

29 Aug 2020

What Is Statistics?

Statistics concerns with the collection of data, organisation, interpretation, analysis and the data presentation of numerical data. In short, statistics is a crucial process, which helps to make the decision based on the data.

Characteristics of Statistics

The important characteristics of Statistics are as follows:

- Statistics are numerically expressed.
- It has an aggregate of facts
- Data are collected in systematic order
- It should be comparable to each other
- Data are collected for a Predetermined purpose
- It is effected by many causes.
- It must be enumerated or estimated accurately.

Uses of Statistics

The important functions of statistics are:

- Statistics helps in gathering information about the appropriate quantitative data
- It depicts the complex data in the graphical form, tabular form and in diagrammatic representation, to understand it easily
- It provides the exact description and better understanding
- It helps in designing the effective and proper planning of the statistical inquiry in any field
- It gives valid inferences with the reliability measures about the population parameters from the sample data
- It helps to understand the variability pattern through the quantitative observations

Basic Unit of Statistics - Data:

Data are individual pieces of factual information recorded and used for the purpose of analysis. It is the raw information from which statistics are created.

Types of Data

Qualitative Data: They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated.

Quantitative Data: These can be measured and not simply observed. They can be numerically represented and calculations can be performed on them.

Quantitative Data (numerical)	Qualitative Data (categorical)
<ul style="list-style-type: none">• Deals with numbers.• Also referred to as Numerical Data.• Data which can be measured.• Height, weight, area, volume, length, time, temperature, speed, cost, etc.• Quantitative → Quantity	<ul style="list-style-type: none">• Deals with names, labels, descriptions.• Also referred to as Categorical Data.• Data which can not measured.• Eye color, smells, car models, textures, tastes, favorites, candy bars, etc.• Qualitative → Quality

DATA

Facts or figures, which are numerical or otherwise, collected with a definite purpose are called data.

Types Of Data

Quantitative Data

These represent numerical value.

These can be numerically computed.

Qualitative Data

These represent some characteristics or attributes.

These depict descriptions that may be observed but cannot be computed.

Primary Data

Data collected for first time.

Secondary Data

Data that is sourced by someone other than the user.

Discrete Data

These are the data that can take only specific value.

Continuous Data

These are the data that can take values from a given range.

Frequency Distribution Table

A list, table or graph that displays the frequency of various outcomes in a sample of data.

Frequency Distribution Table

Ungrouped

It is used for small data set.
For eg.

Marks Obtained	Frequency
16	3
17	4
18	8
19	10
20	12
21	6
22	3

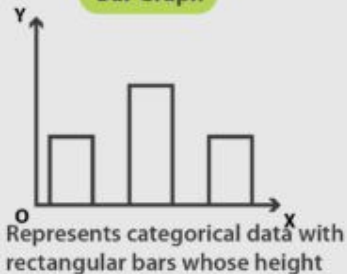
Grouped

It is used for large data set.
For eg.

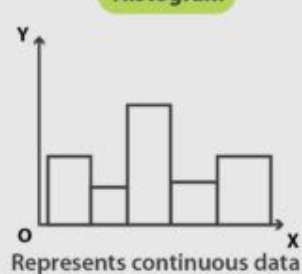
Class Interval	Frequency
0-5	3
5-10	11
10-16	14
15-20	2

Graphical Representation of Frequency Distribution Table

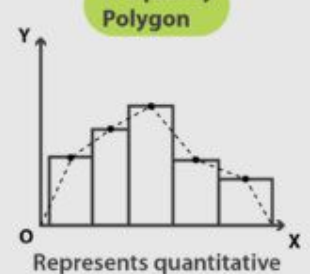
Bar Graph



Histogram



Frequency Polygon



5 Key Words of Statistics

Population

All the members of a group about which you want to draw a conclusion.

Eg: All Singapore citizens who are currently registered to vote,

Sample

The part of the population selected for analysis.

Eg : The registered voters selected to participate in a recent survey concerning their intention to vote in the next election,

Parameter

A numerical measure that describes a characteristic of a population.

EXAMPLES The percentage of all registered voters who intend to vote in the next election.

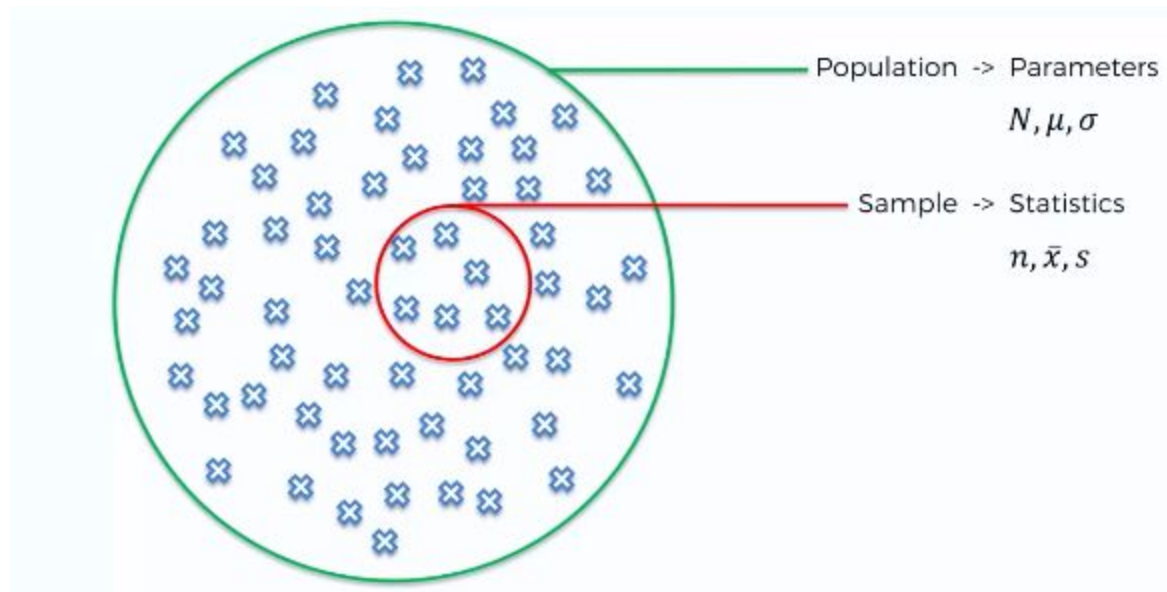
Statistic

A numerical measure that describes a characteristic of a sample.

EXAMPLES The percentage in a sample of registered voters who intend to vote in the next election.

Variable

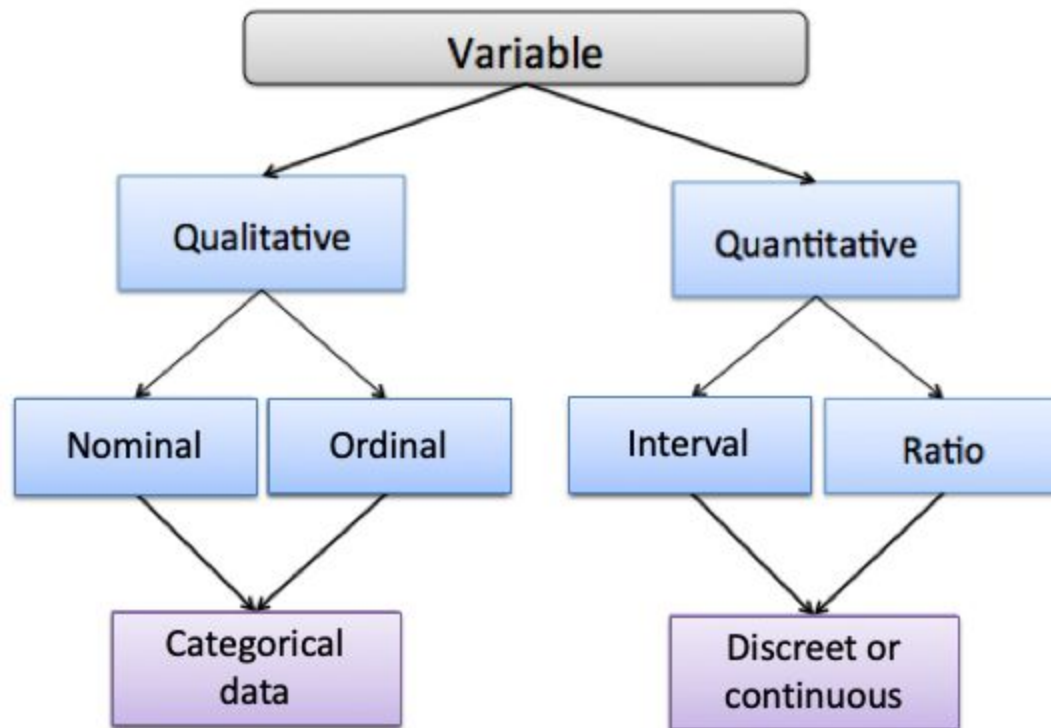
A characteristic of an item or an individual that will be analyzed using statistics.



INTERPRETATION All the variables taken together form the data of an analysis. Although you may have heard people saying that they are analyzing their data, they are, more precisely, analyzing their variables.

You should distinguish between a variable, such as gender, and its value for an individual, such as male. An observation is all the values for an individual item in the sample. For example, a survey might contain two variables, gender and age. The first observation might be male, 40. The second observation might be female, 45. The third observation might be female, 55. A variable is sometimes known as a column of data because of the convention of entering each observation as a unique row in a table of data. (Likewise, you may hear some refer to an observation as a row of data.)

Variables can be divided into the following types:



Categorical

Qualitative data are often termed **categorical data**. Data that can be added into **categories** according to their characteristics.

Nominal Variable (Unordered list)

A variable that has two or more categories, without any implied ordering.

Examples :

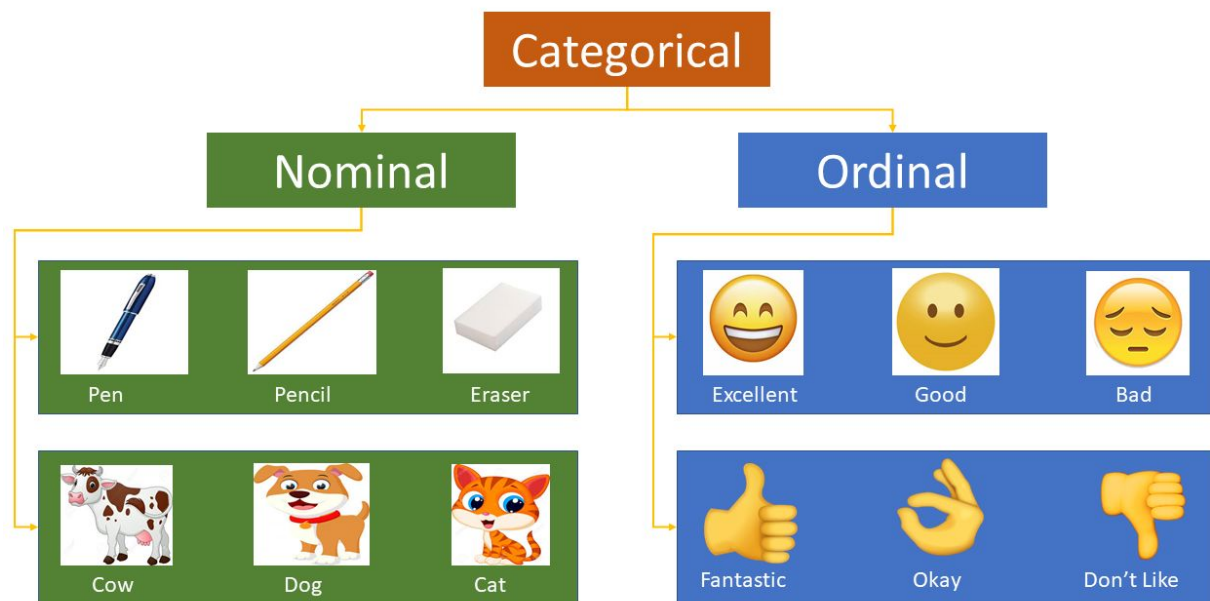
- Gender - Male, Female
- Marital Status - Unmarried, Married, Divorcee
- State - New Delhi, Haryana, Illinois, Michigan

Ordinal Variable (Ordered list)

A variable that has two or more categories, with clear ordering.

Examples :

- Scale - Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
- Rating - Very low, Low, Medium, Great, Very great



Discrete data

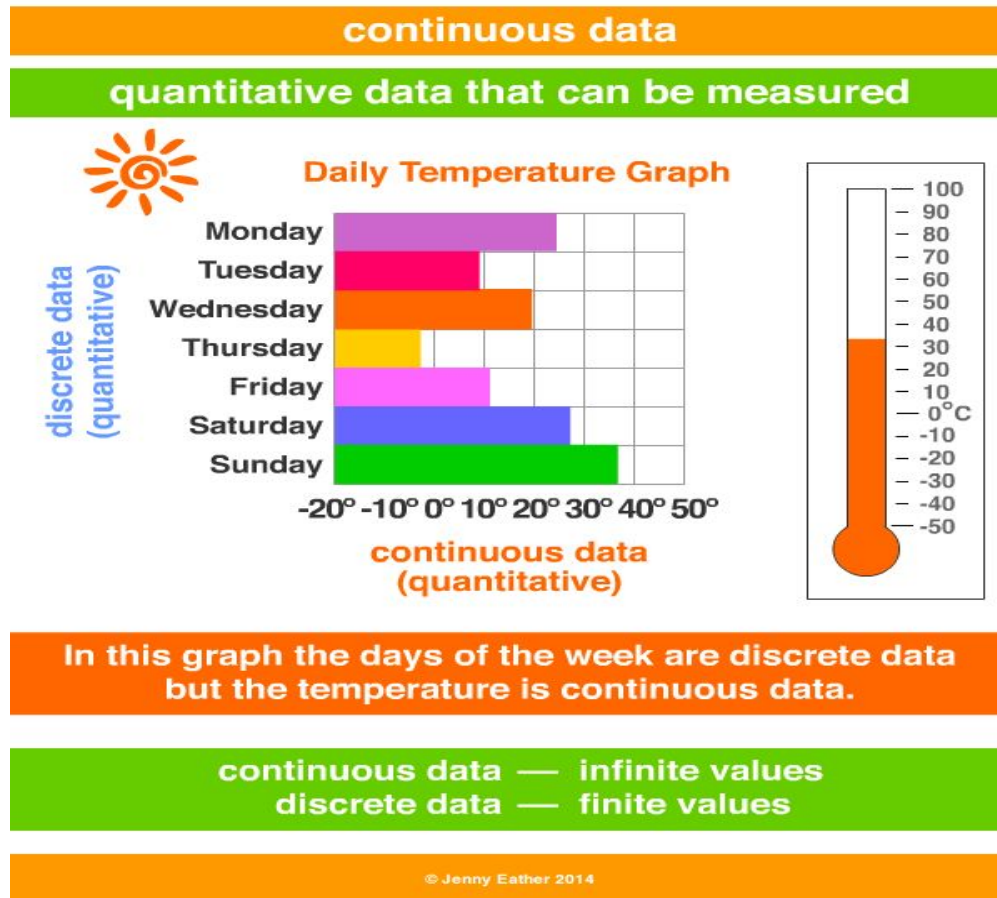
Discrete data is quantitative data that can be counted and has a finite number of possible values

e.g. days of the week.

Continuous data

Continuous data is quantitative data that can be measured. it has an infinite number of possible values within

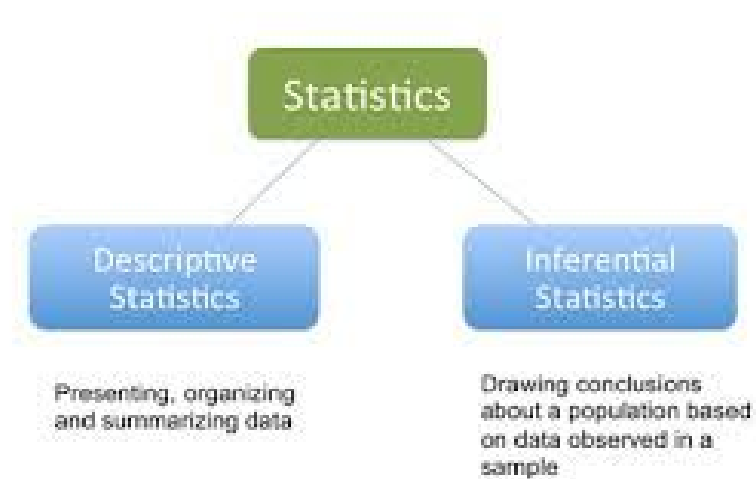
E.g a selected range e.g. temperature range.



Branches of Statistics

The two main branches of statistics are:

- Descriptive Statistics
- Inferential Statistics



DESCRIPTIVE STATISTICS

Descriptive statistics uses the collected data to provide the descriptions of the population.

and provides information on summary statistics that includes *Mean, Standard Error, Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count.*

Descriptive statistics answer the following questions:

- What is the value that best describes the data set?
- How much a data set spreads from its average value?

- What is the smallest and largest number in a data set?

Inferential Statistics – Based on the data sample taken from the population, inferential statistics makes the predictions and inferences. With inferential statistics, you take data from samples and make generalizations about a population.

BASIS FOR COMPARISON	DESCRIPTIVE STATISTICS	INFERENTIAL STATISTICS
Meaning	Descriptive Statistics is that branch of statistics which is concerned with describing the population under study.	Inferential Statistics is a type of statistics, that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
What it does?	Organize, analyze and present data in a meaningful way.	Compares, test and predicts data.
Form of final Result	Charts, Graphs and Tables	Probability
Usage	To describe a situation.	To explain the chances of occurrence of an event.

Function

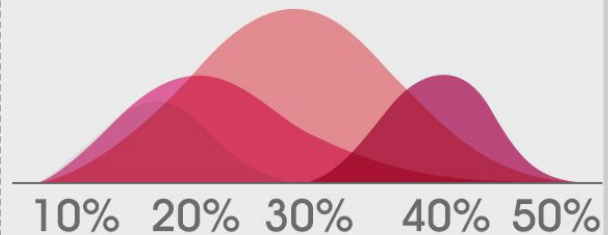
It explains the data, which is already known, to summarize sample.

It attempts to reach the conclusion to learn about the population, that extends beyond the data available.



**DESCRIPTIVE
STATISTICS**
described data.

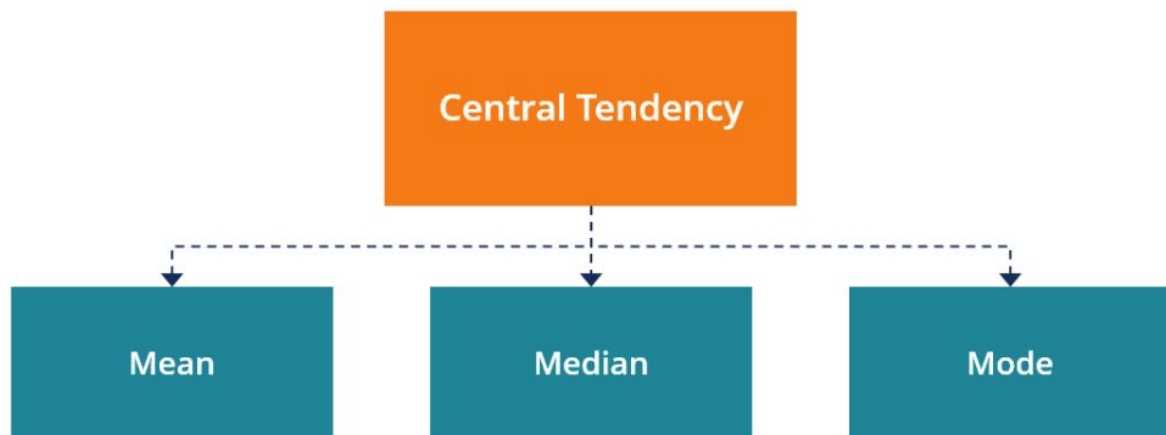
**INFERENCE
STATISTICS**
studies a sample
of the same data.



Measure of Central Tendency

It describes a whole set of data with a single value that represents the centre of its distribution.

There are three main measures of central tendency: the mode, the median and the mean.



Mean, Median and Mode

Central Tendency Measures		
Measure	Formula	Description
Mean	$\sum x/n$	Balance Point
Median	$n+1/2$ Position	Middle Value when ordered
Mode	None	Most frequent

Mean

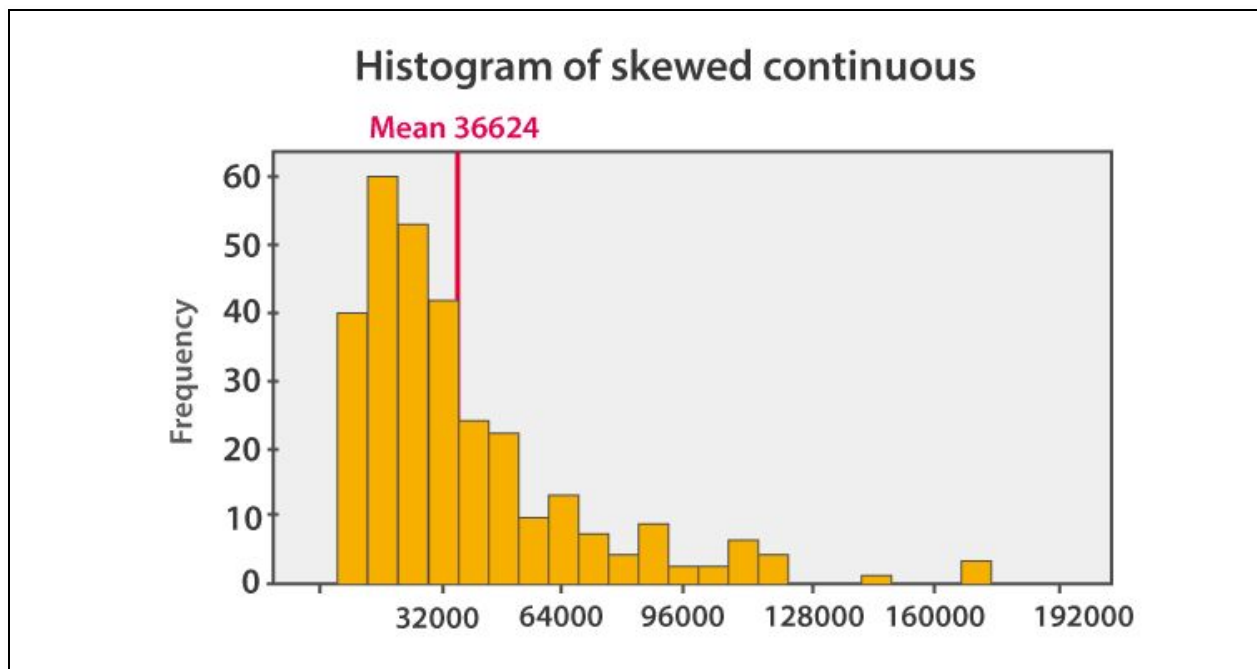
It is the sum of the observations divided by the sample size.

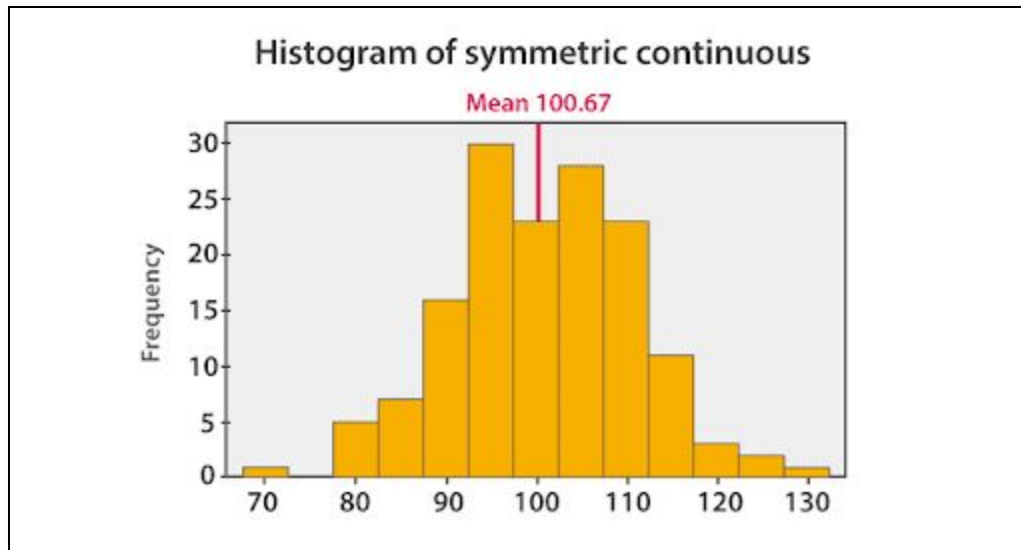
*The mean of the values 5,6,6,8,9,9,9,9,10,10 is
(5+6+6+8+9+9+9+9+10+10)/10 = 8.1*

Limitation :

It is affected by extreme values (Outliers). Very large or very small numbers can distort the answer

The histogram given below shows that the mean value of symmetric continuous data and the skewed continuous data.





In symmetric data distribution, the mean value is located accurately at the centre. But in the skewed continuous data distribution, the extreme values in the extended tail pull the mean value away from the centre. So it is recommended that the mean can be used for the symmetric distributions.

Median

Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order. When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values.


Advantage :

*It is **NOT** affected by extreme values. Very large or very small numbers does not affect it*


Mode

It is the value that occurs most frequently in a dataset

mode



The value that occurs most often in a data set.
Useful for data sets containing outliers.
If there's no mode in the data set, it's of no use.
Not as popular as mean or median.



How to determine the mode in a data set.

Order the values from least to greatest.
Locate the value that occurs the most.

3, 4, 5, 5, 6, 6, 6, 7, 8, 8, 99 mode = 6

3, 4, 5, 5, 5, 6, 6, 6, 8, 8, 99 modes = 5 and 6

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 no mode

one mode ~ unimodal, two modes ~ bimodal, more ~ multimodal

© Jenny Eather 2015

Advantage :

It can be used when the data is not numerical.

Disadvantage :

- 1. There may be no mode at all if none of the data is the same*
- 2. There may be more than one mode*

MEAN

The "mean" is the "average". To find the mean, you add up all the numbers and then divide by the number of numbers.

TO FIND THE MEAN FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13
average the set of numbers:

$$(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) \div 9 = 15$$

Note that the mean isn't a value from the original list. This is a common result.
DO NOT assume that the mean will be one of the original numbers.

MEDIAN

The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in **numerical order**, so you may have to sort the list first.

FOR AN ODD NUMBER OF VALUES: 1, 5, 2, 8, 7

Sort the numbers 1, 2, 5, 7, 8

FOR AN EVEN NUMBER OF VALUES: 1, 5, 2, 10, 8, 7

Sort the numbers: 1, 2, 5, 7, 8, 10.

TAKE THE AVERAGE OF THE TWO MEAN NUMBERS: $(5+7)/2 = 6$

MODE

The "mode" is the value that occurs most often. If no number is repeated, then there is no mode for the list.

TO FIND THE MODE FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13

Sort the numbers: 13, 13, 13, 13, 14, 14, 16, 18, 21

When to use mean, median and mode?

Mean – When your data is not skewed i.e normally distributed. In other words, there are no extreme values present in the data set (Outliers).

Median – When your data is skewed or you are dealing with ordinal (ordered categories) data (e.g. likert scale 1. Strongly dislike 2. Dislike 3. Neutral 4. Like 5. Strongly like)

Mode - When dealing with nominal (unordered categories) data.

The Use of MCT depends on the scale of measurement of the data:

Scale	Mode	Median	Mean
Nominal	√		
Ordinal	√	√	
Interval	√	√	√
Ratio	√	√	√

Example

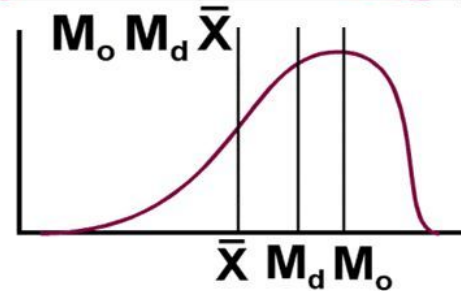
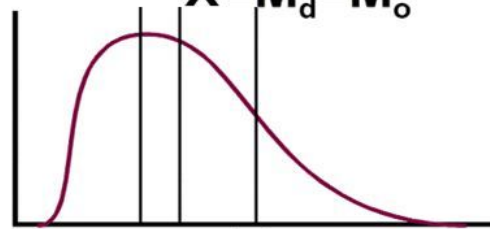
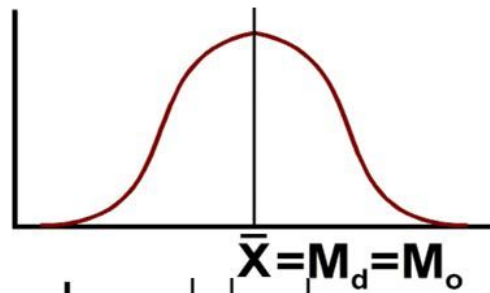
In real life, suppose a company is considering expanding into an area and is studying the size of containers that competitors are offering. They would be more interested in the mode because they want to know what size tends to sell most often.

Measures of Central Tendency

The Shape of Distributions



- With perfectly bell shaped distributions, the mean, median, and mode are identical.
- With positively skewed data, the mode is lowest, followed by the median and mean.
- With negatively skewed data, the mean is lowest, followed by the median and mode.



Measure of Dispersion

It refers to the spread or dispersion of scores. There are four main measures of variability: *Range, Inter quartile range, Standard deviation and Variance.*

Range	Difference between max and min in a distribution
Standard Deviation	Average distance of scores in a distribution from their mean
Variance	Square of the standard deviation
Skewness	Degree to which scores in a distribution are spread out.
Kurtosis	Flatness or peakness of the curve

Range

It is simply the largest observation minus the smallest observation.


Advantage:

It is easy to calculate.

Disadvantage:

It is very sensitive to outliers and does not use all the observations in a data set

range



The range is the difference between the lowest and highest value.

- Find the highest and lowest values.
- Subtract the lowest value from the highest.

2, 2, 3, 5, 5, 7, 8

LowestHighest

$8 - 2 = 6$

The range is 6

Standard Deviation

It is a measure of spread of data about the mean.

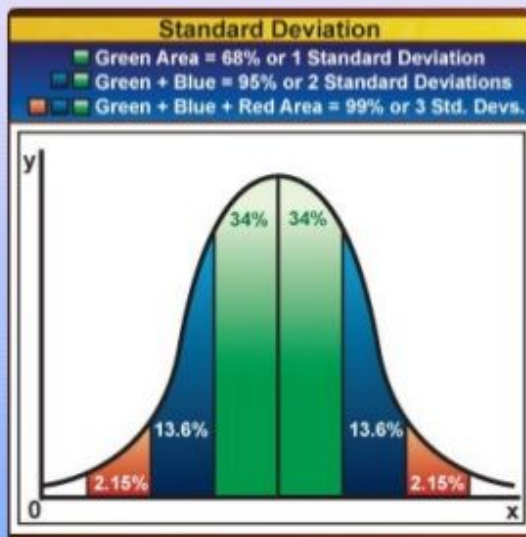
Advantage :

It gives a better picture of your data than just the mean alone.

Disadvantage :

- 1. It doesn't give a clear picture about the whole range of the data.*
- 2. It can give a skewed picture if data contain outliers.*

Standard deviation



- Standard deviation tells us the average distance of each score from the mean.
- 68% of normally distributed data is within 1 sd each side of the mean
- 95% within 2 sd
- Almost all is within 3 sd

Inter Quartile Range (IQR):

The **interquartile range** is a measure of where the “middle fifty” is in a data set. Where a **range** is a measure of where the beginning and end are in a set, **an interquartile range is a measure of where the bulk of the values lie**. That’s why it’s preferred over many other **measures of spread** when reporting things like school performance or SAT scores.

The interquartile range formula is the first **quartile** subtracted from the third **quartile**:

$$IQR = Q3 - Q1.$$

Steps:

- **Step 1: Put the numbers in order.**
1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- **Step 2: Find the median.**
1, 2, 5, 6, 7, **9**, 12, 15, 18, 19, 27.
- **Step 3: Place parentheses around the numbers above and below the median.**
Not necessary **statistically**, but it makes Q1 and Q3 easier to spot.
(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
- **Step 4: Find Q1 and Q3**
Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.
(1, 2, **5**, 6, 7), **9**, (12, 15, **18**, 19, 27). Q1 = 5 and Q3 = 18.
- **Step 5: Subtract Q1 from Q3 to find the interquartile range.**

What if I Have an Even Set of Numbers?

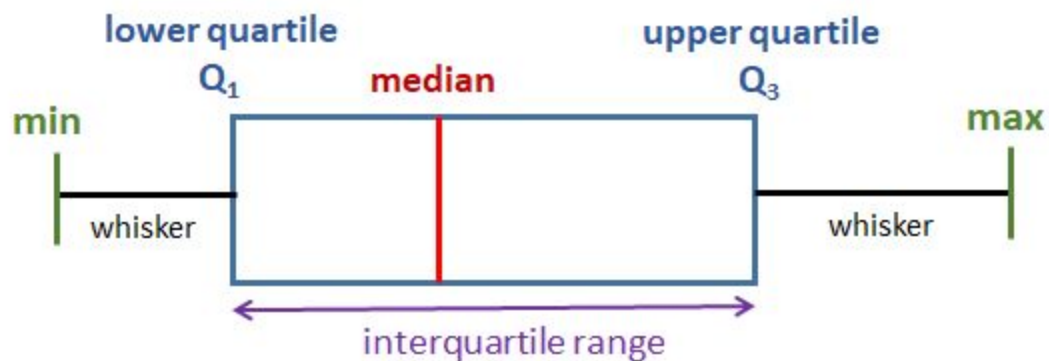
Example question: Find the IQR for the following data set: 3, 5, 7, 8, 9, 11, 15, 16, 20, 21.

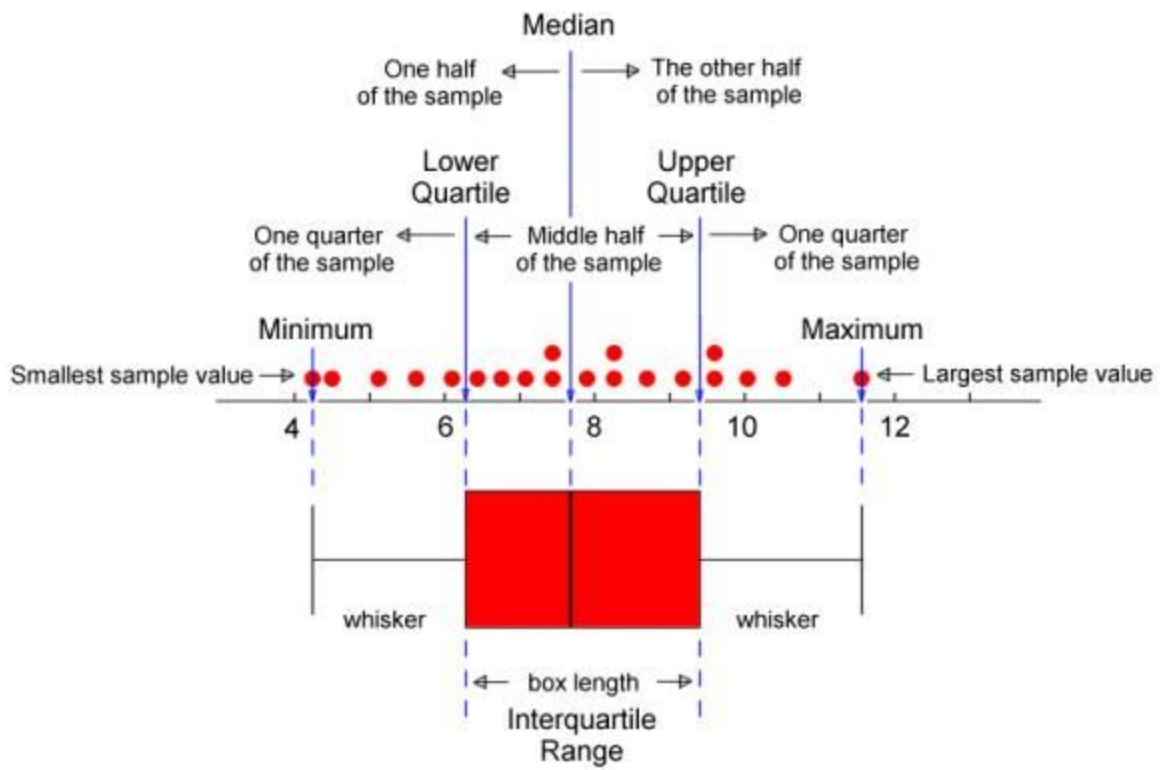
- **Step 1: Put the numbers in order.**
3, 5, 7, 8, 9, 11, 15, 16, 20, 21.
- **Step 2: Make a mark in the center of the data:**
3, 5, 7, 8, 9, | 11, 15, 16, 20, 21.

- **Step 3:** Place parentheses around the numbers above and below the mark you made in Step 2—it makes Q1 and Q3 easier to spot.
(3, 5, 7, 8, 9), | (11, 15, 16, 20, 21).
- **Step 4: Find Q1 and Q3**
Q1 is the median (the middle) of the lower half of the data, and Q3 is the median (the middle) of the upper half of the data.
(3, 5, **7**, 8, 9), | (11, 15, **16**, 20, 21). Q1 = 7 and Q3 = 16.
- **Step 5: Subtract Q1 from Q3.**
 $16 - 7 = 9$.
This is your IQR.

Box and Whisker Plot

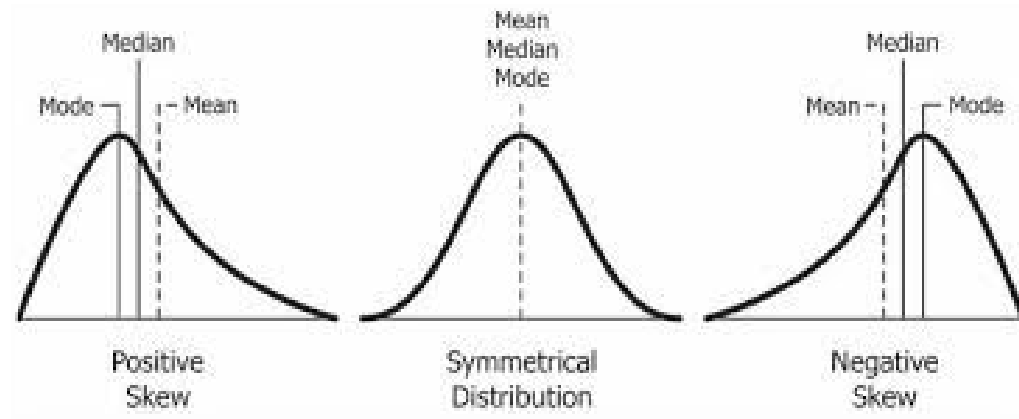
A box and whisker plot (also called a box plot) shows the five-number summary of a set of data: **minimum**, **lower quartile**, **median**, **upper quartile**, and **maximum**.





Skewness

It is a measure of symmetry. A distribution is symmetric if it looks the same to the left and right of the center point.



Kurtosis

It is a measure of whether the data are peaked or flat relative to the rest of the data. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.

Coefficient of Variation

*The coefficient of variation (relative standard deviation) is a statistical measure of the dispersion of data points around the mean. The metric is commonly used to compare the data dispersion between distinct series of data. Unlike the **standard deviation** that must always be considered in the context of the mean of the data, the coefficient of variation provides a relatively simple and quick tool to compare different data series.*

*n finance, the coefficient of variation is important in investment selection. From a financial perspective, the financial metric represents the **risk-to-reward** ratio where the volatility shows the risk of an investment and the mean indicates the reward of an investment.*

*By determining the coefficient of variation of different **securities**, an investor identifies the risk-to-reward ratio of each security and develops an investment decision. Generally, an investor seeks a security with a lower coefficient (of variation) because it provides the most optimal risk-to-reward ratio with low volatility but high returns. However, the low coefficient is not favorable when the average expected return is below zero.*

Formula for Coefficient of Variation

Mathematically, the standard formula for the coefficient of variation is expressed in the following way:

$$\text{Coefficient of Variation} = \frac{\sigma}{\mu} \times 100\%$$

Where:

- σ – the standard deviation
- μ – the mean

In the context of **finance**, we can re-write the above formula in the following way:

$$\text{Coefficient of Variation} = \frac{\text{Volatility}}{\text{Expected Return}} \times 100\%$$

Example of Coefficient of Variation

Fred wants to find a new investment for his portfolio. He is looking for a safe investment that provides stable returns. He considers the following options for investment:

- **Stocks:** Fred was offered stock of ABC Corp. It is a mature company with strong operational and financial performance. The volatility of the stock is 10% and the expected return is 14%.
- **ETFs:** Another option is an **Exchange-Traded Fund (ETF)** which tracks the performance of the S&P 500 index. The ETF offers an expected return of 13% with a volatility of 7%.
- **Bonds:** Bonds with excellent **credit ratings** offer an expected return of 3% with 2% volatility.

In order to select the most suitable investment opportunity, Fred decided to calculate the coefficient of variation of each option. Using the formula above, he obtained the following results:

$$\text{Coefficient of Variation (Stock)} = \frac{10\%}{14\%} \times 100\% = 71.4\%$$

$$\text{Coefficient of Variation (ETF)} = \frac{7\%}{13\%} \times 100\% = 53.8\%$$

$$\text{Coefficient of Variation (Bond)} = \frac{2\%}{3\%} \times 100\% = 66.7\%$$

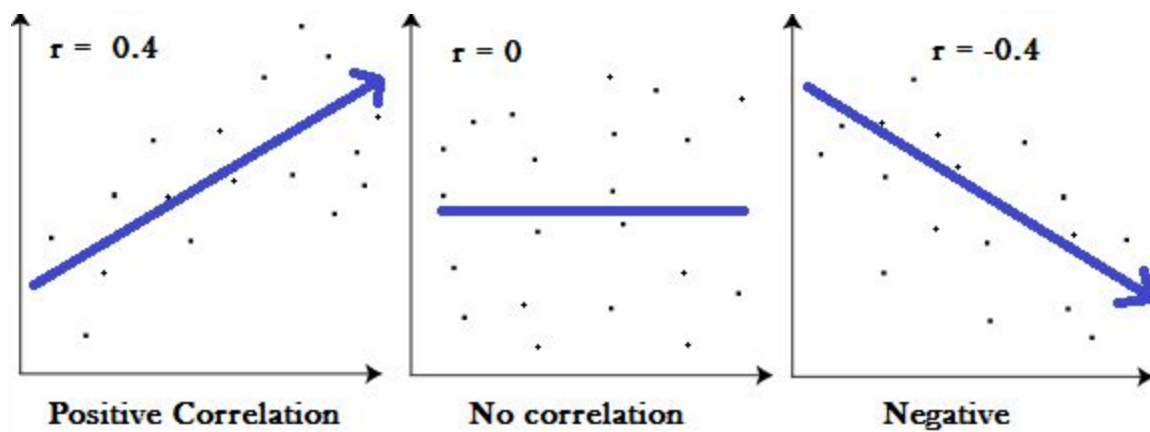
Based on the calculations above, Fred wants to invest in the ETF because it offers the lowest coefficient (of variation) with the most optimal risk-to-reward ratio.

Examine the differences between univariate and bivariate data.

Univariate Data	Bivariate Data
<ul style="list-style-type: none">• involving a single variable	<ul style="list-style-type: none">• involving two variables
<ul style="list-style-type: none">• does not deal with causes or relationships	<ul style="list-style-type: none">• deals with causes or relationships
<ul style="list-style-type: none">• the major purpose of univariate analysis is to describe	<ul style="list-style-type: none">• the major purpose of bivariate analysis is to explain
<ul style="list-style-type: none">• central tendency - mean, mode, median• dispersion - range, variance, max, min, quartiles, standard deviation.• frequency distributions• bar graph, histogram, pie chart, line graph, box-and-whisker plot	<ul style="list-style-type: none">• analysis of two variables simultaneously• correlations• comparisons, relationships, causes, explanations• tables where one variable is contingent on the values of the other variable.• independent and dependent variables
Sample question: How many of the students in the freshman class are female?	Sample question: Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

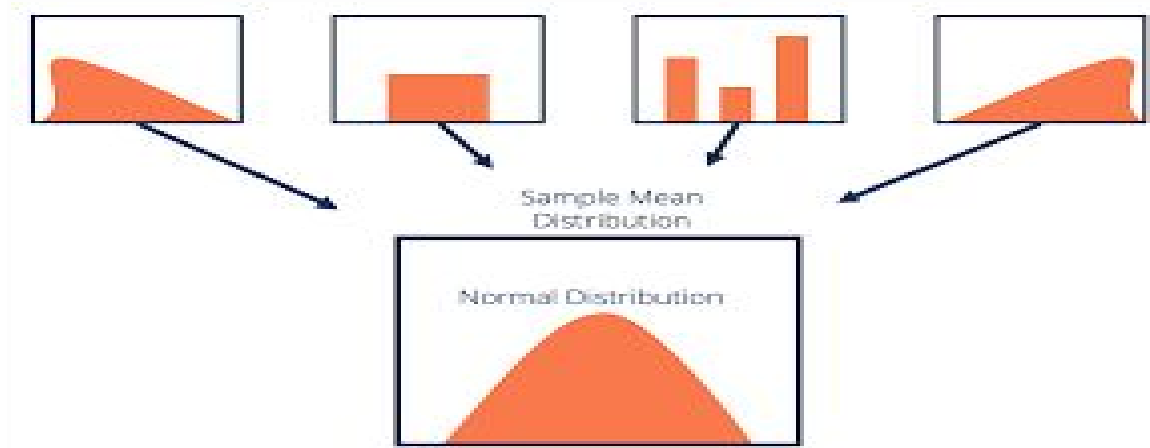
Correlation Coefficient

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect [negative correlation](#), while a correlation of 1.0 shows a perfect [positive correlation](#). A correlation of 0.0 shows no linear relationship between the movement of the two variables.



Central Limit Theorem

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population **with replacement**, then the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually $n \geq 30$). If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that $\min(np, n(1-p)) \geq 5$, where n is the sample size and p is the probability of success in the population. This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.



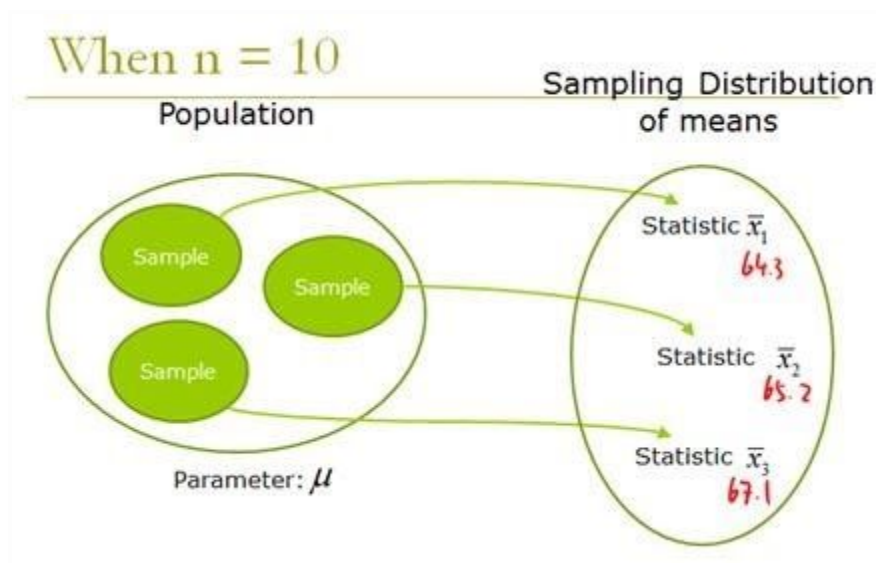
An essential component of the Central Limit Theorem is that the **average** of your sample means will be the population mean. In other words, add up the means from all of your samples, find the average and that average will be your actual population mean. Similarly, if you find the average of all of the **standard deviations** in your **sample**, you'll find the actual standard deviation for your population. It's a pretty useful phenomenon that can help accurately predict characteristics of a **population**.

Significance of the Central Limit Theorem

The central limit theorem has both statistical significance as well as practical applications. Isn't that the sweet spot we aim for when we're learning a new concept?

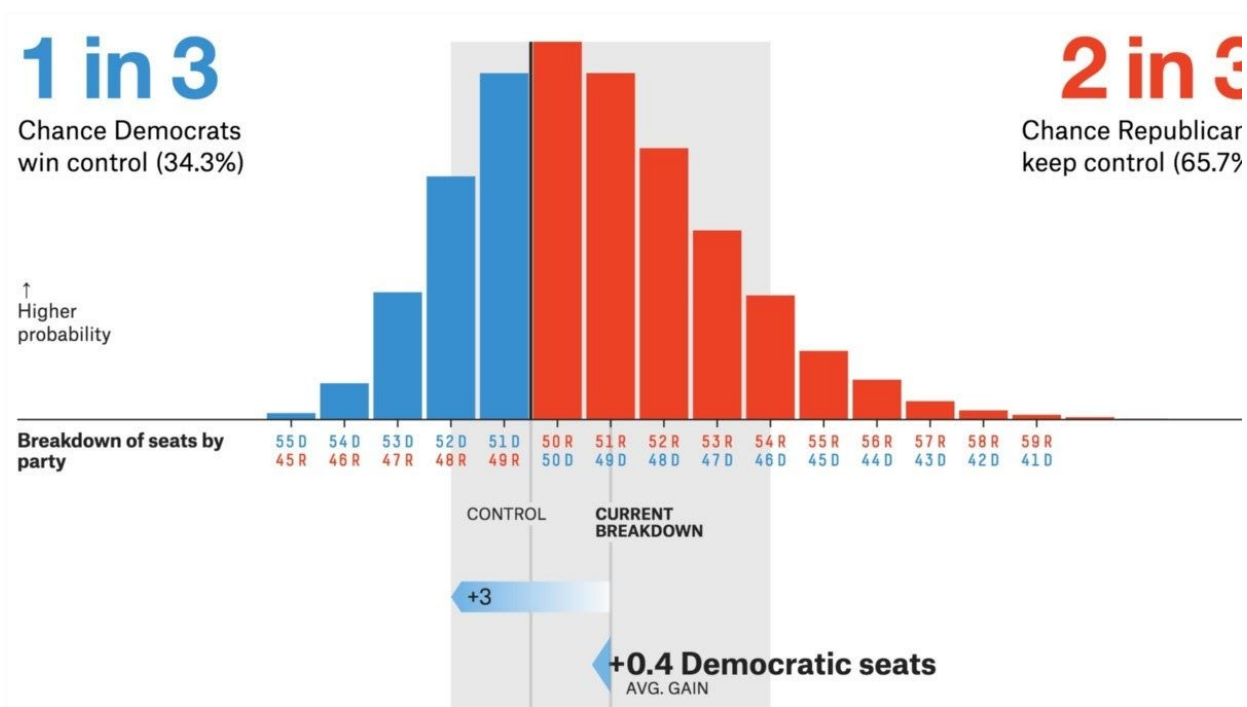
We'll look at both aspects to gauge where we can use them.

Statistical Significance of CLT



- *Analyzing data involves statistical methods like hypothesis testing and constructing confidence intervals. These methods assume that the population is normally distributed. In the case of unknown or non-normal distributions, we treat the sampling distribution as normal according to the central limit theorem*
- *If we increase the samples drawn from the population, the standard deviation of sample means will decrease. This helps us estimate the population mean much more accurately*
- *Also, the sample mean can be used to create the range of values known as a confidence interval (that is likely to consist of the population mean)*

Practical Applications of CLT



Source: projects.fivethirtyeight.com

- *Political/election polls are prime CLT applications. These polls estimate the percentage of people who support a particular candidate. You might have seen*

these results on news channels that come with confidence intervals. The central limit theorem helps calculate that

- *Confidence interval, an application of CLT, is used to calculate the mean family income for a particular region*

The central limit theorem has many applications in different fields. Can you think of more examples? Let me know in the comments section below the article – I will include them here.

Assumptions Behind the Central Limit Theorem

Before we dive into the implementation of the central limit theorem, it's important to understand the assumptions behind this technique:

1. *The data must follow the randomization condition. It must be sampled randomly*
2. *Samples should be independent of each other. One sample should not influence the other samples*
3. *Sample size should be not more than 10% of the population when sampling is done without replacement*
4. *The sample size should be sufficiently large. Now, how we will figure out how large this size should be? Well, it depends on the population. When the population is skewed or asymmetric, the sample size should be large. If the population is symmetric, then we can draw small samples as well*

In general, a sample size of 30 is considered sufficient when the population is symmetric.

The mean of the sample means is denoted as:

$$\mu \bar{X} = \mu$$

where,

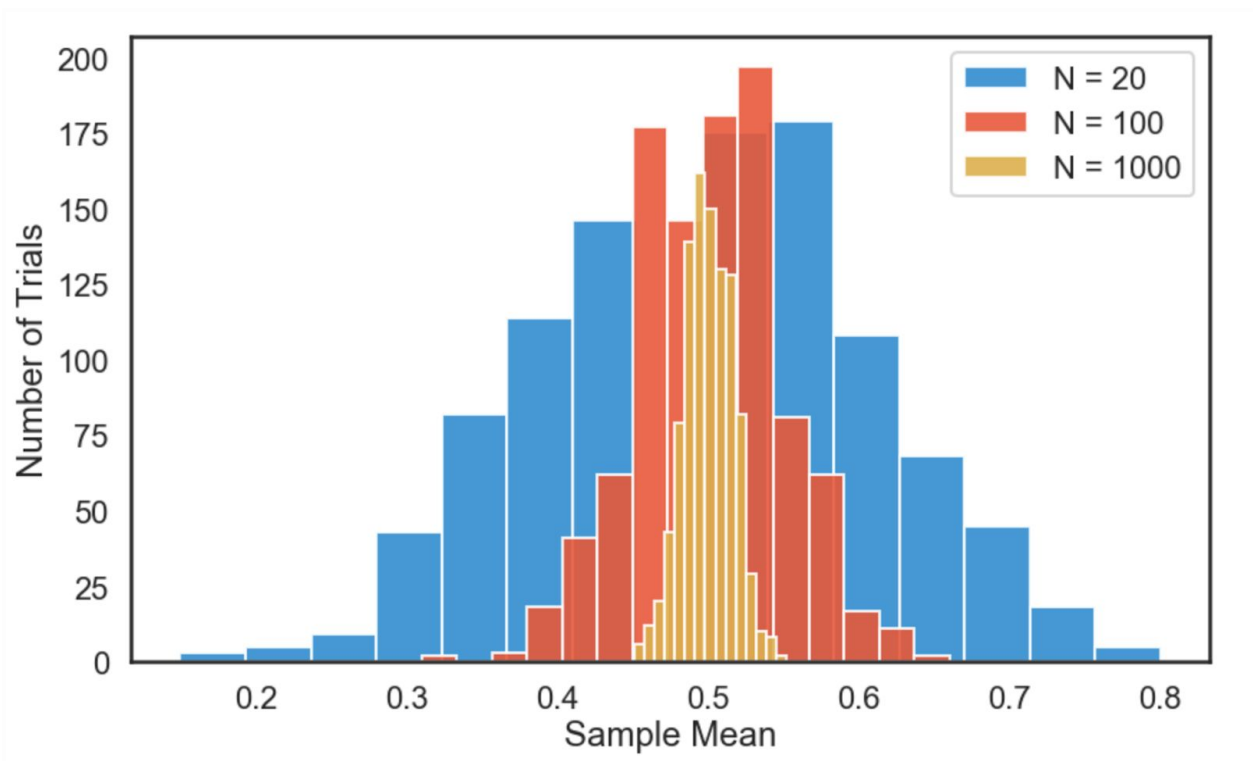
- $\mu \bar{X}$ = Mean of the sample means
- μ = Population mean

And, the standard deviation of the sample mean is denoted as:

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

where,

- $\sigma_{\bar{X}}$ = Standard deviation of the sample mean
- σ = Population standard deviation
- n = sample size



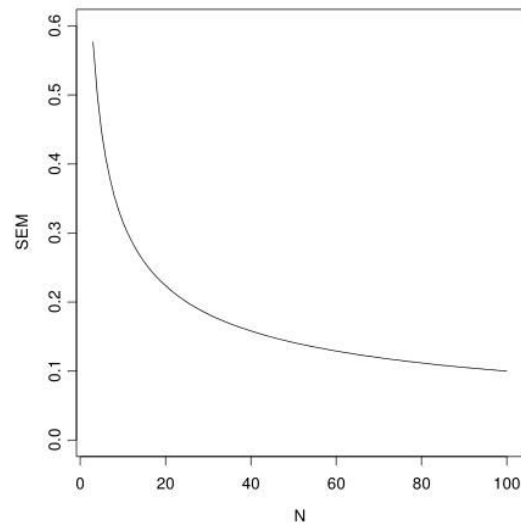
Standar Error

The standard error(SE) is very similar to **standard deviation**. Both are **measures of spread**. The higher the number, the more spread out your data is. To put it simply, the two terms are essentially equal—but there is one important difference. While the standard error uses statistics (sample data) standard deviations use parameters (population data).

Standard Error of the Mean

$$SE_{M_x} = \frac{\sigma}{\sqrt{N}}$$

- This equation implies that sampling error decreases as sample size increases.
- This is important because it suggests that if we want to make sampling error as small as possible, we need to use as large of a sample size as we can manage.



Z score

Z score standardization is one of the most popular method to normalize data.

The idea is to rescale an original variable to have equal range and/or variance. In this case, we rescale an original variable to have a **mean of zero** and **standard deviation of one**.

$$Z = \frac{x - \text{mean}}{\text{std.dev}}$$

Z score

Mathematically, scaled variable would be calculated by subtracting mean of the original variable from raw value and then divide it by standard deviation of the original variable.