Chapter 3 Exercises

3.1

Suppose you conduct an experiment and inject a drug into three mice.

Their times for running a maze are 8, 10, 15 s; the times for two control mice are 5 and 9 s.

a.) Compute the difference in mean times between the treatment group and the control group.

```
mice.t <- c(8, 10, 15)
mice.c <- c(5, 9)

observed <- mean(mice.t) - mean(mice.c)

observed
```

```
[1] 4
```

b.) Write out all the possible permutations of these times to the two groups and calculate the diffence in means.

```
mice <- c(mice.t, mice.c)

# 5 choose 3 for treatment

treatment <- combinations(n = 5, r = 3, mice, repeats.allowed = F)

control <- matrix(nrow = 10, ncol = 2)
for( i in 1:nrow(control))
{
  control[i,] <- mice[!mice %in% treatment[i,]]
}

perms <- data.table(cbind(treatment, control))

stopifnot(nrow(perms) == choose(5, 3))

colnames(perms) <- c("D1", "D2", "D3", "C1", "C2")

perms$Xd <- (perms$D1 + perms$D2 + perms$D3) / 3
perms$Xc <- (perms$C1 + perms$C2) / 2
perms$Diff <- round(perms$Xd - perms$Xc, 2)
```

Table 1: Mice Permutations

D1	D2	D3	C1	C2	Xd	Xc	Diff
5	8	9	10	15	7.33	12.5	-5.17
5	8	10	15	9	7.67	12.0	-4.33
5	8	15	10	9	9.33	9.5	-0.17
5	9	10	8	15	8.00	11.5	-3.50
5	9	15	8	10	9.67	9.0	0.67
5	10	15	8	9	10.00	8.5	1.50
8	9	10	15	5	9.00	10.0	-1.00
8	9	15	10	5	10.67	7.5	3.17
8	10	15	5	9	11.00	7.0	4.00
9	10	15	8	5	11.33	6.5	4.83

c.) What proportion of the differences are as large or larger than the observed differences in mean times?

```
gte.observed <- perms[Diff >= observed]

pretty_kable(gte.observed, "Greater than or Equal to Observed")
```

Table 2: Greater than or Equal to Observed

D1	D2	D3	C1	C2	Xd	Xc	Diff
8	10	15	5	9	11.00	7.0	4.00
9	10	15	8	5	11.33	6.5	4.83

```
p1c <- nrow(gte.observed) / nrow(perms)
```

Proportion of differences greater than or equal to observed: 20%

d.) For each permutation, calculate the mean of the treatment group only.

What proportion of these means are as large or larger than the observed mean of the treatment group?

```
gte.t <- perms[ Xd >= mean(mice.t),]
pretty_kable(gte.t, "Mean Treatment Greater than or Equal to Observed")
```

Table 3: Mean Treatment Greater than or Equal to Observed

D1	D2	D3	C1	C2	Xd	Xc	Diff
8	10	15	5	9	11.00	7.0	4.00
9	10	15	8	5	11.33	6.5	4.83

```
p1d <- nrow(gte.t) / nrow(perms)
```

Proportion of treatment groups greater than observed: 20%

3.2

Your statistics professor comes to class with a big urn that she claims contains 9,999 blue marbels and 1 red marble.

You draw our one marble at random and finds that it is red.

Would you be willing to tell your professor that you think she is wrong about the distribution of colors?

Why or why not?

- Yes, a 1/10,000 chance is pretty rare.

What are you assuming in making your decision?

What if instead, she claims there are nine blue marbles and 1 red one (and you draw out a red marble)?

- A 1/10 chance is fairly common.

3.3

In a hypothesis test comparing two populations means, $H_0 : \mu_1 = \mu_2 \; versus \; H_A : \mu_1 > \mu_2$

a.) Which P-value, *0.03* or *0.006*, provides stronger evidence for the alternative hypothesis?

0.03 provides stronger evidence for the alternative hypothesis.

b.) Which P-value, 0.095 or 0.04, provides stronger evidence that chance alone might account for the observed result?

0.095 provides stronger evidence that chance alone is responsible for the observed result.

3.4

In the algorithms for conducting a permutation test, why do we add 1 to the number of replications N when calculating the P-Value?

Answer: We need to account for the original observed result.

3.5

In the flight delays case study in Section 1.1, the data contain flight delays for two airlines, American Airlines and United Airlines.

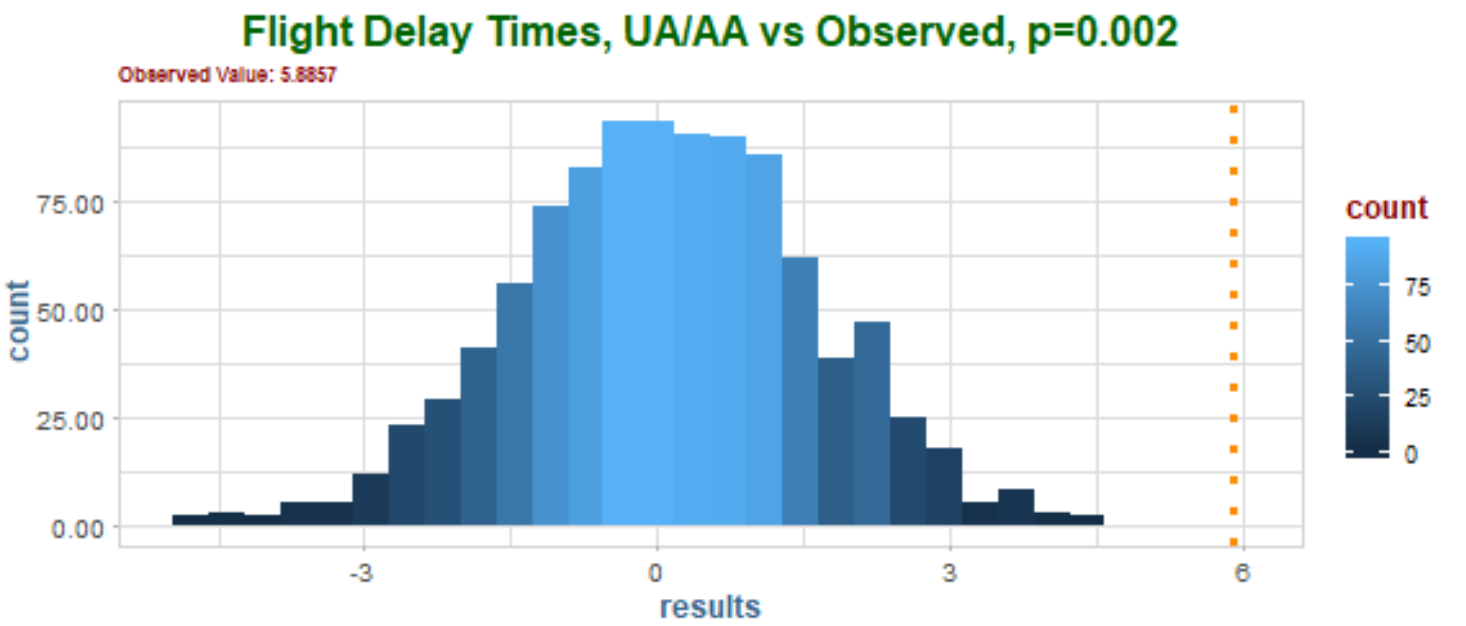
```
Flights <- data.table(read.csv(paste0(data.dir, "FlightDelays.csv"),
                               header = T))
```

a.) Conduct a two-sided permutation test to see if the difference in mean delay times between the two carriers are statistically significant.

```
Flights[, .(Delay = mean(Delay)), by = Carrier]
```

```
   Carrier    Delay
1:      UA 15.98308
2:      AA 10.09738
```

```
observed <- mean(Flights[Carrier == "UA"]$Delay) - mean(Flights[Carrier == "AA"]$Delay)

N <- 10e2 - 1
results <- numeric(N)

for(i in 1:N)
{
   index <- sample(nrow(Flights), nrow(Flights[Carrier == "UA"]), replace = F)
   results[i] <- mean(Flights[index]$Delay) - mean(Flights[-index]$Delay)
}

# two-sided test
p <- 2 * (sum(results[results >= observed]) + 1) / ( N + 1)
v <- p*(1 - p) / ( N + 1 )

ggplot(data.table(results)) +
   geom_histogram(aes(results, fill = ..count..), bins = 30) +
   geom_vline(xintercept = observed, col = "darkorange", linetype = 3, lwd = 1.2) +
   scale_y_continuous(labels = comma) +
   labs(title = paste0("Flight Delay Times, UA/AA vs Observed, p=", round(p, 5)),
        subtitle = paste0("Observed Value: ", round(observed, 4)))
```



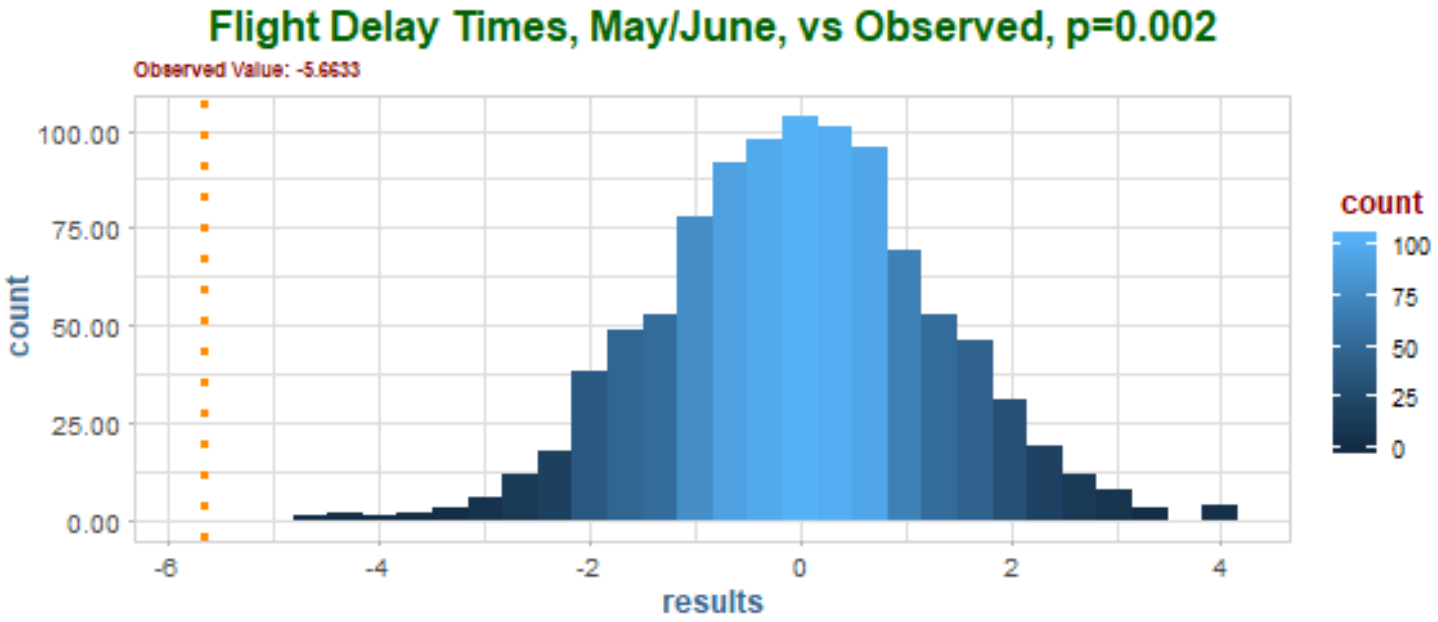
b.) The flights took place in May and June of 2009. Conduct a two-sided permutation test to see if the differences in mean delay times between two months is statistically significant.

```
Flights[, .(Delay = mean(Delay)), by = Month]
```

```
   Month     Delay
1:   May  8.884442
2:  June 14.547783
```

```
observed <- mean(Flights[Month == "May"]$Delay) - mean(Flights[Month == "June"]$Delay)

N <- 10e2 - 1
results <- numeric(N)

for(i in 1:N)
{
   index <- sample(nrow(Flights), nrow(Flights[Month == "May"]), replace = F)
   results[i] = mean(Flights[index]$Delay) - mean(Flights[-index]$Delay)
}

# two-sided test

p <- 2 * (sum(results[results <= observed]) + 1) / ( N + 1)
v <- p*(1 - p) / ( N + 1 )

ggplot(data.table(results)) +
   geom_histogram(aes(results, fill = ..count..), bins = 30) +
```

```
   geom_vline(xintercept = observed, col = "darkorange", linetype = 3, lwd = 1.2) +
   scale_y_continuous(labels = comma) +
   labs(title = paste0("Flight Delay Times, May/June, vs Observed, p=", round(p, 5)),
        subtitle = paste0("Observed Value: ", round(observed, 4)))
```



3.6

In the flight delays case study in Section 1.1, the data contains flight delays for two airlines, American and United.

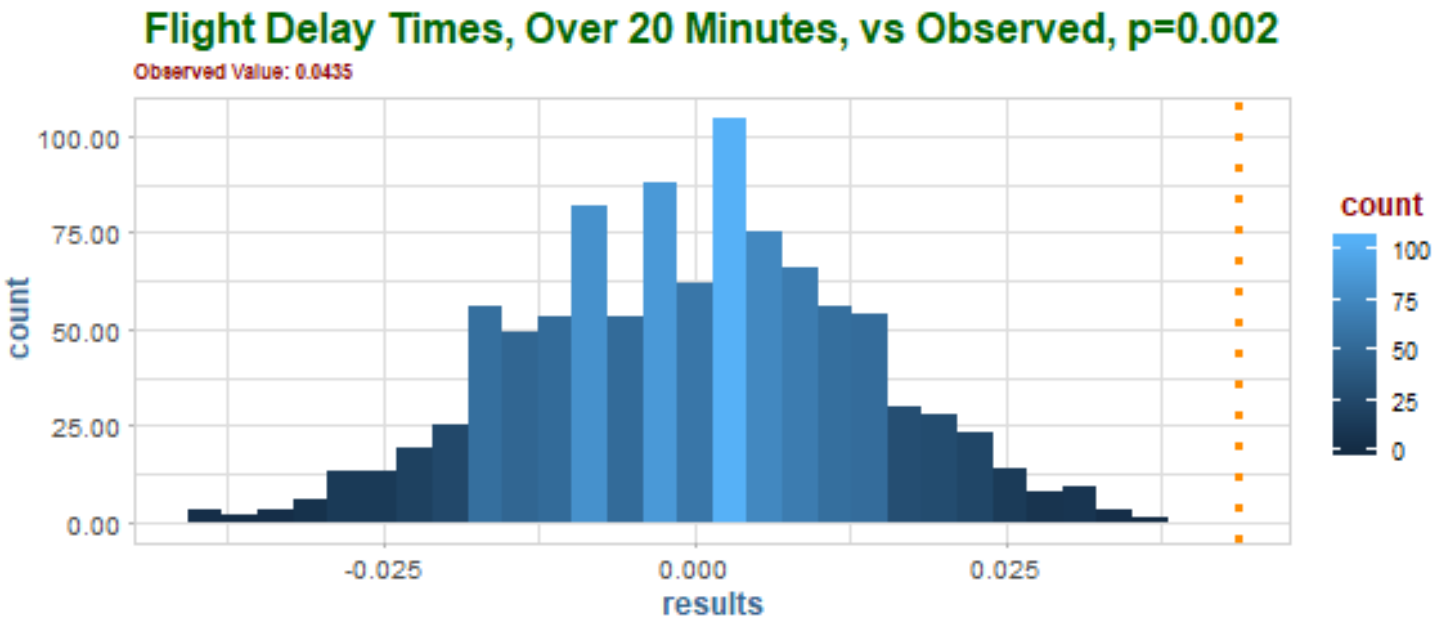
a.) Compute the proportion of times that each carrier's flight was delays more than 20 min.

```
Flights[, .(Delay20 = sum(Delay > 20) / .N), by = Carrier]
```

```
   Carrier   Delay20
1:      UA 0.2128228
2:      AA 0.1693049
```

```
observed <- as.numeric(Flights[Carrier == "UA", .(Delay = sum(Delay > 20)/.N)] - Flights[Carrie

N <- 10e2 - 1
results <- numeric(N)

for(i in 1:N)
{
   index <- sample(nrow(Flights), nrow(Flights[Carrier == "UA"]), replace = F)
```

```
   results[i] <- as.numeric(Flights[index, .(Delay = sum(Delay > 20)/.N)] - Flights[-index, .(D
}

p <- 2 * (sum(results[results >= observed]) + 1) / ( N + 1)
v <- p*(1 - p) / ( N + 1 )

ggplot(data.table(results)) +
   geom_histogram(aes(results, fill = ..count..), bins = 30) +
   geom_vline(xintercept = observed, col = "darkorange", linetype = 3, lwd = 1.2) +
   scale_y_continuous(labels = comma) +
   labs(title = paste0("Flight Delay Times, Over 20 Minutes, vs Observed, p=", round(p, 5)),
        subtitle = paste0("Observed Value: ", round(observed, 4)))
```



- Conduct a two-sided test to see if the difference in these proportions is statistically significant.

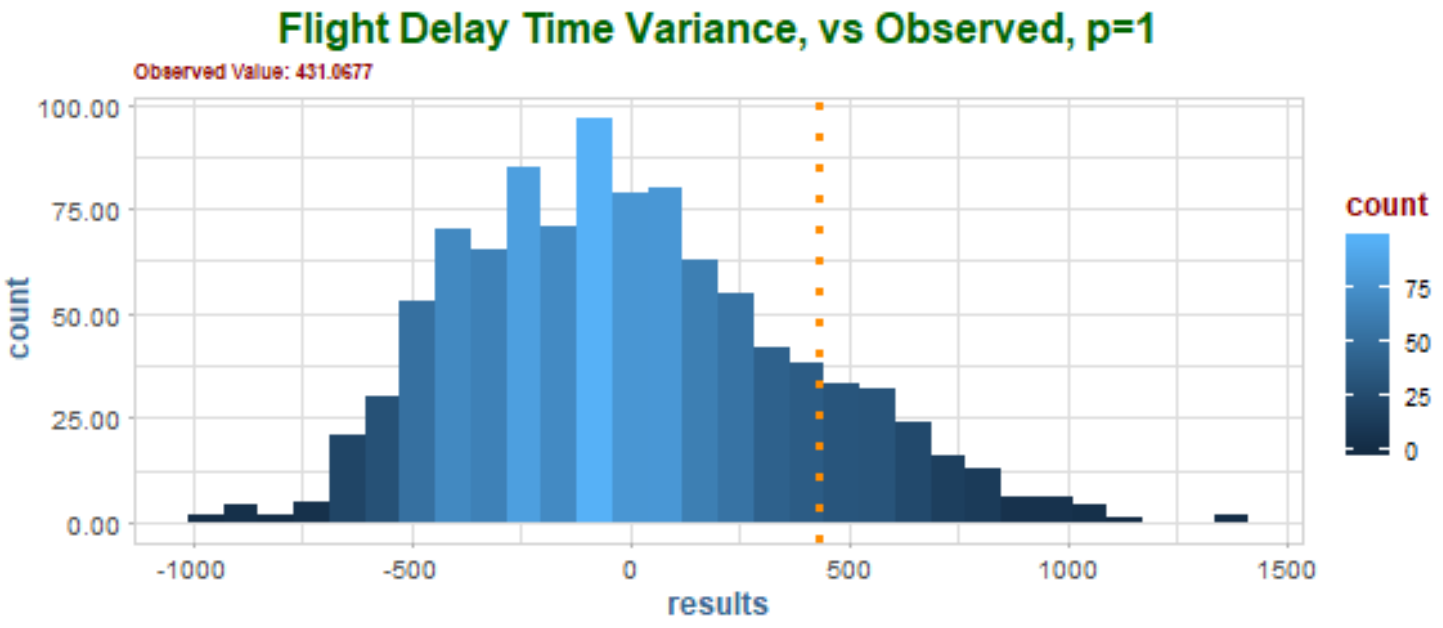
Answer: There is statistical significance with a P-value < 0.0001.

b.) Compute the variance in the flight delay lengths for each carrier.

```
Flights[, .(Variance = var(Delay)), by = Carrier]
```

```
   Carrier Variance
1:      UA 2037.525
2:      AA 1606.457
```

```
observed <- var(Flights[Carrier == "UA"]$Delay) - var(Flights[Carrier == "AA"]$Delay)

N <- 10e2 - 1
results <- numeric(N)

for(i in 1:N)
{
   index <- sample(nrow(Flights), nrow(Flights[Carrier == "UA"]), replace = F)
   results[i] <- var(Flights[index]$Delay) - var(Flights[-index]$Delay)
}

p <- min(1, 2 * (sum(results[results >= observed]) + 1) / ( N + 1))
v <- p*(1 - p) / ( N + 1 )

ggplot(data.table(results)) +
   geom_histogram(aes(results, fill = ..count..), bins = 30) +
   geom_vline(xintercept = observed, col = "darkorange", linetype = 3, lwd = 1.2) +
   scale_y_continuous(labels = comma) +
   labs(title = paste0("Flight Delay Time Variance, vs Observed, p=", round(p, 5)),
        subtitle = paste0("Observed Value: ", round(observed, 4)))
```



- Conduct a test to see if the variance for United Airlines differes from that of American Airlines.

Answer: There does not appear to be a statistically significant difference in the variance in delay times between airlines.

3.7

In the flight delays case study in Section 1.1, repeat Excercise 3.5 part **(a)** using three test statistics,

- i.) The mean of the United Airline delay times
- ii.) The sum of the United Airline delay times
- iii.) The difference in the means

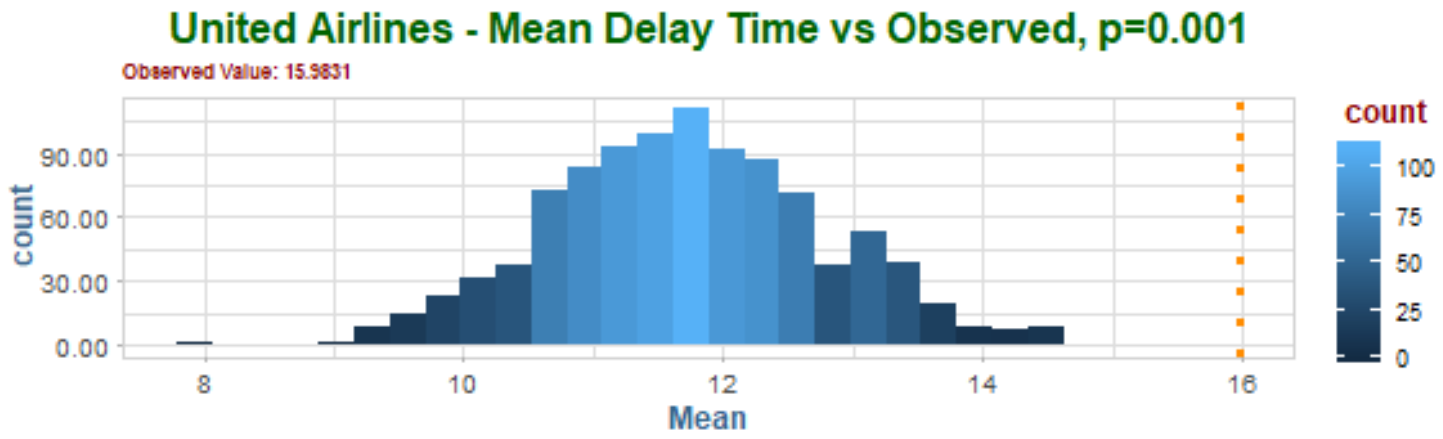
Compare the P-values.

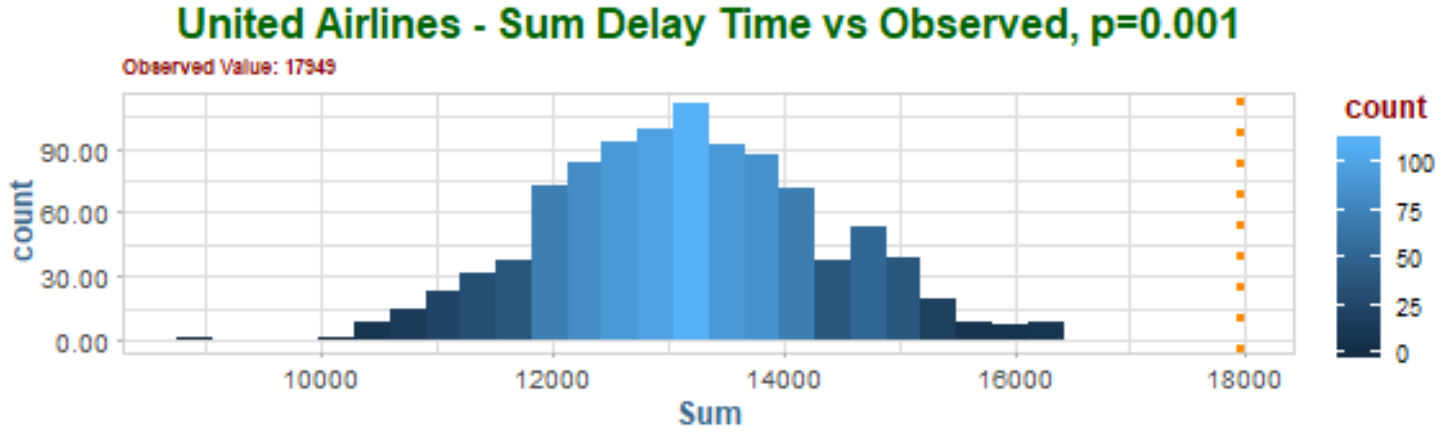
Make sure all three test statistics are computed within the same **for** loop.

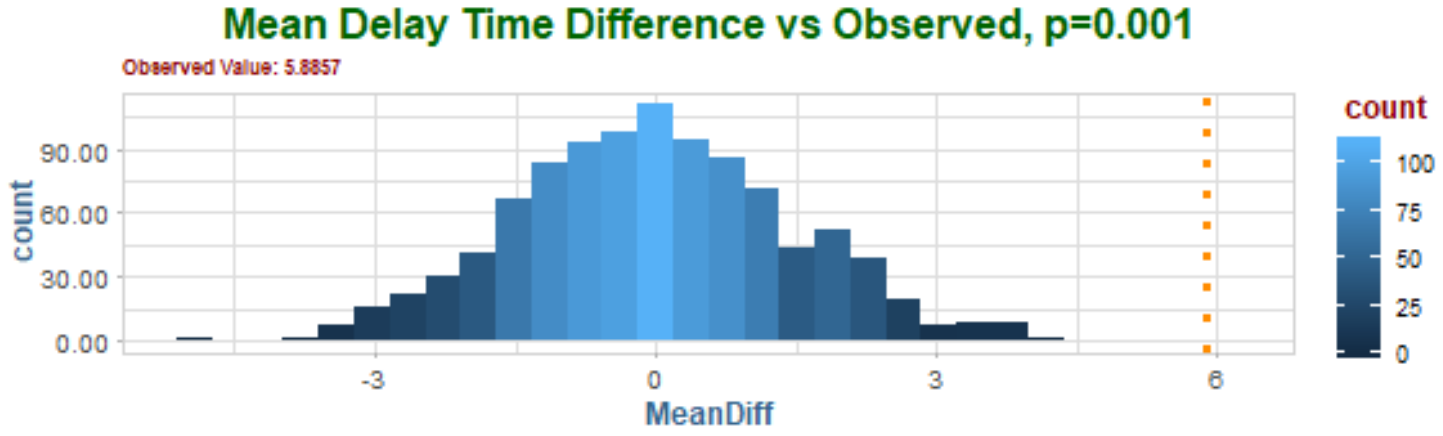
What do you observe?

```
UA.Delay <- Flights[Carrier == "UA"]$Delay
AA.Delay <- Flights[Carrier == "AA"]$Delay

observed.mean <- mean(UA.Delay)
observed.sum <- sum(UA.Delay)
observed.diff <- mean(UA.Delay) - mean(AA.Delay)

N <- 10e2 - 1
results <- matrix(nrow = N, ncol = 3)

for(i in 1:N)
{
   index <- sample(nrow(Flights), length(UA.Delay), replace = F)

   results[i, 1] <- mean(Flights[index]$Delay)
   results[i, 2] <- sum(Flights[index]$Delay)
   results[i, 3] <- mean(Flights[index]$Delay) - mean(Flights[-index]$Delay)
}

dt.results <- data.table(results)
colnames(dt.results) <- c("Mean", "Sum", "MeanDiff")

p.mean <- ( sum( dt.results$Mean >= observed.mean ) + 1 ) / ( N + 1)
p.sum <- ( sum(dt.results$Sum >= observed.sum) + 1) / ( N + 1 )
p.diff <- ( sum(dt.results$MeanDiff >= observed.diff) + 1) / ( N + 1)

p1 <- ggplot(dt.results) +
   geom_histogram(aes(Mean, fill = ..count..), bins = 30) +
   geom_vline(xintercept = observed.mean, col = "darkorange", lwd = 1.2, linetype = 3) +
   scale_y_continuous(labels = comma) +
   labs(title = paste0("United Airlines - Mean Delay Time vs Observed, p=", round(p.mean, 5)),
```

```
        subtitle = paste0("Observed Value: ", round(observed.mean, 4)))

p2 <- ggplot(dt.results) +
   geom_histogram(aes(Sum, fill = ..count..), bins = 30) +
   geom_vline(xintercept = observed.sum, col = "darkorange", lwd = 1.2, linetype = 3) +
   scale_y_continuous(labels = comma) +
   labs(title = paste0("United Airlines - Sum Delay Time vs Observed, p=", round(p.sum, 5)),
        subtitle = paste0("Observed Value: ", round(observed.sum, 4)))

p3 <- ggplot(dt.results) +
   geom_histogram(aes(MeanDiff, fill = ..count..), bins = 30) +
   geom_vline(xintercept = observed.diff, col = "darkorange", lwd = 1.2, linetype = 3) +
   scale_y_continuous(labels = comma) +
   labs(title = paste0("Mean Delay Time Difference vs Observed, p=", round(p.diff, 5)),
        subtitle = paste0("Observed Value: ", round(observed.diff, 4)))

grid.arrange(p1, p2, p3, nrow = 3)
```


United Airlines - Mean Delay Time vs Observed, p=0.001
Observed Value: 15.9831
count
90.00
60.00
30.00
0.00
8
10
12
14
16
Mean
count
100
75
50
25
0


United Airlines - Sum Delay Time vs Observed, p=0.001
Observed Value: 17949
count
90.00
60.00
30.00
0.00
10000
12000
14000
16000
18000
Sum
count
100
75
50
25
0


Mean Delay Time Difference vs Observed, p=0.001
Observed Value: 5.8857
count
90.00
60.00
30.00
0.00
-3
0
3
6
MeanDiff
count
100
75
50
25
0

3.8

In the flight delays case study in Section 1.1,

a.) Find the trimmed mean of the delay times for United Airlines and American Airlines.

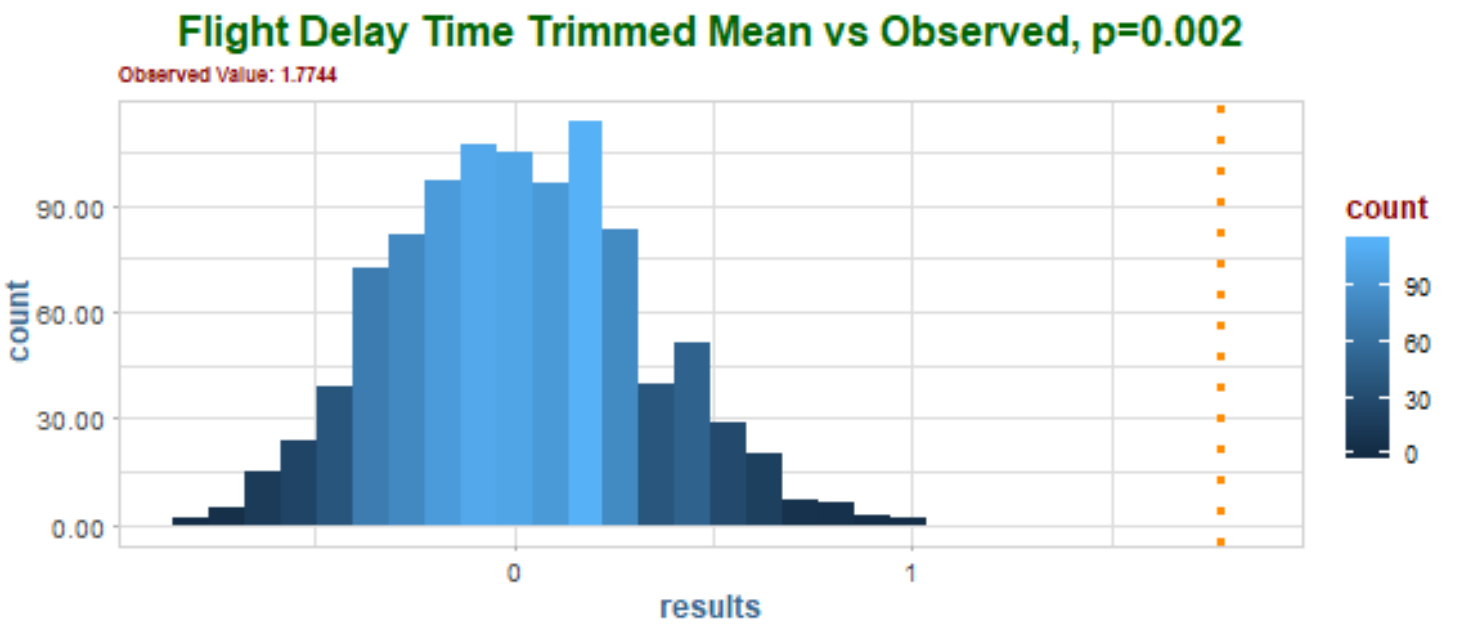
```
trim.amount <- .25
UA.trimmed <- mean(UA.Delay, trim = trim.amount)
AA.trimmed <- mean(AA.Delay, trim = trim.amount)

observed <- UA.trimmed - AA.trimmed

pretty_kable(data.table( UA = UA.trimmed, AA = AA.trimmed), "Trimmed Means")
```

Table 4: Trimmed Means

UA	AA
-0.8	-2.57

b.) Conduct a two-sided test to see if the difference in trimmed means is statistically significant.

```
N <- 10e2 - 1
results <- numeric(N)

for(i in 1:N)
{
   index <- sample(nrow(Flights), nrow(Flights[Carrier == "UA"]), replace = F)
   results[i] <- mean(Flights[index]$Delay, trim = trim.amount) - mean(Flights[-index]$Delay, t
}

p <- min(1, 2 * (sum(results[results >= observed]) + 1) / ( N + 1))
v <- p*(1 - p) / ( N + 1 )

ggplot(data.table(results)) +
   geom_histogram(aes(results, fill = ..count..), bins = 30) +
   geom_vline(xintercept = observed, col = "darkorange", linetype = 3, lwd = 1.2) +
   scale_y_continuous(labels = comma) +
   labs(title = paste0("Flight Delay Time Trimmed Mean vs Observed, p=", round(p, 5)),
        subtitle = paste0("Observed Value: ", round(observed, 4)))
```



3.9

In the flight delays case study in Section 1.1,

a.) Compute the proportion of times the flights in May and in June were delayed more than 20 min.

```
delay20_month <- Flights[, .(Delay20 = sum(Delay > 20) / .N), by = Month]

pretty_kable(delay20_month, "Delayed by more than 20 min, by month")
```

Table 5: Delayed by more than 20 min, by month

Month	Delay20
May	0.17
June	0.20

Conduct a two-sided test for statistical significance.

```
observed <- delay20_month[Month == "May"]$Delay20 - delay20_month[Month == "June"]$Delay20

N <- 10e2 - 1
results <- numeric(N)

for( i in 1:N)
{
   index <- sample(nrow(Flights), nrow(Flights[Month == "May"]), replace = F)
```

```
   results[i] <- as.numeric( Flights[index, .(Delay = sum(Delay > 20) / .N)] - Flights[-index,
}

p <- min(1, 2 * (sum(results[results >= observed]) + 1) / ( N + 1))
v <- p*(1 - p) / ( N + 1 )
```

P-value: 0.0035, which is statistically significant.

b.) Compute the ratio of the variances in the flight delay times in May and in June.

```
observed <- var(Flights[Month == "May"]$Delay) - var(Flights[Month == "June"]$Delay)

N <- 10e2 - 1
results <- numeric(N)

for( i in 1:N)
{
   index <- sample(nrow(Flights), nrow(Flights[Month == "May"]), replace = F)
   results[i] <- var( Flights[index, .(Delay)] ) - var( Flights[-index, .(Delay)] )
}

p <- min(1, 2 * (sum(results[results >= observed]) + 1) / ( N + 1))
v <- p*(1 - p) / ( N + 1 )
```

Is this evidence that the true ratio is not equal to 1, or could this be due to chance variability?

The variance appear to be due to random chance, so there does appear to be a stastical significance between the two months.

Conduct a two-sided test to check.

P-value: **1**

3.10

In the black spruce case study in Section 1.10, seedlings were planted in plots that were either subject to competition (from other plants), or not.

Use the data set *Spruce* to conduct a test to see if the mean difference is how much the seedlings grow (in height) over the corse of the study under these two treatments is stastically significant.

```
Spruce <- data.table(read.csv(paste0(data.dir, "Spruce.csv"),
                              header = T))

observed <- mean( Spruce[Competition == "NC"]$Ht.change ) - mean( Spruce[Competition == "C"]$Ht

N <- 10e2 - 1
```

```
results <- numeric(N)

for(i in 1:N)
{
   index <- sample(nrow(Spruce), nrow(Spruce[Competition == "NC"]), replace = F)
   resutls <- mean( Spruce[index]$Ht.change ) - mean( Spruce[-index]$Ht.change )
}

p <- min(1, 2 * (sum(results[results >= observed]) + 1) / ( N + 1))
```

There is stastical significance in between the heights of the two groups (Competition / No-competition).

P-value: **0.002**

3.11

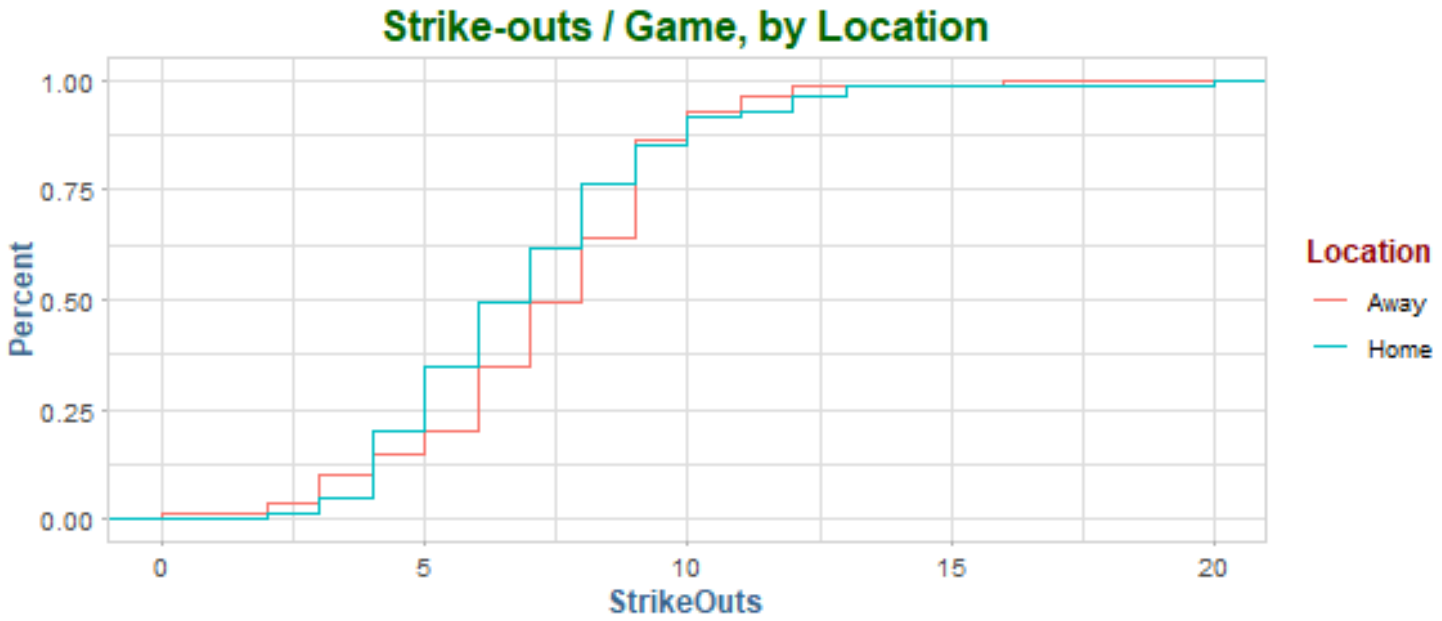
The file *Phillies2009* contains data from the 2009 season for the baseball team the Philadelphia Phillies.

```
Phillies <- data.table(read.csv(paste0(data.dir, "Phillies2009.csv"),
                                header = T))
```

a.) Compare the empirical distribution functions of the number of strike-outs per game (*StrikeOuts*) for games played at home and games played away (*Location*).

```
ggplot(Phillies, aes(StrikeOuts, color = Location)) +
   stat_ecdf(stat = "point") +
   labs(title = "Strike-outs / Game, by Location", y = "Percent")
```

```
Warning: Ignoring unknown parameters: stat
```



b.) Find the mean number of strike-outs per game for the home and the away games.

```
strikeouts <- Phillies[, .(StrikeOuts = mean(StrikeOuts)), by = Location]

pretty_kable(strikeouts, "StrikeOuts by Location")
```

Table 6: StrikeOuts by Location

Location	StrikeOuts
Home	6.95
Away	7.31

```
observed <- mean(strikeouts[Location == "Away"]$StrikeOuts) - mean(strikeouts[Location == "Home
```

c.) Perform a permutation tests to see if the differences in means is statistically significant.

```
N <- 10e2 - 1
results <- numeric(N)

for( i in 1:N)
{
   index <- sample(nrow(Phillies), nrow(Phillies[Location == "Home"]), replace = F)
   results[i] <- mean(Phillies[index]$StrikeOuts) - mean(Phillies[-index]$StrikeOuts)
}

p <- min(1, ( sum(results >= observed) + 1) / ( N + 1) )
```

There does not appear to be a statistically significat relationship between the number of home and away strikeouts.

P-value: **0.204**

3.12

In the Iowa recidivism case study in Section 1.4, offenders had originally been convicted of either a felony or misdemenor.

```
Recidivism <- data.table(read.csv(paste0(data.dir, "Recidivism.csv"),
                                  header = T))
```

a.) Use R to create a table displaying the proportion of felons who recidivated and the propportion of those convictted of a misdemeanor who recidivated.

```
Recidivism[, Recidivated := is.na(Days)]

Recidivated <- Recidivism[, .(RecidPct = sum(Recidivated) / .N), by = Offense]

pretty_kable(Recidivated, "Recidivated")
```

Table 7: Recidivated

Offense	RecidPct
Felony	0.69
Misdemeanor	0.67

b.) Determine whether or not the difference in recidivism proportions computed in (a) is statistically significant.

```
observed <- as.numeric( Recidivated[Offense == "Felony"]$RecidPct - Recidivated[Offense == "Mis

N <- 10e2 - 1
results <- numeric(N)

for( i in 1:N )
{
   index <- sample(nrow(Recidivism), nrow(Recidivism[Offense == "Felony"]), replace = F)
   results[i] <- as.numeric( Recidivism[index, .(Pct = sum(Recidivated) / .N)] - Recidivism[-ir
}

p <- ( sum(results >= observed) + 1 ) / ( N + 1 )
```

It appears there is statistical significance in the recidivism rates between felony and misdemeanor convictions.

P-value: 0.023

3.13

In the Iowa recidivism case study in Section 1.4, for those offenders who recidivated, we have data on the number of days until they reoffended.

For those offenders who did recidivate, determine if the difference in the mean number of days (**Days**) until recidivism between those under 25 years of age and those 25 years of age and older is statistically significant.

```
Recid.Age25 <- Recidivism[, .(RecidDays = mean(Days, na.rm = T)), by = Age25]

pretty_kable(Recid.Age25, "Days Until Recidivism by Age ( < 25 < )")
```

Table 8: Days Until Recidivism by Age (< 25 <)

Age25	RecidDays
Under 25	449.07
Over 25	479.72
NA	NaN

```
observed <- as.numeric( Recid.Age25[Age25 == "Under 25"]$RecidDays - Recid.Age25[Age25 == "Over

N <- 10e2 - 1
results <- numeric(N)

for( i in 1:N )
{
   index <- sample(nrow(Recidivism), nrow(Recidivism[Age25 == "Under 25"]), replace = F)
   results[i] <- mean(Recidivism[index]$Days, na.rm = T) - mean(Recidivism[-index]$Days, na.rm
}

p <- min(1, ( sum(results >= observed) + 1 ) / ( N + 1 ) )
```

There does not appear to be a statistical difference between the days until recidivism by those over/under 25 years of age.

P-value: 0.999

3.14

Does chocolate ice cream have more calories than vanilla ice cream? The data set **IceCream** contains calorie information for a sample of brands of chocolate and vanilla ice cream.

```
IceCream <- data.table(read.csv(paste0(data.dir, "IceCream.csv"),
                                header = T))
```

a.) Inspect the data set, then explain why this is an example of matched pairs data.

```
pretty_kable(IceCream, "Ice Cream Data")
```

This data set is matched pairs becase there are multiple brands associated with each flavor. They are not independent samples.

b.) Compute summary statistics of the number of calories for the two flavors.

```
pretty_kable(IceCream[, .(Vanilla = mean(VanillaCalories), Chocolate = mean(ChocolateCalories))
             "Ice cream Calories by Flavor")
```

c.) Conduct a permutation tests to determine whether or not chocolate ice cream has, on average, more calories than vanilla ice cream.

```
diff <- IceCream$VanillaCalories - IceCream$ChocolateCalories
observed <- mean(diff)

N <- 10e2 - 1
results <- numeric(N)

for( i in 1:N )
{
   Sign <- sample(c(-1, 1), nrow(IceCream), replace = T)
   diff2 <- Sign * diff
   results[i] <- mean(diff2)
}

p <- ( sum( results <= observed ) + 1 ) / ( N + 1 )
```

The data supports the claim that chocolate ice cream has more calories than vanilla.

P-value: 0.001

3.15

Is there a difference in the price of groceries sold by the two retailers Target and Walmart?

The data set Groceries contains a sample of grocery items and their prices advertised on their respective web sites on one specific day.

```
Groceries <- data.table(read.csv(paste0(data.dir, "Groceries.csv"),
                                 header = T))
```

a.) Inspect the data set, then explain why this is an example of matched pairs data.

Table 9: Ice Cream Data

Brand	VanillaCalories	VanillaFat	VanillaSugar	ChocolateCalories	ChocolateFat	ChocolateSugar
Baskin Robbins	260	16.0	26.0	260	14.0	31.0
Ben & Jerry's	240	16.0	19.0	260	16.0	22.0
Blue Bunny	140	7.0	12.0	130	7.0	14.0
Breyers	140	7.0	13.0	140	8.0	16.0
Brigham's	190	12.0	17.0	200	12.0	18.0
Bulla	234	13.5	21.8	266	15.0	22.6
Carvel	240	14.0	21.0	250	13.0	25.0
Cass-Clay	130	7.0	11.0	150	7.0	16.0
Chapman's	120	6.0	11.0	120	5.0	12.0
Cold Stone	270	15.5	23.0	264	16.2	23.6
Culver's	222	13.0	19.0	205	10.0	20.0
Dairy Queen	140	4.5	19.0	150	5.0	17.0
Dove	240	15.0	20.0	290	17.0	27.0
Dreamery	260	15.0	24.0	280	12.0	33.0
Edy's Grand	140	8.0	13.0	150	8.0	15.0
Emack & Bolio's	160	9.0	12.0	170	9.0	13.0
Good Humor	120	6.0	12.0	120	6.0	14.0
Graeter's	260	16.0	24.0	260	16.0	24.0
Green and Black	194	11.6	18.0	227	12.8	22.7
Green's	150	8.0	17.0	140	8.0	15.0
Haagen Dazs	270	18.0	21.0	270	18.0	21.0
Hershey's	140	9.0	14.0	140	8.0	13.0
Hill Station	226	15.6	16.8	235	14.3	21.2
Kemp's	130	7.0	13.0	140	6.0	17.0
Klein's	130	8.0	15.0	140	8.0	14.0
Oberweis Dairy	307	21.0	23.0	320	21.0	19.0
Our Family	130	7.0	11.0	130	6.0	15.0
Perry's	140	8.0	15.0	140	7.0	15.0
Ronnybrook Farm	240	16.0	20.0	260	19.0	21.0
Ruggles	150	8.0	12.0	150	8.0	16.0
Sara Lee	242	15.5	21.5	234	14.4	20.9
Schwan's	140	7.0	12.0	140	7.0	12.0
Sheer Bliss	300	19.0	27.0	320	19.0	29.0
Smith's	150	8.0	13.0	150	8.0	13.0
Stonyfield Farm	240	16.0	20.0	250	17.0	20.0
Tillamook	160	9.0	10.0	170	9.0	13.0
Turkey Hill	140	8.0	16.0	150	8.0	19.0
Value Choice	130	6.0	12.0	130	6.0	15.0
Whitey's	250	14.0	23.0	250	13.0	25.0

Table 10: Ice cream Calories by Flavor

Vanilla	Chocolate
191.41	198.74

```
pretty_kable(Groceries, "Target vs Walmart")
```

Table 11: Target vs Walmart

Product	Size	Target	Walmart
Kellogg NutriGrain Bars	8 bars	2.50	2.78
Quaker Oats Life Cereal Original	18oz	3.19	6.01
General Mills Lucky Charms	11.50z	3.19	2.98
Quaker Oats Old Fashioned	18oz	2.82	2.68
Nabisco Oreo Cookies	14.3oz	2.99	2.98
Nabisco Chips Ahoy	13oz	2.64	1.98
Doritos Nacho Cheese Chips	10oz	3.99	2.50
Cheez-it Original Baked	21oz	4.79	4.79
Swiss Miss Hot Chocolate	10 count	1.49	1.28
Tazo Chai Classic Latte Black Tea	32 oz	3.49	2.98
Annie's Macaroni & Cheese	6oz	1.79	1.72
Rice A Roni Chicken	6.9oz	1.00	1.00
Zatarain's Jambalaya Rice Mix	8oz	1.62	1.54
SPAM Original Lunch Meat	12oz	2.79	2.64
Campbell's Chicken Noodle Soup	10.75oz	0.99	1.58
Dinty Moore Hearty Meals Beef Stew	15oz	1.99	1.98
Hormel Chili with Beans	15oz	1.94	1.88
Dole Pineapple Chunks	20 oz	1.59	1.47
Skippy Creamy Peanut Butter	16.3oz	2.59	2.58
Smucker's Strawberry Preserve	18oz	2.99	2.84
Heinz Tomato Ketchup	32oz	2.99	2.88
Near East Couscous Toasted Pine Nuts mix	5.6oz	2.12	1.98
Barilla Angel Hair Pasta	16oz	1.42	1.38
Betty Crocker Super Moist Chocolate Fudge Cake Mix	15.25oz	1.22	1.17
Kraft Jet-Puffed Marshmllows	16oz	1.99	1.96
Dunkin' Donuts Original Blend Medium Roast Ground Coffee	12oz	7.19	6.98
Dove Promises Milk Chocolate	8.87oz	3.19	3.50
Skittles	41oz	7.99	6.98
Vlasic Kosher Dill Pickle Spears	24oz	2.39	2.18
Vlasic Old Fashioned Sauerkraut	32oz	1.99	1.97

b.) Compute summary statistics of the prices for each store.

```
diff <- Groceries$Target - Groceries$Walmart

observed <- mean(diff)
```

Observed difference: **0.0566667**

c.) Conduct a permutation test to determine whether or not there is a difference in the mean prices.

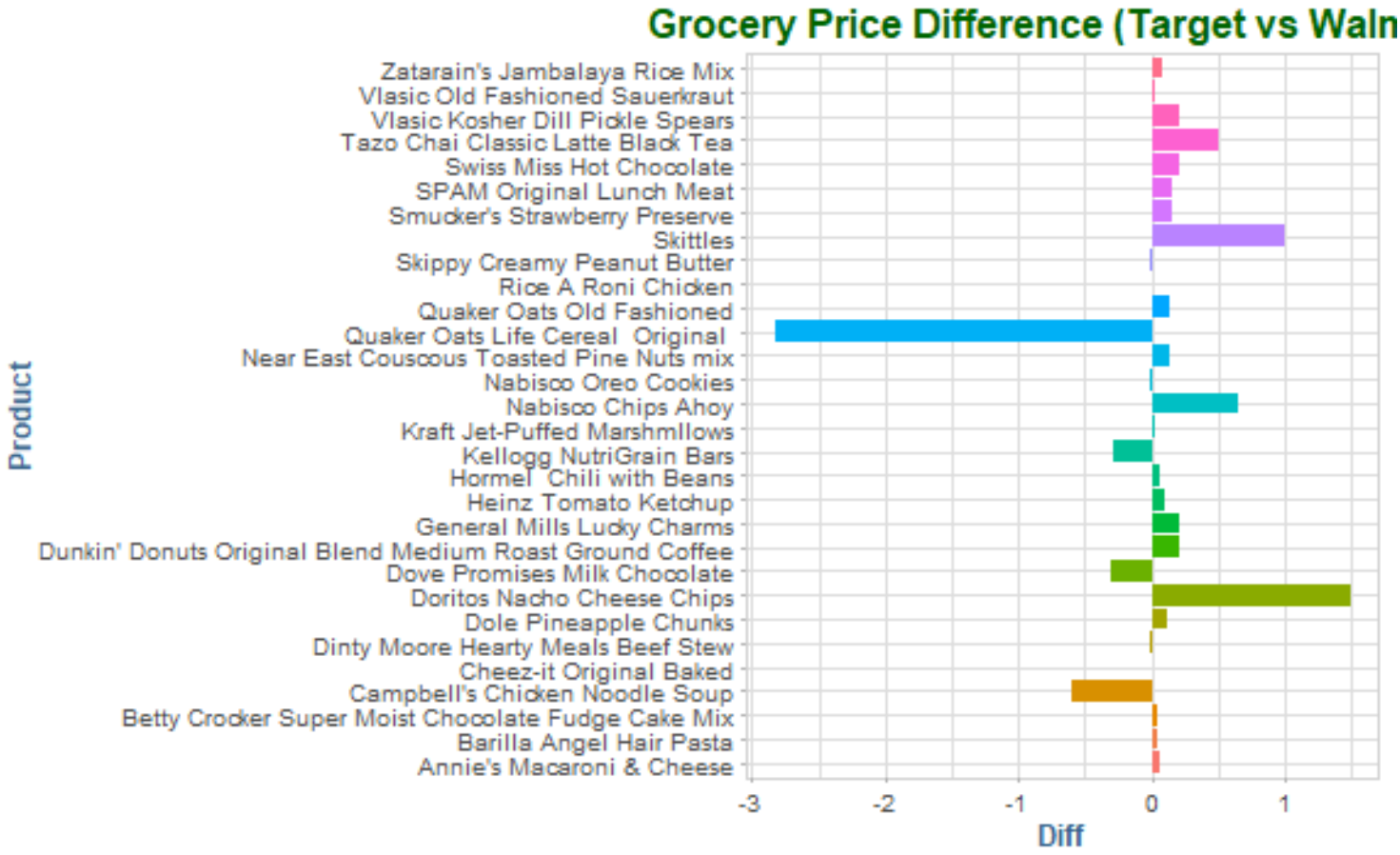
```
N <- 10e2 - 1
results <- numeric(N)

for( i in 1:N )
{
   Sign <- sample(c(-1, 1), nrow(Groceries), replace = T)
   diff2 <- Sign * diff
   results[i] <- mean(diff2)
}

p <- min(1, 2 * ( sum(results >= observed ) + 1 ) / ( N + 1) )
```

There does not appear to be a statistical difference in the prices at the two stores.

P-value: 0.706

d.) Create a histogram of the difference in prices.

```
ggplot(data.table( Product = Groceries$Product, Diff = diff)) +
   geom_bar(aes(Product, Diff, fill = Product), stat = "identity") +
   coord_flip() +
   labs(title = "Grocery Price Difference (Target vs Walmart)") +
   theme(legend.position = "none")
```



What is unusual about Quaker Oats Life cerial?

Price difference is a multiple SD outlier.

e.) Redo the hypothesis test without this observation.

```
diff <- Groceries[Product != "Quaker Oats Life Cereal  Original ", .(Diff = Target - Walmart )

observed <- mean(diff)

N <- 10e2 - 1
results <- numeric(N)

for( i in 1:N )
{
   Sign <- sample(c(-1, 1), length(diff), replace = T)
   diff2 <- Sign * diff
   results[i] <- mean(diff2)
}

p <- min(1, 2 * ( sum(results >= observed ) + 1 ) / ( N + 1) )
```

Do you reach the same conclusion?

No, there is a statistical difference in the prices of products at the two retailers with the outlier removed.

P-value: 0.024

3.16

In the sampling version of permutation testing, the one-sided P-value is

$\hat{P} = \frac{(X+1)}{(N+1)}$, where X is the number of permutation test statistics that are as large or larger than the observed test statistic.

Suppose the true P-value (for the exaustive test, conditional on the observed data) is **p**.

a.) What is the variance of $\hat{P}$?

$\mathbb{V} = p * (1-p)/(N+1)$

b.) What is the variance of $\hat{P}_2$ for the two-sided test (assuming that **p** is not close to 0.5, where **p** is the smaller true one-side P-value)?