Intro to R

Winter 2017

Intro to R

- R is a language and environment that allows:
  - Data management
  - Graphs and tables
  - Statistical analyses
  - You will need: some basic statistics
    - We will discuss these
- R is open source and runs on Windows, Mac, Linux systems

R Environment

- R is an integrated software suite that includes:
  - Effective data handling
  - A suite of operators for array/matrix calculations
  - Intermediate tools for data analysis
  - Graphical facilities
  - Simple and effective programming language which includes conditionals, loops, functions, I/O

R

- Goals for this section of the course include:
  - Becoming familiar with Statistical Packages
  - Creating new Datasets
  - Importing & exporting Datasets
  - Manipulating data in a Dataset
  - Basic analysis of data (mainly descriptive statistics with some inferential statistics)
  - An overview of R's advanced features

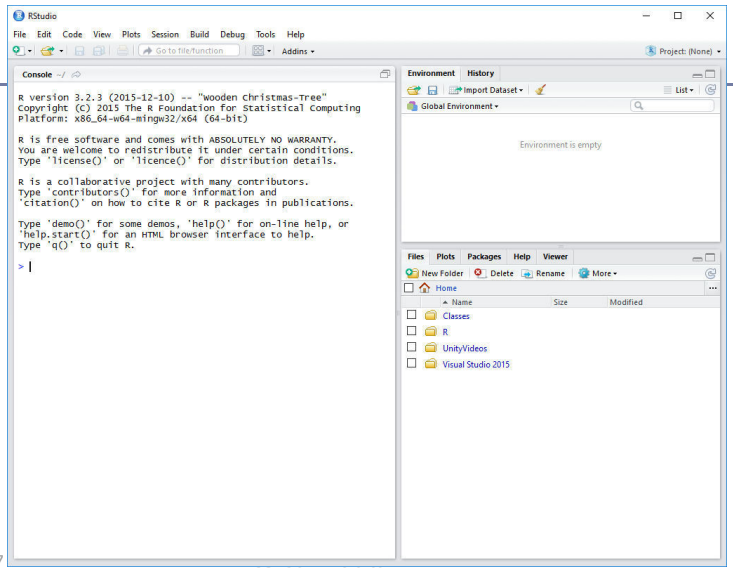
Note: This is not a statistics course such as Math 207. We will only concentrate on basic statistical concepts.

R Resources

- Web site resources:
  - R console application only
    - https://cran.r-project.org/
  - Rstudio IDE
    - https://www.rstudio.com/products/rstudio/download/
    - https://cran.rstudio.com/
  - R documentation
    - http://www.tutorialspoint.com/r/index.htm
    - http://www.cyclismo.org/tutorial/R/index.html

https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

Open RStudio



R Session

- Start an RStudio session
- We will use the console window of RStudio

```
R Console
> 1+2 # calculation
[1] 3
> x=1 # variable assignment
> x # output variable value
[1] 1
> values=c(1,2,3,4,5) # create a vector of values
> values
[1] 1 2 3 4 5
> mean(values) # call function mean using vector values
[1] 3
> help(mean)
```


Basic Datatypes
Numeric

- Numeric – the default datatype for numbers
  - Contains a decimal point

```
> x=10.5 # numeric
> k=1 # still numeric
> is.integer(k)
[1] FALSE
>
```

Basic Datatypes
Logical

- Logical – is either TRUE or FALSE

```
> x = 1; y = 2; z = 1 # assign values to variables
> a = x < y # is x smaller than y ?
> a
[1] TRUE
> b = y == z # is y equal to z ?
> b
[1] FALSE
>
```


Basic Datatypes
Character

- Character – is used to represent **text** values

```
> firstName = "Computer"
> lastName = " Science"
> firstName
[1] "Computer"
> paste (firstName, lastName) # concatenates values together
[1] "Computer  Science"
> pi = as.character (3.14) # force 3.14 to be string
> class (pi)
[1] "character"
> pi * 2 # what happens
```

Basic Datatypes
Integer

- Integer – created using as.integer () function or suffix L as in 2L
  - No decimal point
  - Only use integer in interface with another software package or to save space (memory)

```
> k=as.integer(1)
> k
[1] 1
> is.integer(k)
[1] TRUE
> x=2
> is.integer(x)
[1] FALSE
> j=2L
> is.integer(j)
[1] TRUE
> j
[1] 2
```


Data Structures

http://adv-r.had.co.nz/Data-structures.html

- Combine multiple pieces of data into one variable
- Atomic Vector – often just called *vector*
  - Sequence of data of the same type (1, 2, 3, 9)
- Generic Vector/Lists
  - Sequence of data of many types (100, 200, "oak")
- Matrix
  - Grid of data of the same type $\begin{bmatrix} 1 & 9 \\ 2 & 3 \end{bmatrix}$
- Data Frame
  - Grid of data of many types $\begin{bmatrix} 100 & 200 & "oak" \\ 32 & 40 & "maple" \end{bmatrix}$

Vector

- A sequence of data of the same type
- Six types of atomic vectors
  1. Logical
  2. Integer
  3. Double (Numeric
  4. Character
  5. Complex
  6. Raw
- For now we will concern ourselves with 1-4.

```
> v1=c(1,2,3)
> v2=4:6
> v3=7.1:10.1
> v4=seq(1.1,1.9,by=0.1)
> v3
[1]  7.1  8.1  9.1 10.1
> v4
[1] 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9
```


Measures of Central Tendency

- Used to describe the center of a distribution
- Define each of the following:
  - Mean
  - Median
  - Mode

Problems

- 1) Create a vector of ages in a variable called age with the following integer values: 18, 19, 18, 21, 22, 23, 19, 18
- 2) Compute the mean and median of the age values
- 3) Compute the mean of the first 1000 natural numbers

Problem

- Given the following dataset, find the mean, median, and mode of the Age variable using R

Breed	Age	Weight
Collie	2	23.2
Collie	3	35.7
Setter	5	45.4
Shepard	1	65.9
Setter	2	72.2

An R Solution

- First of all, what do we expect the answers to be?
- Let's use R to check expected results:

1. Create a vector **age** with the Age values
2. Call function mean
3. Call function median
4. Call function mode

Did we get our expected results?

Data Frame

- A data frame is a two-dimensional (2D) structure where
  - column data refers to a variable
  - row data refers to an observation or a case
- Column names are to be unique non-empty.
- Row names are optional but should be unique.
- Allowable types of variable info: numeric, factor or character type.

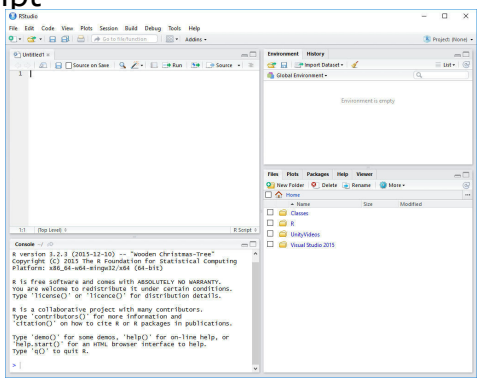
Dog Data Frame Example

- What type is
  Breed?
  Age?
  Weight?

Breed	Age	Weight
Collie	2	23.2
Collie	3	35.7
Setter	5	45.4
Shepard	1	65.9
Setter	2	72.2

Dog Data Frame

- We are going to start creating scripts in Rstudio
- File->New File->R Script

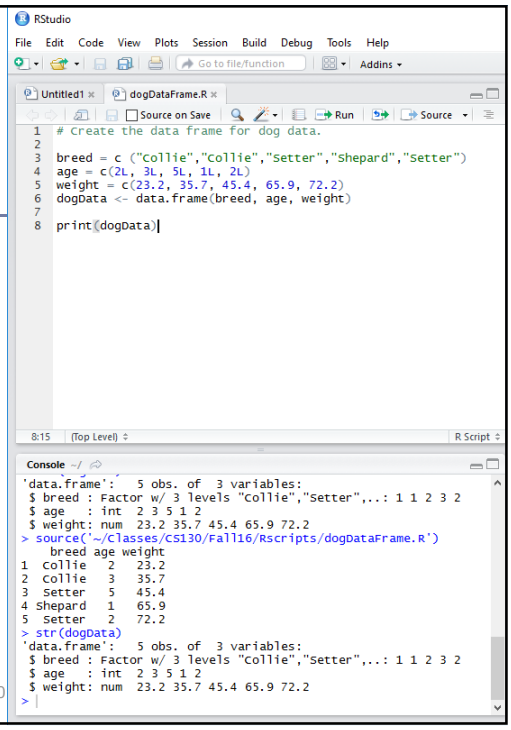


Dog Data Frame

- In the Untitled script window, type the following R script

```
# Create the data frame for dog data.

breed = c("Collie","Collie","Setter","Shepard","Setter")
age = c(2L, 3L, 5L, 1L, 2L)
weight = c(23.2, 35.7, 45.4, 65.9, 72.2)
dogData <- data.frame(breed, age, weight)

print(dogData)
```


Execute the script



Problems

- Find the mean and median of the age and weight variables. Use the console window to do this.

  Hint: Variables of a Data Frame can be specified as dataframe$variable (e.g. dogData$age)

Variables in R

- Let's define the following terms
- Variable
  - Categorical (or Qualitative) Variable
    - Nominal
    - Ordinal
  - Quantitative Variables
    - Numeric
      - Discrete
      - Continuous

Qualitative vs. Quantitative

- Qualitative: classify individuals into categories
- Quantitative: tell how much or how many of something there is

- Which are qualitative and which are quantitative?
  - Person's Age
  - Person's Gender
  - Mileage (in miles per gallon) of a car
  - Color of a car

Qualitative: Ordinal vs. Nominal

- Ordinal variables:
  - One whose categories have a natural ordering
  - Example: grades

- Nominal variables:
  - One whose categories have no natural ordering
  - Example: state of residence

Create Ordinal Values

http://www.statmethods.net/input/valuelabels.html

```
classRank=c(1, 1, 2, 1, 3)

classRankOrdinal = ordered(classRank,
levels=c(1,2,3,4),
labels=c("Fr", "So", "Jr", "Sr") )

print(classRankOrdinal)

barplot(summary(classRankOrdinal))
```


Why do we want ordinal values?

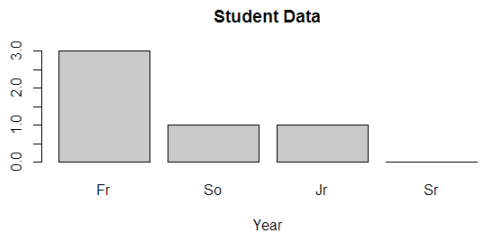
```
classRankNotOrdinal=("Fr", "Fr", "So", "Fr", "Jr")

barplot(table(classRankNotOrdinal))
```

Bar Chart

http://statmethods.net/graphs/bar.html

- A **bar chart** or **bar graph** is a chart that presents grouped data with rectangular bars with lengths proportional to the values that they represent.
- function table returns a vector of frequency data

```
> barplot(table(classRankOrdinal ),
main = "Student Data",
xlab = "Year")
```



Quantitative

- Discrete variables: Variables whose possible values can be listed
  - Example: number of children
- Continuous variables: Variables that can take any value in an interval
  - Example: height of a person

Problem

- Using the command str(dogData), identify:
  - variable name
  - quantitative or qualitative
  - discrete, continuous, neither
  - nominal, ordinal, neither
- A specific variable can be selected and passed to the class function. Pass the variable age of dogData to class. What does the result tell us?

Importing Data into R

- getwd()
- data = read.table("filename.txt", header=FALSE)

- Copy testData.txt from CS130 Public to the location provided by getwd()
- Open testData.txt in a text editor

- testData =read.table("testData.txt", header=TRUE)
- print(testData)
- str(testData)

Candy Dataset Example

http://zeus.cs.pacificu.edu/chadd/cs130w17/candy.txt
This file contains a header

Brand	Name	ServingPerPkg	OzPerPkg	Calories	TotalFatInGrams	SatFatInGrams
M&M/Mars	Snickers Peanut Butter	1.0	2.00	310	20.0	7.0
Hershey	Cookies 'n Mint	1.0	1.55	230	12.0	6.0
Hershey	Cadbury Dairy Milk	3.5	5.00	220	12.0	8.0
M&M/Mars	Snickers	3.0	3.70	170	8.0	3.0
Charms	Sugar Daddy	1.0	1.70	200	2.5	2.5

Write dataframe to file

```
write.table( dataframe, “file.txt”)
getwd()

write.table(candy, “candy.txt”)
```

Go to Documents and open candy.txt in a text editor

Problem

- Identify each of the following for Total Fat in Grams:
  - Minimum:
  - Maximum:
  - Mean:
  - Standard Deviation:

  Use the help feature!

R Visualizing Data

Winter 2017

mtcars Data Frame

- R has a built-in data frame called mtcars
- Useful R functions
  - length(object) # number of variables
  - str(object) # structure of an object
  - class(object) # class or type of an object
  - names(object) # names
  - dim(object) # number of observations and variables
- In the console, call each function using **mtcars** as the object

mtcars Data Frame

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

[1]	mpg	Miles/(US) gallon
[2]	cyl	Number of cylinders
[3]	disp	Displacement (cu.in.)
[4]	hp	Gross horsepower
[5]	drat	Rear axle ratio
[6]	wt	Weight (1000 lbs)
[7]	qsec	1/4 mile time
[8]	vs	V/S (vshape or straight line engine)
[9]	am	Transmission (0 = automatic, 1 = manual)
[10]	gear	Number of forward gears
[11]	carb	Number of carburetors

Recoding Variables

- Copy mtcars to tempMtcars to protect mtcars data
  > tempMtcars = mtcars

- Recode am variable as amCategorical
  > tempMtcars$amCategorical = as.factor (mtcars$am)

  > tempMtcars$amLabels = factor (mtcars$am,
  levels=c('0','1'), labels=c("auto", "manual"))

  > tempMtcars$amOrdered = factor (mtcars$am,
  levels=c('1','0'), labels=c("manual", "auto"),
  ordered=TRUE)

  > barplot(summary(tempMtcars$amOrdered))

  > barplot(summary(tempMtcars$amLabels))

table function

- The table function will return a vector of table counts
- For instance, transmission=table(tempMtcars$am) will return a count of the number of automatic (value is 0) and manual (value is 1) transmission types

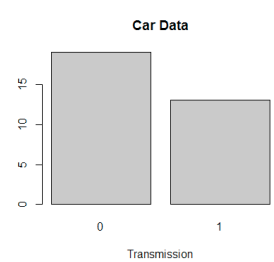
```
> transmission=table(tempMtcars$am)
> transmission

 0  1
19 13
```


Bar Chart

http://statmethods.net/graphs/bar.html

- A **bar chart** or **bar graph** is a chart that presents grouped data with rectangular bars with lengths proportional to the values that they represent.
- function table returns a vector of frequency data

```
> barplot(table(tempMtcars$amCategorical),
main = "Car Data",
xlab = "Transmission")
```



Recoding Variables

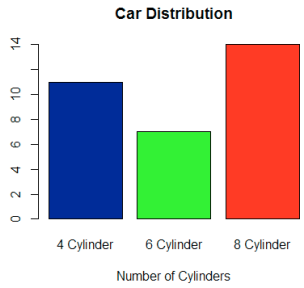
- Create a new variable mpgClass where mpg<=25 is “low”, mpg>25 is “high”

```
> tempMtcars$mpgClass[tempMtcars$mpg <= 25] = "low"
> tempMtcars$mpgClass[tempMtcars$mpg > 25] = "high"
> tempMtcars$mpgClass
[1] "low" "low" "low" "low" "low" "low" "low" "low"
[9] "low" "low" "low" "low" "low" "low" "low" "low"
[17] "low" "high" "high" "high" "low" "low" "low" "low"
[25] "low" "high" "high" "high" "low" "low" "low" "low"
> typeof(tempMtcars$mpgClass)
[1] "character"
```

```
barplot(table(tempMtcars$mpgClass), main = "Car Data",
xlab="MPG")
```


Bar Chart

```
> barplot (table(mtcars$cyl),
main = "Car Distribution",
xlab = "Number of Cylinders",
col = c("darkblue", "green", "red"),
names.arg = c("4 Cylinder", "6 Cylinder", "8 Cylinder"))
```

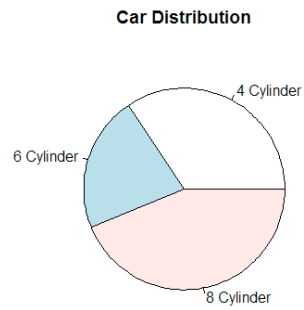


Pie Chart

http://statmethods.net/graphs/pie.html

- A pie chart is a circular graphical representation of data that illustrates a numerical proportion
- A pie chart gives a better visualization of the frequency of occurrence as a percent

```
> pie(table (mtcars$cyl),
labels = c("4 Cylinder", "6 Cylinder", "8 Cylinder"),
main="Car Distribution")
```



Problem

- For the given CS100 class information, create a data frame, cs100DataFrame.R that displays pie and bar chart representations of the Year data properly labeled.

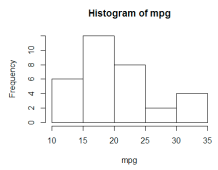
ID	Year	Age
0001	FR	18
0002	FR	18
0003	SR	22
0004	JR	22
0005	SO	19
0006	FR	19
0007	SR	23
0008	SO	19
0009	SR	22

Histogram

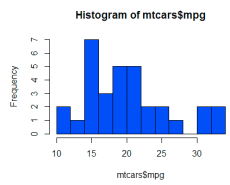
http://statmethods.net/graphs/density.html

- A histogram is a graphical representation of the distribution of numerical data
- Bin – are adjacent intervals usually of equal size
- Notice: breaks <> number of bins and breaks is just a suggestion and not guaranteed

>hist (mtcars$mpg)



> hist (mtcars$mpg, breaks=10, col="blue")



Boxplots

http://statmethods.net/graphs/boxplot.html

- A boxplot is a way of graphically showing numerical data through quartiles
- A box-and-whisker plot is a boxplot that shows variability outside the upper and lower quartiles
- Quartile – the three points that divide the ranked data values into 4 equal sized groups

Quartile Definitions

https://en.wikipedia.org/wiki/Quartile

https://www.mathsisfun.com/data/quartiles.html

http://dsearls.org/other/CalculatingQuartiles/CalculatingQuartiles.htm

- **first quartile/lower quartile/25th percentile/** Q_1
  - splits off the lowest 25% of data from the highest 75%
- **second quartile /median/50th percentile /** Q_2
  - cuts data set in half
- **third quartile/upper quartile/75th percentile /** Q_3
  - splits off the highest 25% of data from the lowest 75%
- **interquartile range /** IQR
  - $IQR = Q_3 - Q_1$

Quartile

https://www.mathsisfun.com/data/quartiles.html

http://dsearls.org/other/CalculatingQuartiles/CalculatingQuartiles.htm

- No universal agreement on computing quartile values.
- We will use the TI-83 method

1. Use the median to divide the ordered data set into two halves.
   - If there are an odd number of data points in the original ordered data set, do not include the median (the central value in the ordered list) in either half.
   - If there are an even number of data points in the original ordered data set, split this data set exactly in half.
2. The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.

Problem Continued

- Using R, show the box-and-whisker plot and quantiles for
  - 6, 7, 19, 20, 42, 100, 200
  - 6, 7, 20, 100, 200

Paint Problem

- Let's put everything together
- A paint manufacturer tested two experimental brands of paint over a period of months to determine how long they would last without fading. Here are the results:

BrandA	**BrandB**	Report on the following
10	25	-Mean
20	35	-Median
60	40	-Mode
40	45	-Std Deviation
50	35	-Minimum
30	30	-Maximum

Paint Problem

1. Using Rstudio, create an R script on your desktop called paintDataFrame.R that creates a data frame paintData for the paint data.
   a) Name the variables brandAPaint and brandBPaint
2. Enter the data
3. Output the data frame
4. Save and run the script. Show me.

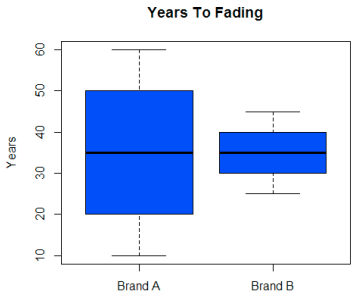
Paint Problem Continued

5. Compute and output the mean, median, std deviation, minimum, and maximum for each brand of paint

```
[1] "Brand A Mean = 35"
[1] "Brand A Median = 35"
[1] "Brand A Std Dev = 18.7082869338697"
[1] "Brand A Minimum = 10"
[1] "Brand A Maximum = 60"
[1] ""
[1] "Brand B Mean = 35"
[1] "Brand B Median = 35"
[1] "Brand B Std Dev = 7.07106781186548"
[1] "Brand B Minimum = 25"
[1] "Brand B Maximum = 45"
```

Paint Problem Continued

5. Output a Box-and-Whisker Plot for each brand of paint as follows. Get as close as possible. This isn't easy but give it a try.
6. What do the descriptive statistics tell us?
7. Which paint would you buy? Justify your answer



HYPOTHESIS TESTING

Winter 2017

Hypothesis Testing

- Hypothesis testing is a decision making process for **evaluating claims about a population.**

- The researcher must:
  - Define the population under study
  - State the hypothesis that is under investigation
  - Give the significance level
  - Select a sample from the population
  - Collect the data
  - Perform the statistical test
  - Reach a conclusion

Population on Samples

- Population: the entire collection of individuals about which information is sought
- Sample: subset of a population, containing the individuals that are actually observed

Population and Samples

Give at least three examples of a population

1.
2.
3.

For the population listed in 1., give an example of a sample from the population

Can you make up some hypothesis about the population in 1.

Hypothesis Tests

- Examples of hypothesis tests include t-test, Chi-Square, and correlation analysis to name a few
- To use this tool properly, you must understand the statistics
- Applying an incorrect test to a given set of data will give incorrect results

Hypothesis Testing

- Hypothesis testing is the formal statistical technique of collecting data to answer questions through the use of a statistical model.
- "In statistics, a result is called **statistically significant** if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the **significance level**."

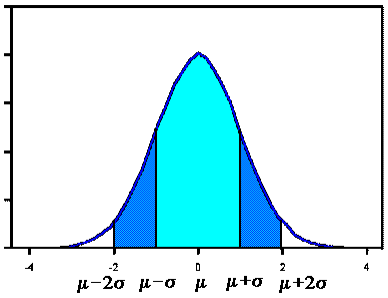
http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

NULL Hypothesis

- The Null Hypothesis refers to a general or default position – denoted H_0
- Null Hypothesis is assumed true until evidence indicate otherwise

The Normal Distribution

- The following Hypothesis Tests assume that the data is normally distributed.
- The standard normal curve in the picture has a mean of 0 and standard deviation of 1. A dataset with a normal distribution has about 68% of the observations within σ of the mean μ which in this case is (-1,1)



http://www.stat.yale.edu/Courses/1997-98/101/normal.htm

The Normal Distribution Continued

- About 95% of the observations will fall within 2 standard deviations of the mean (-2,2)
- About 99.7% of the observations will fall within 3 standard deviations of the mean
- Example: Consider 130 observations of body temperature with the results below. If the data is normal, what must be the case?

Variable	N	Mean	Median	StDev	Min	Max
BODY TEMP	130	98.249	98.300	0.733	96.300	100.800

Hypothesis Tests

- We will be using the following hypothesis tests in this course:
  - One sample t-test
  - Unpaired or independent samples t-test
  - Paired t-test
  - Correlation analysis

One-Sample T-Test

- This is the easiest of the statistical tests to understand
- Compare observed vs hypothesized mean
  - Observed: measured
  - Hypothesized: we choose this value to be meaningful
- T-Test determines the likelihood that the difference between the means occurs by chance
- The chance is reported as the p-value

p-value

- p-value: the probability that the difference occurs due to chance
  - A small p-value means that the difference is unlikely to be the result of chance
  - A large p-value means the difference is likely to be the result of chance
- What do we mean by random chance? Keep this question in mind and we will come back and give an answer.

Statistically Significant Difference

- The lower the p-value, the more certain that we can be that there is a **statistically significant** difference
- Most disciplines look for a p-value of 0.05 or less
  - if $p < 0.05$, reject the null hypothesis
  - if $p >= 0.05$, do not reject the null hypothesis

Problem 11.1

The file LipidData in the CS130 Public directory represents a blood lipid screening of medical students.

1. Grab this Excel file, open it up in Excel.
2. What is the mean Cholesterol value?
3. Is the cholesterol level significantly greater than 190? Can you tell by looking at the data? What do you think?

Problem 11.1

How to import Excel file into R

1. Prepare workspace rm(list=ls())
2. What directory are you working in getwd()

```
> getwd()
[1] "C:/Users/ryandj/Documents"
```

3. Change the location of your data set setwd("location")
4. Install Readxl package and then activate the package

```
> install.packages("readxl")
Installing package into 'C:/Users/ryandj/Documents/R/win-library/3.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.3/readxl_0.1.1.zip'
Content type 'application/zip' length 801404 bytes (782 KB)
downloaded 782 KB

package 'readxl' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ryandj\AppData\Local\Temp\RtmpmOvYkN\downloaded_packages
> library(readxl)
```


Problem 11.1

How to import Excel file into R

5. Copy the LipidData.xlsx from CS130Public to your Desktop
6. Import the data into R

```
> lipiddata=read_excel("LipidData.xlsx")
```

```
Global Environment
Data
lipiddata      95 obs. of 24 variables
  Name : chr "J. Suds" "T. Wilson" "D.S. Quintent" "…
  Gender : chr "male" "female" "male" "female" ...
  Age : num 22 22 22 22 25 22 23 24 23 22 ...
  Weight : num 138 115 190 115 160 150 154 185 178 1…
  Cholesterol : num 197 181 190 131 172 233 194 155 …
  Triglycerides : num 152 59 117 54 93 176 79 89 307…
  HDL : num 43 60 41 58 49 42 49 45 28 50 ...
  LDL : num 151.6 120.1 147.1 72.1 121.5 ...
  idealBodyWtPCT : num 92.8 100 106.7 79.3 87 ...
  Height : num 67.1 63 72 69 73 ...
  Skinfold : num 28 26 30 14 21 32 18 16 5 16 ...
  SystolicBP : num 124 122 124 120 138 100 128 128 1…
  DiastolicBP : num 78 70 80 70 92 72 78 74 82 88 ...
  Weight-3yr : chr "145" "122" "190" "105" ...
  idealweight-3yrPCT : chr "97.633136099999987" "106…
  Trig-3yrs : chr "135" "57" "86" "72" ...
  Chol-3yrs : chr "182" "151" "169" "133" ...
  HDL-3yrs : chr "34" "48" "37" "67" ...
  LDL-3yrs : chr "145.84" "102.08799999999999" "130.…
  ExerciseFreqMinPerWeek: num 180 0 90 120 40 0 0 90…
  CoffeeIntakeCupsPerDay: num 1 2 0 5 2 0 2 0 1 0 ...
  SmokingHistory : chr "no" "no" "no" "no" ...
  HeartHistory : chr "none" "none" "none" "none" ...
  CholesterolLoss : chr "15" "30" "21" "-2" ...
```

Problem 11.1 Continued

- Our first objective is to perform a one-sample t-test on data from blood lipid screening of medical students. Specifically, we will test whether the mean cholesterol level is different than 190 in a *statistically significant* way, the point at which cholesterol levels may be unhealthy.
- What is the NULL hypothesis?
- What is the alternative hypothesis?

Problem 11.1 Results

```
> t.test(df$Cholesterol,mu=190)

        One Sample t-test

data:  df$Cholesterol
t = 0.33649, df = 94, p-value = 0.7373
alternative hypothesis: true mean is not equal to 190
95 percent confidence interval:
 183.9644 198.4988
sample estimates:
mean of x
 191.2316
```

Problem 11.1 Results

- The mean is slightly higher than 190; however, this difference is well within the range of sampling variance.
- A significance level of .737 indicates you would see a difference of this magnitude by chance more than 73% of the time
- Thus the cholesterol level is not significantly different than 190

Paired T-Test

- The most common use of the paired t-test is the comparison of two measurements (typically one measurement occurs “before” a treatment and the other “after” a treatment from the **same** individual or group.
- This test can determine if the treatment had a statistically significant effect.
- The p-value is the primary statistic of concern and the interpretation of the p-value is the same as for the one-sample t-test

Problem 11.2

- Using the LipidData

1. What is the mean for Triglycerides?
2. What is the mean for Trig-3yrs?
3. Does it look like there is a statistically significant difference between Triglycerides and Trig-3yrs?

Problem 11.2 Continued

- Perform the paired t-test using the LipidData file
- State the Null Hypothesis and the alternative hypothesis
- There are only 43 students that have a before and after. How to we create tri43 (the before students) and tri433yrs (the after students)? Notice: These variables are not part of the data frame

```
> t.test(tri43, tri433yrs,paired=TRUE)

        Paired t-test

data:  tri43 and tri433yrs
t = 0.38594, df = 42, p-value = 0.7015
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.45745  21.29466
sample estimates:
mean of the differences
               3.418605
```

- Should we accept the Null Hypothesis? Why?
- State your conclusion

Unpaired T-Test

- One measurement per individual
- Break our population into two natural subgroups
  - Male/Female; Smoker/Non-Smoker; Oak/Maple
  - Do the groups have a difference in measurement?
- Our primary statistic of concern is the p-value
  - How likely to occur by chance?

Problem 11.3

Question: Are the prices of houses near the Charles River more expensive than the prices of houses away from the Charles River.

The file BostonHousingData in the CS130 Public directory contains information about Boston houses.

1. Grab this Excel file, open it up in R
2. State the Null Hypothesis and the alternative hypothesis
3. Perform an unpaired t-test

Problem 11.3

- What is the test variable? Why?
- What is the grouping variable? Why?
- Is the grouping variable in the data set a Factor? If not, make it a factor.

```
        Welch Two Sample t-test

data:  bostonhousing$MedianValue by bostonhousing$Charles
t = -3.1133, df = 36.876, p-value = 0.003567
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.476831  -2.215483
sample estimates:
 mean in group Far mean in group Near
          22.09384           28.44000
```


Problem 11.3

- Do you reject the Null Hypothesis? Why?
- State your conclusion

Correlation Analysis

- Correlation Analysis addresses the following: Is there a statistically significant association between variable X and variable Y?
- Interpreting the Pearson Correlation Coefficient is not an exact science. We might use the following interpretation:
  - -1.0 to -0.7 strong negative association
  - -0.7 to -0.3 weak negative association
  - -0.3 to +0.3 little or no association
  - +0.3 to +0.7 weak positive association
  - +0.7 to +1.0 strong positive association

Correlation Analysis Visual

- Use Scattergrams (Scatterplots) to visually display data analyzed with this test.
- You can also produce a correlation matrix of the relationship of all variables in the matrix.

http://www.statmethods.net/stats/correlations.html

Correlations

```
cor(dataframe)
cor(mtcars, use="complete.obs")
install.packages("Hmisc")
library(Hmisc)
rcorr(as.matrix(mtcars))
x <- mtcars[1:3]
y <- mtcars[4:6]
print(x)
cor(x,y)
rcorr(as.matrix(mtcars[1:6]))
```

http://www.statmethods.net/stats/correlations.html

Problem 11.4

- What is the correlation between Cholesterol and Triglycerides?
- What is the correlation between Cholesterol and LDL?
- How would you graph either of these relationships?