greatlearning
Learning for Life

Statistical Learning – Descriptive Statistics

What do the numbers tell?

Why Study Statistics

- Technological developments, Revolution of Internet and social networks, data generated from electronic devices, produce large amount of data
- Large storage capacity
- Advancement in enormous computing power to effectively process and analyze large amount of data
- Better data visualization from Business Intelligence
- Discovery of patterns and trends from this data can help organizations gain competitive advantage in marketplace

Types of Statistics

- Descriptive statistics is concerned with Data Summarization Graphs/Charts and tables.

- Inferential Statistics is the method used to talk about a population parameter from a sample. It involves point estimation, interval estimation, and hypothesis testing.

Some Key Terms

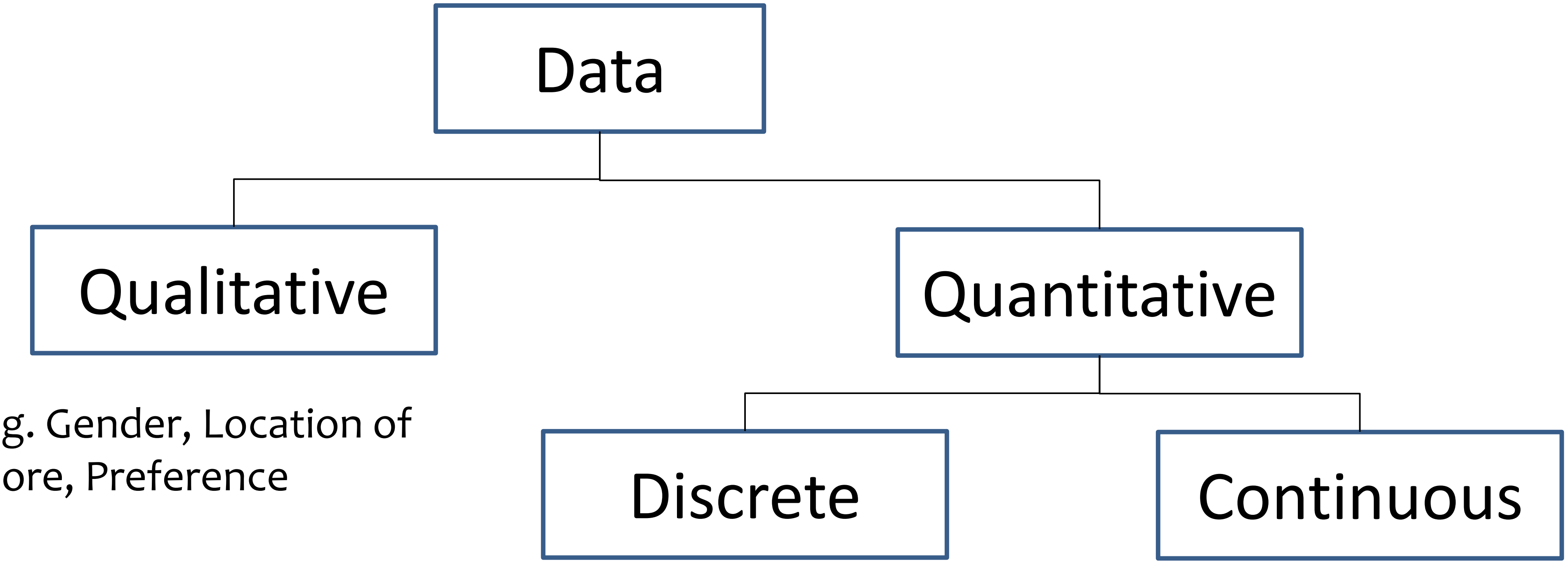
- **Population** is the collection of all possible observations of a specified of characteristic interest.

- **Sample** is a subset of population

- **Parameter** is the population characteristic of interest. For example, you are interested in the average income of a particular class of people. The average income of this entire class of people is called a parameter.

- **Statistic** is based on a sample to make inferences about the population parameter. The average income of population can be estimated by the average income based on the sample. This sample average is called a statistic.

Types of Data



Data

Qualitative

Quantitative

Discrete

Continuous

E.g. Gender, Location of store, Preference

E.g. Family size, Number of rooms in a hotel, number of credit cards issued

E.g. Waiting time, Length of a part produced

Measurement Scales

- Nominal –e.g. Internet service provider
- Ordinal: e.g. Bond rating, employee designation
- Interval: e.g. Temperature in °C or °F
- Ratio: e.g. cost of an item

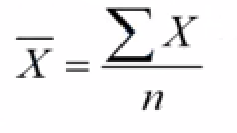
Measure of central Tendency

- As a manager, You need the summary measures of central tendency to draw meaningful conclusions in the functional area of operation.

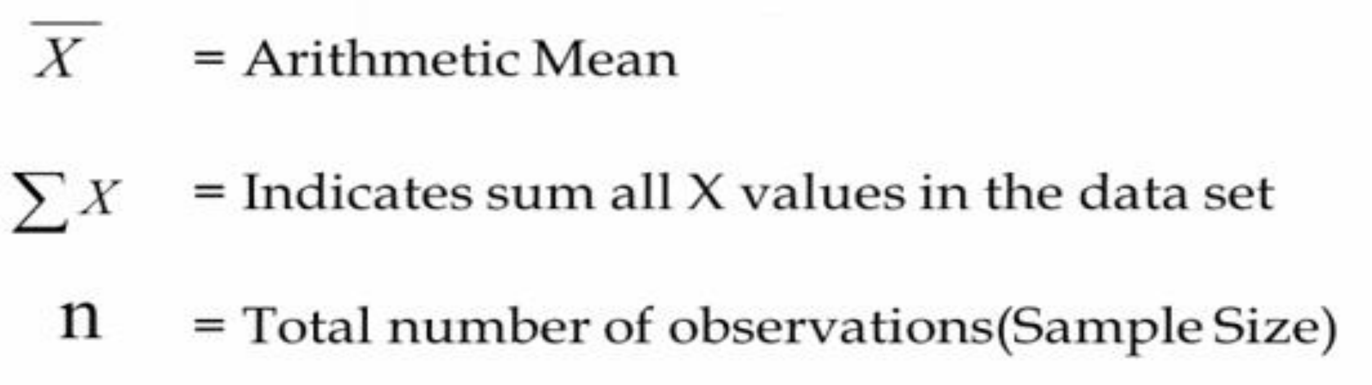
The most widely used measures of central tendency are the Arithmetic Mean, Median and Mode.

Arithmetic Mean

- Arithmetic mean(called Mean) is defined as the sum of all observations in a data set divided by the total number of observations. For example, consider a data set containing the following observations:
- In symbolic form mean is giver



$$\overline{X} = \frac{\sum X}{n}$$



$\overline{X}$ = Arithmetic Mean

$\sum X$ = Indicates sum all X values in the data set

n = Total number of observations(Sample Size)

Arithmetic Mean - Example

- The inner diameter of a particular grade of tire based on 5 sample measurements are as follows: (Figures in millimetres)

565, 570, 572, 568, 585

Applying the formula $\overline{X} = \frac{\sum X}{n}$

We get mean = (565 + 570+572+568+585)/5 =572

- Caution: Arithmetic Mean is affected by extreme values or fluctuations in sampling. It is not the best average to use when the data set contains extreme values (Very high or very low values).

Median

- Median is the middle most observation when you arrange data in ascending order of magnitude. Median is such 50% of the observations are above the median and 50% of the observations are below the median.
- Median is a very useful measure for ranked data in the context Preferences and rating. It is not affected by extreme values (greater resistance to outliers)
- Median = (n+1)/2 th value of ranked data.
- n = Number of observations in the sample

Median - Example

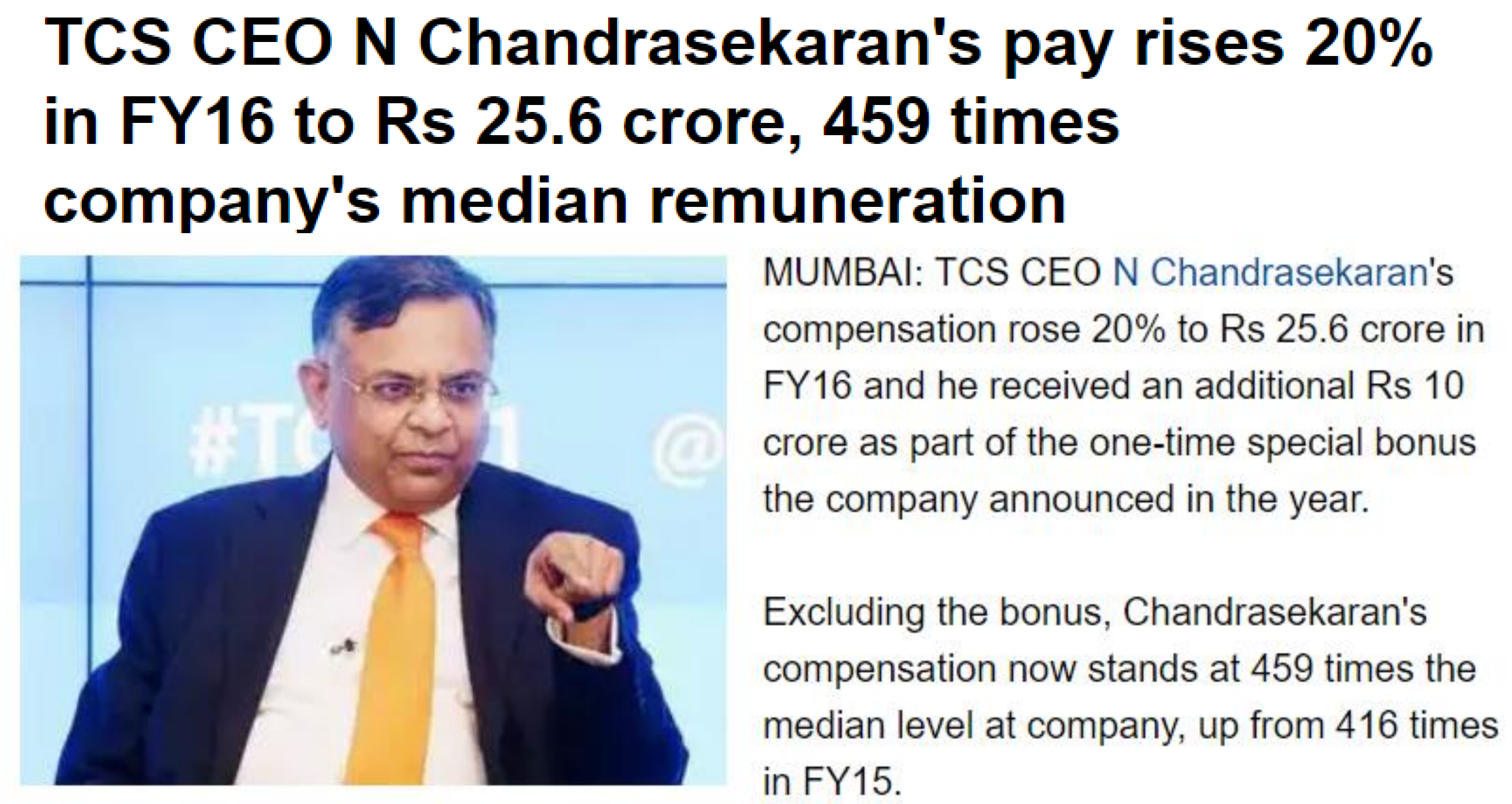
- Marks obtained by 7 students in computer science exam are given below: Compute the median.

45 40 60 80 90 65 55

- Arranging the data after ranking them

40 45 55 60 65 80 90

- Median = (n+1)/2 th value in this set = (7+1)/2 th observation= 4th observation=60
- Hence median = 60 for this problem.



TCS CEO N Chandrasekaran's pay rises 20% in FY16 to Rs 25.6 crore, 459 times company's median remuneration

MUMBAI: TCS CEO N Chandrasekaran's compensation rose 20% to Rs 25.6 crore in FY16 and he received an additional Rs 10 crore as part of the one-time special bonus the company announced in the year.

Excluding the bonus, Chandrasekaran's compensation now stands at 459 times the median level at company, up from 416 times in FY15.

https://economictimes.indiatimes.com/articleshow/52450273.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

Australia batsman Hughes dies from head injury

………. Questions about the response time of ambulances dispatched to the stadium were also raised. The head of New South Wales Ambulance was to be hauled before the state health minister Jillian Skinner on Thursday after the ambulance authority issued conflicting statements about their response times. The arrival of the first ambulance took 15 minutes, NSW Ambulance clarified in a statement on Wednesday. The state's **median response time** for the highest priority life-threatening cases was just under eight minutes in 2013-14, according the authority's statistics……

http://timesofindia.indiatimes.com/articleshow/45292785.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

Mode

- Mode is that value which occurs most often. It has the maximum frequency of occurrence. Mode also has resistance to outliers.
- Mode is a very useful measure when you want to keep in the inventory, the most popular shirt in terms of collar size during festival season.

- Caution: In a few problems in real life, there will be more than one mode such as bimodal and multi-modal values. In these cases mode cannot be uniquely determined.

Mode - Example

- The life in number of hours of 10 flashlight batteries are as follows: Find the mode

- 340 340 350 350 340 340 320 340 330 330

- 340 occurs five times. Hence, mode = 34O.

Comparison of Mean, Median, Mode Cont.

Mean	Median	Mode
Affected by extreme values.	Not affected by extreme values.	Not affected by extreme values.
Can be treated algebraically. That is, Means of several groups can be combined.	Cannot be treated algebraically. That is, Medians of several groups cannot be combined.	Cannot be treated algebraically. That is, Modes of several groups cannot be combined.

Measures of Dispersion.

- In simple terms, of dispersion indicate how large the spread of the distribution is around the central tendency.
- It answers unambiguously the equation

“What is the magnitude of departure from the average value for different groups having identical averages?”.

Range

- Range is the simplest of all the measures of dispersion. It is calculated as the difference between maximum and minimum value in the data set.

$$\text{Range} = X_{\text{Maximum}} - X_{\text{Minimum}}$$

Range -Example

Example for calculating Range

The following data represents the percentage return on the investment for the 9 mutual funds per annum.

Calculate the Range.

12, 14, 11, 18, 11.3, 12, 14, 11, 9

Range = $X_{\text{Maximum}} - X_{\text{minimum}}$ = **18 - 9 = 9**

Caution: If one of the components of range namely the maximum value or minimum value becomes an extreme value, then range should not be used.

Inter-Quartile Range(IQR)

- IQR= Range computed on middle 50% of the observations after eliminating the highest and lowest 25% of observations in a data set that is arranged ascending order. IQR is less affected by outliers.

- **IQR =Q3-Q1**

Interquartile Range-Example

- The following data represents the percentage return on investment for 9 mutual funds per annum. Calculate interquartile range.

- Data set: 12, 14, 11, 18, 11.5, 12, 14, 11, 9
- Arranging in ascending order, the data set becomes

9, 11, 11, 11.5, 12, 12, 14, 14, 18

IQR = Q3 – Q1 = 14 – 11 = 3

Standard deviation

- Standard deviation forms the cornerstone for the inferential statistics.

- To define standard deviation, you need to define another term called variance. In simple terms, standard deviation is the square root of variance.

Key Formulas

Important Terms with Notations

Sample Variance

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1}$$

Sample Standard Deviation

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Population Variance

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Population Standard Deviation $\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$

Where $\overline{X} = \frac{\sum X}{n}$ (Sample Mean) and

$\mu = \frac{\sum X}{N}$ (Population Mean)

n =Number of observations in the sample(Sample size)

N =Number of observations in the Population (Population Size)

Remarks

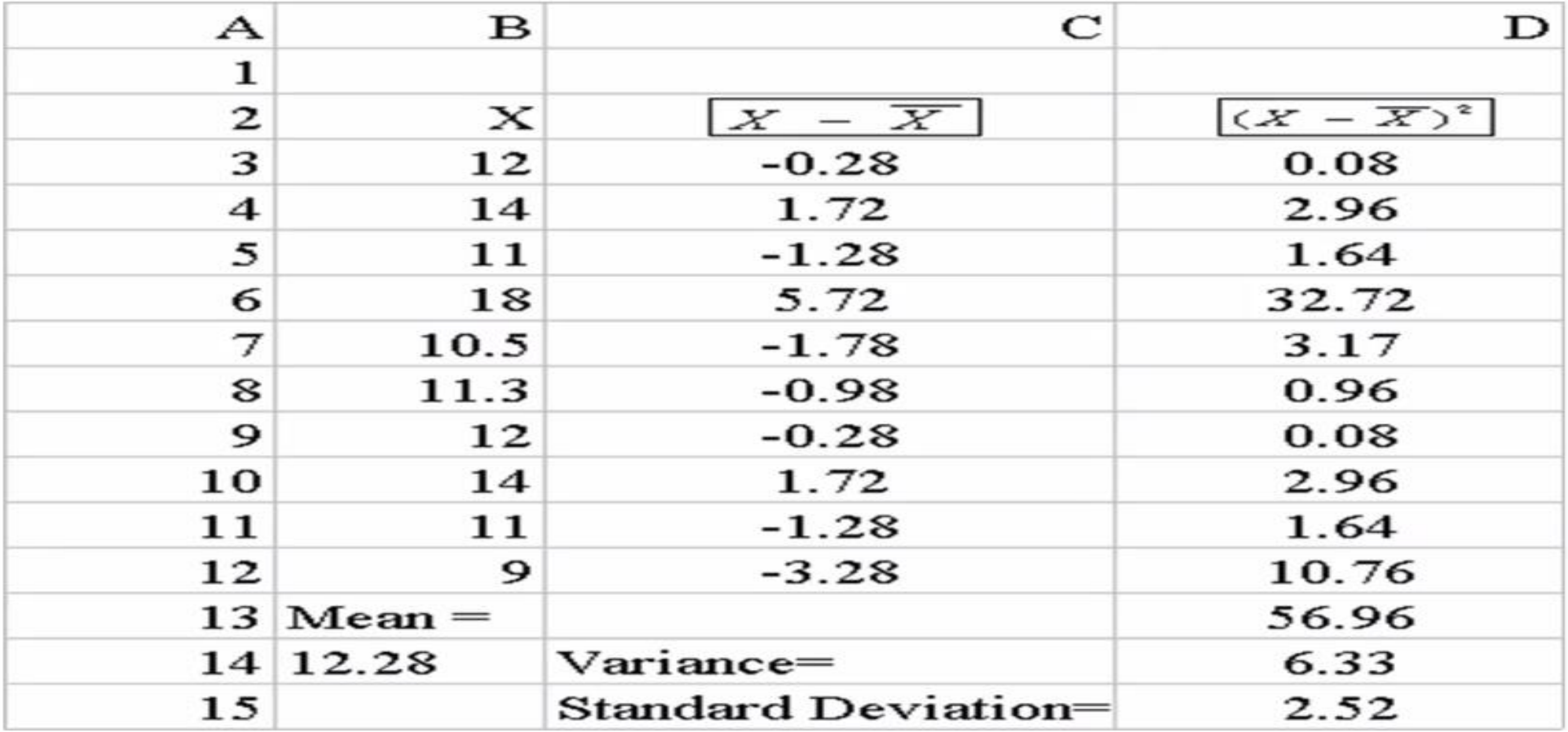
1. $S^2 = \frac{\sum (X - \overline{X})^2}{n-1}$ is an unbiased estimator of $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
2. $\overline{X} = \frac{\sum X}{n}$ is an unbiased estimator of $\mu = \frac{\sum X}{N}$
3. The divisor n-1 is always used while calculating sample variance for ensuring property of being unbiased
4. Standard deviation is always the square root of variance

Example of Standard Deviation

- The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate the sample standard deviation.

- 12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

Solution for the Example



A	B	C	D
1			
2	X	$\boxed{X - \overline{X}}$	$\boxed{(X - \overline{X})^2}$
3	12	-0.28	0.08
4	14	1.72	2.96
5	11	-1.28	1.64
6	18	5.72	32.72
7	10.5	-1.78	3.17
8	11.3	-0.98	0.96
9	12	-0.28	0.08
10	14	1.72	2.96
11	11	-1.28	1.64
12	9	-3.28	10.76
13	Mean =		56.96
14	12.28	Variance=	6.33
15		Standard Deviation=	2.52

Solution for the example cont.

From the spreadsheet of the Microsoft excel in the previous slide, it is easy to see

that Mean = $\bar{X} = \frac{\sum X}{n}$ = 12.28 (In column A and row14, 12.28 is seen)

Sample variance = $S^2 = \frac{\sum (X-\bar{X})^2}{n-1}$ = 6.33 (In column D and row14, 6.33 is seen)

Sample standard deviation = $S = \sqrt{\frac{\sum (X-\bar{X})^2}{n-1}}$ = 2.52 (In column D and row15, 2.52 is seen)

Coefficient of Variation (Relative Dispersion)

- Coefficient Variation (CV) is defined as the ratio of standard deviation to mean.

- In symbolic form

$CV = S/X$ for the sample data and $= \sigma/\mu$ for the population data.

Coefficient of Variation Example

- Following is the performance of two Sales Teams in terms of monthly sales

Comment on the results.

Sales Team 1

- Standard deviation: 10 units

Sales Team 2

- Standard Deviation 12 units

Coefficient of Variation Example

- Additional information

Sales Team 1

- Mean: 70 units

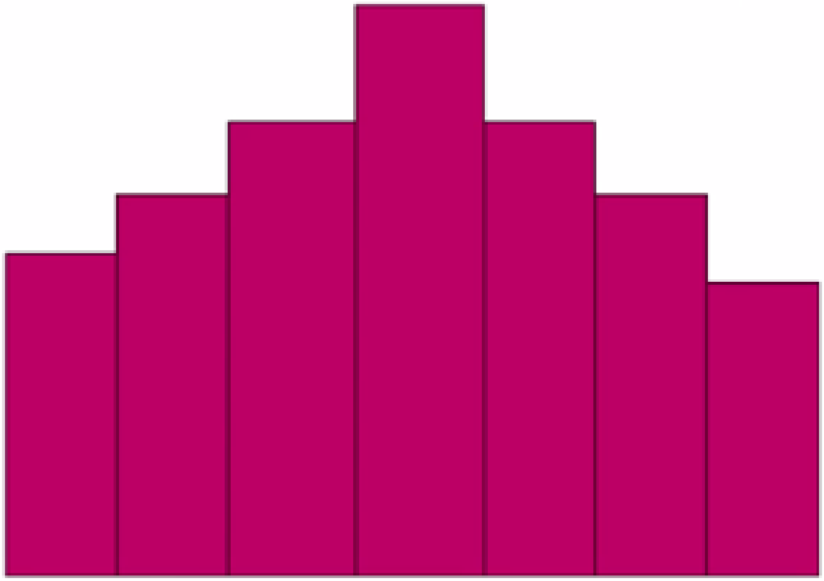
Sales Team 2

- Mean: 120 units

Interpretation for the Example

- The CV for Team 1 is 10/70 = 0.14 or 14%
- The CV for Team 2 is 12/120 = 0.10 or 10%

HISTOGRAM



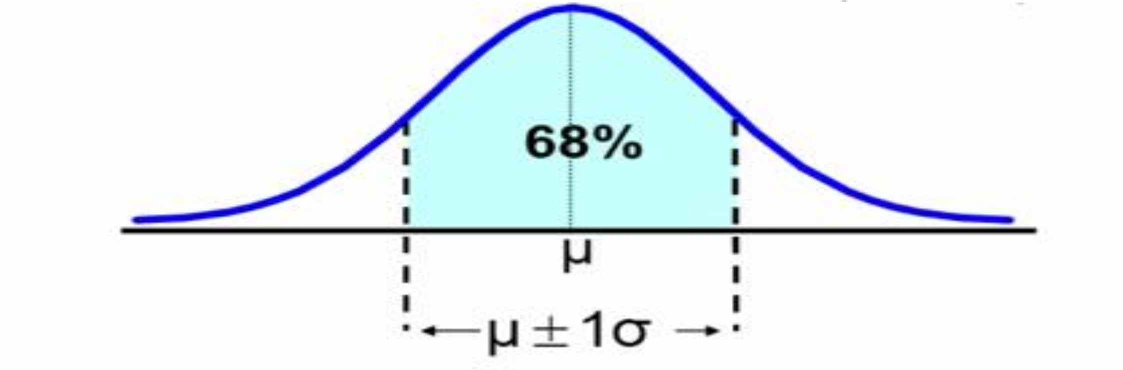
Histogram(also known as frequency histogram) is a snap shot of the frequency distribution.

Histogram is a graphical representation of the frequency distribution in which the X-axis represents the classes and the Y-axis represents the frequencies in bars.

Histogram depicts the pattern of the distribution emerging from the characteristic being measured.

The Empirical Rule

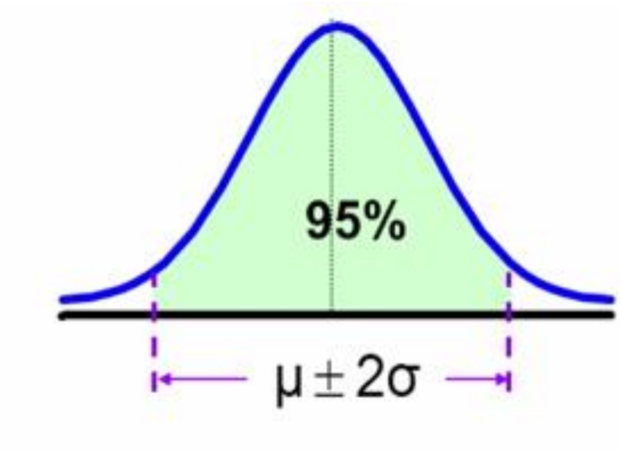
- The empirical rule approximates the variation of data in the bell-shaped distribution.
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$

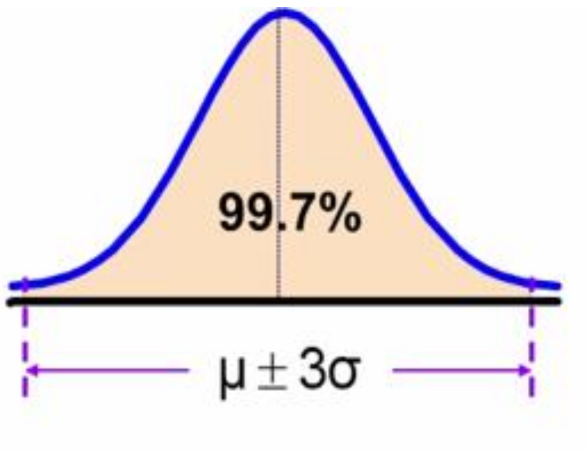


The Empirical Rule

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations $\mu \pm 2\sigma$ e mean, or

- Approximately 99.7% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or



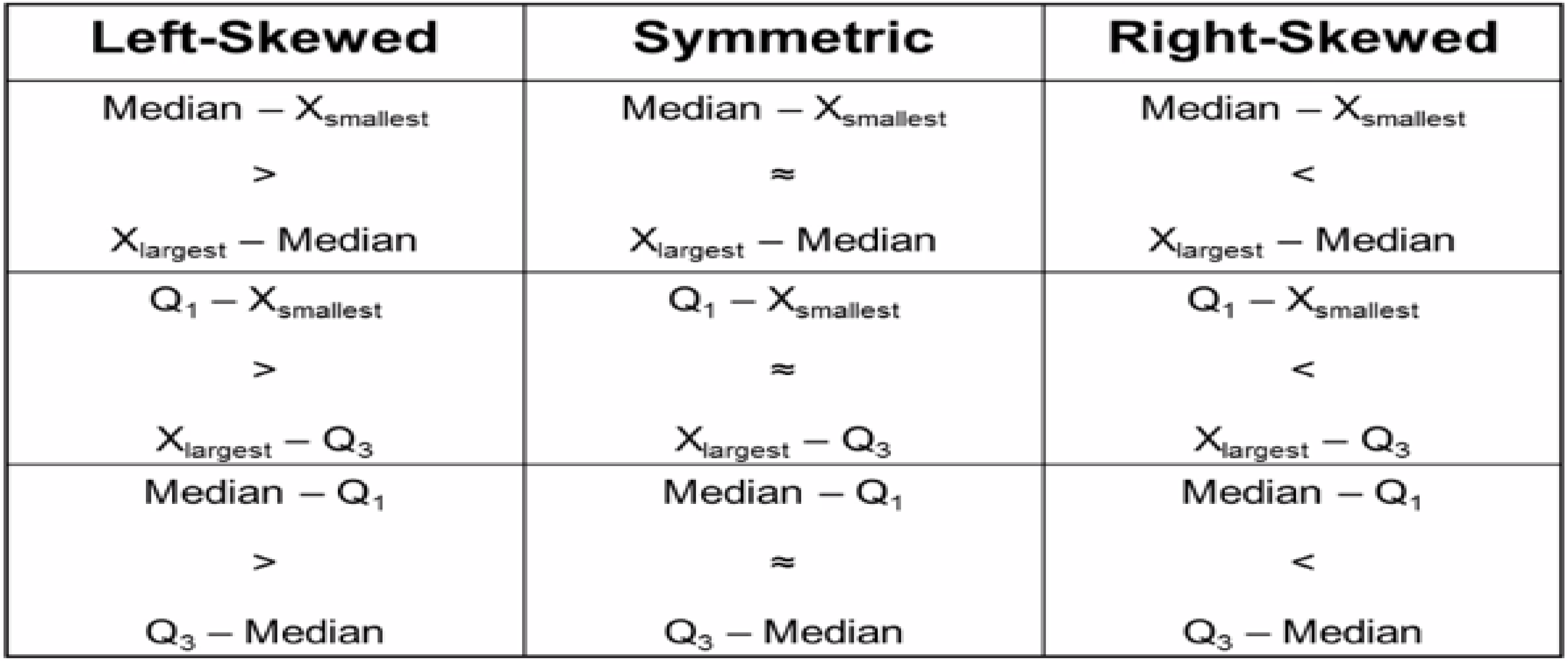


$\mu \pm 3\sigma$

The five number summary

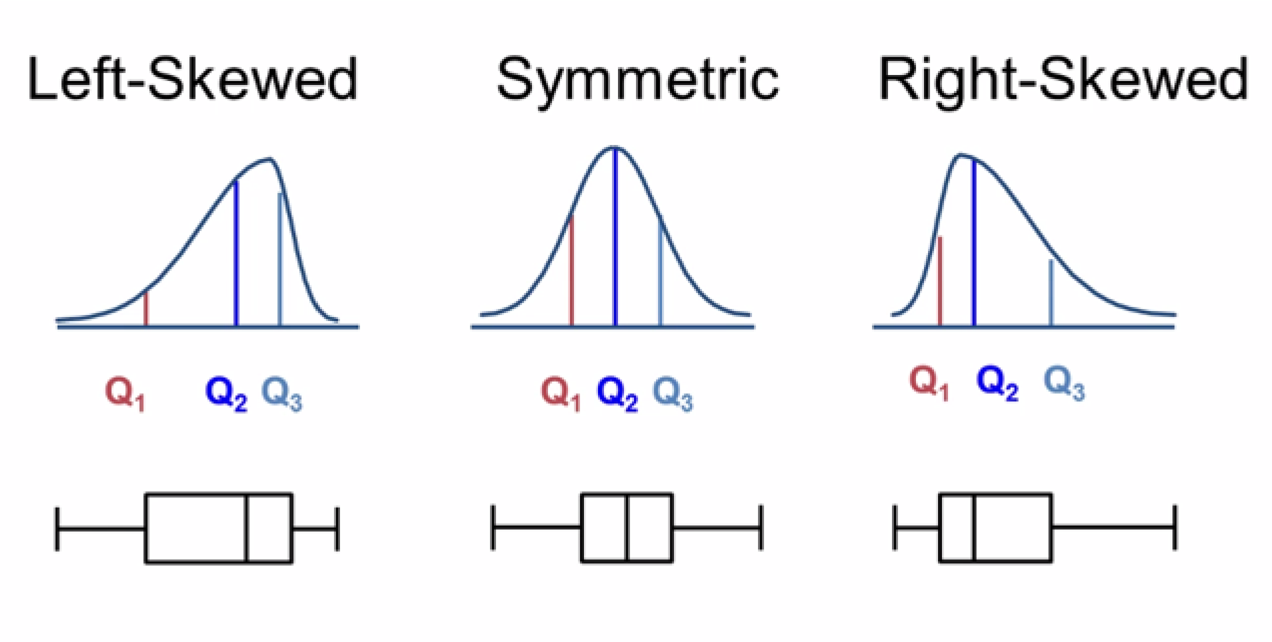
- The five numbers that help describe the center, spread and shape of the data are:
  - $X_{Smallest}$
  - First Quartile (Q_1)
  - Median (Q_2)
  - Third Quartile (Q_3)
  - $X_{Largest}$

Relationships among the five-number summary and distribution shape



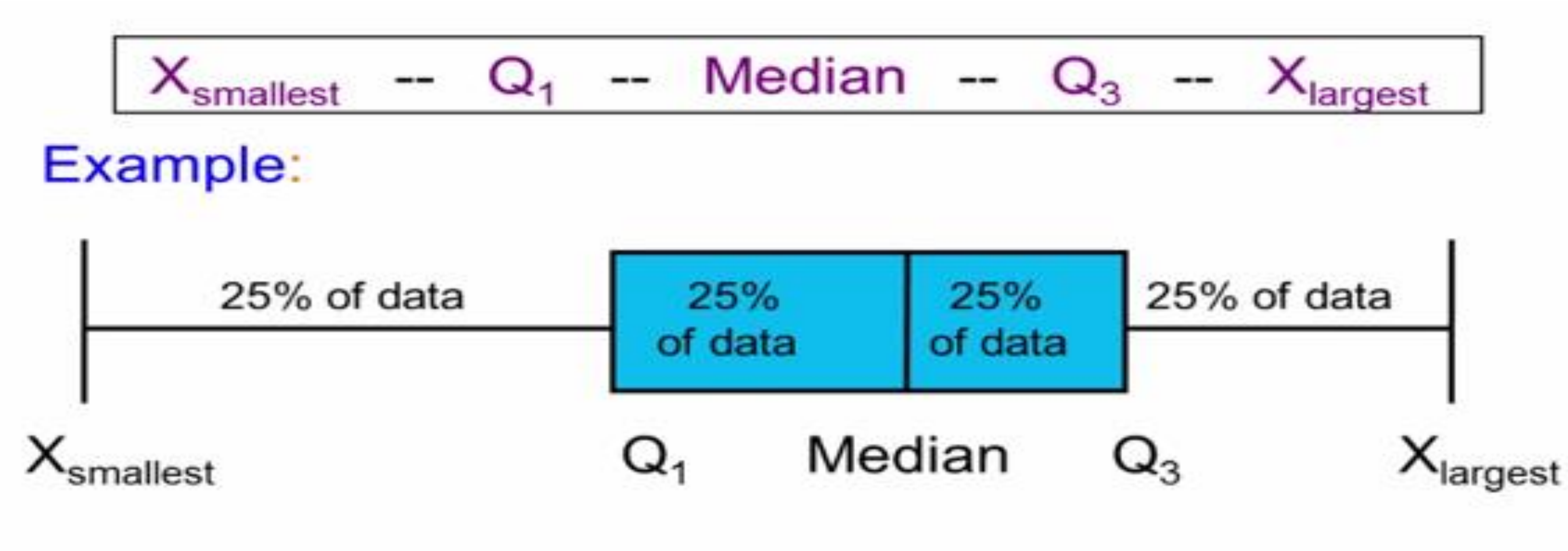
Left-Skewed	Symmetric	Right-Skewed
$\text{Median} - X_{\text{smallest}}$ $>$ $X_{\text{largest}} - \text{Median}$	$\text{Median} - X_{\text{smallest}}$ $\approx$ $X_{\text{largest}} - \text{Median}$	$\text{Median} - X_{\text{smallest}}$ $<$ $X_{\text{largest}} - \text{Median}$
$Q_1 - X_{\text{smallest}}$ $>$ $X_{\text{largest}} - Q_3$	$Q_1 - X_{\text{smallest}}$ $\approx$ $X_{\text{largest}} - Q_3$	$Q_1 - X_{\text{smallest}}$ $<$ $X_{\text{largest}} - Q_3$
$\text{Median} - Q_1$ $>$ $Q_3 - \text{Median}$	$\text{Median} - Q_1$ $\approx$ $Q_3 - \text{Median}$	$\text{Median} - Q_1$ $<$ $Q_3 - \text{Median}$

Distribution shape and The Boxplots



Five number summary and The Boxplot

- The Boxplot: A graphical display of the data on the five-number summary:

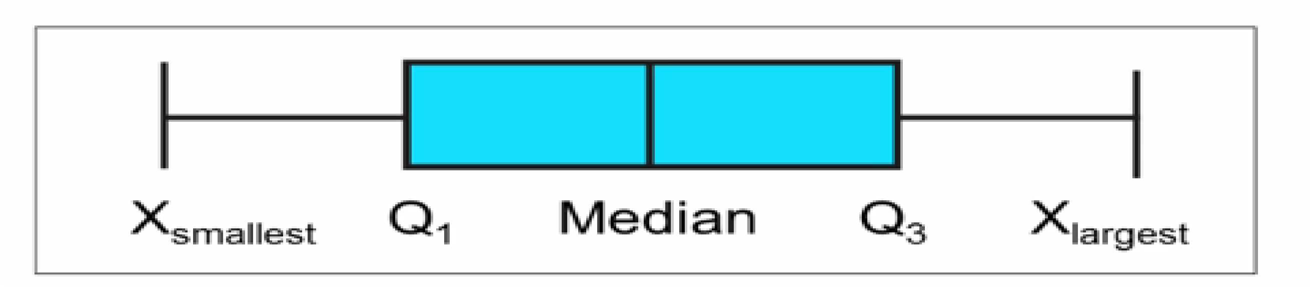


$X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:

Five number summary: Shape of Boxplots

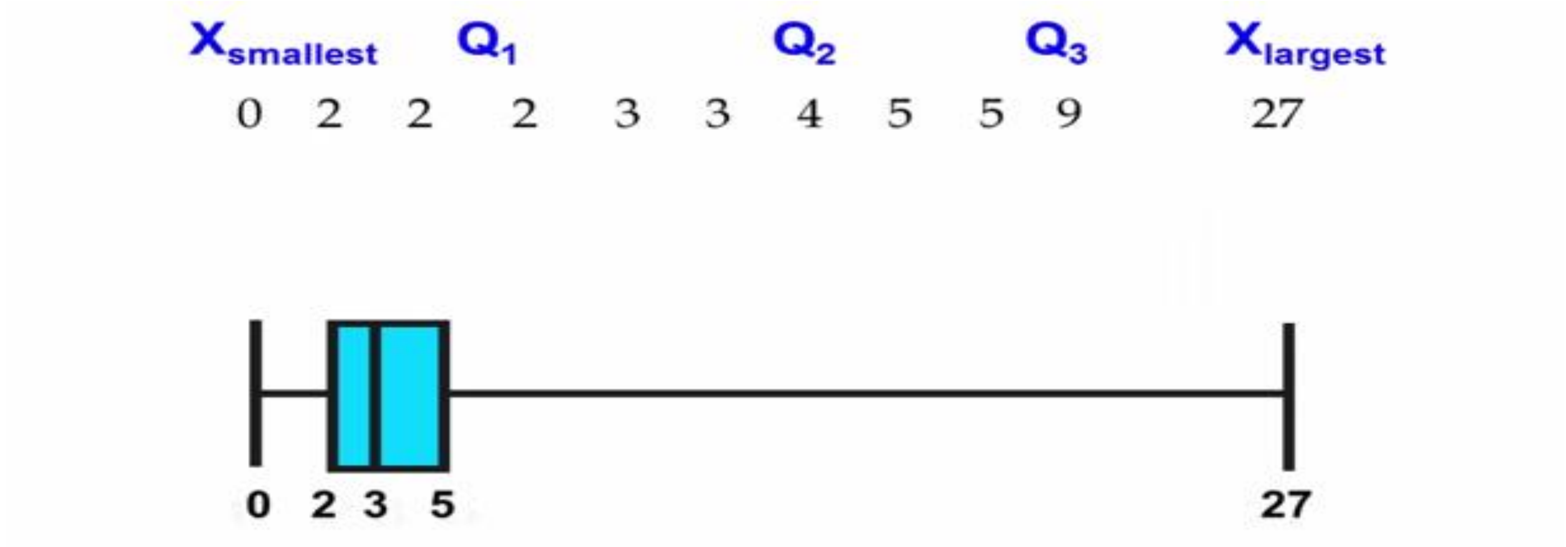
- If data is symmetric around the median then the box and central line are centered between the endpoints.



- A Boxplot can be shown in either a vertical or horizontal orientation.

Boxplot Example

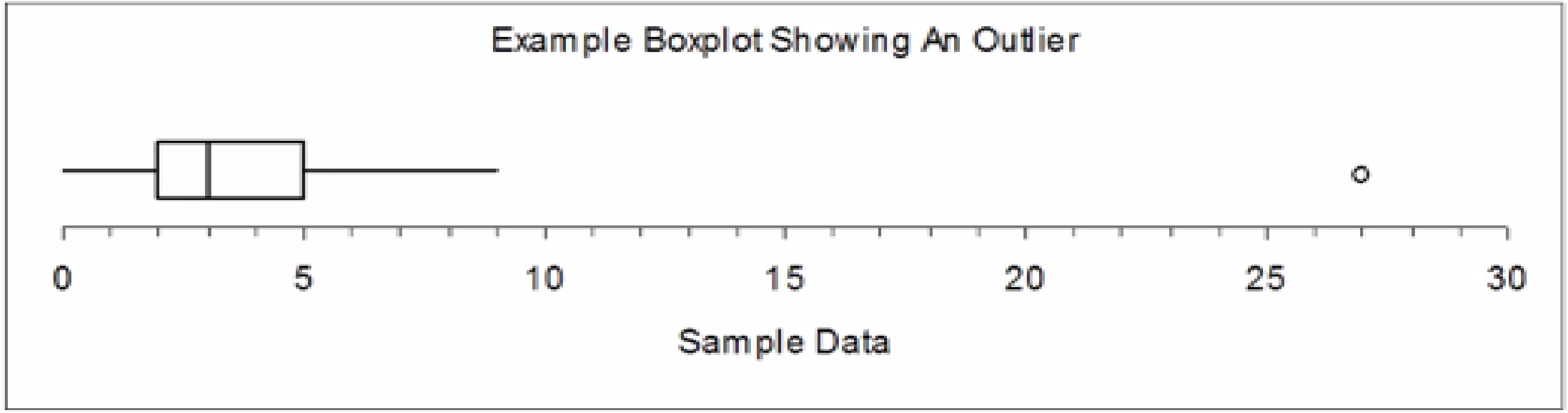
- Below is a Boxplot for the following data:

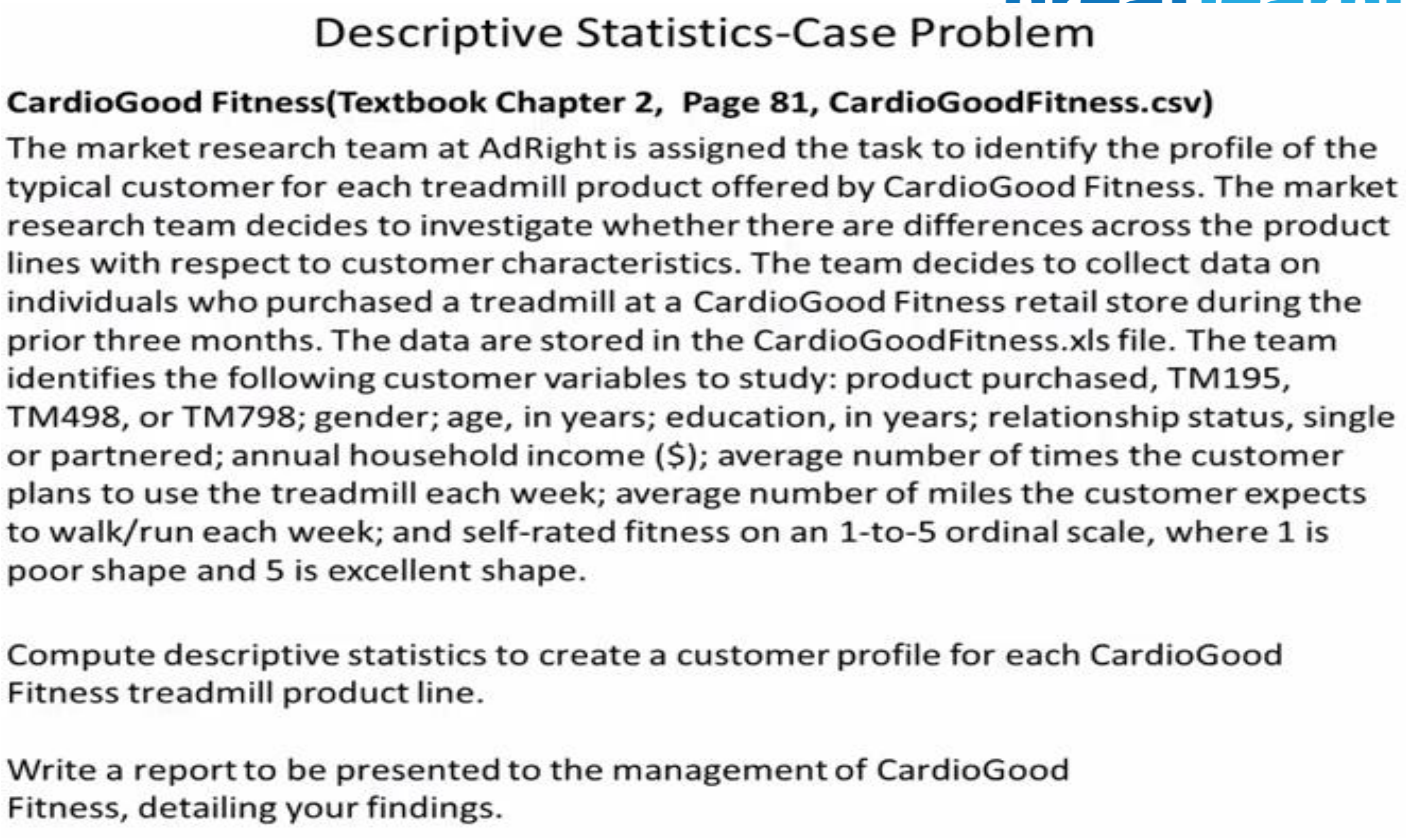


- The data are right skewed, as the plot depicts

Boxplot example showing an outlier

- The Boxplot below of the same data shows the outlier value of 27 plotted separately.
- A value is considered an outlier if it is more than 1.5 times the interquartile range between Q_1 or above Q_3.





Descriptive Statistics-Case Problem

CardioGood Fitness(Textbook Chapter 2, Page 81, CardioGoodFitness.csv)

The market research team at AdRight is assigned the task to identify the profile of the typical customer for each treadmill product offered by CardioGood Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a CardioGood Fitness retail store during the prior three months. The data are stored in the CardioGoodFitness.xls file. The team identifies the following customer variables to study: product purchased, TM195, TM498, or TM798; gender; age, in years; education, in years; relationship status, single or partnered; annual household income ($); average number of times the customer plans to use the treadmill each week; average number of miles the customer expects to walk/run each week; and self-rated fitness on an 1-to-5 ordinal scale, where 1 is poor shape and 5 is excellent shape.

Compute descriptive statistics to create a customer profile for each CardioGood Fitness treadmill product line.

Write a report to be presented to the management of CardioGood Fitness, detailing your findings.

Covariance

- The covariance measures the strength of the linear relationship between two numerical variables (X and Y). Formula for sample Covariance is

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

- Drawback of covariance: It can have any value, so it cannot be used to determine the relative strength of the relationship

Coefficient of Correlation

- Coefficient of correlation (r) measures the relative strength of a linear relationship between two numerical variables.

$$r = \frac{\operatorname{cov}(X, Y)}{S_X S_Y}$$

- The values r range from -1 to + 1

- The value -1 indicates a perfect negative correlation and +1 indicates a perfect positive correlation