Homework 1, Stat 250

Victor Huang

Q1. (Exercise 1.1)

a) The population is all high school students, the sample is the 1500 high school students being surveyed. The 47% is a statistic since it measures the sample and not the population

b) The population is all Americans in 2010 (The US Census), the 9.6% is a parameter since it measures the population and not a sample.

c) The population is all the rosters for the NBA teams for the 2006-2007 season, the average height of 78.93 in for the players is a parameter since it measures the population and not a sample.

d) The population is all American adults (ages 18 and older) and the sample if the 2106 national adults choosen for the poll. The 19% percentage is a statistic since it measures the sample and not the population.

Q2. (Exercise 1.3)

a) The five days of strict rest and the usual care of 1-2 dats of rest followed by stepwise return to activity

b) No, the researcher knows who is receiving the treatment

c) No, they can not.

d) No, we can not.

Q3. (Exercise 1.5)

a) Since no treatment was given, this is an observational study

b) No, they can not. They neglected to consider other random variables (ex. genetic predisposition).

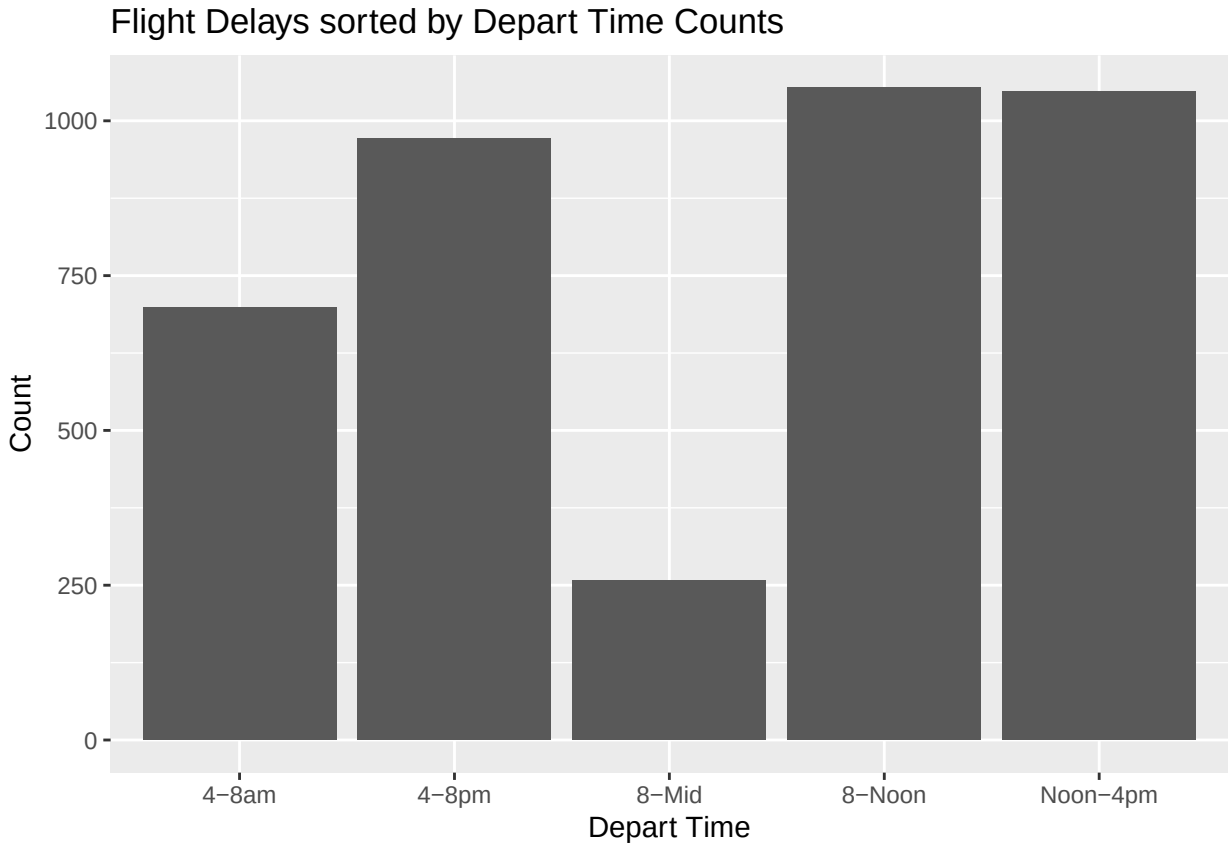
Q4. (Exercise 2.4)

(a)

```
FlightDelays <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/FlightDelays.csv")
tally(~DepartTime, data = FlightDelays)
```

```
## DepartTime
##    4-8am    4-8pm    8-Mid   8-Noon Noon-4pm
##      699      972      257     1053     1048
```

```
gf_bar(~DepartTime, data = FlightDelays, xlab = "Depart Time", ylab = "Count", title = "Flight Delays s
```



(b)

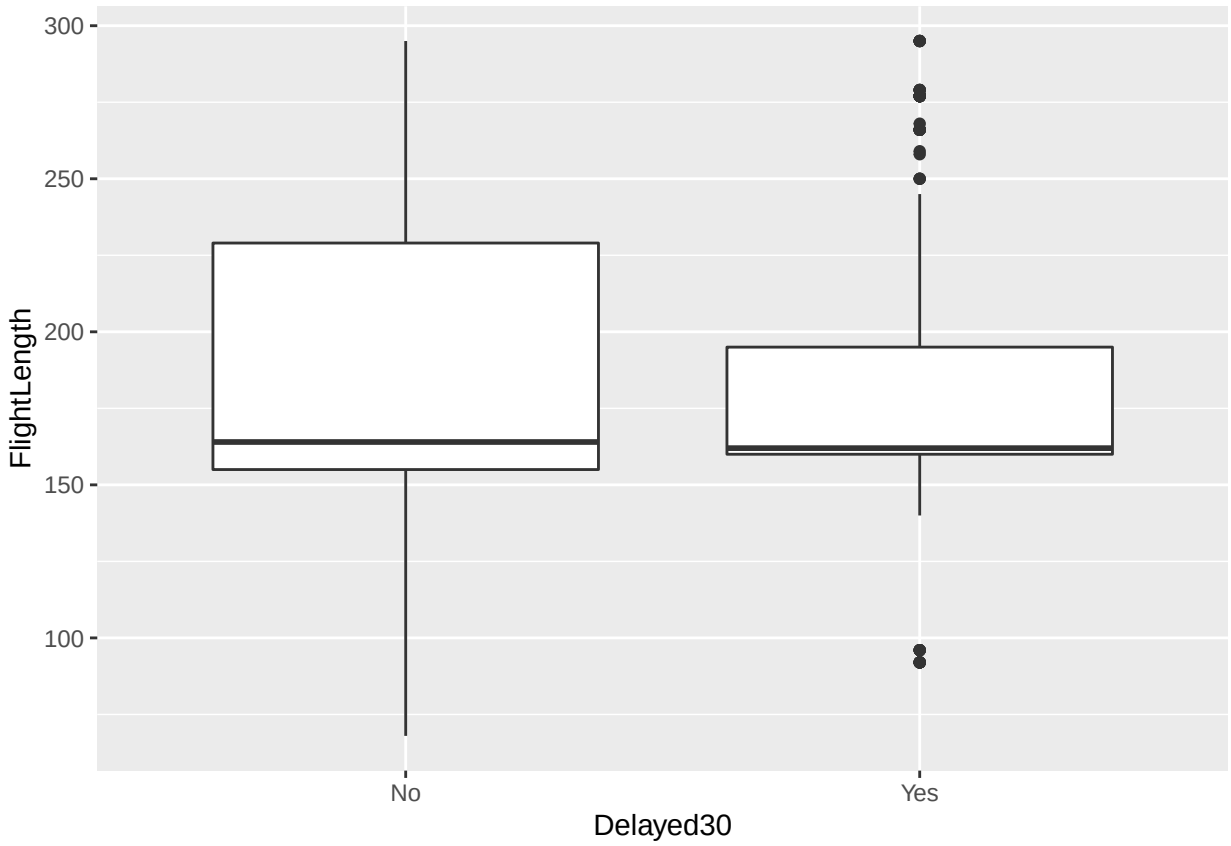
Monday has a proportion of 0.10. Tuesday has a proportion of 0.16 Wednesday has a proportion of 0.13 Thursday has a proportion of 0.22 Friday has a propotion of 0.24 Saturday has a proportion of 0.08 Sunday has a proportion of 0.08

```
tally(Day ~ Delayed30, data = FlightDelays, format = "proportion")
```

```
##      Delayed30
## Day           No        Yes
##   Fri 0.14364802 0.24120603
##   Mon 0.16579254 0.10217755
##   Sat 0.11829837 0.07872697
##   Sun 0.14772727 0.07370184
##   Thu 0.12645688 0.22110553
##   Tue 0.15588578 0.15577889
##   Wed 0.14219114 0.12730318
```

(c)

```
gf_boxplot(FlightLength ~ Delayed30, data = FlightDelays)
```



(d)

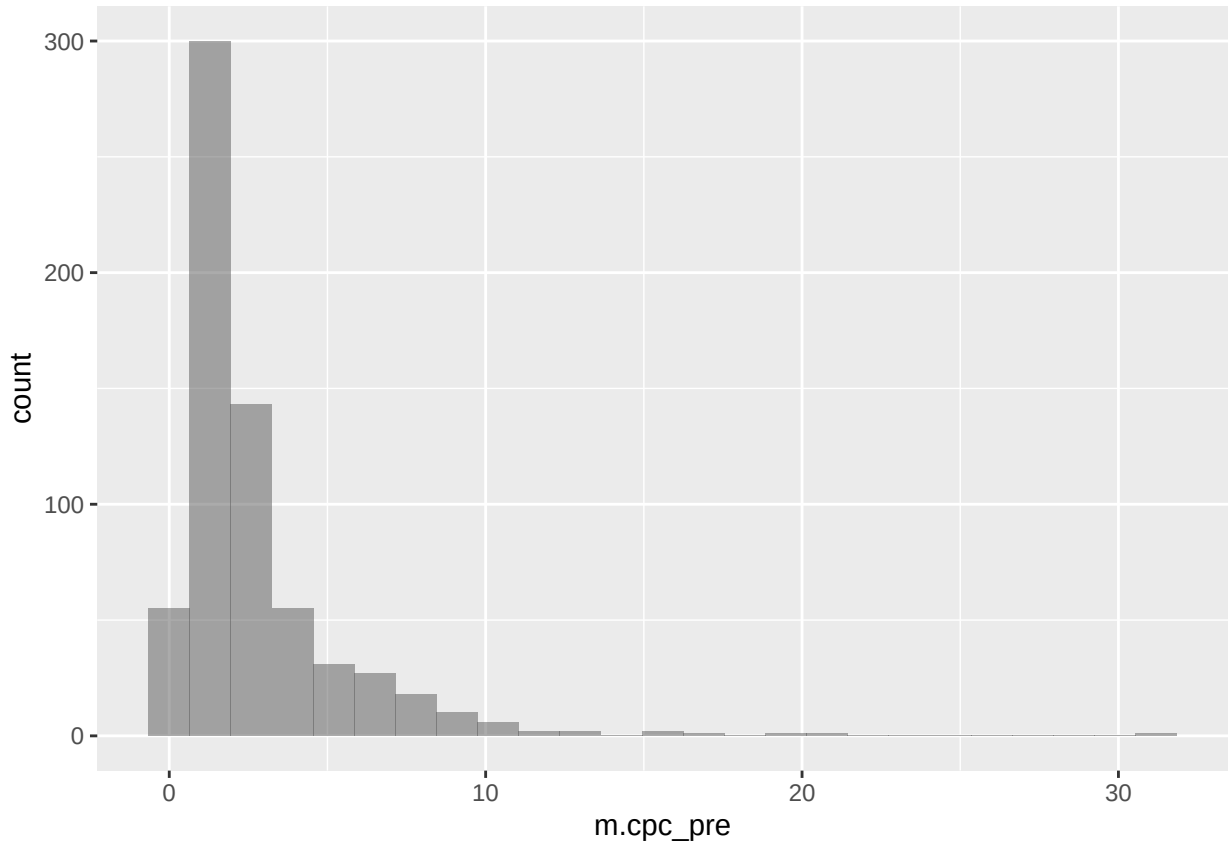
Since the mean is approximately the same, I don't think there is a realtionship between the length of flight

Q5. (Exercise 2.8)

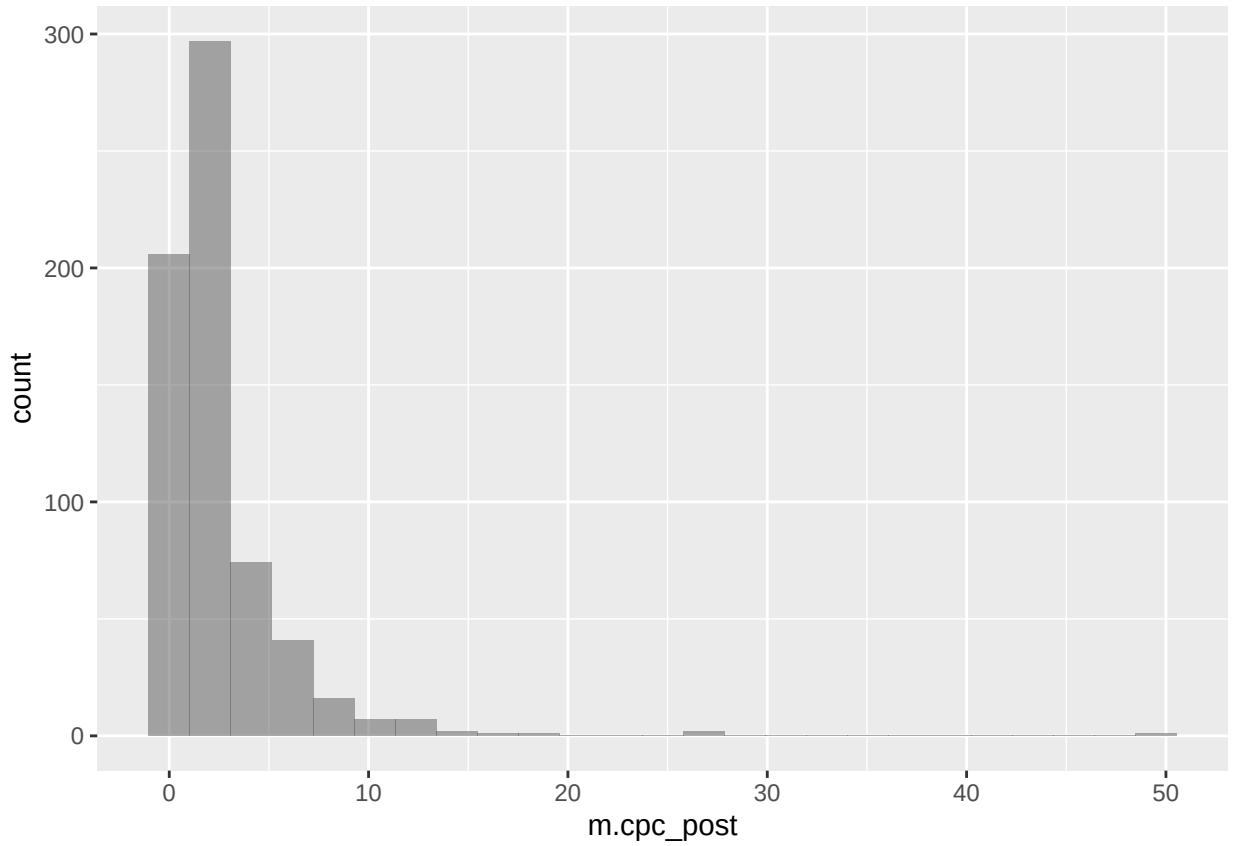
(a)

Both histograms are strongly skewed the right.

```
MobileAds <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/MobileAds.csv")
gf_histogram(~ m.cpc_pre, data = MobileAds)
```



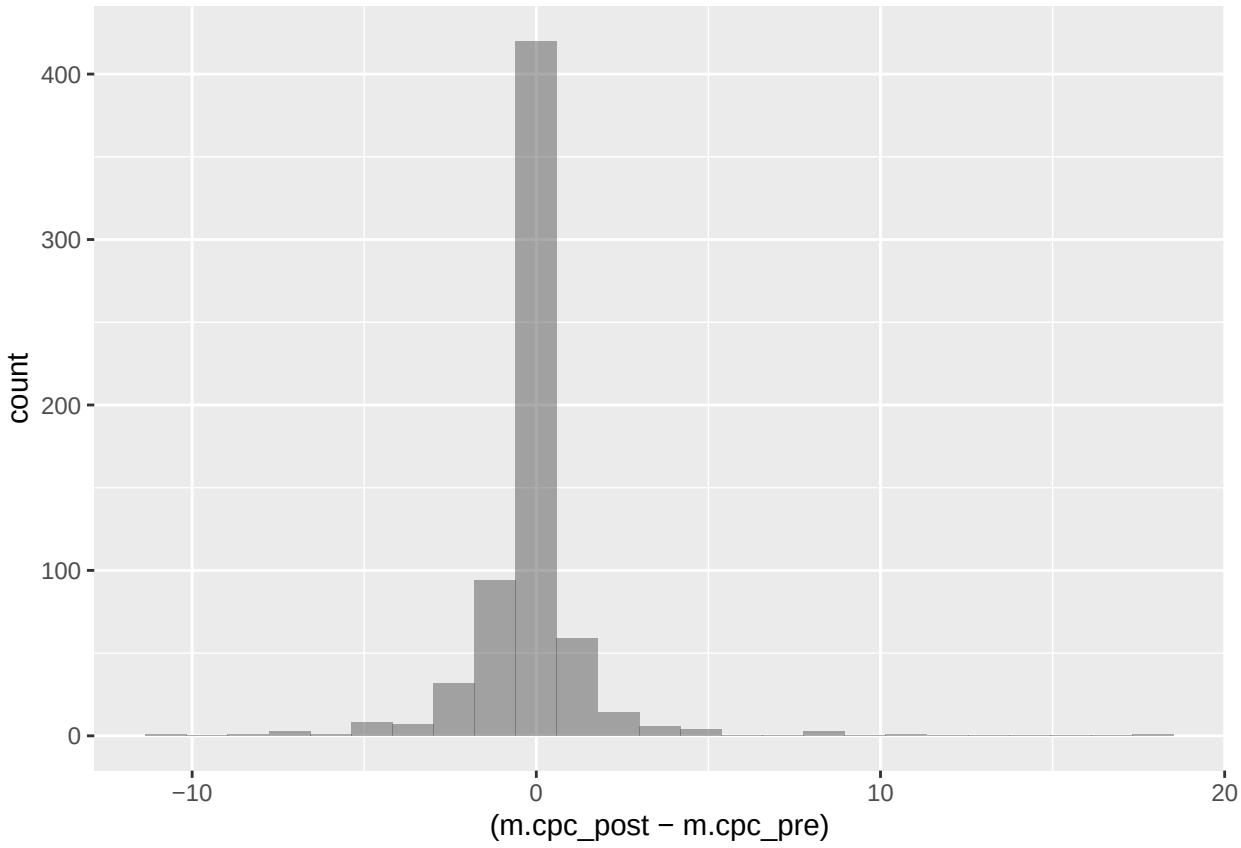

```
gf_histogram(~ m.cpc_post, data = MobileAds)
```



(b)

The distribution of the histogram created by the differnece of the post and pre is normally distributed and bell shaped.

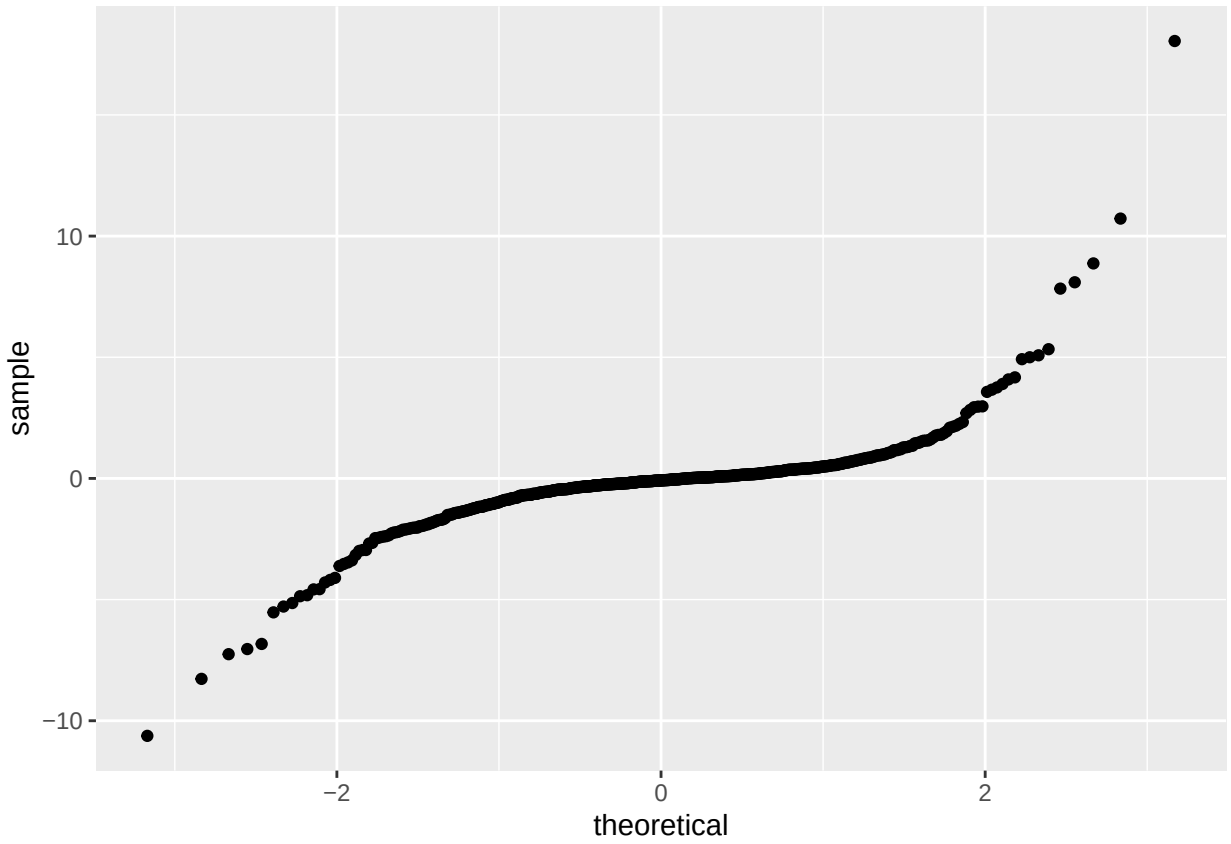
```
gf_histogram(~(m.cpc_post - m.cpc_pre), data = MobileAds)
```



(c)

The concentration of points in the middle is higher, there appears to be little to no outlier cases, and it appears to follow the "s" curve. As such, I would say the normal quantile plot is normally distributed.

```
gf_qq(~(m.cpc_post - m.cpc_pre), data = MobileAds)
```



Q6. (Exercise 2.10 a)

$$f(x) = \lambda e^{-\lambda x}$$

$$P[x \leq x] = \int_0^x \lambda e^{-\lambda x}$$

$$P[x \leq x] = e^{-0} - e^{-\lambda x}$$

$$P[x \leq x] = 1 - e^{-\lambda x}$$

$$P[x \leq m] : 1 - e^{-\lambda x} = \frac{1}{2}$$

$$e^{-\lambda m} = \frac{1}{2}$$

$$-\lambda m = ln^{\frac{1}{2}}$$

$$m = \frac{ln(2)}{\lambda} = 2^{\frac{1}{\lambda}}$$

Median: $\frac{ln(2)}{\lambda} = 2^{\frac{1}{\lambda}}$ First Quantile: $\frac{ln(\frac{4}{3})}{\lambda} = \frac{4}{3}^{\frac{1}{\lambda}}$ Third Quantile: $\frac{ln(4)}{\lambda} = 4^{\frac{1}{\lambda}}$

Q7. (Exercise 2.12)

a. $P_{30} = 1.085191$
$P_{60} = 14.3069$

b. $P_{10} = -16.00965$
$P_{90} = 66.00965$

c. $P_{75} = 46.58367$

```
qnorm(0.3,10,17)
```

```
## [1] 1.085191
```

```
qnorm(0.6,10,17)
```

```
## [1] 14.3069
```

```
qnorm(0.1,25,32)
```

```
## [1] -16.00965
```

```
qnorm(0.9,25,32)
```

```
## [1] 66.00965
```

```
qnorm(0.75,25,32)
```

```
## [1] 46.58367
```

Q8. (Exercise 3.8)

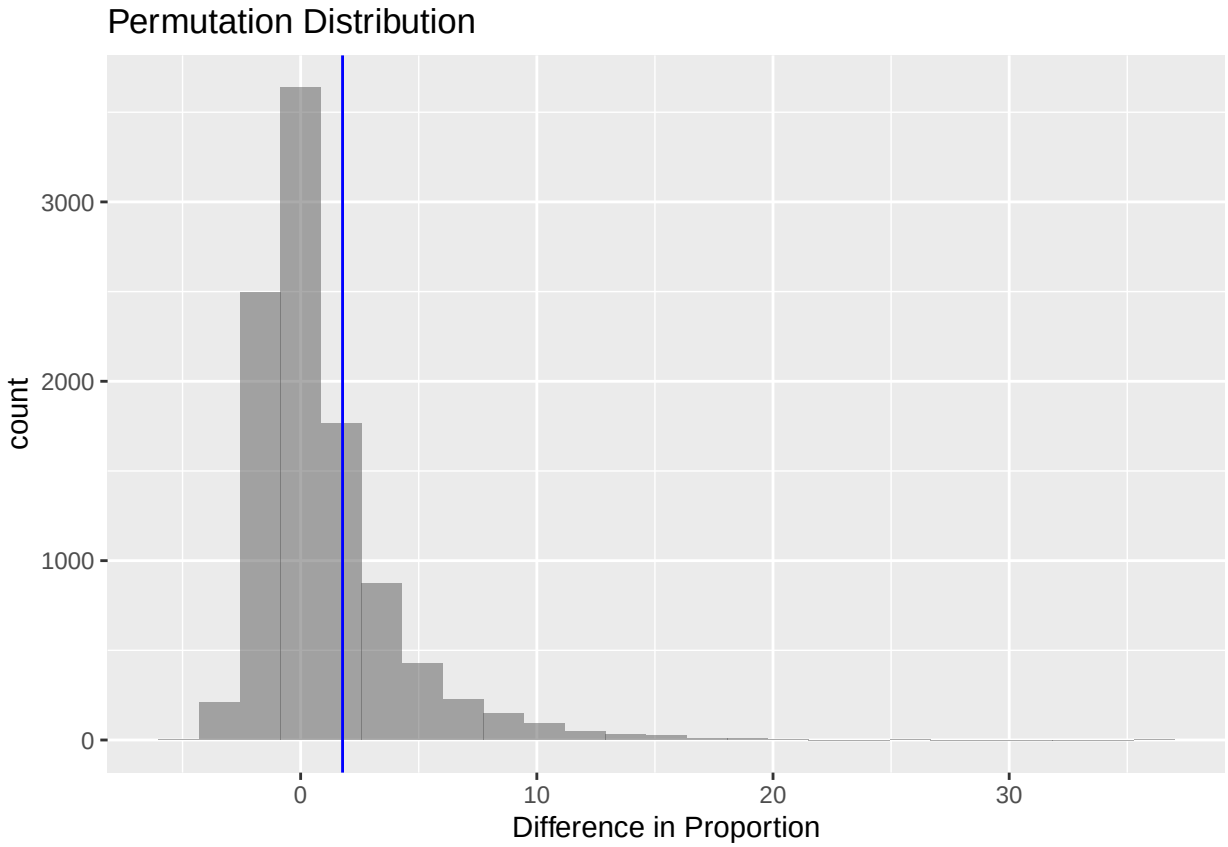
(a)

```
meancarrier <- mean(~ Delay | Carrier, data = FlightDelays, trim = 0.25)
```

(b)

Our null hypothesis is that therre exists no difference in trimmed means between delayed flights The p-value is 0.5308, since the p-value is so high, we fail to reject our null hypothesis. As such, we can conclude that there is no statistically discernable diffference in trimmed means between UA and AA.

```
observed <- meancarrier["UA"] - meancarrier["AA"]
N <- 10^4 - 1
result <- numeric(N)
for (i in 1:N){
  index <- sample(nrow(FlightDelays), size = 24, replace = FALSE)
  result[i] <- mean(FlightDelays$Delay[index], trim = 0.25) - mean(FlightDelays$Delay[-index], trim = 0
}
gf_histogram(~result, title = "Permutation Distribution", xlab = "Difference in Proportion") %>% gf_vli
```




```
2*(sum(result >= observed) + 1)/(N + 1)
```

```
## [1] 0.5096
```

Q9. (Exercise 3.6)

(a)

The proportion of flights that are 20 mins late for UA is 0.222 and 0.171 for AA. The null hypothesis is that there is no difference in proportion, and the alternative is that there is a difference. Our p-value is 8e-04, it is small enough to be statically discernible. As such, we agree with our alternative hypithesis and conclude that there is a statistically discernible difference in proportions between the flights that are delayed 20 mins or more between UA and AA.

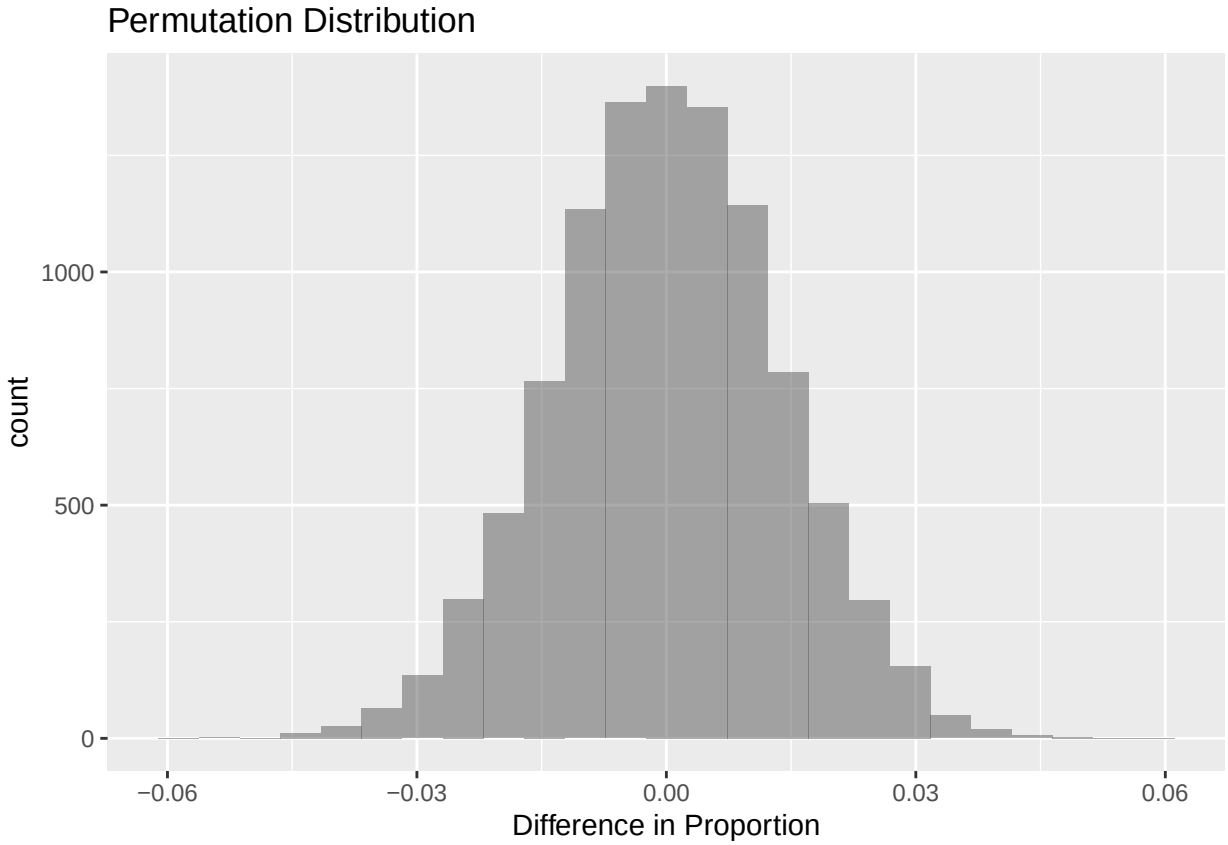
```
UA <- filter(FlightDelays, Carrier == "UA")
AA <- filter(FlightDelays, Carrier == "AA")
propUA <- prop(~Delay >= 20, data = UA)
propUA
```

```
## prop_TRUE
##  0.222618
```

```
propAA <- prop(~Delay >= 20, data = AA)
propAA
```

```
## prop_TRUE
## 0.1713696
```

```
observed1 <- nrow(UA$Delays >= 20) / nrow(UA) - nrow(AA$Delay >= 20) / nrow(UA)
N <- 10^4 - 1
result1 <- numeric(N)
for (i in 1:N){
  index <- sample(nrow(FlightDelays), size = 1123, replace = FALSE)
  result1[i] <- prop(FlightDelays$Delay[index] >= 20) - prop(FlightDelays$Delay[-index] >= 20)
}
gf_histogram(~result1, title = "Permutation Distribution", xlab = "Difference in Proportion") %>% gf_vli
```




```
2*(sum(result1 >= observed1) + 1)/(N + 1)
```

```
## [1] 2e-04
```

(b)

Our null hypthoesis is that there is no difference in variance, our alternative hypothesis is that there is a difference. We get a p-value of 0.2954, which is higher than the treshold. As such, we fail to reject the null hypothesis and concclude that there is no discernable difference in variance between the two carriers.

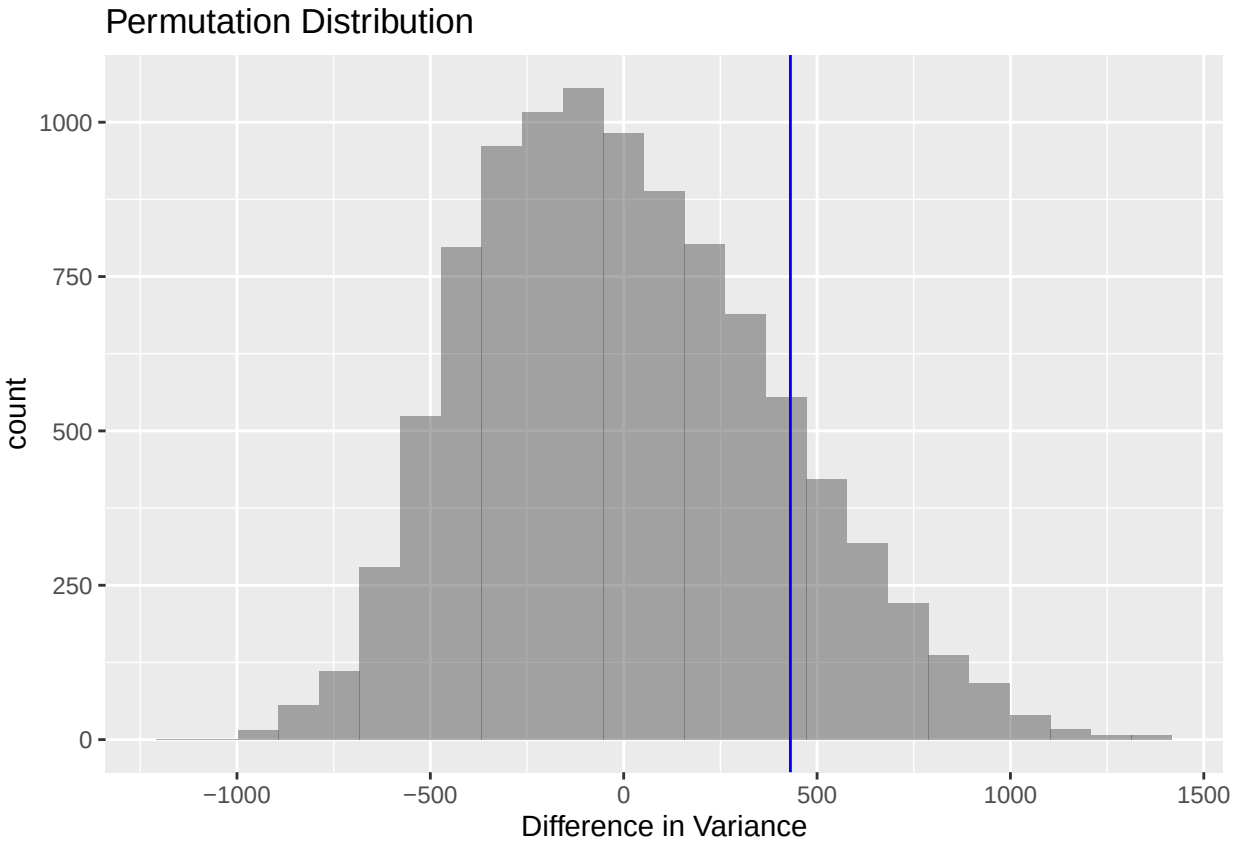
```
UA <- filter(FlightDelays, Carrier == "UA")
AA <- filter(FlightDelays, Carrier == "AA")
varUA <- var(~Delay, data = UA)
varUA
```

```
## [1] 2037.525
```

```
varAA <- var(~Delay, data = AA)
varAA
```

```
## [1] 1606.457
```

```
observed2 <- varUA - varAA
N <- 10^4 - 1
result2 <- numeric(N)
for (i in 1:N){
  index <- sample(nrow(FlightDelays), size = 1123, replace = FALSE)
  result2[i] <- var(FlightDelays$Delay[index]) - var(FlightDelays$Delay[-index])
}
gf_histogram(~result2, title = "Permutation Distribution", xlab = "Difference in Variance") %>% gf_vlin
```




```
2*(sum(result2 >= observed2) + 1)/(N + 1)
```

```
## [1] 0.2952
```

Q10. (Exercise 3.15)

(a)

This is because each cereals price (variable) is being collected at two different retail stores (Walmart and Shoprite). As such, they are linked and related, making them linked data pair.

(b)

The summary statistic for each store is shown below

```
Groceries <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Groceries.csv")
summary(Groceries$Walmart)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.760   2.340   2.706   2.955   6.980
```

```
summary(Groceries$Target)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.990   1.827   2.545   2.762   3.140   7.990
```

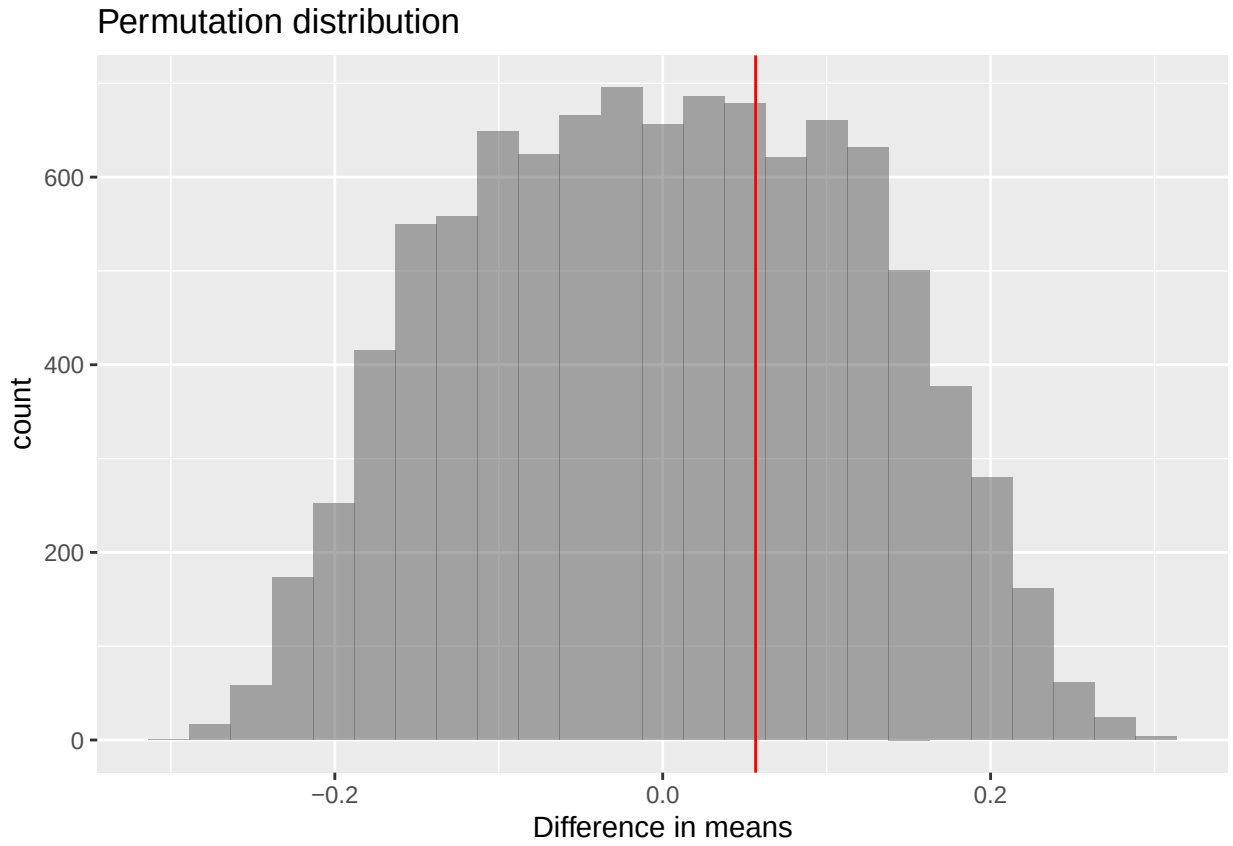
(c)

OUr null hypothesis is that there is no difference while our alternative hypothesis is that there exists a difference. Since our p-value is so small (p-value 0.7034), we failt to reject the null hypothesis and conclude that there is no statistically discernable difference between price points.

```
difference <- Groceries$Target - Groceries$Walmart
observed3 <- mean(difference)


N <- 10^4 - 1
result3 <- numeric(N)
n <- length(difference)
for(i in 1:N) {
index <- sample(c(-1,1), n, replace = TRUE)
result3[i] <- mean(index * difference)
  }

gf_histogram(~result3,title = "Permutation distribution",
xlab = "Difference in means") %>%
gf_vline(xintercept = ~observed3, color = "red")
```



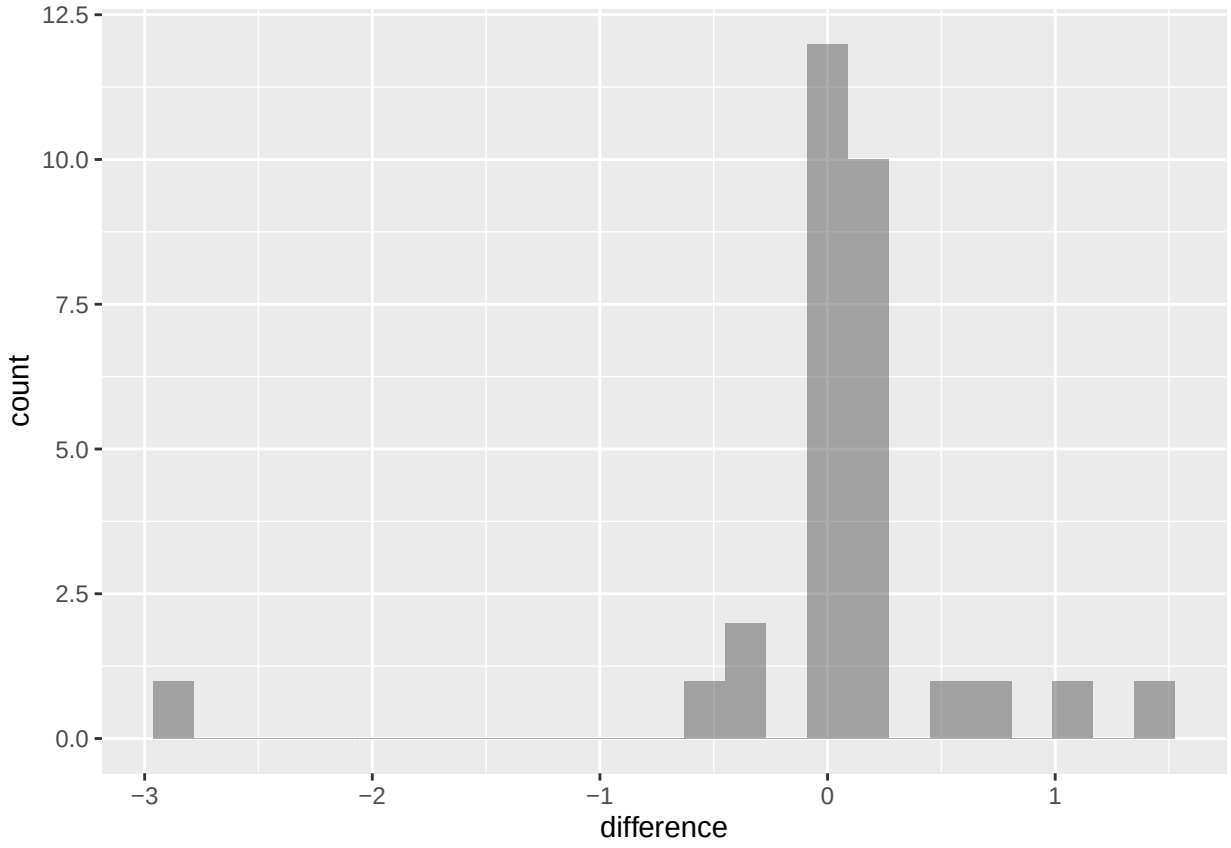

```
2*(sum(result3 >= observed3) + 1)/(N + 1)
```

```
## [1] 0.6968
```

(d)

The Quaker Oats Life cereal price difference is discernably different from other product price differences.

```
gf_histogram(~difference, data = Groceries)
```



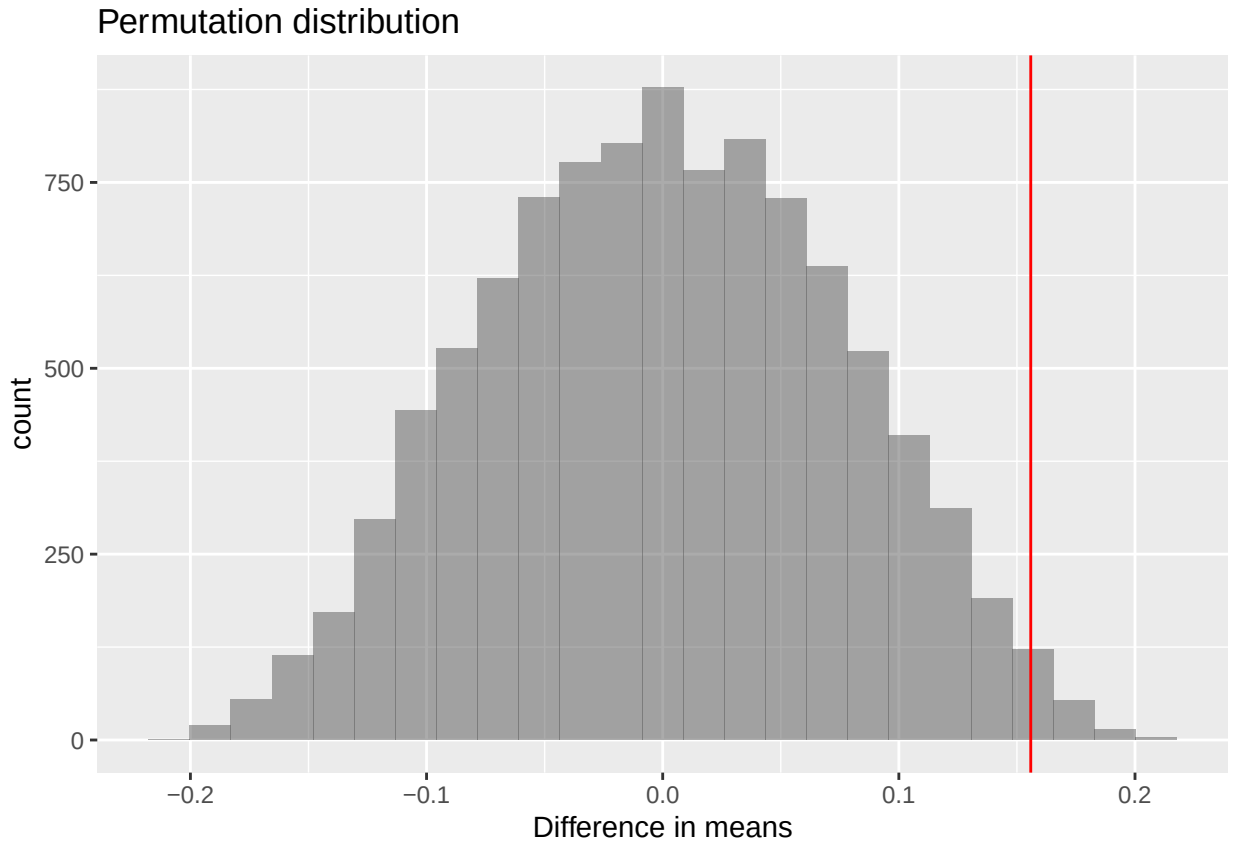
(e)

Keeping the same hypotheses and getting rid of the observation, we still have a small p-value (p-value = 0.0218). As such, out conclusion does not change as we accept the alternative hypothesis and concldue taht there does indeed exists a price point difference between Walmart and Target.

```
Groceries <- Groceries[-2,]
difference <- Groceries$Target - Groceries$Walmart
observed3 <- mean(difference)


N <- 10^4 - 1
result3 <- numeric(N)
n <- length(difference)
for(i in 1:N) {
index <- sample(c(-1,1), n, replace = TRUE)
result3[i] <- mean(index * difference)
  }

gf_histogram(~result3,title = "Permutation distribution",
xlab = "Difference in means") %>%
gf_vline(xintercept = ~observed3, color = "red")
```




```
2*(sum(result3 >= observed3) + 1)/(N + 1)
```

```
## [1] 0.0252
```