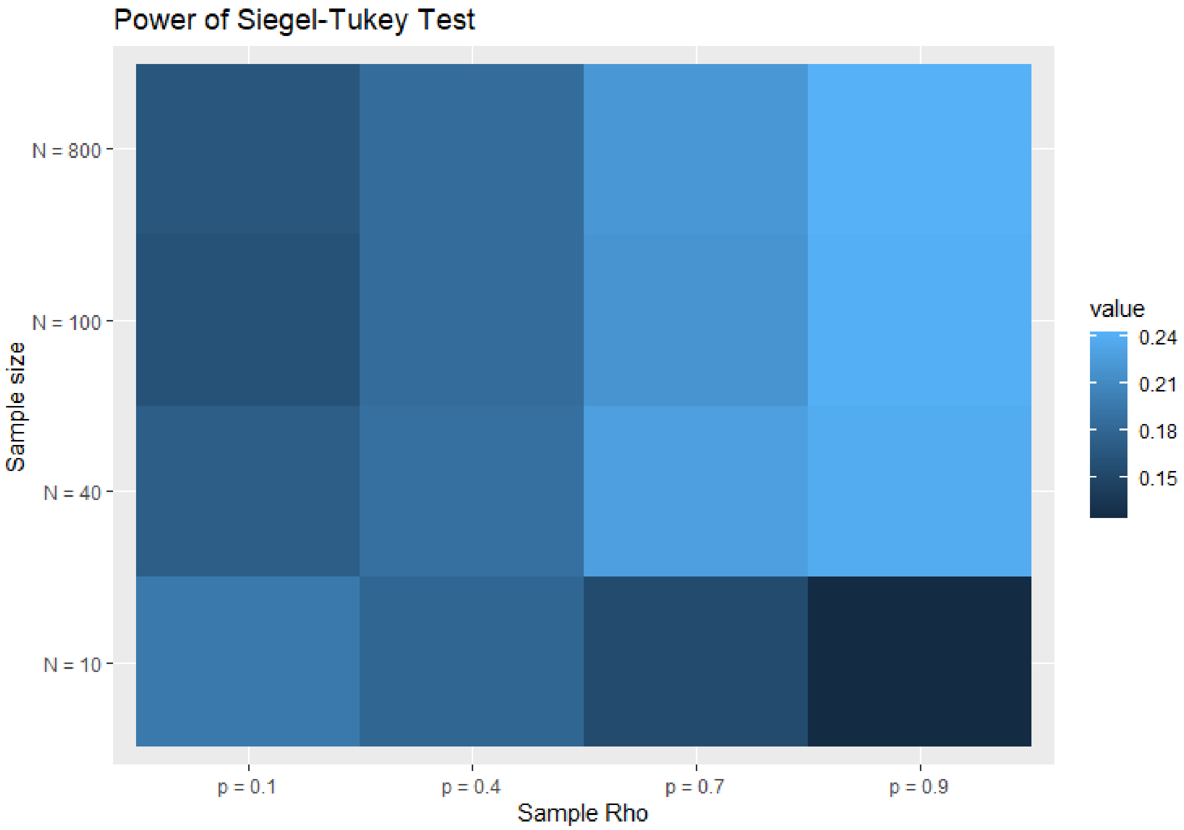
As you can see, the WMW test performs substantially better in situation witha small sample size and low correlation between samples. As either of these factors increase, the relative advantage enjoyed by the WMW test over the t-test dwindles until it is overtaken. It is, however, worth noting that the WMW test is not necessarily a weak test in these conditions, just less powerful than a t-test. In summary, in situation favoring the WMW test, it is the more powerful test by far, but even in relatively unfavorable conditions, the WMW test is still a somewhat powerful test even when compared to a t-test.

5 Additional Analysis

Following the same parameters I used for the t-test and WMW test, I also conducted the Siegel-Tukey test for the variance of the non-parametric distribution with specified ρ and N, and found the power for it. The simulated data was generated in the same way, but when I applied the treatment effect to one of the variables, instead of adding 2 to one of the variables, I multiplied the variable by 0.8 to introduce a difference to be detected by the test. The transformed data is as follows: $Z_1 = Y_1$, and $Z_2 = Y_2(0.8)$. Looking at the table, you can see that the power for the Siegel-Tukey test stays increasing as you go up in correlation and sample size but it does not exceed 0.95 or even 0.3. When looking at the smaller sample sizes and correlations of all the test conducted I see that there are some cases where the Siegel-Tukey test would have higher power, but as sample size and correlation increase you can see that the t-test and WMW test have a higher power. From this I can conclude that the Siegel-Tukey test is not very good at detecting a false null hypothesis. I can also say that the Siegel-Tukey test would not be good for most areas of the research done in the article. This makes sense as the authors of the article did not conduct the Siegel-Tukey test, let alone mention it in the paper at all.

	N = 10	N = 40	N = 100	N = 800
ρ = 0.1	0.19089	0.16788	0.17249	0.17444
ρ = 0.4	0.18271	0.19039	0.19323	0.20087
ρ = 0.7	0.14331	0.23305	0.22794	0.21594
ρ = 0.9	0.12886	0.25290	0.26330	0.23991

Table 5: Power of Siegel-Tukey Test



6 Appendix

```
###Power for the Wilcoxon Mann Whitney test###
#generate bivariate normal distribution of 10 samples and correlation .1

powerWMW10p1 = c()

for(i in 1:1000){
    N = 10
    rho <- .1
    mu <- c(0,0)
    sigma <- matrix(c(1,rho, rho,1),2)
    sam10 <- mvrnorm(N, mu = mu, Sigma = sigma)
    sam10df <- data.frame(sam10)
    x1 = sam10df[,1]
    x2 = sam10df[,2]
    y1 = 14.8 + 16.5*sam10df[,1] + 7.5*(sam10df[,1]^2) - 1.15*(sam10df[,1]^3)
    y2 = 14.8 + 16.5*sam10df[,2] + 7.5*(sam10df[,2]^2) - 1.15*(sam10df[,2]^3)
    z1 = y1
    z2 = y2+2
    wilcox.test(z1,z2)
    s = sd(z1-z2)
    e = abs((mean(z1)-mean(z2))/s)
    powerWMW10p1[i] = pwr.2p.test(h = e, n = N, sig.level = 0.05,
    alternative = ”two.sided”)$power
}
power_WMW10p1 = sum(powerWMW10p1)/1000
power_WMW10p1

#Then Repeat for sample sizes 40,100,and 800 and correlations .4, .7, and .9


###Power for the t-test###
#generate bivariate normal distribution of 10 samples and correlation .1

powerttest10p1 = c()

for(i in 1:1000){
  N = 10
  rho <- .1
  mu <- c(0,0)
  sigma <- matrix(c(1,rho, rho,1),2)
  sam10 <- mvrnorm(N, mu = mu, Sigma = sigma)
  sam10df <- data.frame(sam10)
  x1 = sam10df[,1]
  x2 = sam10df[,2]
  y1 = 14.8 + 16.5*sam10df[,1] + 7.5*(sam10df[,1]^2) - 1.15*(sam10df[,1]^3)
  y2 = 14.8 + 16.5*sam10df[,2] + 7.5*(sam10df[,2]^2) - 1.15*(sam10df[,2]^3)

  z1 = y1
  z2 = y2+2
  d = z1-z2
  wilcox.test(z1,z2, paired = TRUE)
  se = sd(d)/sqrt(N)
  a = 1.96*(se)
  powerttest[i] = 1-pnorm((a-2)/se)

}

power_ttest10p1 = sum(powerttest10p1)/1000

#Then Repeat for sample sizes 40,100,and 800 and correlations .4, .7, and .9


###Power for the Tukey Test###
#generate bivariate normal distribution of 10 samples and correlation .1

powerT = c()

for(i in 1:1000){
  N = 10
  rho <- .1
  mu <- c(0,0)
  sigma <- matrix(c(1,rho, rho,1),2)
  sam10 <- mvrnorm(N, mu = mu, Sigma = sigma)
  sam10df <- data.frame(sam10)
  x1 = sam10df[,1]
  x2 = sam10df[,2]
  y1 = 14.8 + 16.5*sam10df[,1] + 7.5*(sam10df[,1]^2) - 1.15*(sam10df[,1]^3)
  y2 = 14.8 + 16.5*sam10df[,2] + 7.5*(sam10df[,2]^2) - 1.15*(sam10df[,2]^3)

  z1 = y1
  z2 = y2*0.8
  d = z1-z2
  dataT <- data.frame(group = rep(c(”A”, ”B”), each = 10), values = c(z1, z2))
  model <- aov(values~group, data = dataT)
  amr <- anova(model)[”Residuals”, ”Mean■Sq”]
  TK <- TukeyHSD(model, conf.level = .95)
  TKL <- list(TK$group)
  TKD <- data.frame(TKL)
  amd <- TKD["B-A", "diff"]
  pow <- power.tukey.test(n = 10, groups = 2, delta = amd, within.var = amr)
  powerT[i] = pow$power
}
power_T = sum(powerT)/1000

#Then Repeat for sample sizes 40,100,and 800 and correlations .4, .7, and .9

#HeatMaps
library(reshape)
library(ggplot2)

# WMW Heat Map
dataWMW = c(power_WMW10p1, power_WMW40p1, power_WMW100p1, power_WMW800p1,
            power_WMW10p4, power_WMW40p4, power_WMW100p4, power_WMW800p4,
            power_WMW10p7, power_WMW40p7, power_WMW100p7, power_WMW800p7,
            power_WMW10p9, power_WMW40p9, power_WMW100p9, power_WMW800p9)

matWMW = matrix(dataWMW, nrow = 4, ncol = 4, byrow = TRUE)

colnames(matWMW)<- c("N=10", "N=40", "N=100", "N=800")
rownames(matWMW)<- c("p=0.1", "p=0.4", "p=0.7", "p=0.9")
head(matWMW,4)

data_meltWMW <- melt(matWMW)
library(ggplot2)

ggpWMW <- ggplot(data_meltWMW, aes(X2, X1)) +
  geom_tile(aes(fill = value)) +
  labs(x = "Sample■Size", y = "Sample■Rho", title = "Power■of■WMW■Test")
ggpWMW

#t-test Heat Map
datat = c(power_ttest10p1, power_ttest40p1, power_ttest100p1, power_ttest800p1,
          power_ttest10p4, power_ttest40p4, power_ttest100p4, power_ttest800p4,
          power_ttest10p7, power_ttest40p7, power_ttest100p7, power_ttest800p7,
          power_ttest10p9, power_ttest40p9, power_ttest100p9, power_ttest800p9)

matttest = matrix(datat, nrow = 4, ncol = 4, byrow = TRUE)
colnames(matttest)<- c("N=10", "N=40", "N=100", "N=800")
rownames(matttest)<- c("p=0.1", "p=0.4", "p=0.7", "p=0.9")
head(matttest,4)
```

```
data_meltt <- melt(matttest)

ggpt <- ggplot(data_meltt, aes(X1, X2)) +
  geom_tile(aes(fill = value)) +
  labs(x = "Sample■Size", y = "Sample■Rho", title = "Power■of■t■Tests")
ggpt

#Ratio Heat Map
ratio = dataWMW/datat
ratioMat = matrix(ratio, nrow = 4, ncol = 4, byrow = FALSE)
rownames(ratioMat)<- c("N=10", "N=40", "N=100", "N=800")
colnames(ratioMat)<- c("p=0.1", "p=0.4", "p=0.7", "p=0.9")
head(ratioMat,4)

data_meltR <- melt(ratioMat)

ggpR <- ggplot(data_meltR, aes(X1, X2)) +
  geom_tile(aes(fill = value)) +
  labs(x = "Sample■Size", y = "Sample■Rho", title = "Power■Ratio■Between■Tests")
ggpR
```

7 References

1. Vickers, A.J. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. BMC Med Res Methodol 5, 35 (2005). https://doi.org/10.1186/1471-2288-5-35