Statistical Concepts

Data Exploration:

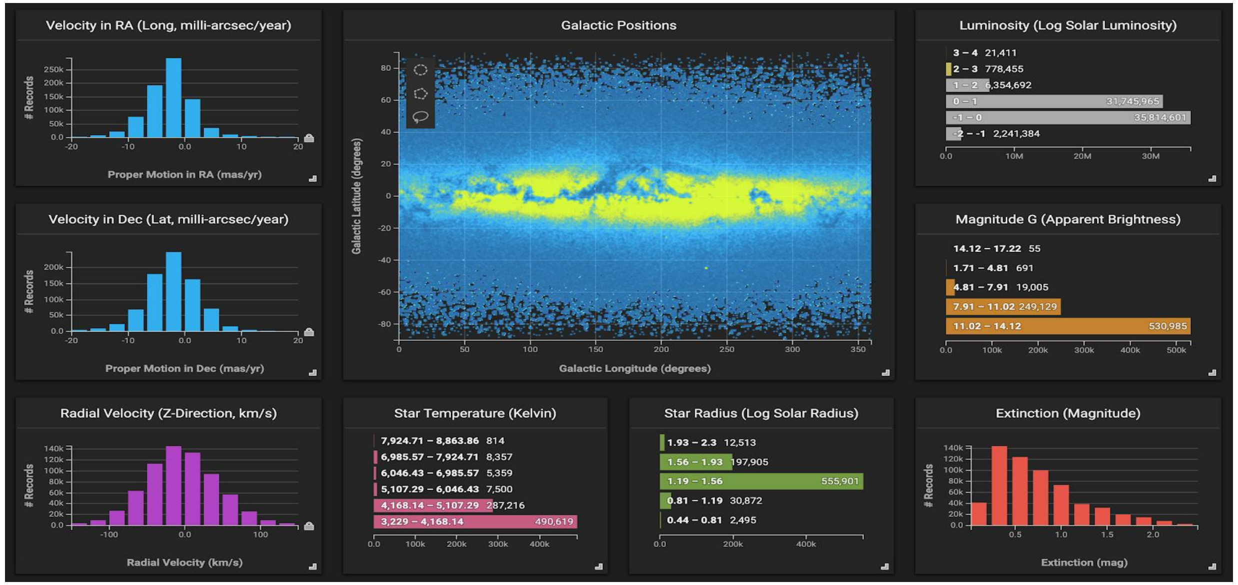
- Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

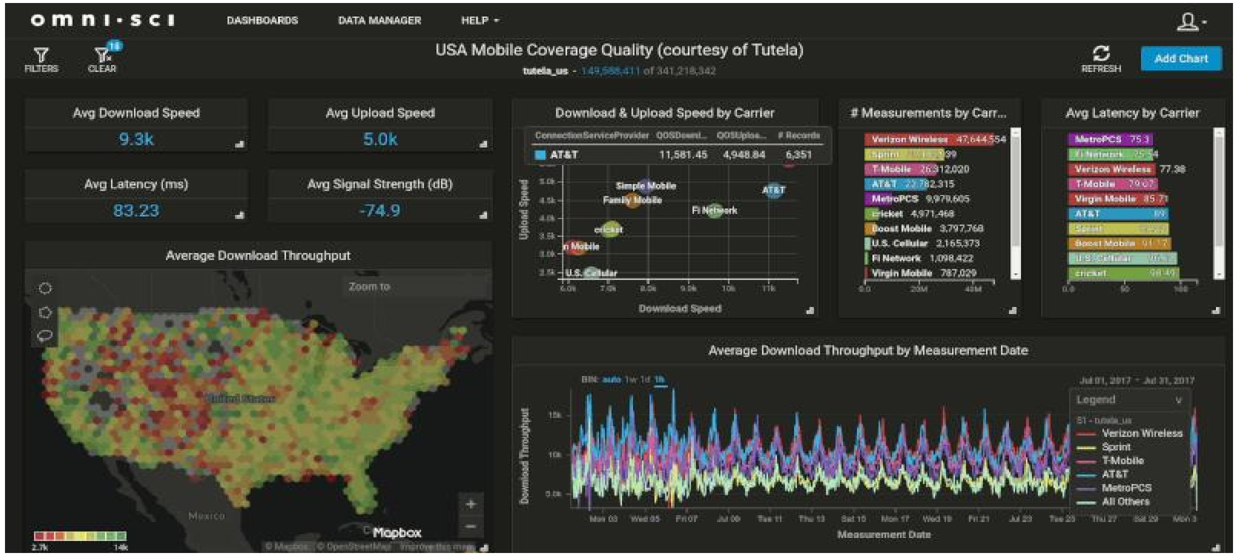
- Raw data is typically reviewed with a combination of manual workflows and automated data-exploration techniques to visually explore data sets, look for similarities, patterns and outliers and to identify the relationships between different variables.

- This is also sometimes referred to as exploratory data analysis, which is a statistical technique employed to analyze raw data sets in search of their broad characteristics.

Why is Data Exploration Important?

- ➢ Starting with data exploration helps users to make better decisions on where to dig deeper into the data and to take a broad understanding of the business when asking more detailed questions later.
- ➢ Performing the initial step of data exploration enables data analysts to better understand and visually identify anomalies and relationships that might otherwise go undetected.
- ➢ Data exploration tools include data visualization software and business intelligence platforms, such as Microsoft Power BI, Qlik and Tableau.

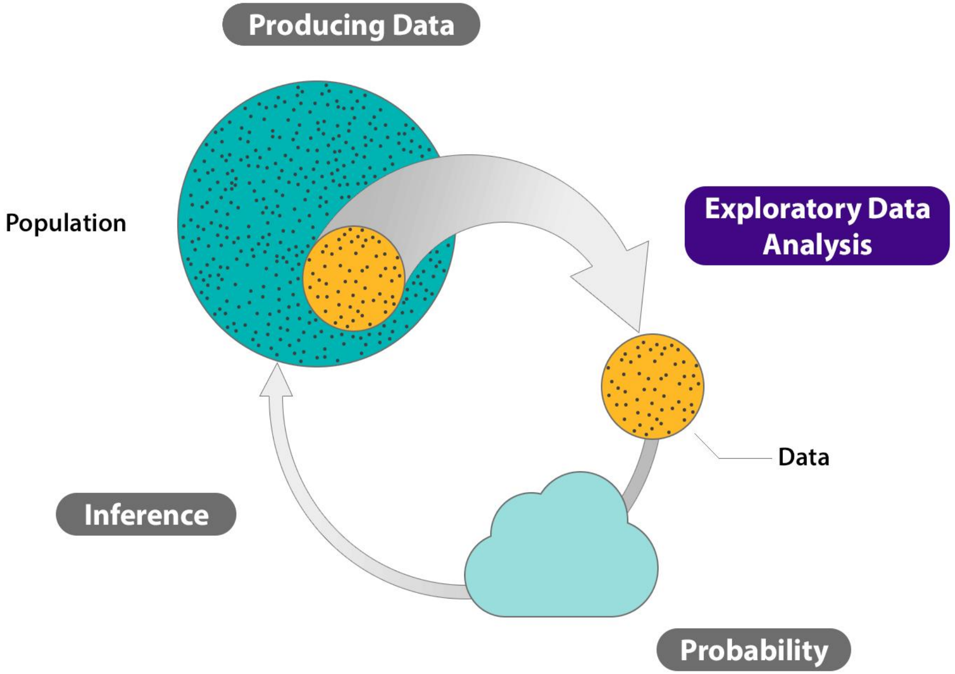




Exploratory Data Analysis

Raw data are not very informative. ***Exploratory Data Analysis (EDA)*** is how we make sense of the data by converting them from their raw form to a more informative one.

- **In particular, EDA consists of:**
  - organizing and summarizing the raw data,
  - discovering important features and patterns in the data and any striking deviations from those patterns, and then
  - interpreting our findings in the context of the problem
- **And can be useful for:**
  - describing the distribution of a single variable (center, spread, shape, outliers)
  - checking data (for errors or other problems)
  - checking assumptions to more complex statistical analyses
  - investigating relationships between variables



Important Features of Exploratory Data Analysis

There are two important features to the structure of the EDA,

- **Examining Distributions —** exploring data one variable at a time (***univariate)***.
- **Examining Relationships —** exploring data two variables at a time (***bivariate)***.

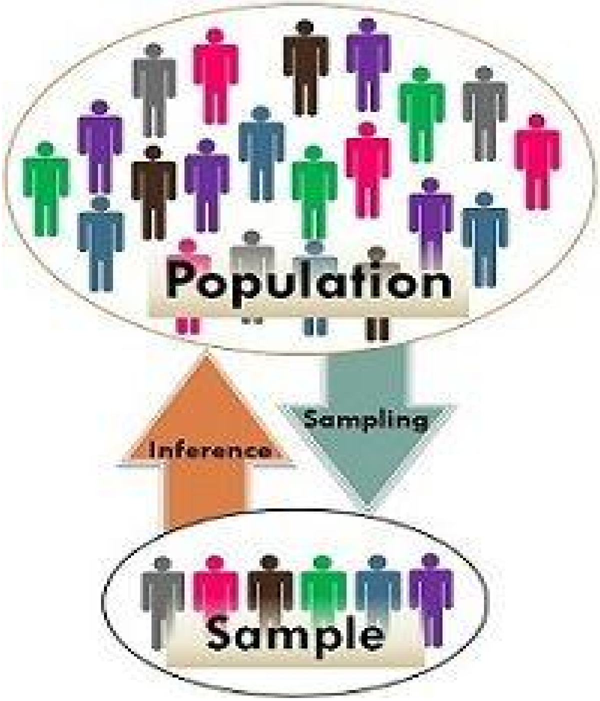
- In Exploratory Data Analysis, our exploration of data will always consist of the following two elements:
  - **visual displays**, supplemented by **numerical measures**.

Populations and Observations/Samples

Normally, when an experiment involving random variables is performed, data are generated. These data points are often measurements of some kind.

- In statistics, all instances of a random variable as called **observations** of that variable.
- Furthermore, the set of data points actually collected is termed a **sample** of a given **population** of observations.

To generalize: the population consists of the entire set of ***all possible observations***, while the sample is a subset of the population.



Definition of Population

The different types of population are discussed as under:

- ***Finite Population:*** When the number of elements of the population is fixed and thus making it possible to enumerate it in totality, the population is said to be finite.
- ***Infinite Population***: When the number of units in a population are uncountable, and so it is impossible to observe all the items of the universe, then the population is considered as infinite.
- ***Existent Population:*** The population which comprises of objects that exist in reality is called existent population.
- ***Hypothetical Population:*** Hypothetical or imaginary population is the population which exists hypothetically.

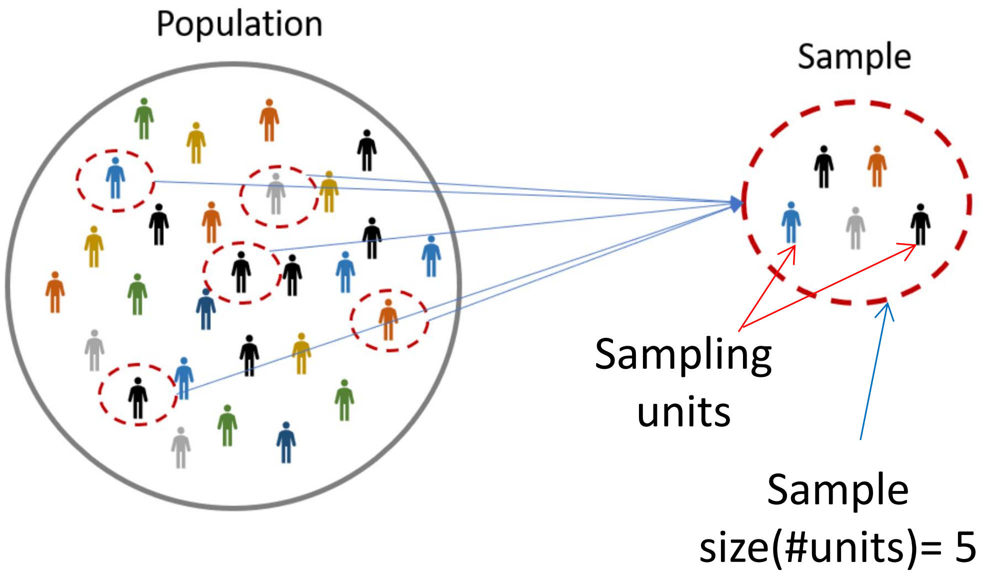
Examples

- *The population of all workers working in the sugar factory.*
- *The population of motorcycles produced by a particular company.*
- *The population of mosquitoes in a town.*
- *The population of tax payers in India.*

Definition of Sample

- A part of population chosen at random for participation in the study.
- The sample so selected should be such that it represent the population in all its characteristics, and it should be free from bias, so as to produce miniature cross-section, as the sample observations are used to make generalisations about the population.

- In other words, the respondents selected out of population constitutes a '**sample**'.
- The process of selecting respondents is known as '**sampling**.'
- The units under study are called **sampling units**.
- The number of units in a sample is called **sample size**.

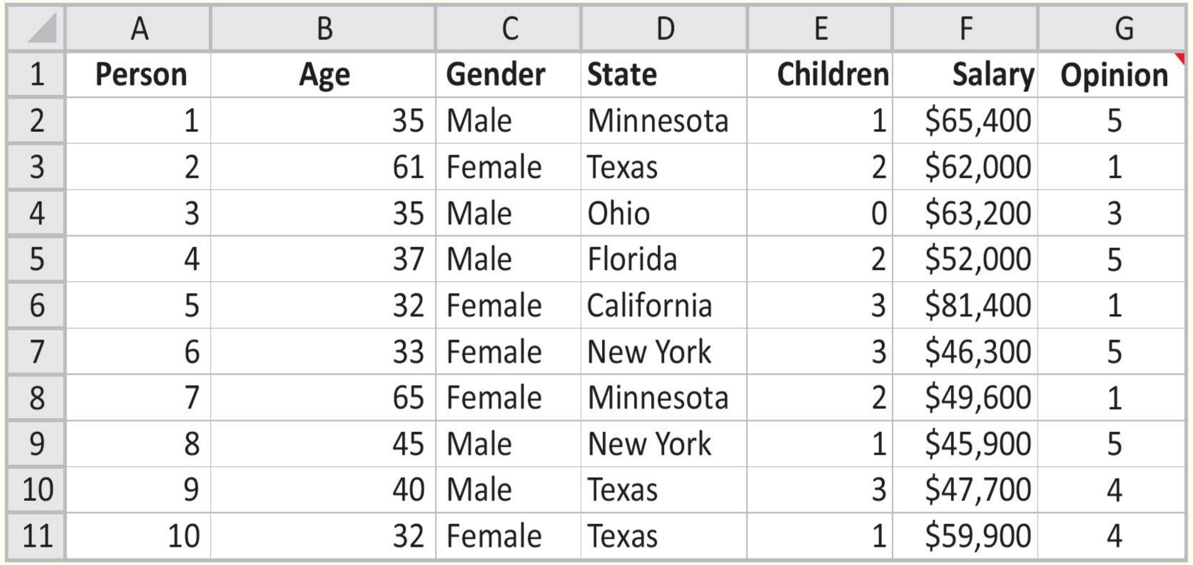


Types of Sampling

Refer sampling slides separatly.....

Data Sets, Variables, and Observations

- A **data set** is usually a rectangular array of data, with variables in columns and observations in rows.
- A **variable** (or **field** or **attribute**) is a characteristic of members of a population, such as height, gender, or salary.
- An **observation** (or **case** or **record**) is a list of all variable values for a single member of a population.



	A	B	C	D	E	F	G
1	**Person**	**Age**	**Gender**	**State**	**Children**	**Salary**	**Opinion**
2	1	35	Male	Minnesota	1	$65,400	5
3	2	61	Female	Texas	2	$62,000	1
4	3	35	Male	Ohio	0	$63,200	3
5	4	37	Male	Florida	2	$52,000	5
6	5	32	Female	California	3	$81,400	1
7	6	33	Female	New York	3	$46,300	5
8	7	65	Female	Minnesota	2	$49,600	1
9	8	45	Male	New York	1	$45,900	5
10	9	40	Male	Texas	3	$47,700	4
11	10	32	Female	Texas	1	$59,900	4

- □ Data set includes observations on 10 people who responded to a questionnaire on the president’s environmental policies.
- □ Variables include age, gender, state, children, salary, and opinion.
  - ❑ Include a row that lists variable names.
  - ❑ Include a column that shows an index of the observation.

Data Types

- **Variables** (or **attributes, dimensions, features**)

A variable is a characteristic that can be measured and that can assume different values. (Height, age, income, province or country of birth, grades obtained at college and type of housing are all examples of variables.)

- **Types of variables**

Variables may be classified into two main categories:

  - **Categorical (Qualitative)**

    (A **categorical** variable (called qualitative variable) refers to a characteristic that can't be quantifiable.)

  - **Numeric (Quantitative)**

    (A variable is **numeric** if meaningful arithmetic can be performed on it.)

A variable is numerical if meaningful arithmetic can be performed on it. Otherwise, the variable is categorical.

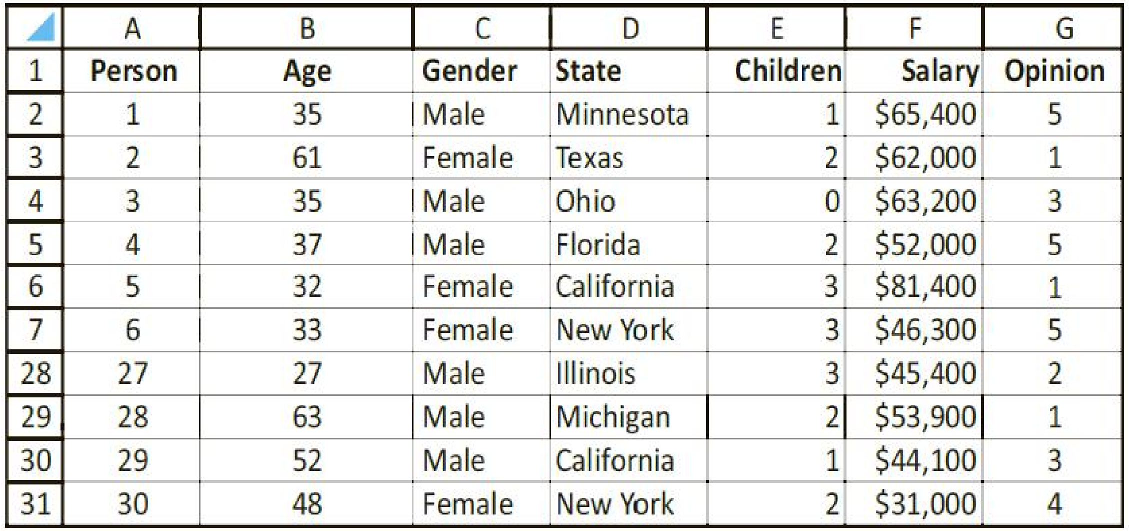
Categorical (Qualitative) Variable/Attribute

- **Nominal:** Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things for example,
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- The values do not have any meaningful order about them.

- **Binary:** Nominal attribute with only 2 states (0 and 1), where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as Boolean if the two states correspond to true and false.
  - ***Symmetric binary:*** both outcomes equally important, e.g., gender
  - ***Asymmetric binary:*** outcomes not equally important, e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

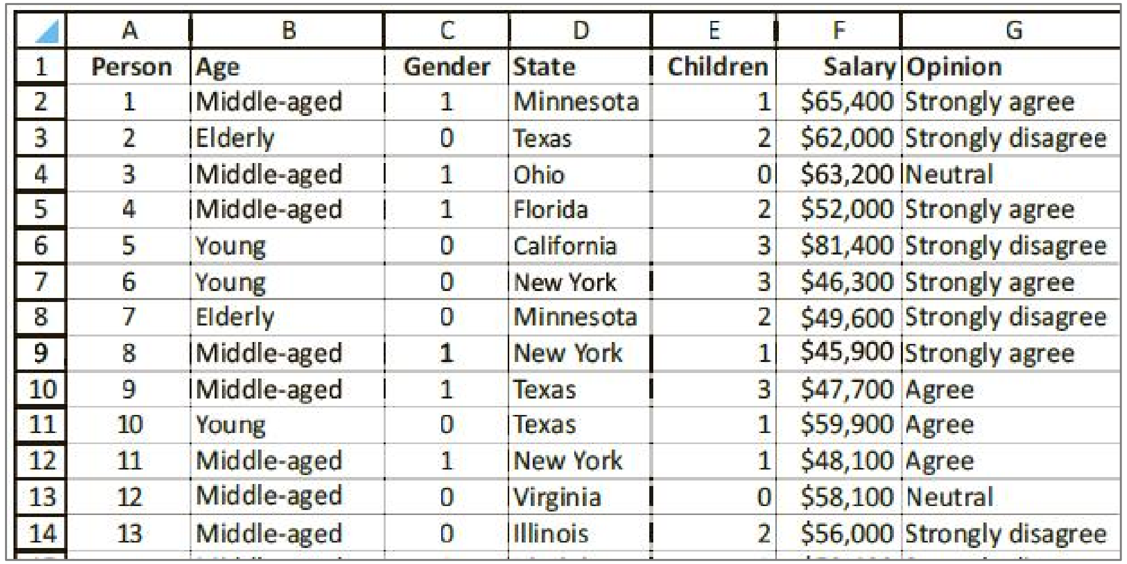
- **Ordinal:** Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}*,* grades, army rankings
  - Other examples of ordinal attributes include Grade (e.g., A+, A, A−, B+, and so on) and
  - Professional rank. Professional ranks can be enumerated in a sequential order, such as assistant, associate, and full for professors,

Categorical (Qualitative) Variable/Attribute (cont..)



	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	35	Male	Minnesota	1	$65,400	5
3	2	61	Female	Texas	2	$62,000	1
4	3	35	Male	Ohio	0	$63,200	3
5	4	37	Male	Florida	2	$52,000	5
6	5	32	Female	California	3	$81,400	1
7	6	33	Female	New York	3	$46,300	5
28	27	27	Male	Illinois	3	$45,400	2
29	28	63	Male	Michigan	2	$53,900	1
30	29	52	Male	California	1	$44,100	3
31	30	48	Female	New York	2	$31,000	4

- Categorical variables can be coded numerically.
- Gender has not been coded, whereas Opinion has been coded.
- This is largely a matter of taste-*coding a categorical variable does not make it numerical and appropriate for arithmetic operations.*

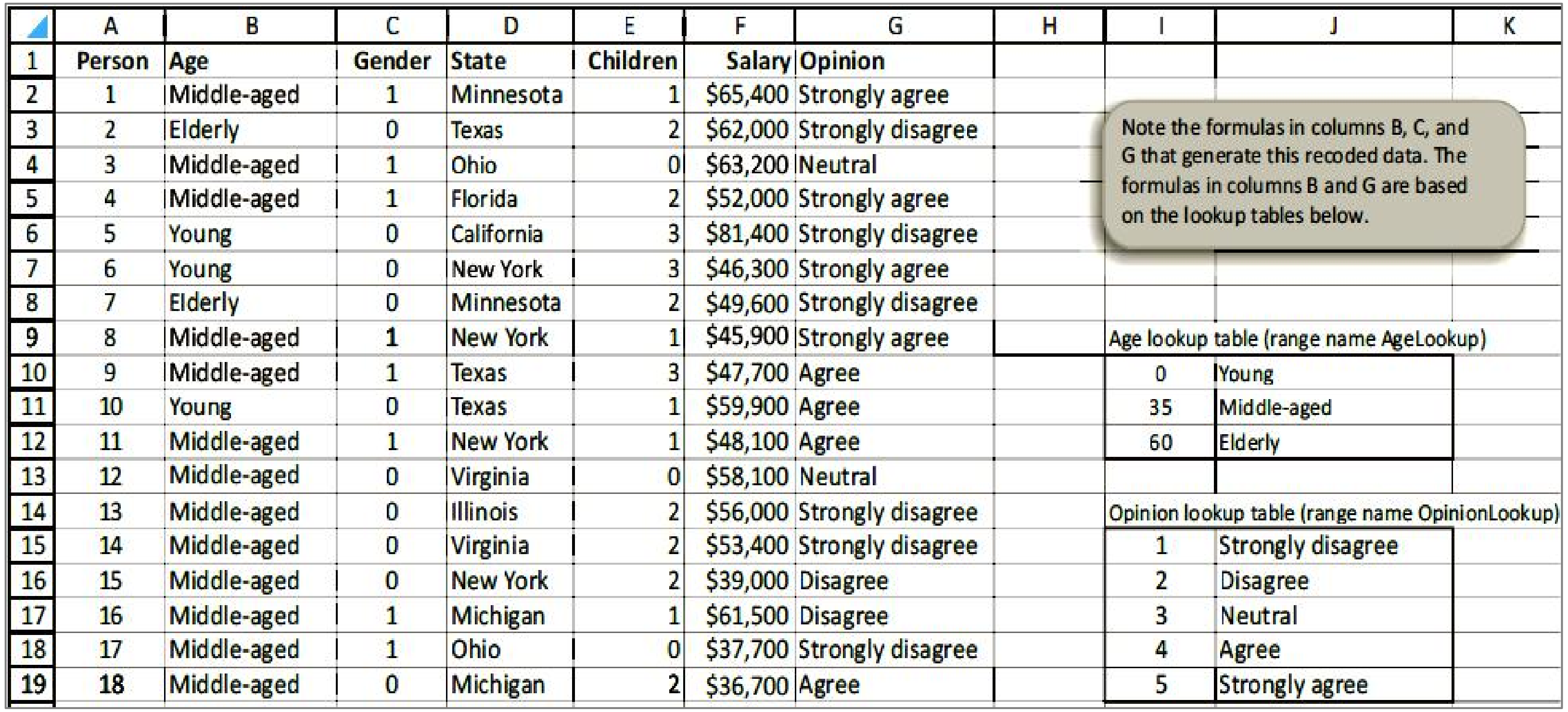


	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	Middle-aged	1	Minnesota	1	$65,400	Strongly agree
3	2	Elderly	0	Texas	2	$62,000	Strongly disagree
4	3	Middle-aged	1	Ohio	0	$63,200	Neutral
5	4	Middle-aged	1	Florida	2	$52,000	Strongly agree
6	5	Young	0	California	3	$81,400	Strongly disagree
7	6	Young	0	New York	3	$46,300	Strongly agree
8	7	Elderly	0	Minnesota	2	$49,600	Strongly disagree
9	8	Middle-aged	1	New York	1	$45,900	Strongly agree
10	9	Middle-aged	1	Texas	3	$47,700	Agree
11	10	Young	0	Texas	1	$59,900	Agree
12	11	Middle-aged	1	New York	1	$48,100	Agree
13	12	Middle-aged	0	Virginia	0	$58,100	Neutral
14	13	Middle-aged	0	Illinois	2	$56,000	Strongly disagree

- Now Opinion has been replaced by text, and Gender has been coded as 1 for males and 0 for females.
- This 0−1 coding for a categorical variable is very common. Such a variable is called a **dummy variable,** it often simplies the analysis.

A dummy variable is a 0−1 coded variable for a specific category. It is coded as 1 for all observations in that category and 0 for all observations not in that category.

Categorical (Qualitative) Variable/Attribute (cont..)



	A	B	C	D	E	F	G	H	I	J	K
1	Person	Age	Gender	State	Children	Salary	Opinion				
2	1	Middle-aged	1	Minnesota	1	$65,400	Strongly agree				
3	2	Elderly	0	Texas	2	$62,000	Strongly disagree				
4	3	Middle-aged	1	Ohio	0	$63,200	Neutral				
5	4	Middle-aged	1	Florida	2	$52,000	Strongly agree				
6	5	Young	0	California	3	$81,400	Strongly disagree				
7	6	Young	0	New York	3	$46,300	Strongly agree				
8	7	Elderly	0	Minnesota	2	$49,600	Strongly disagree				
9	8	Middle-aged	1	New York	1	$45,900	Strongly agree		Age lookup table (range name AgeLookup)		
10	9	Middle-aged	1	Texas	3	$47,700	Agree		0	Young	
11	10	Young	0	Texas	1	$59,900	Agree		35	Middle-aged	
12	11	Middle-aged	1	New York	1	$48,100	Agree		60	Elderly	
13	12	Middle-aged	0	Virginia	0	$58,100	Neutral				
14	13	Middle-aged	0	Illinois	2	$56,000	Strongly disagree		Opinion lookup table (range name OpinionLookup)		
15	14	Middle-aged	0	Virginia	2	$53,400	Strongly disagree		1	Strongly disagree	
16	15	Middle-aged	0	New York	2	$39,000	Disagree		2	Disagree	
17	16	Middle-aged	1	Michigan	1	$61,500	Disagree		3	Neutral	
18	17	Middle-aged	1	Ohio	0	$37,700	Strongly disagree		4	Agree	
19	18	Middle-aged	0	Michigan	2	$36,700	Agree		5	Strongly agree	

Note the formulas in columns B, C, and G that generate this recoded data. The formulas in columns B and G are based on the lookup tables below.

- A **binned** (or **discretized**) **variable** *corresponds to a numerical variable that has been categorized into discrete categories.*
- These categories are usually called **bins**.
- The Age variable has been categorized as “young” (34 years or younger),
- “middle-aged” (from 35 to 59 years), and “elderly” (60 years or older).

Numerical (Quantitative) Variable/Attribute

- Numerical variables can be classified as **discrete** or **continuous**.
- The basic distinction is whether the data arise from counts or continuous measurements.
  - *The variable Children is clearly a count (discrete),*
  - *whereas the variable Salary is best treated as continuous.*
- Numeric attributes can be ***interval-scaled*** or ***ratio-scaled***.
  - ***Interval***

    Measured on a scale of **equal-sized units.**
    The values of interval-scaled attributes have order and can be positive, 0, or negative.
    E.g., *temperature in C˚or F˚, calendar dates*
    No true zero-point
  - ***Ratio***

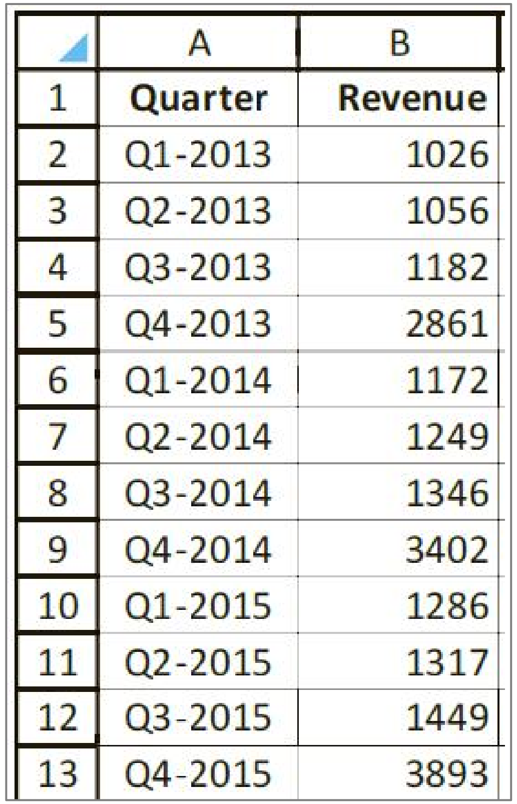
    Inherent **zero-point.** We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    e.g., *temperature in Kelvin, length, counts, monetary quantities*

Data sets can also be categorized as **cross-sectional** or **time series**

- ***Cross-sectional*** data are data on a cross section of a population at a distinct point in time.
  - ➢ *The opinion data set is cross-sectional.*
- ***Time series*** data are data collected over time.
  - ➢ *a time series data set tracks one or more variables through time.*

A time series data set generally has the same layout—variables in columns and observations in rows—but now each variable is a time series. Also, one of the columns usually indicates the time period.

It has quarterly observations on revenues from toy sales over a four-year period in column B, with the time periods listed chronologically in column A.



	A	B
1	**Quarter**	**Revenue**
2	Q1-2013	1026
3	Q2-2013	1056
4	Q3-2013	1182
5	Q4-2013	2861
6	Q1-2014	1172
7	Q2-2014	1249
8	Q3-2014	1346
9	Q4-2014	3402
10	Q1-2015	1286
11	Q2-2015	1317
12	Q3-2015	1449
13	Q4-2015	3893

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:

• **Distinctness :**	= and ≠
• **Order :**	<, ≤, >, and ≥
• **Addition :** (Differences are meaningful)	+ and -
• **Multiplication :** (Ratios are meaningful)	* and /

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & meaningful differences
- Ratio attribute: all 4 properties/operations

~~Descriptive measures for categorical variables~~

Descriptive statistics are the first pieces of information used to understand and represent a dataset.
There goal, in essence, is to describe the main features of numerical and categorical information with simple summaries.

Initially interested in understanding for categorical data, to include:

Frequencies: The number of observations for a particular category

Proportions: The percent that each category accounts for out of the whole

Marginals: The totals in a cross tabulation by row or column

Visualization: We should understand these features of the data through statistics and visualization

Descriptive measures for categorical variables

Frequencies

- To produce ***contingency tables*** which calculate counts for each combination of categorical variables
- For instance, we may want to get the total count of female and male customers.

Female	Male
7170	6889

- If we want to understand the number of married and single females and male customers we can produce a cross classification table:

	Female	Male
Married	3602	3264
Single	3568	3625

- We can also produce multidimensional tables based on three or more categorical variables. In this case we assess the count of customers by marital status, gender, and location:

		Place-1	Place-2	Place-3
Married	Female	190	638	188
	Male	197	692	210
Single	Female	183	686	175
	Male	239	717	242

Descriptive measures for categorical variables

Proportions

- Contingency tables that present the proportions (percentages) of each category or combination of categories.
- The following reproduces the previous tables but calculates the proportions rather than counts:

Female	Male
0.5099936	0.4900064

- Percentages for gender by marital status

	Female	Male
Married	0.2562060	0.2321644
Single	0.2537876	0.2578420

- Customer percentages across location by gender and marital status

		Place-1	Place-2	Place-3
Married	Female	0.014	0.045	0.013
	Male	0.014	0.049	0.015
Single	Female	0.013	0.049	0.012
	Male	0.017	0.051	0.017

Descriptive measures for categorical variables

Marginals

- Marginals show the total counts or percentages across columns or rows in a contingency table.

	Female	Male
Married	3602	3264
Single	3568	3625

- We can compute the marginal frequencies and the percentages for these marginal frequencies

Married	Single
6866	7193

ROW

Female	Male
7170	6889

COLUMN

FREQUENCY MARGINALS

	Female	Male
Married	0.5246140	0.4753860
Single	0.4960378	0.5039622

ROW

	Female	Male
Married	0.5023710	0.4737988
Single	0.4976290	0.5262012

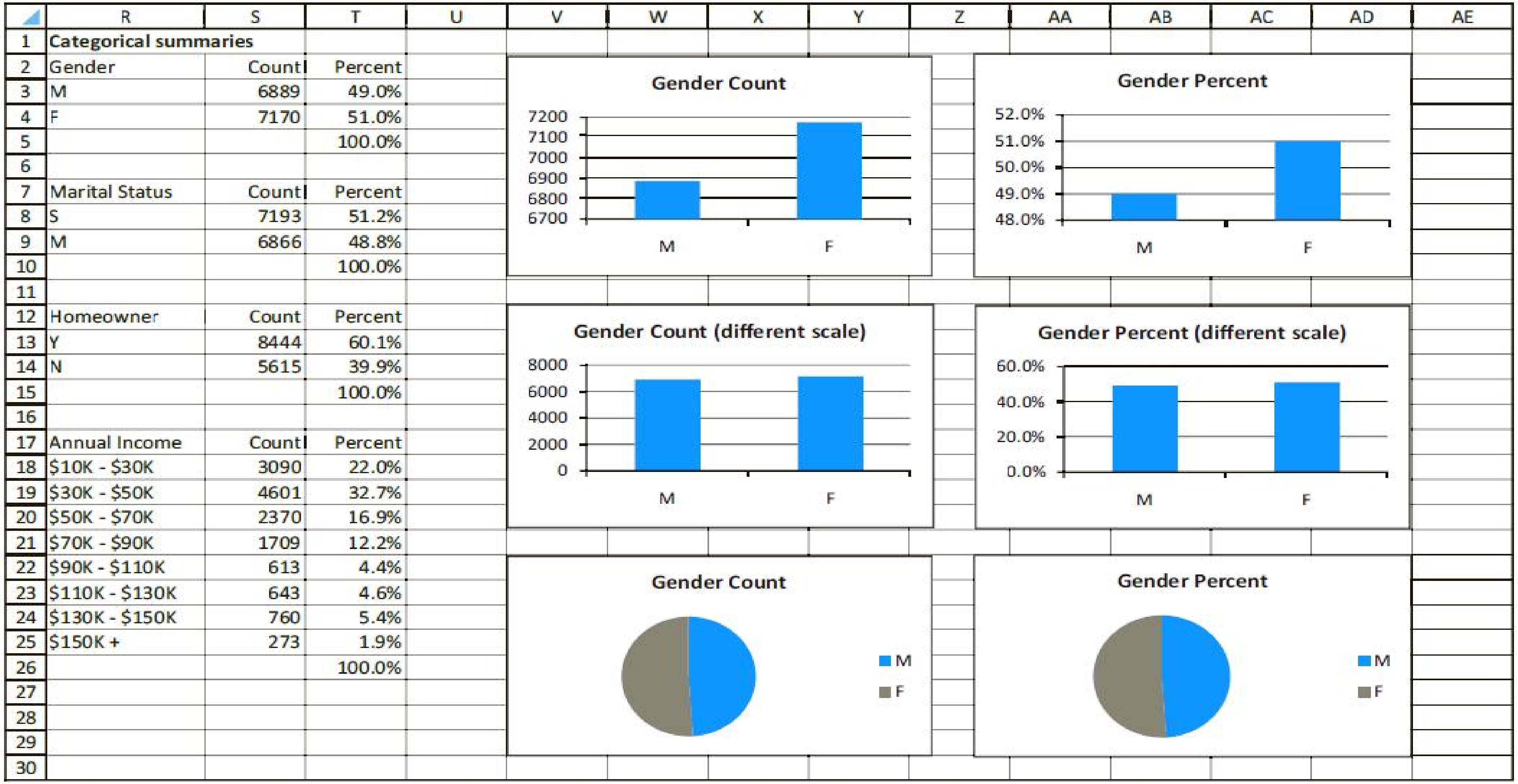
COLUMN

PERCENTAGE MARGINALS

Descriptive measures for categorical variables

Visualization

- Bar charts are most often used to visualize categorical variables.



	R	S	T
1	Categorical summaries		
2	Gender	Count	Percent
3	M	6889	49.0%
4	F	7170	51.0%
5			100.0%
6			
7	Marital Status	Count	Percent
8	S	7193	51.2%
9	M	6866	48.8%
10			100.0%
11			
12	Homeowner	Count	Percent
13	Y	8444	60.1%
14	N	5615	39.9%
15			100.0%
16			
17	Annual Income	Count	Percent
18	\$10K - \$30K	3090	22.0%
19	\$30K - \$50K	4601	32.7%
20	\$50K - \$70K	2370	16.9%
21	\$70K - \$90K	1709	12.2%
22	\$90K - \$110K	613	4.4%
23	\$110K - \$130K	643	4.6%
24	\$130K - \$150K	760	5.4%
25	\$150K +	273	1.9%
26			100.0%
27			
28			
29			
30			

Descriptive measures for Numerical variables

- Measure of Central Tendency.
- Measure of Variability.
- Measure of Shape.(Kurtosis and Skewness)

Measures of Central Tendency

There are three common measures of central tendency, all of which try to answer the basic question of which value is most “typical.”

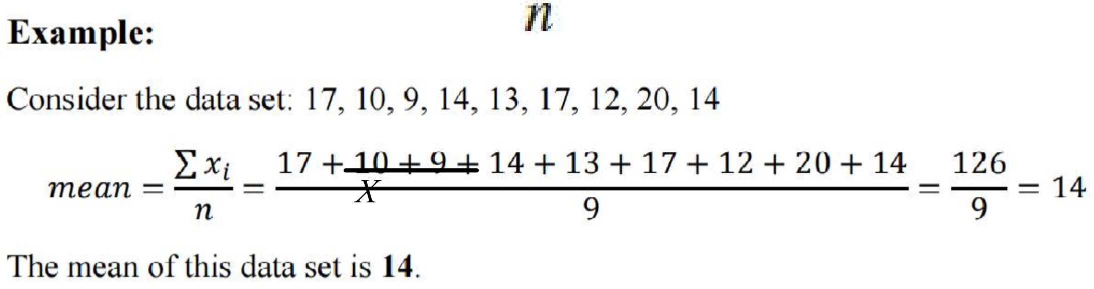
- These are the **mean, the median,** and **the mode.**

- **Mean of the Sample.**

A measure of central tendency is a number that represents the typical value in a collection of numbers.

Mean = sum of all data points/n (The mean is also known as the "average" or the "arithmetic average.")

$$\text{Mean} = \frac{\sum_{i=1}^{n} X_i}{n}$$



Example:

Consider the data set: 17, 10, 9, 14, 13, 17, 12, 20, 14

$$mean = \frac{\sum x_i}{n} = \frac{17 + \underset{X}{\cancel{10 + 9 +}} 14 + 13 + 17 + 12 + 20 + 14}{9} = \frac{126}{9} = 14$$

The mean of this data set is **14**.

- If the data set represents a sample from some larger population, this measure is called the sample mean and is denoted by $\overline{X}$.
- If the data set represents the entire population, it is called the population mean and is denoted by μ

Measures of Central Tendency

- A **trimmed mean** (sometimes called a ***truncated mean***) is similar to a mean, but it trims any outliers. Outliers can affect the mean (especially if there are just one or two very large values), so a trimmed mean can often be a better fit for data sets with erratic high or low values or for extremely skewed distributions. Even a small number of extreme values can corrupt the mean.
- For example, the mean salary at a company may be substantially pushed up by that of a few highly paid managers. Similarly, the mean score of a class in an exam could be pulled down quite a bit by a few very low scores.
- Which is the mean obtained after chopping off values at the high and low extremes.
- Example: Find the trimmed 20% mean for the following test scores: 60, 81, 83, 91, 99.
    - Step 1: Trim the top and bottom 20% from the data. That leaves us with the middle three values: ~~60~~, 81, 83, 91, ~~99~~.
    - Step 2: Find the mean with the remaining values. The mean is (81 + 83 + 91) / 3) = 85.

Measures of Central Tendency (cont..)

- **Median of a Simple.**
- The median of a set of data is the “middle element” when the data is arranged in ascending order. To determine the median:-

  1. Put the data in order from smallest to largest.
  2. Determine the number in the exact center.

  If there are an odd number of data points, the median will be the number in the absolute middle.

  If there is an **even number of data points, the median is the mean of the two center data points,** meaning the two center values should be added together and divided by 2.

- Example:
  - Consider the data set: 17, 10, 9, 14, 13, 17, 12, 20, 14
  - Step 1: Put the data in order from smallest to largest. 9, 10, 12, 13, 14, 14, 17, 17, 20
  - Step 2: Determine the absolute middle of the data. 9, 10, 12, 13, 14, 14, 17, 17, 20

Measures of Central Tendency (cont..)

The Mode of Sample:-

The mode is the most frequently occurring measurement in a data set.

There may be **one mode; multiple modes**, if more than one number occurs most frequently; or no mode at all, if every number occurs only once.

Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal,** and **trimodal**.

To determine the mode:

1. Put the data in order from smallest to largest, as you did to find your median.
2. Look for any value that occurs more than once.
3. Determine which of the values from Step 2 occurs most frequently.

- Example:-

Consider the data set: 17, 10, 9, 14, 13, 17, 12, 20, 14

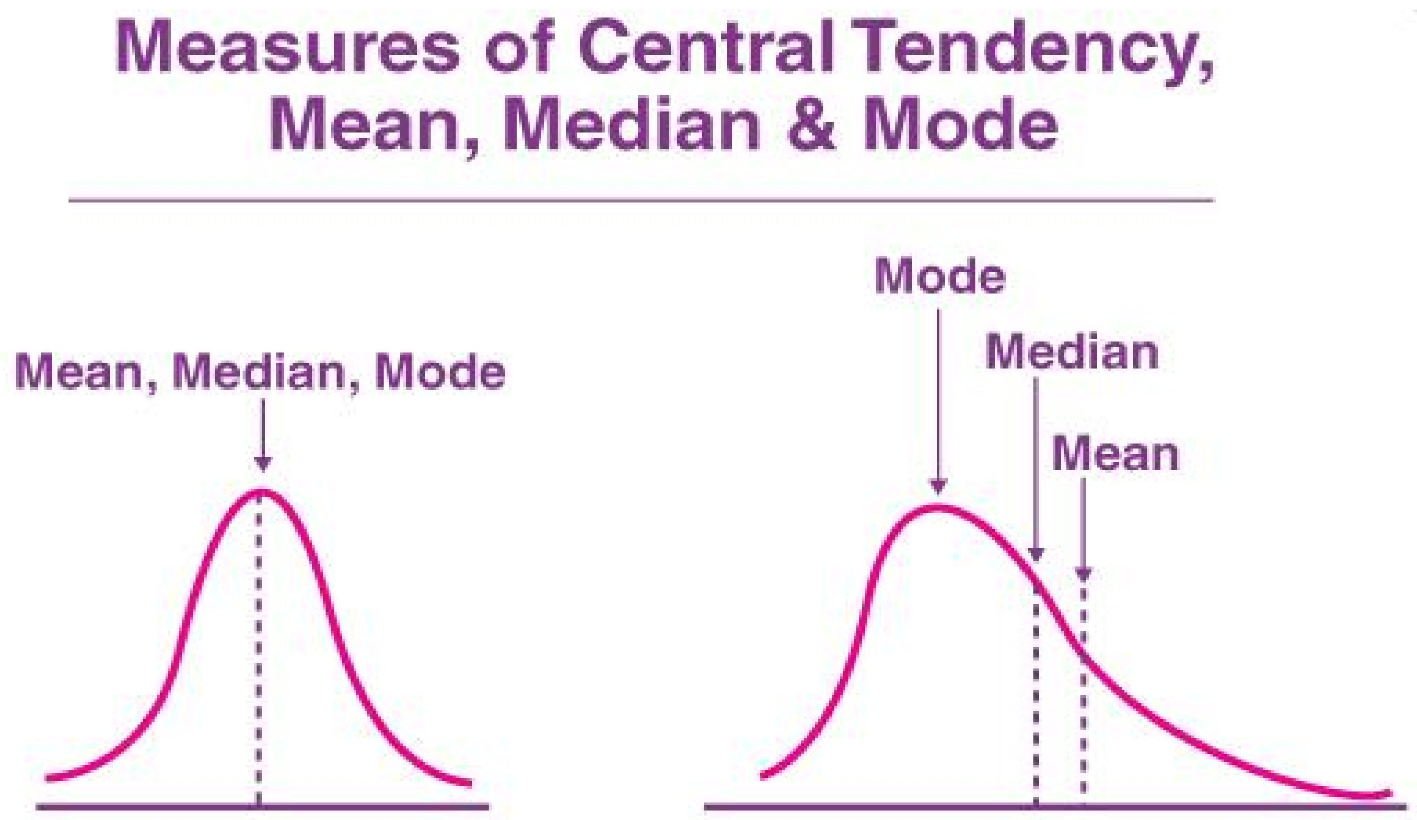
Step 1: Put the data in order from smallest to largest. 9, 10, 12, 13, 14, 14, 17, 17, 20

Step 2: Look for any number that occurs more than once. 9, 10, 12, 13, 14, 14, 17, 17, 20

Step 3: Determine which of those occur most frequently. 14 and 17 both occur twice.

The modes of this data set are 14 and 17.

Measures of Central Tendency (cont..)

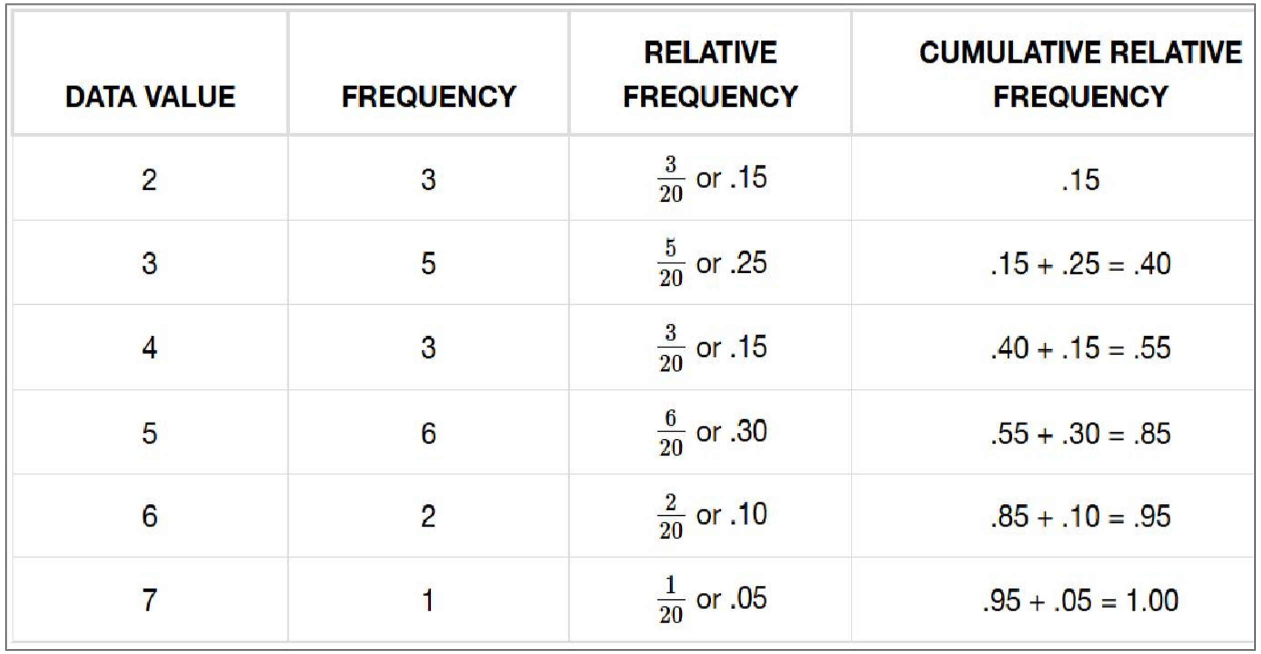


Measures of Central Tendency, Mean, Median & Mode

Frequency , Relative Frequency and Cumulative Relative Frequency.

- **Frequency** (or event) recording is a way to measure the number of times a behavior occurs within a given period.
- The **advantage** of using frequency distributions is that they present raw data in an organized, easy-to-read format. The most frequently occurring scores are easily identified, as are score ranges, lower and upper limits, cases that are not common, outliers, and total number of observations between any given scores.

- A **relative frequency** distribution shows the proportion of the total number of observations associated with each value or class of values and is related to a probability distribution
- **Advantage:** *Within the overall number of observations, a relative frequency reflects how frequently a given type of event occurs within that total number of observations.*

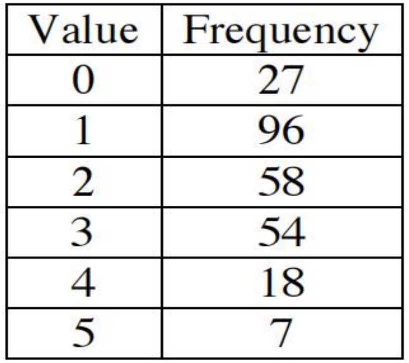


DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or .15	.15
3	5	$\frac{5}{20}$ or .25	.15 + .25 = .40
4	3	$\frac{3}{20}$ or .15	.40 + .15 = .55
5	6	$\frac{6}{20}$ or .30	.55 + .30 = .85
6	2	$\frac{2}{20}$ or .10	.85 + .10 = .95
7	1	$\frac{1}{20}$ or .05	.95 + .05 = 1.00

- **Cumulative frequency** represents the sum of the relative frequencies.
- **Cumulative frequency** is used to determine the number of observations that lie above (or below) a particular value in a data set.

Class room exercises

1. Find the Central tendency Mean , Median and Mode.



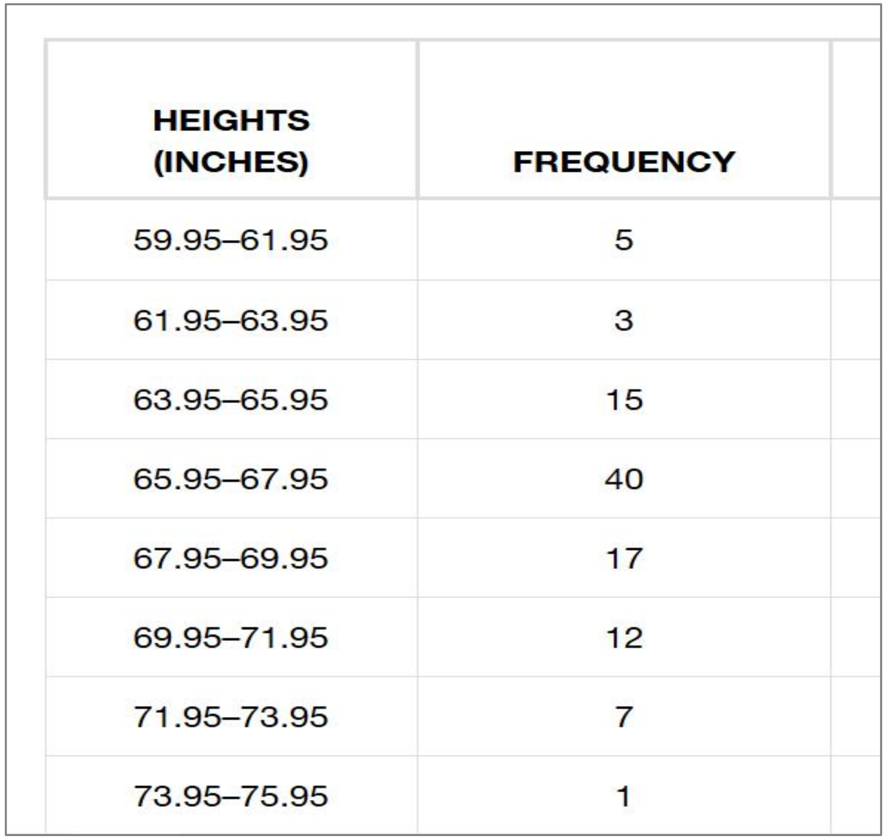
Value	Frequency
0	27
1	96
2	58
3	54
4	18
5	7

2. Table represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

   a) From Table , find the percentage of heights that are less than 65.95 inches.

   b) Find the percentage of heights that fall between 61.95 and 65.95 inches.

   c) The number of players in the sample who are between 61.95 and 71.95 inches tall is _____.



HEIGHTS (INCHES)	FREQUENCY
59.95–61.95	5
61.95–63.95	3
63.95–65.95	15
65.95–67.95	40
67.95–69.95	17
69.95–71.95	12
71.95–73.95	7
73.95–75.95	1

Mean, Median and Mode from Grouped Frequencies

Refer the “**Mean, Median and Mode from Grouped Frequencies**“ slides separatly..........

Measures of Variablility.

- Measures of variability give a sense of how spread out the response values are.
- The range, standard deviation and variance each reflect different aspects of spread.
- **Percentiles** and **quartiles** certainly tell you something about variability.
- Specifically, for any percentage ***p***, the ***p***th percentile is the value such that a percentage ***p*** of all values are less than it. Similarly, the first, second, and third quartiles are the percentiles corresponding to ***p = 25%***, ***p = 50%***, and ***p = 75%***. These three values divide the data into four groups, each with (approximately) a quarter of all observations.
- Note that the second quartile is equal to the median by definition.

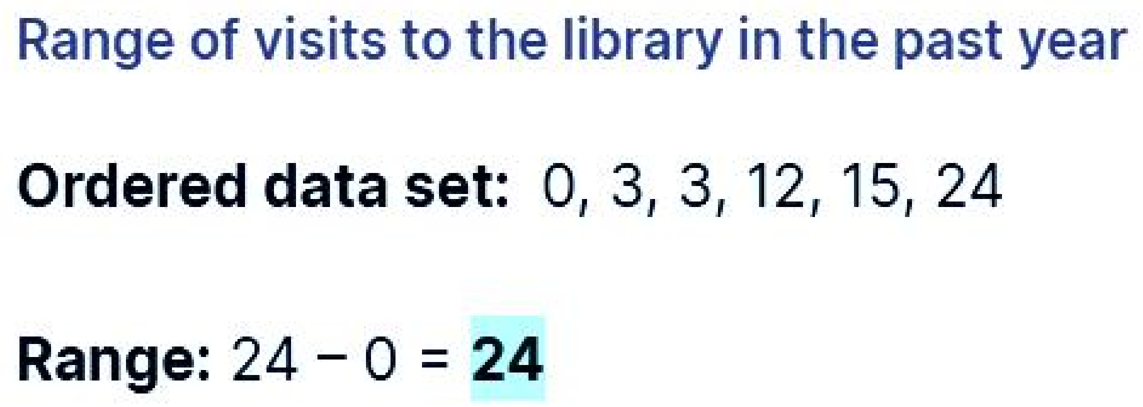
For example, if you learn that your score in the verbal SAT test is at the 93rd percentile, this means that you scored better than 93% of those taking the test.

Measures of Variablility (cont..)

- Range
- Interquartile Range
- Variance
- Standard Deviation

Range

The range gives you an idea of how far apart the most extreme response scores are. To find the range, simply subtract the lowest value from the highest value.



Range of visits to the library in the past year

Ordered data set: 0, 3, 3, 12, 15, 24

Range: 24 − 0 = **24**

Measures of Variablility (cont..)

Interquartile range

- The Interquartile range measures the variability, based on dividing an ordered set of data into quartiles.
- Quartiles are three values or cuts that divide each respective part as the first, second, and third quartiles, denoted by Q1, Q2, and Q3

Q1= It is the cut in the first half of the rank-ordered data set

Q2= It is the median value of the set

Q3= It is the cut in the second half of the rank-ordered data set.

So it is really the range of the middle 50% of the data.

Measures of Variablility (cont..)

Interquartile range(IQR) = Upper Quartile – Lower Quartile

IQR=Q3–Q1

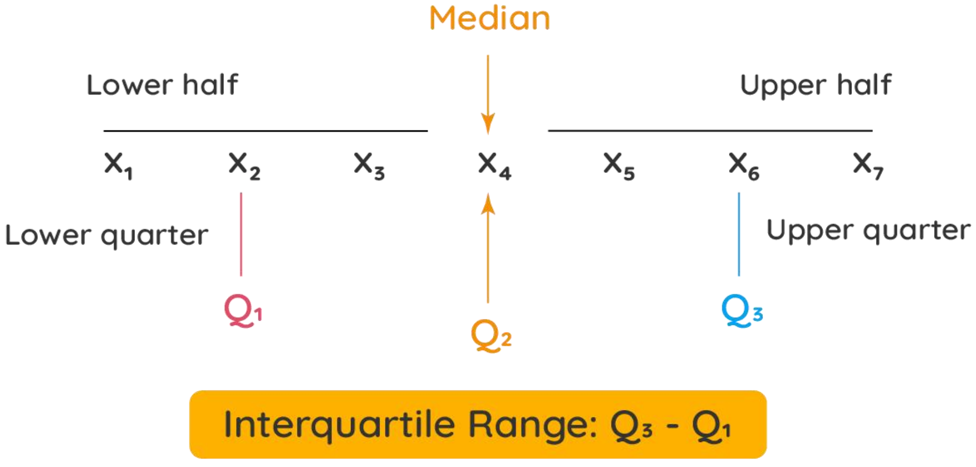
Where,

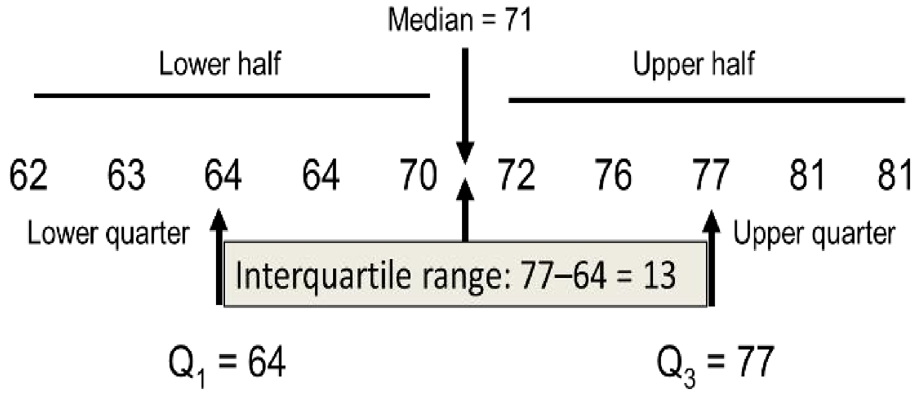
IQR = Interquartile range

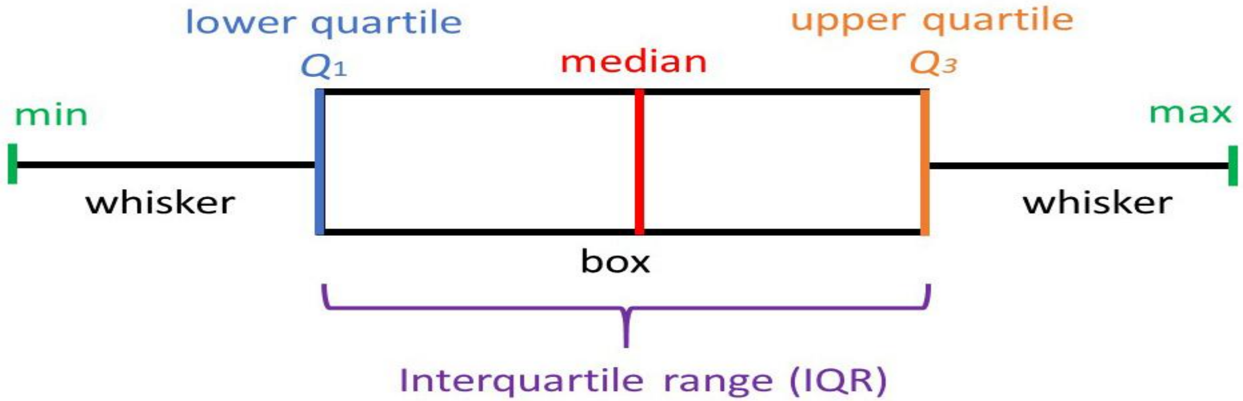
Q1 = (1/4)[(n + 1)]th term)

Q3= (3/4)[(n + 1)]th term)

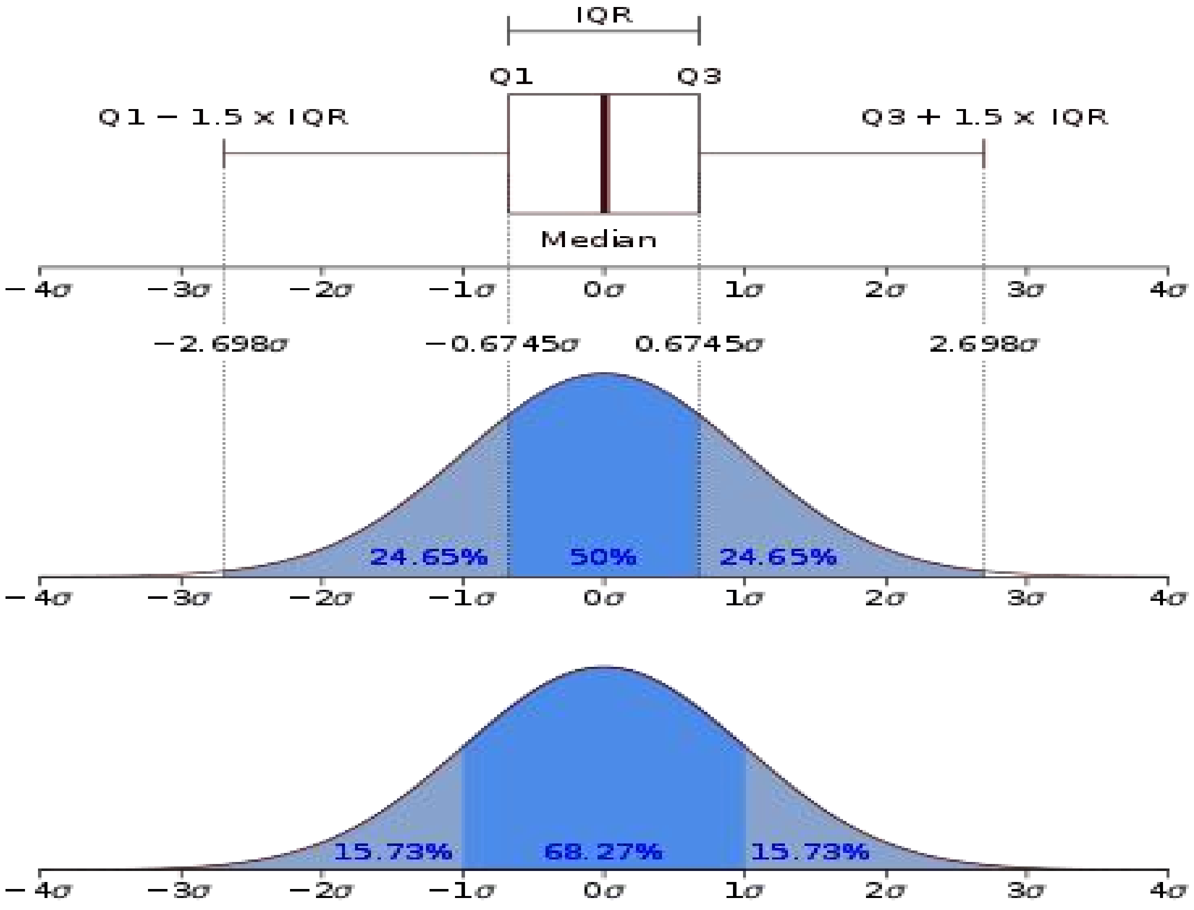
n = number of data points

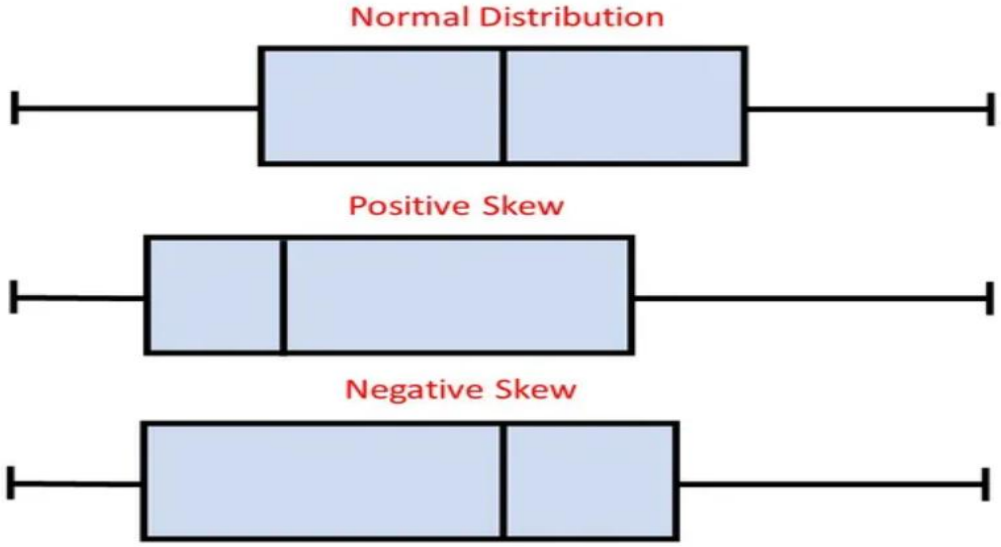






Measures of Variablility (cont..)





- The median of the data set is located to the right of the center of the box, which indicates that the distribution is **negatively skewed**.

- The median of the data set is located to the left of the center of the box, which indicates that the distribution is **positively skewed.**

Measures of Variablility (cont..)

Variance

The variance is essentially the average of the squared deviations from the mean, where if ***X_i*** is a typical observation, its squared deviation from the mean is **$(X_i$ – mean$)^2$**.

As in the discussion of the mean, there is a sample variance, denoted by ***s^2***, and a population variance, denoted by ***σ^2***

Formula for Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n - 1}$$

Formula for Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \text{mean})^2}{n}$$

The use variance to see how individual numbers relate to each other within a data set. Variance analysis helps an organization to be proactive in achieving their business targets

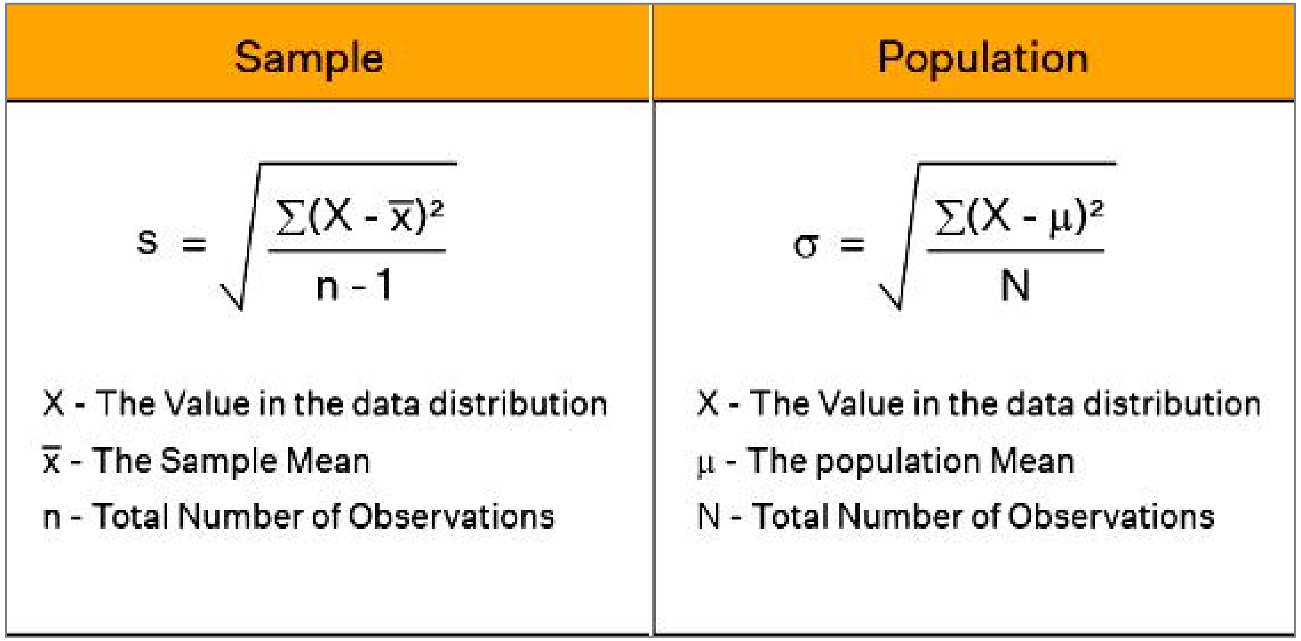
Measures of Variablility (cont..)

Standard deviation

A fundamental problem with variance is that it is in squared units.

A more natural measure is the **standard deviation**, which is the square root of the variance.

- The **sample standard deviation**, denoted by ***s***, is the square root of the sample variance.
- The **population standard deviation**, denoted by ***σ***, is the square root of the population variance.



Sample	Population
$s = \sqrt{\dfrac{\sum(X - \overline{x})^2}{n - 1}}$	$\sigma = \sqrt{\dfrac{\sum(X - \mu)^2}{N}}$
X - The Value in the data distribution $\overline{x}$ - The Sample Mean n - Total Number of Observations	X - The Value in the data distribution μ - The population Mean N - Total Number of Observations

Measures of Variablility (cont..)

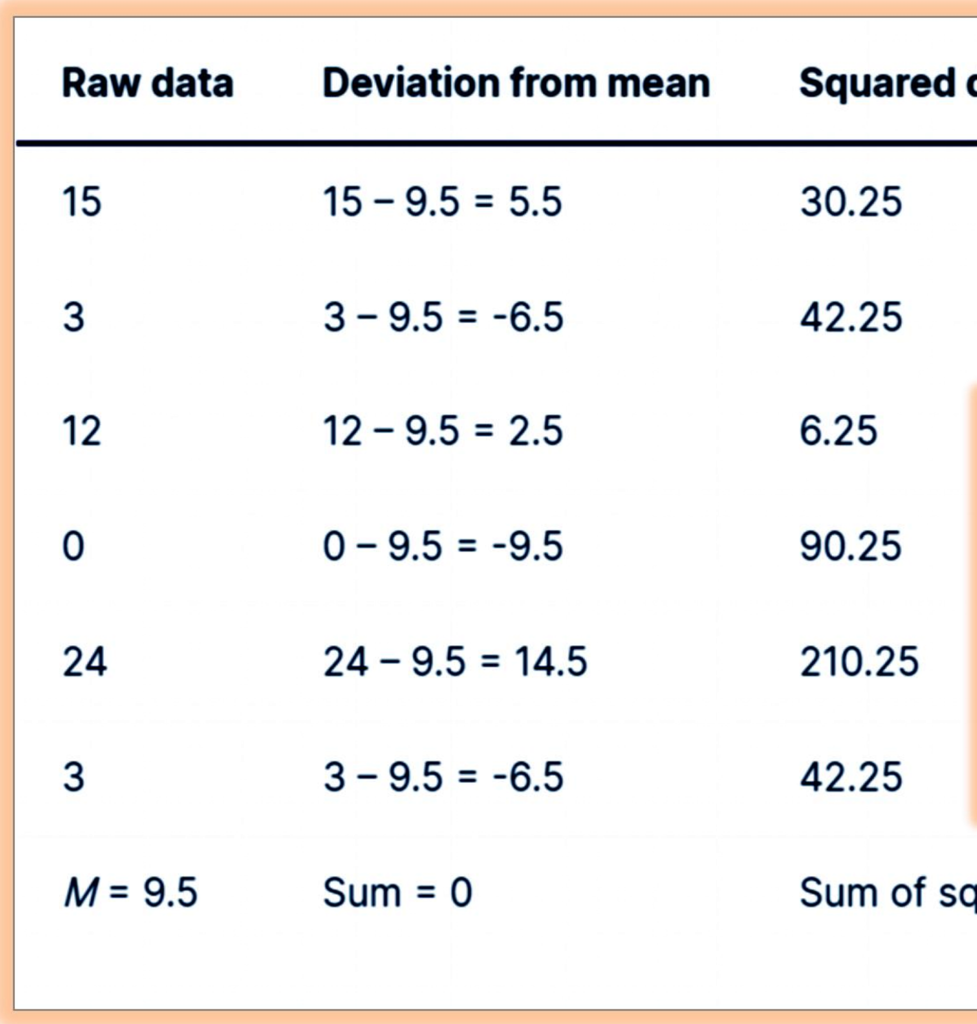
Standard deviation

Important Points

- Standard deviation is sensitive to extreme values. A single very extreme value can increase the standard deviation and misrepresent the dispersion.
- For two data sets with the same mean, the one with the larger standard deviation is the one in which the data is more spread out from the center.
- Standard deviation is equal to 0 if all values are equal (because all values are then equal to the mean).

Measures of Variablility (cont..)

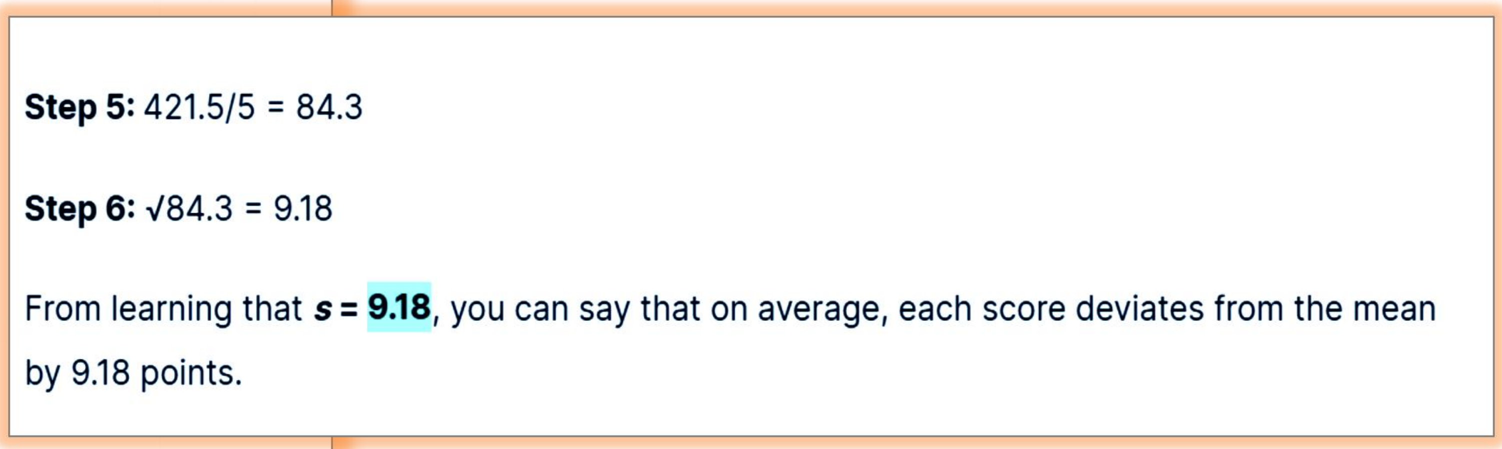
Standard deviation



Raw data	Deviation from mean	Squared deviation
15	15 − 9.5 = 5.5	30.25
3	3 − 9.5 = -6.5	42.25
12	12 − 9.5 = 2.5	6.25
0	0 − 9.5 = -9.5	90.25
24	24 − 9.5 = 14.5	210.25
3	3 − 9.5 = -6.5	42.25
M = 9.5	Sum = 0	Sum of squares = 421.5

There are six steps for finding the standard deviation:

1. **List each score and find their mean.**
2. **Subtract the mean from each score to get the deviation from the mean.**
3. **Square each of these deviations.**
4. **Add up all of the squared deviations.**
5. **Divide the sum of the squared deviations by N – 1.**
6. **Find the square root of the number you found.**



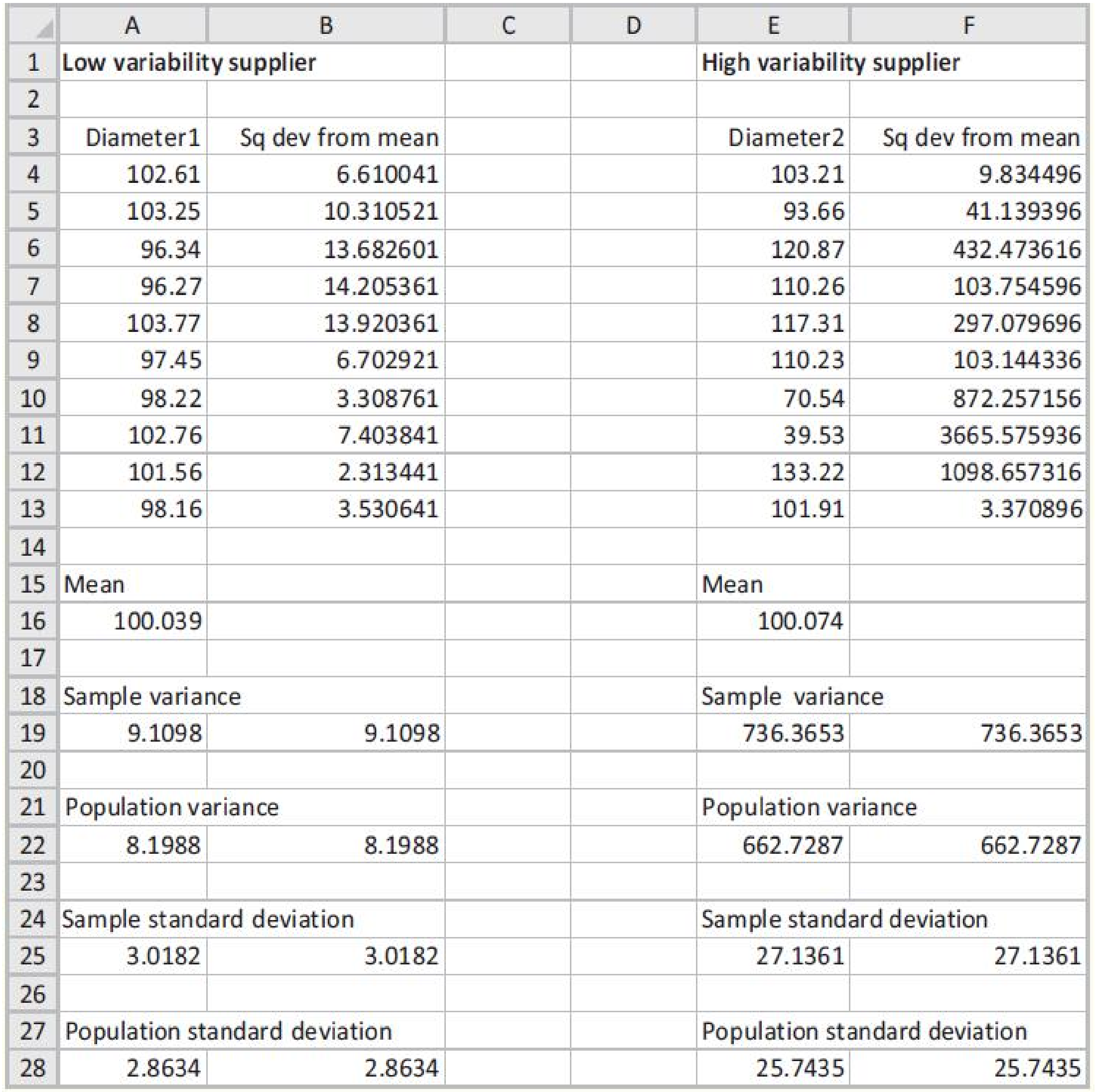
Step 5: 421.5/5 = 84.3

Step 6: √84.3 = 9.18

From learning that ***s*** **= 9.18**, you can say that on average, each score deviates from the mean by 9.18 points.

Measures of Variablility (cont..)

Standard deviation



	A	B	C	D	E	F
1	Low variability supplier				High variability supplier	
2						
3	Diameter1	Sq dev from mean			Diameter2	Sq dev from mean
4	102.61	6.610041			103.21	9.834496
5	103.25	10.310521			93.66	41.139396
6	96.34	13.682601			120.87	432.473616
7	96.27	14.205361			110.26	103.754596
8	103.77	13.920361			117.31	297.079696
9	97.45	6.702921			110.23	103.144336
10	98.22	3.308761			70.54	872.257156
11	102.76	7.403841			39.53	3665.575936
12	101.56	2.313441			133.22	1098.657316
13	98.16	3.530641			101.91	3.370896
14						
15	Mean				Mean	
16	100.039				100.074	
17						
18	Sample variance				Sample variance	
19	9.1098	9.1098			736.3653	736.3653
20						
21	Population variance				Population variance	
22	8.1988	8.1988			662.7287	662.7287
23						
24	Sample standard deviation				Sample standard deviation	
25	3.0182	3.0182			27.1361	27.1361
26						
27	Population standard deviation				Population standard deviation	
28	2.8634	2.8634			25.7435	25.7435

Empirical Rules for Interpreting Standard Deviation

- The interpretation of the standard deviation can be stated as three **empirical rules**.
  - “Empirical” means that they are based on commonly observed data, as opposed to theoretical mathematical arguments.
  - If the values of a variable are approximately ***normally*** **distributed (symmetric and bell-shaped)**, then the following rules hold:
    - Approximately 68% of the observations are within one standard deviation of the mean, that is interval $\overline{X} \pm s$.
    - Approximately 95% of the observations are within two standard deviations of the mean, that is interval $\overline{X} \pm 2s$.
    - Approximately 99.7% of the observations are within three standard deviations of the mean, that is interval $\overline{X} \pm 3s$.

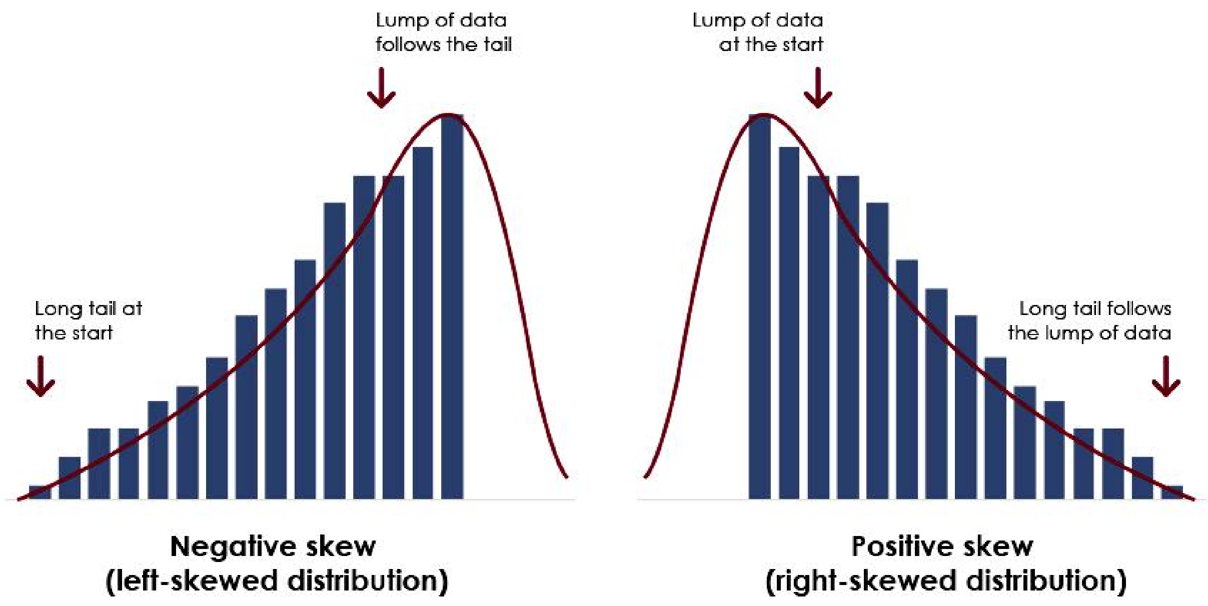
Measures of Shape

There are two final measures of a distribution you will hear occasionally:

- **skewness** and
- **kurtosis**.

Skewness

- Skewness refers to the degree of symmetry, or more precisely, the degree of lack of symmetry.
- Skewness is used to measure the level of asymmetry in our graph.
- It is the measure of asymmetry that occurs when our data deviates from the norm.



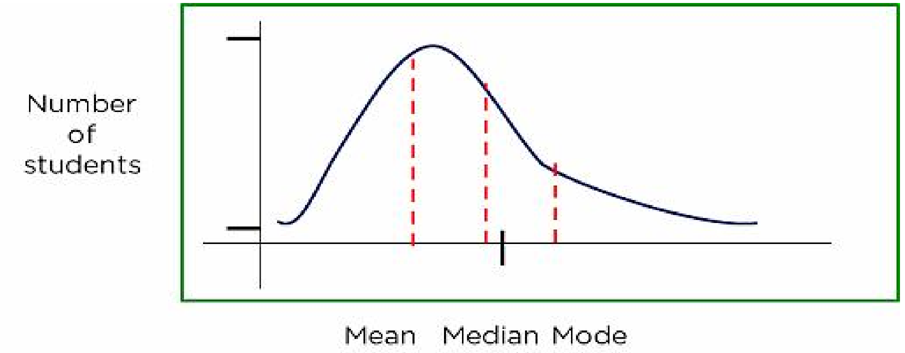
Measures of Shape (cont..)

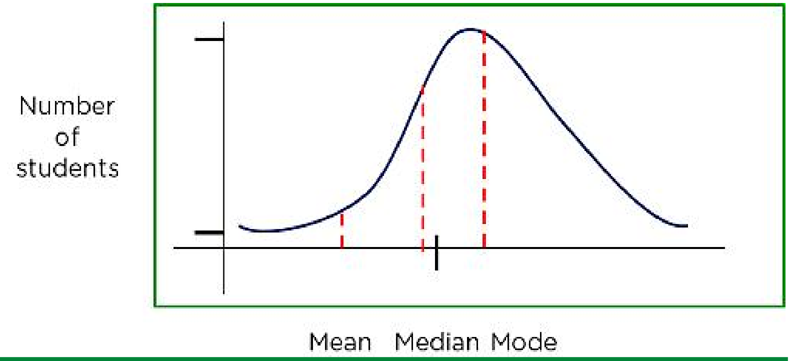
Skewness

- Sometimes, the normal distribution tends to tilt more on one side.
- This is because the probability of data being more or less than the mean is higher and hence makes the distribution asymmetrical.
- This also means that the data is not equally distributed.
- The skewness can be on two types:

1. **Positively Skewed:** In a distribution that is Positively Skewed, the values are more concentrated towards the right side, and the left tail is spread out. Hence, the statistical results are bent towards the left-hand side. Hence, that the mean, median, and mode are always positive.
In this distribution, Mean > Median > Mode.

2. **Negatively Skewed:** In a Negatively Skewed distribution, the data points are more concentrated towards the right-hand side of the distribution. This makes the mean, median, and mode bend towards the right. Hence these values are always negative.
In this distribution, Mode > Median > Mean.





Measures of Shape (cont..)

Skewness

How these central tendency measures tend to spread when the normal distribution is distorted.

For the nomenclature just follow the direction of the tail

— For the right graph has the tail to the right, so it is right-skewed (positively skewed) and

— For the left graph since the tail is to the left, it is left-skewed (negatively skewed).

How about deriving a measure that captures the horizontal distance between the Mode and the Mean of the distribution?

It's intuitive to think that the higher the skewness, the more apart these measures will be.

Pearson's Coefficient of Skewness : This method is most frequently used for measuring skewness.

The formula for measuring coefficient of skewness is given by

$$\text{Pearson's First Coefficient} = \frac{Mean - Mode}{Standard\ deviation}$$

$$\text{Skewness} = \frac{\sum (x - \bar{x})^3}{(n - 1) \cdot S^3}$$

Where:

S: standard deviation

$\bar{X}$: Mean

The value of this coefficient would be **zero** in a ***symmetrical distribution***. If mean is greater than mode, coefficient of skewness would be positive otherwise negative. The value of the Pearson's coefficient of skewness usually lies between ±1 for moderately skewed distubution.

Measures of Shape (cont..)

Skewness

$$\text{Pearson's First Coefficient} = \frac{Mean - Mode}{Standard\ deviation}$$

If mode is not well defined, we use the formula

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$

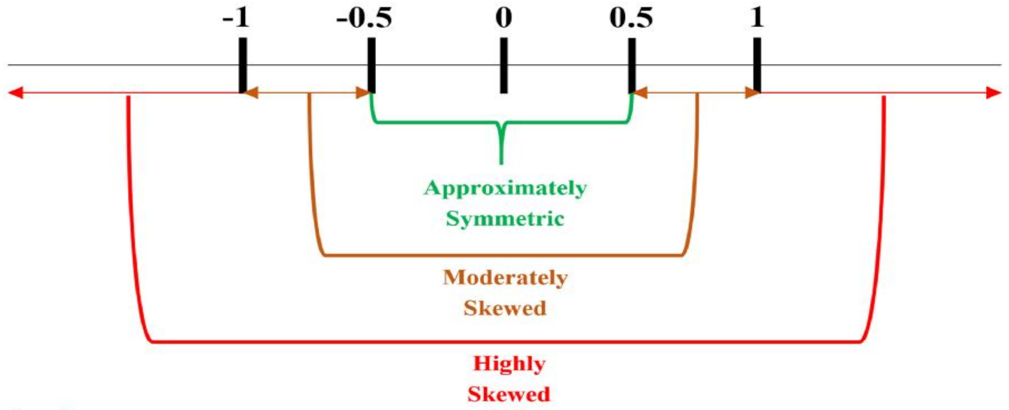
Substituting this in Pearson’s first coefficient gives us Pearson’s second coefficient and the formula for skewness:

$$\text{Pearson's Second Coefficient} = \frac{3(Mean - Median)}{Standard\ deviation}$$

If this value is between:

- -0.5 and 0.5, the distribution of the value is almost symmetrical
- -1 and -0.5, the data is negatively skewed, and if it is between 0.5 to 1, the data is positively skewed. The skewness is moderate.
- If the skewness is lower than -1 (negatively skewed) or greater than 1 (positively skewed), the data is highly skewed.

Scale of Skewness:



(-0.5, 0.5) = Low

(-1, -0.5) U (0.5, 1) = Moderate

(-1 & beyond) U (1 & beyond) = High

Measures of Shape (cont..)

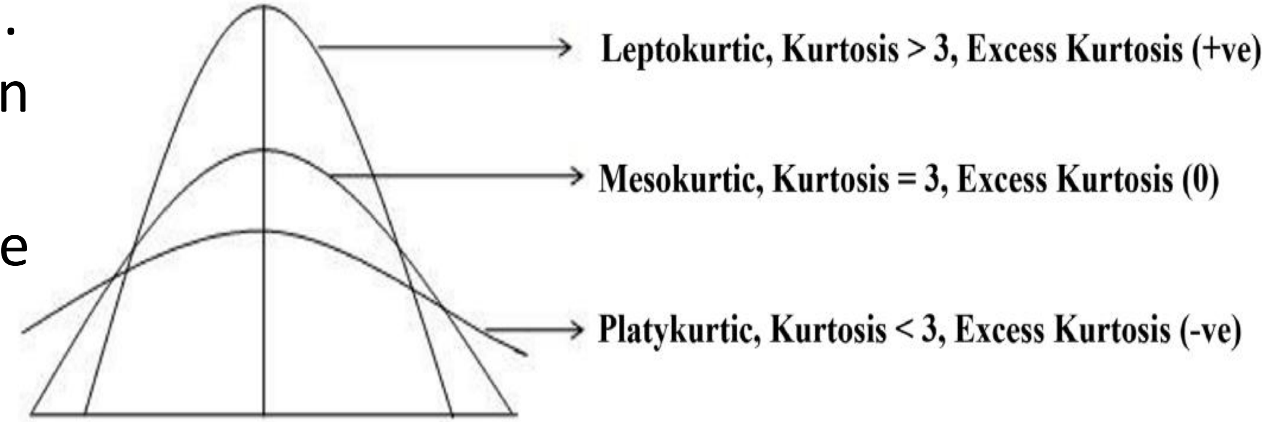
What Is Kurtosis?

- Kurtosis gives a measure of flatness of distribution.
- We need to know another measure to get the complete idea about the shape of the distribution which can be studied with the help of Kurtosis.
- Kurtosis is associated with the ***"movement of probability mass from the shoulders of a distribution into its center and tails."***
- The degree of kurtosis of a distribution is measured relative to that of a normal curve.
- The curves with greater peakedness than the normal curve are called ***"Leptokurtic"***.
- The curves which are more flat than the normal curve are called ***"Platykurtic"***.
- The normal curve is called ***"Mesokurtic."***

For sample:

$$\boxed{\text{Kurtosis} = \frac{\sum (x - \bar{x})^4}{(n - 1) \cdot S^4}}$$

Where:
S: standard deviation $\overline{X}$: Mean

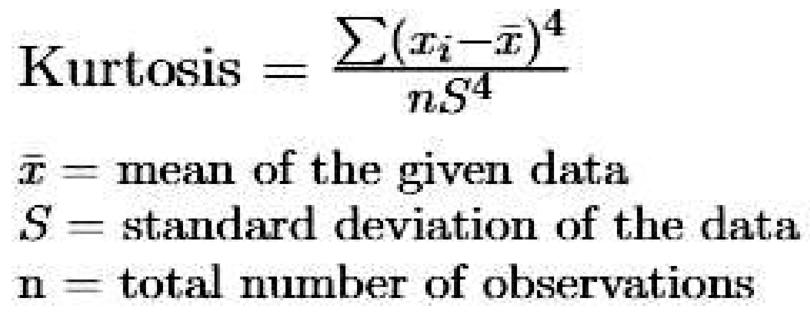

Leptokurtic, Kurtosis > 3, Excess Kurtosis (+ve)
Mesokurtic, Kurtosis = 3, Excess Kurtosis (0)
Platykurtic, Kurtosis < 3, Excess Kurtosis (-ve)

Measures of Shape (cont..)

Kurtosis

Kurtosis is used to find the presence of outliers in data. It gives us the total degree of outliers present.

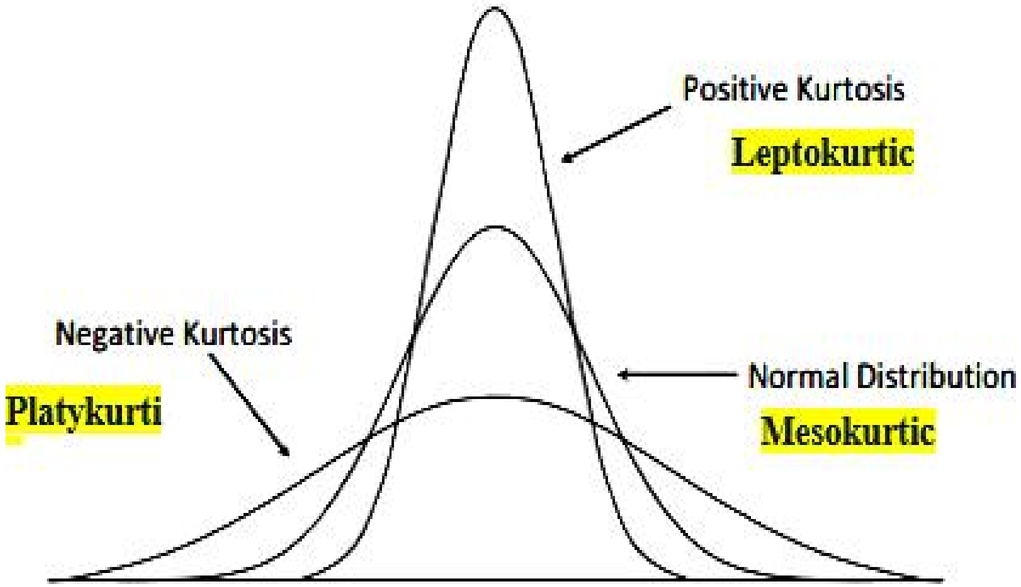
- A normal distribution has kurtosis **exactly 3** (excess kurtosis exactly 0).
- Any distribution with **kurtosis ≈3** (excess ≈0) is called ***mesokurtic***.
- A distribution with **kurtosis <3** (excess kurtosis <0) is called ***platykurtic***. (Its tails are shorter and thinner, and often its central peak is lower and broader).
- A distribution with ***kurtosis >3*** (excess kurtosis >0) is called ***leptokurtic.*** (Its tails are longer and fatter, and often its central peak is higher and sharper).



$$\text{Kurtosis} = \frac{\sum(x_i - \bar{x})^4}{nS^4}$$

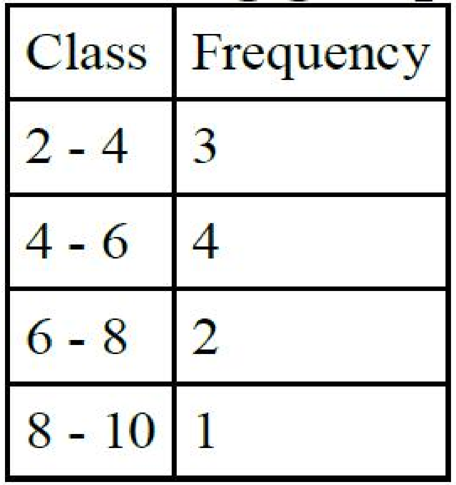
$\bar{x}$ = mean of the given data
S = standard deviation of the data
n = total number of observations

Excess Kurtosis = Kurtosis – 3



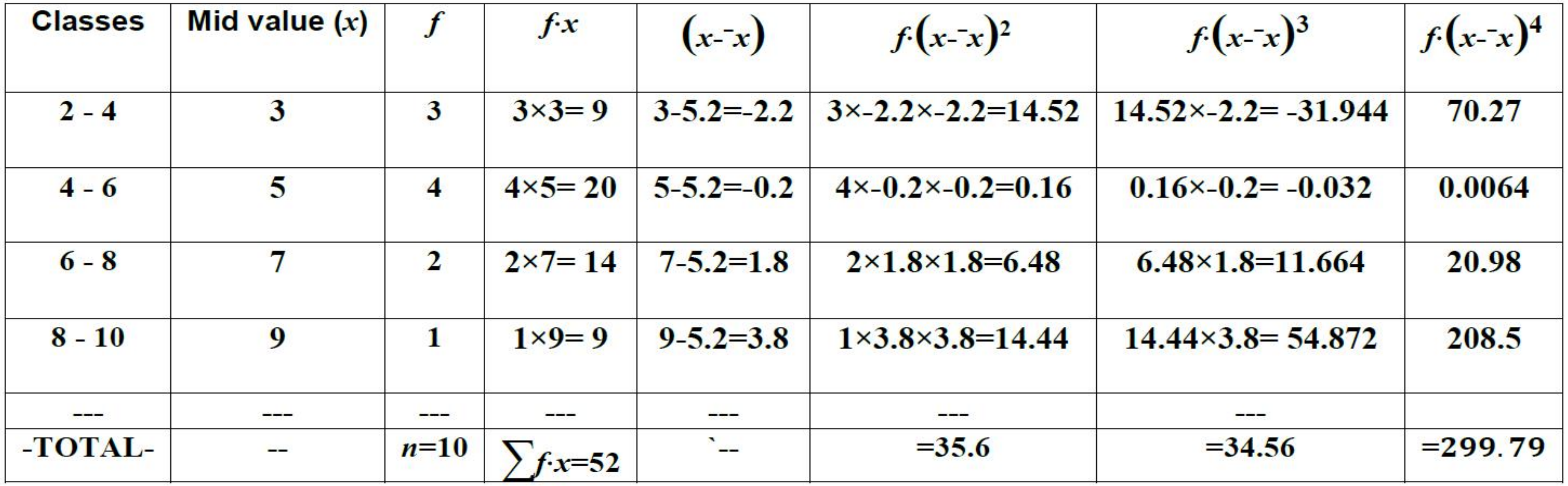
Measures of Sample Skewness and Kurtosis

Examples: Calculate Sample Skewness and Sample Kurtosis from the following grouped data



Class	Frequency
2 - 4	3
4 - 6	4
6 - 8	2
8 - 10	1

Solution:



Classes	Mid value (x)	f	$f\cdot x$	$(x-\bar{x})$	$f\cdot(x-\bar{x})^2$	$f\cdot(x-\bar{x})^3$	$f\cdot(x-\bar{x})^4$
2 - 4	3	3	3×3= 9	3-5.2=-2.2	3×-2.2×-2.2=14.52	14.52×-2.2= -31.944	70.27
4 - 6	5	4	4×5= 20	5-5.2=-0.2	4×-0.2×-0.2=0.16	0.16×-0.2= -0.032	0.0064
6 - 8	7	2	2×7= 14	7-5.2=1.8	2×1.8×1.8=6.48	6.48×1.8=11.664	20.98
8 - 10	9	1	1×9= 9	9-5.2=3.8	1×3.8×3.8=14.44	14.44×3.8= 54.872	208.5
---	---	---	---	---	---	---	
-TOTAL-	--	n=10	$\sum f\cdot x$=52	`--	=35.6	=34.56	=299.79

$$\text{Mean} = \sum \boldsymbol{f}\cdot\boldsymbol{x} = \frac{\sum f\cdot x}{\sum \text{f}} = \frac{52}{10} = 5.2$$

Standard deviation (S.D)

$$S.D = \sqrt{\frac{\sum_{i=1}^{n} fi(Xi-\bar{X})^2}{\sum_{i=1}^{n} fi}}$$

$$S.D = \sqrt{\frac{35.6}{10}} = 1.88$$

Calculate the Skewness

$$\boxed{\text{Skewness} = \frac{\sum (x - \bar{x})^3}{(n - 1)\cdot S^3}}$$

$$\text{Skewness} = \frac{34.56}{9*(1.88)^3} = 0.48$$

Calculate the Kurtosis:

$$\boxed{\text{Kurtosis} = \frac{\sum (x - \bar{x})^4}{(n - 1)\cdot S^4}}$$

$$\text{Kurtosis} = \frac{299.79}{9*(1.88)^4} = 2.12$$

Outliers and Missing values

Outliers

- An **outlier** is a value or an entire observation (row) that lies well outside of the norm.
  - Some statisticians define an outlier as any value more than three standard deviations from the mean, but this is only a rule of thumb.
- Even if values are not unusual by themselves, there still might be unusual *combinations* of values.
- When dealing with outliers, it is best to run the analyses two ways: with the outliers and without them.
- Let’s just agree to define outliers as ***extreme*** values, and then for any particular data set, you can decide how ***extreme*** a value needs to be to qualify as an outlier.

For example, let us consider a row of data [10,15,22,330,30,45,60]. In this dataset, we can easily conclude that 330 is way off from the rest of the values in the dataset, thus 330 is an outlier. It was easy to figure out the outlier in such a small dataset, but when the dataset is huge, we need various methods to determine whether a certain value is an outlier or necessary information.

Outliers (cont..)

Types of outliers :

There are three types of outliers

- **Global Outliers:** *The data point or points whose values are far outside everything else in the dataset are global outliers. Suppose we look at a taxi service company's number of rides every day. The rides suddenly dropped to zero due to the pandemic-induced lockdown. This sudden decrease in the number is a global outlier for the taxi company.*
- **Collective Outliers:** *Some data points collectively as a whole deviates from the dataset. These data points individually may not be a global or contextual outlier, but they behave as outliers when aggregated together. For example, closing all shops in a neighborhood is a collective outlier as individual shops keep on opening and closing, but all shops together never close down; hence, this scenario will be considered a collective outlier.*
- **Contextual Outliers:** *Contextual outliers are those values of data points that deviate quite a lot from the rest of the data points that are in the same context, however, in a different context, it may not be an outlier at all. For example, a sudden surge in orders for an e-commerce site at night can be a contextual outlier.*

Outliers can lead to vague or misleading predictions while using machine learning models. Specific models like linear regression, logistic regression, and support vector machines are susceptible to outliers. Outliers decrease the mathematical power of these models, and thus the output of the models becomes unreliable.

Outliers (cont..)

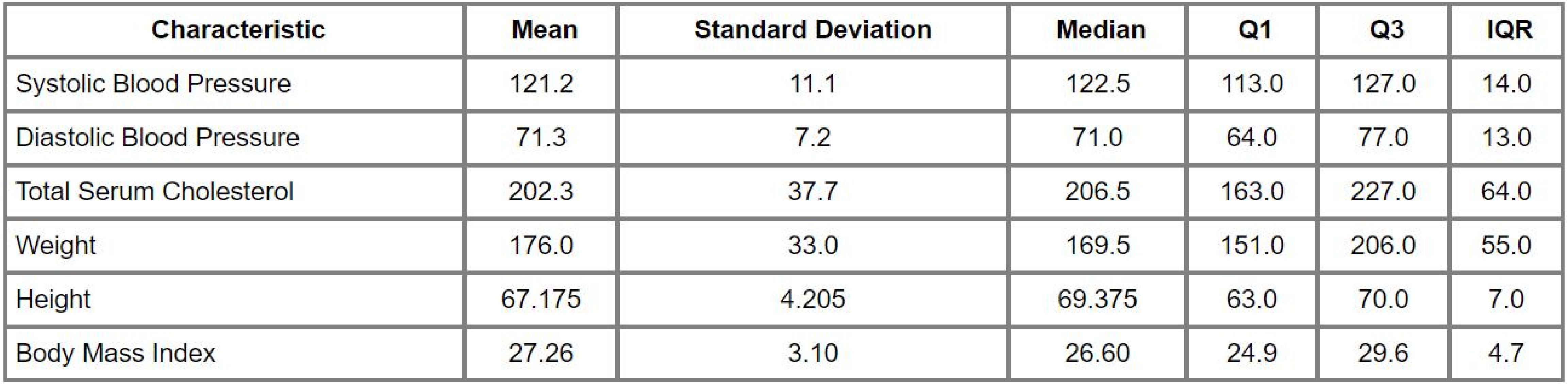
- **Tukey Fences**
- When there are no outliers in a sample, the mean and standard deviation are used to summarize a typical value and the variability in the sample, respectively.
- When there are outliers in a sample, the median and interquartile range are used to summarize a typical value and the variability in the sample, respectively.
- Outliers are values **below Q1-1.5(Q3-Q1)** or **above Q3+1.5(Q3-Q1)** or equivalently, values below **Q1-1.5 IQR** or above **Q3+1.5 IQR**.
- In previous example, for the diastolic blood pressures, the lower limit is 64 - 1.5(77-64) = 44.5 and the upper limit is 77 + 1.5(77-64) = 96.5. The diastolic blood pressures range from 62 to 81. Therefore there are no outliers.

Outliers (cont..)

Example : The Full Framingham Cohort Data

- The Framingham Heart Study is a long-term, ongoing cardiovascular cohort study on residents of the city of Framingham, Massachusetts. The study began in 1948 with 5,209 adult subjects from Framingham, and is now on its third generation of participants.
- Table 1 displays the means, standard deviations, medians, quartiles and interquartile ranges for each of the continuous variables in the subsample of n=10 participants who attended the seventh examination of the Framingham Offspring Study.

Table 1 - Summary Statistics on n=10 Participants

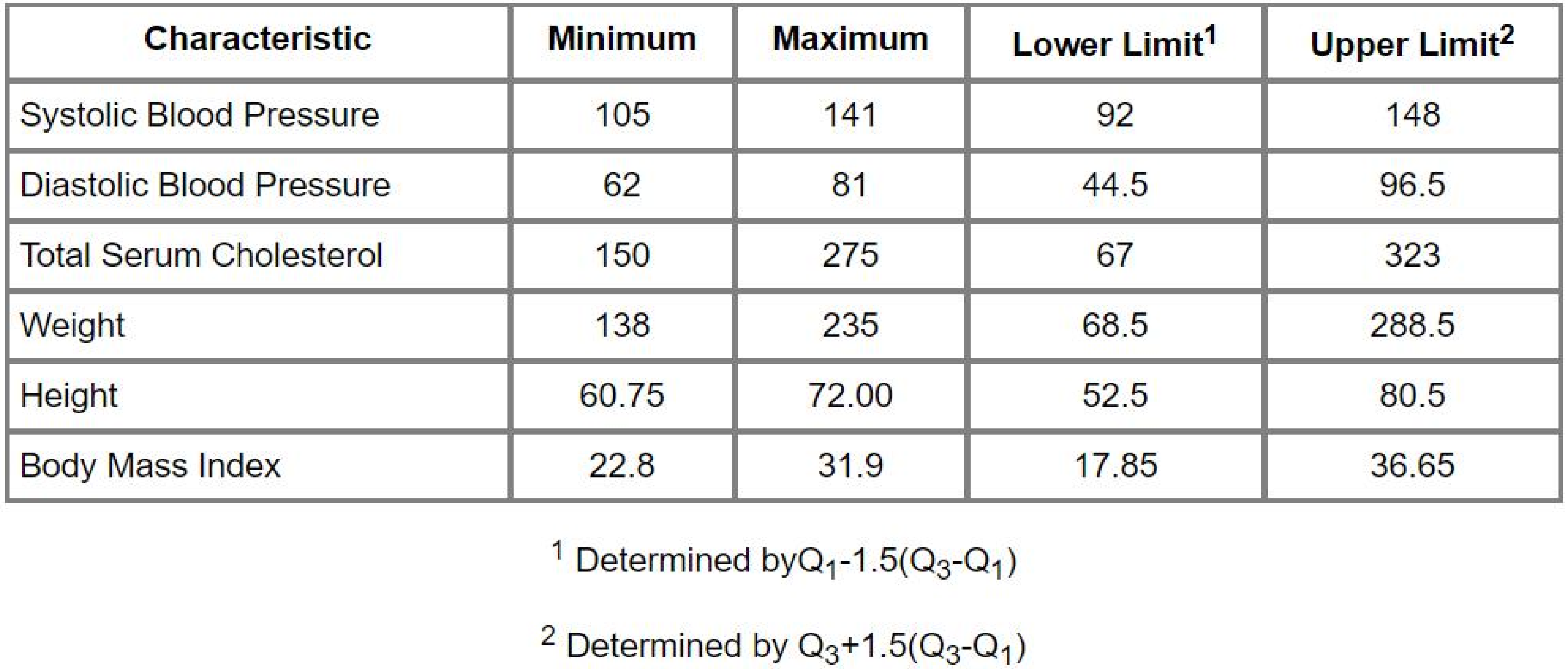


Characteristic	Mean	Standard Deviation	Median	Q1	Q3	IQR
Systolic Blood Pressure	121.2	11.1	122.5	113.0	127.0	14.0
Diastolic Blood Pressure	71.3	7.2	71.0	64.0	77.0	13.0
Total Serum Cholesterol	202.3	37.7	206.5	163.0	227.0	64.0
Weight	176.0	33.0	169.5	151.0	206.0	55.0
Height	67.175	4.205	69.375	63.0	70.0	7.0
Body Mass Index	27.26	3.10	26.60	24.9	29.6	4.7

Outliers (cont..)

- Table 2 displays the observed minimum and maximum values along with the limits to determine outliers using the quartile rule for each of the variables in the subsample of n=10 participants.
- Are there outliers in any of the variables? Which statistics are most appropriate to summarize the average or typical value and the dispersion?

Table 2 - Limits for Assessing Outliers in Characteristics Measured in the n=10 Participants



Characteristic	Minimum	Maximum	Lower Limit[1]	Upper Limit[2]
Systolic Blood Pressure	105	141	92	148
Diastolic Blood Pressure	62	81	44.5	96.5
Total Serum Cholesterol	150	275	67	323
Weight	138	235	68.5	288.5
Height	60.75	72.00	52.5	80.5
Body Mass Index	22.8	31.9	17.85	36.65

[1] Determined byQ_1-1.5(Q_3-Q_1)

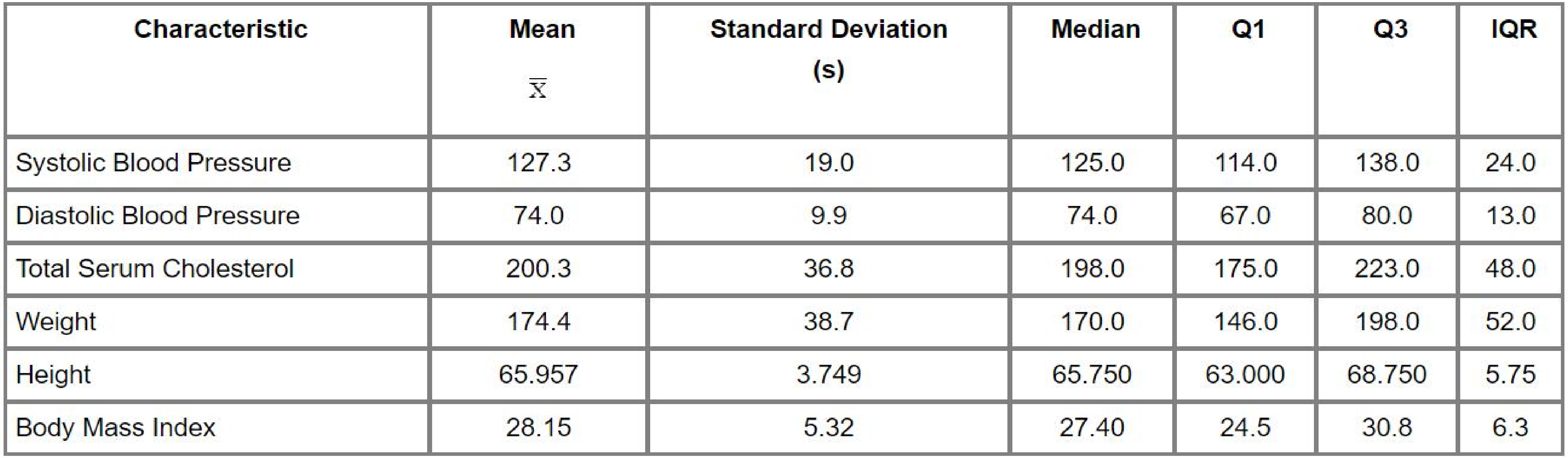
[2] Determined by Q_3+1.5(Q_3-Q_1)

Since there are no suspected outliers in the subsample of n=10 participants, the mean and standard deviation are the most appropriate statistics to summarize average values and dispersion, respectively, of each of these characteristics.

Outliers (cont..)

- For clarity, we have so far used a very small subset of the Framingham Offspring Cohort to illustrate calculations of summary statistics and determination of outliers. For your interest, Table 3 displays the means, standard deviations, medians, quartiles and interquartile ranges for each of the continuous variable displayed in Table 1 in the full sample (n=3,539) of participants who attended the seventh examination of the Framingham Offspring Study.

Table 3-Summary Statistics on Sample of (n=3,539) Participants

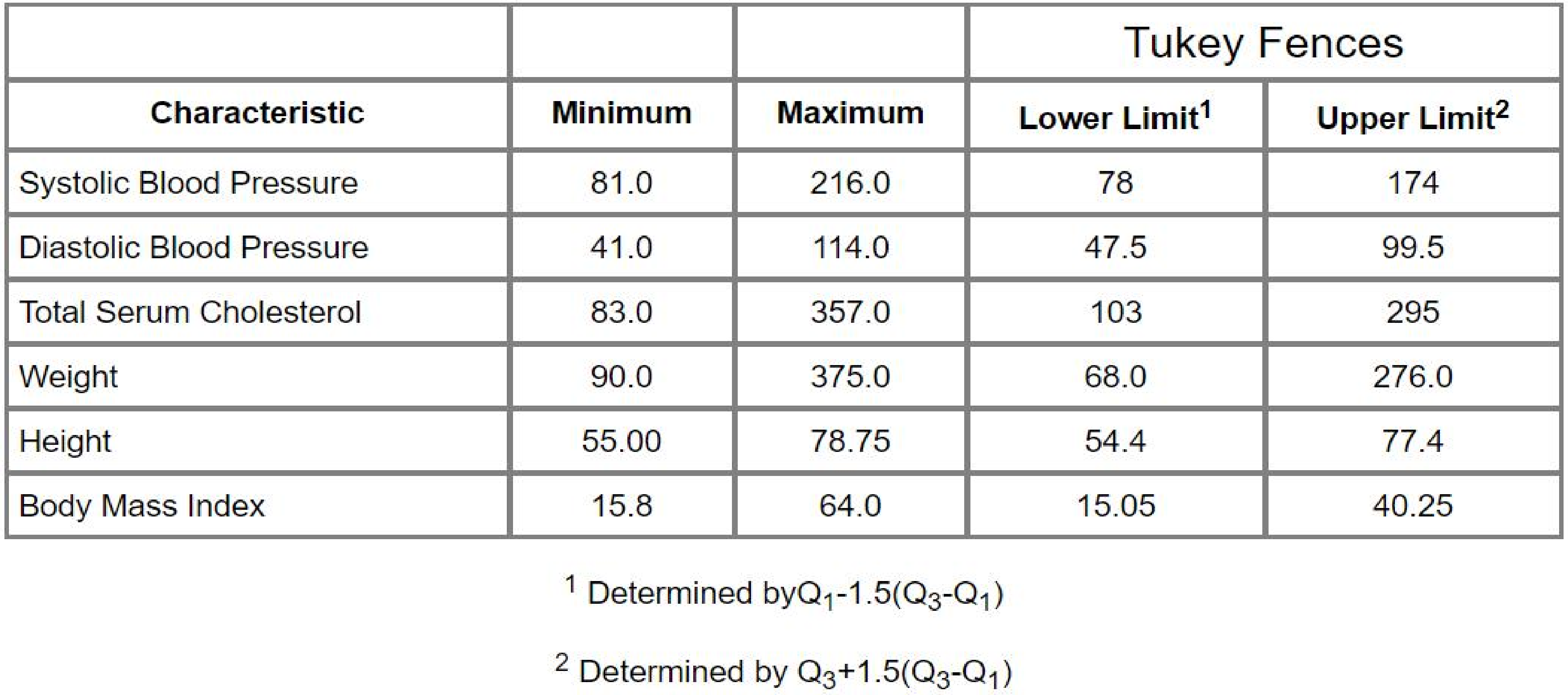


Characteristic	Mean $\bar{X}$	Standard Deviation (s)	Median	Q1	Q3	IQR
Systolic Blood Pressure	127.3	19.0	125.0	114.0	138.0	24.0
Diastolic Blood Pressure	74.0	9.9	74.0	67.0	80.0	13.0
Total Serum Cholesterol	200.3	36.8	198.0	175.0	223.0	48.0
Weight	174.4	38.7	170.0	146.0	198.0	52.0
Height	65.957	3.749	65.750	63.000	68.750	5.75
Body Mass Index	28.15	5.32	27.40	24.5	30.8	6.3

Outliers (cont..)

- Table 4 displays the observed minimum and maximum values along with the limits to determine outliers using the quartile rule for each of the variables in the full sample (n=3,539).

Table 4 - Limits for Assessing Outliers in Characteristics Presented in Table 3



			Tukey Fences	
Characteristic	**Minimum**	**Maximum**	**Lower Limit**[1]	**Upper Limit**[2]
Systolic Blood Pressure	81.0	216.0	78	174
Diastolic Blood Pressure	41.0	114.0	47.5	99.5
Total Serum Cholesterol	83.0	357.0	103	295
Weight	90.0	375.0	68.0	276.0
Height	55.00	78.75	54.4	77.4
Body Mass Index	15.8	64.0	15.05	40.25

[1] Determined byQ_1-1.5(Q_3-Q_1)

[2] Determined by Q_3+1.5(Q_3-Q_1)

Are there outliers in any of the variables?

Which statistics are most appropriate to summarize the average or typical values and the dispersion for each variable?

Outliers (cont..)

Observations on example……

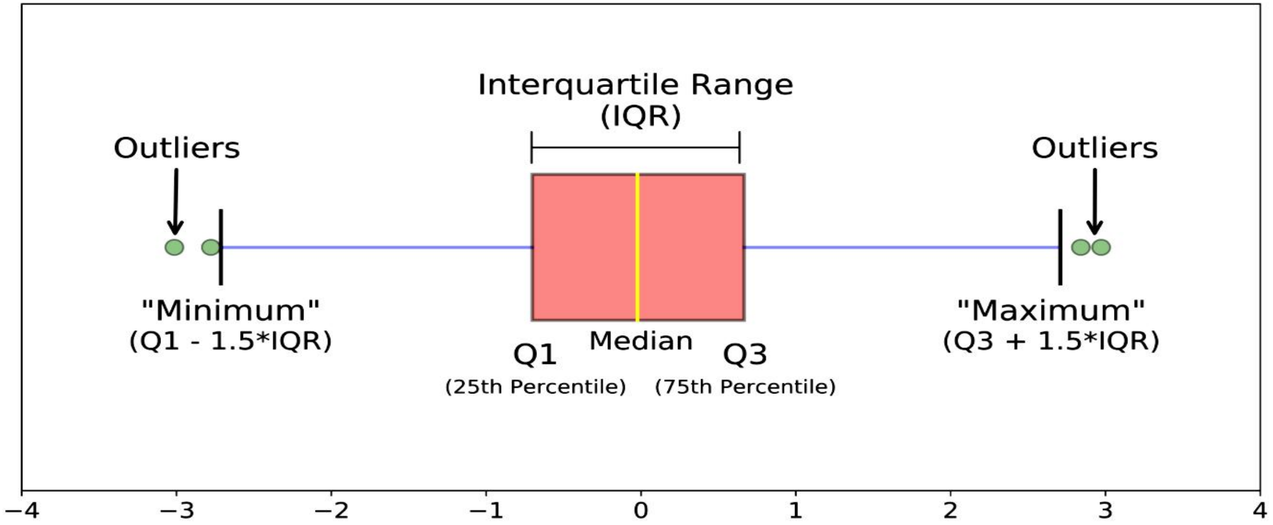
- In the full sample, each of the characteristics has outliers on the upper end of the distribution as the maximum values exceed the upper limits in each case. There are also outliers on the low end for diastolic blood pressure and total cholesterol, since the minimums are below the lower limits.
- For some of these characteristics, the difference between the upper limit and the maximum (or the lower limit and the minimum) is small (e.g., height, systolic and diastolic blood pressures), while for others (e.g., total cholesterol, weight and body mass index) the difference is much larger. This method for determining outliers is a popular one but not generally applied as a hard and fast rule. In this application it would be reasonable to present means and standard deviations for height, systolic and diastolic blood pressures and medians and interquartile ranges for total cholesterol, weight and body mass index.

Outliers (cont..)

Boxplot Analysis

Box plots are a simple way to visualize data through quantiles and detect outliers. **A boxplot incorporates the five-number summary as follows: Minimum, Q1, Median, Q3, Maximum**

- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR. IQR(Interquartile Range) is the basic mathematics behind boxplots.
  - The median is marked by a line within the box
  - Two lines (called **whiskers**) outside the box extend to the smallest (Minimum) and largest (Maximum) observations. The top and bottom whiskers can be understood as the boundaries of data, and any data lying outside it will be an outlier.
  - **Outliers:** points beyond a specified outlier threshold, plotted individually



Statistical detection :*Removing and modifying the outliers using statistical detection techniques is a widely followed method.*

- *Z-Score*
- *Density-based spatial clustering*
- *Regression Analysis*
- *Proximity-based clustering*
- *IQR Scores*

Outliers (cont..)

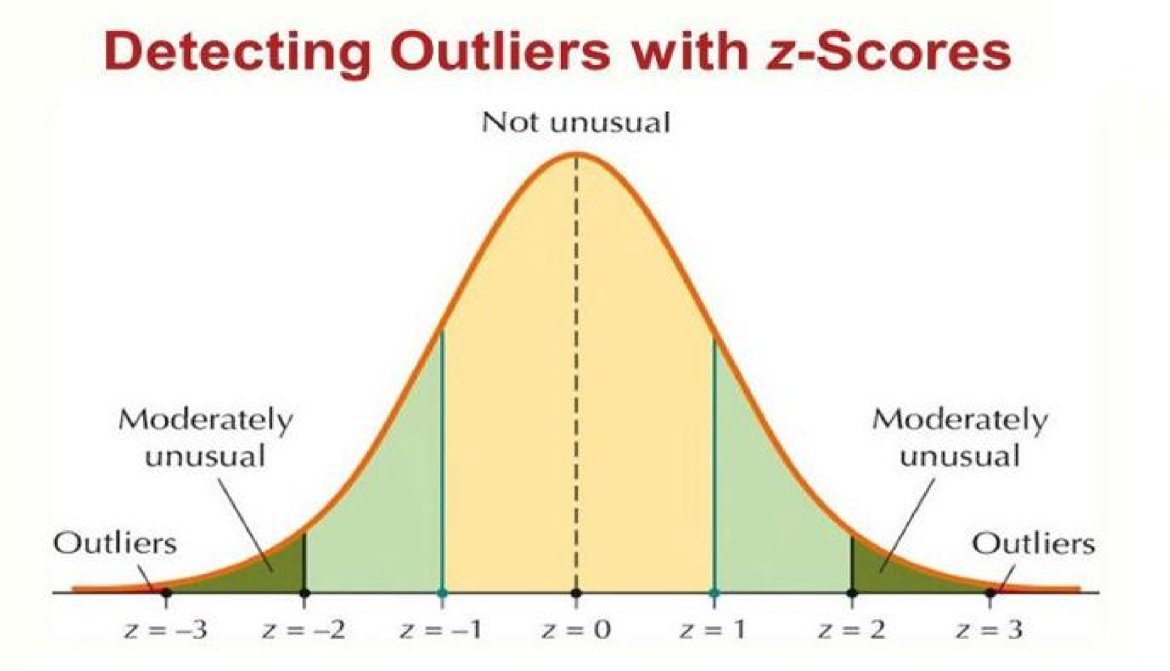
- **Z-Score** for Outlier Detection.

While data points are referred to as x in a normal distribution, they are called z or z scores in the z distribution. A z score is a standard score that tells you how many standard deviations away from the mean an individual value (x) lies:

- **A positive z score means that your x value is greater than the mean.**
- **A negative z score means that your x value is less than the mean.**
- **A z score of zero means that your x value is equal to the mean.**

$$Z\ Score = \frac{(\text{Observation} - \text{Mean})}{\text{Standard Deviation}}$$

$$Z\ Score = \frac{X - \mu}{\sigma}$$



Detecting Outliers with z-Scores

Outliers and Missing values

Missing Values Detection and Handling.

- Related to Pre-processing.
- Consumes most of the time in Data Analytics.
- Handling missing values is one of the challenges of data analysis

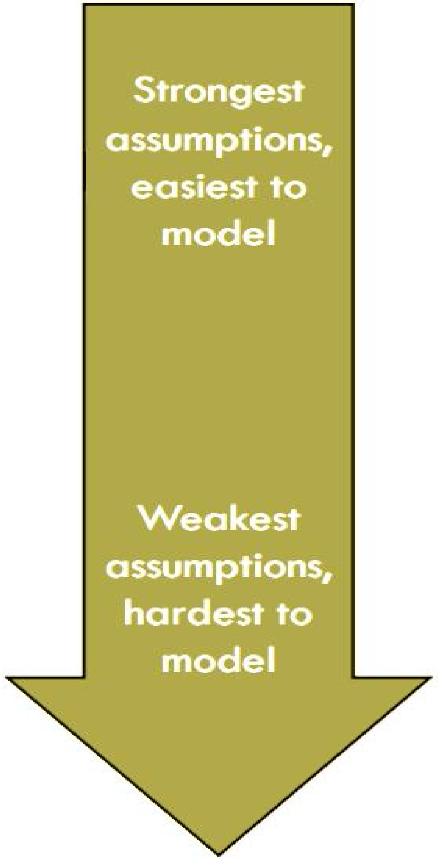
- **Reasons for Missing Values.**
  - ➢ Improper maintenance of past data.
  - ➢ Observations are not recorded for certain fields due to some reasons.
  - ➢ Failure in recording the values due to human error.
  - ➢ The user has not provided the values intentionally.
  - ➢ Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
  - ➢ Information is not collected (e.g., people decline to give their age and weight)

- **Handling missing values**
  - ➢ Eliminate data objects or variables
  - ➢ Estimate missing values
  - ➢ Ignore the missing value during analysis
  - ➢ Replace with all possible values (weighted by their probabilities)

Types of Missing Values

Some definitions are based on representation: Missing data is the lack of a recorded answer for a particular field.

- **Missing completely at random (MCAR)**
- **Missing at Random (MAR)**
- **Missing Not at Random (MNAR)**



Strongest assumptions, easiest to model

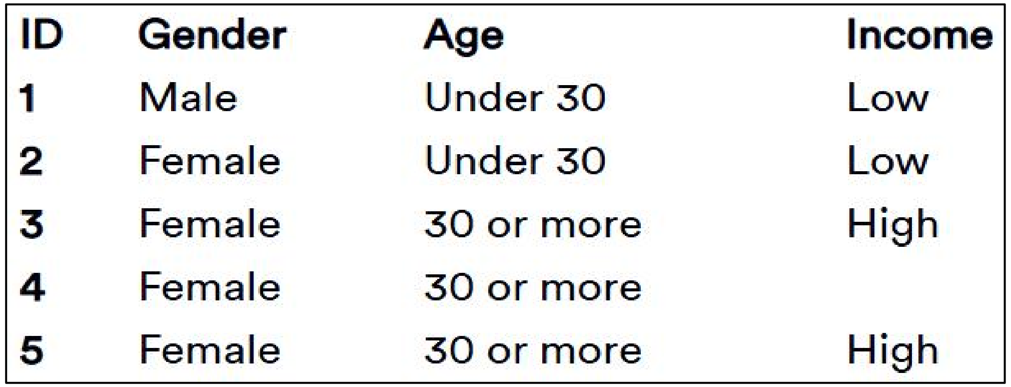
Weakest assumptions, hardest to model

Types of Missing Values

Missing Completely at Random (MCAR)

- Missingness of a value is independent of attributes
  - Fill in values based on the attribute
  - Analysis may be unbiased overall
- The missingness on the variable is completely unsystematic.

Example when we take a random sample of a population, where each member has the same chance of being included in the sample.



ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

When we make this assumption, we are assuming that whether or not the person has missing data is completely unrelated to the other information in the data.

When data is missing completely at random, it means that we can undertake analyses using only observations that have complete data (provided we have enough of such observations).

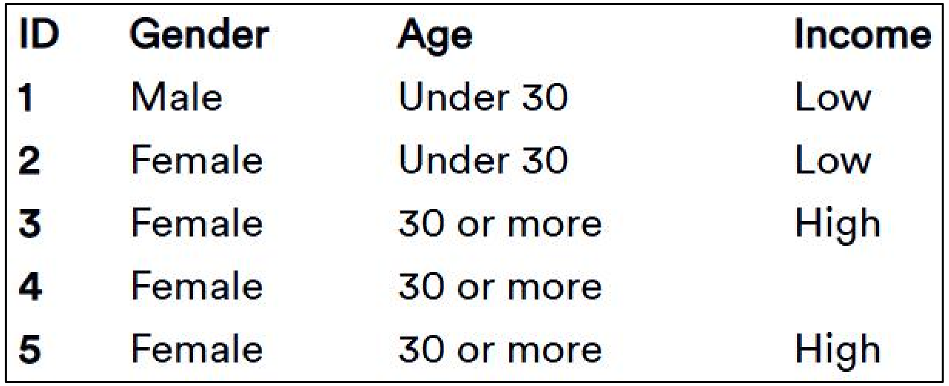
Types of Missing Values

Missing at Random (MAR)

- Missingness is related to other variables
- Fill in values based other values
- Almost always produces a bias in the analysis

Example of MAR is when we take a sample from a population, where the probability to be included depends on some known property.

A simple predictive model is that income can be predicted based on gender and age. Looking at the table, we note that our missing value is for a Female aged 30 or more, and observations says the other females aged 30 or more have a High income. As a result, we can predict that the missing value should be High.



ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

There is a systematic relationship between the inclination of missing values and the observed data, but not the missing data. All that is required is a probabilistic relationship

Types of Missing Values (cont..)

Missing not at Random (MNAR) - Nonignorable

- Missingness is related to unobserved measurements
- When the missing values on a variable are related to the values of that variable itself, even after controlling for other variables.

MNAR means that the probability of being missing varies for reasons that are unknown to us.

Data was obtained from 31 women, of whom 14 were located six months later. Of these, three had exited from homelessness, so the estimated proportion to have exited homelessness is 3/14 = 21%. As there is no data for the 17 women who could not be contacted, it is possible that none, some, or all of these 17 may have exited from homelessness. This means that potentially the proportion to have exited from homelessness in the sample is between 3/31 = 10% and 20/31 = 65%. As a result, reporting 21% as being the correct result is misleading. In this example the missing data is nonignorable.

Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Types of Missing Values (cont..)

Can Formalize these Definitions..

Let X represent a matrix of the data we “expect” to have; X = {X_o,X_m} where Xo is the observed data and Xm the missing data.

Let’s define R as a matrix with the same dimensions as X where Ri,j = 1 if the datum is missing, and 0 otherwise.

1. MCAR: $P(R \mid X_o , X_m) = P(R)$

2. MAR: $P(R \mid X_o , X_m) = P(R \mid X_o)$

3. MNAR: No simplification.

Finding Relationships among Variables

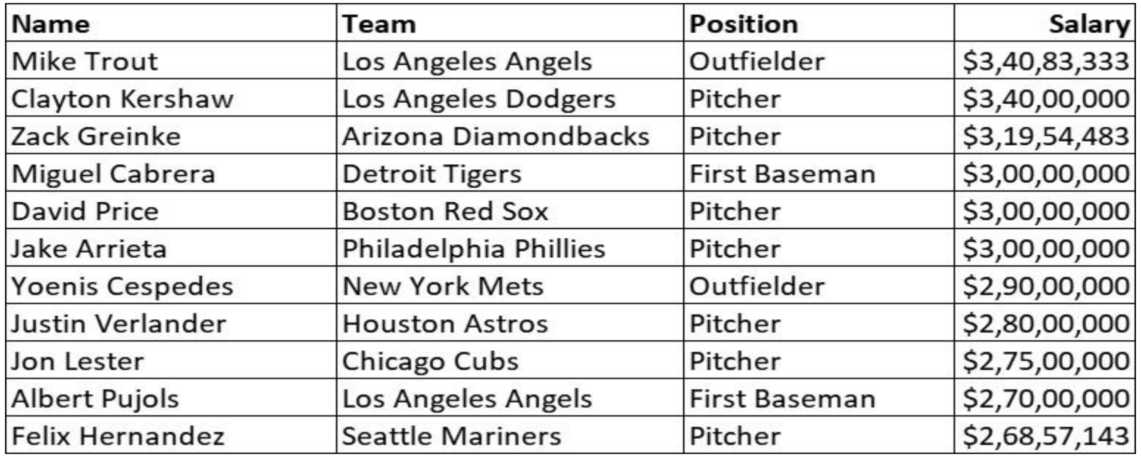
This is an important first step in any exploratory data analysis.

- To look closely at variables one at a time, but it is almost never the last step.
- The primary interest is usually in relationships between variables.

For a variable such as baseball salary, the entire focus was on how salaries were distributed over some range.

It is natural to ask what drives baseball salaries.

- *Does it depend on qualitative factors, such as;*
  - ❖ *Player's team or position?*
- *Does it depend on quantitative factors, such as;*
  - ❖ *Number of hits the player gets or the number of strikeouts?*



Name	Team	Position	Salary
Mike Trout	Los Angeles Angels	Outfielder	$3,40,83,333
Clayton Kershaw	Los Angeles Dodgers	Pitcher	$3,40,00,000
Zack Greinke	Arizona Diamondbacks	Pitcher	$3,19,54,483
Miguel Cabrera	Detroit Tigers	First Baseman	$3,00,00,000
David Price	Boston Red Sox	Pitcher	$3,00,00,000
Jake Arrieta	Philadelphia Phillies	Pitcher	$3,00,00,000
Yoenis Cespedes	New York Mets	Outfielder	$2,90,00,000
Justin Verlander	Houston Astros	Pitcher	$2,80,00,000
Jon Lester	Chicago Cubs	Pitcher	$2,75,00,000
Albert Pujols	Los Angeles Angels	First Baseman	$2,70,00,000
Felix Hernandez	Seattle Mariners	Pitcher	$2,68,57,143

To answer these questions, you have to examine relationships between various variables and salary.

Types of Relationships among Variables

a. **Categorical vs Categorical**
b. **Categorical vs Numerical**
c. **Numerical vs Numerical**

Relationships Among Categorical Variables

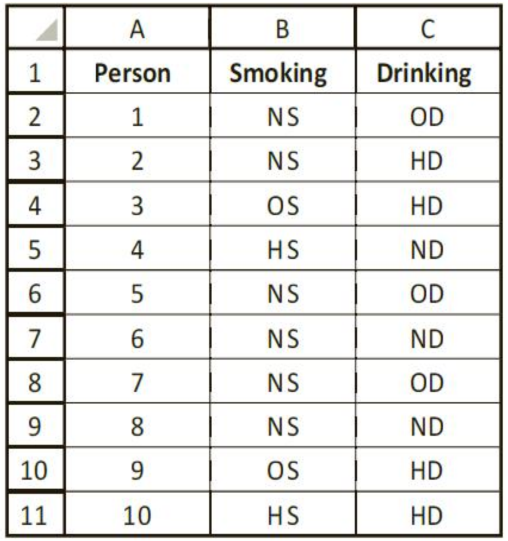
(Categorical vs Categorical)

Consider a data set with at least two categorical variables, Smoking and Drinking.

Smoking	Drinking
Non Smoker (NS)	Non Drinker (ND)
Occasional Smoker (OS)	Occasional Drinker (OD)
Heavy Smoker (HS)	Heavy Drinker (HD)

Do the data indicate that smoking and drinking habits are related? For example,

- *Do nondrinkers tend to be nonsmokers?*
- *Do heavy smokers tend to be heavy drinkers?*



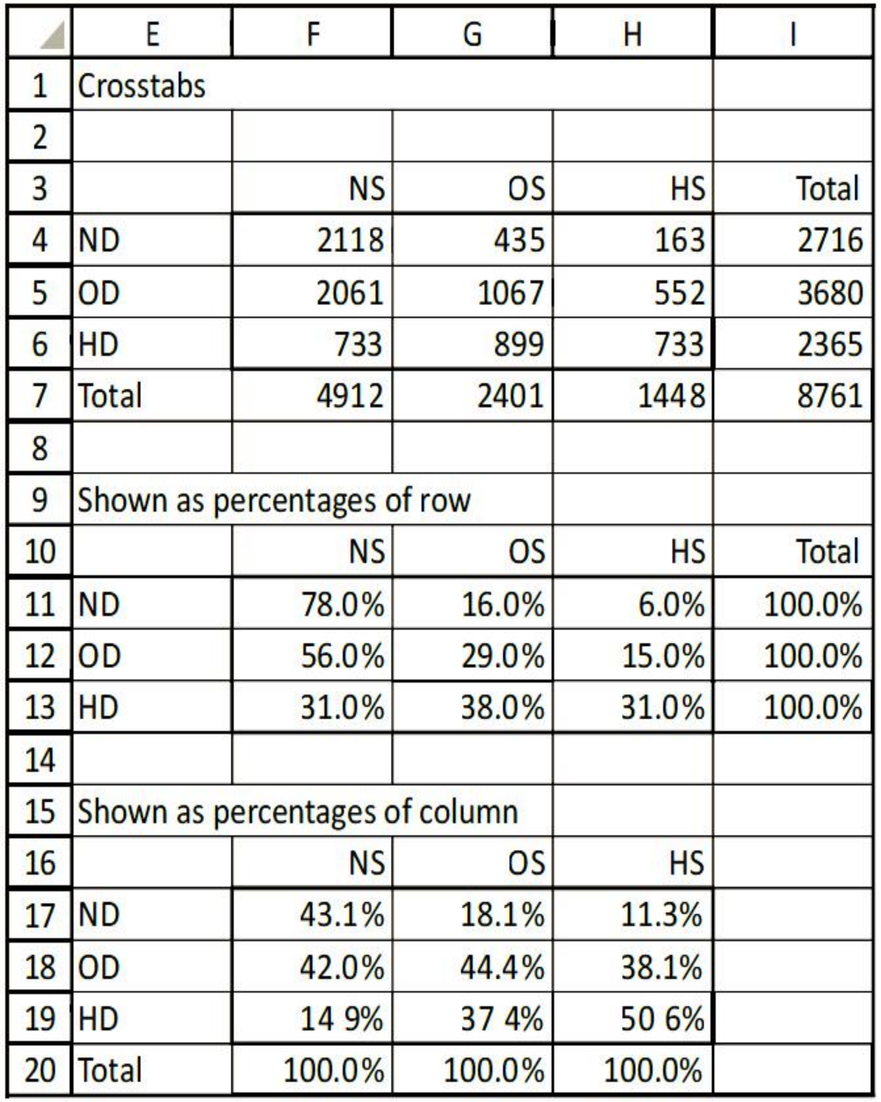
	A	B	C
1	Person	Smoking	Drinking
2	1	NS	OD
3	2	NS	HD
4	3	OS	HD
5	4	HS	ND
6	5	NS	OD
7	6	NS	ND
8	7	NS	OD
9	8	NS	ND
10	9	OS	HD
11	10	HS	HD

- The most meaningful way to describe a categorical variable is with counts, possibly expressed as percentages of totals, and corresponding
- charts of the counts.
- We can find the counts of the categories of either variable separately, and more
- importantly, we can find counts of the joint categories of the two variables, such as the
- count of all nondrinkers who are also nonsmokers.

It is customary to display all such counts in a table called a crosstabs (for crosstabulations). This is also sometimes called a contingency table.

Relationships Among Categorical Variables (cont..)

(Categorical vs Categorical)



	E	F	G	H	I
1	Crosstabs				
2					
3		NS	OS	HS	Total
4	ND	2118	435	163	2716
5	OD	2061	1067	552	3680
6	HD	733	899	733	2365
7	Total	4912	2401	1448	8761
8					
9	Shown as percentages of row				
10		NS	OS	HS	Total
11	ND	78.0%	16.0%	6.0%	100.0%
12	OD	56.0%	29.0%	15.0%	100.0%
13	HD	31.0%	38.0%	31.0%	100.0%
14					
15	Shown as percentages of column				
16		NS	OS	HS	
17	ND	43.1%	18.1%	11.3%	
18	OD	42.0%	44.4%	38.1%	
19	HD	14 9%	37 4%	50 6%	
20	Total	100.0%	100.0%	100.0%	

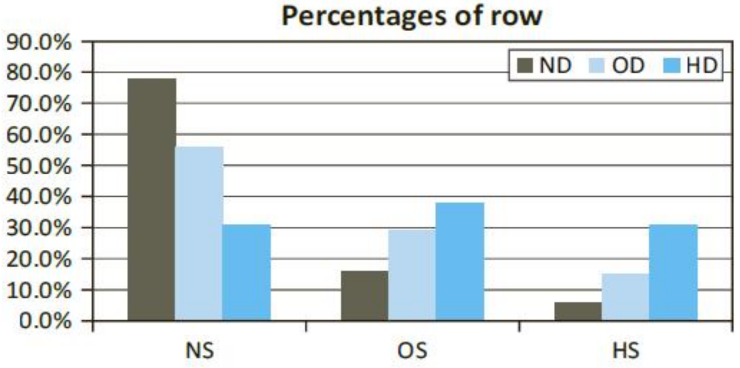
Do the data indicate that smoking and drinking habits are related? For example,

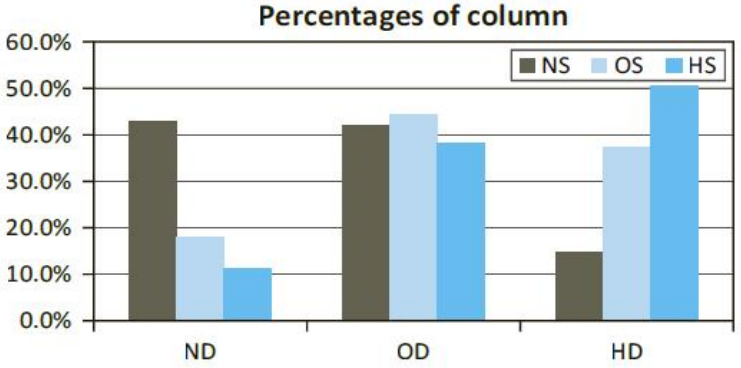
- *Do nondrinkers tend to be nonsmokers?*
- *Do heavy smokers tend to be heavy drinkers?*

The 1st two arguments are for the condition on smoking; the 2nd two are for the condition on drinking.

- ➢ *You can then sum across rows and down columns to get the totals.*
- ➢ *It is useful to express the counts as percentages of row in the middle table and as percentages of column in the bottom table.*
- ➢ *The latter two tables indicate, in complementary ways, **that there is definitely a relationship between smoking and drinking.***

These tables indicate that smoking and drinking habits tend to go with one another. These tendencies are reinforced by the column charts of the two percentage tables





Relationships Among Categorical & Numerical Variables

(Categorical vs Numerical)

It describes a very common situation where the goal is to break down a numerical variable such as salary by a categorical variable such as gender.

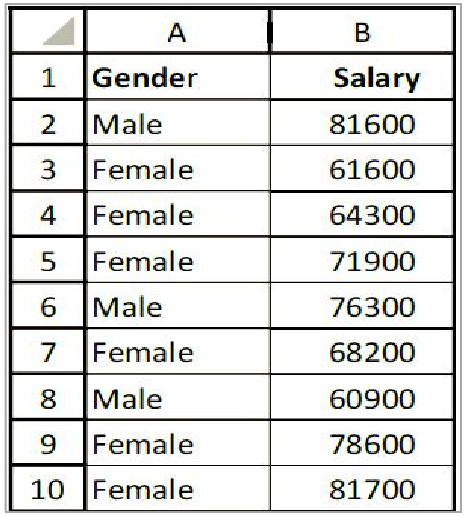
- This general problem, typically referred to as the **comparison problem**, is one of the most important problems in data analysis.
- It occurs whenever you want to compare a numerical measure across two or more subpopulations. Here are some examples:
  - ➢ *The subpopulations are males and females, and the numerical measure is salary.*
  - ➢ *The subpopulations are different regions of the country, and the numerical measure is the cost of living.*
  - ➢ *The subpopulations are different days of the week, and the numerical measure is the number of customers going to a particular fast-food chain.*
  - ➢ *The subpopulations are different machines in a manufacturing plant, and the numerical measure is the number of defective parts produced per day.*
  - ➢ *The subpopulations are patients who have taken a new drug and those who have taken a placebo, and the numerical measure is the recovery rate from a particular disease.*
  - ➢ *The subpopulations are undergraduates with various majors (business, English, history, and so on), and the numerical measure is the starting salary after graduating.*

Relationships Among Categorical & Numerical Variables (cont..)

(Categorical vs Numerical)

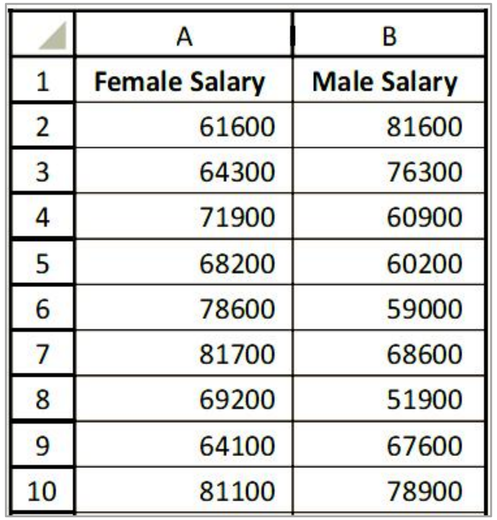
There are two possible data formats you will see,

stacked



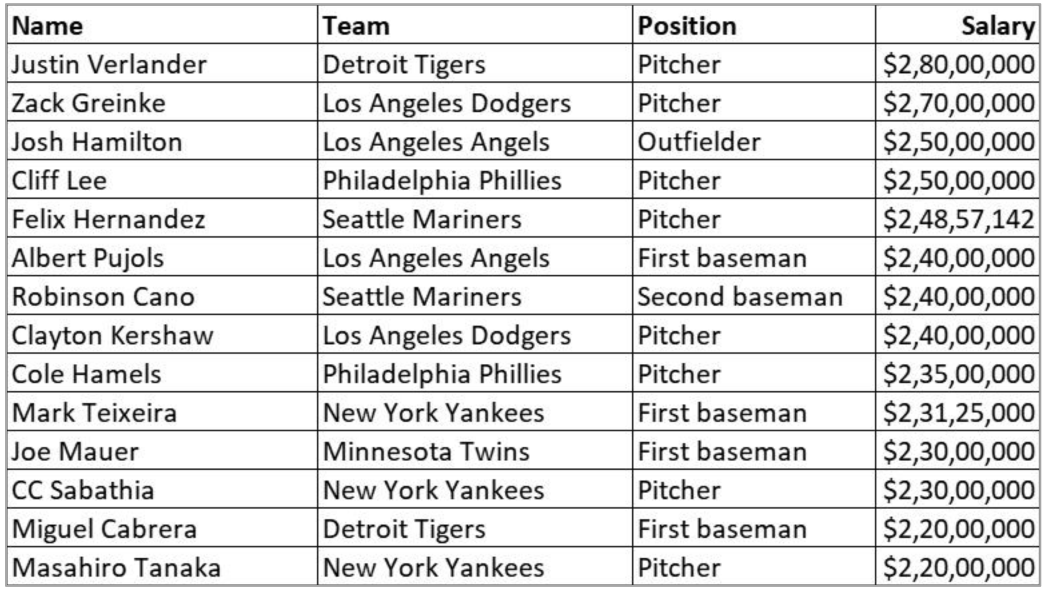
	A	B
1	**Gender**	**Salary**
2	Male	81600
3	Female	61600
4	Female	64300
5	Female	71900
6	Male	76300
7	Female	68200
8	Male	60900
9	Female	78600
10	Female	81700

unstacked



	A	B
1	**Female Salary**	**Male Salary**
2	61600	81600
3	64300	76300
4	71900	60900
5	68200	60200
6	78600	59000
7	81700	68600
8	69200	51900
9	64100	67600
10	81100	78900

The data are **stacked** if there are two "long" variables, Gender and Salary, as indicated in Figure. Occasionally will see data in **unstacked** format. (Note that both tables list exactly the same data)



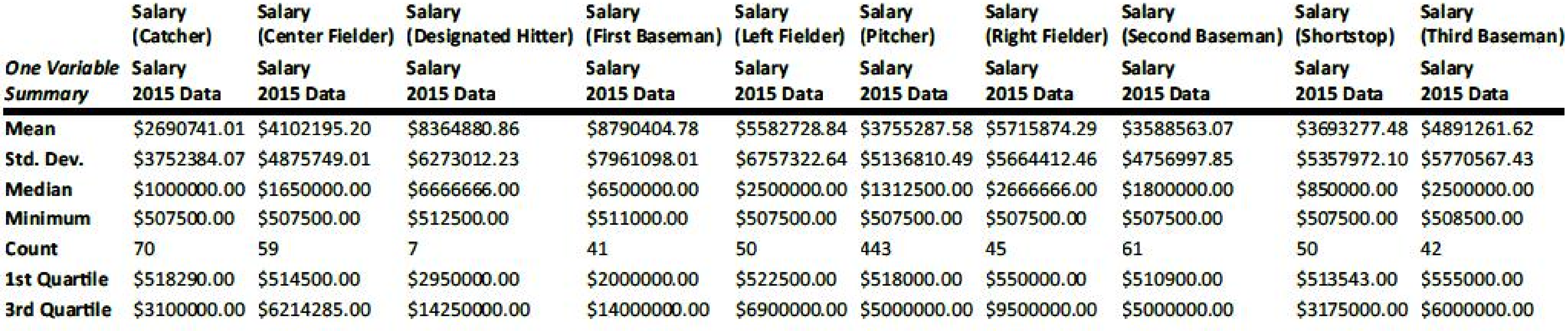
Name	Team	Position	Salary
Justin Verlander	Detroit Tigers	Pitcher	$2,80,00,000
Zack Greinke	Los Angeles Dodgers	Pitcher	$2,70,00,000
Josh Hamilton	Los Angeles Angels	Outfielder	$2,50,00,000
Cliff Lee	Philadelphia Phillies	Pitcher	$2,50,00,000
Felix Hernandez	Seattle Mariners	Pitcher	$2,48,57,142
Albert Pujols	Los Angeles Angels	First baseman	$2,40,00,000
Robinson Cano	Seattle Mariners	Second baseman	$2,40,00,000
Clayton Kershaw	Los Angeles Dodgers	Pitcher	$2,40,00,000
Cole Hamels	Philadelphia Phillies	Pitcher	$2,35,00,000
Mark Teixeira	New York Yankees	First baseman	$2,31,25,000
Joe Mauer	Minnesota Twins	First baseman	$2,30,00,000
CC Sabathia	New York Yankees	Pitcher	$2,30,00,000
Miguel Cabrera	Detroit Tigers	First baseman	$2,20,00,000
Masahiro Tanaka	New York Yankees	Pitcher	$2,20,00,000

Data of baseball salaries

- Do pitchers (or any other positions) earn more than others?
- Does one league pay more than the other, or do any divisions pay more than others?
- How does the notoriously high Yankees payroll compare to the others?

Relationships Among Categorical & Numerical Variables (cont..)

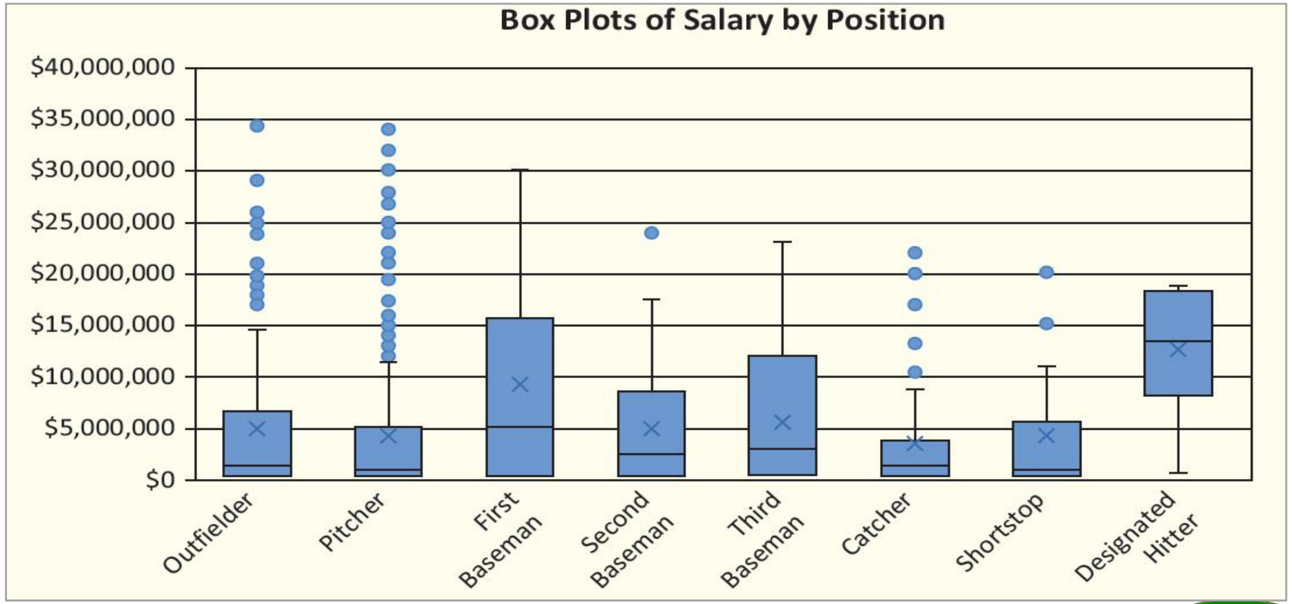
(Categorical vs Numerical)



	Salary (Catcher)	Salary (Center Fielder)	Salary (Designated Hitter)	Salary (First Baseman)	Salary (Left Fielder)	Salary (Pitcher)	Salary (Right Fielder)	Salary (Second Baseman)	Salary (Shortstop)	Salary (Third Baseman)
One Variable Summary	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data	Salary 2015 Data
Mean	$2690741.01	$4102195.20	$8364880.86	$8790404.78	$5582728.84	$3755287.58	$5715874.29	$3588563.07	$3693277.48	$4891261.62
Std. Dev.	$3752384.07	$4875749.01	$6273012.23	$7961098.01	$6757322.64	$5136810.49	$5664412.46	$4756997.85	$5357972.10	$5770567.43
Median	$1000000.00	$1650000.00	$6666666.00	$6500000.00	$2500000.00	$1312500.00	$2666666.00	$1800000.00	$850000.00	$2500000.00
Minimum	$507500.00	$507500.00	$512500.00	$511000.00	$507500.00	$507500.00	$507500.00	$507500.00	$507500.00	$508500.00
Count	70	59	7	41	50	443	45	61	50	42
1st Quartile	$518290.00	$514500.00	$2950000.00	$2000000.00	$522500.00	$518000.00	$550000.00	$510900.00	$513543.00	$555000.00
3rd Quartile	$3100000.00	$6214285.00	$14250000.00	$14000000.00	$6900000.00	$5000000.00	$9500000.00	$5000000.00	$3175000.00	$6000000.00

This table lists each of the requested summary measures for each of the nine positions in the data set.

If you want to see salaries broken down by team or any other categorical variable, you can easily run this analysis again and choose a different Cat variable.



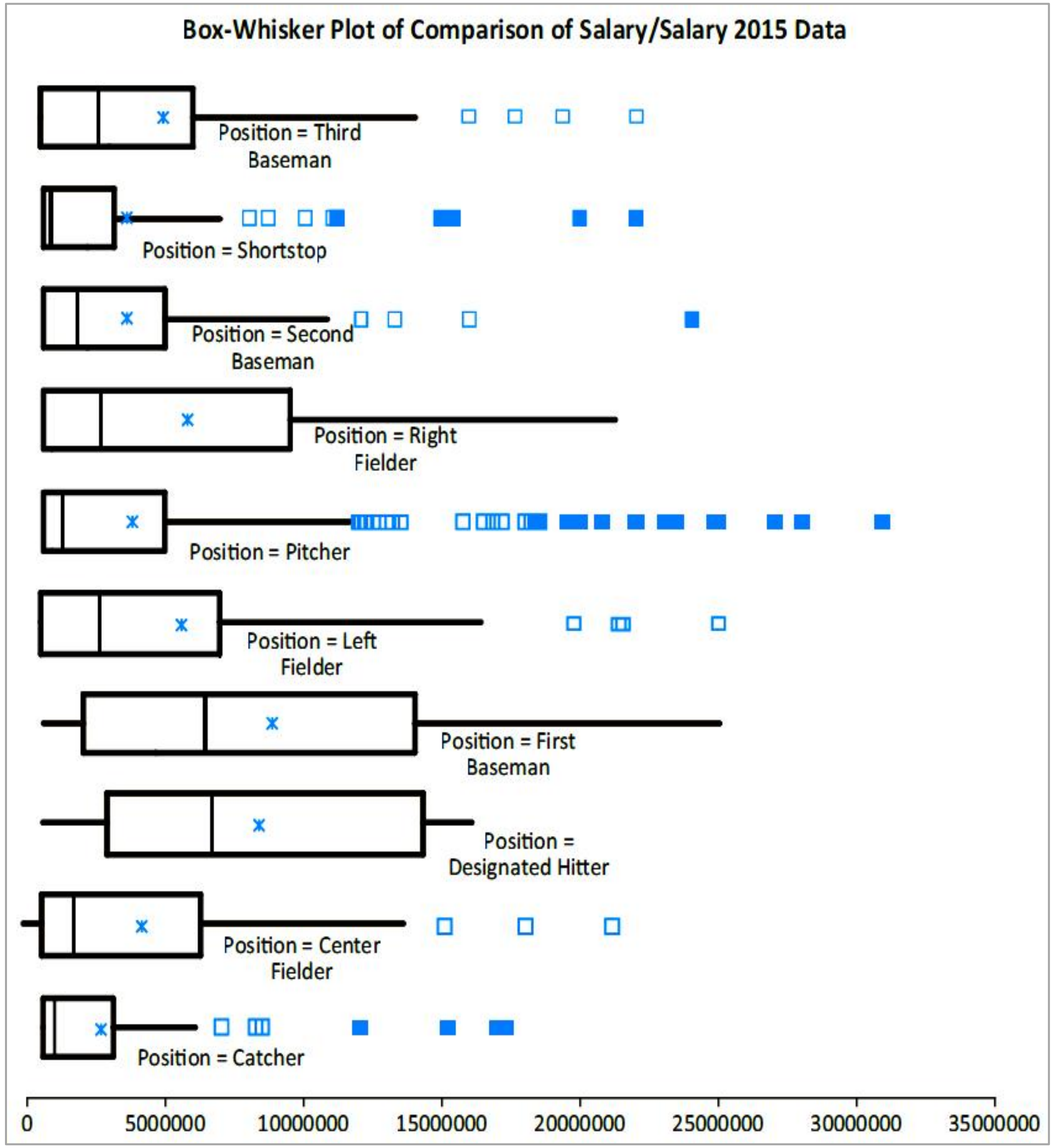
Relationships Among Numerical & Numerical Variables (cont..)

(Numerical vs Numerical)

- To study relationships among numeric variables, a new type of chart, called a ***scatterplot***, and two new summary measures, **correlation** and **covariance**, are used.
- These measures can be applied to any variables that are displayed numerically.
- However, they are appropriate only for truly numerical variables, not for categorical variables that have been coded numerically.

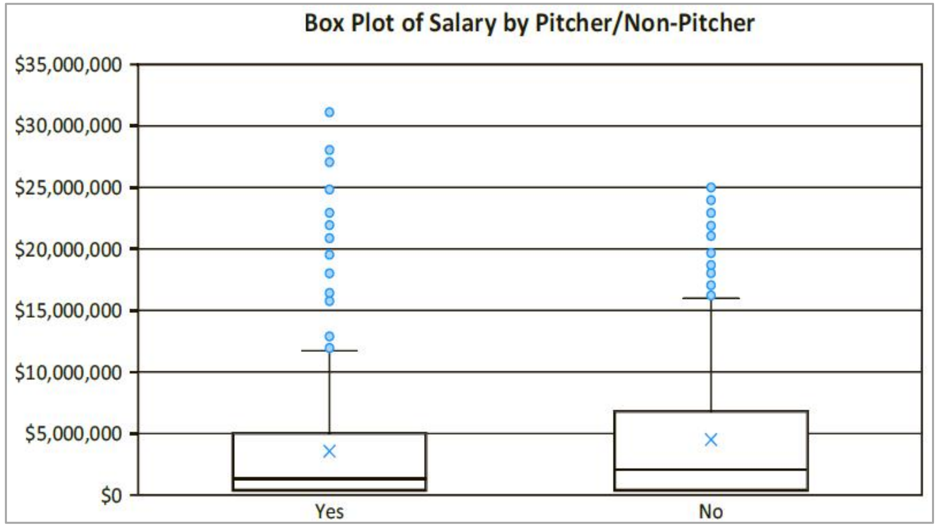
Relationships Among Categorical & Numerical Variables (cont..)

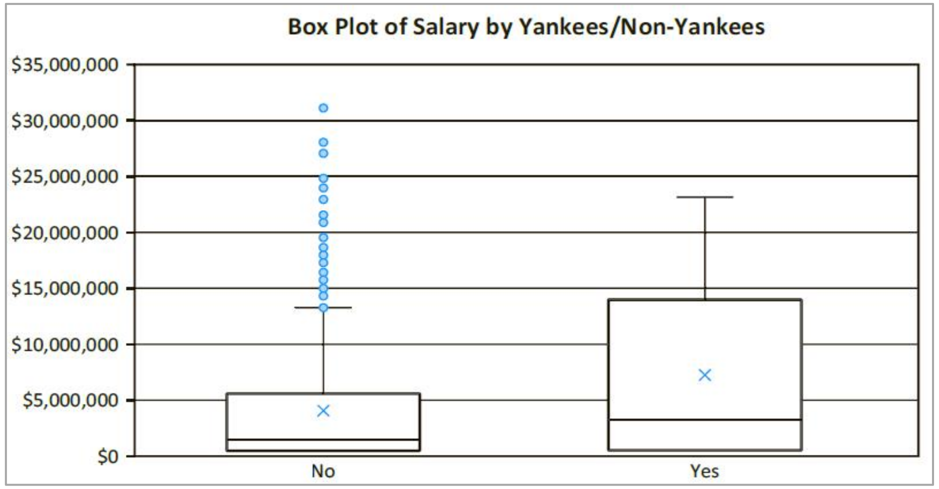
(Categorical vs Numerical)



From these box plots, we can conclude the following:

- Pitchers make somewhat less than other players, although there are many outliers in each group.
- The Yankees payroll is indeed much larger than the payrolls for the rest of the teams. In fact, it is so large that its stars' salaries aren't even considered outliers relative to the rest of the team.





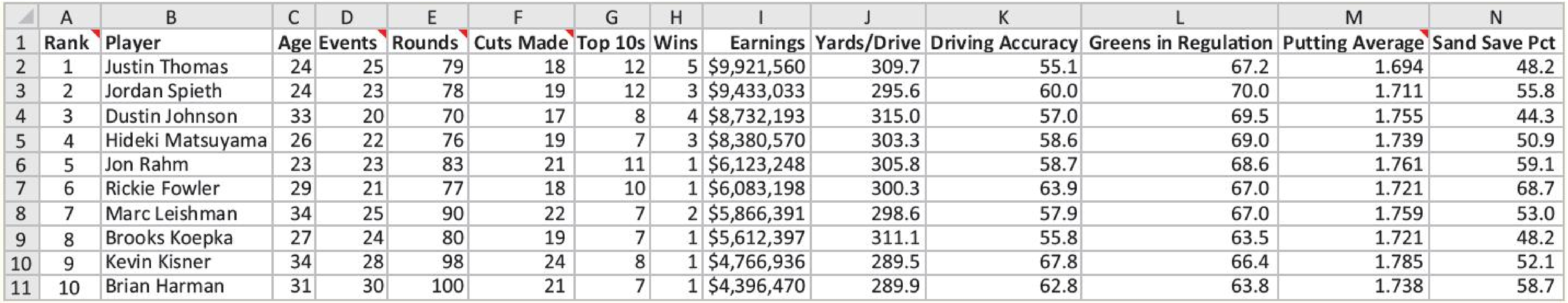
These side-by-side box plots are so easy to obtain, you can generate a lot of them to provide insights into the salary data.

Relationships Among Numerical & Numerical Variables (cont..)

(Scatterplot)

- A **scatterplot** is a scatter of points, where each point denotes the values of an observation for two selected variables.
  - It is a graphical method for detecting relationships between two numerical variables.
  - The two variables are often labeled generically as *X* and *Y*, so a scatterplot is sometimes called an **X-Y chart**.
  - The purpose of a scatterplot is to make a relationship (or the lack of it) apparent.

Data set includes an observation (Golf Stats) for each of the top 200 earners on the PGA Tour.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Rank	Player	Age	Events	Rounds	Cuts Made	Top 10s	Wins	Earnings	Yards/Drive	Driving Accuracy	Greens in Regulation	Putting Average	Sand Save Pct
2	1	Justin Thomas	24	25	79	18	12	5	$9,921,560	309.7	55.1	67.2	1.694	48.2
3	2	Jordan Spieth	24	23	78	19	12	3	$9,433,033	295.6	60.0	70.0	1.711	55.8
4	3	Dustin Johnson	33	20	70	17	8	4	$8,732,193	315.0	57.0	69.5	1.755	44.3
5	4	Hideki Matsuyama	26	22	76	19	7	3	$8,380,570	303.3	58.6	69.0	1.739	50.9
6	5	Jon Rahm	23	23	83	21	11	1	$6,123,248	305.8	58.7	68.6	1.761	59.1
7	6	Rickie Fowler	29	21	77	18	10	1	$6,083,198	300.3	63.9	67.0	1.721	68.7
8	7	Marc Leishman	34	25	90	22	7	2	$5,866,391	298.6	57.9	67.0	1.759	53.0
9	8	Brooks Koepka	27	24	80	19	7	1	$5,612,397	311.1	55.8	63.5	1.721	48.2
10	9	Kevin Kisner	34	28	98	24	8	1	$4,766,936	289.5	67.8	66.4	1.785	52.1
11	10	Brian Harman	31	30	100	21	7	1	$4,396,470	289.9	62.8	63.8	1.738	58.7

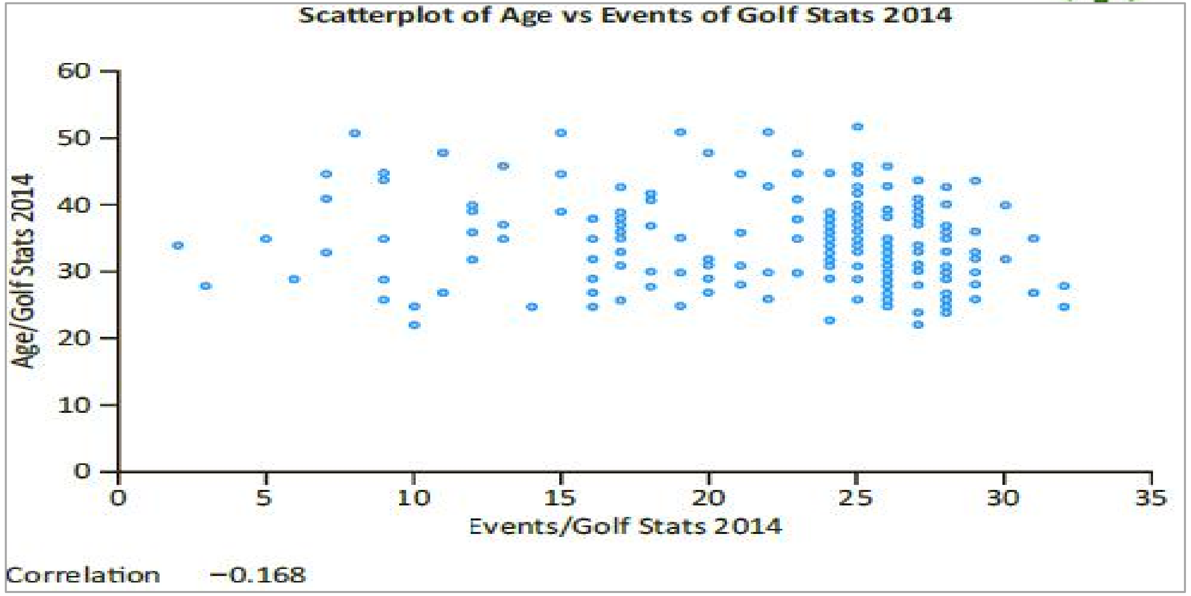
Relationships Among Numerical & Numerical Variables (cont..)

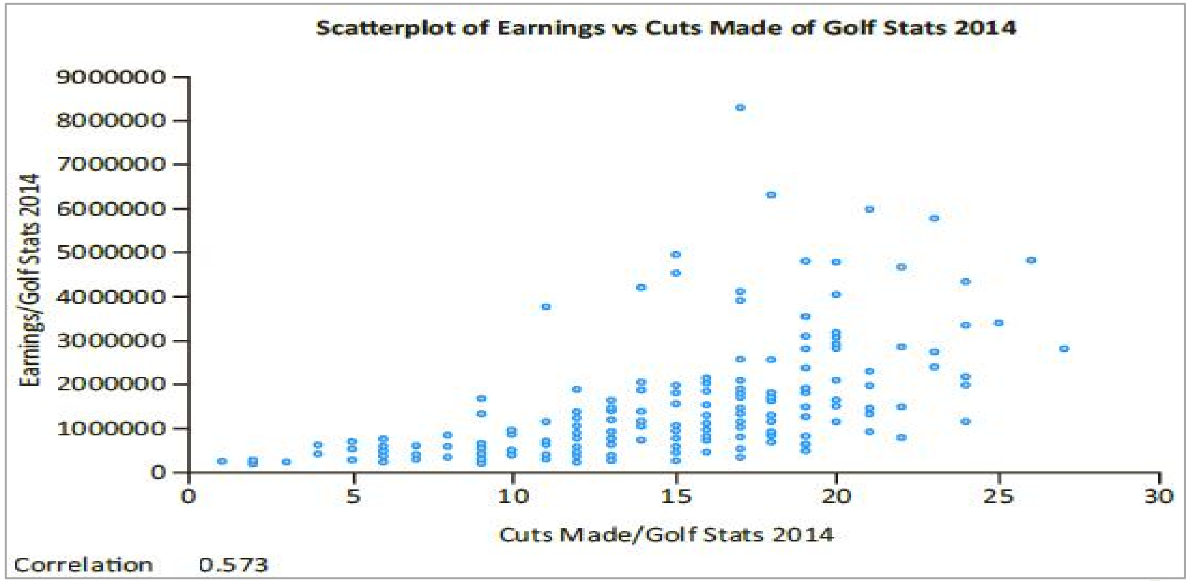
(Scatterplot)

This example is typical in that there are many numerical variables, and it is up to you to search for possible relationships. A good first step is to ask some interesting questions and then try to answer them with scatterplots. For example,

- Do younger players play more events?
- Are earnings related to age?
- Which is related most strongly to earnings: driving, putting, or greens in regulation?
- Do the answers to these questions remain the same from year to year?

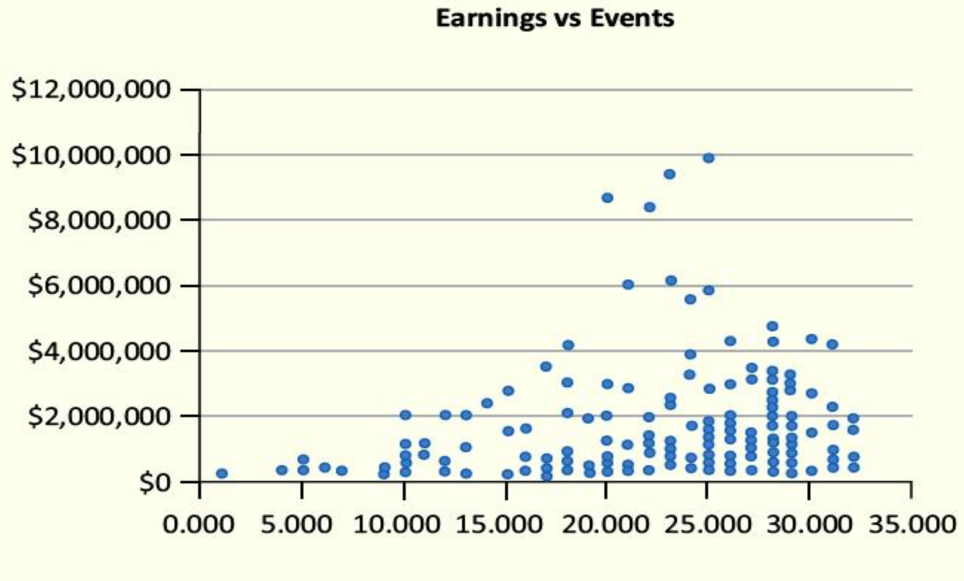
This example is all about exploring the data,

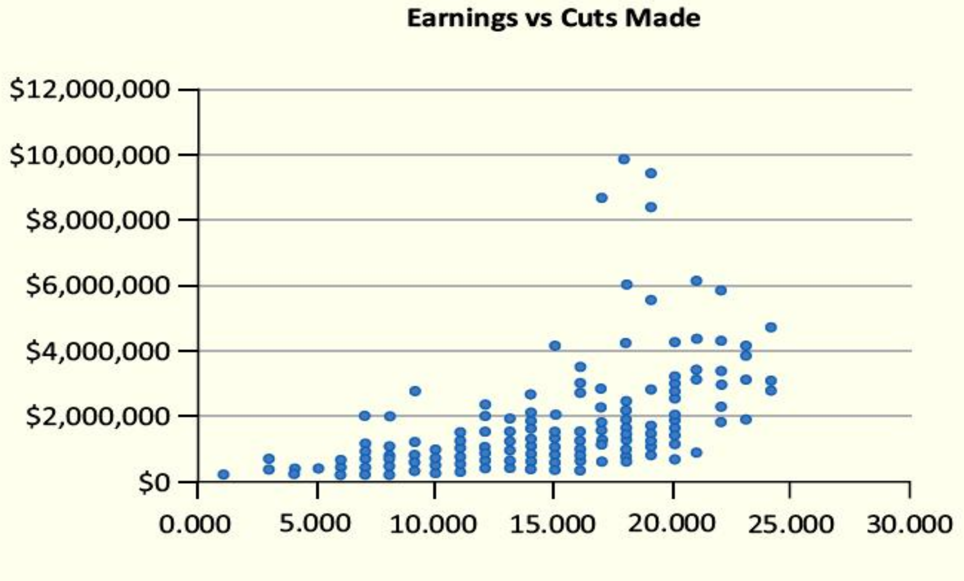


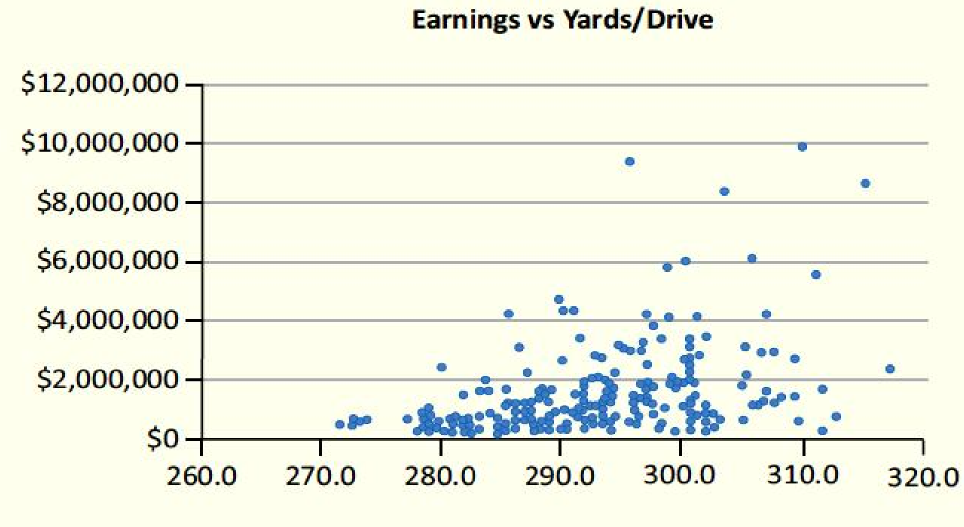


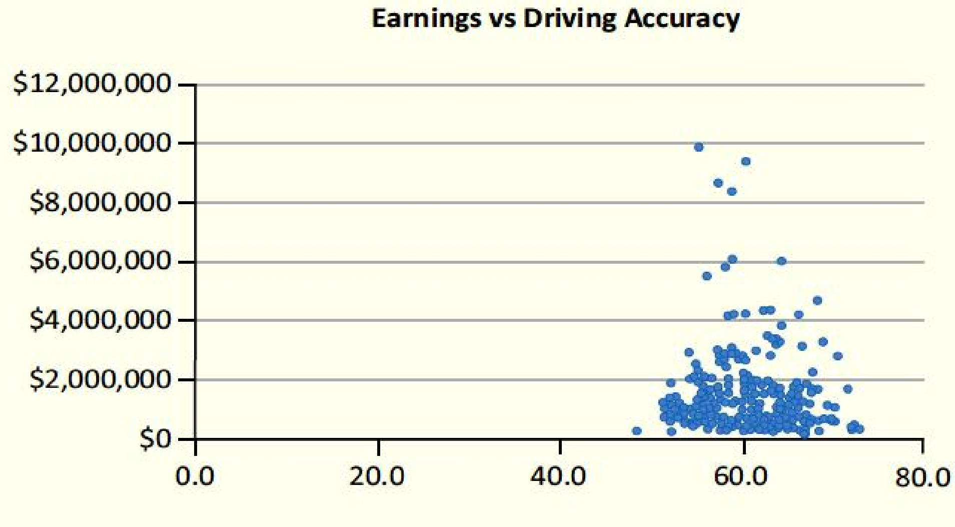
Relationships Among Numerical & Numerical Variables (cont..)

(Scatterplot)





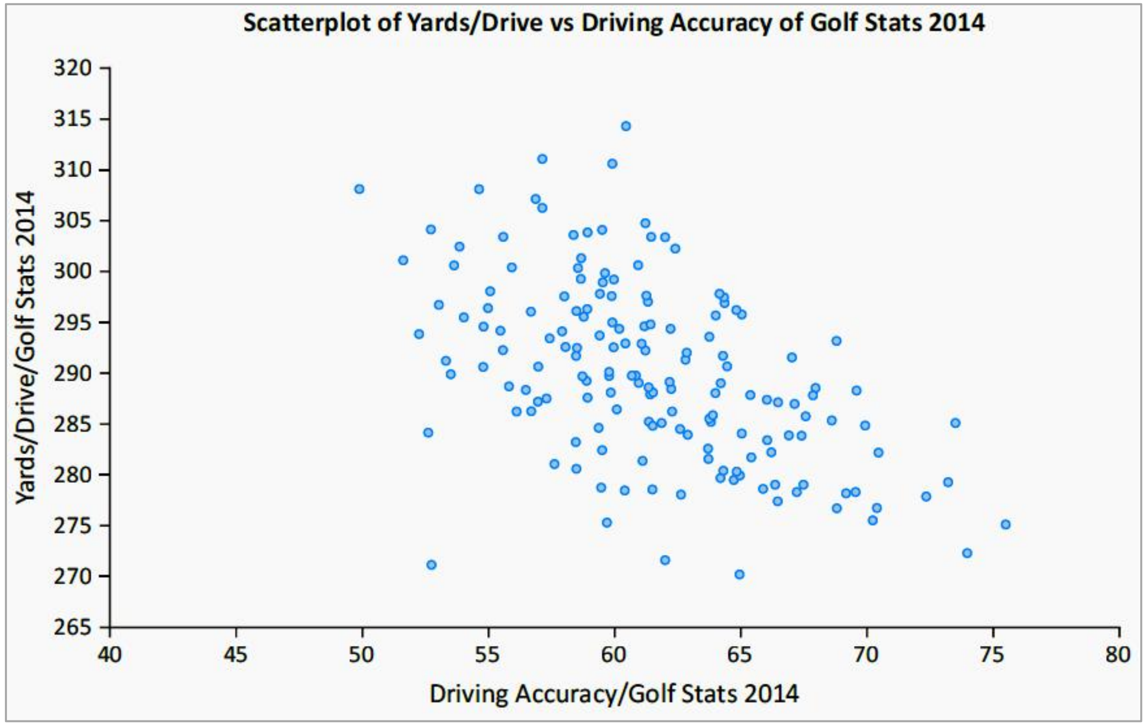


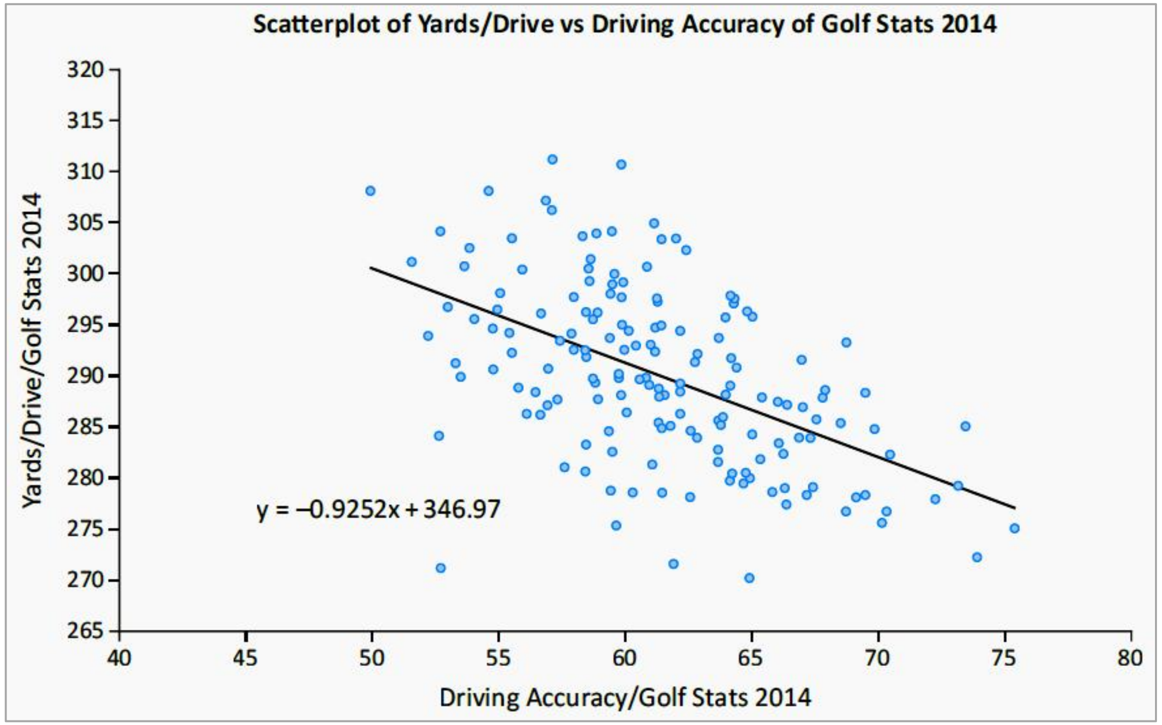


Relationships Among Numerical & Numerical Variables (cont..)

(Scatterplot)

- Once you have a scatterplot, it enables you to superimpose one of several trend lines on the scatterplot.
  - A **trend line** is a line or curve that "fits" the scatter as well as possible.
  - This could be a straight line, or it could be one of several types of curves.





Relationships Among Numerical & Numerical Variables (cont..)

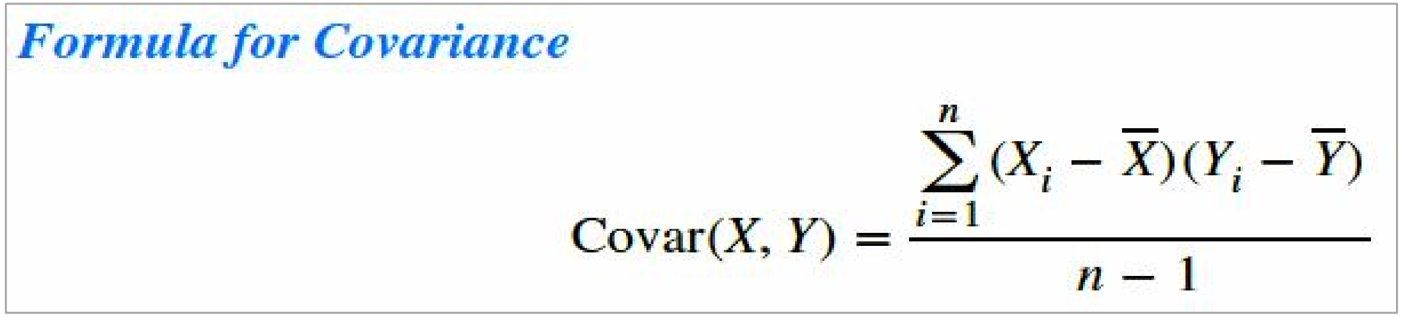
Correlation and Covariance

- **Correlation** and **covariance** measure the strength and direction of a linear relationship between two numerical variables. (**Bi-Variate Measures)**
  - The relationship is "strong" if the points in a scatterplot cluster tightly around some straight line.
    - ➢ *If this straight line rises from left to right, the relationship is positive and the measures will be positive numbers.*
    - ➢ *If it falls from left to right, the relationship is negative and the measures will be negative numbers.*
  - The two numerical variables must be "paired" variables.
    - ➢ *They must have the same number of observations, and the values for any observation should be naturally paired.*

Specifically, each measures the strength and direction of a linear relationship between two numerical variables.

Relationships Among Numerical & Numerical Variables (cont..)

- **Covariance** is essentially an average of products of deviations from means.



Formula for Covariance

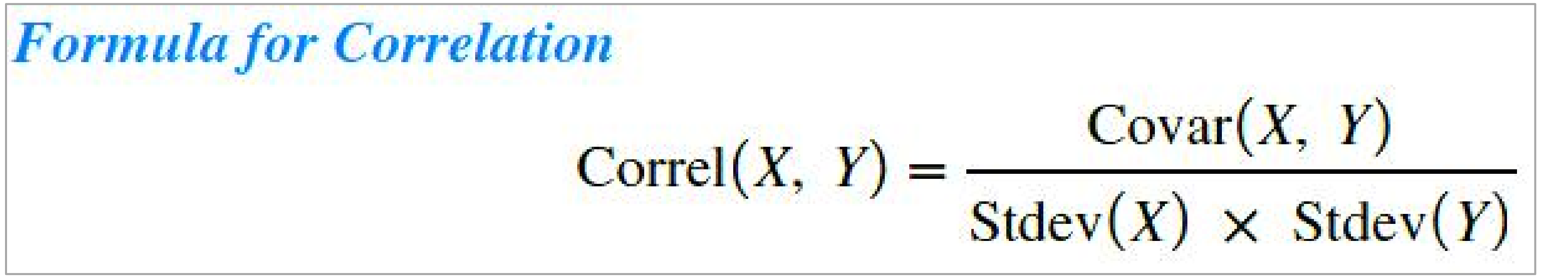
$$\text{Covar}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

- With this in mind, let X_i and Y_i be the paired values for observation *i*, and let *n* be the number of observations. Then the covariance between *X* and *Y*, denoted by *Covar*(*X*, *Y*).
- Covariance has a serious limitation as a descriptive measure because it is very sensitive to the *units* in which *X* and *Y* are measured.

For example, the covariance can be inflated by a factor of 1000 simply by measuring X in dollars rather than thousands of dollars. In contrast, the ***correlation****, denoted remedies this problem.*

Relationships Among Numerical & Numerical Variables (cont..)

- **Correlation** is a *unitless* quantity that is unaffected by the measurement scale.
- For example, the correlation is the same regardless of whether the variables are measured in dollars, thousands of dollars, or millions of dollars.



Formula for Correlation

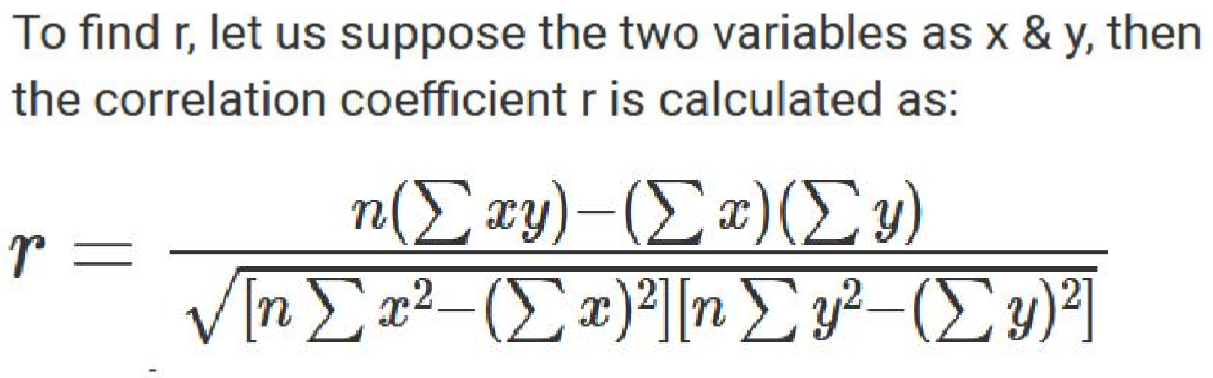
$$\text{Correl}(X,\ Y) = \frac{\text{Covar}(X,\ Y)}{\text{Stdev}(X)\ \times\ \text{Stdev}(Y)}$$

- The correlation is defined by Equation, where Stdev(X) and Stdev(Y) denote the standard deviations of X & Y, and Covar(X,Y) denote the covariance of X & Y.

Relationships Among Numerical & Numerical Variables (cont..)

Correlation

Pearson’s correlation coefficient formula



To find r, let us suppose the two variables as x & y, then the correlation coefficient r is calculated as:

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

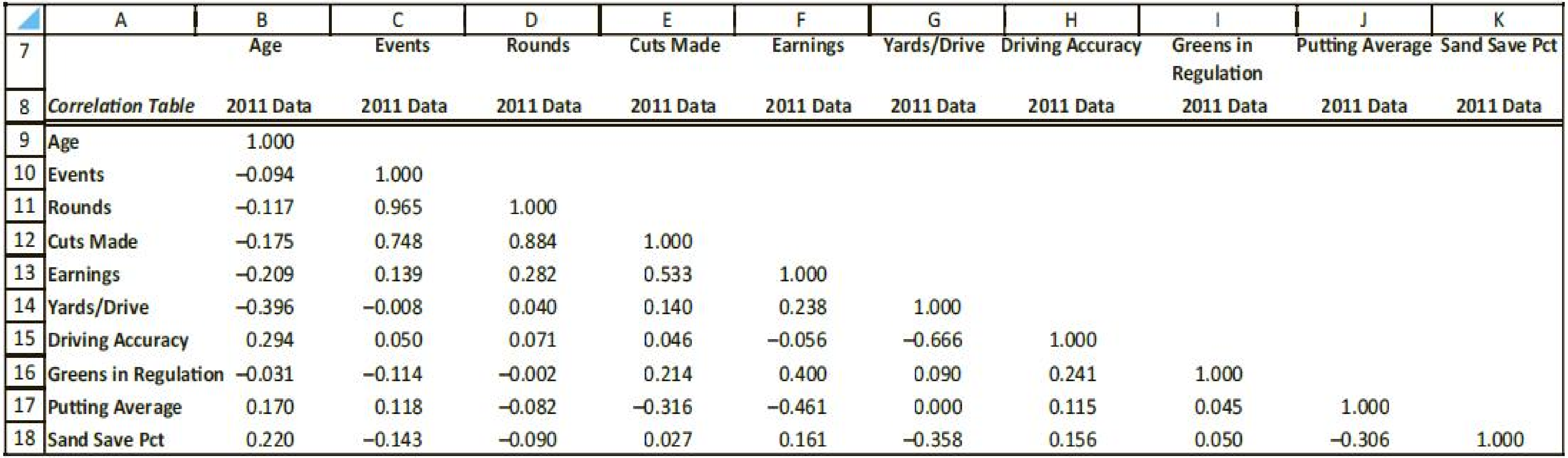
Both variables are quantitative and normally distributed with no outliers, so we calculate a Pearson’s r correlation coefficient.

- The closer **r (Correlation)** is to **zero**, the weaker the linear relationship.
- **Positive r (Correlation)** values indicate a ***positive correlation***, where the values of both variables tend to increase together.
- **Negative r (Correlation)** values indicate a ***negative correlation***, where the values of one variable tend to increase when the values of the other variable decrease.

Relationships Among Numerical & Numerical Variables (cont..)

Correlation

The resulting table of correlations appears in Figure. You can ignore the 1.000 values along the diagonal because a variable is always perfectly correlated with itself.

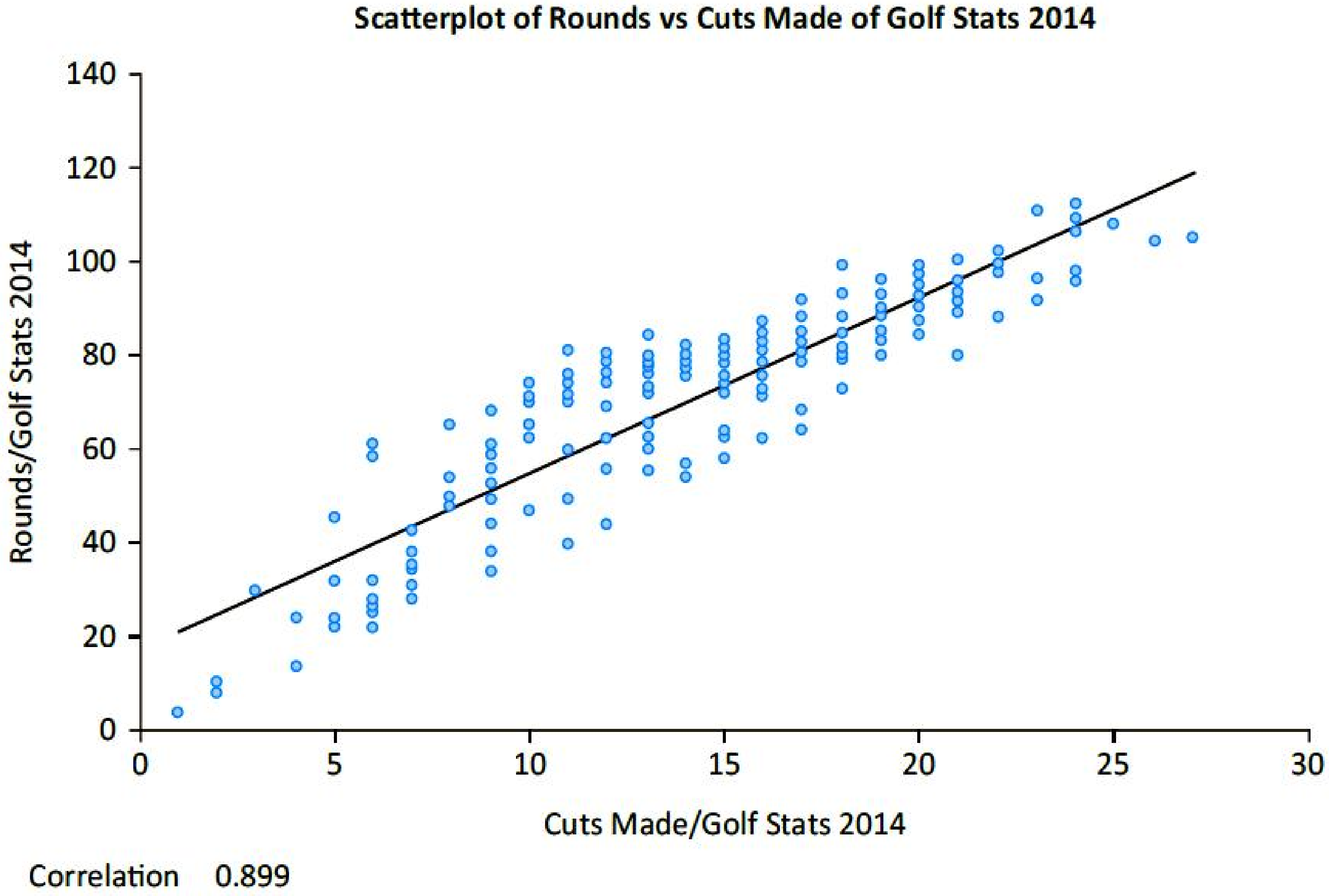


	A	B	C	D	E	F	G	H	I	J	K
7		Age	Events	Rounds	Cuts Made	Earnings	Yards/Drive	Driving Accuracy	Greens in Regulation	Putting Average	Sand Save Pct
8	*Correlation Table*	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data	2011 Data
9	Age	1.000									
10	Events	–0.094	1.000								
11	Rounds	–0.117	0.965	1.000							
12	Cuts Made	–0.175	0.748	0.884	1.000						
13	Earnings	–0.209	0.139	0.282	0.533	1.000					
14	Yards/Drive	–0.396	–0.008	0.040	0.140	0.238	1.000				
15	Driving Accuracy	0.294	0.050	0.071	0.046	–0.056	–0.666	1.000			
16	Greens in Regulation	–0.031	–0.114	–0.002	0.214	0.400	0.090	0.241	1.000		
17	Putting Average	0.170	0.118	–0.082	–0.316	–0.461	0.000	0.115	0.045	1.000	
18	Sand Save Pct	0.220	–0.143	–0.090	0.027	0.161	–0.358	0.156	0.050	–0.306	1.000

Finally, correlations (and covariances) are symmetric in that the correlation between any two variables X and Y is the same as the correlation between Y and X.

Relationships Among Numerical & Numerical Variables (cont..)

Correlation



- For example, the scatterplot corresponding to the 0.884 correlation between Cuts Made and Rounds appears in Figure. (We also superimposed a trend line.)
- This chart shows the strong linear relationship between cuts made and rounds played, but it also shows that there is still considerable variability around the best-fitting straight line, even with a correlation as large as 0.899.