Homework 10 Solutions

JMG

2020-05-06

```r
library(resampledata)
library(fastR2)
```

Problem 1

Suppose you conduct an experiment and inject a drug into three mice. Their times for running a maze are 8, 10, and 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.

Solution

mean group 1 $= \frac{8+10+15}{3} = 11$

mean group 2 $= \frac{5+9}{2} = 7$

difference in means $= 11 - 7 = 4$

b) Write out all possible permutations of these times to the two groups and calculate the difference in means for each permutation.

Solution

We need to permute $\{8, 10, 15, 5, 9\}$, divide into one group of three and one group of two, and compute the difference in means between the two groups. Since there are $n = 5$ total elements there are $n! = 120$ permutations but not all of these lead to distinct difference in means. Thus, we only list ones that do lead to distinct differences in means.

1) 4
2) 1.5
3) 4.8333333
4) -0.1666667
5) 3.1666667
6) -3.3333333
7) -1
8) 0.6666667
9) -5.1666667
10) -3.5

For later, we store these difference in means in a vector:

```
diff_means_one <- c(4,1.5,4.83,-0.17,3.17,-3.33,-1,0.67,-5.17,-3.5)
```

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

Solution

```
sum(diff_means_one >= 4)/length(diff_means_one)
```

```
## [1] 0.2
```

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?

Solution

1) 11
2) 10
3) 11.3333333
4) 9.3333333
5) 10.6666667
6) 8.6666667
7) 9
8) 9.6666667
9) 7.3333333
10) 8

We store these means in a vector and copmute the proportion as large or larger than the observed:

```
treat_means_one <- c(11,10,11.33,9.33,10.67,8.67,9,9.67,7.33,8)
sum(treat_means_one >= 11)/length(treat_means_one)
```

```
## [1] 0.2
```

Problem 2

In a hypothesis test comparing two population 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?

Solution

Recall that a smaller p-value provides stronger evidence for rejecting the null-hypothesis thus 0.006 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?

Solution

Recall that the null-hypothesis corresponds to the absence of any true effect and a larger p-value provides lack of evidence for rejecting the null hypothesis. Therefore, 0.095 provides stronger evidence that chance alone might account for the observed result.

Problem 3

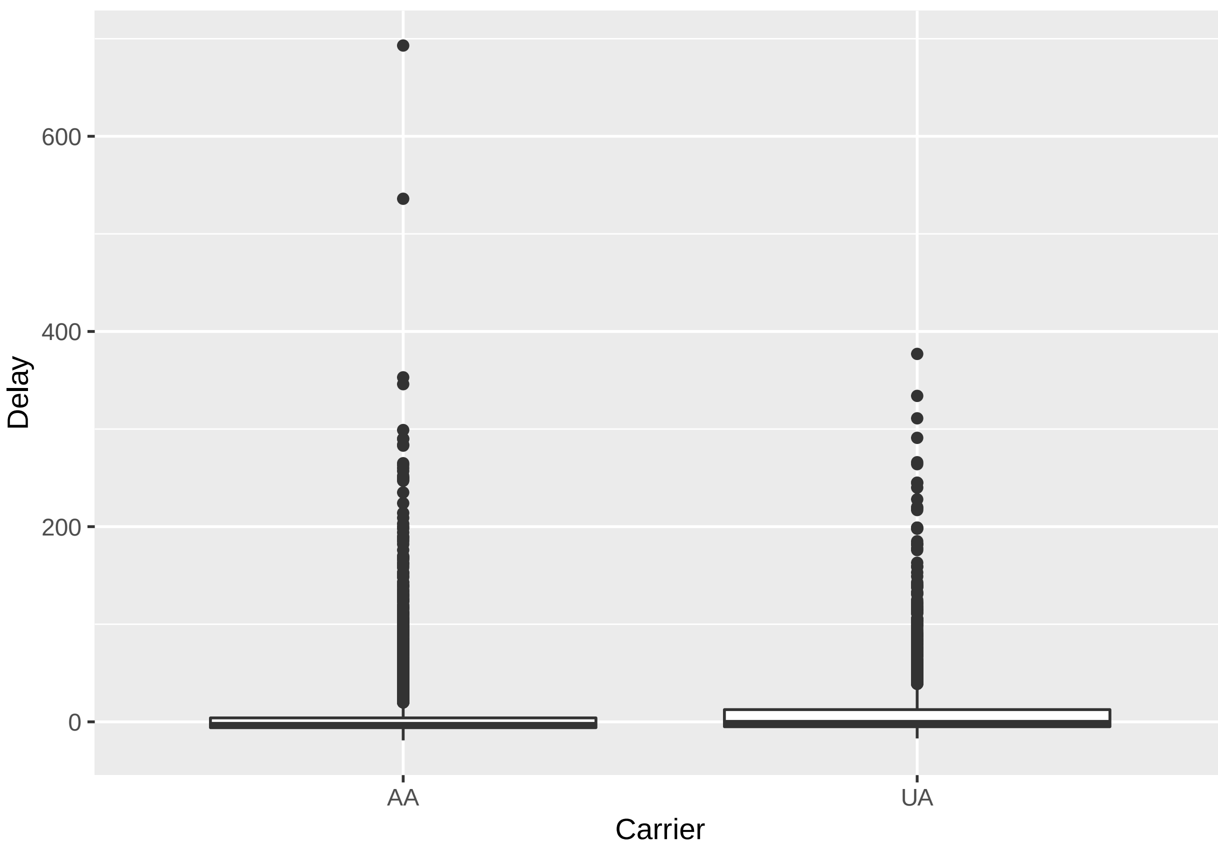
Consider the flight delays data from the resample package, the first few rows of which are shown below

```r
head(FlightDelays)
```

```
##   ID Carrier FlightNo Destination DepartTime Day Month FlightLength Delay
## 1  1      UA      403         DEN      4-8am Fri   May          281    -1
## 2  2      UA      405         DEN     8-Noon Fri   May          277   102
## 3  3      UA      409         DEN      4-8pm Fri   May          279     4
## 4  4      UA      511         ORD     8-Noon Fri   May          158    -2
## 5  5      UA      667         ORD      4-8am Fri   May          143    -3
## 6  6      UA      669         ORD      4-8am Fri   May          150     0
##   Delayed30
## 1        No
## 2       Yes
## 3        No
## 4        No
## 5        No
## 6        No
```

This data records flight information for two airlines: United (UA), and American (AA). We might be intersted to see if one airline has flights that are consistently delayed more often than the other. To begin to study this, we can plot the delay times for each airline:

```r
FlightDelays %>% gf_boxplot(Delay~Carrier)
```



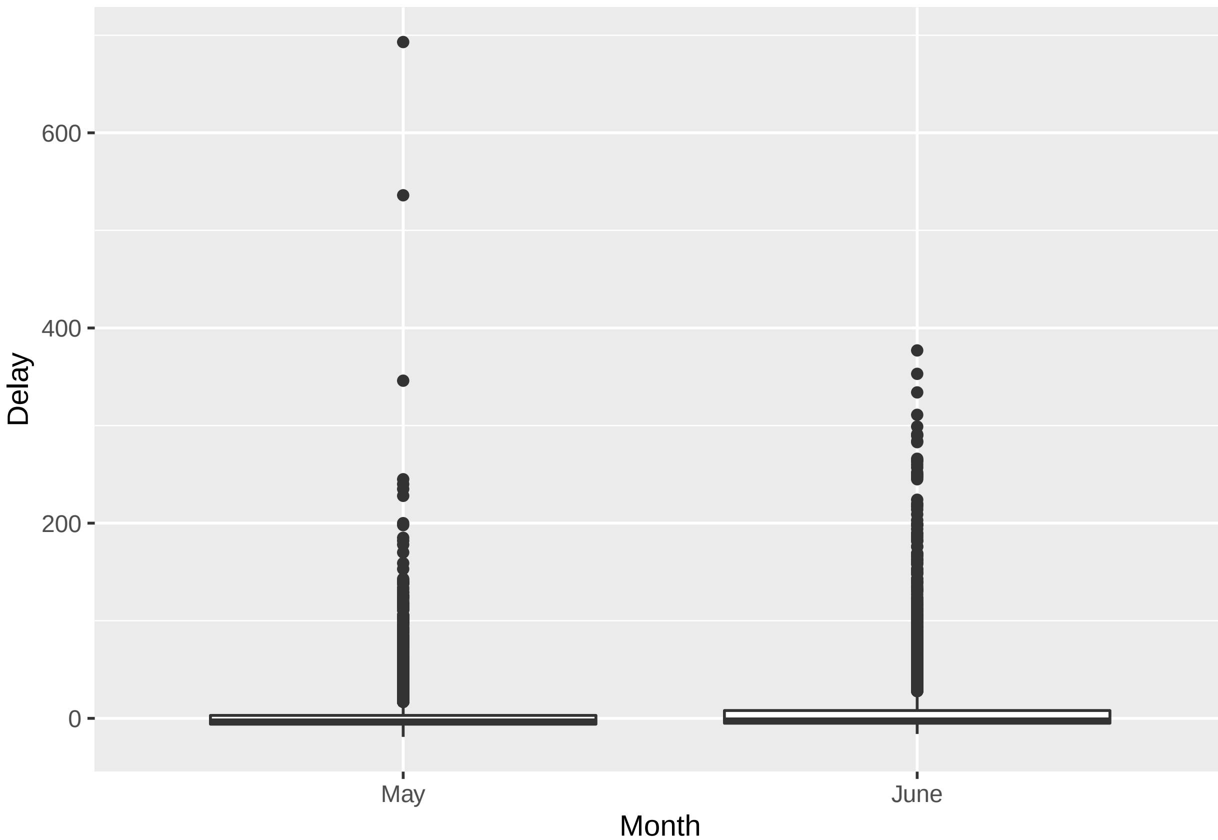
Note the number of times each airline appears in the data:

```
table(FlightDelays$Carrier)
```

```
##
##   AA   UA
## 2906 1123
```

Alternatively, we might be interested to see if there is a difference in delay time between flights in the month of May and flights in the month of June. Again, a plot may be helpful:

```
FlightDelays %>% gf_boxplot(Delay~Month)
```



Note the number of times each month appears in the data:

```
table(FlightDelays$Month)
```

```
##
##  May June
## 1999 2030
```

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

Solution

First, compute the observed difference in means and place all of the delays into a single vector:

```
UA_delays <- FlightDelays %>% filter(Carrier == "UA") %>%
  select(Delay) %>% unlist()
Num_UA <- length(UA_delays)
AA_delays <- FlightDelays %>% filter(Carrier == "AA") %>%
```

```
  select(Delay) %>% unlist()
Num_AA <- length(AA_delays)
All_delays <- c(UA_delays,AA_delays)
(ob_diff_mean <- mean(UA_delays) - mean(AA_delays))
```

```
## [1] 5.885696
```

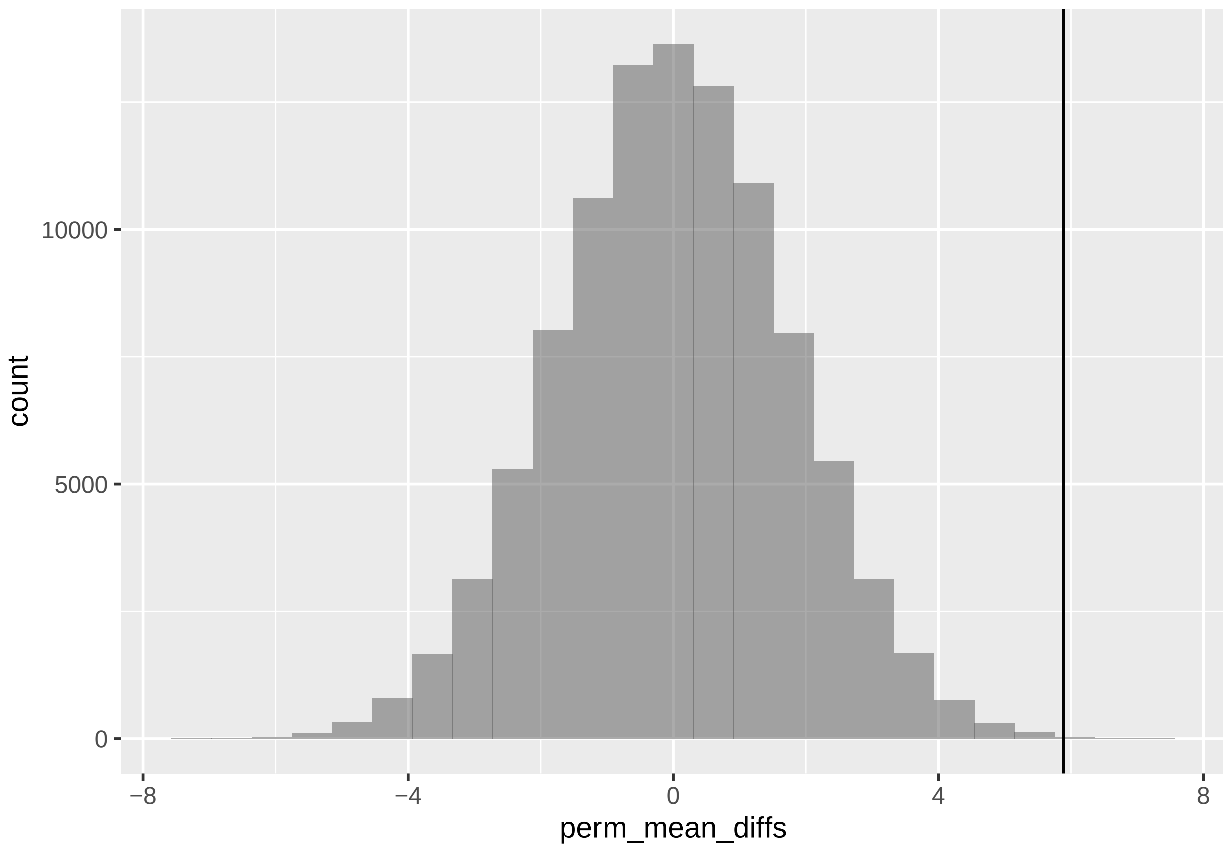
Now carry out the permutations:

```
N <- 10^5 - 1
perm_mean_diffs <- numeric(N)
for (i in 1:N){
  perm_data <- sample(All_delays)
  perm_UA <- perm_data[1:Num_UA]
  perm_AA <- perm_data[-(1:Num_AA)]
  perm_mean_diffs[i] <- mean(perm_UA) - mean(perm_AA)
}
```

Let's plot the result:

```
gf_histogram(~perm_mean_diffs) %>% gf_vline(xintercept = ob_diff_mean)
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```



Now compute the two-sided p-value:

```
p_less <- (sum((perm_mean_diffs <= ob_diff_mean)) + 1)/(N + 1)
p_greater <- (sum((perm_mean_diffs >= ob_diff_mean)) + 1)/(N + 1)
(p_value_3a <- 2*min(p_less,p_greater))
```

```
## [1] 7e-04
```

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

Solution

First, compute the observed difference in means and place all of the delays into a single vector:

```r
May_delays <- FlightDelays %>% filter(Month == "May") %>%
  select(Delay) %>% unlist()
Num_May <- length(May_delays)
June_delays <- FlightDelays %>% filter(Month == "June") %>%
  select(Delay) %>% unlist()
Num_June <- length(June_delays)
All_delays <- c(May_delays,June_delays)
(ob_diff_mean <- mean(May_delays) - mean(June_delays))
```

```
## [1] -5.663341
```

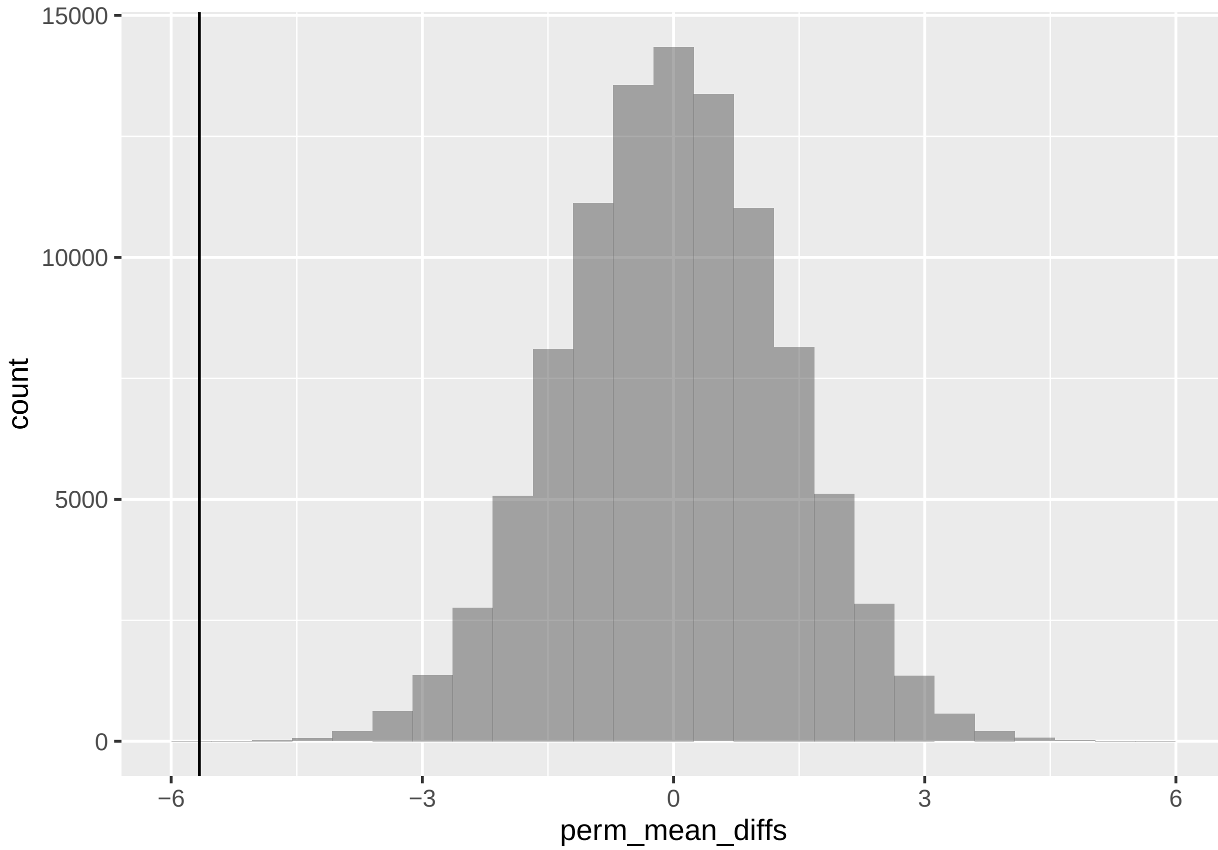
Now carry out the permutations:

```r
N <- 10^5 - 1
perm_mean_diffs <- numeric(N)
for (i in 1:N){
  perm_data <- sample(All_delays)
  perm_May <- perm_data[1:Num_May]
  perm_June <- perm_data[-(1:Num_June)]
  perm_mean_diffs[i] <- mean(perm_May) - mean(perm_June)
}
```

Let's plot the result:

```r
gf_histogram(~perm_mean_diffs) %>% gf_vline(xintercept = ob_diff_mean)
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```



Now compute the two-sided p-value:

```
p_less <- (sum((perm_mean_diffs <= ob_diff_mean)) + 1)/(N + 1)
p_greater <- (sum((perm_mean_diffs >= ob_diff_mean)) + 1)/(N + 1)
(p_value_3b <- 2*min(p_less,p_greater))
```

```
## [1] 2e-05
```

Problem 4

One might often test the hypothesis in problem 3a using a method implemented with the following code:

```
t.test(Delay~Carrier,data=FlightDelays)
```

```
##
##  Welch Two Sample t-test
##
## data:  Delay by Carrier
## t = -3.8255, df = 1843.8, p-value = 0.0001349
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.903198 -2.868194
## sample estimates:
## mean in group AA mean in group UA
##         10.09738         15.98308
```

Notice that this returns a p-value:

```r
t.test(Delay~Carrier,data=FlightDelays)$p.value
```

```
## [1] 0.0001348746
```

a) How do the result returned by using the t.test function above compare with what you obtained using a permutation test for 3a?

Solution

The p-value obtained in 3a is 7×10^{-4} which is a little larger than that obtained above but both are really small and strongly suggest that the data provides evidence for rejecting the null hypothesis.

b) Modify the above code to apply the t.test to test the hypothesis from problem 3b. Compare the results with what you obtained using a permutation test as done in problem 3b.

Solution

We can obtain a p-value as:

```r
t.test(Delay~Month,data=FlightDelays)$p.value
```

```
## [1] 1.504322e-05
```

The p-value obtained in 3b is 2×10^{-5} which is a little larger than that obtained above but both are really small and strongly suggest that the data provides evidence for rejecting the null hypothesis.