Chapter 2: Exploring Data with Tables and Graphs

p. 40-74

Chapter objectives:

Both Ch.2 and Ch.3 focus on important characteristics of data including:

1. Center: Shows the middle of the data
2. Variation: Measuring the amount the data vary
3. Distribution: The nature or shape of the spread of the data
4. Outliers: Sample values that lie very far away from the vast majority of the other sample values
5. Time: Any change in the characteristics of the data over time.

Section 1: Frequency Distributions for Organizing and Summarizing Data

Objectives:

- Develop an ability to summarize data in the format of a frequency distribution and a relative frequency distribution.
- For a frequency distribution, identify values of class width, midpoint, limits, and boundaries.

Frequency distribution

- Or frequency table
- Helpful for organizing and summarizing data
- Shows how data are partitioned among several categories
- Lists categories along with the number (frequency) of data value in each of them

Grade	50-59	60-69	70-79	80-89	90-100
Frequency	1	2	4	5	2

Definitions

- **Lower class limits**: the smallest numbers that can belong to each of the different classes.

  Grades Example: 50, 60, 70, 80, 90

- **Upper class limits**: the largest numbers that can belong to each of the different classes.

  Grades Example: 59, 69, 79, 89, 100

- **Class boundaries**: the numbers used to separate the classes, but without the gaps created by class limits

  Grades Example: 59.5, 69.5, 79.5, 89.5

Definitions

- **Class midpoints**: the values in the middle of the classes.

  Grades Example: 54.5, 64.5, 74.5, 84.5, 95

- **Class width**: the differences between the two consecutive lower class limits (or boundaries).

  Grades Example: 10, 10, 10, 10

Procedure for constructing a Frequency Distribution

1. Select the number of classes

2. Calculate the class width.

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

3. Choose the value for the first lower class limit.

Procedure for constructing a Frequency Distribution

4. Using the first lower class limit and the class width, list the other lower class limits.

5. List the lower class limits in a vertical column and then determine and enter the upper class limits.

6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.

Relative frequency distribution

- A variation of the basic frequency distribution
- Each class frequency is replaced by a relative frequency (or proportion or percentage).

$$\text{Relative frequency for a class} = \frac{\text{frequency for a class}}{\text{sum of all frequencies}}$$

$$\text{Percentage for a class} = \frac{\text{frequency for a class}}{\text{sum of all frequencies}} \times 100\%$$

Cumulative frequency distribution

- A variation of a basic frequency distribution
- Frequency for each class is the sum of the frequencies for that class and all previous classes.

Example!

Grade	50-59	60-69	70-79	80-89	90-100
Frequency	1	2	4	5	2
Rel. Freq.					
Cum. Freq.					
Cum. Rel. Freq.					

Understanding the data distribution

- Frequency distributions can help!
- In statistics, we often want to know if data is “normal”

Understanding the data distribution

- Frequency distributions can help!
- In statistics, we often want to know if data is “normal”

Characteristics of Frequency Distributions for Normally Distributed Data

1. The frequencies start low, then increase to one or two high frequencies, and then decrease to a low frequency.
2. The distribution is approximately symmetric.

Example!

Do you think our grades frequency distribution is normal?

Grade	50-59	60-69	70-79	80-89	90-100
Frequency	1	2	4	5	2
Rel. Freq.	7.1%	14.3%	28.6%	35.7%	14.3%
Cum. Freq.	1	3	7	12	14
Cum. Rel. Freq.	7.1%	21.4%	50%	85.7%	100%

Think about it…

- What could it mean if we see a frequency distribution with two maximum, e.g. one in the middle and one towards the end?

- It is said that lies on tax returns typically over use the digit 6 (According to the TV show NUMB3RS). How could we use frequency distributions to test if the report is falsified?

Section 1 Homework

1, 4, 7

9, 11, 13, 17, 19, 21,

23, 25, 27

This is not to be turned in, but beneficial for your understanding.

Section 2: Histograms

Objectives:

- Develop the ability to picture the distribution of data in the format of a histogram or relative frequency histogram.
- Examine a histogram and identify common distributions

A Histogram…

- Is a graph consisting of bars of equal width drawn adjacent to each other (unless there are gaps in the data).
- The horizontal scale (x axis) represents quantitative data values
- The vertical scale (y axis) represents frequencies
- The heights of the bars correspond to frequency distributions

…

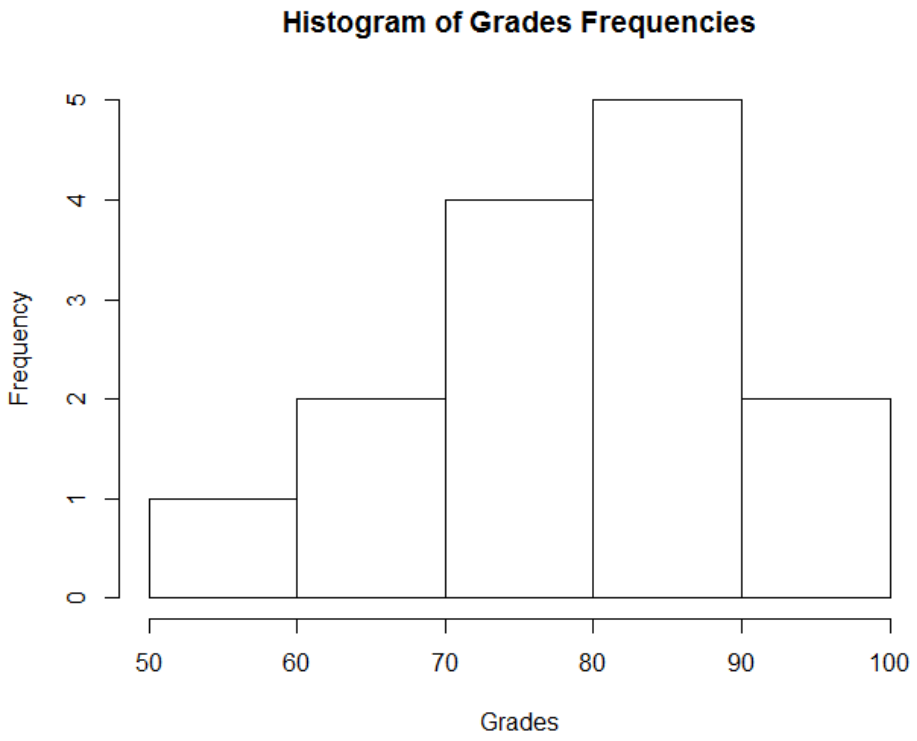
- Is essentially a graph of a frequency distribution
  - A relative frequency histogram can also be drawn

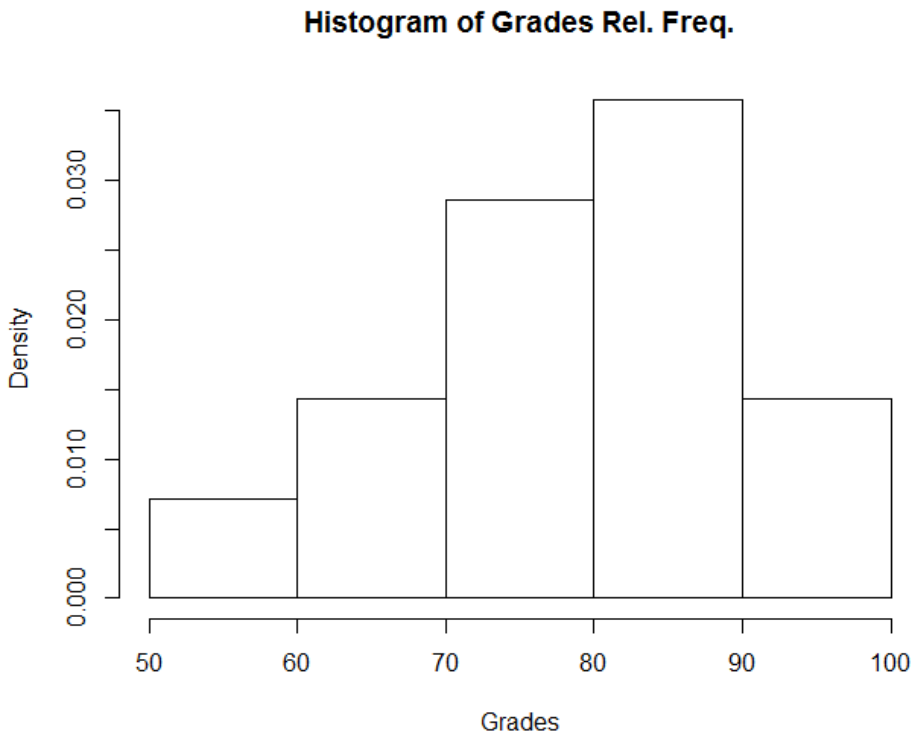
Importance of a histogram:

- Visually displays the shape of the *distribution* of the data
- Shows the location of the *center* of the data
- Shows the *spread* of the data
- Identifies *outliers*

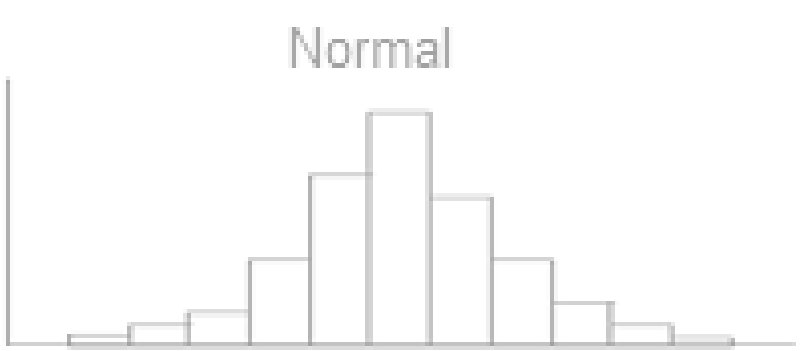
Example!

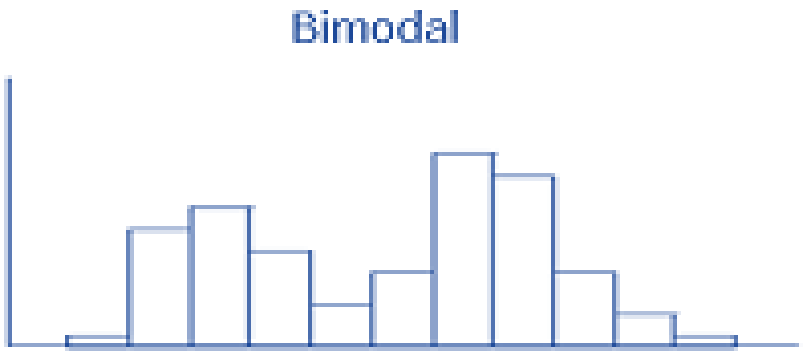
- Histogram of the grades data: {55,61,66,72,73,75,78,81,83,85,87,89,93,95}

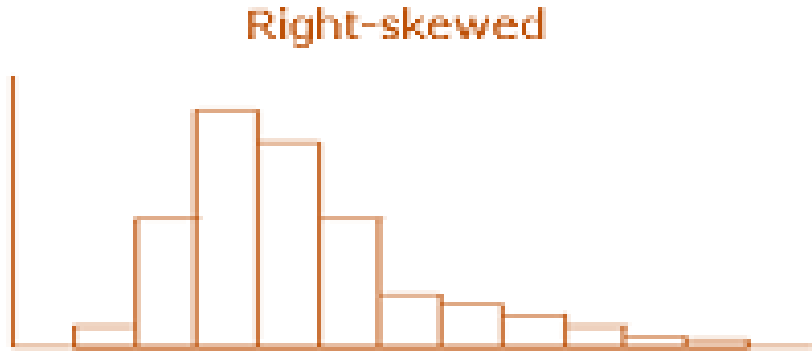


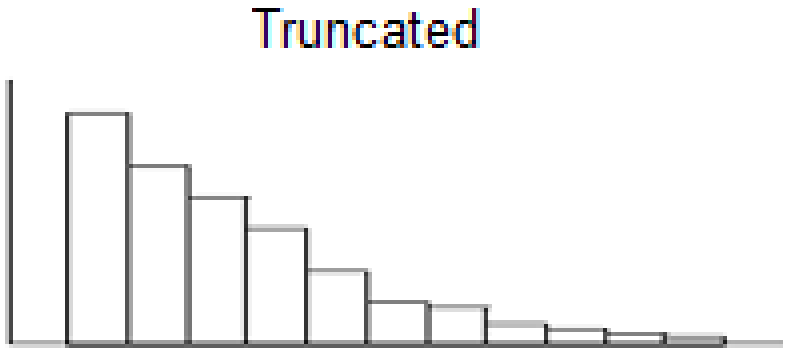


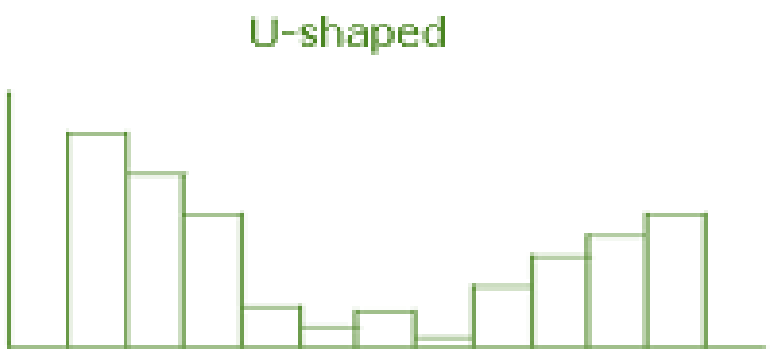
Common histogram shapes

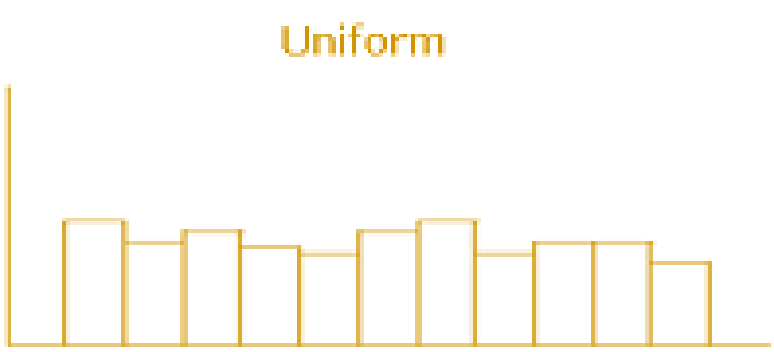






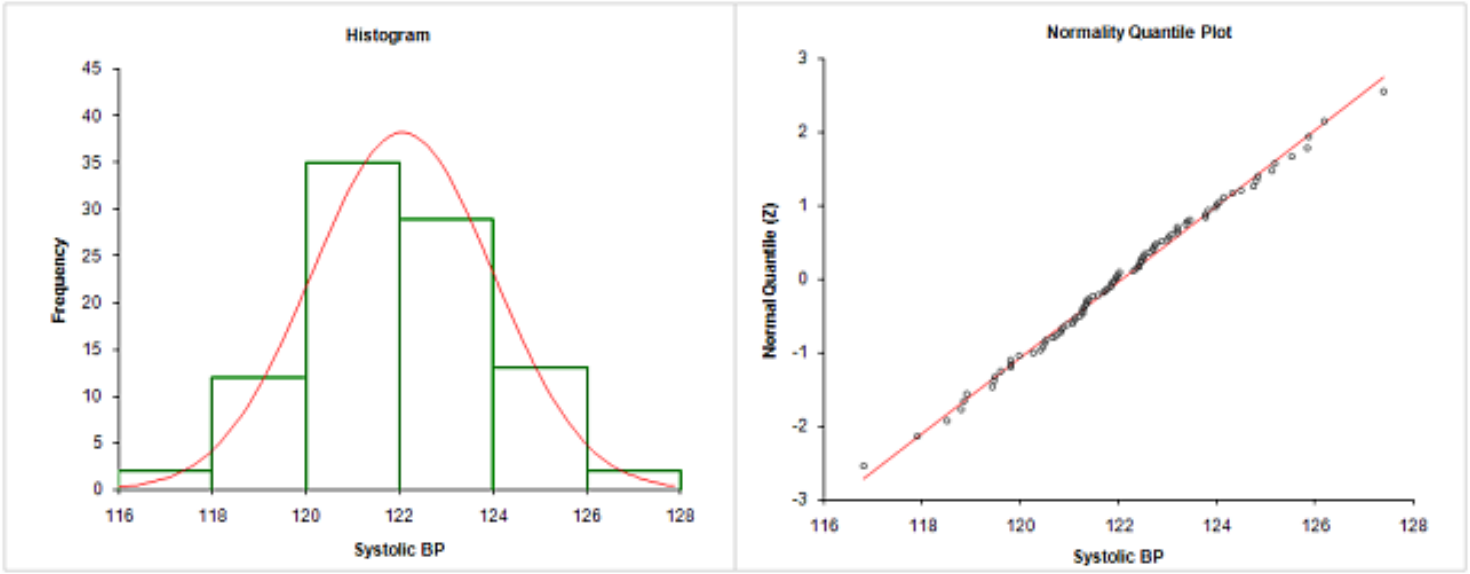






Assessing normality with Normal Quantile Plots

- Still subjective
- Better than histograms, especially for small data
- Essentially, the graph will resemble a straight line if the data is normal



Section 2 Homework

1-4

2 of 5-8

5 of 9-18

19

This is not to be turned in, but beneficial for your understanding.

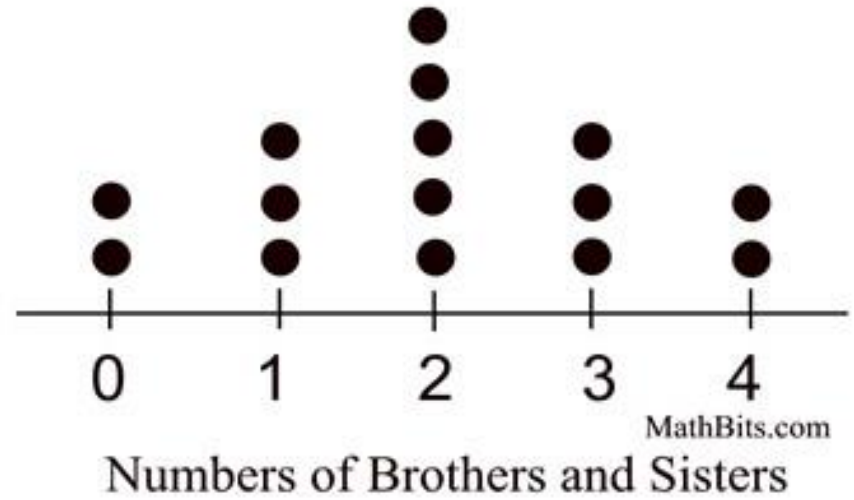
Section 3: Graphs that Enlighten and Graphs that Deceive

Objectives:

- Develop an ability to graph data using a dotplot, stemplot, time-series graph, Pareto chart, pie chart, and frequency polygon.
- Determine when a graph is deceptive through the use of a nonzero axis or a pictograph that uses an object of area or volume for one-dimensional data.

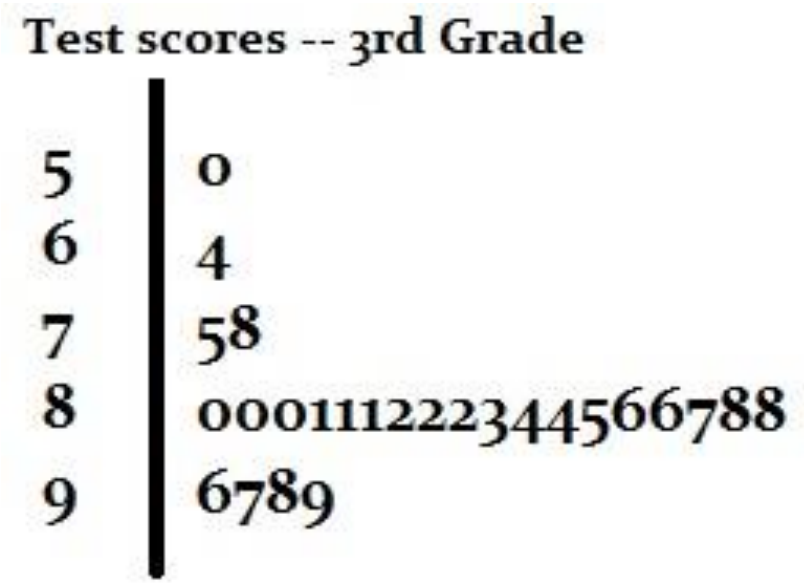
Dotplots

- A graph of quantitative data in which each data value is plotted as a point (or dot) above a horizontal axis of values.
- Dots of equal value are stacked
- Features
  - Displays the shape of a distribution
  - It is usually possible to recreate the original list of data values



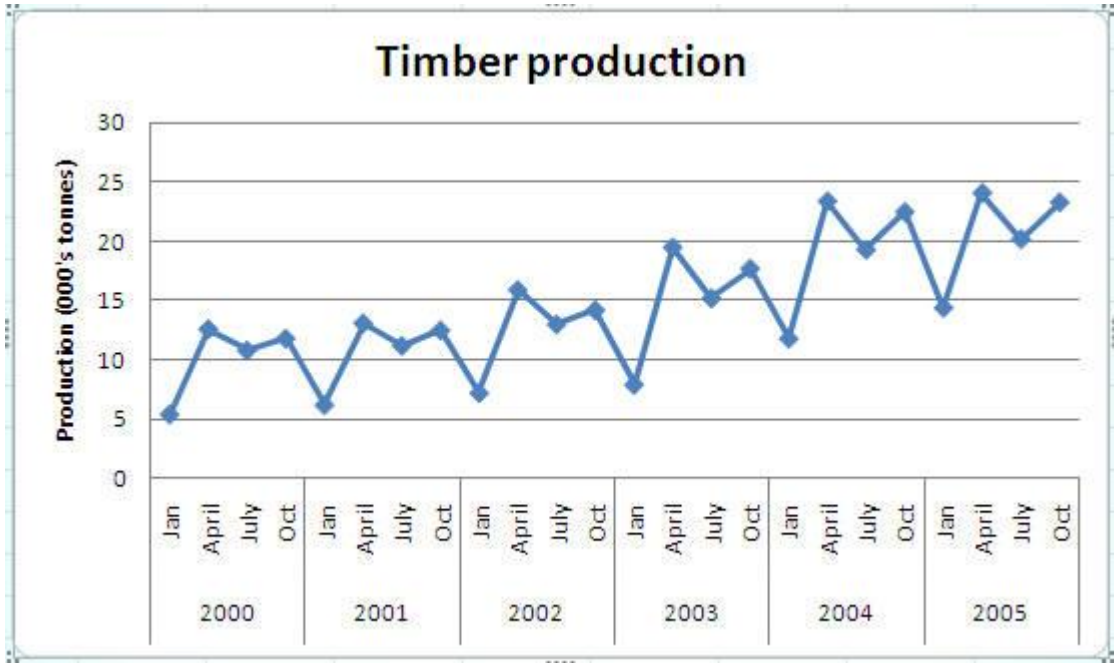
Stemplot (stem-and-leaf)

- Represents quantitative data
- Separates the data values into 2 parts:
  - The stem (the leftmost digit(s))
  - The leaf (the rightmost digit)
- Features:
  - Shows the shape of the distribution
  - Retains the original data values
  - The sample data are ordered (sorted)



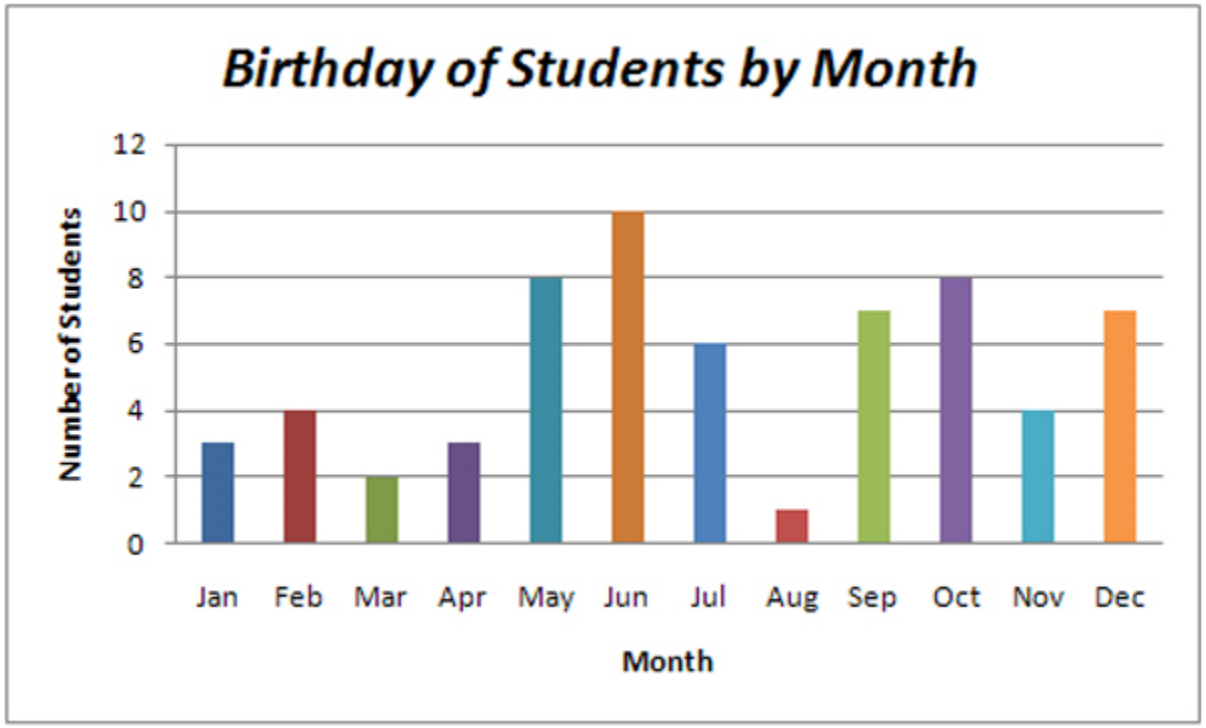
Time-series graph

- A graph of time-series data (quantitative data that have been collected at different points in time, i.e. monthly)
- Features
  - It reveals information about trends over time



Bar graph

- Uses bars of equal width to show frequencies of categories of categorical (qualitative) data.
- The bars may or may not be separated by small gaps
- Features:
  - Shows the relative distribution of categorical data so that it is easier to compare different categories.

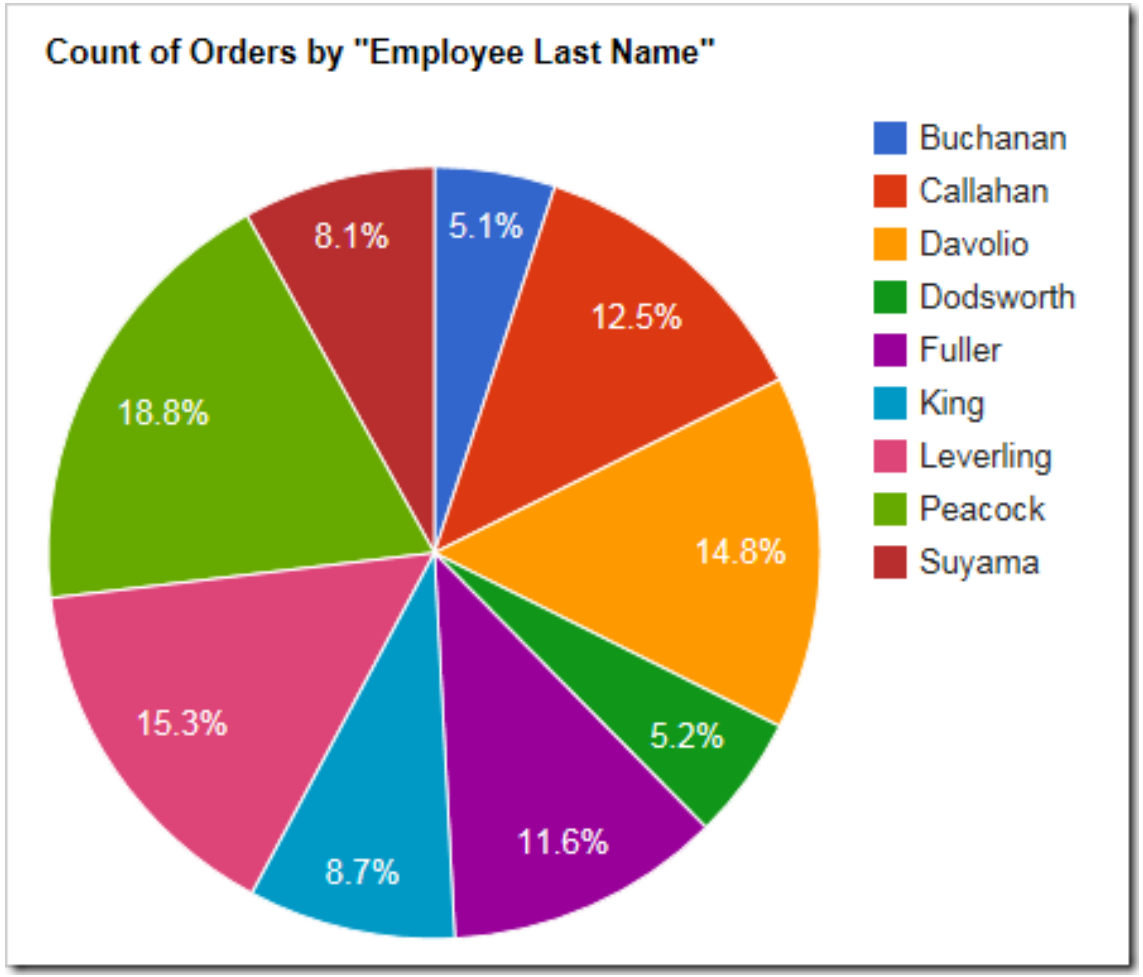


Pareto Charts

- A bar chart for categorical data where the bars are arranged in descending order by frequency.
- Features
  - Shows the relative distribution of categorical data so that it is easier to compare the different categories.
  - Draws attention to the more important categories (most/least frequent).

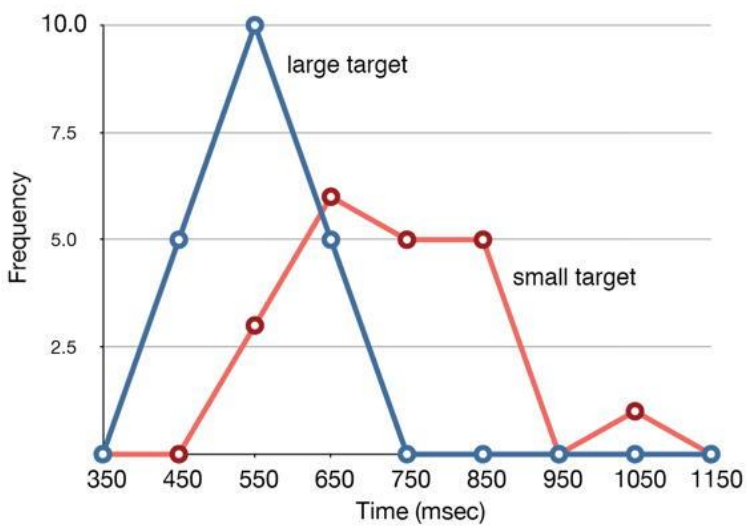
Pie chart

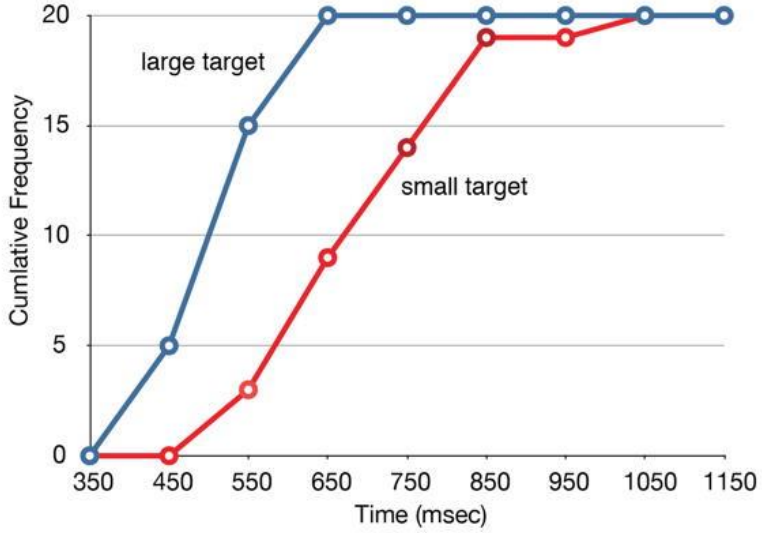
- A very common graph
- Depicts categorical data as proportional slices of a circle
- Not the most effective chart available
- Features:
  - Shows the distribution of categorical data in a commonly used format



Frequency polygon

- Uses line segments connected to points located directly above class midpoint values.
- Very similar to a histogram, with line segments instead of bars
- Relative frequency polygons are also used
- These are easier to plot together than histograms



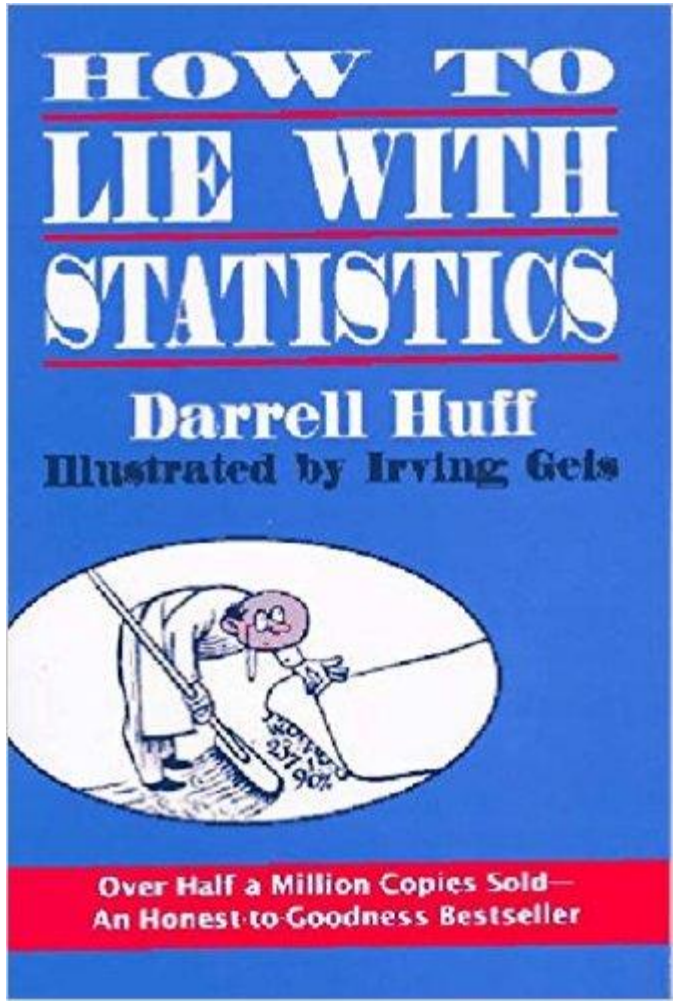


Draw a histogram, pie chart, and dotplot of the grades frequency distribution

Think: Is a Pareto chart useful here?

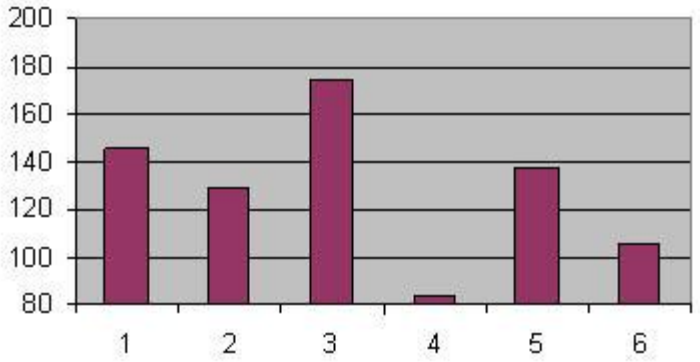
Graphs that Deceive

- Used to mislead people
- This information should help prevent you from being mislead
- There are many whom are skeptical about statistics, and they should be!

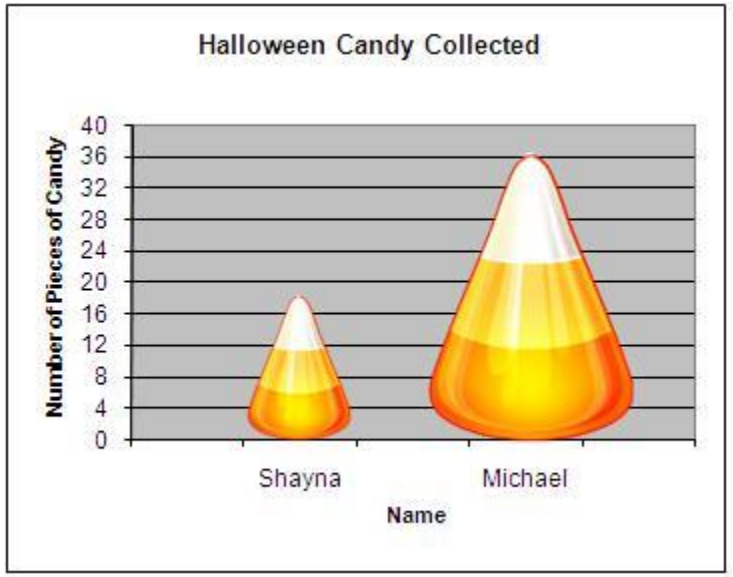


Common tools for deception

- Nonzero vertical axis
  - Typically graph that deals with frequencies
  - This exaggerated the differences between groups



- Pictographs
  - Drawings of 2 dimensional data with a 3 dimensional image.
  - Tricks of geometry create misleading images.



Concluding thoughts…

- These are the most commonly used graphs, but there are many more useful options available.
- Principals of graphics from Edward Tufte:
  - For small data sets of 20 or less, use a table instead of a graph.
  - A graph of data should make us focus on the true nature of the data
  - Do not distort the data
  - Almost all of the ink on a graph should be used for the data

Section 3 Homework

1-4

Odd of 5-15

17,18

This is not to be turned in, but beneficial for your understanding.

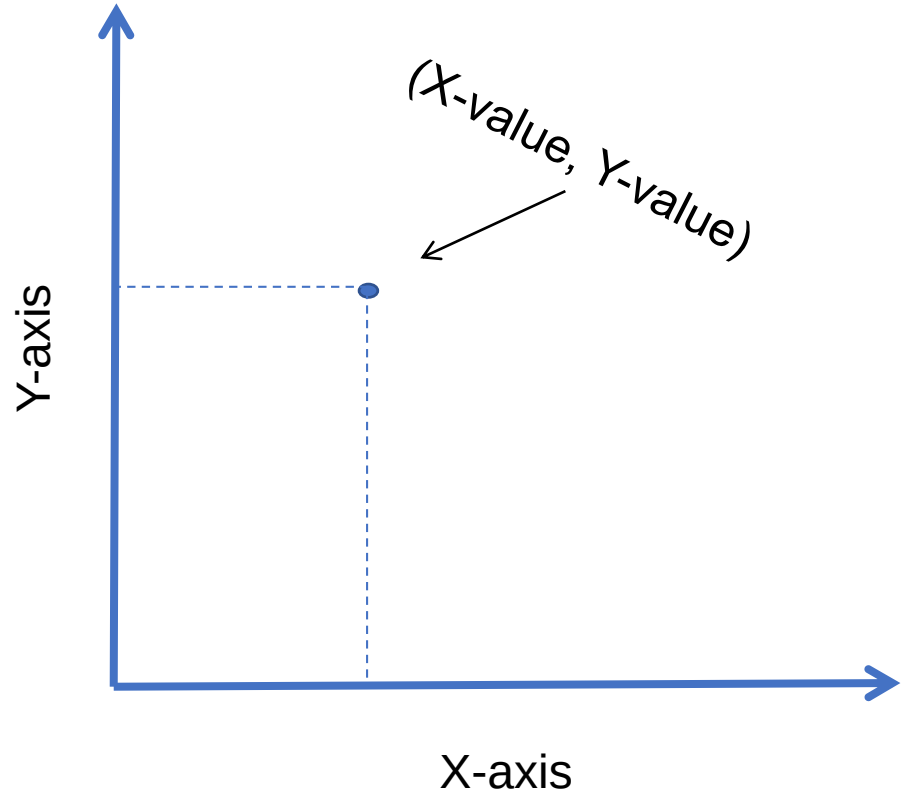
Section 4: Scatterplots, Correlation, and Regression

Objectives:

- Develop an ability to construct a scatterplot of paired data.
- Analyze a scatterplot to determine whether there appears to be a correlation between two variables.

Definitions

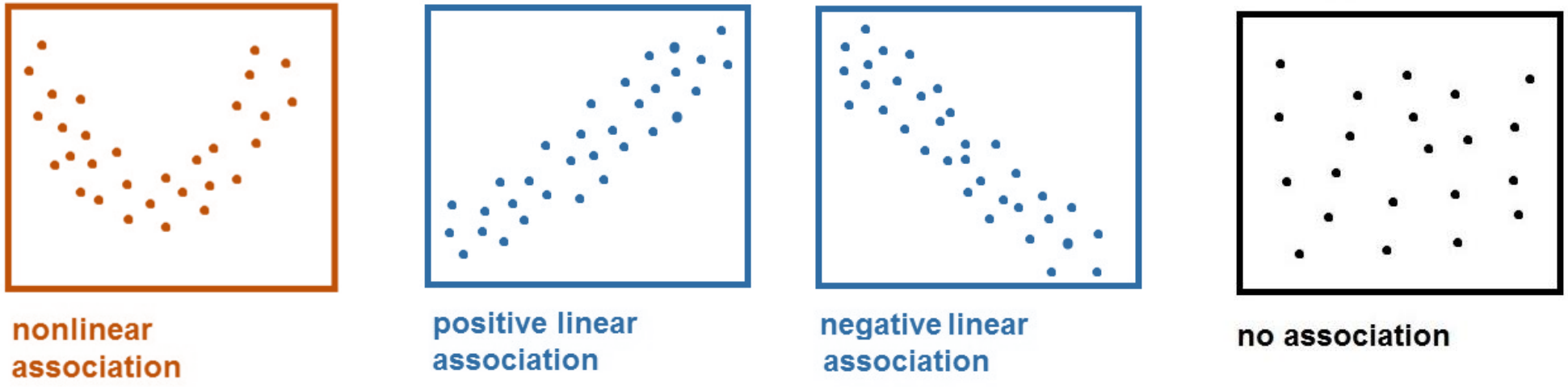
A **scatter plot** or **scatter diagram** is a plot pf paired (x, y) quantitative data with a horizontal x-axis and vertical y-axis. The horizontal axis is used for the first variable (x), and the vertical axis is used for the second variable (y).



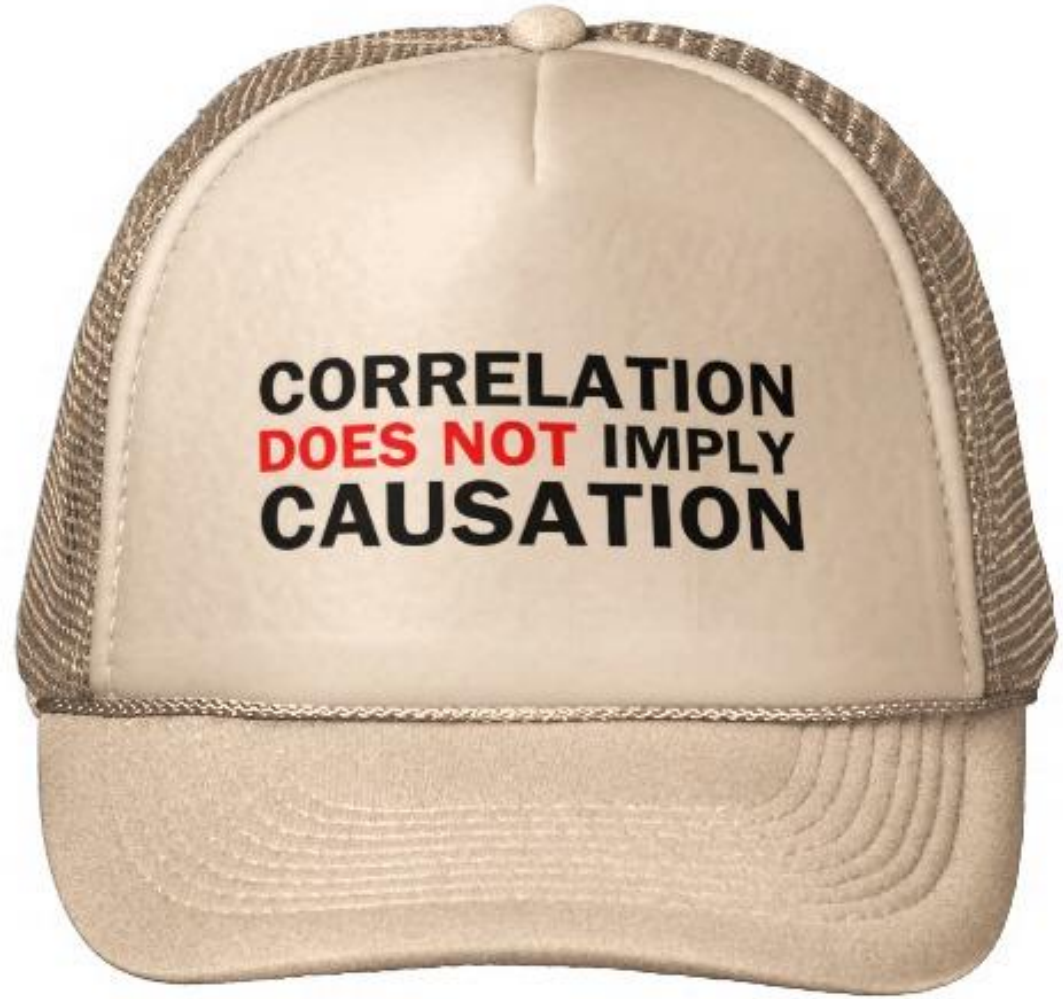
Definitions

A **correlation** exists between two variables when the values of one variable are somehow *associate* with the value of the other variable.

- A **linear correlation** exists the plotted points of paired data result in a pattern that can be approximated by a straight line.




STOP
AND
THINK
lifesroadsigns.com

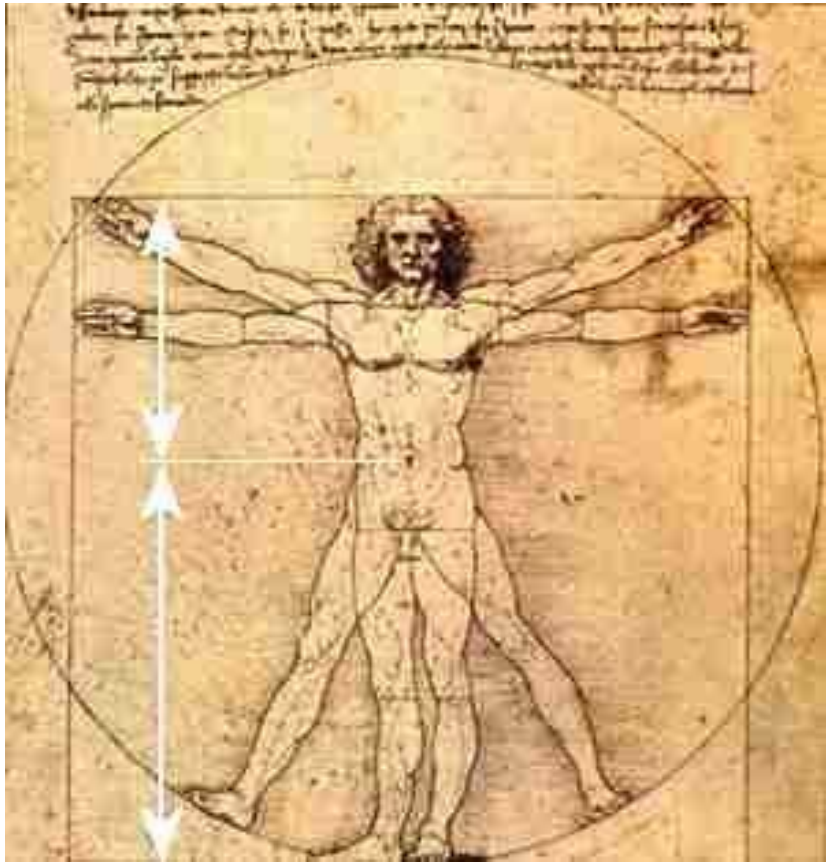

CORRELATION
DOES NOT IMPLY
CAUSATION

Example: StatCrunch

https://www.statcrunch.com/books/?book=triola_statbs2t

Use the Body data set and examine scatter plots of:

- Weight vs. height
- Height vs. systolic
- Weight vs. waist circumference
- Plots by gender



Linear correlation coefficient

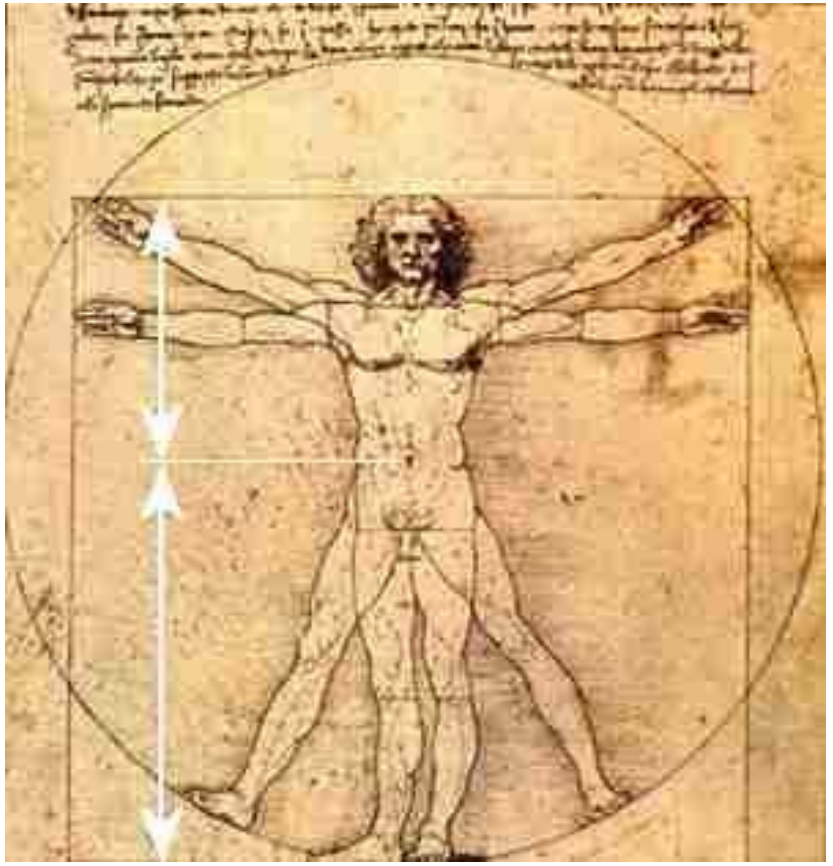
- Denoted by r or ρ
- Measures strength of association of the linear association between two variables
- $-1 < r < 1$
- Can be computed manually or via technology

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

Example: StatCrunch

https://www.statcrunch.com/books/?book=triola_statbs2t

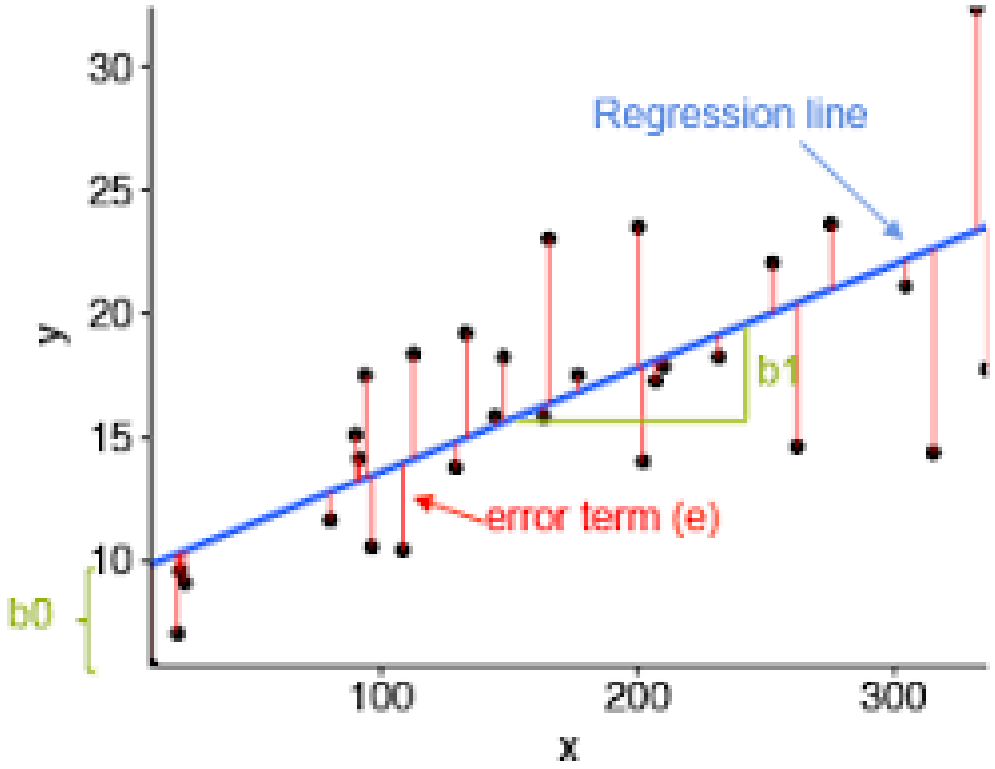
Use the Body data set and examine correlations between different variables.



Linear Regression

- Forms a straight line to describe the relationship between two variables that are correlated.
- This **regression line** is also called the line of best fit.
- Has the following equation:

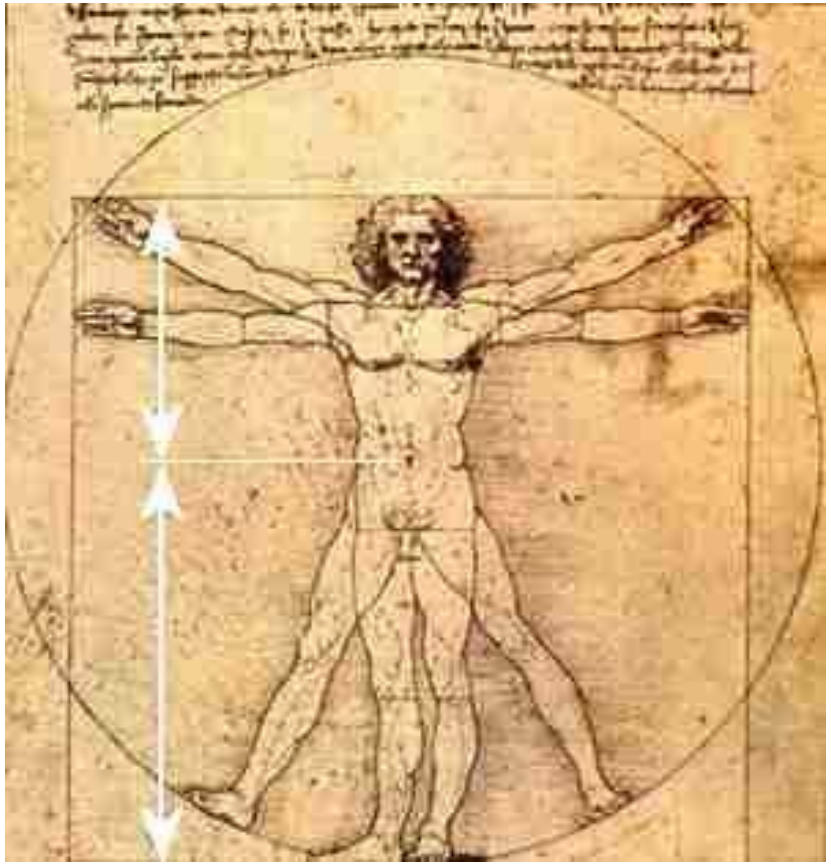
$$\hat{y} = b_0 + b_1 x$$



Example: StatCrunch

https://www.statcrunch.com/books/?book=triola_statbs2t

Use the Body data set and examine linear regression for weight vs. height and waist vs. BMI.



Section 3 Homework

1-3, 5, 9, 10

This is not to be turned in, but beneficial for your understanding.