

Data Basics

Chapter 1

Some slides developed by Mine Çetinkaya-Rundel of OpenIntro
The slides may be copied, edited, and/or shared via the [CC BY-SA license](#)

EXAMPLE 1

How Can Statistics Help Us Learn about the World?



How can we evaluate evidence against global warming?

Are cell phones dangerous?

What are the chances of a tax return being audited?

How likely are we to win the lottery?

Is there bias against women in appointing managers?

Statistics



Statistics is the art and science of

1. **Designing** studies,
2. **Analyzing** data that those studies produce.

The ultimate goal is to translate data into knowledge and understanding.

Statistics is the art and science of learning from data.

Three Aspects of a Study



www.icts.uiowa.edu

1. **Design:** Planning how to obtain data
2. **Description:** Summarizing the data
3. **Inference:** Making decisions and predictions

1st Aspect of a Study: Design

How do we **conduct the experiment** or **select people** for the **survey** to insure trustworthy results?

Design Examples:

1. Planning data collection to study effects of Vitamin E on athletic strength
2. For a marketing survey, selecting people to provide proper coverage



fineartamerica.com

2nd Aspect of a Study: Description



www.emecogroup.org

Summarize raw data and present in useful formats (e.g., average, charts or graphs)

Description Examples:

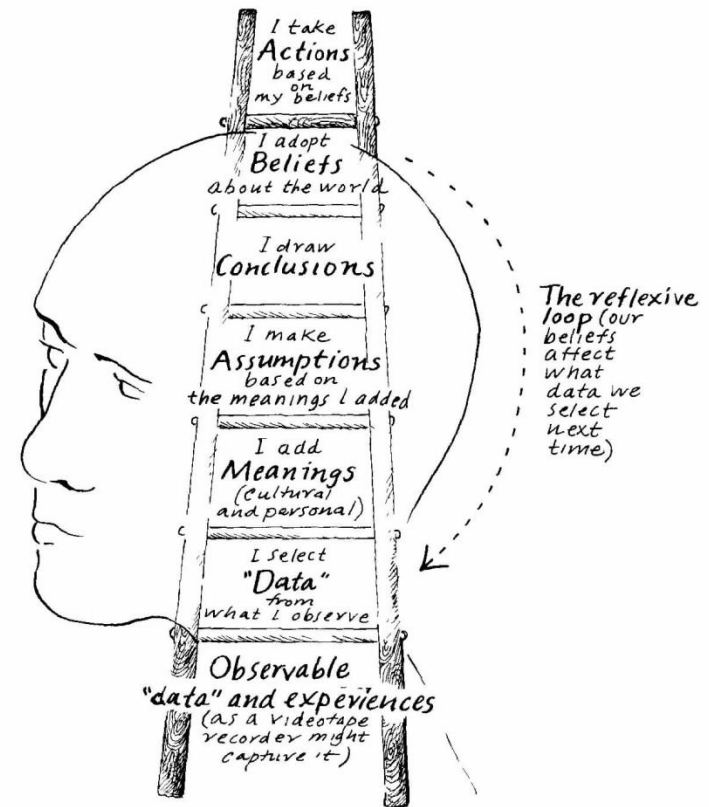
- ▣ A graph showing total precipitation in Clarksville for each month of 2005
- ▣ Average age of students in a statistics class is 25 years

3rd Aspect of a Study: Inference

Make decisions or predictions
based on the data

Inference Examples:

- ❑ Relationship between smoking cigarettes and getting emphysema
- ❑ 47% of the registered voters in Illinois will vote in the primary



Subjects

Subjects - The entities that we measure in a study

Subjects could be

1. individuals,
2. schools,
3. rats,
4. counties,
5. widgets



Mr. Ages from the Rats of NIMH
kiriko-moth.com

Population and Samples

- **Population:** All subjects of interest
- **Sample:** Subset of the population for whom we have data
- We observe samples, but we are interested in populations.

Sample & Population for an Exit Poll

- In California in 2003, a special election was held to consider whether Governor Gray Davis should be recalled from office.
- An exit poll sampled 3160 of the 8 million people who voted. Define the **sample** and the **population** for this exit poll.



ltn.co.uk

Descriptive vs. Inferential Statistics

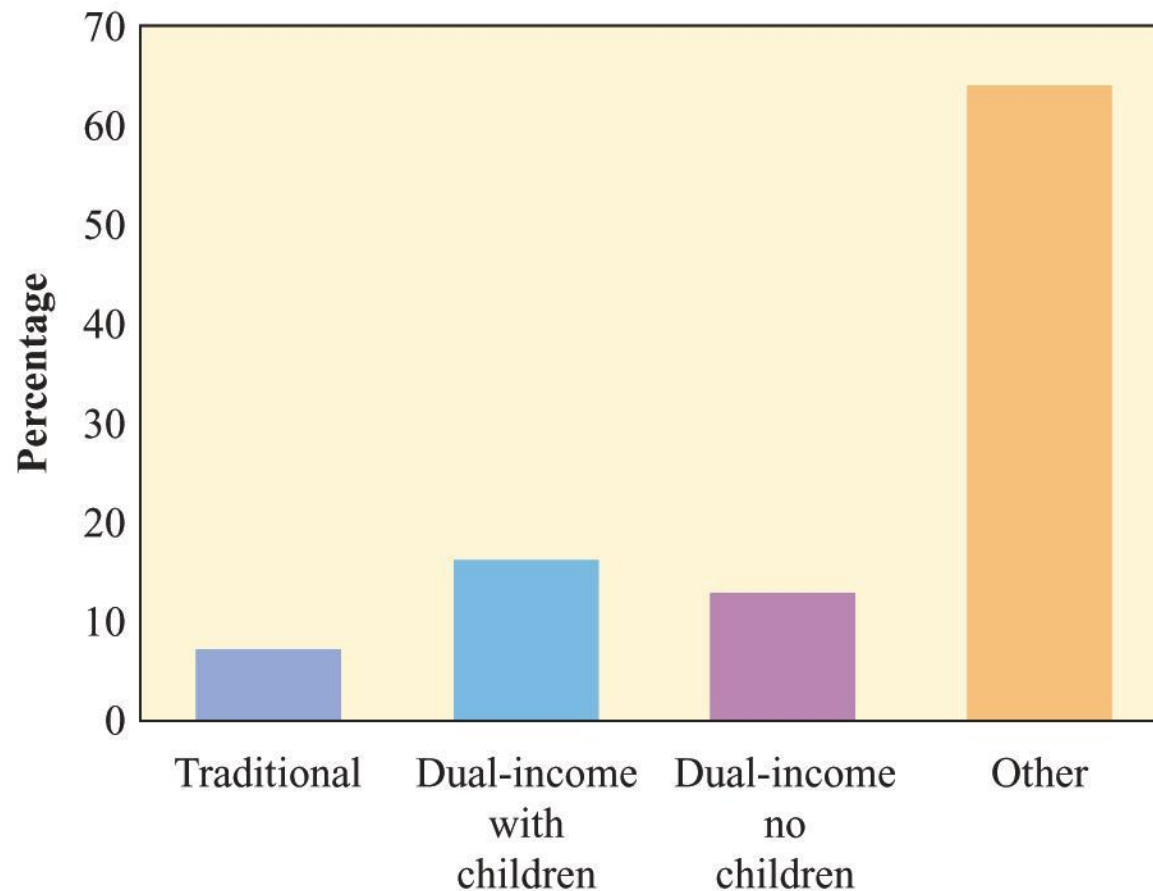


mallimages.mallfinder.com

- **Descriptive** statistics summarize data – graphs and numbers such as averages and percentages
- **Inferential** statistics make decisions or predictions about a population based on data obtained from a sample of that population.

Descriptive Statistics Example

Types of U.S. Households



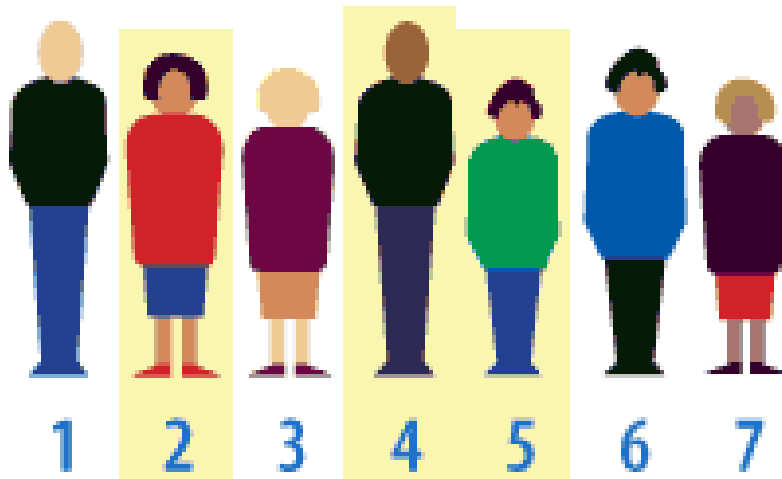
Inferential Statistics Example



The results for the sample of 834 Florida residents are summarized by the percentage, **54.0%**, who favored handgun control.

The poll reports that the “margin of error” for how close this number falls to the population percentage is **3.4%**.

Randomness



Assign Numbers,
Auto-Generate Random
Selections

- **Simple Random Sampling:**
each subject in the population has the same chance of being included in the sample
- Randomness is crucial to insuring that the sample is representative of the population so that powerful inferences can be made

Variability

- Measurements may vary from subject to subject, and
- Measurements may vary from sample to sample.

Predictions are therefore likely to be more accurate for larger samples.



What Role Do Computers Play in Statistics?



www.masternewmedia.org

- **Data files** - Large data sets organized in a spreadsheet format known as a data file
 - ▣ Each row contains measurements for a particular subject and column for a particular characteristic
- **Databases** – An existing archive collection of data files

Sources should always be checked for reliability.



Introduction to RStudio

Introduction

The facts:

- R is a language and environment for statistical computing and graphics
- Freely available and maintained by volunteers
- R is extensible; can be expanded by installing “packages”

How to get it:

- <http://www.r-project.org/> (or Google “Download R”)
- Available for Windows, Mac, Linux
- Free to install, no catches

Also highly recommended:

- R Studio: a free IDE for R
- <http://www.rstudio.com/>
- If you install R and R Studio, then you only need to run R Studio

For R related tutorials and/or resources see the following link:

<http://www.cs.utexas.edu/~cannata/dataVis/Class%20Notes/Getting%20Started%20with%20RStudio.pdf>

Using R

- R is command-line driven (*very* little point-and-click)
- You use “functions” to work with data
- Most analyses require writing a script, which is sourced into the R console
- traditional R command line interface is not very user friendly
- R Studio makes this process easier

What's so special about RStudio ?

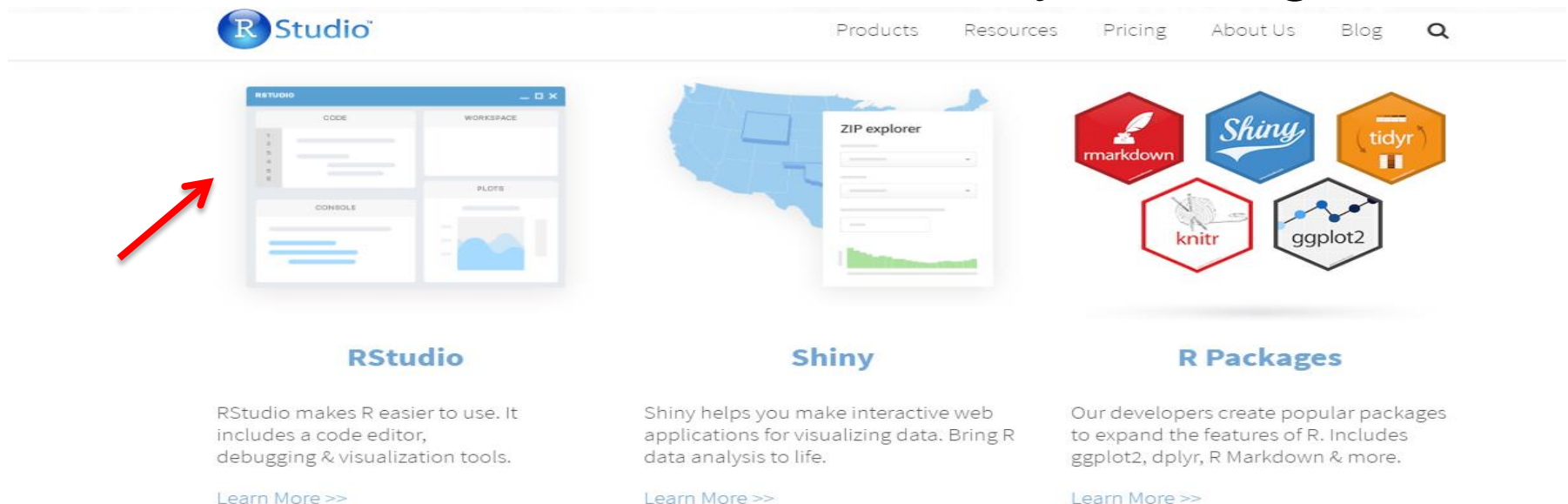
- Free
- Over 4000 packages that add functionality (about 25 come with R)
- Produces nice print-ready graphics
- Open-source (you can see *how* it does what it does)
- Easy to install and non-invasive

Install RStudio

- To install RStudio go to:

<http://www.rstudio.org/>

- You can download RStudio by clicking on



The screenshot shows the RStudio website homepage. At the top is the RStudio logo and a navigation menu with links for Products, Resources, Pricing, About Us, and Blog. Below the navigation bar are three main sections:

- RStudio**: Features a red arrow pointing to a thumbnail image of the RStudio IDE interface. The text below states: "RStudio makes R easier to use. It includes a code editor, debugging & visualization tools." and provides a "Learn More >>" link.
- Shiny**: Features a thumbnail image of a Shiny application titled "ZIP explorer" overlaid on a map of the United States. The text below states: "Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life." and provides a "Learn More >>" link.
- R Packages**: Features a cluster of five hexagonal icons representing popular R packages: rmarkdown, Shiny, tidy, knitr, and ggplot2. The text below states: "Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more." and provides a "Learn More >>" link.

Install RStudio

- Next you click to download RStudio for your desktop



Products

Resources

Pricing

About Us

Blog



	RStudio Desktop (Free License)	RStudio Desktop (Commercial License)	RStudio Server (Free License)	RStudio Server Pro (Commercial License)
Integrated Development Environment for R	✓	✓	✓	✓
Priority support		✓		✓
Access via Web Browser			✓	✓
Enterprise Security and Access Controls				✓
Project Sharing				✓
Access to Multiple Versions of R				✓

Install RStudio

- Finally, click **DOWNLOAD RSTUDIO DESKTOP**



Linux server and want to enable users to remotely

Products

Resources

Pricing

About Us

Blog



Do you need support or a commercial license? [Check out our commercial offerings](#)

RStudio Desktop 0.99.903 — [Release Notes](#)

RStudio requires R 2.11.1+. If you don't already have R, download it [here](#).

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 0.99.903 - Windows Vista/7/8/10	77.1 MB	2016-07-18	716f28f2143c5e21f4acea5752e284f8
RStudio 0.99.903 - Mac OS X 10.6+ (64-bit)	60 MB	2016-07-18	d14a1585b5a5ac0839507b9c04d460d6
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (32-bit)	81.6 MB	2016-07-18	761eae80b0ba4d4cd9051a802a2c44e2
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (64-bit)	88.3 MB	2016-07-18	98ea59d3db00e0083d3e4053514f764d
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	81 MB	2016-07-18	ce2ea1023d99175cb909def0fe66eba7
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	81.9 MB	2016-07-18	152f247255e86904cf3354afbc7b3b99

Starting RSTUDIO

- Go to your Start menu and in programs start RStudio by clicking on the RStudio icon:



- When you open the GUI, you will the RStudio screen

RStudio screen

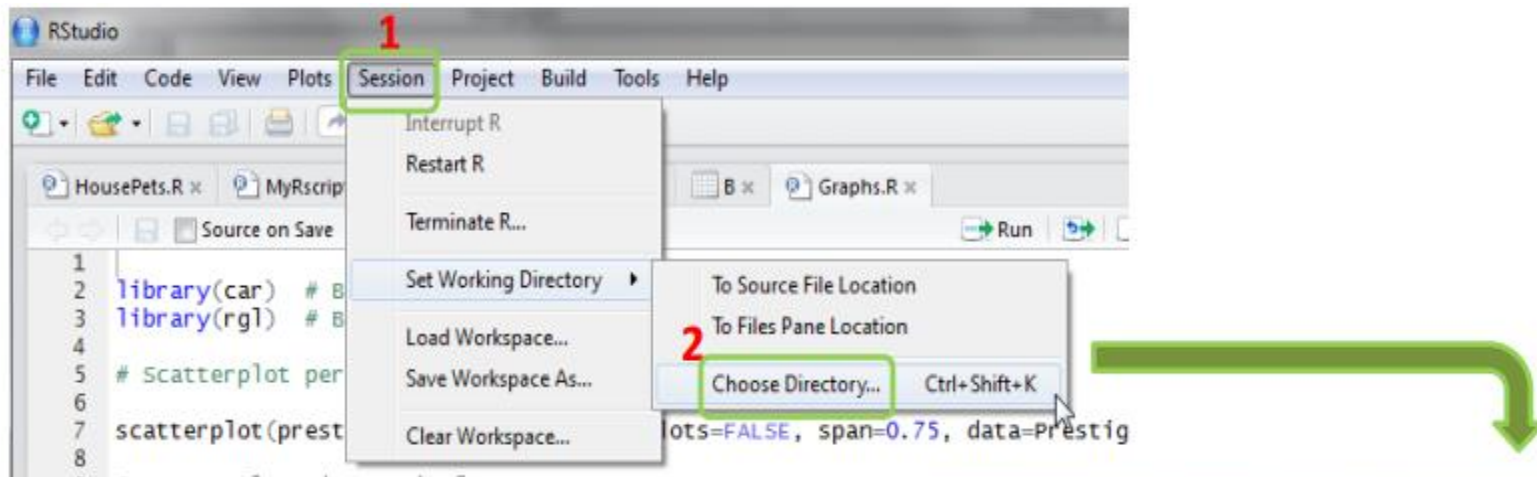
The image shows the RStudio desktop application window. The interface is divided into several panes, each with a specific function. Callouts in blue speech bubbles identify the following components:

- Script editor:** The top-left pane where R code is written. It contains a script named 'Untitled1.R' with the following code:

```
2 # plot(x[,1]~x[,2])
3 plot(x[,3]~x[,2])
4 # apply(x,1,function(x){
5 #   sum(x,na.rm=T)
6 # })
7 y=matrix(1:21,nrow=3)
8
9 z=as.data.frame(y)
10 ?mean
11 ?sqrt
12 ?factorial
13 mean(5,6,55,4,27)
14 sqrt(c(9,16))
15 length(paste("hello","you"))
16 length(c("hello","you"))
17 b=1:6
```
- Workspace and history file:** The top-right pane, labeled 'Workspace' and 'History'. It shows the current workspace containing variables 'x' (3x3 integer matrix) and 'y' (7x3 integer matrix). The 'History' tab shows the execution history of the script.
- Files, plots & files; manage packages:** The bottom-left pane, labeled 'Files', 'Plots', 'Packages', and 'Help'. It shows the current project files, including 'R: Complex Vectors'. The 'Plots' tab is active, showing a plot of 'Complex Vectors'.
- R console:** The bottom-right pane, labeled 'Console'. It shows the output of the R code executed in the script editor. The output is as follows:

```
> length(c("hello"," you",sep=""))
[1] 3
> length(paste("hello","you"))
[1] 1
> paste("hello","you")
[1] "hello you"
> length(c("hello"," you"))
[1] 2
> length(c("hello","you"))
[1] 2
> b=1:6
> fix(b)
fix(b)
> fix(b)
> View(y)
```

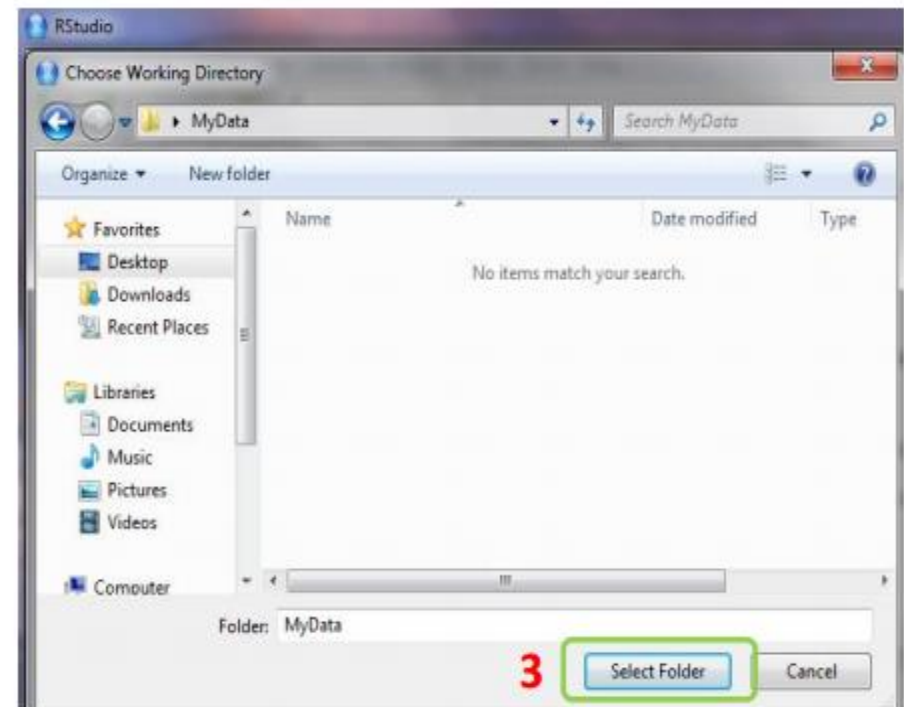
Changing the working directory



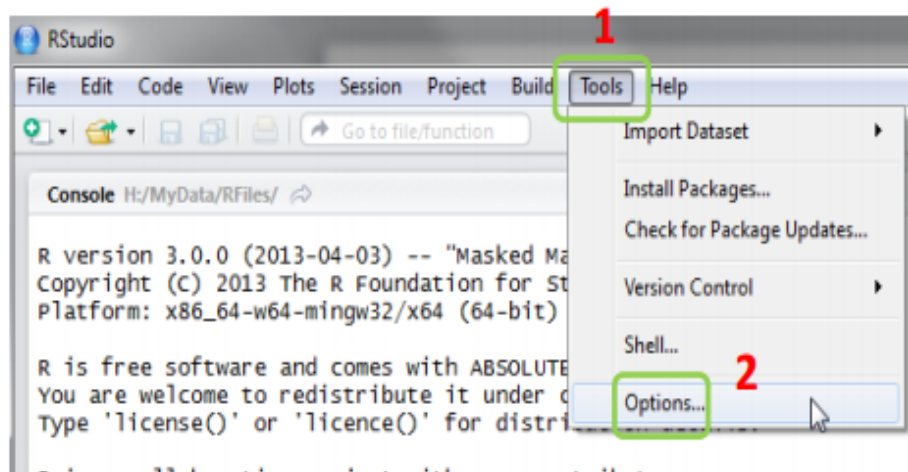
If you have different projects you can change the working directory for that session, see above. Or you can type:

```
# Shows the working directory (wd)  
getwd()
```

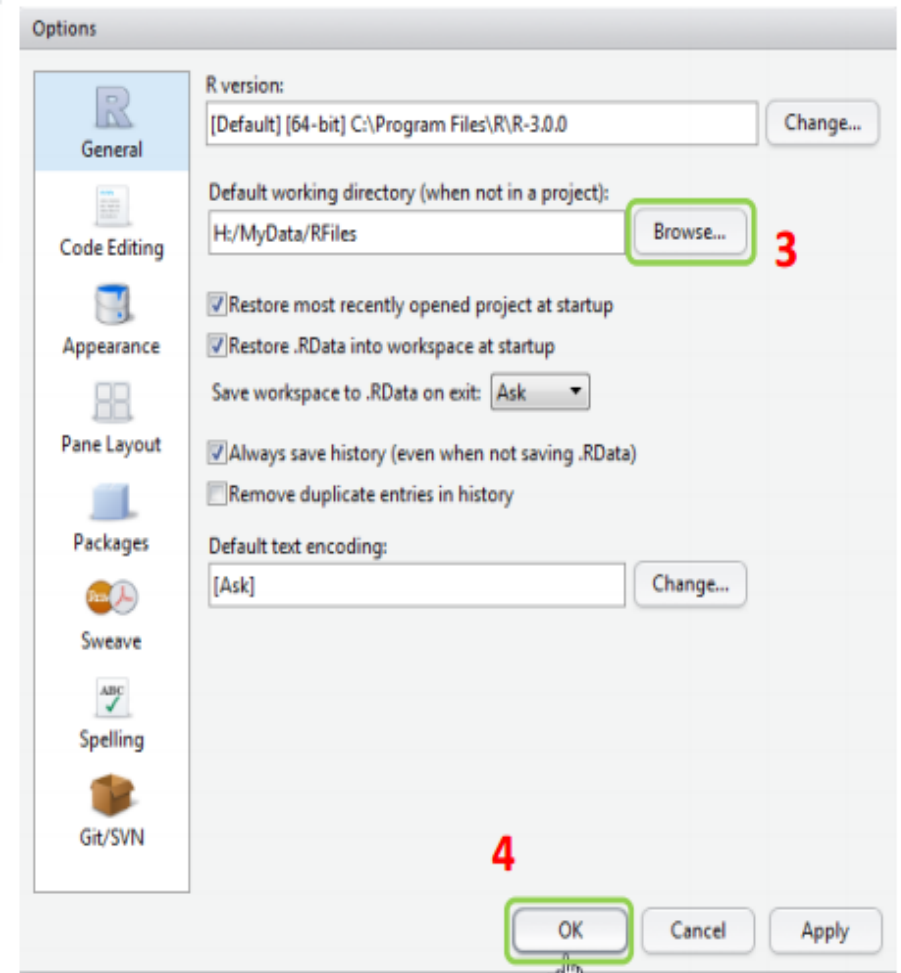
```
# Changes the wd  
setwd("C:/myfolder/data")
```



Setting a default working directory

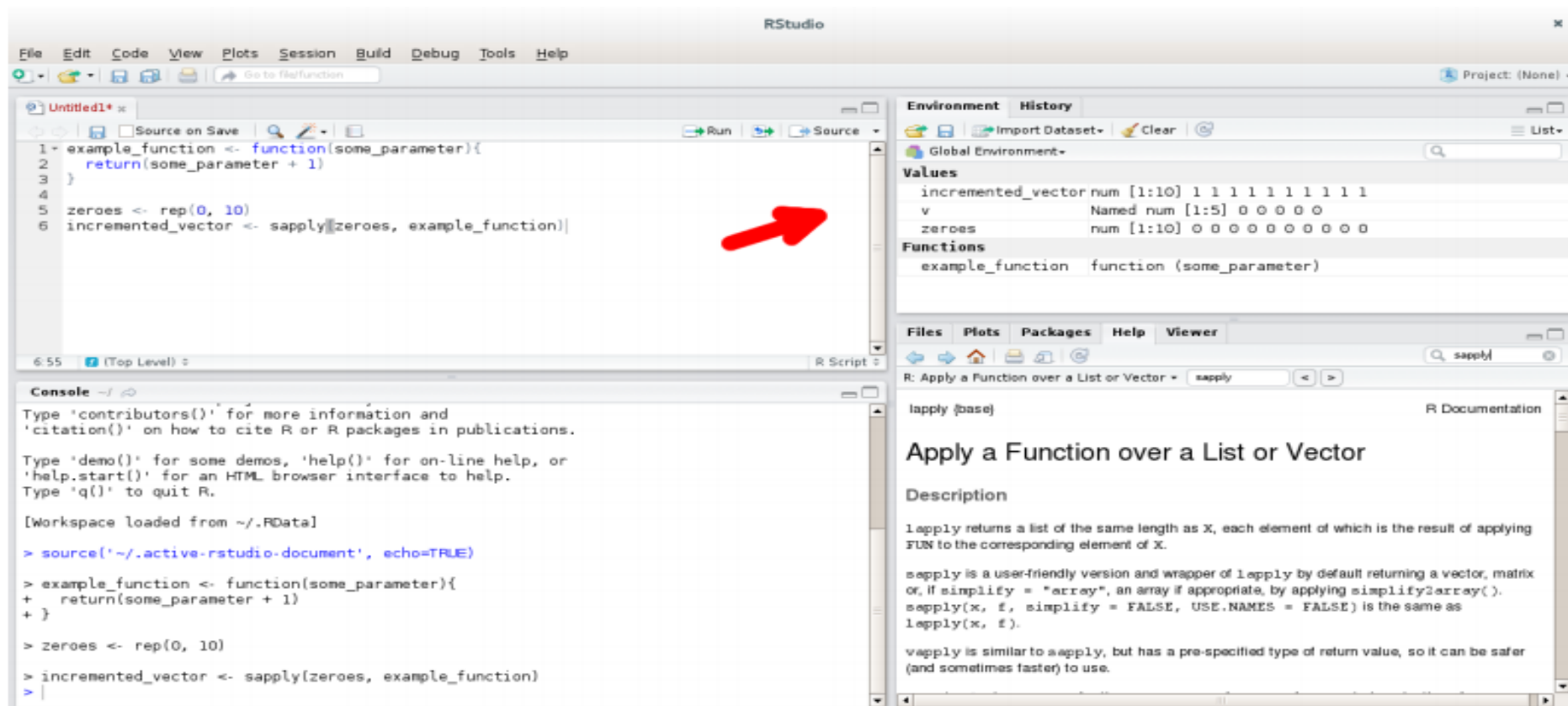


Every time you open RStudio, it goes to a default directory. You can change the default to a folder where you have your datafiles so you do not have to do it every time. In the menu go to Tools->Options



Global Environment

- The `globalenv()`, or global environment, is the interactive workspace. This is the environment in which you normally work.
- The enclosing environment for functions created in it, and the binding of the names to the values in this environment



The screenshot displays the RStudio interface. The main editor window shows an R script with the following code:

```
1- example_function <- function(some_parameter){  
2-   return(some_parameter + 1)  
3- }  
4-  
5- zeroes <- rep(0, 10)  
6- incremented_vector <- sapply(zeroes, example_function)
```

A red arrow points to the 'Global Environment' entry in the Environment pane on the right. The Environment pane shows the following values:

Variable	Value
incremented_vector	num [1:10] 1 1 1 1 1 1 1 1 1 1
v	Named num [1:5] 0 0 0 0 0
zeroes	num [1:10] 0 0 0 0 0 0 0 0 0 0

The Functions pane shows the `example_function` defined as `function (some_parameter)`.

The Console pane shows the output of the script execution:

```
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
[Workspace loaded from ~/.RData]  
  
> source('~/.active-rstudio-document', echo=TRUE)  
  
> example_function <- function(some_parameter){  
+   return(some_parameter + 1)  
+ }  
  
> zeroes <- rep(0, 10)  
  
> incremented_vector <- sapply(zeroes, example_function)  
>
```

The Files pane shows the `sapply` function being used in the script.

Data matrix

Data collected on students in a statistics class on a variety of variables:

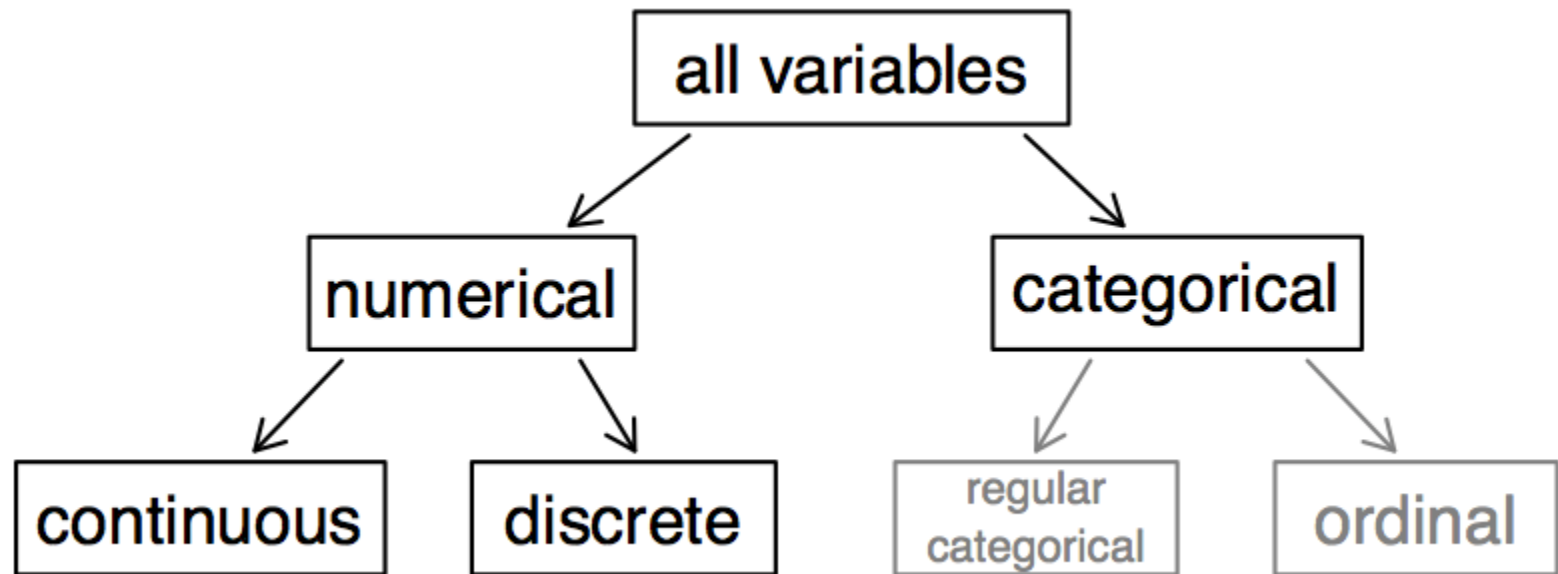
variable
↓

Stu.	gender	intro_extra	...	dread
1	male	extravert	...	3
2	female	extravert	...	2
3	female	introvert	...	4
4	female	extravert	...	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	...	3



←
observation

Types of variables



Quantitative Variable

A variable is called **quantitative** if observations take numerical values for different magnitudes of the variable.

Examples: Age, Number of siblings, and Annual Income



Categorical Variable



A variable is **categorical** if each observation belongs to one of a set of categories.

Examples:

Gender (Male or Female)

Religion (Catholic Jewish ...)(Catholic, Jewish,)

Type of residence (house, condominium, apartment, dormitory, other)

Belief in life after death (Yes, No)

Discrete Quantitative Variable

A quantitative variable is **discrete** if its possible values form a set of separate numbers:
 $0, 1, 2, 3, \dots$

Examples:

1. **Number** of pets in a household
2. **Number** of children in a family
3. **Number** of foreign languages spoken by an individual



Continuous Quantitative Variable



A quantitative variable is **continuous** if its possible values form an interval

Measurements

Examples:

Height/Weight

Age

Blood pressure

Types of variables (cont.)

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6.068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6.140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8.752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7.122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5.131	13.4	82.0	3.7	21070	45549	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Washakie	WY	8289	8533	8.714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6.695	7.9	77.9	6.5	28463	53853	none

Table 1.5: Seven rows from the county data set.

State- categorical

Federal- numerical, continuous

Smoking-ban- categorical, ordinal (could also be used as numerical)

Pop 2010- numerical, discrete

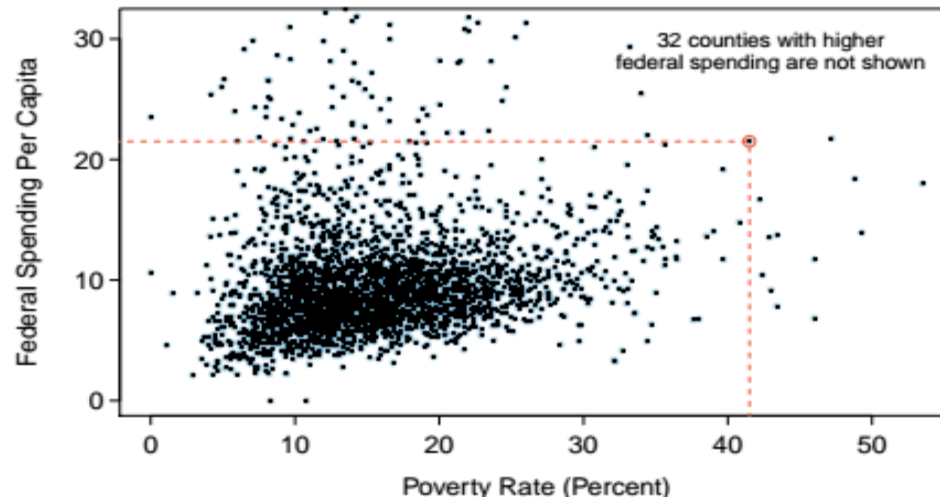
Associated vs. independent



- When two variables show some connection with one another, they are called **associated** variables.
 - Associated variables can also be called **dependent** variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be **independent**.

Relationships among variables

Does there appear to be a relationship between the fed-spend of study and the poverty ?



A scatterplot showing fed_sepend against poverty

Response and Explanatory Variables



Response variable

(dependent, y)

outcome variable

Explanatory variable

(independent, x)

defines groups

Response/Explanatory

1. Blood alcohol level/
of beers consumed
2. Grade on test/Amount
of study time
3. Yield of corn/Amount
of rainfall

Association

Association – When a value for one variable is more likely with certain values of the other variable

Data analysis with two variables

1. Tell whether there is an association and
2. Describe that association



Contingency Table

- Displays two categorical variables
- The rows list the categories of one variable; the columns list the other
- Entries in the table are frequencies

Food Type	Pesticides	
	Yes	No
Organic	29	98
Conventional	19485	7086
cell ↑		



Marginal proportion

The question, “Do organic and conventionally grown foods differ in the proportion of food items with pesticide residues?”

TABLE 3.1: Frequencies for Food Type and Pesticide Status.

The row totals and the column totals are the frequencies for the categories of each variable. The counts inside the table give information about the association.

Food Type	Pesticide Status		Total
	Present	Not Present	
Organic	29	98	127
Conventional	19485	7086	26571
Total	19514	7184	26698

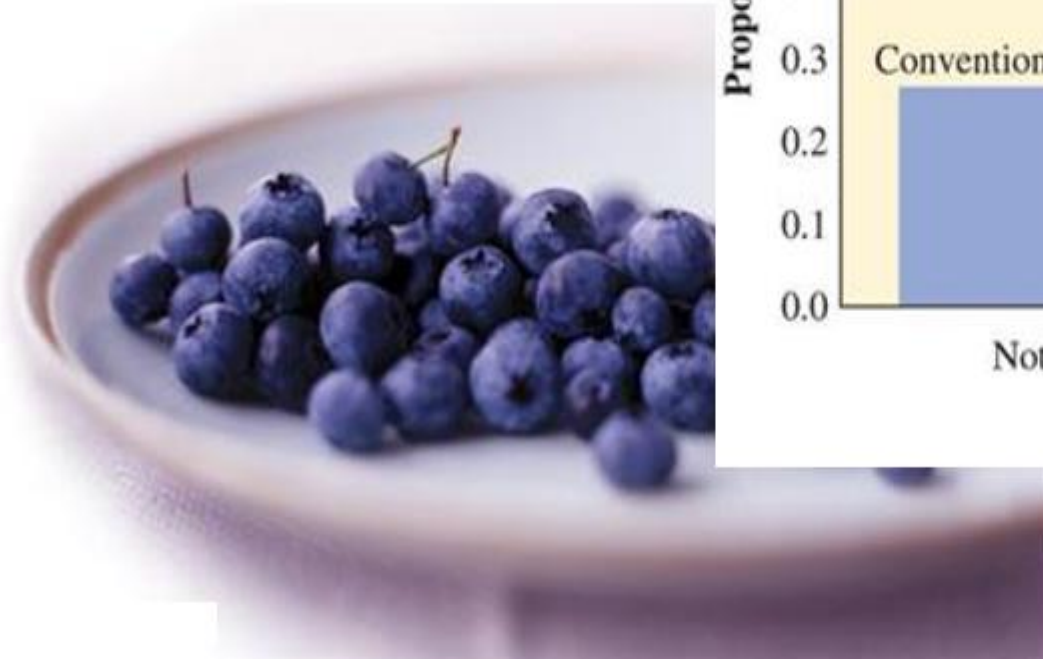
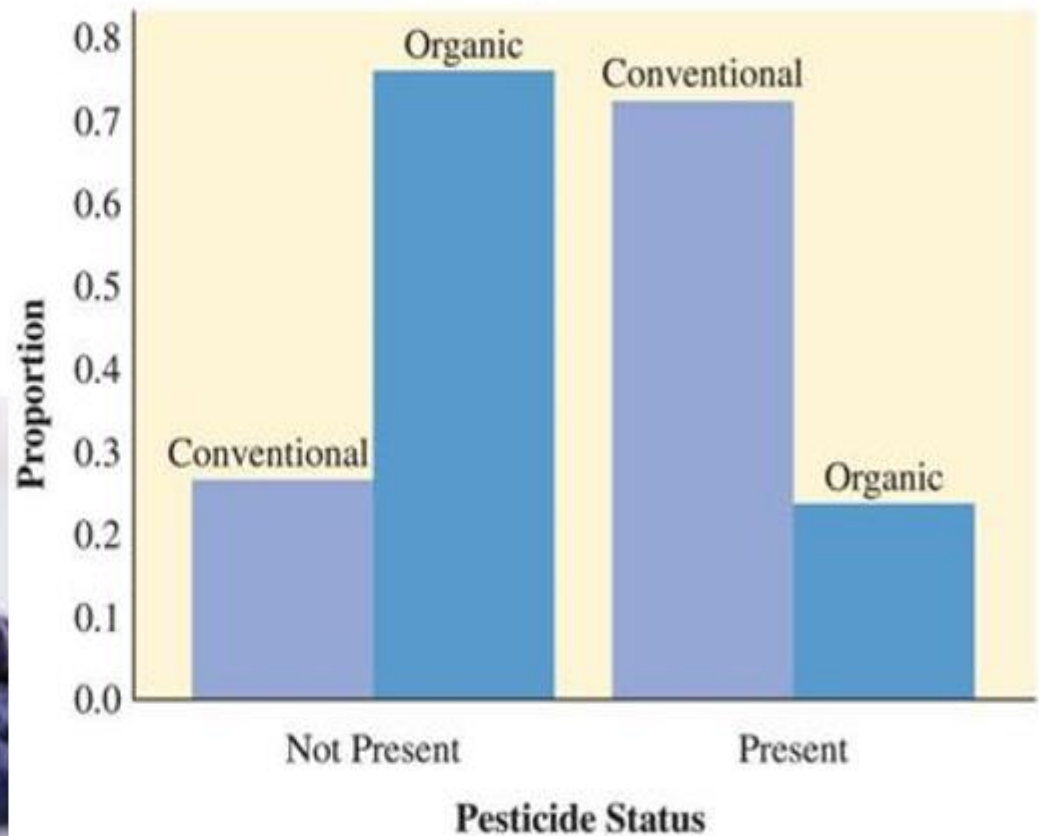
What is the **response** variable?

What is the **explanatory** variable

Proportions & Conditional Proportions

Side by side bar charts show conditional proportions and allow for easy comparison

Pesticide Status for Organic vs. Conventional Foods

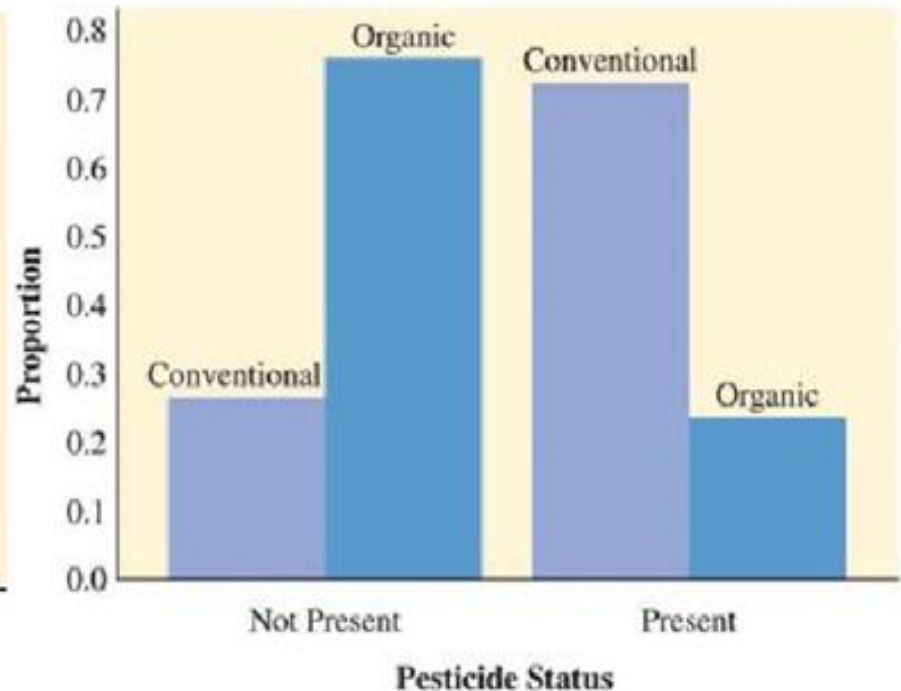
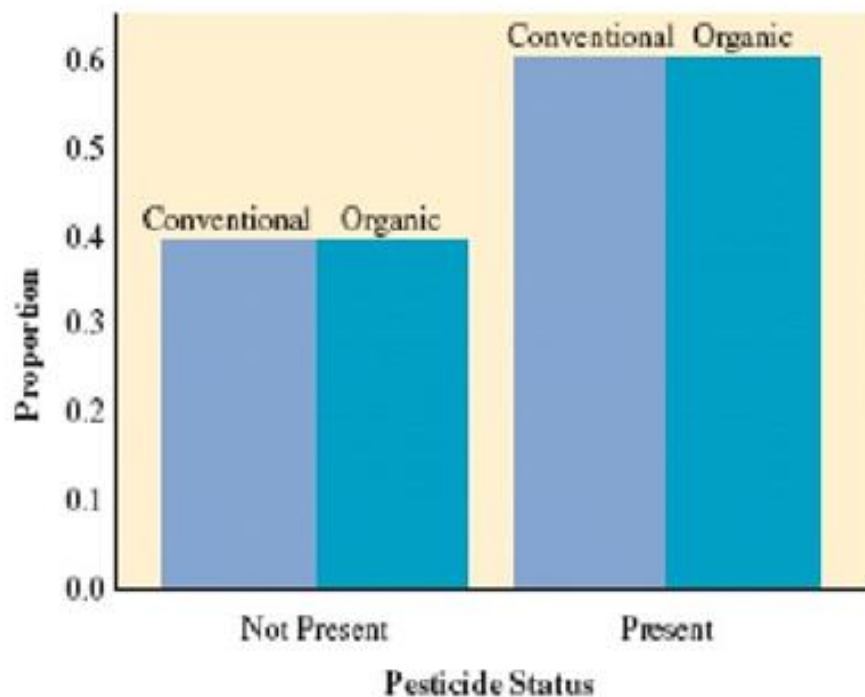


Proportions & Conditional Proportions

If no association, then
proportions would be
the same

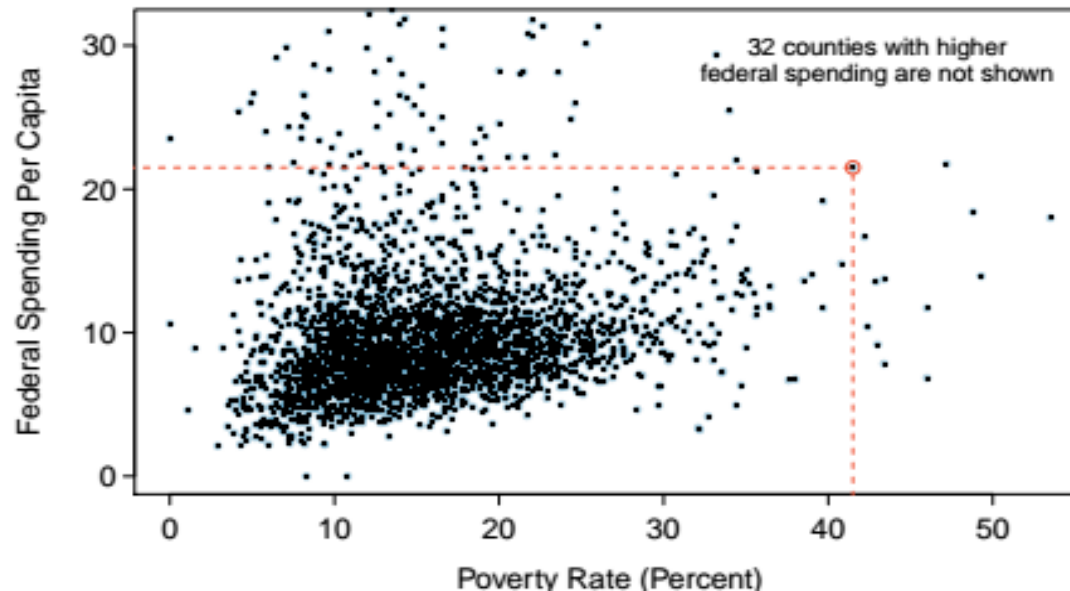
Since there is association,
then proportions are
different

Pesticide Status for Organic vs. Conventional Foods



Examining numerical data

Scatterplots provides a case-by-case view data for two numerical variable.



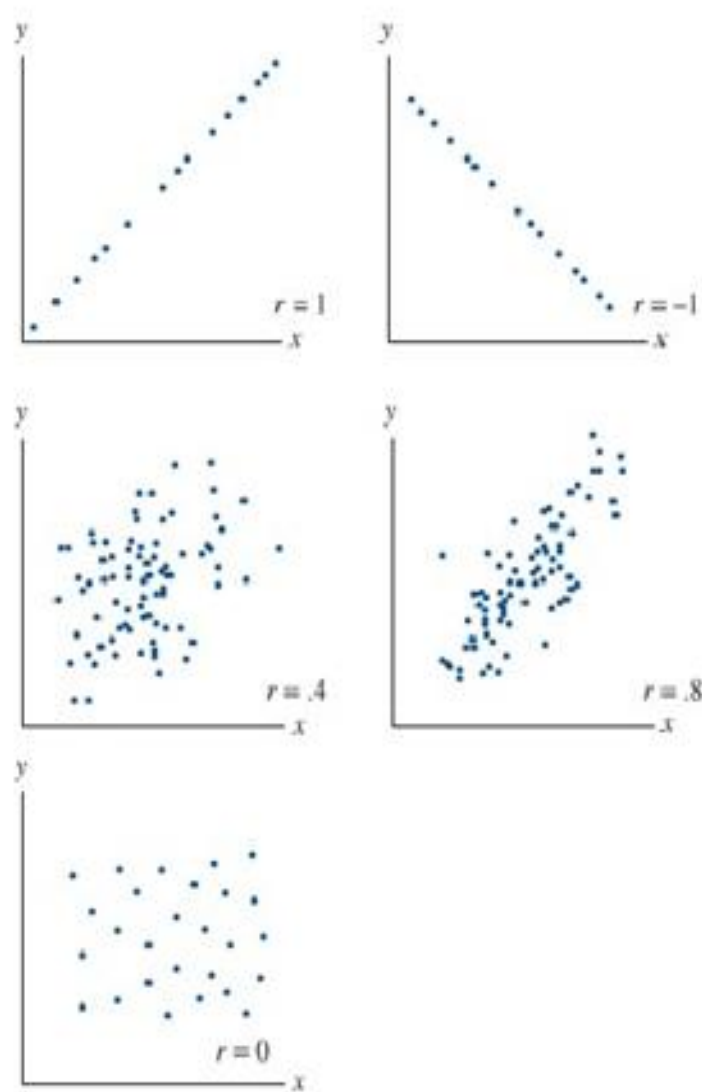
A scatterplot showing fed_sepend against poverty

Interpreting Scatterplots

The overall pattern includes trend, direction, and strength of the relationship

- **Trend**: linear, curved, clusters, no pattern
- **Direction**: positive, negative, no direction
- **Strength**: how closely the points fit the trend

Also look for outliers from the overall trend



Linear Correlation, r

Measures the strength and direction of the linear association between x and y

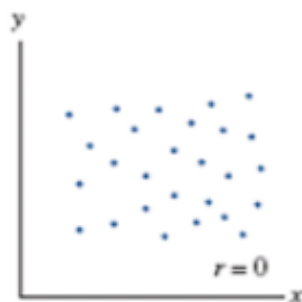
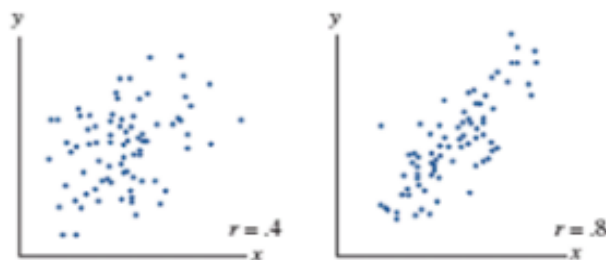
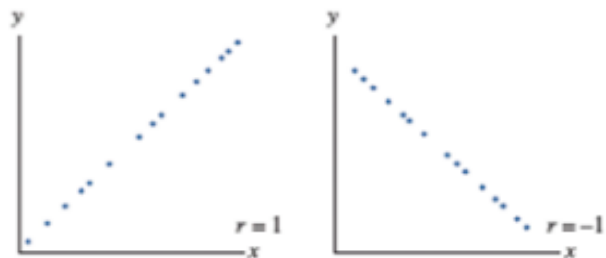
Calculating the Correlation r

$$r = \frac{1}{n-1} \sum z_x z_y = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

where n is the number of points, \bar{x} and \bar{y} are means, and s_x and s_y are standard deviations for x and y . The sum is taken over all n observations.

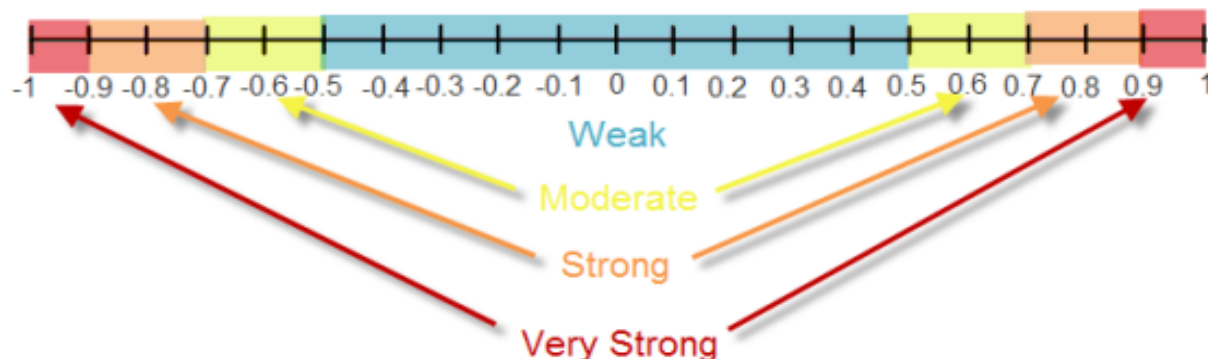
- ❖ **Positive** $r \Rightarrow$ positive association
- ❖ **Negative** $r \Rightarrow$ negative association
- ❖ r close to $+1$ or -1 indicates **strong** linear association
- ❖ r close to 0 indicates **weak** association

Correlation coefficient: Measuring Strength & Direction of a Linear Relationship



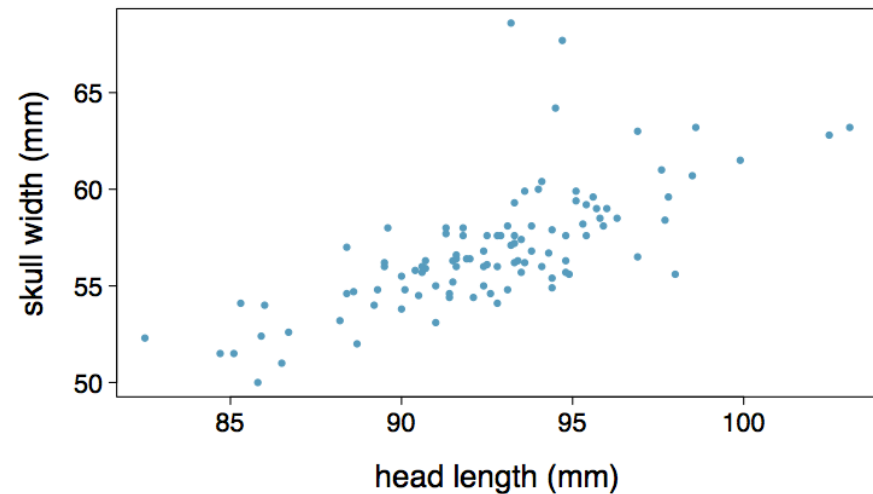
- ❖ **Positive** $r \Rightarrow$ positive association
- ❖ **Negative** $r \Rightarrow$ negative association
- ❖ r close to $+1$ or -1 indicates **strong** linear association
- ❖ r close to 0 indicates **weak** association

Strength of Correlation by r-Values



Practice

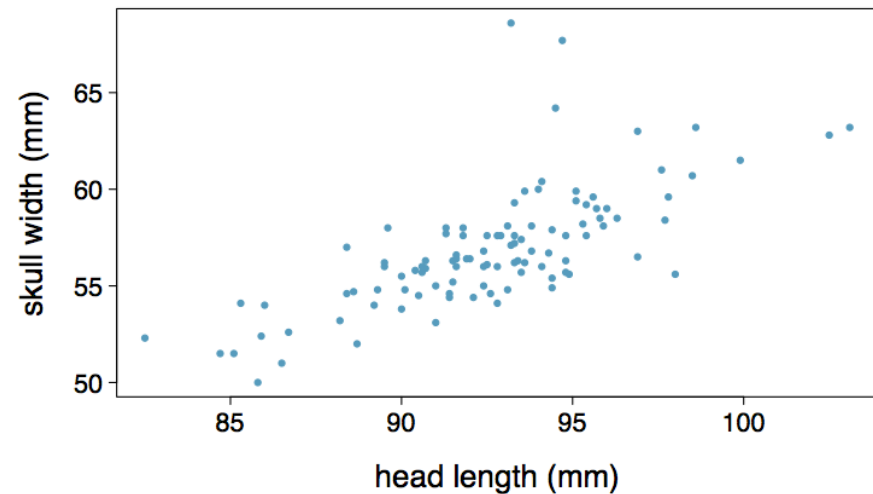
Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



- (a) There is no relationship between head length and skull width, i.e. the variables are independent.
- (b) Head length and skull width are positively associated.
- (c) Skull width and head length are negatively associated.
- (d) A longer head causes the skull to be wider.
- (e) A wider skull causes the head to be longer.

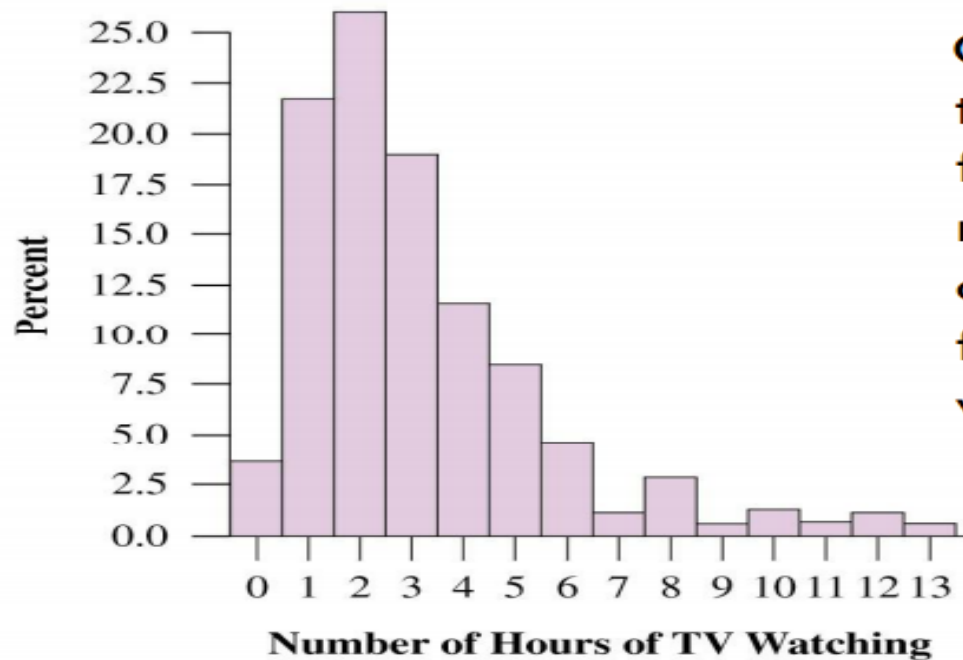
Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



- (a) There is no relationship between head length and skull width, i.e. the variables are independent.
- (b) *Head length and skull width are positively associated.***
- (c) Skull width and head length are negatively associated.
- (d) A longer head causes the skull to be wider.
- (e) A wider skull causes the head to be longer.

Histograms



Graph that uses bars to portray frequencies or relative frequencies of possible outcomes for a quantitative variable

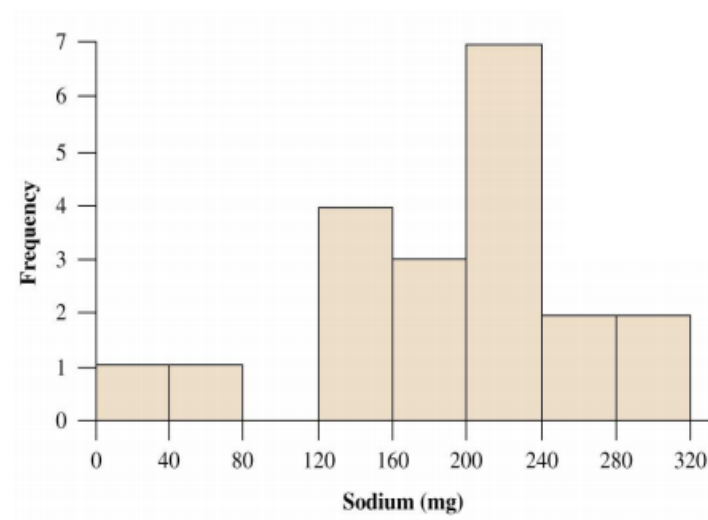
The 2004 General Social Survey asked, “On an average day, about how many hours do you personally watch television?” Figure shows the histogram of the 899 responses

Questions to Explore

- What was the most common outcome?
- What percentage of people reported watching TV no more than 2 hours per day?

Constructing a Histogram

1. Divide into intervals of equal width
2. Count # of observations in each interval
3. Label endpoints of intervals on horizontal axis
4. Draw a bar over each value or interval with height equal to its frequency (or percentage)
5. Label and title



Sodium in Cereals

Cereal	Sodium
Frosted Mini Wheats	0
Raisin Bran	340
All Bran	70
Apple Jacks	140
Cap'n Crunch	200
Cheerios	180
Cinnamon Toast Crunch	210
Crackling Oat Bran	150
Fiber One	100
Frosted Flakes	130
Froot Loops	140
Honey Bunches of Oats	180
Honey Nut Cheerios	190
Life	160
Rice Krispies	290
Honey Smacks	50
Special K	220
Wheaties	180
Corn Flakes	200
Honeycomb	210

Sodium in Cereals

0	210
260	125
220	290
210	140
220	200
125	170
250	150
170	70
230	200
290	180

Questions to Explore

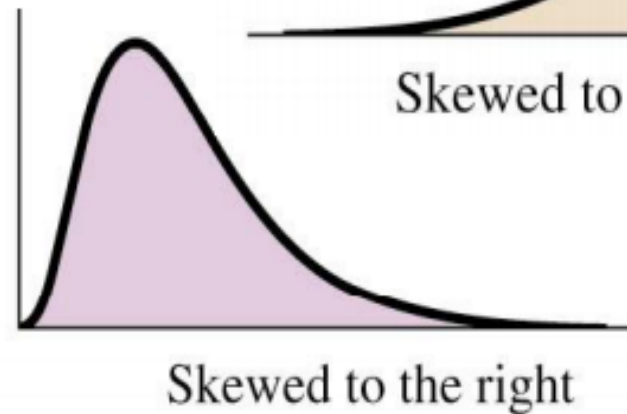
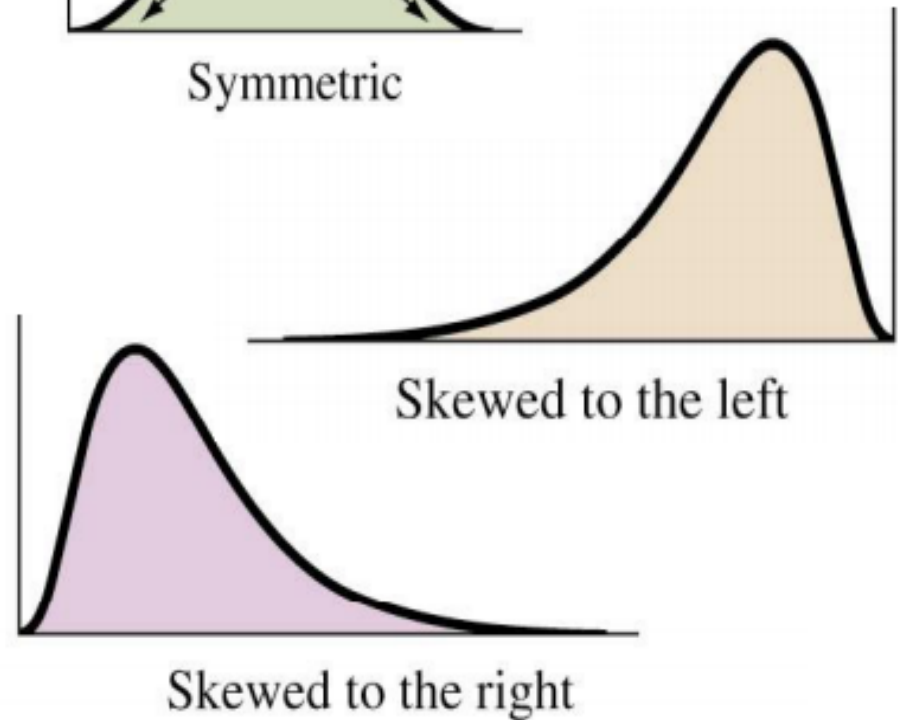
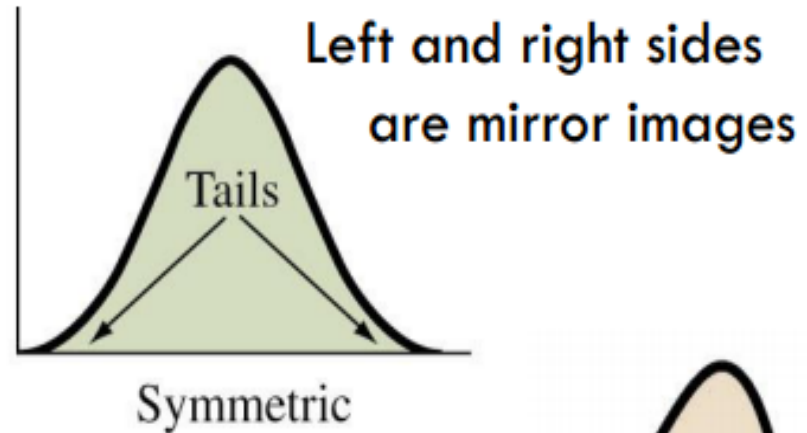
- A. Construct a histogram.
- B. Insight

The table summarizes the sodium values using eight intervals and lists the number of observations in each, as well as the proportions and percentages.

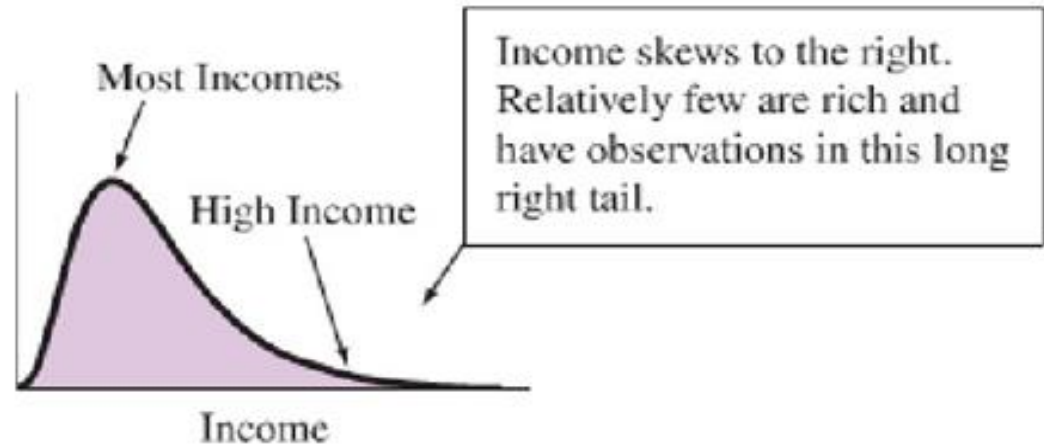
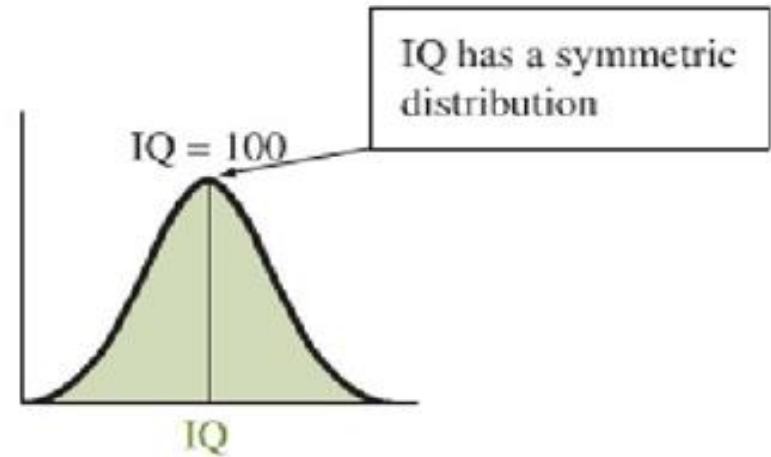
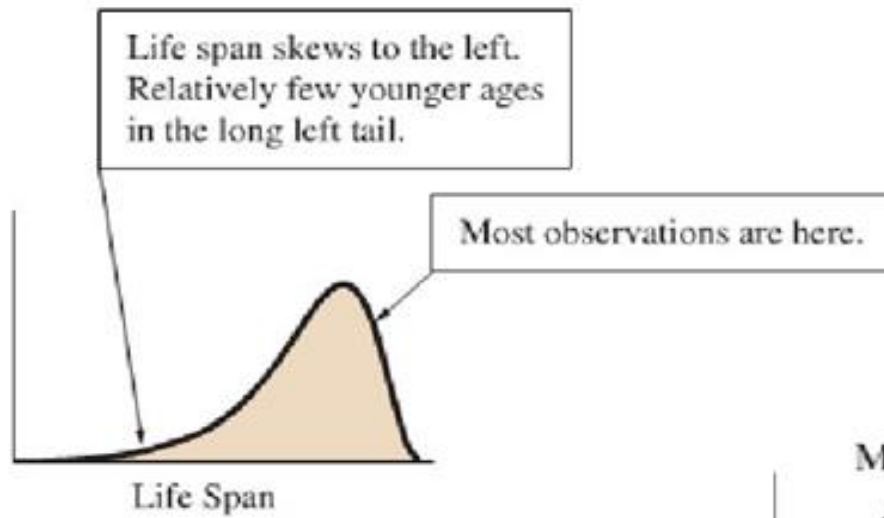
Interval	Frequency	Proportion	Percentage
0 to 39	1	0.05	5%
40 to 79	2	0.10	10%
80 to 119	1	0.05	5%
120 to 159	4	0.20	20%
160 to 199	5	0.25	25%
200 to 239	5	0.25	25%
240 to 279	0	0.00	0%
280 to 319	1	0.05	5%
320 to 359	1	0.05	5%

Interpreting Histograms

- ❑ Assess where a distribution is **centered** by finding the median
- ❑ Assess the **spread** of a distribution
- ❑ **Shape** of a distribution: roughly symmetric, skewed to the right, or skewed to the left



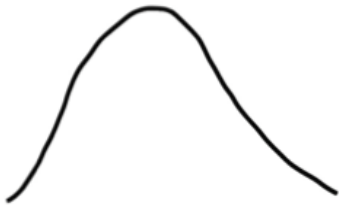
Examples of Skewness



Commonly observed shapes of distributions

Modality

unimodal



bimodal



multimodal



uniform



Skewness

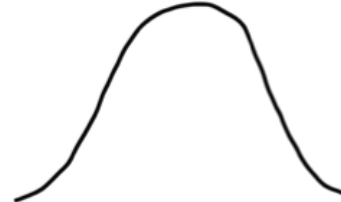
right skew



left skew

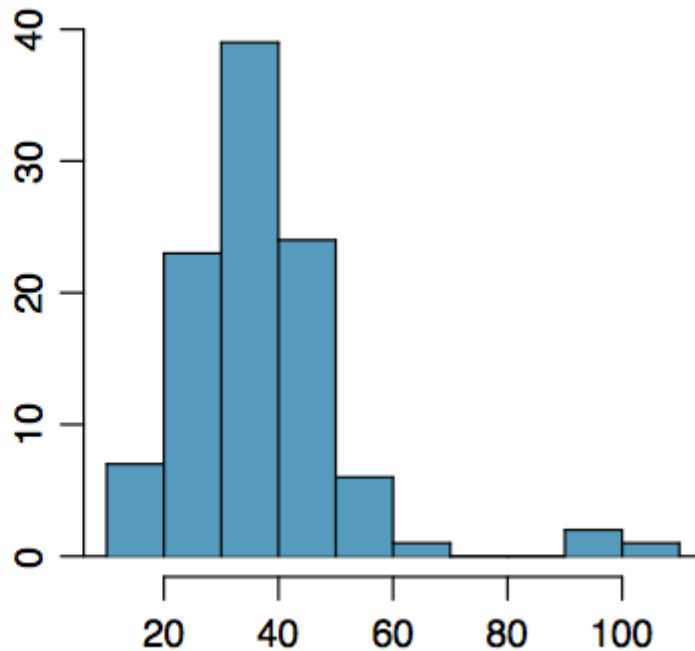
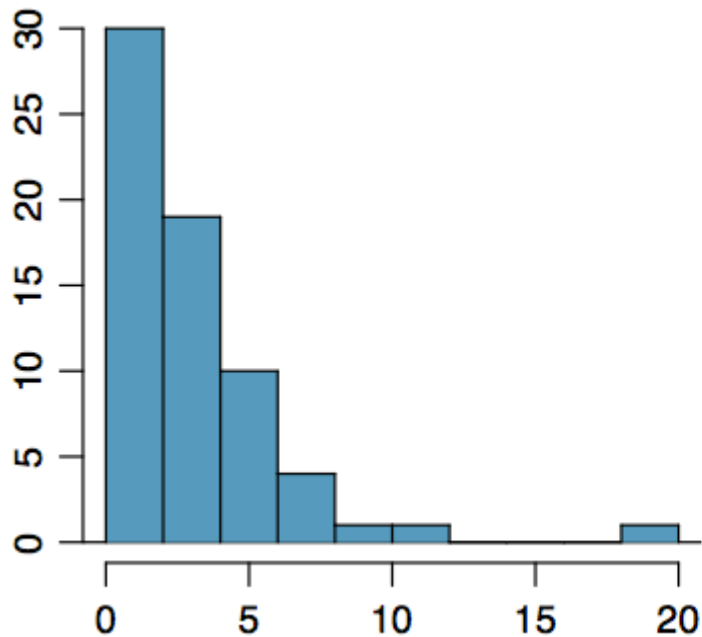


symmetric



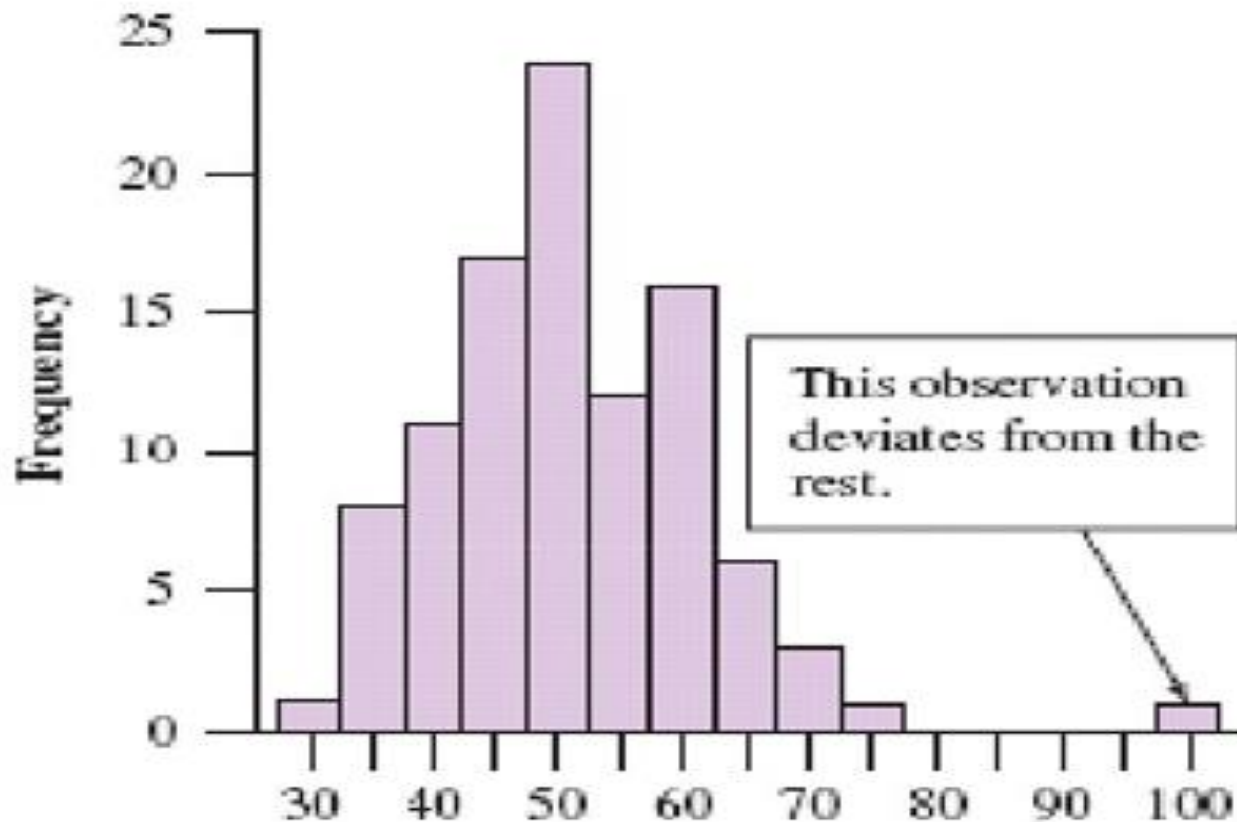
Shape of a Distribution: Unusual Observations

Are there any unusual observations or potential outliers?



Outlier

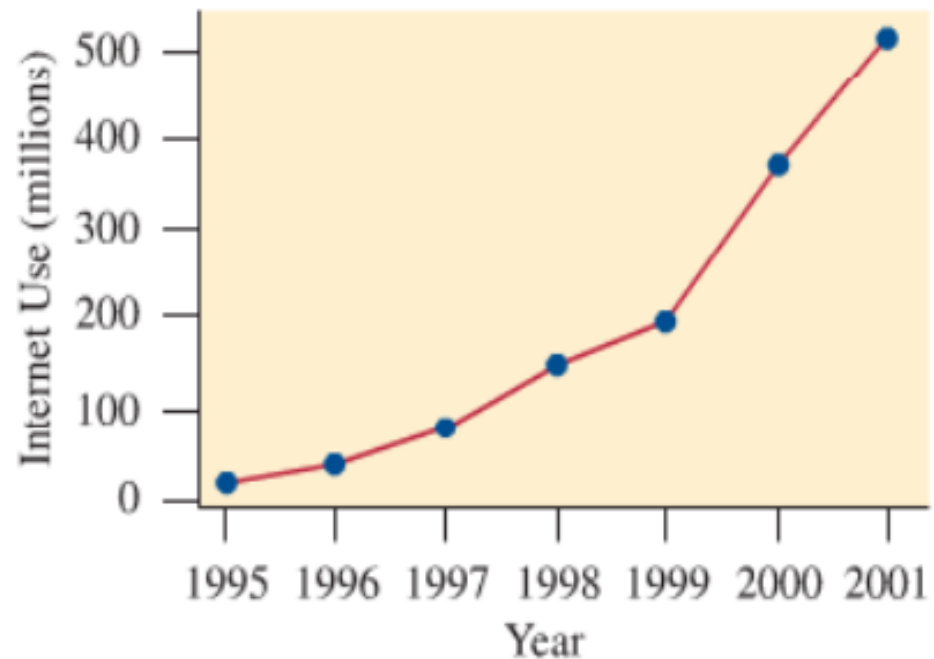
- An outlier falls far from the rest of the data



Time Plots

- Display a **time series**, data collected over time
- Plots observation on the vertical against time on the horizontal
- Points are usually connected
- Common patterns should be noted

Time Plot from 1995 – 2001
of the # worldwide who
use the Internet



Practice



Which of these variables do you expect to be uniformly distributed?

- (a) weights of adult females
- (b) salaries of a random sample of people from North Carolina
- (c) house prices
- (d) birthdays of classmates (day of the month)

Practice

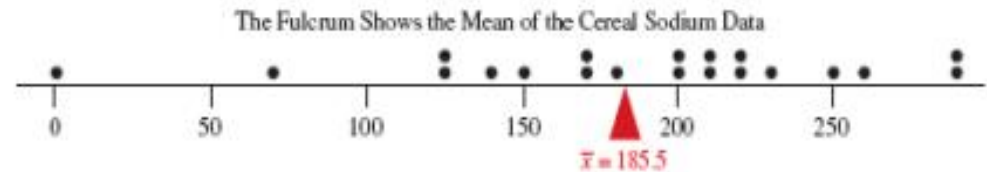
Which of these variables do you expect to be uniformly distributed?

- (a) weights of adult females
- (b) salaries of a random sample of people from North Carolina
- (c) house prices
- (d) birthdays of classmates (day of the month)*

Mean

- The mean is the sum of the observations divided by the number of observations
- It is the center of mass

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



Cereal	Sodium
Frosted Mini Wheats	0
Raisin Bran	210
All Bran	260
Apple Jacks	125
Capt Crunch	220
Cheerios	290
Cinnamon Toast	210
Crackling Oat Bran	140
Crispix	220
Frosted Flakes	200
Fruit Loops	125
Grape Nuts	170
Honey Nut Cheerios	250
Life	150
Oatmeal Raisin Crisp	170
Sugar Smacks	70
Special K	230
Wheaties	200
Corn Flakes	290
Honeycomb	180

Median

The **median** is the value that splits the data in half when ordered in ascending order.

0, 1, **2**, 3, 4

If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2}, \underline{3}, 4, 5 \rightarrow \frac{2 + 3}{2} = \mathbf{2.5}$$

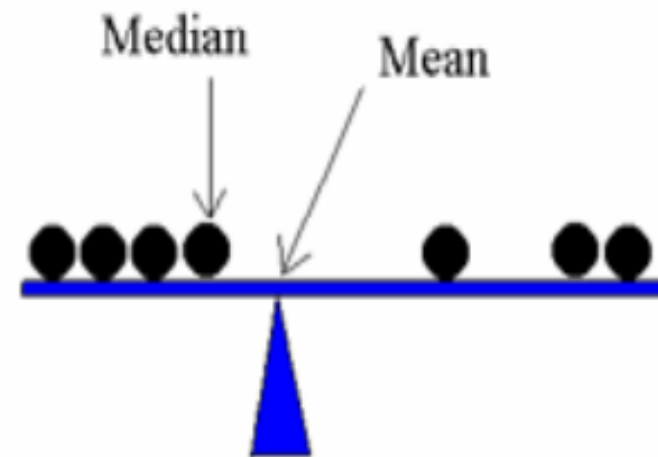
Since the median is the midpoint of the data 50% of the values are below it. Hence, it is also the **50th percentile**.

Order	Data
1	78
2	91
3	94
4	98
5	99
6	101
7	103
8	105
9	114

Order	Data
1	78
2	91
3	94
4	98
5	99
6	101
7	103
8	105
9	114
10	121

Resistant Measures

- A measure is **resistant** if extreme observations (outliers) have little, if any, influence on its value
 - ▣ Median is resistant to outliers
 - ▣ Mean is not resistant to outliers

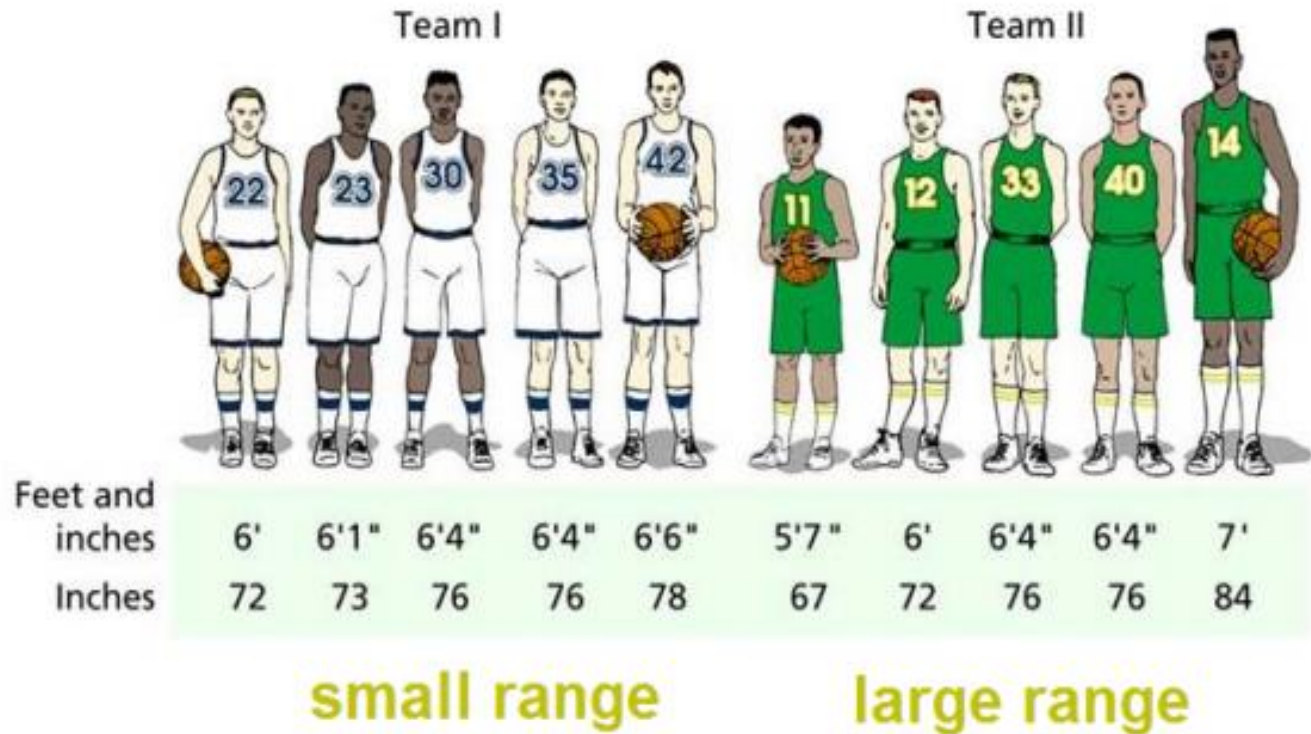


www.stat.psu.edu

Range

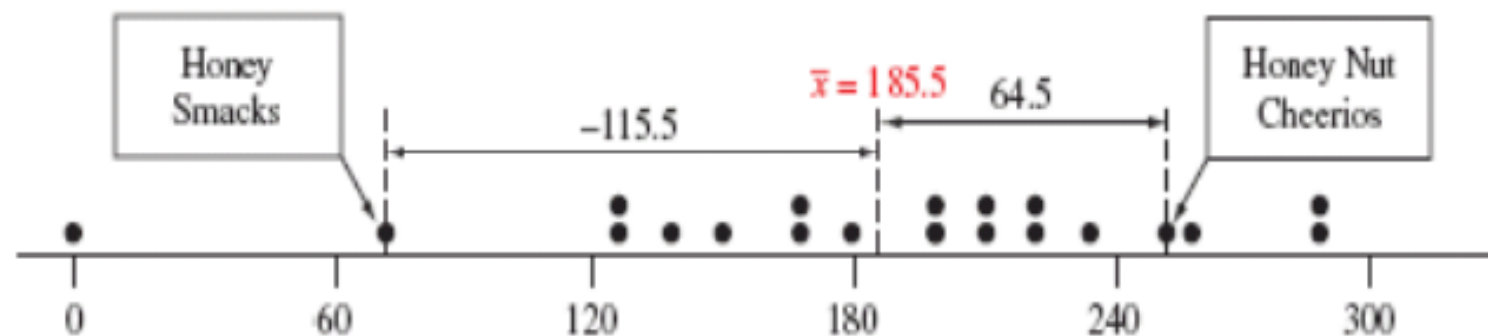
$$\text{Range} = \text{max} - \text{min}$$

The range is strongly affected by outliers.



Standard Deviation

- Each data value has an associated **deviation** from the mean, $x - \bar{x}$
- A deviation is **positive** if it falls **above** the mean and **negative** if it falls **below** the mean
- The sum of the deviations is always **zero**

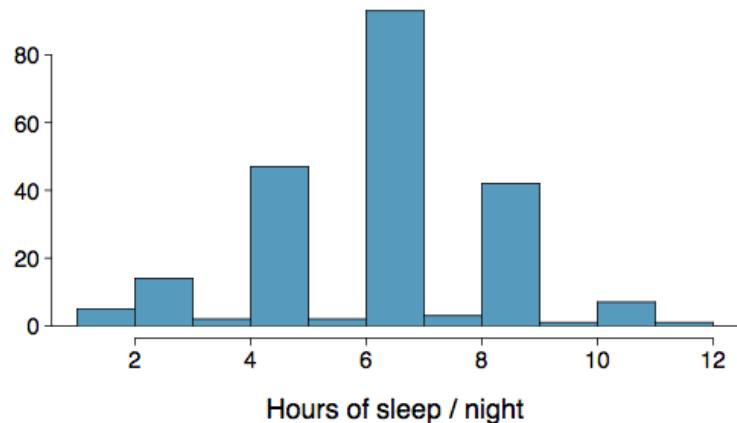


Variance

Variance is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.
- The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5 - 6.71)^2 + (9 - 6.71)^2 + \dots + (7 - 6.71)^2}{217 - 1} = 4.11 \text{ hours}^2$$

Variance (cont.)



Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

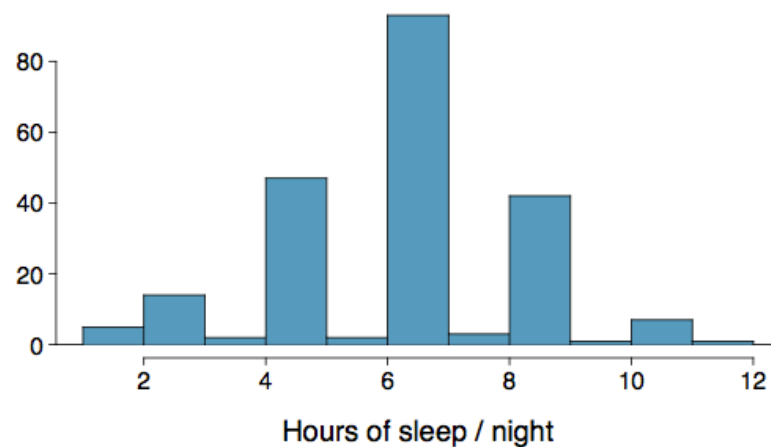
Standard Deviation

The **standard deviation** is the square root of the variance, and has the same units as the data.

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

The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$



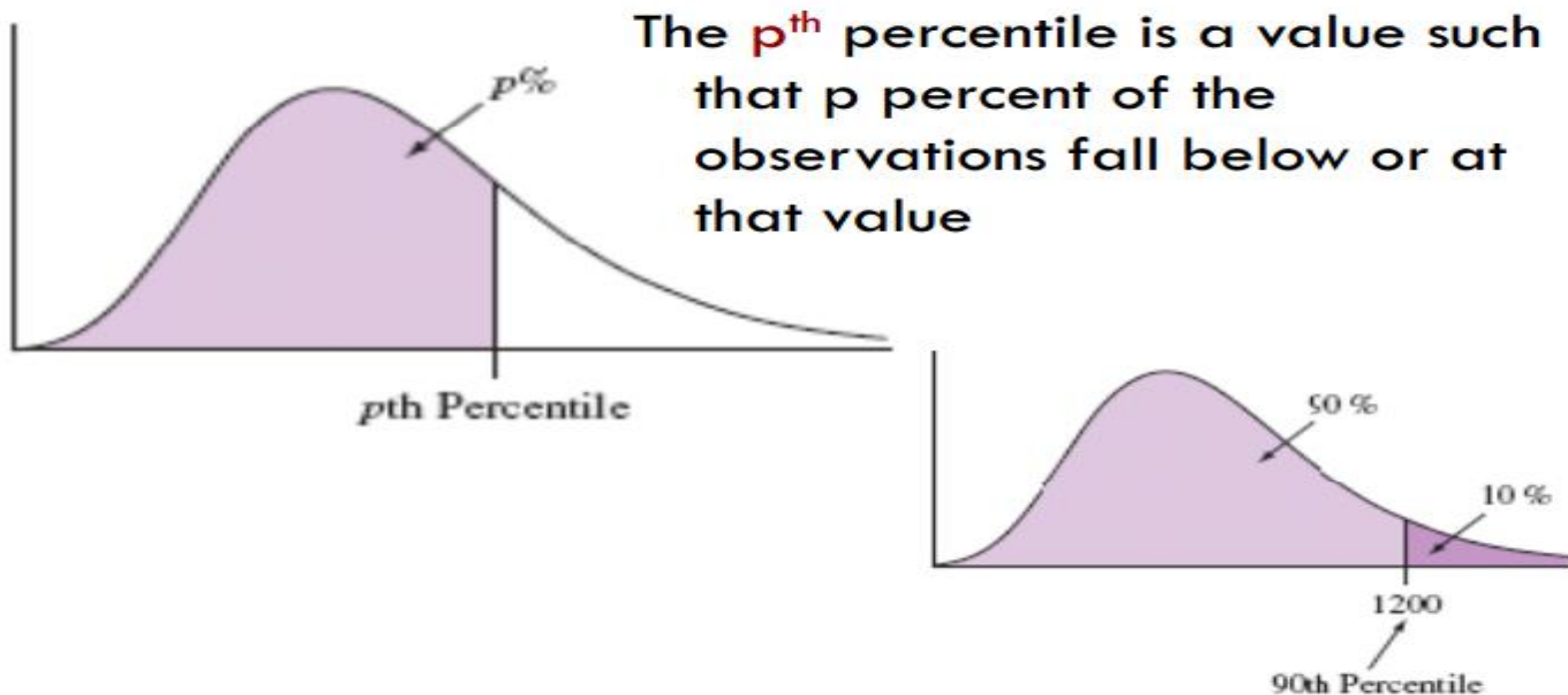
We can see that all of the data are within 3 standard deviations of the mean.

Properties of Sample Standard Deviation

1. Measures spread of data
2. Only zero when all observations are same; otherwise, $s > 0$
3. As the spread increases, s gets larger
4. Same units as observations
5. Not resistant
6. Strong skewness or outliers greatly increase s

Percentile

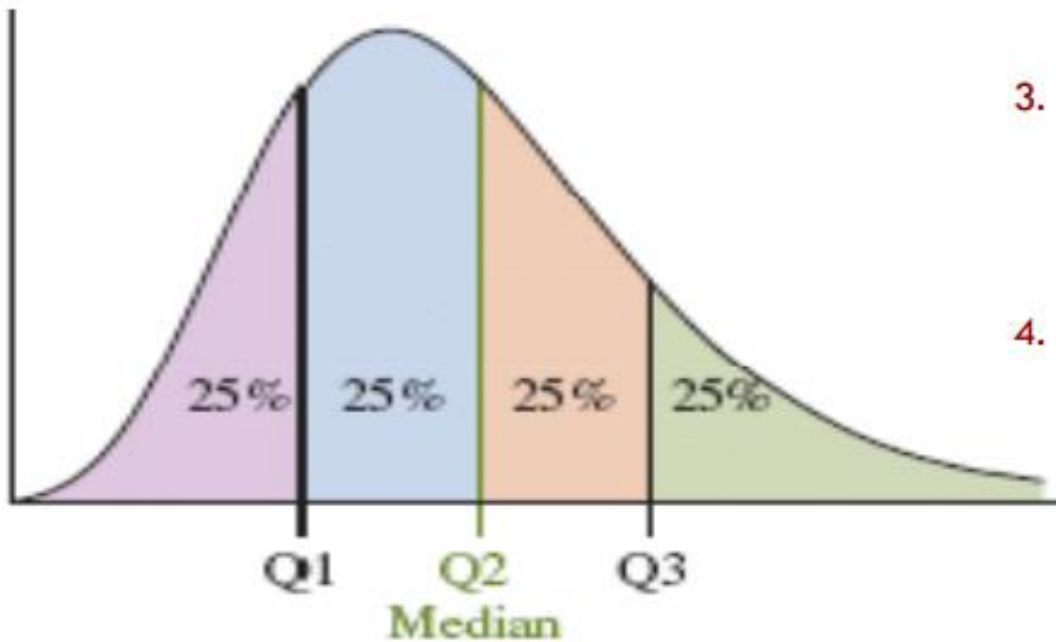
Suppose you're informed that your score of 1200 (out of 1600) on the SAT college entrance exam falls at the 90th percentile. Set $p=90$ in this definition. Then, 90% of those who took the exam scored between the minimum score and 1200. Only 10% of the scores were higher than yours.



Finding Quartiles

Splits the data into four parts

1. Arrange data in order
2. The **median** is the second quartile, Q_2
3. Q_1 is the **median of the lower half** of the observations
4. Q_3 is the **median of the upper half** of the observations



Q1, Q3, and IQR

Quartiles divide a ranked data set into four equal parts:

1. 25% of the data at or below Q_1 and 75% above
2. 50% of the obs are above the **median** and 50% are below
3. 75% of the data at or below Q_3 and 25% above

$Q_1 = \text{first quartile} = 2.2$

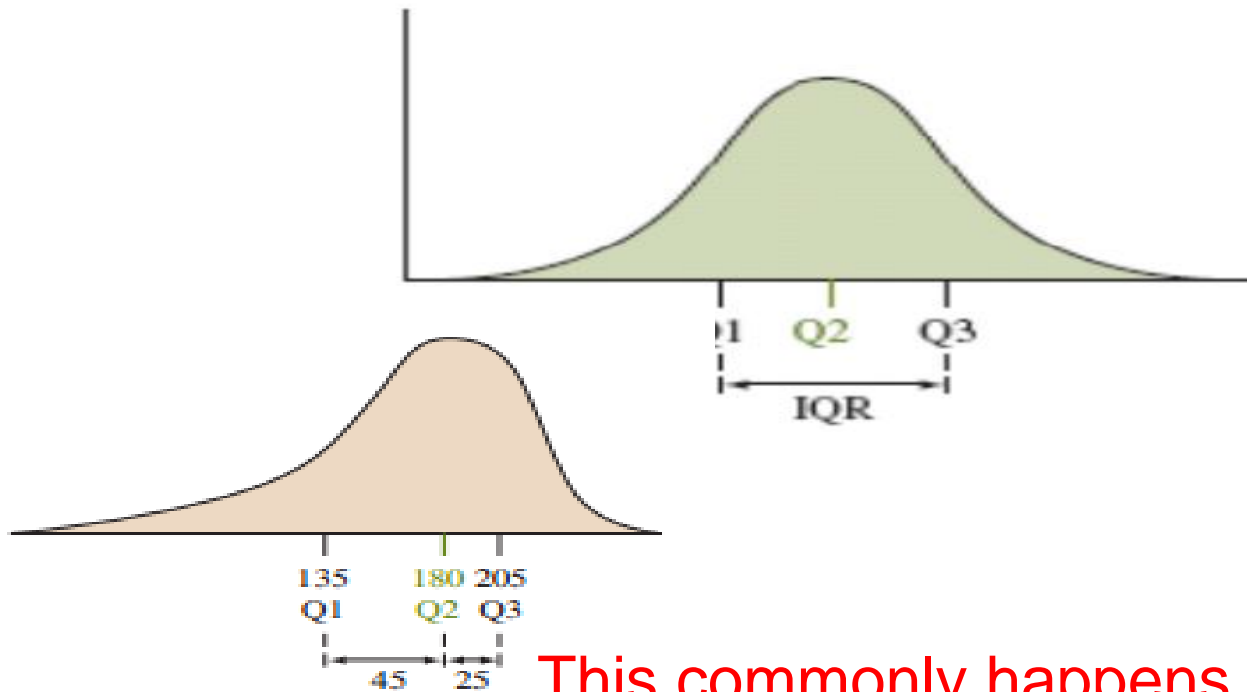
$M = \text{median} = 3.4$

$Q_3 = \text{third quartile} = 4.35$

1	0.6
2	1.2
3	1.6
4	1.9
5	1.5
6	2.1
7	2.3
8	2.3
9	2.5
10	2.8
11	2.9
12	3.3
13	3.4
14	3.6
15	3.7
16	3.8
17	3.9
18	4.1
19	4.2
20	4.5
21	4.7
22	4.9
23	5.3
24	5.6
25	6.1

Calculating Interquartile Range

The **interquartile range** is the distance between the third and first quartile, giving spread of middle 50% of the data: **$IQR = Q3 - Q1$**



This commonly happens when the distribution is skewed to the left, as shown in the margin figure

Criteria for Identifying an Outlier

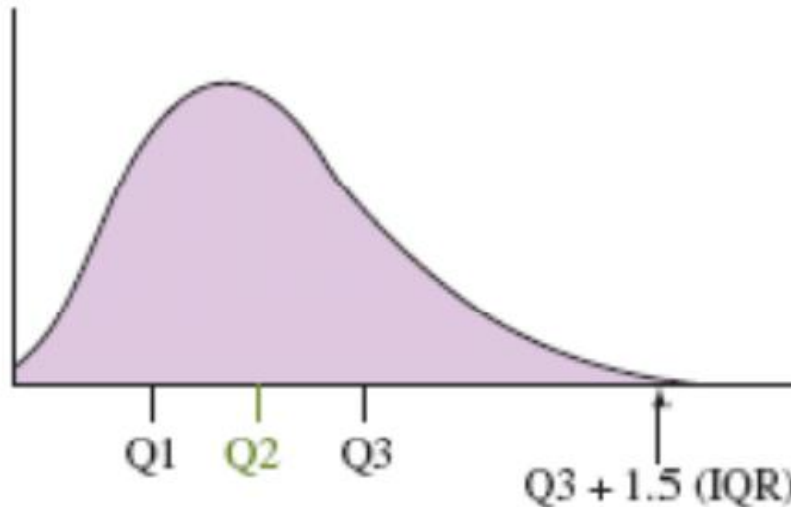
An observation is a **potential outlier** if it falls more than $1.5 \times IQR$ below the first or more than $1.5 \times IQR$ above the third quartile.

First quartile $Q1 = 135$

Median = 180

Third quartile $Q3 = 205$

Maximum value = 340

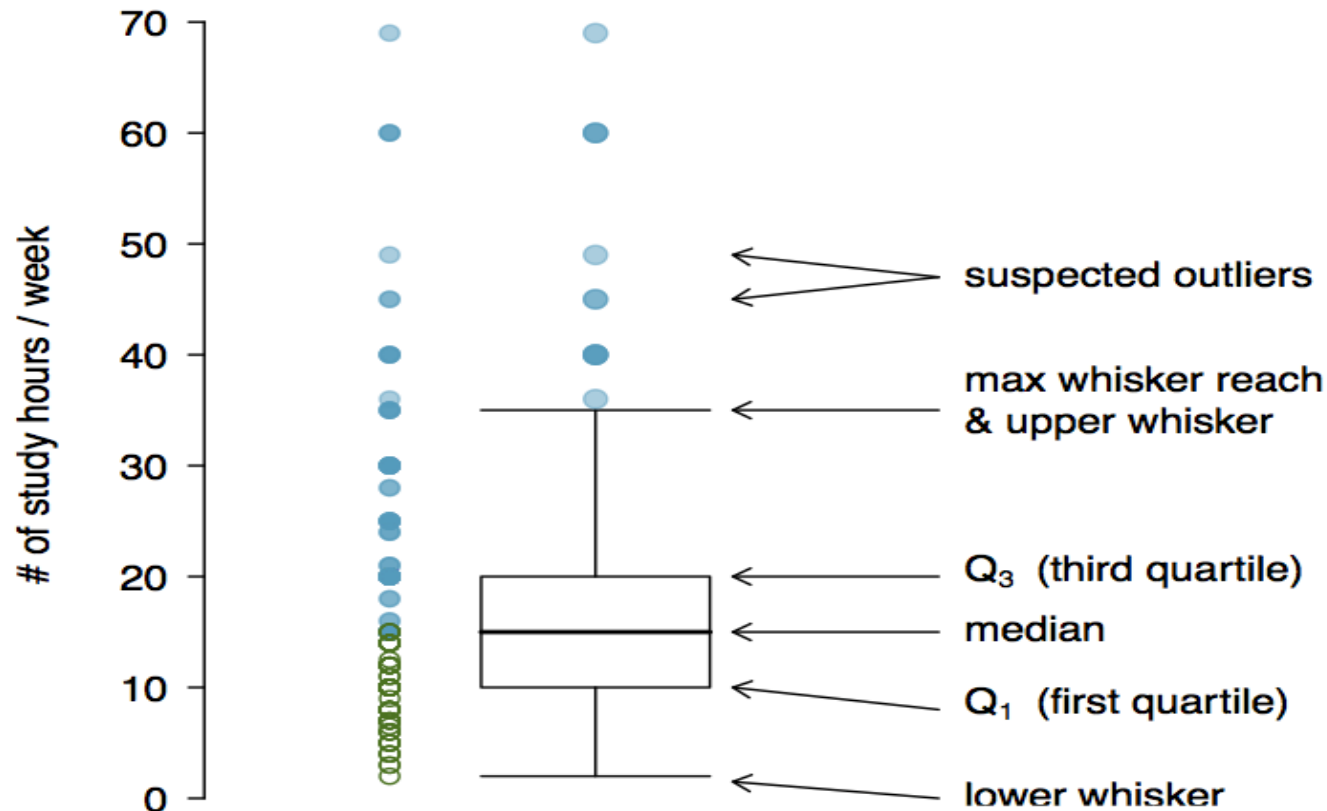


the breakfast cereal sodium data has $Q1=135$ and $Q3=205$. So, $IQR=Q3-Q1=205-135=70$. For those data $1.5 \times IQR = 1.5 \times 70 = 105$. (lower boundary, potential outliers below), $Q3 + 1.5 \times IQR = 205 + 105 = 310$ (upper boundary, potential outliers above).

observations below 30 or above 310 are potential outliers.

Anatomy of a Box Plot

The box in a **box plot** represents the middle 50% of the data, and the thick line in the box is the median.

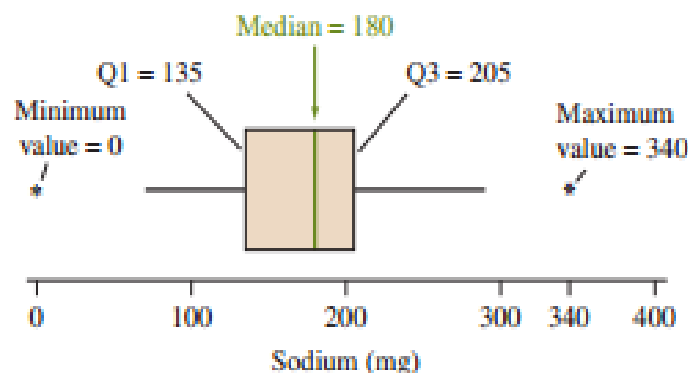


Whiskers and Outliers

Whiskers of a box plot can extend up to $1.5 \times \text{IQR}$ away from the quartiles.

max upper whisker reach = $Q3 + 1.5 \times \text{IQR}$

max lower whisker reach = $Q1 - 1.5 \times \text{IQR}$



$Q1 = 135$ mg and $Q3 = 205$ mg. Thus $\text{IQR} = 205 - 135 = 70$ mg, and the lower and upper boundaries are $135 - 1.5 \times 70 = 30$ mg and $205 + 1.5 \times 70 = 310$ mg,

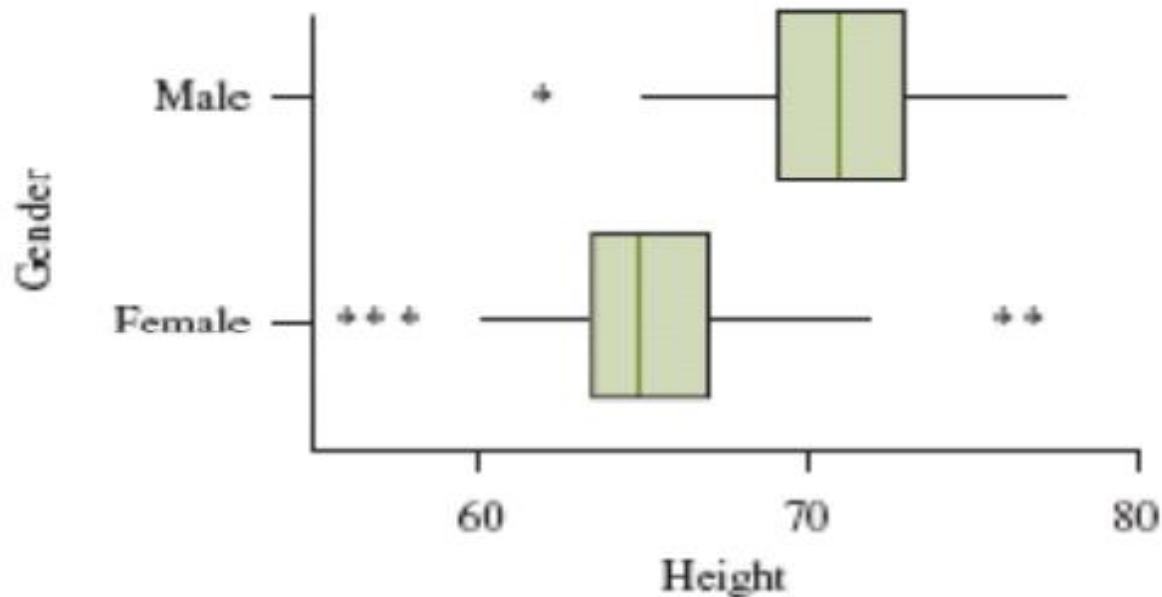
Outliers (cont.)

Why is it important to look for outliers?

- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Comparing Distributions

Boxplots do not display the shape of the distribution as clearly as histograms, but are useful for making graphical comparisons of two or more distributions



Box Plots of Male and Female College Student Heights

Robust Statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

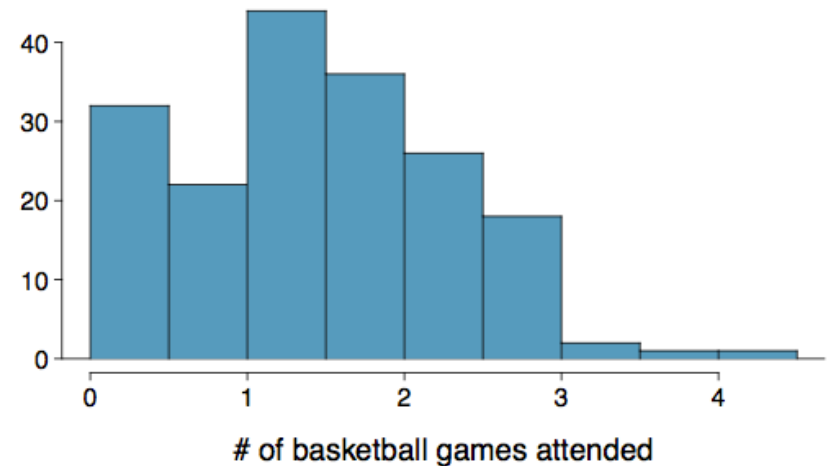
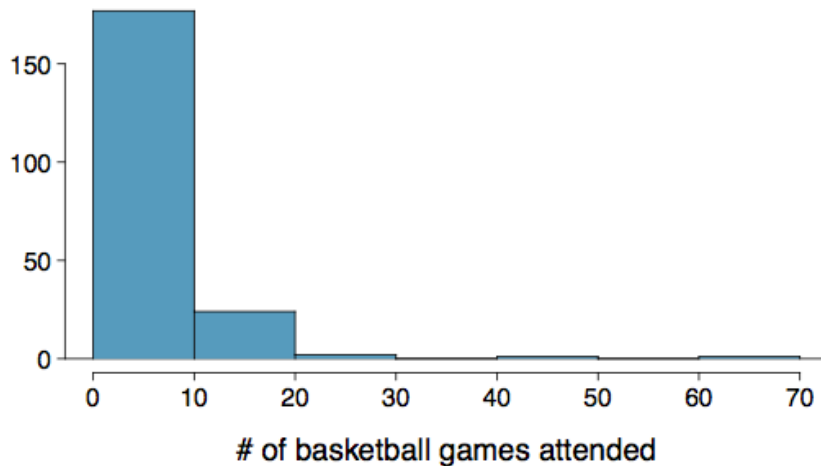
If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Median

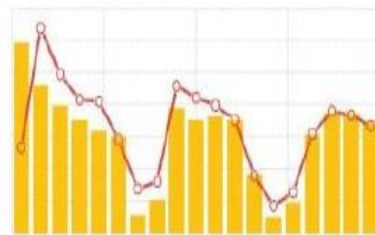
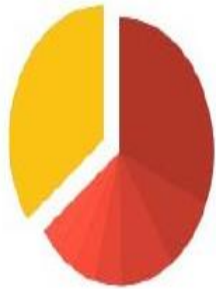
Extremely Skewed Data

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the [log transformation](#).

The histograms on the left shows the distribution of number of basketball games attended by students. The histogram on the right shows the distribution of log of number of games attended.



Graphs for Categorical Variables



2.

Use pie charts and bar graphs to summarize categorical variables

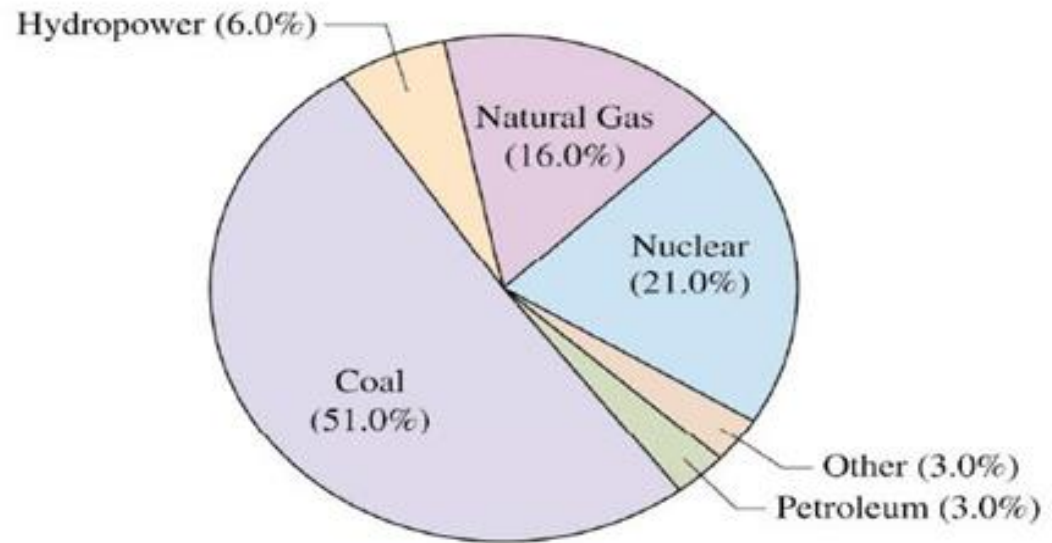
1. **Pie Chart:** A circle having a “slice of pie” for each category
2. **Bar Graph:** A graph that displays a vertical bar for each category

Pie Charts

- Summarize categorical variable
- Drawn as circle where each category is a slice
- The size of each slice is proportional to the percentage in that category

Source	U.S. Percentage
Coal	51
Hydropower	6
Natural gas	16
Nuclear	21
Petroleum	3
Other	3
Total	100

Percentage Use for Sources of Electricity

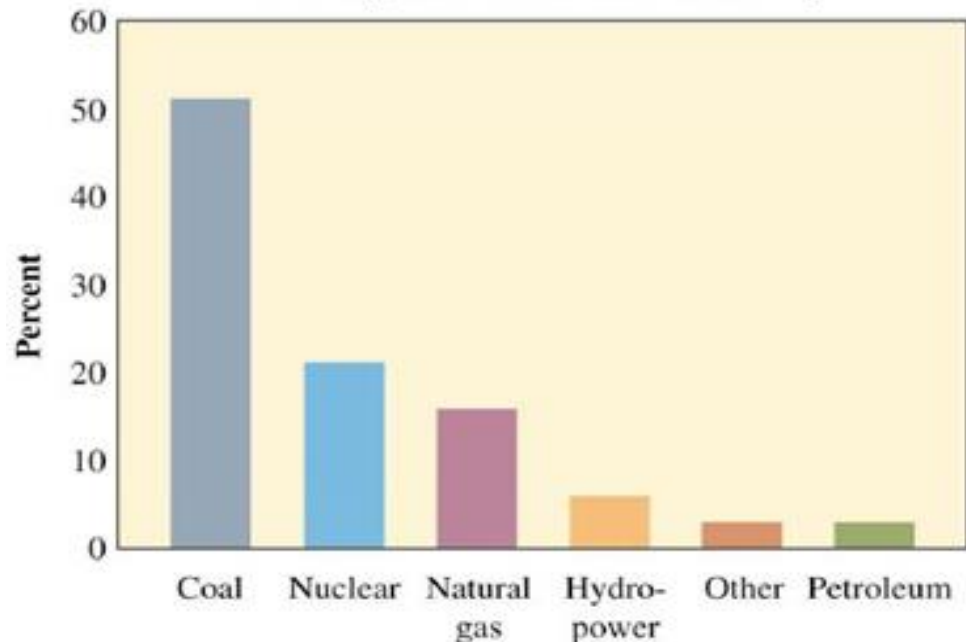


Bar Graphs

- Summarizes categorical variable
- Vertical bars for each category
- Height of each bar represents either counts or percentages
- Easier to compare categories with bar graph than with pie chart
- Called **Pareto Charts** when ordered from tallest to shortest

Source	U.S. Percentage
Coal	51
Hydropower	6
Natural gas	16
Nuclear	21
Petroleum	3
Other	3
Total	100

Percentage use for sources of electricity



Experimental Design Terminology...

Placebo: fake treatment, often used as the control group for medical studies

Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment

Blinding: when experimental units do not know whether they are in the control or treatment group

Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Prospective vs. Retrospective Studies

A **prospective study** identifies individuals and collects information as events unfold.

- Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.

Retrospective studies collect data after events have taken place.

- Example: Researchers reviewing past events in medical records.

More on Blocking



We would like to design an experiment to investigate if energy gels makes you run faster:

- Treatment: energy gel
- Control: no energy gel

It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:

- Divide the sample to pro and amateur
- Randomly assign pro athletes to treatment and control groups
- Randomly assign amateur athletes to treatment and control groups
- Pro/amateur status is equally represented in the resulting treatment and control groups

Why is this important? Can you think of other variables to block for?

Obtaining Good Samples

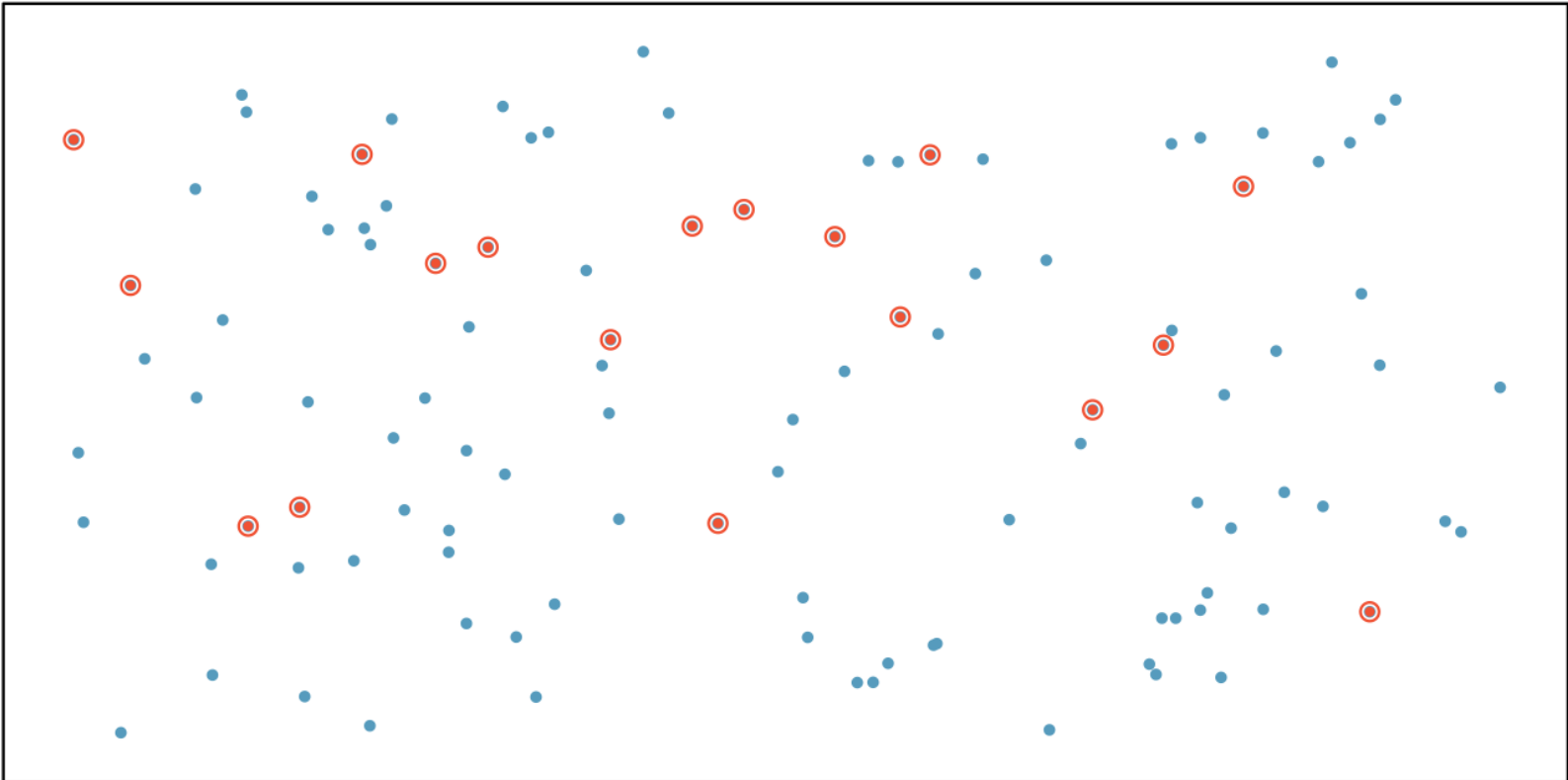
Almost all statistical methods are based on the notion of implied randomness.

If observational data are not collected in a random framework from a population, these statistical methods -- the estimates and errors associated with the estimates -- are not reliable.

Most commonly used random sampling techniques are **simple**, **stratified**, and **cluster** sampling.

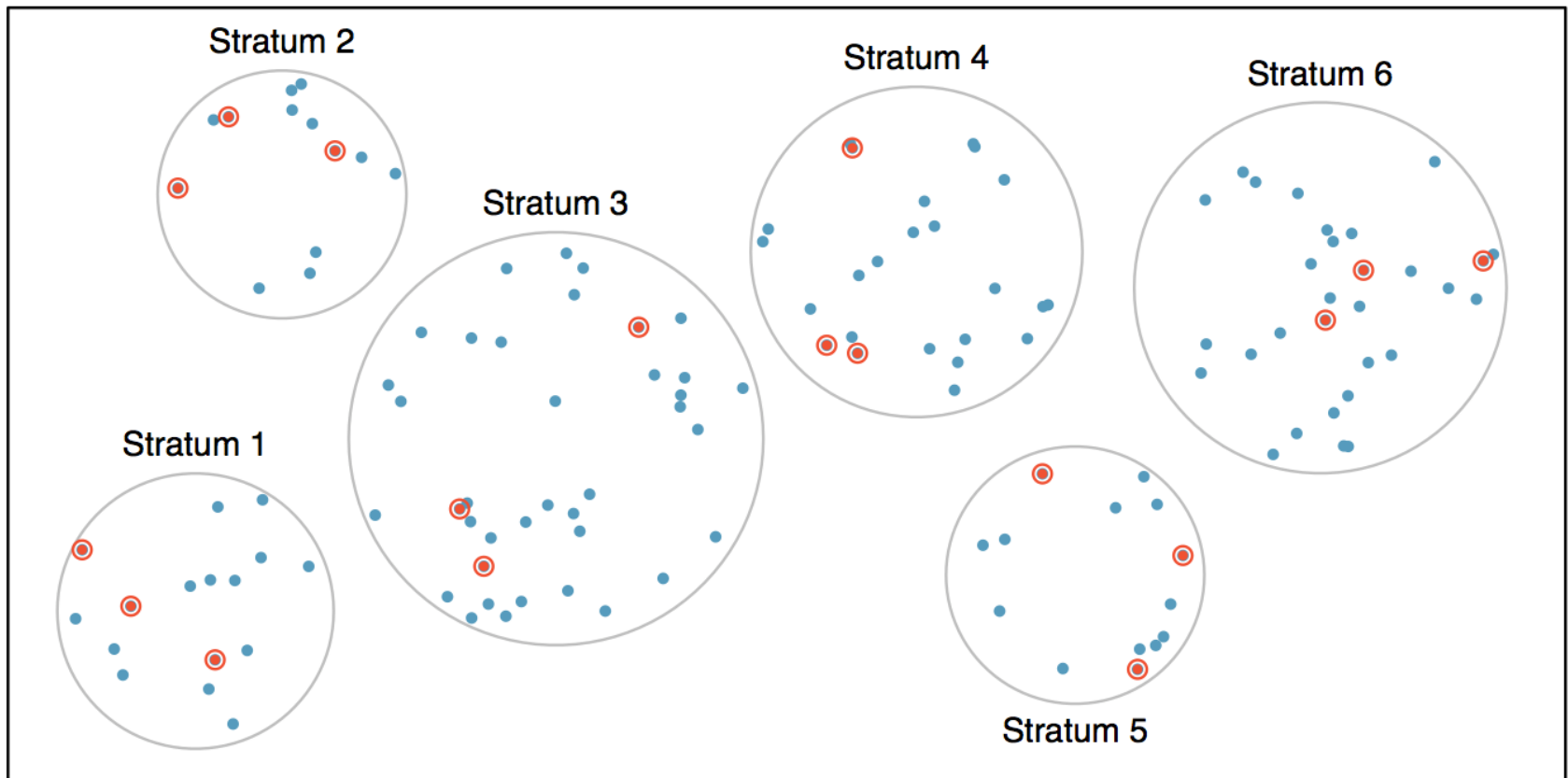
Simple Random Sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



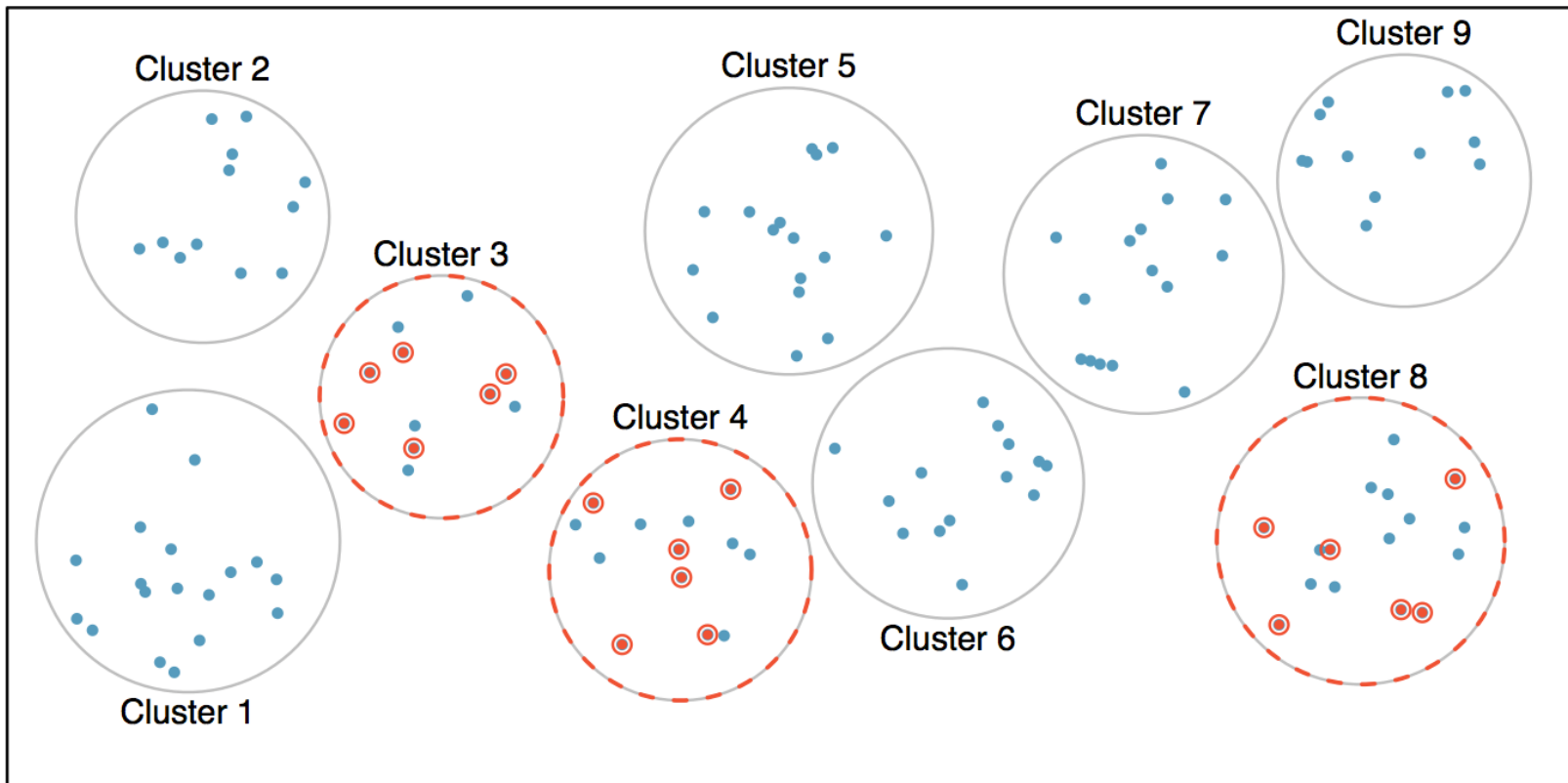
Stratified Sample

Strata are made up of similar observations. We take a simple random sample from each stratum.



Cluster Sample

Clusters are usually not made up of homogeneous observations, and we take a simple random sample from a random sample of clusters. Usually preferred for economical reasons.



Anecdotal evidence and early smoking research

- Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- Anti-smoking research was faced with resistance based on anecdotal evidence such as "My uncle smokes three packs a day and he's in perfectly good health", evidence based on a limited sample size that might not be representative of the population.
- It was concluded that "smoking is a complex human behavior, by its nature difficult to study, confounded by human variability."
- In time researchers were able to examine larger samples of cases (smokers), and trends showing that smoking has negative health