Chapter 1 - Data

Read sections 1.1 - 1.3

Statistics consist of three major areas:

- Data Collection (sampling plans and experimental designs)
- Descriptive Statistics (numerical and graphical summaries)
- Inferential Statistics (confidence intervals and hypothesis testing)

Statistical procedures are part of the **Scientific Method** (steps 2-5 below) first espoused by Sir Francis Bacon (1561-1626), who wrote "to learn the secrets of nature involves collecting data and carrying out experiments." The modern methodology:

1. Observe some phenomenon
2. State a hypothesis explaining the phenomenon
3. Collect data
4. Analyze the data and Test: Do the data support the hypothesis?
5. Conclusion. If the test fails, go back to step 2.

If you encounter a "scientific claim" that you disagree with, scrutinize the steps of the scientific method used. "Statistics don't lie, but liars do statistics." - Mark Twain.

Individuals or **Cases** or **Units**: The objects from which data is collected. Individuals may be people, places, animals, things, or time periods.

Variable: Any characteristic of an individual that can be measured.

Two Types of Variables:

- **Categorical** or **Qualitative** - The possible values are *categories* or *levels*. Beware, some category names are actually numbers (e.g. zip codes and dates)
- **Numerical** or **Quantitative** - The possible values are *numbers* so that mathematical operations, such as averaging, make sense!

QUESTION: Categorical or Numerical?

1. Lifetime of a battery:
2. Type of battery:
3. Distance to school:
4. UPC code on a box of cereal:

Two Types of Numerical Variables:

- **Discrete** - The possible values are isolated points on the number line. Discrete variables can be either:
  - **finite** (e.g. the number of beers left in a six pack: 0, 1, 2, 3, 4, 5 or 6)
  - **infinite** (e.g. the number of (full) minutes until the next terrorist attack: 0, 1, 2, 3, ... , ∞).
- **Continuous** - The possible values are an interval on the number line (e.g. the distance between any two students in this classroom (in feet) is in the interval [0,50) - all real numbers between 0 and 50, including 0 and excluding 50).

QUESTION: Discrete or Continuous?

1. Amount of money on you:
2. Your height:
3. Reaction time:
4. Number of children you have:

Population: The entire group of individuals that we want information about. For example: all grizzly bears in Yellowstone National Park; all G.E. light bulbs (made now and in the future); all tosses with a weighted die

Sample: A part of the population from which data is collected. For example: 22 tagged grizzly bears in Yellowstone National Park; 1 box of G.E. light bulbs; 100 tosses with a weighted die.

Typically, it is unrealistic to obtain a **census** (i.e., data from the entire population of interest). So one collects data from a sample and uses the sample results to draw conclusions about the population. This process is called **Inference**.

Explanatory Variable vs. Response Variable: One or more variables (**explanatory variables**) are used to predict or explain the values of another variable (**response variable**).

Obtaining and Installing R

1. Visit http://cran.r-project.org. This is the website for The Comprehensive R Archive Network (CRAN) from which you can download R and R packages.

2. The first box on this page is labeled *Download and Install R*. In that box, click on the appropriate link. For example, MAC users will click on *Download R for (Mac) OS X* and Microsoft Windows users will click on the link *Download R for Windows*. The rest of these instructions are specific to Windows users.

3. On the new page, click on the link named *base*.

4. On the new page, the link *README.R-2.4.1* provides a brief synopsis on installation and other instructions for R version 2.4.1 for Windows. You shouldn't need to look at this file, but take a look if you get into trouble.

5. Click on the link *Download R 3.3.1 for Windows* to download the executable file R-3.3.1-win.exe to the hard drive on your computer.

6. Exit from your Internet Browser. Open Windows Explorer. Go to the folder in which you saved R-3.3.1-win.exe and run the program.

7. You will be guided through the installation by a Setup Wizard.

8. There are many excellent resources for using R. One interactive site is at http://www.math.csi.cuny.edu/Statistics/R/simpleR, called "Simple R" by John Veranzi.

9. Special-purpose software routines are bundled as separate "packages." Some packages are automatically downloaded when R is downloaded. To download additional packages, execute R on your PC and then click on the tab *Packages* from one of the tabs at the top of the screen.

From the drop down menu, click on *Install package(s) ...* and then choose the package(s) that you want to download. The packages that you will need to download for this course are the following:

- lattice
- pastecs

MASS is another package which we will be using which you do NOT need to download because it is a part of "base R."

Entering Data into R

Lactococcus lactis and *Leuconostoc citrovorum* are two common bacteria used for making cotttage cheese. While developing a new type of cottage cheese, a large dairy producer has added the fungus *Penicillium candidiuma* (PC), typically used to make Brie, to their cottage cheese recipe. One part of the process used to make cottage cheese involves cooking curdled milk for about an hour. A researcher is interested in determining whether adding PC increases the cooking time. Seven dairiy facilities (referred to as A, B, ..., G) make two batches of cottage cheese, one with and one without the fungus PC. The cooking time in minutes was recorded for each batch. The results of the experiment are in a text file called "dairy.txt" which is shown below:

```
Dairy Treatment Time
A withPC 68
A withoutPC 61
B withPC 75
B withoutPC 69
C withPC 62
C withoutPC 64
D withPC 86
D withoutPC 76
E withPC 52
E withoutPC 52
F withPC 46
F withoutPC 38
G withPC 72
G withoutPC 68
```

Text data files that are tab or space delimited can be imported into R. This means that the names of the variables in the file cannot have spaces in them (e.g. don't use "Cook Time"). To get dairy.txt into R, execute the following command:

```
> D = read.table("dairy.txt",header=TRUE)
```

read.table is a *function*, and the *parameter* **header=TRUE** tells R that the first line of the file contains the variable names of each of the columns of data. You could end up with an error like:

```
Error in file(file, "r") : unable to open connection
In addition:
Warning message: cannot open file 'dairy.txt'
```

The above error occurred because dairy.txt was not in the **working directory**. To change the working directory to the one where dairy.txt resides, in R, click on tab **File** → **(Change dir ...)** and you will see a **Choose Directory** window appear. In this window, you can directly enter the directory that contains diary.txt on your computer, or you can hit the Browse button to find the directory. Once you find the directory that contains dairy.txt, then (click OK in the Browser Window if you hit the Browse button and then ...) click OK in the **Choose Directory** window. Now we can try to read the data into R again.

```
> D = read.table("dairy.txt",header=TRUE)
```

The R-variable **D** that contains the data is called a **data frame**. We could have used any variable name like "DairyData" "CCheese", but I don't like to type much, so I used "D". Note that you can not have spaces in your R-variable names! Type the variable name at the R prompt to see what the data looks like:

```
> D
   Dairy Treatment     Time
1      A    withPC       68
2      A withoutPC       61
3      B    withPC       75
4      B withoutPC       69
5      C    withPC       62
6      C withoutPC       64
7      D    withPC       86
8      D withoutPC       76
9      E    withPC       52
10     E withoutPC       52
11     F    withPC       46
12     F withoutPC       38
13     G    withPC       72
14     G withoutPC       68
```

To access the individual columns of the data in D, type

```
> D$Dairy
 [1] A A B B C C D D E E F F G G
Levels: A B C D E F G
> D$Treatment
 [1] withPC    withoutPC withPC    withoutPC withPC    withoutPC withPC
 [8] withoutPC withPC    withoutPC withPC    withoutPC withPC    withoutPC
Levels: withoutPC withPC
> D$Time
 [1] 68 61 75 69 62 64 86 76 52 52 46 38 72 68
```

Or you can execute

```
>attach(D)
> Dairy
 [1] A A B B C C D D E E F F G G
```

```
Levels: A B C D E F G
> Treatment
 [1] withPC    withoutPC withPC    withoutPC withPC    withoutPC withPC
 [8] withoutPC withPC    withoutPC withPC    withoutPC withPC    withoutPC
Levels: withoutPC withPC
> Time
 [1] 68 61 75 69 62 64 86 76 52 52 46 38 72 68
```

R is case-sensitive! The upper and lower-case letters in the variable name must be EXACTLY as given in the data file or R will not find it. For example,

```
> TIME
Error: object "TIME" not found
> D$time
NULL
```

Notice that R recognizes that **Dairy** and **Treatment** are categorical variables and gives the *levels* or categories associated with each. The variable **Time** is recognized as a quantitative variable.

In addition to **read.table**, we will be using many other functions that R has available. For example, **mean()** calculates the mean and **median()** calculates the median. The functions **sd()** and **var()** calculate the standard deviation and variance respectively. For example:

```
> mean(Time)
[1] 63.5
> median(Time)
[1] 66
> sd(Time)
[1] 12.91243
> sd(Dairy)
Error in var(as.vector(x), na.rm = na.rm) :
        missing observations in cov/cor
In addition: Warning message: NAs introduced by coercion
```

The command **sd(Dairy)** yields an error because **Dairy** is a categorical variable.

Oftentimes, it is a good idea to store a result in an R-variable so that you can refer to it later. Then you can type the new variable name to see what is stored in it. For example,

```
> Time.mean = mean(Time)
> Time.mean
[1] 63.5
> Time.mean/10 +100
[1] 106.35
```

The last command shows that R-variables can be used with the mathematical operators +, -, * and /. To compute the mean and standard deviation of the cook times of cottage cheese with PC and without PC, execute

```
> tapply(Time,Treatment,mean)
withoutPC    withPC
 61.14286  65.85714
> tapply(Time,Treatment,sd)
withoutPC    withPC
 12.62839  13.74080
```

Does this suggest that adding the fungus PC increases the cook time of cottage cheese?

Exercises

Starting on page 56, do problems: 1.3, 1.5, 1.7