ST 551 Homework 7

Student's Name Goes Here

Fall 2022

Instructions

This assignment is due by December 2nd 11:59 PM, on Canvas via Gradescope. **You should submit your assignment as a typed PDF which you can compile using the provide .Rmd (R Markdown) template.** Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, grammatically correct sentences for your solutions.

General Pearson's Chi-squared Test

Question 1 (5 points)

Suppose we are comparing two categorical variables: S = Favorite cards suit {Clubs, Diamonds, Hearts, Spades} and C = Favorite primary color {Red, Yellow, Blue}. In the particular population we are interested in, the true joint distribution of these two variables is as shown in the table below:

	Clubs	Diamonds	Hearts	Spades	Row Totals
Red	0.02	0.04	0.06	0.08	0.20
Yellow	0.03	0.06	0.09	0.12	0.30
Blue	0.05	0.10	0.15	0.20	0.50
Column Totals	0.10	0.20	0.30	0.40	1.00

Part A. (1 point) Are these two variables independent in this population? In other words, is the null hypothesis of Pearson's Chi-squared test (that there is no association between these two categorical variables) true?

I would argue yes. The reason behind this is that knowing one variable, doesn't the probability of another. For example, since they are independent then $P(\text{Red} \mid \text{Clubs}) = P(\text{Red})$. However, In this case $P(\text{Red} \mid \text{Clubs}) = \frac{.02}{.10} = .20 = P(\text{Red}) = .20$. This can be shown for all others as well.

Part B. (4 points) Simulate 10,000 datasets from this population for each of the sample sizes in the table below, and perform (1) 1. Pearson's Chi-squared Test without continuity correction (Just so that we are all doing the same, simple test), and (2) 2. Fisher's Exact Test. The following code might be helpful to get you started

```r
n <- 5000
ptab_a <- rbind(c(0.02, 0.04, 0.06, 0.08),
                c(0.03, 0.06, 0.09, 0.12),
                c(0.05, 0.10, 0.15, 0.20))

N <- 10000

chi_p <- rep(0,N)
fish_p <- rep(0,N)

for (i in 1:N) {
  samp_tab <- matrix(rmultinom(1, n, ptab_a), nrow=3)
  chi_p[i] <- chisq.test(samp_tab, correct = FALSE)$p.val
  fish_p[i] <- fisher.test(samp_tab, simulate.p.value = TRUE)$p.val

}

sum(chi_p <= .05, na.rm = TRUE)/N
sum(fish_p <= .05, na.rm = TRUE)/N
```

> Note that in small sample sizes, it is quite possible to get an entire row or entire column of zeros, which will cause an `NA` result from the `chisq.test()` function. I would suggest just ignoring those cases, and calculate the rejection rate based on the non-NA results. The functions `mean()`

and `sum()` have an optional argument called `na.rm` (standing for NA remove). Set this argument to `TRUE` to calculate the mean or sum of the non-NA elements for a vector that has some NA values.

Fill in the rejection rates for a level $\alpha = 0.05$ test. What do these rejection rates tell us about how the two tests perform? Which test has better performance?

Sample Size	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
Pearson's rejection rate	.0462	.0480	.0458	.0512	.0506
Fisher's rejection rate	.0502	.0504	.0474	.0500	.0495

It seems like both perform very well and have rejection rates that we would expect. Although, it does seem at least with smaller samples that pearson's test seems to vary a bit more, so we might prefer fishers test.

Question 2 (5 points)

Now suppose we are studying the same two categorical variables, but in the population we are studying the true joint distribution of these two variables is as shown in this new table:

	Clubs	Diamonds	Hearts	Spades	Row Totals
Red	0.02	0.04	0.05	0.09	0.20
Yellow	0.02	0.07	0.10	0.11	0.30
Blue	0.06	0.09	0.15	0.20	0.50
Column Totals	0.10	0.20	0.30	0.40	1.00

Part A. (1 point) Are these two variables independent in this population? In other words, is the null hypothesis of Pearson's Chi-squared test (that there is no association between these two categorical variables) true?

No, because if we look at: $P(\text{Yellow} \mid \text{Clubs}) = \frac{P(\text{Yellow and Clubs})}{\text{Clubs}} = \frac{.02}{.10} = .20 \neq P(\text{Yellow}) = .30$

There are other examples throughout the table.

Part B. (4 points) Simulate 10,000 datasets from this population for each of the sample sizes in the table below, and perform (1) 1. Pearson's Chi-squared Test without continuity correction (Just so that we are all doing the same, simple test), and (2) 2. Fisher's Exact Test. Fill in the rejection rates for a level $\alpha = 0.05$ test. What do these rejection rates tell us about how the two tests perform?

```
n <- 5000
ptab_a <- rbind(c(0.02, 0.04, 0.05, 0.09),
                c(0.02, 0.07, 0.10, 0.11),
                c(0.06, 0.09, 0.15, 0.20))

N <- 10000

chi_p <- rep(0,N)
fish_p <- rep(0,N)

for (i in 1:N) {
  samp_tab <- matrix(rmultinom(1, n, ptab_a), nrow=3)
  chi_p[i] <- chisq.test(samp_tab, correct = FALSE)$p.val
  fish_p[i] <- fisher.test(samp_tab, simulate.p.value = TRUE)$p.val

}

sum(chi_p <= .05, na.rm = TRUE)/N
```

```
## [1] 1
```

```
sum(fish_p <= .05, na.rm = TRUE)/N
```

```
## [1] 1
```

Sample Size	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
Pearson's rejection rate	.0592	.0927	.4217	.7733	1.000
Fisher's rejection rate	.0660	.0982	.4267	.7774	1.000

Again, both test perform similarly overall! They both seem to be decent options. They both reject that the categories are independent with higher probability with higher sample sizes.

Wilcoxon Rank-sum Test

Question 3 (6 points)

Is the Wilcoxon Rank-sum Test an *asymptotically exact* test of equality for medians? If not, give an example of distributions F and G with **the same median**, for which the test will always **reject too often**, no matter the sample size. Demonstrate with some simulations that the rejection rate is incorrect if we view the Wilcoxon rank sum test as a test of equality of medians.

As seen in lecture, the Wilcoxon Rank-Sum Test (WRST) is an exact test of the dsitributions, but not exact for medians. This is true unless we have a location shift (additive effect) assumption. An example of rejecting too often:

$$F_X = N(\log(2), 1), m_X = \log(2)$$

$$F_Y = Exp(1), m_Y = \log(2)$$

```
N <- 10000
n1 <- 30
n2 <- 100
n3 <- 500

p_val1 <- rep(0,N)
p_val2 <- rep(0,N)
p_val3 <- rep(0,N)

for (i in 1:N) {

  sample1A <- rnorm(n1, log(2), 1)
  sample1B <- rexp(n1,1)
  sample2A <- rnorm(n2, log(2), 1)
  sample2B <- rexp(n2,1)
  sample3A <- rnorm(n3, log(2), 1)
  sample3B <- rexp(n3,1)

  p_val1[i] <- wilcox.test(sample1A,sample1B)$p.value
  p_val2[i] <- wilcox.test(sample2A,sample2B)$p.value
  p_val3[i] <- wilcox.test(sample3A,sample3B)$p.value


}

sum(p_val1 <= .05)/N
```

```
## [1] 0.1214
```

```
sum(p_val2 <= .05)/N
```

```
## [1] 0.2904
```

```
sum(p_val3 <= .05)/N
```

```
## [1] 0.8769
```

We can see even at small and large sample sizes, we reject quite often that they are not the same. However, we can clearly see that the medians are the same.

Question 4 (6 points)

Is the Wilcoxon Rank-sum Test a *consistent* test of $H_0 : F = G$? If not, give an example of $F \neq G$ for which the test **will not have power tending to one**, no matter how large the sample size. Demonstrate with some simulations that the power is limited even in very large sample sizes.

As seen in lecture, the Wilcoxon Rank-Sum Test (WRST) is not a consistent test for population distributions. It only works if we have a location shift assumption about the distributions. An example of power being limited (lower rejection rate than expected no matter sample size):

$$F_X = N(0,1) \text{ and } F_Y = \begin{cases} N(-2,1) & \text{with probability } 1/2 \\ N(2,1) & \text{with probability } 1/2 \end{cases} \text{with } n_X < 0.5(n_Y - 1)$$

with $n_X < 0.5(n_Y - 1)$

```
N <- 10000
n1X <- 30
n2X <- 100
n3X <- 200
n1Y <- 80
n2Y <- 250
n3Y <- 500

p_val1 <- rep(0,N)
p_val2 <- rep(0,N)
p_val3 <- rep(0,N)

sample1Y <- rep(0,n1Y)
sample2Y <- rep(0,n2Y)
sample3Y <- rep(0,n3Y)


for (i in 1:N) {

  for(j in 1:n1Y) {
    p <- runif(1,0,1)
    if(p < .5) {sample1Y[j] <- rnorm(1,-2,1)}
    if(p >= .5) {sample1Y[j] <- rnorm(1,2,1)}
  }
  for(j in 1:n2Y) {
    p <- runif(1,0,1)
    if(p < .5) {sample2Y[j] <- rnorm(1,-2,1)}
    if(p >= .5) {sample2Y[j] <- rnorm(1,2,1)}
  }
  for(j in 1:n3Y) {
    p <- runif(1,0,1)
    if(p < .5) {sample3Y[j] <- rnorm(1,-2,1)}
    if(p >= .5) {sample3Y[j] <- rnorm(1,2,1)}
  }
  sample1X <- rnorm(n1X, 0, 1)
  sample2X <- rnorm(n2X, 0, 1)
  sample3X <- rnorm(n3X, 0, 1)
  p_val1[i] <- wilcox.test(sample1X,sample1Y)$p.value
  p_val2[i] <- wilcox.test(sample2X,sample2Y)$p.value
  p_val3[i] <- wilcox.test(sample3X,sample3Y)$p.value


}

sum(p_val1 <= .05)/N
```

```
## [1] 0.0194
```

```
sum(p_val2 <= .05)/N
```

```
## [1] 0.0258
```

```
sum(p_val3 <= .05)/N
```

```
## [1] 0.0243
```

As seen, no matter the sample size the rejection rate stayed too small.

Question 5 (6 points)

Is the Wilcoxon Rank-sum Test an *asymptotically exact* test of $H_0 : P(X > Y) = 0.5$ for independent observations $X \sim F$ and $Y \sim G$? If not, give an example of distributions F and G with $X \sim F$ and $Y \sim G$ such that $P(X > Y) = 0.5$ for which the test will not reject at the target level, no matter the sample size. Demonstrate with some simulations that the rejection rate is incorrect — try a variety of sample sizes, and explore keeping the ratio of sample sizes n_X/n_Y fixed while increasing both.

I would argue that it is not asymptotically exact of that test under these assumptions. The reasoning behind this is there is no additive effect assumption. Here the example is similar to the last:

$$F_X = N(0,1) \text{ and } F_Y = \begin{cases} N(-2,1) & \text{with probability } 1/2 \\ N(2,1) & \text{with probability } 1/2 \end{cases} \text{ with } n_X > 0.5(n_Y - 1)$$

with $n_X > 0.5(n_Y - 1)$

```
N <- 10000
n1Y <- 30
n2Y <- 100
n3Y <- 200
n1X <- 80
n2X <- 250
n3X <- 500

p_val1 <- rep(0,N)
p_val2 <- rep(0,N)
p_val3 <- rep(0,N)

sample1Y <- rep(0,n1Y)
sample2Y <- rep(0,n2Y)
sample3Y <- rep(0,n3Y)


for (i in 1:N) {

  for(j in 1:n1Y) {
    p <- runif(1,0,1)
    if(p < .5) {sample1Y[j] <- rnorm(1,-2,1)}
    if(p >= .5) {sample1Y[j] <- rnorm(1,2,1)}
  }
  for(j in 1:n2Y) {
    p <- runif(1,0,1)
    if(p < .5) {sample2Y[j] <- rnorm(1,-2,1)}
    if(p >= .5) {sample2Y[j] <- rnorm(1,2,1)}
  }
  for(j in 1:n3Y) {
    p <- runif(1,0,1)
    if(p < .5) {sample3Y[j] <- rnorm(1,-2,1)}
    if(p >= .5) {sample3Y[j] <- rnorm(1,2,1)}
  }
  sample1X <- rnorm(n1X, 0, 1)
  sample2X <- rnorm(n2X, 0, 1)
  sample3X <- rnorm(n3X, 0, 1)
  p_val1[i] <- wilcox.test(sample1X,sample1Y)$p.value
  p_val2[i] <- wilcox.test(sample2X,sample2Y)$p.value
  p_val3[i] <- wilcox.test(sample3X,sample3Y)$p.value


}

sum(p_val1 <= .05)/N
```

```
## [1] 0.1385
```

```
sum(p_val2 <= .05)/N
```

```
## [1] 0.1279
```

```
sum(p_val3 <= .05)/N
```

```
## [1] 0.1321
```

As seen the rejection rate is incorrect for multiple sample sizes.

Question 6 (6 points)

Construct three distributions F, G, and H such that for $X \sim F$, $Y \sim G$, and $Z \sim H$, we have

$$P(X > Y) > 0.5$$

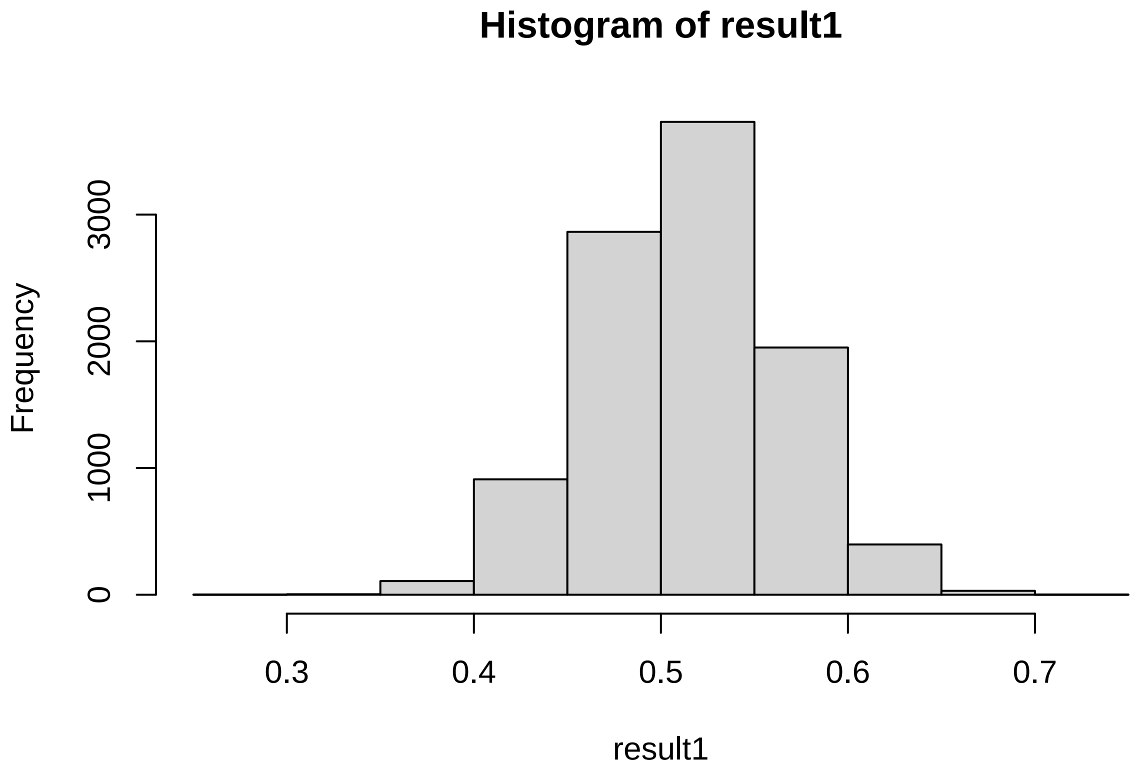
$$P(Y > Z) > 0.5$$

$$P(Z > X) > 0.5$$

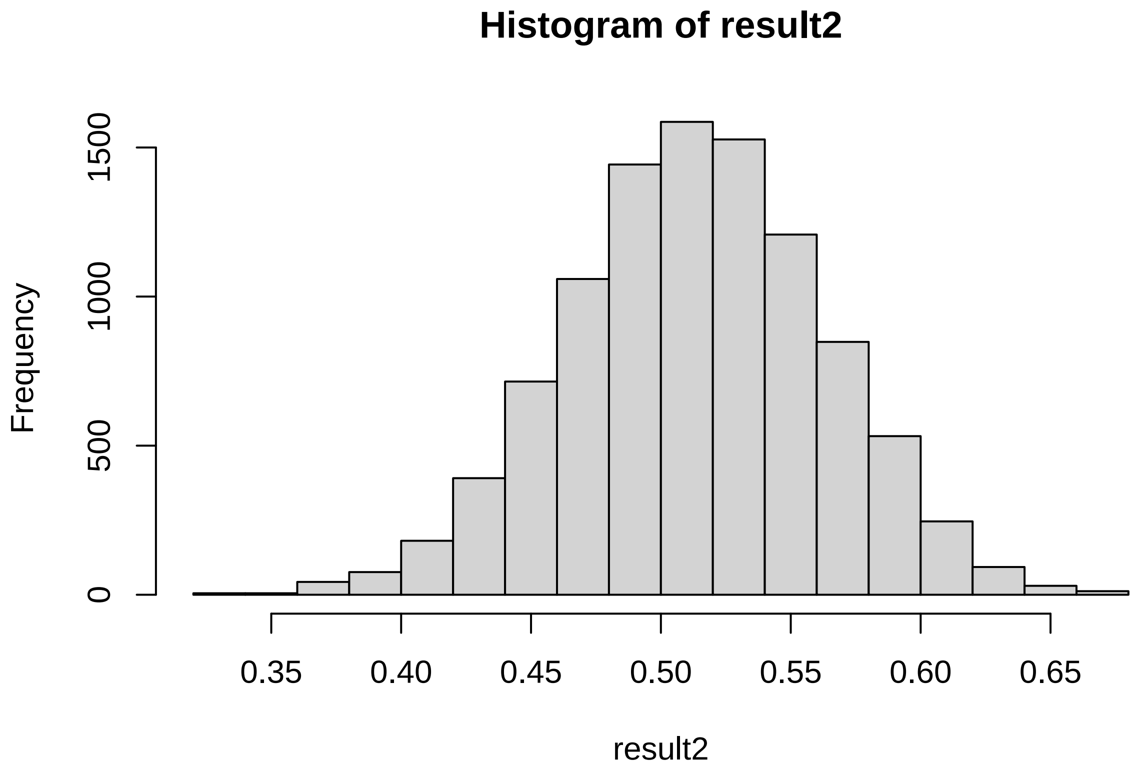
> Hints: One way to accomplish this is to have G skewed one way and H skewed the other way, and then shift them up and down the real number line (by adding or subtracting a constant). So for instance, maybe $F = N(0, 1)$, $G = \text{Chi-squared}(3) + c_1$, $H = -\text{Chi-squared}(3) + c_2$. Other skewed distributions would work as well or possibly better.

How large can you get each of the probabilities $P(X > Y)$, $P(Y > Z)$, and $P(Z > X)$?

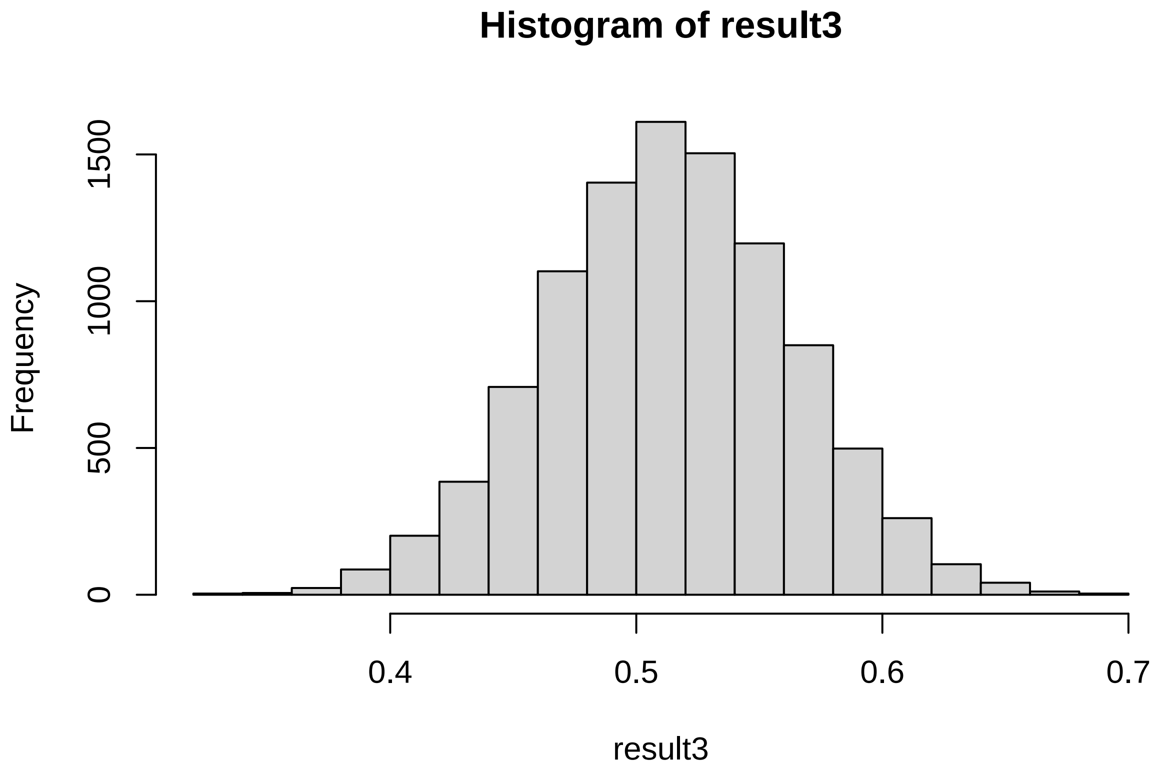
```
N <- 10000
n <- 100

result1 <- rep(0,N)
result2 <- rep(0,N)
result3 <- rep(0,N)

for( i in 1:N ) {
  x <- rnorm(n, 0, 1)
  y <- rchisq(n, 3) - 2.6
  z <- -rchisq(n, 3) + 2.6

  result1[i] <- sum(x > y)/n
  result2[i] <- sum(y > z)/n
  result3[i] <- sum(z > x)/n

}

hist(result1)
```



```
hist(result2)
```



```
hist(result3)
```



```
mean(result1)
```

```
## [1] 0.518998
```

```
mean(result2)
```

```
## [1] 0.518402
```

```
mean(result3)
```

```
## [1] 0.518509
```

I decided to simulate to answer this question. I simply had to simulate to find c that all statements hold true. I tried this starting at 0, and then going up by integers and decimals. It seems like around 2.6 is a good answer.

Other tests

Question 7 (12 points)

The following data are based on a comparison between two diets: a low-carb diet (similar to the Atkins diet), and a Mediterranean diet. The study ran for two years, with 10 subjects randomly assigned to the low-carb diet, and 12 subjects random assigned to the Mediterranean diet. Each subject's change in weight after 24 months was recorded (initial weight – final weight) in kg, as shown in the table below:

Mediterranean Diet	-6.7	-1.9	-0.6	0.3	0.9	1.0	2.4	4.7	7.8	9.7	10.3	12.0
Low-Carb Diet	-5.6	0.3	2.0	2.0	4.6	5.0	6.5	7.5	11.8	17.0		

Part A. (4 points) Perform a level 0.05 test that the median weight change for the two diets is equal. Give the test statistic and state your conclusion.

```
x <- c(-6.7, -1.9, -0.6, 0.3, 0.9, 1.0, 2.4, 4.7, 7.8, 9.7, 10.3, 12.0)
y <- c(-5.6, 0.3, 2.0, 2.0, 4.6, 5.0, 6.5, 7.5, 11.8, 17.0)


new <- c(x,y)
median <- median(new)

sampx <- c(sum(x>median),sum(x<=median))
sampy <- c(sum(y>median),sum(y<=median))
table <- rbind(sampx,sampy)


p_c <- (sampx[1] + sampy[1])/(length(x)+length(y))
p_x <- sampx[1]/length(x)
p_y <- sampy[1]/length(y)

z <- (p_x - p_y)/sqrt(p_c*(1-p_c)*((1/length(x))+(1/length(y))))
z
```

```
## [1] -0.8563488
```

```
p <- 2*(1-pnorm(abs(z)))
p
```

```
## [1] 0.3918049
```

I got a test statistic of -.85634. I would say that we would fail to reject that the medians are the same.

Part B. (4 points) Perform the Wilcoxon Rank-Sum test to test that the probability that a person loses more weight on the low-carb diet than on the Mediterranean diet is 1/2. Give the test statistic and state your conclusion.

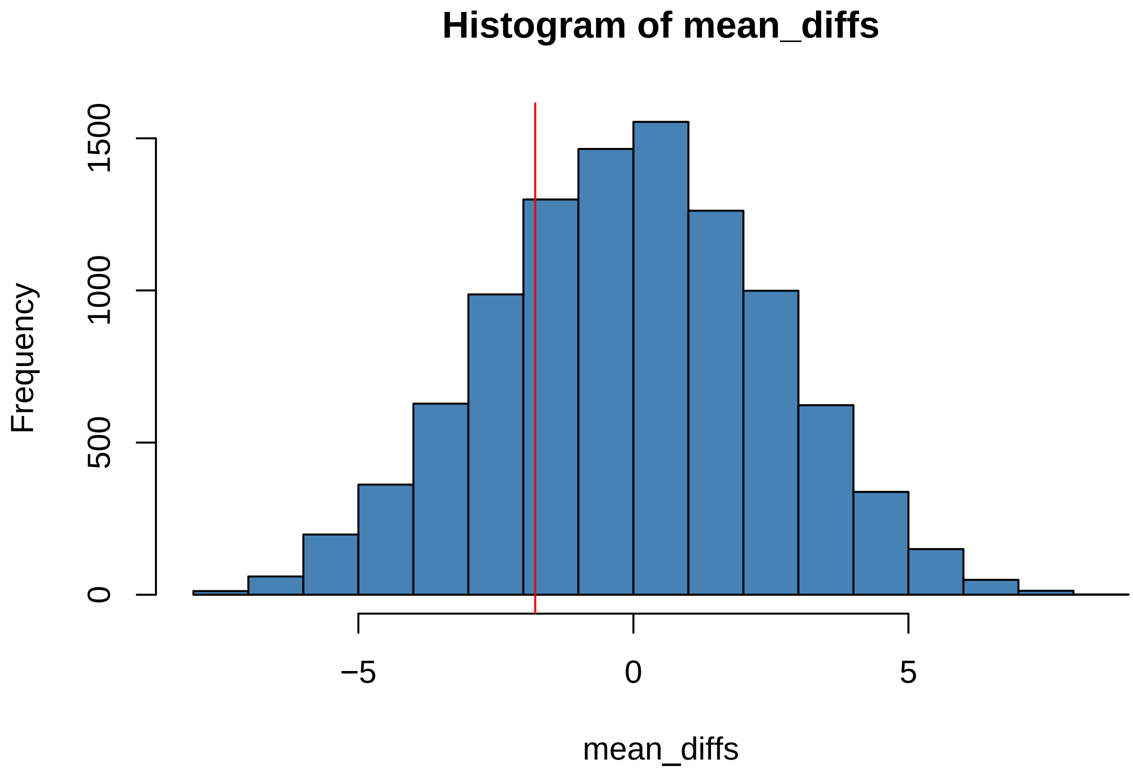
```
wilcox.test(x,y)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 49.5, p-value = 0.5094
## alternative hypothesis: true location shift is not equal to 0
```

I got a W = 49.5 and I would fail to reject the null hypothesis that the probability that a low car diet loses more weight than a Mediterranean diet is 1/2.

Part C. (4 points) Perform a permutation test, using the difference in sample means as your test statistic, to test that the mean weight change for the two diets is equal. Give a two-sided p-value, and be sure to include your R code and output.

```
x <- c(-6.7, -1.9, -0.6, 0.3, 0.9, 1.0, 2.4, 4.7, 7.8, 9.7, 10.3, 12.0)
y <- c(-5.6, 0.3, 2.0, 2.0, 4.6, 5.0, 6.5, 7.5, 11.8, 17.0)
actual_diff <- mean(x) - mean(y)
x <- data.frame(x)
x$label <- "M"
colnames(x) <- c("difference", "label")
y <- data.frame(y)
y$label <- "L"
colnames(y) <- c("difference", "label")
data <- rbind(x,y)
N <- 10000
mean_diffs <- rep(0,N)
data2 <- data
for ( i in 1:N ) {
  data2$label <- sample(data$label)

  mean_diffs[i] <- mean(data2$difference[data2$label == "M"]) - mean(data2$difference[data2$label == "L

}

hist(mean_diffs, col = "steelblue")
# Include a vertical line for our observed value
abline(v = actual_diff, col = "red")
```



```
mean(mean_diffs < -abs(actual_diff)) + mean(mean_diffs > abs(actual_diff))
```

```
## [1] 0.4909
```

```
t.test(difference ~ label, data)
```

```
##
##  Welch Two Sample t-test
##
## data:  difference by label
## t = 0.69685, df = 18.438, p-value = 0.4946
## alternative hypothesis: true difference in means between group L and group M is not equal to 0
## 95 percent confidence interval:
##  -3.587382  7.157382
## sample estimates:
## mean in group L mean in group M
##           5.110           3.325
```

Fairly close to t.test. In this case it also agrees to t.test. p value of .4928. We would fail to reject.

Part D. (4 points) Perform Levene's test to test that the variances are equal in the two groups.

```
x <- c(-6.7, -1.9, -0.6, 0.3, 0.9, 1.0, 2.4, 4.7, 7.8, 9.7, 10.3, 12.0)
y <- c(-5.6, 0.3, 2.0, 2.0, 4.6, 5.0, 6.5, 7.5, 11.8, 17.0)

u_vals <- (x - mean(x))^2
v_vals <- (y - mean(y))^2

t.test(u_vals, v_vals)
```

```
##
##  Welch Two Sample t-test
##
## data:  u_vals and v_vals
## t = -0.31607, df = 14.466, p-value = 0.7565
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -45.62895  33.87690
## sample estimates:
## mean of x mean of y
##  29.24688  35.12290
```

Again, we would fail to reject the null that the variances are equal in the two groups.