

Exploratory Data Analysis

- Exploratory Data Analysis involves both [graphical displays of data](#) and [numerical summaries of data](#).
- A common situation is for a data set to be represented as a [matrix](#).
- There is a [row for each unit](#) and a [column for each variable](#).
- A [unit](#) is an object that can be measured, such as a person, or a thing.
- A [variable](#) is a characteristic of a unit that can be assigned a number or a category.
- In the data set I collected from the survey of students in the class, each student who responded to the survey is a unit. Variables include [sex](#), [major](#), [year in school](#), [miles from home](#), [height](#), and [blood type](#).

Variables

- We can also make a categorical variable from a continuous variable by dividing the range of the variable into classes (So, for example, height could be categorized as [short](#), [average](#), or [tall](#)).
- Identifying the types of variables can be important because some methods of statistical analysis are appropriate only for a specific type of variable.

Exploratory Data Analysis

Bret Larget

Department of Statistics

University of Wisconsin - Madison

September 8, 2003

Variables

- Variables are either [quantitative](#) or [categorical](#).
- In a [categorical variable](#), the measurement for each unit is a category.
For example, for the data I collected in the course survey, the variables [blood type](#) and [sex](#) are categorical.
- The variable [year in school](#) is an example of an [ordinal](#) categorical variable, because there is a meaningful ordering of the levels of the variable.
- [Quantitative variables](#) record a number for each unit.
- The variable height is an example of a [continuous variable](#) because it could take on a continuum of possible values.
- The variable number of sisters is [discrete](#) because we can list the possible values.
- Often, continuous variables are rounded to a discrete set of values. Most people round their height to the nearest inch (or half inch).

Summaries of Categorical Variables

- A **frequency distribution** is a list of the observed categories and a count of the number of observations in each.
- A frequency distribution may be displayed with a **table** or with a **bar chart**.
- For **ordinal** categorical random variables, it is conventional to order the categories in the display (table or bar chart) in the meaningful order.
- For non-ordinal variables, two conventional choices are alphabetical and by size of the counts.
- The vertical axis of a bar chart may show **frequency** or **relative frequency**.
- It is conventional to leave space between bars of a bar chart of a categorical variable.

Summary of Majors

```
> cbind(summary(Major))
           [,1]
Animal Science      2
Bacteriology        1
Biochemistry        3
Biological Aspects of Conservation 2
Biology             19
Biology/Zoology     1
Biomedical Engineering 9
Botany              1
Dairy Science       1
Forestry            1
Genetics            31
Graduate student in Bacteriology 1
Kinesiology         4
Medical MicroBiology and Immunology 1
Molecular Biology  1
Nursing             1
Plant Pathology     1
Possibly Kinesiology 1
Undecided           1
Wildlife Ecology    5
Wildlife Ecology - Natural Resources 1
Zoology            11
```

Samples

- A **sample** is a collection of units on which we have measured one or more variables.
- The number of observations in a sample is called the **sample size**.
- Common notation for the sample size is n .
- The textbook adopts the convention of using **uppercase letters** for variables and **lower case letters** for observed values.

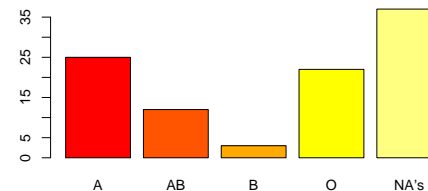
Summary of Blood Type Data

Here is a frequency table.

```
> summary(BloodType)
 A  AB  B  O NA's
25  12  3  22  37
```

Here is a bar chart.

```
> barplot(summary(BloodType))
```

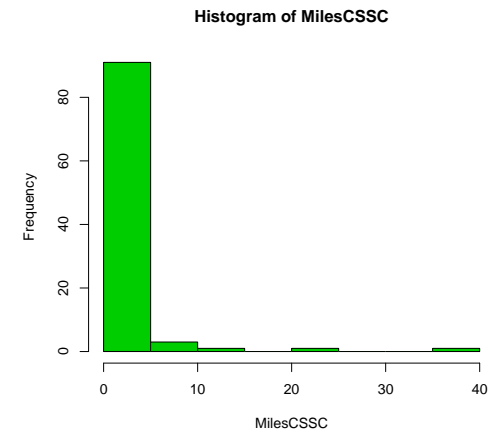


Summaries of Quantitative Variables

- Quantitative variables from very small samples can be displayed with a **dotplot**.
- **Histograms** are a more general tool for displaying the distribution of quantitative variables.
- A histogram is a bar graph of counts of observations in each **class**, but no space is drawn between classes.
- If classes are of different widths, the bars should be drawn so that **areas** are proportional to frequencies.
- Selection of classes is arbitrary. Different choices can lead to different pictures.
- Too few classes is an over-summary of the data.
- Too many classes can cloud important features of the data with noise.

Summary of Miles from CSSC

```
> hist(MilesCSSC, col = 3)
```



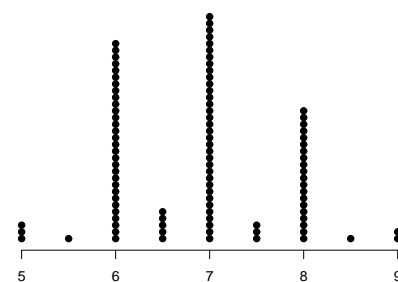
Summary of Second Majors

```
> cbind(summary(Major2))
```

	[,1]
Anthropology	1
Bacteriology	4
Biochemistry	2
Biological Aspects of Conservation	2
Biology	2
Dietetics	1
Economics	1
Entomology	1
French	1
Genetics	1
Horticulture	1
IES	2
Life Science Communication	1
Math	2
Physics	1
Plant Pathology	1
Political Science	1
Psychology	1
Scandinavian Studies	1
Spanish	1
Wildlife Ecology	2

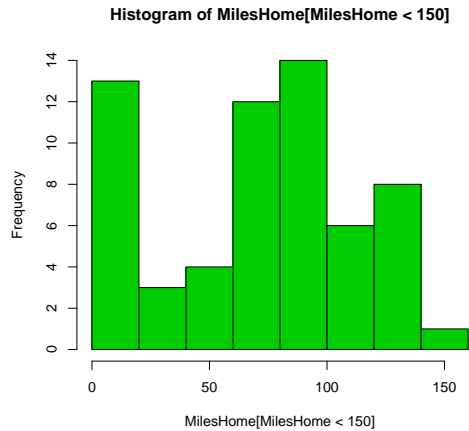
A Dotplot of Hours of Sleep

```
> source("../R/dotplot.R")  
> dotplot(Sleep)
```



Summary of Miles from Home for Students within 150 miles

```
> hist(MilesHome[MilesHome < 150], col = 3)
```



Statistics 371, Fall 2003

12

Stem-and-Leaf Diagrams

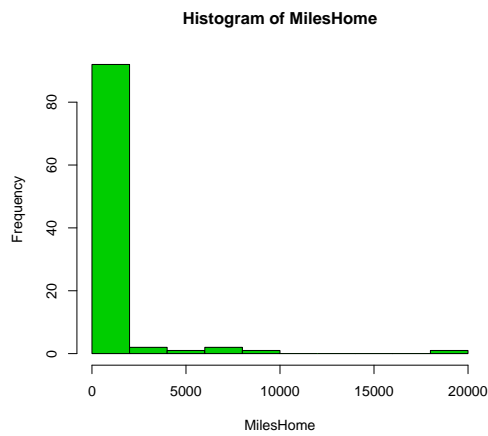
- **Stem-and-Leaf diagrams** are useful for showing the shape of the distribution of small data sets without losing any (or much) information.
- Begin by rounding all data to the same precision.
- The last digit is the **leaf**.
- Anything before the last digit is the **stem**.
- In a stem-and-leaf diagram, each observation is represented by a single digit to the right of a line.
- Stems are shown only once.
- Show stems to fill gaps!
- Combining or splitting stems can lead to a better picture of the distribution.

Statistics 371, Fall 2003

14

Summary of Miles from Home

```
> hist(MilesHome, col = 3)
```

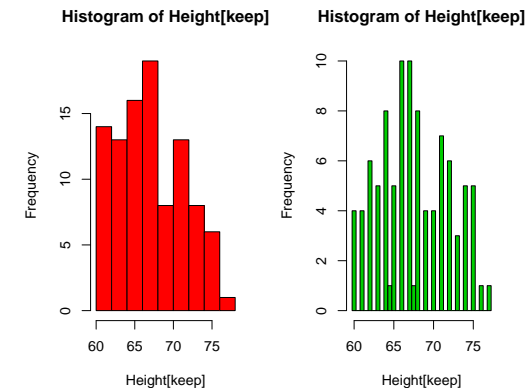


Statistics 371, Fall 2003

11

Summary of Height

```
> par(mfrow = c(1, 2))  
> hist(Height[keep], col = 2)  
> hist(Height[keep], breaks = breaks, col = 3)
```



Statistics 371, Fall 2003

13

Quantiles

- **Quantiles** are a generic name for positions in the distribution of a quantitative variable.
- For example, **percentiles** divide a distribution into 100 equal sized parts.
- **Quartiles**, which divide a distribution into four equal sized parts are a common statistical tool.
- The first quartile, Q_1 , is the location that separates the smallest quarter of the data from the rest. This is also known as the **25th percentile**.
- The third quartile, Q_3 , is the location that separates the top quarter of the data and is also known as the 75th percentile.
- The **median** is the second quartile.
- Different authors and statistical software packages have different definitions of quantiles.
- The definition I prefer is that **a value x is a p quantile of a sample if the proportion of observations less than or equal to x is at least p and if the proportion of observations greater than or equal to x is at least $1 - p$.**

Boxplots

- In a simple boxplot, a box extending from the first to third quartiles represents the middle half of the data. The box is divided at the median, and whiskers extend from each end to the maximum and minimum.
- It is common to draw more sophisticated boxplots in which the whiskers extend to the most extreme observations within **upper and lower fences** and individual observations outside these fences are labeled with individual points as potential **outliers**.
- The most common rule defining the fences are that they are 1.5 IQR below the first quartile and 1.5 IQR above the third quartile.

Comparing the mean and the median

- The mean and median of a symmetric distribution are the same.
- The median is **more resistant to outliers** than the mean. For example, the mean and median of the numbers 1, 2, 3 are 2, but for the data set 1, 2, 30, the median is still 2, but the mean is 11, far away from each observation.
- The median can be a better measure of a 'typical value' than the mean **especially for strongly skewed variables**.
- If a variable is **skewed to the right**, the mean will typically be larger than the median.
- The opposite is true if the variable is **skewed to the left**.

Example:

```
> mean(MilesHome)
[1] 760.2917
> median(MilesHome)
[1] 117.55
```

5 Number Summary

- The **minimum**, **first quartile**, **median**, **third quartile**, and **maximum** are called the **five-number summary** of a quantitative variable.
- The **interquartile range (IQR)** is the difference between the third and first quartiles.

$$\text{IQR} = Q_3 - Q_1$$

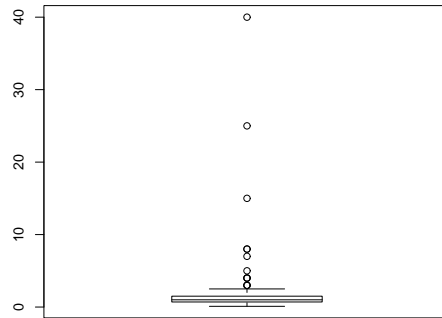
- The **range** is the difference between the maximum and the minimum.
- Graphical displays of five-number summaries are called **boxplots**.

Example:

```
> fivenum(MilesCSSC)
[1] 0.1 0.7 1.0 1.5 40.0
```

Boxplot of Miles from CSSC

```
> boxplot(MilesCSSC)
```



Measures of Dispersion

- The **standard deviation** or **SD** is the most common statistical measure of dispersion.
- A **deviation from the mean** is the signed distance of an observation from the mean.

$$\text{deviation} = \text{value of observation} - \text{mean}$$

Observations greater than the mean have positive deviations while those less than the mean have negative deviations.

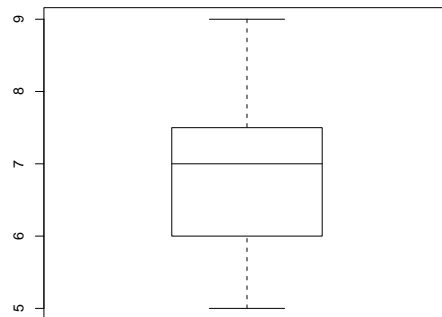
- The standard deviation is a special type of average deviation

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

This is **almost the square root of the mean squared deviation from the mean**.

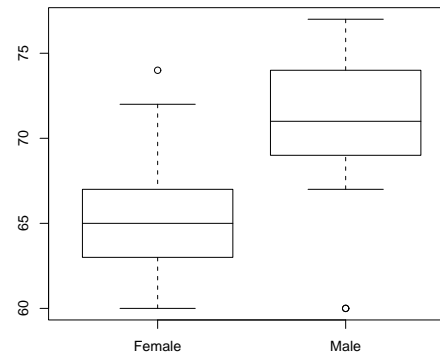
Boxplot of Hours of Sleep

```
> boxplot(Sleep)
```



Side-by-side boxplot of height versus sex

```
> boxplot(split(Height, Sex))
```



The Empirical Rule

For many variables (especially those that are nearly symmetric and bell-shaped), the following empirical rule is often a very good approximation.

- About 68% of the observations are within 1 SD of the mean.
- About 95% of the observations are within 2 SDs of the mean.
- Nearly all observations are within 3 SDs of the mean.

Example:

```
> m = mean(Sleep)
> s = sd(Sleep)
> m
[1] 6.868687
> s
[1] 0.8736093
> sum(abs(Sleep - m) < s)/length(Sleep)
[1] 0.7272727
> sum(abs(Sleep - m) < 2 * s)/length(Sleep)
[1] 0.949495
> sum(abs(Sleep - m) < 3 * s)/length(Sleep)
[1] 1
```

Statistics 371, Fall 2003

27

Measures of Dispersion

- Statisticians use $n - 1$ instead of n in the denominator for a technical mathematical reason of historical, if not practical, importance.
- The standard deviation can often be interpreted as the size of a [typical deviation from the mean](#).

Statistics 371, Fall 2003

26

Samples and Populations

- The previous techniques are useful for describing a data set, or a [sample](#) of data.
- It is often of interest to generalize findings from a sample to a larger group that statisticians call a [population](#).
- This generalization is called [statistical inference](#).
- Statistical inference is often concerned with using [statistics](#), characteristics that can be calculated from sample data, to estimate [parameters](#), characteristics of populations.
- Examples:
 - p = population proportion, \hat{p} = sample proportion
 - μ = population mean, \bar{y} = sample mean
 - σ = population standard deviation, s = sample standard deviation

Statistics 371, Fall 2003

28