# STAT 3021 Lab 2

*Ganghua Wang*

Do the following exercise in R to practice concepts we learned in Chapter 1 using R.

## 1. Import data

Doownload "Survey.csv" from Canvas Lab handout module and read the dataset in R. * Learn how to set the working directory : [Session]-[Set Working Directory] - [Choose Directory] *Learn the command to import the data set in R.

```
# getwd()  # Current working directory
# setwd()  # Set wd.

# Several ways to read a file
## 1. Change working directory
Survey<-read.csv("SurveyFall2019.csv")
## 2. Load from Files window
## 3. Load from Environment window
## 4. Use absolute path
Survey<-read.csv("~/MyWebsite/keywgh.github.io/teach/courses/STAT3011/SurveyFall2019.csv")
## 5. Import from online website
Survey<-read.csv("https://keywgh.github.io/teach/courses/STAT3011/SurveyFall2019.csv")

View(Survey)  # Open a new tab to view data
```

## 2. Glance at data

- Use **head** function to see the top 6 rows of your dataset.
- Find the total number of students who took this survey.
- Identify three numeric and three categorical variables in this dataset.

- Identify the number of students in each class (Freshman, Sophomore, Junior, Senior, others)

```
dim(Survey)  #dimention of the dataset
head(Survey)  # top 6 rows
table(Survey$year)  # Summary table for variable year.
```

## 3. Barplot

Use the following command to make a bargraph of the variable year. What are some similarities and differences between two barplots?

```
barplot(table(Survey$year), main="Number of students in each year")
#option main=" " for title of plot

barplot(table(Survey$year)/length(Survey$year), main="Proportion of students in each year")
#length(Survey$year) gives the total number of observations in the data set
```

## 4. Histogram

Draw a histogram for ideal weight of students in the survey. Comment on the shape of the distribution (unimodal, skewed, symmetric, etc.)

```
hist(Survey$ideal.weight)

hist(Survey$ideal.weight, main="Histogram of Ideal Weight", xlab="in pounds", xlim=c(50, 350))
#what do options main = " ", xlab=" ", xlim() do?

hist(Survey$ideal.weight, freq=FALSE, main="Relative frequency of Ideal Weight",
     xlab="in pounds")
# what does freq=FALSE do?

hist(Survey$ideal.weight, freq=FALSE, main="Relative frequency of Ideal Weight",
     xlab="in pounds", breaks=20)
#what does 'breaks=20' do?
```

The behavior of **funtion** is controlled by **keywords**. To learn more about the **hist()** command, type ?hist in the console and hit enter.

```
?hist
```

## 5. Selection by condition.

This time, we will observe the variable ideal weight by gender. Use the following command to construct a histogram of female students' ideal weight. What is the mean and median of female students' ideal weight? What about standard deviation?

```
hist(Survey$ideal.weight[Survey$gender=="Female"])
# R uses an observed ideal.weight value only if the condition inside [] is satisfied.


mean(Survey$ideal.weight[Survey$gender=="Female"])


median(Survey$ideal.weight[Survey$gender=="Female"])


sd(Survey$ideal.weight[Survey$gender=="Female"])
```

What is the total number of female students in this dataset? How many female students' ideal weights are less than or equal to 130 pounds (mean)? What about less than 125 (median)?

```
table(Survey$gender) #to check total number of male and female students

sum(Survey$ideal.weight[Survey$gender=="Female"]<=130)

sum(Survey$ideal.weight[Survey$gender=="Female"]<=125)
#argument inside () returns "TRUE/FALSE"
```

## Summary

Need to know:

- Import data
- Basic summary information about data
- Select specific variable, or observations satisfy some conditions
- Barplot and histogram

Good to know:

- Learn to use keywords, such as control the labels and color of your plot.
- Select by index

For those who are able to learn by themselves(the followings totally won't be mentioned in any exam, but if you're going to take further (advanced) statistics class, you may find them helpful):

- LaTeX
- R markdown