R Demonstration: Exploratory Data Analysis

Abstract

This demonstration note contains related R codes for the first unit (Exploratory Data Analysis) of STAT 35000. Please get data sets from Canvas. We will use the `class` data to illustrate materials used in this unit.

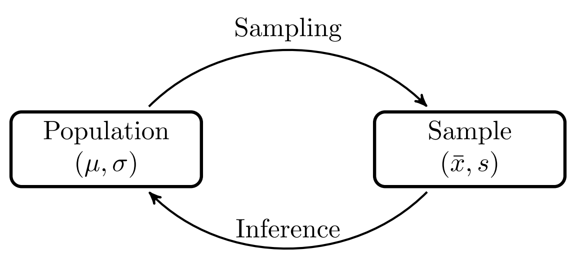
You should download all the files to some folder in your local drive and then set it as the working directory in `RStudio`.

1 Introduction to R and Rstudio

Torfs and Brauer [2014] is a nice short introduction to R. You may check it out to learn the basics of R and Rstudio.

2 Basics

What does statistics study? Statistics is a mathematical science pertaining collection, presentation, analysis and interpretation of data.



Example: We are interested in the *students from Clay Middle School* in Carmel, which is the population of our interest. And we randomly select *19 students* from Clay Middule School as the sample. In particular, we record several characteristics, `name`, `sex`, `age`, `height` and `weight`. And we store them in the following data table `classdata`.

```
classdata = read.table("class.txt", header = TRUE)
```

You can see `class` data set by directly type it

```
classdata
##       name sex age height weight
## 1    Alice   F  13   56.5   84.0
## 2    Becka   F  13   65.3   98.0
## 3     Gail   F  14   64.3   90.0
## 4    Karen   F  12   56.3   77.0
## 5    Kathy   F  12   59.8   84.5
## 6     Mary   F  15   66.5  112.0
## 7    Sandy   F  11   51.3   50.5
## 8   Sharon   F  15   62.5  112.5
## 9    Tammy   F  14   62.8  102.5
## 10  Alfred   M  14   69.0  112.5
## 11    Duke   M  14   63.5  102.5
## 12   Guido   M  15   67.0  133.0
## 13   James   M  12   57.3   83.0
## 14 Jeffrey   M  13   62.5   84.0
## 15    John   M  12   59.0   99.5
## 16  Philip   M  16   72.0  150.0
## 17  Robert   M  12   64.8  128.0
## 18  Thomas   M  11   57.5   85.0
## 19 William   M  15   66.5  112.0
```

When the data is huge, we may only want to take a look of the first several rows, as well as the dimensions of the data table. The following functions `dim` and `head` will be very useful

```
dim(classdata)
## [1] 19  5
head(classdata)
##    name sex age height weight
## 1 Alice   F  13   56.5   84.0
## 2 Becka   F  13   65.3   98.0
## 3  Gail   F  14   64.3   90.0
## 4 Karen   F  12   56.3   77.0
## 5 Kathy   F  12   59.8   84.5
## 6  Mary   F  15   66.5  112.0
```

3 Data Visualization

1. Stem-and-leaf plot
2. Histogram
3. Boxplot

The goal of graphic display is to see the data **distribution** from the following aspects, number of peaks, skewness and outliers.

Example: The number of touchdown passes thrown by each of the 31 teams in the National Football League in 2000 is given below:

```
touchdown = c(14, 29, 22, 18, 20, 15, 6, 9, 32, 18, 19, 18, 23, 28, 37, 21,
    14, 19, 21, 20, 16, 22, 33, 28, 12, 18, 22, 14, 33, 21, 12)
```

How to make a stem-and-leaf plot?

1. Select one or more leading digits for the stem values (any value appropriate). The trailing digits become the leaves.
2. List possible stem values in a vertical column.
3. Put the leaf for each observation besides the corresponding stem.
4. Indicate the units for stems and leaves.

Note that stem-and-leaf display is suitable for a data set with a *small or moderate size*.

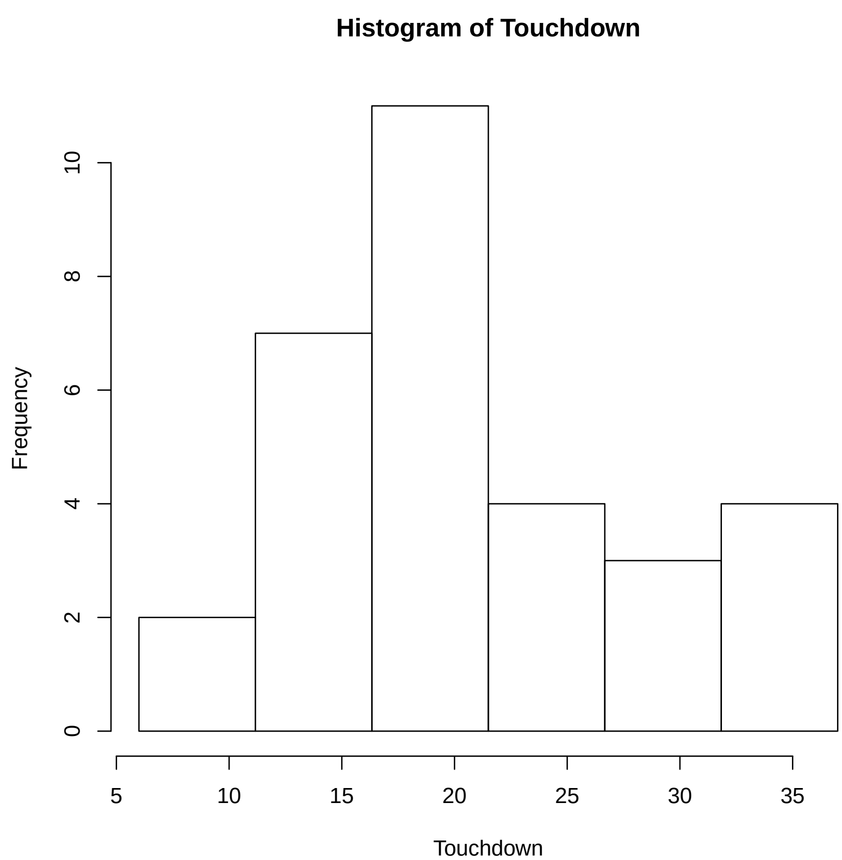
```
stem(touchdown)
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 69
##   1 | 22444
##   1 | 56888899
##   2 | 001112223
##   2 | 889
##   3 | 233
##   3 | 7
stem(touchdown, scale = 0.5)
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 69
##   1 | 2244456888899
##   2 | 001112223889
##   3 | 2337
```

Note that the parameter `scale` can be used to expand the scale of the plot. A value of `scale` = 2 will cause the plot to be roughly twice as long as the default.

How to make a histogram?

1. Decide number of intervals b by Sturge's Rule: $2^{b-1} = n$.
2. Divide the measurement axis into b intervals with equal width such that each obs falls into exactly one interval.
3. Calculate frequency for each interval.
4. Draw a rectangle above each interval with rectangle height=frequency.

```
b = round(log2(length(touchdown))) + 1
b
## [1] 6
hist(touchdown, breaks = seq(min(touchdown), max(touchdown), length = b + 1),
    include.lowest = TRUE, right = TRUE, xlab = "Touchdown", main = "Histogram of Touchdown")
```



Note that `right`=TRUE, just means right-closed (left open) intervals $(a, b]$. And `include.lowest`=TRUE just means the lowest value will be included in the first bar.

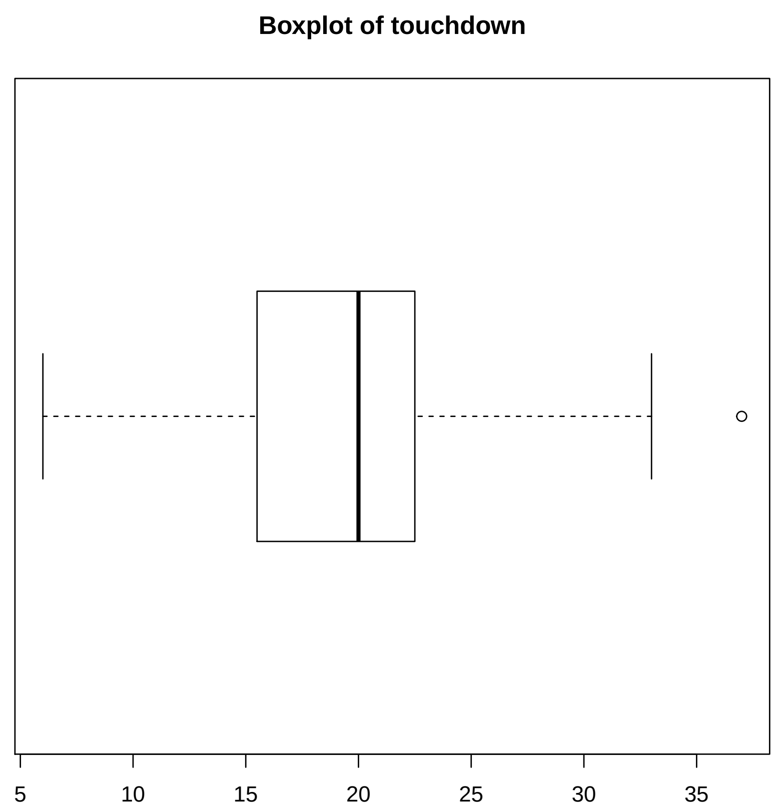
Note that although histogram can not recover the individual data points as the stem-and-leaf does, it's suitable for large data sets.

Boxplot It is very useful in describing several of a data set's important features such as: center, spread, symmetry and outliers.

1. Draw a horizontal axis, find Q_1, Q_2 and Q_3 and calculate IQR.
2. Place a rectangle above the axis, with the left edge at Q_1, right edge at Q_3. And place a vertical line segment inside the rectangle at the location of Q_2.
3. Identify the outliers: any obs farther than 1.5IQR from the nearest quartile is an outlier. And label outliers with star or circle.
4. Drawing a whisker out from the rectangle to the smallest and largest obs that are not outliers.

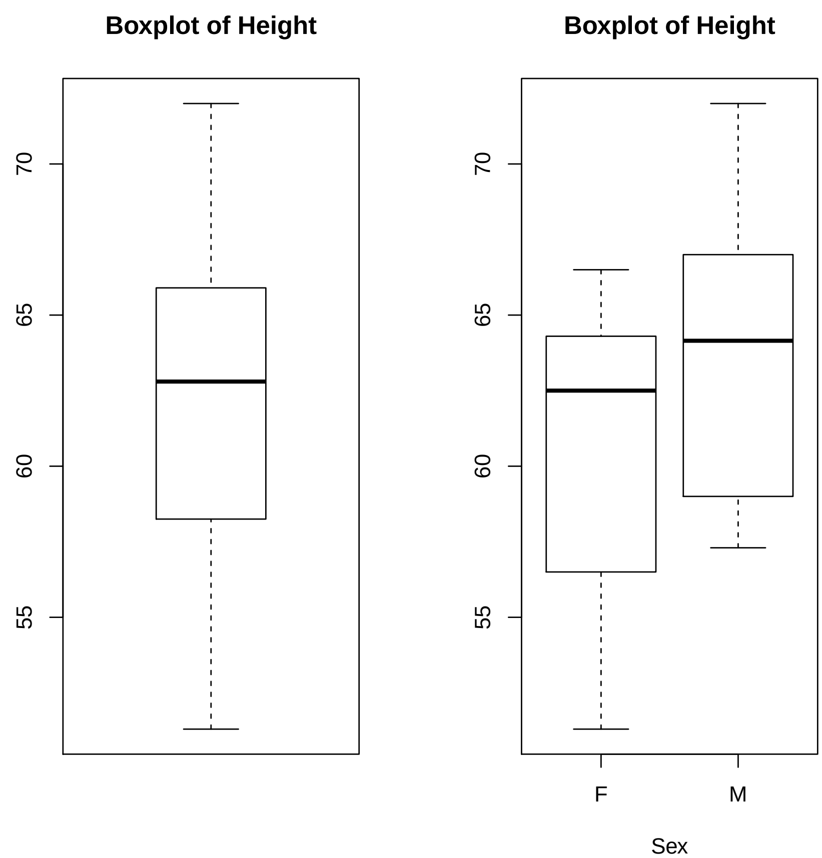
We then can draw boxplot for the touchdown data directly by using `boxplot` function in `R`, which is the same as we draw by hand with the 5-number-summary statistics.

```
summary(touchdown)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   15.50   20.00   20.45   22.50   37.00
boxplot(touchdown, main = "Boxplot of touchdown", horizontal = TRUE)
```



side-by-side boxplot Compare the boxplots of height with respect to gender for the `class` data by using the side-by-side boxplot

```
par(mfrow = c(1, 2))
boxplot(classdata$height, main = "Boxplot of Height")
boxplot(classdata$height ~ classdata$sex, xlab = "Sex", main = "Boxplot of Height")
```



4 Descriptive Statistics

Visual displays give us general ideas about the shape of data distribution. Descriptive statistics give us quantitative measures instead.

1. Measures of center
   - Mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.
   - Median: $\tilde{x}$ = the value in the middle of **sorted** sample. Note that if there are two date points sitting in the middle, then $\tilde{x}$ is the average of the two.
   - Trimmed Mean: Mean of the trimmed off data set.
2. Measures of variability
   - Inter Quartile Range (IQR): IQR= $Q_3 - Q_1$ with Q_1 as the first quartile and Q_3 as the third quartile.

- Variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}x_i^2-n(\bar{x})^2}{n-1}$; standard deviation: $s = \sqrt{s^2}$.

Example: The following sample contains weights (lbs) of basses in a specific lake: $x_1 = 1.22, x_2 = 1.51, x_3 = 1.34, x_4 = 1.60, x_5 = 0.98, x_6 = 1.71, x_7 = 1.82, x_8 = 1.04, x_9 = 1.10, x_{10} = 0.85, x_{11} = 1.08$. Then

1. $\bar{x} = \frac{1.22+1.51+\cdots+1.08}{11} = 1.295455$.
2. Order the data set from smallest to largest: $x_{(1)} = 0.85, x_{(2)} = 0.98, \cdots, x_{(11)} = 1.82$. And $\tilde{x} = x_{(6)} = 1.22$.
3. $\bar{x}_{10\%} = \frac{x_{(2)}+\cdots+x_{(10)}}{9} = 1.286667$.
4. $Q_1 = \frac{x_{(3)}+x_{(4)}}{2} = 1.06, Q_3 = \frac{x_{(8)}+x_{(9)}}{2} = 1.555$
5. IQR$=Q_3 - Q_1 = 1.555 - 1.060 = 0.495$.
6. $\sum_{i=1}^{n} x_i^2 = 19.5015$, $s^2 = \frac{\sum_{i=1}^{n}x_i^2-n(\bar{x})^2}{n-1} = \frac{19.5015-11\times1.295455^2}{10} = 0.1041273$, and $s = \sqrt{s^2} = \sqrt{0.1041273} = 0.3226876$.

```
basses = c(1.22, 1.51, 1.34, 1.6, 0.98, 1.71, 1.82, 1.04, 1.1, 0.85, 1.08)
mean(basses)
## [1] 1.295455
sort(basses)
##  [1] 0.85 0.98 1.04 1.08 1.10 1.22 1.34 1.51 1.60 1.71 1.82
median(basses)
## [1] 1.22
mean(basses, 0.1)
## [1] 1.286667
quantile(basses, c(0.25, 0.5, 0.75))
##   25%   50%   75%
## 1.060 1.220 1.555
summary(basses)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.850   1.060   1.220   1.295   1.555   1.820
var(basses)
## [1] 0.1041273
sd(basses)
## [1] 0.3226876
```

5 Linear Regression

1. Correlation Coefficient: $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)\ s_x s_y} = \frac{s_{xy}}{s_x s_y}$.
2. Coefficient of Determination: $R^2 = r^2$.

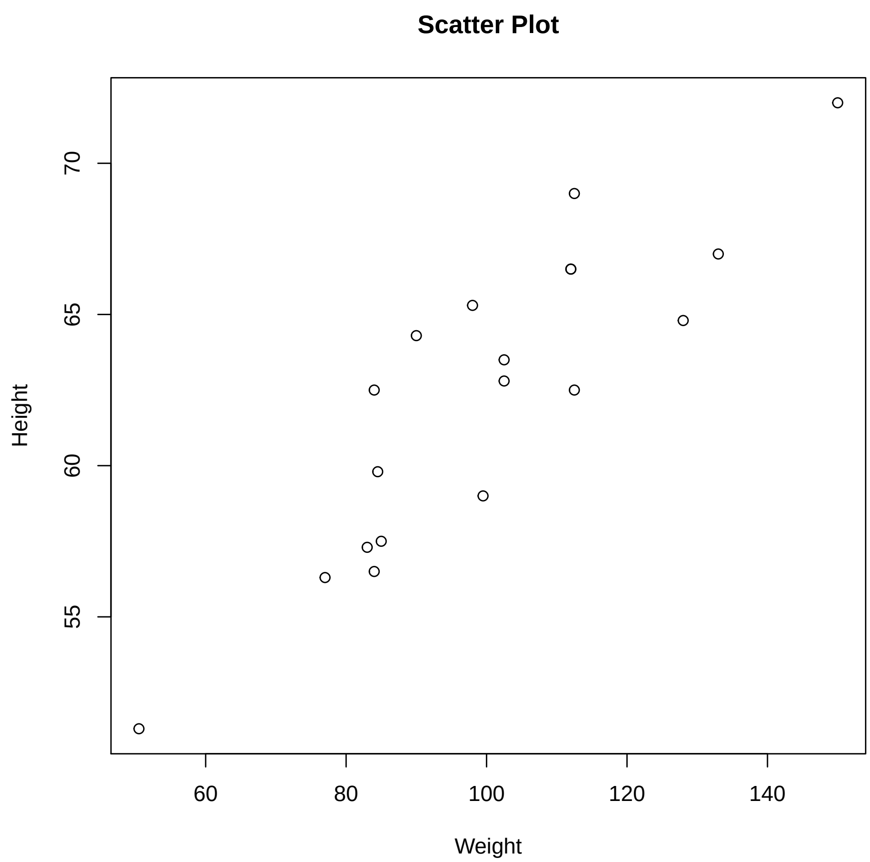
3. Best fitted linear line using least square method: $\hat{y} = \hat{\alpha} + \hat{\beta}x$ with $\hat{\beta} = r\frac{s_y}{s_x}$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

Relationship between two continous variables can be shown by scatter plot. And we need to learn two aspects of the linear relationship between the predictor and response:

1. Direction: positive or negative. It is determined by sign(r). sign $= 1$ positive and sign $= -1$ negative.

2. Strength: weak, moderate or strong. It is determined by $|r|$. $|r| \in [0, 0.3]$: weak; $|r| \in (0.3, 0.8]$: moderate; $|r| \in (0.8, 1]$: strong.

```
plot(classdata$height ~ classdata$weight, xlab = "Weight", ylab = "Height",
    main = "Scatter Plot")
```



The least square regression line can be fitted by using `lm` function

```
fit = lm(classdata$height ~ classdata$weight)
summary(fit)
##
## Call:
## lm(formula = classdata$height ~ classdata$weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2328 -1.8602 -0.2124  1.7970  4.1982
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      42.57014    2.67989  15.885 1.24e-11 ***
## classdata$weight  0.19761    0.02616   7.555 7.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.527 on 17 degrees of freedom
## Multiple R-squared:  0.7705,Adjusted R-squared:  0.757
## F-statistic: 57.08 on 1 and 17 DF,  p-value: 7.887e-07
```

From the results shown above, we can see that the prediction line equation is

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = 42.57014 + 0.19761x.$$

And the coefficient of determination R^2=0.7705, which means there are 77.05% of the variation in the response variable `height` explained by the predictor variable `weight`. And $r = \sqrt{(R^2)} = 0.8777813$, which indicates that the linear relationship between `height` and `weight` is positive and strong.

6 Frequencey Table for Discrete Case

Flip a fair coin four times and we care about the number of Heads. Regard the population as the outcomes of infinite independent trials of flipping the coin 4 times. We are interested in the distribution of the variable: number of Heads by flipping a fair coin 4 times.

```
result = replicate(1000, sum(sample(0:1, 4, replace = TRUE)))
table(result)/1000
## result
##     0     1     2     3     4
## 0.063 0.257 0.351 0.270 0.059
```

References

Paul Torfs and Claudia Brauer. A (very) short introduction to r, 2014.