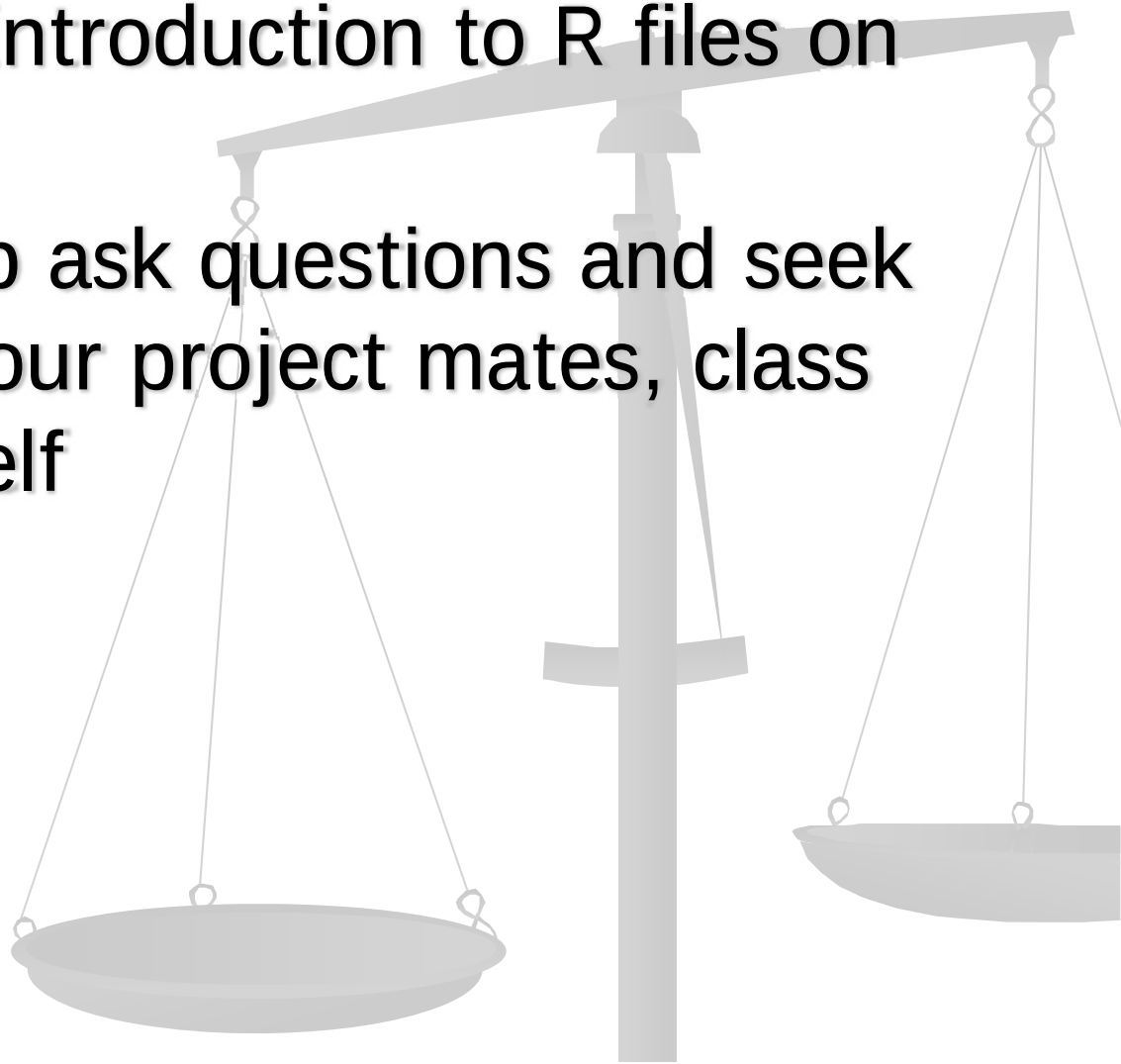
MA 222 Statistics - lecture 1

Spring 2008

R

- Please see the Introduction to R files on the website.
- If you need help ask questions and seek answers from your project mates, class mates and myself



Data, Data, Data, all around us !

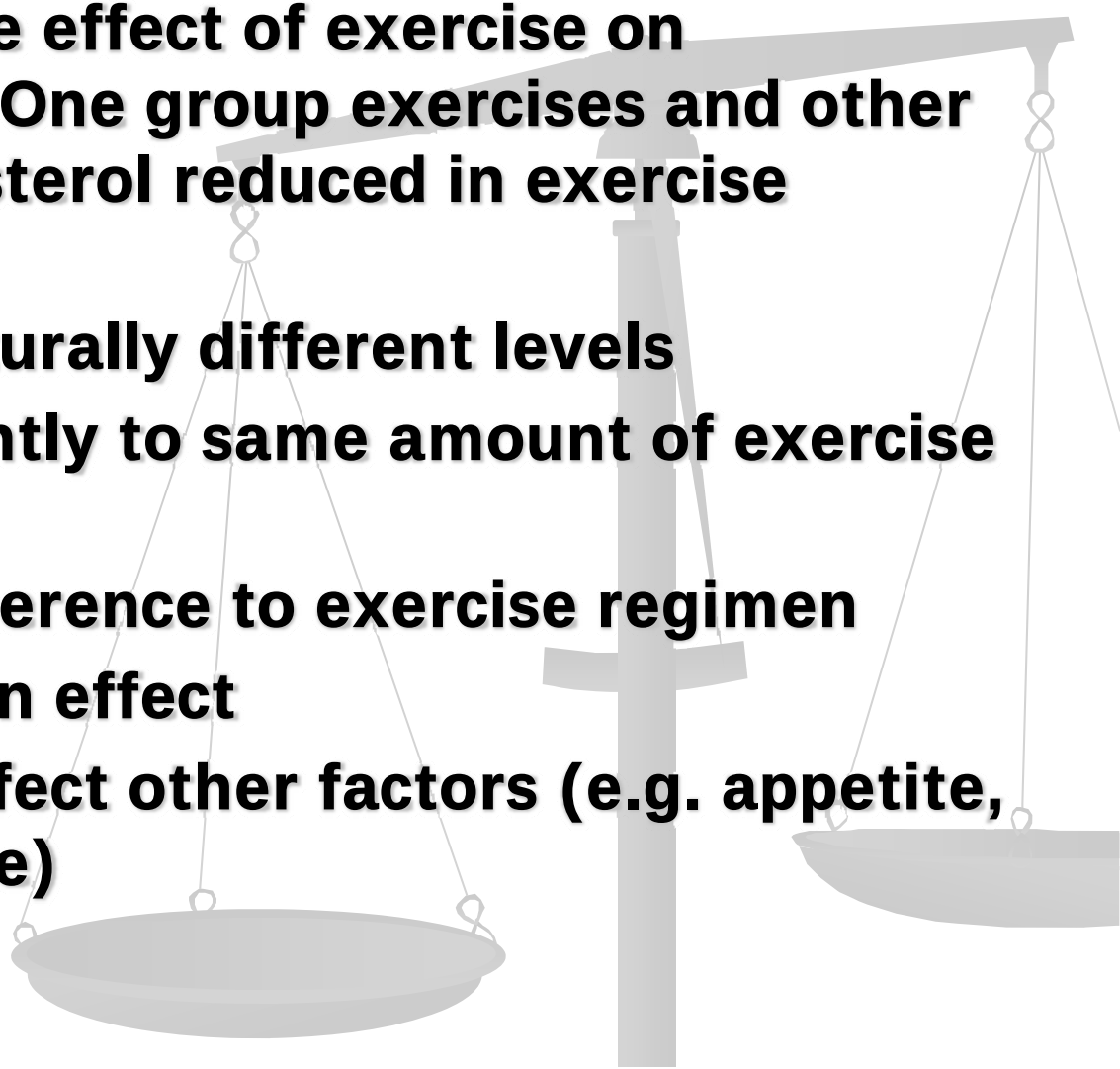
- We have already learned about types of random variables.
- In practice we do not observe variables instead we see data:

	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						

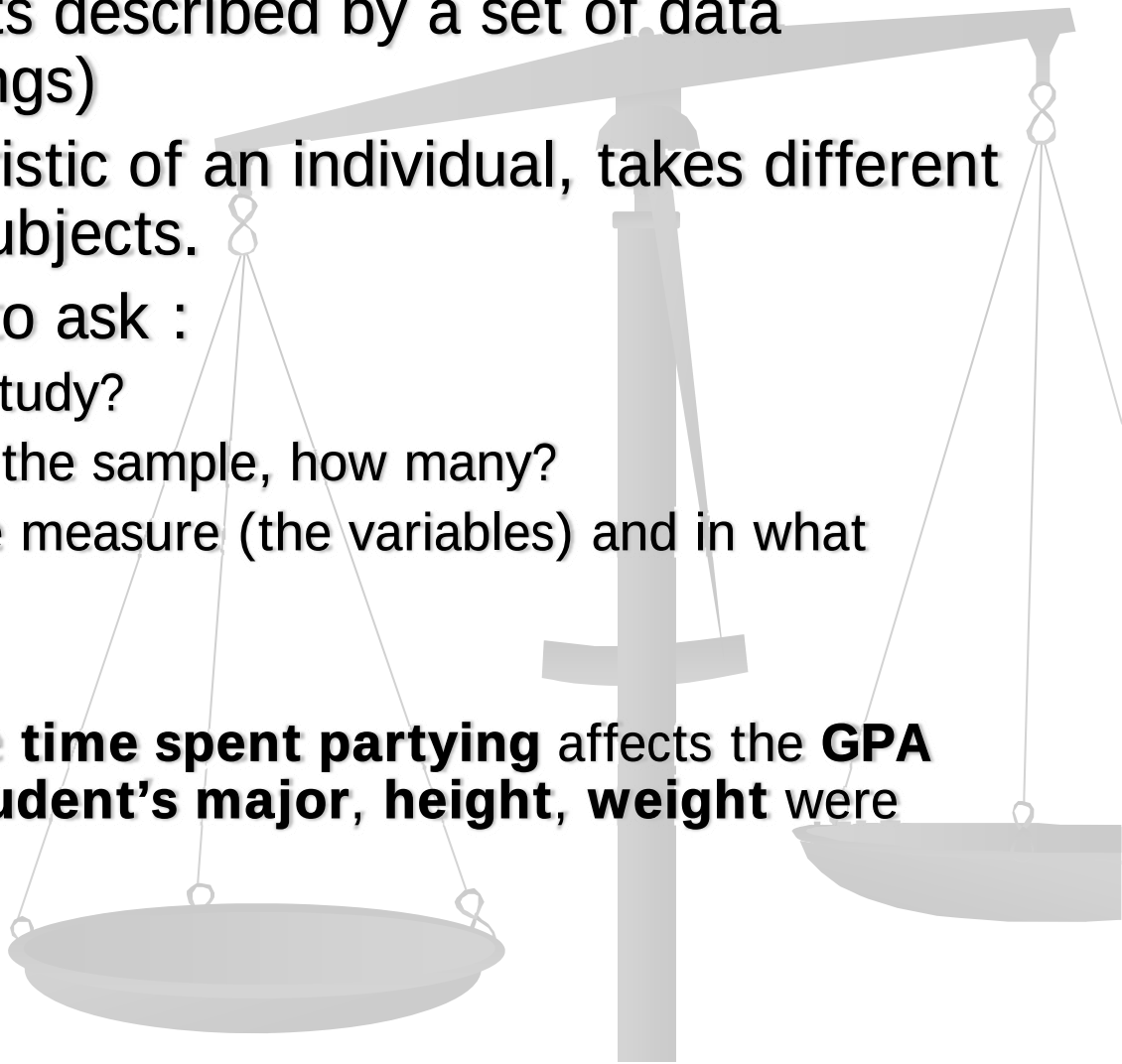
Ready NUM

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)**



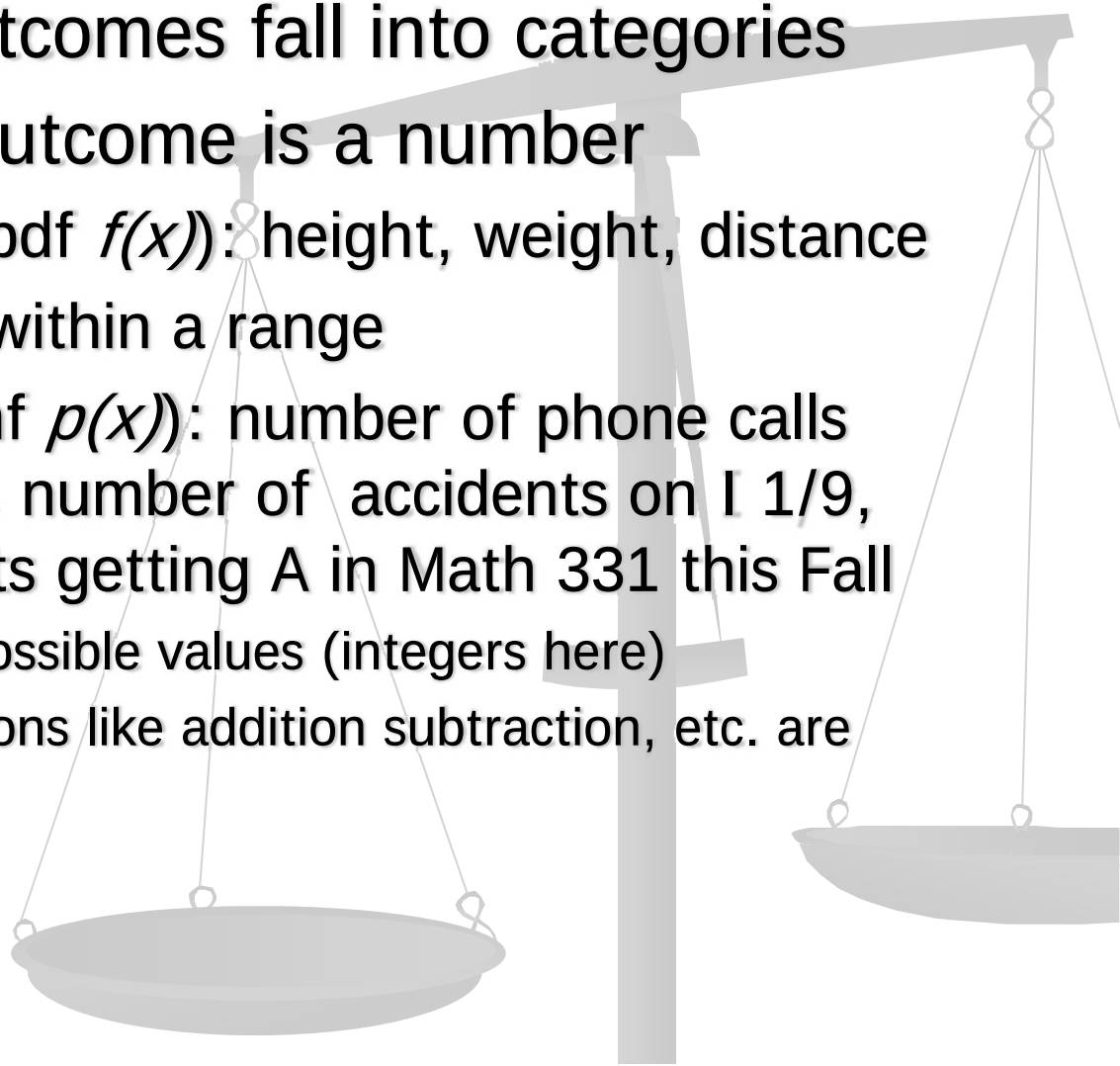
- **Individuals** – objects described by a set of data (people, animals, things)
- **Variable** – characteristic of an individual, takes different values for different subjects.
- The three questions to ask :
  - Why: Purpose of study?
  - Who: Members of the sample, how many?
  - What: What did we measure (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* (with pdf *f(x)*): height, weight, distance

  Can take any value within a range
  - *Discrete* (with pmf *p(x)*): 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

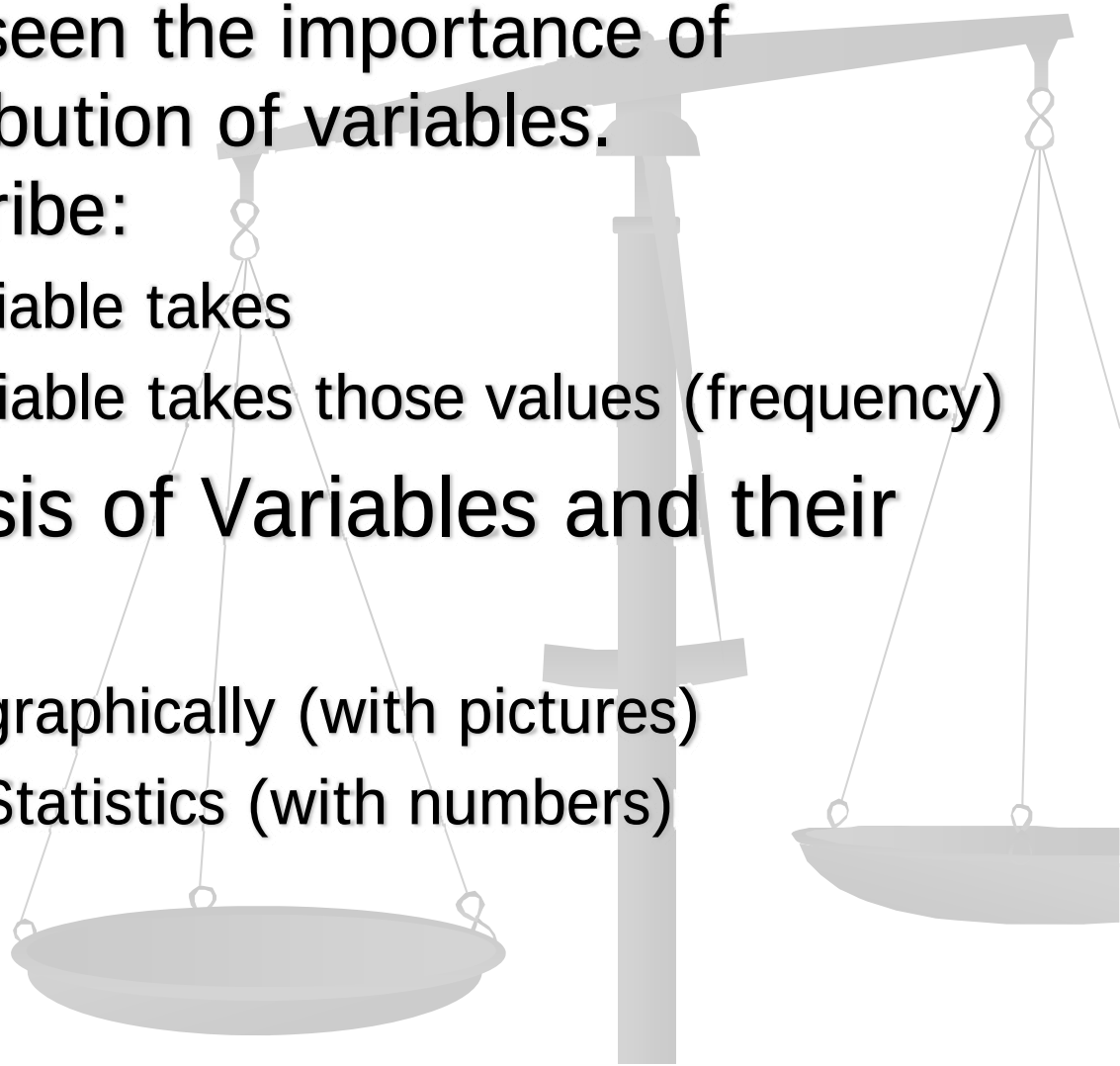


Distribution of a variable.

- We already have seen the importance of knowing the distribution of variables. Distributions describe:
  - What values a variable takes
  - How often the variable takes those values (frequency)

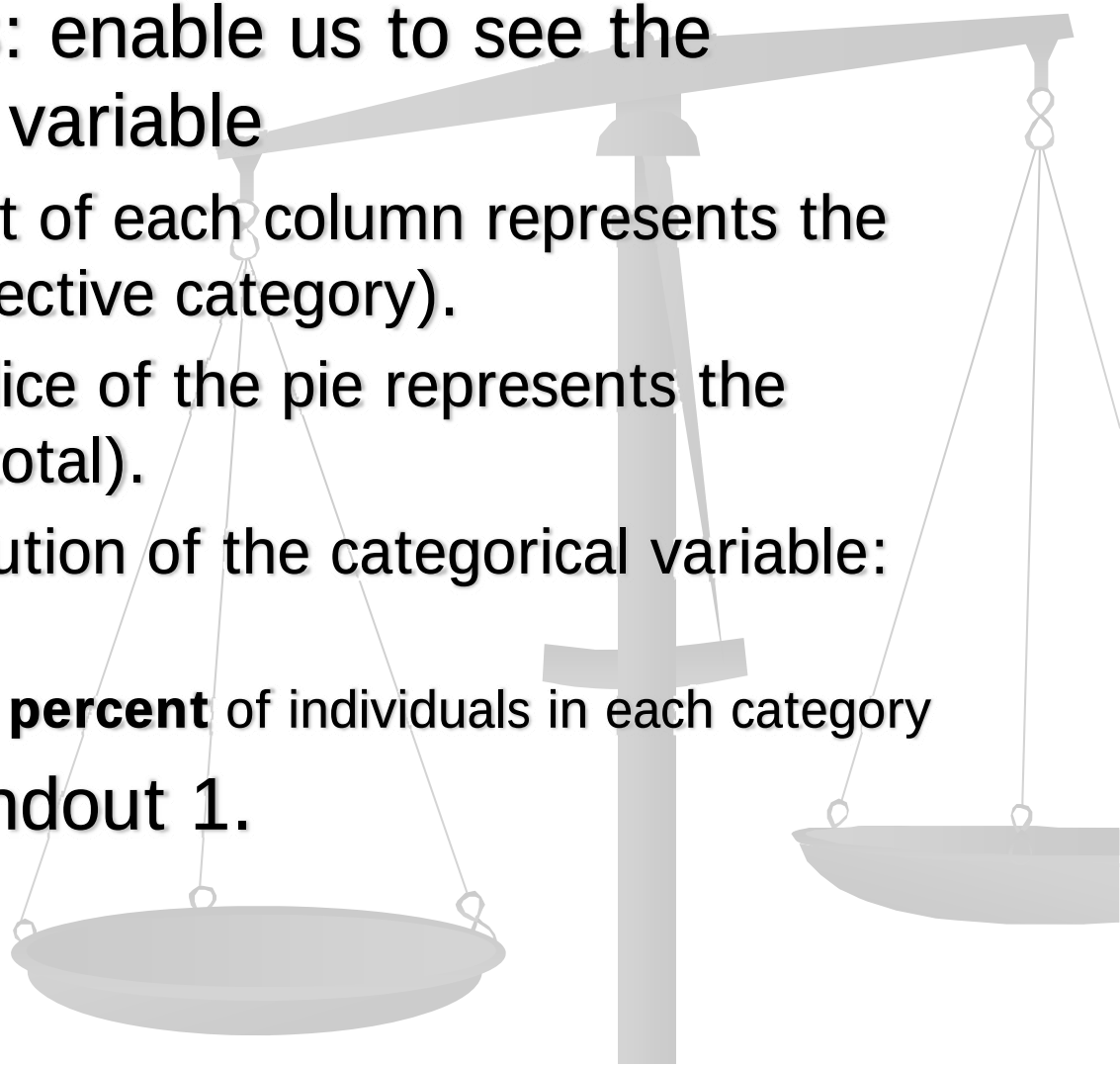
Preliminary Analysis of Variables and their distributions:

- 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
- Please see the handout 1.



Bar Graph

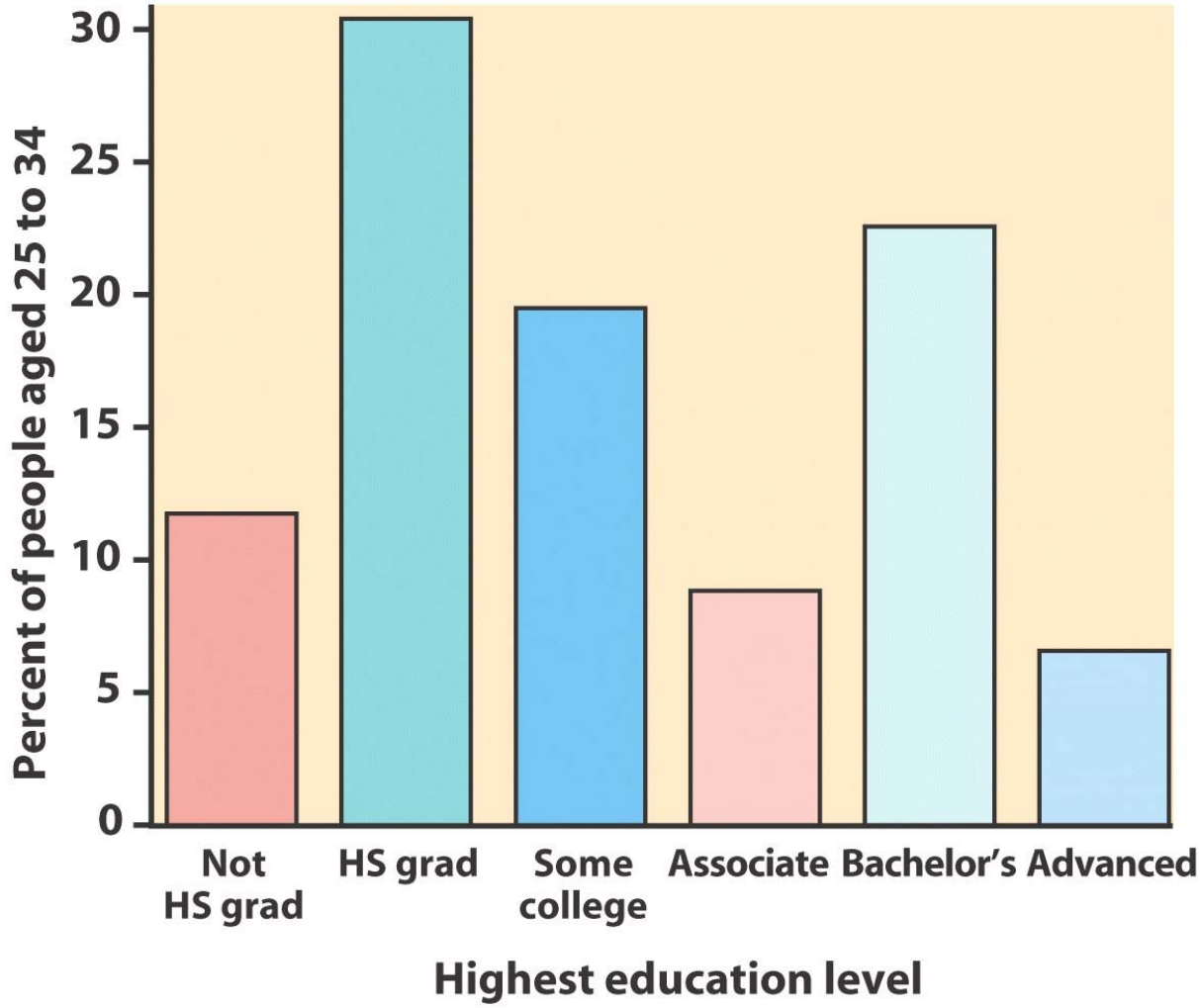


Figure 1-1a
Introduction to the Practice of Statistics, Fifth Edition

Pie Chart

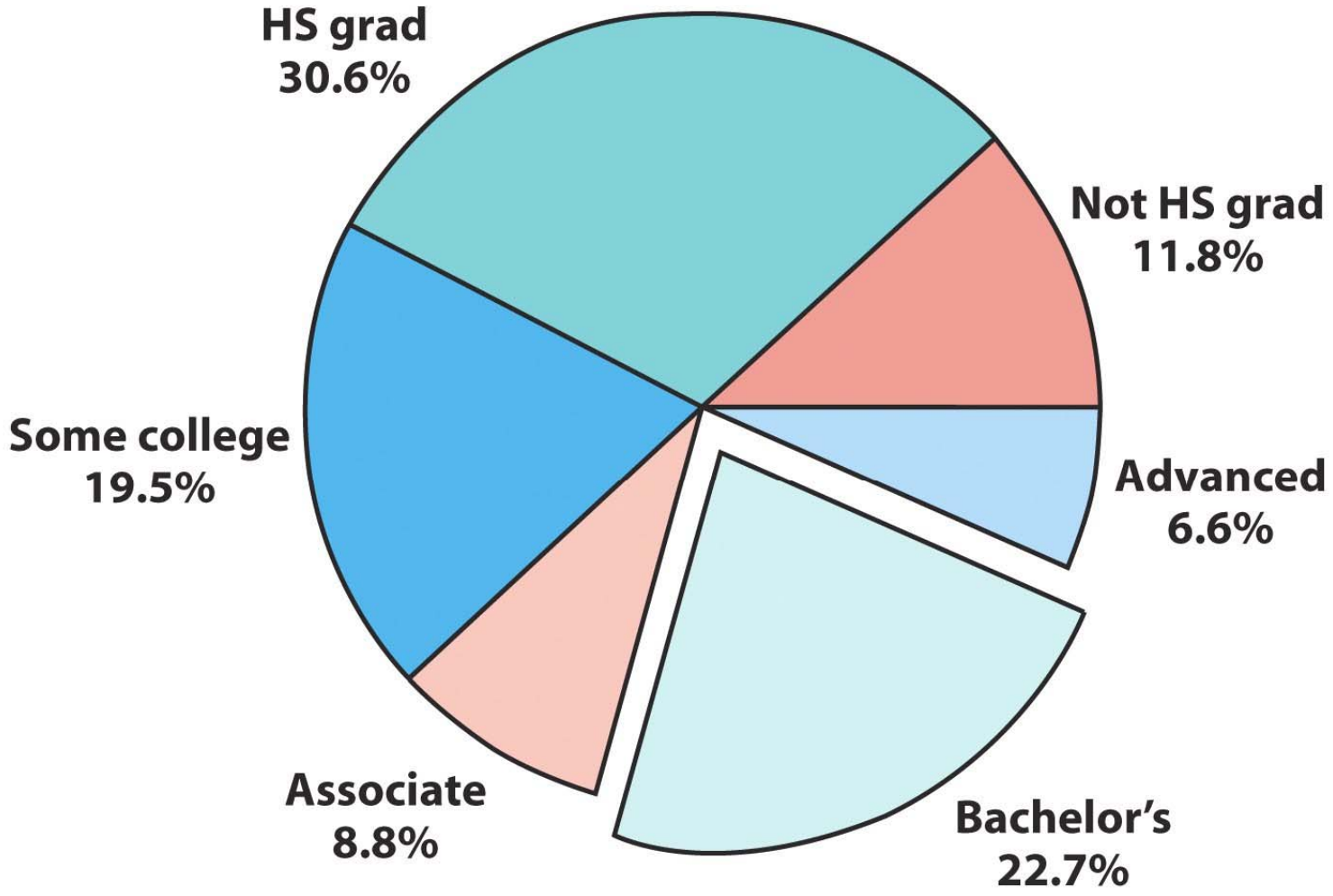


Figure 1-1b
Introduction to the Practice of Statistics, Fifth Edition

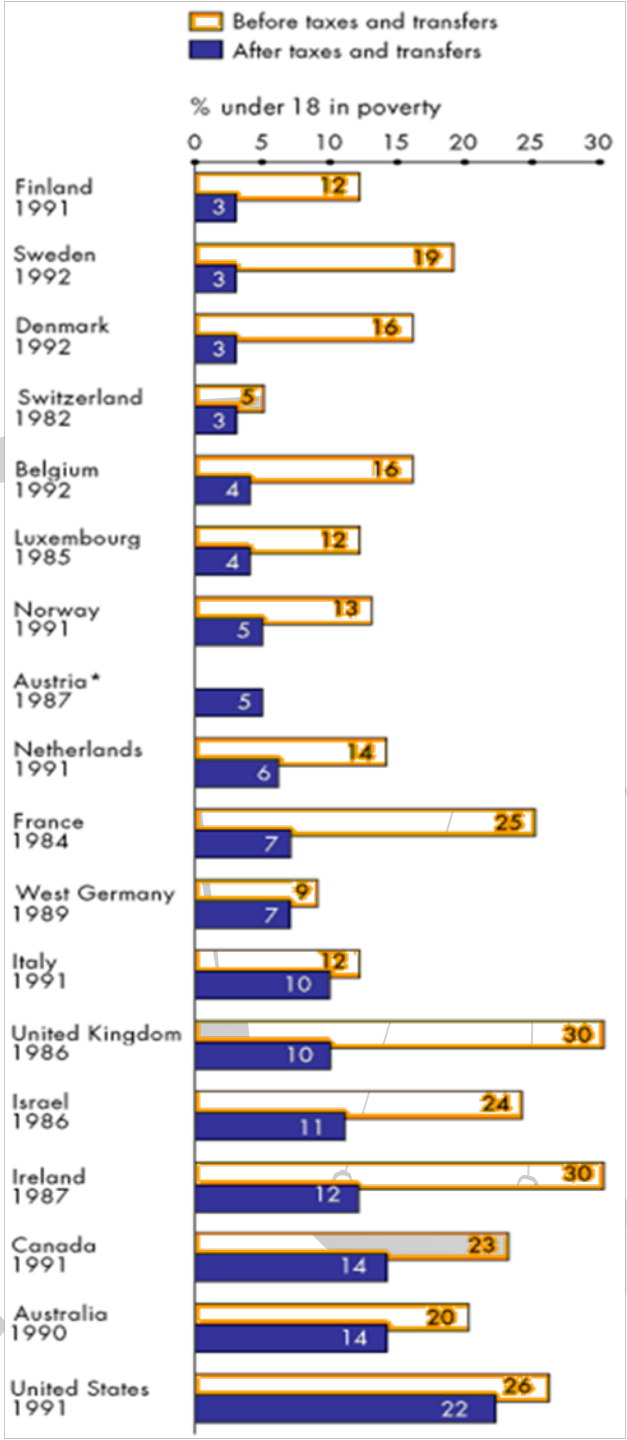
EXAMPLE - Child poverty before and after government intervention—UNICEF, 1996

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:**

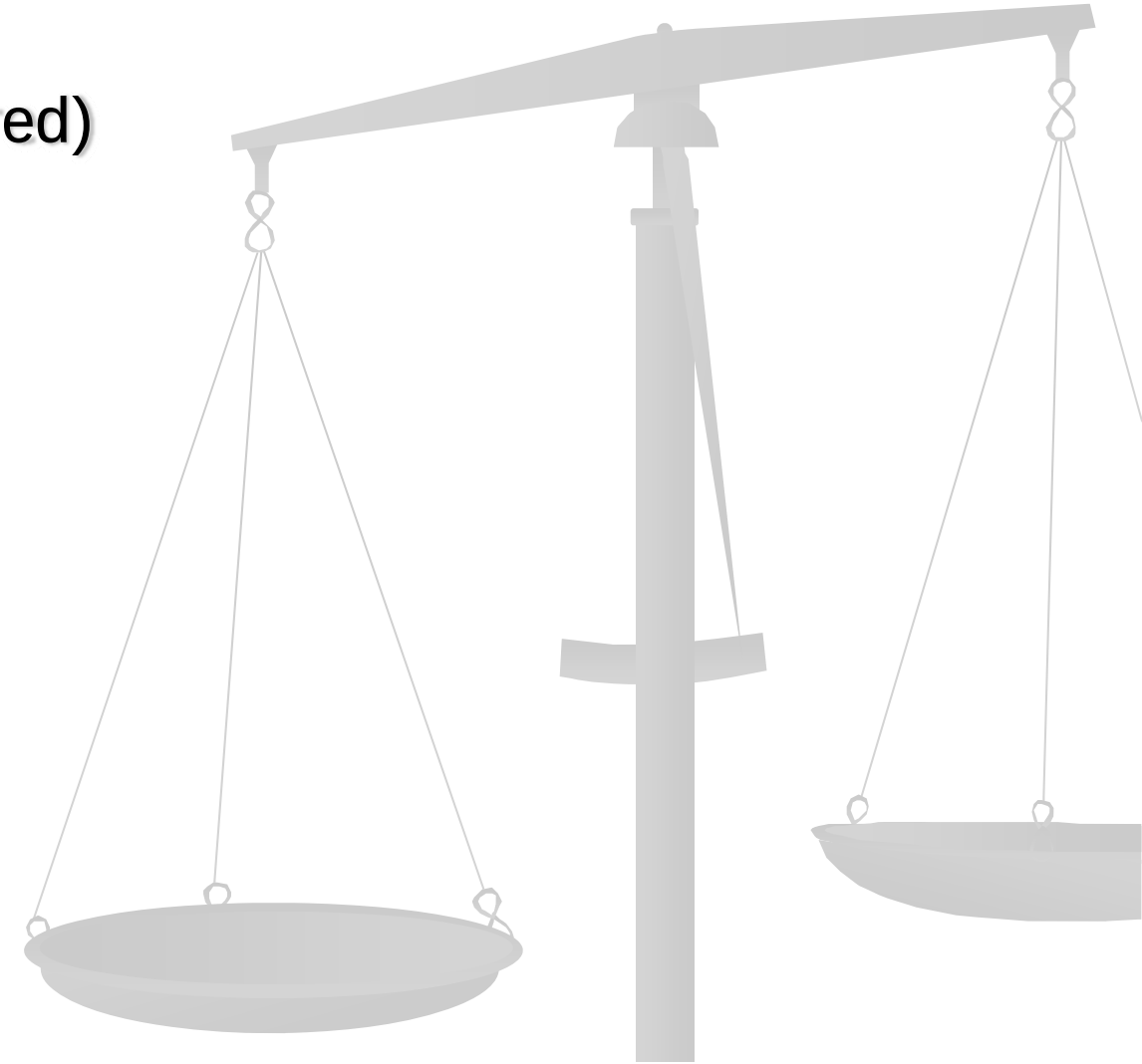
  You are interested in studying the distribution of various majors of 400 students enrolled in an undergraduate program at a small university.

- 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- this includes both discrete and continuous random variables.

- Stemplots (not covered)
- Histograms



Histograms (example)

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

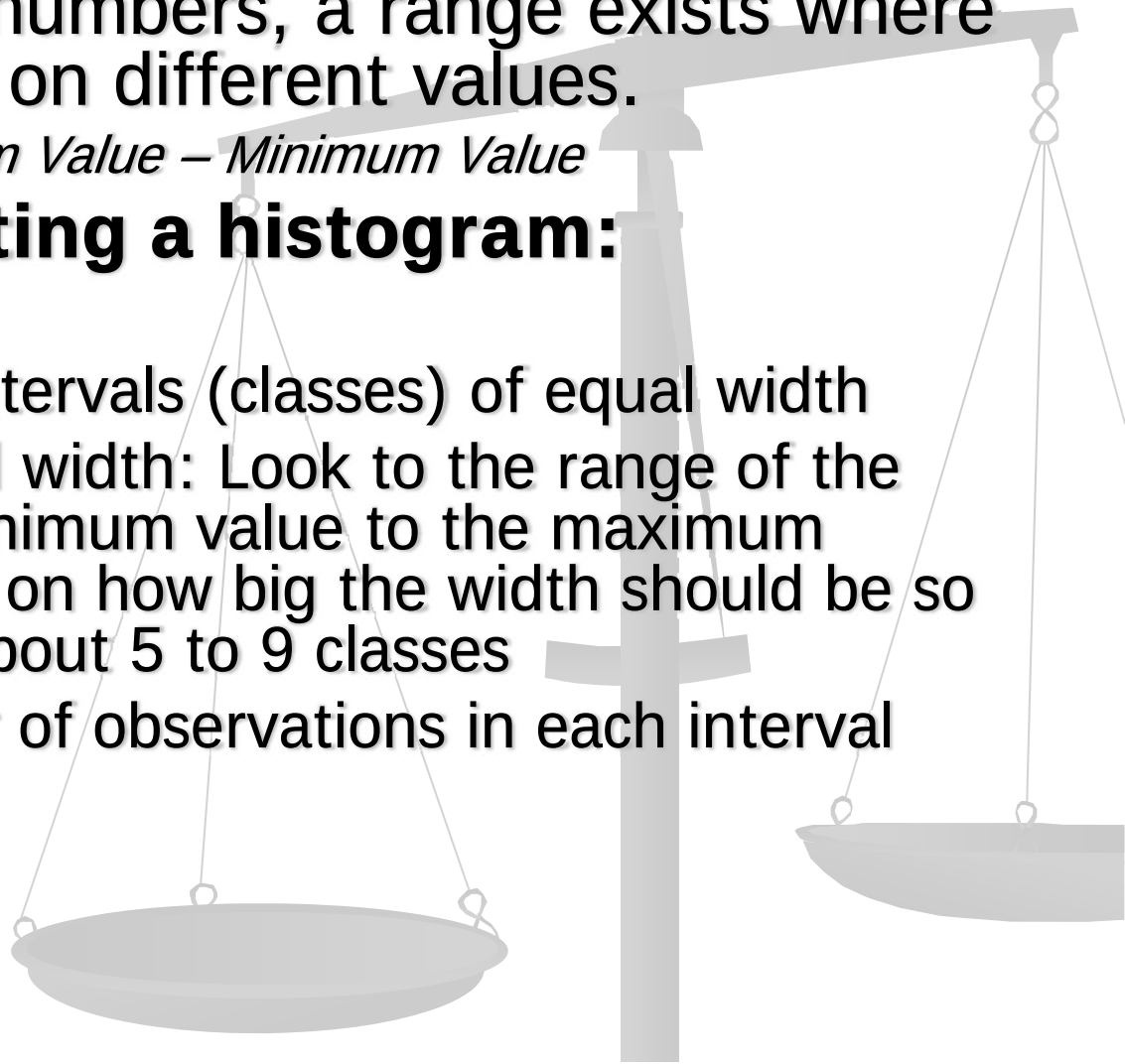
State	Percent	State	Percent	State	Percent
Alabama	1.5	Louisiana	2.4	Ohio	1.6
Alaska	3.6	Maine	0.6	Oklahoma	4.3
Arizona	21.3	Maryland	4.0	Oregon	6.5
Arkansas	2.8	Massachusetts	5.6	Pennsylvania	2.6
California	28.1	Michigan	2.7	Rhode Island	7.0
Colorado	14.9	Minnesota	2.4	South Carolina	2.2
Connecticut	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.
  - *Range = Maximum Value – 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

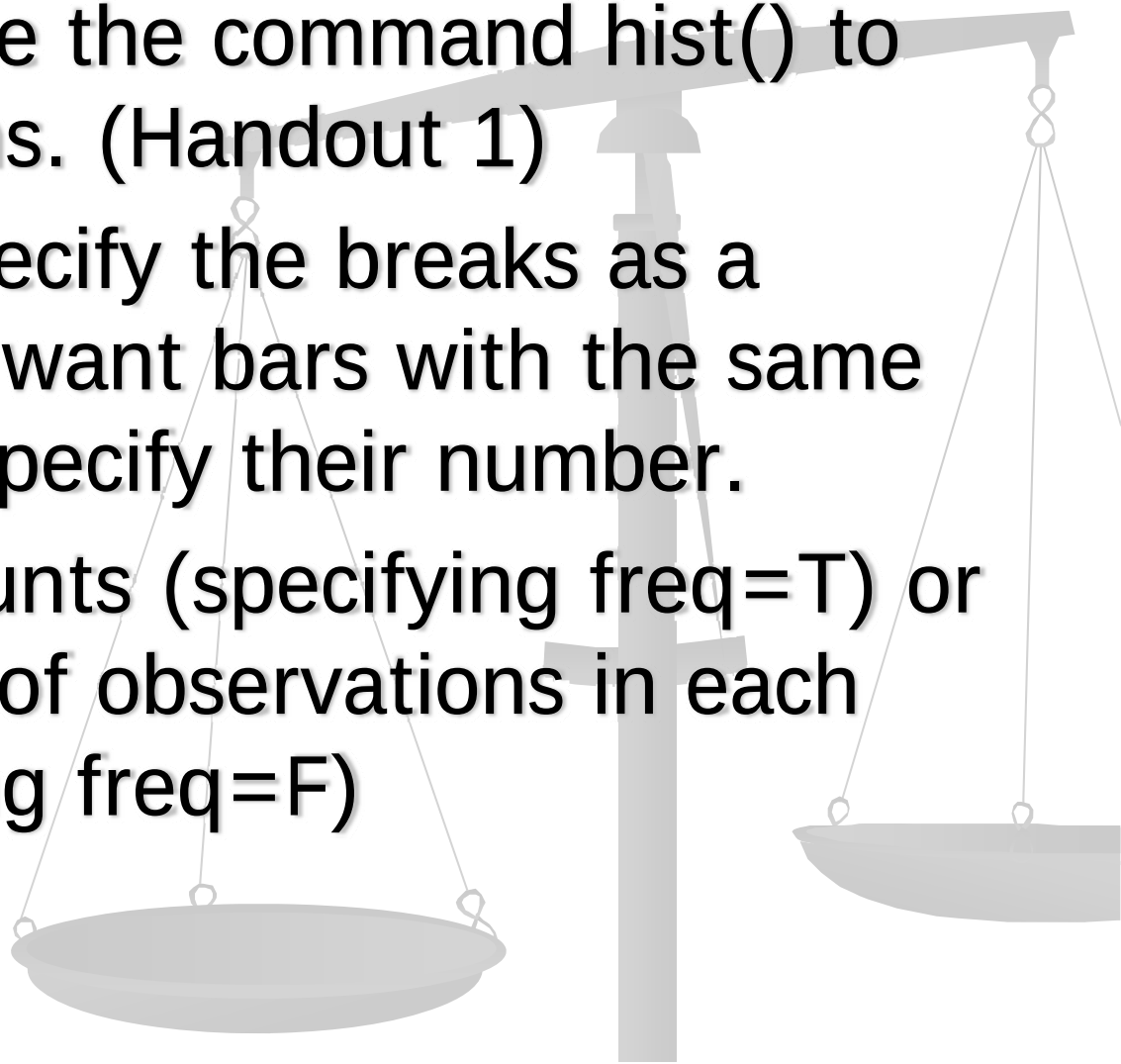


Frequency Table

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

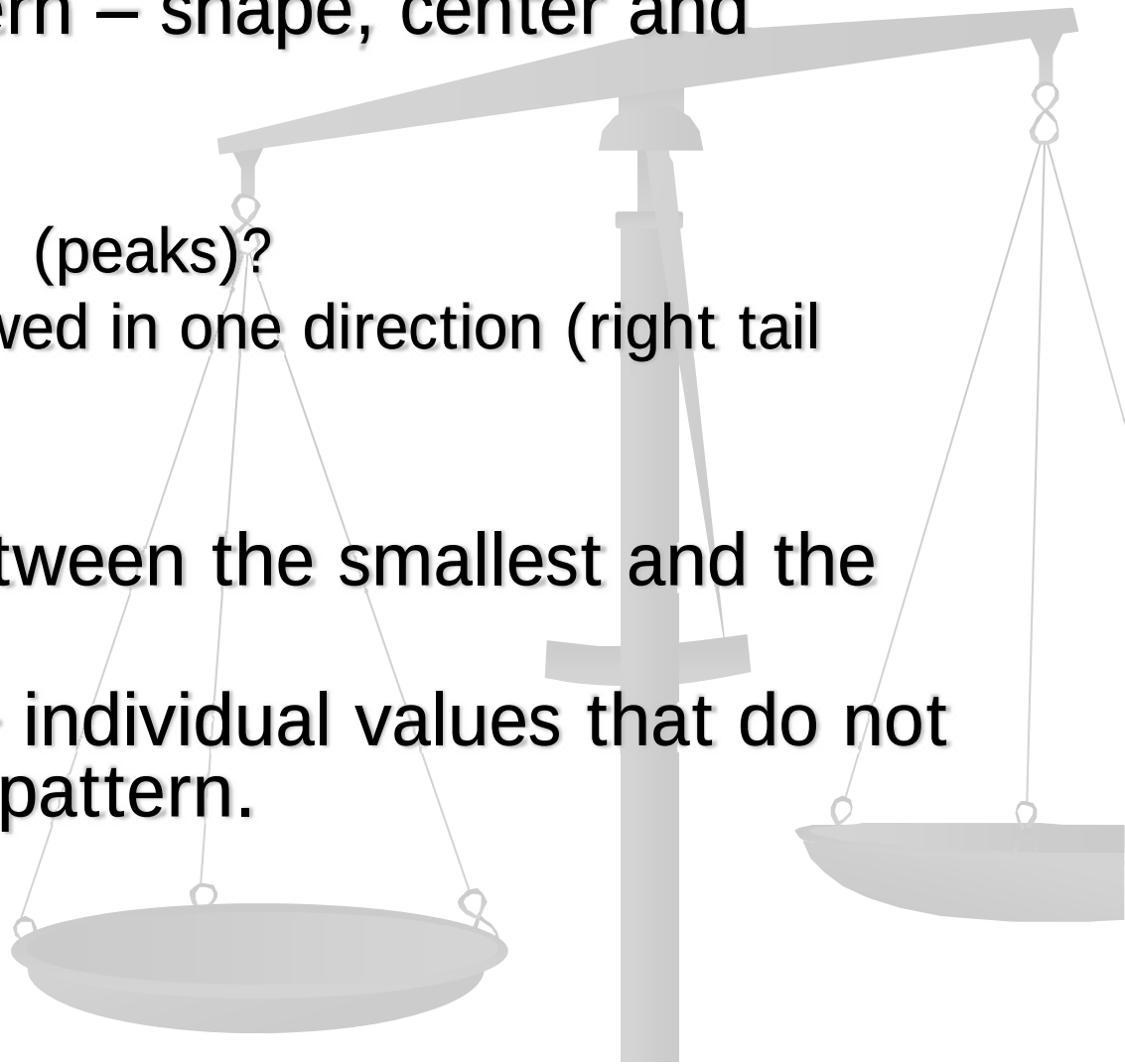
Using R

- In R you can use the command hist() to make histograms. (Handout 1)
- 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)



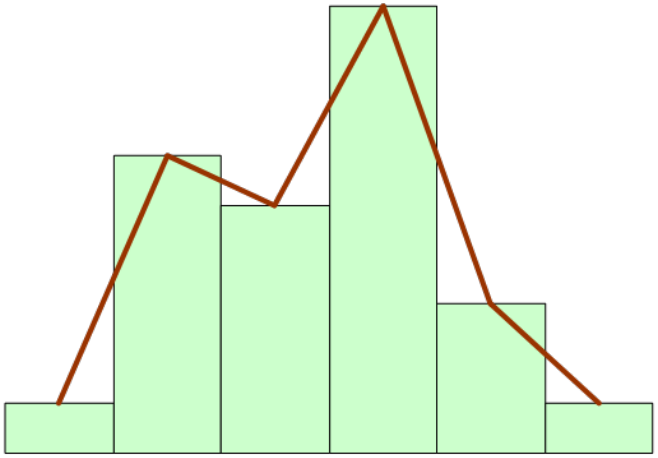
Examining distributions

- 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.

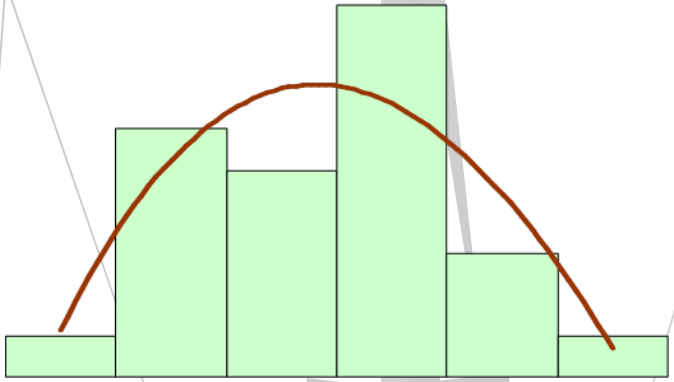


Interpreting histograms

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.**



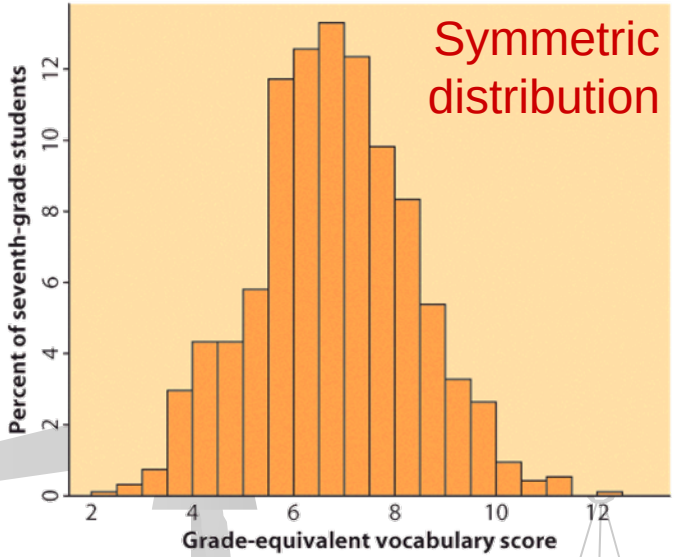
Histogram with a line connecting each column → too detailed

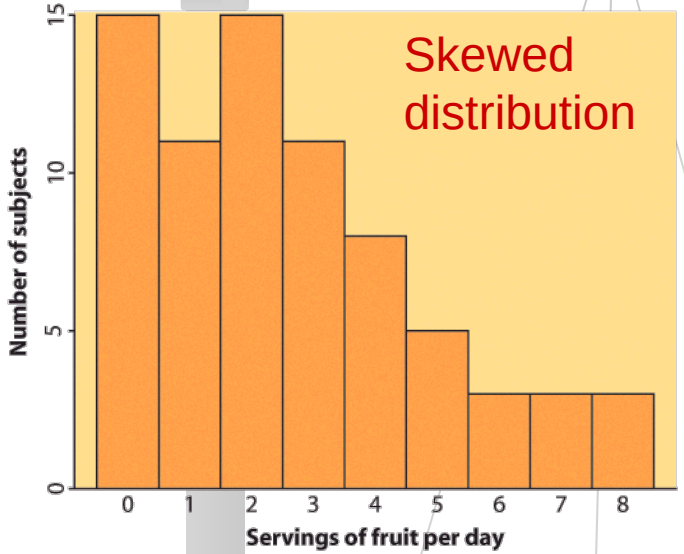


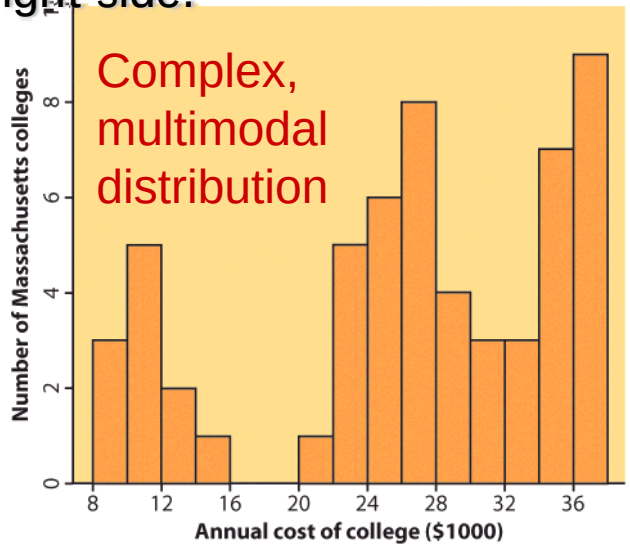
Histogram with a smoothed curve highlighting the overall pattern of the distribution

Most common distribution shapes

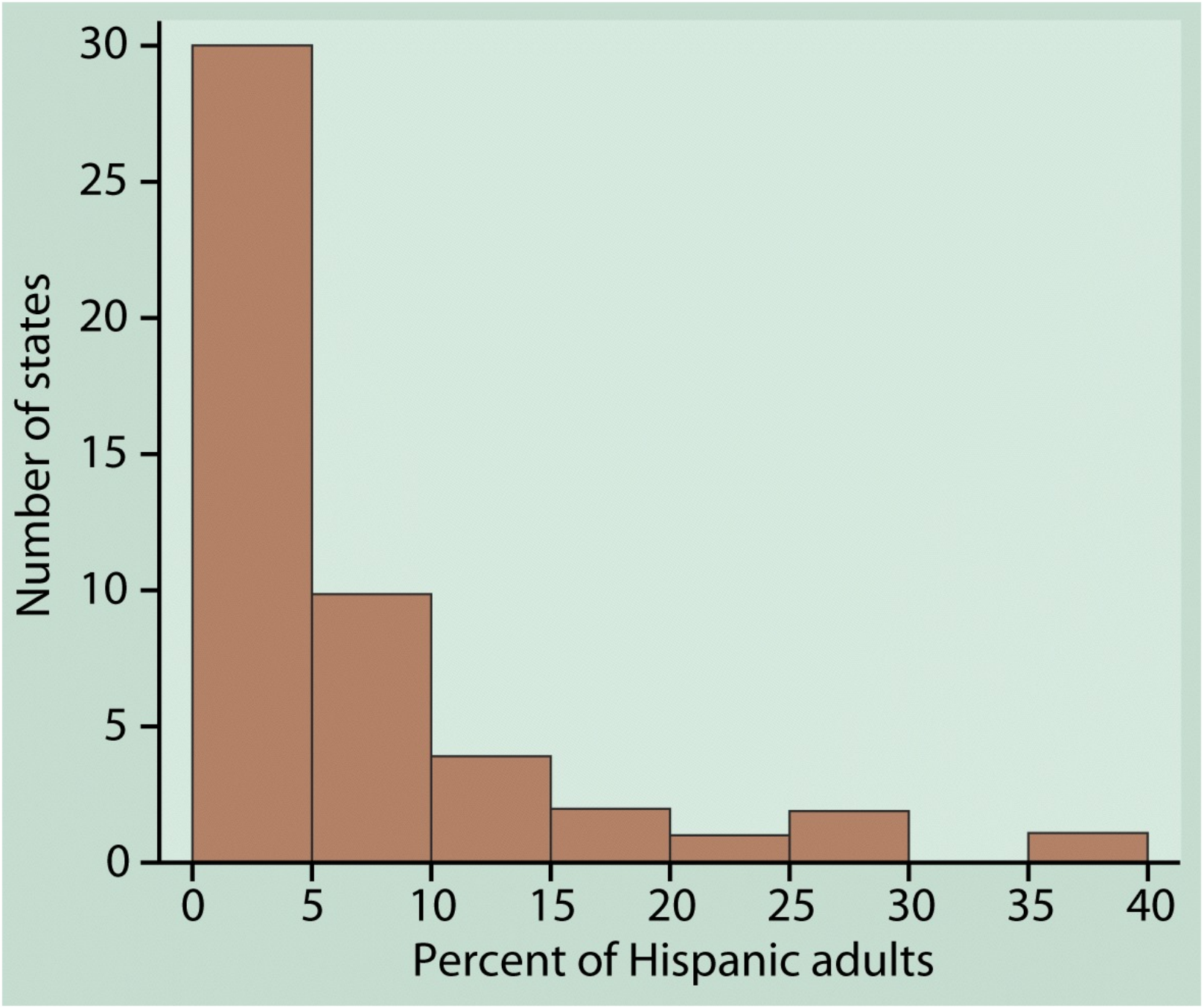
- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side. It is **skewed to the left**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





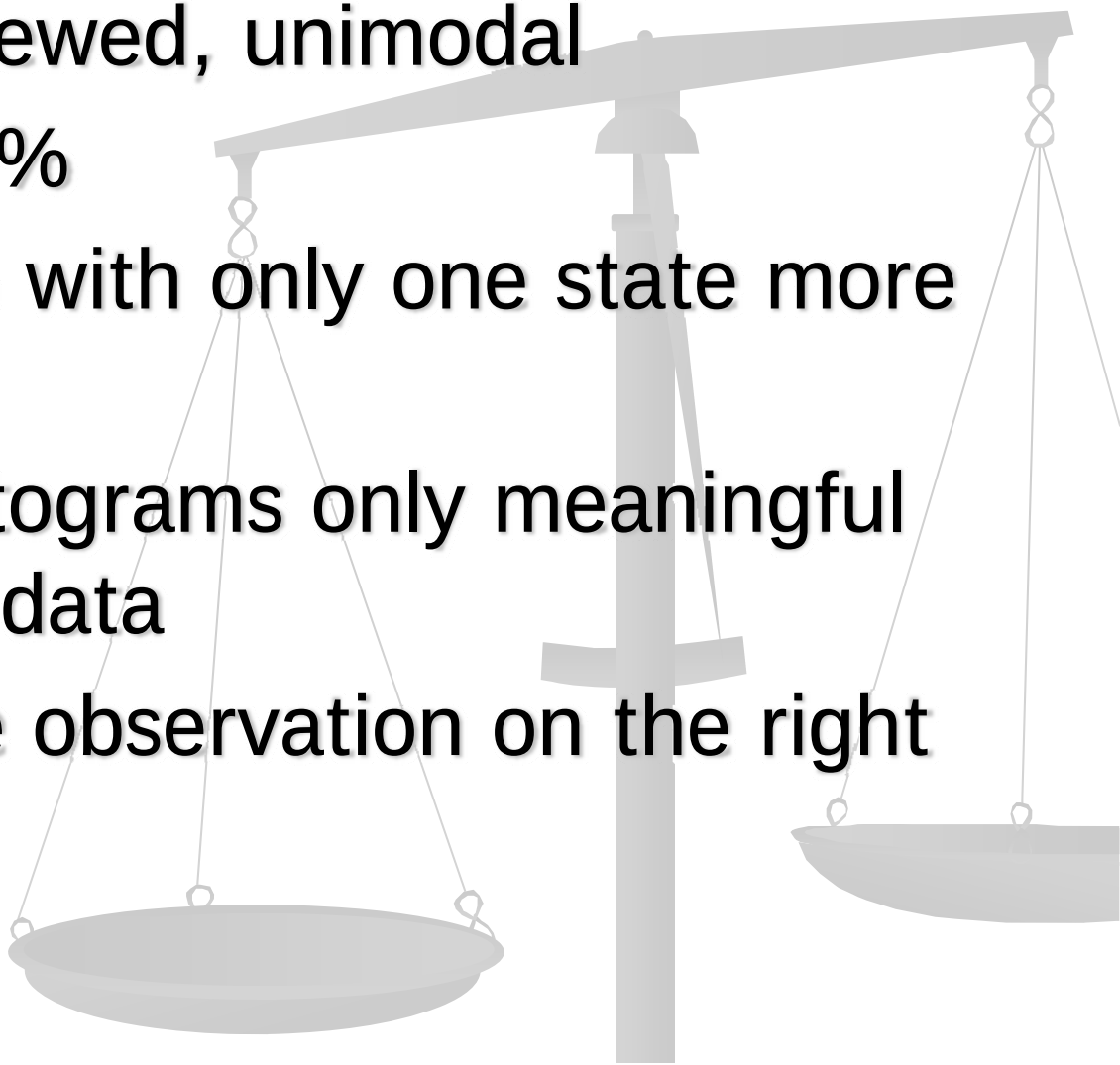


- Not all distributions have a simple overall shape, especially when there are few observations.


Number of states
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35
40
Percent of Hispanic adults

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?



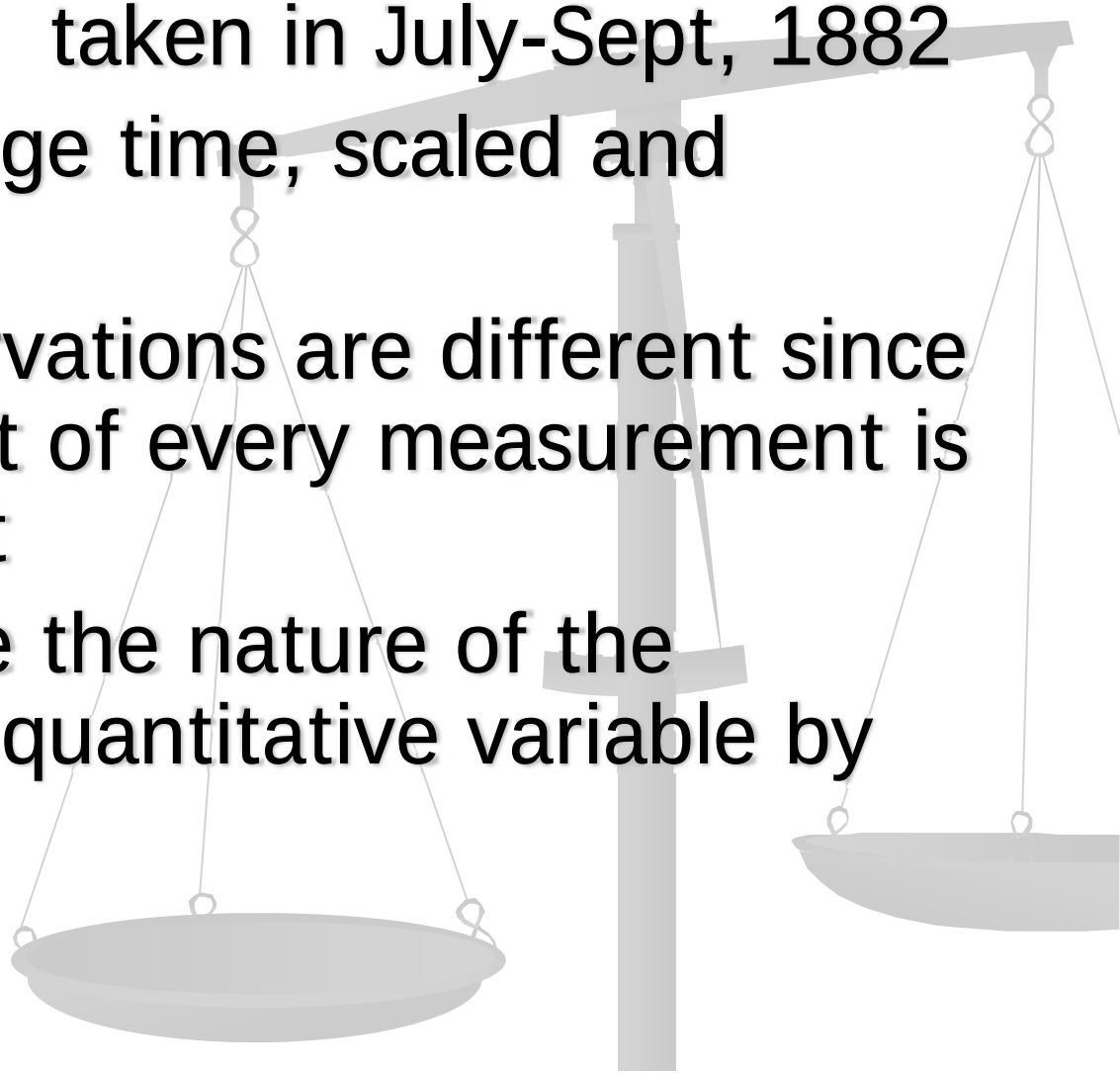
Quantitative Variables-Graphical Display

- Deviations from 24,800 nanoseconds

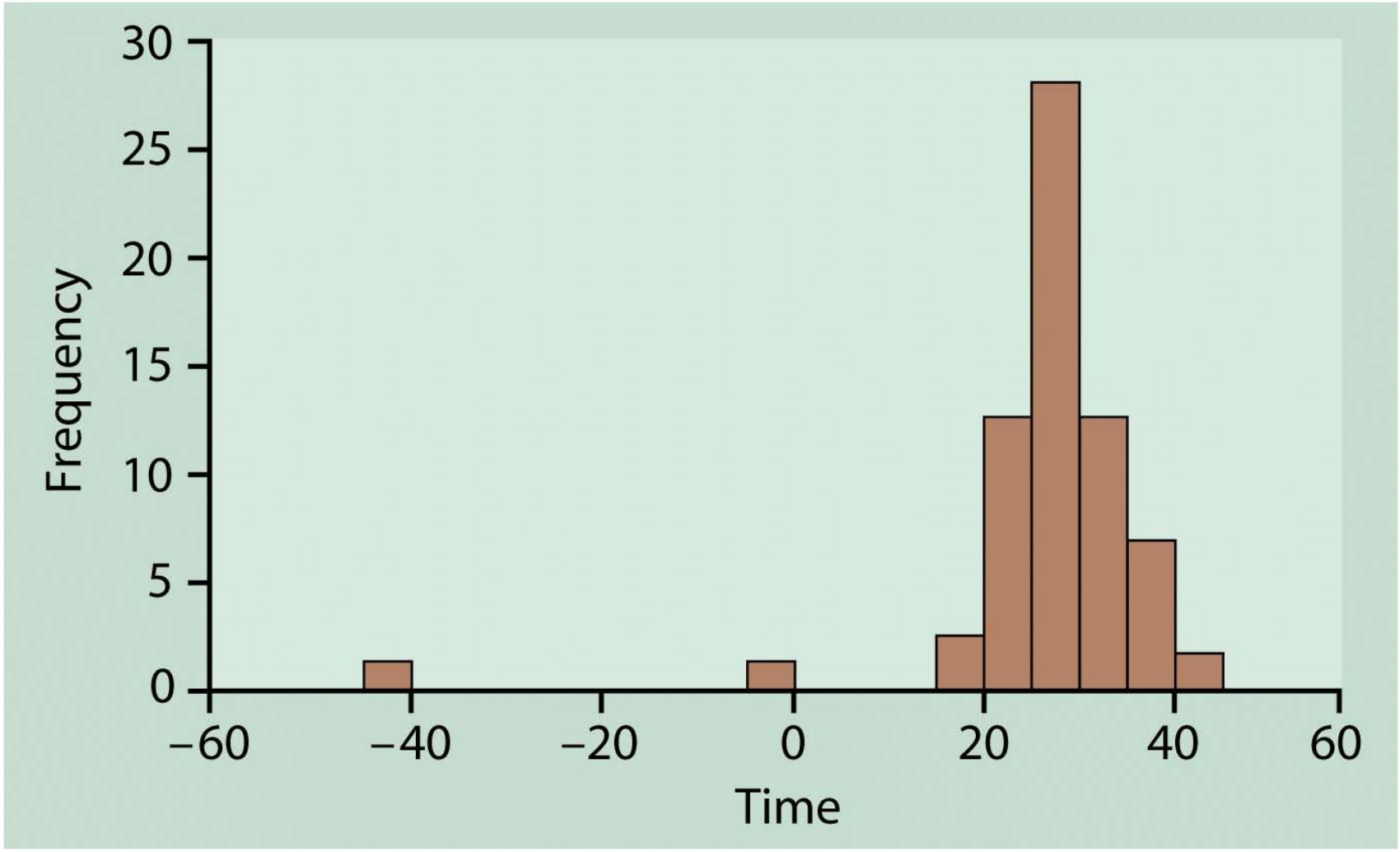
TABLE 1.1 Newcomb's measurements of the passage time of light

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

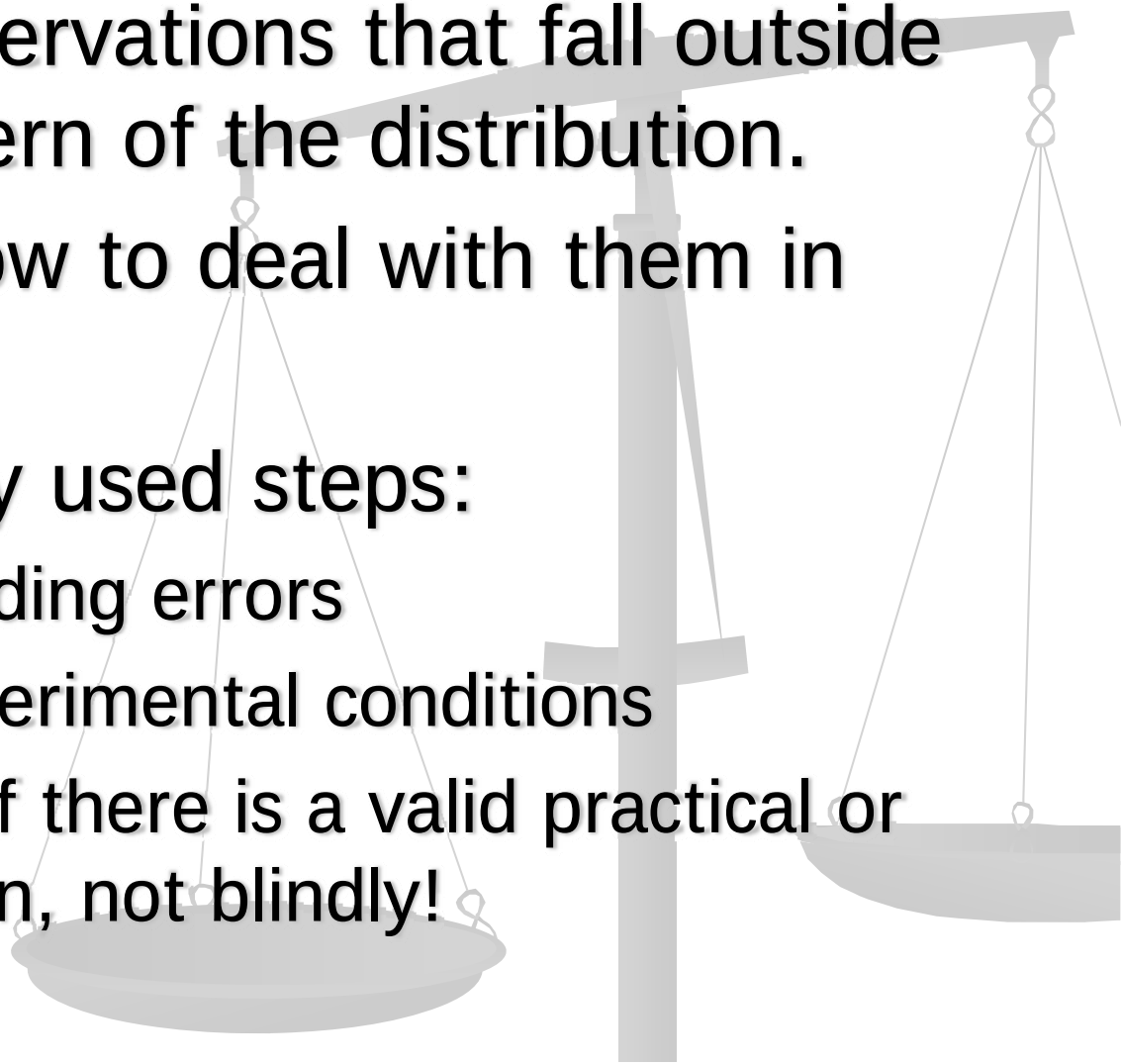


Newcomb's data (dealing with outliers)

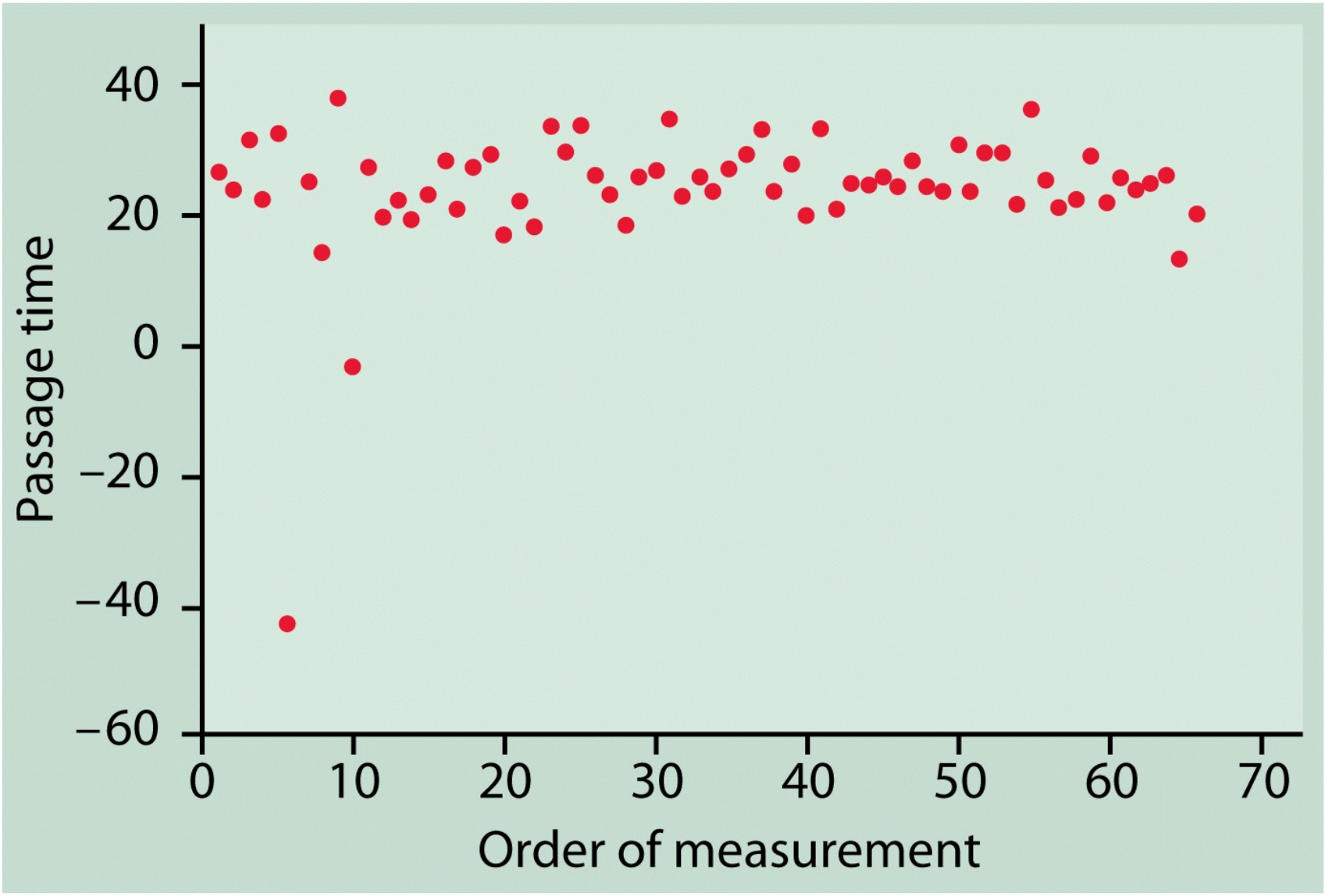


Outliers

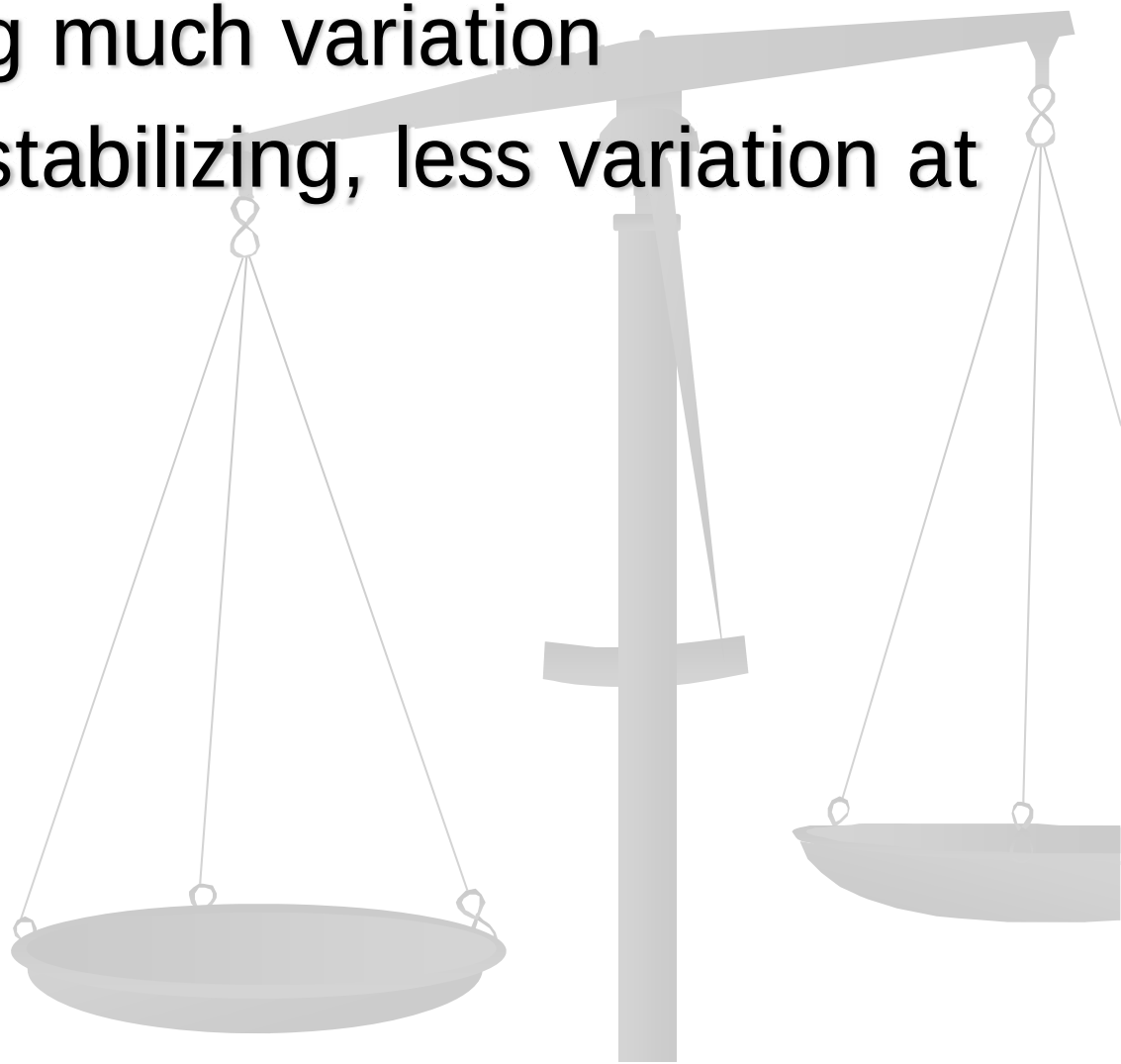
- Outliers are observations that fall outside the overall pattern of the distribution.
- We will learn how to deal with them in MA331
- Some commonly used steps:
  - Check for recording errors
  - Violation of experimental conditions
  - Discard it only if there is a valid practical or statistical reason, not blindly!



Time plots. Newcomb's data.



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

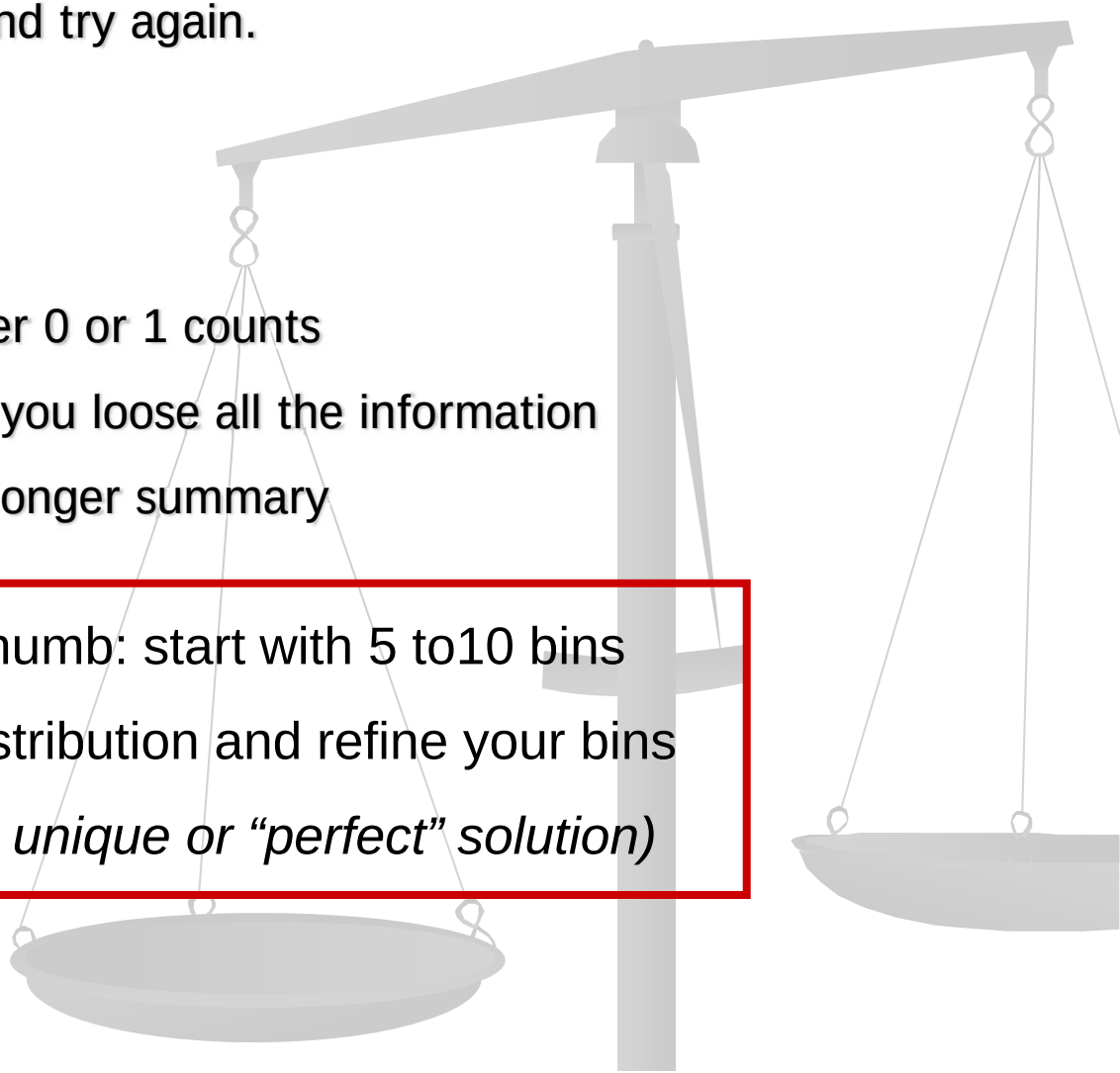


How to create a histogram

It is an iterative process – try and try again.

What bin size should you use?

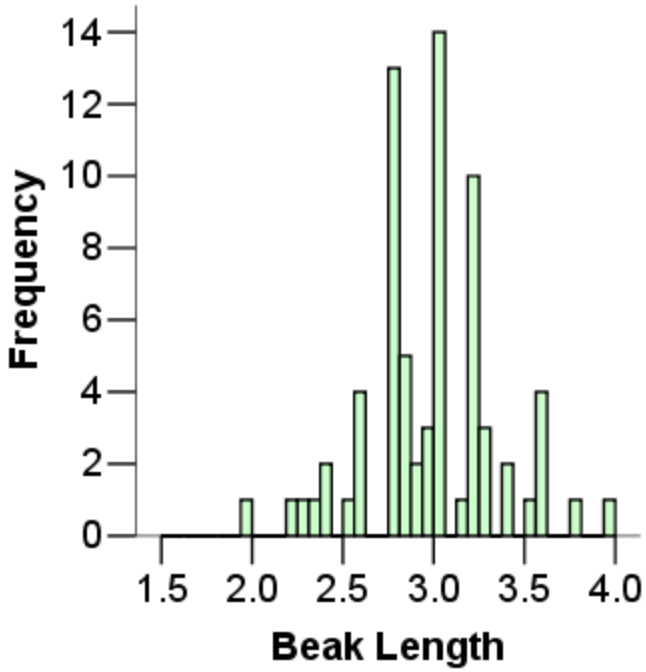
- Not too many bins with either 0 or 1 counts
- Not overly summarized that you loose all the information
- Not so detailed that it is no longer summary

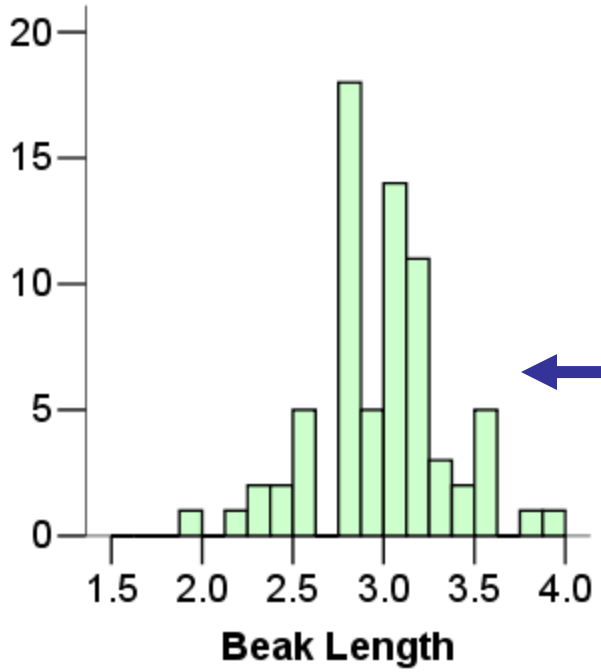


➔ rule of thumb: start with 5 to10 bins

Look at the distribution and refine your bins

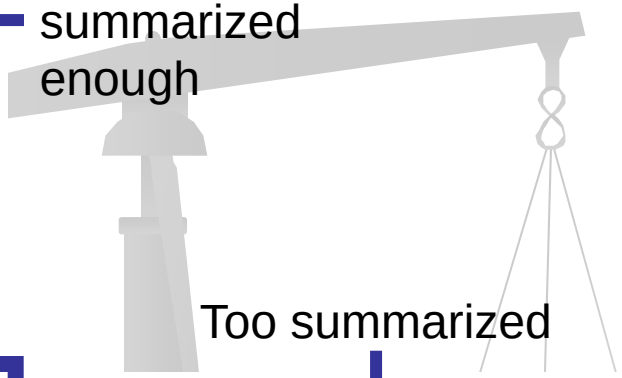
(There isn't a unique or "perfect" solution)



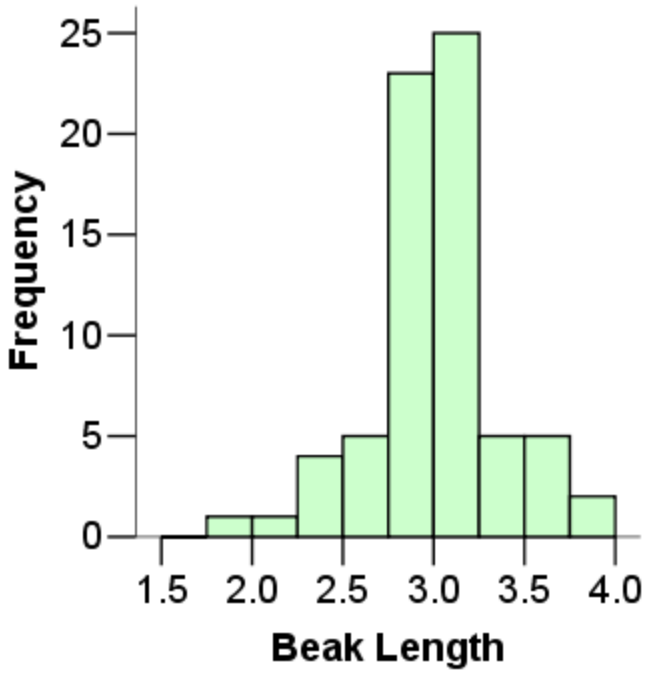


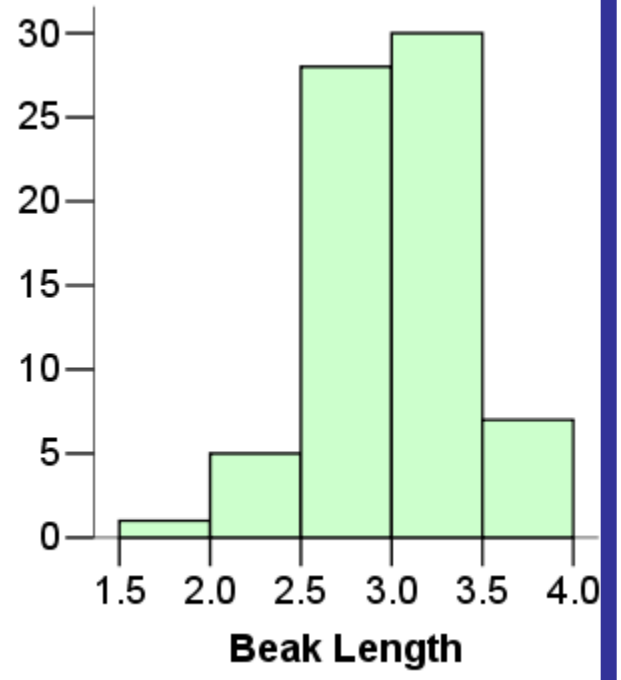
Same data set

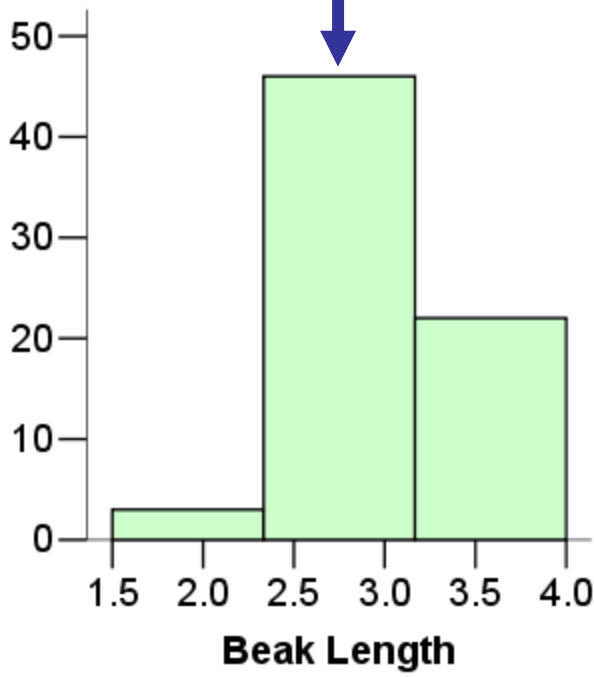
Not summarized enough



Too summarized



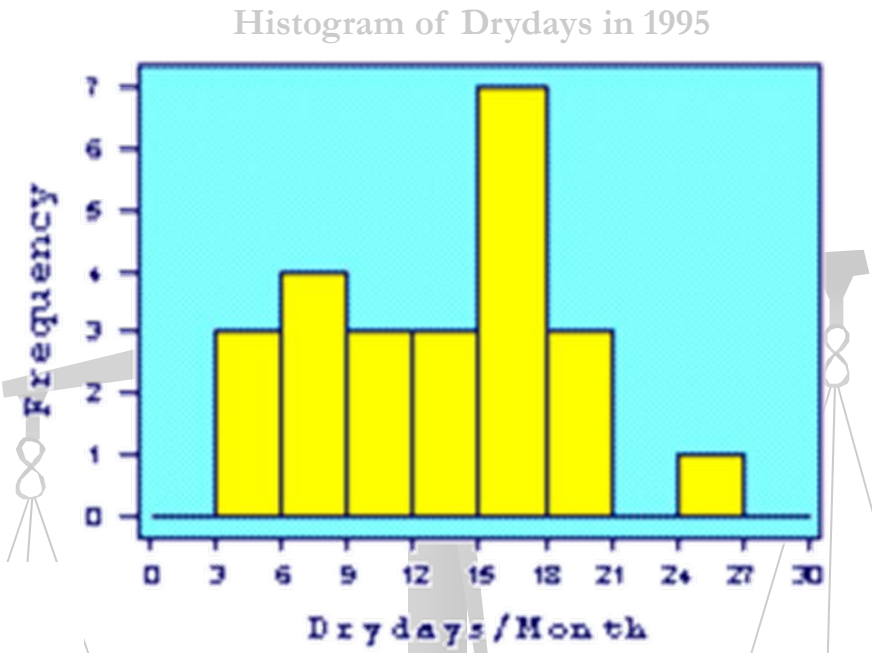


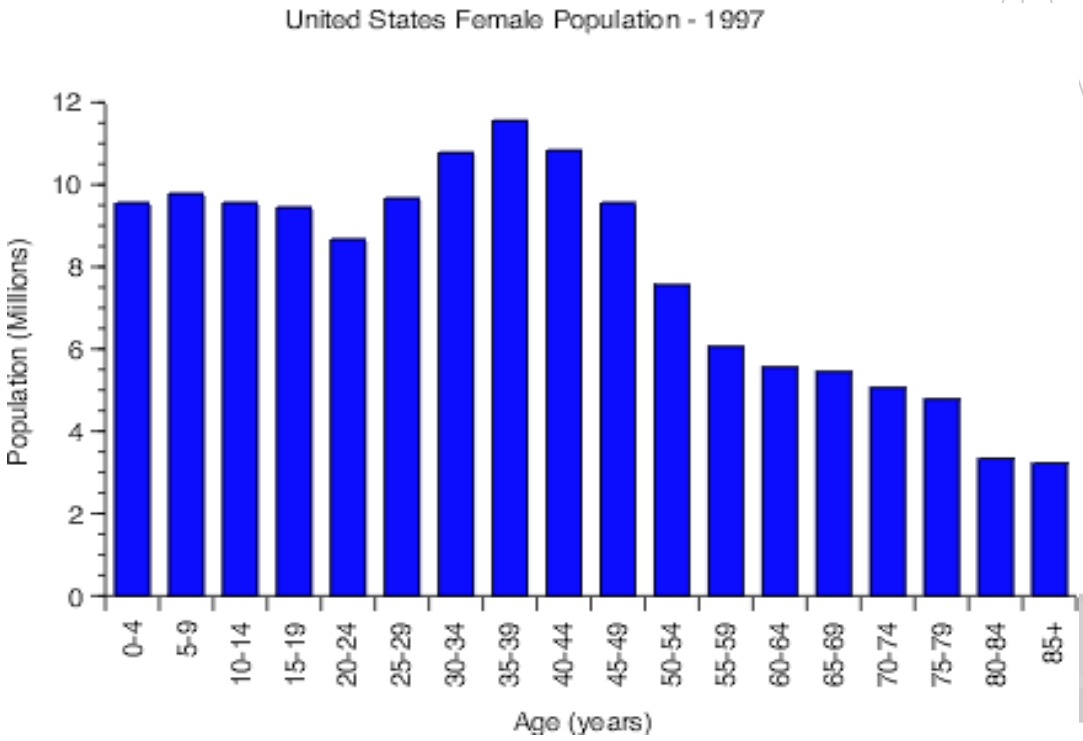


IMPORTANT NOTE:

Your data are the way they are.

Do not try to force them into a particular shape.

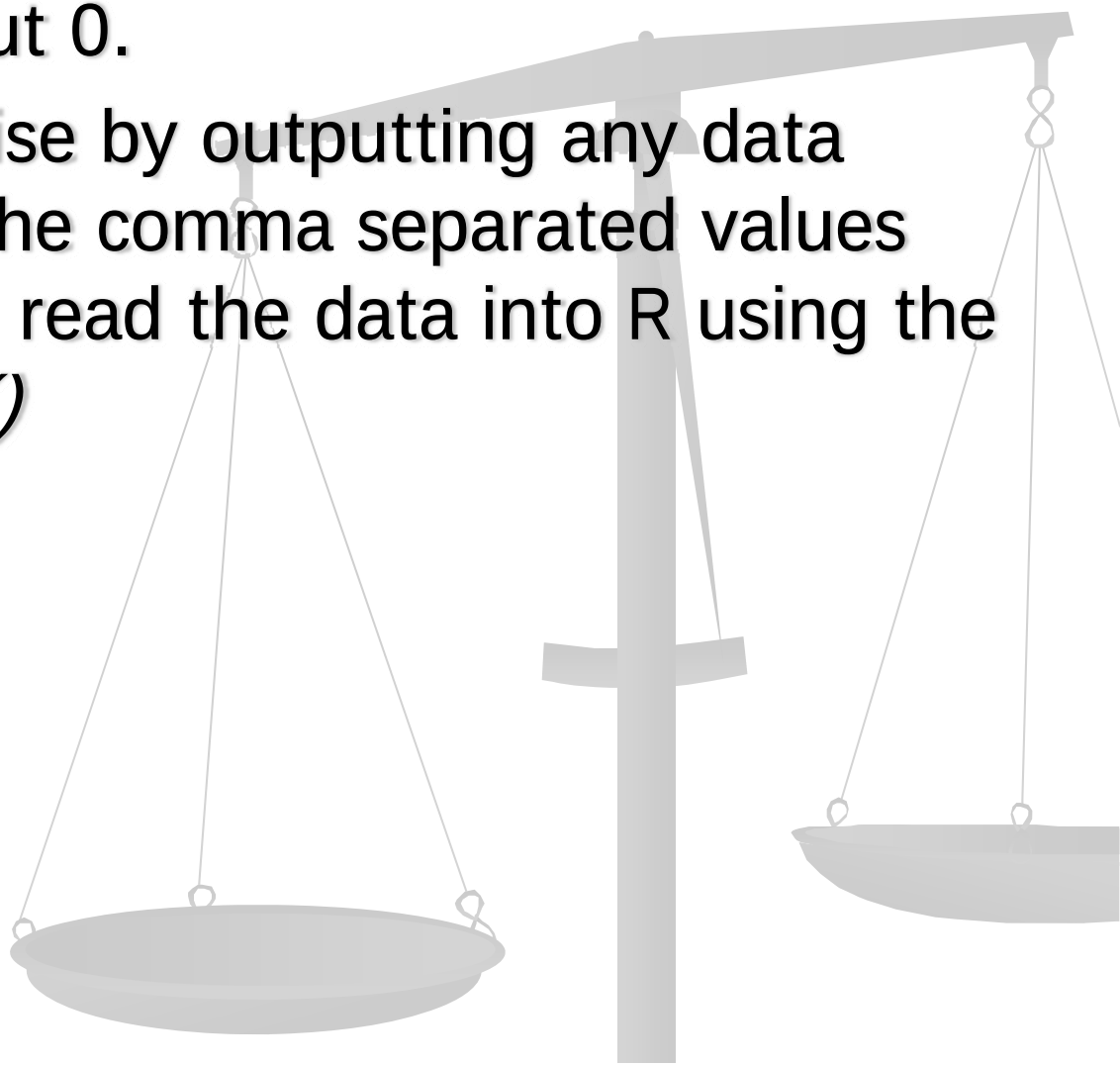




It is a common misconception that if you have a large enough data set, the data will eventually turn out nice and symmetrical.

Exercises: Learn to input data in R

- Please see Handout 0.
- Do a simple exercise by outputting any data from Excel using the comma separated values (.csv) format, and read the data into R using the function *read.csv()*



Summary

- Categorical and Quantitative variable
- Graphical tools for categorical variable

Bar Chart, Pie Chart

- For quantitative variable: Histogram
- Describe: Shape, center, spread
- Watch out for patterns and deviations from patterns.