PES UNIVERSITY ONLINE

STATISTICS FOR DATA SCIENCE

Statistics Types and Summary

D. Uma
Department of Computer Science and Engineering
umaprabha@pes.edu

STATISTICS FOR DATA SCIENCE

Descriptive & Inferential Statistics

D. Uma

Department of Computer Science and Engineering

Statistics

1. WHAT IS STATISTICS?

2. TYPES OF STATISTICS

3. DESCRIPTIVE STATISTICS

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

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

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

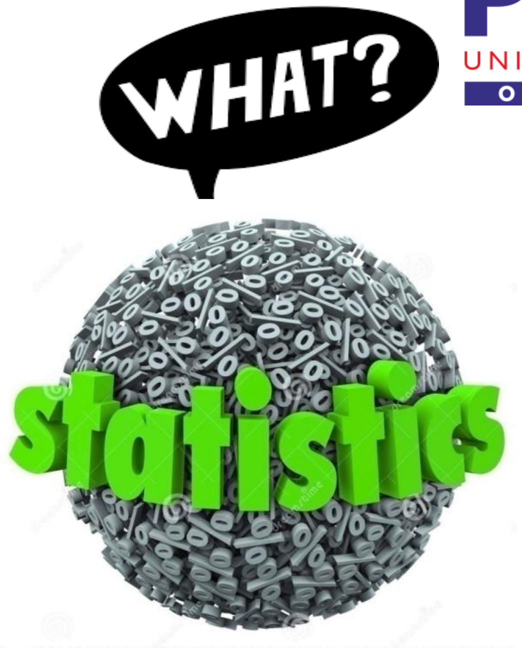
The word statistics comes from a **Italian** word **Statista** meaning **statement**

and

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

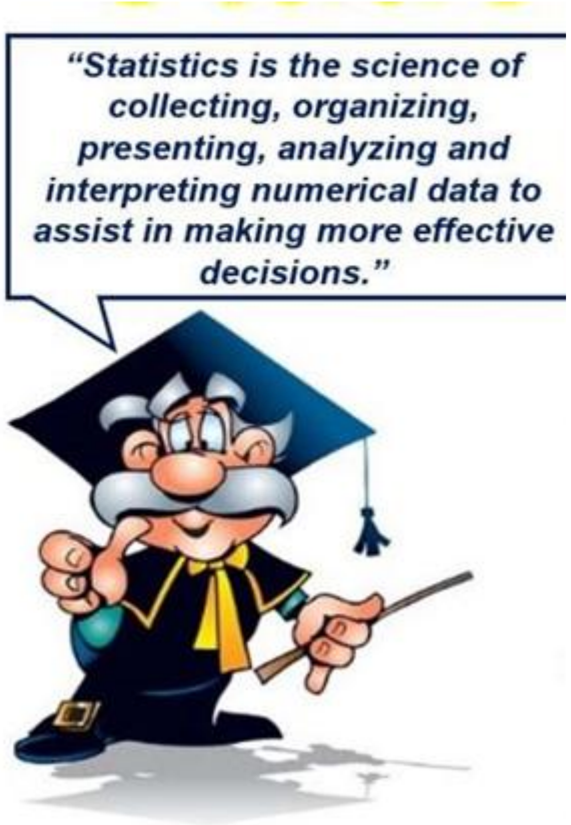
Statistics is a science of data.

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



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

Statistics



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

Statistics

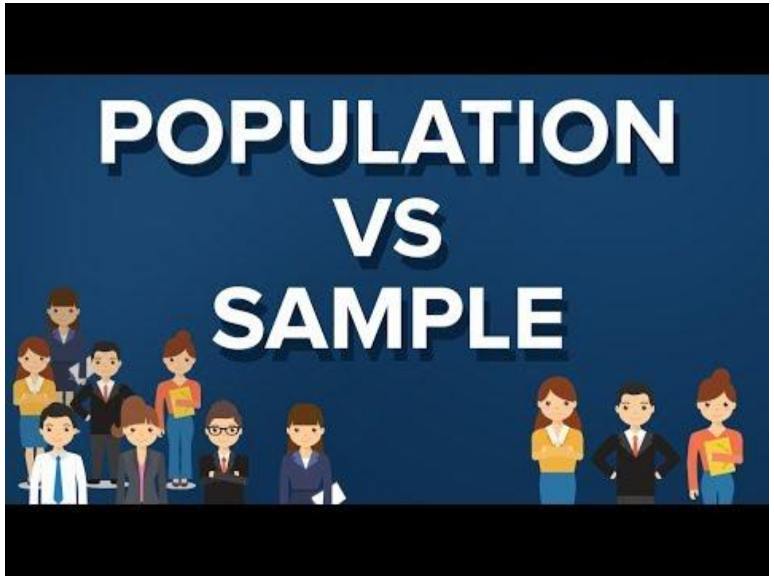
Statistics is the **branch of mathematics** that **transforms data** into **useful information** for decision makers.

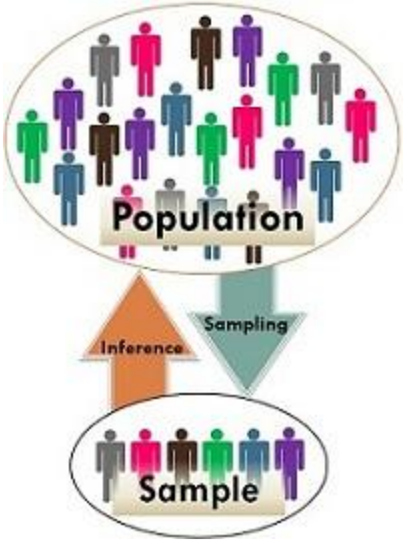
Sources: www.youtube.com, www. Colourbox.com

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.



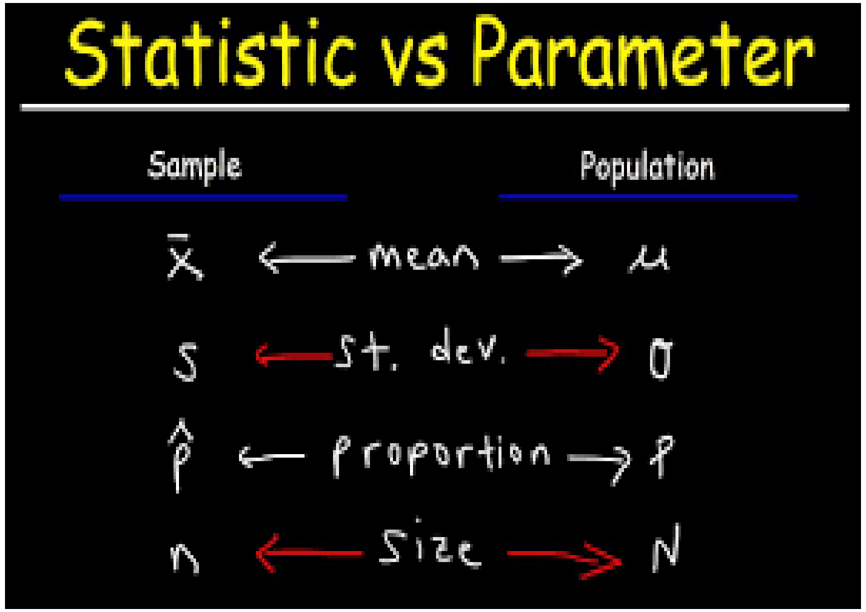


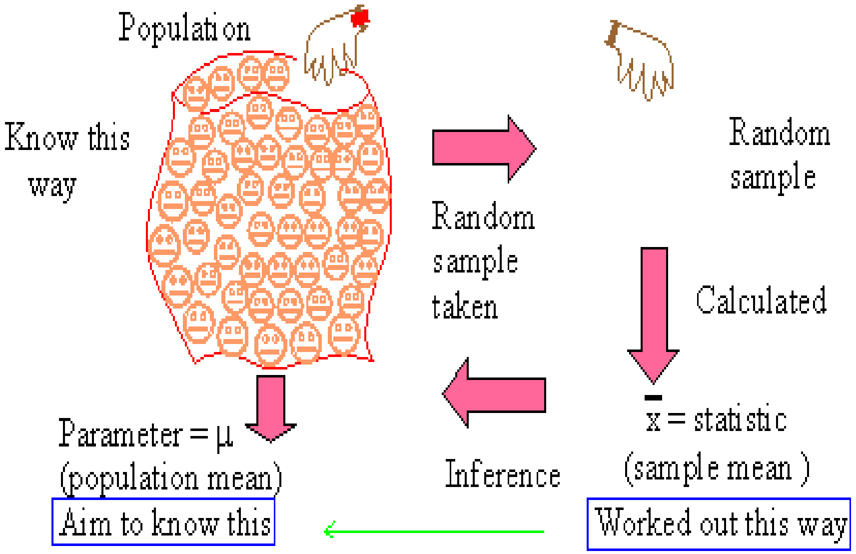
Source Image : 365datascience.com

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

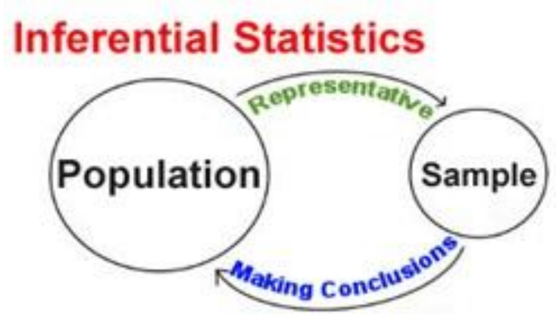
Processes of statistics

Statistics comprises of two processes.

1. **Describing set of data**

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

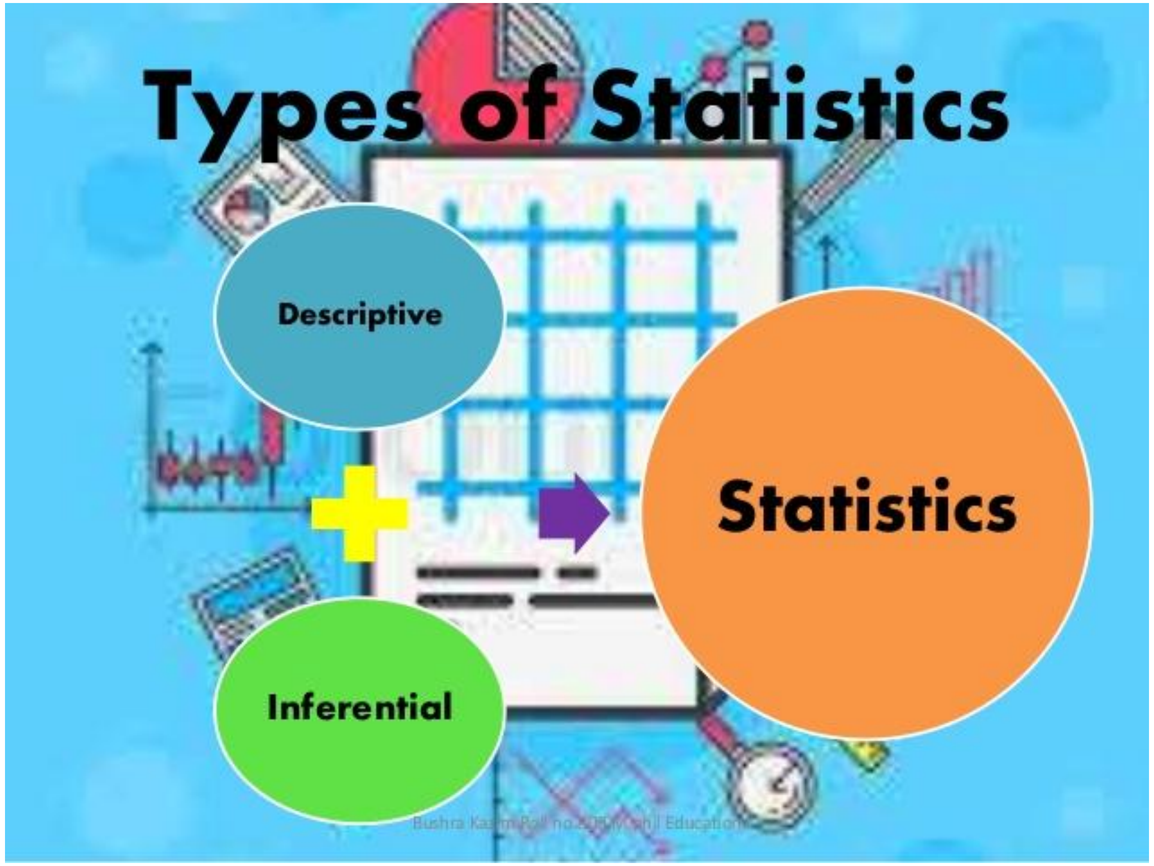
- Measures of Central Tendency
- Measures of Dispersion/Spread
- How it gets accumulates?



Types of Statistics

Numerical methods to organize, summarize and display data

Draws inferences from a population using sample



Source: www.slideshare.net

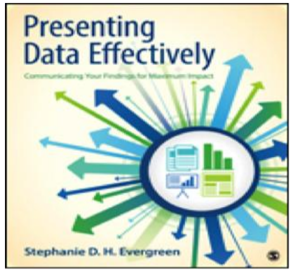
Descriptive statistics

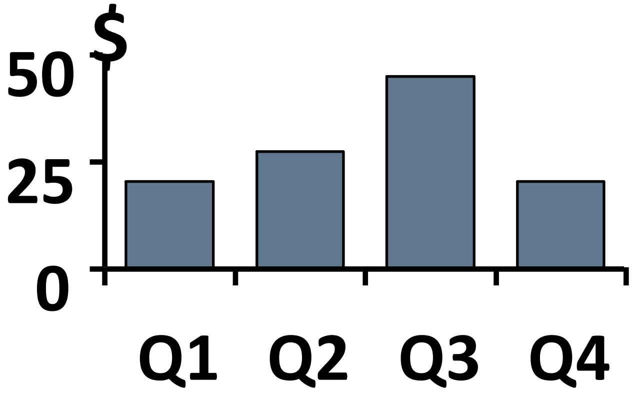
Purpose

- Describe Data

- Collect Data
  - e.g. Survey

- Present Data
  - e.g. Tables and graphs





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

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

Why Descriptive Statistics?

An Illustration : Which Group is Smarter?

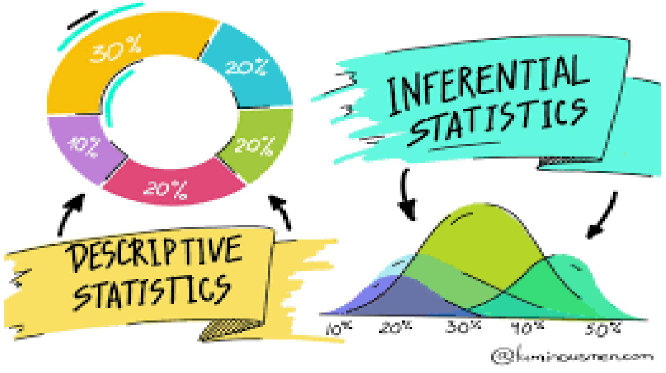
Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Source: www,slideshare.net

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	



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.

Figure speaks it all !!!

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.

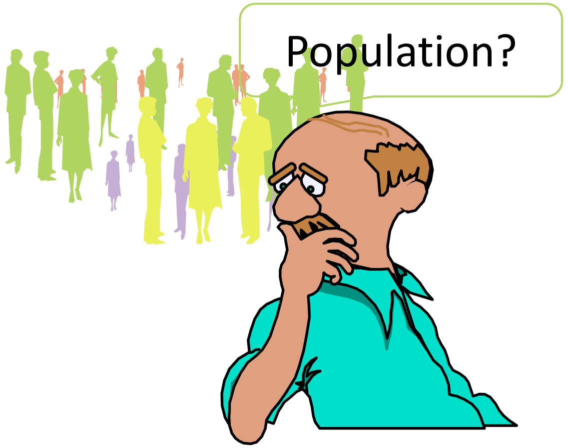
Inferential statistics

- 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/

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/

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

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:

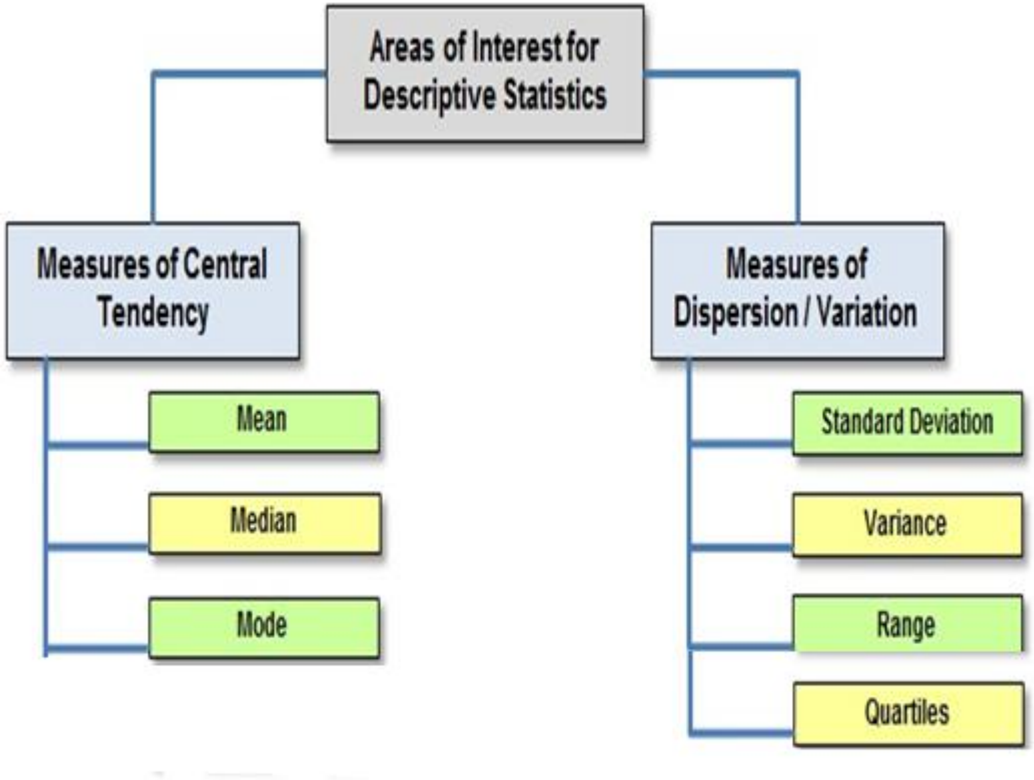
- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode
- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

Source: www.slideshare.net

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

Descriptive statistics

INDICATORS OF CENTRAL TENDENCY

- Mode
  - Most Frequently Occurring Score
- Median
  - Middle Score
- Mean
  - Arithmetic Average, *etc.*

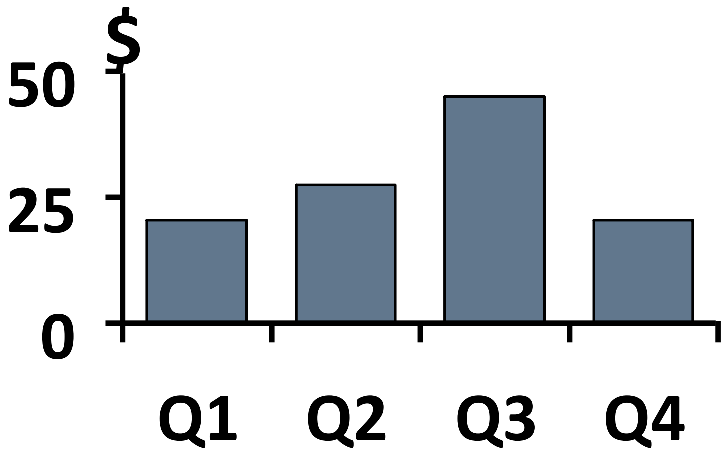
13/07/2018 Descriptive Statistics 10

Source: www.slideshare.net

Descriptive statistics

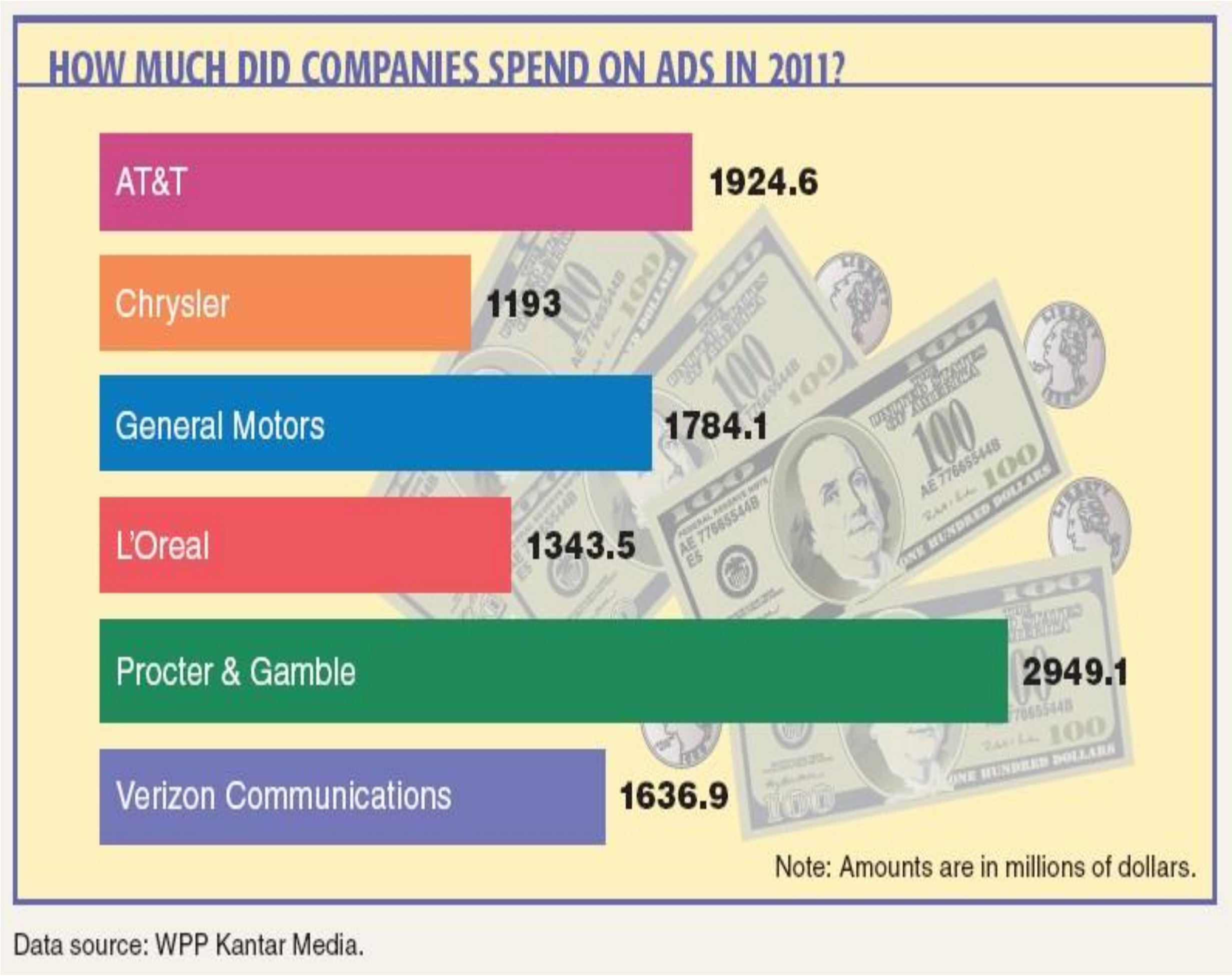
- **Descriptive statistics** are methods for organizing and summarizing data.

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



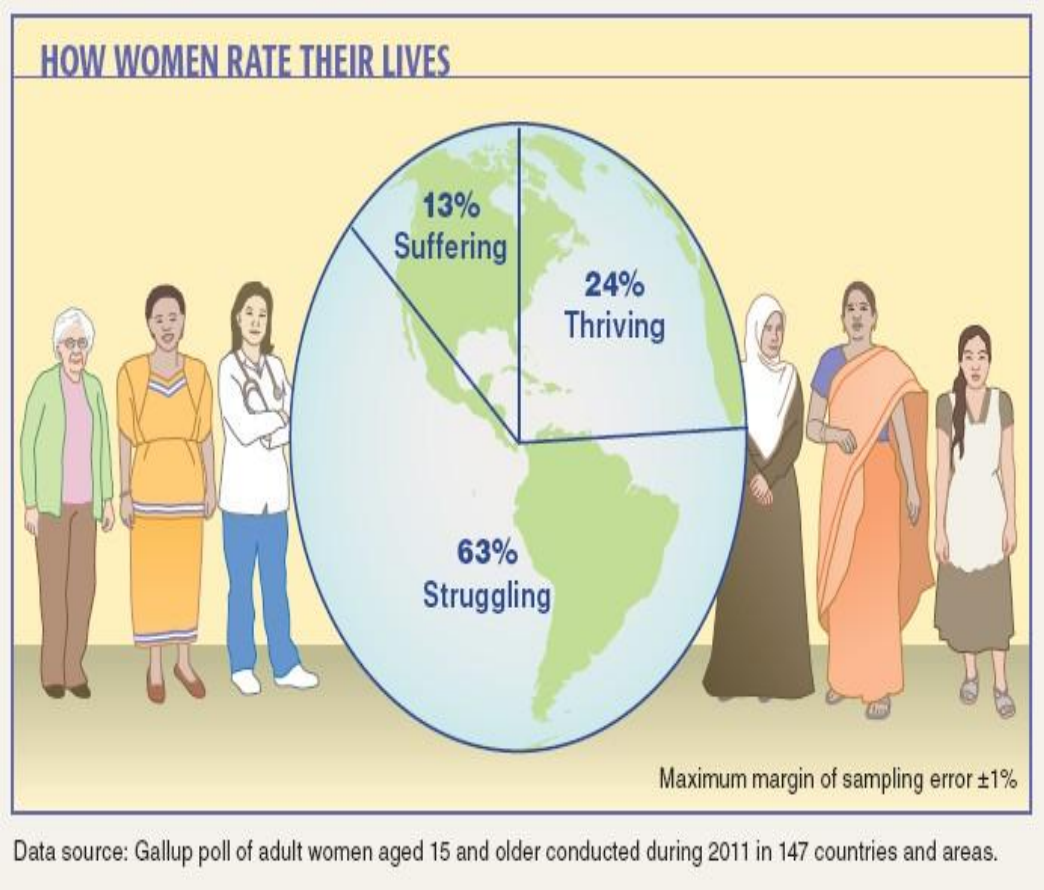
$\bar{X} = 30.5 \quad S^2 = 113$

Descriptive statistics



Source: luminousmen.com/post/

Descriptive statistics



Source: luminousmen.com/post/

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

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

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:

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

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

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

Source: sixsigma-institute.org

Measures of Central Tendency: Mean

Mean

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

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

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

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

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

Measures of Central Tendency: Median

Median

Annual Salary: 10k 11k 11k 15k 15k 15k $\boxed{\text{19k 20k}}$ 21k 21k 22k 22k 24k 25k

For an even number of scores take the mean of the middle two:

$$\frac{(19 + 20)}{2} = 19.5$$

$$50^{th}\ \text{Percentile} = \frac{n+1}{2}$$

Advantages

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

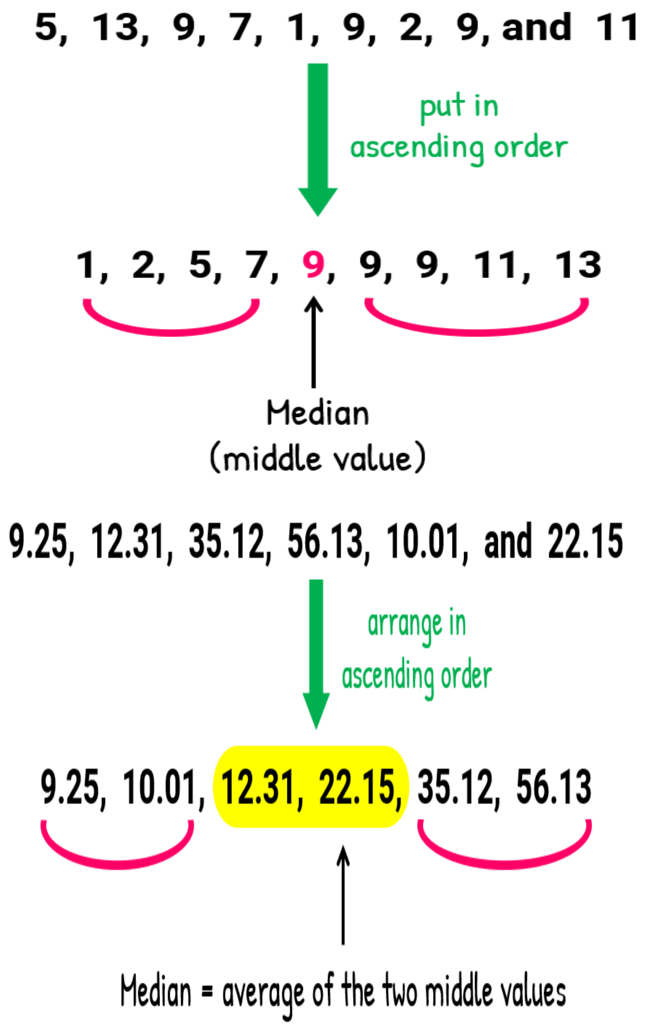
Disadvantages

- 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.



Source: Chilimath

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

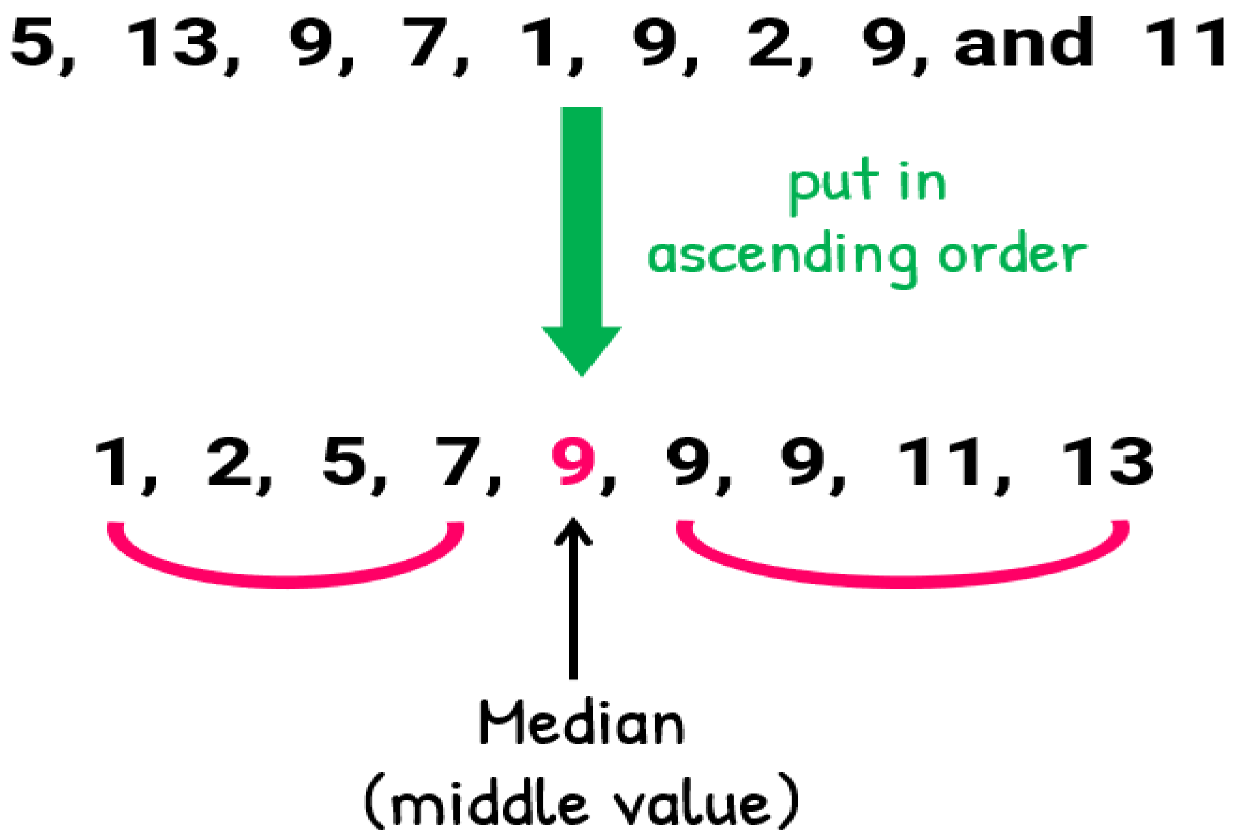
To calculate the median, we need to put the numbers in order and find the middle value.

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.

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.

Measures of Central Tendency: Median- Example

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

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:

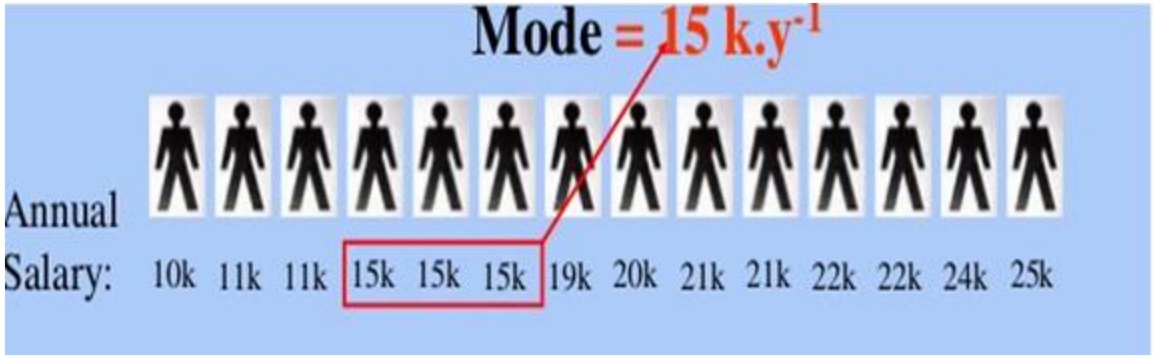
Its value is perceived as it is. It cannot be utilized for further algebraic treatment.

Source : Chilimath

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

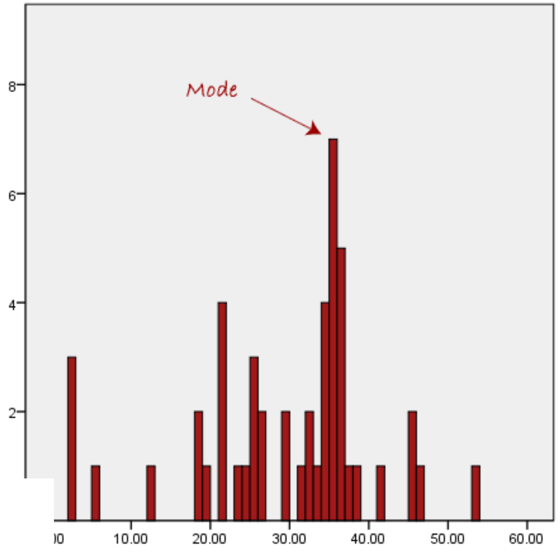
- Terminal Statistic
- A given sub-group could make this measure unrepresentative.

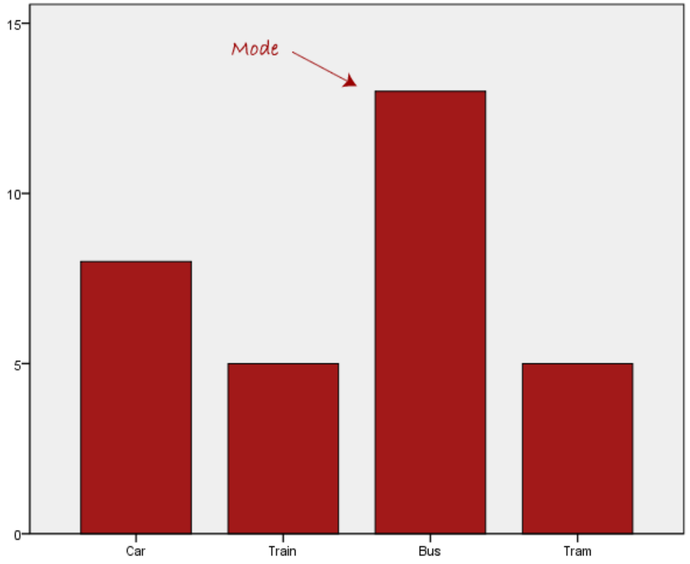
Source: slideshare.net

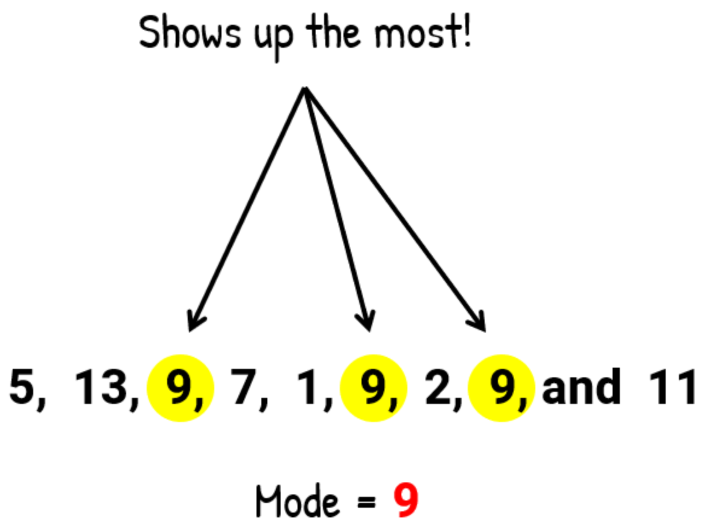
Measures of Central Tendency: Mode

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

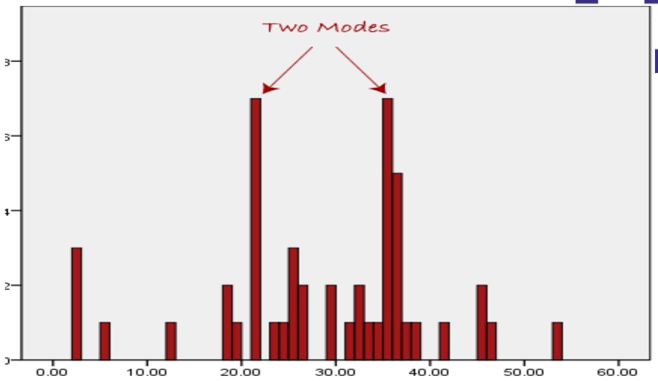
Measures of Central Tendency: Mode - Example

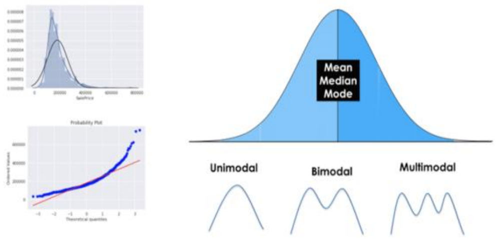
Consider the data given below:

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

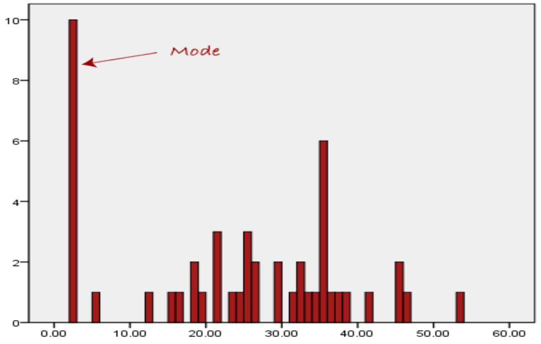
3 4 **5** **5** **5** 7 9 12 14 16 19

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

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**".

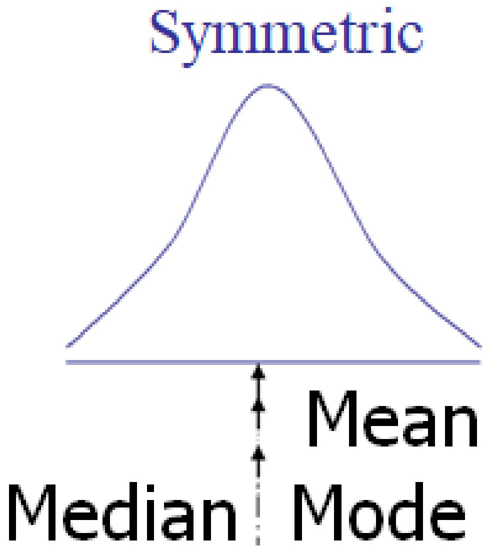
- Depends on whether or not data have one ("**unimodal**") or more ("**multimodal**") modes.

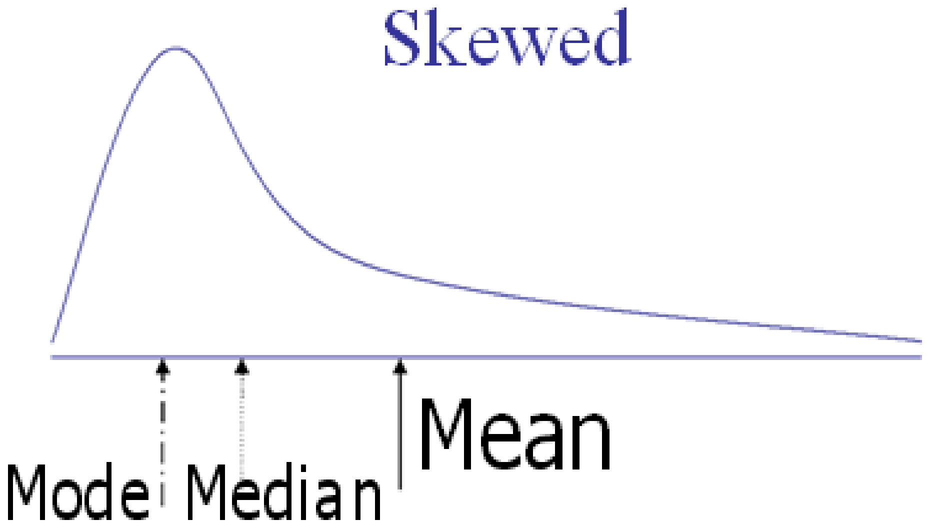
Source:www.slideshare.net

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.

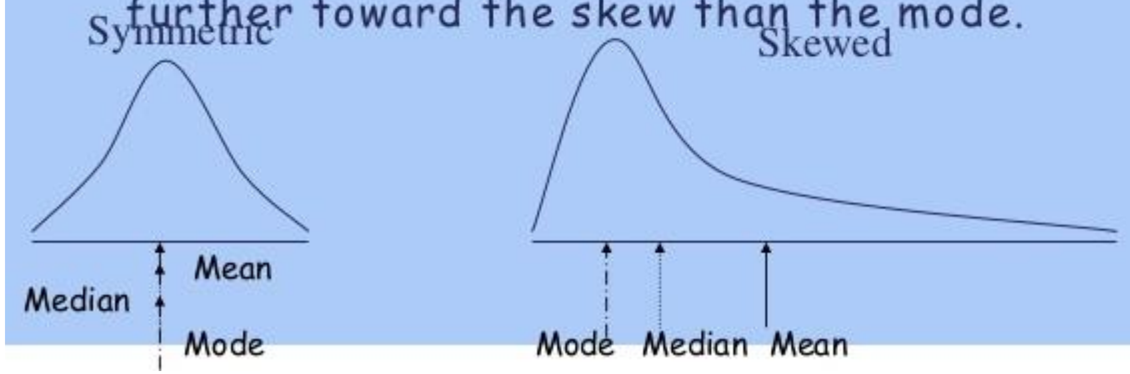




Source:www.slideshare.net

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:

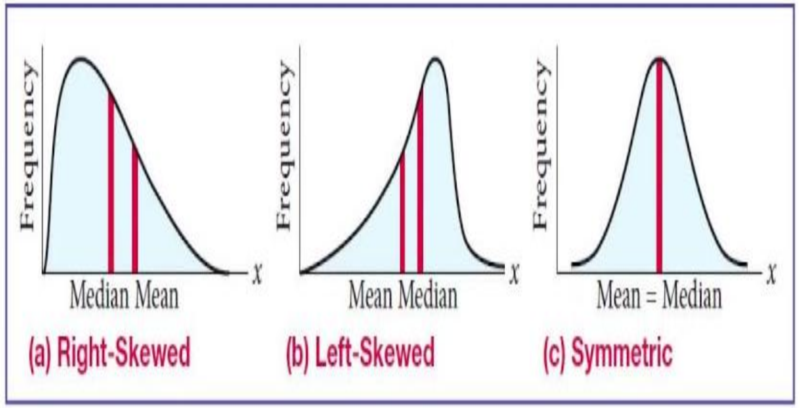
9, 15, 11, 12, 3, 5, 10, 20, 14, 6, 8, 8, 12, 12, 18, 15, 6, 9, 18, 11

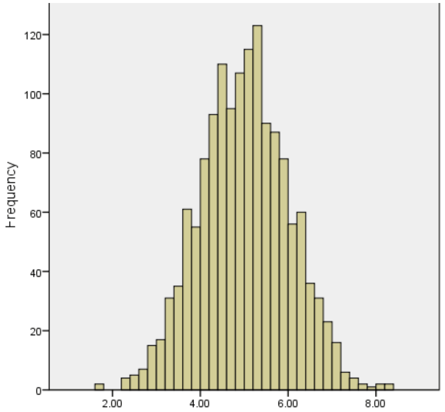
Find the mean, median and mode

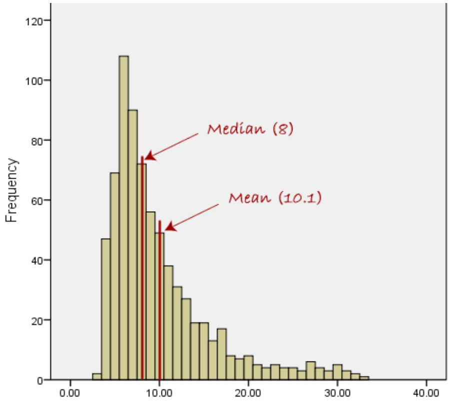
Symmetric and Skewed Distributions:

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”.

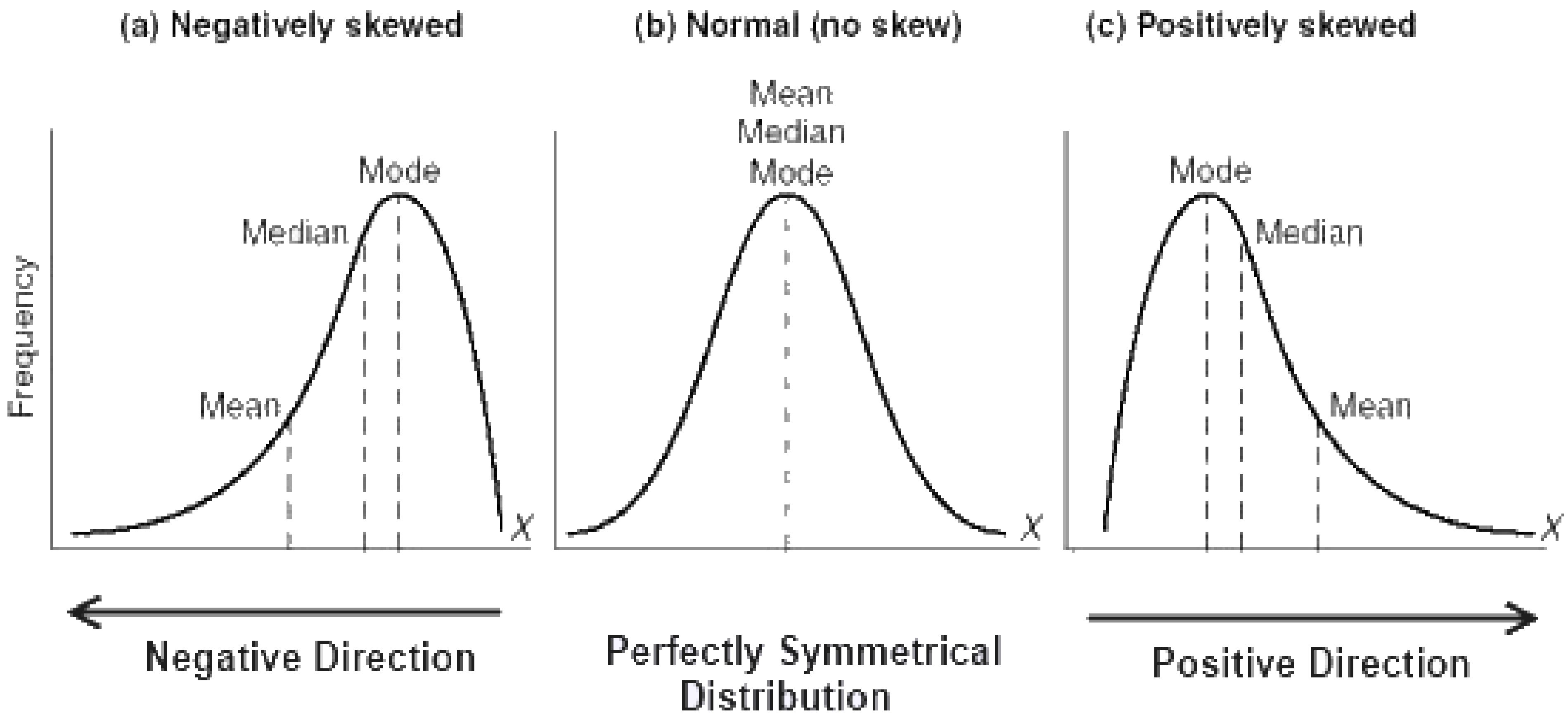
If **mean = median = mode**, the shape of the distribution is **symmetric**.

- 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**)

Symmetrical vs Skewed data



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

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

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

Measures of spread: Range

Range = Maximum Value – Minimum Value

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

put in ascending order

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

lowest highest

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?

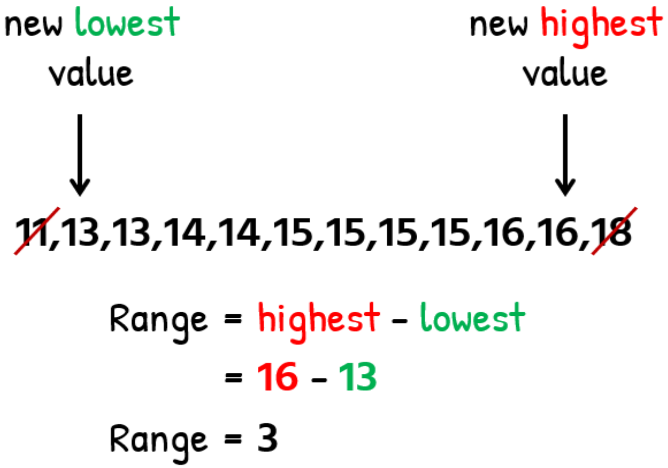
Source: Chilimath.com

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).



Case 3

Source: Chilimath.com

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

new lowest value

new highest value

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

Range = highest - lowest

= 16 - 13

Range = 3

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.

Measures of Dispersion: Range - Example

AGES OF STUDENTS

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

Range = highest - lowest

= 16 - 13

Range = 3

Source: Plantlet

Measures of Dispersion: Range - Example

new lowest value

new highest value

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

Range = highest − lowest

= 16 − 13

Range = 3

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.

Measures of Dispersion: Quartiles

As the name ‘**quartile**’ suggests, we want to divide the data into **four equal parts**

The first quartile is the 25th percentile

The median is the 50th percentile

The third quartile is the 75th percenile

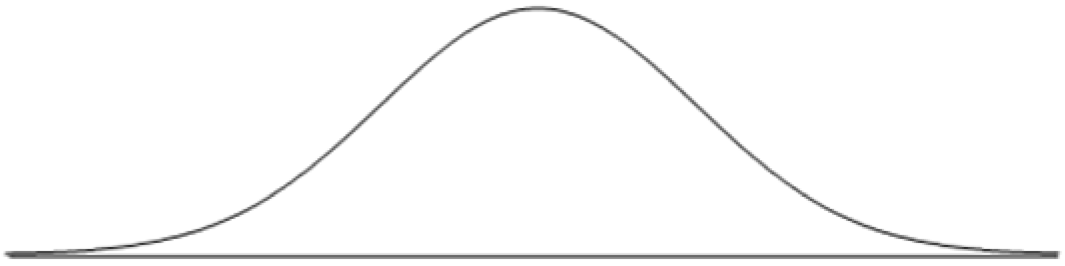


Figure : Graph representing heights of adult males.

In the above graph, we want to divide the area under our curve into four equal areas.

First put the list of numbers in order

Then cut the list into four equal parts

The **Quartiles** are at the "**cuts**"

Measures of Dispersion: Quartiles

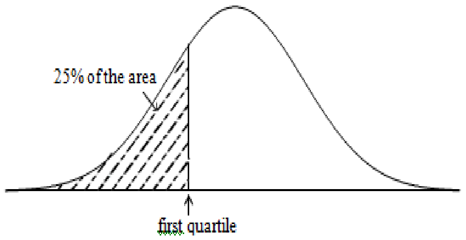
The first quartile is the 25^{th} percentile

The median is the 50^{th} percentile

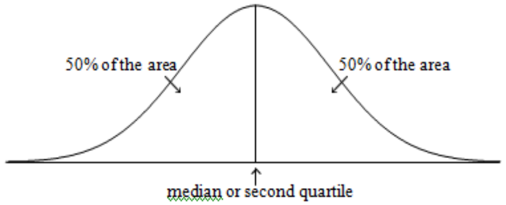
The third quartile is the 75^{th} percenile

Measures of Dispersion: Quartiles

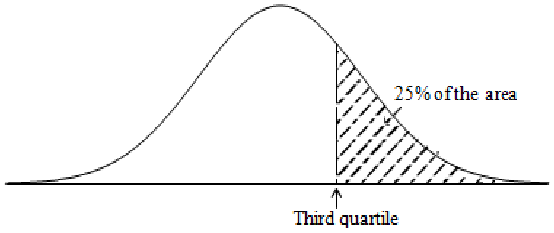
The first quartile is the 25th percentile



The median is the 50th percentile



The third quartile is the 75th percentile

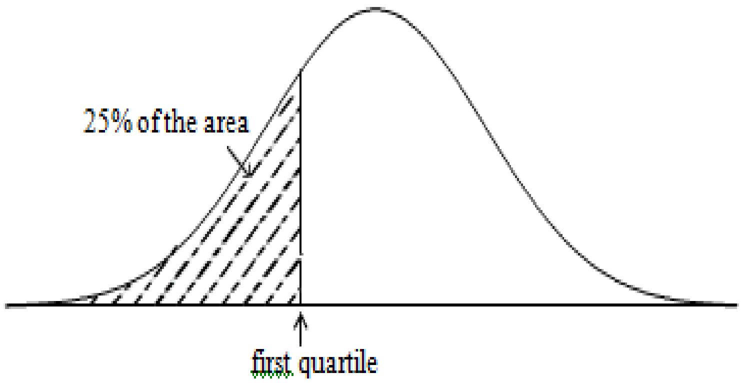


The first quartile

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.



Measures of Dispersion: 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.

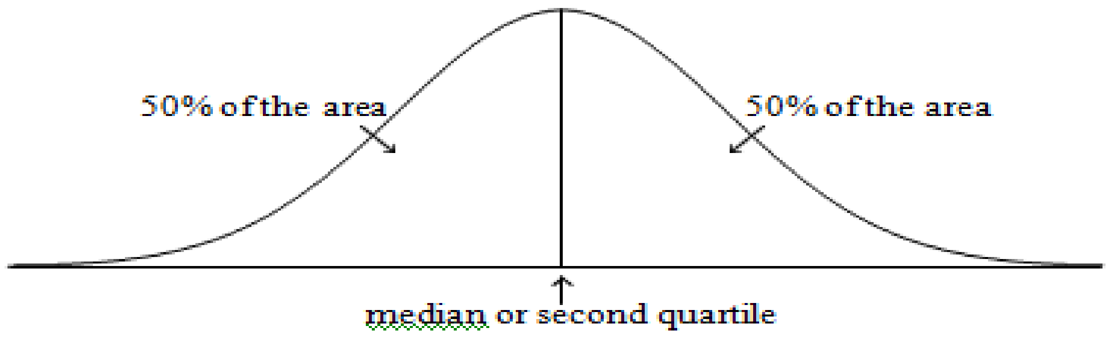
If not, then take the average of the sample values on either side of this value.

The second quartile or median

It is easy to see how to divide the area in Figure 9 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)



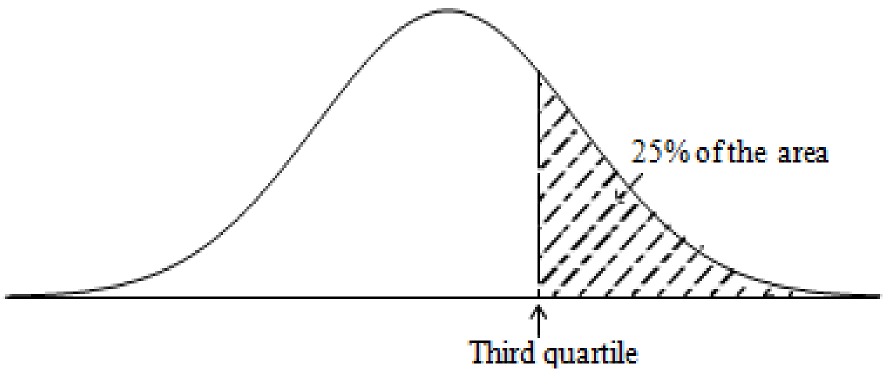
The third quartile

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.

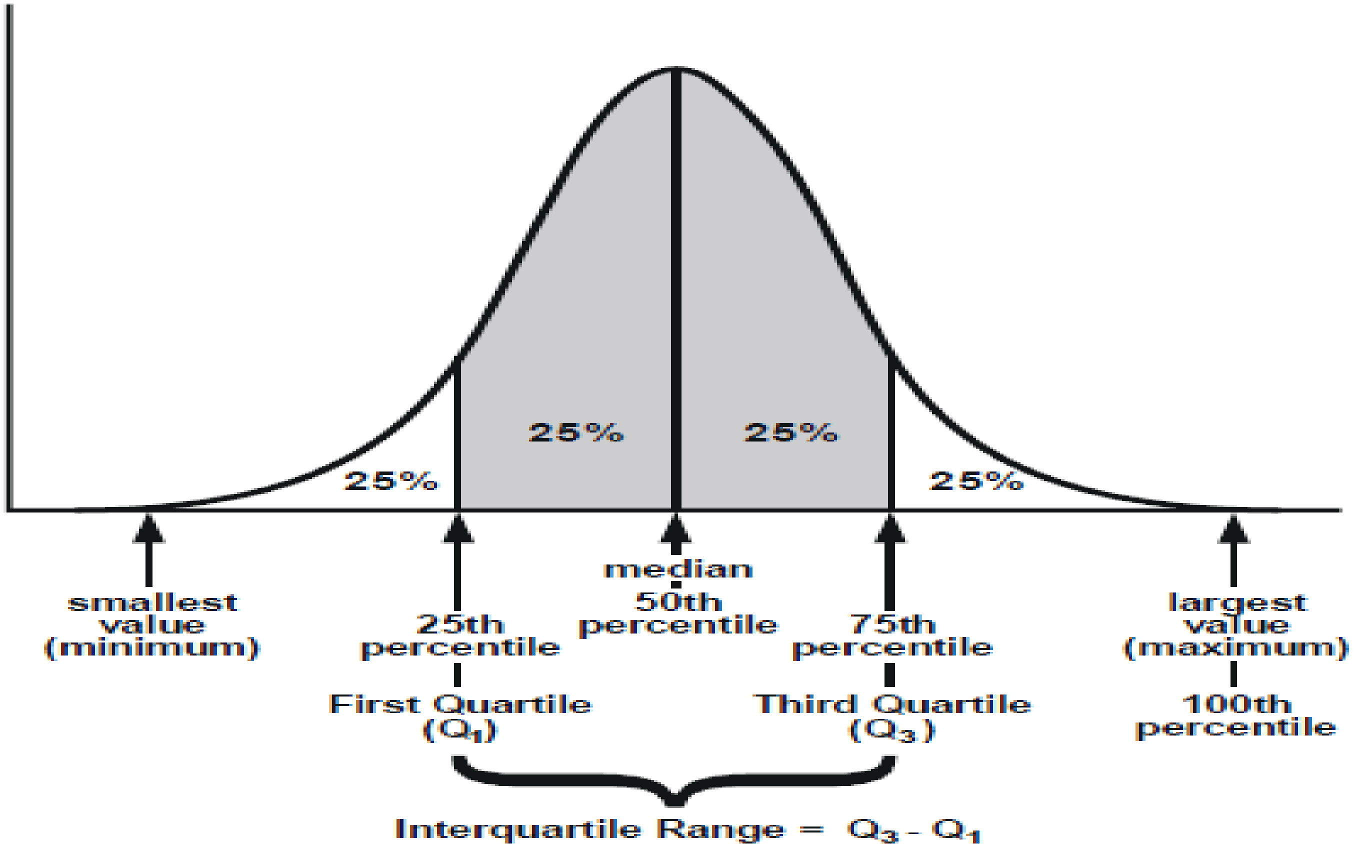
The third quartile is also called the 75th percen

The third quartile is computed in the same way the value 0.75(n+1) is used.



Measures of Dispersion: Quartiles Summary

The first (Q_1), second (Q_2) and third (Q_3) quartiles divide the distribution into four equal parts.



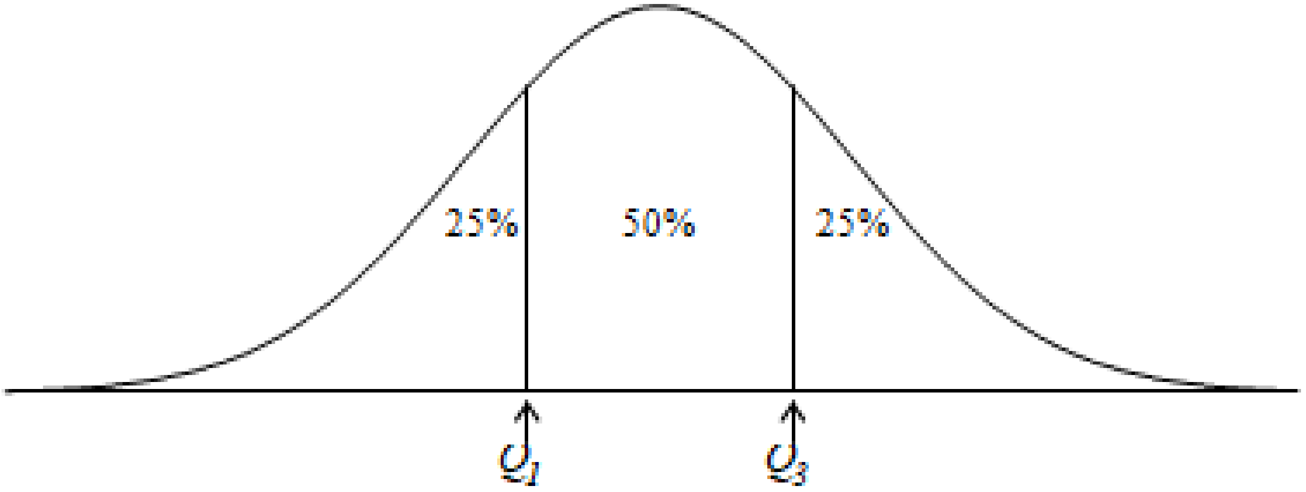
Source: https://medium.com/analytics-vidhya/descriptive-statistics-in-data-science-with-illustrations-in-python-efd5ccc152f1

Measures of Spread: InterQuartile Range

Interquartile Range = Upper Quartile(Q3) – Lower Quartile(Q1)

IQR = Q3 –Q1

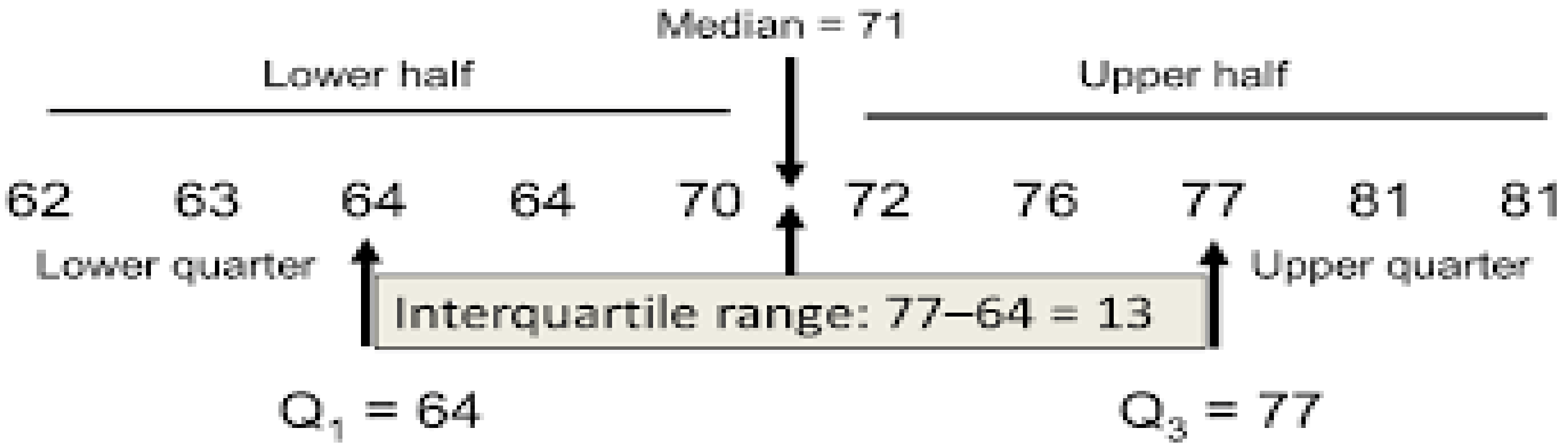
The interquartile range quantifies the difference between the third and first quartiles.



Measures of Spread: InterQuartile Range

Interquartile Range = Upper Quartile(Q3) – Lower Quartile(Q1)

IQR = Q3 –Q1



Source: sphweb.bumc.bu.edu/

Measures of Spread: InterQuartile Range

Interquartile Range = Upper Quartile(Q3) – Lower Quartile(Q1)

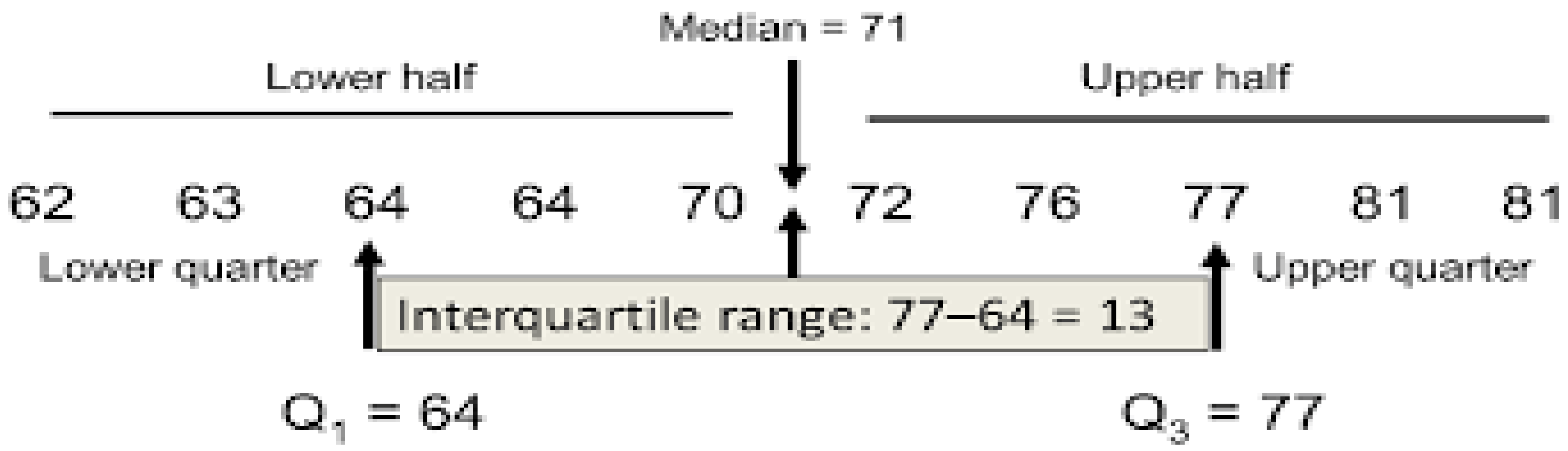
IQR = Q3 –Q1

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.

Measures of Spread: IQR-Example



Source: sphweb.bumc.bu.edu/

Measures of Spread: IQR-Example

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

Measures of spread: Variance

Average of the distance that each score is from the mean (Squared deviation from the mean).

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 step4 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

Source: standard-deviation-calculator.com

Measures of spread: Variance And Standard Deviation

Consider these two list of numbers:-
28,29,30,31,32 and
10,20,30,40,50. Find their means.

What did u found ??

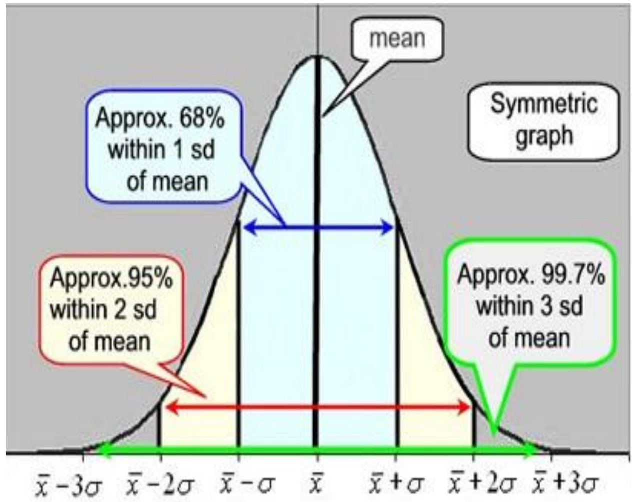
Both the list have same mean i.e. 30
But,
Clearly list differs which is not captured by mean

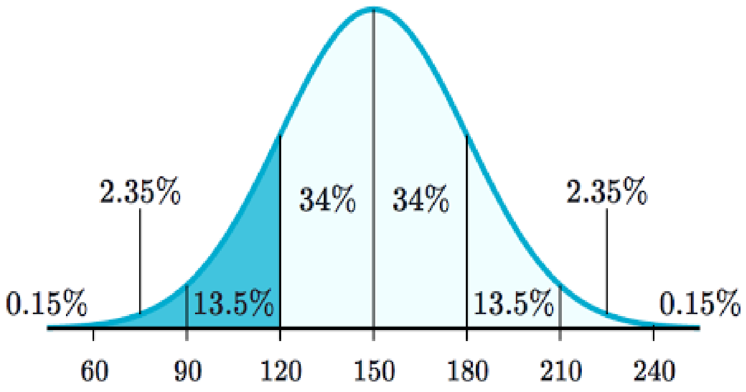
Source: standard-deviation-calculator.com

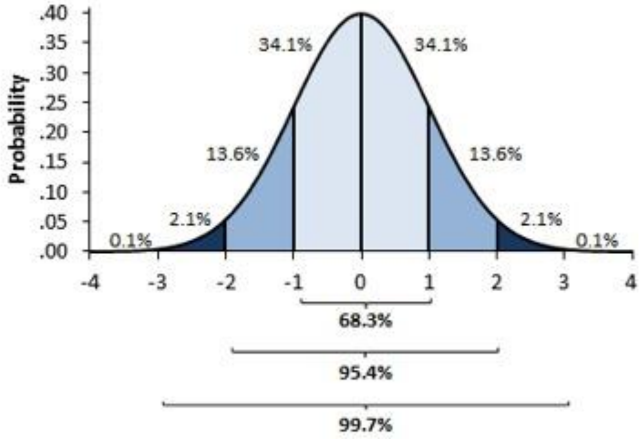
Measures of spread: Standard Deviation

Standard deviation signifies the deviation of the terms from the mean value of the distribution.

It quantifies the amount of variation of a set of data values.







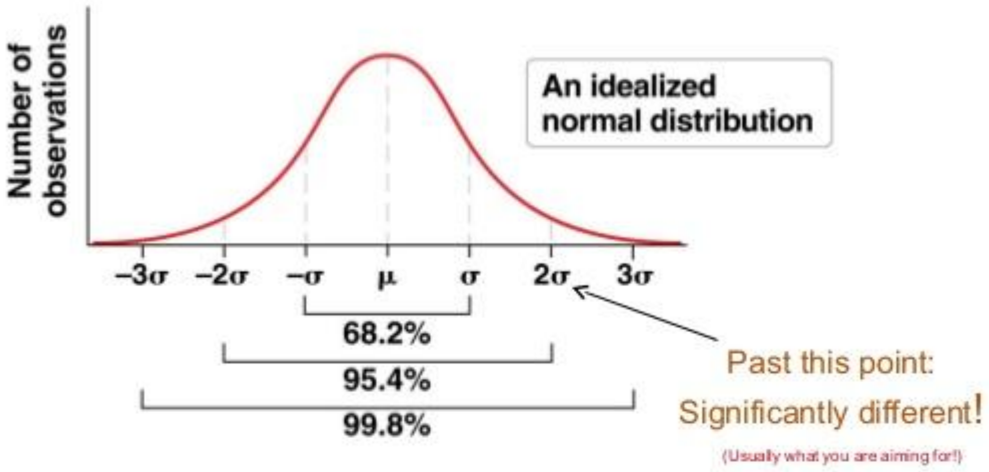
Measures of spread: Standard Deviation

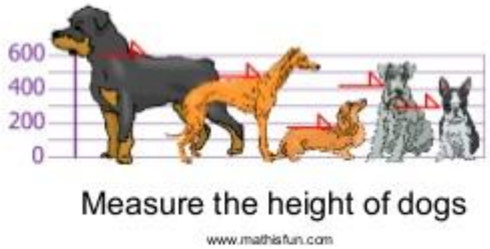
Standard deviation signifies the deviation of the terms from the mean value of the distribution.

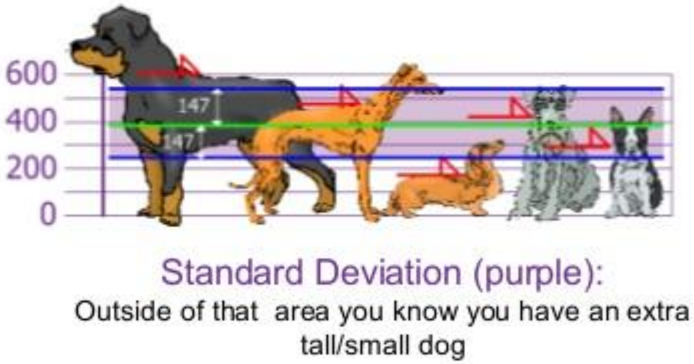
It quantifies the amount of variation of a set of data values.

Source: exceluser.com, MathBitsNotebook

Measures of spread: Standard Deviation







Source: standard-deviation-calculator.com

Measures of spread: Variance and 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	0
22	-4	16
78	Sum = 0	32

Mean = 26

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

Problem:

The heights of the players (in centimeters) from a basketball team are represented by the table:

Height	[170, 175)	[175, 180)	[180, 185)	[185, 190)	[190, 195)	[195, 2.00)
No. of players	1	3	4	8	5	2

Calculate standard deviation.

Measures of spread : Variance - Example

Population $N = 9$

$$\mu_x = 59.11 \quad \boxed{46 \;\; 64 \;\; 54 \;\; 77 \;\; 67 \;\; 68 \;\; 62 \;\; 56 \;\; 38}$$

$$\sigma^2 = \frac{\sum (x - \mu_x)^2}{N} = \frac{1146.88}{9} = 127.43$$

Random Sample $n = 4$

$$\boxed{38 \;\; 62 \;\; 67 \;\; 62} \quad \bar{x} = 57.25$$

$$s_x^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{510.75}{3} = 170.25$$

Source: Chilimath

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.

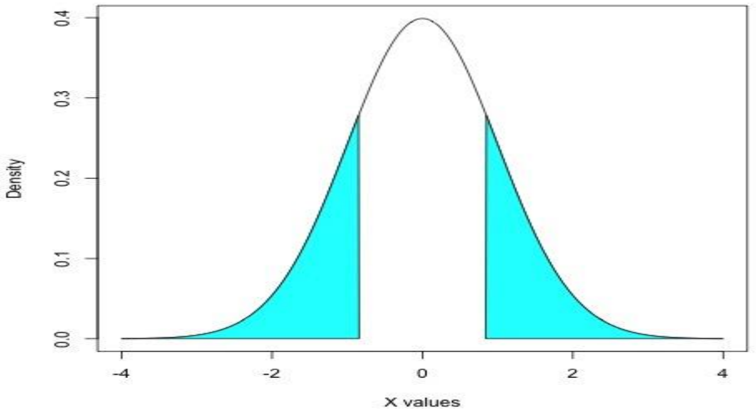
Source: exceluser.com, MathBitsNotebook

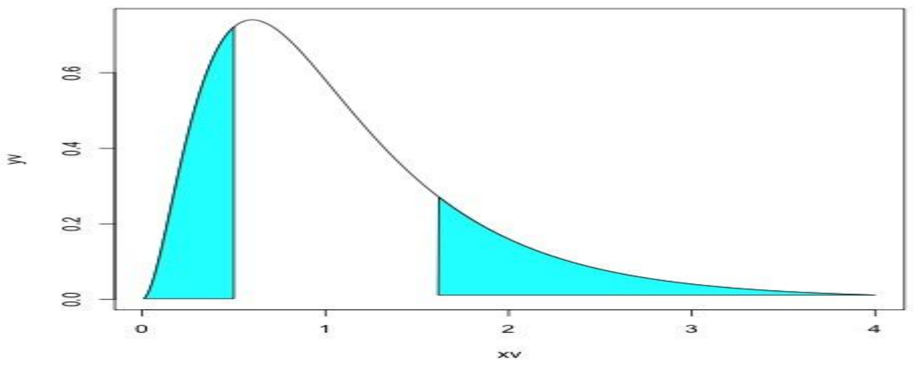
Trimmed Mean

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

Trimmed Mean

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:-

………………………

Trimmed Mean

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.

Trimmed Mean

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

Source:

The pth percentile of a sample, for a number p between 0 and 100, divides the sample such that

1. p% of the sample values are less than the pth percentile

2. And (100-p%) are greater.

Percentile

Steps:-

Order the n samples values from smallest to largest

Compute the quantity (p/100)(n+1), where n is the sample size.

If this quantity is an integer, the sample value in this position is the percentile.

Otherwise, average the two sample values pn either side.

Steps:-

Order the n samples values from smallest to largest

Compute the quantity (p/100)(n+1), where n is the sample size.

If this quantity is an integer, the sample value in this position is the percentile.

Otherwise, average the two sample values pn either side.

PES UNIVERSITY ONLINE

THANK YOU

D. Uma
Department of Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335