PES UNIVERSITY

CELEBRATING 50 YEARS

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS
UE23MA242A

Unit 1: Types of Statistics & Summary Statistics

Mamatha.H.R

Department of Computer Science and Engineering

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Unit 1: Types of Statistics & Summary Statistics

Mamatha H R

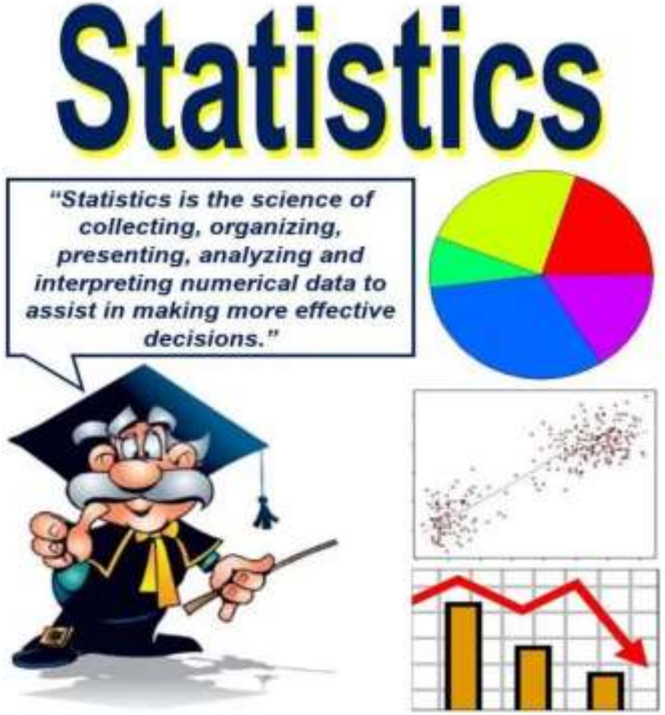
Department of Computer Science and Engineering

Topics to be covered

- Statistics
- Types of Statistics
- Summary Statistics

Statistics

- Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical information.
- It involves study and manipulation of data, including ways to gather, review, analyze, and draw conclusions from data.


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

Source: i0.wp.com

Slide Courtesy:Dr.Uma

Statistics

Statistics involves:

- Collecting Data
  - Ex: Survey
- Presenting Data
  - Ex: Charts & Tables
- Characterizing Data
  - Ex: Average

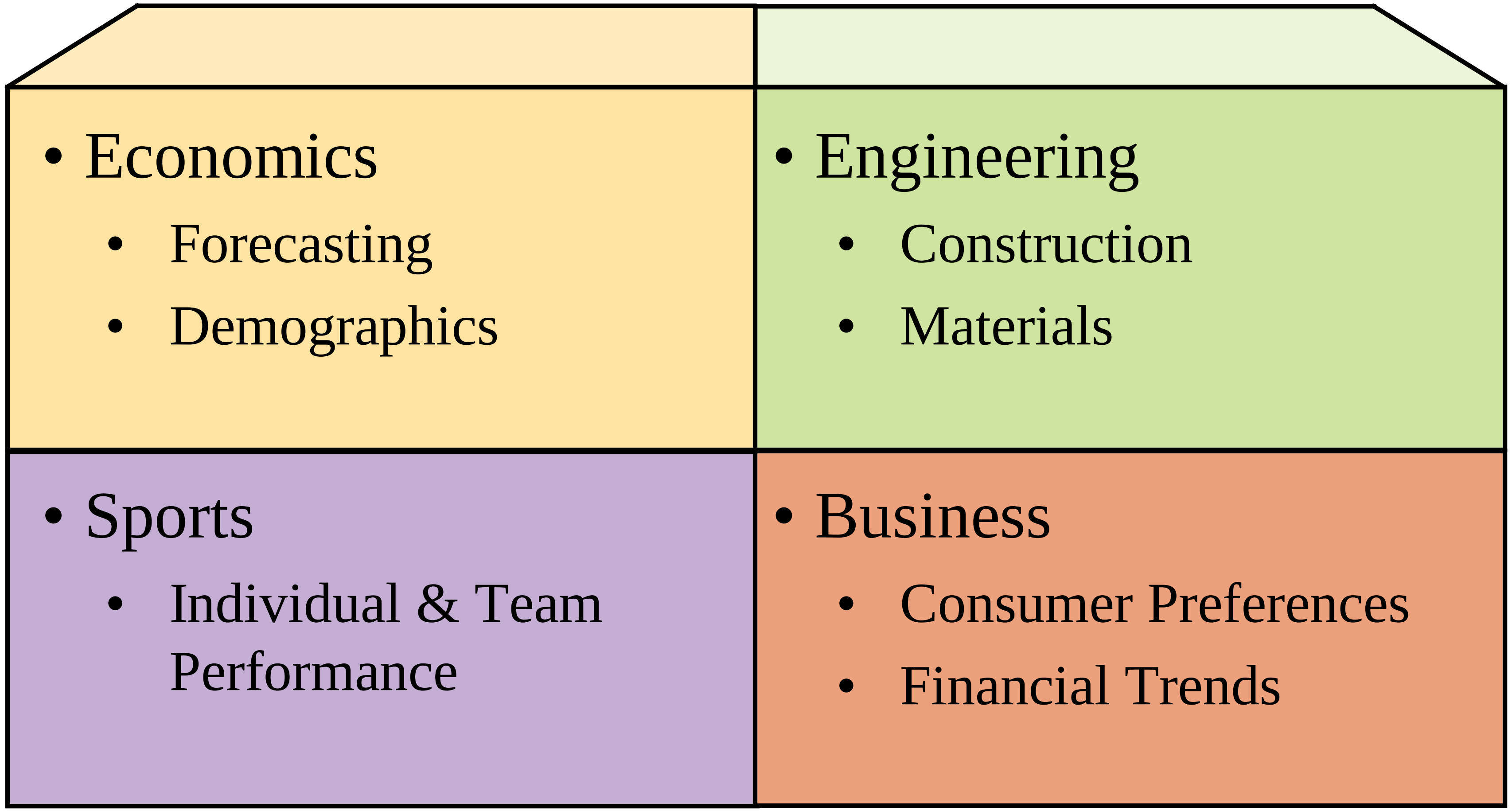
Slide Courtesy:Dr.Uma

Why do we need to know about statistics

- To know how to properly present information.
- To know how to draw conclusions about populations based on sample information.
- To know how to improve processes.
- To know how to obtain reliable forecasts.
- To find out why a process behaves the way it does.
- To find out 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.

Slide Courtesy:Dr.Uma

Applications of Statistics



- Economics
  - Forecasting
  - Demographics

- Engineering
  - Construction
  - Materials

- Sports
  - Individual & Team Performance

- Business
  - Consumer Preferences
  - Financial Trends

Processes of statistics

Statistics involves 2 main processes:

1. **Describing** sets of data.

1. **Drawing conclusions** (making estimates, decisions, predictions, etc) about sets of data based on sampling.

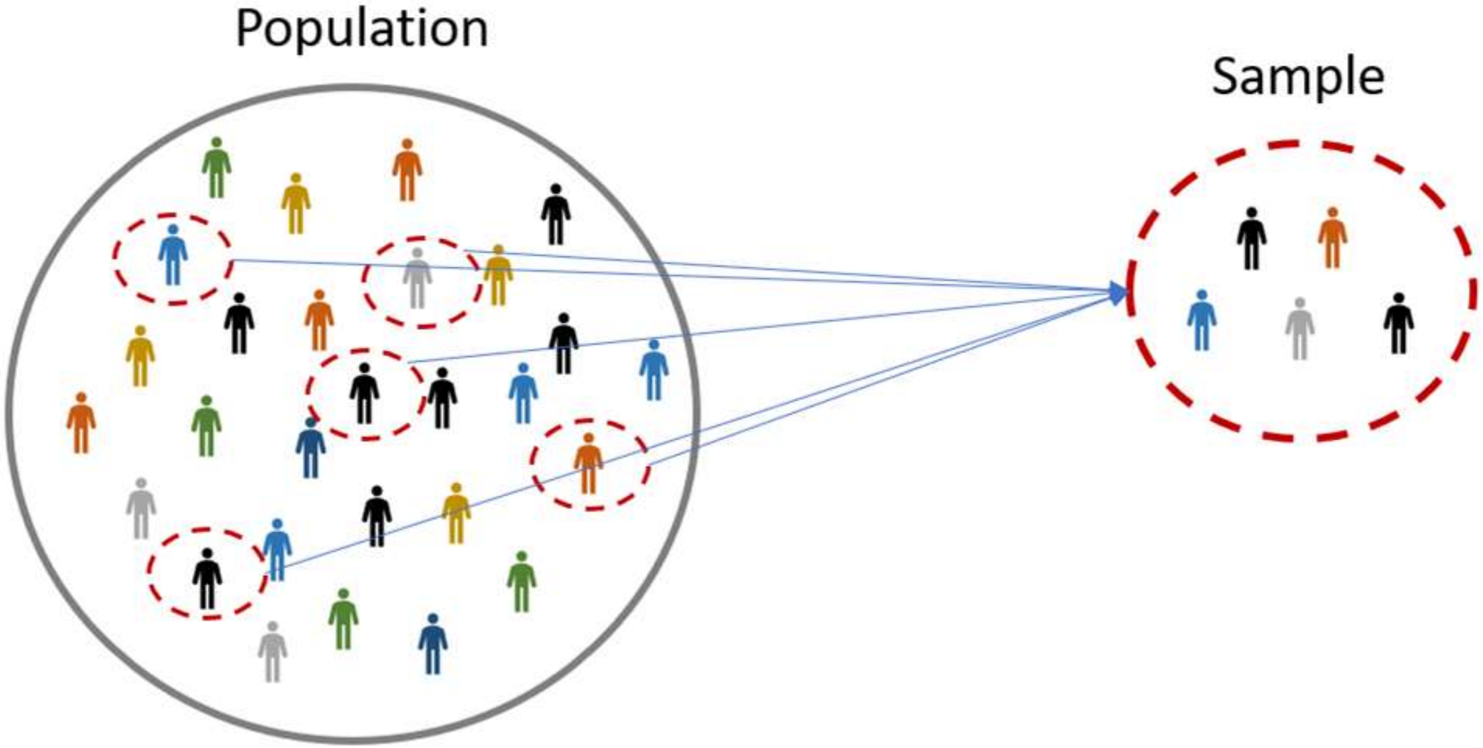
Population and Sample

POPULATION

A population is the **entire collection of objects or outcomes** about which information is sought.

SAMPLE

A sample is a **subset of a population**, containing the objects or outcomes that are actually observed.



Source: sigmamagic.com

Sample Statistic and Population Parameter

➔**Sample statistic:**

- It is a numerical measurement describing some **characteristic** of a **sample.**
- Example: sample average, median, sample standard deviation, and percentiles.

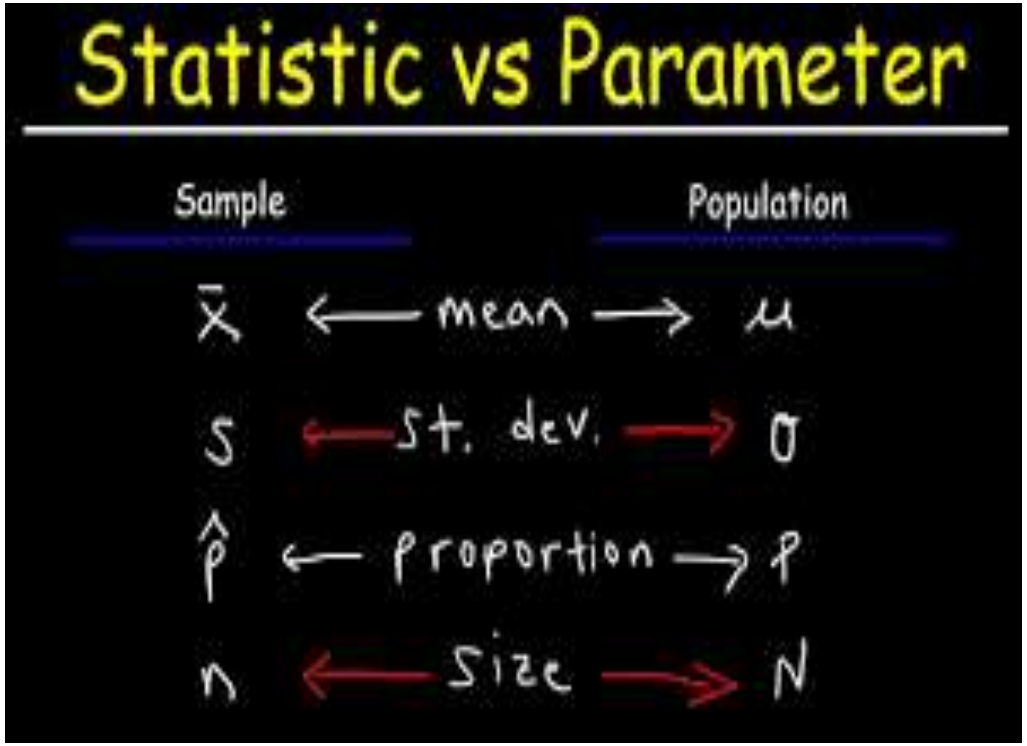
➔**Population parameter:**

- It is a numerical measurement describing some **characteristic** of a **population**.
- Example: mean and variance of a population are population parameters.

Slide Courtesy:Dr.Uma

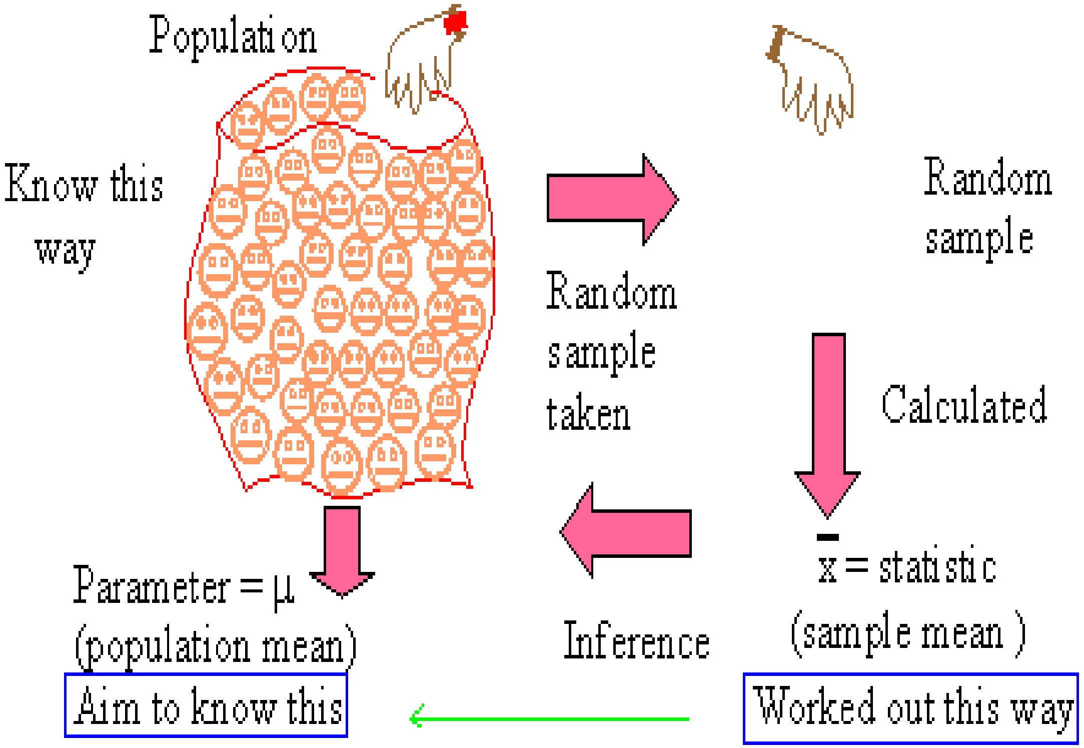
Sources: youtube..com,
Pinkmonkey.com

Sample Statistic and Population Parameter



Statistic vs Parameter

Sample		Population
$\bar{x}$	← mean →	μ
s	← st. dev. →	σ
$\hat{p}$	← proportion →	p
n	← size →	N



Population

Know this way

Random sample taken

Random sample

Calculated

$\bar{x}$ = statistic (sample mean)

Worked out this way

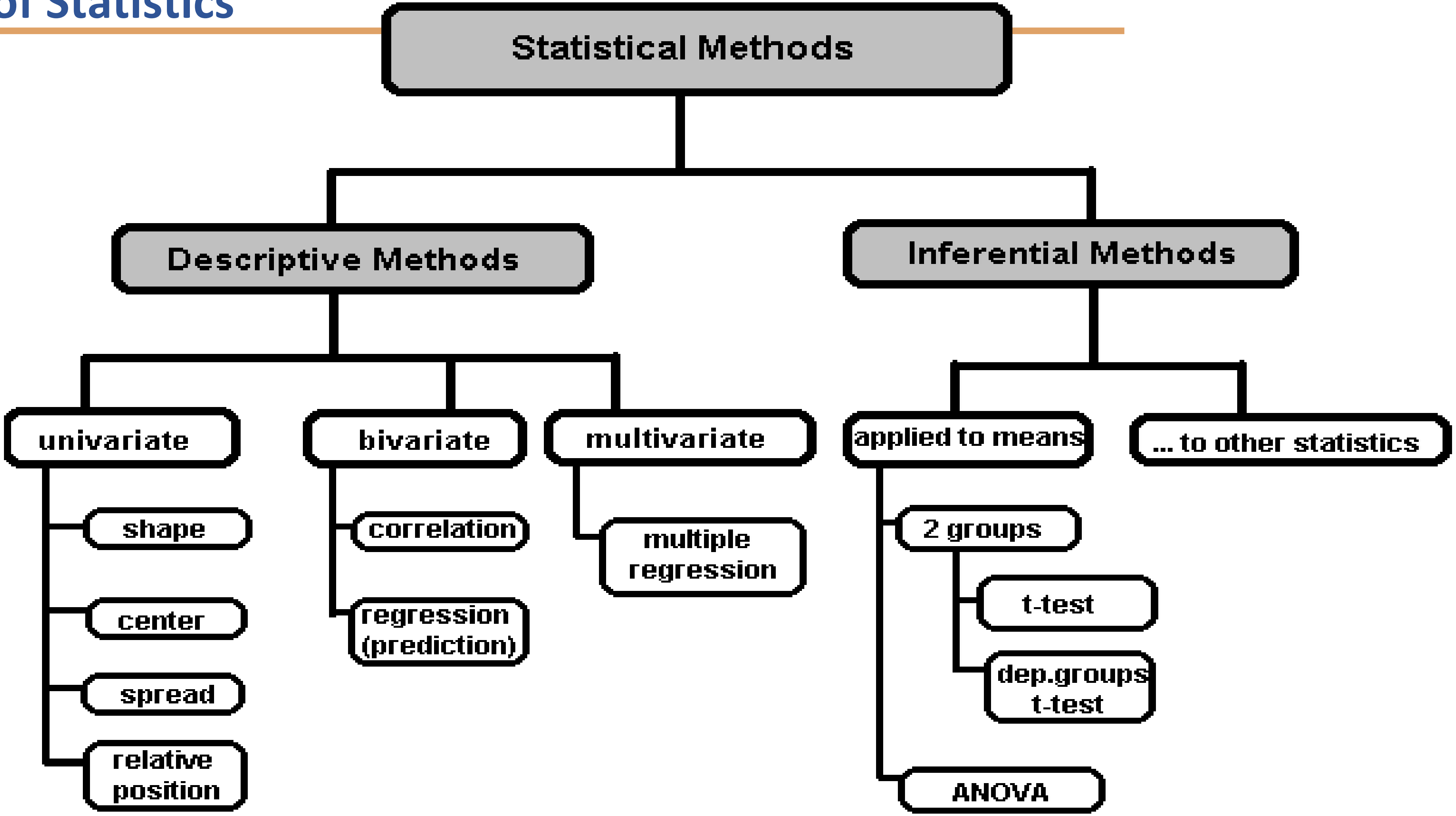
Inference

Parameter = μ (population mean)

Aim to know this

Sources: youtube..com,
Pinkmonkey.com

Taxonomy of Statistics

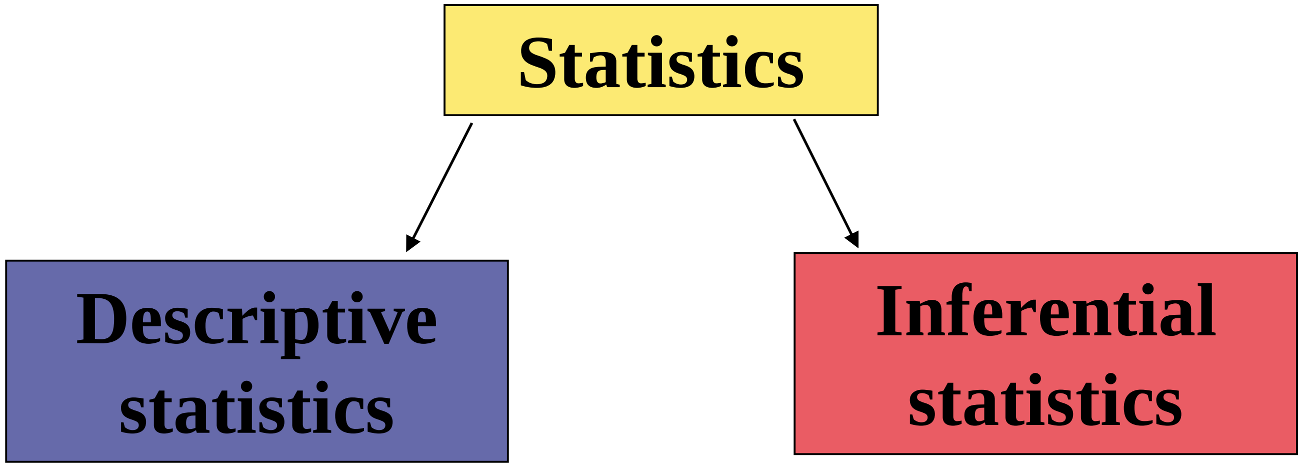


Source: image.slidesharecdn.com

Branches of Statistics

The study of statistics has **two** major branches:

1) **Descriptive statistics**
2) **Inferential statistics**



Statistics

Descriptive statistics

Inferential statistics

Involves organization, summarization, and display of data.

Involves using a sample to draw conclusions about a population.

Descriptive Statistics

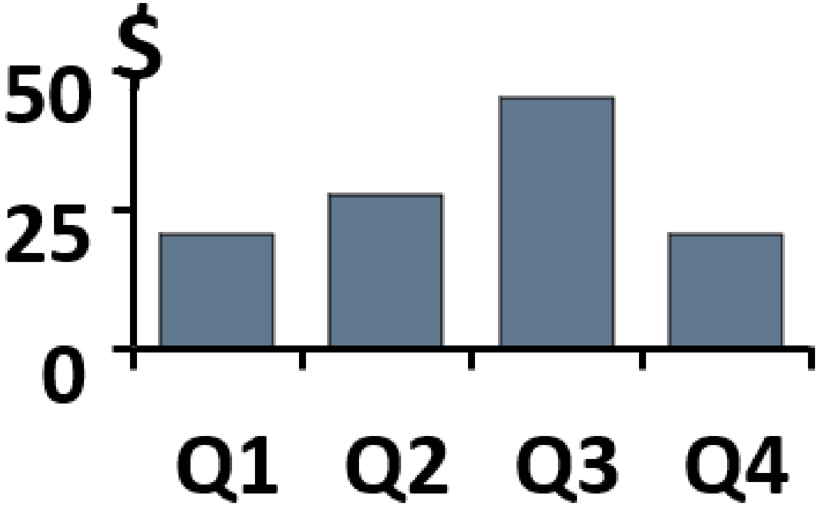
- Descriptive statistics are **methods for organizing and summarizing data**.
- Descriptive statistics utilizes numerical and graphical methods to look for **patterns in a data set**, to **summarize the information** revealed in a data set and to **present that information** in a convenient form.
- A descriptive value for a population is called a parameter and a descriptive value for a sample is called a statistic.
- For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.

Descriptive Statistics

- Purpose: To describe data
- Collect Data
  - e.g. Survey

- Present Data
  - e.g. Tables and graphs

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



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

Source: luminousmen.com/post/

Slide Courtesy:Dr.Uma

Descriptive Statistics

Types of descriptive statistics:

- **Organize Data**
  - Tables
  - Graphs
- **Summarize Data**
  - Central Tendency
  - Variation

Descriptive Statistics

➔ Organizing Data

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

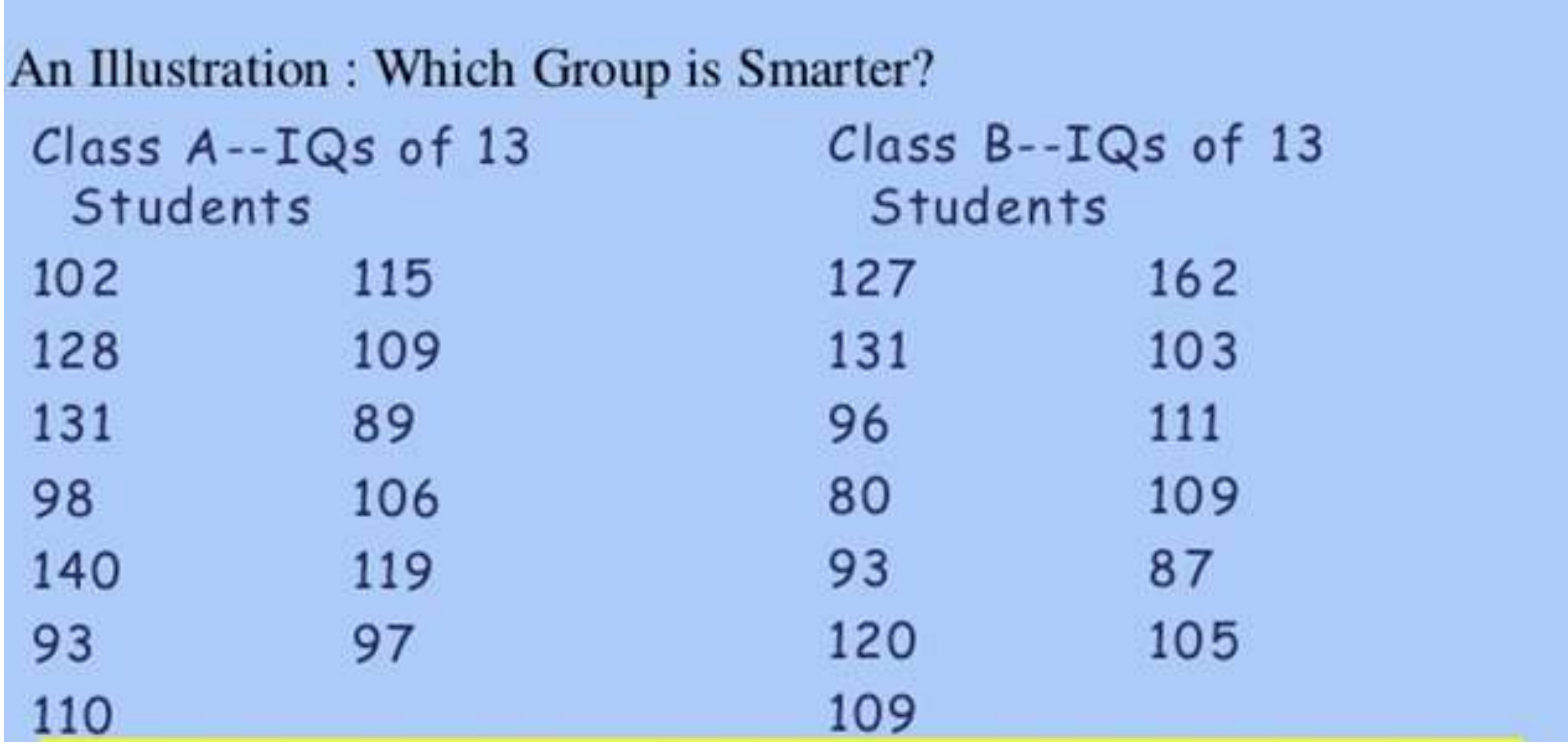
Descriptive Statistics

➔ Summarizing Data:

- **Central Tendency** (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- **Variation** (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

Why is Descriptive Statistics used?

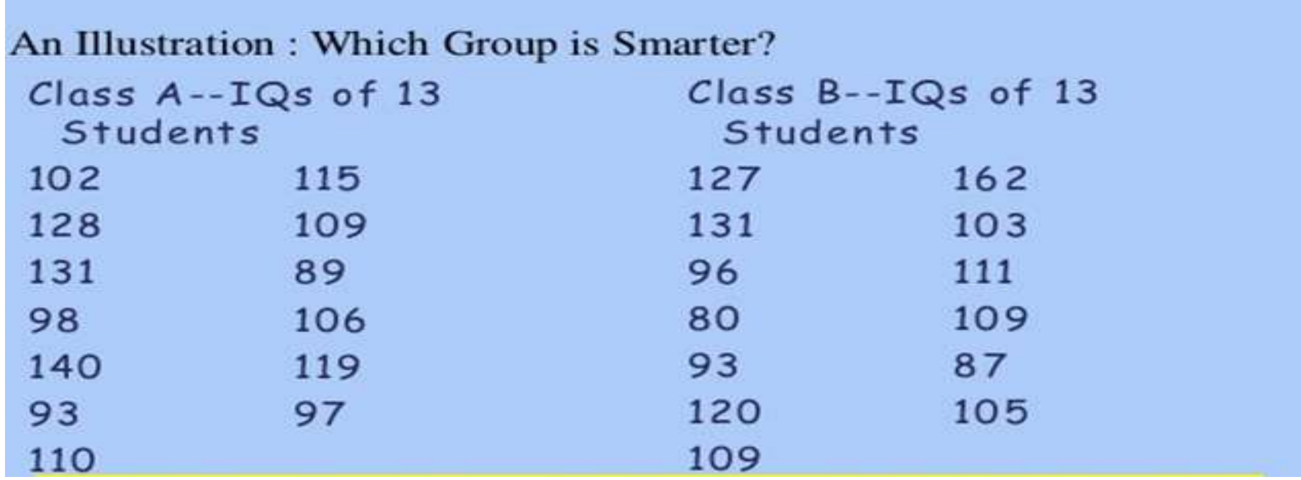


An Illustration : Which Group is Smarter?

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

Why is Descriptive Statistics used?



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	

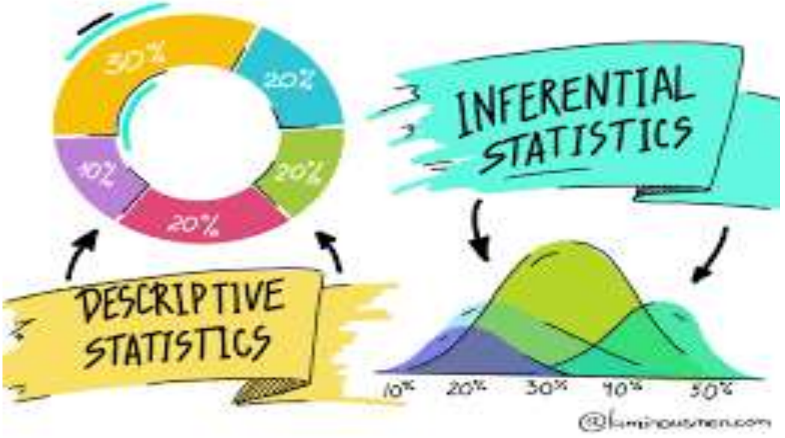
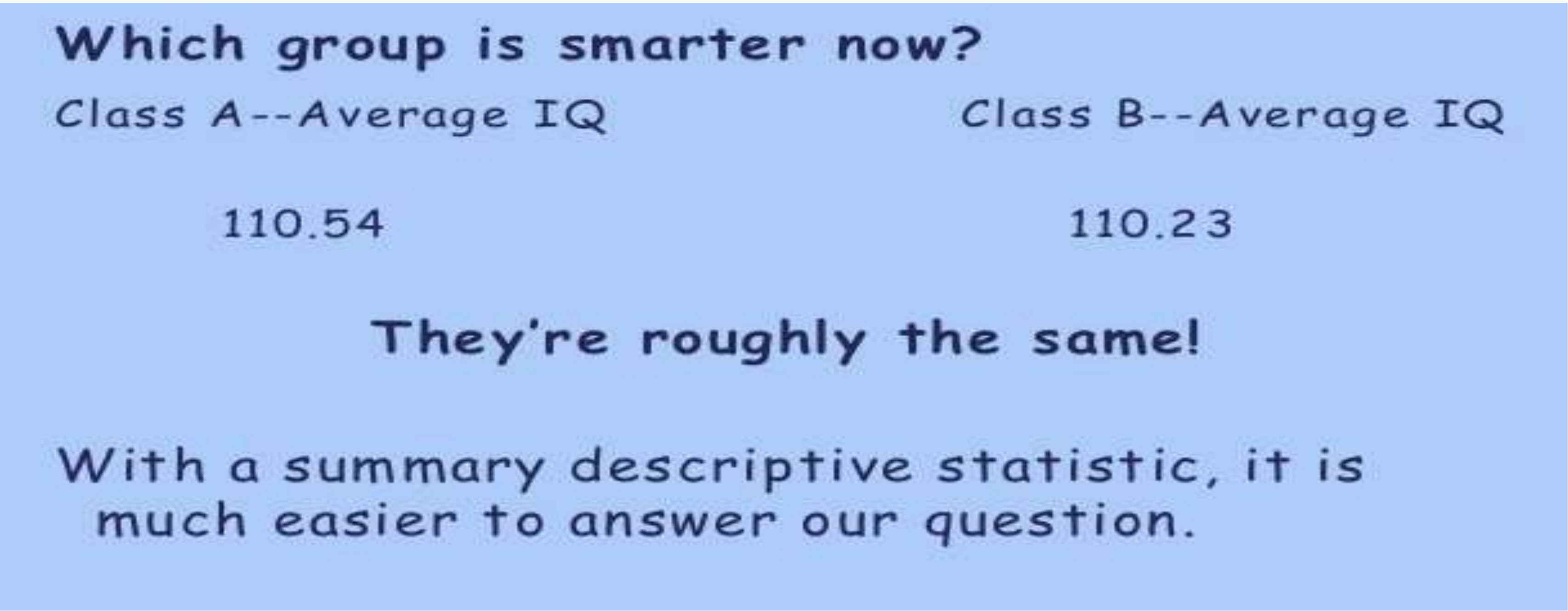


Figure speaks it all !!!



Which group is smarter now?

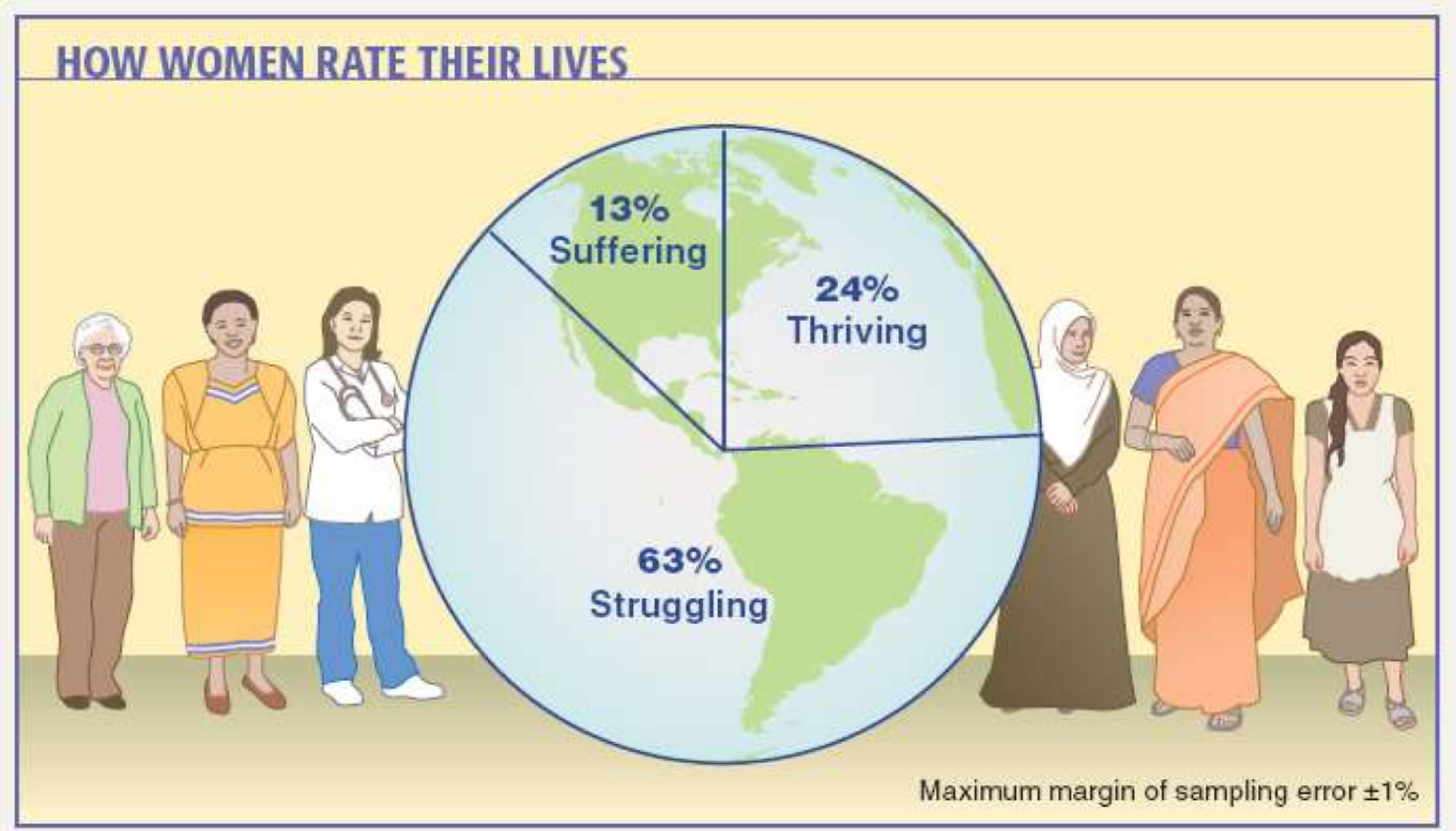
Class A--Average IQ	Class B--Average IQ
110.54	110.23

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

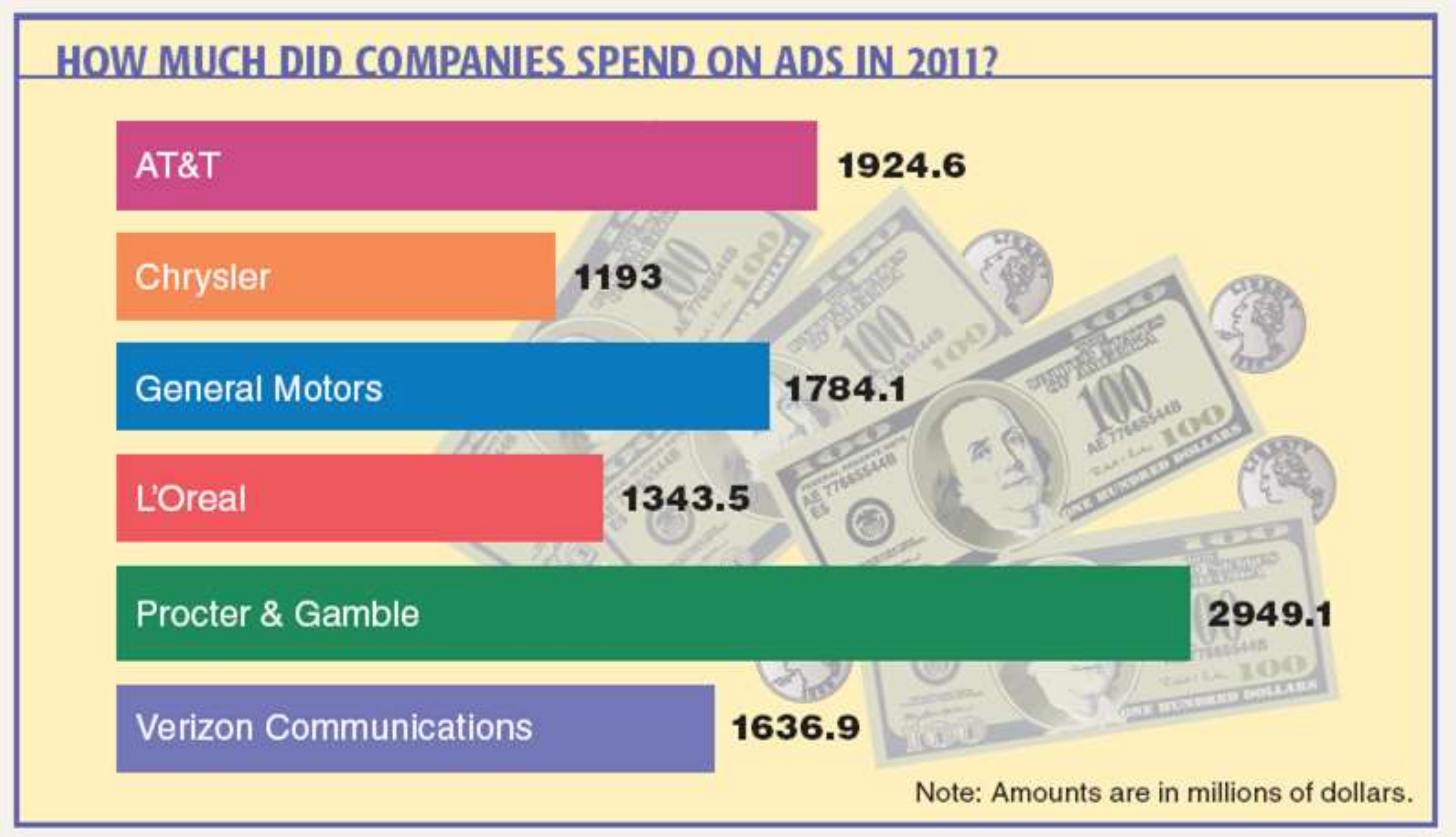
Why is Descriptive Statistics used?



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

Source: slidetodoc.com

Descriptive Statistics Examples



Source: WPP Kantar media

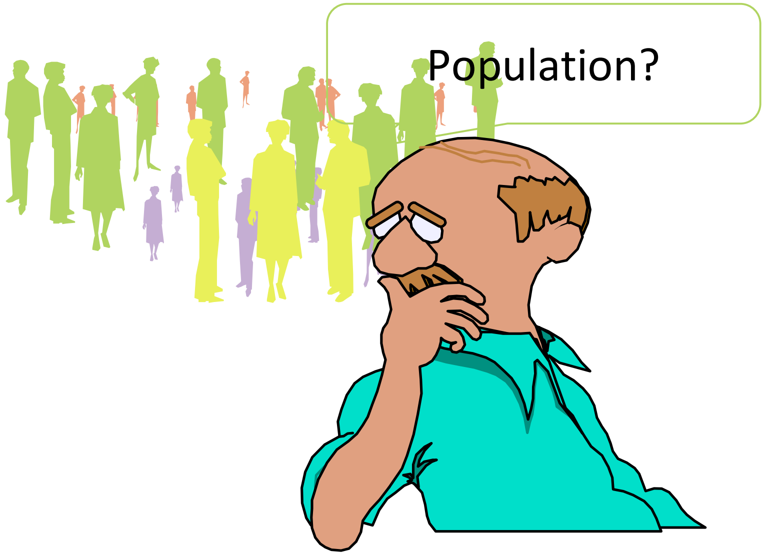
Inferential Statistics

- Inferential statistics utilizes sample data to make estimates, decisions, predictions or other generalizations about a larger set of data.
- There are two main areas of inferential statistics:

1. **Estimating parameters:** This means taking a statistic from the sample data (for example the sample mean) and using it to say something about a population parameter (for example the population mean).
2. **Hypothesis tests:** This is where sample data can be used to answer research questions. For example, one might be interested in knowing if a new cancer drug is effective; or if breakfast helps children perform better in schools.

Inferential Statistics

- Purpose: Make decision about population characteristics.
- Inferential statistics involves:
  - Estimation: e.g. Population Parameters
  - Hypothesis Testing



Source: luminousmen.com/post/

Why is Inferential Statistics used?

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 summary statistic).
- Conclusion : You conclude that the population mean income μ is likely to be close to $27,500 as well.
- This is an example of statistical inference.

Slide Courtesy:Dr.Uma

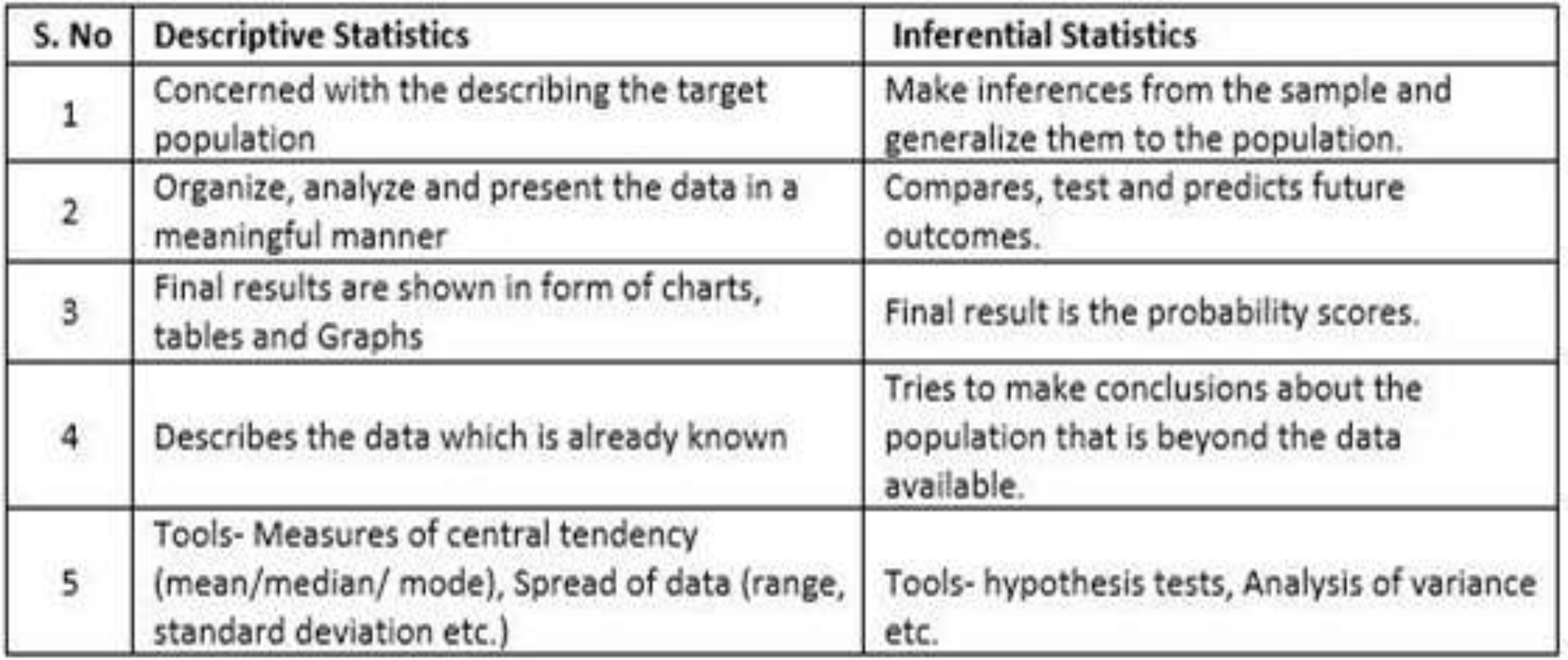
Inferential Statistics examples

- You randomly select a sample of 12th graders in your state and collect data on their JEE scores and other characteristics.

  You can use inferential statistics to make estimates and test hypotheses about the whole population of 11th graders in the state based on your sample data.

- To find out the average salary of IT engineers across the country:
  We can have a predefined selective number of IT engineers from a particular city, say Mumbai. We can gather data about their salaries much more easily and then use the data to evaluate the average income of IT engineers across the country.

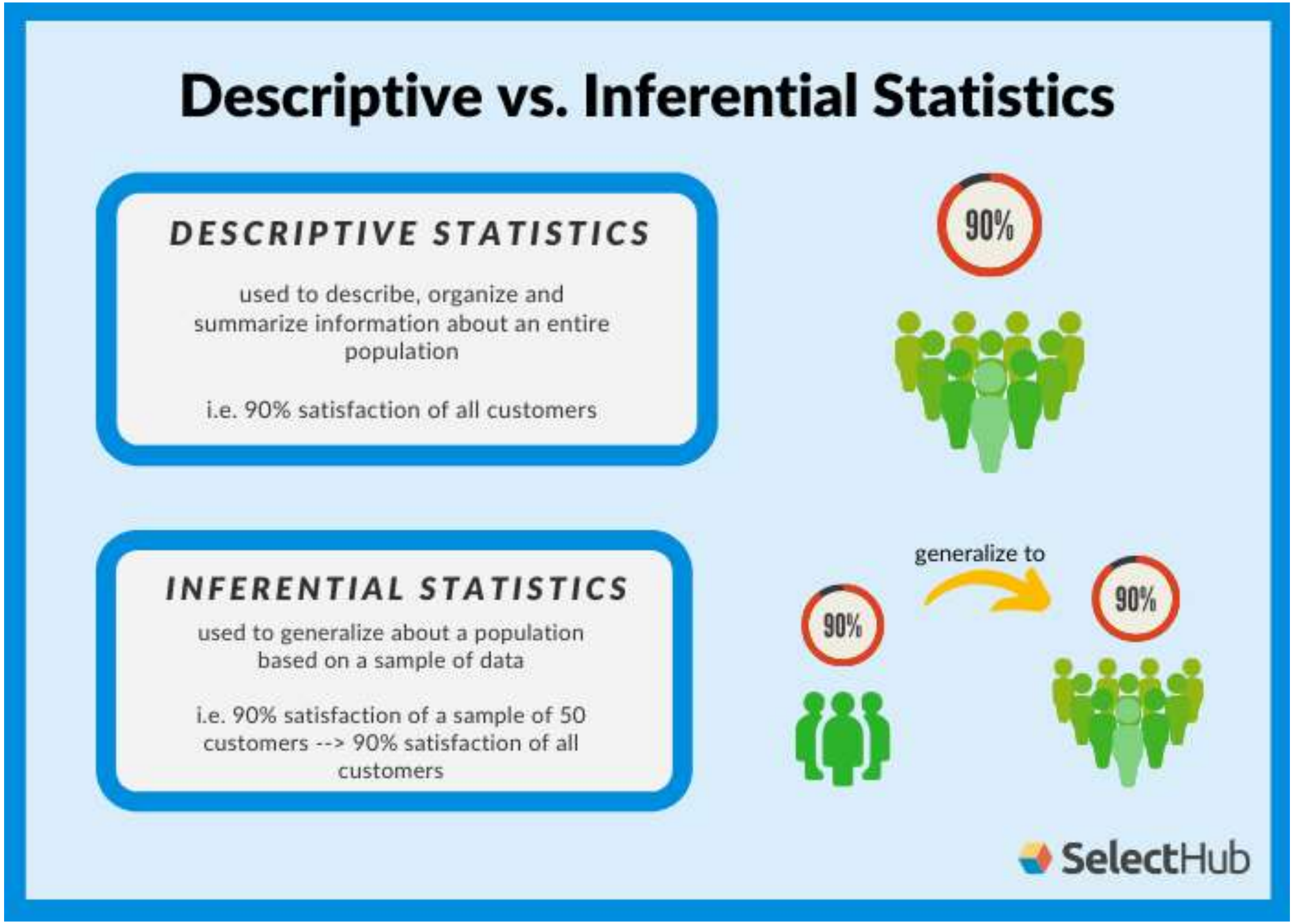
Descriptive Statistics vs Inferential Statistics



S. No	Descriptive Statistics	Inferential Statistics
1	Concerned with the describing the target population	Make inferences from the sample and generalize them to the population.
2	Organize, analyze and present the data in a meaningful manner	Compares, test and predicts future outcomes.
3	Final results are shown in form of charts, tables and Graphs	Final result is the probability scores.
4	Describes the data which is already known	Tries to make conclusions about the population that is beyond the data available.
5	Tools- Measures of central tendency (mean/median/ mode), Spread of data (range, standard deviation etc.)	Tools- hypothesis tests, Analysis of variance etc.

Source: miro.medium.com

Descriptive Statistics vs Inferential Statistics



Descriptive vs. Inferential Statistics

DESCRIPTIVE STATISTICS

used to describe, organize and summarize information about an entire population

i.e. 90% satisfaction of all customers

INFERENTIAL STATISTICS

used to generalize about a population based on a sample of data

i.e. 90% satisfaction of a sample of 50 customers --> 90% satisfaction of all customers

Source: selecthub.com

Descriptive Statistics 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

Slide Courtesy:Dr.Uma

Questions

Q1. 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.

Ans: 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.

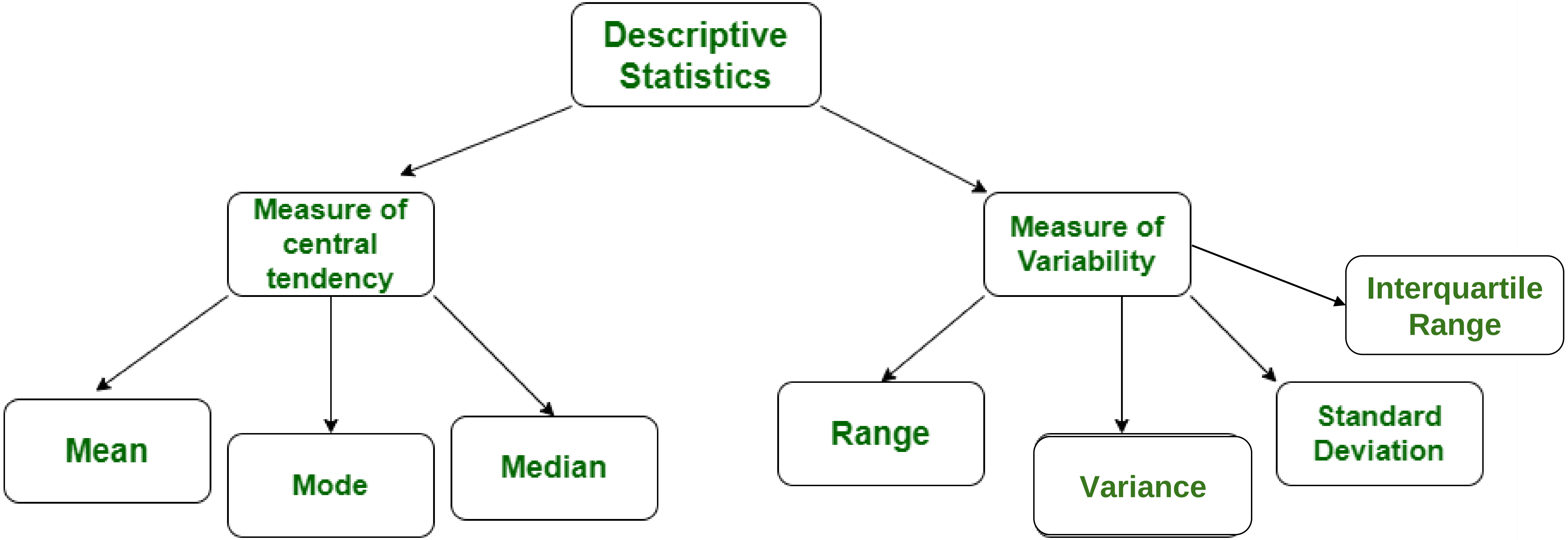
Questions

Q2. A burger outlet wanted to perform market research to determine what type of chicken burgers their customers liked. The outlet is researching to figure out the favorite tastes of their customers to provide better services and dishes to the customers. The outlet gathered a customer sample size of a 100 customers in different age groups and regular nearby customers at the outlet. The outlet was able to determine that 80% of the customers liked their chicken burgers to be spicy and crispy while the rest liked it non-crunchy and non-spicy.

What type of statistics was applied to arrive that conclusion?

Ans: Inferential statistics

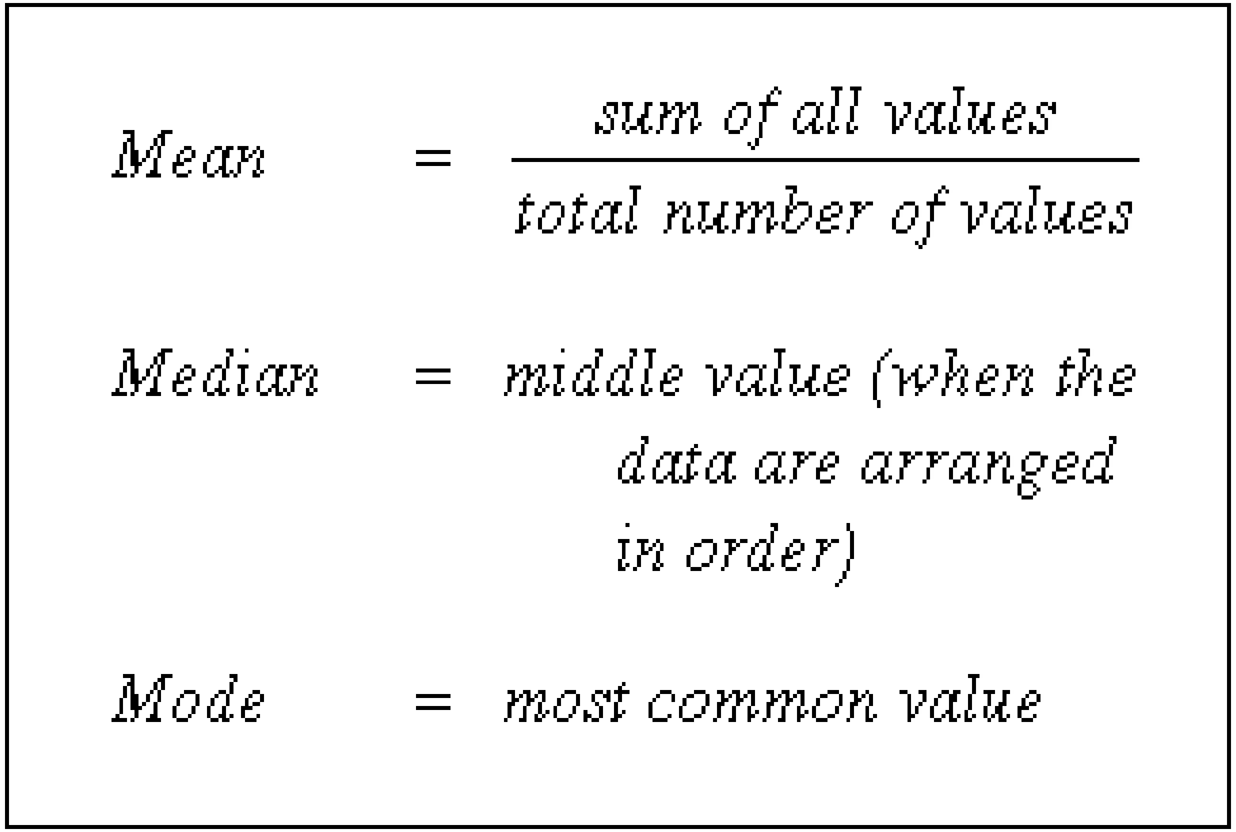
Types of Descriptive Statistics



Source: geeksforgeeks

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.



$$\text{Mean} = \frac{\text{sum of all values}}{\text{total number of values}}$$

Median = middle value (when the data are arranged in order)

Mode = most common value

Source: sixsigma-institute.org

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:

a) Population mean:

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

where N is the population size

a) Sample mean:

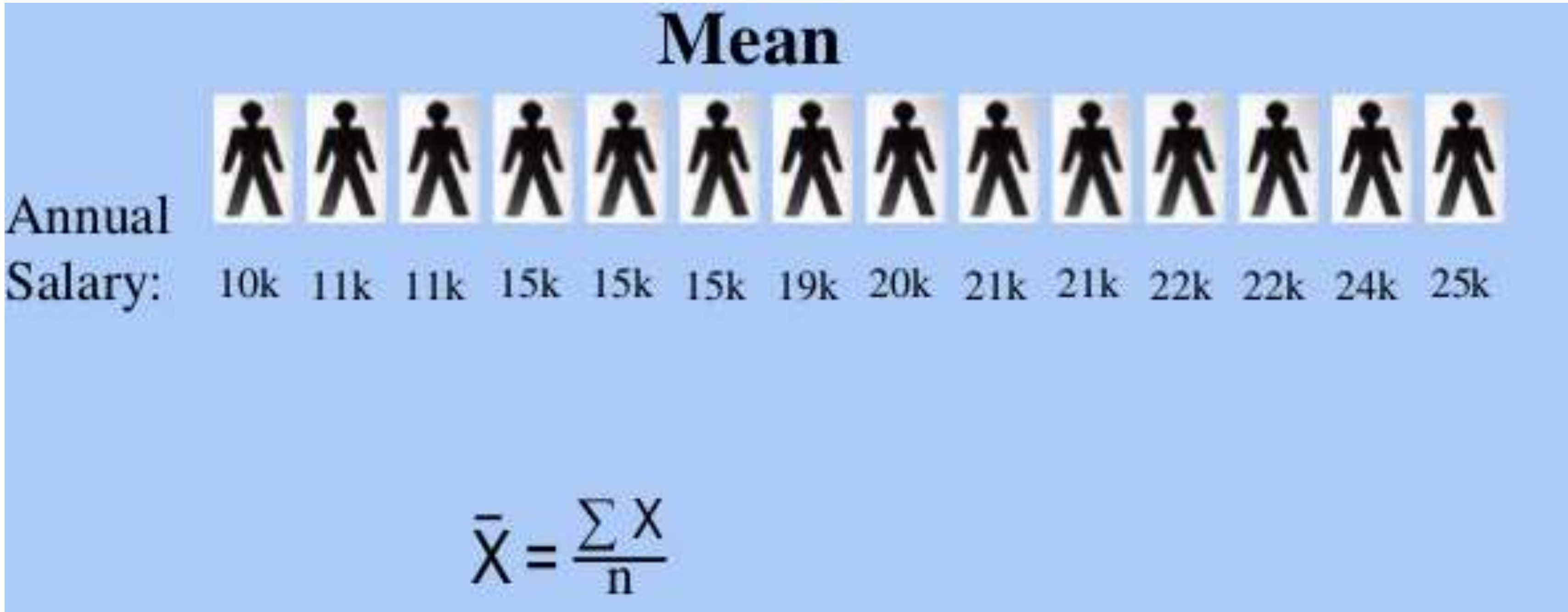
$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where n is the sample size

Source: sixsigma-institute.org

Slide Courtesy:Dr.Uma

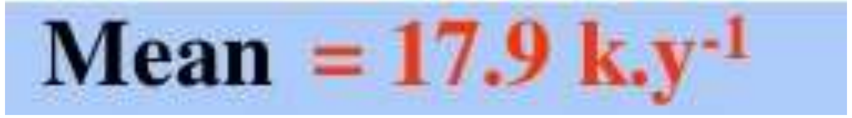
Measures of central tendency: Mean



Mean

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

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



Mean $= 17.9$ k.y^{-1}

Source: slideshare.net

Slide Courtesy:Dr.Uma

Measures of central tendency: Weighted mean

- Weighted mean is an average where certain values of the data set contribute more to the mean value.
- For a finite set of dataset with measurement values X1, X2, …., Xn

(a set of n numbers), and the corresponding weights w1, w2,….wn

it is defined by the formula:

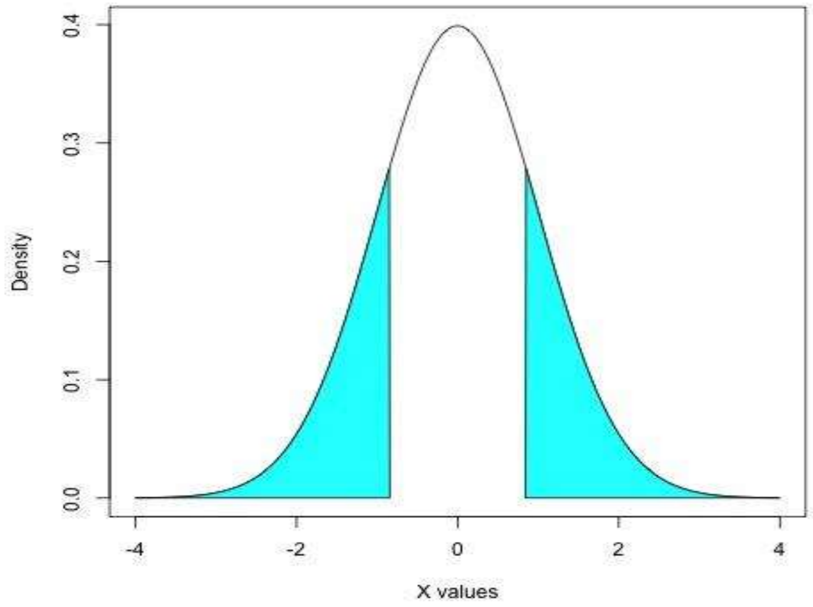
$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

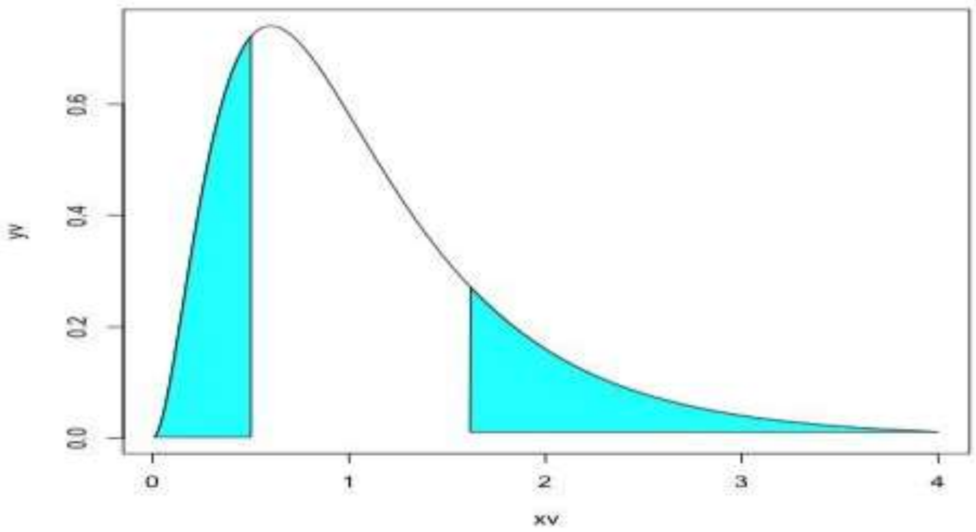
Measures of central tendency: Trimmed mean

- The trimmed mean is computed by arranging the sample values in order, “trimming” an equal number of them from each end, and computing the mean of those remaining.
- If p% of the data are trimmed from each end, the resulting trimmed mean is called the **“p% trimmed mean”.**
- There are no hard-and-fast rules on how many values to trim.
- The most commonly used trimmed means are the 5%, 10%, and 20% trimmed means.

Measures of central tendency: Trimmed mean

- If the sample size is denoted by n, and a p% trimmed mean is desired, the number of data points to be trimmed is np/100
- It is used to reduce the effects of outliers on the calculated average.
- This method is best suited for data with large, erratic deviations or extremely skewed distributions.





Source: exceluser.com, MathBitsNotebook

Question

A simple random sample of five men is chosen from a large population of men, and their heights are measured. The five heights (in inches) are 65.51, 72.30, 68.31, 67.05, and 70.68. Find the sample mean.

Ans: .

The sample mean is

$$X = \frac{1}{5}(65.51 + 72.30 + 68.31 + 67.05 + 70.68) = 68.77 \text{ in.}$$

Measures of central tendency: Mean

➔ **Advantages:**

- It takes into account all the available information.
- It can be combined with means of other groups to give the overall 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 data set has a unique mean value.

➔ **Disadvantages:**

- It is a very sensitive measure.
- Thus, its value is easily affected by extreme values known as the outliers.
- It can only be used on interval or ratio data.

Source: slideshare.net

Slide Courtesy:Dr.Uma

Measures of central tendency: Median

- **Median** is the value separating the higher half from the lower half of a data sample, a population, or a probability distribution.
- It is the middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the average.
- For a data set, it may be thought of as "the middle" value.
- The basic feature of the median in describing data compared to the mean (often simply described as the "average") is that it is not affected by a small proportion of extremely large or small values, and therefore provides a better representation of a "typical" value.

Measures of central tendency: Median

➔ Process of calculating median:

1. Arrange all the values of the data set in ascending order.

   X1,X2,X3,....,Xn

1. Find the **middle position**.

3. The element corresponding to middle position is considered as median if odd number of elements are present.

   i.e. if n is odd, median = (n+1/2)th element’s value

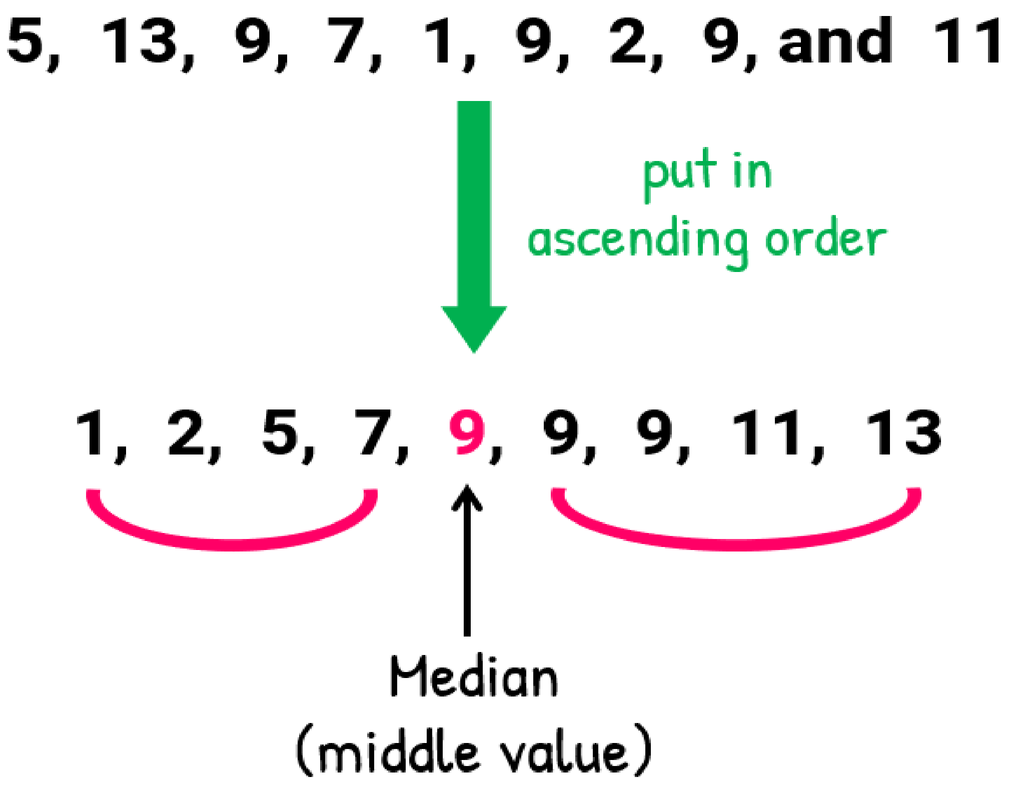
4. If there are even number of elements present then the average of the elements present in the middle positions is considered as median. i.e. if n is even,

median = ((n/2)th element’s value + (n/2 + 1)th element’s value))/2

Measures of central tendency: Median

Odd no. of elements

Even no. of

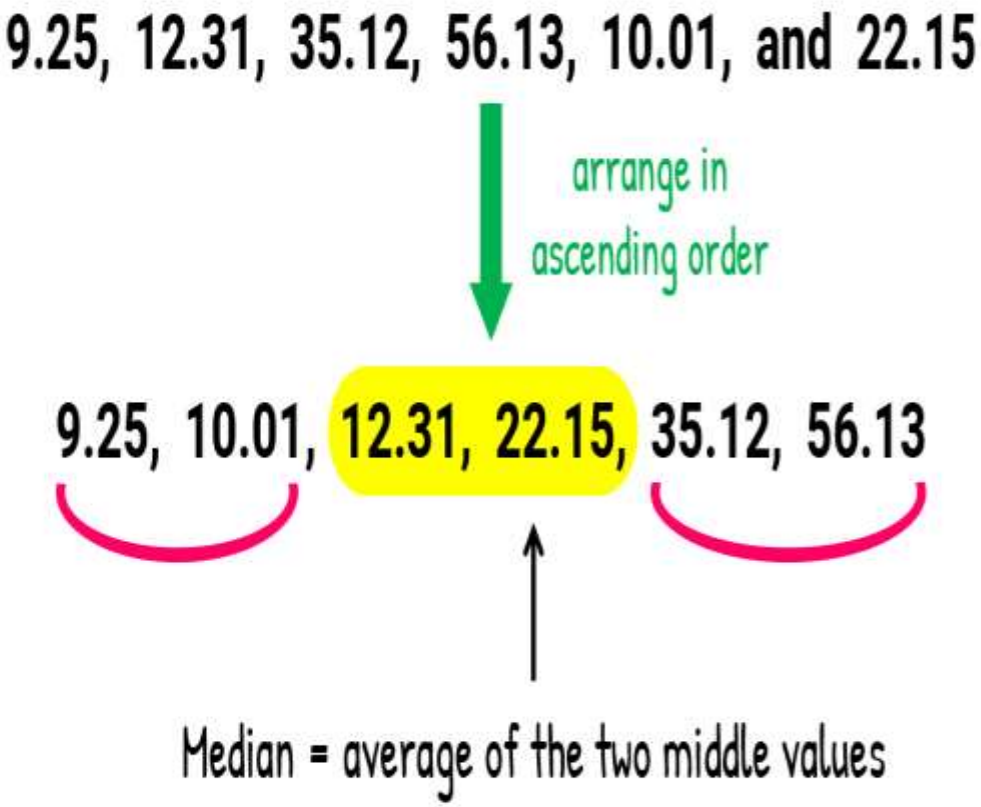


5, 13, 9, 7, 1, 9, 2, 9, and 11

put in ascending order

1, 2, 5, 7, 9, 9, 9, 11, 13

Median (middle value)



9.25, 12.31, 35.12, 56.13, 10.01, and 22.15

arrange in ascending order

9.25, 10.01, 12.31, 22.15, 35.12, 56.13

Median = average of the two middle values

Source: Chilimath

Slide Courtesy:Dr.Uma

Measures of central tendency: Median-example

Consider the data given below:

A simple random sample of five men is chosen from a large population of men, and their heights are measured. The five heights (in inches) are

65.51, 72.30, 68.31, 67.05, 70.68.

Calculate the median.

Ans:

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

The five heights, arranged in increasing order, are

65.51 67.05 68.31 70.68 72.30.

n = 5

- The sample median is the middle number, which is 68.31.
- Half of the other values in the list are below 68.31 and half are above 68.31.

Slide Courtesy:Dr.Uma

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 and 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.

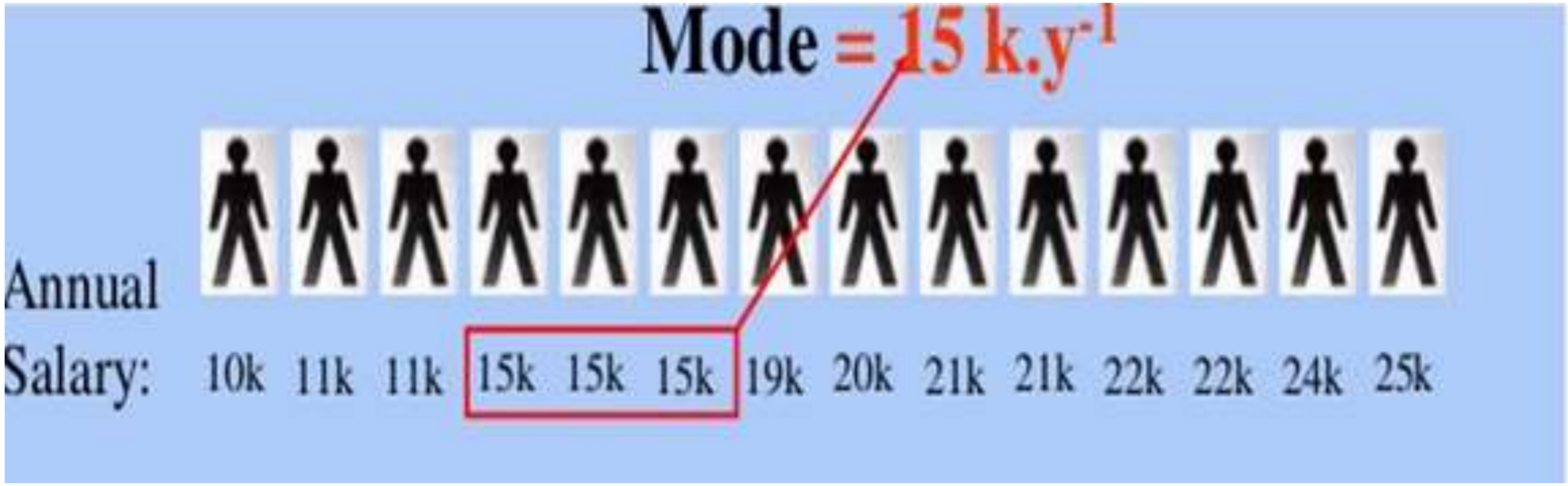
Slide Courtesy:Dr.Uma

Measures of central tendency: Mode

- The **mode** is the value that appears most often in a set of data values
- Like the statistical mean and median, the mode is a way of expressing, in a (usually) single number.
- To calculate the mode, we need to look at which **value appears the most often**.
- Example, the mode of the sample [1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17] is 6.
- Given the list of data [1, 1, 2, 4, 4] its mode is not unique.
  It has 2 modes: 1 and 4
- A dataset, in such a case, is said to be **bimodal**, while a set with_more than two modes may be described as **multimodal**.
- **Empirical formula:**

$$mean - mode = 3 \times (mean - median)$$

Measures of central tendency: Mode-example



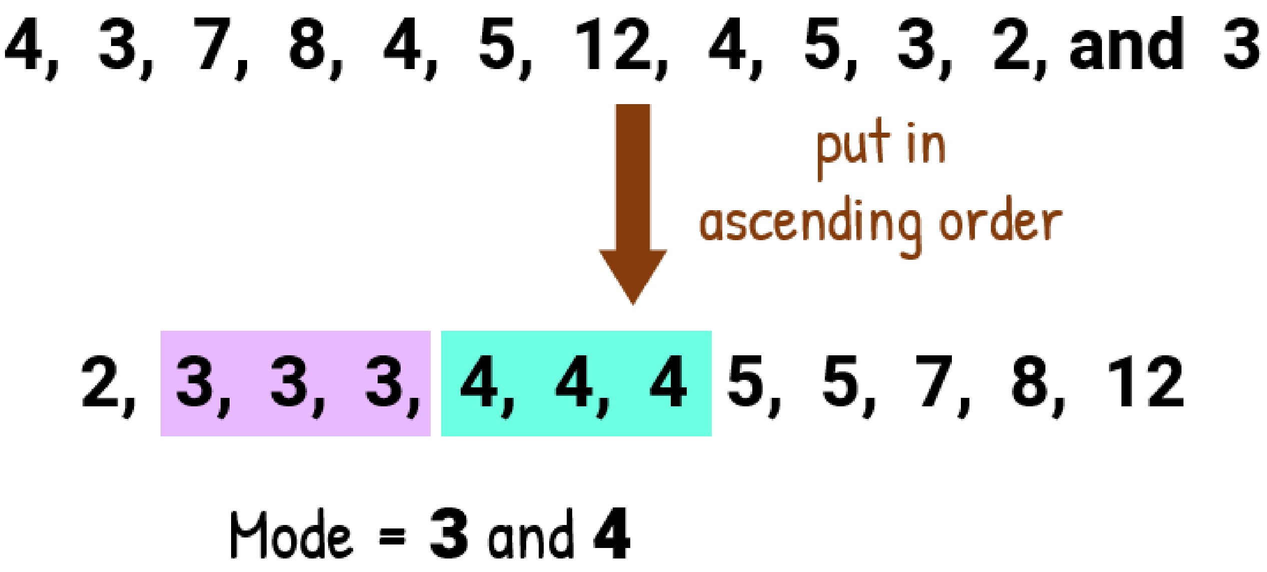
Mode = 15k

Source: slideshare.net

Slide Courtesy:Dr.Uma

Measures of central tendency: Mode-example

Consider the data given below:



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

put in ascending order

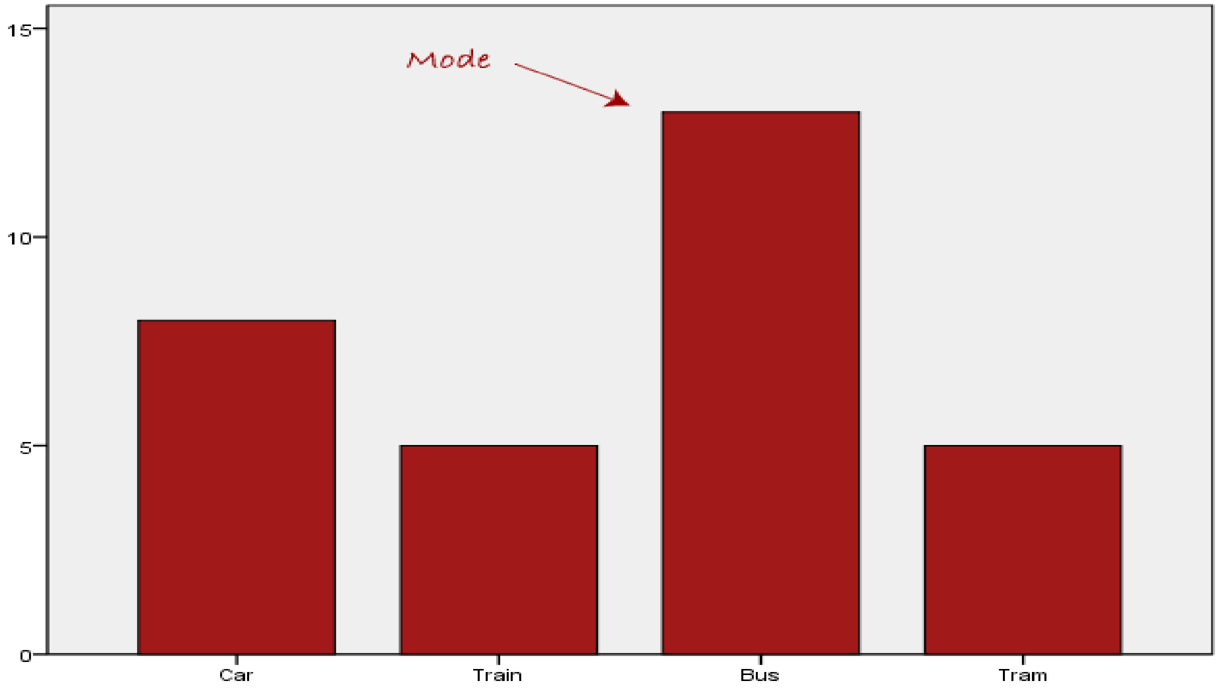
2, 3, 3, 3, 4, 4, 4 5, 5, 7, 8, 12

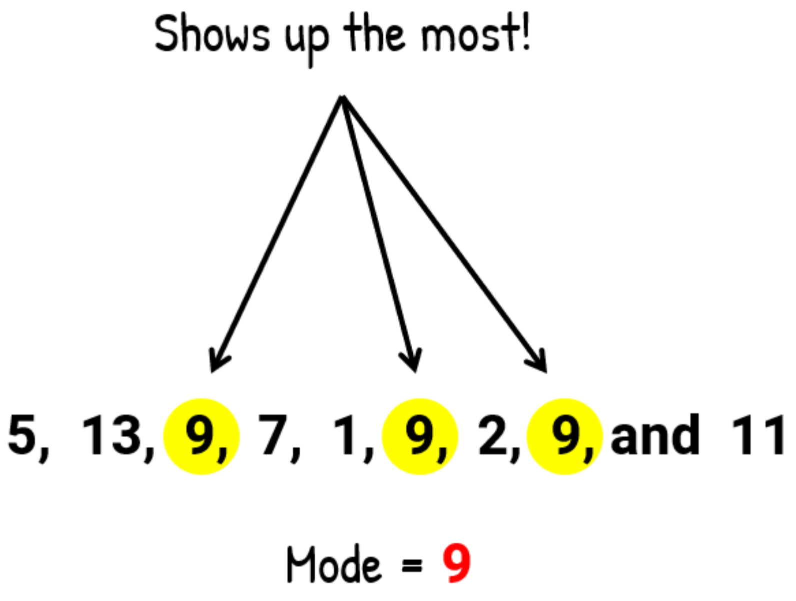
Mode = **3** and **4**

The values 3 and 4 appear the most number of times in the above data.
Since the above data has 2 modes, it is **bimodal**.

Slide Courtesy:Dr.Uma

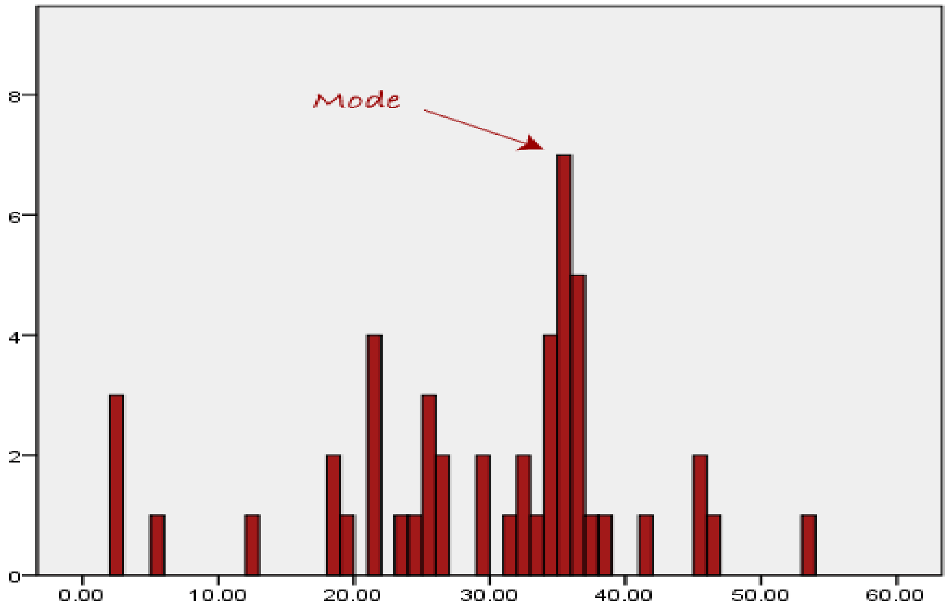
Measures of central tendency: Mode-example




Shows up the most!

5, 13, 9, 7, 1, 9, 2, 9, and 11

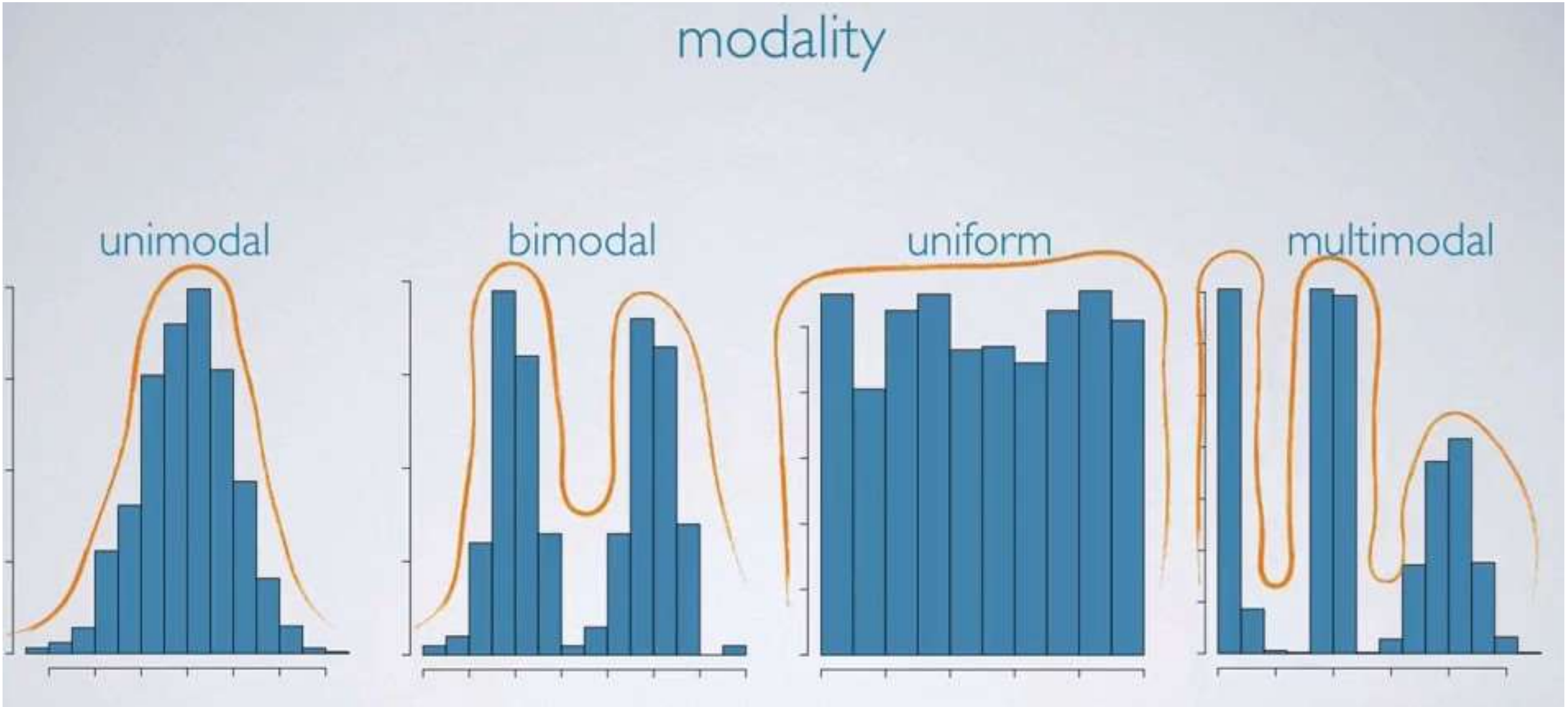
Mode = 9



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

Slide Courtesy:Dr.Uma

Measures of central tendency: Mode



Source: miro.medium.com

Measures of central tendency: Mode

➔ **Advantages:**

- Quick and easy to compute.
- Unaffected by extreme values.
- Can be used at any level of measurement.
- Useful to find the most “popular” or common item. This includes data sets that do not involve numbers.

➔ **Disadvantages:**

- It is a terminal statistic.
- A given subgroup could make this measure unrepresentative of the population’s centre.
- 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.

Slide Courtesy:Dr.Uma

Measures of central tendency: Questions

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.

Ans:

Sorting in ascending order:

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

- Mean = 222/20 = 11.1
- Median = (11+11)/2 = 11
- Mode = 12

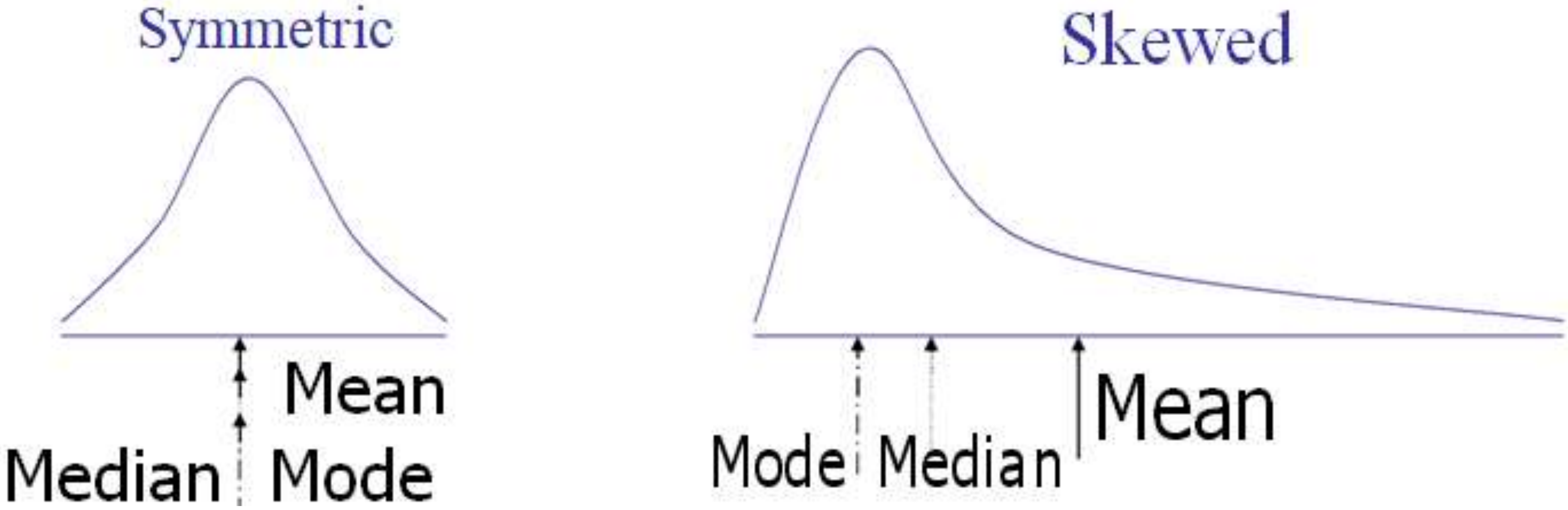
Skewed and Symmetric distributions

- **Skewness** is a **measure of the asymmetry** of the distribution of about its mean.
- The skewness value can be positive, zero, negative, or undefined.

- **Symmetric Distribution:** A symmetric distribution is one where the left and right hand sides of the distribution are roughly equally balanced around the mean.
- **In symmetric distributions, the mean, median, and mode are the same**.

- **Skewed Distribution:** A skewed distribution is one where the left and right hand sides of the distribution are not balanced around the mean.
- In skewed data, the mean and median lie further toward the skew than the mode.
- **The greater the distance of mean and median, the greater is the skewness of the distribution.**

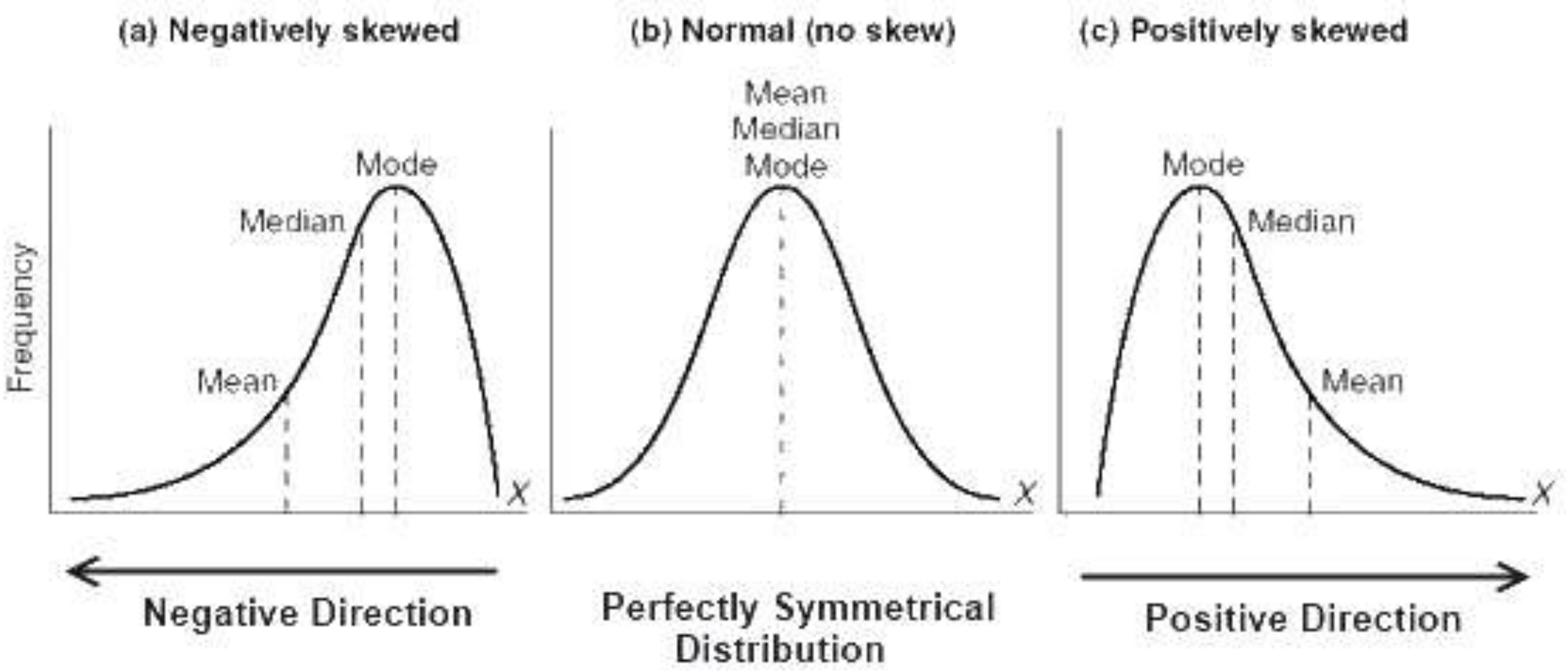
Skewed and Symmetric distributions



Source:www.slideshare.net

Slide Courtesy:Dr.Uma

Skewed and Symmetric distributions



Left skewed

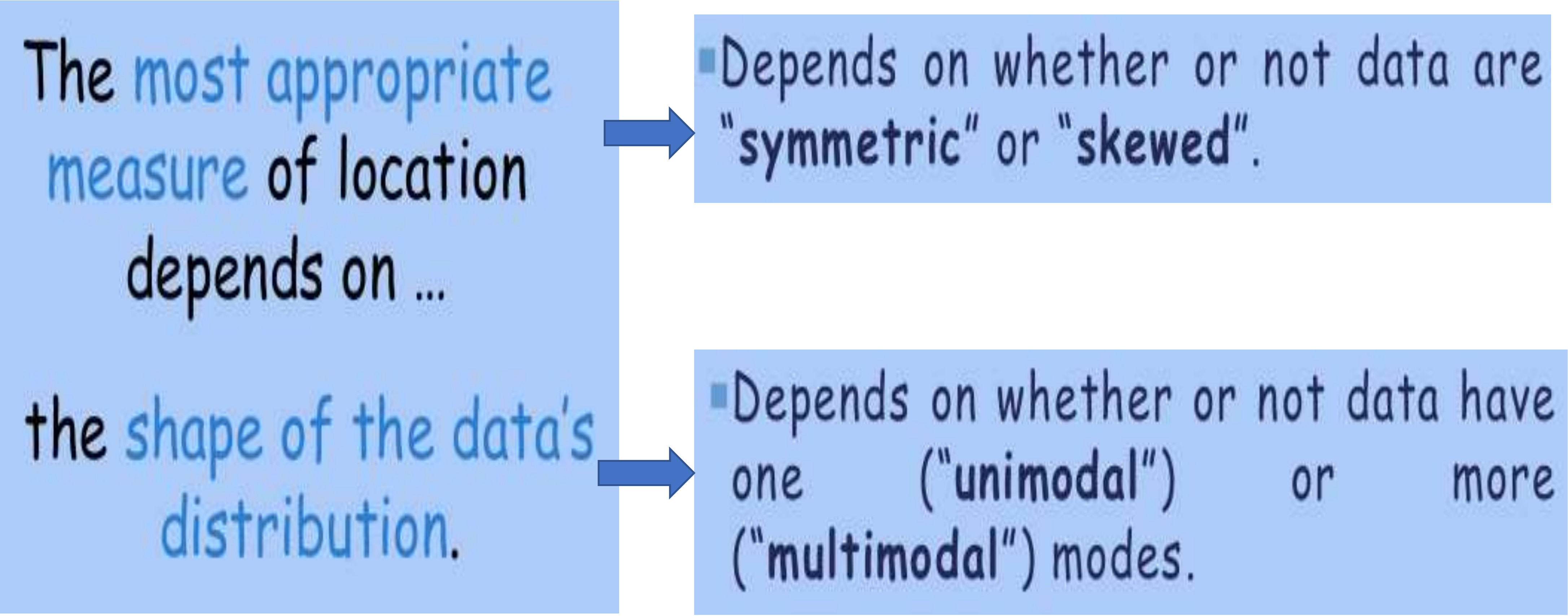
Right skewed

Source: https://www.fromthegenesis.com/skewness/

Skewed and Symmetric distributions

- **Distribution of a variable:** tells us what values the variable takes and how often it takes these values.
- **Shape:** It is the "shape" of the distribution of the data.
- 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 distribution**)

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

Slide Courtesy:Dr.Uma

Measures of central tendency

Various central tendency measures can be applied on different types of data.

- **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

Slide Courtesy:Dr.Uma

Measures of central tendency: Question

For the following data

30 75 79 80 80 105 126 138 149 179 179 191

223 232 232 236 240 242 245 247 254 274 384 470

Compute the mean, median, and the 5%, 10%, and 20% trimmed means.

Solution:

- The mean is found by averaging together all 24 numbers, which produces a value of 195.42.
- The median is the average of the 12th and 13th numbers, which is (191 + 223)/2 = 207.00.
- To compute the 5% trimmed mean, we must drop 5% of the data from each end.
- This comes to (0.05)(24) = 1.2 observations.
- We round 1.2 to 1, and trim one observation off each end.

Measures of central tendency: Question

- The 5% trimmed mean is the average of the remaining 22 numbers: 75 + 79 +⋯+ 274 + 384/22= 190.45

- To compute the 10% trimmed mean, round off (0.1)(24) = 2.4 to 2. Drop 2 observations from each end, and then average the remaining 20: 79 + 80 +⋯+ 254 + 274/20= 186.55

- To compute the 20% trimmed mean, round off (0.2)(24) = 4.8 to 5.

- Drop 5 observations from each end, and then average the remaining 14: 105 + 126 +⋯+ 242 + 245/14= 194.07

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

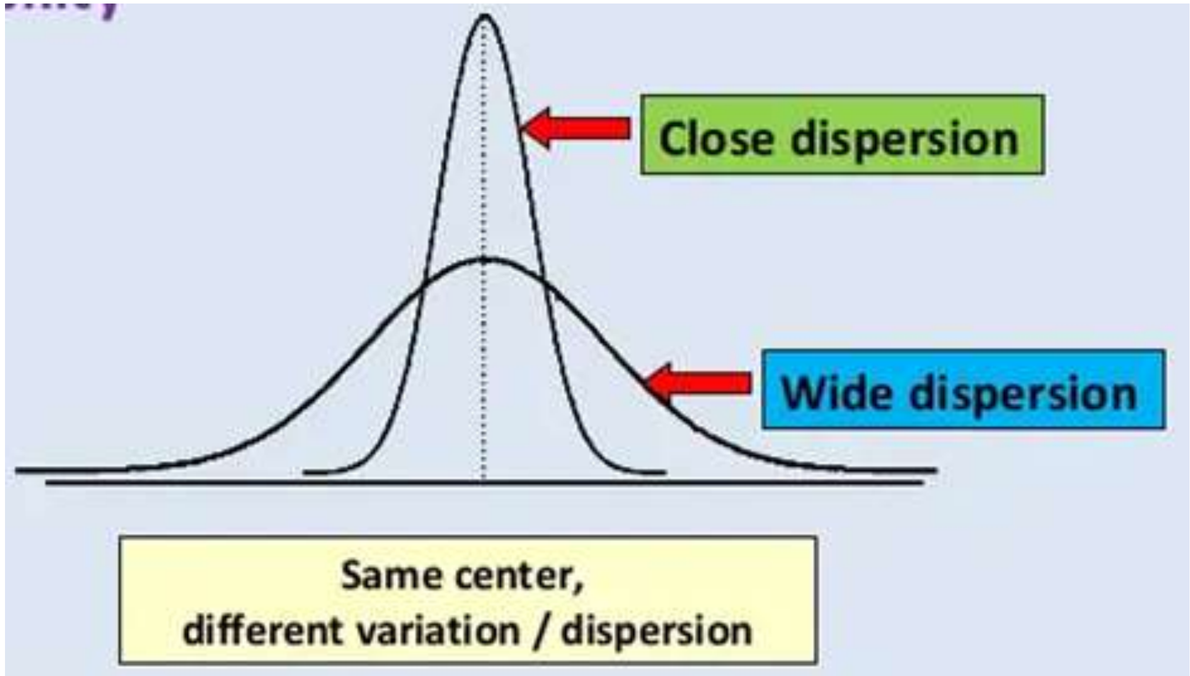
Slide Courtesy:Dr.Uma

Measures of Spread/Dispersion

- In statistics, the measures of dispersion help to interpret the variability of data
- It helps to know how much homogeneous or heterogeneous the data is.
- In simple terms, it shows how squeezed or scattered the variable is
- There are two main types of dispersion methods in statistics which are:

  (i) Absolute Measure of Dispersion

  (ii) Relative Measure of Dispersion



Source: image.slidesharecdn.com

Measures of Spread/Dispersion

- **Absolute Measure of Dispersion:**

  It contains the same unit as the original data set. Absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or mean deviations. It includes range, standard deviation, quartile deviation, etc.

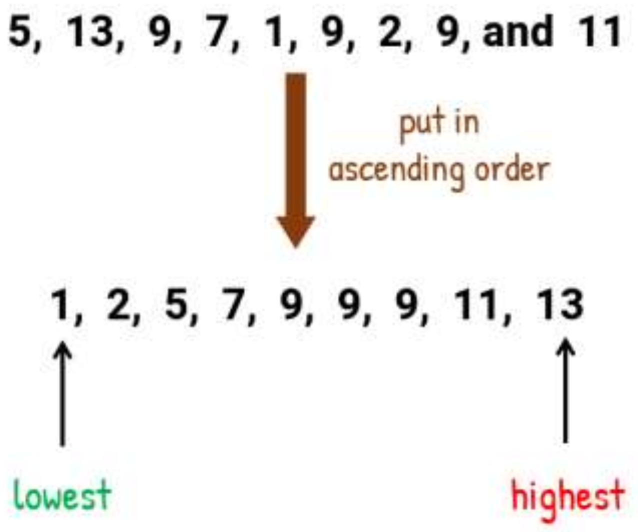
- **Relative Measure of Dispersion:**

  The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include: Coefficient of Range, Coefficient of Variation, Coefficient of Standard Deviation, Coefficient of Quartile Deviation, Coefficient of Mean Deviation.

Measures of Spread: Range

- Range is the most common and easily understandable measure of dispersion.
- It is the difference between two extreme observations of the data set.
- If X max and X min are the two extreme observations then

$$\text{Range} = X\text{ max} – X\text{ min}$$



5, 13, 9, 7, 1, 9, 2, 9, and 11

1, 2, 5, 7, 9, 9, 9, 11, 13

Source: Chilimath.com

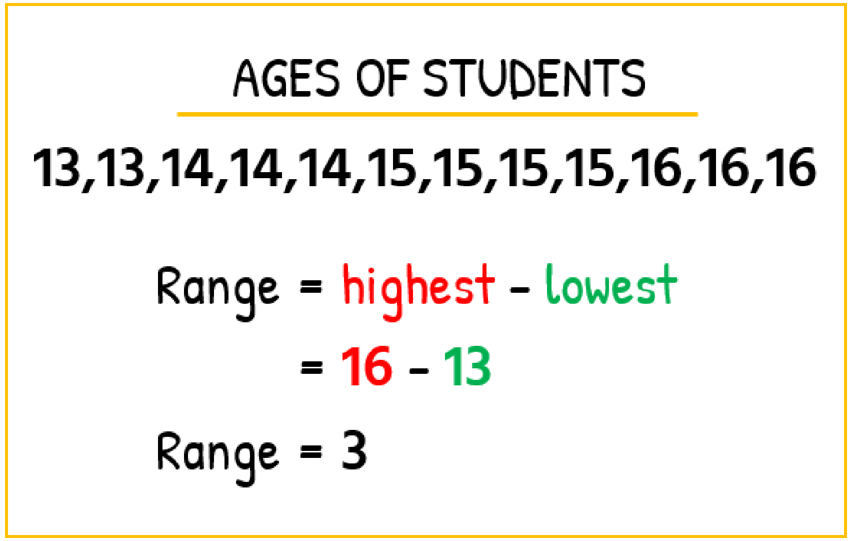
Here, range = 13 - 1

= 12

Slide Courtesy:Dr.Uma

Measures of Spread: Range

Class A



AGES OF STUDENTS

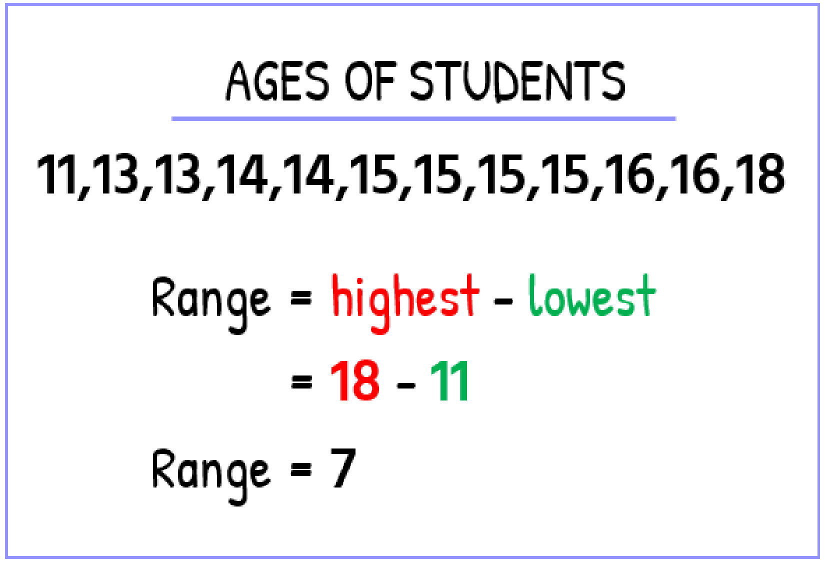
13,13,14,14,14,15,15,15,15,16,16,16

Range = highest - lowest

= 16 - 13

Range = 3

Class B



AGES OF STUDENTS

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

Range = highest - lowest

= 18 - 11

Range = 7

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

Measures of Spread: Range

Range Can Be Misleading:

- The range can sometimes be misleading when there are extremely high or low values.
- Example: **{8, 11, 5, 9, 7, 6, 3616}**

  lowest value : 5

  highest 3616,
- So the range is 3616 - 5 = **3611**.
- The single value of 3616 makes the range large, but most values are around 10.

Measures of Spread: Range

➔ **Advantages:**

- It is the simplest of the measure of dispersion
- Easy to calculate
- Easy to understand
- Independent of change of origin

➔ **Disadvantages:**

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on change of scale
- It can drastically be affected by outliers (values that are not typical as compared to the rest of the elements in the set).

Measures of Spread:

When presenting or analysing measurements of a continuous variable it is sometimes helpful to group subjects into several equal groups.

For example, to create four equal groups we need the values that split the data such that 25% of the observations are in each group.

The cut off points are called quartiles, and there are three of them (the middle one also being called the median).

Likewise, we use two tertiles to split data into three groups, four quintiles to split them into five groups, and so on.

Measures of Spread:

The general term for such cut off points is quantiles;

other values likely to be encountered are deciles, which split data into 10 parts,

and centiles, which split the data into 100 parts (also called percentiles).

Values such as quartiles can also be expressed as centiles; for example, the lowest quartile is also the 25th centile and the median is the 50th centile.

Measures of Spread:

- A quintile is a statistical value of a data set that represents 20% of a given population, so the first quintile represents the lowest fifth of the data (1% to 20%); the second quintile represents the second fifth (21% to 40%) and so on.

Example:

- Quintiles are used to create cut-off points for a given population; a government-sponsored socio-economic study may use quintiles to determine the maximum wealth a family could possess in order to belong to the lowest quintile of society. This cut-off point can then be used as a prerequisite for a family to receive a special government subsidy aimed to help society's less fortunate.

Measure of Spread:Percentile

- A **percentile** is a comparison measure between a particular value and the values of the rest of the data set.
- It shows the percentage of values that a particular element has surpassed.
- For example, if you score 75 points on a test, and are ranked in the 85th percentile, it means that the score 75 is higher than 85% of the scores.
- The percentile rank is calculated using the formula

  R= (P/100)* (N+1) where P is the desired percentile and N is the number of data points.

- The pth percentile of a sample, for a number p between 0 and 100, divides the sample such that,
  - p% of the sample values are less than the pth percentile
  - (100-p%) are greater than the pth percentile

Percentile

Steps to calculate the percentile rank:

1. Order the n samples values from smallest to largest.
2. Compute the quantity (P/100)(n+1), where n is the sample size.
3. If the above quantity is an integer, the sample value in this position is the percentile.
4. Otherwise, average the two sample values at the preceding and succeeding integer positions with respect to the quantity obtained in step 3.

Slide Courtesy:Dr.Uma

Percentile example

If the scores of a set of students in a math test are 25, 7, 9, 13, 2 and 8 what is the 15th percentile and 75th percentile?

Ans: Arrange the numbers in ascending order and give the rank ranging from 1(the lowest number) to 5 (the highest number)

Score	2	7	8	9	13	25

R = (P/100)(N+1)

= (15/100) (6+1)

= 1.05 (is it not an integer)

Percentile value = (1st element + 2nd element value)/2

= (2+7)/2

= 4.5 is the 15th percentile

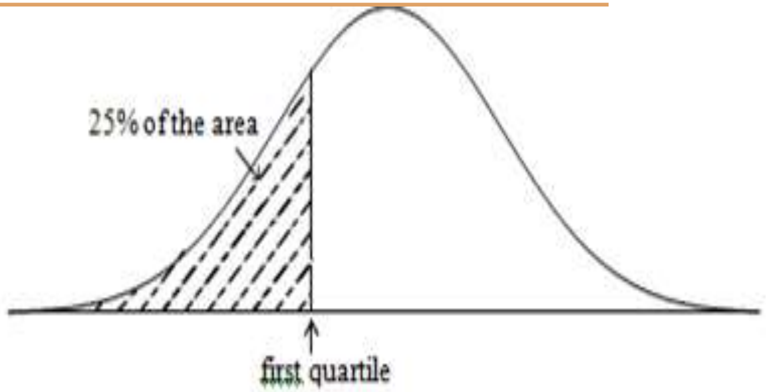
Thus, score 19 is the 75th percentile

Measures of Spread: Quartiles

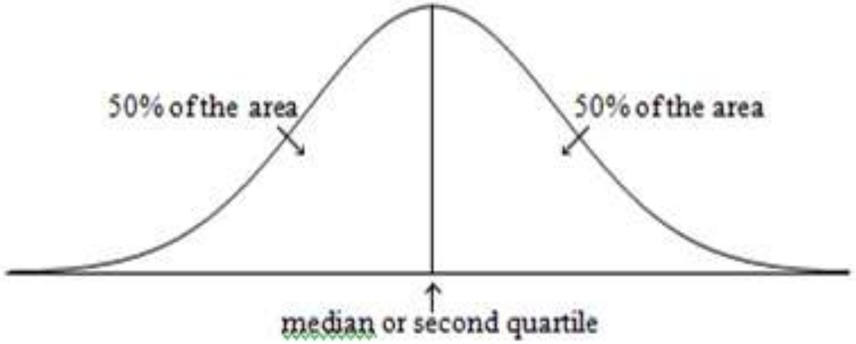
- Quartiles are the values that **divide a list of numbers into quarters**.
- Quartiles are obtained by first putting the list of numbers in order and then cutting the list into four equal parts.
- The Quartiles are at the "cuts" in the data.
- The first quartile, (Q1) is the middle number between the smallest number and the median of the data.
- The second quartile, (Q2) is the median of the data set.
- The third quartile, (Q3) is the middle number between the median and the largest number.

Measures of Spread: Quartiles

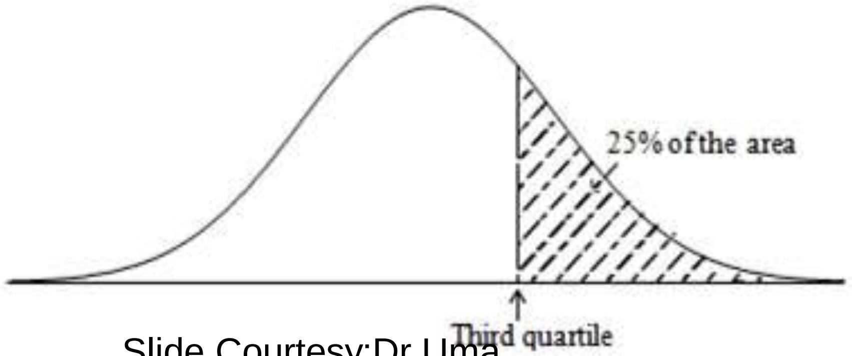
- The first quartile is the 25th percentile



- The median is the 50th percentile



- The third quartile is the 75th percentile

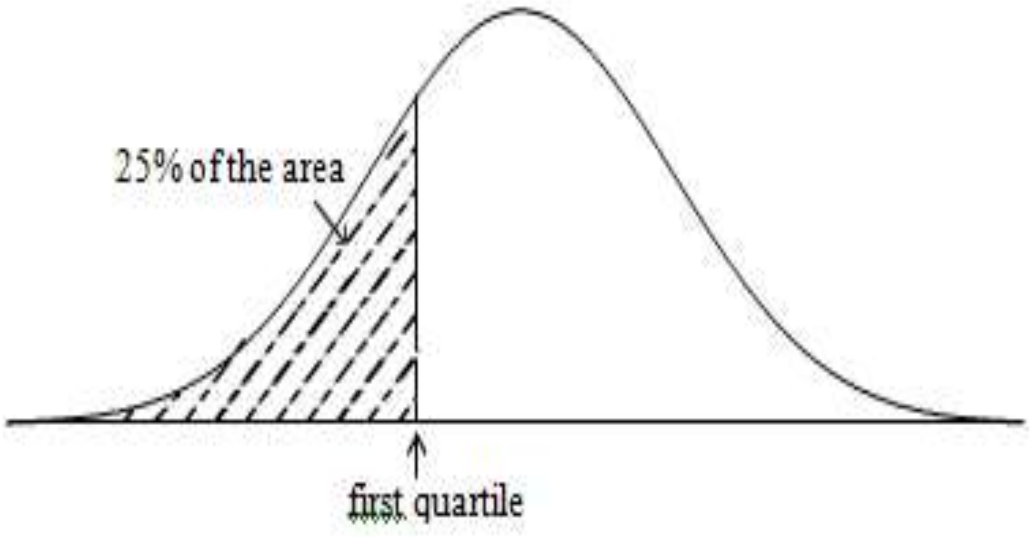


Slide Courtesy:Dr.Uma

Measures of Spread: Quartiles

➔ **The First Quartile:**

- The first quartile is the point which gives us 25% of the area to the left of it and 75% to the right of it.
- This means that 25% of the observations are less than or equal to the first quartile and 75% of the observations greater than or equal to the first quartile.
- The first quartile is also called the 25th percentile.



Slide Courtesy:Dr.Uma

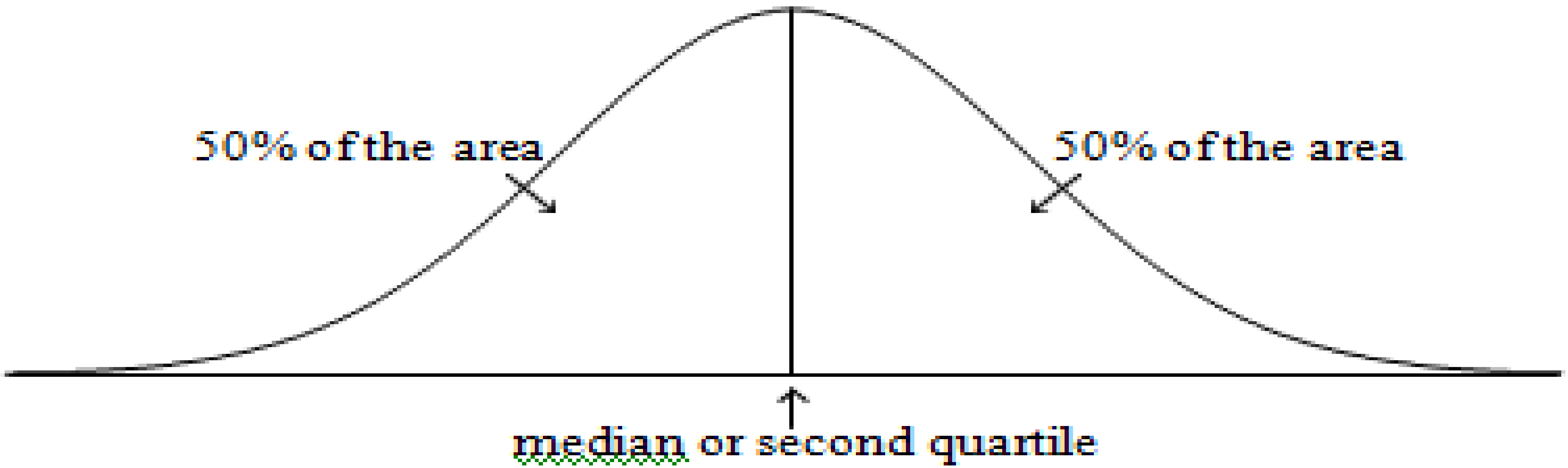
Measures of Spread: Quartiles

- To find the first quartile, compute the value 0.25(n +1).
- If this is an integer, then the sample value in that position is the first quartile.
- If not, then take the average of the sample values on either side of this value.

Measures of Spread: Quartiles

➔ **The Second Quartile or median:**

- It is easy to see how to divide the area in Figure into two equal parts, since the graph is symmetric.
- The point which gives us 50% of the area to the left of it and 50% to the right of it is called the second quartile or median
- Second quartile is calculated using the value 0.5(n+1)



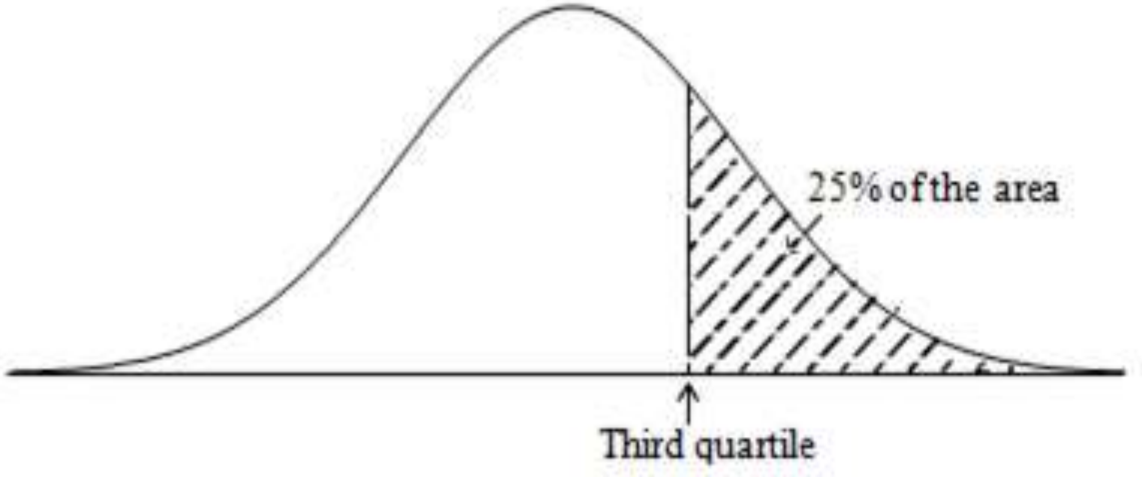
Slide Courtesy:Dr.Uma

Measures of Spread: Quartiles

➔ **The Third Quartile:**

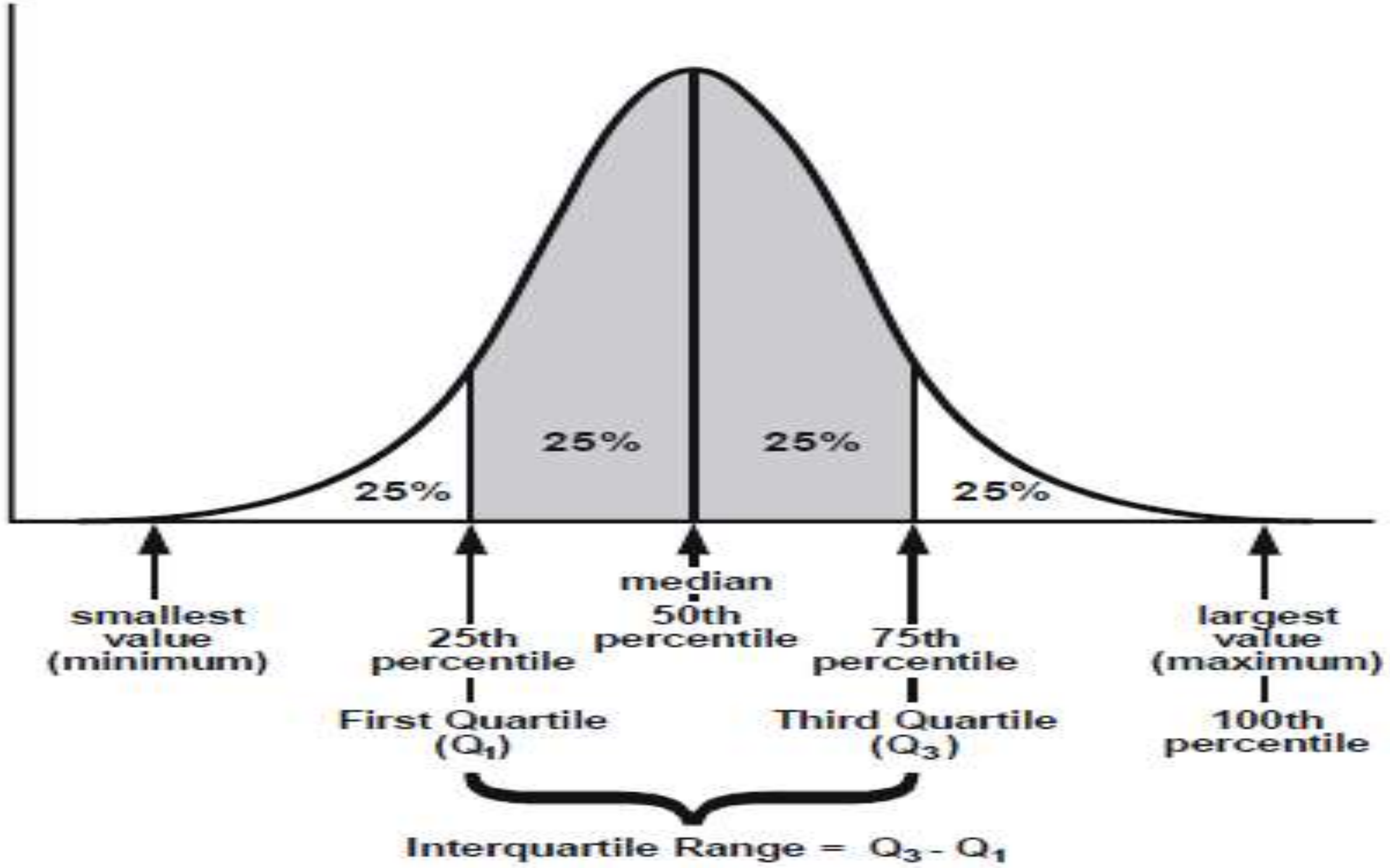
- The third quartile is the point which gives us 75% of the area to the left of it and 25% of the area to the right of it.
- This means that 75% of the observations are less than or equal to the third quartile and 25% of the observation are greater than or equal to the third quartile.
- The third quartile is also called the 75th percentile.
- The third quartile is computed in the same way, except that

  the value 0.75(n+1) is used.



Slide Courtesy:Dr.Uma

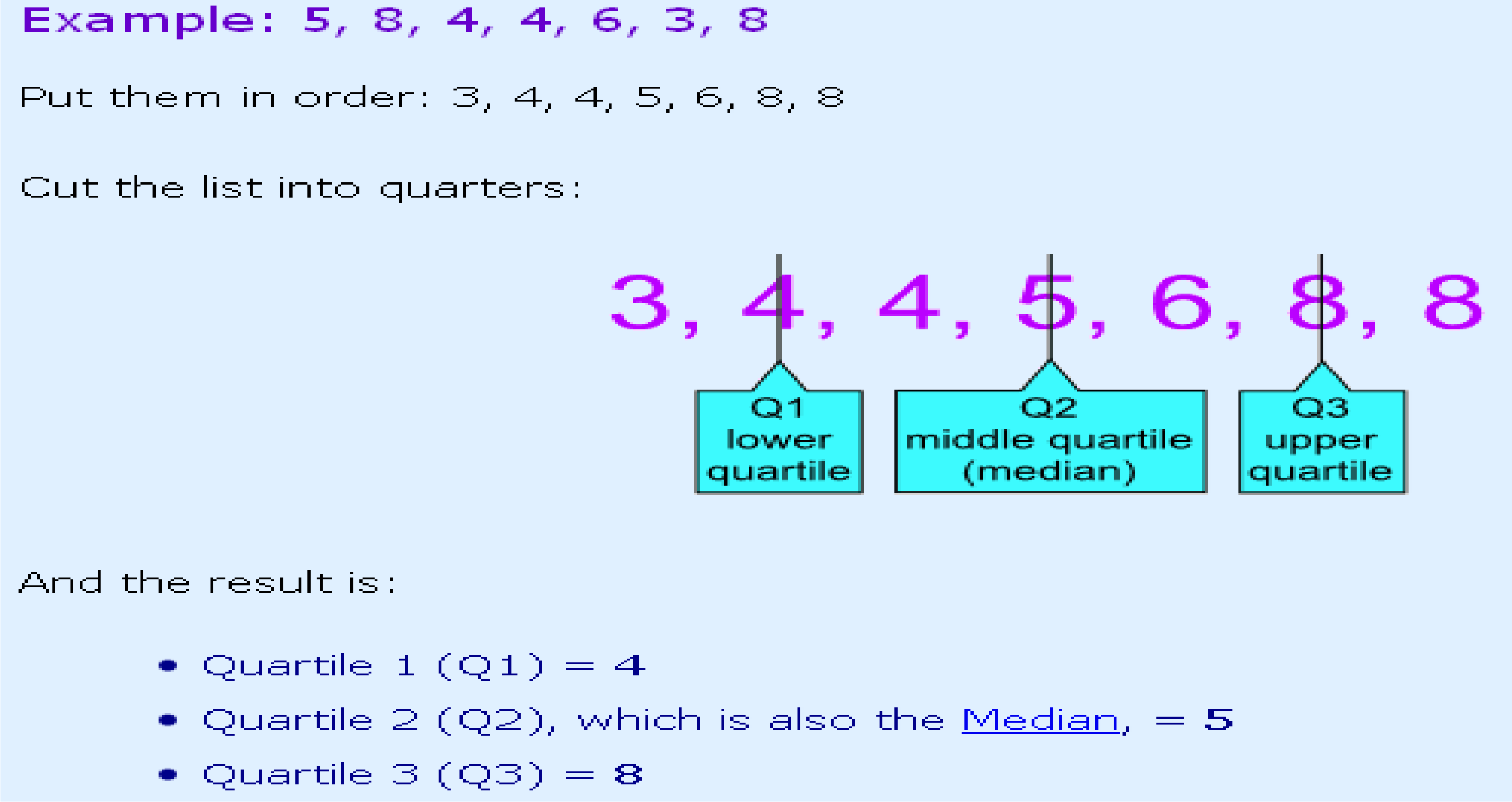
Measures of Spread: Quartile Summary



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

Slide Courtesy:Dr.Uma

Measures of Spread: Quartile example



Example: 5, 8, 4, 4, 6, 3, 8

Put them in order: 3, 4, 4, 5, 6, 8, 8

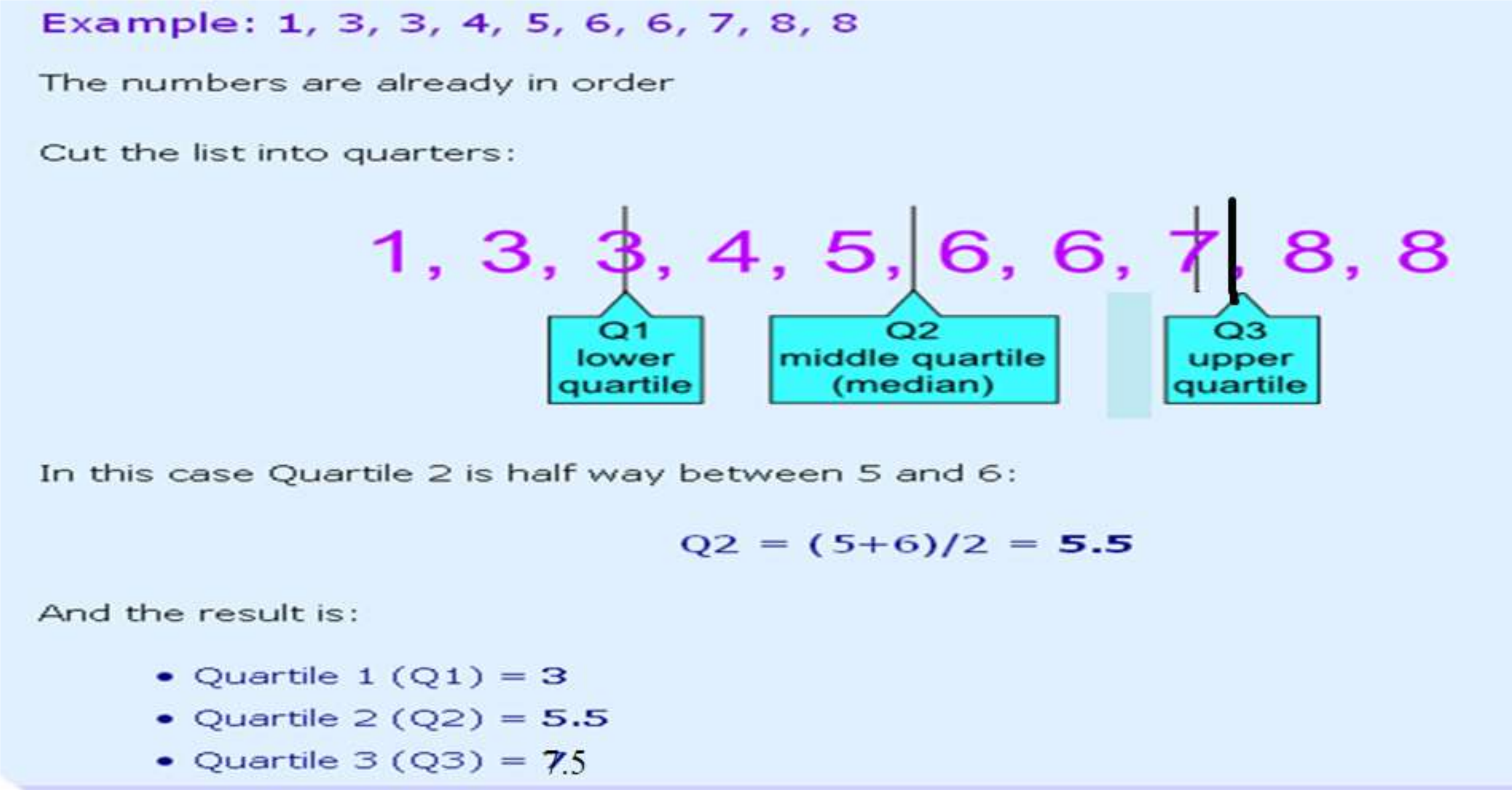
Cut the list into quarters:

And the result is:

- Quartile 1 (Q1) = 4
- Quartile 2 (Q2), which is also the Median, = 5
- Quartile 3 (Q3) = 8

Source: mathsisfun.com

Measures of Spread: Quartile example



Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are already in order

Cut the list into quarters:

In this case Quartile 2 is half way between 5 and 6:

Q2 = (5+6)/2 = **5.5**

And the result is:

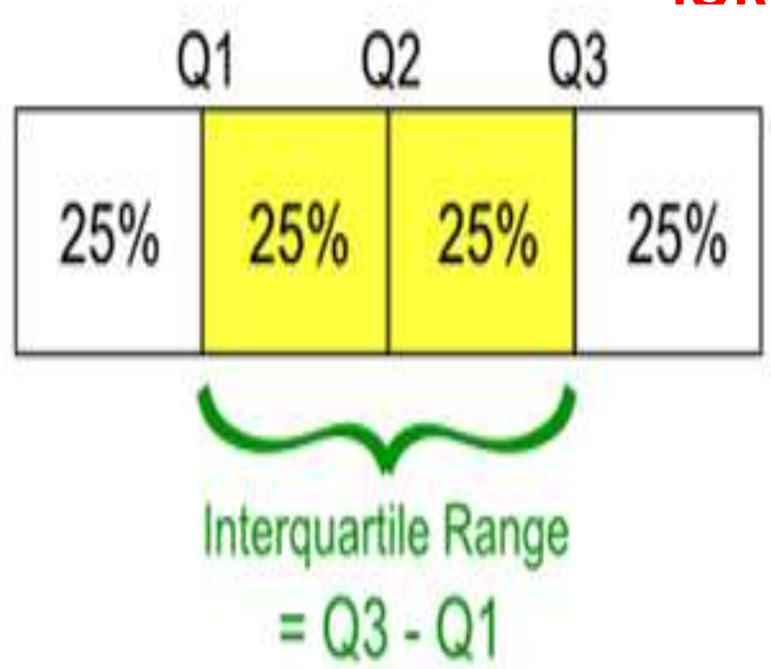
- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = 7.5

Source: mathsisfun.com

Measures of Spread: Inter-quartile Range

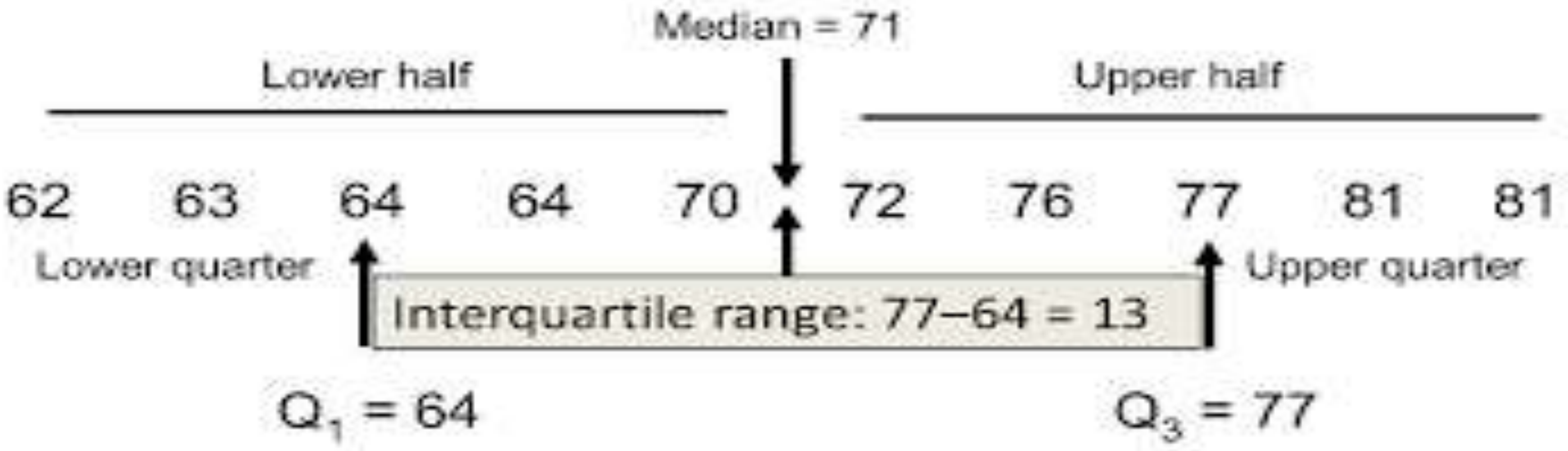
- Interquartile range is the distance or range between the 25th percentile and the 75th percentile.
- That is, quantifies the difference between the third and first quartiles.
- **Interquartile Range** = Upper Quartile(Q3) – Lower Quartile(Q1)

IQR = Q3 –Q1



Source: mathsisfun.com

Measures of Spread: Inter-quartile Range



Median = 71

Lower half

Upper half

62 63 64 64 70 72 76 77 81 81

Lower quarter

Upper quarter

Interquartile range: 77–64 = 13

$Q_1 = 64$

$Q_3 = 77$

Source: sphweb.bumc.bu.edu/

Slide Courtesy:Dr.Uma

Measures of Spread: Inter-quartile Range

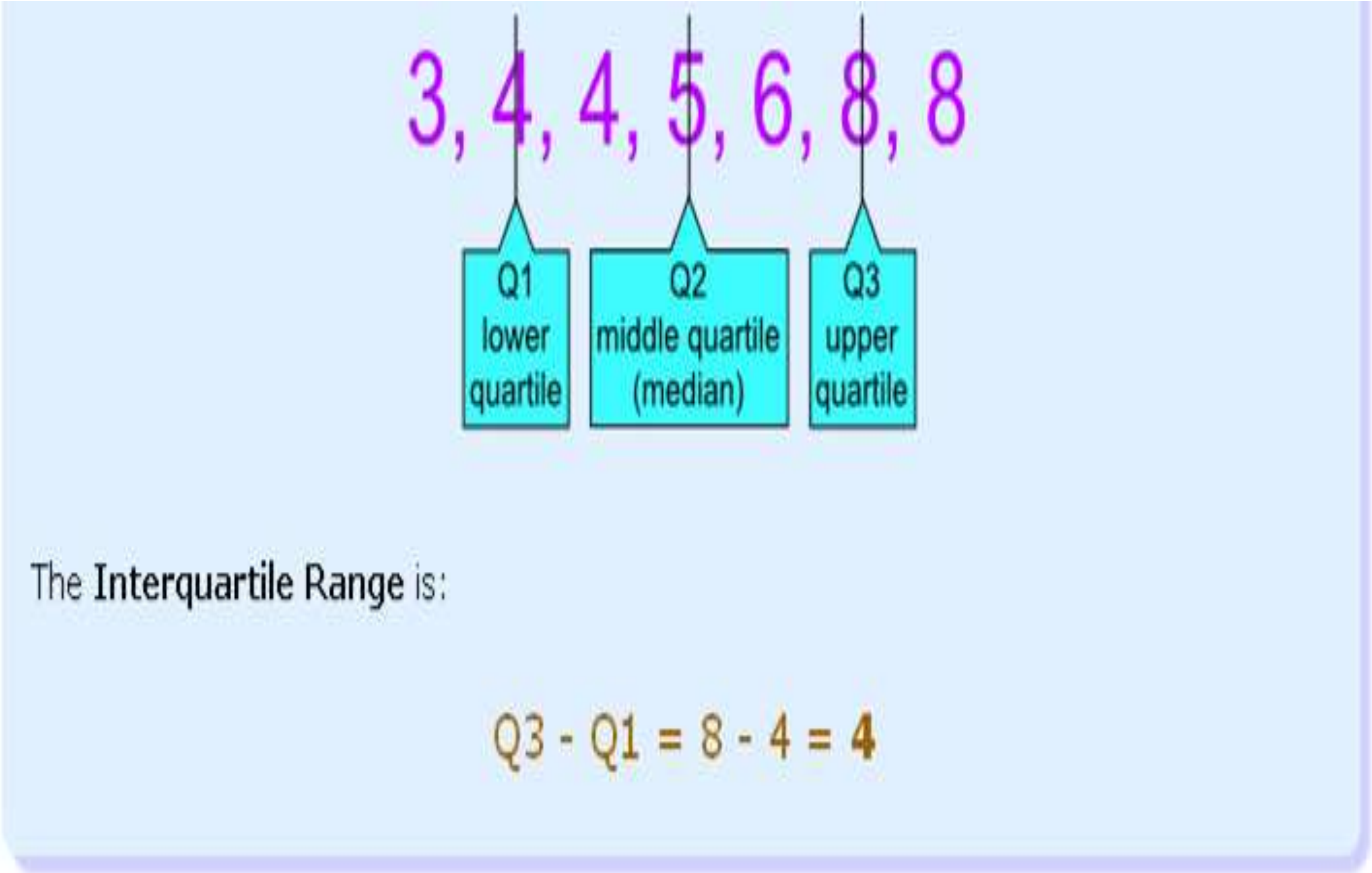
➔ Steps to find IQR :

1. Arrange the data scores in ascending order.
2. Find the median of the data set(the number in the middle).
3. Find the median of the lower half of the scores (Q1).
4. Find the median of the upper half of the scores (Q3).

Note: If the number of scores is even, the median is the
average of the two middle scores.

Slide Courtesy:Dr.Uma

Measures of Spread: Inter-quartile Range example



The **Interquartile Range** is:

$$Q3 - Q1 = 8 - 4 = \mathbf{4}$$

Source: mathsisfun.com

Measures of Spread: Inter-quartile Range question

For the following data sets, calculate the quartiles and find the interquartile range.

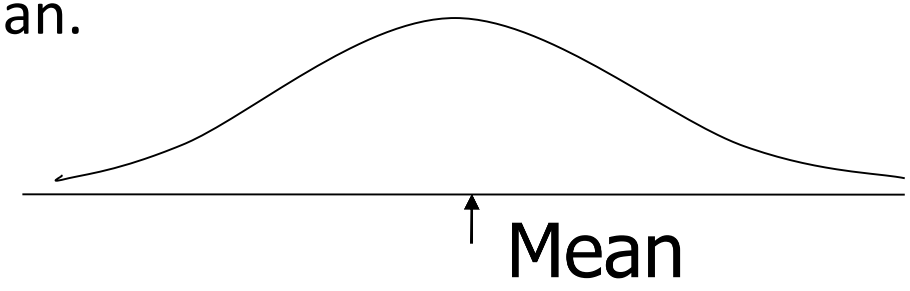
The following numbers represent the time in minutes that twelve employees took to get to work on a particular day.

18 34 68 22 10 92 46 52 38 29 45 37

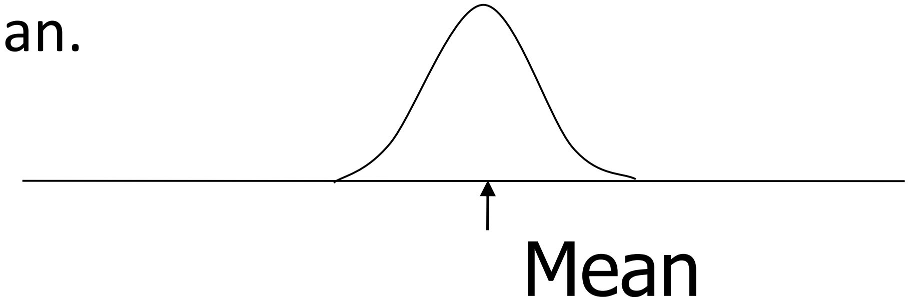
Slide Courtesy:Dr.Uma

Measures of Spread: Variance

- Variance is a measure of the spread of the recorded values on a variable.
- It is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value.
- The larger the variance, the further the individual cases are from the mean.



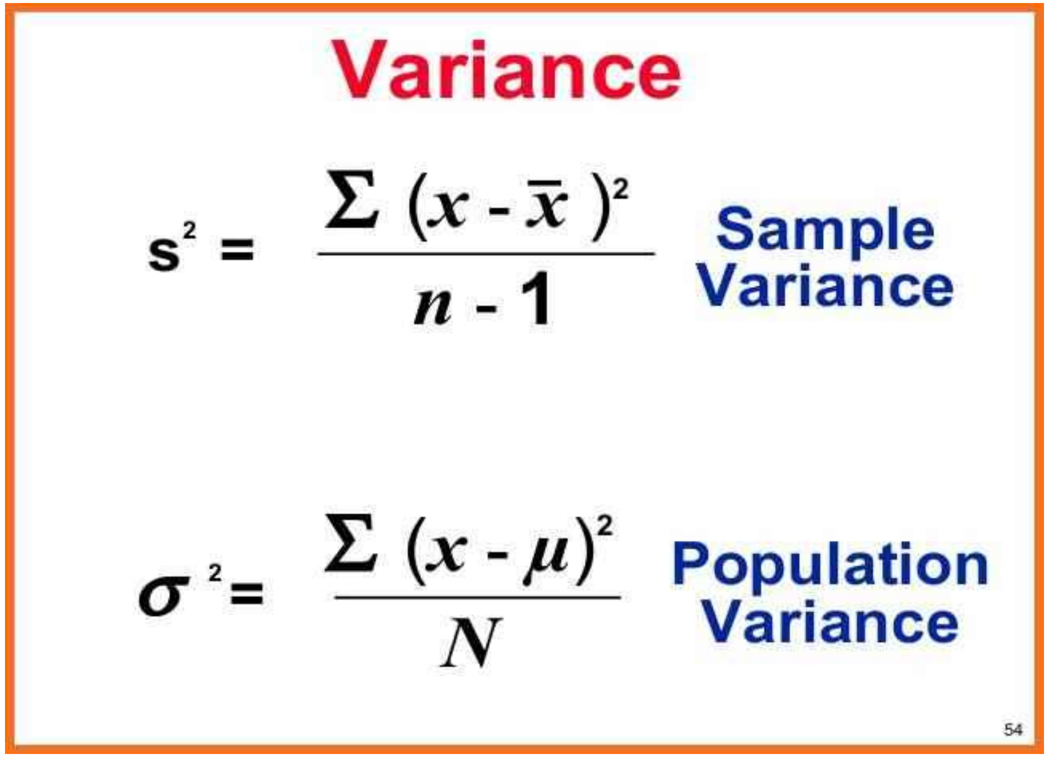
- The smaller the variance, the closer the individual scores are to the mean.



Measures of Spread: Variance

- It is the average of the distance that each score is from the mean (Squared deviation from the mean)
- Steps to calculate variance:

1. Find the mean value of the given data values.
2. Subtract mean from each data value.
3. Square each value that is obtained from step2.
4. Find the sum of all values that is obtained from step 3.
5. Divide the result that is obtained from step 4 by N(for population) and n-1(for sample).



Variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

54

Source: standard-deviation-calculator.com

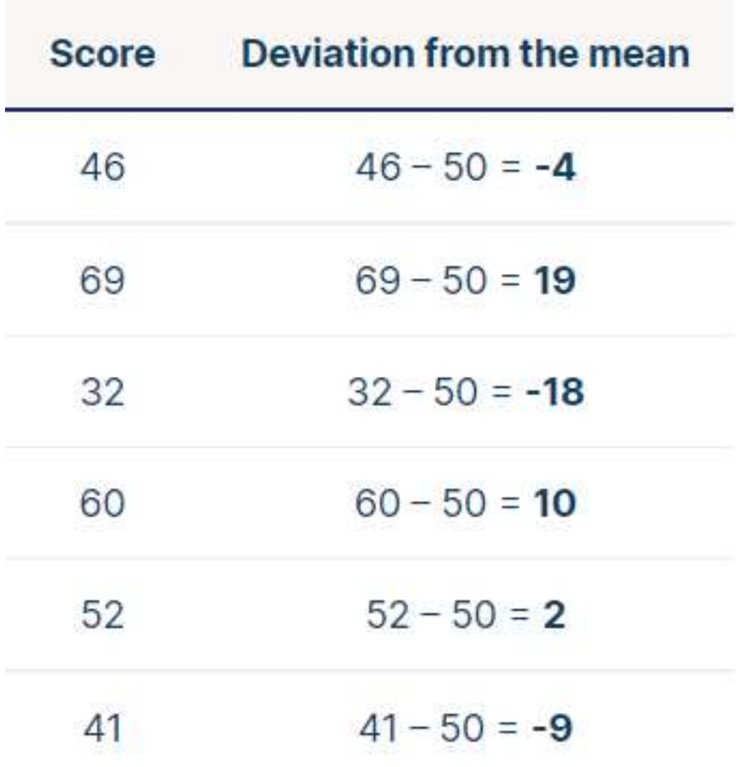
Slide Courtesy:Dr.Uma

Measures of Spread: Variance - example

Calculate the variance of the following dataset:46, 69, 32, 60, 52, 41

Solution:

- Find the mean (x̅): x̅ = (46 + 69 + 32 + 60 + 52 + 41) ÷ 6 = 50
- **Find each score's deviation from the mean.** Subtract the mean from each score to get the deviations from the mean. Since x̅ = 50, take away 50 from each score.



Score	Deviation from the mean
46	46 – 50 = **-4**
69	69 – 50 = **19**
32	32 – 50 = **-18**
60	60 – 50 = **10**
52	52 – 50 = **2**
41	41 – 50 = **-9**

Source: scribbr.com

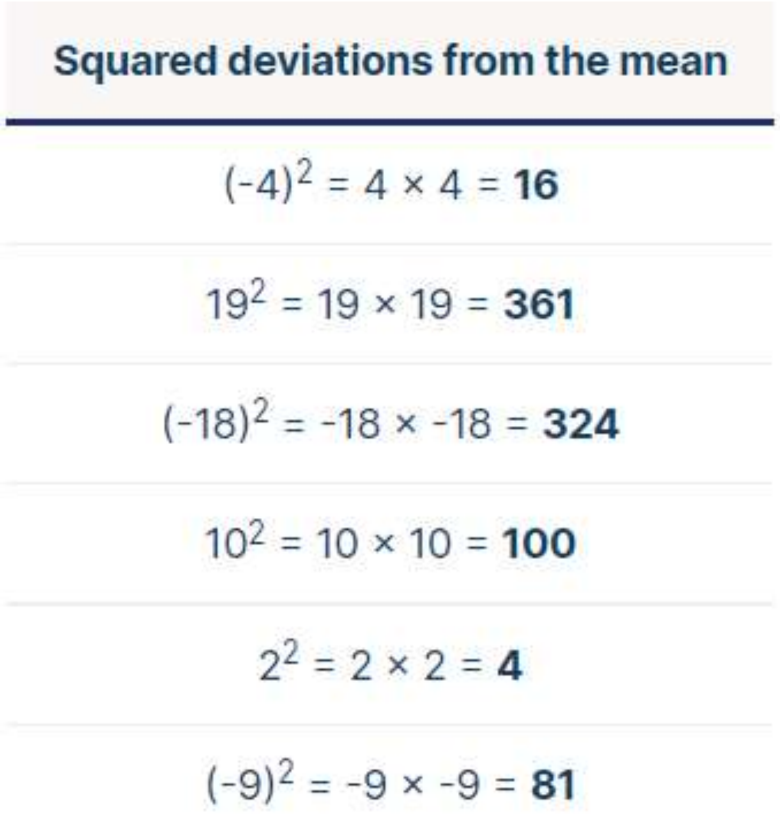
Measures of Spread: Variance - example

- Square each deviation from the mean
- Add up all of the squared deviations. This is called the sum of squares. 16 + 361 + 324 + 100 + 4 + 81 = 886
- Divide the sum of the squares by *n* – 1 (for a sample) or *N* (for a population).

  Since we're working with a sample, we'll use *n* – 1, where *n* = 6.

  886 ÷ (6 – 1) = 886 ÷ 5 = 177.2

- Variance = 177.2

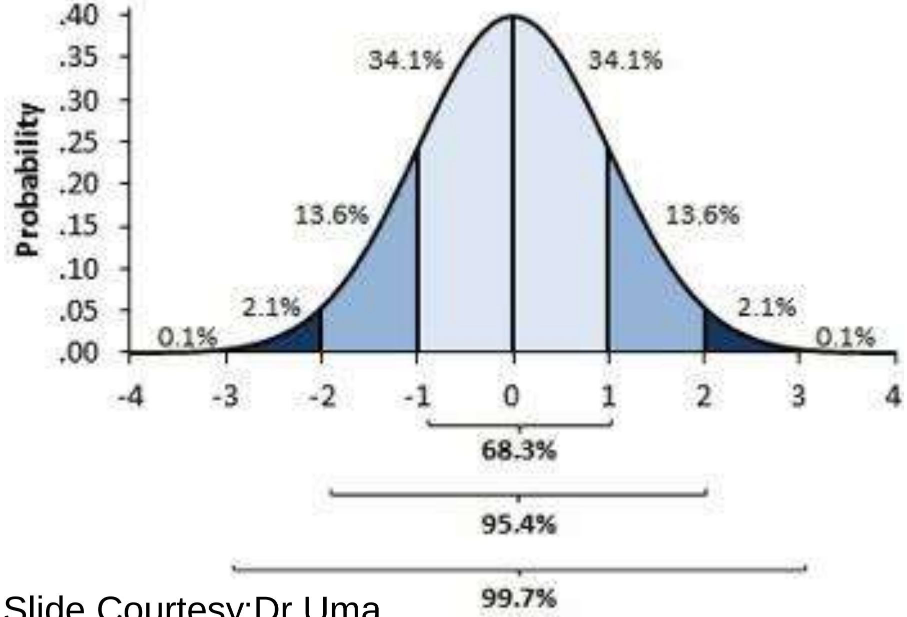


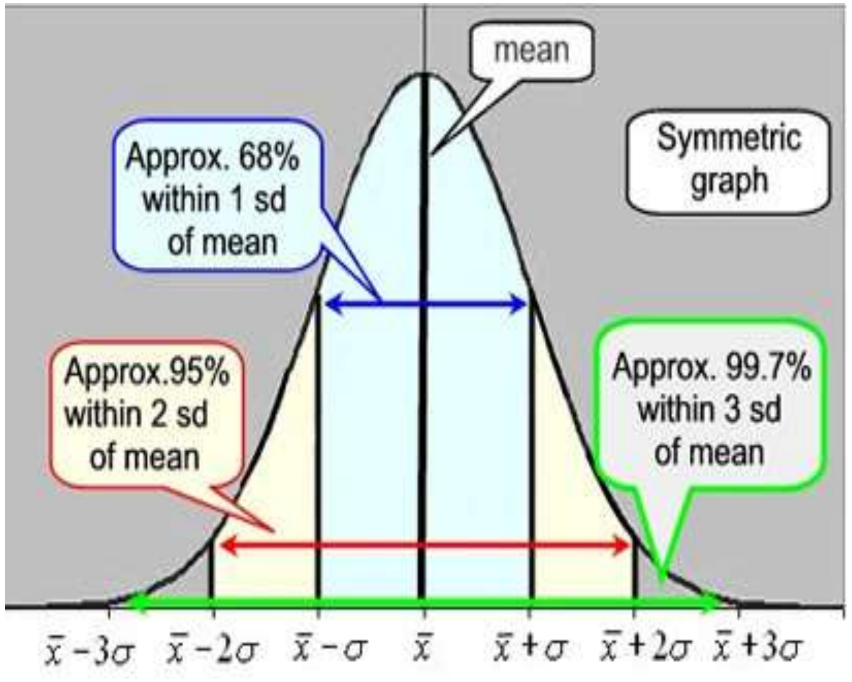
Squared deviations from the mean
$(-4)^2 = 4 \times 4 = \mathbf{16}$
$19^2 = 19 \times 19 = \mathbf{361}$
$(-18)^2 = -18 \times -18 = \mathbf{324}$
$10^2 = 10 \times 10 = \mathbf{100}$
$2^2 = 2 \times 2 = \mathbf{4}$
$(-9)^2 = -9 \times -9 = \mathbf{81}$

Source: scribbr.com

Measures of Spread: Standard Deviation

- Standard deviation signifies the deviation of the elements of the data set from the mean value of the distribution.
- It quantifies the amount of variation of a set of data values.
- It is a measure of the variability of a single item.
- The standard deviation does not decline as the sample size increases.
- The estimate of the standard deviation becomes more stable as the sample size increases.





Source: exceluser.com, MathBitsNotebook

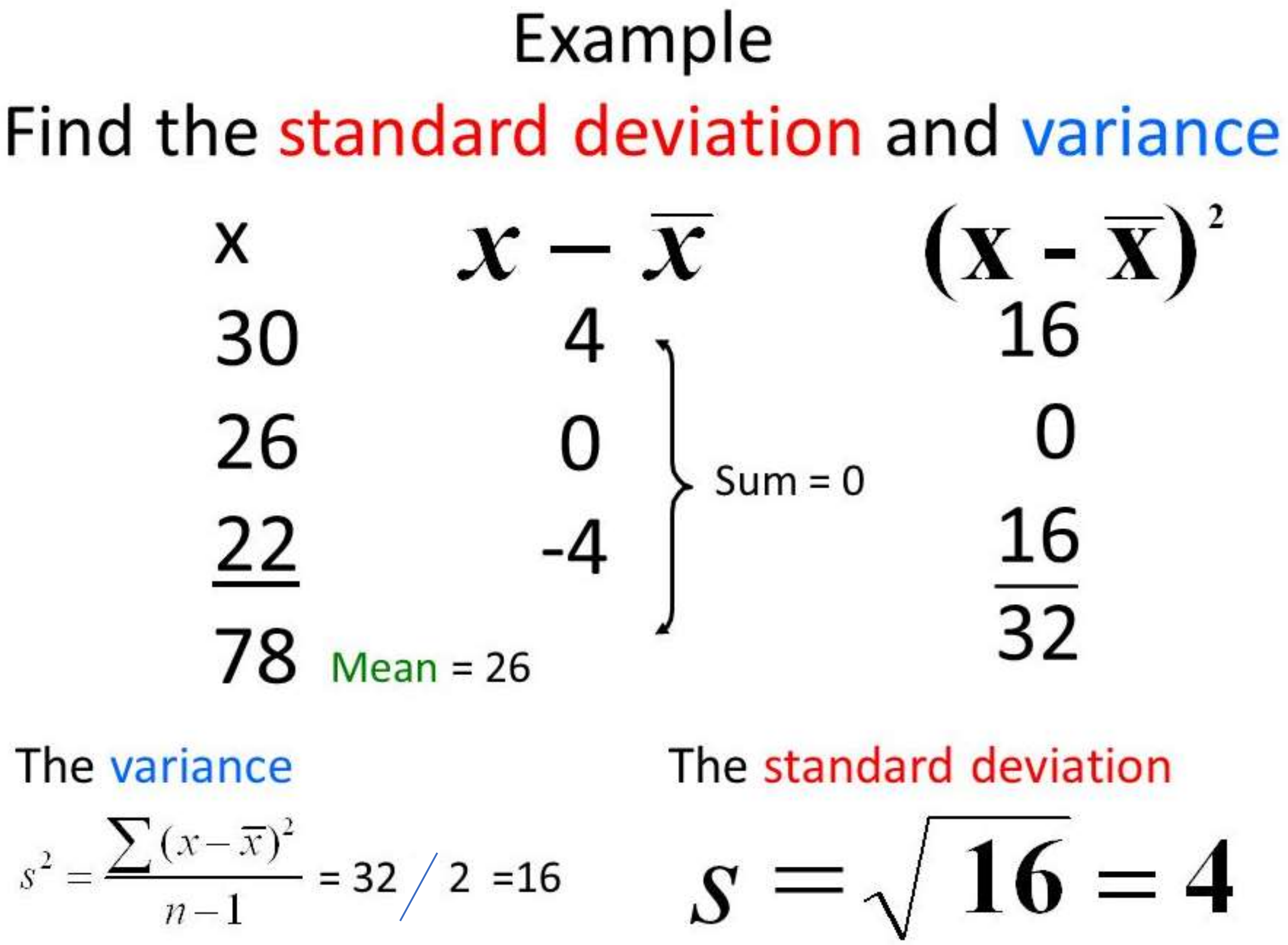
Slide Courtesy:Dr Uma

Measures of Spread: Standard Deviation

- Larger the standard deviation, greater amounts of variation around the mean.
- Std deviation = 0 only when all values are the same (only when you have a constant and not a “variable”)
- If you were to “rescale” a variable, the s.d. would change by the same magnitude.
- Like the mean, the standard deviation will be inflated by an outlier case value.

Measures of Spread: Standard Deviation

Standard Deviation = Square root of Variance



Example

Find the standard deviation and variance

x	$x-\bar{x}$		$(x-\bar{x})^2$
30	4		16
26	0	Sum = 0	0
22	-4		16
78 Mean = 26			32

The variance

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = 32/2 = 16$$

The standard deviation

$$s = \sqrt{16} = 4$$

Source: standard-deviation-calculator.com

Measures of Spread: Standard Deviation example

Calculate Standard Deviation for the following discrete data

Items	5	15	25	35
Frequency	2	1	1	3

Mean

$\bar{x}=5\times2+15\times1+25\times1+35\times3=10+15+25+105\ /7=22.15$

https://www.tutorialspoint.com/statistics/

Measures of Spread: Standard Deviation example

Calculate Standard Deviation for the following discrete data:

Items x	Frequency f	x¯	x−x¯	f(x−x¯)2
5	2	22.15	-17.15	588.245
15	1	22.15	-7.15	51.12
25	1	22.15	2.85	8.12
35	3	22.15	12.85	495.36
	N=7			∑f(x−x¯)2= 1142.845

https://www.tutorialspoint.com/statistics/

Measures of Spread: Standard Deviation example

Calculate Standard Deviation for the following discrete data:

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n} f_i (x_i - \bar{x})^2}{n-1}}$$

$$=\sqrt{1142.845/6}$$

$$\sigma = 13.80123$$

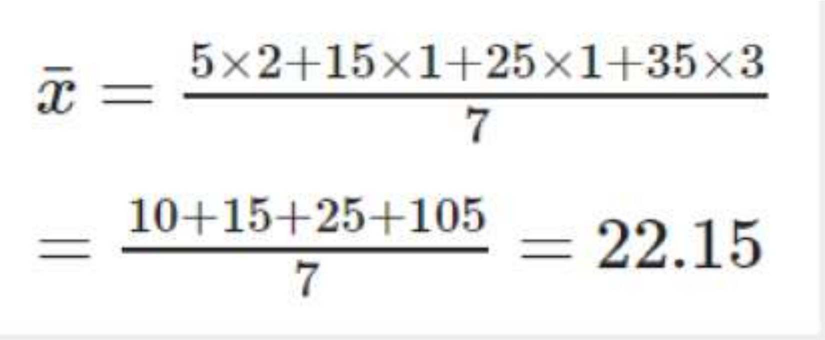
https://www.tutorialspoint.com/statistics/

Measures of Spread: Standard Deviation example

Calculate Standard Deviation for the following continuous data for a given Population :

Items	0-10	10-20	20-30	30-40
Frequency	2	1	1	3

In case of continous series, a mid point is computed as lower−limit+upper−limit/2



$$\bar{x} = \frac{5\times2+15\times1+25\times1+35\times3}{7}$$

$$= \frac{10+15+25+105}{7} = 22.15$$

https://www.tutorialspoint.com/statistics/

Measures of Spread: Standard Deviation example

Calculate Standard Deviation for the following data:

Midpoint x	Frequency f	x¯	x−x¯	f(x−x¯)2
5	2	22.15	-17.15	588.245
15	1	22.15	-7.15	51.12
25	1	22.15	2.85	8.12
35	3	22.15	12.85	495.36
	N=7			∑f(x−x¯)2=1 1142.845

https://www.tutorialspoint.com/statistics/588

Measures of Spread: Standard Deviation example

Calculate Standard Deviation for the following data:

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n} f_i (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{1142.845/6}$$

$$\sigma = 13.80123$$

https://www.tutorialspoint.com/statistics/

Practical Application for Understanding Variance and Standard Deviation

Even though we live in a world where we pay real dollars for goods and services (not percentages of income), most American employers issue raises based on percent of salary. Why do supervisors think the most fair raise is a percentage raise?

Answer:

1) Because higher paid persons win the most money.

2) The easiest thing to do is raise everyone's salary by a fixed percent.

If your budget went up by 5%, salaries can go up by 5%.

The problem is that the flat percent raise gives unequal increased rewards.

References

Text Book:

Statistics for Engineers and Scientists, William Navidi.

PES
UNIVERSITY
CELEBRATING 50 YEARS

THANK YOU

Dr.Mamatha H R

Professor, Department of Computer Science
mamathahr@pes.edu

+91 80 2672 1983 Extn 834