Class Activity#8

Sunny Lee

2021-03-03

Nonparametric Methods: Use the C1 Mice.csv data set inorder to understand nonparametric statistics.

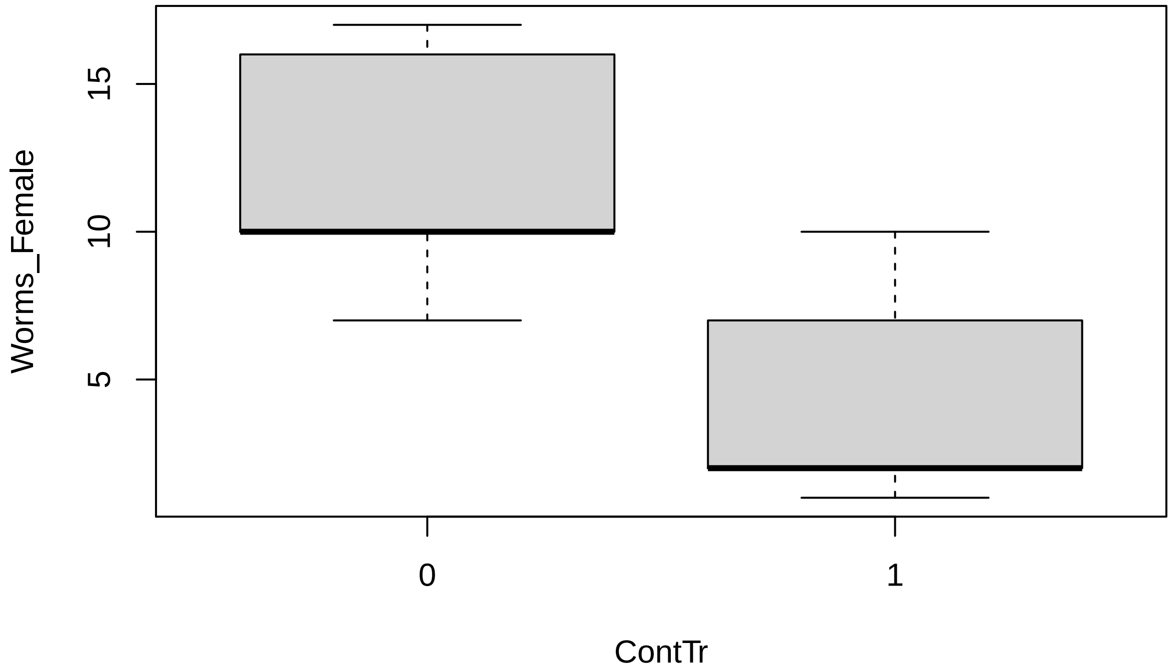
1) Compare the number of worms for the treatment and control groups for both the male and female mice. Does each of the four groups appear to have a similar center and a similar spread? Are there any outliers?

```
Mice<-read.csv("C1 Mice.csv")
head(Mice)
```

```
##   Female.Trt Female.Ctl Male.Trt Male.Ctl
## 1          1         16        3       31
## 2          2         10        5       26
## 3          2         10        9       28
## 4         10          7       10       13
## 5          7         17        6       47
```

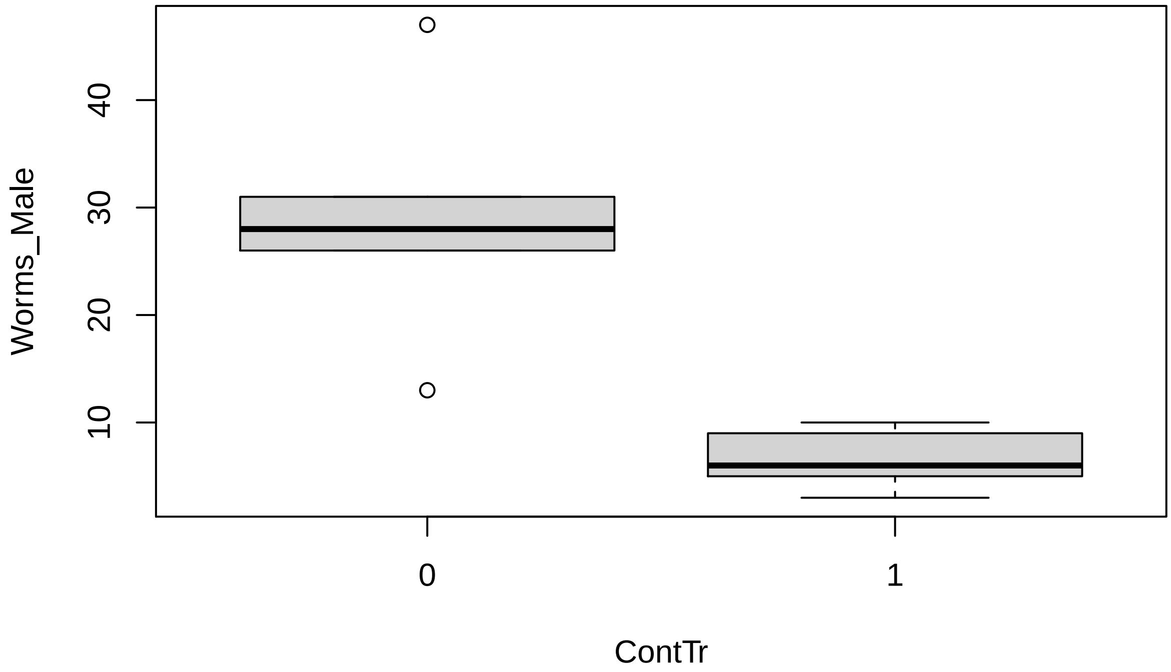
```
Female.Tr <- Mice$Female.Trt
Female.Con <- Mice$Female.Ctl

Observed_diff <- mean(Female.Con) - mean(Female.Tr)
#assume treatment = 1 and control = 0
ContTr <- c(rep(1, 5) , rep(0, 5))
Worms_Female <- c(Female.Tr, Female.Con)
boxplot(Worms_Female~ContTr)
```



```
Male.Tr <- Mice$Male.Trt
Male.Con <- Mice$Male.Ctl

ContTr <- c(rep(1, 5) , rep(0, 5))
Worms_Male <- c(Male.Tr, Male.Con)
boxplot(Worms_Male~ContTr)
```



For the male rats, we see that the control group has two outliers and that they have a different center and spread. For the female rats, we find that there are no outliers, have different centers, but have similar spread.

2) Calculate appropriate summary statistics (e.g., mean, median, standard deviation, and range) for each of the four group.

```
summary(Female.Con)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       7      10      10      12      16      17
```

```
sd(Female.Con)
```

```
## [1] 4.301163
```

```
summary(Female.Tr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     2.0     2.0     4.4     7.0    10.0
```

```
sd(Female.Tr)
```

```
## [1] 3.911521
```

```
summary(Male.Con)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      13      26      28      29      31      47
```

```
sd(Male.Con)
```

```
## [1] 12.18606
```

```
summary(Male.Tr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     3.0     5.0     6.0     6.6     9.0    10.0
```

```
sd(Male.Tr)
```

```
## [1] 2.880972
```

For our Female.Con: mean = 12, median = 10, sd = 4.301163 and our range = 10. For our Female.Tr: mean = 4.4, median = 2, sd = 3.911521 and our range = 9. For our Male.Con: mean = 29, median = 28, sd = 12.18606 and our range = 34. For our Male.Tr: mean = 6.6, median = 6, sd = 2.880972 and our range = 7.

3) To get a feel for the concept of p-value, write each of the female worm counts on an index card. Shuffle the 10 index cards, and then draw five cards at random (without replacement). Call these five cards the treatment group and the five remaining cards the control group. Under the null hypothesis this allocation mimics precisely what actually happened in our experiment, since the only cause of group differences isthe random allocation.

```
Worms_Female <- c(Female.Tr, Female.Con)
diff <- rep(0, 10)

sample <- sample(Worms_Female, 10, replace = FALSE)
treatment <- sample[6:10]
control <- sample[1:5]
diff[1] <- mean(treatment) - mean(control)
```

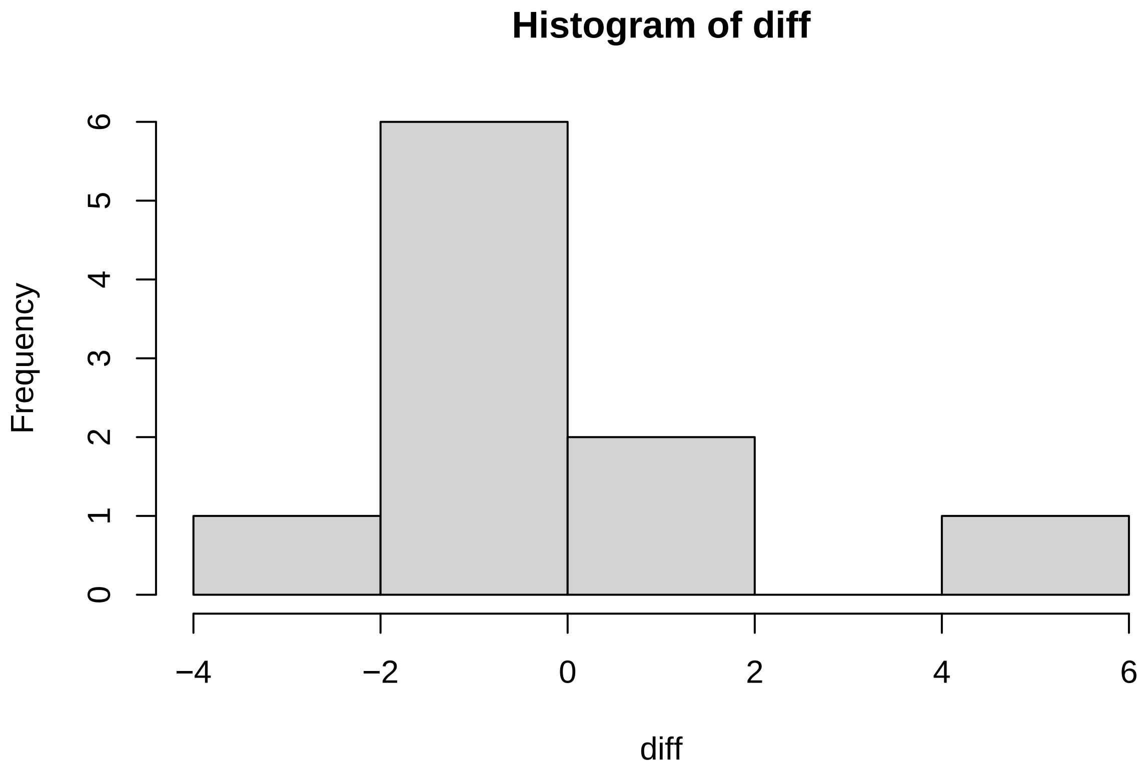
4) If you were to do another allocation, would you get the same difference in means? Explain.

Since we are not replacing any of our numbers and we are randomly choosing which numbers go into which group, we should get a different difference in means.

5) Now, perform nine more random allocations, each time computing and writing down the difference in mean worm count between the control group and the treatment group. Make a dotplot of the 10 differences. What proportion of these differences are 7.6 or larger?

```
for (i in 2:10){
  sample <- sample(Worms_Female, 10, replace = FALSE)
  treatment <- sample[6:10]
  control <- sample[1:5]
  diff[i] <- mean(treatment) - mean(control)
}

hist(diff)
```



```
p_value <- length(diff[diff > 7.6]) / 10
p_value
```

```
## [1] 0
```

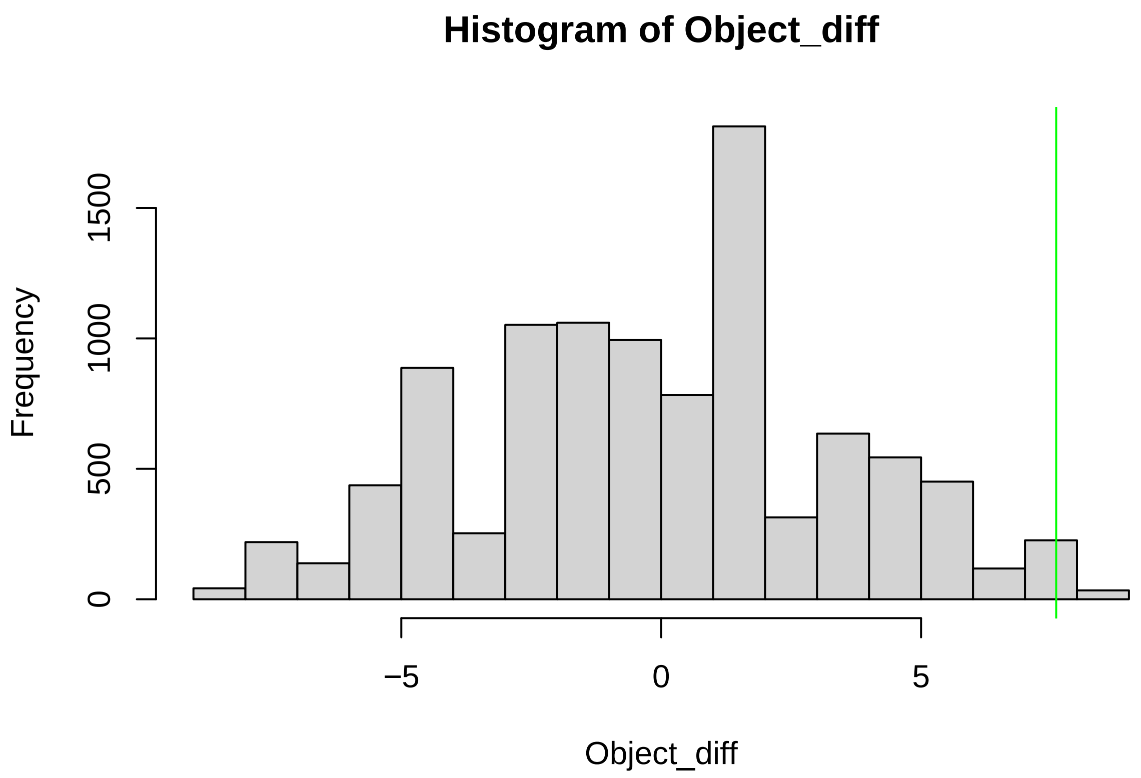
After running the differences of the means multiple times, we find that there are no differences in mean which are higher than 7.6.

6) If you performed the simulation many times, would you expect a large percentage of the simulations to result in a mean differece greater than 7.6? Explain.

```
set.seed(4)
p <- 10000
Worms_Female <- c(Female.Tr, Female.Con)

Object_diff <- rep(0, p)
for(i in 1:p){
  Sample_1 <- sample(Worms_Female, 10, replace = FALSE)
  Object_diff[i] <- mean(Sample_1[6:10]) - mean(Sample_1[1:5])
}

hist(Object_diff)
abline(v = 7.6, col = "green")
```



```
table(abs(Object_diff) > 7.6)
```

```
##
## FALSE  TRUE
##  9924    76
```

```
p_value <- length(Object_diff[Object_diff > 7.6]) / p
p_value
```

```
## [1] 0.0034
```

After running the simulation 10000 times, we find that a very low percent of our mean differences are actually above 7.6. Thus, we would not expect very many mean differences above 7.6.

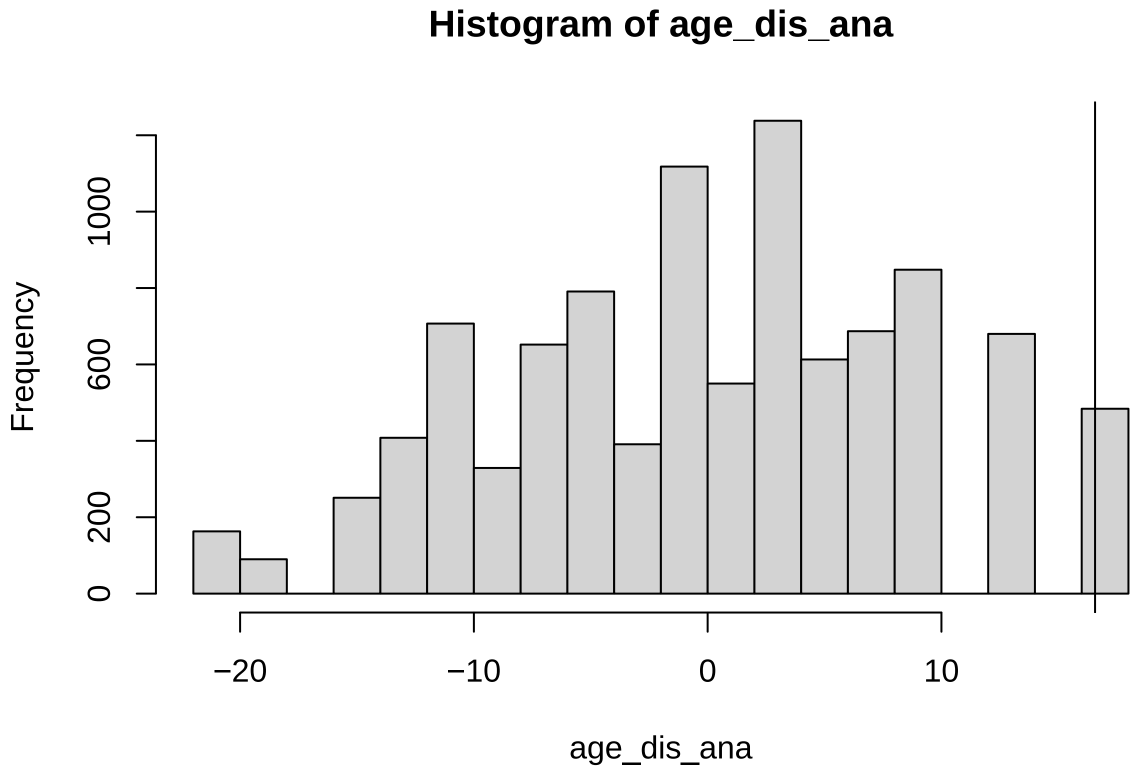
7) Conduct a permutation test to determine whether test to determine whether the observed difference between means is likely to occur just by chance. Use Age as the response variable and Layoff as the explanatory variable. Here we are interested in only a one- sided hypothesis test to determine if the mean age of people who were laid off is higher than the mean age of people who were not laid off.

```
age <- read.csv("C1 Age.csv")

no <- age[age$layoff. == "no", ]
yes <- age[age$layoff. == "yes", ]
Obs_diff <- mean(yes$age) - mean(no$age)

p <- 10000
age_dis_ana <- rep(0, p)
yes_no <- c(no$age, yes$age)
```

```r
for (i in 1:p){
  temp <- sample(yes_no, 10, replace = FALSE)
  age_dis_ana[i] <- mean(temp[8:10]) - mean(temp[1:7])
}
hist(age_dis_ana)
abline(v = Obs_diff)
```



```r
p_value <- mean(age_dis_ana > Obs_diff)
p_value
```

```
## [1] 0.0248
```

After running a permutation test 10000 times, we find the chance that our difference in means is not very likely to be higher than the actual observed difference in means. This gives us a very low p value for our hypothesis test which was that the mean age of those who were laid off was equal to the mean age of those who were not. Thus, we reject the null hypothesis and conclude that the mean age for those who were laid off is higher than the mean age of those who were not laid off.