Lesson 03: Descriptive Statistics

References

- Black, Chapter 3 Descriptive Statistics (pp. 52-95)
- Kabakoff, Chapter 5.2 Numerical and Character Functions (pp. 91-93), Chapter 6.5 Box Plots (pp. 129)
- Davies, Chapter 13 Elementary Statistics (pp. 261-279)
- Stowell, Chapter 5 Summary Statistics for Continuous Variables (pp. 59-62)

Data sets: mileage.csv, shoppers.csv, pontus.csv, geyser.csv

Exercises:

Description: mileage.csv is derived from a 1991 U.S EPA study of passenger car mileage. This file includes information on sixty cars: HP (engine horsepower), MPG (average miles per gallon) WT (vehicle weight in 100 lb units) and CLASS (vehicle weight class C1,.,C6).

```
# Read the comma-delimited text file creating a data frame object in R,
# then examine its structure:

mileage <- read.csv("mileage.csv")
str(mileage)
```

```
## 'data.frame':    60 obs. of  5 variables:
##  $ MAKE : Factor w/ 49 levels "Audi200QuatroWag",..: 17 24 41 27 11 23 25 26 24 31 ...
##  $ HP   : int  90 92 74 95 81 95 92 92 92 103 ...
##  $ MPG  : num  42.2 40.9 40.7 40 39.3 38.8 38.4 38.4 38.4 36.3 ...
##  $ WT   : num  25 25 25 25 25 25 25 25 25 27.5 ...
##  $ CLASS: Factor w/ 6 levels "C1","C2","C3",..: 1 1 1 1 1 1 1 1 1 2 ...
```

1) For each weight class determine the mean and standard deviation of MPG. What can you conclude from these calculations?

```
mpg_class <- aggregate(MPG ~ CLASS, mileage, mean)
mpg_class$SD <- aggregate(MPG ~ CLASS, mileage, sd)[, 2] # [, 2] std devs in second column

mpg_class # low variability within classes, large mean differences between classes
```

```
##   CLASS      MPG        SD
## 1    C1 39.67778 1.3608617
## 2    C2 35.55000 0.5291503
## 3    C3 32.01667 0.6293335
## 4    C4 29.65833 2.1124989
## 5    C5 23.85000 0.7728342
## 6    C6 19.18571 2.7008817
```

2) For each weight class determine the mean and standard deviation of HP. What can you conclude from these calculations?

```
hp_class <- aggregate(HP ~ CLASS, mileage, mean)
hp_class$SD <- aggregate(HP ~ CLASS, mileage, sd)[, 2]

hp_class
```

```
##   CLASS        HP        SD
## 1    C1  89.22222  7.049429
## 2    C2  92.00000  9.086882
## 3    C3 103.50000 12.767145
## 4    C4 123.83333 25.672176
## 5    C5 171.58333 45.350069
## 6    C6 224.71429 74.017372
```

Description: shoppers.csv contains the dollar amounts spent in a store by individual shoppers during one day.

Find the mean, median, range, standard deviation, variance, Q1, Q3 and P10. Plot the histogram and describe the distribution.

```
shoppers <- read.csv("shoppers.csv", header = TRUE)
str(shoppers)
```

```
## 'data.frame':    50 obs. of  1 variable:
##  $ Spending: num  2.32 6.61 6.9 8.04 9.45 ...
```

```
range <- function(x) {max(x, na.rm = TRUE) - min(x, na.rm = TRUE)}

# We'll create a user-defined function to return all our desired summary statistics
# in a data frame.

summary_stats <- function(x) {
  stats <- data.frame(rbind(mean(x, na.rm = TRUE),
                   median(x, na.rm = TRUE),
                   range(x),
                   sd(x, na.rm = TRUE),
                   var(x, na.rm = TRUE),
                   quantile(x, probs = c(0.25), na.rm = TRUE),
                   quantile(x, probs = c(0.75), na.rm = TRUE),
                   quantile(x, probs = c(0.10), na.rm = TRUE)),
             row.names = c("Mean", "Median", "Range", "StdDev", "Var",
                          "Q1", "Q3", "P10"))
  colnames(stats) <- "Value"
  return(stats)
}

summary_stats(shoppers$Spending)
```

```
##            Value
## Mean    25.43640
## Median  20.73500
## Range   61.53000
## StdDev  15.20959
## Var    231.33166
## Q1      14.39250
## Q3      33.66500
## P10     10.17900
```

Description: pontus.csv lists the ages of USA Presidents at the time of their inauguration. Also listed are the heights of the Presidents and their opponents.

```
pontus <- read.csv("pontus.csv")
str(pontus)
```

```
## 'data.frame':    38 obs. of  6 variables:
##  $ President: Factor w/ 37 levels "Buchanan","Carter",..: 36 16 20 23 25 18 19 35 13 29 ...
##  $ Age      : int  57 61 57 57 58 57 61 54 68 49 ...
##  $ Days     : int  2864 1460 2921 2921 2921 1460 2921 1460 31 1460 ...
##  $ Years    : int  10 29 26 28 15 23 17 25 0 4 ...
##  $ Ht       : int  188 170 189 163 183 171 185 168 173 173 ...
##  $ HtOpp    : int  NA 189 170 NA NA 191 171 180 168 185 ...
```

1) Find the mean, median, range, standard deviation, Q1, Q3 and P10 of the Presidents' ages.

```
# We'll use our summary_stats() function from the previous section:
summary_stats(pontus$Age)
```

```
##            Value
## Mean   54.763158
## Median 54.500000
## Range  27.000000
## StdDev  6.561288
## Var    43.050498
## Q1     51.000000
## Q3     57.750000
## P10    46.700000
```

2) Find the mean, median, range, standard deviation, Q1, Q3 and P10 of the heights of the Presidents and also their opponents.

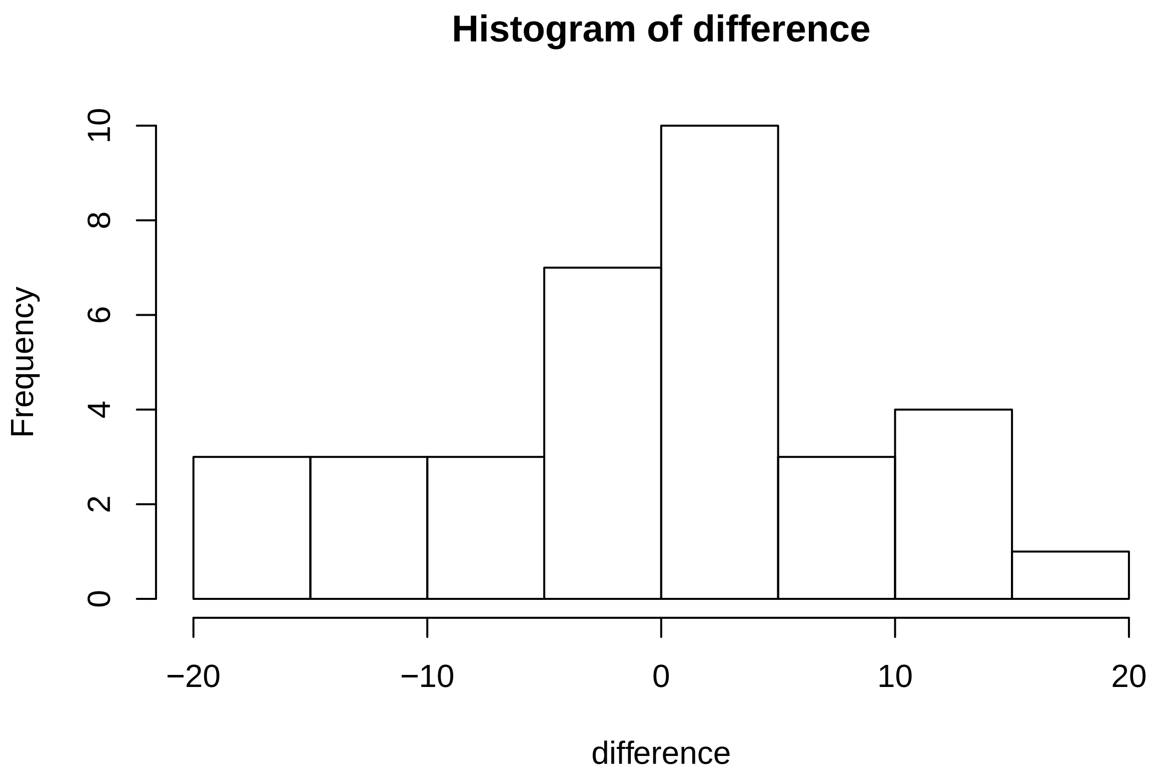
```
apply(pontus[, 5:6], 2, summary_stats) # [, 5:6] Pres. and opponent heights
```

```
## $Ht
##             Value
## Mean   179.684211
## Median 181.000000
## Range   30.000000
## StdDev   7.308289
## Var     53.411095
## Q1     173.000000
## Q3     184.500000
## P10    170.000000
```

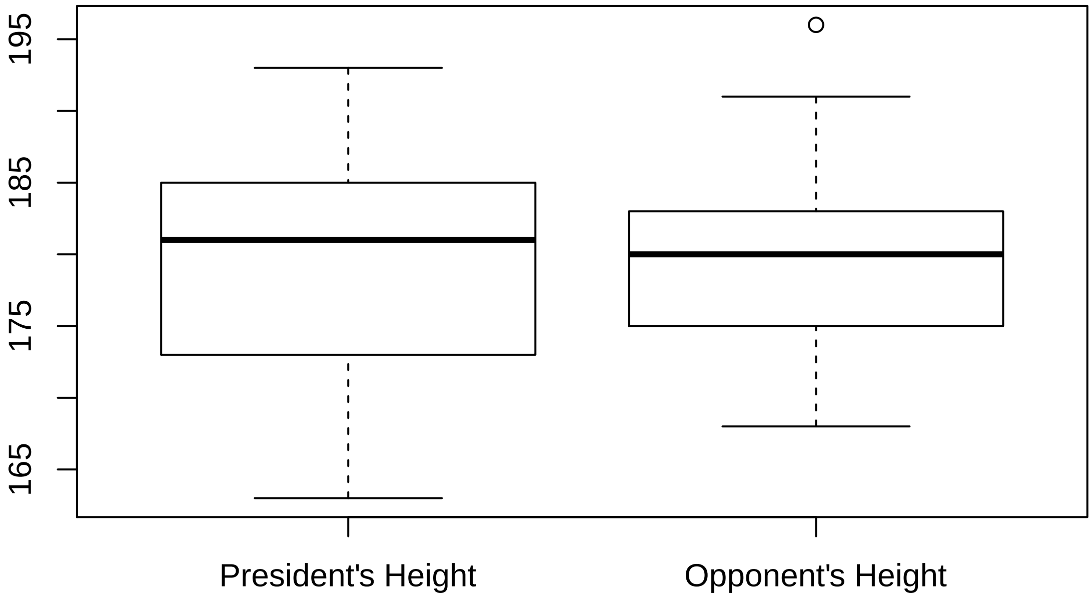
```
##
## $HtOpp
##             Value
## Mean   179.970588
## Median 180.000000
## Range   28.000000
## StdDev   6.201101
## Var     38.453654
## Q1     175.500000
## Q3     182.750000
## P10    173.000000
```

3) Calculate the difference between each President's height and that of his opponent. Plot a histogram of these differences. Construct a boxplot. What do you conclude from your calculations? Why is the difference of average heights calculated in (2) different from the average of the pairwise differences calculated in (3)?

```
difference <- pontus$Ht - pontus$HtOpp
hist(difference)
```



```
with(pontus, boxplot(Ht, HtOpp, names = c("President's Height", "Opponent's Height")))
```



Description: geyser.csv contains the intervals (in minutes) between eruptions of Old Faithful Geyser in Yellowstone National Park. The data were taken on two consecutive weeks: WEEK1 and WEEK2.

Compare the two sets of data using summary(), hist() and boxplot(). What do you conclude?

```
geyser <- read.csv("geyser.csv")
str(geyser)
```

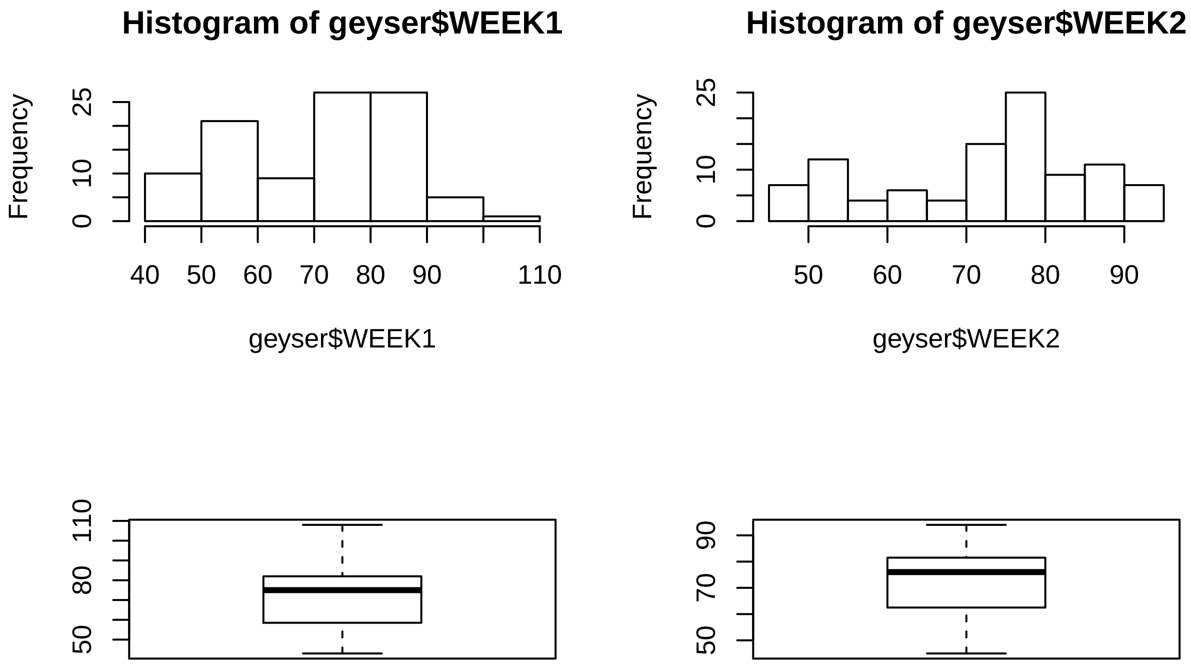
```
## 'data.frame':    100 obs. of  2 variables:
##  $ WEEK1: int  80 71 57 80 75 77 60 86 77 56 ...
##  $ WEEK2: int  56 89 51 79 58 82 52 88 52 78 ...
```

```
apply(geyser, 2, summary)
```

```
##          WEEK1 WEEK2
## Min.     43.00 45.00
## 1st Qu.  58.75 63.25
## Median   75.00 76.00
## Mean     71.62 72.76
## 3rd Qu.  82.00 81.25
## Max.    108.00 94.00
```

```
par(mfrow = c(2, 2))
hist(geyser$WEEK1)
hist(geyser$WEEK2)
```

```r
boxplot(geyser$WEEK1)
boxplot(geyser$WEEK2)
```



```r
par(mfrow = c(1, 1))
```