STATISTICS FOR DATA SCIENCE

Statistics Types and Summary

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Descriptive & Inferential Statistics

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

STATISTICS FOR DATA SCIENCE

Statistics

1. WHAT IS STATISTICS?
2. TYPES OF STATISTICS
3. DESCRIPTIVE STATISTICS

source:en.wikibooks.org,www.fullcirclereal.com

STATISTICS FOR DATA SCIENCE

Why Statistics?

To find a way a process behaves the way it does.

Why a process produces defective goods and services?

To check various performance measures of a process.

To prevent problems caused by various causes of variation in process.

To analyze the real world.

STATISTICS FOR DATA SCIENCE

Statistics

The word **statistics** convey a **variety of meaning** to people in different walks of life.

The word statistics comes from a **Italian** word **Statista** meaning **statement**

and

German word **statistik** meaning **political state**.

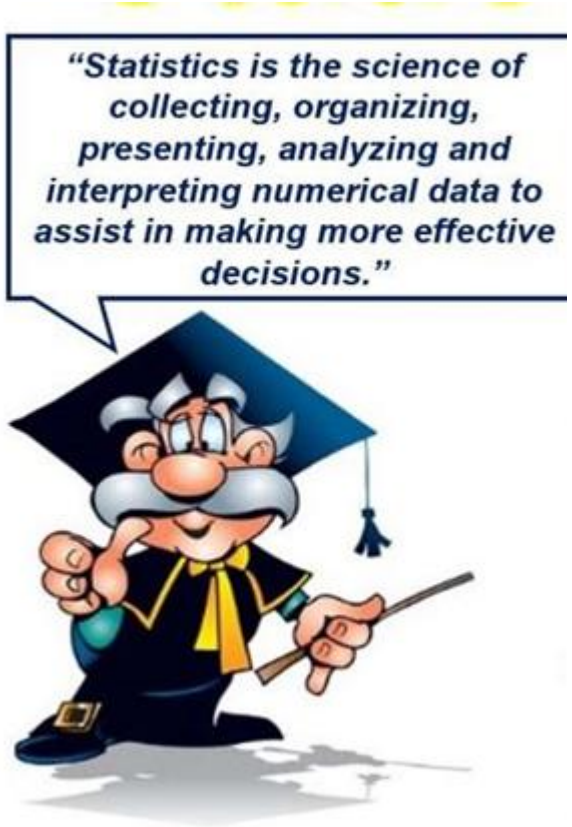
Statistics is a science of data.

It is a **method** of
dealing with **quantitative or qualitative information**.

source:en.wikibooks.org,www.fullcirclereal.com

STATISTICS FOR DATA SCIENCE

Statistics



Statistics is the **science of collecting**, **organizing**, **presenting**, **analyzing** and **interpreting numerical data** to assist in making **more effective decisions.**

Sources: colourbox.com, luminousmen.com/post/

STATISTICS FOR DATA SCIENCE

Statistics

Statistics is the **branch of mathematics** that **transforms data** into **useful information** for decision makers.

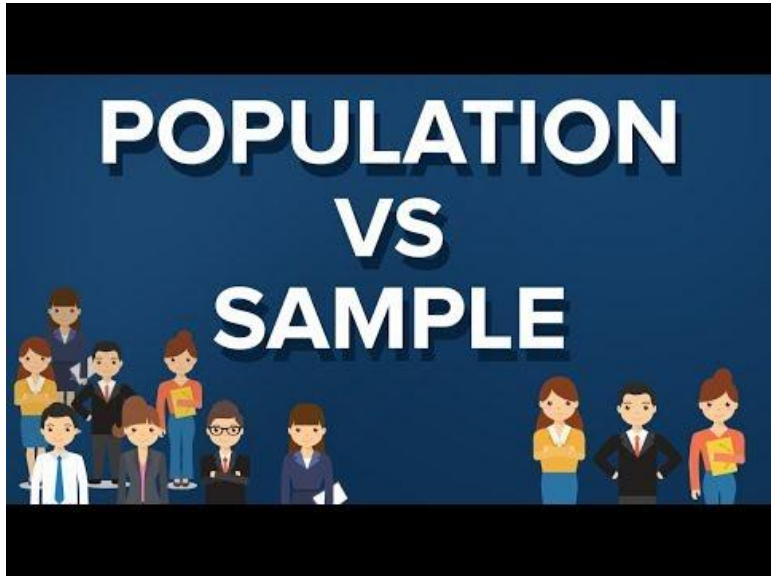
Sources: www.youtube.com, www. Colourbox.com

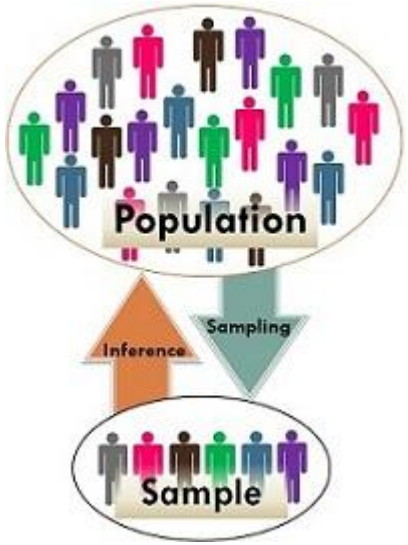
STATISTICS FOR DATA SCIENCE

Population vs. Sample

A **population** is the entire collection of all items(or objects) of interest to our study.

A **sample** is a subset of a population.





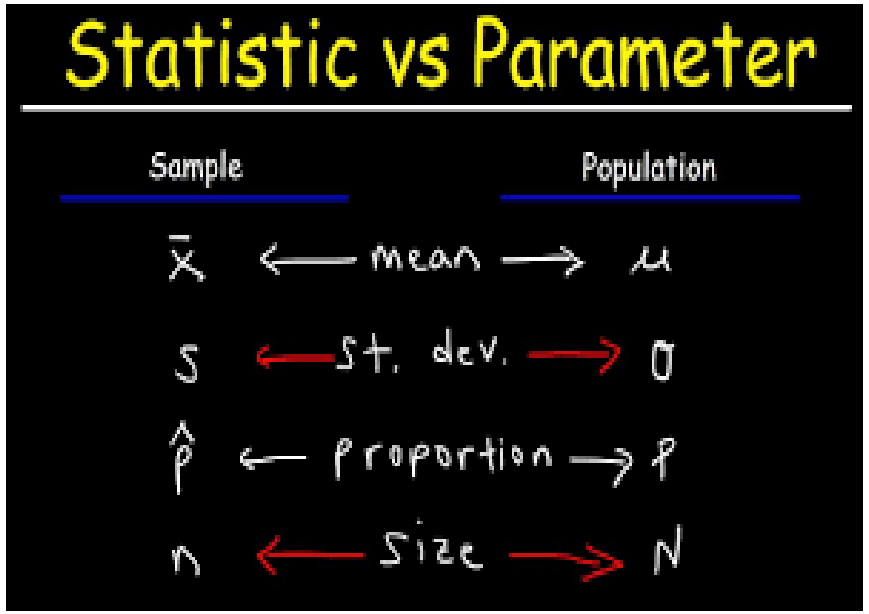
Source Image : 365datascience.com

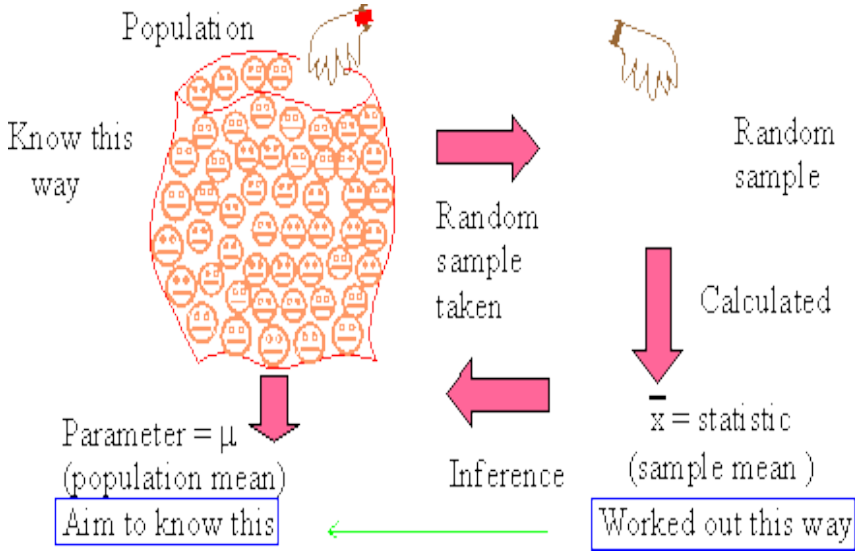
STATISTICS FOR DATA SCIENCE

Parameter vs. Statistic

Parameter is a numerical measurement describing some **characteristic** of a **population**.

Statistic is a numerical measurement describing some **characteristic** of a **sample.**





Sources: youtube..com, Pinkmonkey.com

STATISTICS FOR DATA SCIENCE

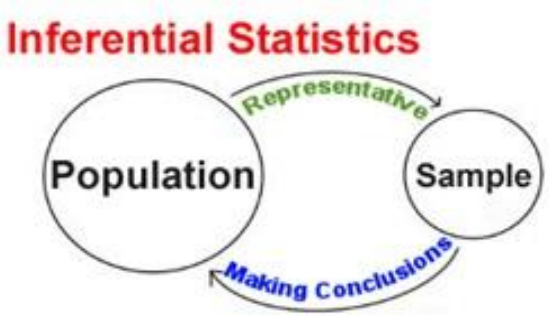
Processes of statistics

Statistics comprises of two processes.

1. **Describing set of data**

2. **Drawing conclusions**
   (making estimates, decisions,
   predictions, about
   set of data based on sampling)

- Measures of Central Tendency
- Measures of Dispersion/Spread
- How it gets accumulates?



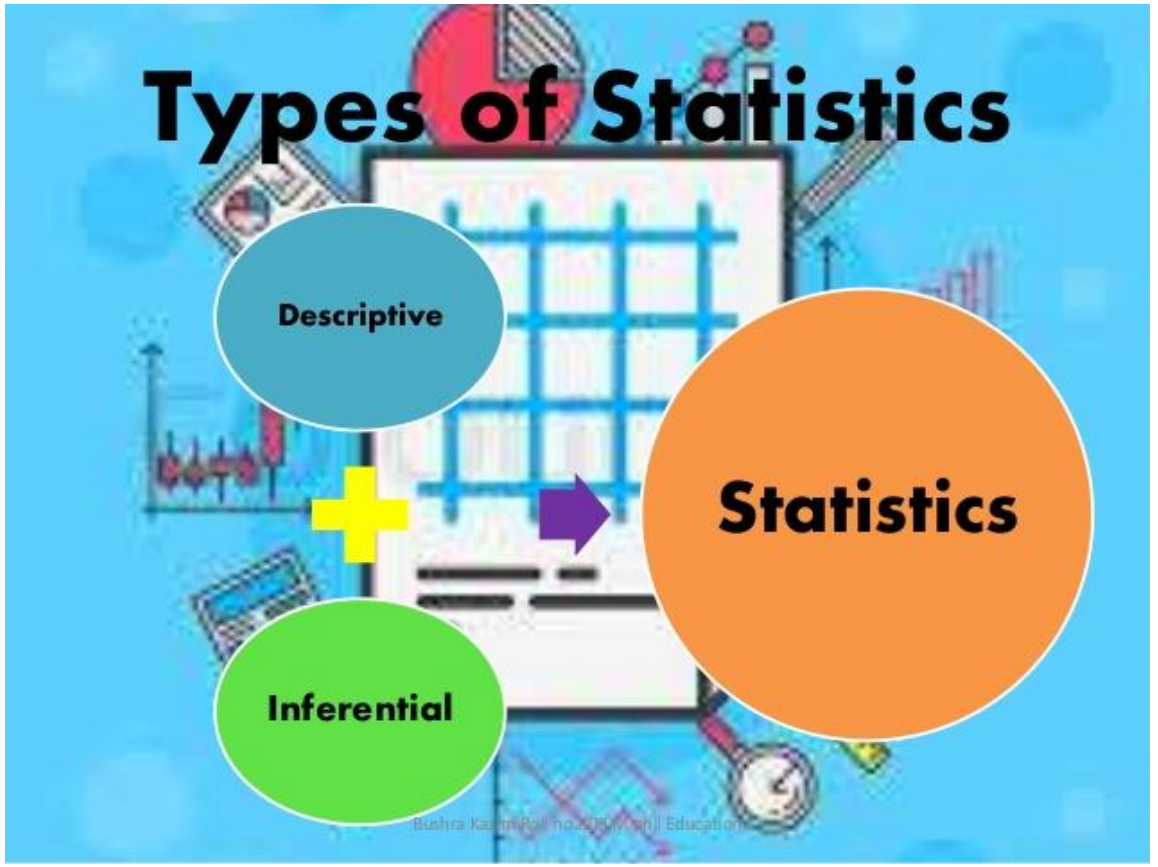
Sources: www.analyticsvidhya.com

STATISTICS FOR DATA SCIENCE

Types of Statistics

Numerical methods to organize, summarize and display data

Draws inferences from a population using sample



Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

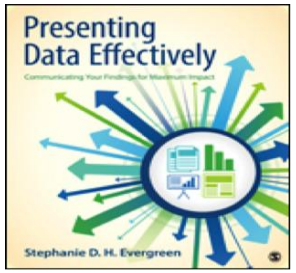
Descriptive statistics

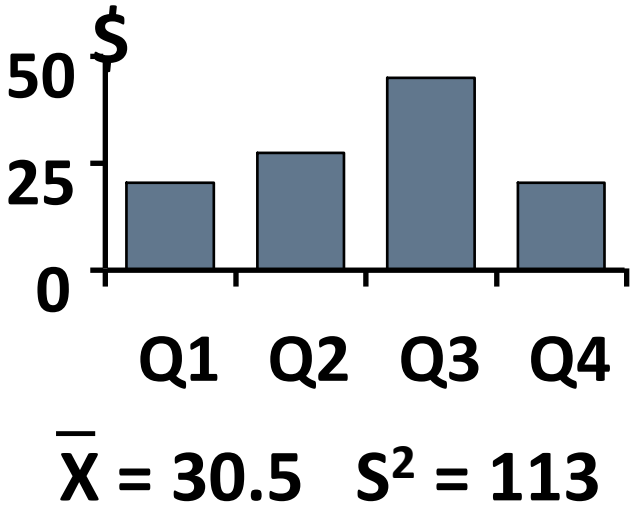
- Collect Data
  - e.g. Survey

Purpose

- Describe Data

- Present Data
  - e.g. Tables and graphs





$\bar{X} = 30.5 \quad S^2 = 113$

- Characterize Data
  - e.g. Sample mean $\bar{X} = \frac{\sum x}{n}$

Source: luminousmen.com/post/

STATISTICS FOR DATA SCIENCE

Why Descriptive Statistics?

An Illustration : Which Group is Smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Source: www,slideshare.net

STATISTICS FOR DATA SCIENCE

Descriptive statistics

An Illustration : Which Group is Smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

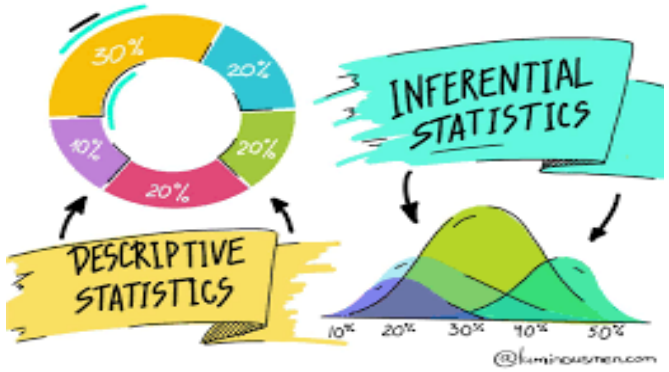


Figure speaks it all !!!

Which group is smarter now?

Class A--Average IQ	Class B--Average IQ
110.54	110.23

They're roughly the same!

With a summary descriptive statistic, it is much easier to answer our question.

Source: luminousmen.com/post/, www.slideshare.net

Example-Descriptive Statistics

In a recent study, volunteers who had less than 6 hours of sleep were four times more likely to answer incorrectly on a science test than were participants who had at least 8 hours of sleep. Decide which part is the descriptive statistic and what conclusion might be drawn using inferential statistics.

The statement "four times more likely to answer incorrectly" is a descriptive statistic. An inference drawn from the sample is that all individuals sleeping less than 6 hours are more likely to answer science question incorrectly than individuals who sleep at least 8 hours.

STATISTICS FOR DATA SCIENCE

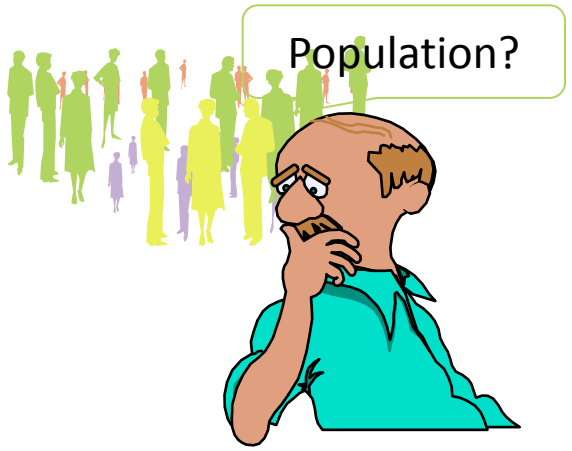
Inferential statistics

- Involves Estimation
  - e.g. Population Parameters
- Hypothesis Testing

Inferential Statistics: Making decisions and drawing conclusions about populations.

Purpose

- Make decision about population characteristics.



Inferential statistics utilizes sample data to make estimates, decisions, predictions or other generalizations about a larger set of data.

Source: luminousmen.com/post/

STATISTICS FOR DATA SCIENCE

Why inferential statistics?

Suppose you want to know the mean income of the subscribers of Netflix

Mean (μ) — a parameter of a population.

You draw a random sample of 100 subscribers and determine that their mean income is $27,500.

Mean($\bar{x}$) = $27,500 (a statistic).

Conclusion : You conclude that the population mean income μ is likely to be close to $27,500 as well.

This example is one of statistical inference.

Source: luminousmen.com/post/

STATISTICS FOR DATA SCIENCE

Descriptive vs. Inferential statistics

Descriptive Statistics

- Organize
- Summarize
- Simplify
- Presentation of data

↓

Describing data

Inferential Statistics

- Generalize from samples to population
- Hypothesis testing
- Relationships among variables

↓

Make predictions

Something to know about !!!!

When we gather data, we want to uncover the "information" in it. One easy way to do that is to think of: "Shape –Position-Spread"

Shape – What is the shape of the histogram?

Position – What is the mean or median?

Spread – What is the range or standard deviation?

Source: sixsigma-institute.org

STATISTICS FOR DATA SCIENCE

Types of Descriptive Statistics

- Organize Data
  - Tables
  - Graphs

→

- Organize Data
  - Tables
    - Frequency Distributions
    - Relative Frequency Distributions
  - Graphs
    - Bar Chart or Histogram
    - Stem and Leaf Plot
    - Frequency Polygon

- Summarize Data
  - Central Tendency
  - Variation

→

Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode
- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

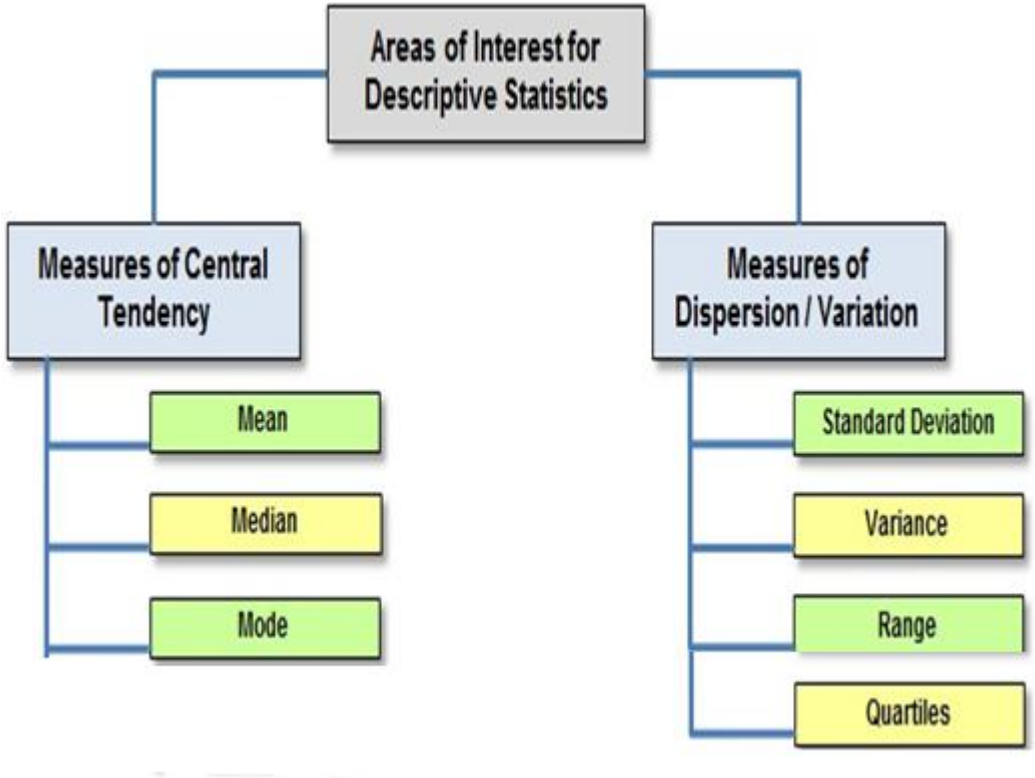
Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Descriptive statistics

- Descriptive Statistics is a method of organizing, summarizing, and presenting data in a convenient and informative way.

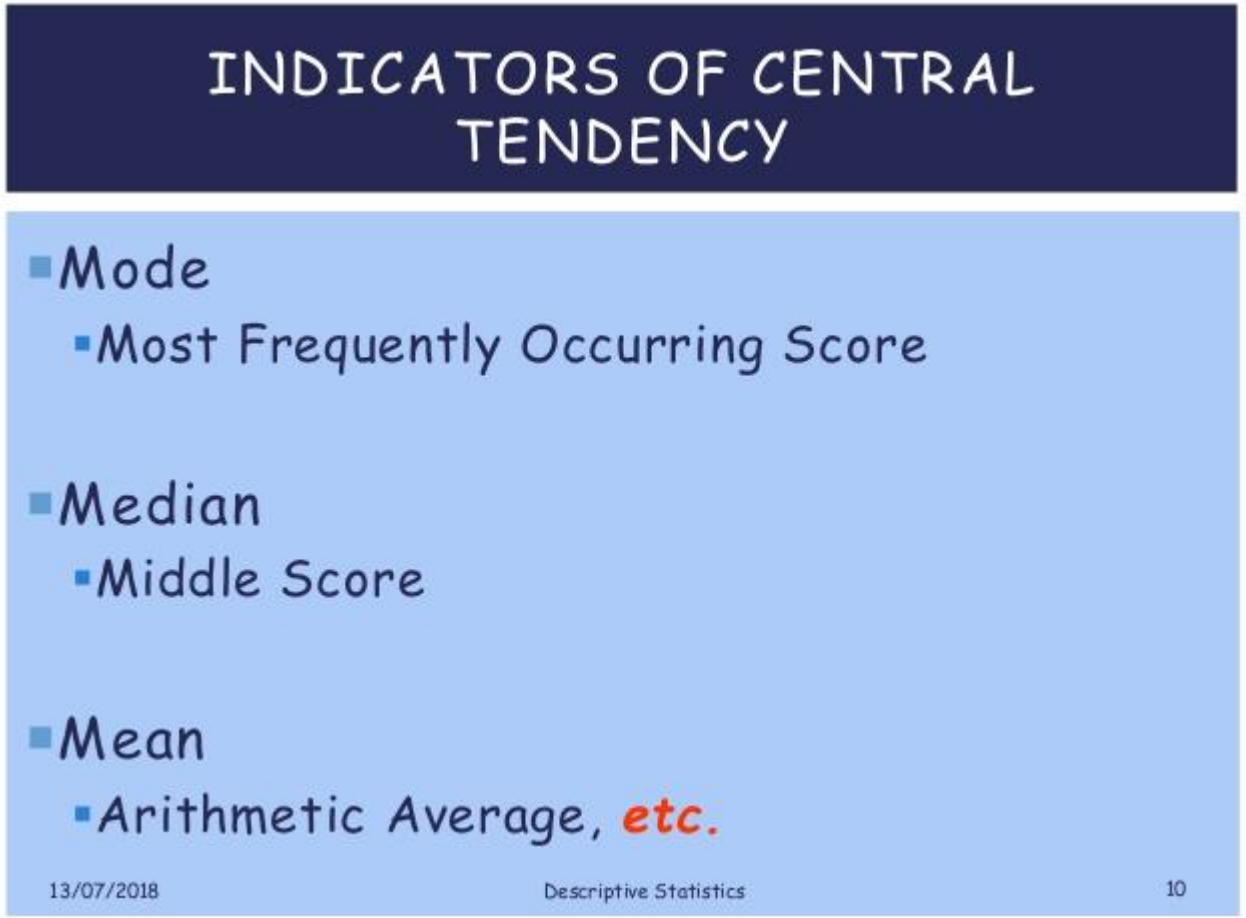
- The actual method used depends on what information we would like to extract.



Source: sixsigma-institute.org

STATISTICS FOR DATA SCIENCE

Descriptive statistics



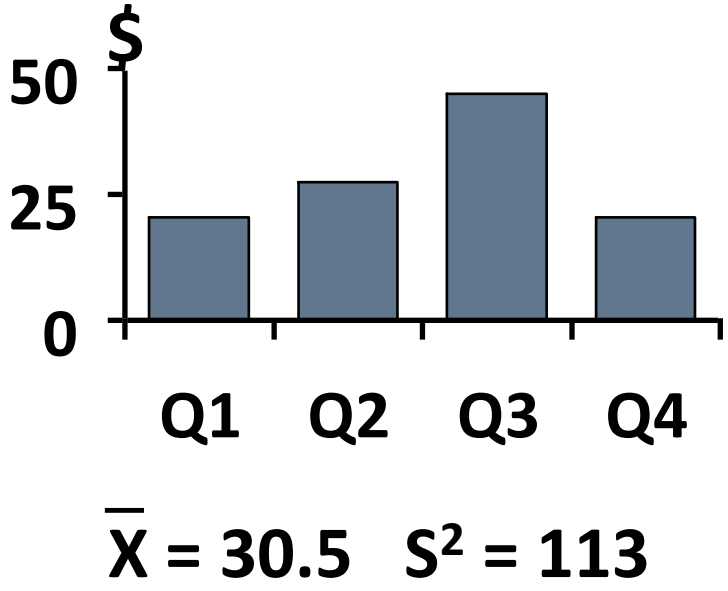
Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Descriptive statistics

- **Descriptive statistics** are methods for organizing and summarizing data.

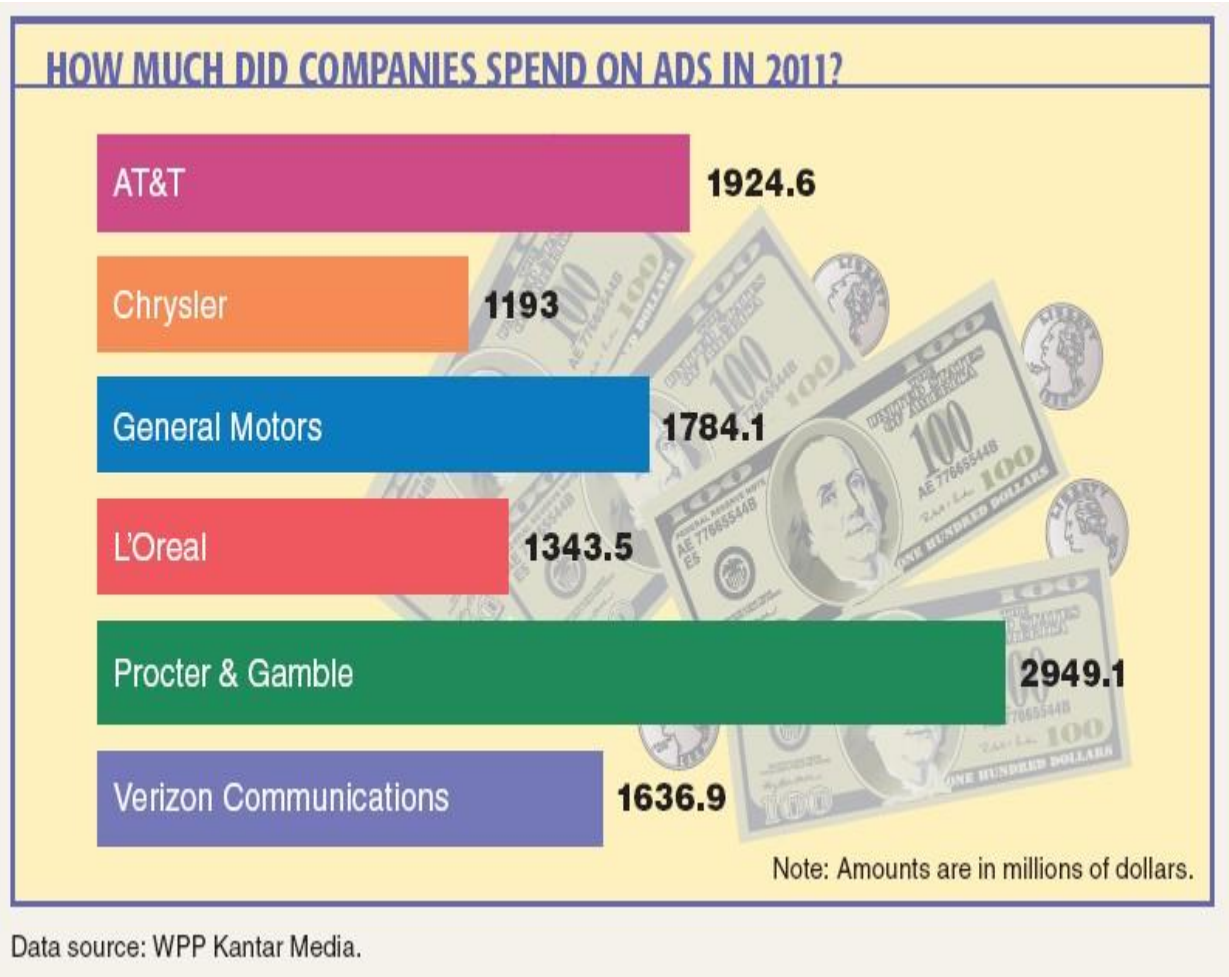
- For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.



Source: luminousmen.com/post/

STATISTICS FOR DATA SCIENCE

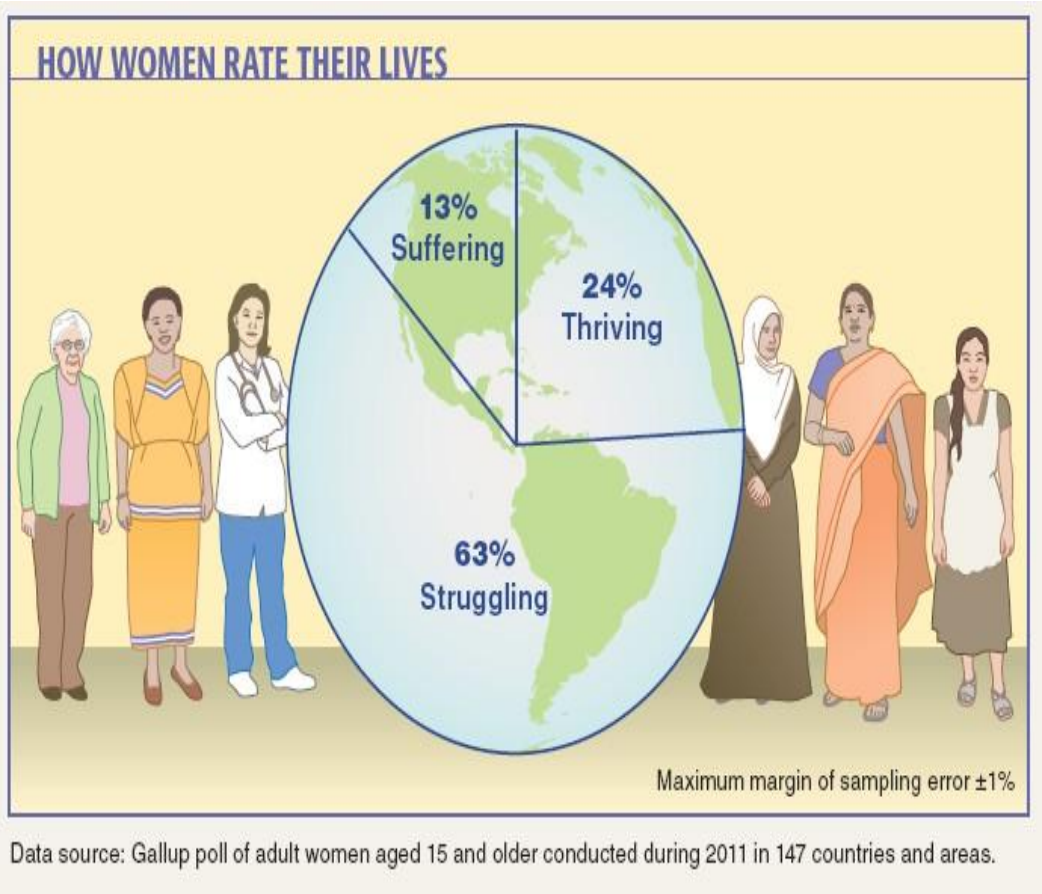
Descriptive statistics - Example



Source: luminousmen.com/post/

STATISTICS FOR DATA SCIENCE

Descriptive statistics - Example



Data source: Gallup poll of adult women aged 15 and older conducted during 2011 in 147 countries and areas.

Source: luminousmen.com/post/

Descriptive statistics - Problem

Problem:

Calculate the average number of truck shipments from the United States to five Canadian cities for the following data given in thousands of bags:

Montreal, 64.0; Ottawa, 15.0; Toronto, 285.0; Vancouver, 228.0; Winnipeg, 45.0

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency

There are three different types of 'average'. These are the *mean*, the *median* and the *mode*.

They are used by statisticians as a way of summarizing where the 'centre' of the data is.

$$\begin{aligned} Mean &= \frac{sum\ of\ all\ values}{total\ number\ of\ values} \\ Median &= middle\ value\ (when\ the\ data\ are\ arranged\ in\ order) \\ Mode &= most\ common\ value \end{aligned}$$

Source: sixsigma-institute.org

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency: Mean

- Mean is the arithmetic average computed by summing all the values in the dataset and dividing the sum by the number of data values.

- The population mean is represented by Greek letter μ.

- For a finite set of dataset with measurement values X1, X2, …., Xn (a set of n numbers), it is defined by the formula:

$$\mu_x = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

$$\mu_x = \frac{\Sigma X}{N} \quad \textbf{mean of a population}$$

$$\bar{X} = \frac{\Sigma X}{n} \quad \textbf{mean of a sample}$$

Source: sixsigma-institute.org

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency: Mean

Mean

Annual Salary: 10k 11k 11k 15k 15k 15k 19k 20k 21k 21k 22k 22k 24k 25k

$$\bar{X} = \frac{\sum X}{n} \qquad \bar{X} = \frac{251}{14}$$

Mean = 17.9 k.y^{-1}

Disadvantages

- Very sensitive measure
- Can only be used on interval or ratio data

Advantages

- Very sensitive measure
- Takes into account all the available information
- Can be combined with means of other groups to give the overall mean

Source: slideshare.net

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency: Mean

1. Add all the values to get the sum.

2. To find the mean, divide the sum by the number of data values (i.e. n).

Consider the data given below:

5, 9, 12, 4, 5, 14, 19, 16, 3, 5, 7

Find the Mean:

1. sum= 5 + 9 + 12 + 4 + 5 + 14 + 19 + 16 + 3 + 5 + 7=99
2. mean = sum/ no. of values= 99/11= 9.

Sometimes the mean will not appear in the original list.
It might even be a decimal value.

Measures of Central Tendency: Mean

Advantages:

Takes into account every number in the data set. That means all numbers are included in calculating the mean.

Easy and quick way to represent the entire data values by a single or unique number due to its straightforward method of calculation.

Each set has a unique mean value.

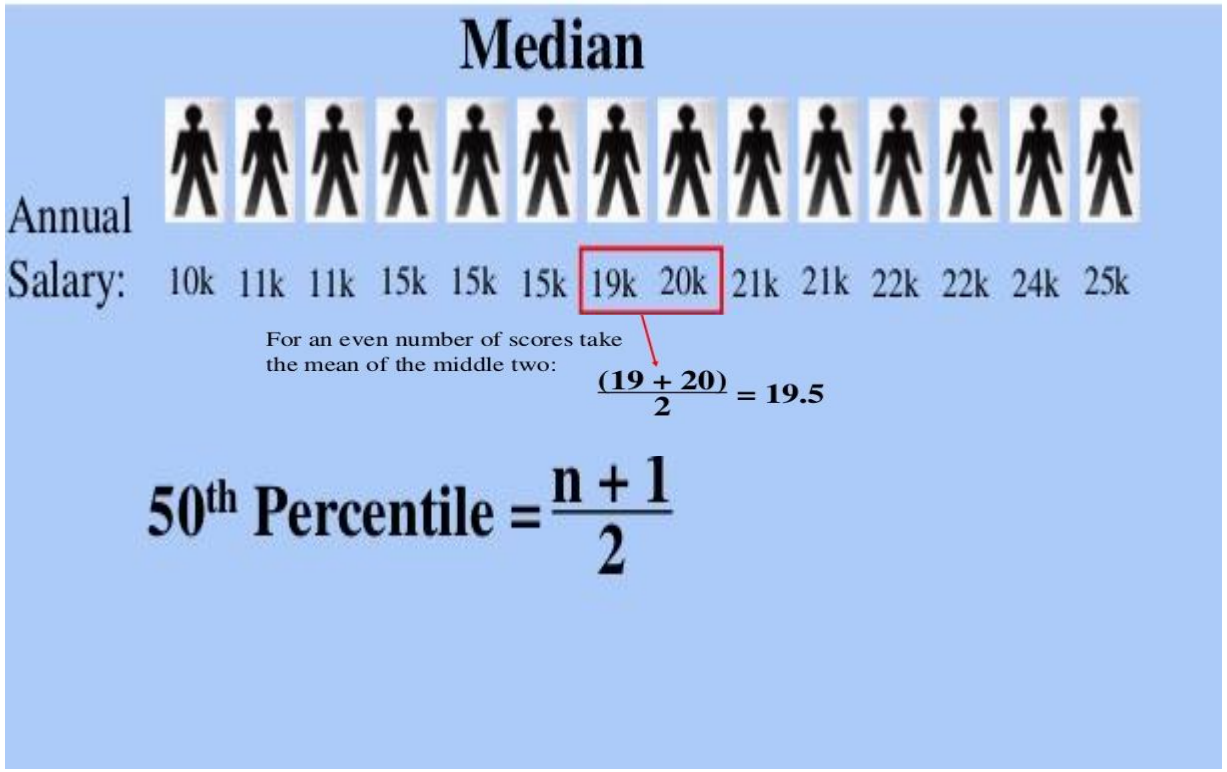
Disadvantages:

Its value is easily affected by extreme values known as the outliers.

Source : Chilimath

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency: Median



Advantages

- Unaffected by extreme scores
- Can be used at all levels above nominal.

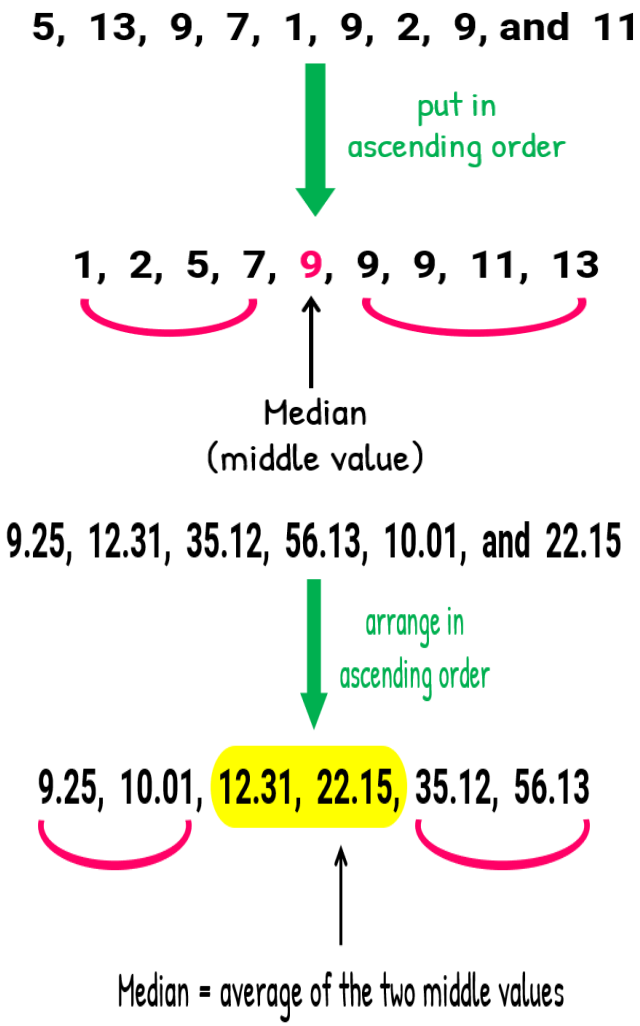
Disadvantages

- Only considers order- value ignored.

Source: slideshare.net

Measures of Central Tendency: Median

1. Arrange all the values in ascending order.
2. Find the middle position.
3. The element corresponding to middle position is considered as median(if odd number of elements are present).
4. If there are even number of elements present then the average of the elements present in the middle positions is considered as median.



Measures of Central Tendency: Median- Example

Consider the data given below:

5, 9, 12, 4, 5, 14, 19, 16, 3, 5, 7 **(n=11)**

The Median

To calculate the median, we need to put the numbers in order and find the middle value.

3 4 5 5 5 **7** 9 12 14 16 19

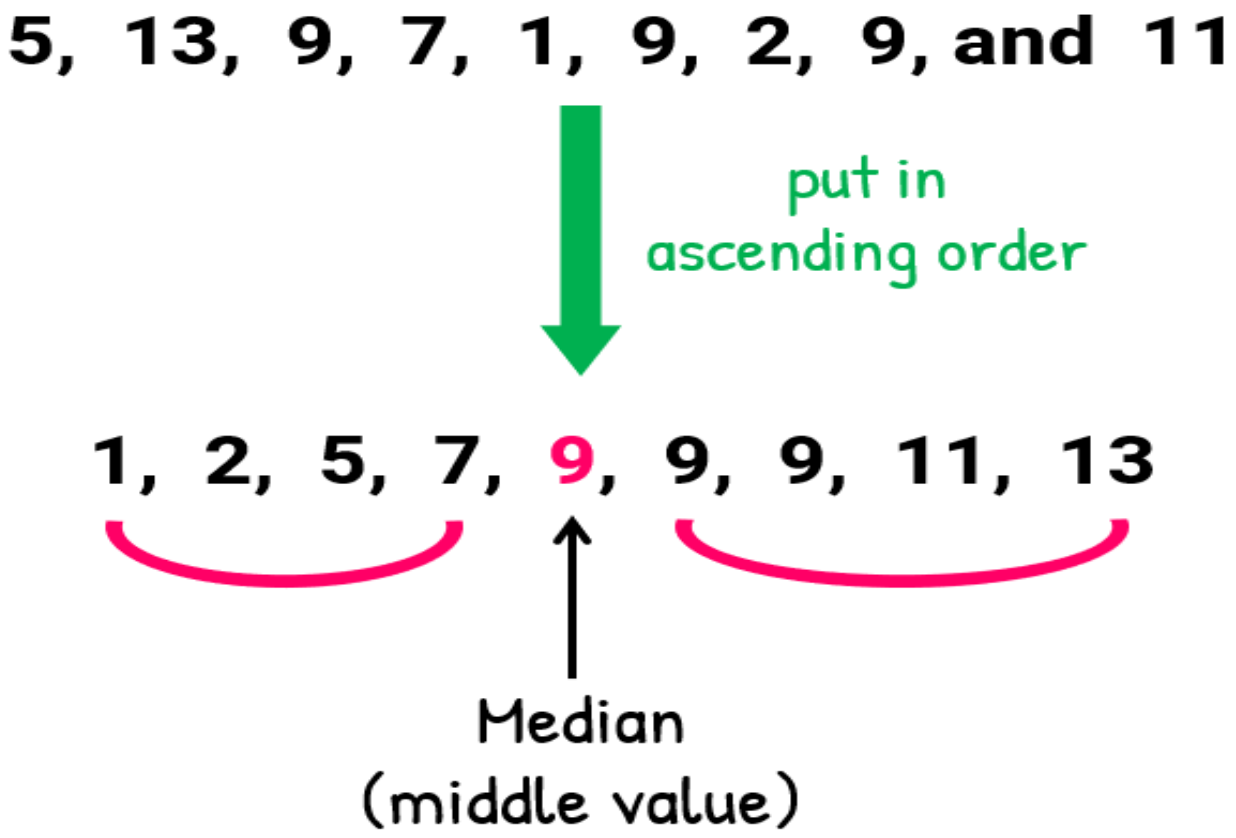
Here the *median* is 7 because this is the middle value.

Half of the other values in the list are below 7 and half are above 7.

Source: sixsigma-institute.org

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency: Median- Example



Source: Chilimath

Measures of Central Tendency: Median- Example

Consider the data given below:

3, 6, 7, 8, 11, 15 **(n=6)**

When there are an even number of values, there is no clear middle value.
In this case, there are two middle values.

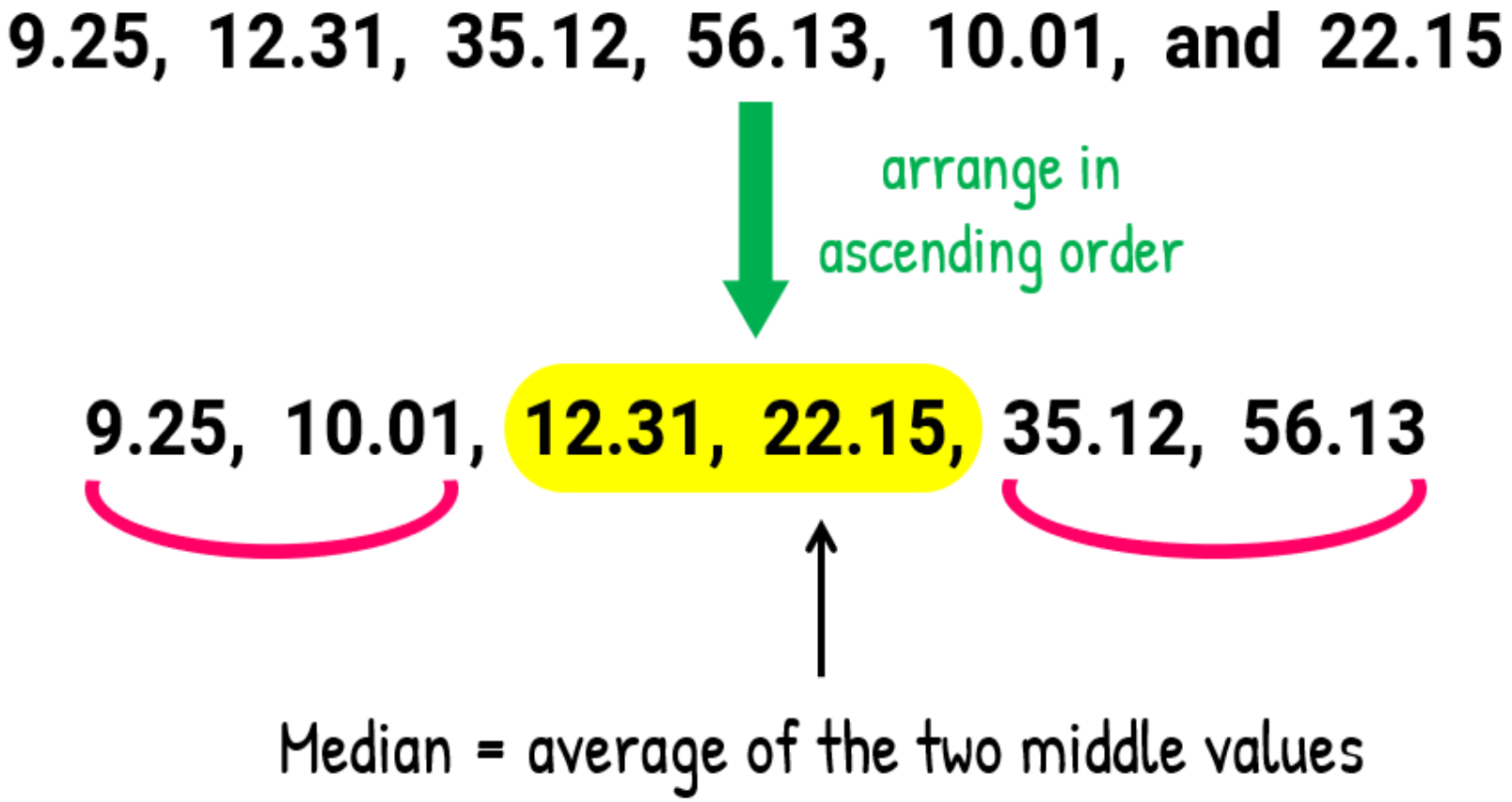
3 6 **7** **8** 11 15

The median is the *mean* of these two middle numbers.7 + 8 / 2 =7.5
So the median for this set of values is **7.5**.

Like the mean, the median value does not always appear in the original list of values.

Source: sixsigma-institute.org

Measures of Central Tendency: Median- Example



Source: Chilimath

Measures of Central Tendency: Median

Advantages:

Not affected by the outliers in the data set.

An outlier is a data point that is radically “distant” or “away” from common trends of values in a given set.

It does not represent a typical number in the set.

The concept of the median is intuitive thus can easily be explained as the center value.

Each set has a unique median value.

Disadvantages:

Its value is perceived as it is. It cannot be utilized for further algebraic treatment.

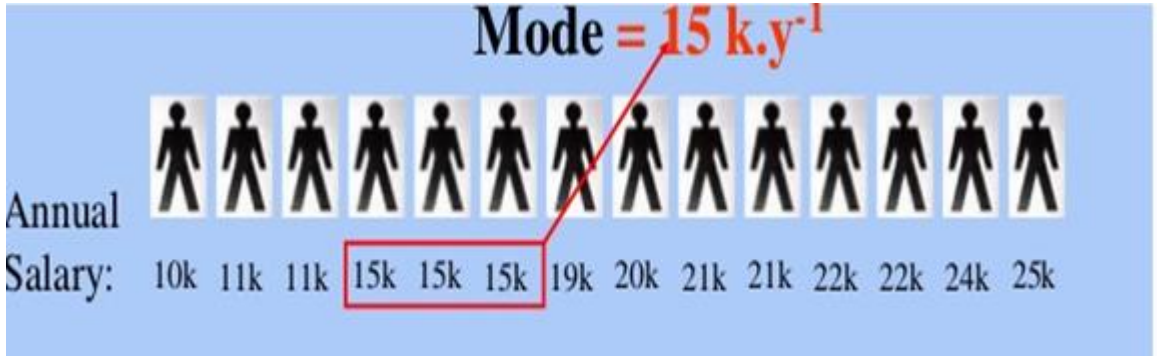
Source : Chilimath

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency: Mode

Mode: Most often value in the data set.

To calculate the mode, we need to look at which value appears the most often.



Advantages

- Quick and easy to compute
- Unaffected by extreme scores
- Can be used at any level of measurement.

Disadvantages

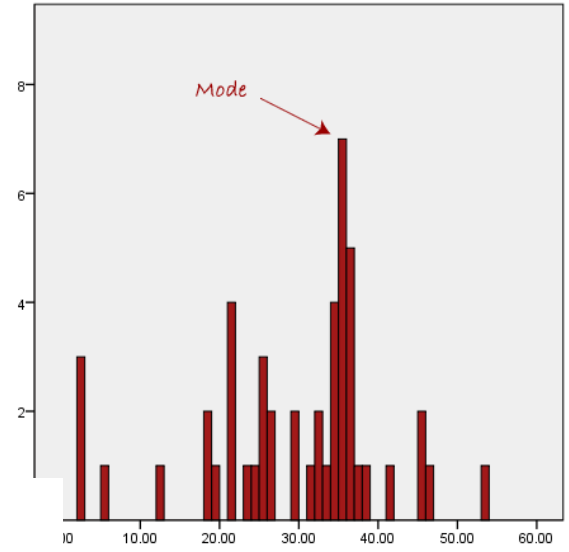
- Terminal Statistic
- A given sub-group could make this measure unrepresentative.

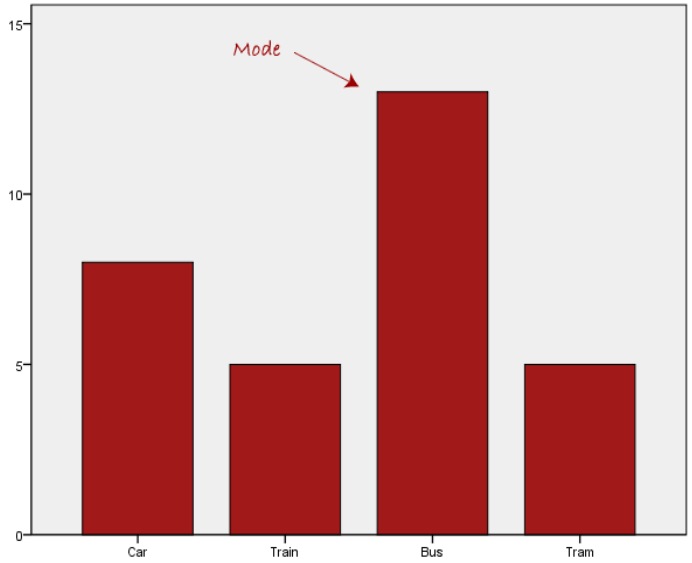
Source: slideshare.net

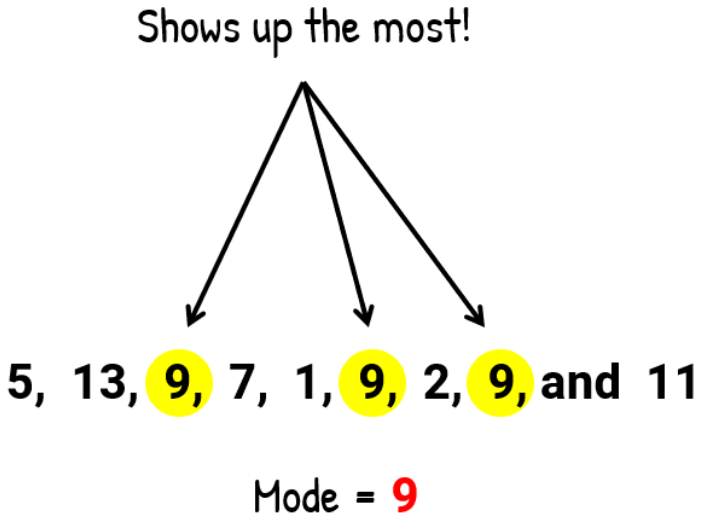
Measures of Central Tendency: Mode

Mode: Most often value in the data set.

To calculate the mode, we need to look at which value appears the most often.







Sources: sixsigma-institute.org, statistics.laerd.com

STATISTICS FOR DATA SCIENCE

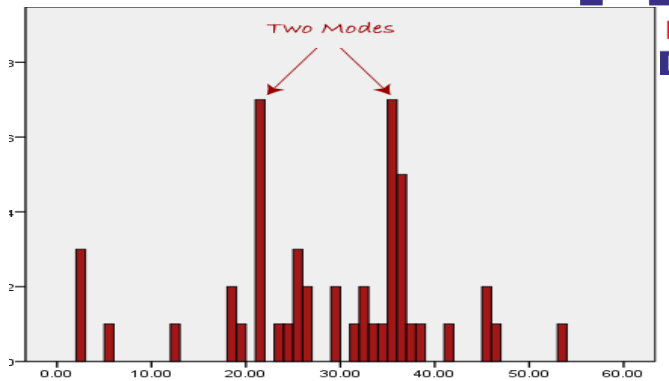
Measures of Central Tendency: Mode - Example

Consider the data given below:

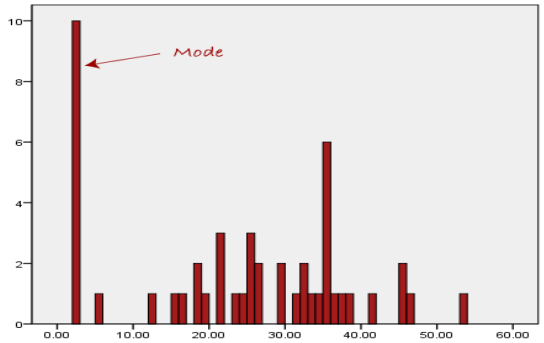
5, 9, 12, 4, 5, 14, 19, 16, 3, 5, 7

3 4 **5** **5** **5** 7 9 12 14 16 19

In this list the *mode* is 5, because it appears most often.



Sometimes there will be more than one mode, because two or more values appear the same number of times.

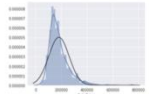


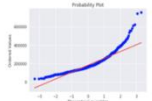
4, 3, 7, 8, 4, 5, 12, 4, 5, 3, 2, and 3

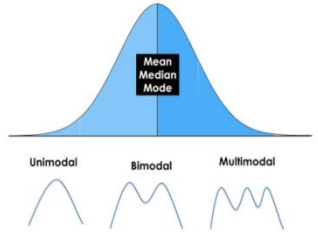
put in ascending order

2, 3, 3, 3, 4, 4, 4 5, 5, 7, 8, 12

Mode = **3** and **4**







Sources: sixsigma-institute.org, statistics.laerd.com, www.towardsdatascience.com, Chilimath.com

Measures of Central Tendency: Mode

Advantages:

Just like the median, the mode is not affected by outliers.

Useful to find the most “popular” or common item. This includes data sets that do not involve numbers.

Disadvantages:

If the set contains no repeating values, the mode is irrelevant.

In contrast, if there are many values that have the same count, then mode can be meaningless.

Source : Chilimath

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency

The most appropriate measure of location depends on ...

the shape of the data's distribution.

- Depends on whether or not data are "**symmetric**" or "**skewed**".

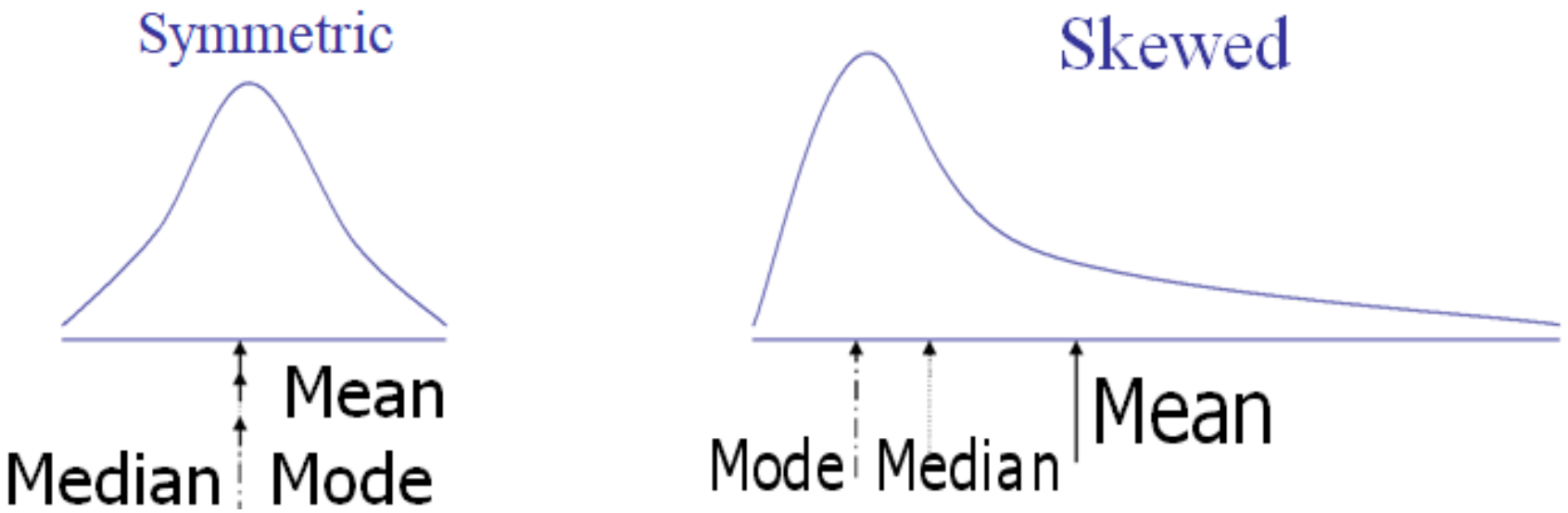
- Depends on whether or not data have one ("**unimodal**") or more ("**multimodal**") modes.

Source:www.slideshare.net

Measures of Central Tendency: Median

In symmetric distributions, the mean, median, and mode are the same.

In skewed data, the mean and median lie further toward the skew than the mode.

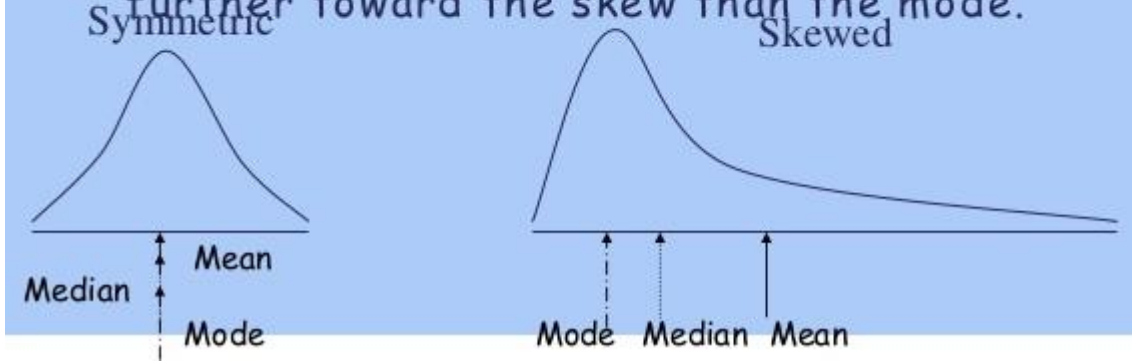


Source:www.slideshare.net

STATISTICS FOR DATA SCIENCE

Measures of Central Tendency: Mode

1. It may give you the most likely experience rather than the "typical" or "central" experience.
2. In symmetric distributions, the mean, median, and mode are the same.
3. In skewed data, the mean and median lie further toward the skew than the mode.



- If the skewness is extreme, the researcher should either transform the data to make them better resemble a normal curve or else use a different set of statistics—nonparametric statistics—to carry out the analysis

- When the median and the mean are different, the distribution is skewed. The greater the difference, the greater the skew.

Source:www.slideshare.net

Problem - Mean, Median and Mode

Alex did a survey of how many games each of his 20 friends owned, and got this:

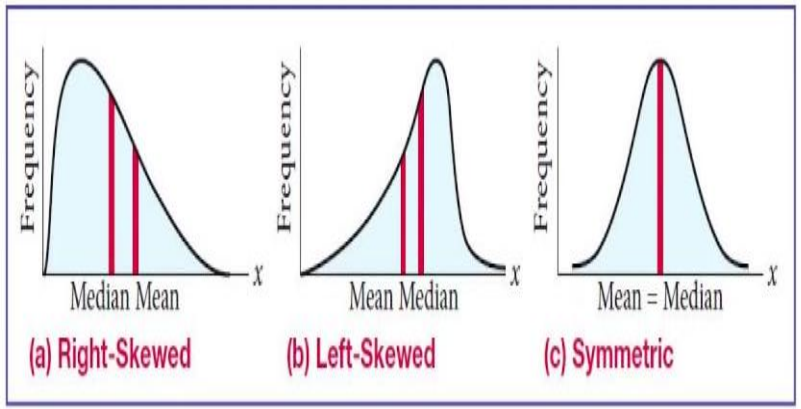
9, 15, 11, 12, 3, 5, 10, 20, 14, 6, 8, 8, 12, 12, 18, 15, 6, 9, 18, 11

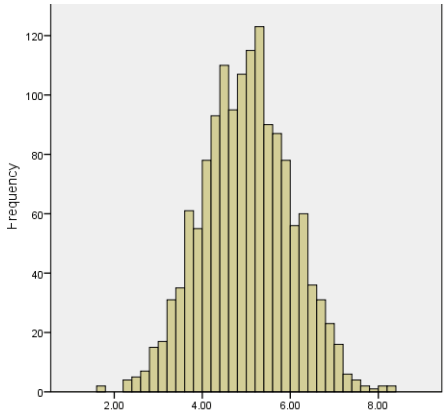
Find the mean, median and mode

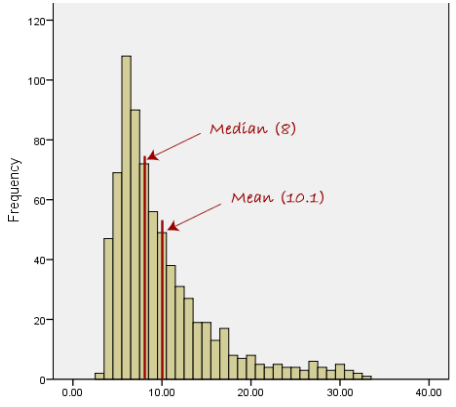
Symmetric and Skewed Distributions:

Symmetric Data: Data sets whose values are evenly spread around the center.

Skewed Data: Data sets that are not symmetric.







Source: sixsigma-institute.org, statistics.laerd.com

Shape: The "shape" of the data is called its "distribution".

If **mean = median = mode**, the shape of the distribution is **symmetric**.

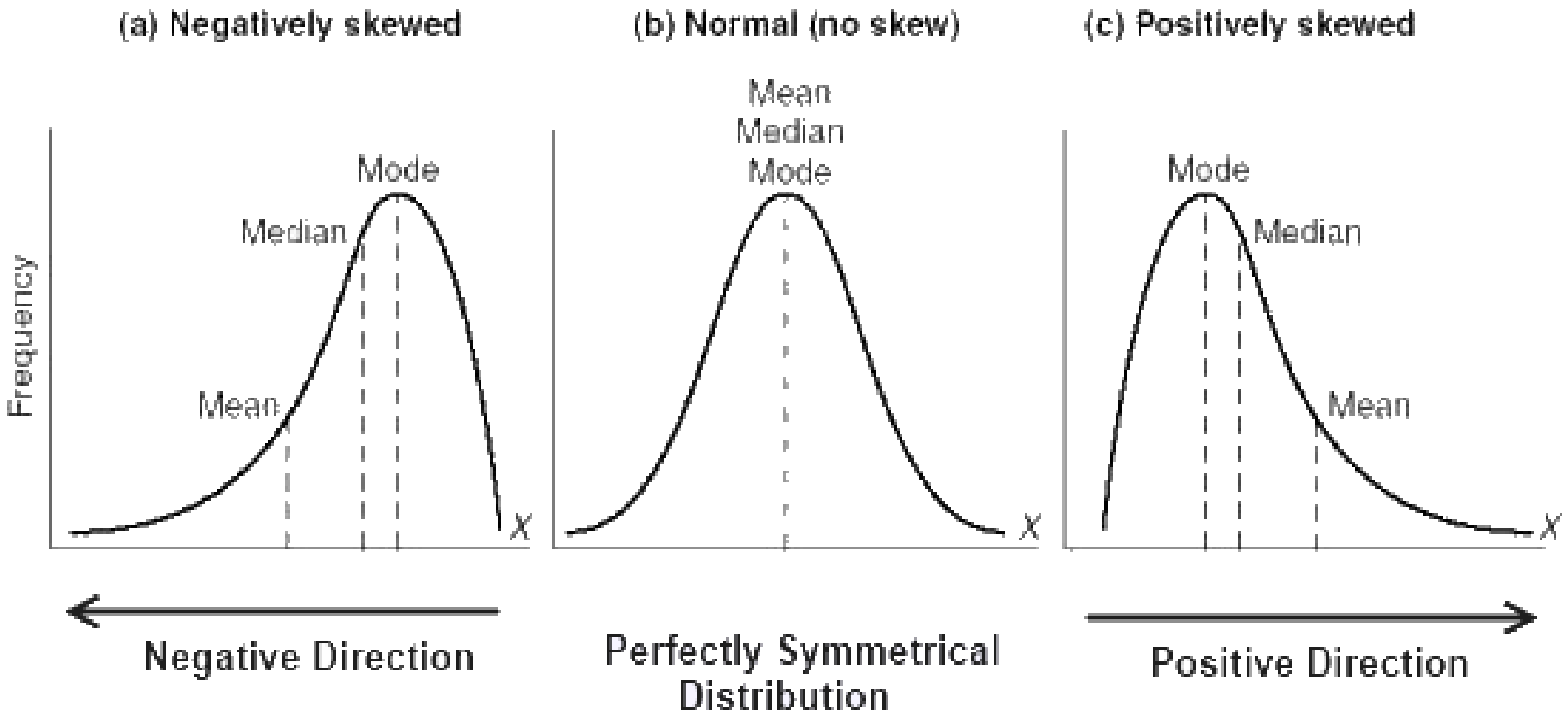
- If **mode < median < mean**, the shape of the distribution trails to the right, is **positively skewed**.

- If **mean < median < mode**, the shape of the distribution trails to the left, is **negatively skewed.**

- Distributions of various "shapes" have different properties and names such as the "**normal" distribution**, which is also known as the "**bell curve**" (among mathematicians it is called the **Gaussian**)

STATISTICS FOR DATA SCIENCE

Symmetrical vs Skewed data



Source: https://medium.com/analytics-vidhya/descriptive-statistics-in-data-science-with-illustrations-in-python-efd5ccc152f1

Measures of Central Tendency

- **Quantitative data**:
  - Mode – the most frequently occurring observation
  - Median – the middle value in the data
  - Mean – arithmetic average

- **Qualitative data:**
  - Mode – always appropriate

    Ex : Maximum Type of Color
  - Mean – never appropriate

    Ex : Average value of Yellow color

STATISTICS FOR DATA SCIENCE

When to use Mean, Median and Mode?

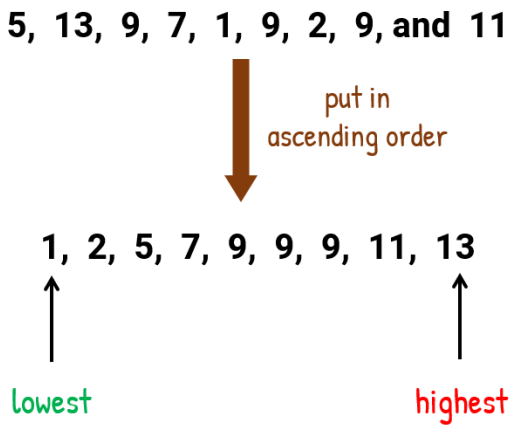
TYPE OF VARIABLE	BEST MEASURE OF CENTRAL TENDENCY
Nominal	Mode
Ordinal	Median
Interval / Ratio (not skewed)	Mean
Interval / Ratio (skewed)	Median

Source: sixsigma-institute.org, statistics.laerd.com

STATISTICS FOR DATA SCIENCE

Measures of spread: Range

Range = Maximum Value – Minimum Value



AGES OF STUDENTS

13,13,14,14,14,15,15,15,15,16,16,16

Range = highest - lowest

= 16 - 13

Range = 3

Case 1

AGES OF STUDENTS

11,13,13,14,14,15,15,15,15,16,16,18

Range = highest - lowest

= 18 - 11

Range = 7

Case 2

Observations:

Since the range of Class A is **smaller** than in Class B, can we claim that the age distribution in Class A is more clustered (closely related) than in Class B? In other words, are the ages listed in Class A more uniform than in Class B?

Source: Chilimath.com

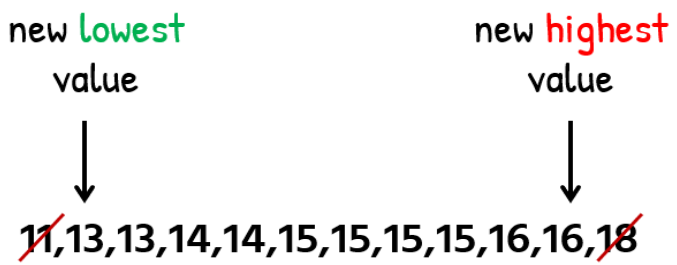
STATISTICS FOR DATA SCIENCE

Measures of spread: Range

Limitations:

1. Using the range to describe the spread of data within a set.

2. It can drastically be affected by outliers (values that are not typical as compared to the rest of the elements in the set).



Range = highest - lowest
= 16 - 13
Range = 3

Case 3

Source: Chilimath.com

STATISTICS FOR DATA SCIENCE

Measures of spread: Range

Range (in statistics) is the difference between the maximum and minimum value of the set. What the range provides is a quick and rough estimate of the spread of data values within a set.

Here we have two classes taking Data Science and the ages of the students in each class.

AGES OF STUDENTS

13,13,14,14,14,15,15,15,15,16,16,16

Range = highest - lowest

= 16 - 13

Range = 3

Case 1

AGES OF STUDENTS

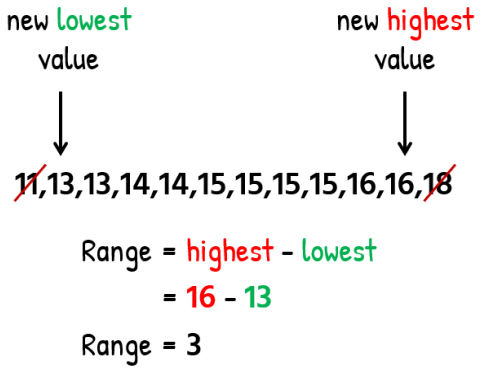
11,13,13,14,14,15,15,15,15,16,16,18

Range = highest - lowest

= 18 - 11

Range = 7

Case 2



Case 3

Source: Chilimath.com

Measures of spread: Range

Advantages:

Just like the median, the mode is not affected by outliers.

Useful to find the most “popular” or common item. This includes data sets that do not involve numbers.

Disadvantages:

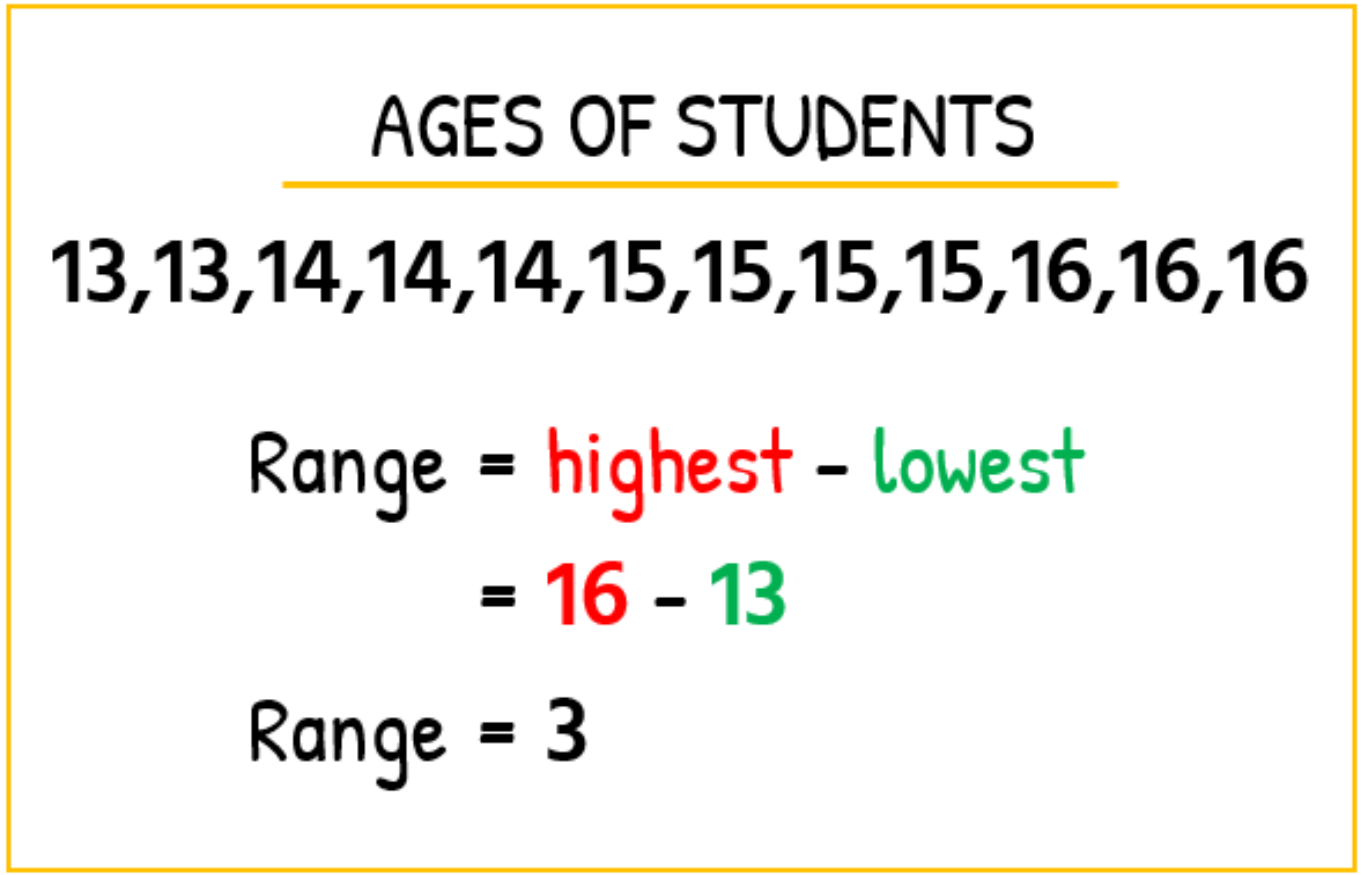
If the set contains no repeating values, the mode is irrelevant.

In contrast, if there are many values that have the same count, then mode can be meaningless.

Source: Chilimath.com

STATISTICS FOR DATA SCIENCE

Measures of Dispersion: Range - Example



Source: Plantlet

STATISTICS FOR DATA SCIENCE

Measures of Dispersion: Range - Example

Source: Chilimath

The Range Can Be Misleading

The range can sometimes be misleading when there are extremely high or low values.

Example: In **{8, 11, 5, 9, 7, 6, 3616}**:

The lowest value is 5,

and the highest is 3616,

So the range is 3616-5 = **3611**.

The single value of 3616 makes the range large, but most values are around 10.

THANK YOU

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering