

Chapter 1 Notes

1 Data

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 (steps 2-5 below) of the Scientific Method 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. Test: Does 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: The objects from which data is collected. Individuals may be people, places, animals, things, *even* time periods.

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

Two Types of Variables:

- **Categorical** (or Qualitative) - The possible values are *categories*. 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:

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 G.E. light bulbs; 100 tosses with a weighted die.

Typically, it is unrealistic to obtain 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**).

2 Obtaining and Installing R

The following is a revised version of what appears in Chapter 7.1 of your Course Notes: Statistics for Researchers STAT401 FALL 2006:

1. Get on the Internet and go to the web address <http://cran.r-project.org>. This is the “official” site of the The Comprehensive R Archive Network (CRAN). Bookmark this address. Lots of information (manuals, answers to frequently asked questions, etc) can be downloaded from this site.
2. There is a box labelled “Precompiled Binary Distributions.” In that box, click on the link Windows (95 and later).
3. Click on the subdirectory link named base. The standard collection of R functions and packages is called Base R.
4. The link README.R-2.3.1 contains a brief synopsis on installation and other instructions for R version 2.3.1 for Windows. This file contains information on installing R. You shouldn’t need to look at this file, but take a look if you get into trouble.
5. Click on the link R-2.3.1-win32.exe. Download this setup program to your hard drive.
6. Exit from the Internet and open Windows Explorer. Go to the folder in which you saved R-2.3.1-win32.exe and run the program.
7. You will be guided through the installation by a Setup Wizard.

8. CRAN contains many excellent resources for using R. One resource in particular is of special note. At the CRAN home page click on the link Contributed. This link is on the left-hand-side of the page. Under the heading “Documents with more than 100 pages,” there are links that allow you to download or to use online a nifty text on using R for introductory statistics. The text is “Simple R,” by John Veranzi. To get the complete text, click on the “PDF” link. To use the interactive Simple R site, click on [Simple R Homepage](#).
9. Special-purpose software routines are bundled as separate “packages.” Some packages are automatically downloaded when base R is downloaded. To download additional packages, execute R on your PC and then click on “Packages” (top menu). Click on “Install package(s) from CRAN” and then choose the package(s) that you want to download. The packages that you will need for this course are the following: (a) MASS (this package does not need to be downloaded separately because it is part of base R), (b) lattice (this package does need to be downloaded separately), and (c) pastecs (this package does need to be downloaded separately).

3 Entering Data into R

A researcher is interested in determining whether adding a certain type of bacteria, called PC, helps increase the firmness of cottage cheese. Seven dairies make two identical batches of cottage cheese, one with and one without the bacteria PC. The results of the experiment are in a text file called “dairy.txt” which is shown below:

```
Farm Treatment Firmness
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 can not have spaces in them (e.g. don’t use “Cheese Firmness”). 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 dairy.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** which 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
  Farm Treatment Firmness
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$Farm
[1] A A B B C C D D E E F F G G
Levels: A B C D E F G
> D$Treatment
[1] withPC   withoutPC withPC   withoutPC withPC   withoutPC withPC
[8] withoutPC withPC   withoutPC withPC   withoutPC withPC   withoutPC
Levels: withoutPC withPC
> D$Firmness
[1] 68 61 75 69 62 64 86 76 52 52 46 38 72 68
```

Or you can execute

```
>attach(D)
> Farm
```

```

[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
> Firmness
[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,

```

> FIRMNESS
Error: object "FIRMNESS" not found
> D$FirmneSS
NULL

```

Notice that R recognizes that **Farm** and **Treatment** are categorical variables and gives the *levels* or categories associated with each. The variable **Firmness** 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(Firmness)
[1] 63.5
> median(Firmness)
[1] 66
> sd(Firmness)
[1] 12.91243
> sd(Farm)
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(Farm)** yields an error because **Farm** 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,

```

> firm.mean = mean(Firmness)
> firm.mean
[1] 63.5
> firm.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 firmness of cottage cheese without PC and of the firmness of cottage cheese with PC, execute

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

Does this suggest that adding PC might increase the firmness of cottage cheese?

4 Exercises

1.3 on p10: 1, 2, 3, 5, 7

1.4 on p18: 9, 11, 15, 19

5 Reading

All sections of Chapter 1