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.

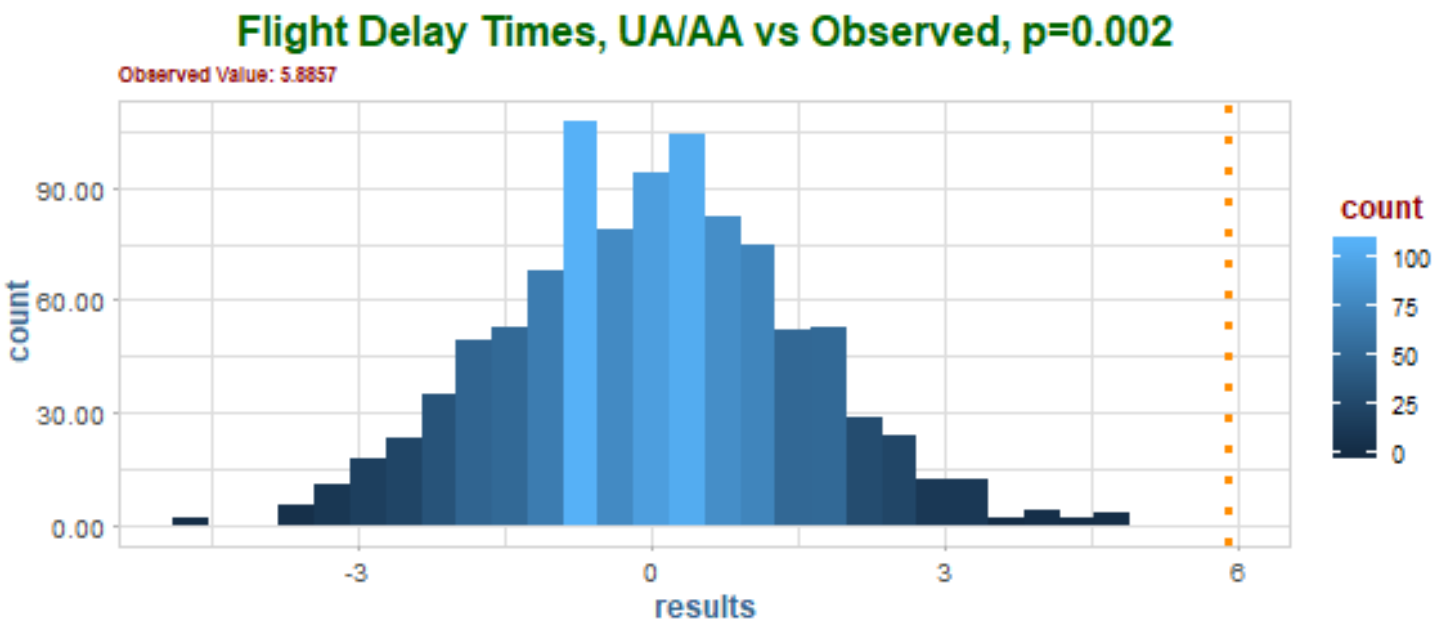
```
Flights <- data.table(read.csv(paste0(data.dir, "FlightDelays.csv"),
                               header = T))
```

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)))
```



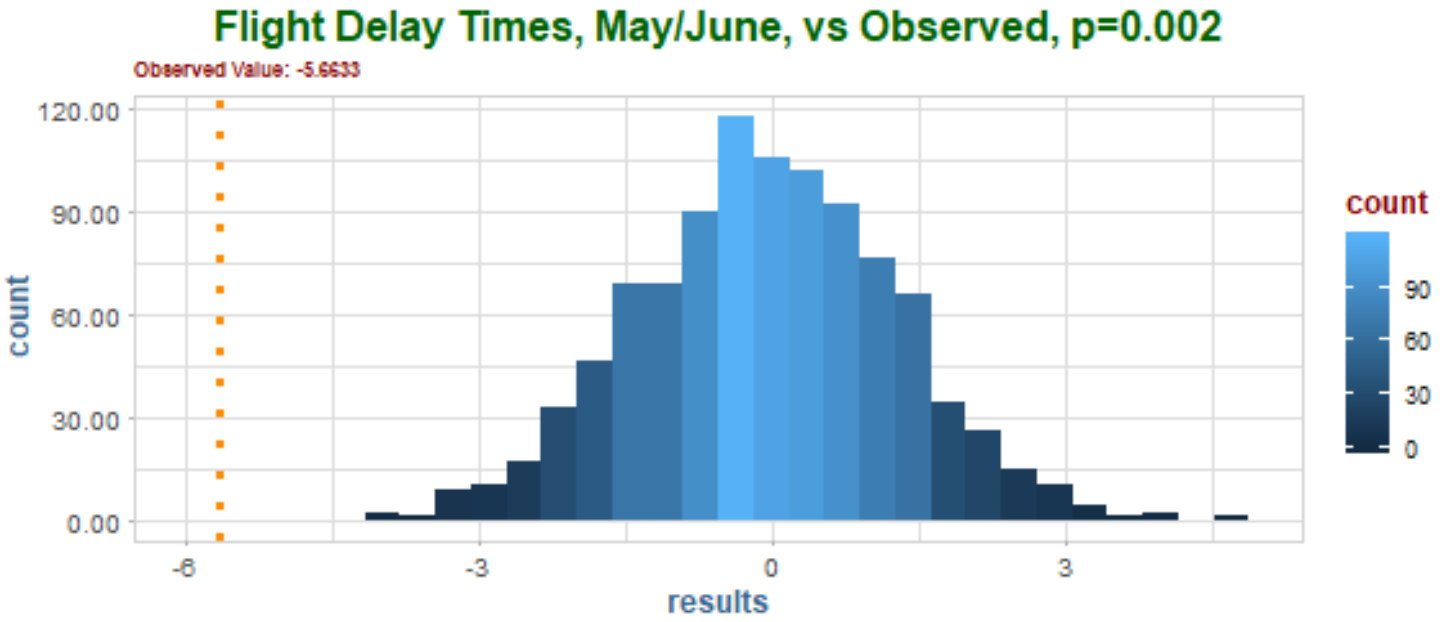
b.) The flights took place in May and June of 2009. Conduct a two-sided permutation test to see if the differences in mean delay times between two months is statistically significant.

```
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.

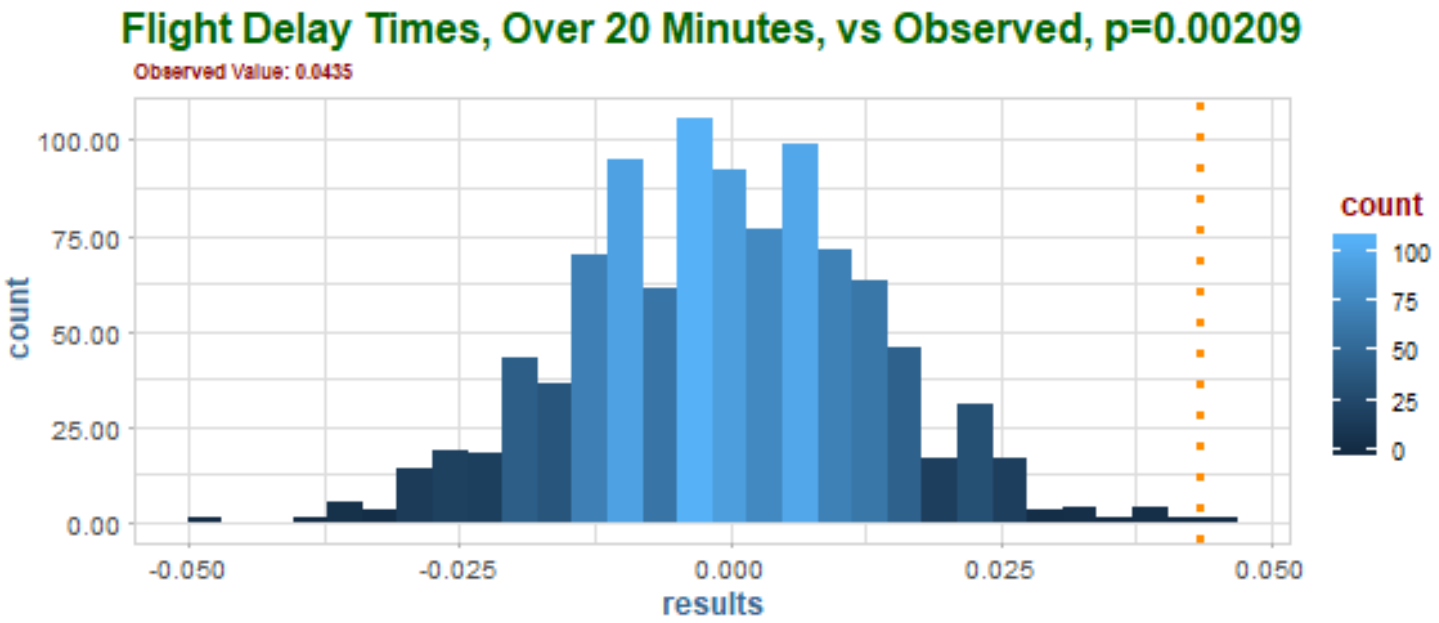
a.) Compute the proportion of times that each carrier's flight was delays more than 20 min.

```
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.

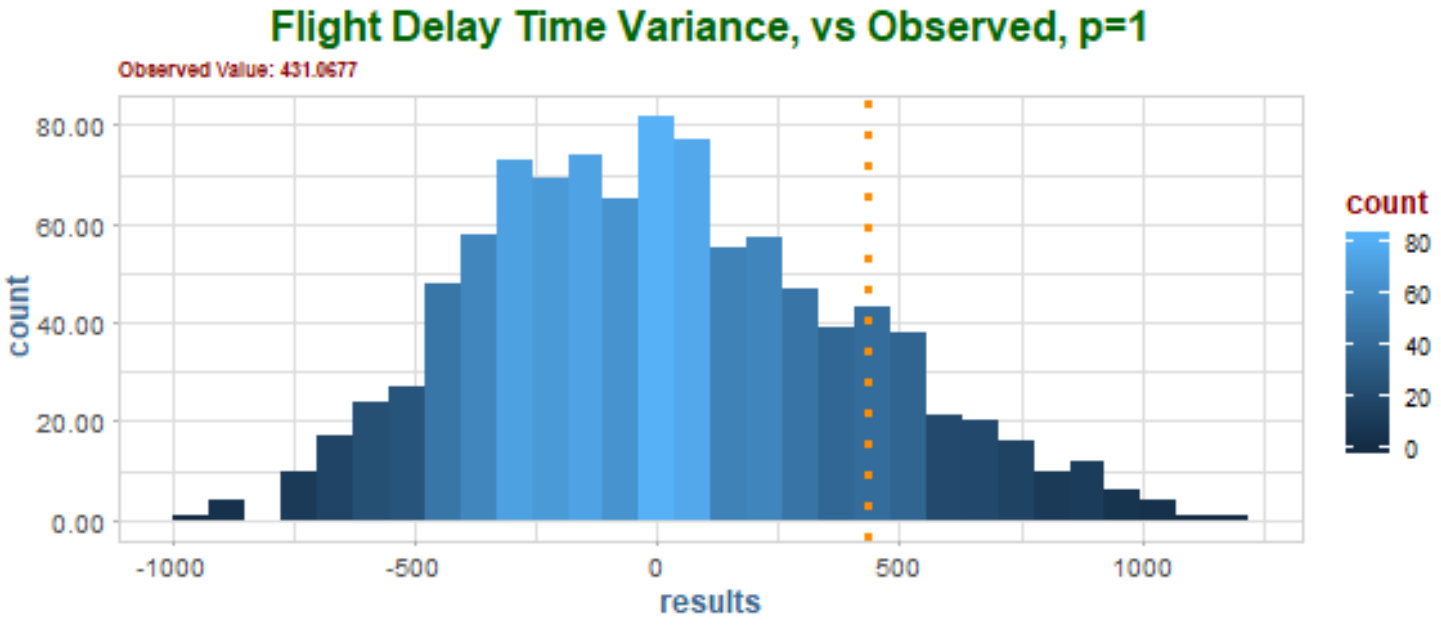
Answer: There is statistical significance with a P-value < 0.0001.

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

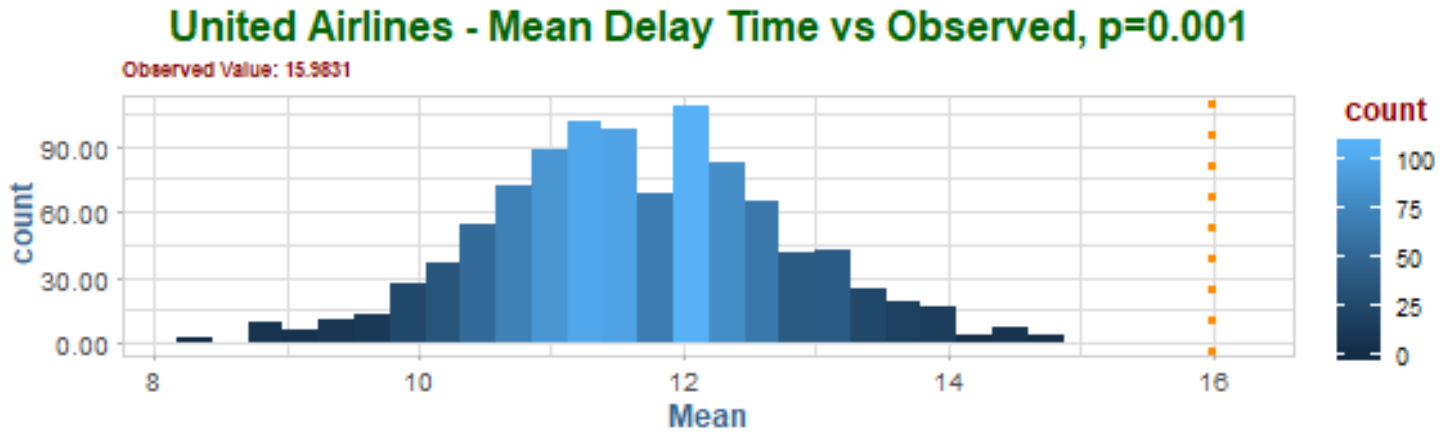
Compare the P-values.

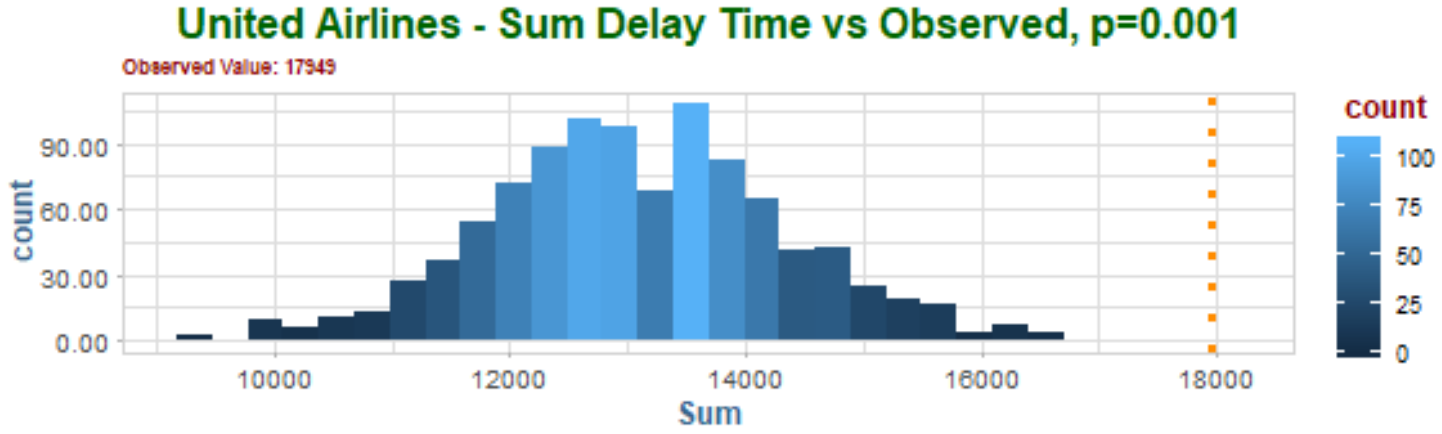
Make sure all three test statistics are computed within the same **for** loop.

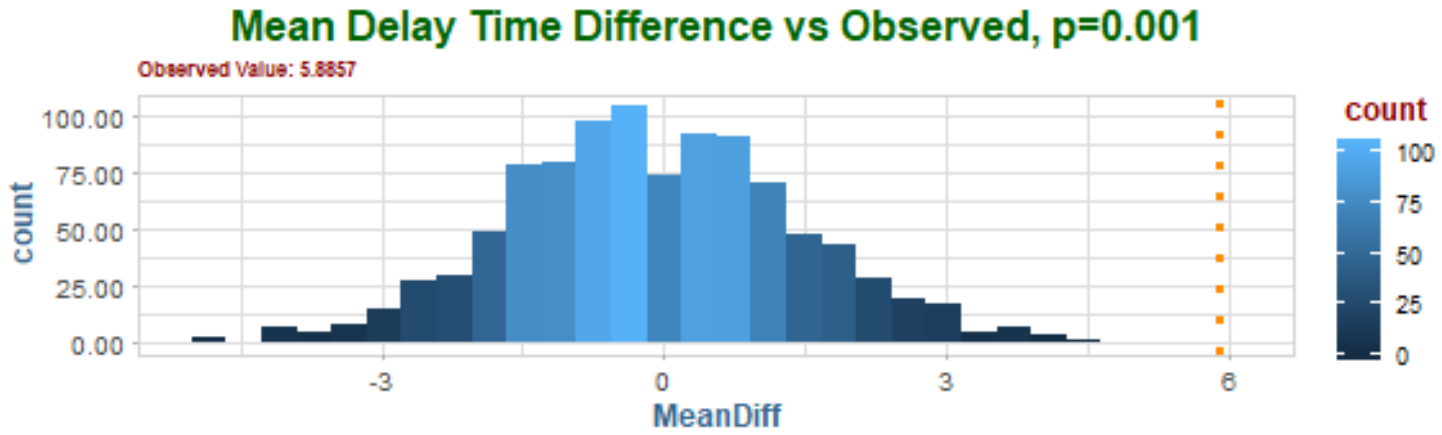
What do you observe?

```
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
100.00
75.00
50.00
25.00
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.

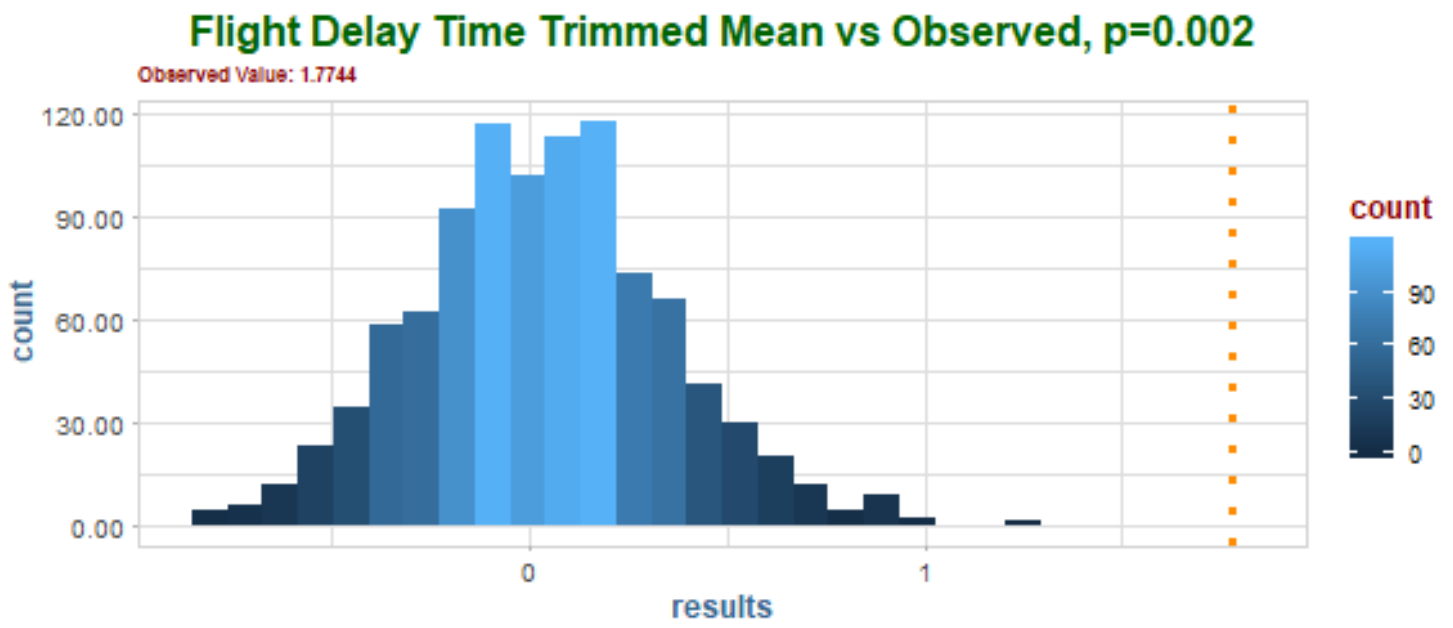
```
trim.amount <- .25
UA.trimmed <- mean(UA.Delay, trim = trim.amount)
AA.trimmed <- mean(AA.Delay, trim = trim.amount)

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: 0.0023, 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: **-4.2220358**

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

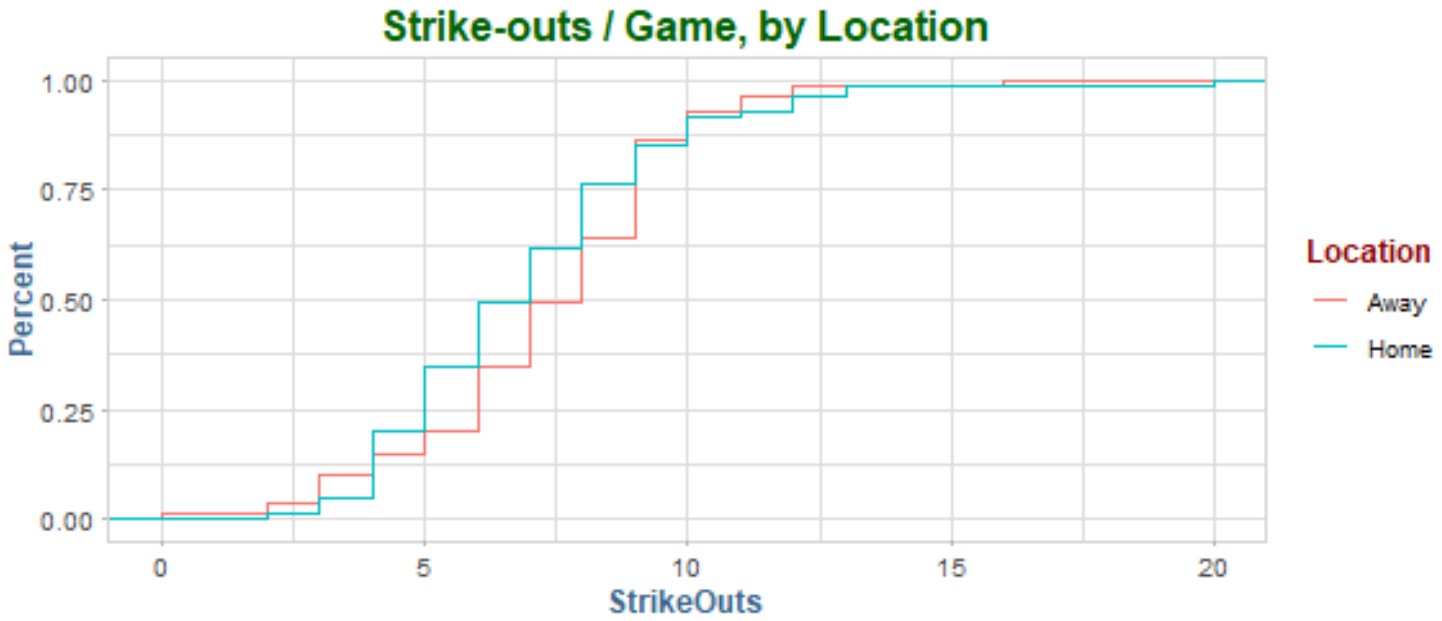
The file *Phillies2009* contains data from the 2009 season for the baseball team the Philadelphia Phillies.

```
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: **0.223**