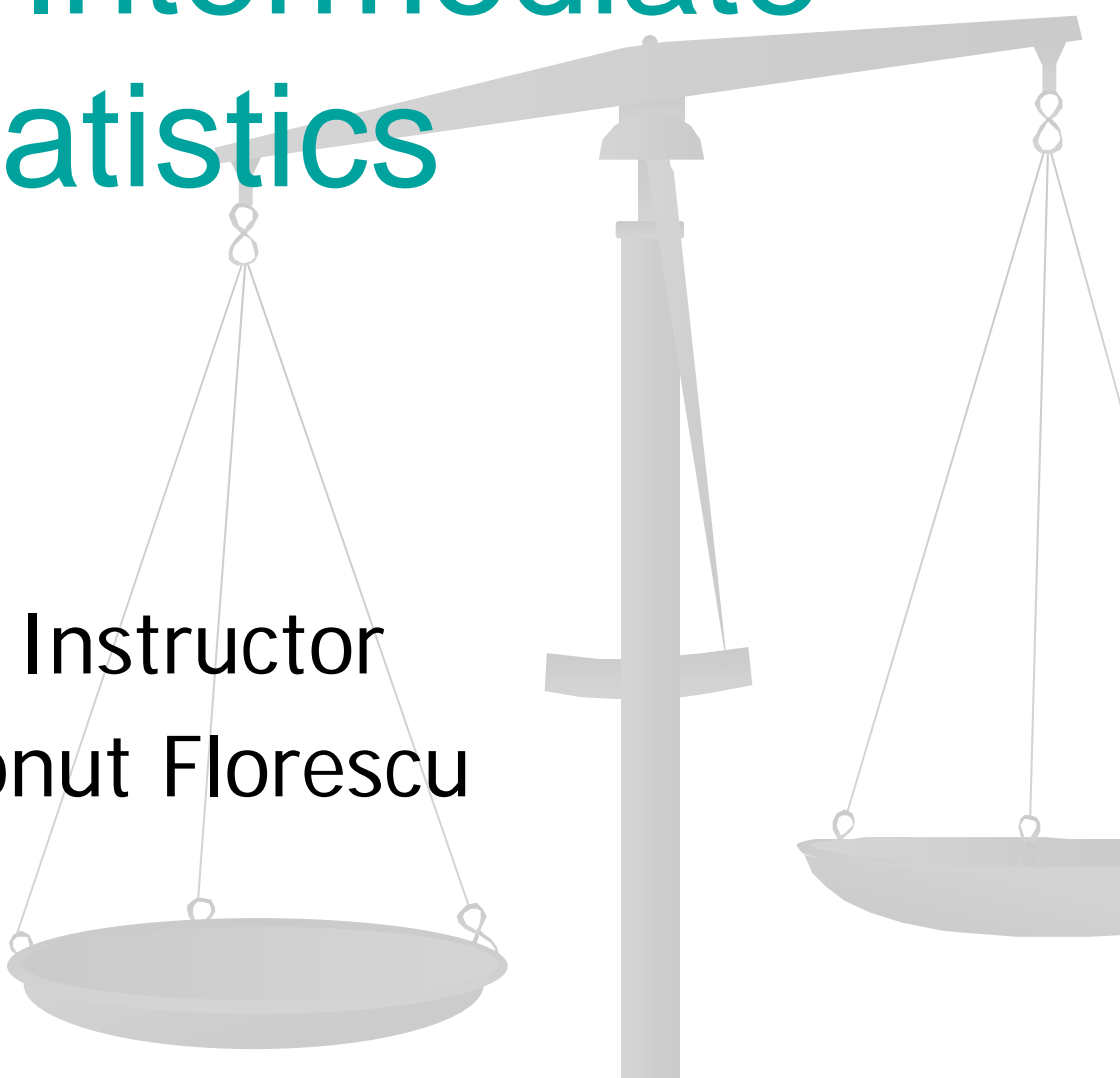
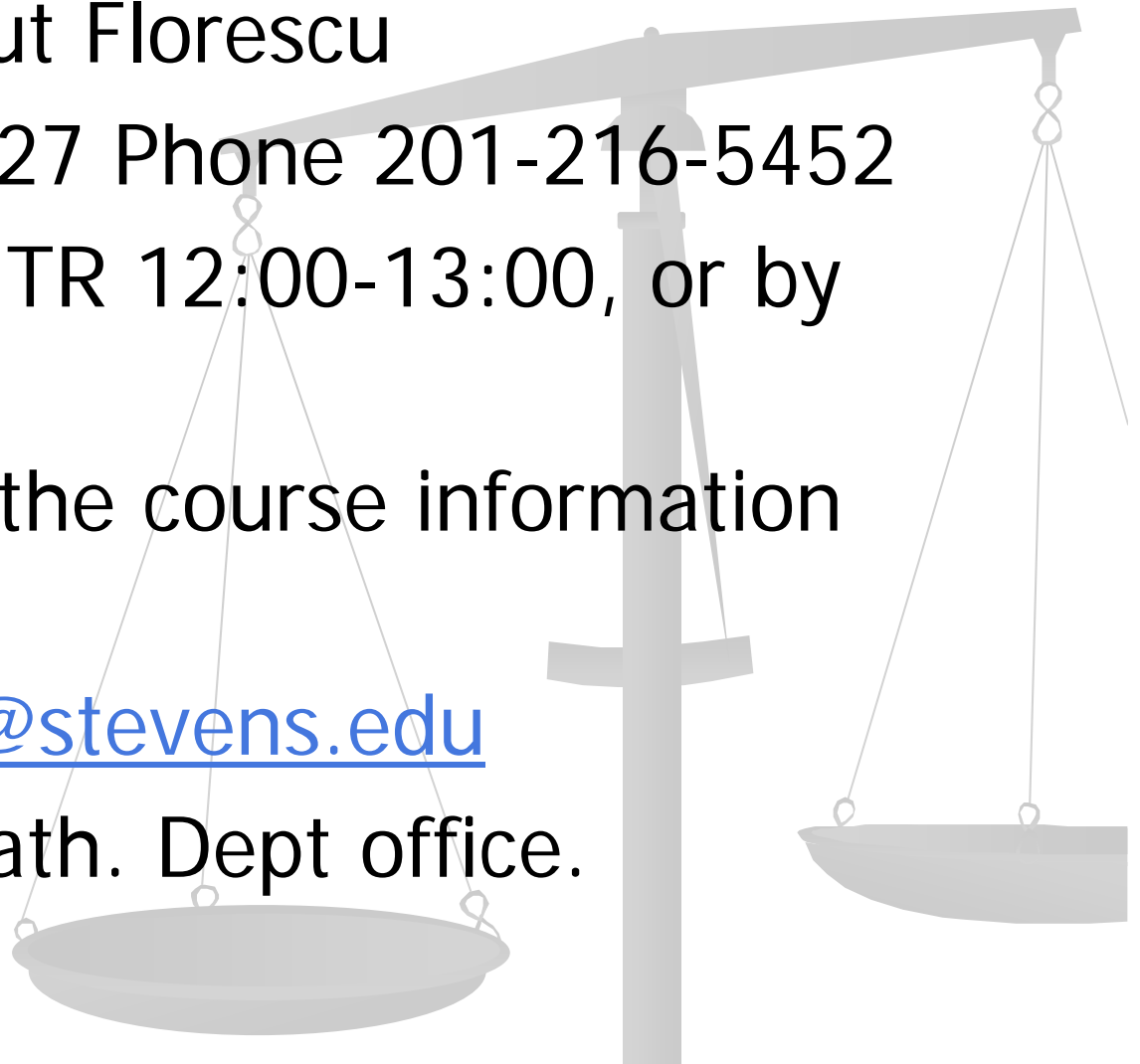


MA 331 Intermediate Statistics



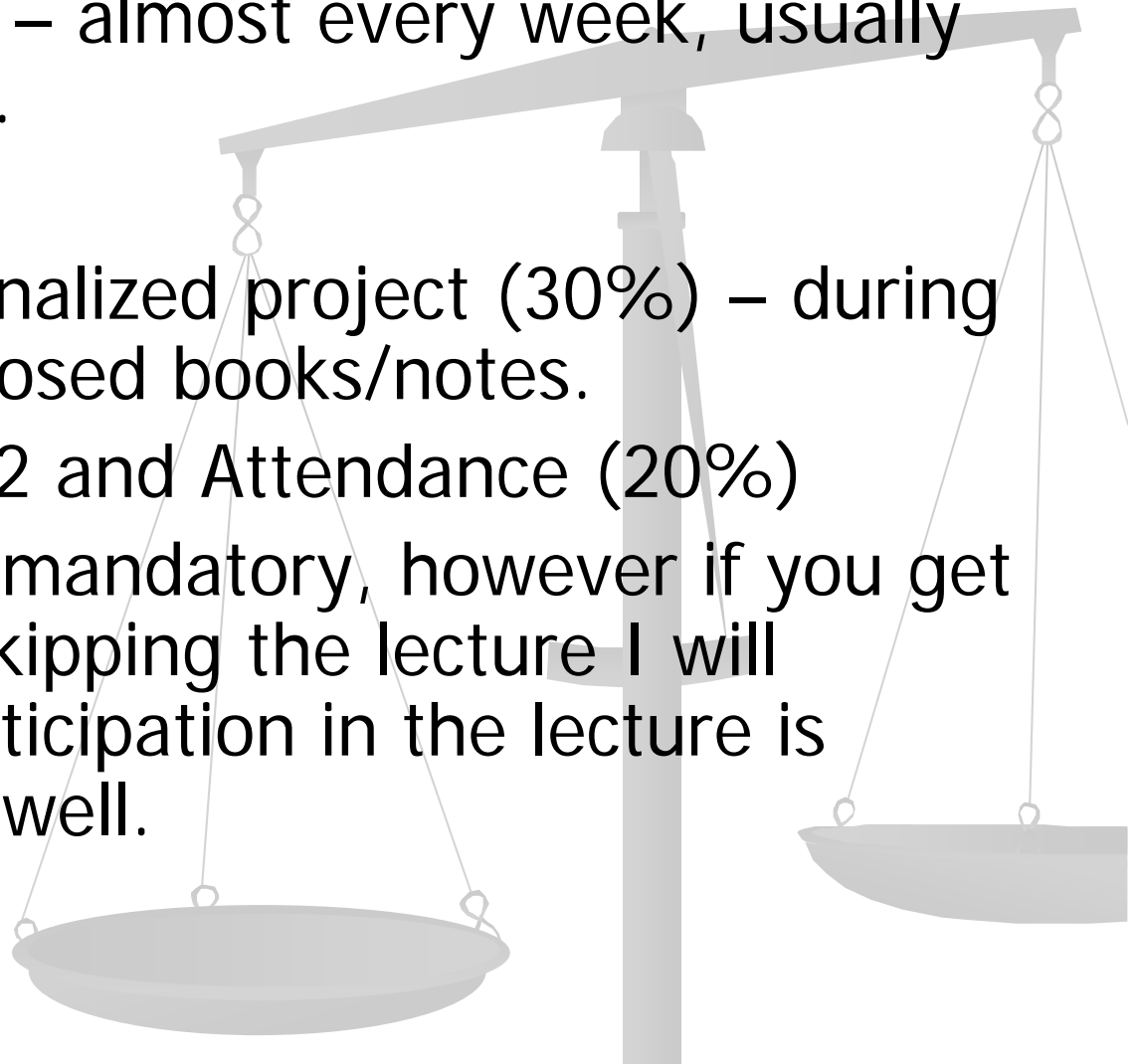
Instructor
Ionut Florescu

- Instructor : Ionut Florescu
- **Office:** Kidde 227 Phone 201-216-5452
- **Office hours:** TR 12:00-13:00, or by appointment.
- Please print off the course information posted on web.
- **Email:** ifloresc@stevens.edu
- **Mailbox:** in Math. Dept office.



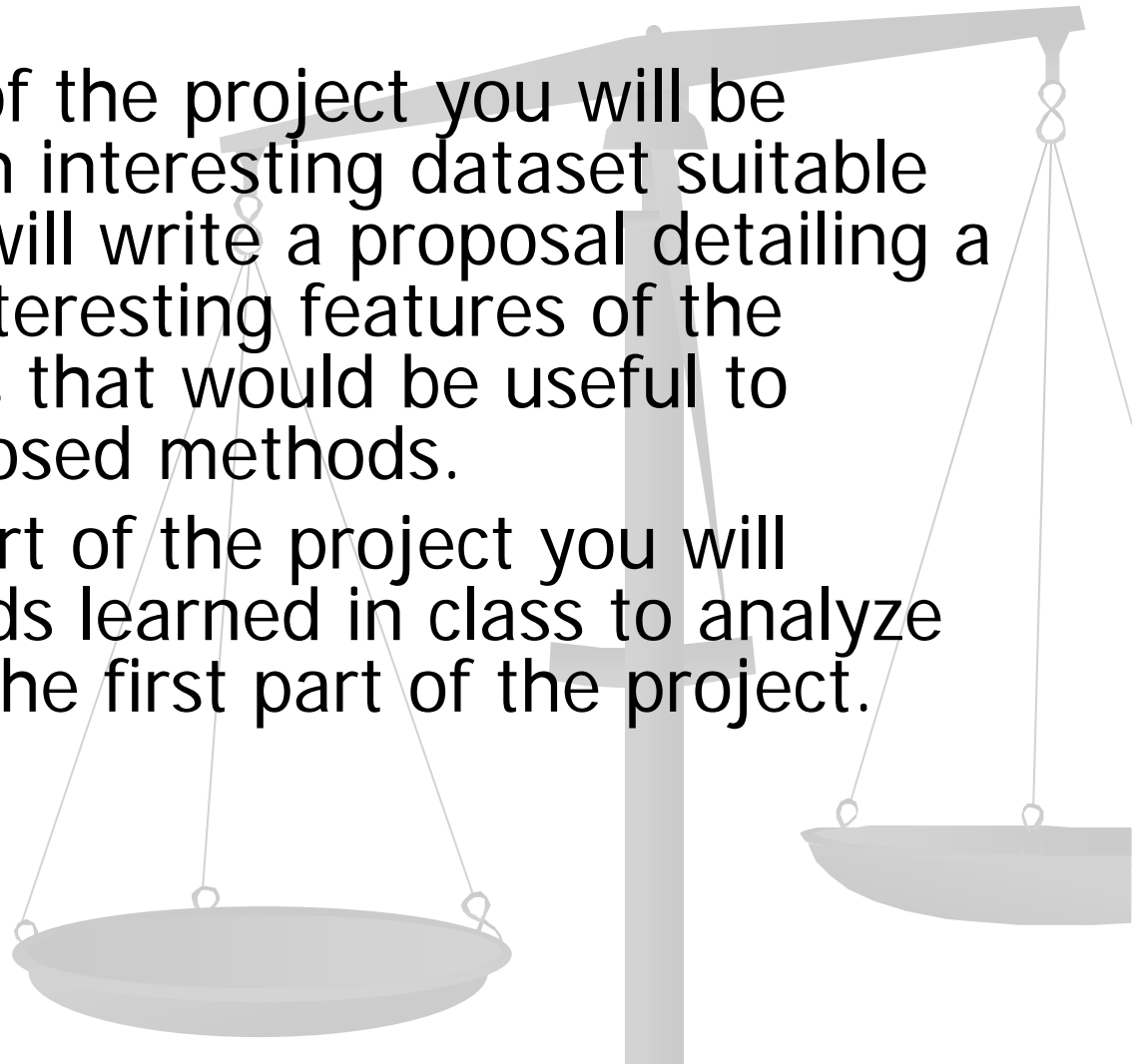
Grades

- Homework (30%) – almost every week, usually due on Thursdays.
- Midterm (20%)
- Final paper and Finalized project (30%) – during the finals week, closed books/notes.
- Project 1, Project 2 and Attendance (20%)
- Attendance is not mandatory, however if you get into the habit of skipping the lecture I will deduct points. Participation in the lecture is rewarded here as well.



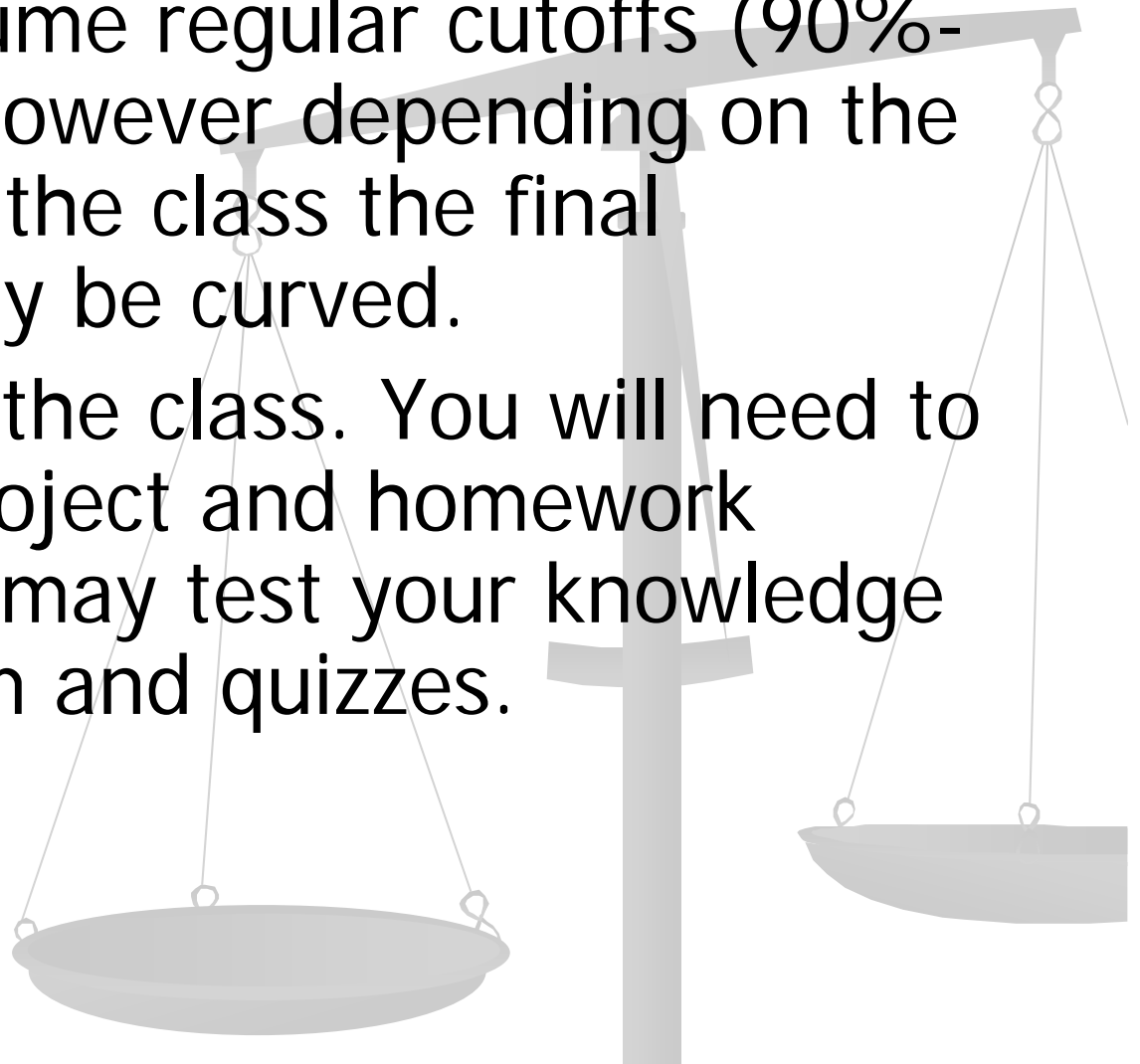
Grades (cont.)

- For the first part of the project you will be required to find an interesting dataset suitable for analysis. You will write a proposal detailing a description and interesting features of the dataset, questions that would be useful to answer, and proposed methods.
- For the second part of the project you will implement methods learned in class to analyze the dataset from the first part of the project.



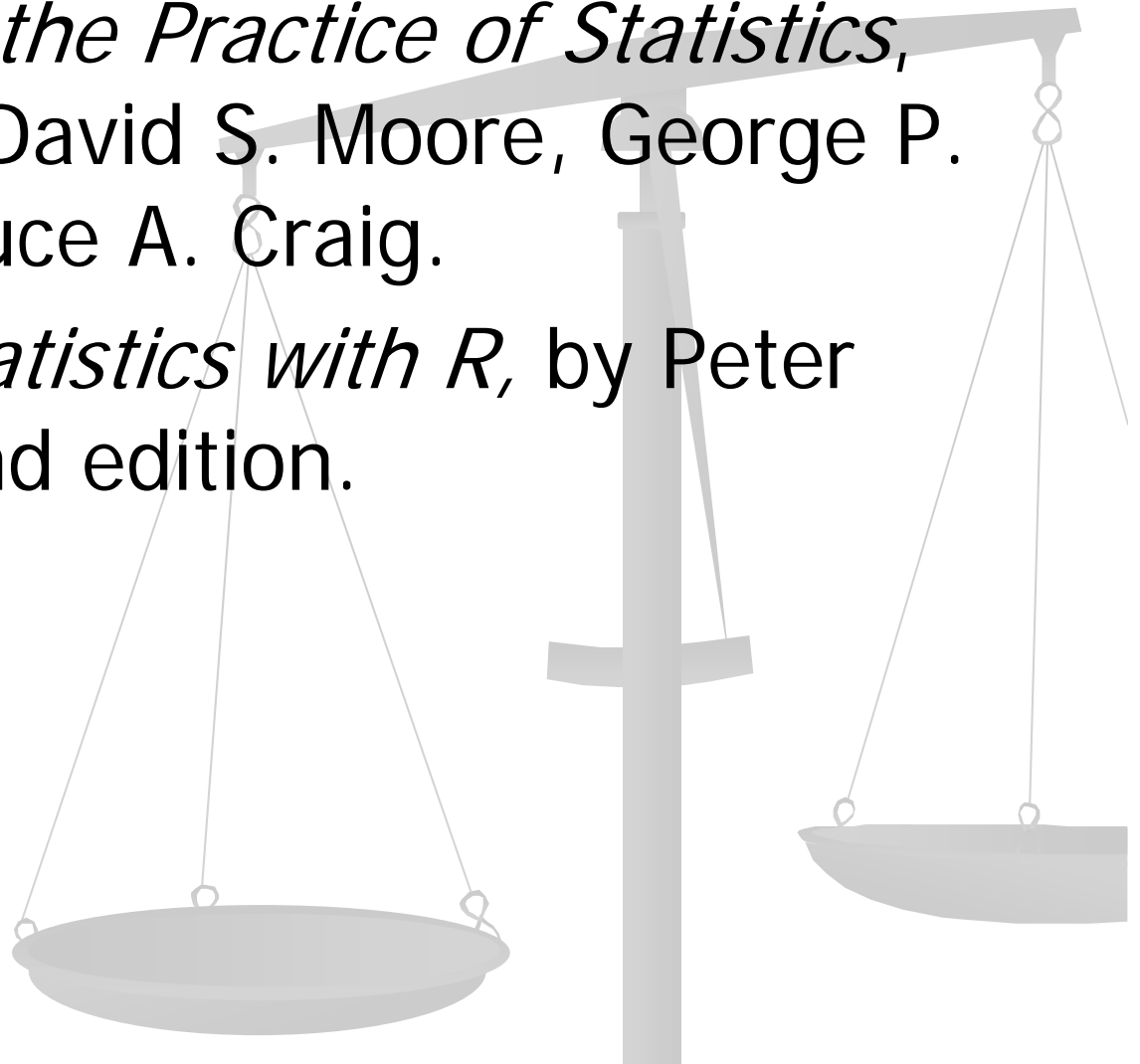
Grades (cont.)

- You should assume regular cutoffs (90%-100% A etc.), however depending on the performance of the class the final percentages may be curved.
- R is needed for the class. You will need to use it for the project and homework problems and I may test your knowledge of R in the exam and quizzes.



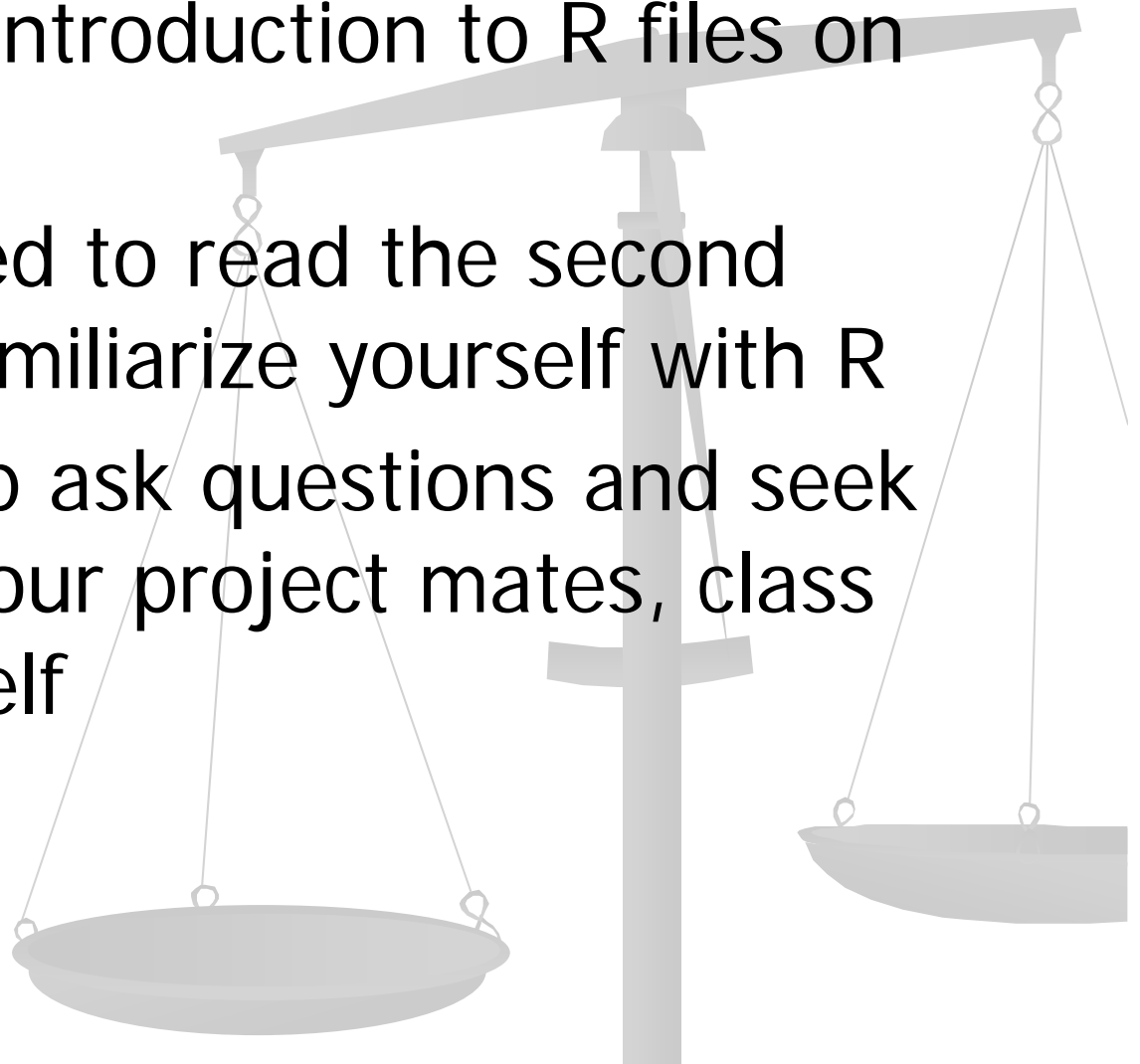
Textbooks

- *Introduction to the Practice of Statistics, 6th edition*, by David S. Moore, George P. McCabe and Bruce A. Craig.
- *Introductory Statistics with R*, by Peter Dalgaard, second edition.



R

- Please see the Introduction to R files on the website.
- You are expected to read the second textbook and familiarize yourself with R
- If you need help ask questions and seek answers from your project mates, class mates and myself



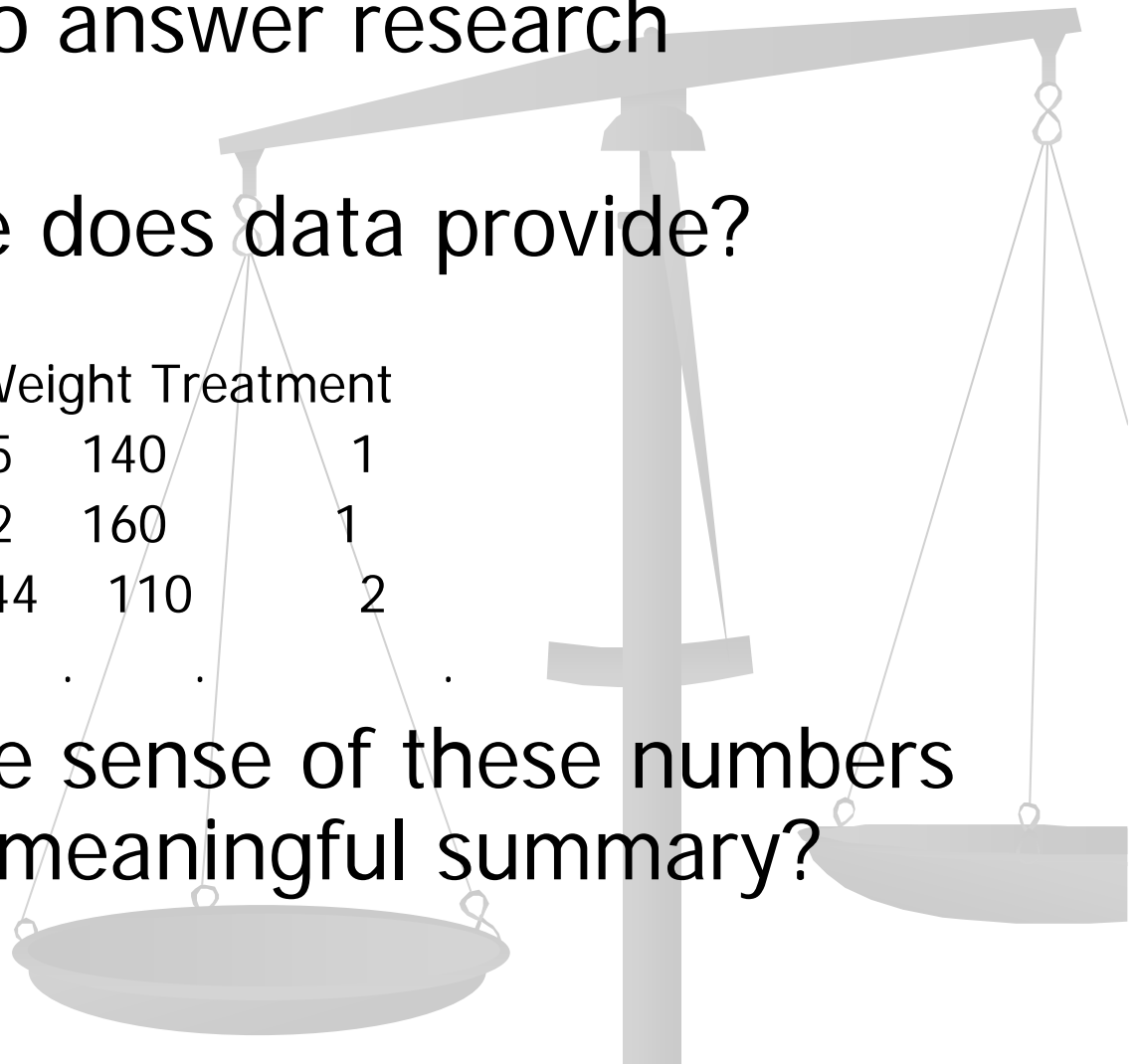
Data, Data, Data, all around us !

- We use data to answer research questions
- What evidence does data provide?

Example 1:

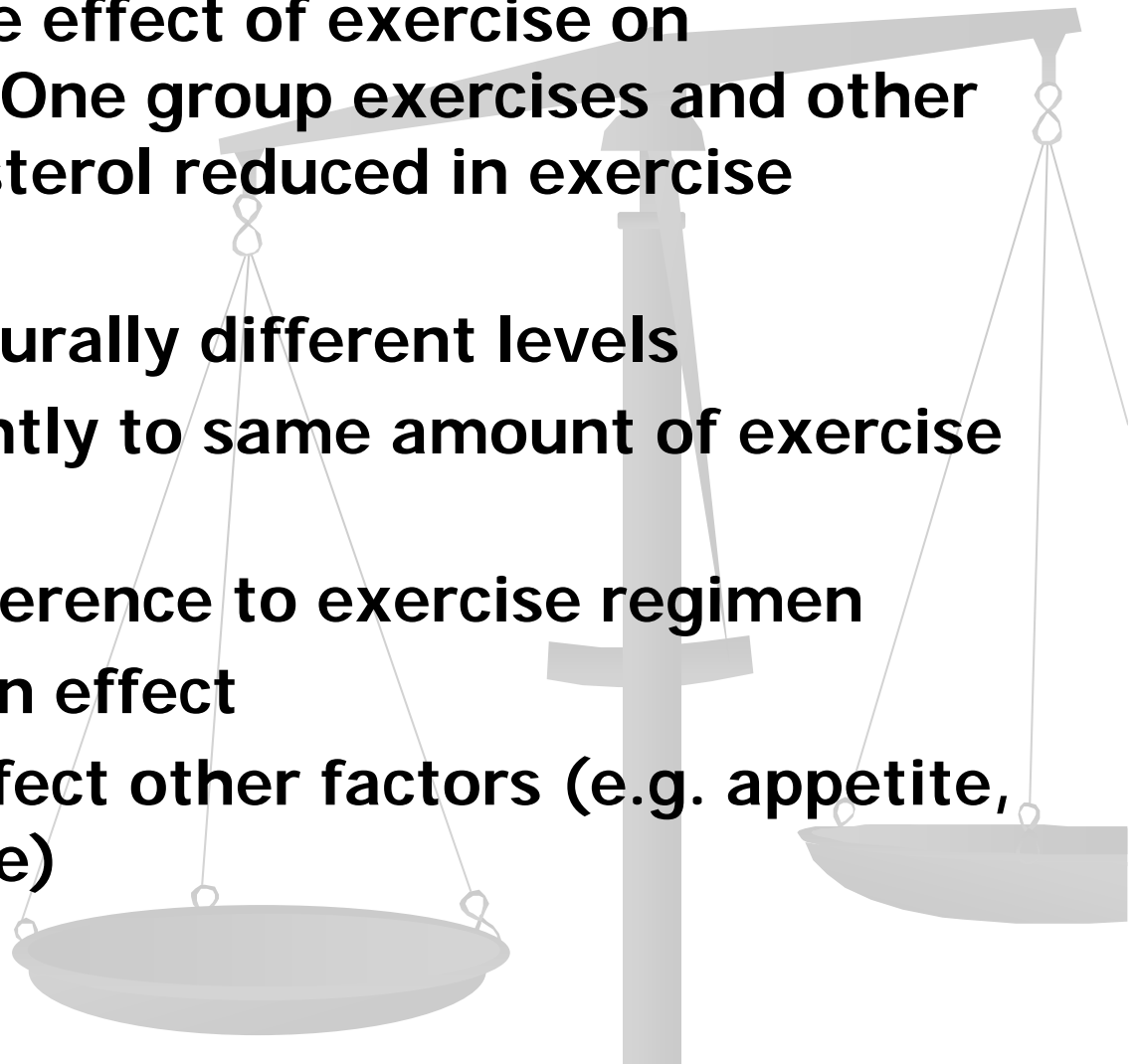
| Subject | SBP | HR | BG | Age | Weight | Treatment |
|---------|-----|----|-----|-----|--------|-----------|
| 1 | 120 | 84 | 100 | 45 | 140 | 1 |
| 2 | 160 | 75 | 233 | 52 | 160 | 1 |
| 3 | 95 | 63 | 92 | 44 | 110 | 2 |
| . | . | . | . | . | . | . |

- How do I make sense of these numbers without some meaningful summary?



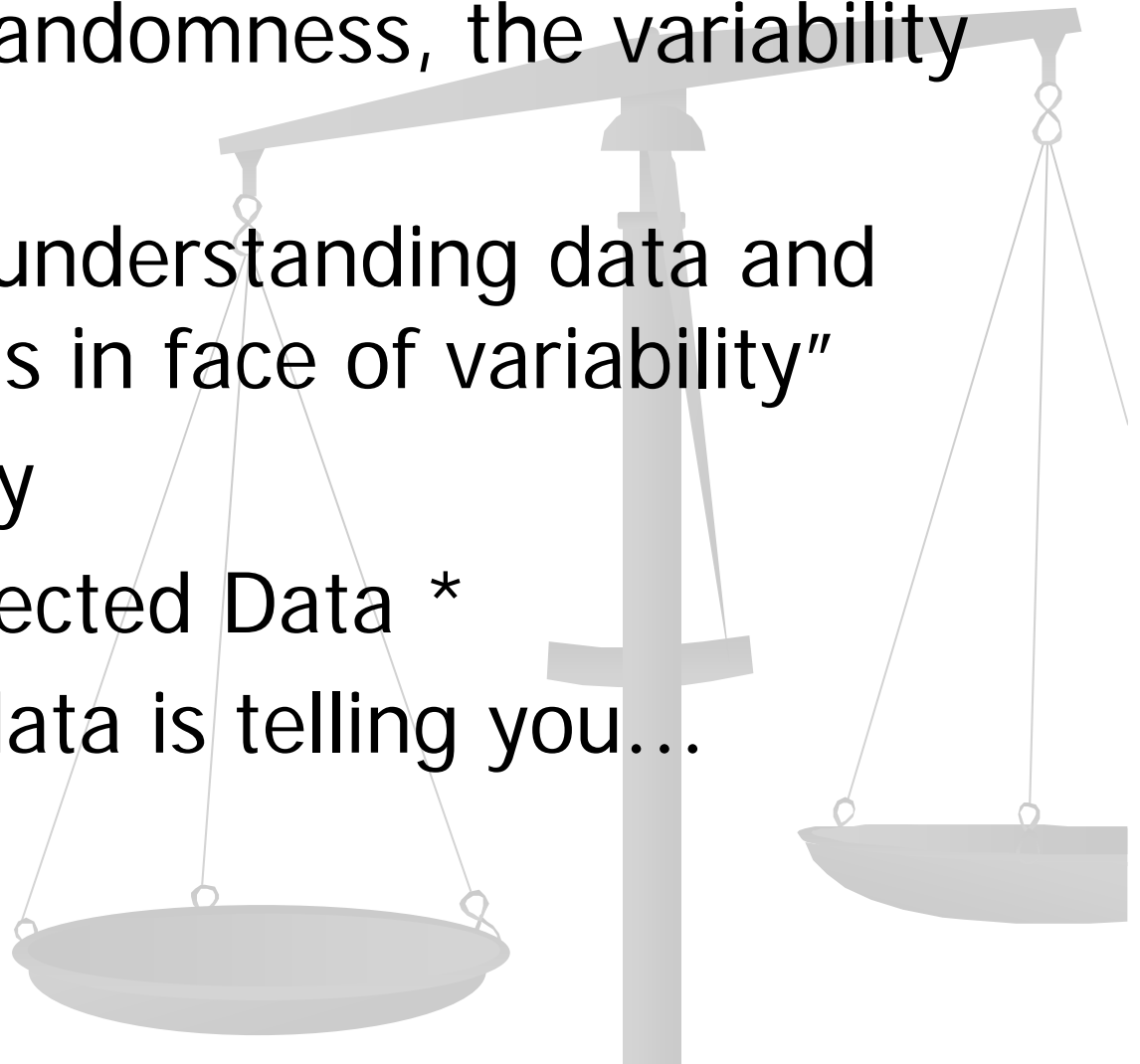
Example

- **Study to assess the effect of exercise on cholesterol levels. One group exercises and other does not. Is cholesterol reduced in exercise group?**
 - **people have naturally different levels**
 - **respond differently to same amount of exercise (e.g. genetics)**
 - **may vary in adherence to exercise regimen**
 - **diet may have an effect**
 - **exercise may affect other factors (e.g. appetite, energy, schedule)**



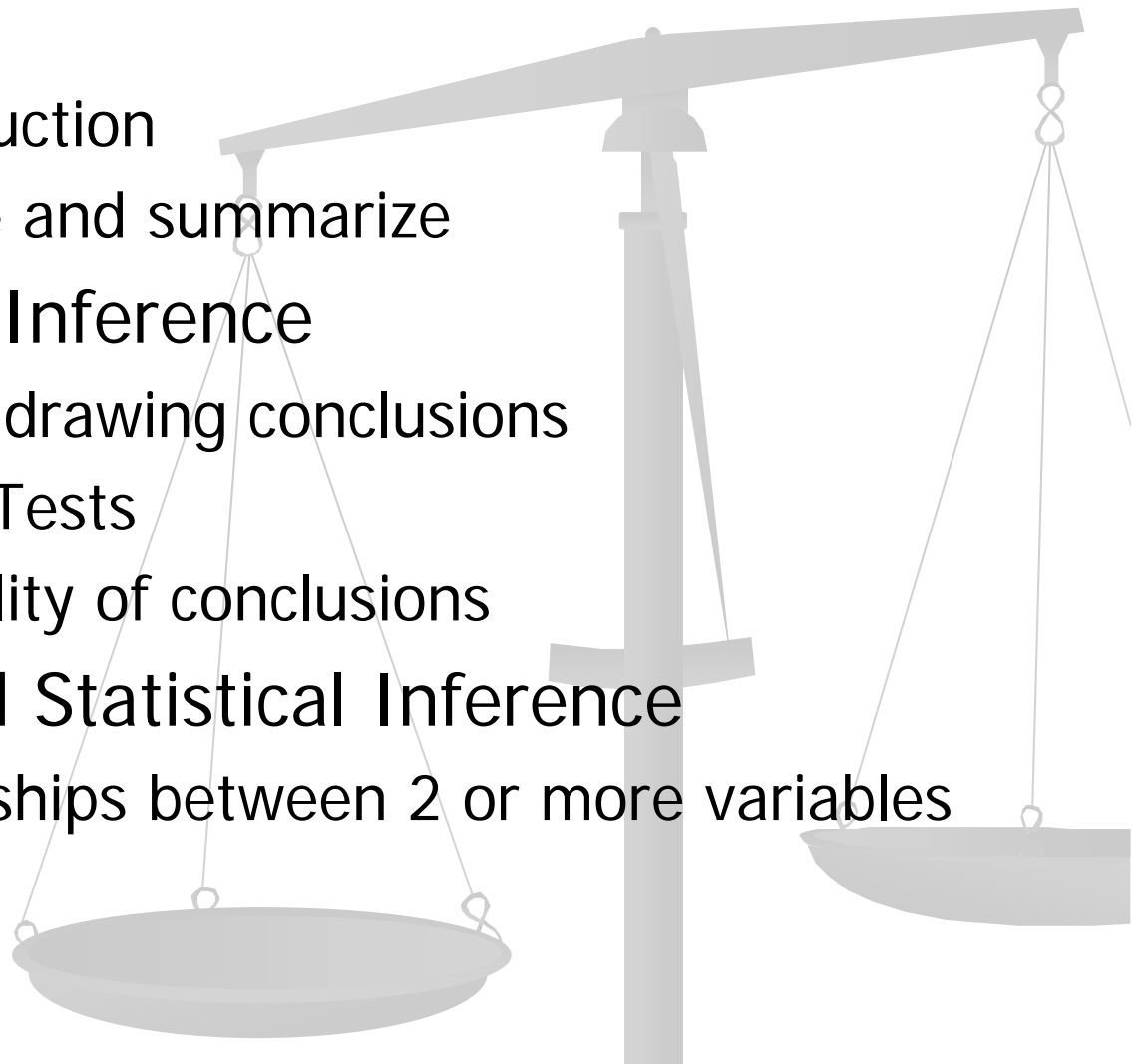
What is statistics?

- Recognize the randomness, the variability in data.
- “the science of understanding data and making decisions in face of variability”
- Design the study
- Analyze the collected Data *
- Discover what data is telling you...



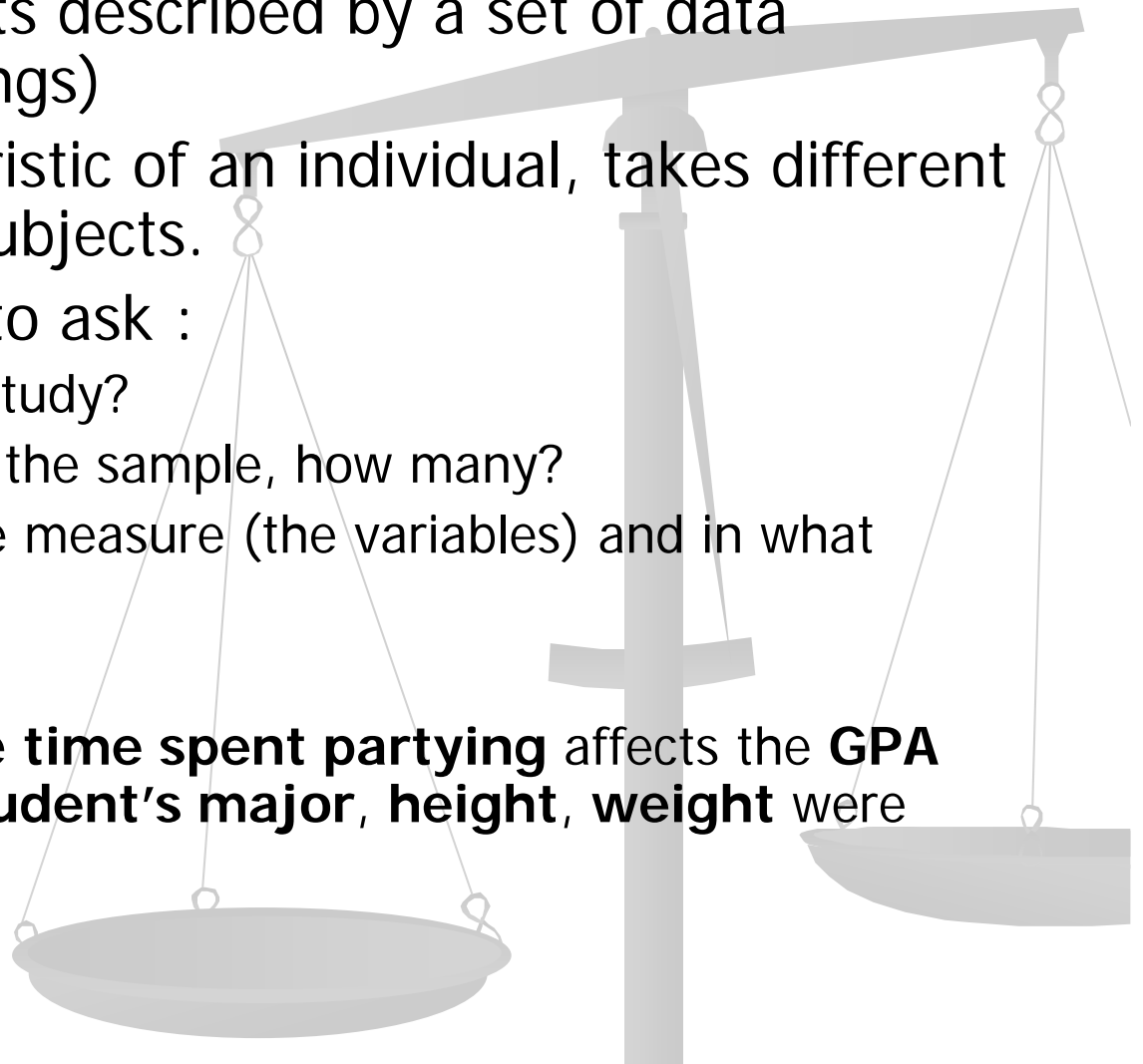
Structure of the course

- Part I: Data:
 - Analysis and production
 - Examine, organize and summarize
- Part II: Statistical Inference
 - Formal Method of drawing conclusions
 - Formal Statistical Tests
 - Testing the reliability of conclusions
- Part III: Advanced Statistical Inference
 - Analyzing relationships between 2 or more variables



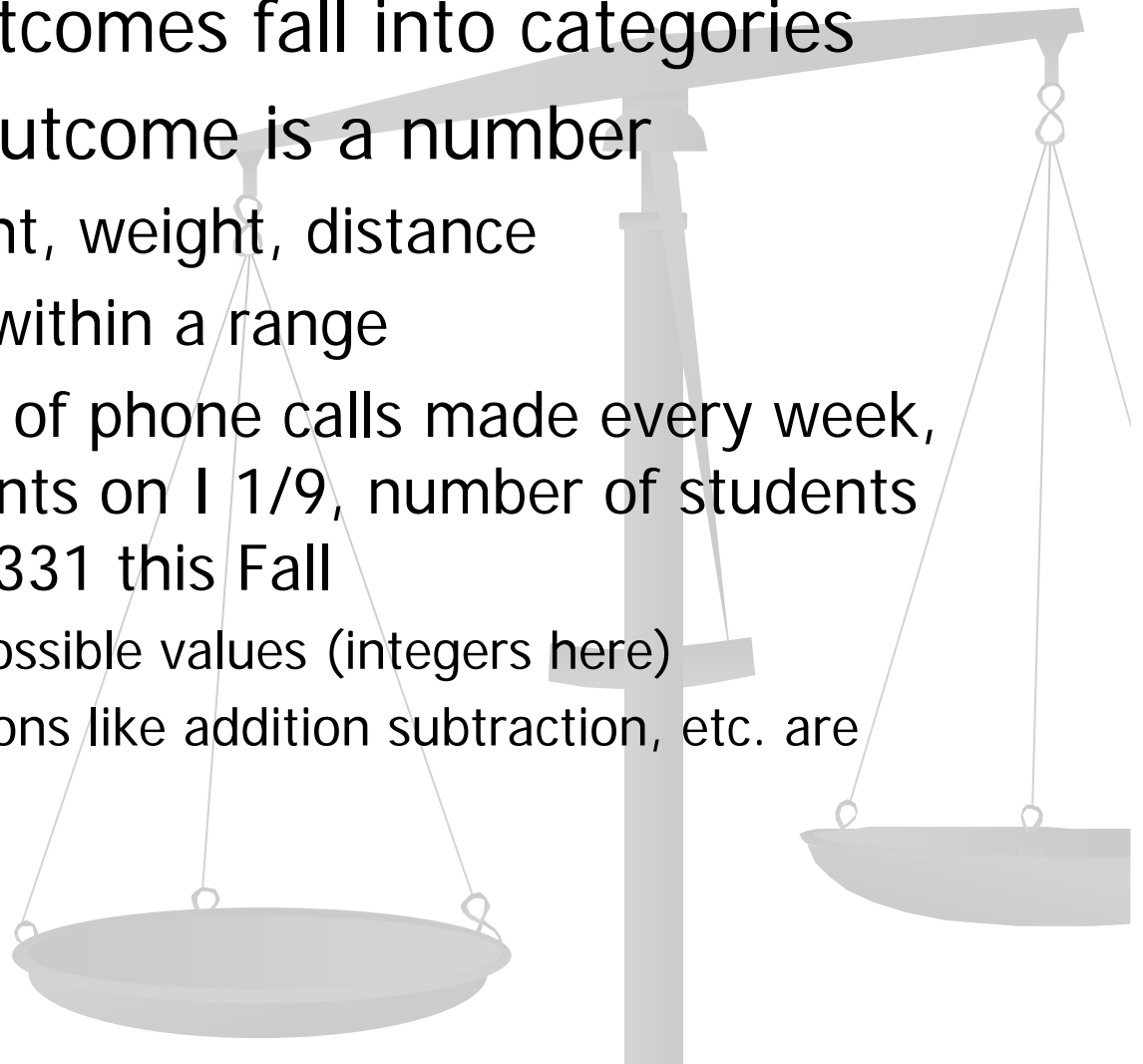
Chapter 1

- **Individuals** – objects described by a set of data (people, animals, things)
- **Variable** – characteristic of an individual, takes different values for different subjects.
- The three questions to ask :
 - Why: Purpose of study?
 - Who: Members of the sample, how many?
 - What: What did we measure (the variables) and in what units?
- **Example:**
 - In a study on how the **time spent partying** affects the **GPA** variables like **age, student's major, height, weight** were also recorded...

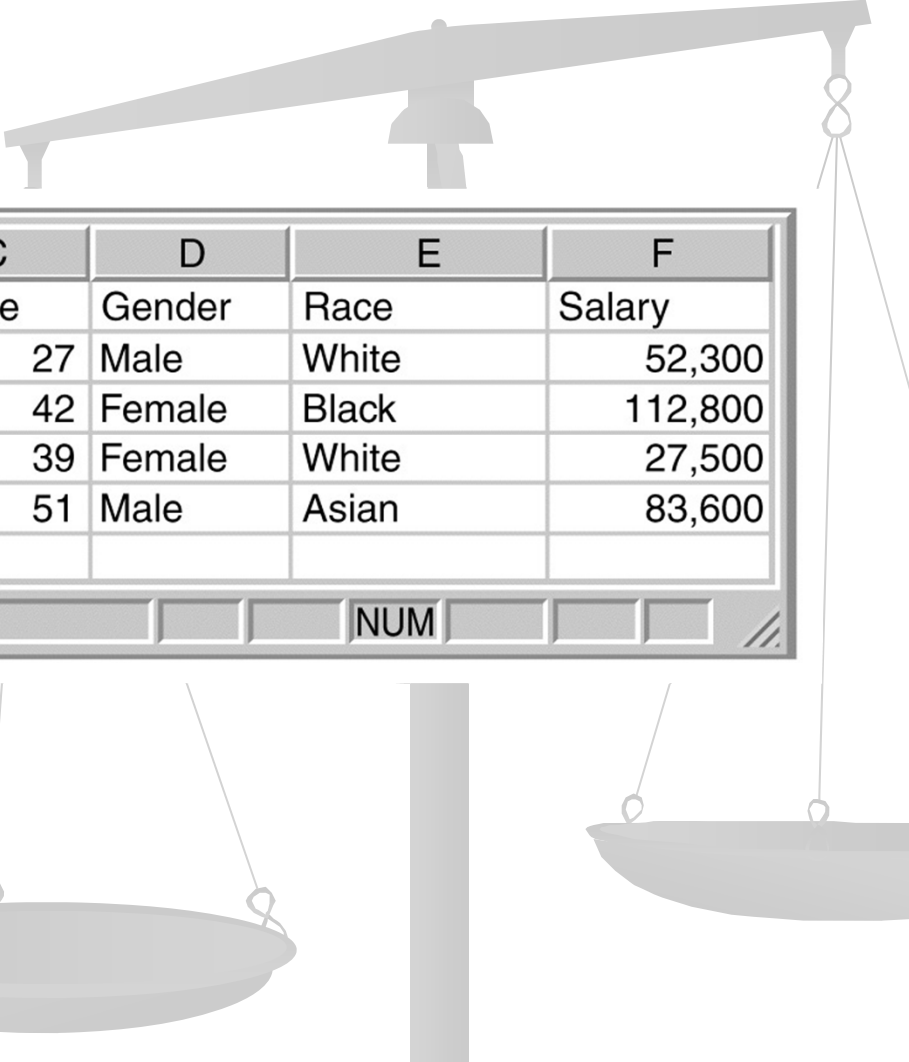


Variable types:

- **Categorical** – outcomes fall into categories
- **Quantitative** – outcome is a number
 - *Continuous* : height, weight, distance
Can take any value within a range
 - *Discrete* : number of phone calls made every week, number of accidents on I 1/9, number of students getting A in Math 331 this Fall
 - Can not take all possible values (integers here)
 - Arithmetic operations like addition subtraction, etc. are meaningful



Information on employees of Cyberstatnet



| | A | B | C | D | E | F |
|---|------------------|------------|-----|--------|-------|---------|
| 1 | Name | Job Type | Age | Gender | Race | Salary |
| 2 | Cedillo, Jose | Technical | 27 | Male | White | 52,300 |
| 3 | Chambers, Tonia | Management | 42 | Female | Black | 112,800 |
| 4 | Childers, Amanda | Clerical | 39 | Female | White | 27,500 |
| 5 | Chen, Huabang | Technical | 51 | Male | Asian | 83,600 |
| 6 | | | | | | |

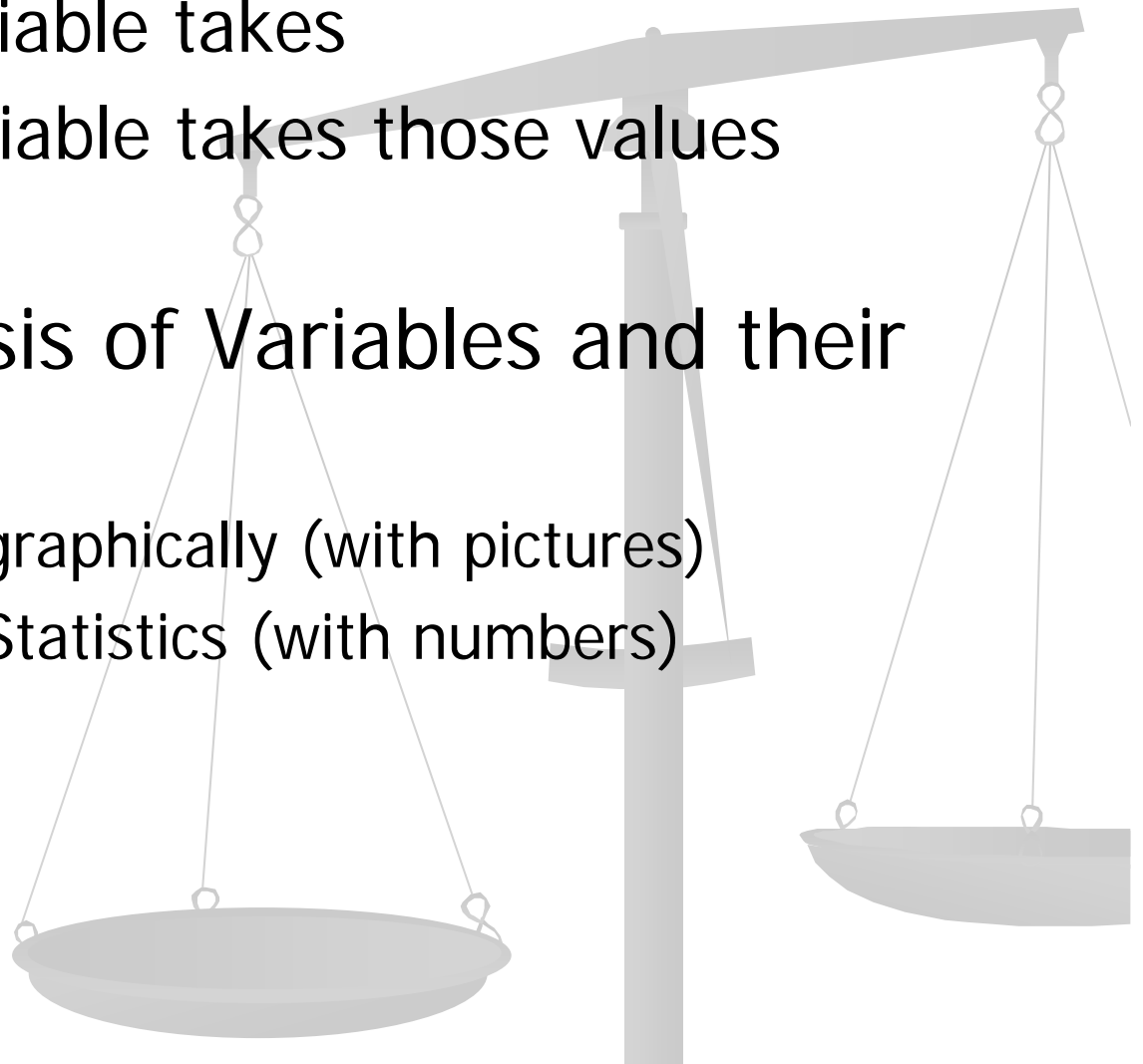
Ready NUM

Distribution of a variable:

- What values a variable takes
- How often the variable takes those values (frequency)

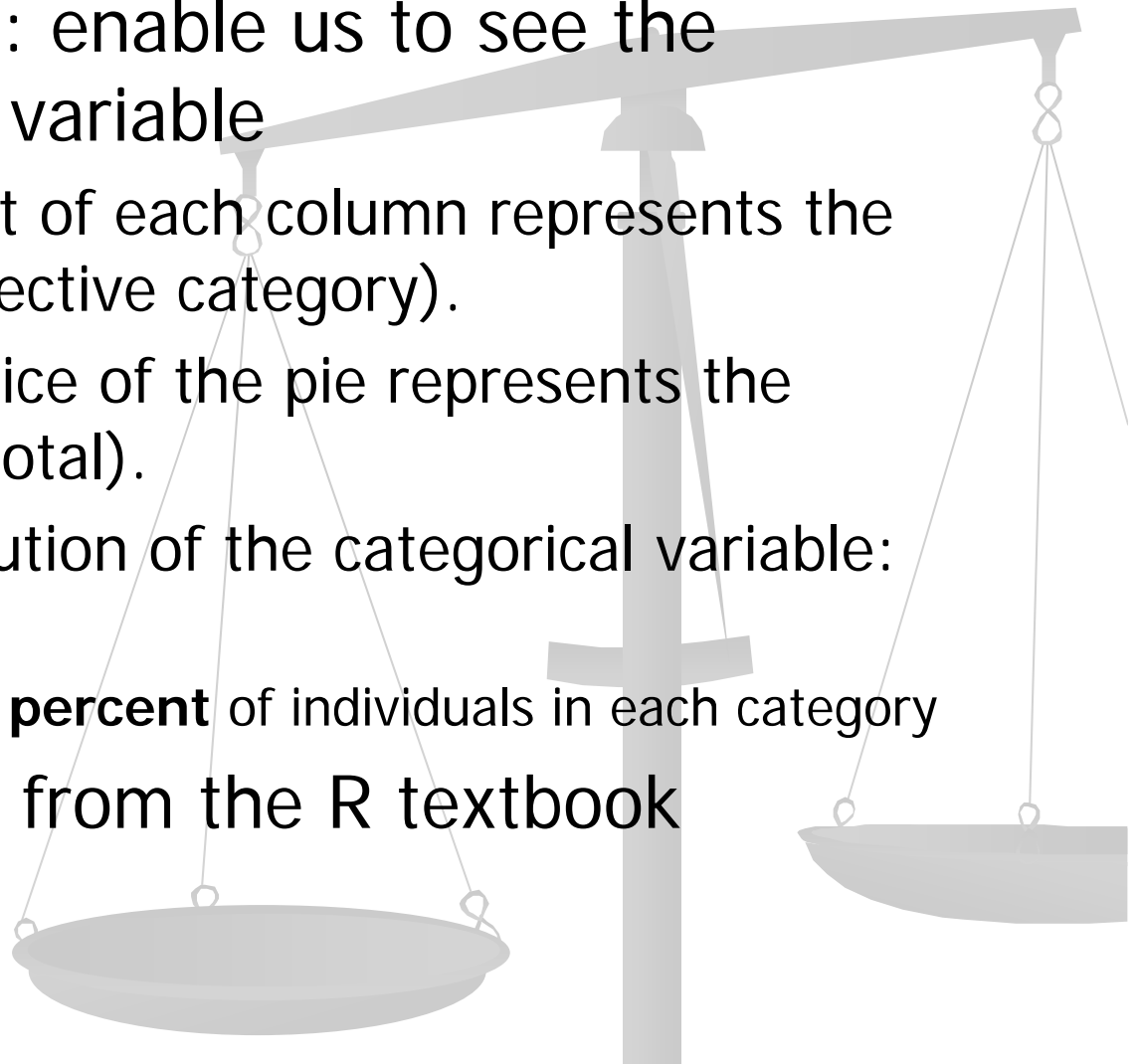
Preliminary Analysis of Variables and their distributions:

- Display variables graphically (with pictures)
- Basic Descriptive Statistics (with numbers)



For the Categorical Variables

- Graphical Displays: enable us to see the distribution of the variable
 - Bar Graphs (height of each column represents the counts in the respective category).
 - Pie charts (each slice of the pie represents the percent from the total).
 - To find the distribution of the categorical variable:
 - List Categories
 - Indicate **count** or **percent** of individuals in each category
- Read pages 75-80 from the R textbook



Bar Graph

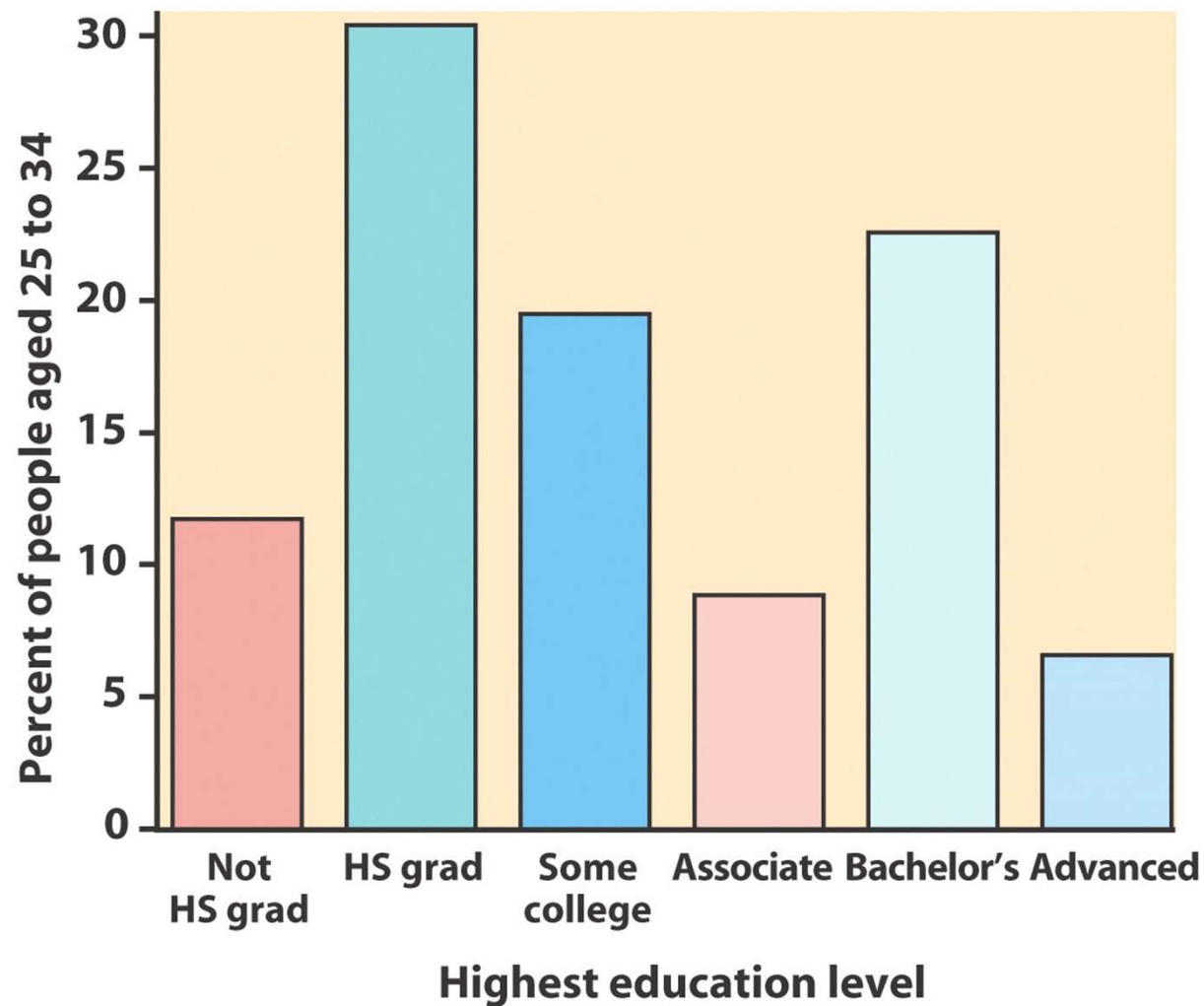
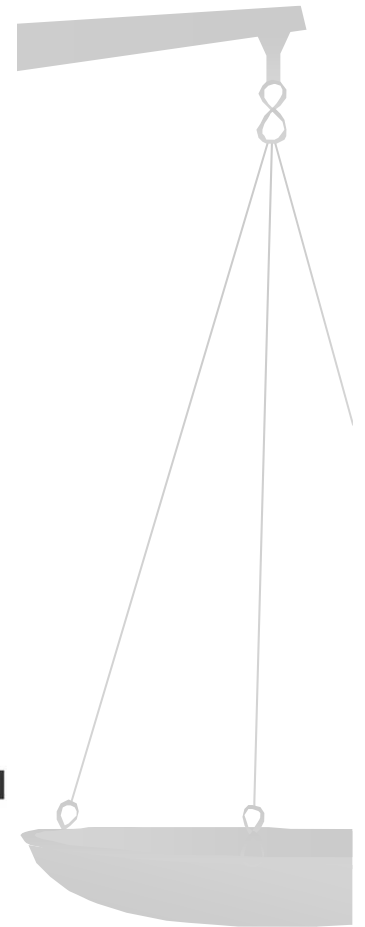


Figure 1-1a
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company



Pie Chart

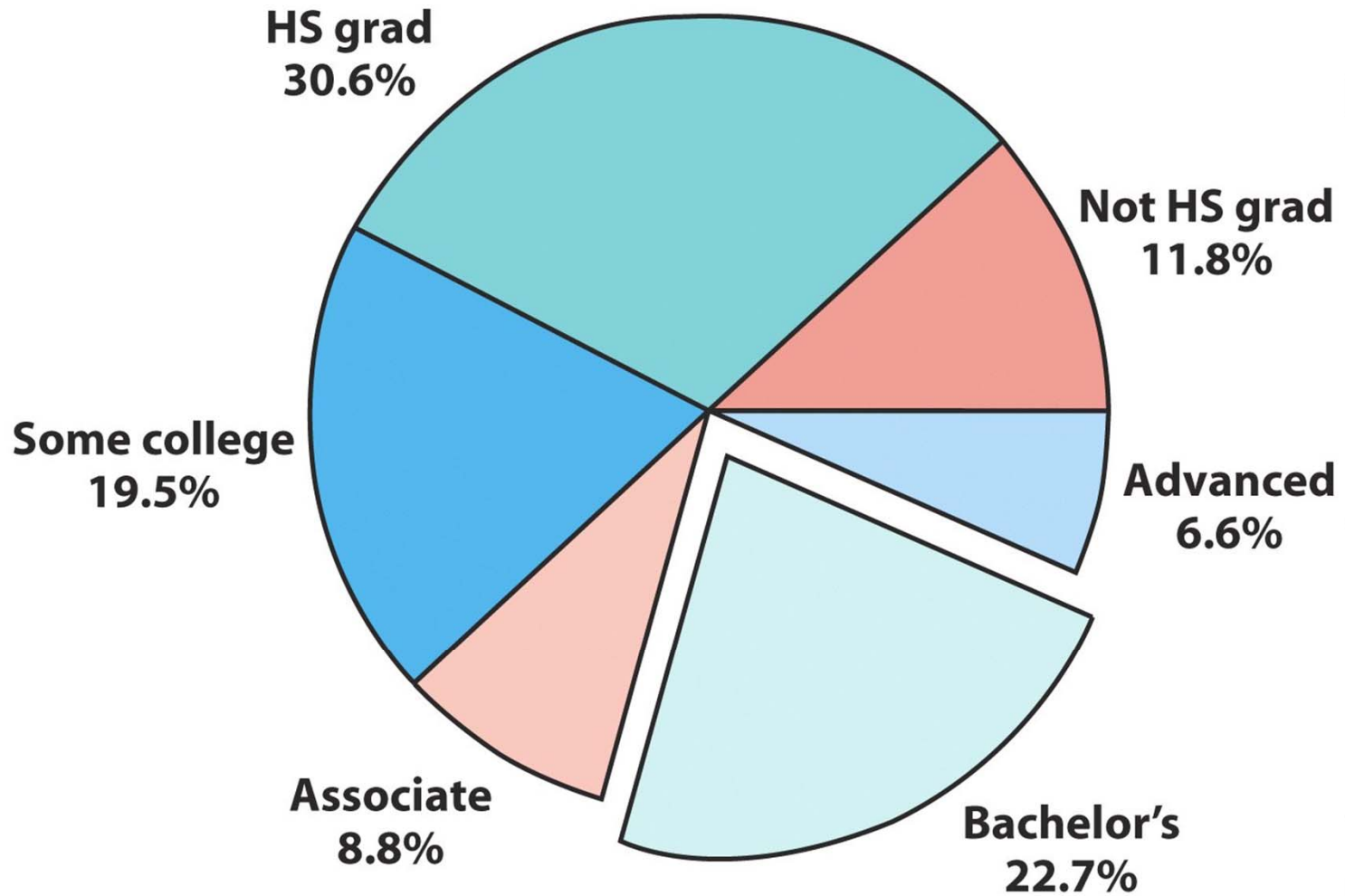


Figure 1-1b
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

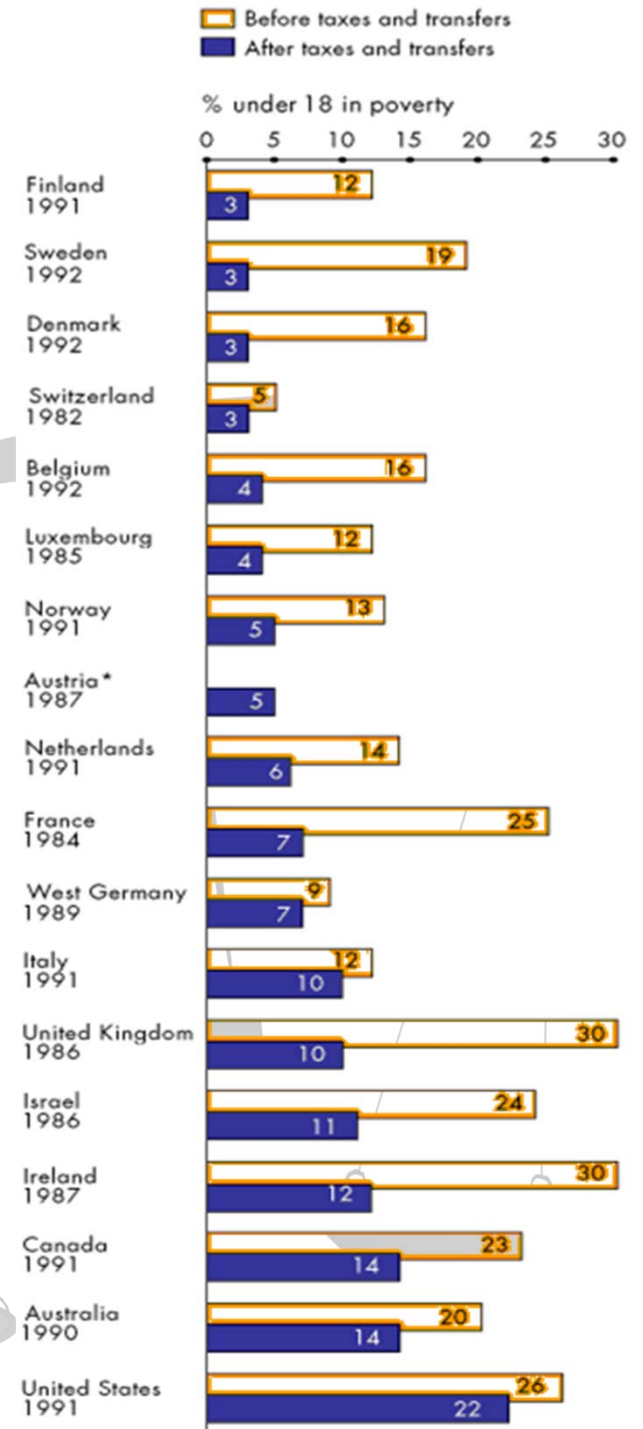
EXAMPLE - Child poverty before and after government intervention—UNICEF, 1996

What does this chart tell you?

- The United States has the highest rate of child poverty among developed nations (22% of under 18).
- Its government does the least—through taxes and subsidies—to remedy the problem (size of orange bars and percent difference between orange/blue bars).

Could you transform this bar graph to fit in 1 pie chart? In two pie charts? Why?

The poverty line is defined as 50% of national median income.



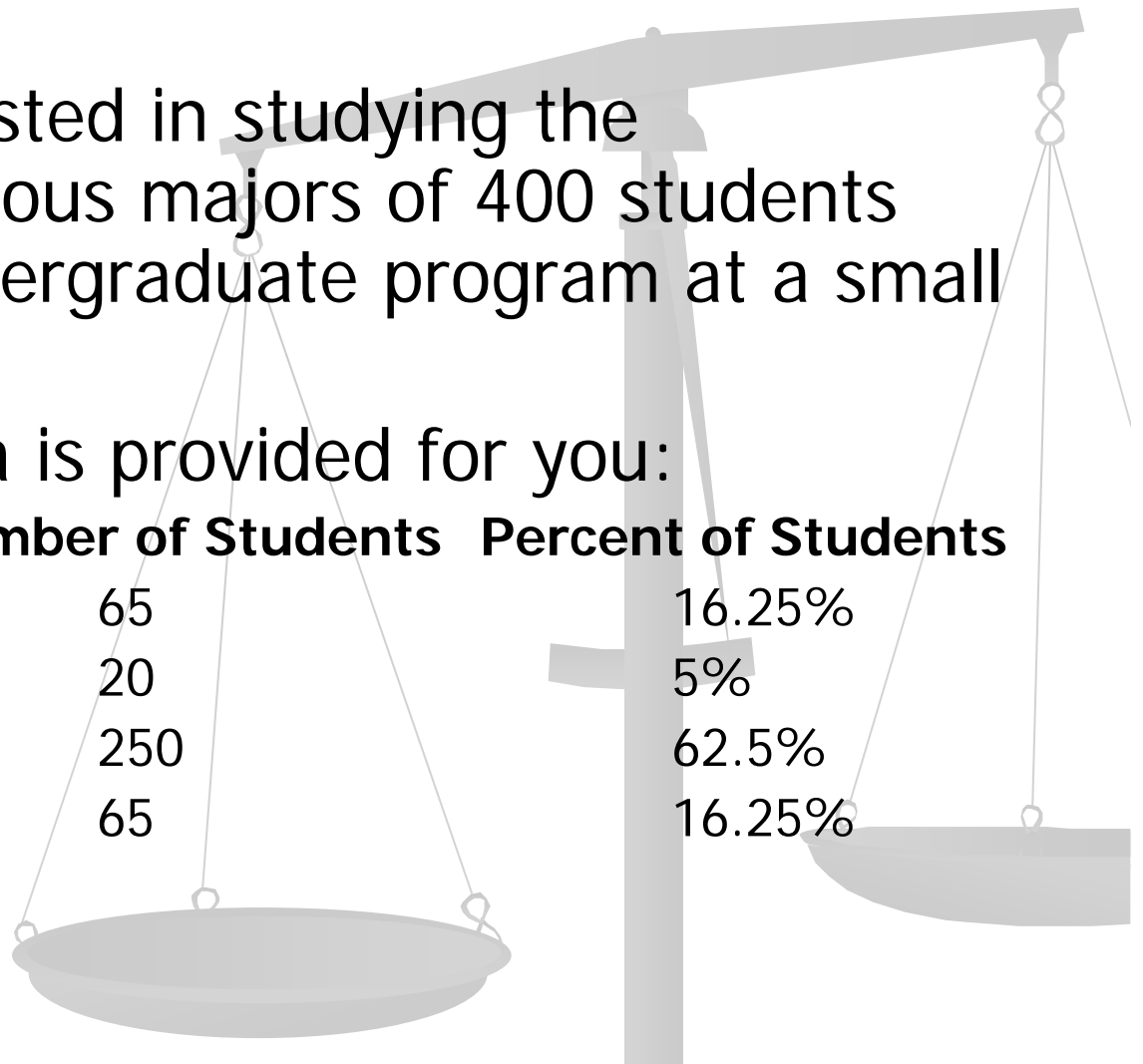
Exercise

- **Example:**

You are interested in studying the distribution of various majors of 400 students enrolled in an undergraduate program at a small university.

- The following data is provided for you:

| ■ Major | Number of Students | Percent of Students |
|-------------------|--------------------|---------------------|
| ■ Math | 65 | 16.25% |
| ■ Stat | 20 | 5% |
| ■ Engineering | 250 | 62.5% |
| ■ Health Sciences | 65 | 16.25% |

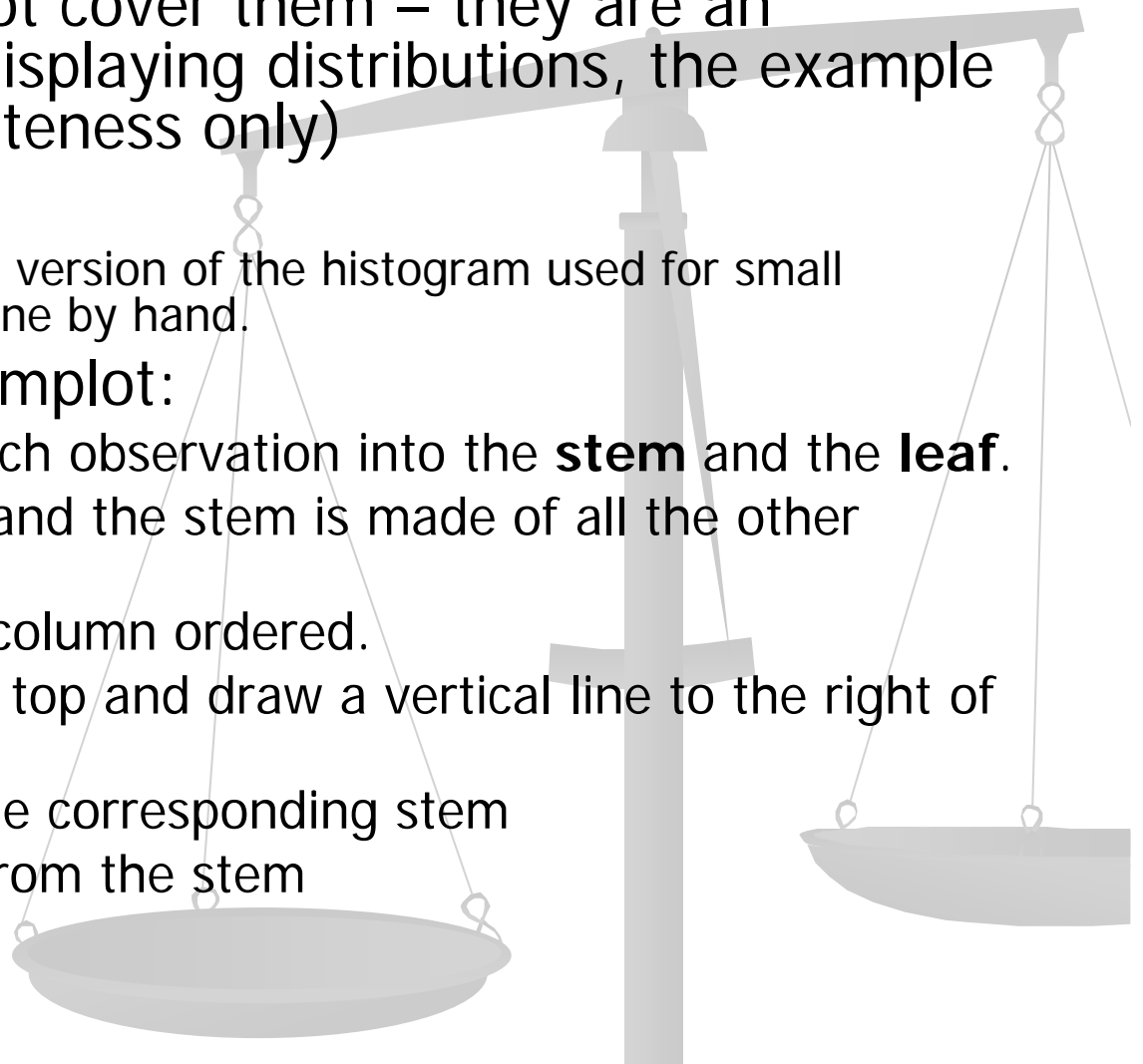


Graphical tools for quantitative data

- Stemplots (We will not cover them – they are an obsolete method of displaying distributions, the example we give is for completeness only)
- Histograms
 - The stemplot is a simple version of the histogram used for small datasets, that can be done by hand.

Steps to construct a stemplot:

- Separate the value for each observation into the **stem** and the **leaf**.
- The leaf is the final digit and the stem is made of all the other digits.
- Write stems in a vertical column ordered.
- Write the smallest on the top and draw a vertical line to the right of the column.
- Write each leaf next to the corresponding stem
- Write them increasingly from the stem



Stemplot example (FYI)

Example 1.4

Numbers of home runs that Babe Ruth hit in each of his 15 years with the New York Yankees:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

- Step 1: Sort the data, sort the stems.

2 3 4 5 6

- Step 2: Write the stems in increasing order

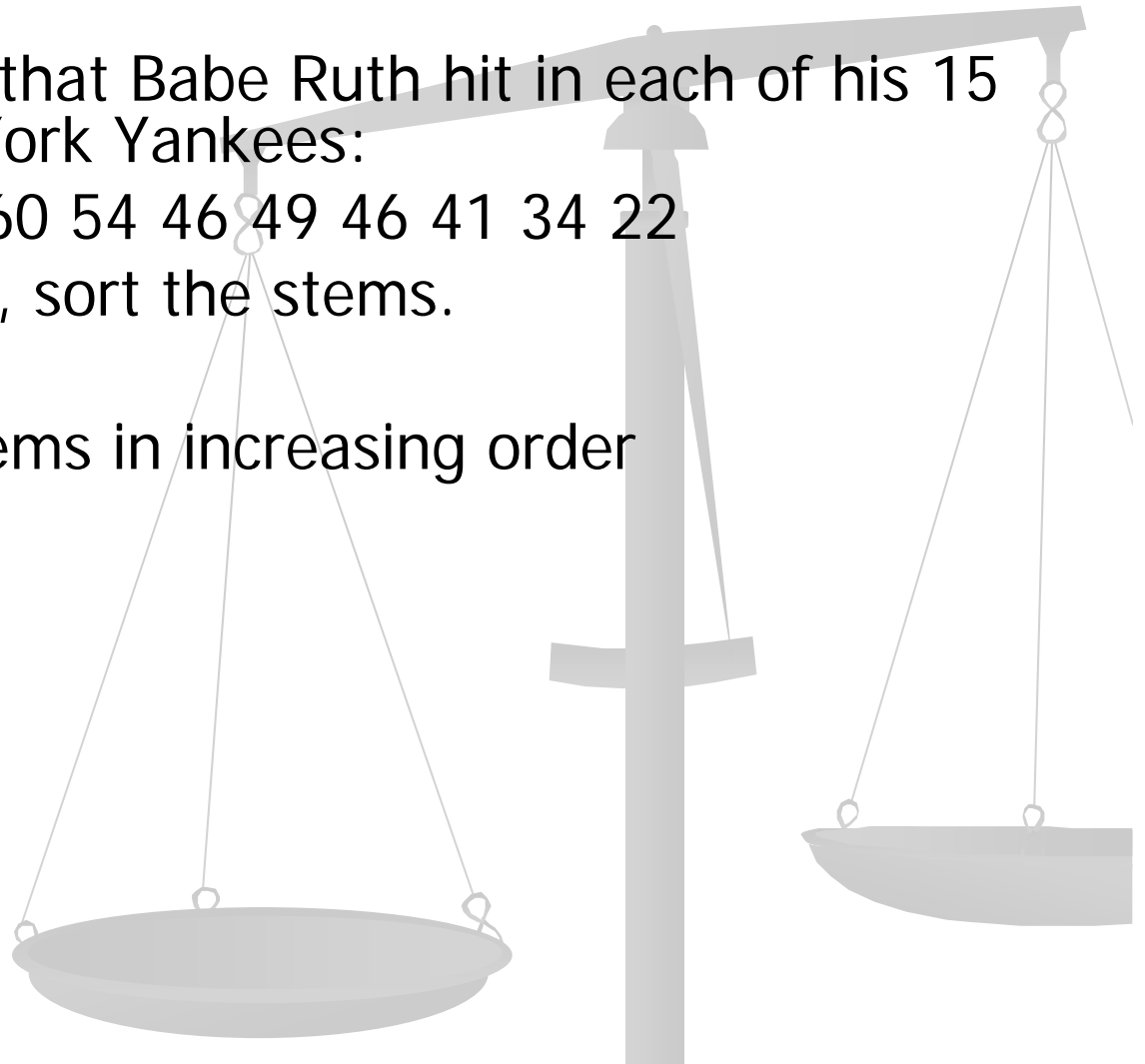
2

3

4

5

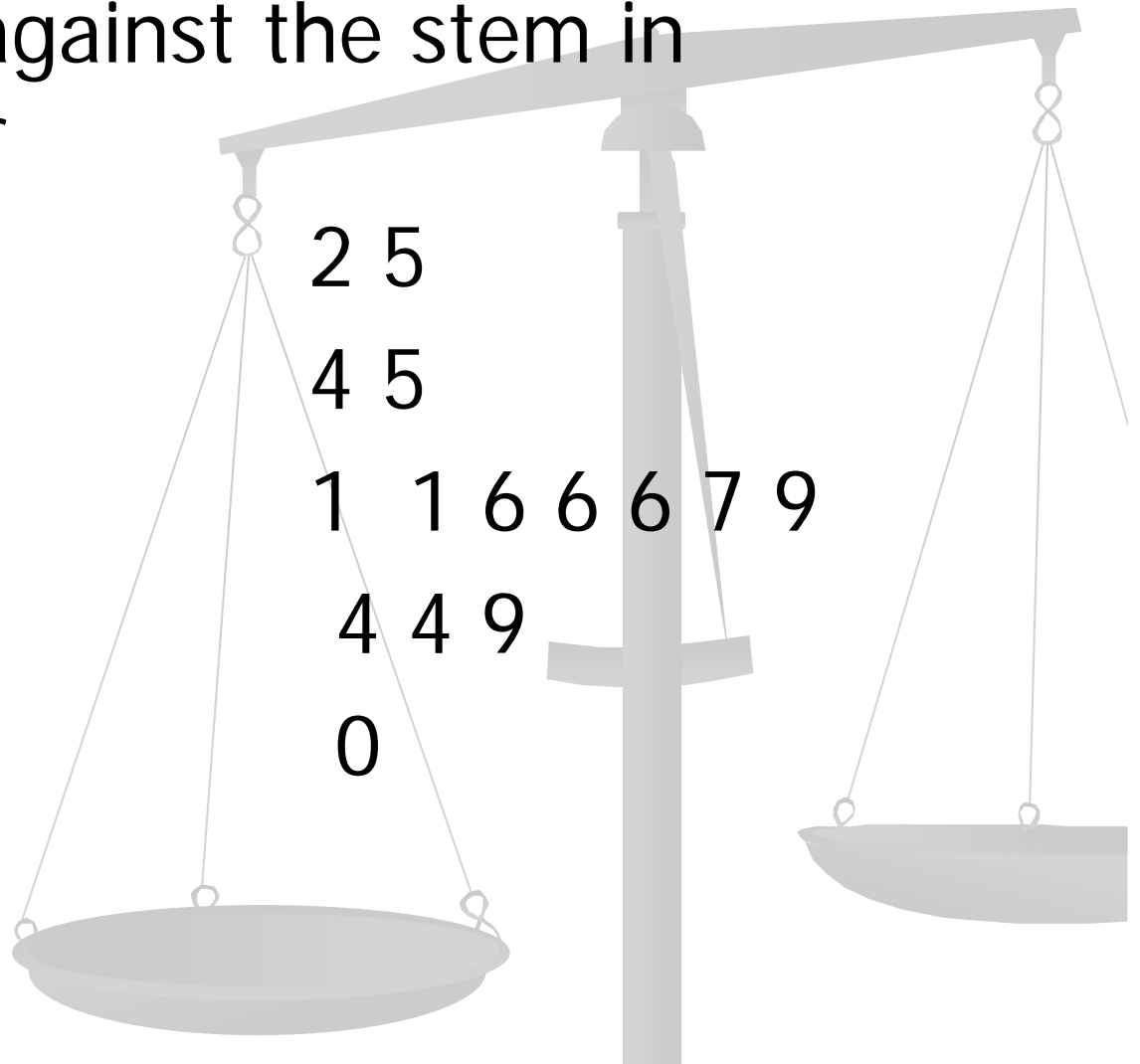
6



Stemplot (FYI)

Write the leaves against the stem in increasing order

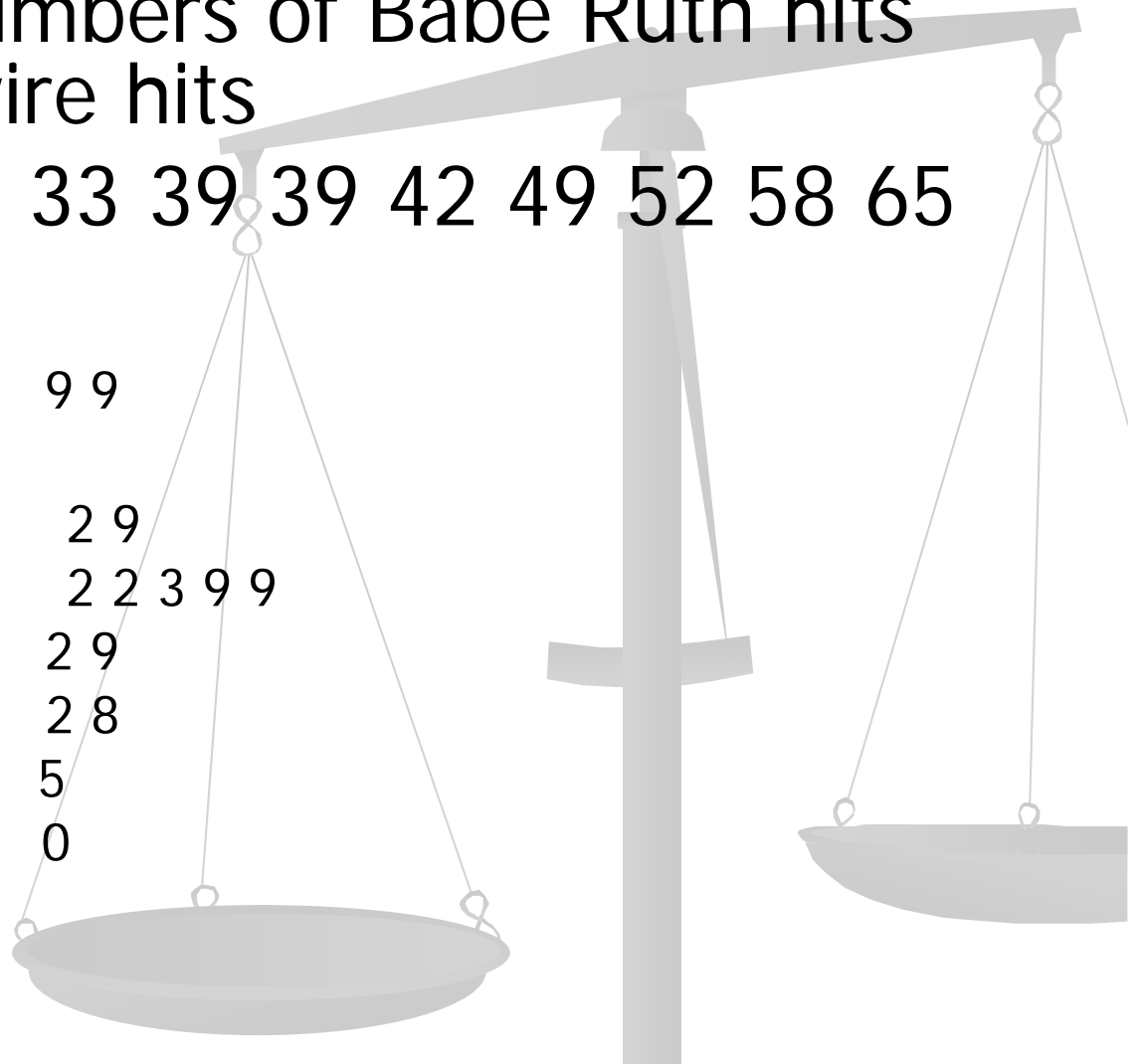
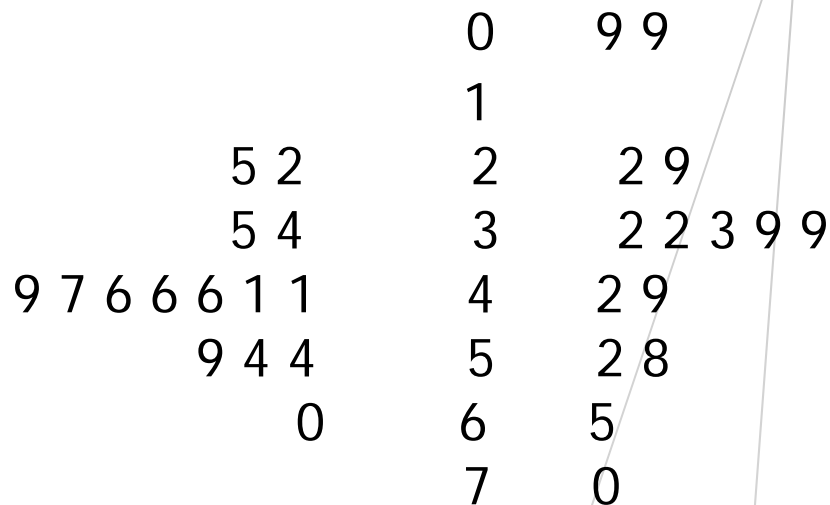
| | |
|---|---------------|
| 2 | 2 5 |
| 3 | 4 5 |
| 4 | 1 1 6 6 6 7 9 |
| 5 | 4 4 9 |
| 6 | 0 |



Back-to-back stemplot (FYI)

- Compare the numbers of Babe Ruth hits and Mark McGwire hits

- 9 9 22 29 32 32 33 39 39 42 49 52 58 65
70



Histograms (example)

TABLE 1.2 Percent of Hispanics in the adult population, by state (2000)

| State | Percent | State | Percent | State | Percent |
|-------------|---------|----------------|---------|----------------|---------|
| Alabama | 1.5 | Louisiana | 2.4 | Ohio | 1.6 |
| Alaska | 3.6 | Maine | 0.6 | Oklahoma | 4.3 |
| Arizona | 21.3 | Maryland | 4.0 | Oregon | 6.5 |
| Arkansas | 2.8 | Massachusetts | 5.6 | Pennsylvania | 2.6 |
| California | 28.1 | Michigan | 2.7 | Rhode Island | 7.0 |
| Colorado | 14.9 | Minnesota | 2.4 | South Carolina | 2.2 |
| Connecticut | 8.0 | Mississippi | 1.3 | South Dakota | 1.2 |
| Delaware | 4.0 | Missouri | 1.8 | Tennessee | 2.0 |
| Florida | 16.1 | Montana | 1.6 | Texas | 28.6 |
| Georgia | 5.0 | Nebraska | 4.5 | Utah | 8.1 |
| Hawaii | 5.7 | Nevada | 16.7 | Vermont | 0.8 |
| Idaho | 6.4 | New Hampshire | 1.4 | Virginia | 4.2 |
| Illinois | 10.7 | New Jersey | 12.3 | Washington | 6.0 |
| Indiana | 3.1 | New Mexico | 38.7 | West Virginia | 0.6 |
| Iowa | 2.3 | New York | 13.8 | Wisconsin | 2.9 |
| Kansas | 5.8 | North Carolina | 4.3 | Wyoming | 5.5 |
| Kentucky | 1.3 | North Dakota | 1.0 | | |

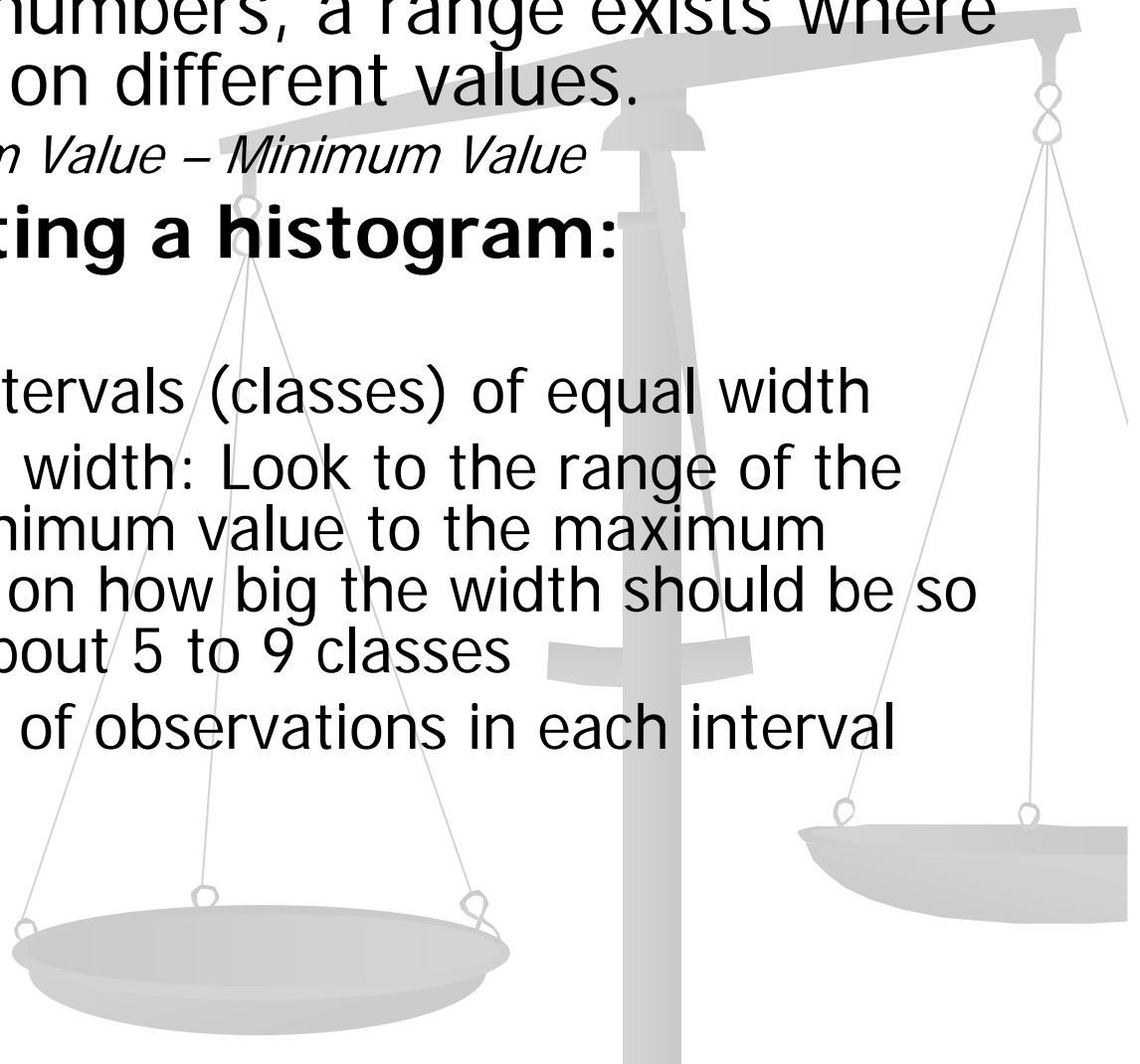
Histogram (cont)

- Within any set of numbers, a range exists where the variable takes on different values.

- $\text{Range} = \text{Maximum Value} - \text{Minimum Value}$

Steps to constructing a histogram:

- Order data
- Divide data into intervals (classes) of equal width
- To choose interval width: Look to the range of the data (from the minimum value to the maximum value) and decide on how big the width should be so you would have about 5 to 9 classes
- Count the number of observations in each interval (class)
- Graph

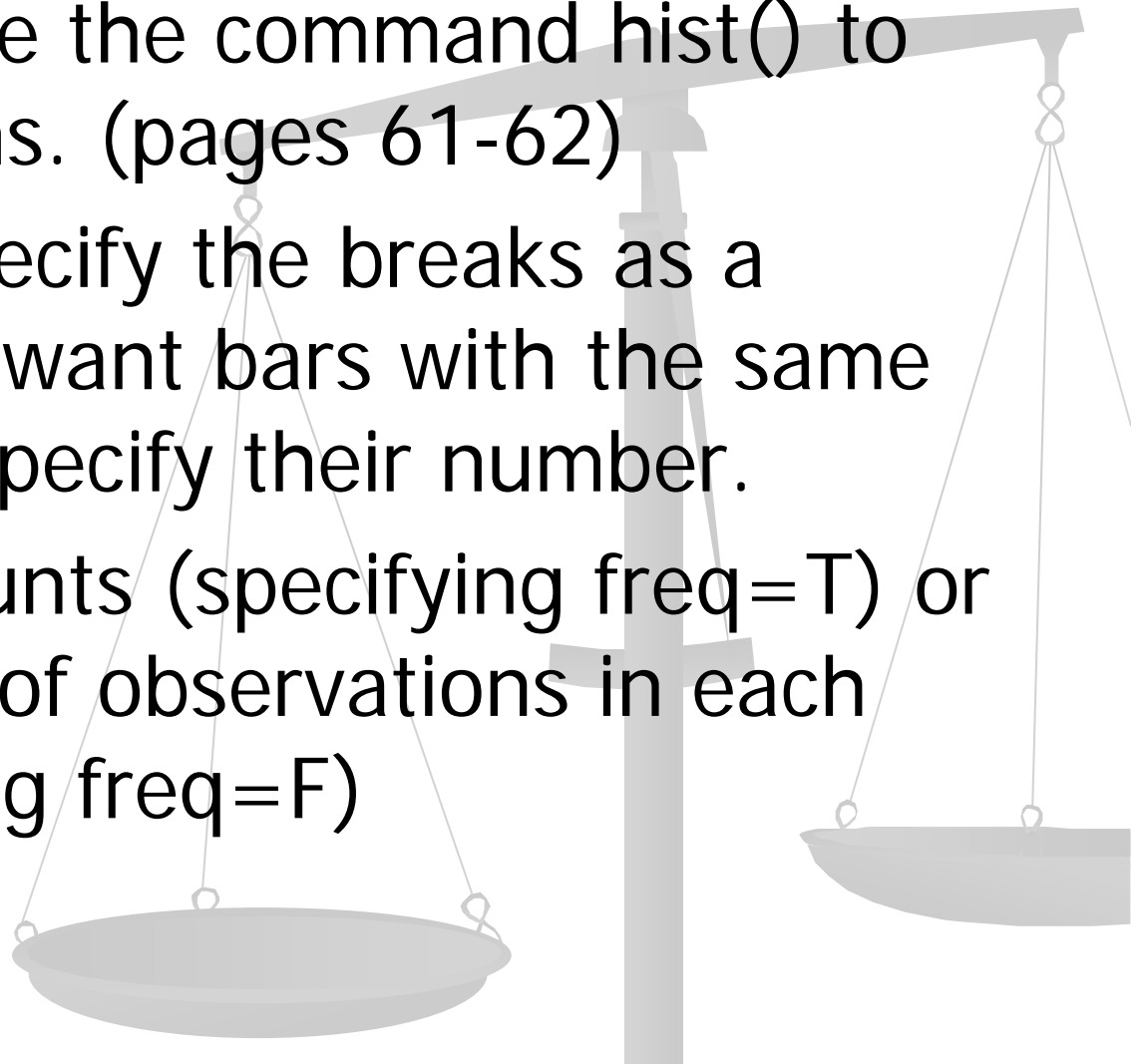


Frequency Table

| Class | Count | Percent | Class | Count | Percent |
|----------|-------|---------|---------|-------|---------|
| 0.1-5.0 | 30 | 60 | 20.1-25 | 1 | 2 |
| 5.1-10.0 | 10 | 20 | 25.1-30 | 2 | 4 |
| 10.1-15 | 4 | 8 | 30.1-35 | 0 | 0 |
| 15.1-20 | 2 | 4 | 35.1-40 | 1 | 2 |

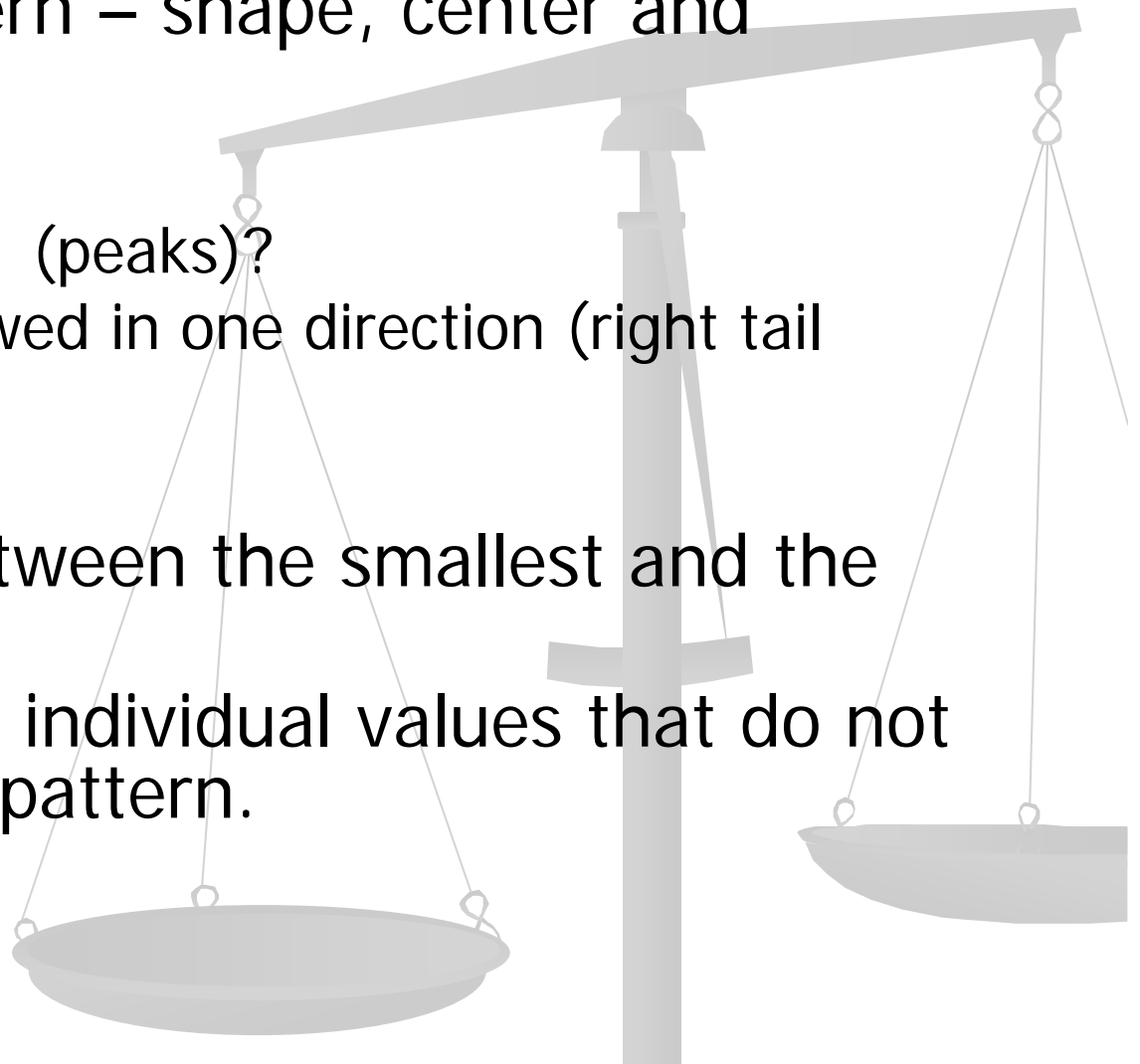
Using R

- In R you can use the command `hist()` to make histograms. (pages 61-62)
- You can also specify the breaks as a vector or if you want bars with the same width you can specify their number.
- You can use counts (specifying `freq=T`) or the percentage of observations in each range (specifying `freq=F`)



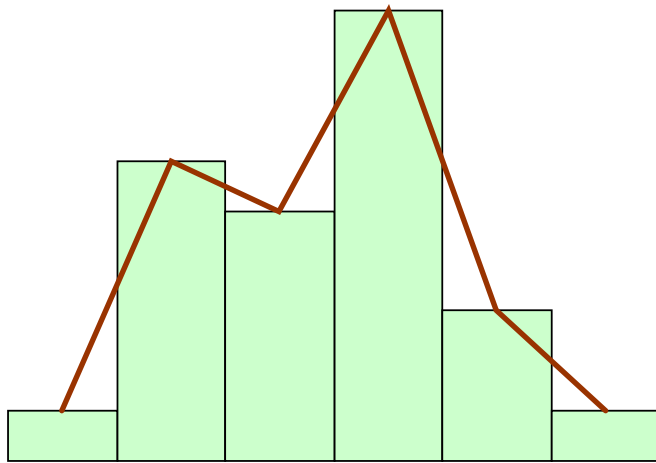
Examining distributions

- Describe the pattern – shape, center and spread.
- Shape –
 - How many modes (peaks)?
 - Symmetric or skewed in one direction (right tail longer or left)
- Center – midpoint
- Spread – range between the smallest and the largest values.
- Look for outliers – individual values that do not match the overall pattern.

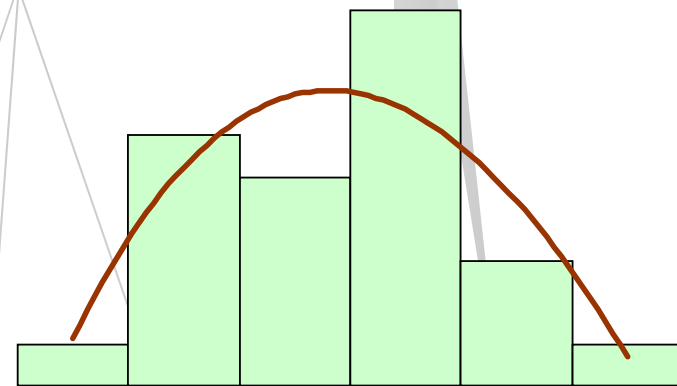


Interpreting histograms

When describing the distribution of a quantitative variable, we look for the overall pattern and for striking deviations from that pattern. We can describe the *overall* pattern of a histogram by its **shape**, **center**, and **spread**.



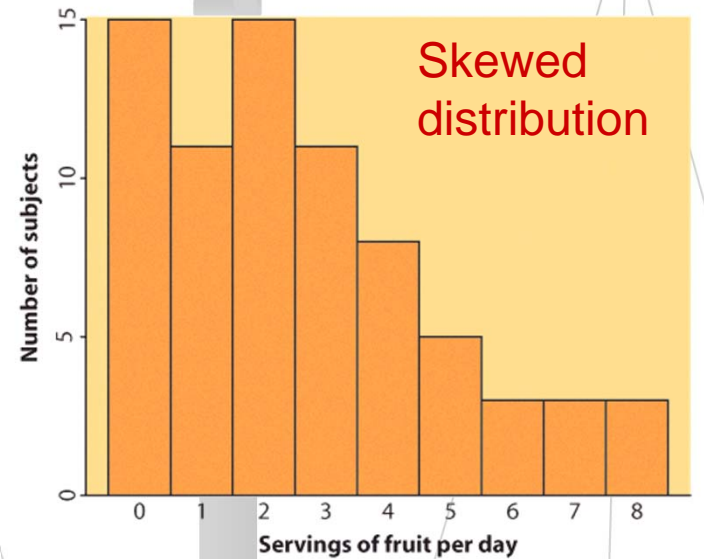
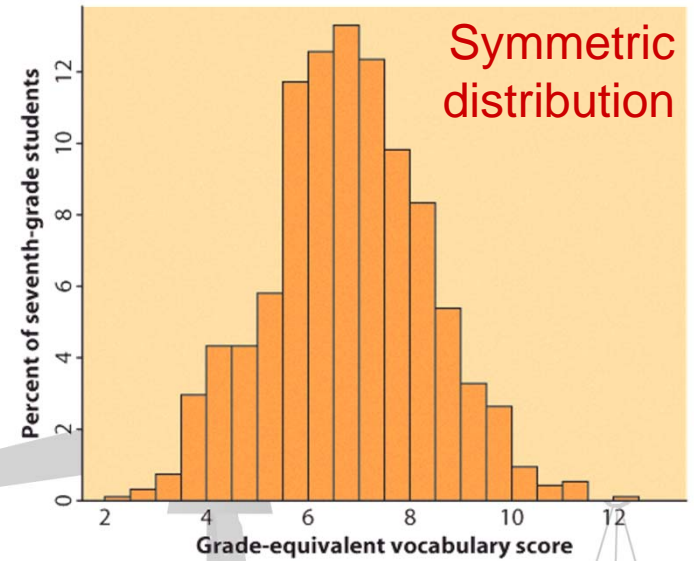
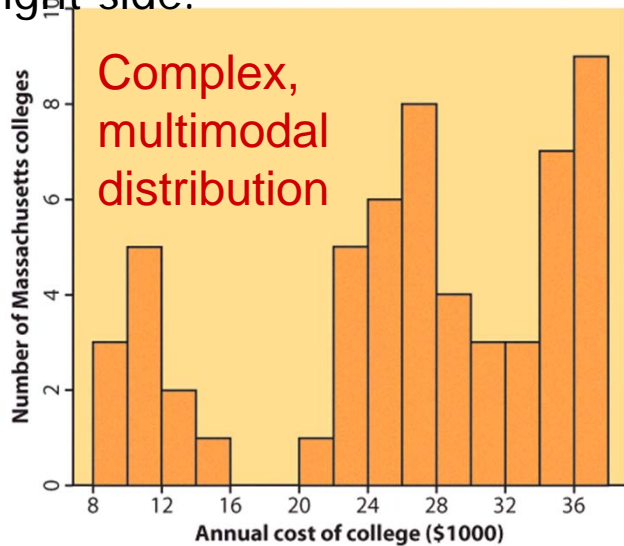
Histogram with a line connecting each column → too detailed



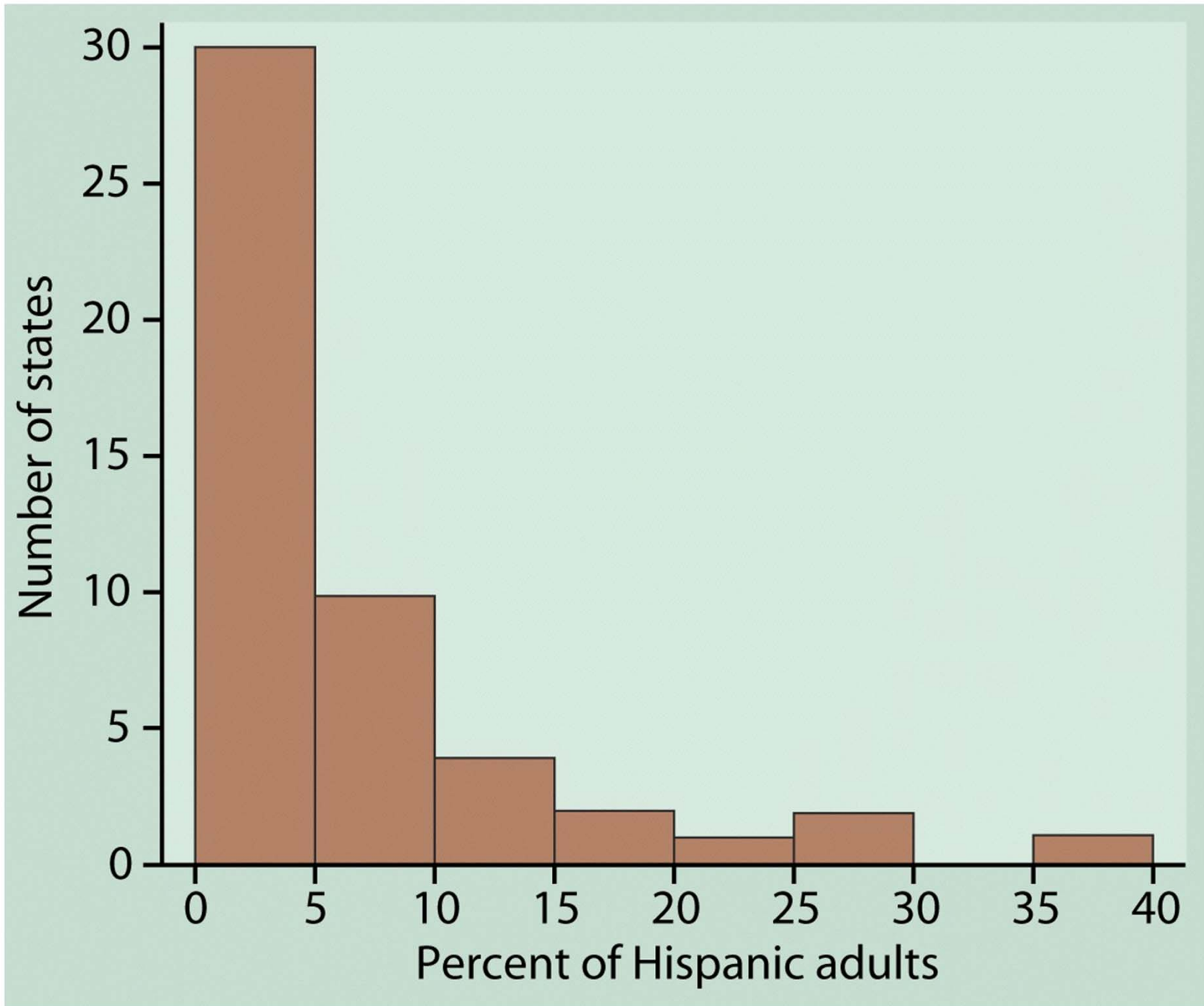
Histogram with a smoothed curve highlighting the overall pattern of the distribution

Most common distribution shapes

- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

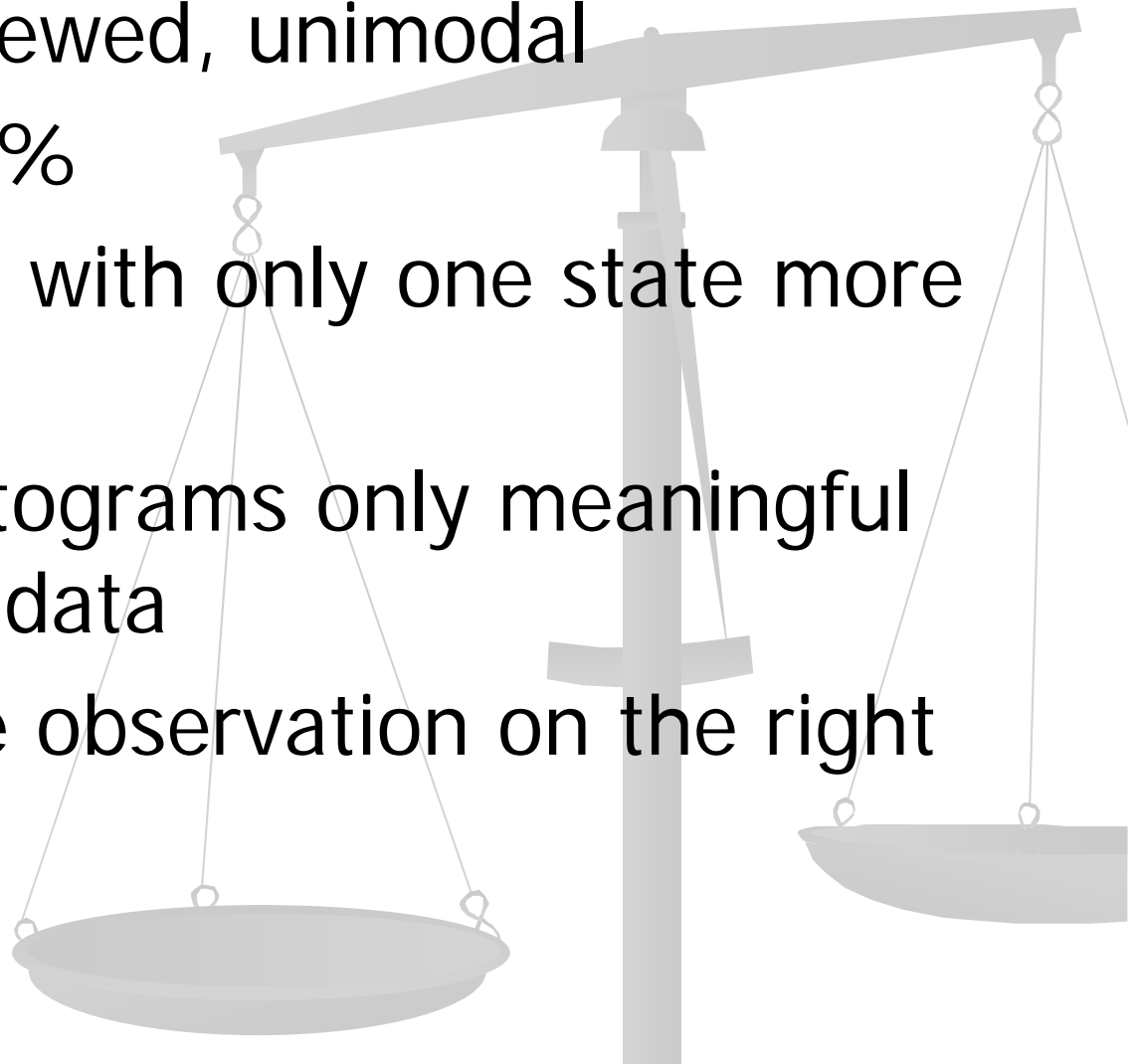


■ Not all distributions have a simple overall shape, especially when there are few observations.



What do you see?

- Shape: Right skewed, unimodal
- Center: about 5%
- Spread : 0-40% with only one state more than 30%
- Remember: Histograms only meaningful for quantitative data
- Is that extreme observation on the right an outlier?



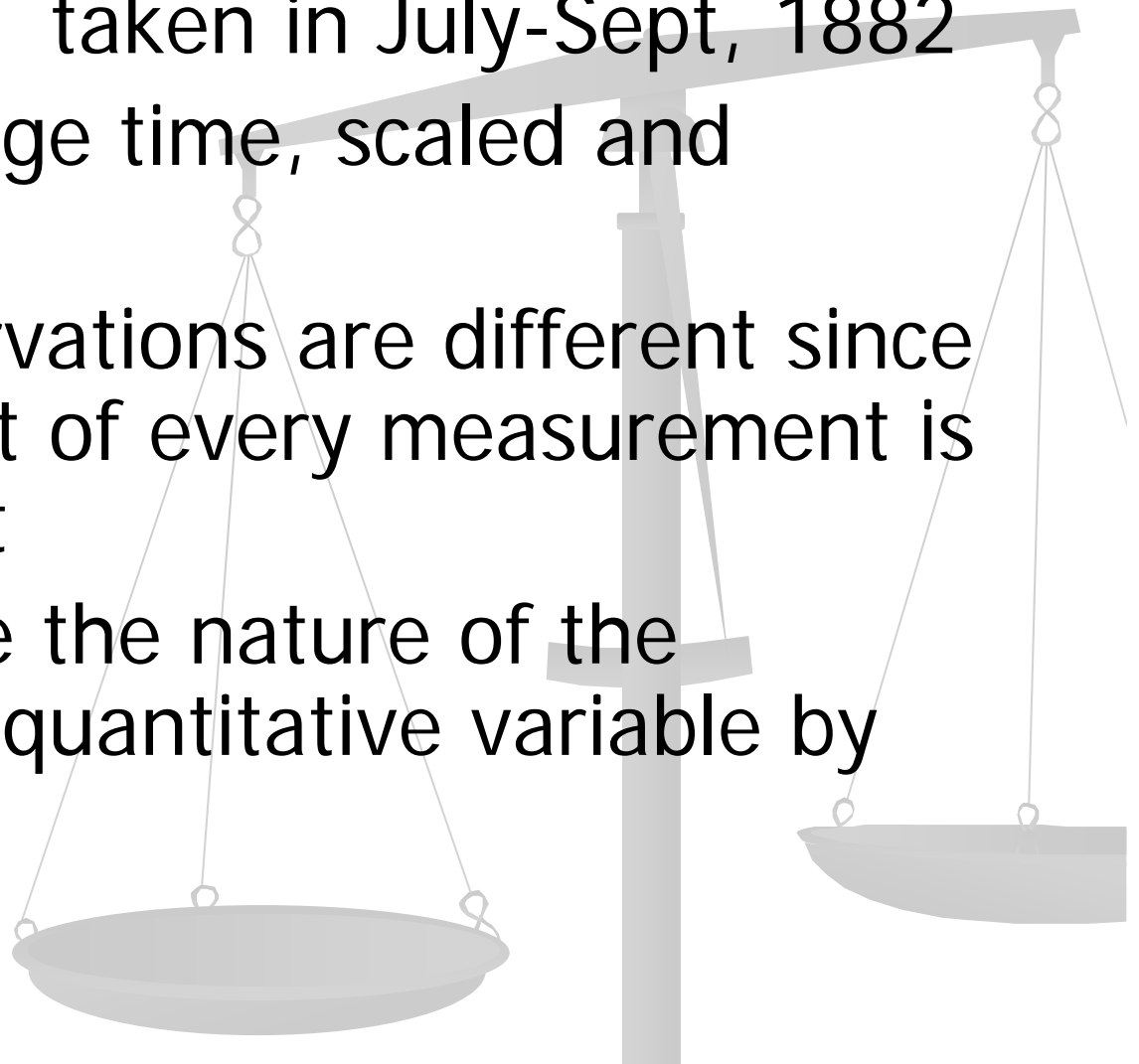
Quantitative Variables-Graphical Display

- Deviations from 24,800 nanoseconds

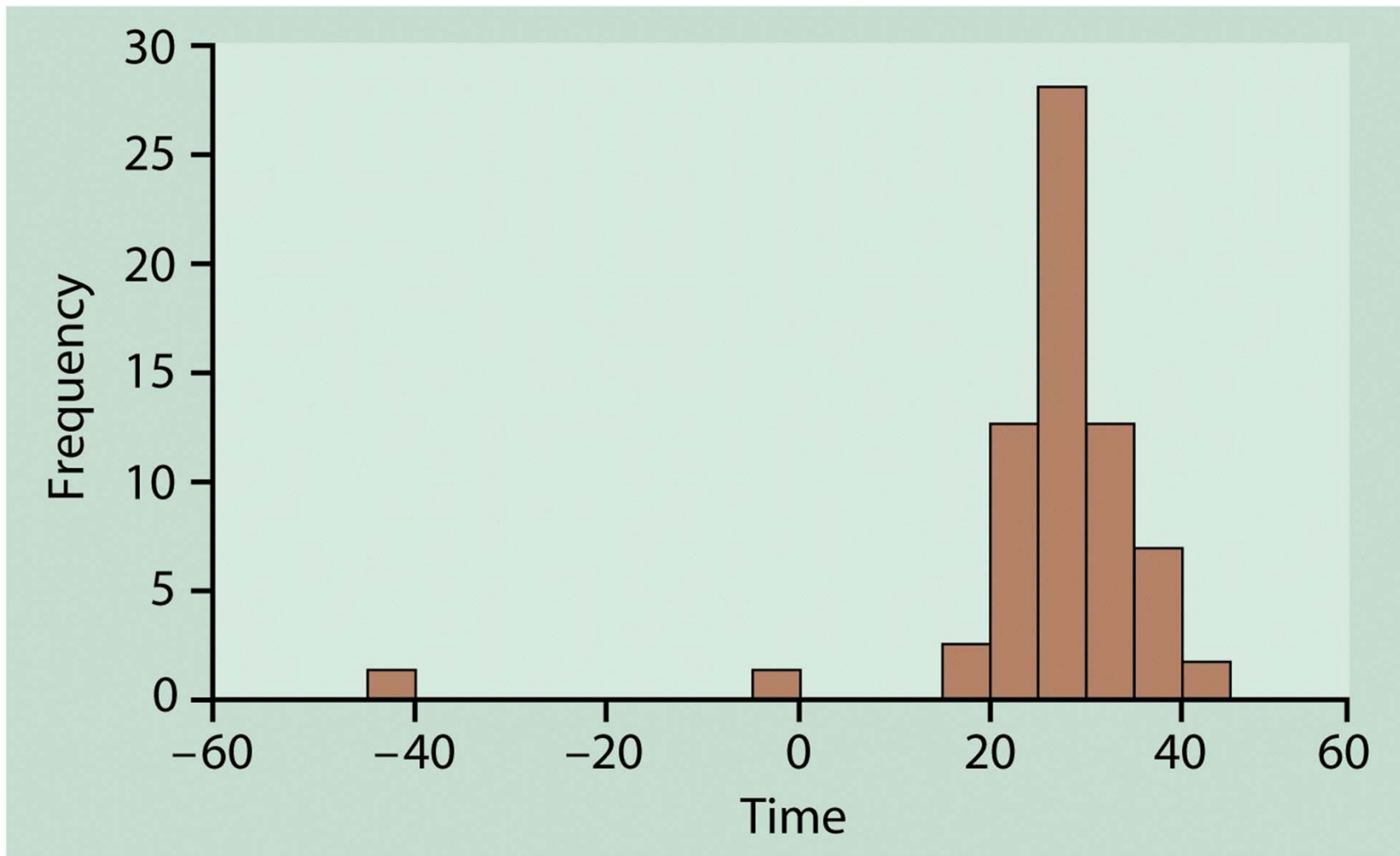
TABLE 1.1 Newcomb's measurements of the passage time of light

| | | | | | |
|-----|----|----|----|----|----|
| 28 | 22 | 36 | 26 | 28 | 28 |
| 26 | 24 | 32 | 30 | 27 | 24 |
| 33 | 21 | 36 | 32 | 31 | 25 |
| 24 | 25 | 28 | 36 | 27 | 32 |
| 34 | 30 | 25 | 26 | 26 | 25 |
| -44 | 23 | 21 | 30 | 33 | 29 |
| 27 | 29 | 28 | 22 | 26 | 27 |
| 16 | 31 | 29 | 36 | 32 | 28 |
| 40 | 19 | 37 | 23 | 32 | 29 |
| -2 | 24 | 25 | 27 | 24 | 16 |
| 29 | 20 | 28 | 27 | 39 | 23 |

- 66 observations taken in July-Sept, 1882
- Variable: passage time, scaled and centered.
- Individual observations are different since the environment of every measurement is slightly different
- We will examine the nature of the variation of the quantitative variable by drawing graphs

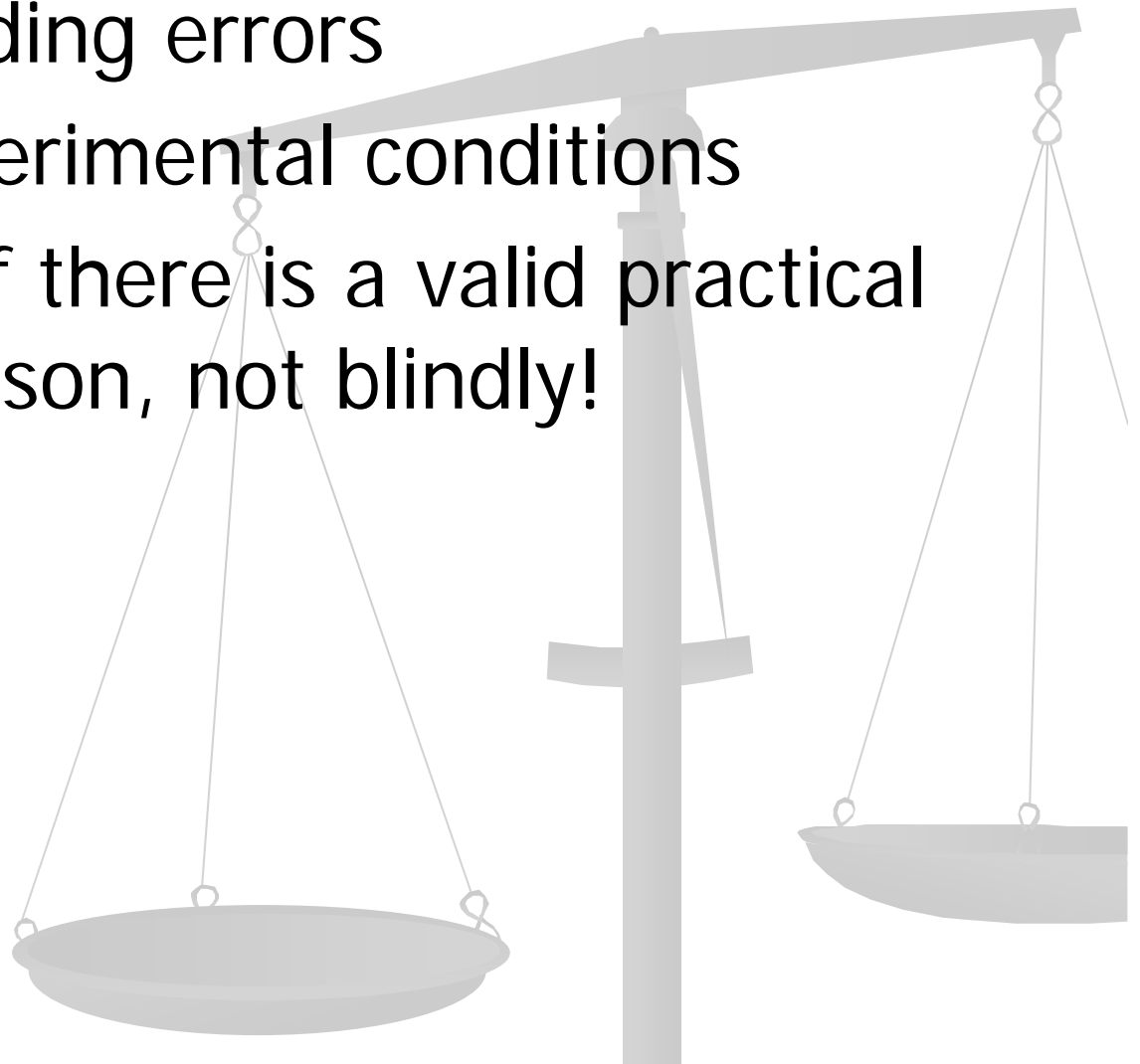


Newcomb's data (dealing with outliers)

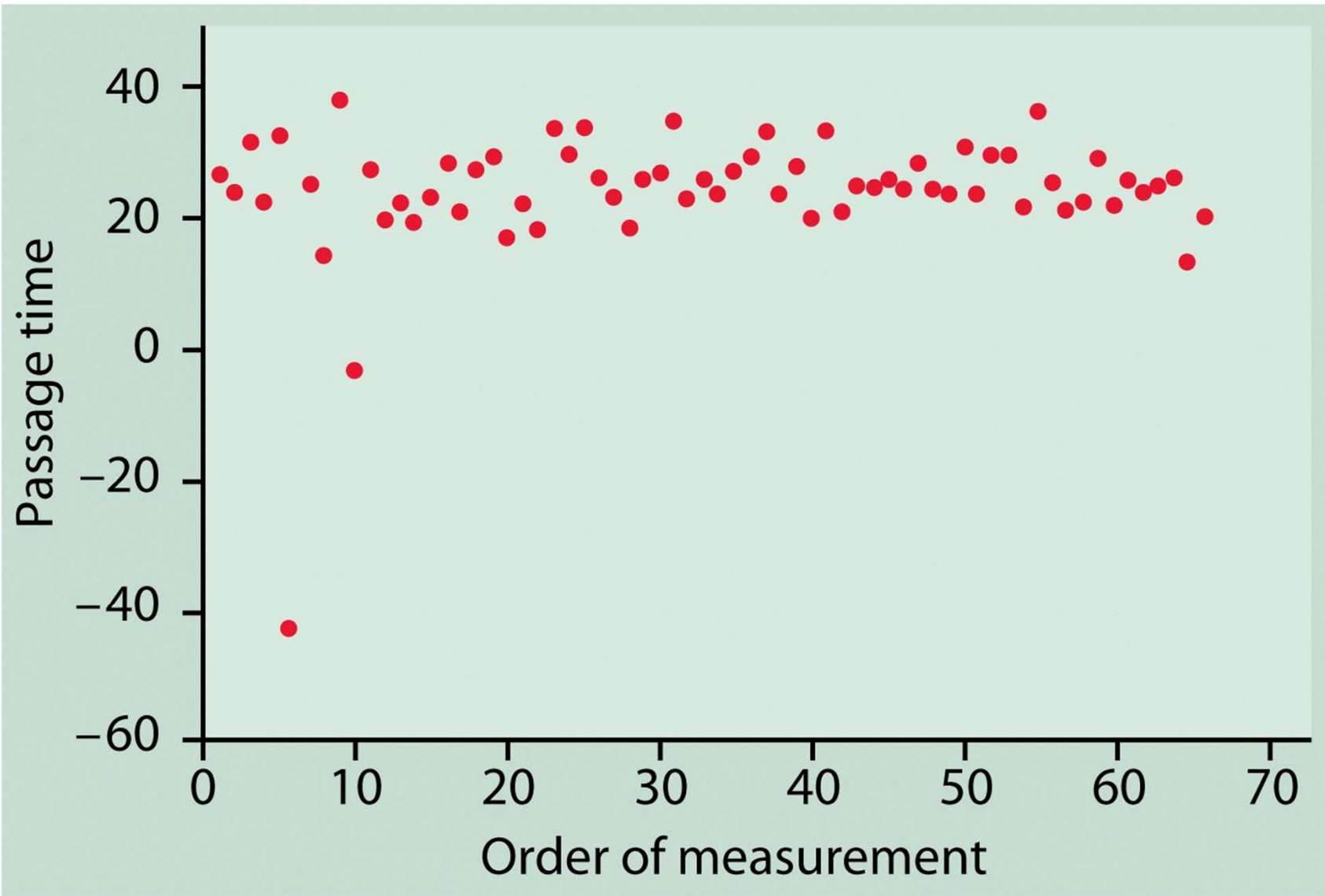


Outliers

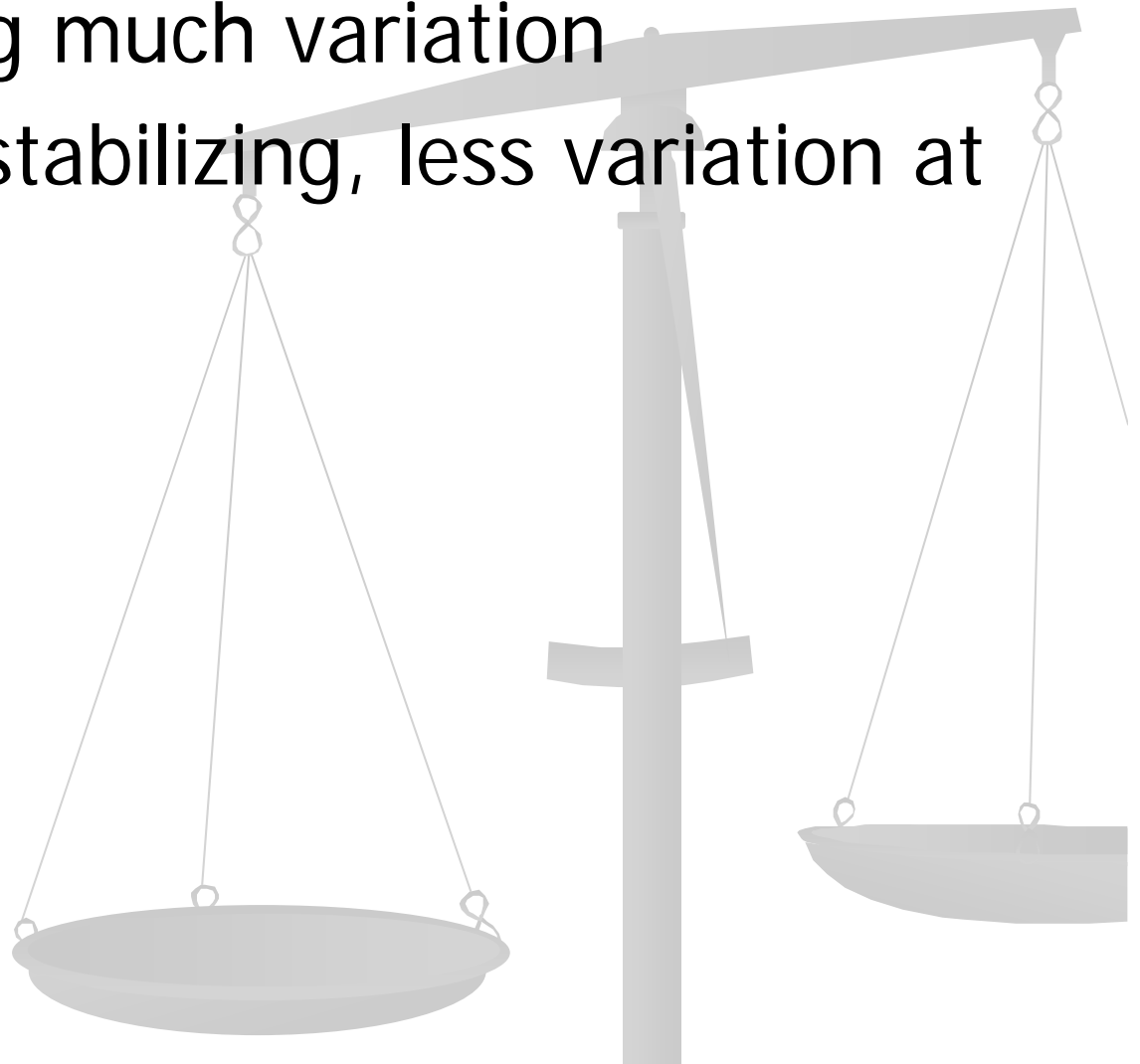
- Check for recording errors
- Violation of experimental conditions
- Discard it only if there is a valid practical or statistical reason, not blindly!



Time plots. Newcomb's data.



- At the beginning much variation
- Measurements stabilizing, less variation at a later time.

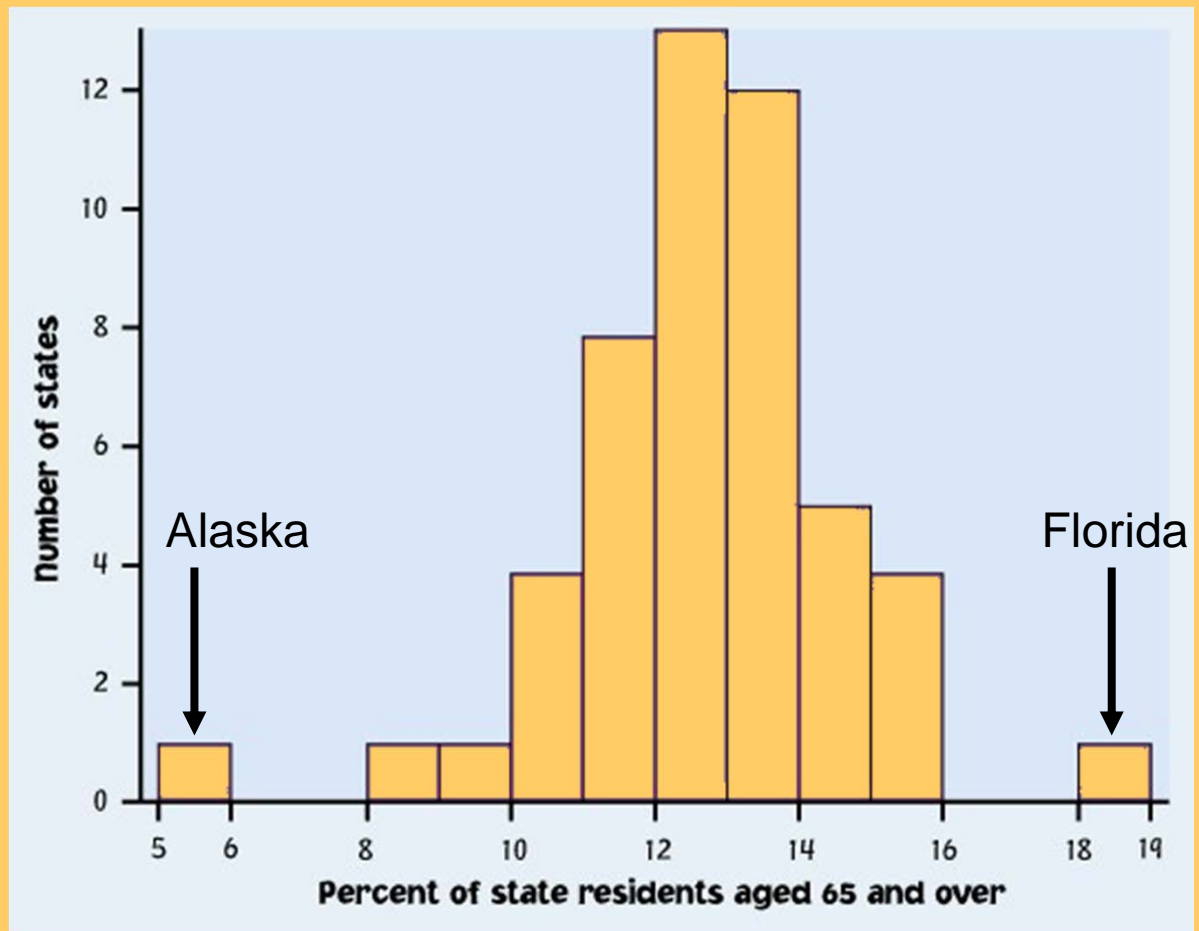


Outliers

An important kind of deviation is an **outlier**. Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

The overall pattern is fairly symmetrical except for 2 states clearly not belonging to the main trend. Alaska and Florida have unusual representation of the elderly in their population.

A large gap in the distribution is typically a sign of an outlier.



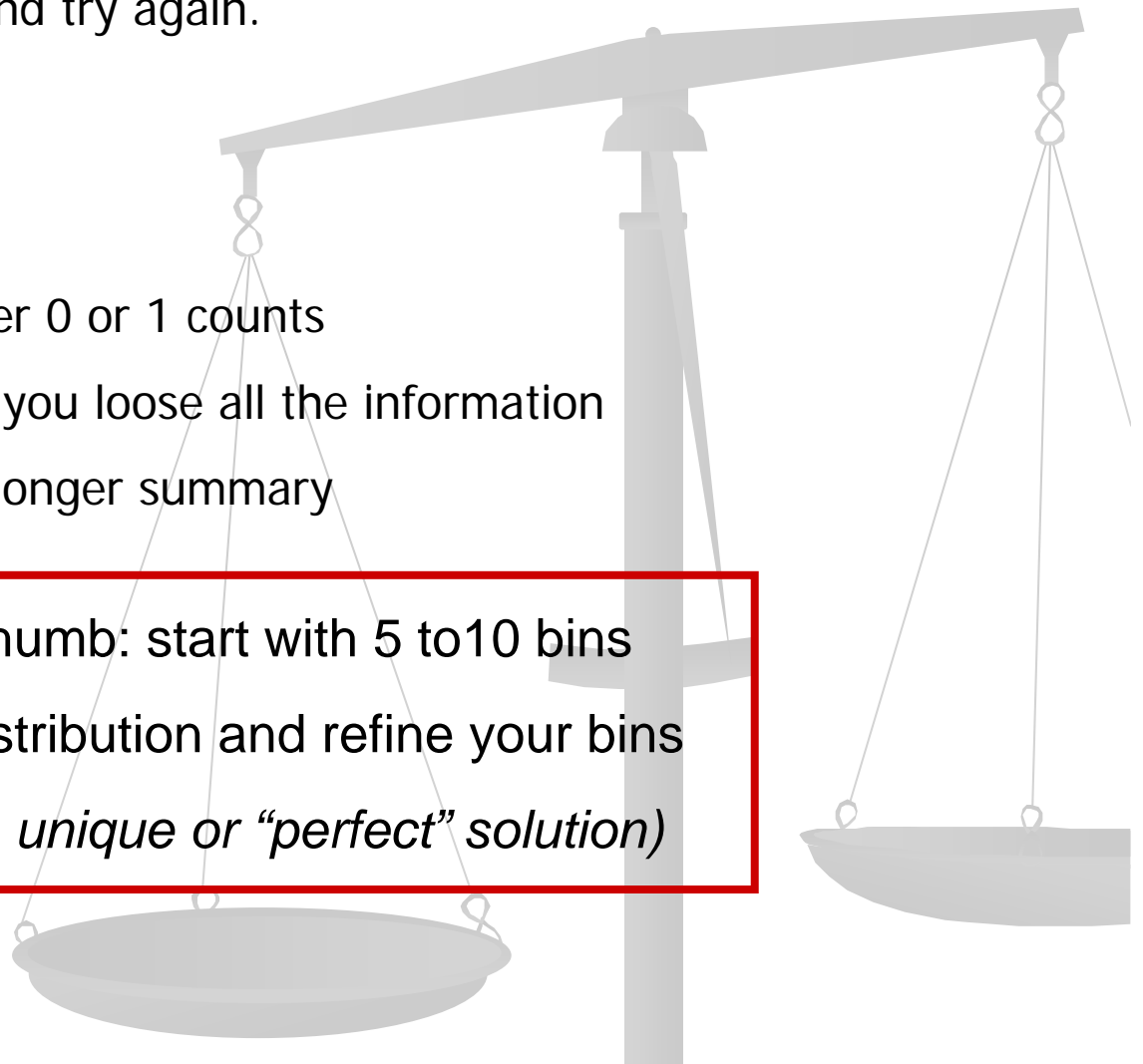
How to create a histogram

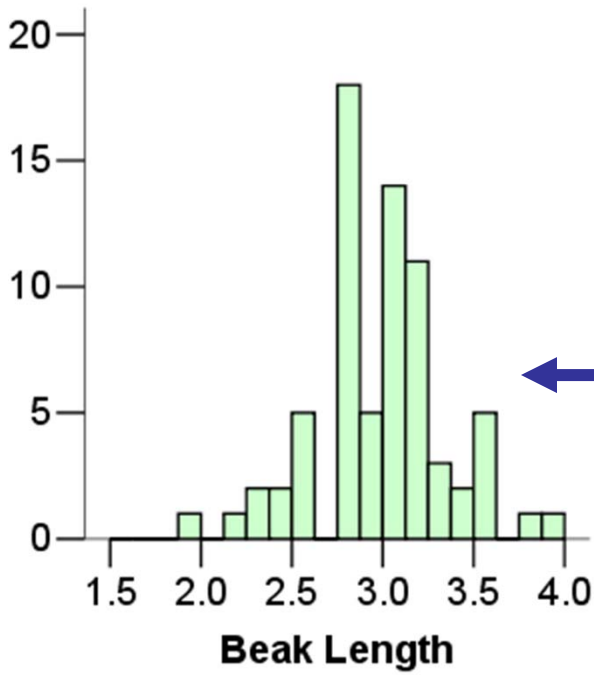
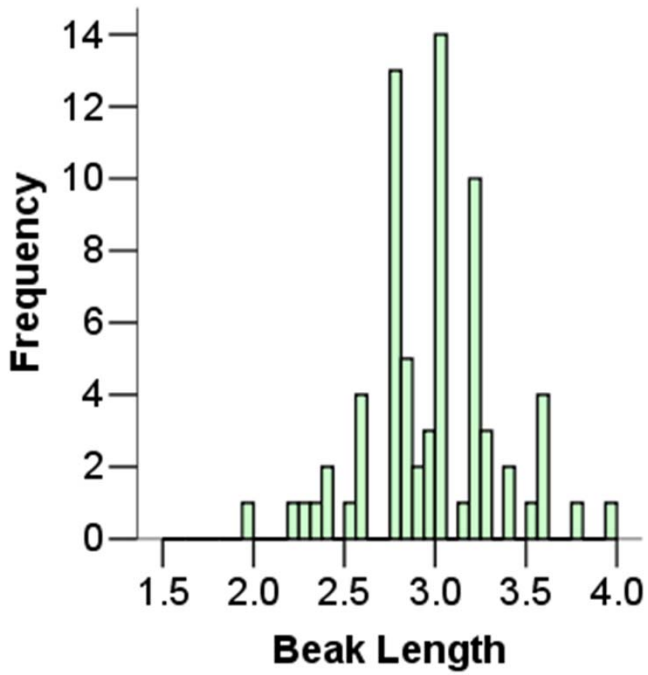
It is an iterative process – try and try again.

What bin size should you use?

- Not too many bins with either 0 or 1 counts
- Not overly summarized that you lose all the information
- Not so detailed that it is no longer a summary

→ rule of thumb: start with 5 to 10 bins
Look at the distribution and refine your bins
(There isn't a unique or "perfect" solution)

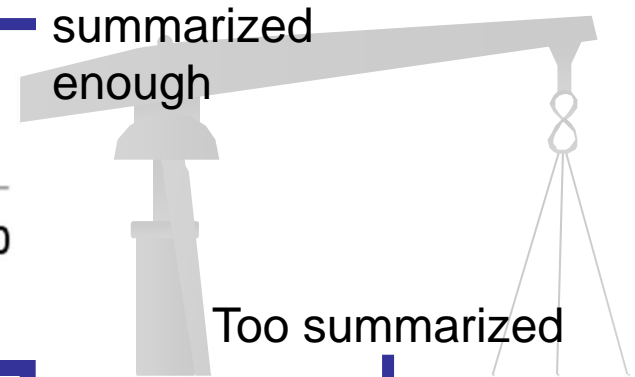




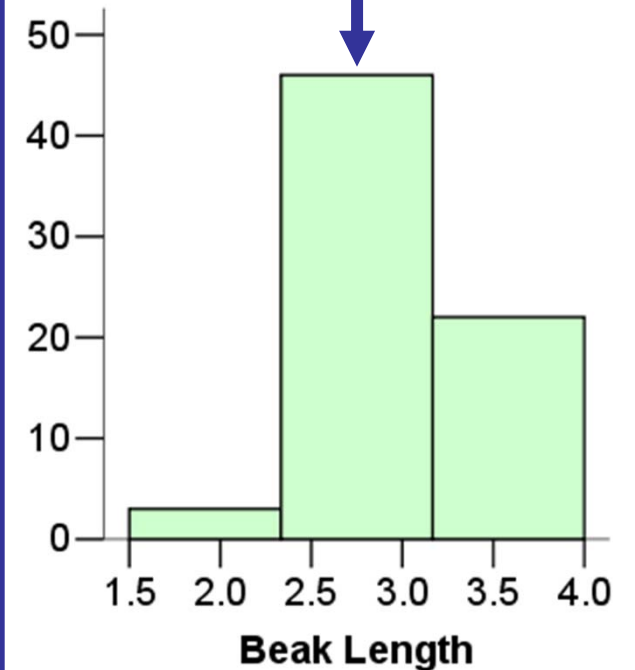
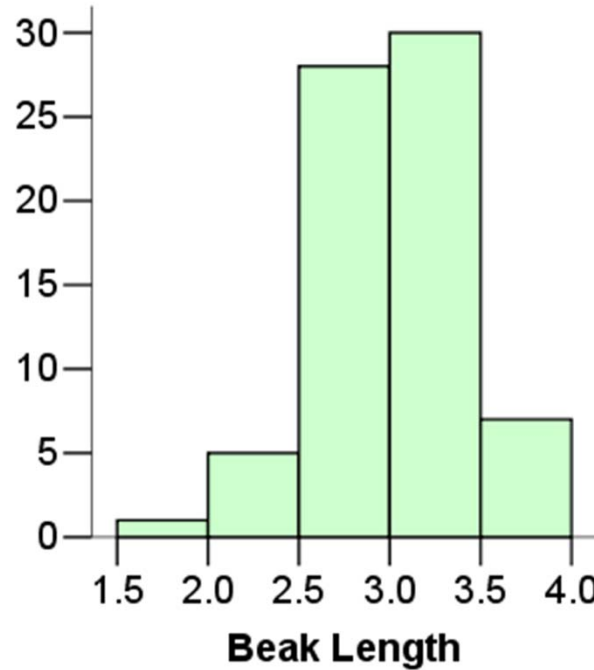
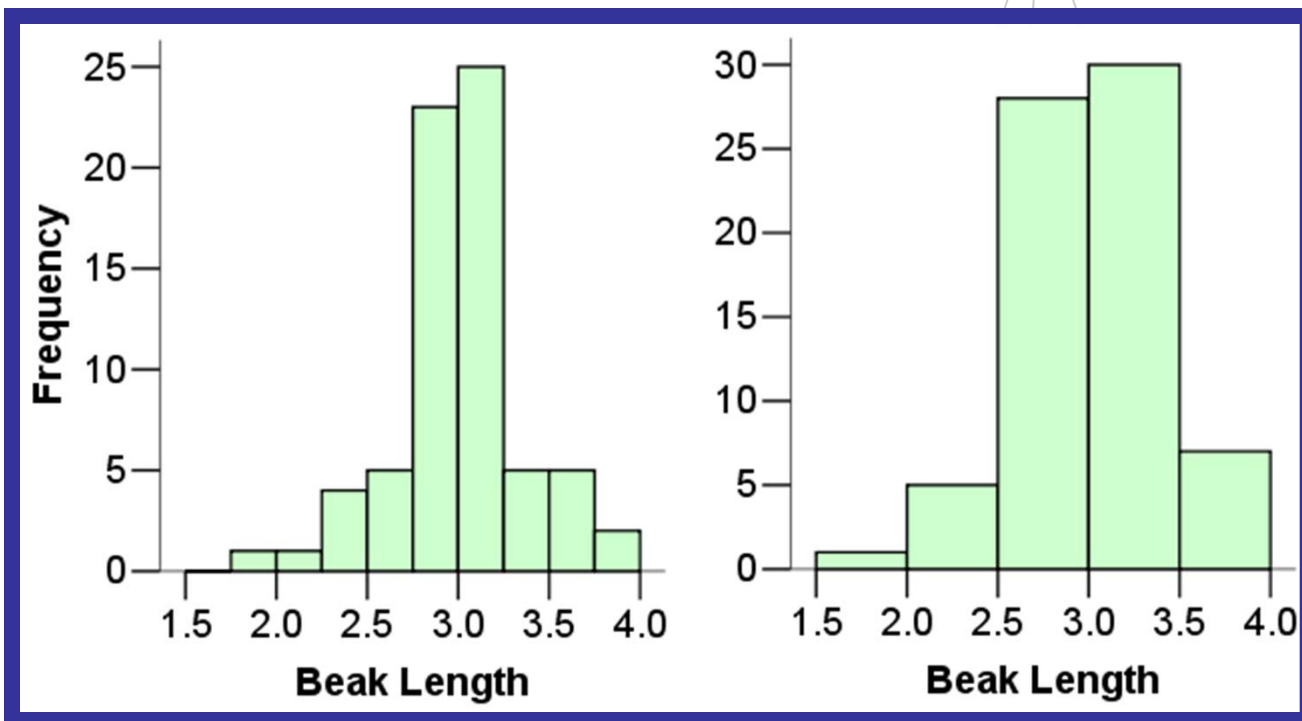
Same data set



Not summarized enough



Too summarized

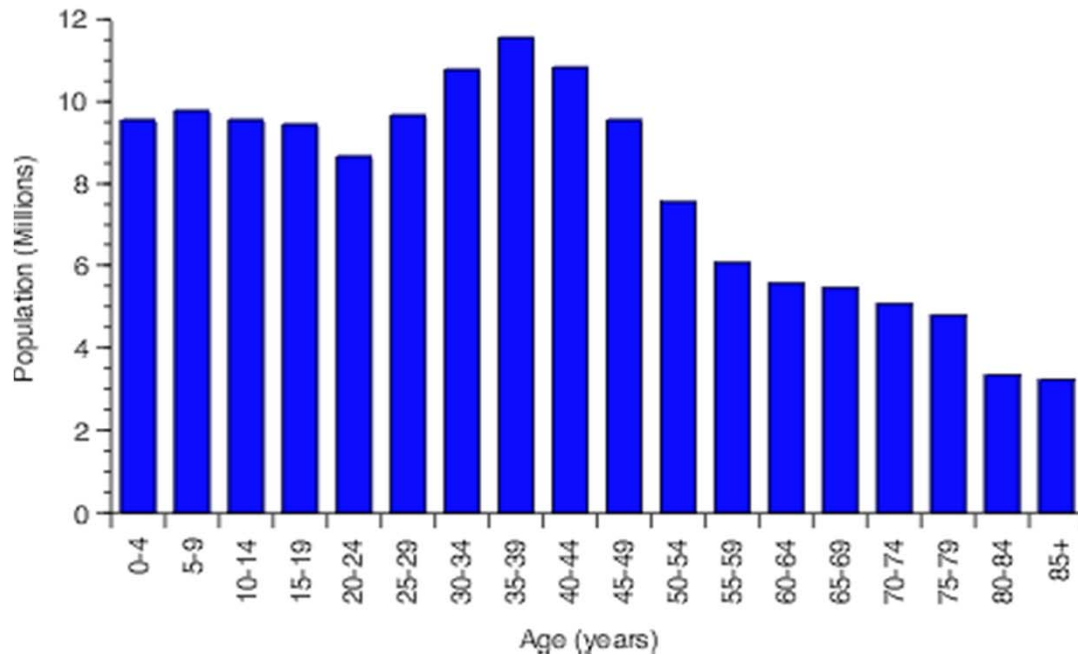


IMPORTANT NOTE:

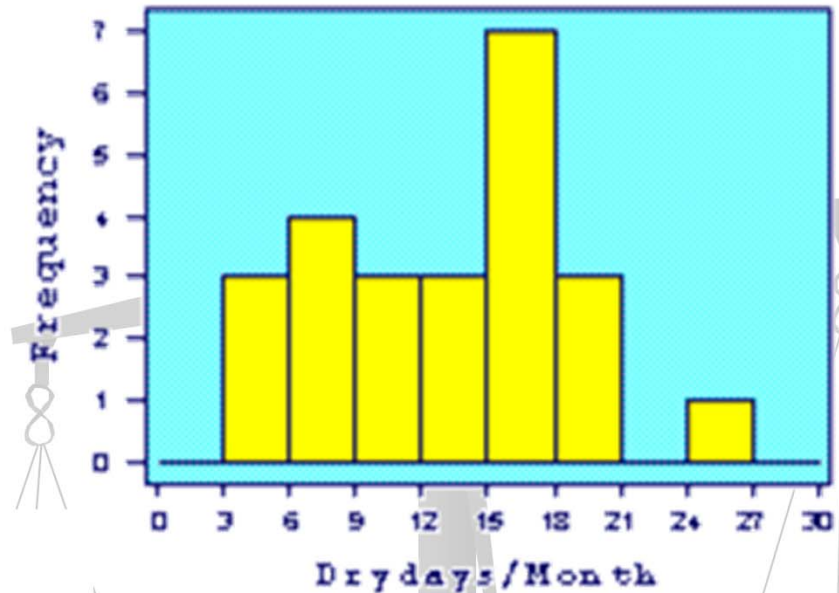
Your data are the way they are.

Do not try to force them into a particular shape.

United States Female Population - 1997



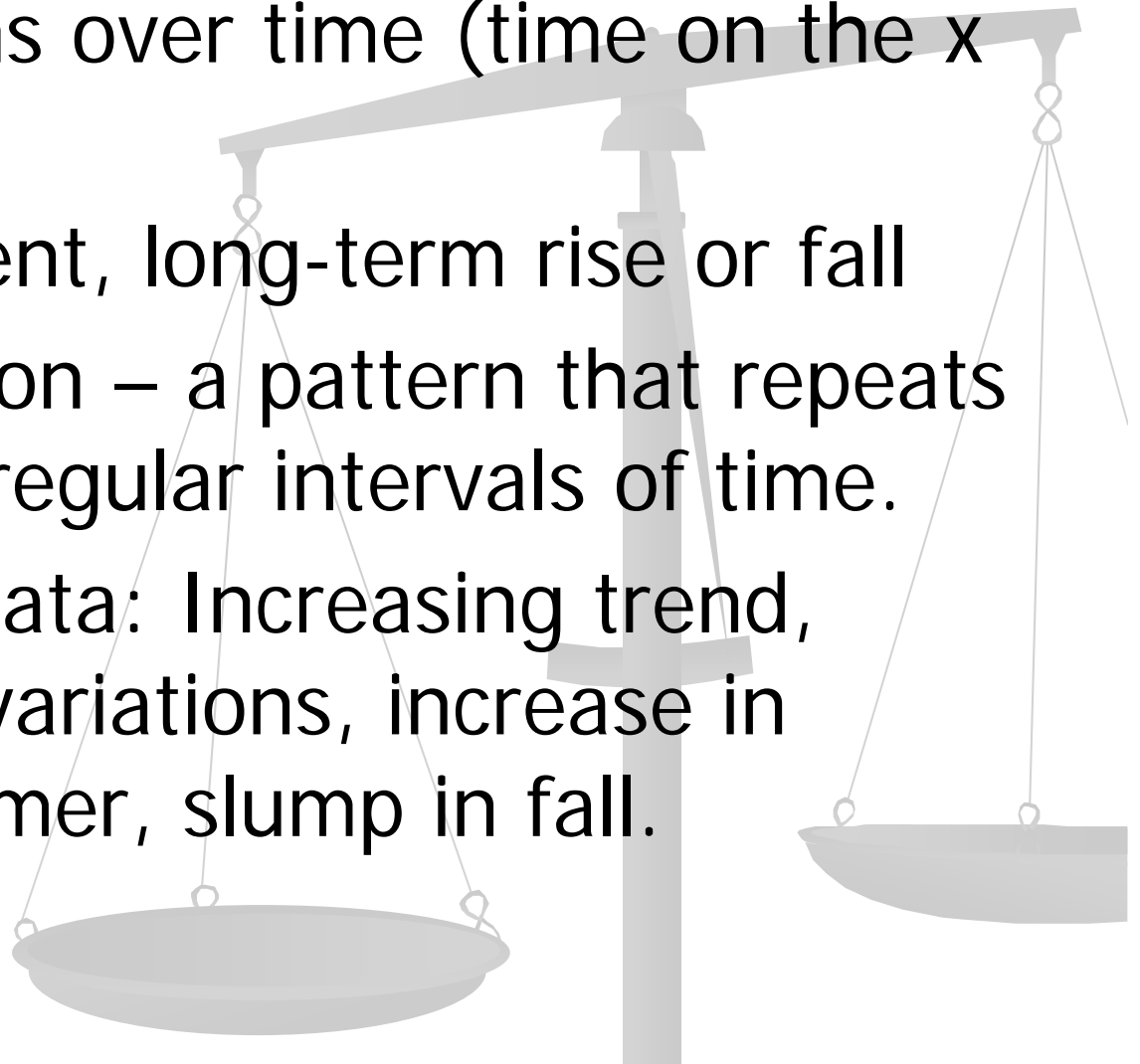
Histogram of Drydays in 1995



It is a common misconception that if you have a large enough data set, the data will eventually turn out nice and symmetrical.

Time series

- Plot observations over time (time on the x axis)
- Trend – persistent, long-term rise or fall
- Seasonal variation – a pattern that repeats itself at known regular intervals of time.
- Gasoline price data: Increasing trend, small seasonal variations, increase in spring and summer, slump in fall.



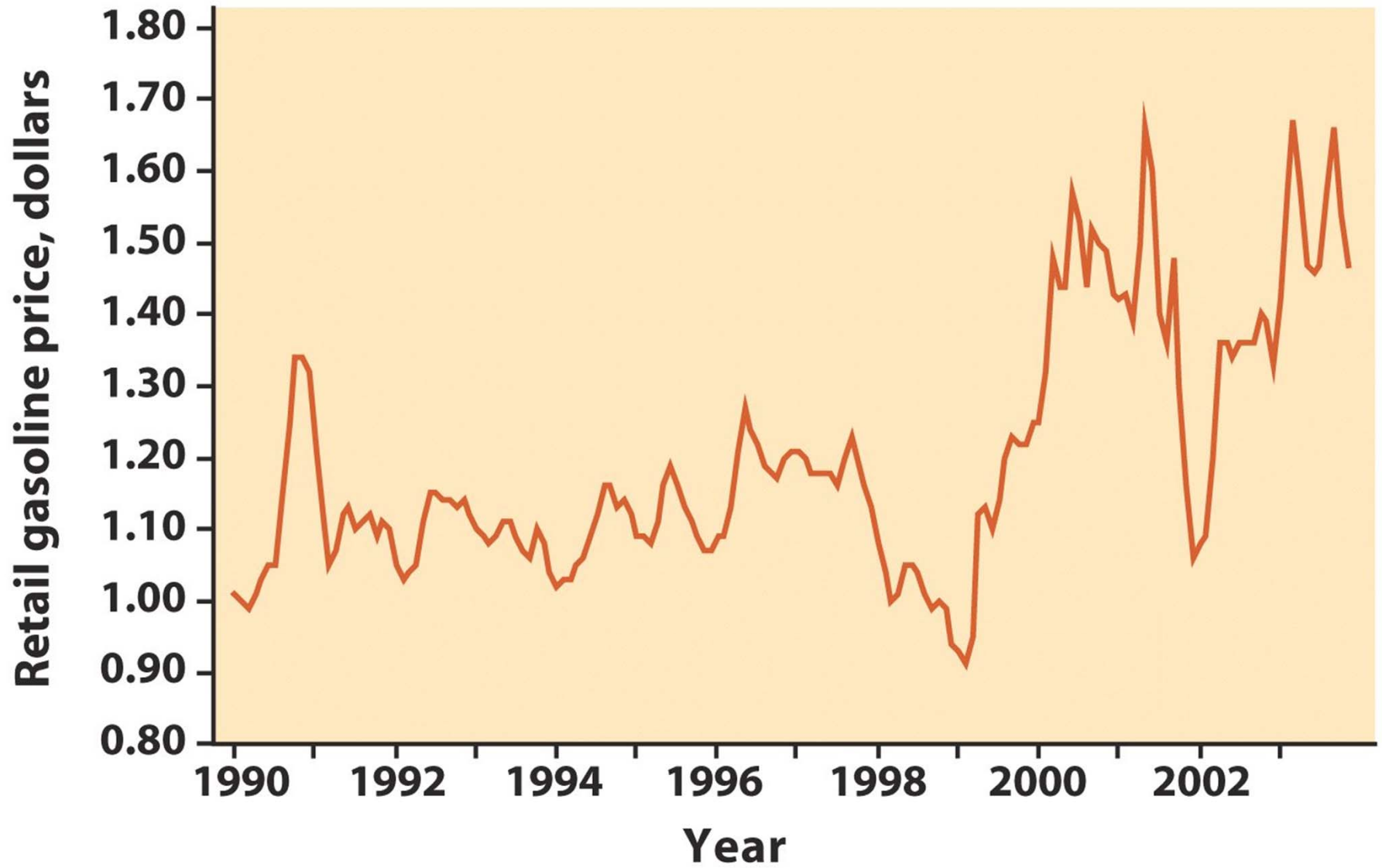


Figure 1-9
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

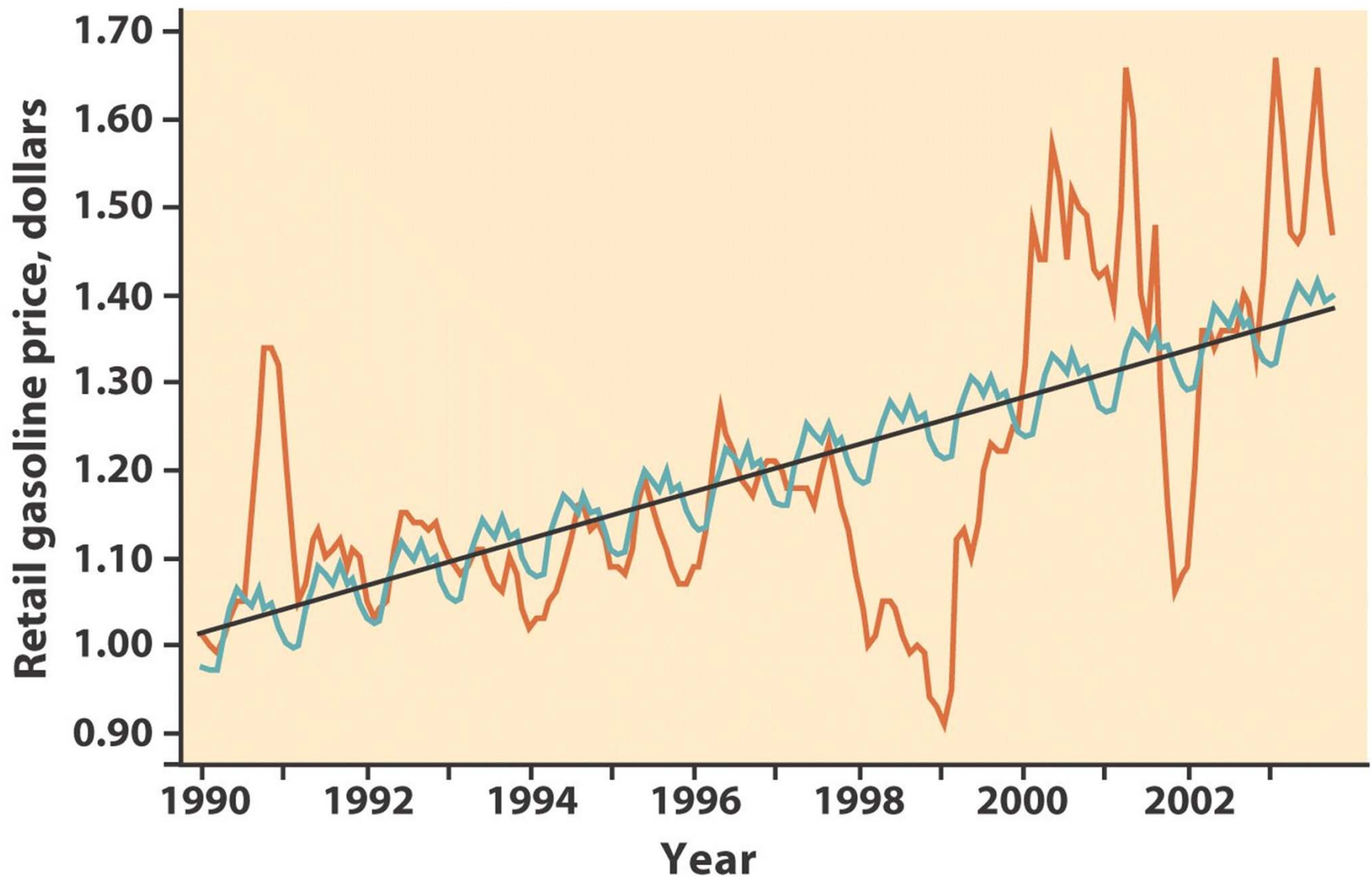


Figure 1-10
Introduction to the Practice of Statistics, Fifth Edition
 © 2005 W. H. Freeman and Company

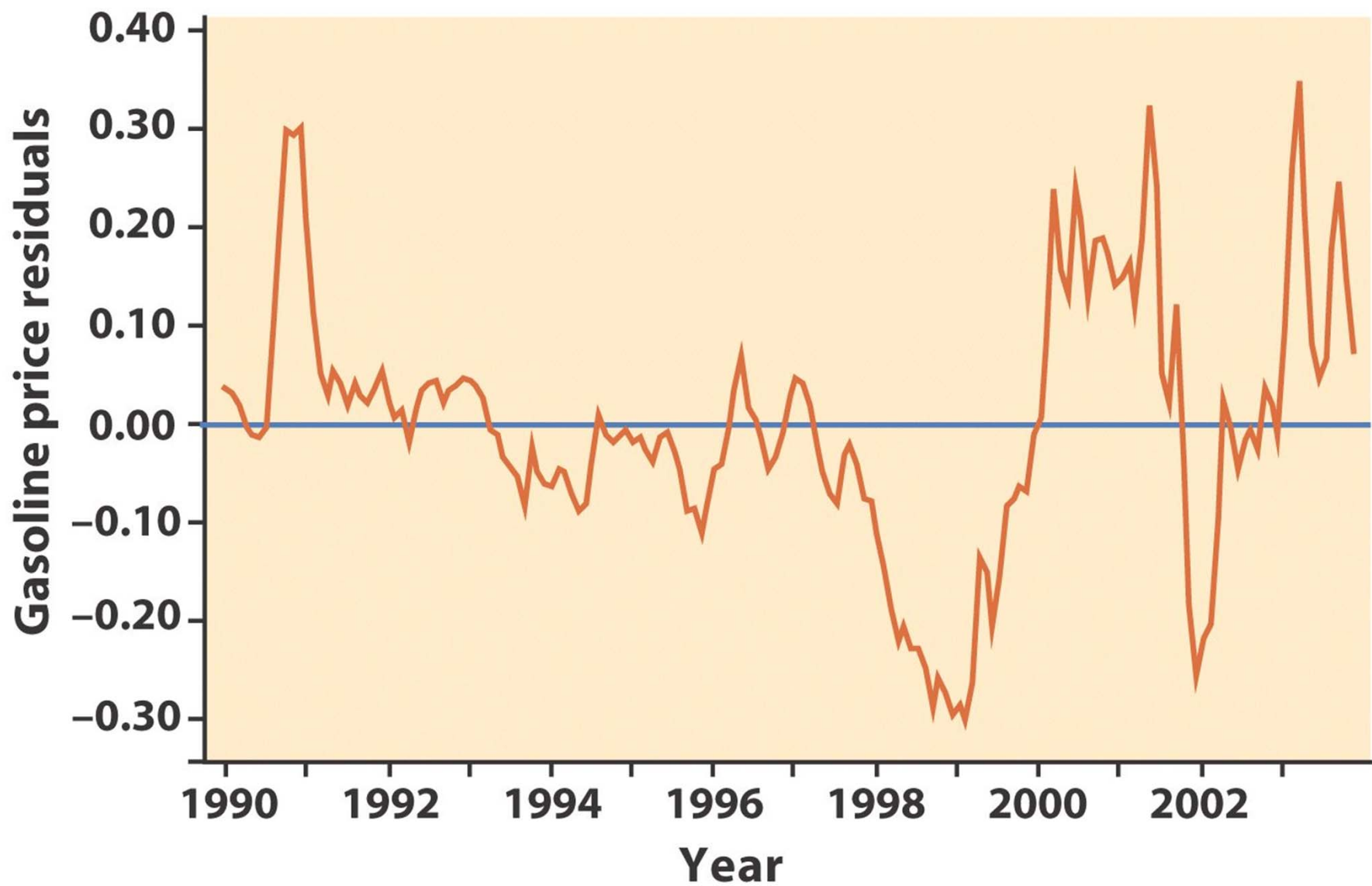
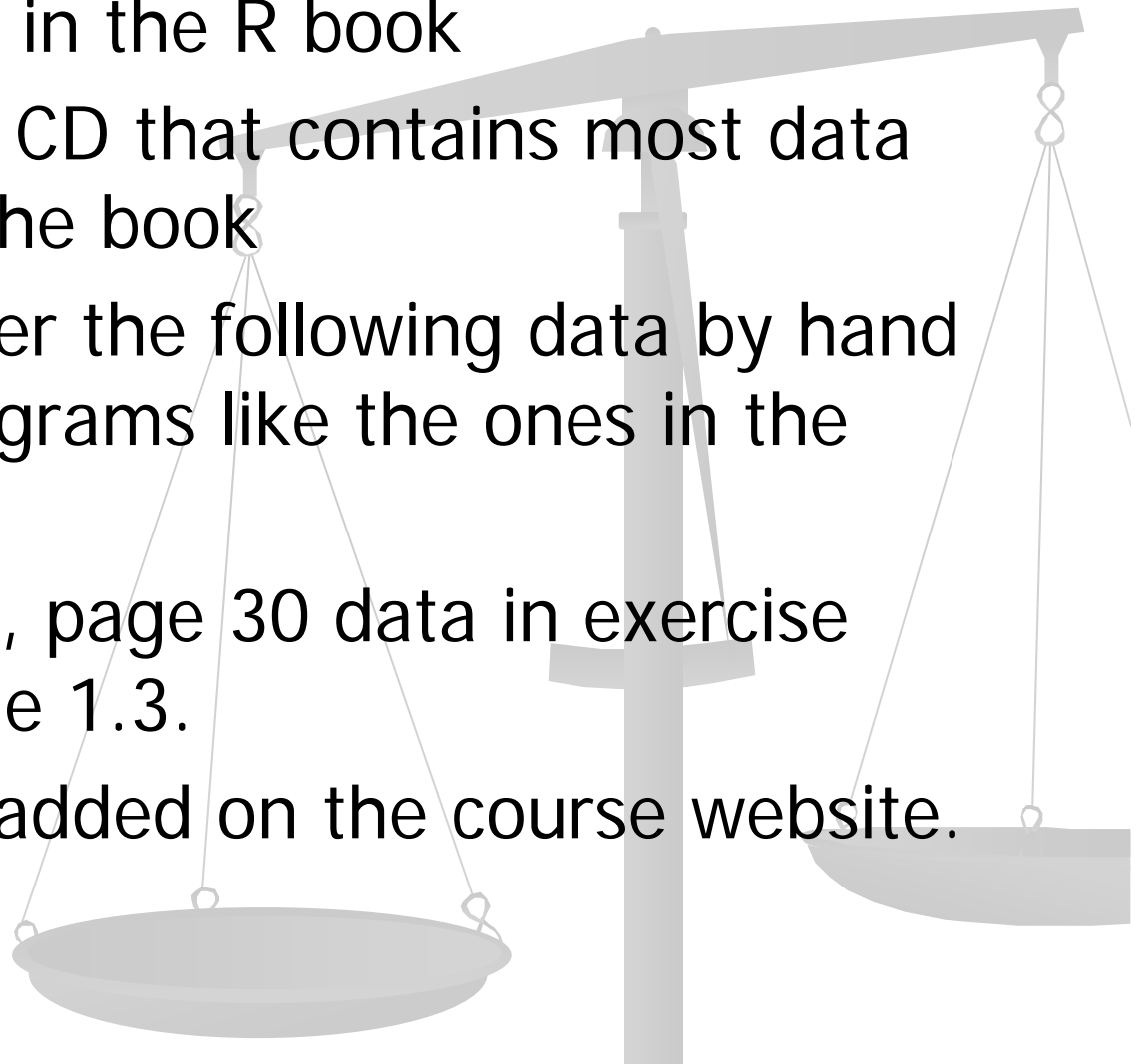


Figure 1-11
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

Exercises: Learn to input data in R

- Read pages 39-44 in the R book
- The manual has a CD that contains most data used throughout the book
- Use the CD or enter the following data by hand and produce histograms like the ones in the book
- Page 15 Table 1.2, page 30 data in exercise 1.26, page 31 table 1.3.
- More data will be added on the course website.



Summary

- Categorical and Quantitative variable
- Graphical tools for categorical variable
Bar Chart, Pie Chart
- For quantitative variable:
Stem and leaf plot, histogram
- Describe: Shape, center, spread
- Watch out for patterns and deviations
from patterns.

