Intro to Social Science Data Analysis

Lecture 5: Descriptive Statistics

Christopher Gandrud

September 26, 2012

Review Questions 1

- What is reproducible research?
- Why is it important?

Review Questions 1

- What is reproducible research?
- Why is it important?

Review Questions 2

- What is the Markdown markup language?
- What does the *knitr* package do?

Review Questions 2

- What is the Markdown markup language?
- What does the *knitr* package do?

Review Questions 3

- What is a code chunk?
- How do you make a code chunk in a Markdown document?

Review Questions 3

- What is a code chunk?
- How do you make a code chunk in a Markdown document?

Assignment 2

Due: Friday 19 October

Describe at least **3** variables in a data set.

You need to select a **range of descriptive statistical tools**. The tools should include both **numerical descriptive statistics** and **graphics**.

These tools should describe the variables':

- central tendency,
- variation,
- their relationships with the other variables.

The descriptions need to be discussed **in paragraph form**.

The description must be **reproducible**. So you should email me the link to a Dropbox folder with:

- the `.csv` data set,
- the `.Rmd` R markdown file,
- the final `.html` file.

So far we have learned how to gather data and get it into R.

Today we will start to learn tools for **describing** our data.

We will learn **descriptive statistics**.

Why do we need tools for describing our data?

Why do we need tools for describing our data?

- To find the **patterns** we are interested but too difficult to find by just looking at the raw data.
- Find potential **data biases**.

Why do we need tools for describing our data?

- To find the **patterns** we are interested but too difficult to find by just looking at the raw data.
- Find potential **data biases**.

Always look at the descriptive statistics before starting your data analysis.

When describing data, **ALWAYS** look at **BOTH**

- The Central Tendency,
- The Variability (dispersion).

When describing data, **ALWAYS** look at **BOTH**

- The Central Tendency,
- The Variability (dispersion).

Central Tendency

The central value around which the data clusters.

Examples of descriptive statistics for the central tendency include: the mode, median, and mean (average).

Variability How the values vary around the central tendency.

Examples of descriptive statistics for the variability include: the range, interquartile range, standard deviation.

Describing Numerical Data

Data that is at the **highest measurment level** (numerical continuous) can be described using **all** of the descriptive statistics.

Populations, Samples, and Descriptive Statistics

Remember that our data is a **sample** of the **population**.

Today we are going to be describing **samples**.

From week 7 we will start to use statisics that help us **infer** things from our samples about the population.

The Data

Most of the examples for this section use World Bank data for 2009 on:

- GDP per capita (current US$)
- Mortality rate, infant (per 1,000 live births)
- World Bank region classification
- World Bank income level classification

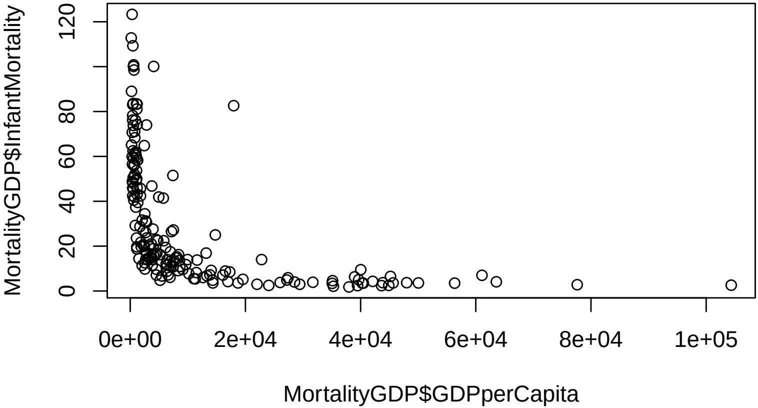
The sample includes 199 jurisdictions.

You can get the data set using the source code file at:
`http://bit.ly/OTWEGS`

You can actually run this source code directly from R using the `source_url` command in the *devtools* package.

```r
# Load package
library(devtools)

# Gather data using source code at:
# http://bit.ly/OTWEGS

# Data is stored in a data frame: MortalityGDP
source_url("http://bit.ly/OTWEGS")

# See contents of MortalityGDP
names(MortalityGDP)

## [1] "country"          "GDPperCapita"
## [3] "InfantMortality" "region"
## [5] "income"
```

```r
# Create scatterplot of GDP/Capita & Infant Mortality
plot(MortalityGDP$GDPperCapita,
     MortalityGDP$InfantMortality)
```



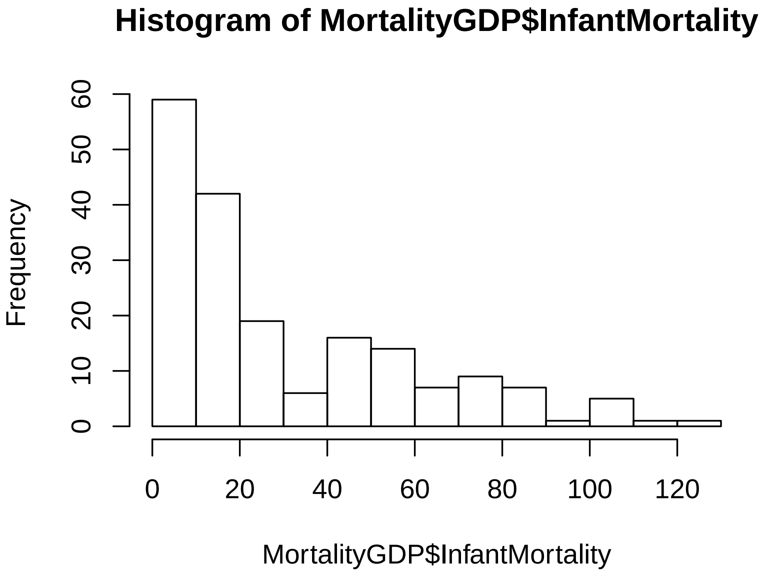
Mode

The most common value in a distribution.

One way to find the mode of a numeric continuous variable is with a **histogram**.

In R you can use the `hist` command.

```r
hist(MortalityGDP$InfantMortality)
```



Uni, Bi, and Multi Modal Distributions

A distribution can have **multiple modes**.

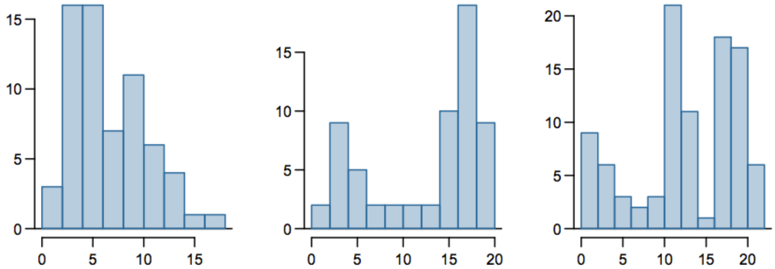


Figure 1.15: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

Diez (2011, 12)

Median

The middle value of a distribution.

You can find the median with the `median` command.

```r
# Create data with no missing values of infant mortality
InfantNoMiss <- subset(MortalityGDP,
                           !is.na(InfantMortality))

# Find the median infant morality rate
median(InfantNoMiss$InfantMortality)

## [1] 17.2
```

Mean (average)

The sum of all data values (x) divided by the number of data valuse (n).

Population Mean (μ_x)

$$\mu_x = \frac{\sum x}{n}$$

Sample Mean ($\bar{x}$)

$$\bar{x} = \frac{\sum x}{n}$$

You can find the mean with the `mean` command.

```
# Find the mean of InfantMortality
mean(InfantNoMiss$InfantMortality)

## [1] 30.23
```

What is the Central Tendency of Infant Mortality?

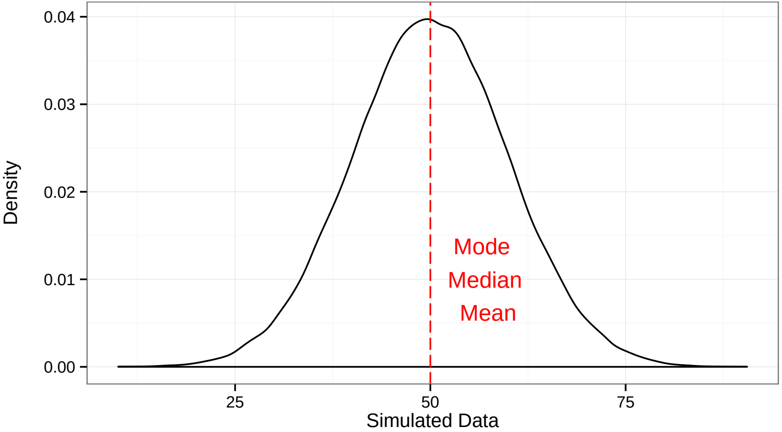
- **Mode:** 0-10
- **Median:** 17.2
- **Mean:** 30.2

The reason that these three measures of central tendency are **not the same** is that the distribution of Infant Mortality in the sample is **highly skewed**.

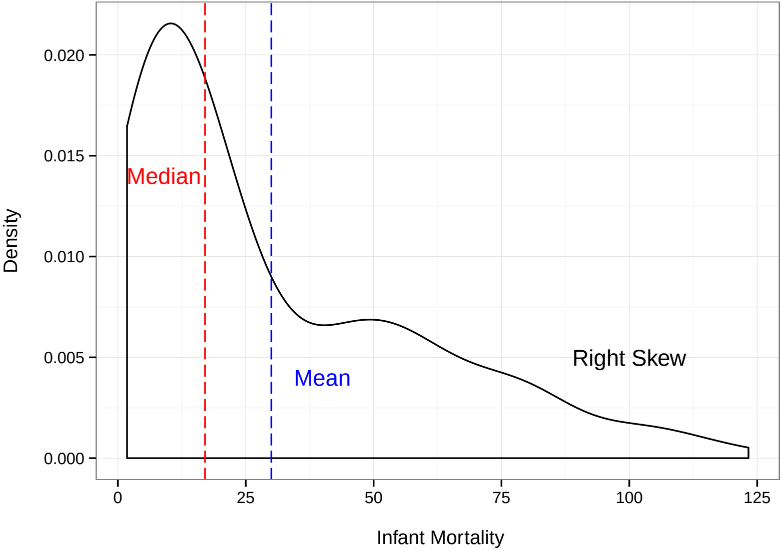
Data that is **normally distributed** has the same mode, median, and mean.

Normally distributed data also is **not skewed**. It has the same variance on the right and left of the central tendency.

```r
# Simulate normally distributed data
Normal <- rnorm(1e+05, mean = 50, sd = 10)
```



The Infant Mortality data is very right skewed.



A distribution can be:

- Right skewed (positively skewed)
  - Right skewed data pulls the mean up.
- Left skewed (negatively skewed)
  - Left skewed data pulls the mean down.

A distribution can be:

- Right skewed (positively skewed)
  - Right skewed data pulls the mean up.
- Left skewed (negatively skewed)
  - Left skewed data pulls the mean down.

A distribution can be:

- Right skewed (positively skewed)
  - Right skewed data pulls the mean up.
- Left skewed (negatively skewed)
  - Left skewed data pulls the mean down.

A distribution can be:

- Right skewed (positively skewed)
  - Right skewed data pulls the mean up.
- Left skewed (negatively skewed)
  - Left skewed data pulls the mean down.

So, the central tendency does not adequately describe distributions by itself.

We also need descriptive statistics of the **variability**

Range

The range is the simplest way to describe variability.

It is the lowest and highest value.

We can find the range with the `range` command.

```
range(InfantNoMiss$InfantMortality)

## [1]   1.8 123.3
```

Problems with the Range

The range is highly influenced by **outliers**–extreme values.

It also **ignores** all of the data between the minimum and maximum values.

```r
#### Find infant mortality outliers
# Reorder data based on infant mortality
OrderMort <- InfantNoMiss[
                      order(InfantNoMiss$InfantMortality,
                           decreasing = TRUE), ]
# Keep country & InfantMortality
OrderMort <- OrderMort[, c("country",
                          "InfantMortality")]

# Show high values
head(OrderMort)

##                      country InfantMortality
## 187             Sierra Leone           123.3
## 38          Congo, Dem. Rep.           112.8
## 39  Central African Republic           109.3
## 190                  Somalia           108.3
## 134                     Mali           100.8
## 12                    Angola           100.1
```

Interquartile Range

One way to deal with outliers is to look at the interquartile range.

The interquartile range is the difference between the upper and lower quartiles.

A quartile is 25% of the data.

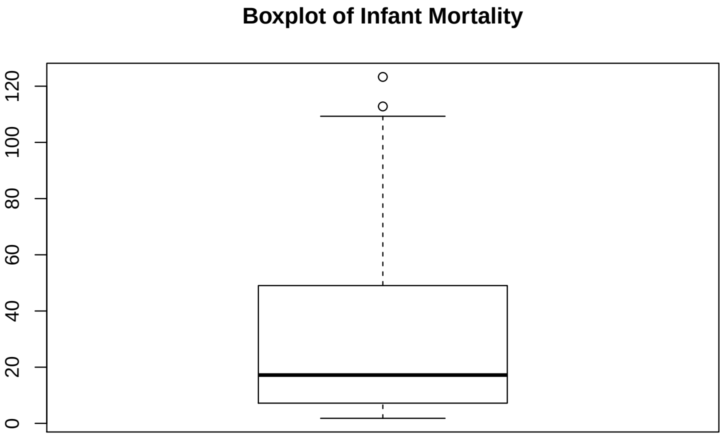
The **lower quartile** is the point up to the lower 25% of the data.
The **upper quartile** is the point up to the upper 75% of the data.

```r
# Find what the quartile points are
summary(InfantNoMiss$InfantMortality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.8     7.2    17.2    30.2    49.0   123.0
```

$$48.92 - 7.175 =$$

```r
# Find the interquartile range of InfantMortality
IQR(InfantNoMiss$InfantMortality)

## [1] 41.85
```

```r
# Boxplot showing interquartile range
boxplot(InfantNoMiss$InfantMortality,
        main = "Boxplot of Infant Mortality")
```



See Diaz (2011, 16) for more boxplot details.

A bigger range means **more variablity**.

Note: big in terms of the variable's scale.

Standard Deviation.

The interquartile range describes variation in terms of the median.

The standard deviation describes **variation in terms of the mean**.

The standard deviation is made of the following parts.

Deviation: the distance of an observation x from the mean $\bar{x}$.

$$\text{Deviation} = x - \bar{x}$$

Sum of Squares: the sum of the squared deviations (they have to be squared or the sum will $= 0$)

$$\text{Sum of Squares} = \sum (x - \bar{x})$$

Degrees of Freedom: Sample size n minus the number of parameters. Today the number of parameters $= 1$. (See Crawley 2005, 36-37 for a good explanation.)

$$\text{df} = n - 1$$

The standard deviation is made of the following parts.

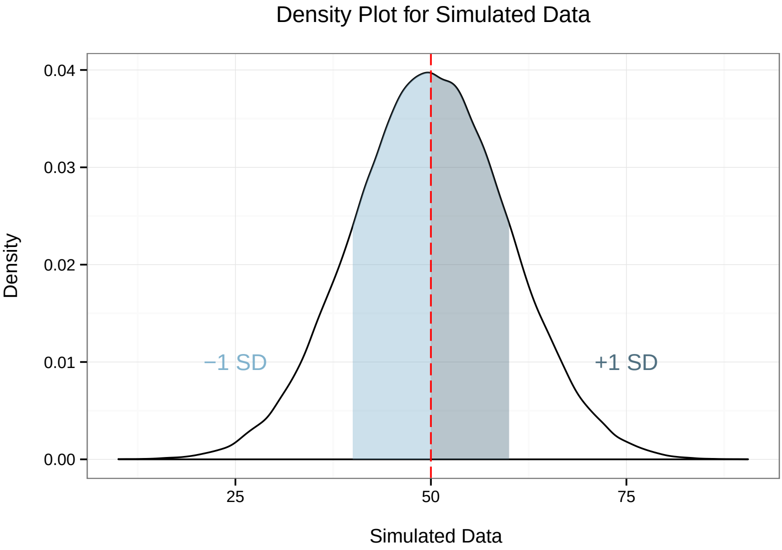
Variance (s^2): roughly the average deviation.

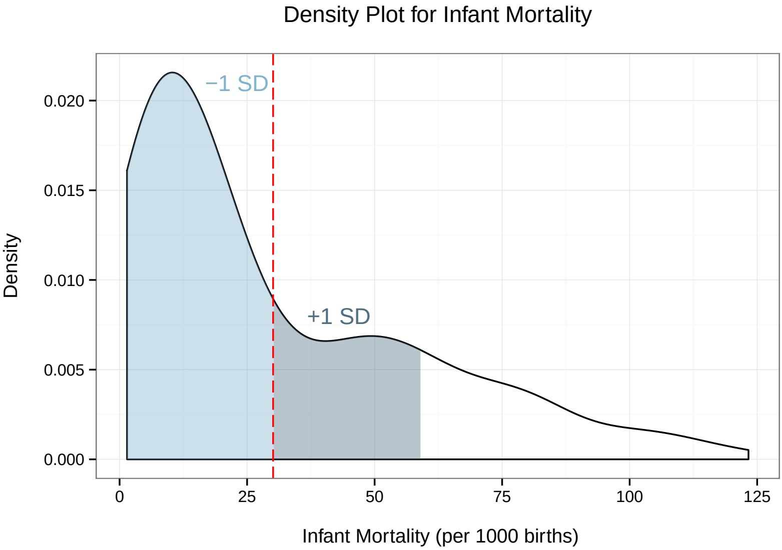
$$s^2 = \frac{\text{Sum of Squares}}{\text{Degrees of Freedom}} = \frac{\sum(x - \bar{x})}{n - 1}$$

Standard Deviation (s): square root of the variance

$$s = \sqrt{s^2}$$

```r
# Find the varience of InfantMortality
var(InfantNoMiss$InfantMortality)

## [1] 827.3


# Find the standard deviation of InfantMortality
sd(InfantNoMiss$InfantMortality)

## [1] 28.76
```


Density Plot for Simulated Data
0.04
0.03
0.02
0.01
0.00
Density
−1 SD
+1 SD
25
50
75
Simulated Data


Density Plot for Infant Mortality
−1 SD
+1 SD
0.020
0.015
0.010
0.005
0.000
Density
0
25
50
75
100
125
Infant Mortality (per 1000 births)

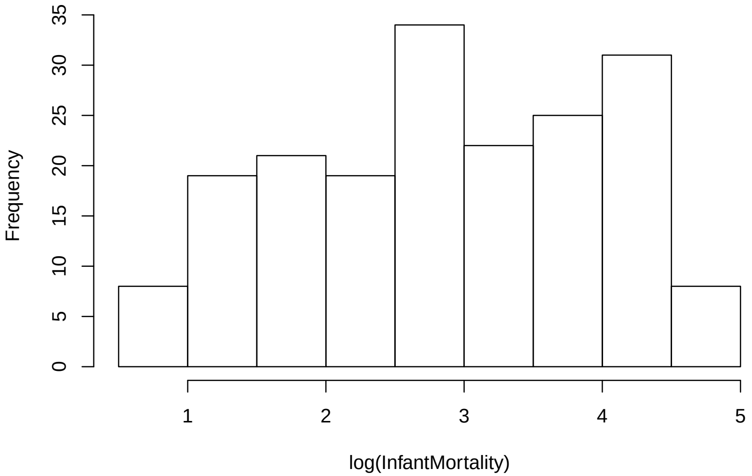
Transforming Data

Transforming data can make **highly skewed** data easier to work with.

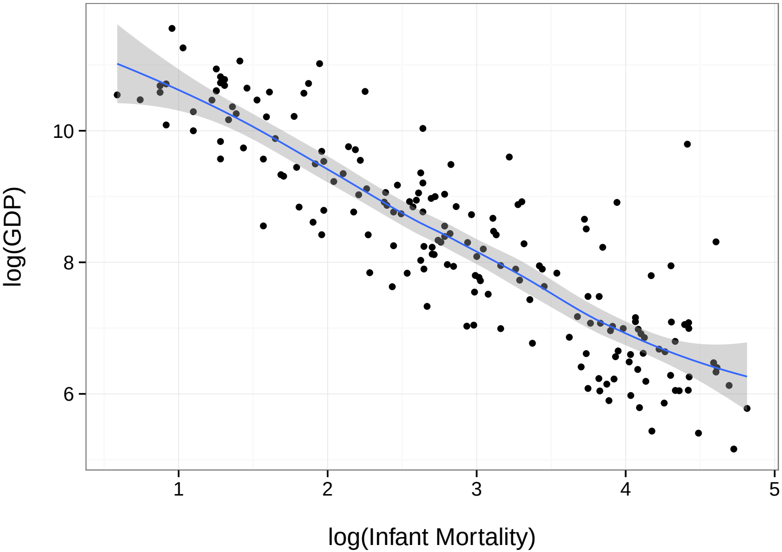
Transforming data just means to **rescale** the data using some function.

For example, we can **log-transform** our Infant Mortality data to see the relationship between the two variables better.

```r
# Log Transform InfantMortality
InfantNoMiss$logInf <- log(
                        InfantNoMiss$InfantMortality)
```



Preview!



Describing Categorical Data

Question

What descriptive statistics can you use for:

- Ordinal data
- Categorical data

Question

What descriptive statistics can you use for:

- Ordinal data
- Categorical data

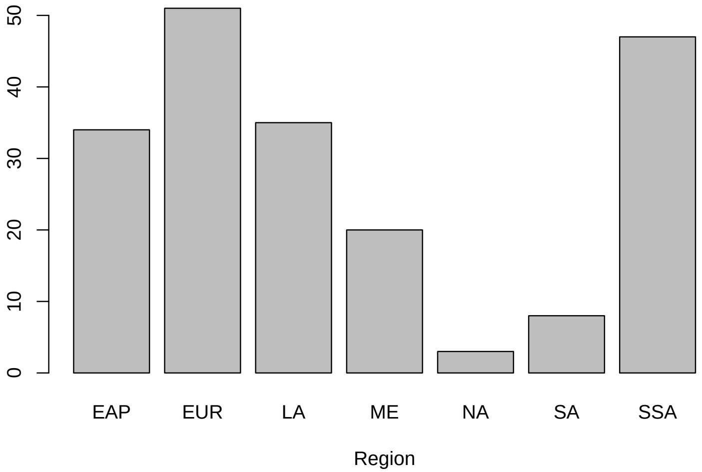
You can use

- Ordinal data; mode, median, range, interquartile range
- Categorical data: mode, frequency tables/barplots

You can use

- Ordinal data; mode, median, range, interquartile range
- Categorical data: mode, frequency tables/barplots

```r
# Use cars data, loaded in R by default
# Create bar plot
plot(MortalityGDP$region, xlab = "Region")
```

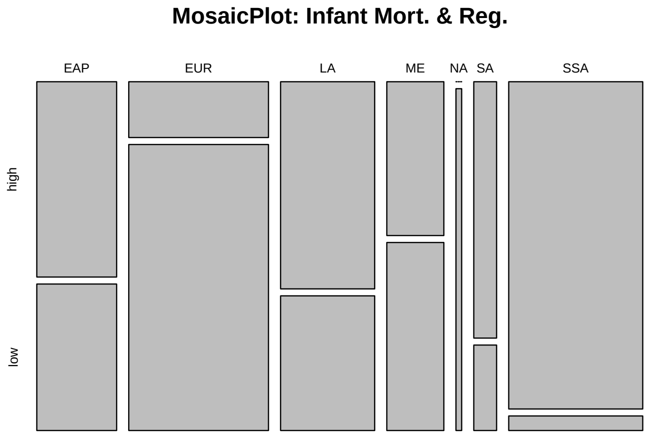


Scatterplot-like Options for Categorical Data

You can use **contingency tables** and **mosaic plots** like scatter plots when you have categorical data.

```
# Create High/Low Income Variable
InfantNoMiss$DumMort[InfantNoMiss$InfantMortality
                     >= 15] <- "high"
InfantNoMiss$DumMort[InfantNoMiss$InfantMortality
                     < 15] <- "low"
# Create contingency table
table(InfantNoMiss$region, InfantNoMiss$DumMort)

##
##        high low
##   EAP    16  12
##   EUR     8  41
##   LA     20  13
##   ME      9  11
##   NA      0   2
##   SA      6   2
##   SSA    45   2
```

```r
mosaicplot(table(InfantNoMiss$region,
            InfantNoMiss$DumMort),
            main = "MosaicPlot: Infant Mort. & Reg.")
```



References I

Crawley, Michael J. 2005. Statistics: An Introduction Using R. Chichester: John Wiley Sons. Ltd.

Diaz, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. OpenIntro Statistics. 1st ed. http://www.openintro.org/stat/downloads.php.