

A hand is shown writing on a graph with a pen. A ruler is placed over the graph, and the background is a light blue gradient.

Lecture 2

Describing data with graphs and numbers.
Normal Distribution. Data relationships.

Describing distributions with numbers

- Mean
- Median
- Quartiles
- Five number summary. Boxplots
- Standard deviation

Mean

- The mean
- The arithmetic mean of a data set (average value)
- Denoted by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum x_i$$

TABLE 1.8 Fuel economy (miles per gallon) for model year 2001 cars

Minicompact cars			Two-seater cars		
Model	City	Highway	Model	City	Highway
Audi TT Coupe	22	31	Acura NSX	17	24
BMW 325CI Convertible	19	27	Audi TT Roadster	22	30
BMW 330CI Convertible	20	28	BMW Z3 Coupe	21	28
BMW M3 Convertible	16	23	BMW Z3 Roadster	20	27
Jaguar XK8 Convertible	17	24	BMW Z8	13	21
Jaguar XKR Convertible	16	22	Chevrolet Corvette	18	26
Mercedes-Benz CLK320	20	28	Dodge Viper	11	21
Mercedes-Benz CLK430	18	24	Ferrari Modena	11	16
Mitsubishi Eclipse	22	30	Ferrari Maranello	8	13
Porsche 911 Carrera	17	25	Honda Insight	61	68
Porsche 911 Turbo	15	22	Honda S2000	20	26
			Lamborghini Diablo	10	13
			Mazda Miata	22	28
			Mercedes-Benz SL500	16	23
			Mercedes-Benz SL600	13	19
			Mercedes-Benz SLK320	21	27
			Plymouth Prowler	17	23
			Porsche Boxster	19	27
			Toyota MR2	25	30

- Mean highway mileage for 19 2-seaters:
Sum: $24+30+\dots+30=490$
Divide by $n=19$
Average: 25.8 miles/gallon

Problem: Honda Insight 68miles/gallon!
If we exclude it, mean mileage: 23.4
miles/gallon

- Mean can be easily influenced by outliers. It is not a robust measure of center.

Median

- Median is the midpoint of a distribution.
- Median is a resistant or robust measure of center.
- Not sensitive to extreme observations
- In a symmetric distribution mean=median
- In a skewed distribution the mean is further out in the long tail than is the median.
- Example: house prices: usually right skewed
 - The mean price of existing houses sold in 2000 in Indiana was 176,200. (Mean chases the right tail)
 - The median price of these houses was 139,000.

Measure of center: the median

The **median** is the midpoint of a distribution—the number such that half of the observations are smaller and half are larger.

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

1. Sort observations by size.
 n = number of observations
-

2.a. If n is **odd**, the median is observation $(n+1)/2$ down the list

← $n = 25$
 $(n+1)/2 = 26/2 = 13$
Median = 3.4

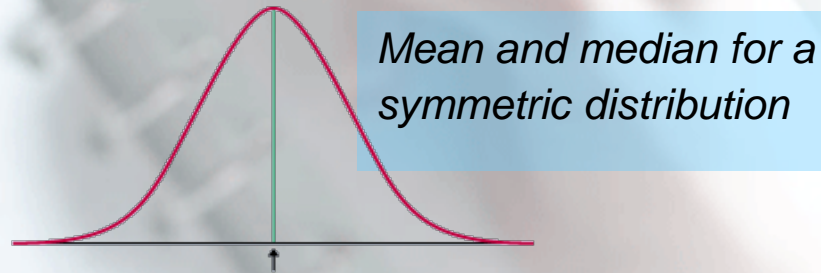
2.b. If n is **even**, the median is the mean of the two middle observations.

$n = 24$ →
 $n/2 = 12$
Median = $(3.3+3.4) / 2 = 3.35$

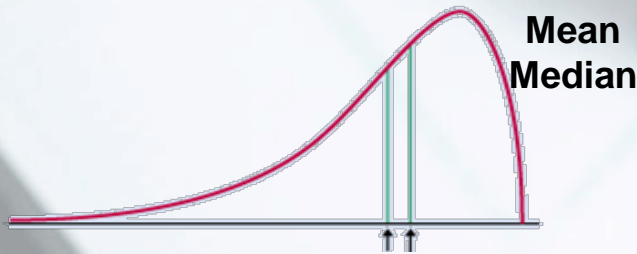
1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	8	2.3
9	9	2.5
10	10	2.8
11	11	2.9
12	12	3.3
13	13	3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	8	4.7
22	9	4.9
23	10	5.3
24	11	5.6

Comparing the mean and the median

The mean and the median are the same only if the distribution is symmetrical. The median is a measure of center that is resistant to skew and outliers. The mean is not.

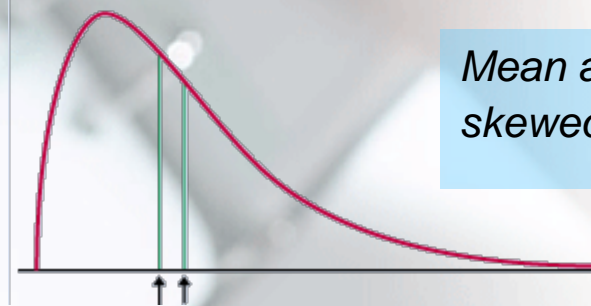


Mean and median for a symmetric distribution



Left skew

Mean
Median

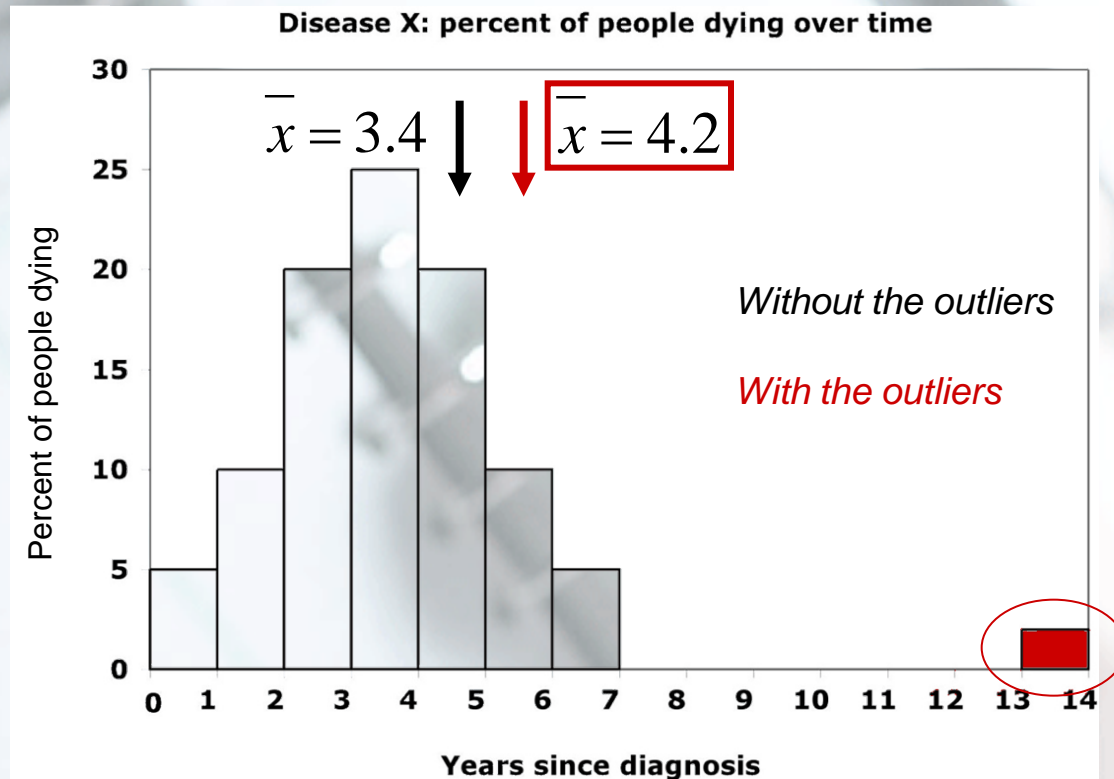


Mean and median for skewed distributions

Mean
Median

Right skew

Mean and median of a distribution with outliers



The mean is pulled to the right a lot by the outliers (from 3.4 to 4.2).

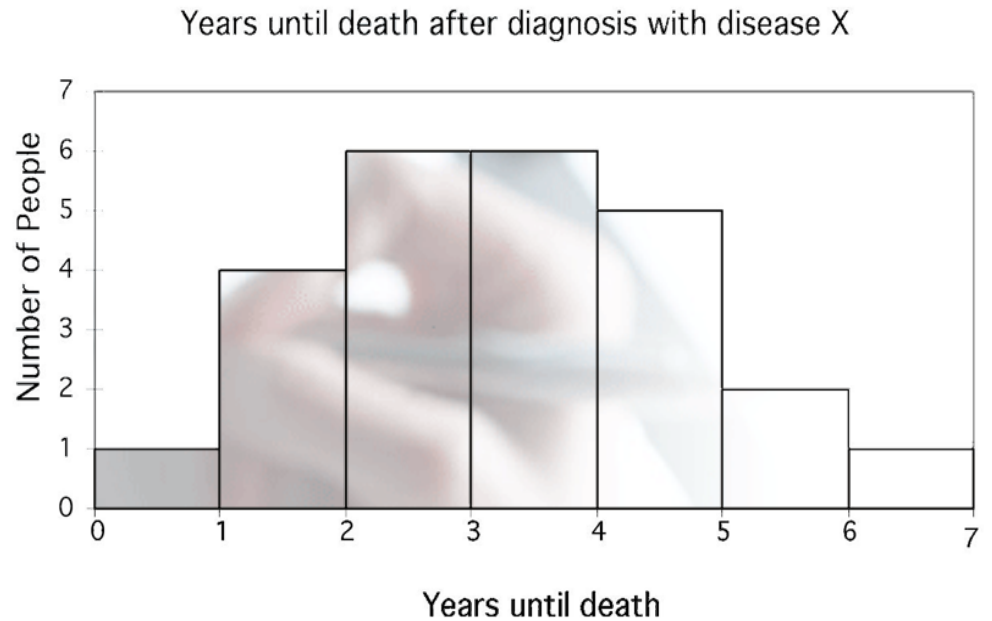
The median, on the other hand, is only slightly pulled to the right by the outliers (from 3.4 to 3.6).

Impact of skewed data

Mean and median of a symmetric

Disease X: $\bar{x} = 3.4$
 $M = 3.4$

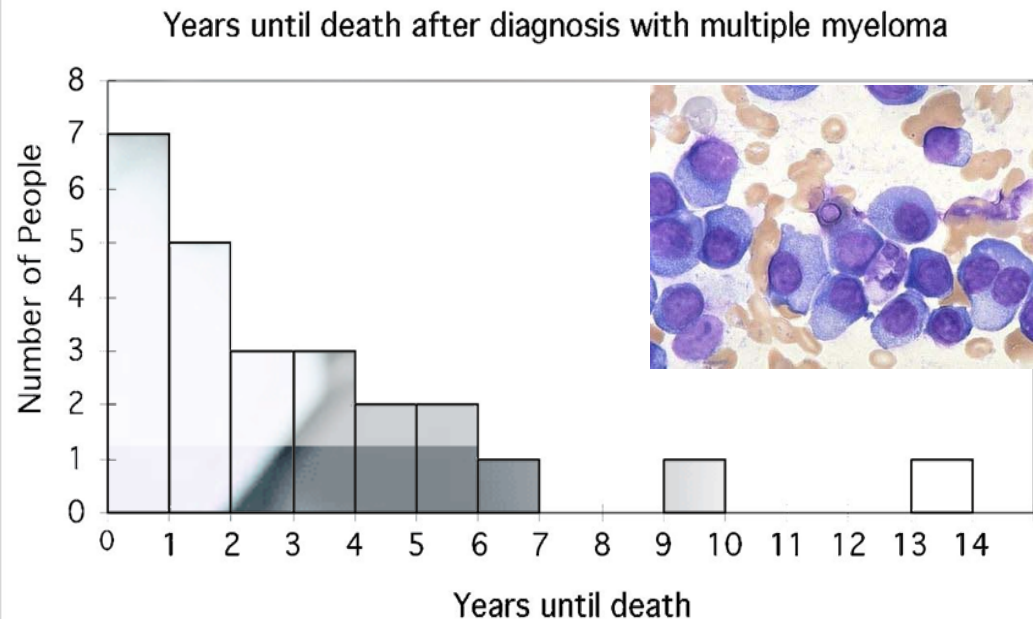
Mean and median are the same.



... and a right-skewed distribution

Multiple myeloma: $\bar{x} = 3.4$
 $M = 2.5$

The mean is pulled toward the skew.



Measures of spread: Quartiles

- Quartiles: Divides data into four parts
- p -th percentile – p percent of the observations fall at or below it.
- Median – 50-th percentile
- Q1-first quartile – 25-th percentile (median of the lower half of data)
- Q3-third quartile – 75-th percentile (median of the upper half of data)

Using R:

- First thing first: import the data. I prefer to use Excel first to save data into a .csv file (comma separated values).
- Read the file TA01_008.XLS from the website and save it as TA01_008.csv
- Now R: I like to use tinn-R as the editor. Open tinn-R and save a file in the same directory that you put the .csv file.
- Now go to R/Rgui/ and click Initiate preferred. If everything is configured fine an R window should open

- Now type and send line to R:
 - `table1.08=read.csv("TA01_008.csv",header=TRUE)`
 - This will import the data into R also telling R that the first line in the data contains the variable names.
 - Table1.08 has a “table” structure. To access individual components in it you have to use `table1.08$nameofvariable`, for example:
 - `table1.08$CarType`
 - Produces:
 - `[1] Two Two Two Two Two Two Two Two Two Two Two Two Two Two`
 - `[16] Two Two Two Two Mini Mini Mini Mini Mini Mini Mini Mini Mini Mini`
 - `Levels: Mini Two`
 - This is a vector and notice that R knows it is a categorical variable.

- `mean(x)` calculates the mean of variable `x`
- `median(x)` will give the median
- In fact you should read section 3.1 in the R textbook for all the functions you will need
- `summary(data.object)` is another useful function.

In fact:

- `summary(table1.08)`
 - CarType City Highway
 - Mini:11 Min. : 8.00 Min. :13.00
 - Two :19 1st Qu.:16.00 1st Qu.:22.25
 - » Median :18.00 Median :25.50
 - » Mean :18.90 Mean :25.80
 - » 3rd Qu.:20.75 3rd Qu.:28.00
 - » Max. :61.00 Max. :68.00

- Lastly if you wish to apply functions only for the part of the dataframe that contains Mini cars:
 - `tapply(table1.08$City,table1.08$CarType,mean)`
 - Mini Two
 - 18.36364 19.21053
- The `tapply` call takes the *table1.08\$City* variable, splits it according to *table1.08\$CarType* variable levels and calculates the function *mean* for each group.
- In the same way you can try:
 - `tapply(table1.08$City,table1.08$CarType,summary)`

Doing it by hand:

The **first quartile, Q_1** , is the value in the sample that has 25% of the data at or below it (\Leftrightarrow it is the median of the lower half of the sorted data, excluding M).

$$M = \text{median} = 3.4$$

The **third quartile, Q_3** , is the value in the sample that has 75% of the data at or below it (\Leftrightarrow it is the median of the upper half of the sorted data, excluding M).

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

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

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

Five-number summary and boxplot

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

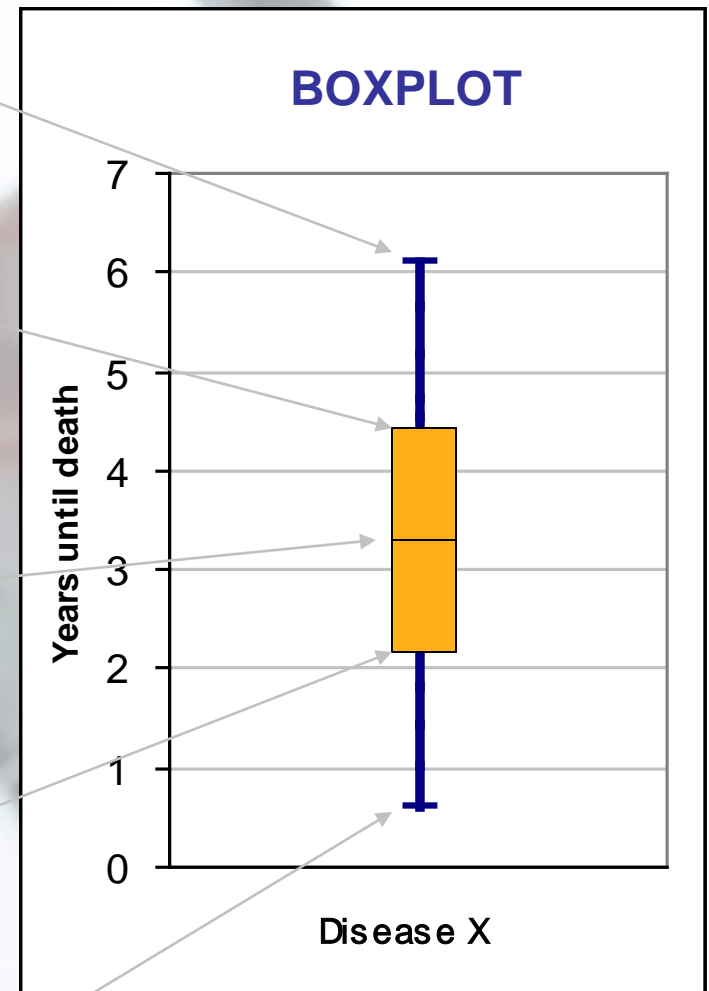
Largest = max = 6.1

Q_3 = third quartile
= 4.35

M = median = 3.4

Q_1 = first quartile
= 2.2

Smallest = min = 0.6

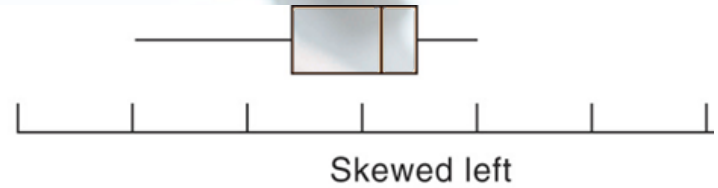
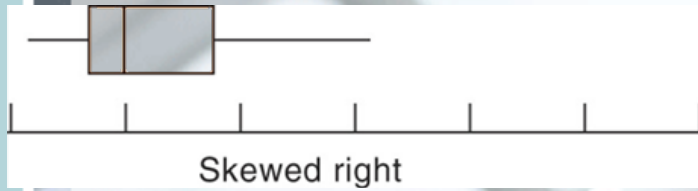


Five-number summary:
min Q_1 M Q_3 max

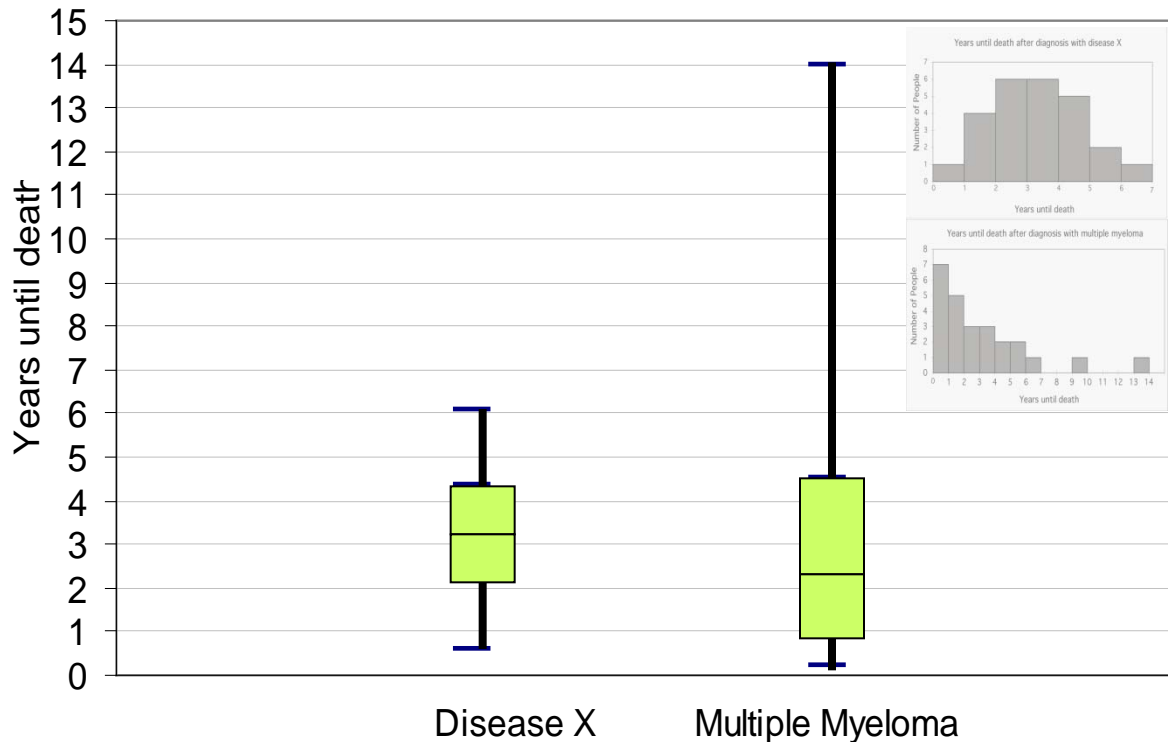
Five-Number Summary

- Minimum Q1 Median Q3 Maximum
- **Boxplot** – visual representation of the five-number summary.
 - Central box: Q1 to Q3.
 - Line inside box: Median
 - Extended straight lines: lowest to highest observation, except outliers
 - Outliers marked as circles or stars.
- To make Boxplots in R use function
 - `boxplot(x)`

Boxplots for skewed data



Comparing box plots for a normal and a right-skewed distribution



Boxplots remain true to the data and depict clearly symmetry or skew.

R code:

- `boxplot(table1.08$City)`
- `boxplot(table1.08$Highway)`

- `boxplot(table1.08$City~table1.08$CarType)`
- `boxplot(table1.08$Highway~table1.08$CarType)`
- `par(mfrow=c(1,2))`
- `boxplot(table1.08$City~table1.08$CarType)`
- `boxplot(table1.08$Highway~table1.08$CarType)`
- `par(mfrow=c(1,1))`

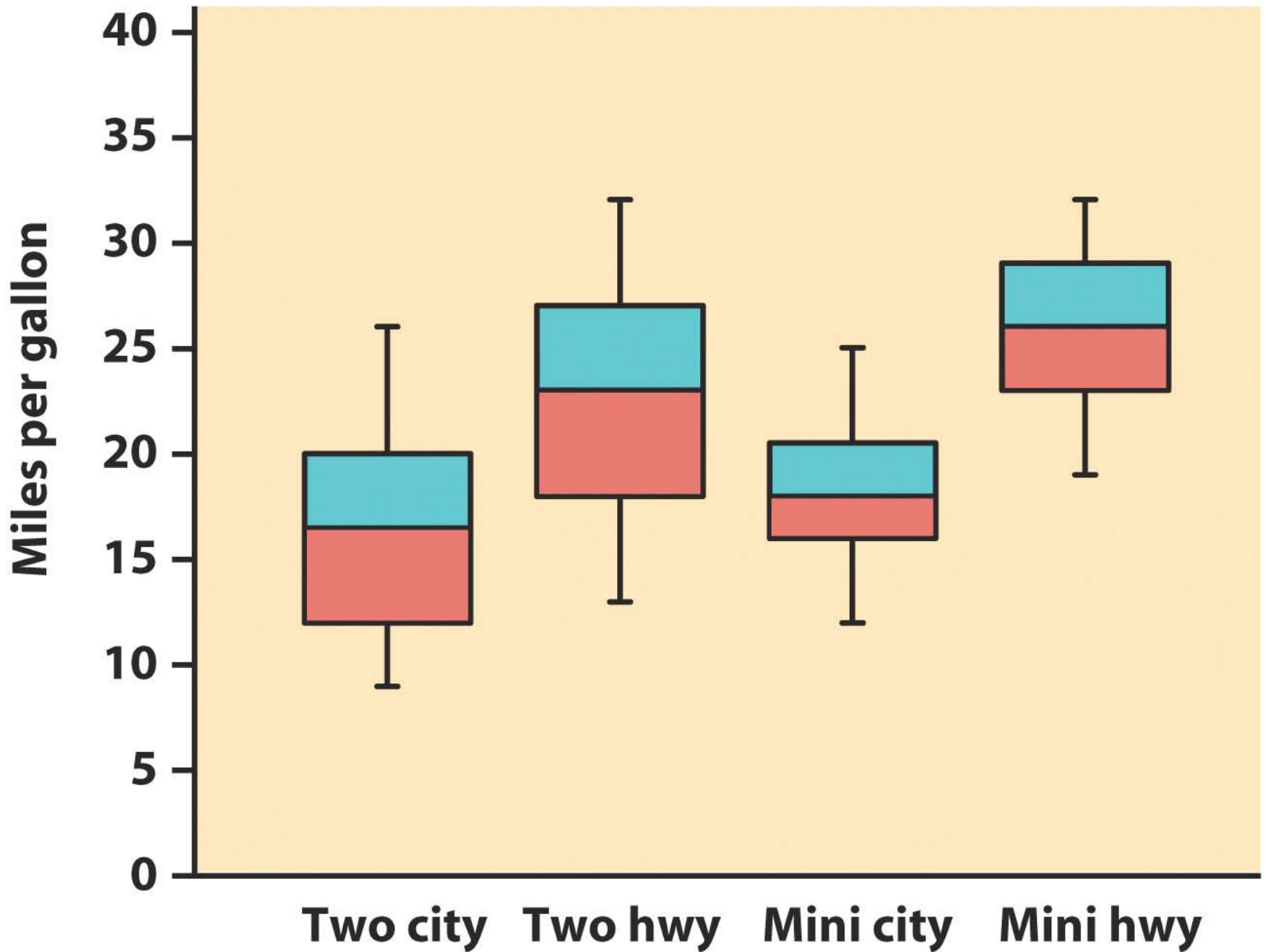


Figure 1-17
Introduction to the Practice of Statistics, Fifth Edition

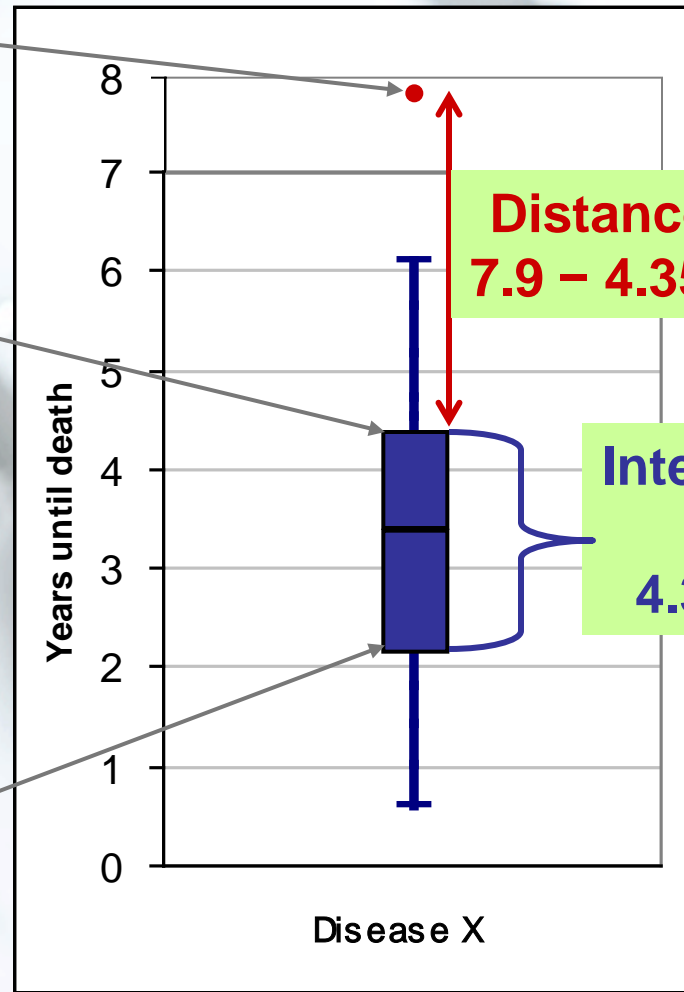
The criterion for suspected outliers

- Outliers are troublesome data points, and it is important to be able to identify them.
- The interquartile range – $IQR=Q3-Q1$
- An observation is a **suspected outlier** if it falls more than $1.5 \cdot IQR$ above the third quartile or below the first quartile.
- This is called the “ **$1.5 \cdot IQR$ rule for outliers.**”

25	6	7.9
24	5	6.1
23	4	5.3
22	3	4.9
21	2	4.7
20	1	4.5
19	6	4.2
18	5	4.1
17	4	3.9
16	3	3.8
15	2	3.7
14	1	3.6
13		3.4
12	6	3.3
11	5	2.9
10	4	2.8
9	3	2.5
8	2	2.3
7	1	2.3
6	6	2.1
5	5	1.5
4	4	1.9
3	3	1.6
2	2	1.2
1	1	0.6

$Q_3 = 4.35$

$Q_1 = 2.2$



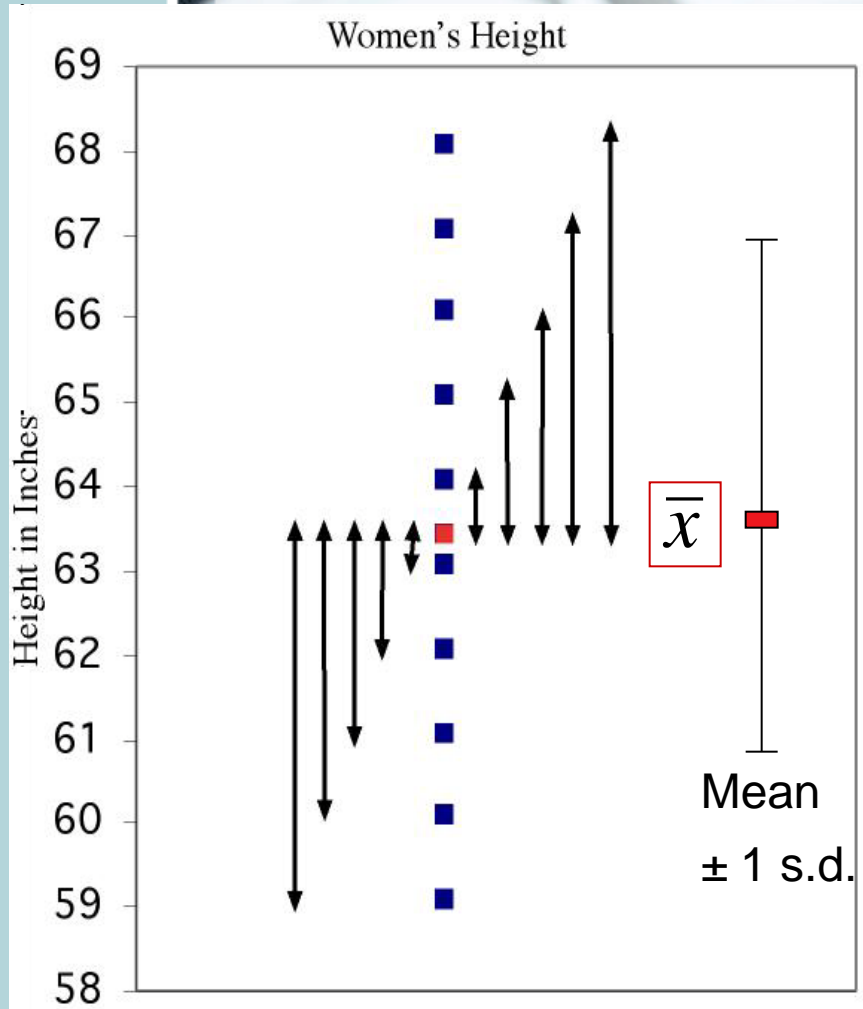
Distance to Q_3
 $7.9 - 4.35 = 3.55$

Interquartile range
 $Q_3 - Q_1$
 $4.35 - 2.2 = 2.15$

Individual #25 has a value of 7.9 years, which is 3.55 years above the third quartile. This is more than 3.225 years, $1.5 * IQR$. Thus, individual #25 is a suspected outlier.

Measure of spread: the **standard deviation**

The standard deviation “s” is used to describe the variation around the mean. Like the mean, it is not resistant to skew or outliers.



1. First calculate the **variance s^2** .

$$s^2 = \frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2$$

2. Then take the square root to get the **standard deviation s** .

$$s = \sqrt{\frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2}$$

Calculations ...

$$s = \sqrt{\frac{1}{df} \sum_1^n (x_i - \bar{x})^2}$$

Mean = 63.4

Sum of squared deviations from mean = 85.2

Degrees freedom (df) = (n - 1) = 13

s^2 = variance = 85.2/13 = 6.55 inches squared

s = standard deviation = $\sqrt{6.55}$ = 2.56 inches

Women height

i	x_i	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	59	63.4	-4.4	19.0
2	60	63.4	-3.4	11.3
3	61	63.4	-2.4	5.6
4	62	63.4	-1.4	1.8
5	62	63.4	-1.4	1.8
6	63	63.4	-0.4	0.1
7	63	63.4	-0.4	0.1
8	63	63.4	-0.4	0.1
9	64	63.4	0.6	0.4
10	64	63.4	0.6	0.4
11	65	63.4	1.6	2.7
12	66	63.4	2.6	7.0
13	67	63.4	3.6	13.3
14	68	63.4	4.6	21.6
	Mean 63.4		Sum 0.0	Sum 85.2

We'll never calculate these by hand, so make sure to know how to get the standard deviation using your calculator.

Properties of the standard deviation

- Standard deviation is always non-negative
- $s=0$ when there is no spread
- s is not resistant to presence of outliers
- The five-number summary usually better describes a skewed distribution or a distribution with outliers.
- Mean and standard deviation are usually used for reasonably symmetric distributions without outliers.

Linear Transformations: changing units of measurements

- $x_{\text{new}} = a + bx_{\text{old}}$
- Common conversions
- $x_{\text{miles}} = 0.62 x_{\text{km}}$

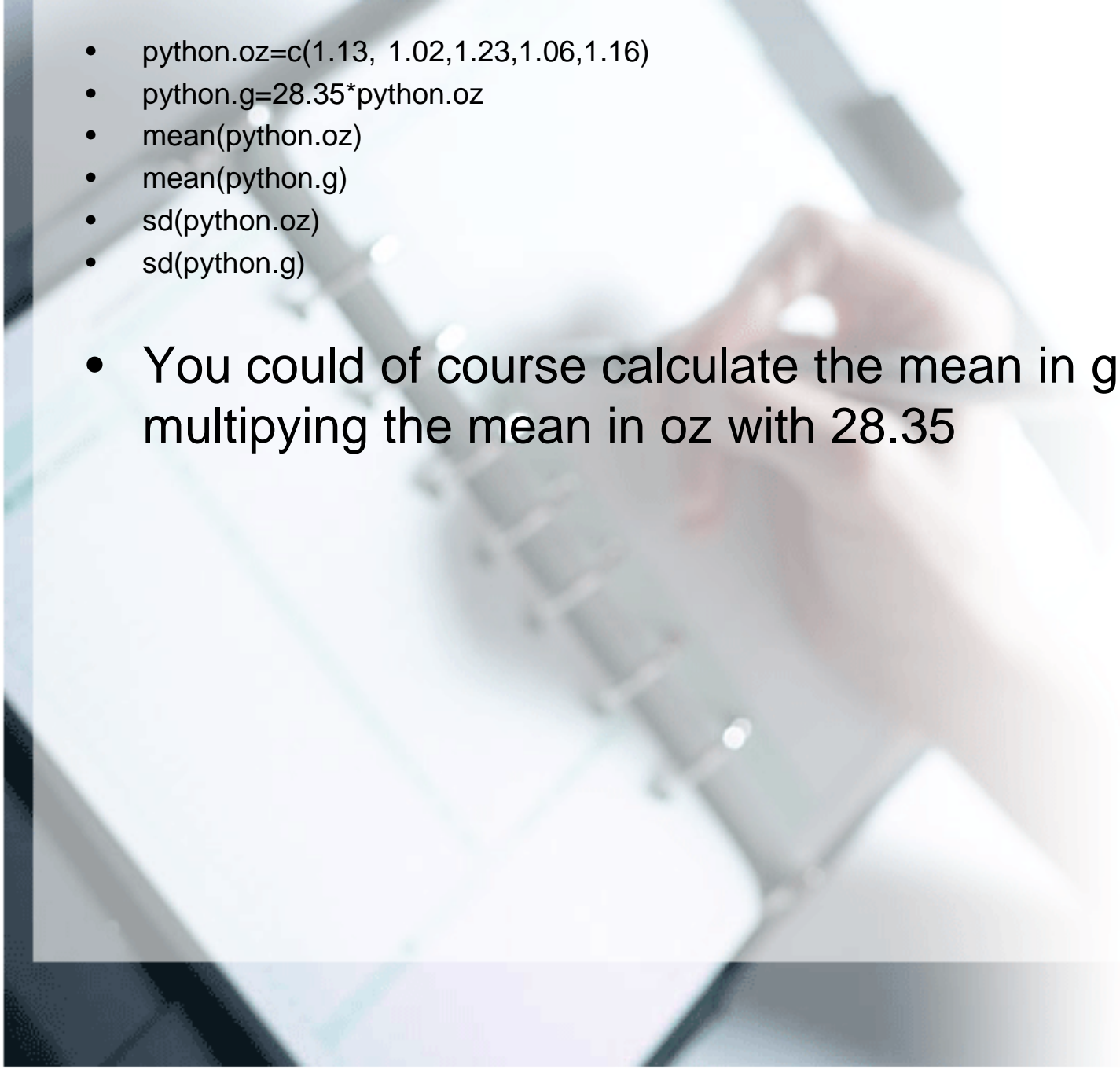
Distance=100km is equivalent to 62 miles

- $x_{\text{g}} = 28.35 x_{\text{oz}}$,

$$x_{\text{celsius}} = \frac{5}{9} (x_{\text{fah}} - 32) = -\frac{160}{9} + \frac{5}{9} x_{\text{fah}}$$

- Linear transformations do not change the shape of a distribution.
- They however change the center and the spread e.g: weights of newly hatched pythons (Example 1.21)

Python Weight	1	2	3	4	5
oz	1.13	1.02	1.23	1.06	1.16
g	32	29	35	30	33

- 
- `python.oz=c(1.13, 1.02,1.23,1.06,1.16)`
 - `python.g=28.35*python.oz`
 - `mean(python.oz)`
 - `mean(python.g)`
 - `sd(python.oz)`
 - `sd(python.g)`
-
- You could of course calculate the mean in g by multiplying the mean in oz with 28.35

Effect of a linear transformation

- Multiplying each observation by a positive number **b** multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by **b**.
- Adding the same number **a** to each observation adds **a** to measures of center and to quartiles and other percentiles but does not change measures of spread (IQR and s.d.)

• Your Transformation: $x_{\text{new}} = a + b * x_{\text{old}}$

• $\text{mean}_{\text{new}} = a + b * \text{mean}_{\text{old}}$

• $\text{median}_{\text{new}} = a + b * \text{median}_{\text{old}}$

• $\text{s.d}_{\text{new}} = |b| * \text{s.d}_{\text{old}}$

• $\text{IQR}_{\text{new}} = |b| * \text{IQR}_{\text{old}}$

$|b|$ = absolute value of b (value without sign)

The normal distribution

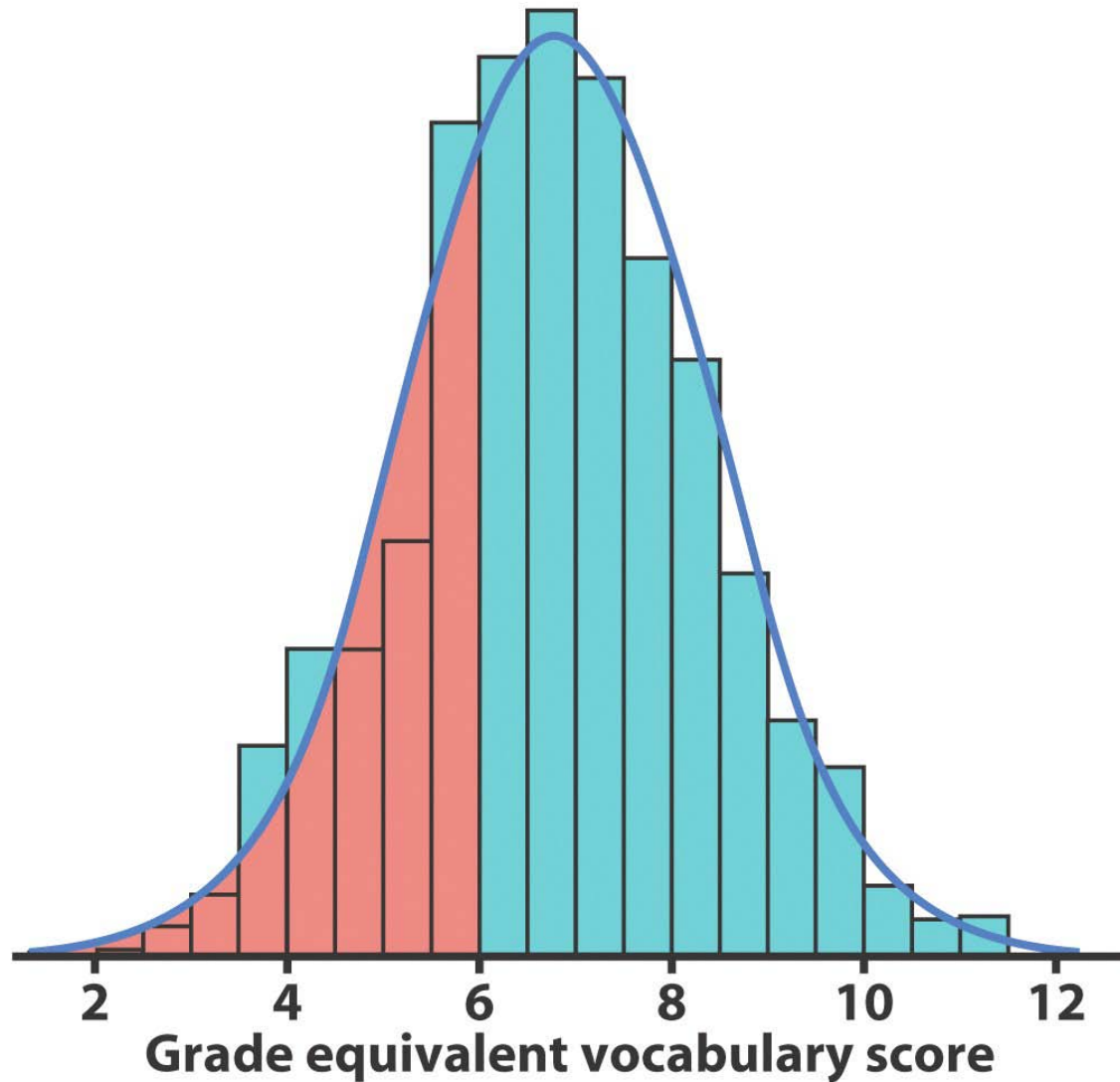
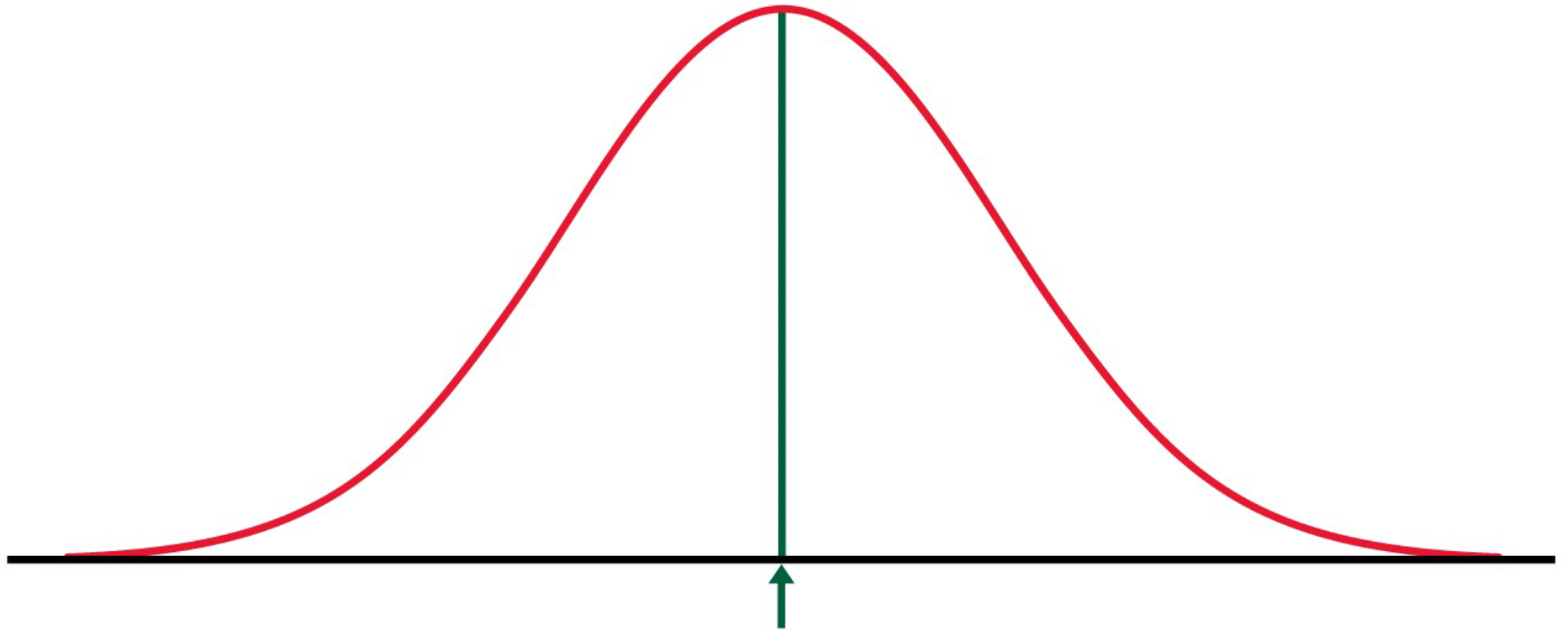


Figure 1-23b
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

Normal density curve



Median and mean

A right skewed density curve

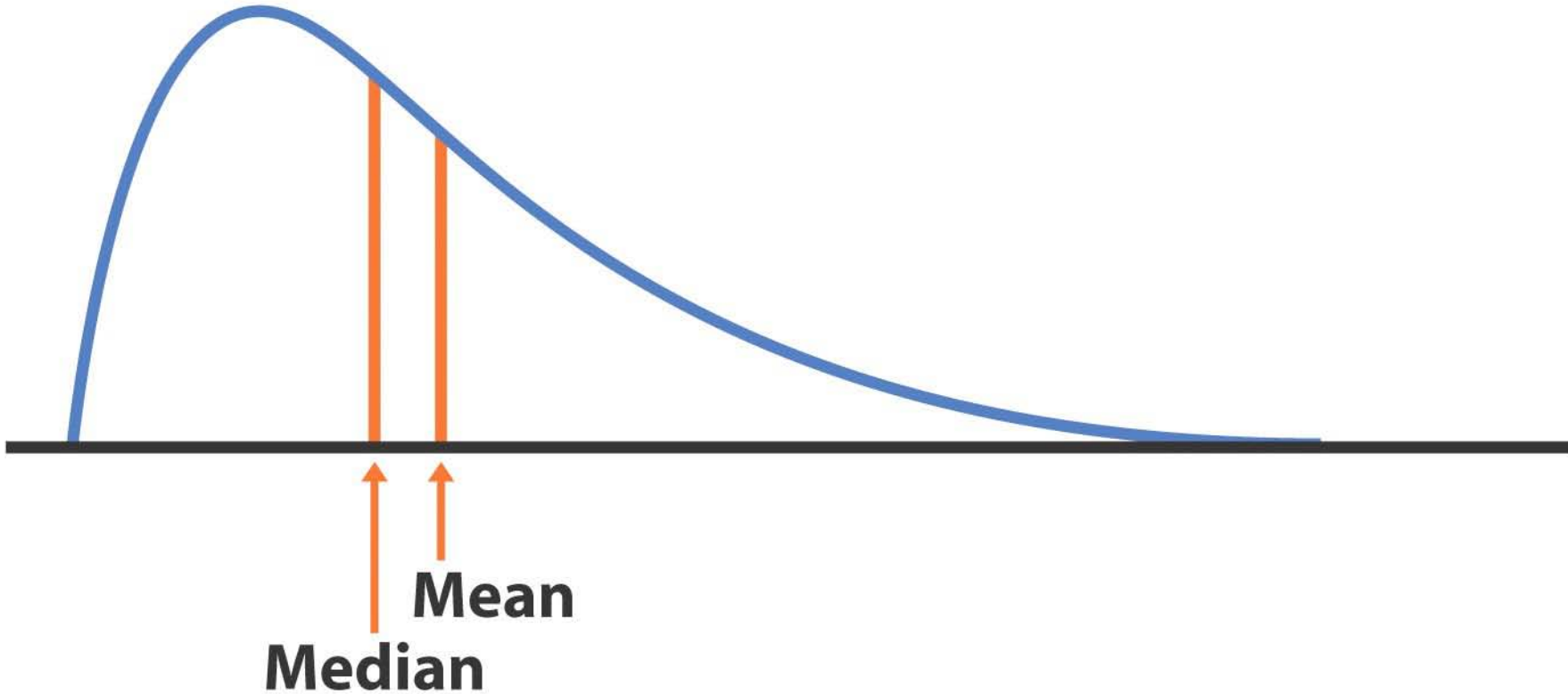


Figure 1-24b

Introduction to the Practice of Statistics, Fifth Edition

© 2005 W. H. Freeman and Company

- μ – mean of the idealized distribution (of the density curve)
- σ – standard deviation of the idealized distribution
- \bar{X} - mean of the actual observations (sample mean)
- s – standard deviation of the actual observations (sample standard deviation)

Mean is the balance point of the density curve.

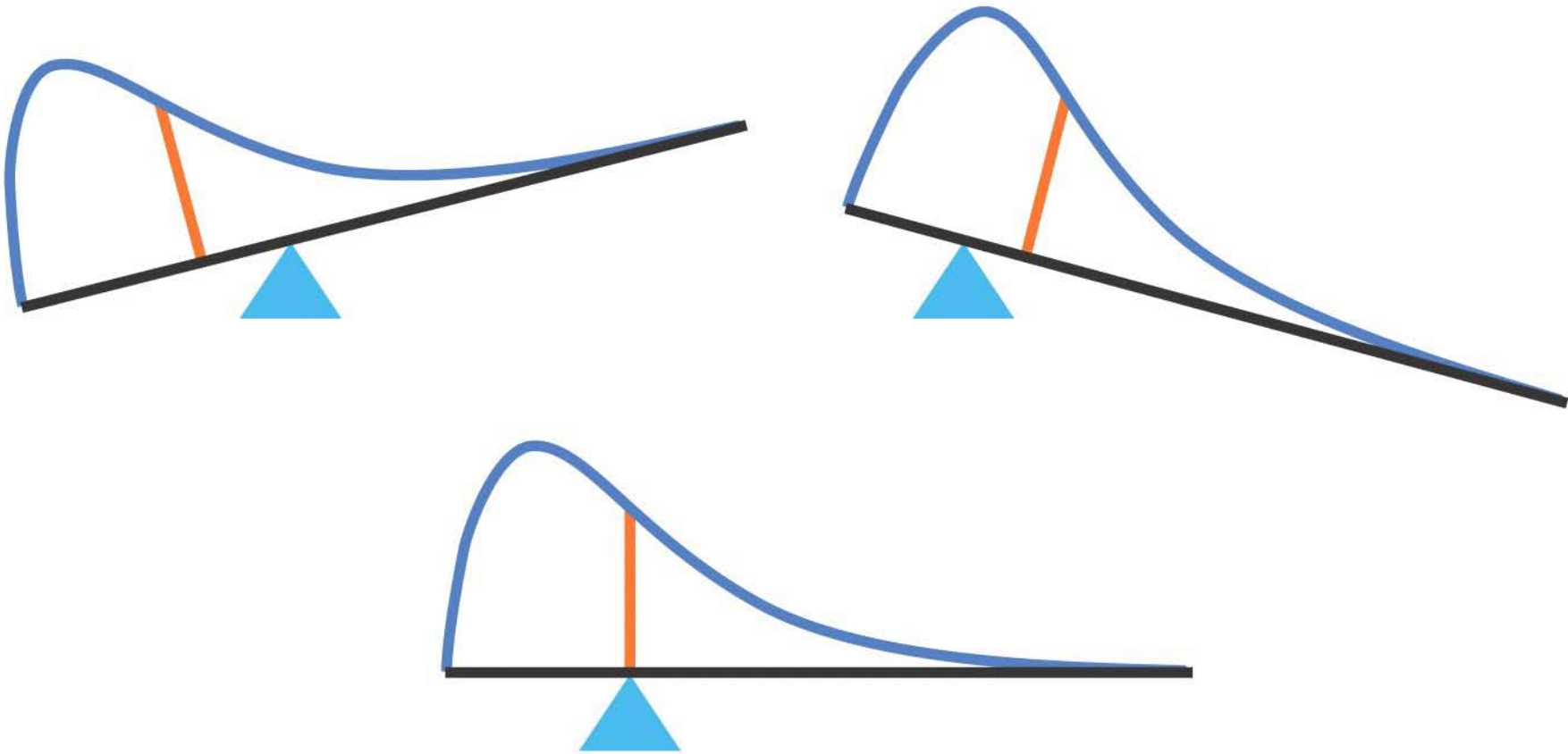
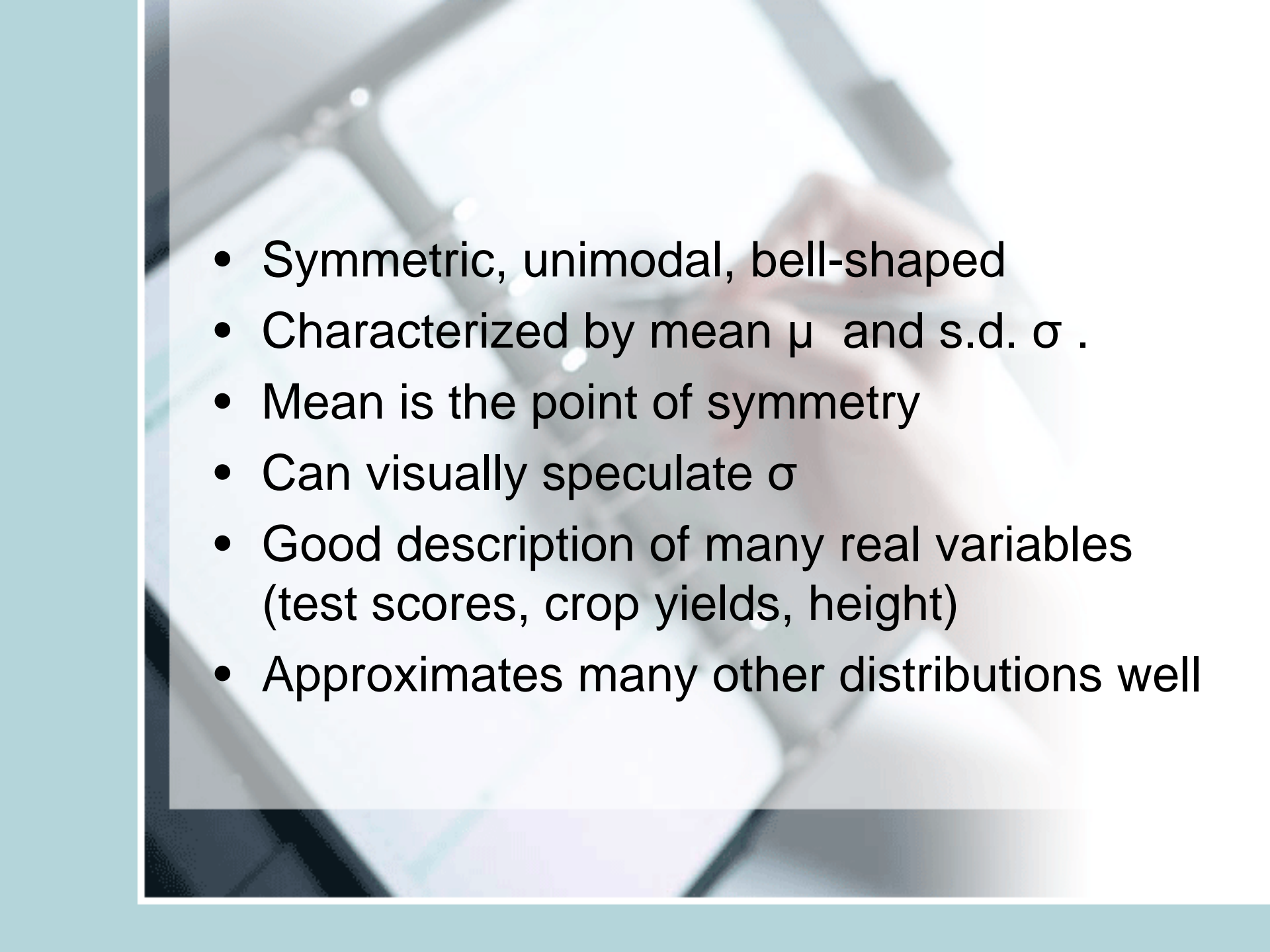


Figure 1-25
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

- 
- Symmetric, unimodal, bell-shaped
 - Characterized by mean μ and s.d. σ .
 - Mean is the point of symmetry
 - Can visually speculate σ
 - Good description of many real variables (test scores, crop yields, height)
 - Approximates many other distributions well

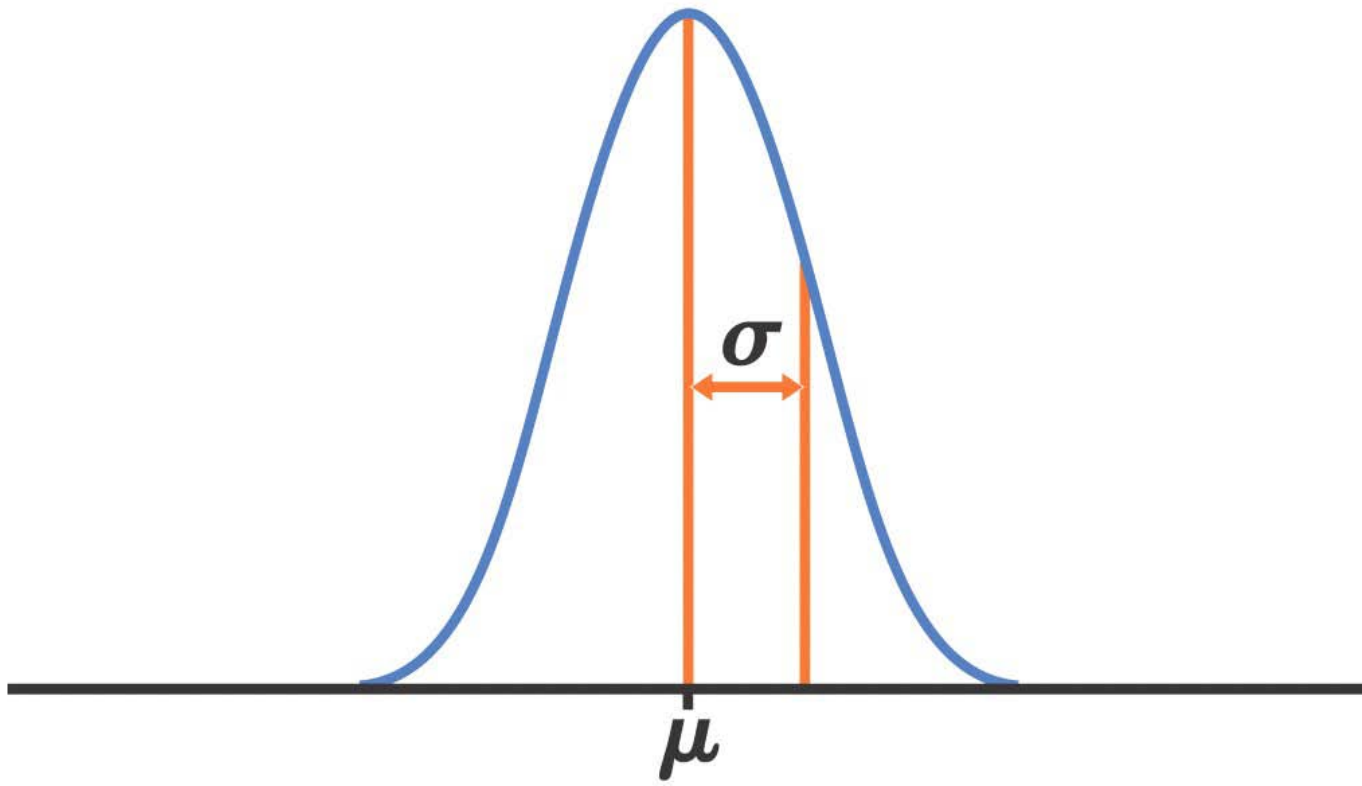
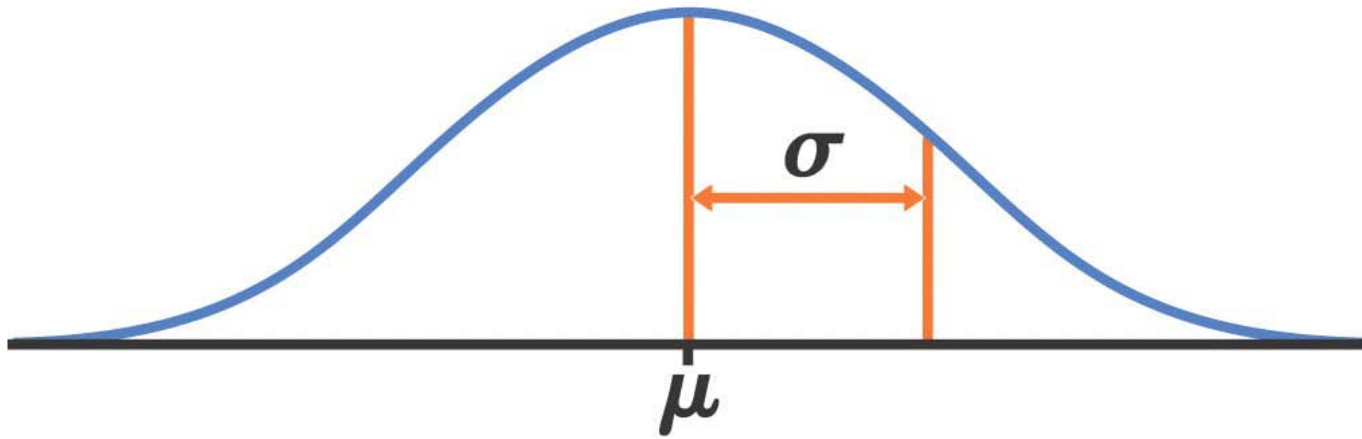


Figure 1-26
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company

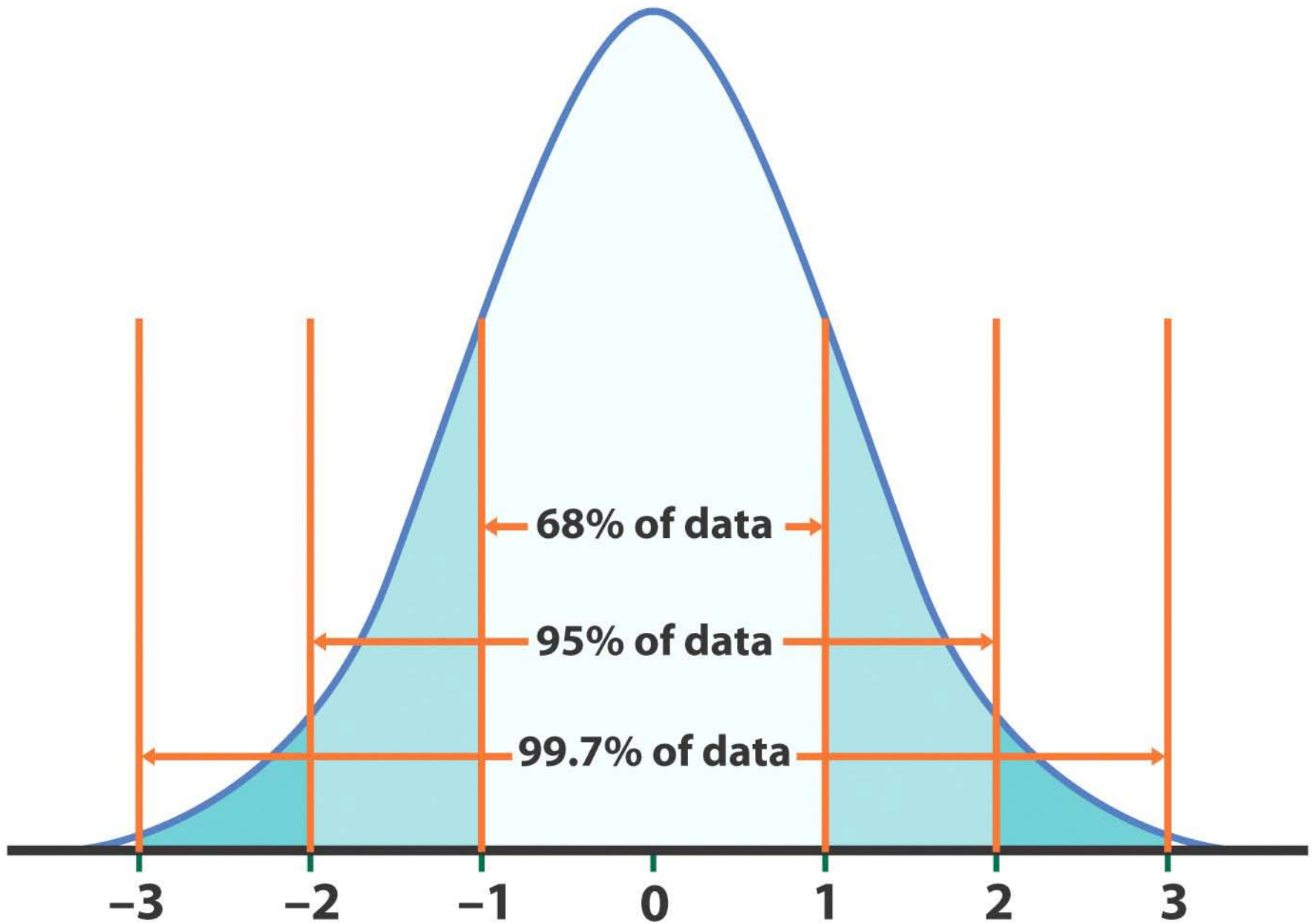


Figure 1-27
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company

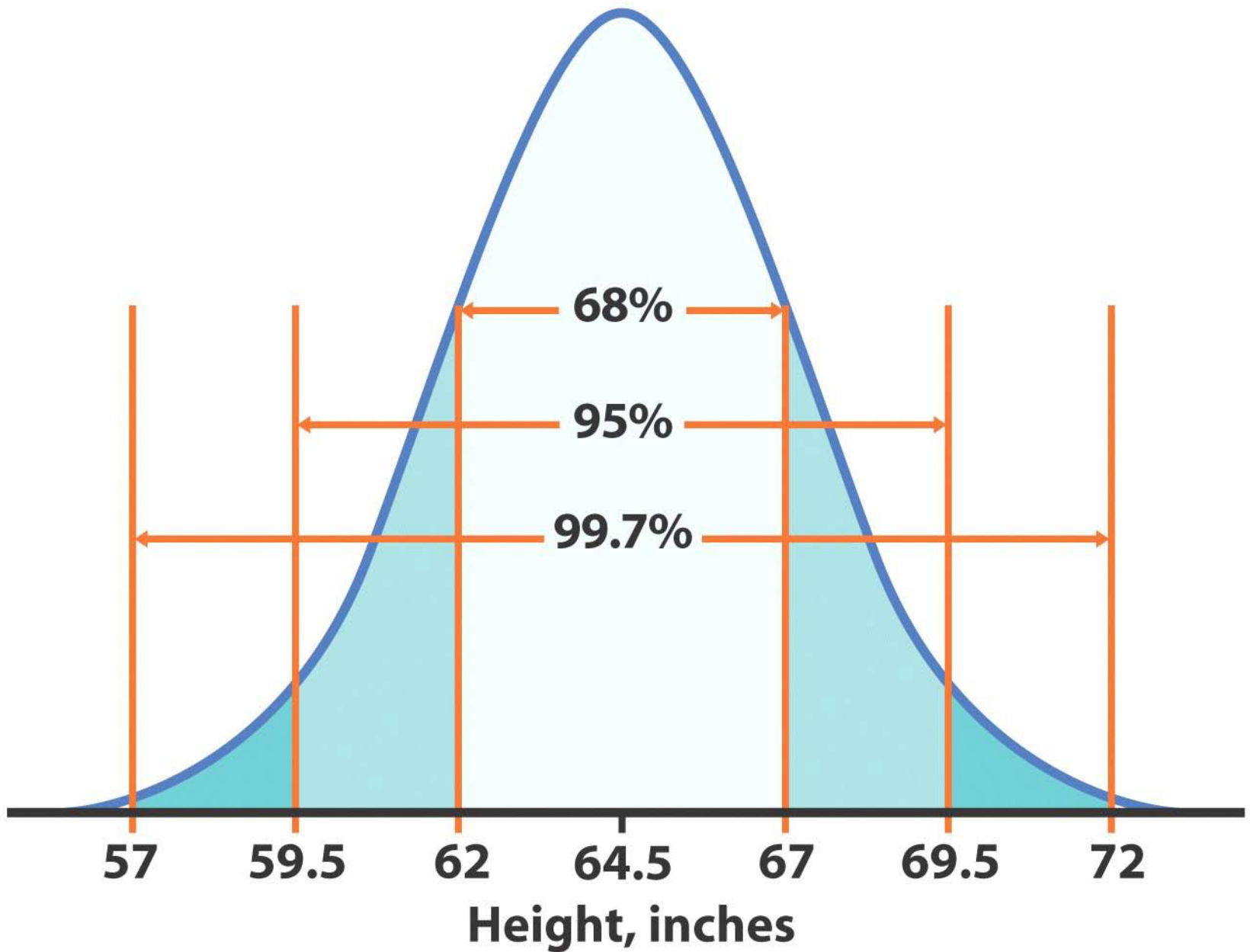
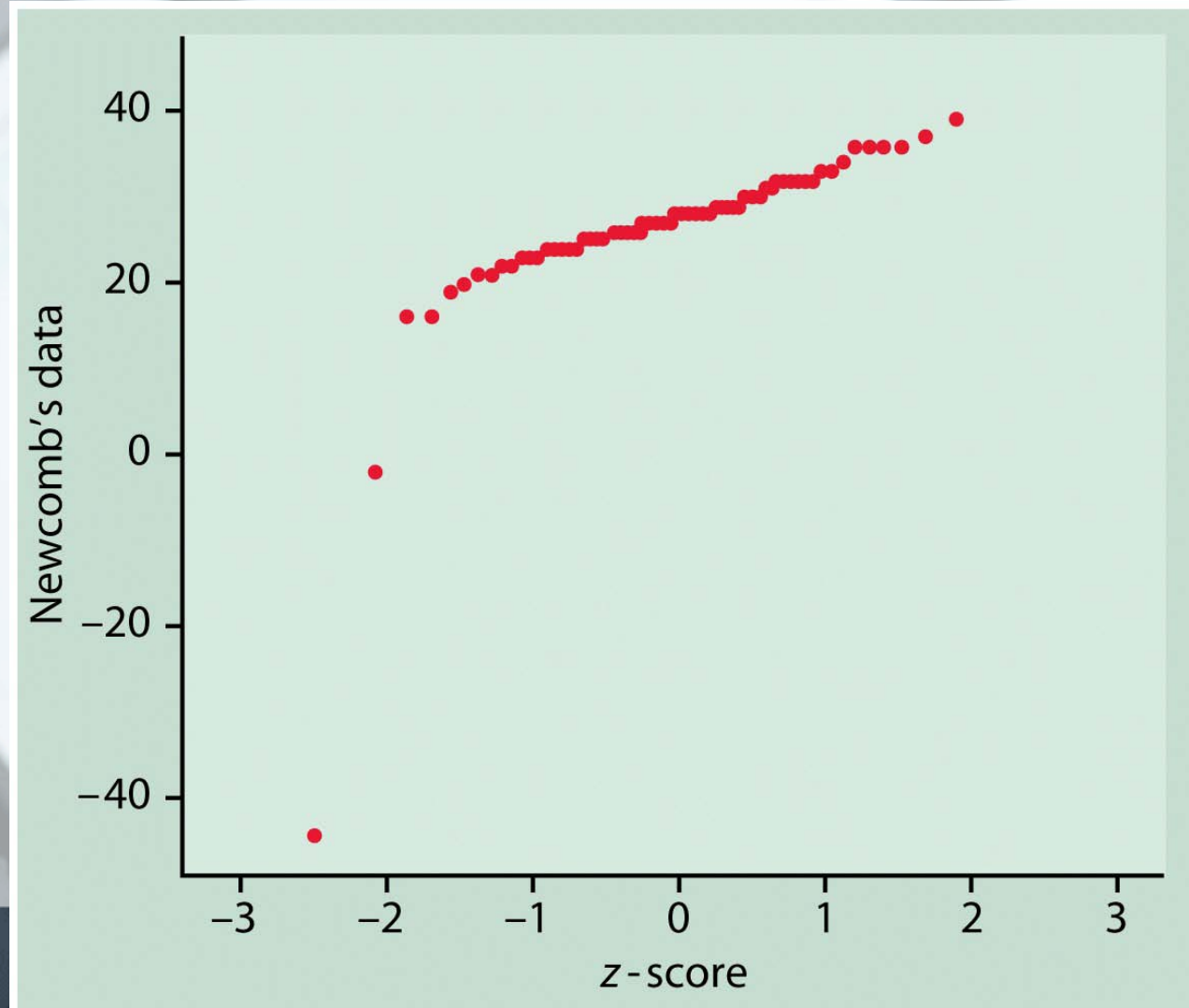


Figure 1-28
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

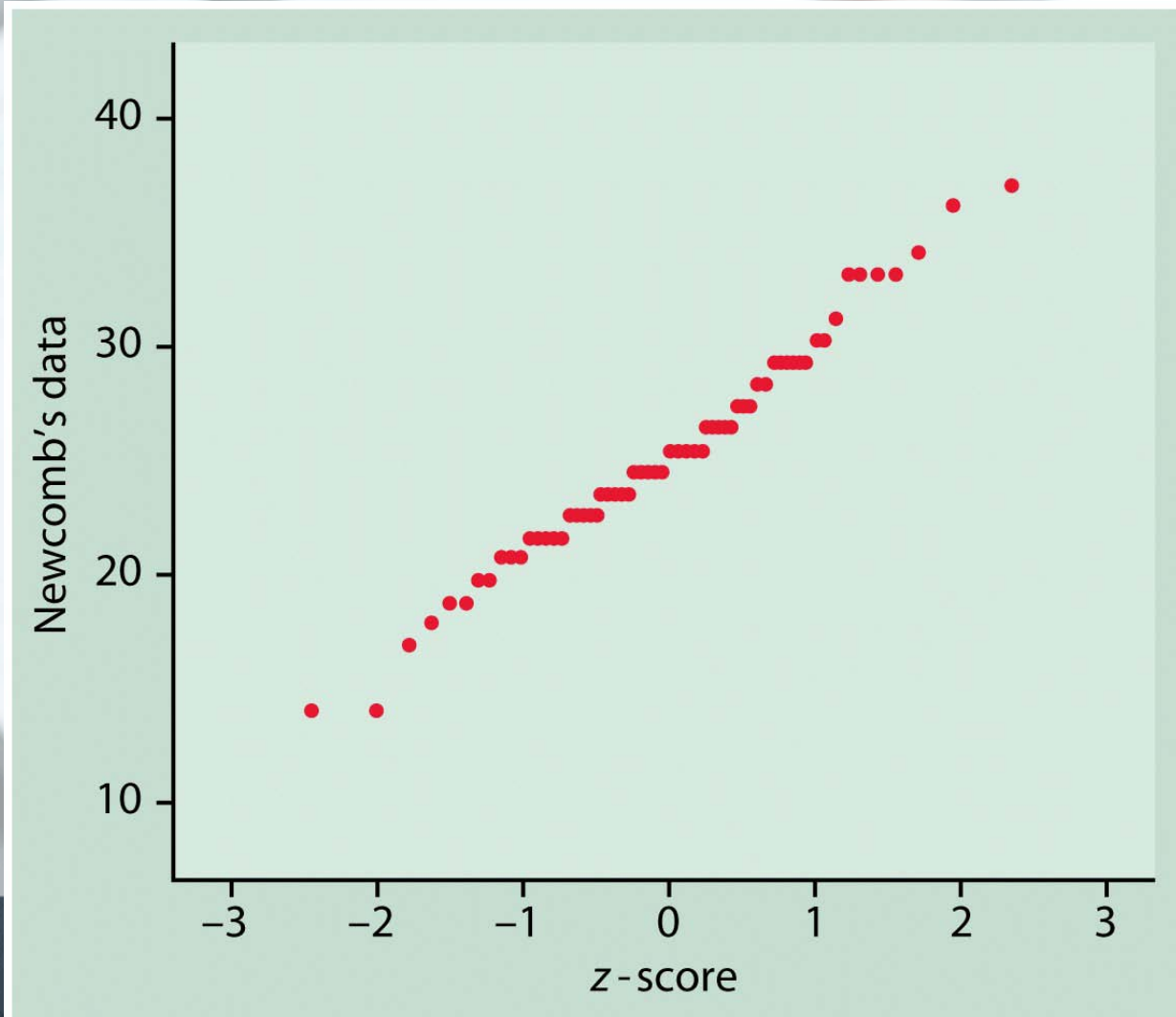
Normal quantile plots R- qqnorm()

- Also named Q-Q plots (quantile-quantile plots)
- USED to determine if the data is close to the normal distribution
 - Arrange the data from smallest to largest and record corresponding percentiles.
 - Find z-scores for these percentiles (for example z-score for 5-th percentile is $z=-1.645$.)
 - Plot each data point against the corresponding z.
- If the data distribution is close to normal the plotted points will lie close to the 45 degree straight line.

Newcomb's data



Newcomb's data without outliers.



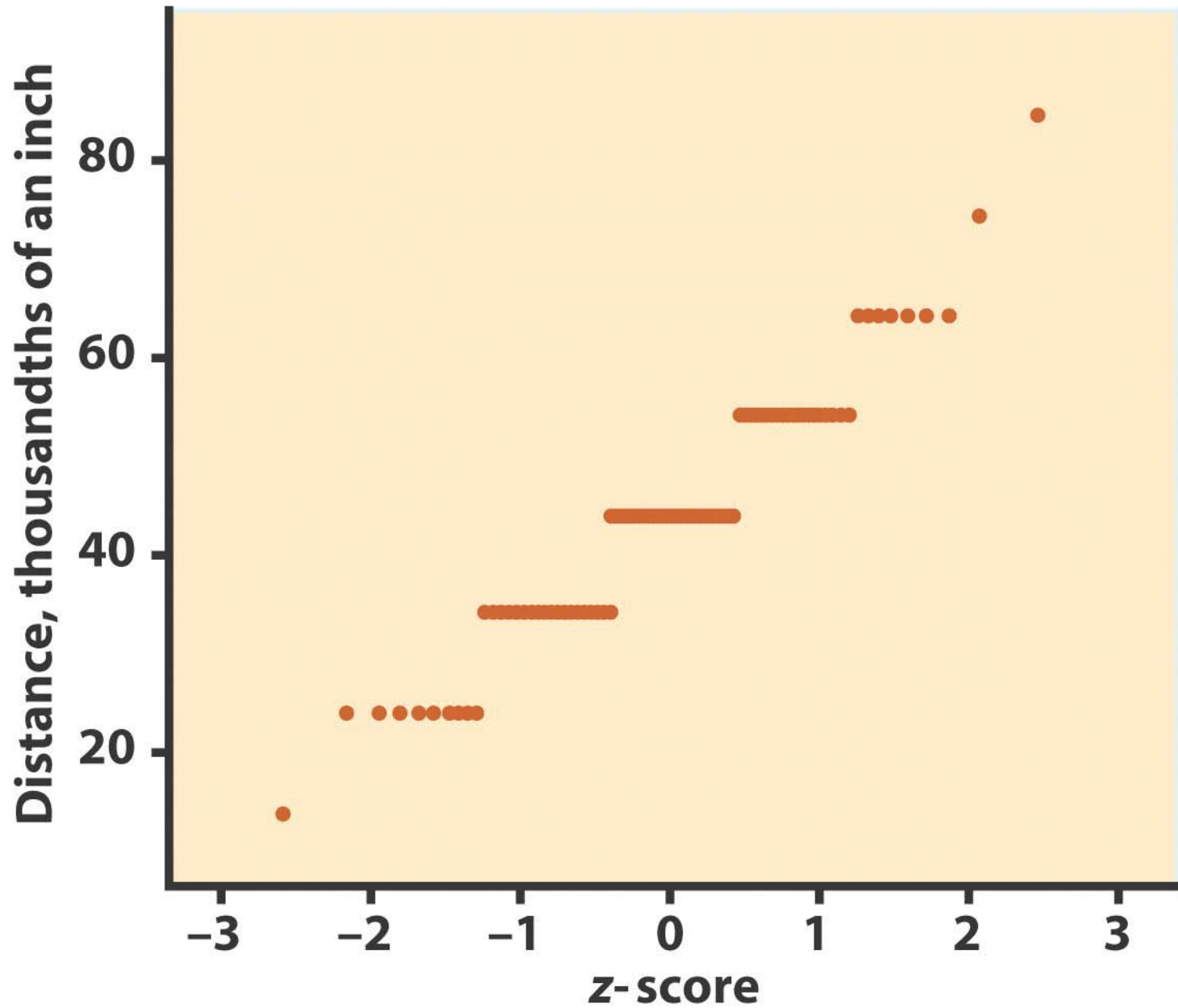


Figure 1-39
Introduction to the Practice of Statistics, Fifth Edition
 © 2005 W.H. Freeman and Company

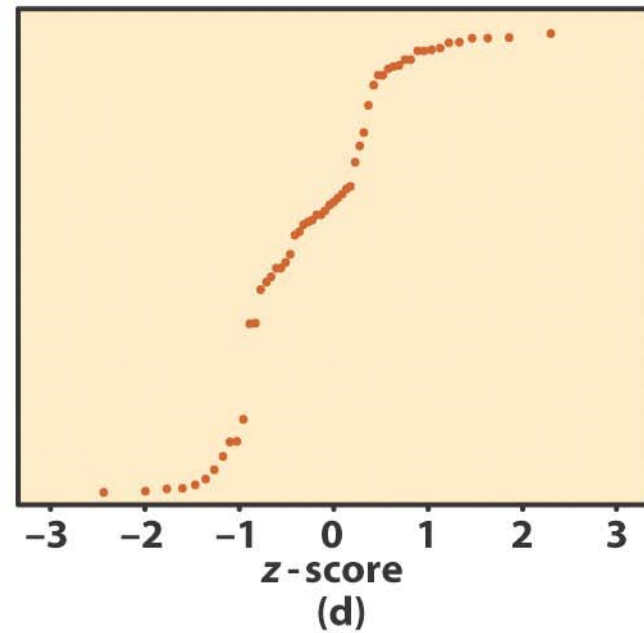
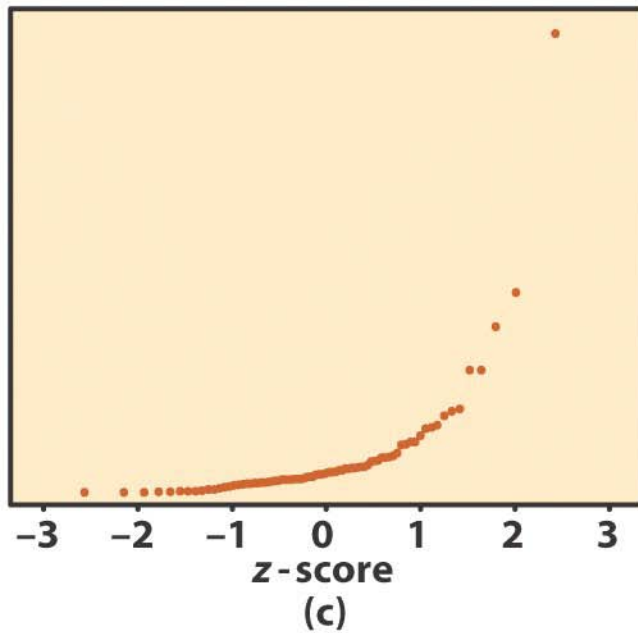
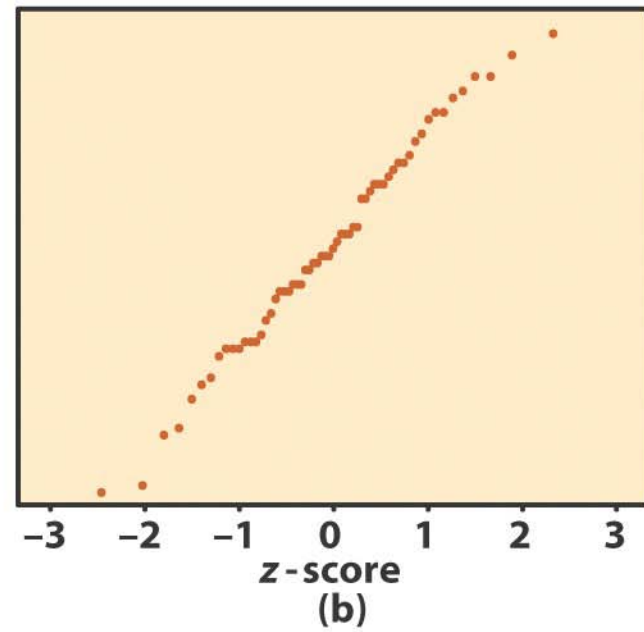
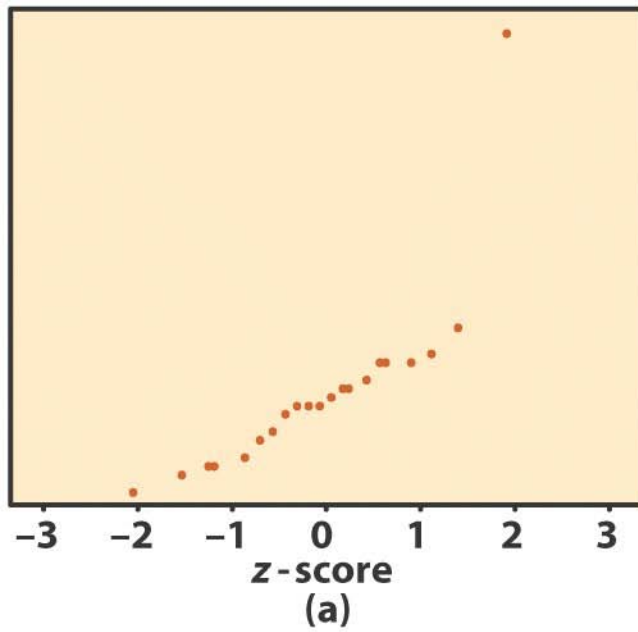
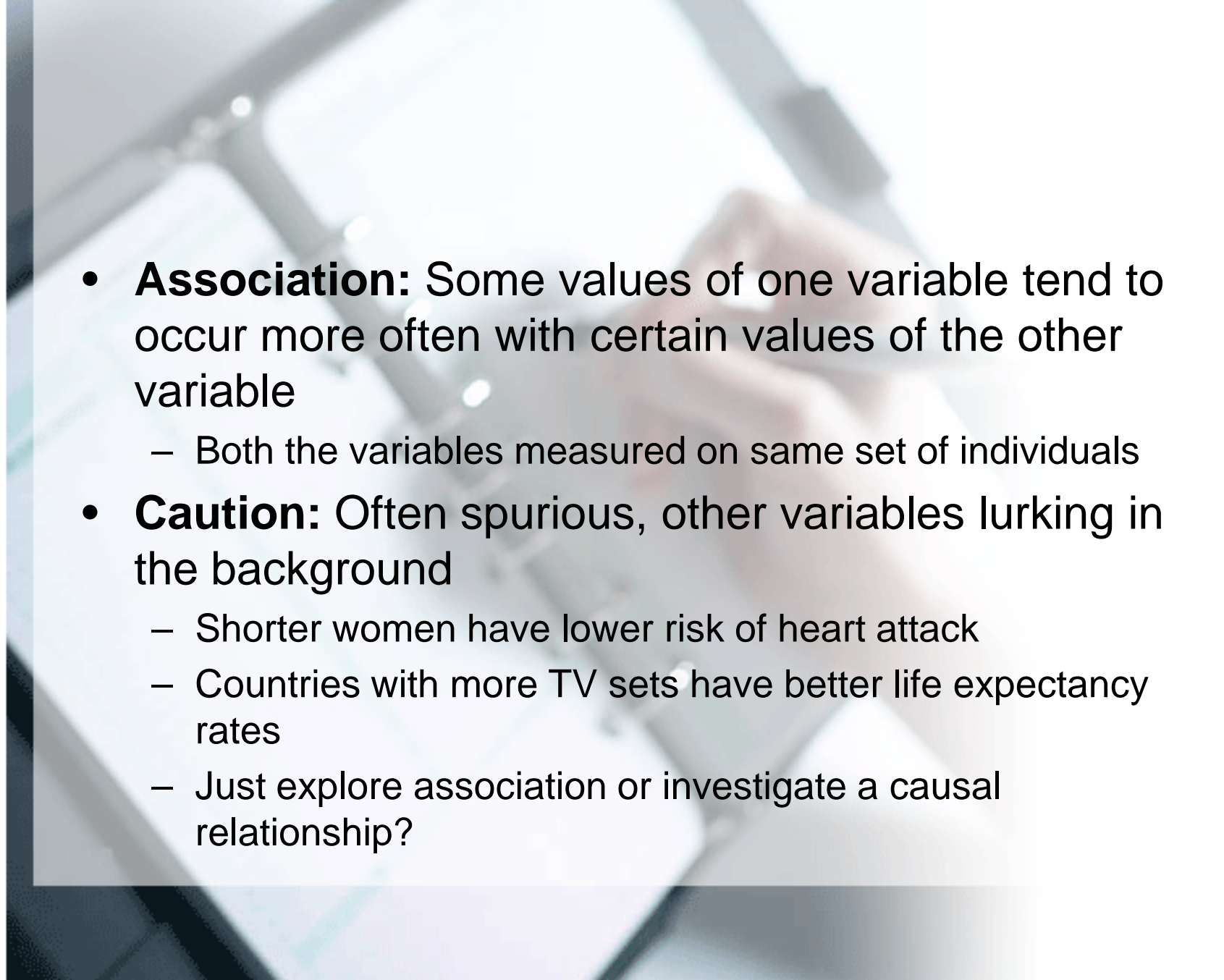


Figure 1-40
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

Looking at Data-Relationships

This is on data with **two** or more **variables**:

- **Response vs Explanatory** variables
- **Scatterplots**
- **Correlation**
 - Height and weight of same individual
 - Smoking habits and life expectancy
 - Age and bone-density of individuals
 - Gender and political affiliation
 - Gender and Smoking

- 
- **Association:** Some values of one variable tend to occur more often with certain values of the other variable
 - Both the variables measured on same set of individuals
 - **Caution:** Often spurious, other variables lurking in the background
 - Shorter women have lower risk of heart attack
 - Countries with more TV sets have better life expectancy rates
 - Just explore association or investigate a causal relationship?

- Who are the **individuals** observed?
- What **variables** are present?
- **Quantitative** or **categorical**?
- Association measures depend on types of variables.
- We will assume Quantitative in this chapter.
- **Response (Y)** measures outcome of interest.
Explanatory (X) explains and *sometimes* causes changes in response variable.

- Different amount of alcohol given to mice, body temperature noted (belief: drop in body temperature with increasing amount of alcohol)

Response: ?

Explanatory: ?

- SAT scores used to predict college GPA

Response: ?

Explanatory: ?

Y: dependent variable

X: independent variable

Here, we have two quantitative variables for each of 16 students.

- 1) How many beers they drank, and
- 2) Their blood alcohol level (BAC)

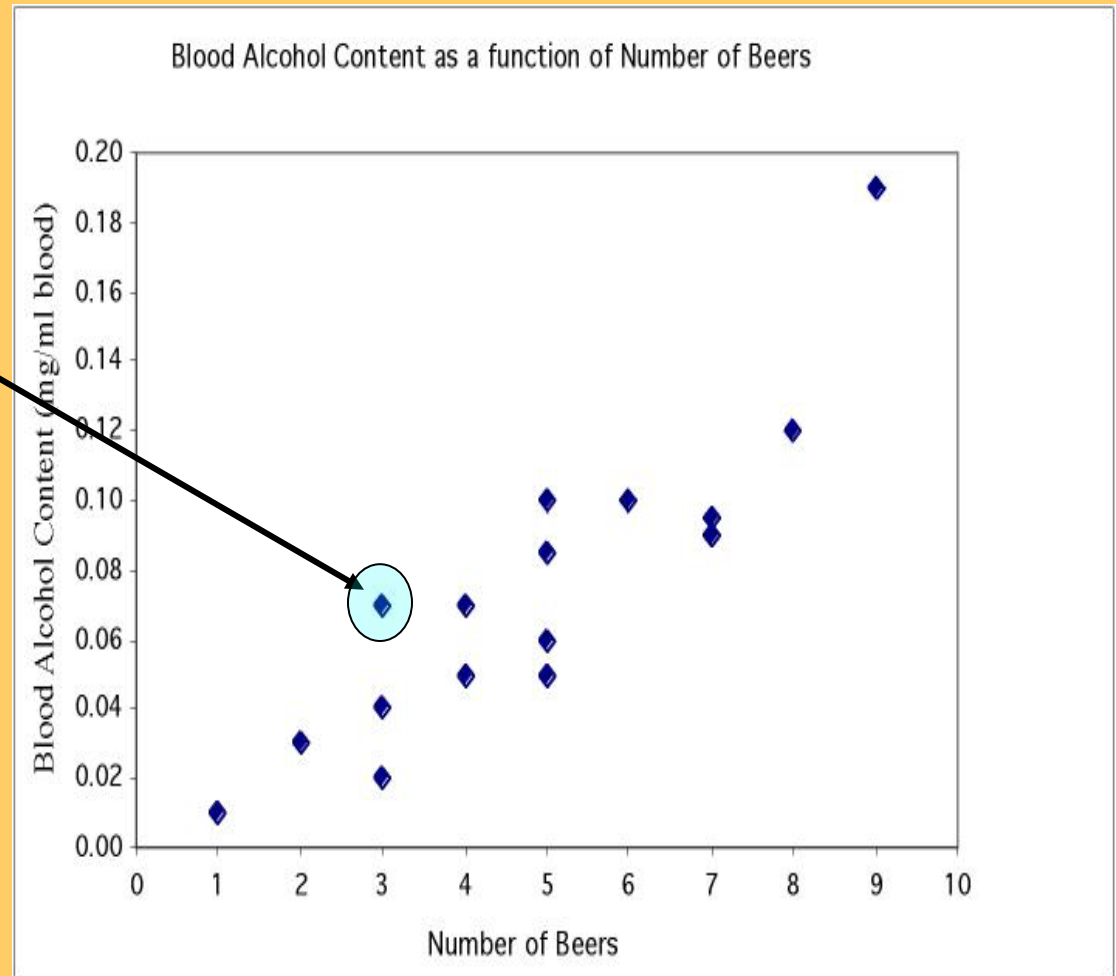
We are interested in the relationship between the two variables: How is one affected by changes in the other one?

Student	Beers	Blood Alcohol
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09

Scatterplots

In a **scatterplot**, one axis is used to represent each of the variables, and the data are plotted as points on the graph.

Student	Beers	BAC
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05

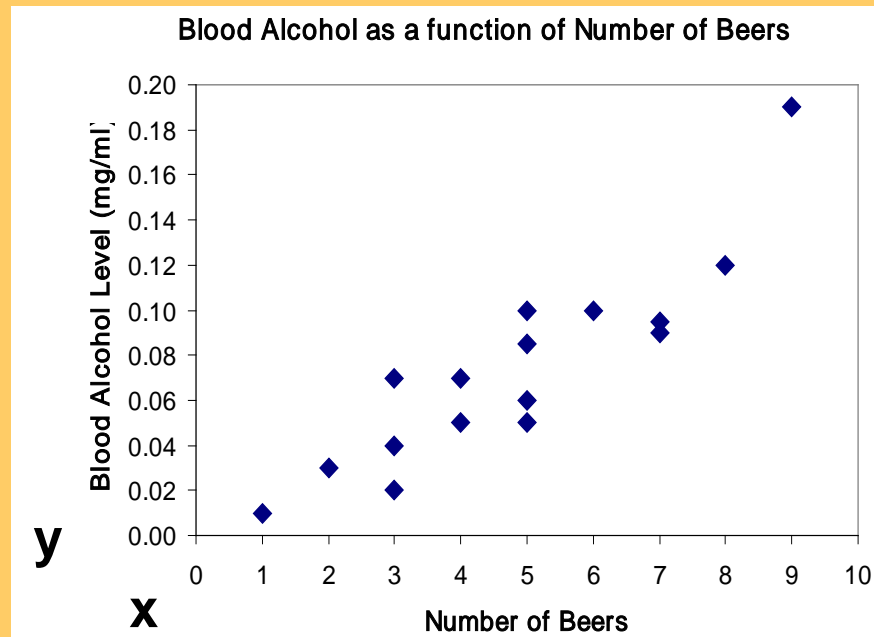


Explanatory and response variables

A **response variable** measures or records an outcome of a study. An **explanatory variable** explains changes in the response variable.

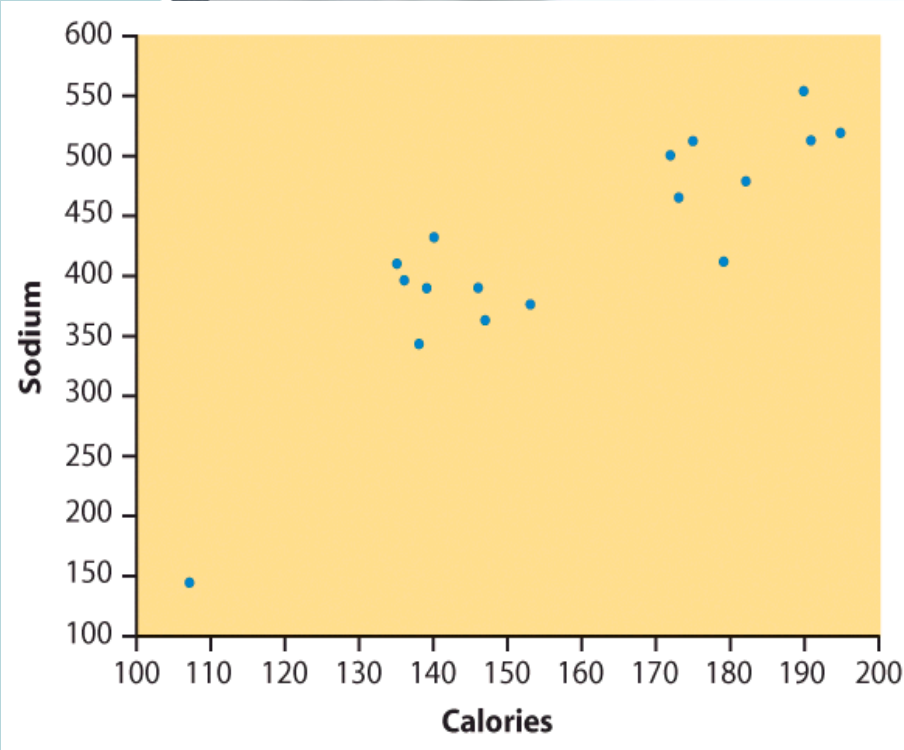
Typically, the *explanatory* or *independent variable* is plotted on the x axis, and the *response* or *dependent variable* is plotted on the y axis.

**Response
(dependent)
variable:**
*blood alcohol
content*



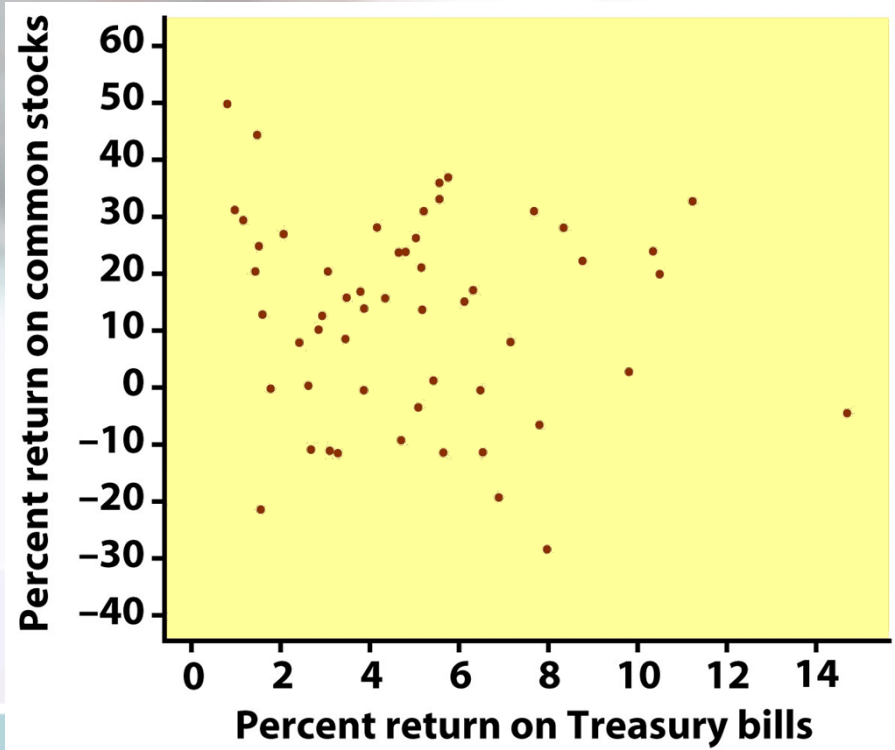
Explanatory (independent) variable:
number of beers

Some plots don't have clear explanatory and response variables.



Do calories explain sodium amounts?

Does percent return on Treasury bills explain percent return on common stocks?

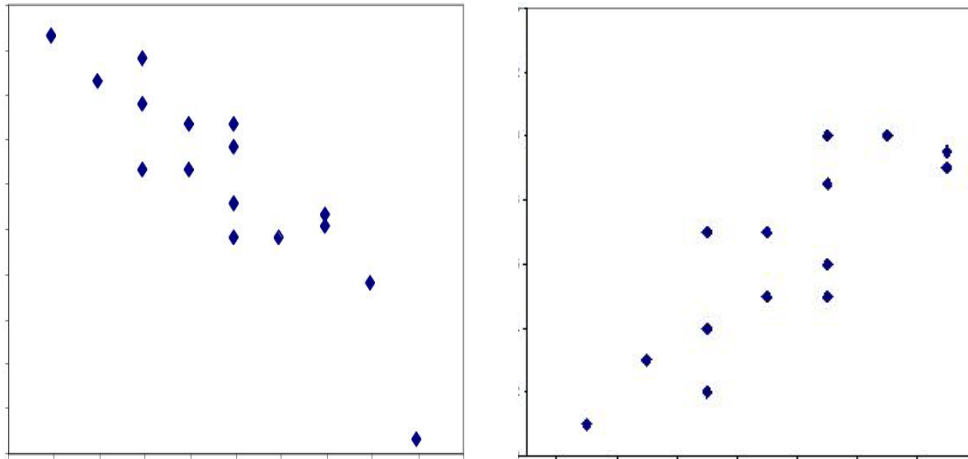


Interpreting scatterplots

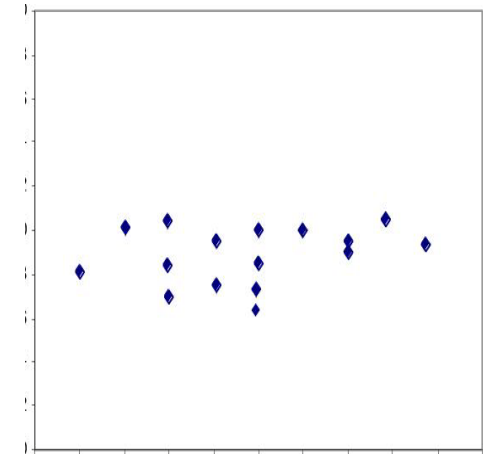
- After plotting two variables on a scatterplot, we describe the relationship by examining the **form**, **direction**, and **strength** of the association. We look for an overall pattern ...
 - Form: linear, curved, clusters, no pattern
 - Direction: positive, negative, no direction
 - Strength: how closely the points fit the “form”
- ... and deviations from that pattern.
 - Outliers

Form and direction of an association

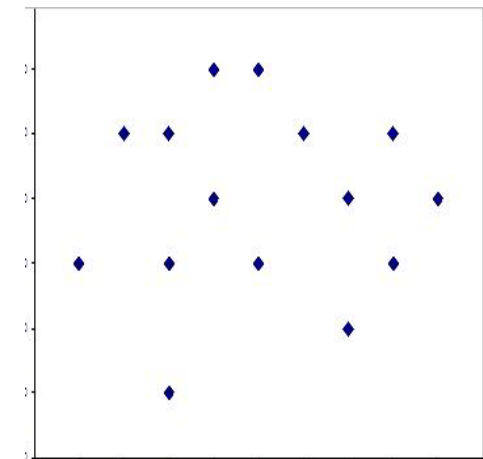
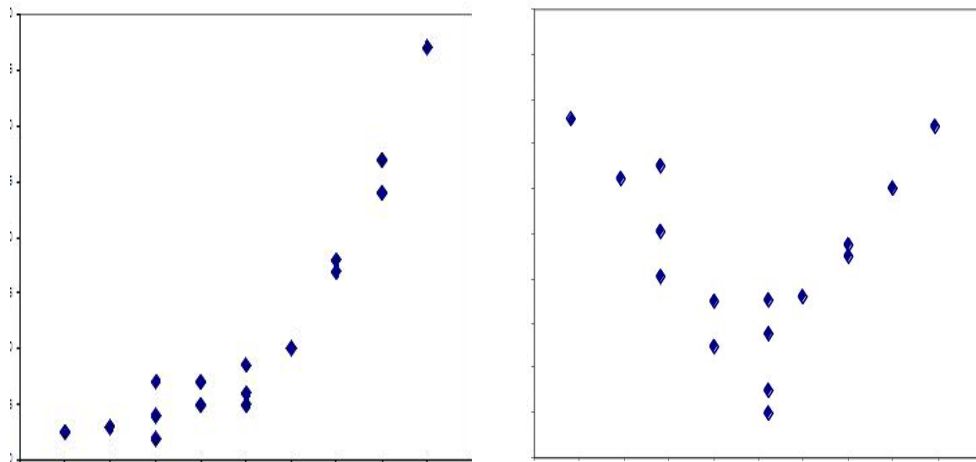
Linear



No relationship

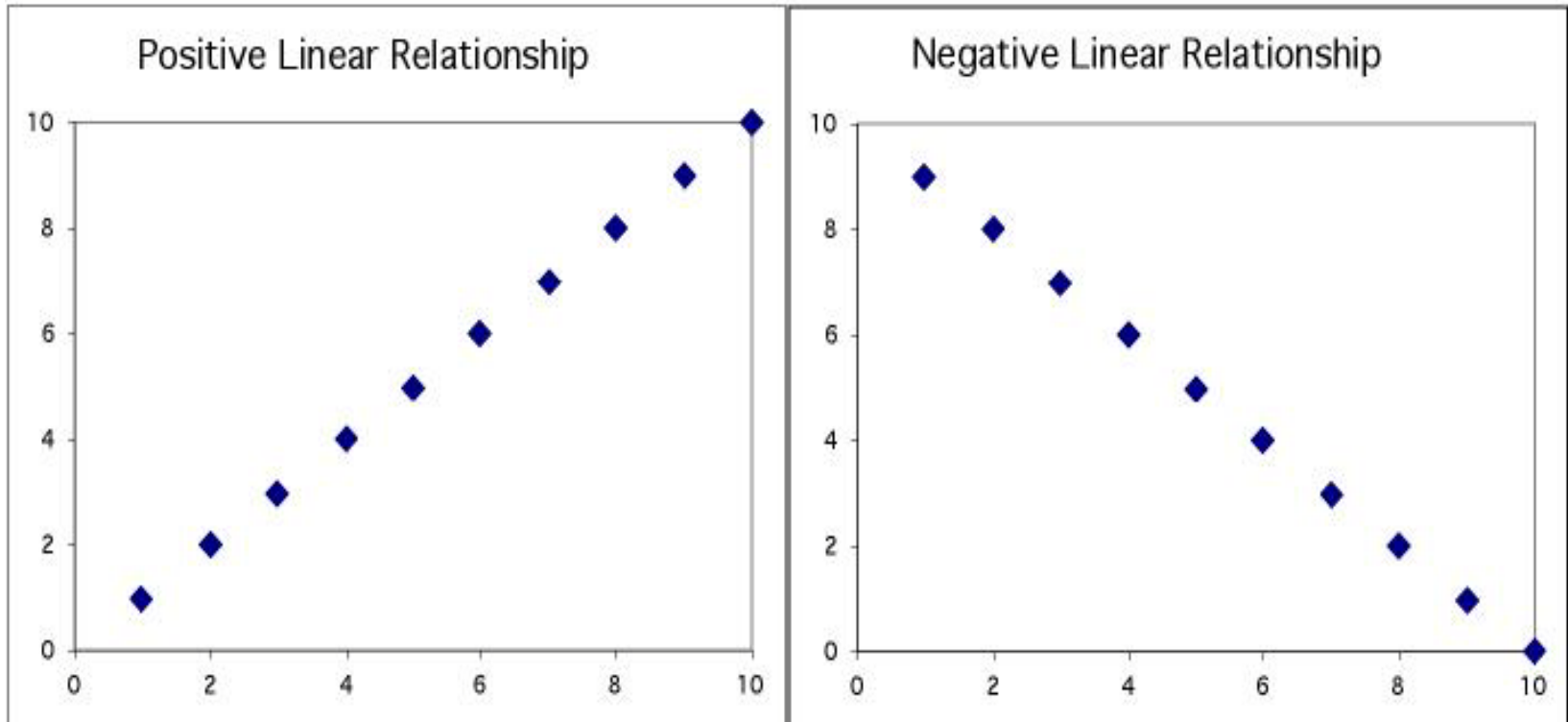


Nonlinear

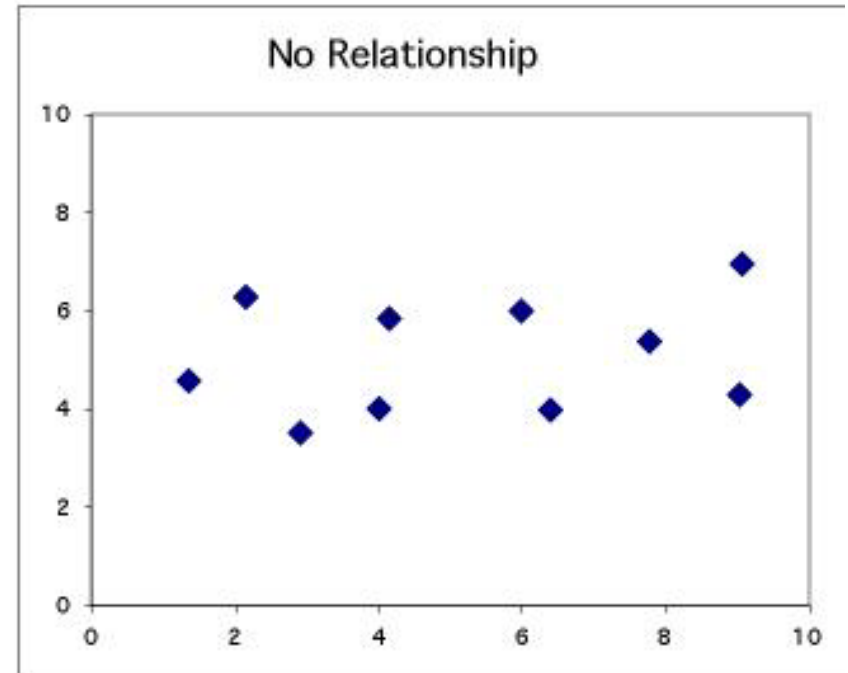
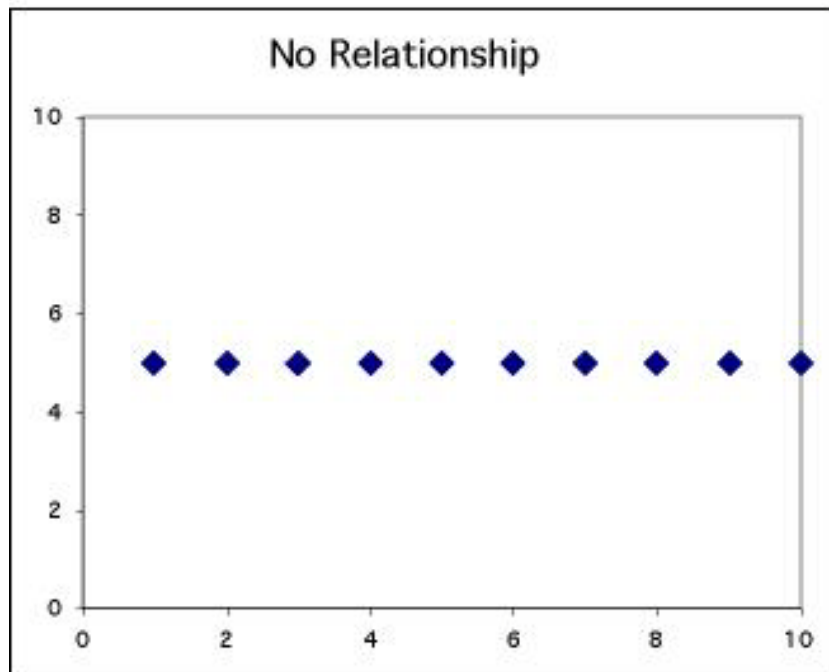


Positive association: High values of one variable tend to occur together with high values of the other variable.

Negative association: High values of one variable tend to occur together with low values of the other variable.



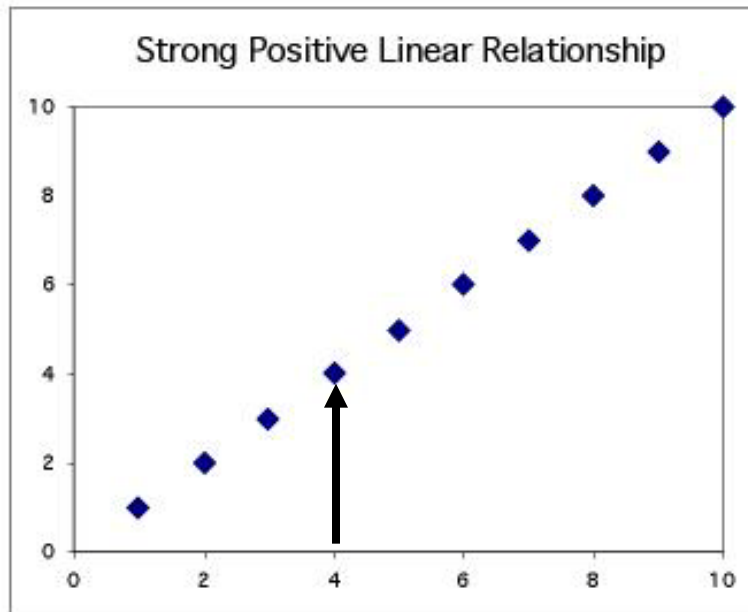
No relationship: X and Y vary independently. Knowing X tells you nothing about Y .



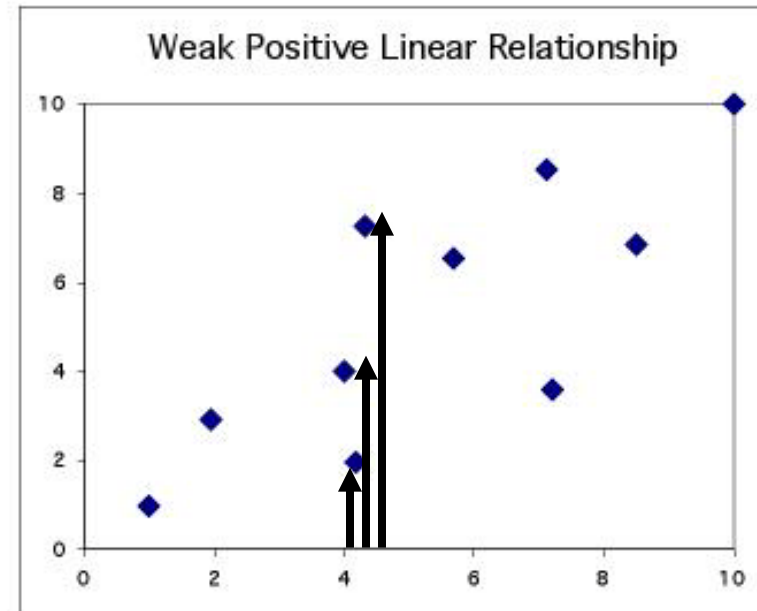
*One way to think about this is to remember the following:
The equation for this line is $y = 5$.
 x is not involved.*

Strength of the association

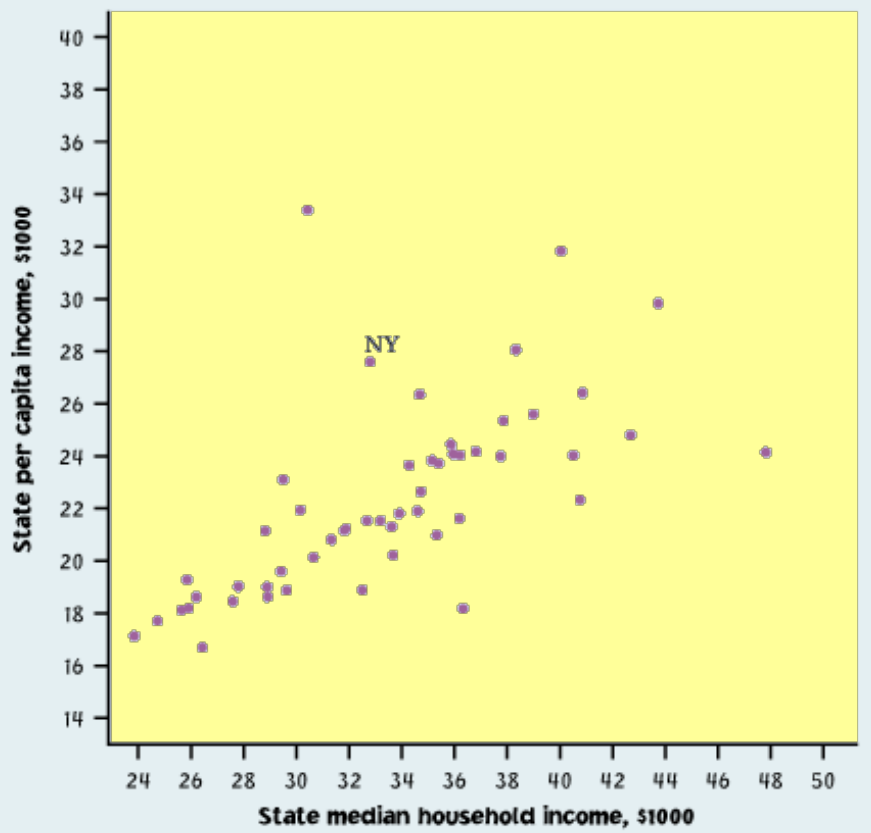
The **strength** of the relationship between the two variables can be seen by how much variation, or **scatter**, there is around the main form.



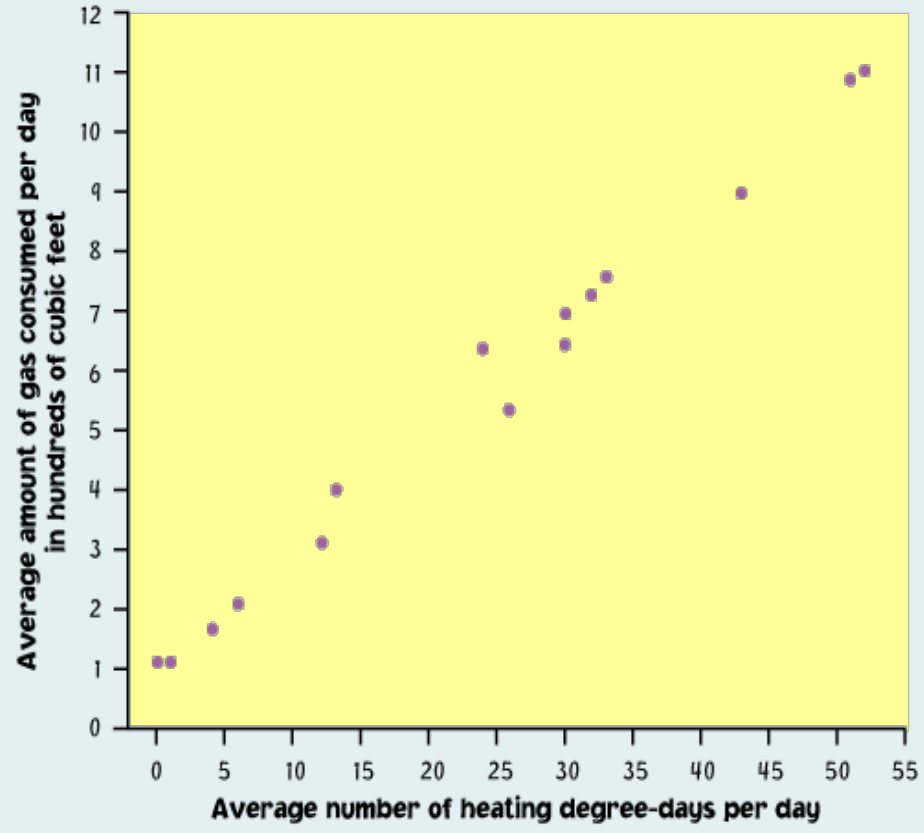
With a strong relationship, you can get a pretty good estimate of y if you know x .



With a weak relationship, for any x you might get a wide range of y values.



This is a **weak** relationship. For a particular state median household income, you can't predict the state per capita income very well.

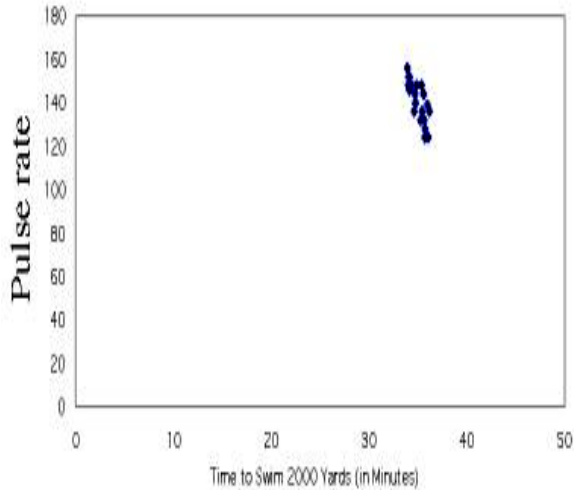


This is a **very strong** relationship. The daily amount of gas consumed can be predicted quite accurately for a given temperature value.

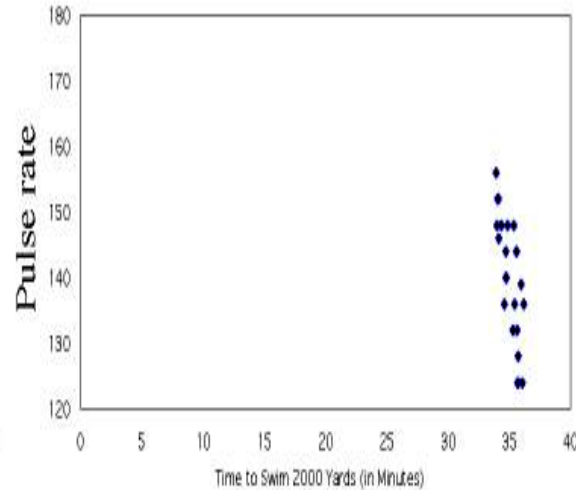
How to scale a scatterplot

Same data in all four plots

Pulse Rate vs Time Spent Swimming

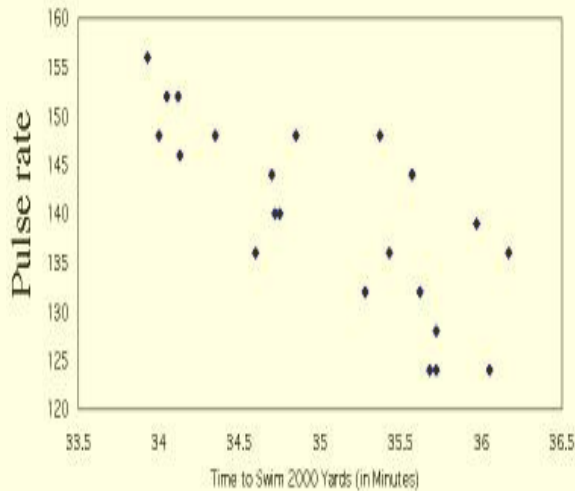


Pulse Rate vs Time Spent Swimming

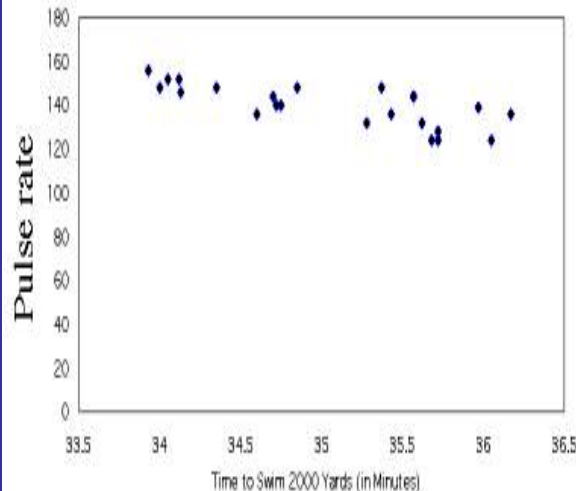


Using an inappropriate scale for a scatterplot can give an incorrect impression.

Pulse Rate vs Time Spent Swimming



Pulse Rate vs Time Spent Swimming



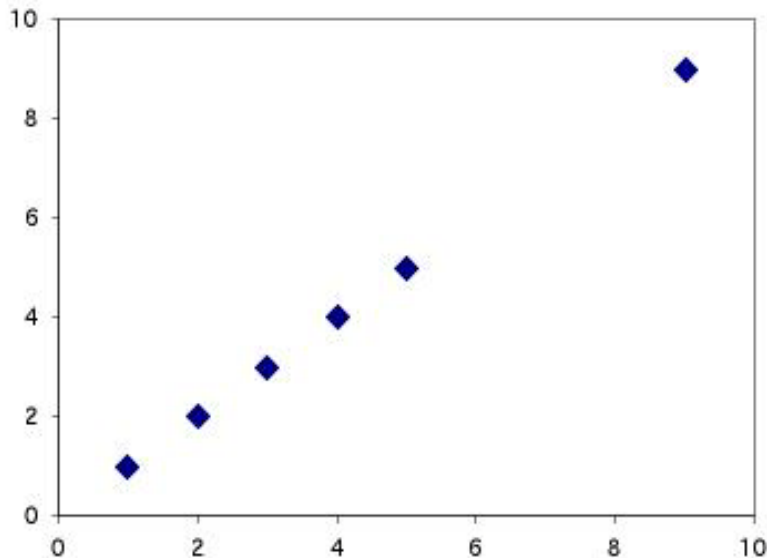
Both variables should be given a similar amount of space:

- Plot roughly square
- Points should occupy all the plot space (no blank space)

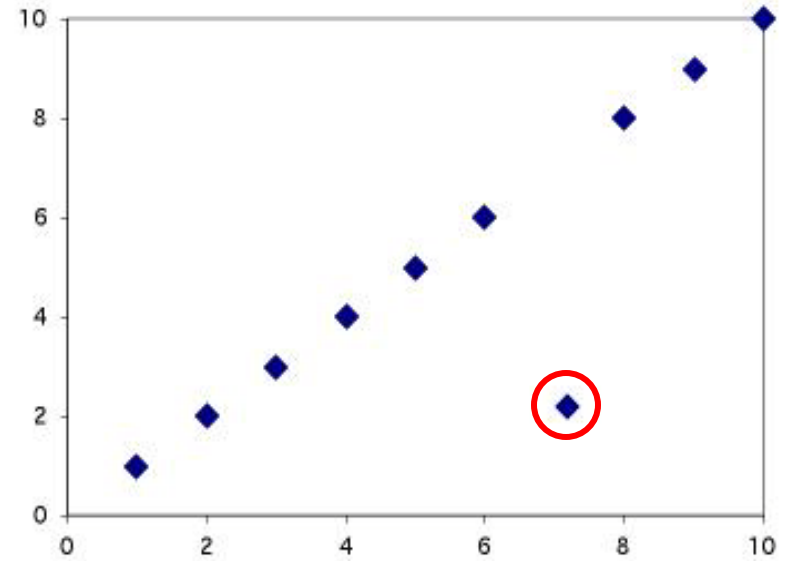
Outliers

An **outlier** is a data value that has a very low probability of occurrence (i.e., it is unusual or unexpected).

Positive Linear Relationship - No Outlier



Positive Linear Relationship with Outlier

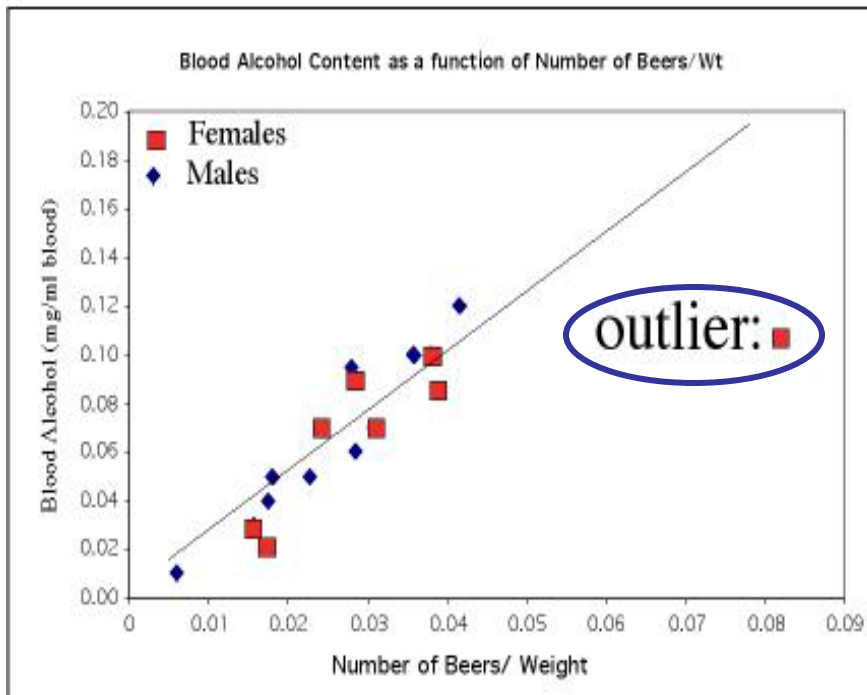
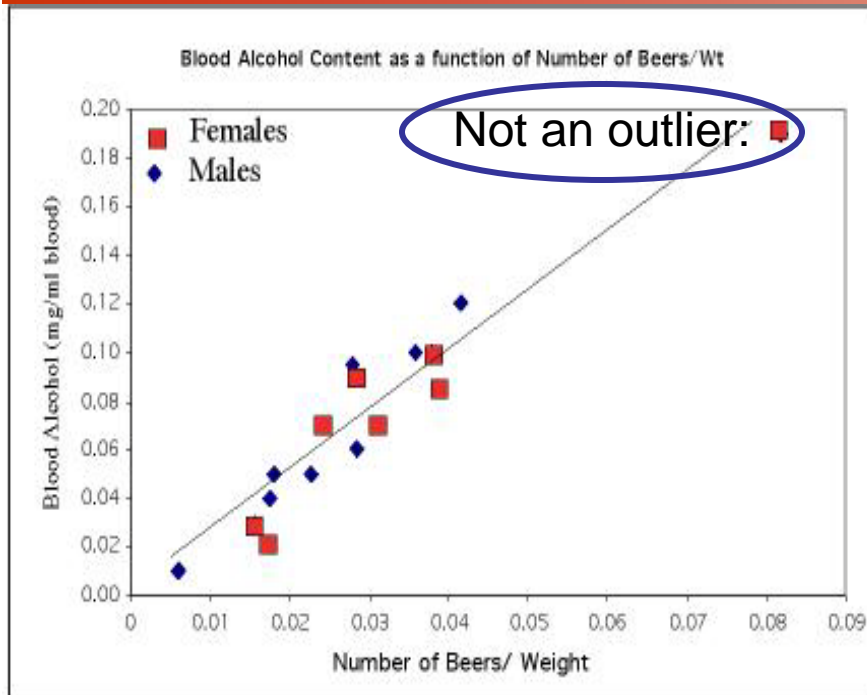


In a scatterplot, outliers are points that fall outside of the overall pattern of the relationship.

Outliers

The upper right-hand point here is not an outlier of the relationship—It is what you would expect for this many beers given the linear relationship between beers/weight and blood alcohol.

This point is not in line with the others, so it is an outlier of the relationship.

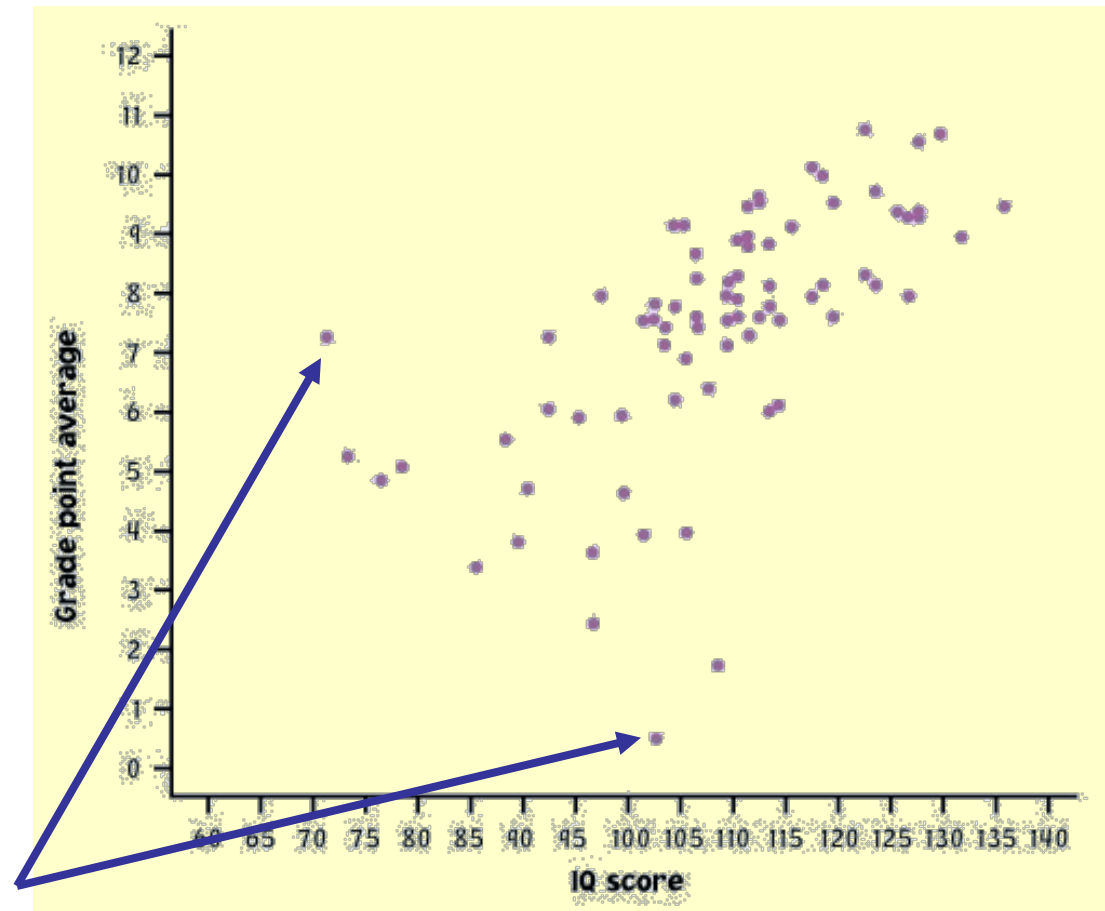


IQ score and Grade point average

a) Describe in words what this plot shows.

b) Describe the direction, shape, and strength. Are there outliers?

c) What is the deal with these people?



R Graphical system

- R is one of the most powerful programs when it comes to drawing and customizing plots. Learning the tricks is not immediate like it is the case with some MS programs, but the rewards are much more significant.
- To make (scatter)plots in R use the function `plot(x,y)` where `x` is the vector of explanatory values and `y` is the vector of responses
- Section 1.3 in the R manual details the basics of making plots
- In addition read about the `lines()` command that adds lines to an existing plot
- One can also make 3D plots using commands:
 - `persp`, `scatterplot3d`, and `wireframe`

Mean SAT total score

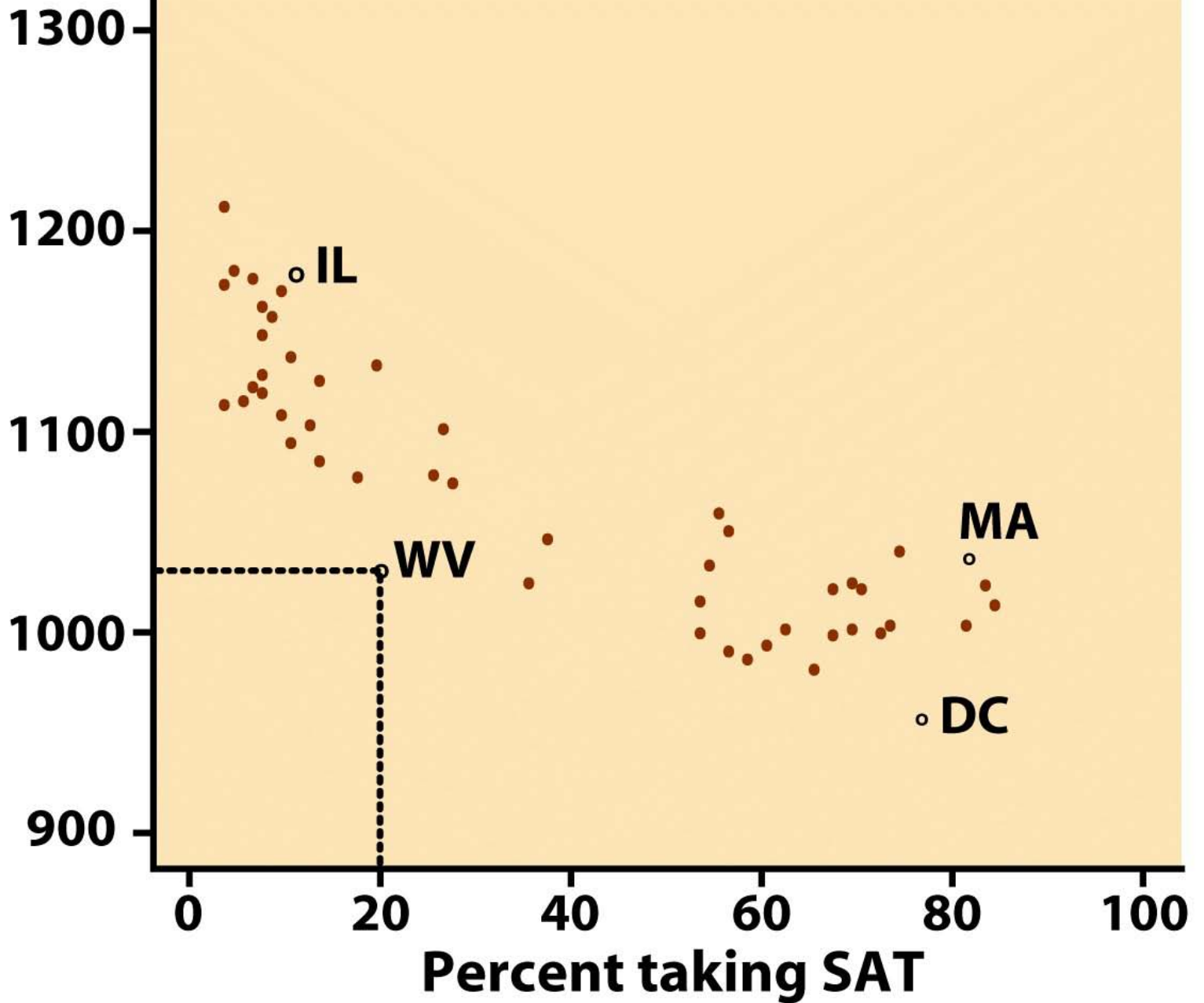


Figure 2-1
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H. Freeman and Company

Example 2: Adding categorical variable/grouping (region): **e** is for northeastern states and **m** is for midwestern states (others excluded). May enhance understanding of the data.

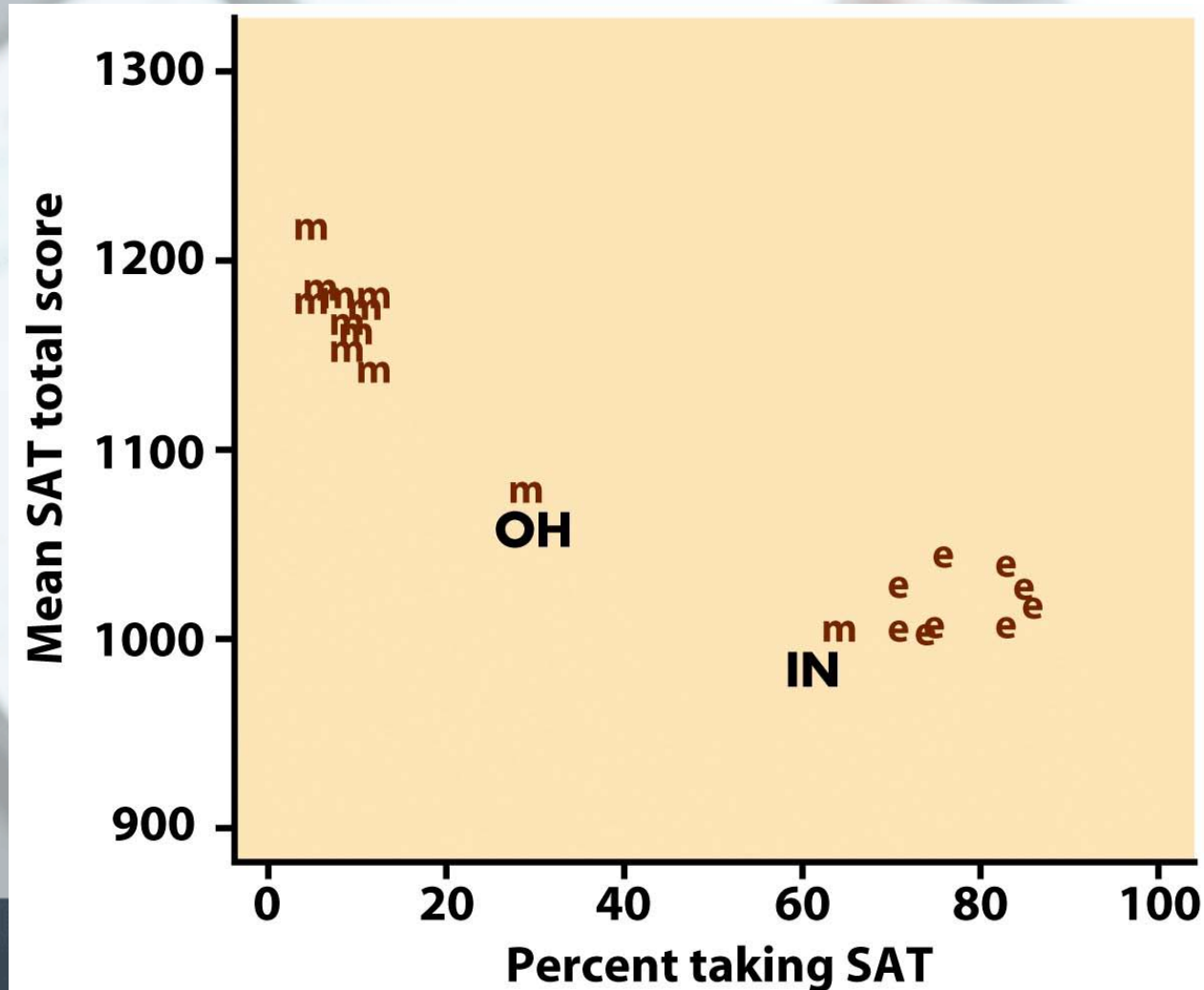
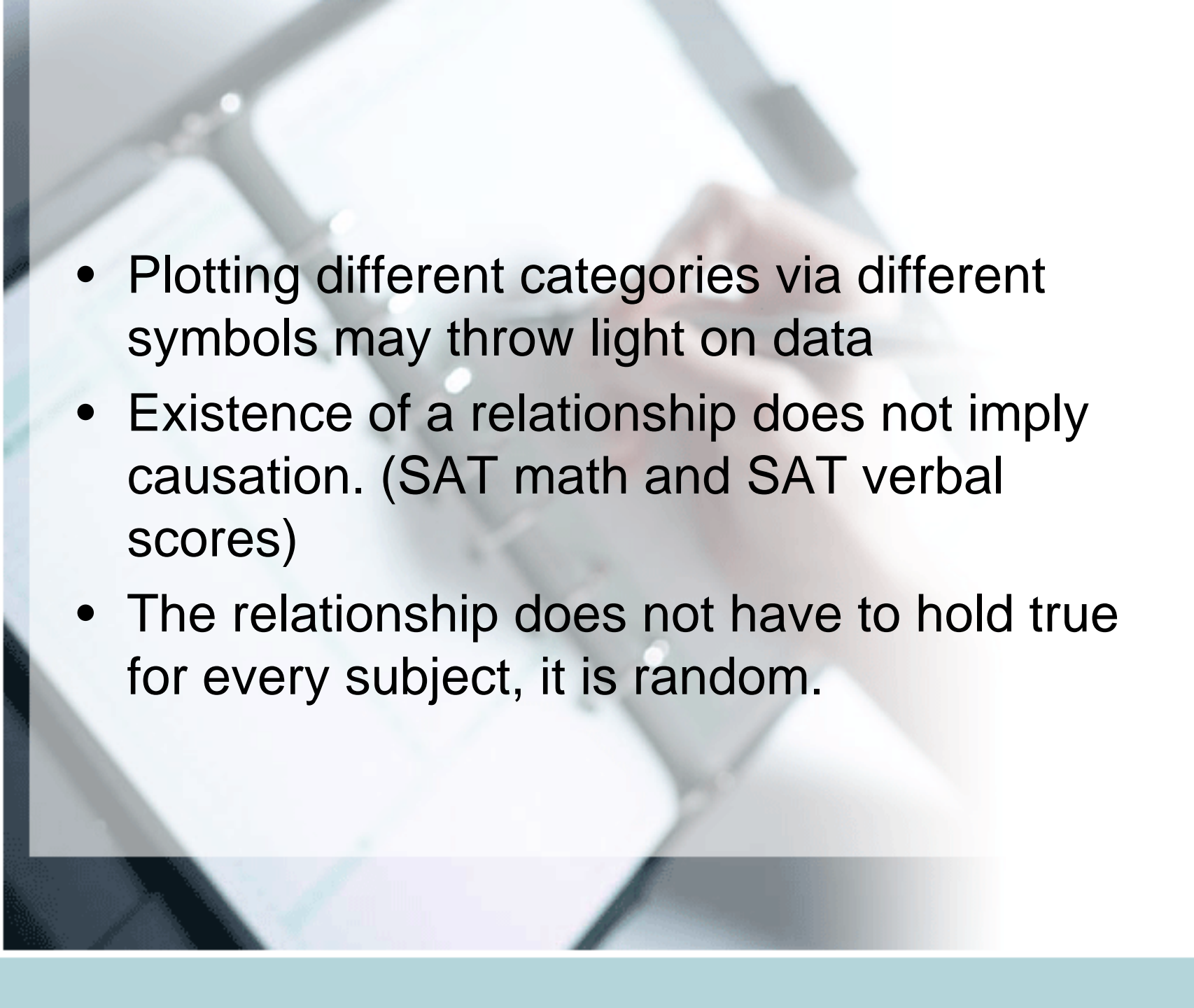


Figure 2-2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

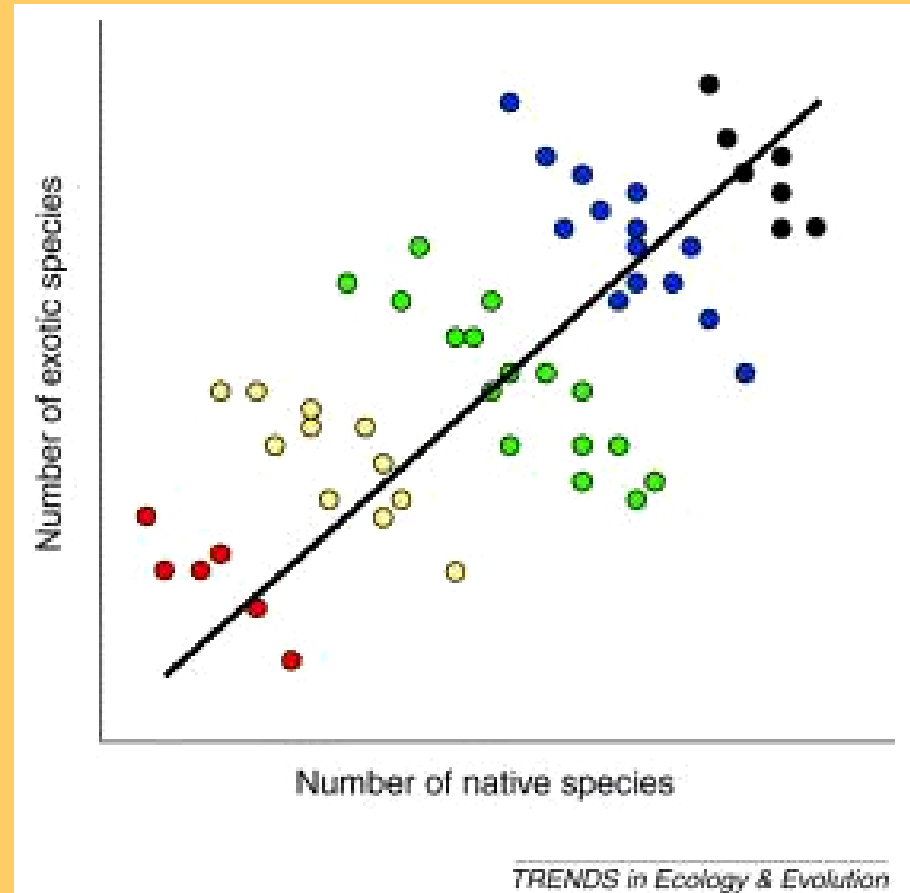
- 
- Plotting different categories via different symbols may throw light on data
 - Existence of a relationship does not imply causation. (SAT math and SAT verbal scores)
 - The relationship does not have to hold true for every subject, it is random.

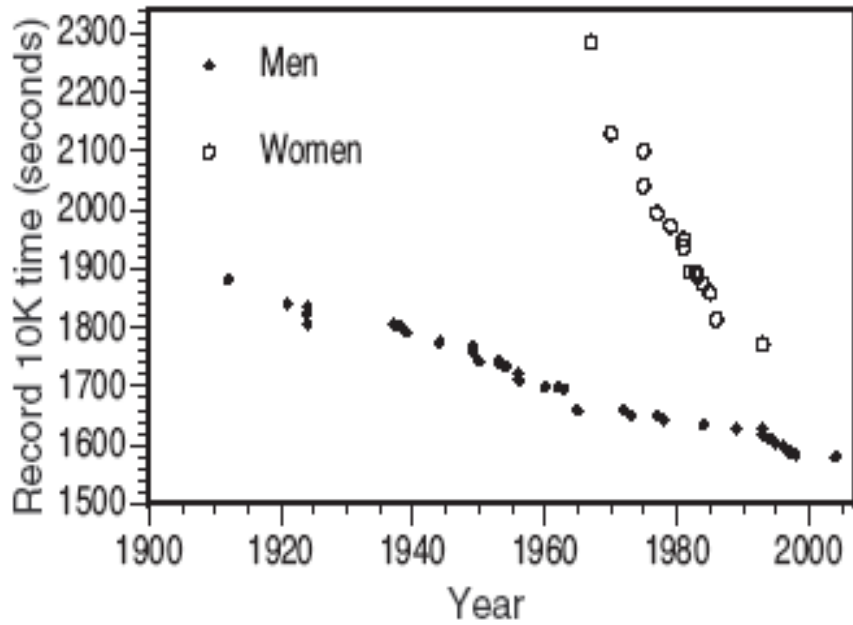
Categorical variables in scatterplots

Often, things are not simple and one-dimensional. We need to group the data into categories to reveal trends.

What may look like a positive linear relationship is in fact a series of negative linear associations.

Plotting different habitats in different colors allows us to make that important distinction.

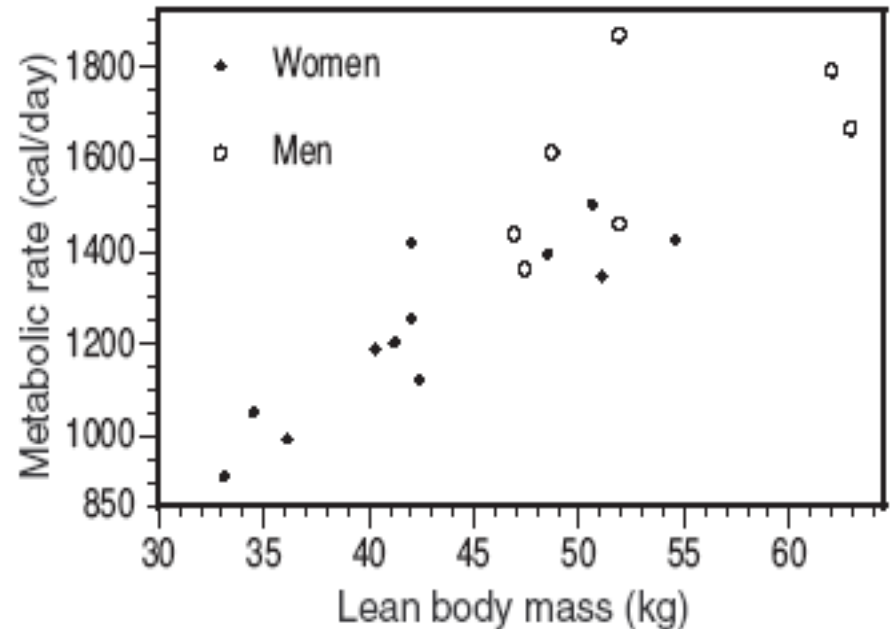




Comparison of men and women racing records over time.

Each group shows a very strong negative linear relationship that would not be apparent without the gender categorization.

Relationship between lean body mass and metabolic rate in men and women. Both men and women follow the same positive linear trend, but women show a stronger association. As a group, males typically have larger values for both variables.

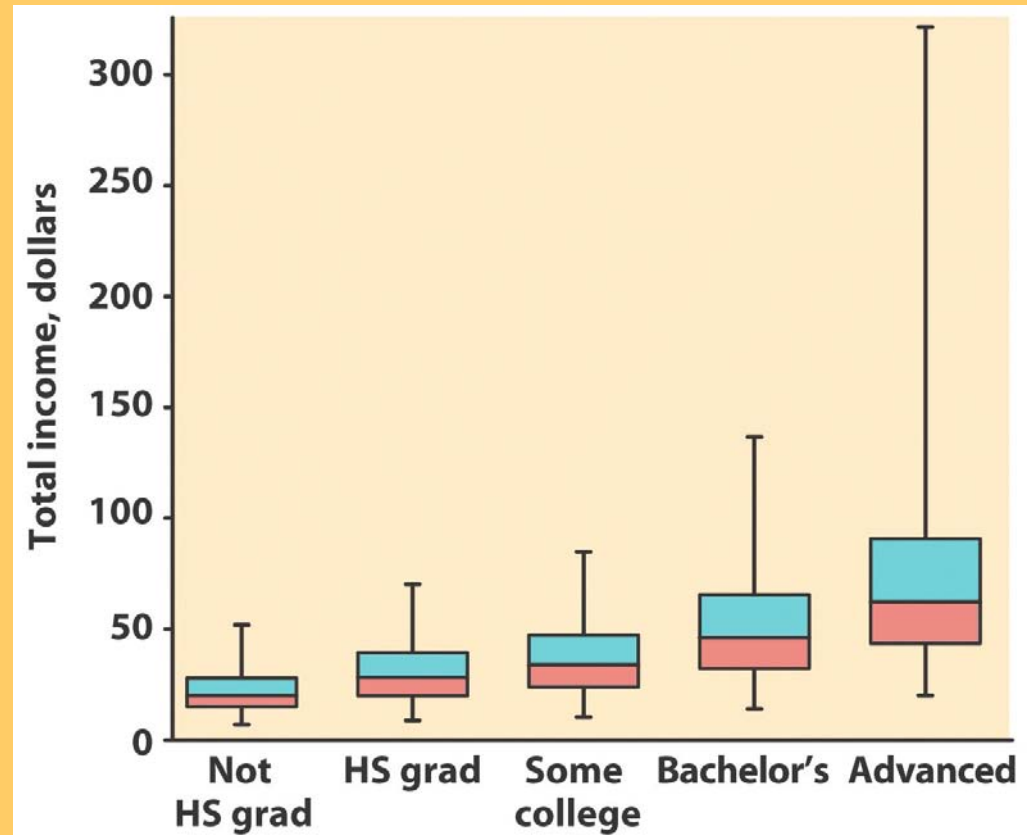


Categorical explanatory variables

When the explanatory variable is categorical, you cannot make a scatterplot, but you can compare the different categories side by side on the same graph (boxplots, or mean \pm standard deviation).

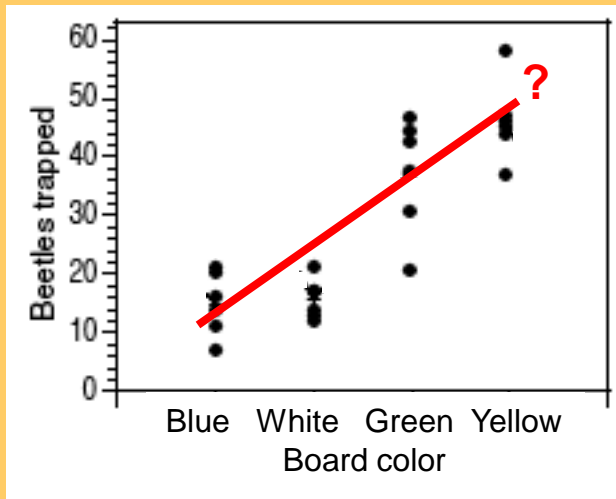
Comparison of income
(quantitative response variable)
for different education levels (five
categories).

**But be careful in your
interpretation: This is NOT a
positive association, because
education is not quantitative.**

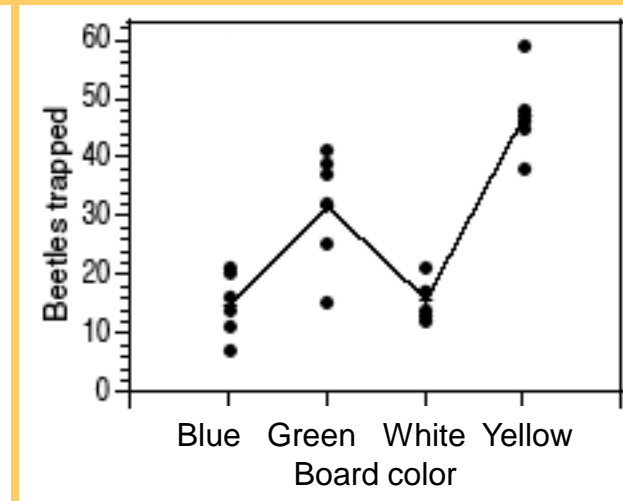
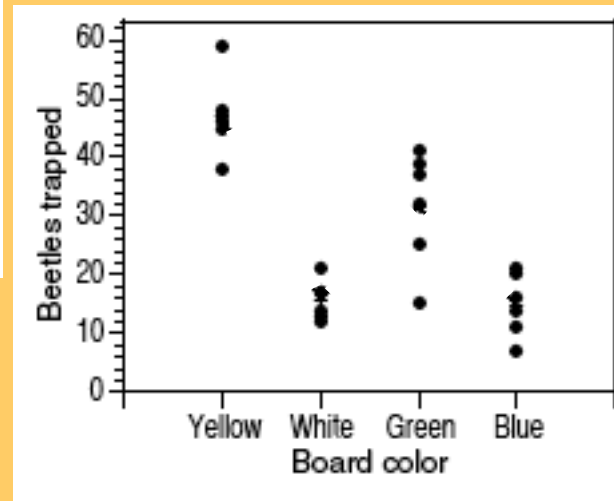


Example: Beetles trapped on boards of different colors

Beetles were trapped on sticky boards scattered throughout a field. The sticky boards were of four different colors (categorical explanatory variable). The number of beetles trapped (response variable) is shown on the graph below.



What association? What relationship?



→ Describe one category at a time.

When both variables are quantitative, the order of the data points is defined entirely by their value. This is not true for categorical data.

Correlation Coefficient

- Linear relationships are quite common.
- Correlation coefficient r measures strength and direction of a linear relationship between two quantitative variables X and Y .
- Data structure: (X, Y) pairs measured on n individuals
- (weight, blood pressure) or (age, bone-density) measured on a set of subjects

The correlation coefficient "r"

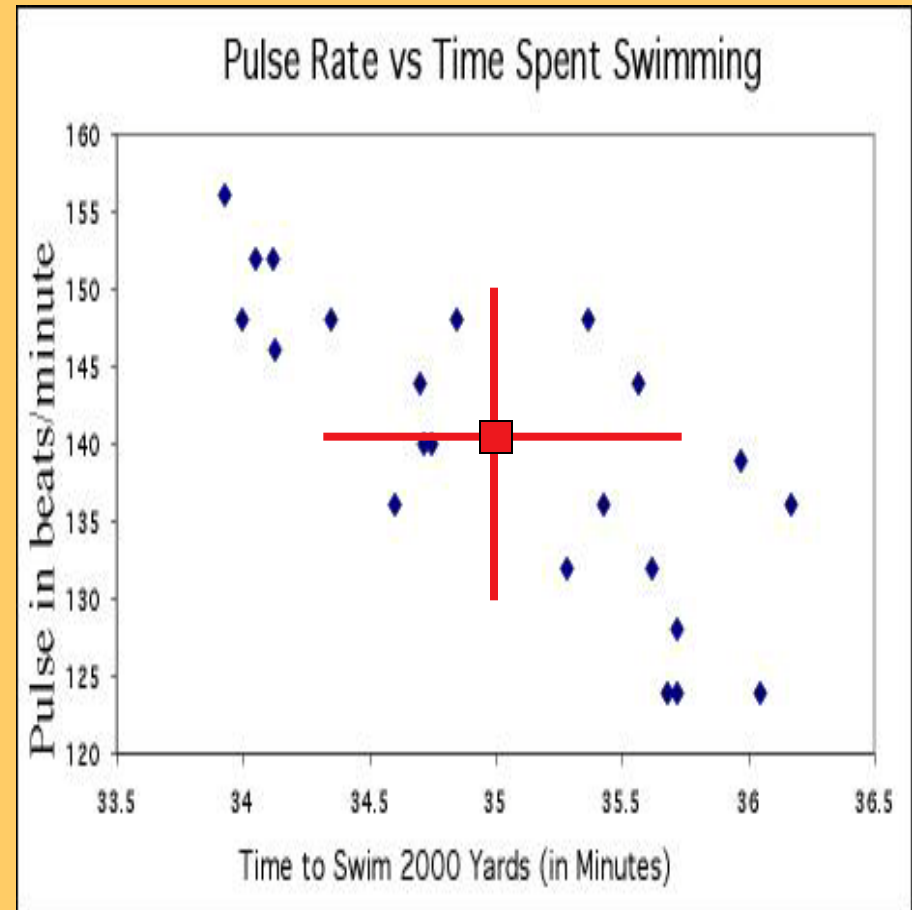
The correlation coefficient is a measure of the direction and strength of a relationship. It is calculated using the mean and the standard deviation of both the x and y variables.

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Time to swim: $\bar{x} = 35$, $s_x = 0.7$

Pulse rate: $\bar{y} = 140$, $s_y = 9.5$

*Correlation can only be used to describe **quantitative** variables. Categorical variables don't have means and standard deviations.*

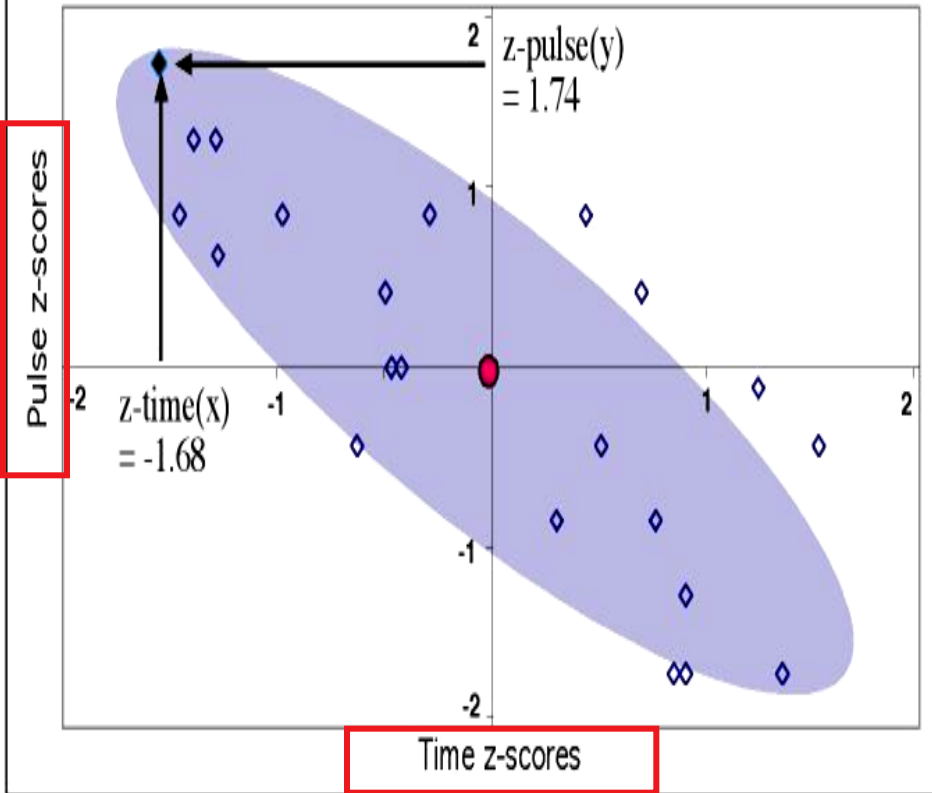




$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

z for time z for pulse

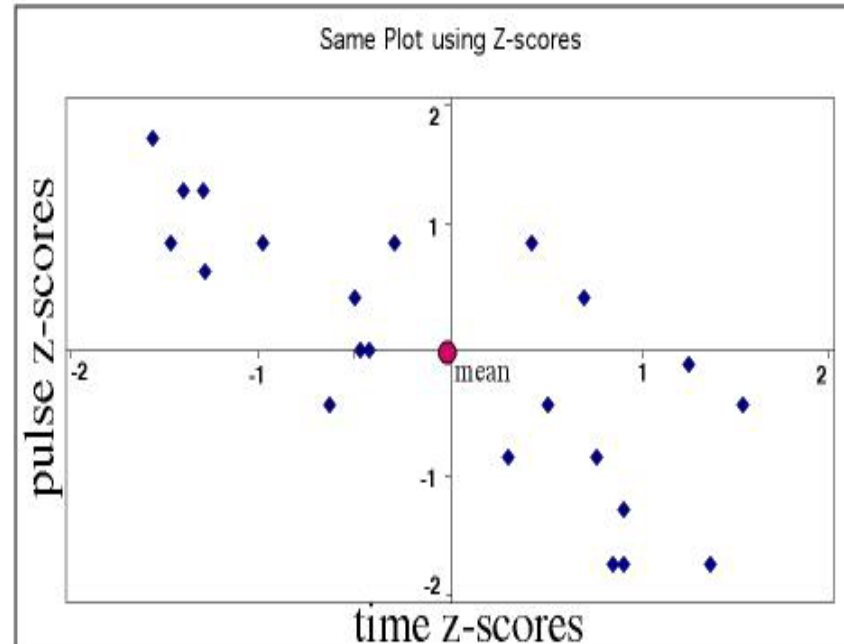
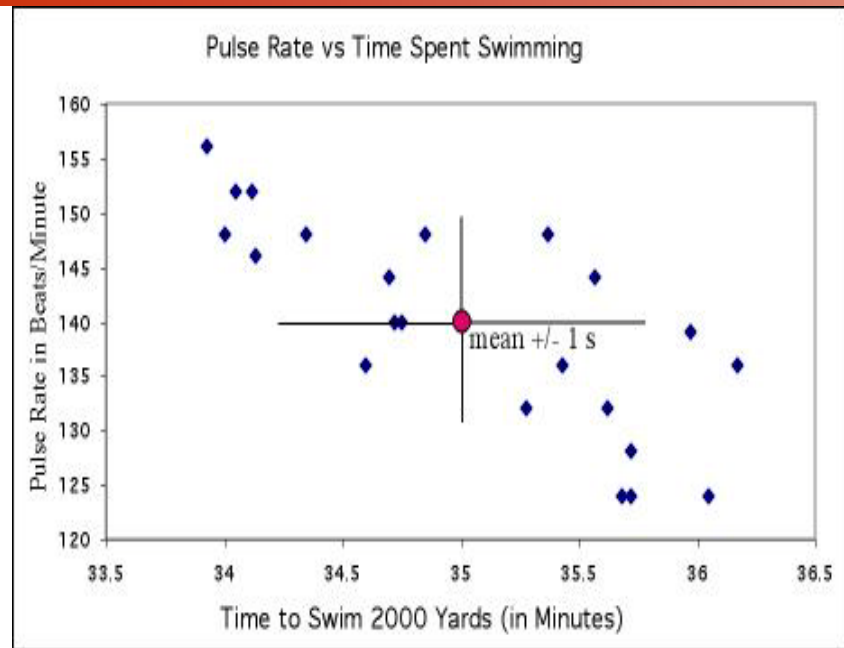
Product of z-scores for this point =
(z-pulse)(z-time) = (-1.74)(1.68) = -2.92



Part of the calculation involves finding z, the standardized score we used when working with the normal distribution.

In R use `cor()`
Read Section 5.4 in your R manual (called there Pearson Correlation).

*You DON'T want to do this by hand.
Make sure you learn how to use your calculator!*



Standardization:

Allows us to compare correlations between data sets where variables are measured in different units or when variables are different.

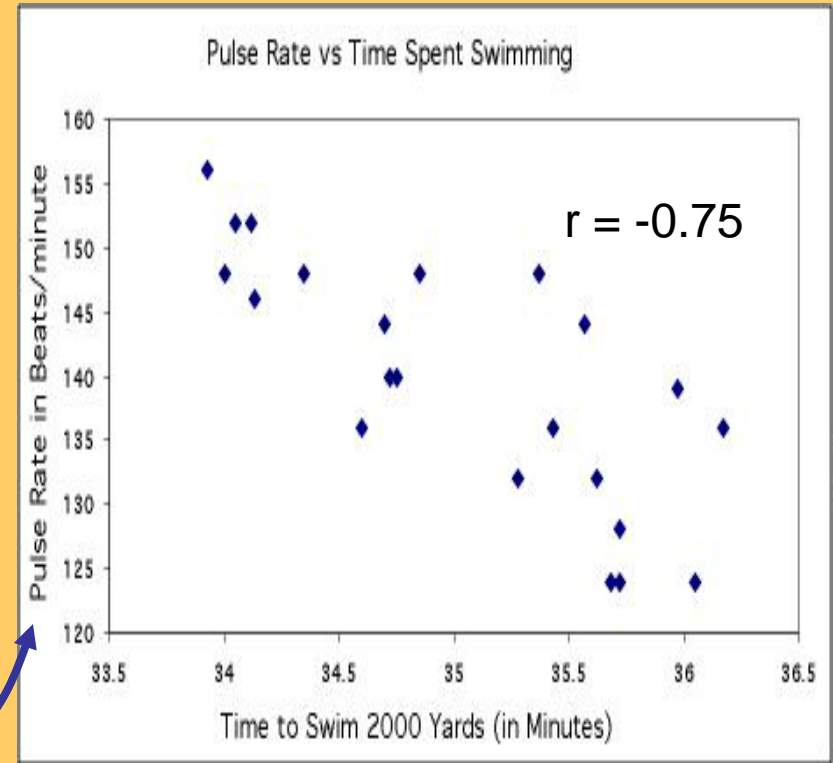
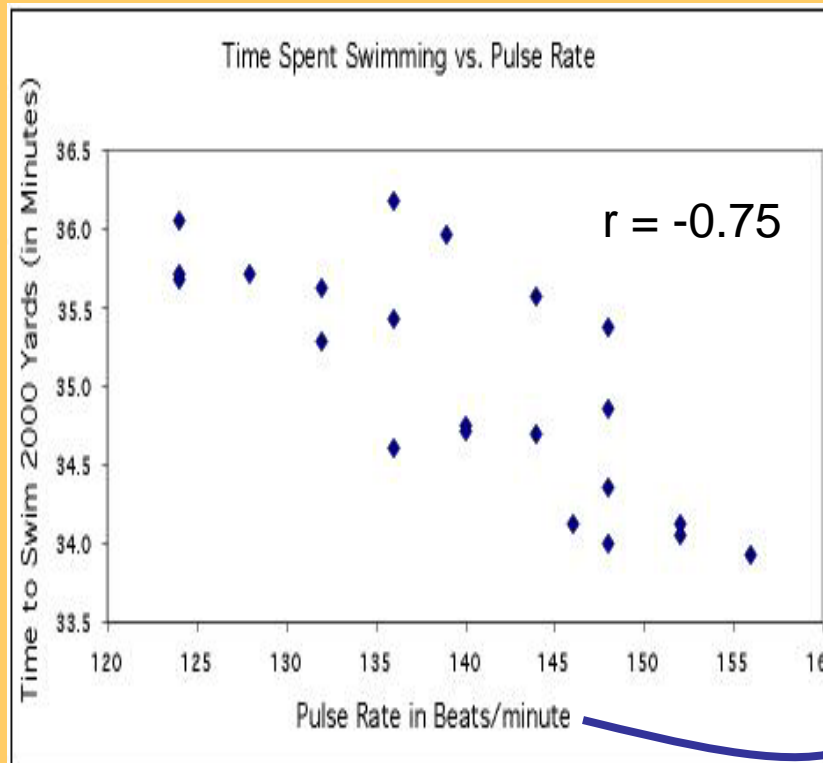
For instance, we might want to compare the correlation between [swim time and pulse], with the correlation between [swim time and breathing rate].



“r” does not distinguish x & y

The correlation coefficient, r, treats x and y symmetrically.

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$



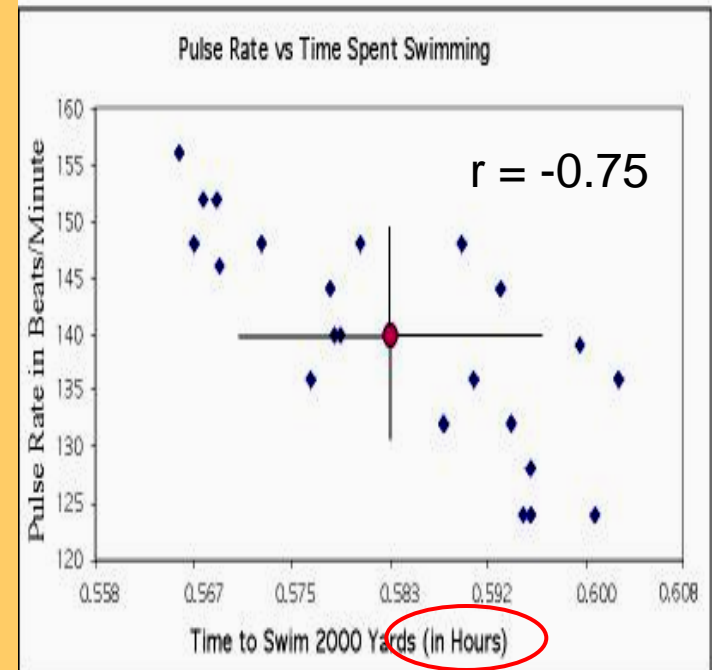
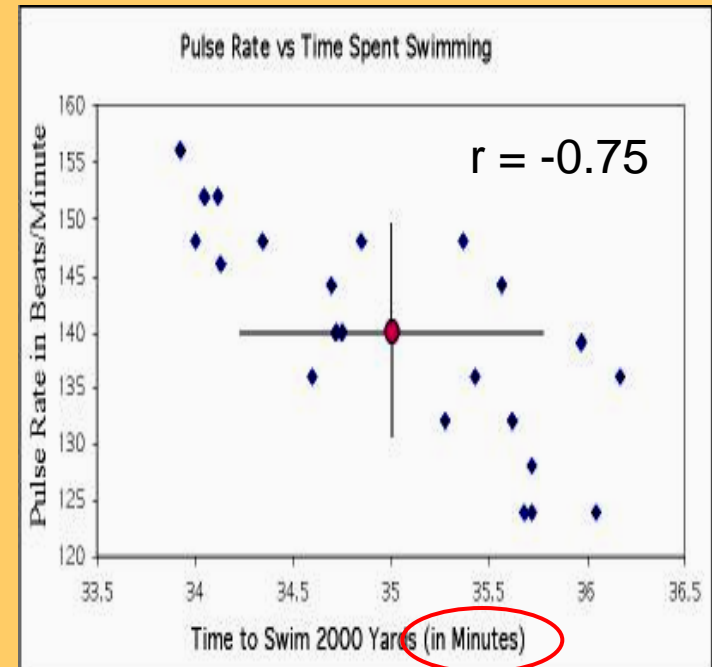
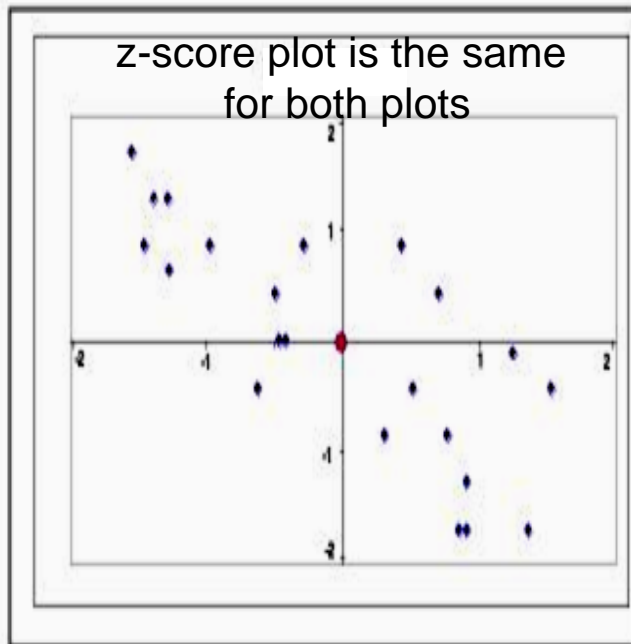
"Time to swim" is the explanatory variable here, and belongs on the x axis. However, in either plot r is the same (r=-0.75).

"r" has no unit

Changing the units of variables does not change the correlation coefficient "r", because we get rid of all our units when we standardize (get z-scores).

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

z for time z for pulse

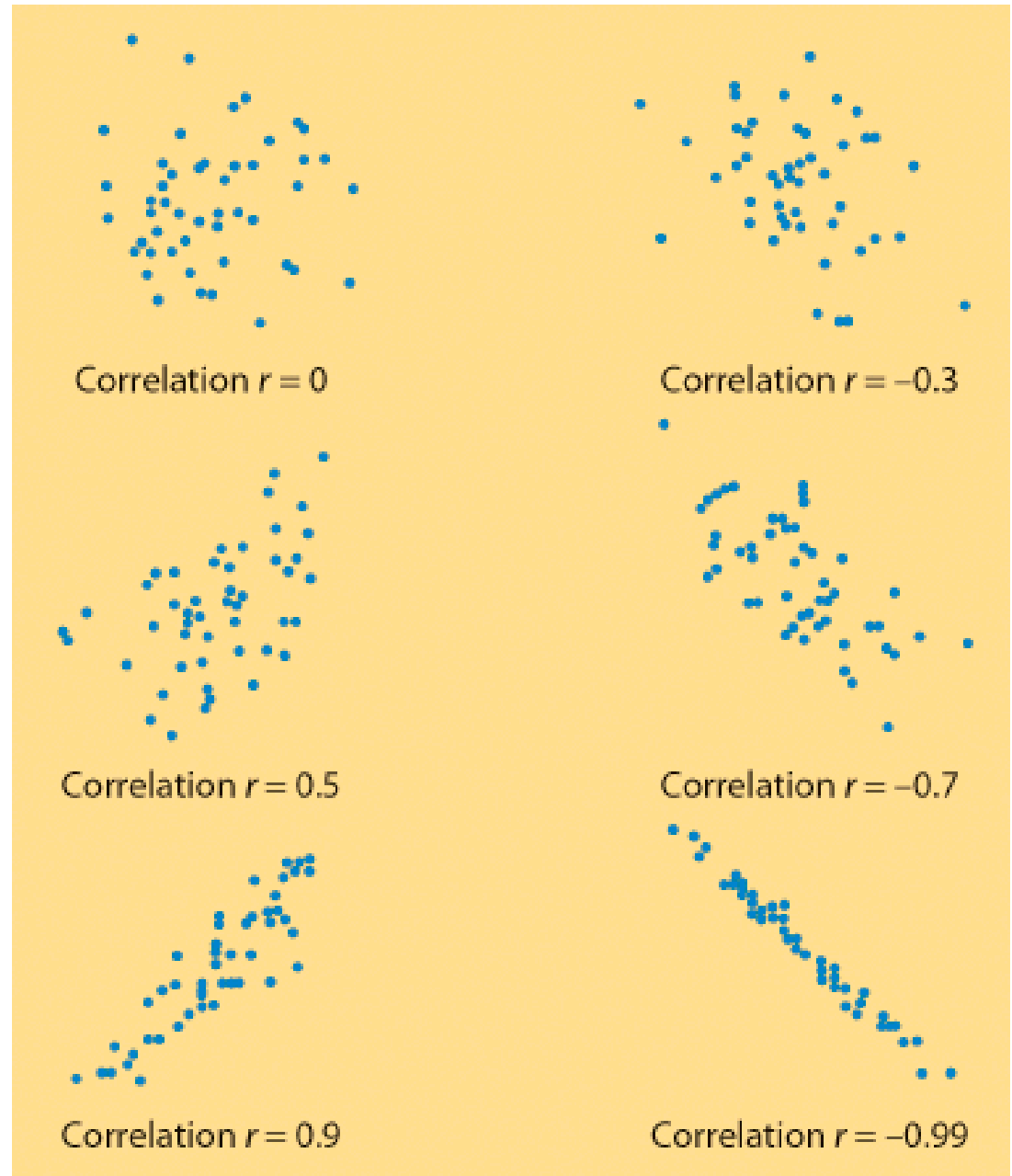


"r" ranges from -1 to +1

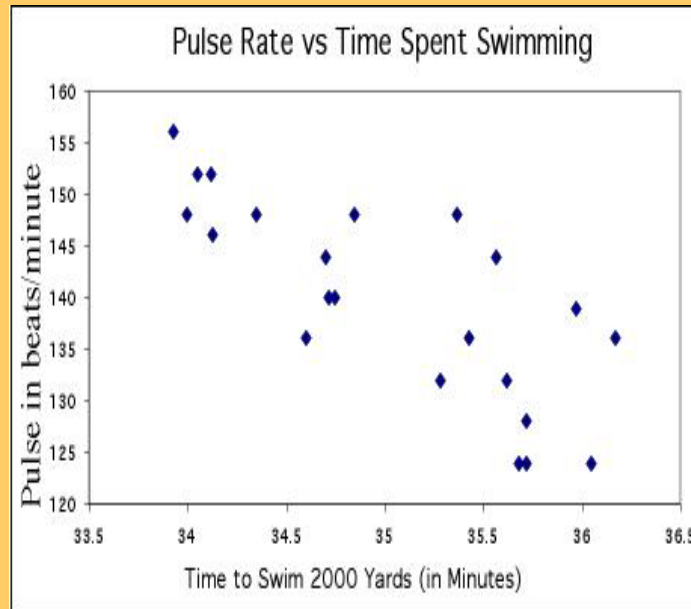
"r" quantifies the **strength** and **direction** of a linear relationship between 2 quantitative variables.

Strength: how closely the points follow a straight line.

Direction: is positive when individuals with higher X values tend to have higher values of Y .

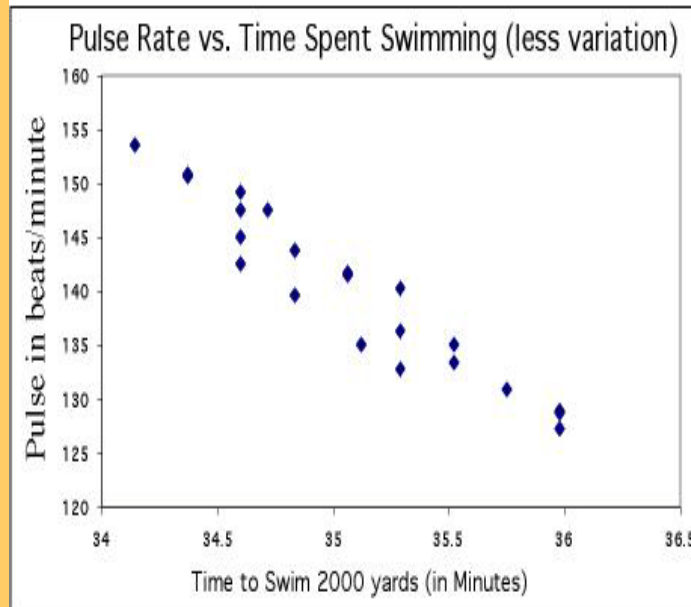


When variability in one or both variables decreases, the correlation coefficient gets stronger (→ closer to +1 or -1).



Pulse: mean = 140.0, $s = 9.5$
Time: mean = 35.0, $s = 0.7$

$r = -0.75$

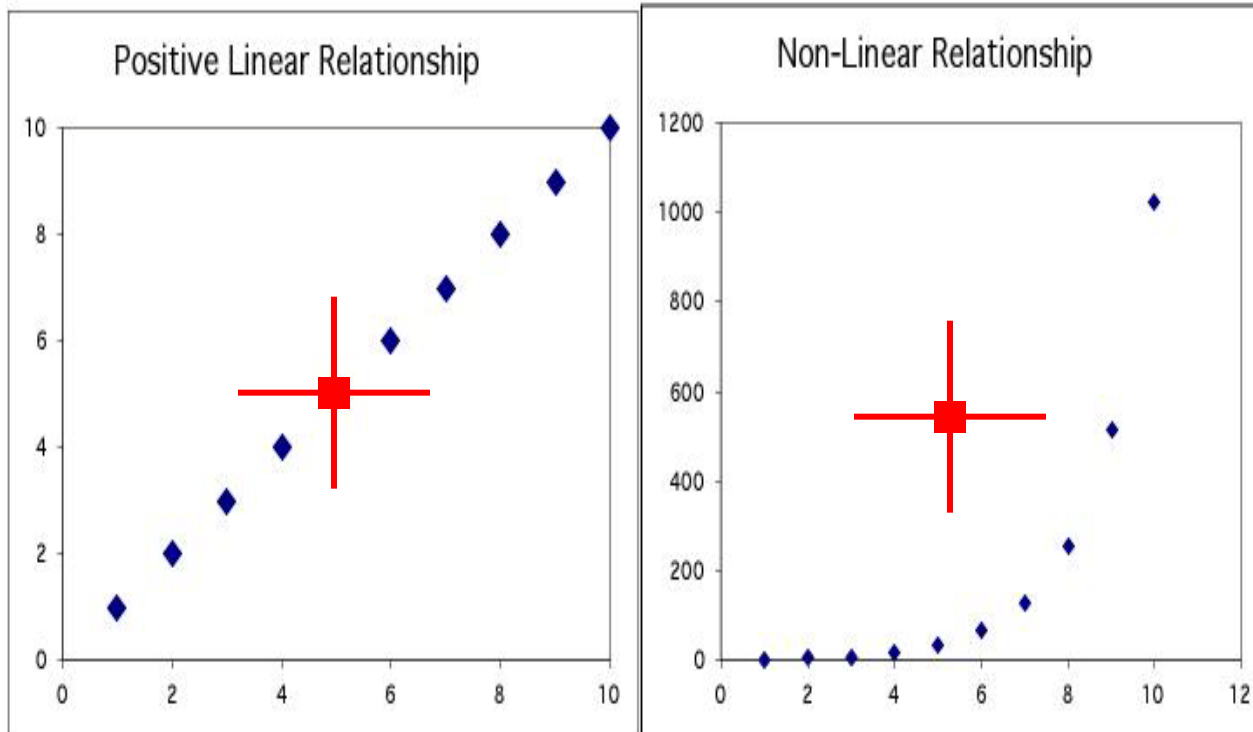


Pulse: mean = 140.0, $s = 7.7$
Time: mean = 35.0, $s = 0.5$

$r = -0.91$

Correlation only describes linear relationships

No matter how strong the association, r does not describe curved relationships.

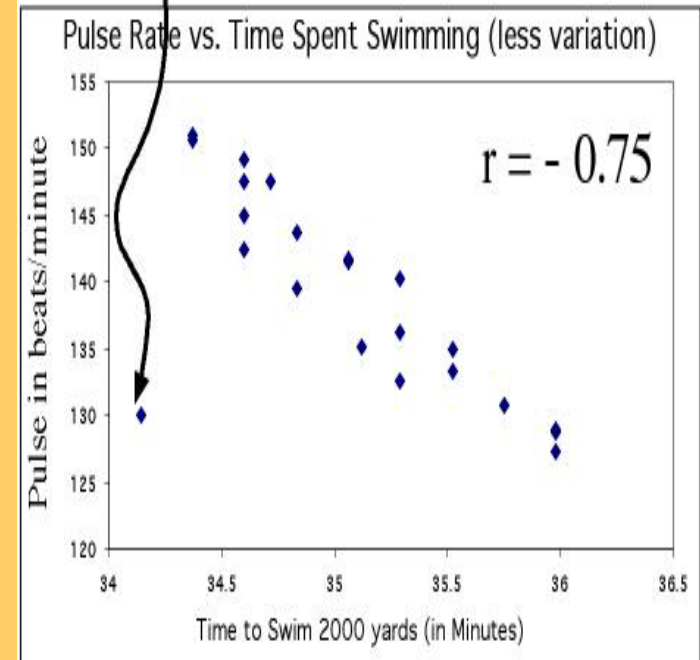
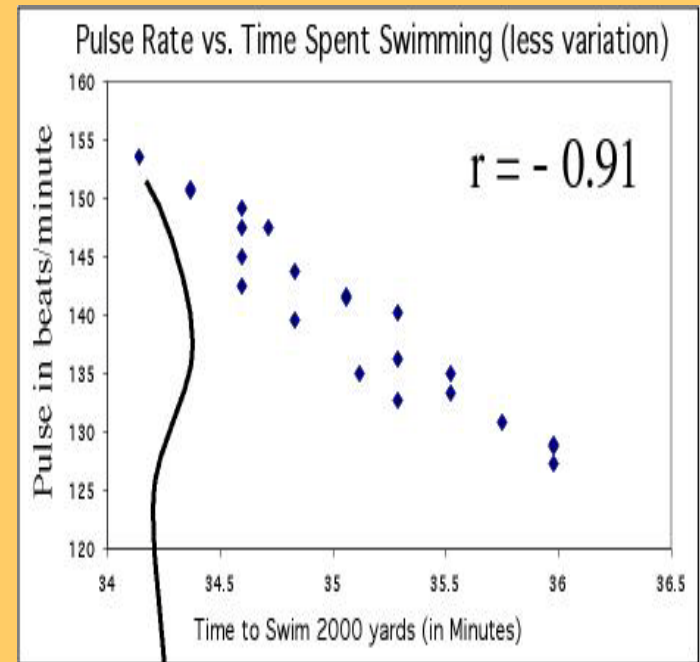


Note: You can sometimes transform a non-linear association to a linear form, for instance by taking the logarithm. You can then calculate a correlation using the transformed data.

Influential points

Correlations are calculated using means and standard deviations, and thus are NOT resistant to outliers.

Just moving one point away from the general trend here decreases the correlation from -0.91 to -0.75

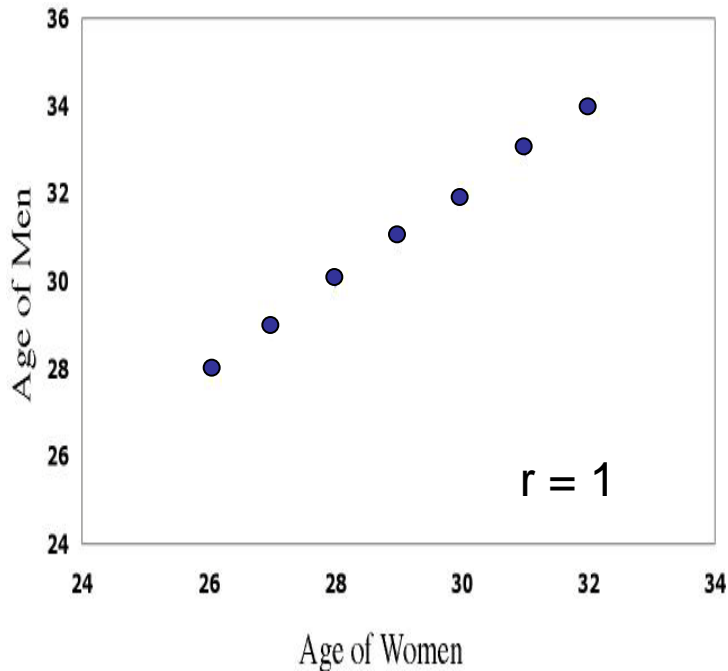
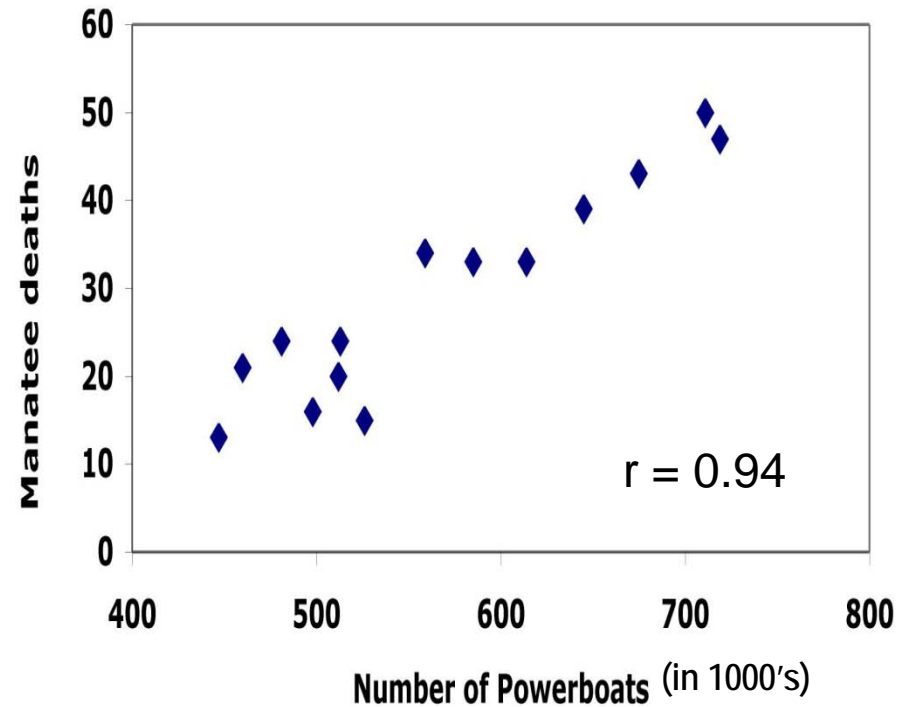


Review examples

1) What is the explanatory variable?

Describe the form, direction and strength of the relationship?

Estimate r .



2) If women always marry men 2 years older than themselves, what is the correlation of the ages between husband and wife?

$$\text{age}_{\text{man}} = \text{age}_{\text{woman}} + 2$$

equation for a straight line

Thought quiz on correlation

1. Why is there no distinction between explanatory and response variable in correlation?
2. Why do both variables have to be quantitative?
3. How does changing the units of one variable affect a correlation?
4. What is the effect of outliers on correlations?
5. Why doesn't a tight fit to a horizontal line imply a strong correlation?