Comparing Samples With Permutation

eCornell

7/26/2021

Scenario: Use this R Markdown file to help you determine whether two numerical groups of data are different. The example used here compares two job categories from the Prestige data set in R.

Step 1: Load the Prestige data set and create colors.

```r
library(carData) # The Prestige data set is available in the carData library
data(Prestige)

#eCornell Hex Codes:
crimson = '#b31b1b' #crimson
lightGray = '#cecece' #lightGray
darkGray = '#606366' #darkGray
skyBlue = '#92b2c4' #skyblue
gold = '#fbb040' #gold
ecBlack = '#393f47' #ecBlack
```

##Step 2: Specify the hypotheses and create a boxplot of the sample.

Null: Prestige score distributions of blue collar (bc) and white collar (wc) jobs are the same.

Alternative: Prestige score distributions of blue collar and white collar jobs are different.

Sample statistic: mean diff in prestige scores

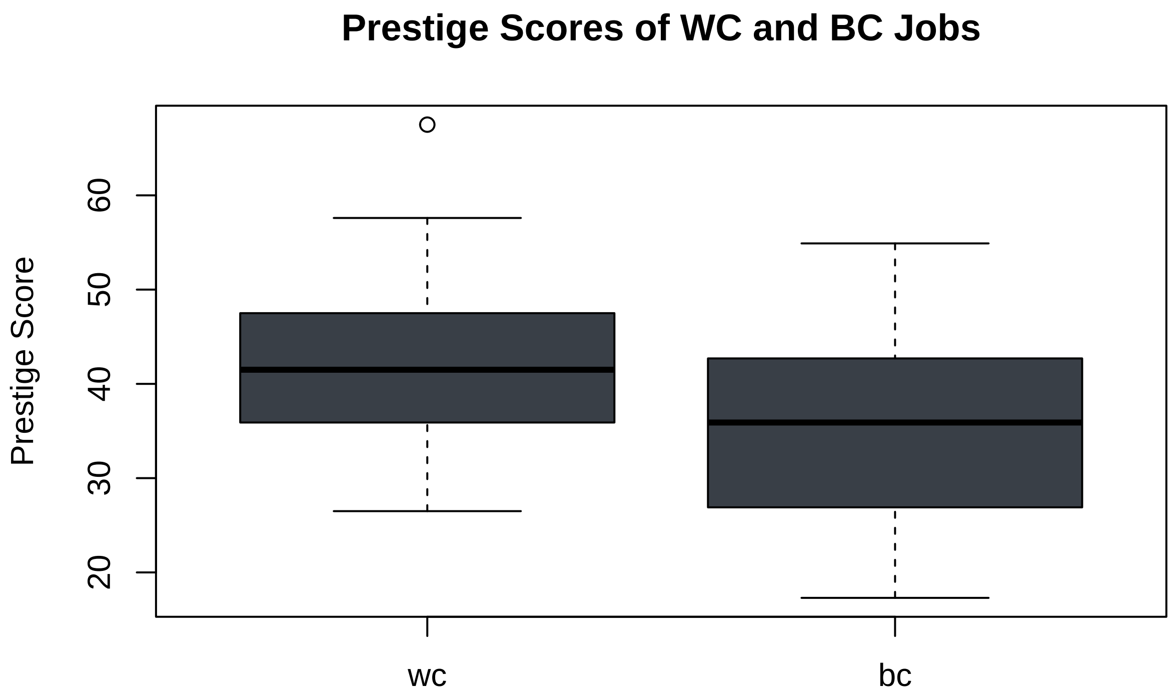
```r
# Exclude rows with 'Professional' in the 'type' column so that we only have two types: wc and bc
Prestige.wb = Prestige[Prestige$type != 'prof',]
# Create a data set that includes only wc jobs:
prestige.wc = Prestige.wb$prestige[Prestige.wb$type == 'wc']
prestige.wc
```

```
##  [1] 67.5 57.6   NA 46.0 41.9 49.4 42.3 47.7 30.9 32.7 38.7 36.1 37.2 38.1 29.4
## [16] 51.1 35.7 35.6 41.5 40.2 26.5   NA 47.3 47.1 51.1   NA   NA
```

```r
# Create a data set that includes only bc jobs:
prestige.bc = Prestige.wb$prestige[Prestige.wb$type == 'bc']
prestige.bc
```

```
##  [1] 34.9   NA   NA 23.3 43.5 51.6 29.7 20.2 54.9   NA 20.8 17.3 20.1   NA 21.5
## [16] 35.3 38.9 25.2 34.8 23.2 33.3 28.8 42.5 44.2 35.9 41.8 35.9 43.7 50.8 37.2
## [31] 28.2 38.1 50.3 27.3 40.9 50.2 51.1 38.9 36.2 29.9 42.9 26.5 48.9 35.9 25.1
## [46] 26.1 42.2 35.2
```

```r
# Create a boxplot of wc and bc jobs:
boxplot(prestige.wc, prestige.bc, names = c('wc', 'bc'),
        ylab = 'Prestige Score',
        main = 'Prestige Scores of WC and BC Jobs',
        col = ecBlack)
```



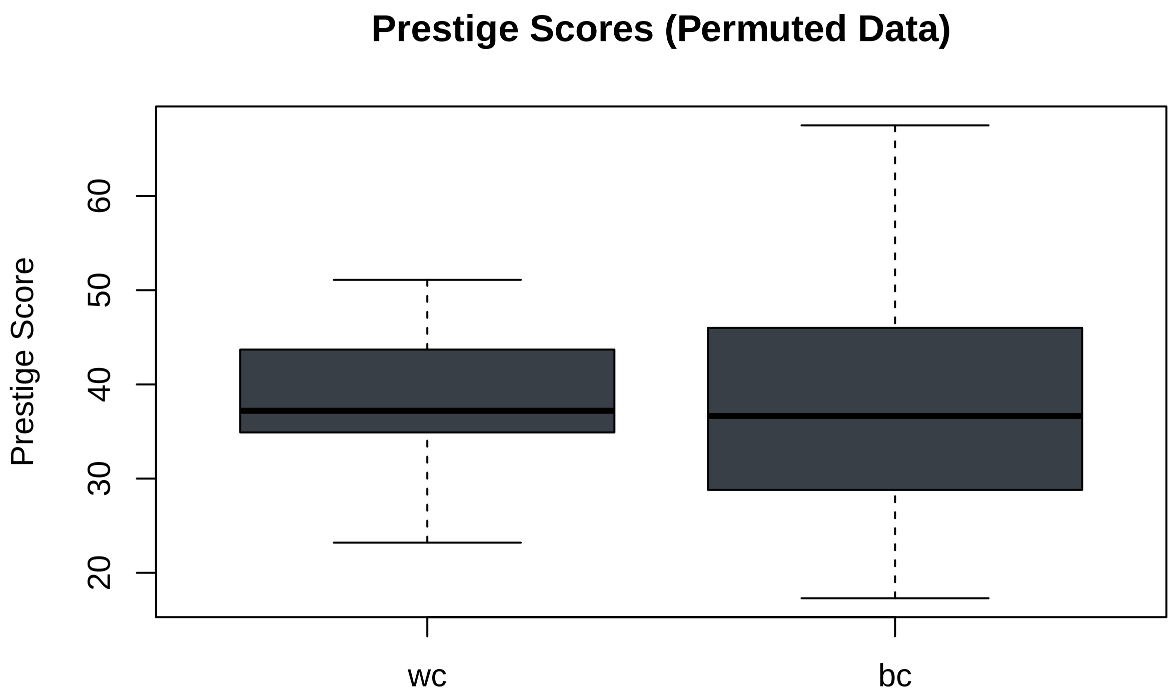
```
obs_stat = mean(prestige.wc, na.rm=TRUE) - mean(prestige.bc, na.rm=TRUE) # na.rm=TRUE takes out any NA
obs_stat
```

```
## [1] 6.716206
```

##Step 3: Calculate the null distribution of the sample statistic. The sample statistic here is the difference between the prestige score of bc and wc jobs.

If there were no difference in the prestige score distributions of bc and wc jobs, randomly permuting the prestige scores across the two groups would not make a difference.

```
set.seed(1) # Set the seed for reproducibility

# Create a data set called Prestige.wb.perm that will be permuted
Prestige.wb.perm = Prestige.wb
# Permute the data set by randomly shuffling prestige values
Prestige.wb.perm$prestige = sample(Prestige.wb$prestige, replace = FALSE)
# Replace the prestige values of the wc jobs with the permuted wc data
prestige.perm.wc = Prestige.wb.perm$prestige[Prestige.wb.perm$type == 'wc']
# Replace the prestige values of the bc jobs with the permuted bc data
prestige.perm.bc = Prestige.wb.perm$prestige[Prestige.wb.perm$type == 'bc']

# Create a boxplot of the permuted data
boxplot(prestige.perm.wc, prestige.perm.bc, names = c('wc', 'bc'),
        ylab = 'Prestige Score',
        main = 'Prestige Scores (Permuted Data)', col = ecBlack)
```



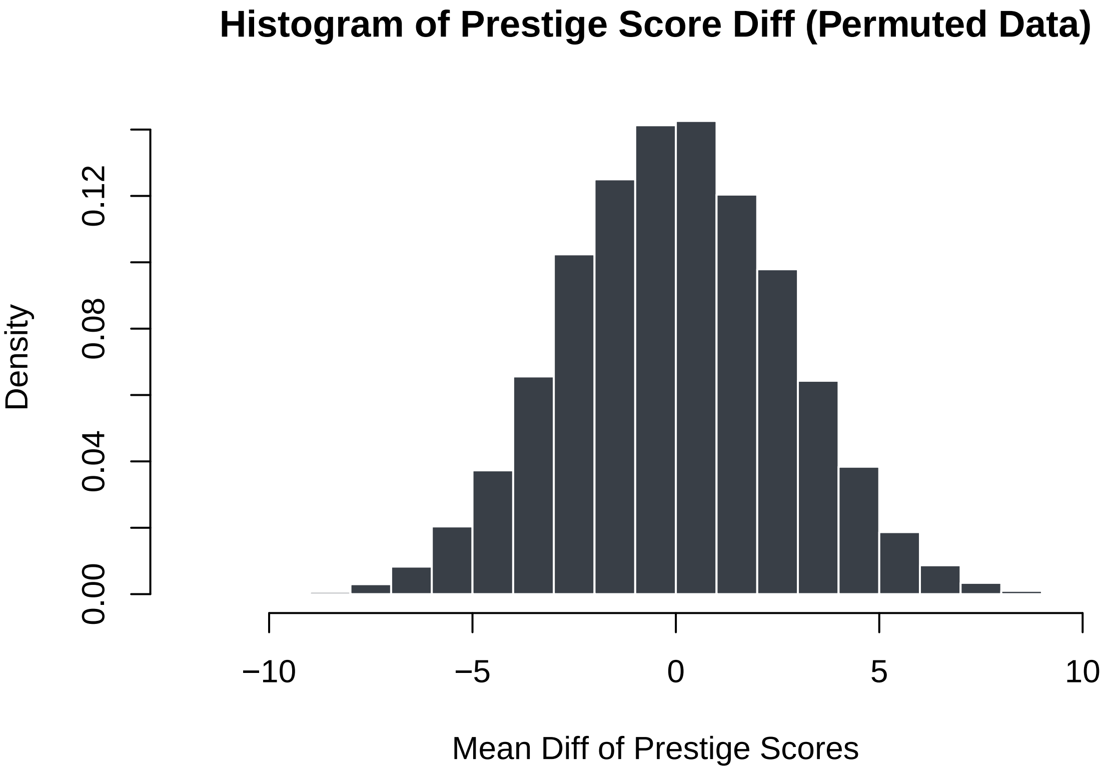
```
mean(prestige.perm.wc, na.rm=TRUE) - mean(prestige.perm.bc, na.rm=TRUE)
```

```
## [1] 0.7452381
```

Step 4: Permute the data set many times and create a histogram of the mean sample statistic.

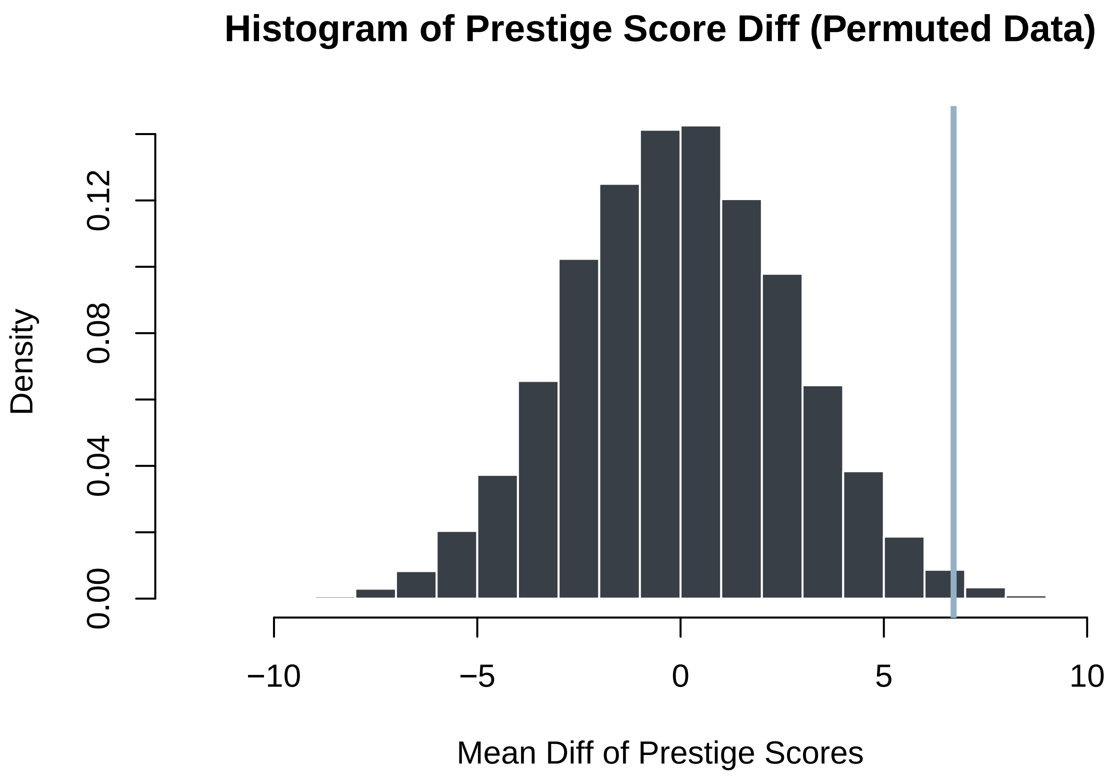
Create many, many permuted data sets and make a histogram of mean prestige score difference.

```
set.seed(1)
P = 10000
store_mean_diff = rep(0, P)

for (n in 1:P){
  Prestige.wb.perm = Prestige.wb
  Prestige.wb.perm$prestige = sample(Prestige.wb$prestige, replace = FALSE)
  prestige.perm.wc = Prestige.wb.perm$prestige[Prestige.wb.perm$type == 'wc']
  prestige.perm.bc = Prestige.wb.perm$prestige[Prestige.wb.perm$type == 'bc']

  store_mean_diff[n] = mean(prestige.perm.wc, na.rm=TRUE) - mean(prestige.perm.bc, na.rm=TRUE)
}

hist(store_mean_diff, breaks = 20, freq = FALSE, col = ecBlack, border = 'white',
     xlab = 'Mean Diff of Prestige Scores',
     main = 'Histogram of Prestige Score Diff (Permuted Data)')
```



##Step 5: Plot the observed statistic on the null distribution.

```r
hist(store_mean_diff, breaks = 20, freq = FALSE, col = ecBlack, border = 'white',
     xlab = 'Mean Diff of Prestige Scores',
     main = 'Histogram of Prestige Score Diff (Permuted Data)')

abline(v = obs_stat, col = skyBlue, lwd = 3)
```



##Step 6: Calculate the p-value and choose the appropriate hypothesis.

```r
# Calculate the p-value:
mean(abs(store_mean_diff) >= abs(obs_stat))
```

```
## [1] 0.0118
```

If null were true, chances of seeing a 6.7 unit or larger difference in average prestige scores is 1%.

So it is reasonable to reject the null hypothesis and conclude that the prestige score distributions of blue and white collar jobs in the population are different.