**Statistics**

***Statistics is kind of methods specially adapted to the collection, Classification, analysis and interpretation of data for making effective decisions in all functional areas of Management***.

Statistics is also a processing of data into meaningful information like grouping, filtering, sorting etc.

Statistician were used to project the number based on the data’s available. But those data’s can be sometimes misleading, So before projecting, data should be checked for correctness.

Data’s used for projection or prediction should be more presentable to get perfect information.

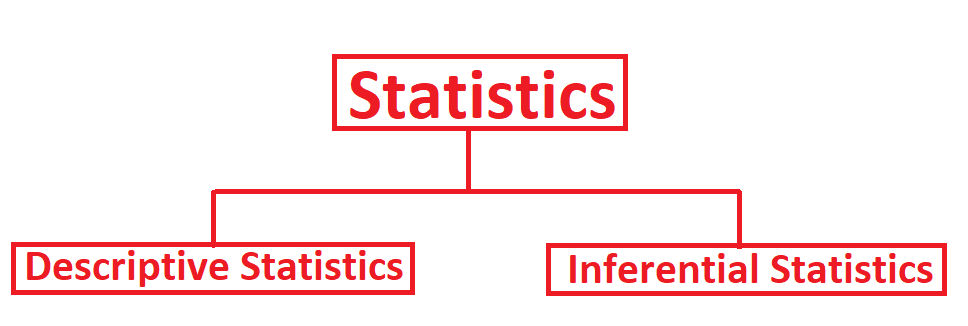
**Projection or Forecast**

*A forecast refers to a calculation or an estimation which uses data from previous events, combined with recent trends to come up a future event outcome*

**Prediction**

*A prediction is an actual act of indicating that something will happen in the future with or without prior information. A prediction is often, but not always, based upon experience or knowledge.*

**Branches of Statistics**



Descriptive Statistics is concerned with Data Summarization, Graphs/Charts and Tables. It processes raw data into information. Information is key to decision making. (Data what is in hand)

Inferential Statistics is a method used to talk about a population parameter from a Sample. It involves Point Estimation, Interval Estimation and Hypothesis Testing. (With the help of Descriptive data., like kind of projection)

**Population**

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

Ex: All students taking the statistics course in a business school is an example of population

**Sample**

*Sample is a subset of population.*

Ex: Selecting a team of 20 students from 100 students of a class (100 students is population)

**Variable**

A variable is an item of interest that can take on many different numerical values. Variable can be of 1) Dependent

2) Independent

Ex: Number of defective items produced in a factory.

If variable takes different values with associated probability it is called random variable.

Ex: Number of times head turns up in a toss of coin ten times is random variable

**Parameter**

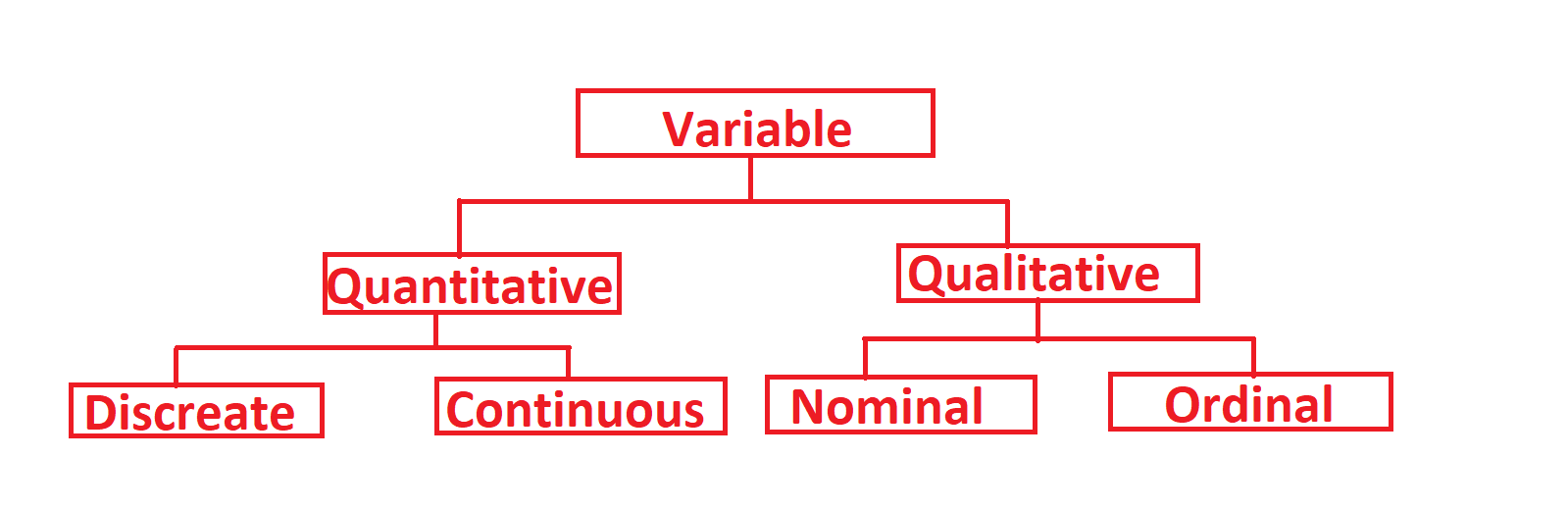
A population characteristic of interest is called a parameter.

Ex: Average Income of particular class of people.

**Statistics**

Statistics is a type of average that is based on a sample. An inference about population parameter.

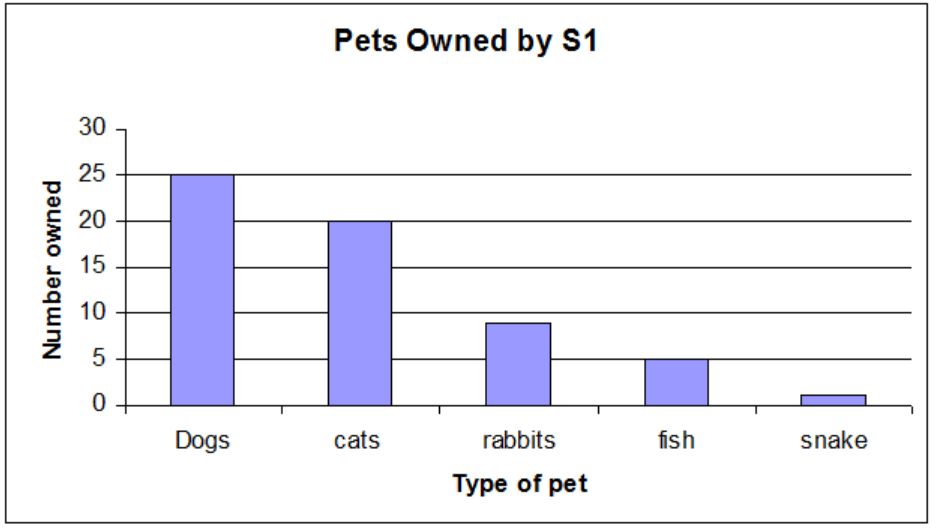
Ex: Average Income of particular Class of people based on sample. This sample average is called statistic.



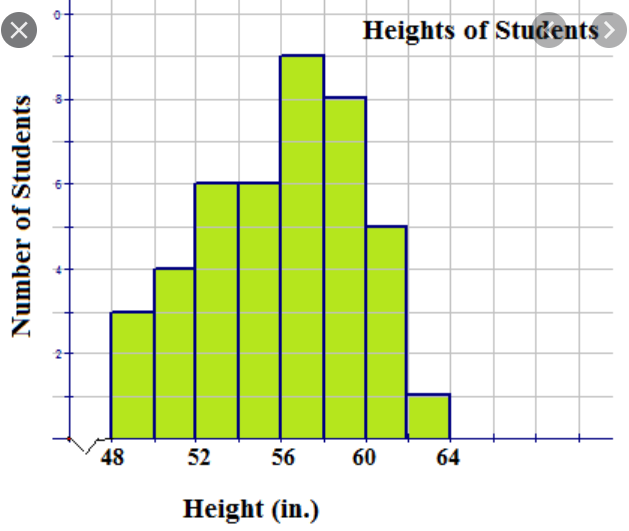
Quantitative can be of Discreate or Continuous, Ex: 1,2,3,4,5 etc.

Qualitative is Discreate, Ex: Hair colour (Black, Brown, Grey etc)

**Quantitative Discreate**



**Quantitative Continuous**



**Qualitative Nominal (Unique)**

Ex: Emp ID, Gender, Religion, Hair colour etc

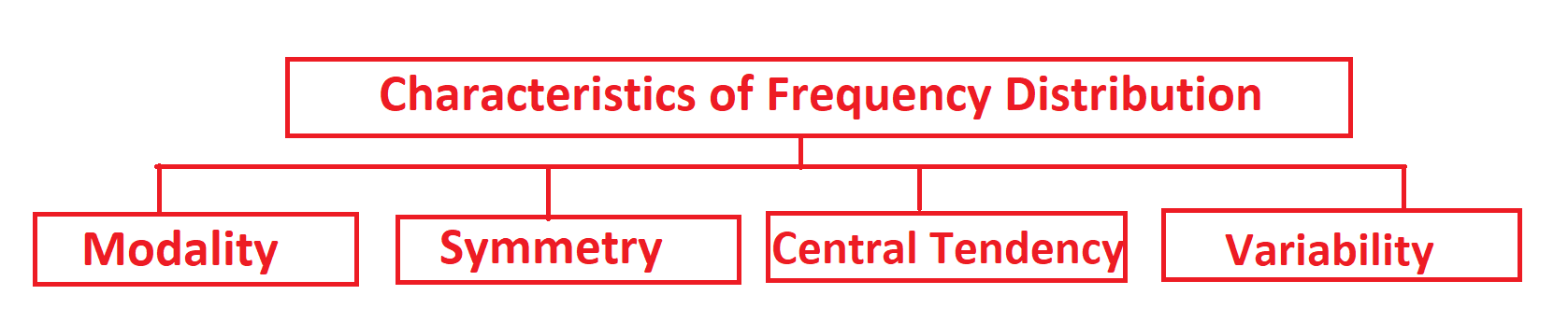
**Qualitative Ordinal**

Putting in orders or groups Ex: (Good, Bad etc)

**Frequency Distribution**

In statistics, a frequency distribution is a list, table or graph that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval.

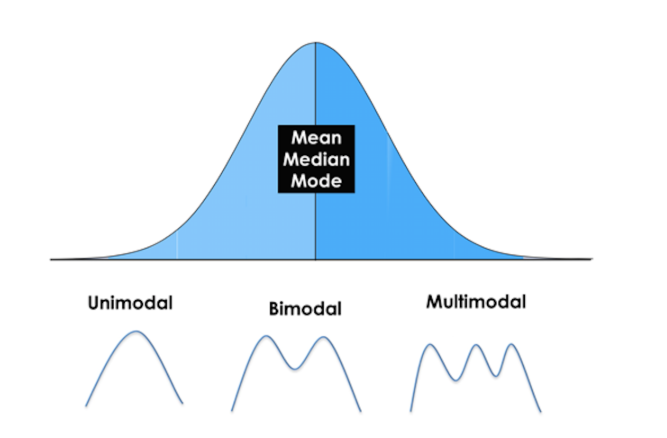
**Characteristics of Frequency Distribution**



**Modality**

The modality of a distribution is determined by the number of peaks it contains. The mode is the number that appears most frequently in a set.

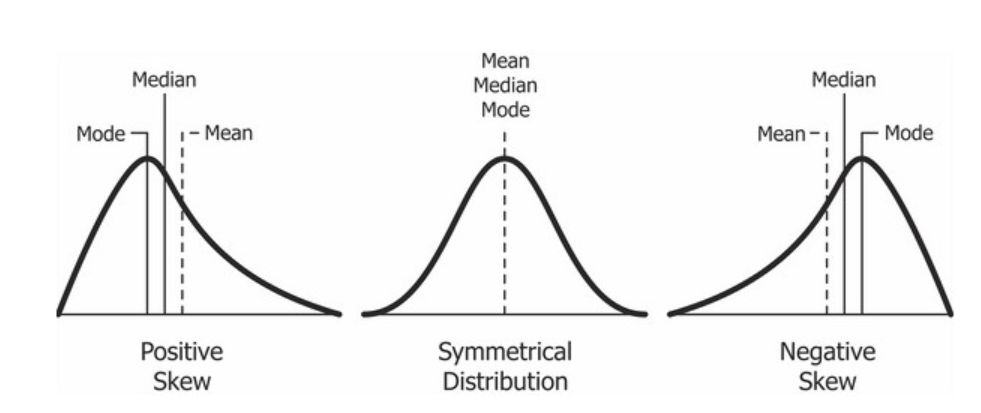
**Single Head, Bi Modal, Multi Modal Modality**



**Symmetry**

**Symmetry, Asymmetry (Skewed), Positive Skewed, Negative Skewed**

Skewness is a measurement of the symmetry of a distribution. Thus, it describes how much a distribution differs from a normal distribution, either to the left or to the right. The skewness value can be either positive, negative or zero. Note that a perfect normal distribution would have a skewness of zero because the mean equals the median.



Positives skew occurs if the data is piled up to the left, which leaves the tail pointing to the right.

Negative skew occurs if the data is piled up to the right, which leaves the tail pointing to the left. Note that positive skews are more frequent than negative ones.

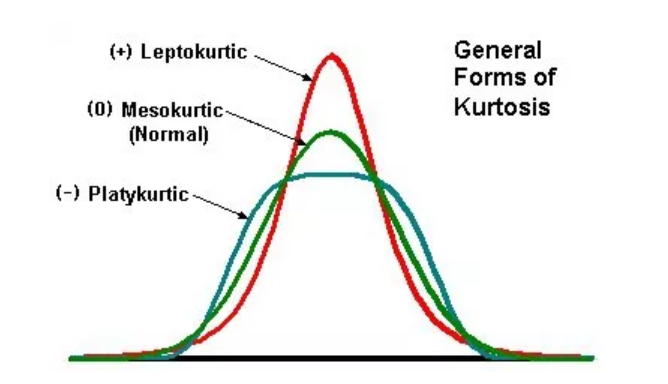
**Kurtosis**

The kurtosis parameter is a measure of the combined weight of the tails relative to the rest of the distribution*.*

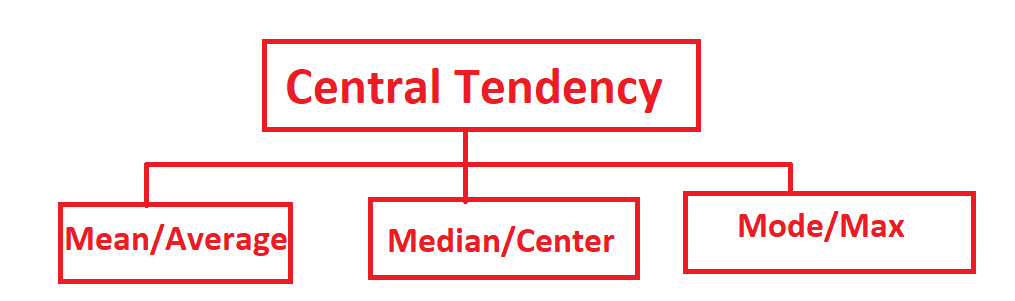
So, kurtosis is all about the tails of the distribution – not the peak or flatness.

A normal random variable has a kurtosis of 3 irrespective of its mean or standard deviation. If a random variable’s kurtosis is greater than 3, it is said to be **Leptokurtic**. If its kurtosis is less than 3, it is said to be**Platykurtic**.

A large value of kurtosis indicates a more serious outlier issue and hence may lead the researcher to choose alternative statistical methods.



**Central Tendency**

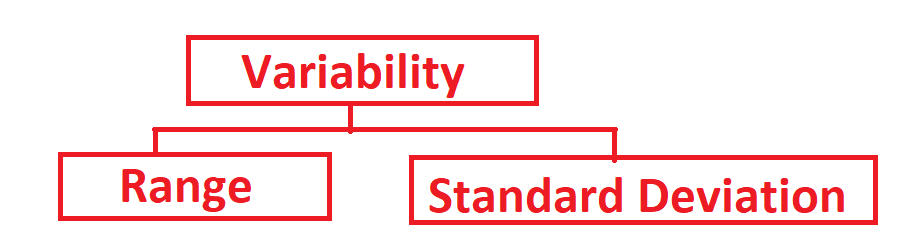
****

***Mean****is the average of a data set.*

***Mode****is the most common number in a data set.*

***Median****is the middle of the set of numbers*

**Variability**

****

***Range*** *is the difference between the largest and smallest values in a set of data.*

***Standard Deviation*** *is a measure of how far the data points are spread out. How far the value is deviated from the mean.*

For Ex: Need to find the mean median and mode of following values

10, 12, 24, 3, 2, 8, 23, 22, and 9

Rearranging we will get (sorting from ascending to descending)

2, 3, 8, 9, 10, 12, 22, 23, 24

Mean = Summation of all values/Total number of values

= ∑x/n

= (10+12+24+3+2+8+23+22+9)/9

=113/9

=12.56

**Median of Odd numbers**

Median (since n is odd number) = (n+1)/2

= (9+1)/2

= 10/2

= 5

So 5th positioned number after rearranging will be median

Here median is 10 (which is in the 5th position)

Mode is not available as there are no repetitive values (Frequently occurring)

**Median of Even numbers**

If n is even number the median will be

Average of n/2th item and (n+1)/2th item

Find the Median of 25, 28, 20, 8, 10, 15

Rearranging 8, 10, 15, 20, 25, 28

Mean = (8+10+15+20+25+28)/6

= 106/6 = 17.67

Median (Since n is even number) = ((n/2)th term + ((n+2)/2))th term/2

= ((6/2) + ((6+2)/2))/2

= (3th term + 4th term)/2

= (15+20)/2

= 35/2

= 17.5

Range

From above example Range = Max – Min

Range = 28 – 8

= 20

Standard Deviation

From above example

Standard Deviation = √Variance

Variance = (Summation (x-mean) ^2)/n

= ((8-17.67)^2 + (10-17.67)^2 + (15-17.67)^2 + (20-17.67)^2 + (25-17.67)^2 + (28-17.67)^2)/6

= ((-9.67)^2 + (-7.67)^2 + (-2.67)^2 + (2.33)^2 + (7.33)^2 + (10.33)^2)/6

= (93.51 + 58.83 + 7.13 + 5.43 + 53.73 + 106.71)/6

= (325.34)/6

= 54.22

SD = √54.22

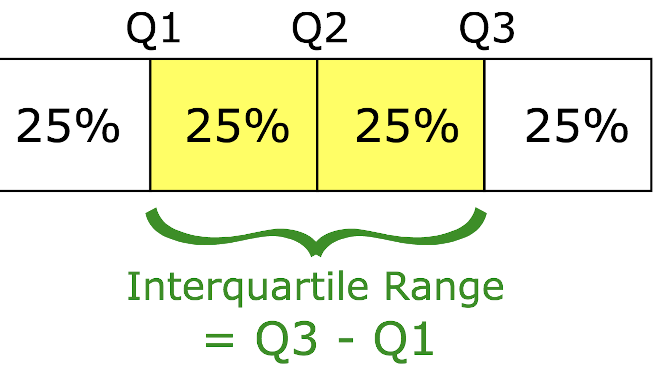
= 7.36

**Exploratory Data Analysis**

**Percentile and Quartile**

**Measures of Variability**

The most popular variability measures are the range, interquartile range (IQR), variance, and standard deviation. These are used to measure the amount of spread or variability within your data. The range describes the difference between the largest and the smallest points in your data. The interquartile range (IQR) is a measure of statistical dispersion between upper (75th) and lower (25th) quartiles.



While the range measures where the beginning and end of your datapoint are, the interquartile range is a measure of where the majority of the values lie.

IQR = Q3-Q1

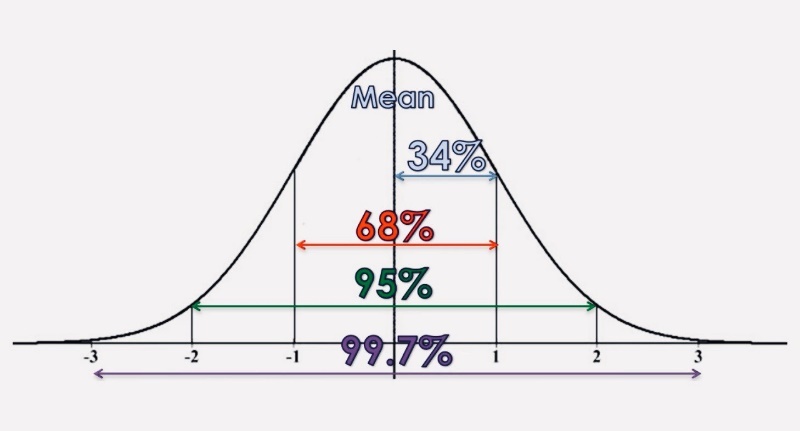
Lower Whisker = Q1-1.5

Upper Whisker = Q3+1.5

**Standard Deviation and Variance**

*When you have a low standard deviation, your data points tend to be close to the mean. A high standard deviation means that your data points are spread out over a wide range.*

*Standard deviation is best used when data is unimodal. In a normal distribution, approximately 34% of the data points are lying between the mean and one standard deviation above or below the mean. Since a normal distribution is symmetrical, 68% of the data points fall between one standard deviation above and one standard deviation below the mean. Approximately 95% fall between two standard deviations below the mean and two standard deviations above the mean. And approximately 99.7% fall between three standard deviations above and three standard deviations below the mean.*



**Coefficient of Variation**

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a

data series around the mean. The coefficient of variation represents the ratio of the standard

deviation to the mean, and it is a useful statistic for comparing the degree of variation from

one data series to another, even if the means are drastically different from one another.

When SD, Mean, Variance etc doesn’t support to select value, we go for Coefficient of variation.

Coefficient of variation = (SD/Mean) \* 100

The coefficient of variation shows the extent of [variability](https://www.investopedia.com/terms/v/variability.asp) of data in a sample in relation to the mean of the population. In finance, the coefficient of variation allows investors to determine how much volatility, or risk, is assumed in comparison to the amount of return expected from investments. Ideally, the coefficient of variation formula should result in a lower ratio of the [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp) to mean return, meaning the better risk-return trade-off. Note that if the expected return in the denominator is negative or zero, the coefficient of variation could be misleading.

The coefficient of variation is helpful when using the risk/reward ratio to select investments. For example, an investor who is risk-averse may want to consider assets with a historically low degree of [volatility](https://www.investopedia.com/terms/v/volatility.asp) and a high degree of return, in relation to the overall market or its industry. Conversely, risk-seeking investors may look to invest in assets with a historically high degree of volatility.

Coefficient of variation is less means more consistent.

**Covariance and Correlation**

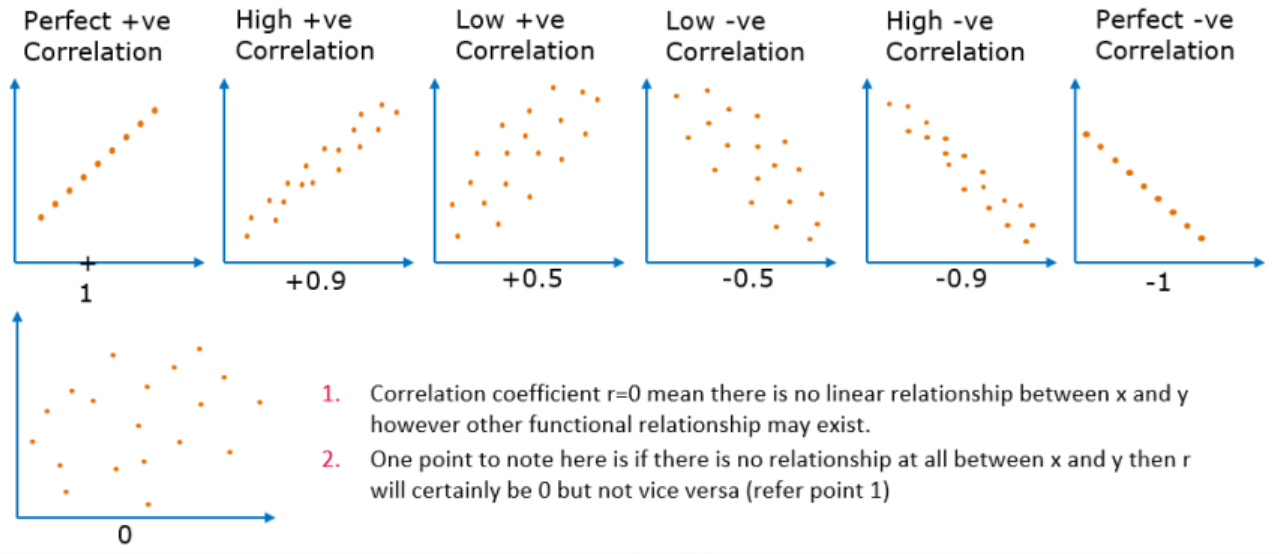
Covariance and Correlation are very helpful in understanding the relationship between two continuous variables.

**Covariance**

Covariance tells whether both variables vary in the same direction (positive covariance) or in the opposite direction (negative covariance). There is no meaning of covariance numerical value only sign is useful. ***Covariance varies between -∞ to +∞.*** **Covariance causes Causation**

**Correlation**

Correlation explains the change in one variable leads how much proportion change in the second variable. ***Correlation varies between -1 to +1***. If the correlation value is 0 then it means there is no Linear Relationship between variables however other functional relationship may exist. **Correlation doesn’t cause causation.**

****

**Survey Sampling Methods**

Sampling method refers to the way that observations are selected from a [population](https://stattrek.com/Help/Glossary.aspx?Target=Population) to be in the [sample](https://stattrek.com/Help/Glossary.aspx?Target=Sample) for a [sample survey](https://stattrek.com/Help/Glossary.aspx?Target=Sample%20survey).

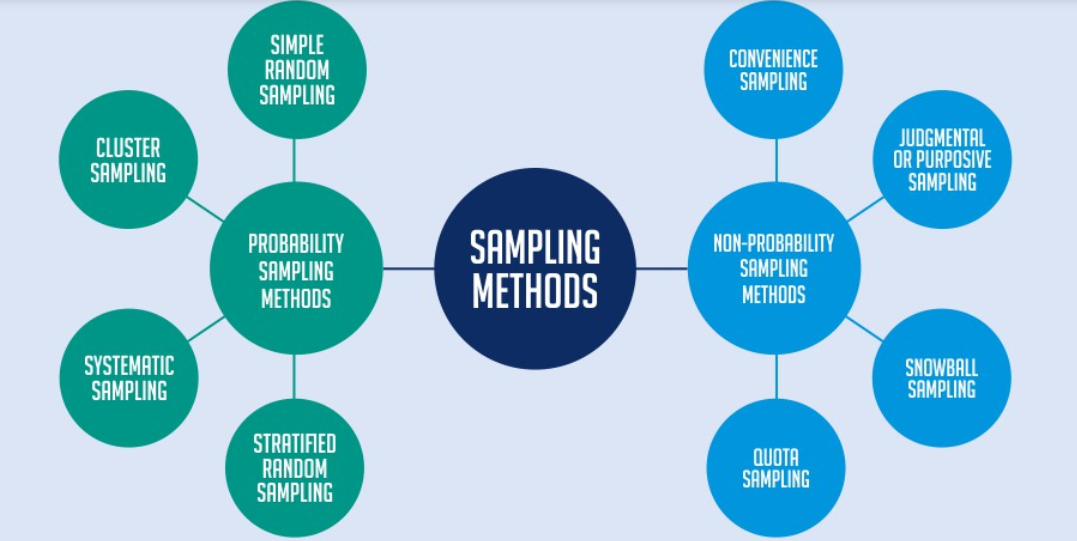
**Population Parameter vs. Sample Statistic**

The reason for conducting a sample survey is to estimate the value of some attribute of a population.

* Population parameter. A population parameter is the true value of a population attribute.
* Sample statistic. A sample statistic is an estimate, based on sample data, of a population parameter.

Consider this example. A public opinion pollster wants to know the percentage of voters that favour a flat-rate income tax. The actual percentage of all the voters is a population parameter. The estimate of that percentage, based on sample data, is a sample statistic.

The quality of a sample statistic (i.e., accuracy, precision, representativeness) is strongly affected by the way that sample observations are chosen; that is., by the sampling method.



**Probability vs. Non-Probability Samples**

As a group, sampling methods fall into one of two categories.

* Probability samples. With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample.
* Non-probability samples. With non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen.

Non-probability sampling methods offer two potential advantages - convenience and cost. The main disadvantage is that non-probability sampling methods do not allow you to estimate the extent to which sample statistics are likely to differ from population parameters. Only probability sampling methods permit that kind of analysis.

**Non-Probability Sampling Methods**

Two of the main types of non-probability sampling methods are voluntary samples and convenience samples.

* **Voluntary sample.** A voluntary sample is made up of people who self-select into the survey. Often, these folks have a strong interest in the main topic of the survey.

Suppose, for example, that a news show asks viewers to participate in an online poll. This would be a volunteer sample. The sample is chosen by the viewers, not by the survey administrator.

* **Convenience sample.** A convenience sample is made up of people who are easy to reach.

Consider the following example. A pollster interviews shoppers at a local mall. If the mall was chosen because it was a convenient site from which to solicit survey participants and/or because it was close to the pollster's home or business, this would be a convenience sample.

* **Convenience sampling:**

This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as [convenience sampling](https://www.questionpro.com/blog/convenience-sampling/), as it’s carried out on the basis of how easy is it for a researcher to get in touch with the subjects. Researchers have nearly no authority over selecting elements of the sample and it’s purely done on the basis of proximity and not representativeness. This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.

For example, startups and NGOs usually conduct convenience sampling at a mall to distribute leaflets of upcoming events or promotion of a cause – they do that by standing at the entrance of the mall and giving out pamphlets randomly.

* **Judgmental or Purposive Sampling:**

In [judgemental or purposive sampling](https://www.questionpro.com/blog/judgmental-sampling/), the sample is formed by the discretion of the judge purely considering the purpose of study along with the understanding of target audience. Also known as deliberate sampling, the participants are selected solely on the basis of research requirements and elements who do not suffice the purpose are kept out of the sample. For instance, when researchers want to understand the thought process of people who are interested in studying for their master’s degree. The selection criteria will be: “Are you interested in studying for Masters in …?” and those who respond with a “No” will be excluded from the sample.

* **Snowball sampling:** [Snowball sampling](https://www.questionpro.com/blog/snowball-sampling/) is a sampling method that is used in studies which need to be carried out to understand subjects which are difficult to trace. For example, it will be extremely challenging to survey shelterless people or illegal immigrants. In such cases, using the snowball theory, researchers can track a few of that particular category to interview and results will be derived on that basis. This sampling method is implemented in situations where the topic is highly sensitive and not openly discussed such as conducting surveys to gather information about HIV Aids. Not many victims will readily respond to the questions but researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.
* **Quota sampling:**  In [Quota sampling](https://www.questionpro.com/blog/quota-sampling/), selection of members in this sampling technique happens on basis of a pre-set standard. In this case, as a sample is formed on basis of specific attributes, the created sample will have the same attributes that are found in the total population. It is an extremely quick method of collecting samples.

#### **Use of the Non-Probability Sampling Method**

There are multiple uses of the non-probability sampling method. They are:

* Create a hypothesis: The [non-probability sampling method](https://www.questionpro.com/blog/non-probability-sampling/) is used to create a hypothesis when limited to no prior information is  available. This method helps with immediate return of data and helps to build a base for any further research.
* Exploratory research: This sampling technique is widely used when researchers aim at conducting qualitative research, pilot studies or [exploratory research](https://www.questionpro.com/blog/what-is-research/).
* Budget and time constraints: The non-probability method when there are budget and time constraints and some preliminary data has to be collected. Since the [survey design](https://www.questionpro.com/features/survey-design/) is not rigid, it is easier to pick respondents at random and have them take the [survey](https://www.questionpro.com/tour/surveys.html) or [questionnaire](https://www.questionpro.com/blog/what-is-a-questionnaire/).

**Probability Sampling Methods**

The main types of probability sampling methods are simple random sampling, stratified sampling, cluster sampling, multistage sampling, and systematic random sampling. The key benefit of probability sampling methods is that they guarantee that the sample chosen is representative of the population. This ensures that the statistical conclusions will be valid.

* **Simple random sampling**. Simple random sampling refers to any sampling method that has the following properties.
  + The population consists of N objects.
  + The sample consists of n objects.
  + If all possible samples of n objects are equally likely to occur, the sampling method is called simple random sampling.

There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the N population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects n numbers. Population members having the selected numbers are included in the sample.

* **Stratified sampling**. With stratified sampling, the population is divided into groups, based on some characteristic. Then, within each group, a probability sample (often a simple random sample) is selected. In stratified sampling, the groups are called strata.

As a example, suppose we conduct a national survey. We might divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.

* **Cluster sampling**. With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen, using a probability method (often simple random sampling). Only individuals within sampled clusters are surveyed.

Note the difference between cluster sampling and stratified sampling. With stratified sampling, the sample includes elements from each stratum. With cluster sampling, in contrast, the sample includes elements only from sampled clusters.

* **Multistage sampling**. With multistage sampling, we select a sample by using combinations of different sampling methods.

For example, in Stage 1, we might use cluster sampling to choose clusters from a population. Then, in Stage 2, we might use simple random sampling to select a subset of elements from each chosen cluster for the final sample.

* **Systematic random sampling**. With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first k elements on the population list. Thereafter, we select every kth element on the list. This method is different from simple random sampling since every possible sample of n elements is not equally likely.

#### **Use of the Probability Sampling Method**

There are multiple uses of the probability sampling method. They are:

* **Reduce Sample Bias:** Using the probability sampling method, the bias in the sample derived from a population is negligible to non-existent. The selection of the sample largely depicts the understanding and the inference of the researcher. Probability sampling leads to higher quality [data collection](https://www.questionpro.com/blog/data-collection/) as the population is appropriately represented by the sample.
* **Diverse Population:** When the population is large and diverse, it is important to have adequate representation so that the data is not skewed towards one [demographic](https://www.questionpro.com/blog/demographic-examples/). For example, if Square would like to understand the people that could their point-of-sale devices, a survey conducted from a sample of people across US from different industries and socio-economic backgrounds, helps.
* **Create an Accurate Sample:** Probability sampling helps the researchers plan and create an accurate sample. This helps to obtain well-defined data.

|  |  |  |
| --- | --- | --- |
|  | Probability Sampling Methods | Non-Probability Sampling Methods |
| Definition | Probability Sampling is a sampling technique in which sample from a larger population are chosen using a method based on the theory of probability. | Non-probability sampling is a sampling technique in which the researcher selects samples based on the subjective judgment of the researcher rather than random selection. |
| Alternatively Known as | Random sampling method. | Non-random sampling method |
| Population selection | The population is selected randomly. | The population is selected arbitrarily. |
| Market Research | The research is conclusive in nature. | The research is exploratory in nature. |
| Sample | Since there is method to deciding the sample, the population demographics is conclusively represented. | Since the sampling method is arbitrary, the population demographics representation is almost always skewed. |
| Time Taken | Take a longer time to conduct since the research design defines the selection parameters before the market research study begins. | This type of sampling method is quick since neither the sample or selection criteria of the sample is undefined. |
| Results | This type of sampling is entirely unbiased and hence the results are unbiased too and conclusive. | This type of sampling is entirely biased and hence the results are biased too rendering the research speculative. |
| Hypothesis | In probability sampling, there is an underlying hypothesis before the study begins and the objective of this method is to prove the hypothesis. | In non-probability sampling, the hypothesis is derived after conducting the research study. |

Problem

An auto analyst is conducting a satisfaction survey, sampling from a list of 10,000 new car buyers. The list includes 2,500 Ford buyers, 2,500 GM buyers, 2,500 Honda buyers, and 2,500 Toyota buyers. The analyst selects a sample of 400 car buyers, by randomly sampling 100 buyers of each brand.

Is this an example of a simple random sample?

1. Yes, because each buyer in the sample was randomly sampled.
2. Yes, because each buyer in the sample had an equal chance of being sampled.
3. Yes, because car buyers of every brand were equally represented in the sample.
4. No, because every possible 400-buyer sample did not have an equal chance of being chosen.
5. No, because the population consisted of purchasers of four different brands of car.

Solution

The correct answer is (D). A [simple random sample](https://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling) requires that every [sample](https://stattrek.com/Help/Glossary.aspx?Target=Sample) of size n (in this problem, n is equal to 400) has an equal chance of being selected. In this problem, there was a 100 percent chance that the sample would include 100 purchasers of each brand of car. There was zero percent chance that the sample would include, for example, 99 Ford buyers, 101 Honda buyers, 100 Toyota buyers, and 100 GM buyers. Thus, all possible samples of size 400 did not have an equal chance of being selected; so this cannot be a simple random sample.

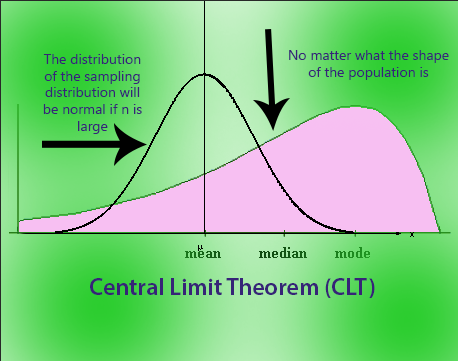
The fact that each buyer in the sample was randomly sampled is a necessary condition for a simple random sample, but it is not sufficient. Similarly, the fact that each buyer in the sample had an equal chance of being selected is characteristic of a simple random sample, but it is not sufficient. The sampling method in this problem used random sampling and gave each buyer an equal chance of being selected; but the sampling method was actually [stratified random sampling](https://stattrek.com/Help/Glossary.aspx?Target=Stratified%20sampling).

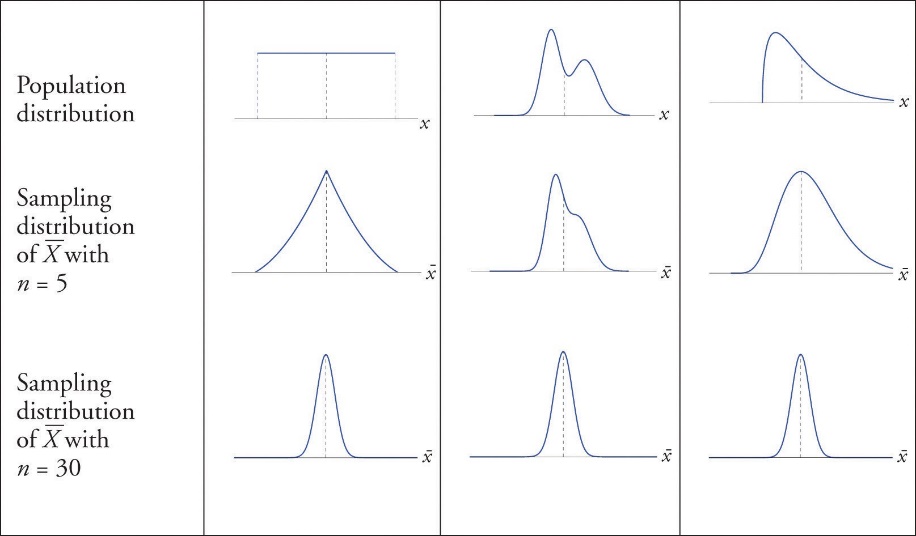
The fact that car buyers of every brand were equally represented in the sample is irrelevant to whether the sampling method was simple random sampling. Similarly, the fact that population consisted of buyers of different car brands is irrelevant.

**Central Limit Theorem**

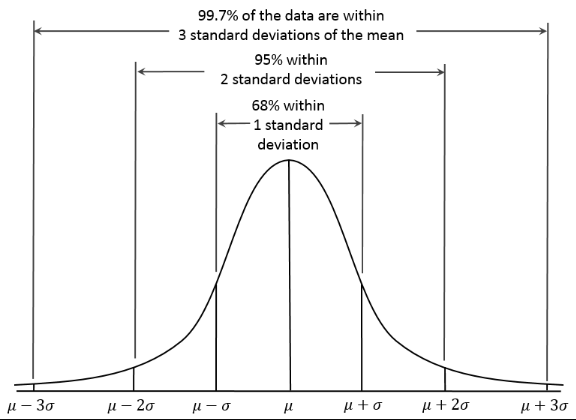
The normal distribution is used to help measure the accuracy of many statistics, including the sample mean, using an important result called the Central Limit Theorem. This theorem gives you the ability to measure how much the means of various samples will vary, without having to take any other sample means to compare it with. By taking this variability into account, you can use your data to answer questions about a population.

The Central Limit Theorem (CLT for short) basically says that for non-normal data, the distribution of the sample means has an approximate normal distribution, no matter what the distribution of the original data looks like, as long as the sample size is large enough (usually at least 30) and all samples have the same size. And it doesn’t just apply to the sample mean; the CLT is also true for other sample statistics, such as the sample proportion. Because statisticians know so much about the normal distribution, these analyses are much easier.





Normal Distribution also known as Gaussian distribution is nothing but symmetrically distributed curve from Mean where Mean=Median=Mode=Zero and 1 unit standard deviated.



Probability Distribution is nothing but probabilities of occurrence of different possible outcomes. A probability distribution is specified in terms of an underlying Sample space. Probability Distribution can be of

1. Continuous or (Temperature on given day)
2. Discrete (Coin flip)

Normal Distribution is commonly encountered continuous probability distribution.

Since Probability is a Sample space and N number of Sample space contribute to Normal distribution which is continuous

**Confidence Interval**

A confidence interval, in statistics, refers to the probability that a population parameter will fall between two set values for a certain proportion of times. Confidence intervals measure the degree of uncertainty or certainty in a sampling method. A confidence interval can take any number of probabilities, with the most common being a 95% or 99% confidence level

Statisticians use confidence intervals to measure uncertainty. For example, a researcher selects different samples randomly from the same population and computes a confidence interval for each sample. The resulting datasets are all different; some intervals include the true population parameter and others do not.

A Confidence interval is a range of values that likely would contain an unknown population parameter. Confidence level refers to the percentage of probability, or certainty, that the confidence interval would contain the true population parameter when you draw a random sample many times. Or, in the vernacular, "We are 99% certain (confidence level) that most of these datasets (confidence intervals) contain the true population parameter."

### **Factors that Affect Confidence Intervals (CI)**

* Population size: this does not usually affect the CI but can be a factor if you are working with small and known groups of people.
* [Sample Size](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/): the smaller your sample, the less likely it is you can be confident the results reflect the true population [parameter](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/).
* Percentage: Extreme answers come with better [accuracy](https://www.statisticshowto.datasciencecentral.com/accuracy-and-precision/). For example, if 99 percent of voters are for Voting system, the chances of error are small. However, if 49.9 percent of voters are “for” and 50.1 percent are “against” then the chances of error are bigger

**Standard Error**

Standard Error is how far the sample mean of data is likely to be from true population mean.

**Z Score**

The z score tells you how many standard deviations from the mean your score is

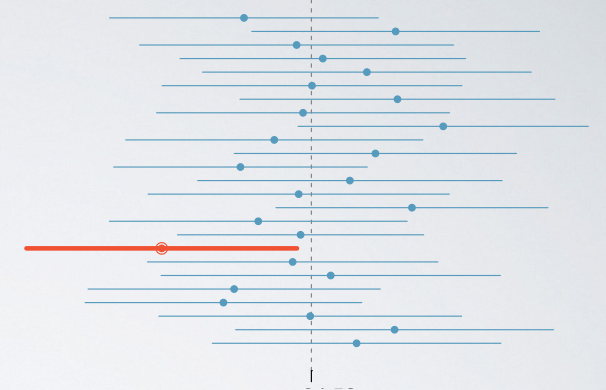
Z Score = (X-µ)/σ

Where µ is mean, σ is Standard Deviation

Note: Z Scores can be easily got from Z score table. So, there is no need to calculate.

**More Confidence Level is More accurate or Not**

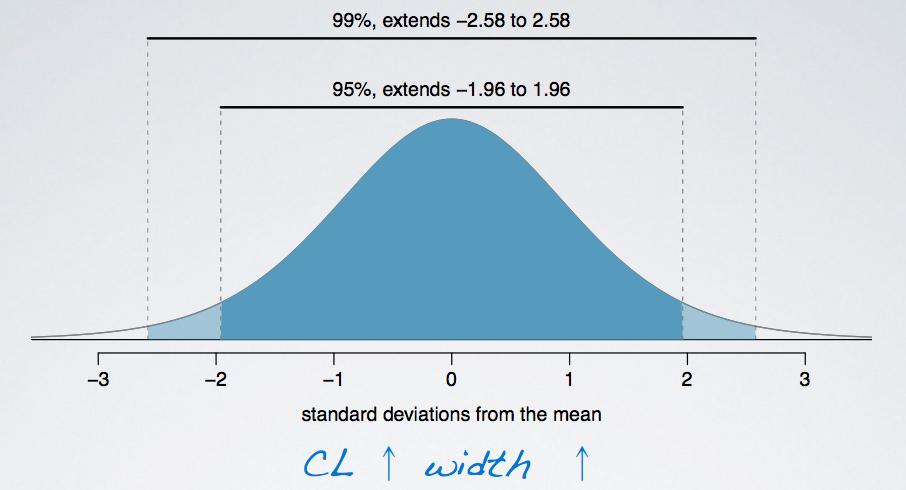
Looking at this figure, it seems like a wider interval would indeed be much better. You can think about the red interval that is plotted on this figure and imagine that it extends even further. It would be much likely for it to then capture the true population parameter which is shown here as the vertical dashed line.



Therefore, as the confidence level increase, so does the width of the confidence interval.

Another way of thinking about this is the width of the area that captures the middle 95% or 99% of the distribution.

* The middle 99% will inevitably span a larger area, and hence the 99% confidence interval is going to be wider. Therefore, as we increase the confidence level, the width of the interval increases as well.
* More accurate means a higher confidence level. So if we are saying that we want to increase accuracy, we also need to increase the confidence level, but this might come at a cost.



What is the drawback when using a wider interval

As the confidence level increase, the width of the confidence interval increase as well. Which then increase the accuracy. However, the precision goes down.

CL↑ Width↑ Accuracy↑, but Precision↓

Example:

Suppose you are watching the weather forecast, and you are told that the next day, low is -20F and high is 110F.

* Is this accurate? Most likely, yes.
* Tomorrow's temperature is probably going to be somewhere between -20F and 100F, however is it informative? Or, in other wards, is it precise? Not really. It is nearly impossible to figure out what to wear tomorrow according to this information.

Example

The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the US. In 2010, the survey collected responses from 1,154 US residents. Based on the survey results, a 95% confidence interval for the average number of hours Americans have to relax or pursue activities that they enjoy after an average work day was found to be 3.53 to 3.83 hours. Determine if each of the following statements are true or false.

* (a) 95% of Americans spend 3.53 to 3.83 hours relaxing after a work day.
* (b) 95% of random samples of 1,154 Americans will yield confidence intervals that contain the true average number of hours Americans spend relaxing after a work day.
* (c) 95% of the time the true average number of hours Americans spend relaxing after a work day is between 3.53 and 3.83 hours.
* (d) We are 95% confident that Americans in this sample spend on average 3.53 to 3.83 hours relaxing after a work day.
* (a) is False, because the confidence interval is not about individuals in the population. But in stead, about the true population parameter.
* (b) is True, because it is the definition of the confidence level. The percentage of random samples that will yield confidence intervals that contain the true population parameter.
* (c) is False, because the population parameter is not this moving target that is sometimes within an interval and sometime outside of it.
* (d) is False, because the confidence interval is not about the sample mean but instead about the population mean.

**Hypothesis Test**

**Z Test**

A Statistical hypothesis is a statement about a population parameter. It may or may not be true.

H0🡪 Null Hypothesis (or) Default Hypothesis – (always start Null Hypothesis is right)

Ex: Coin is Fair

H1🡪 Alternate Hypothesis

Ex: Coin is Biased

Testing Procedure

1. Formulate the Null and the alternative Hypothesis
2. Choose Level of Significance (Confidence Level)
3. Find Critical Value
4. Find test Statistic (Since it is from Sample, Sample size > 30)
5. Draw your conclusion

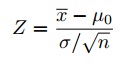
Example:

Blood G[lucose levels](https://www.nlm.nih.gov/medlineplus/ency/article/003482.htm) for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw corn starch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw corn starch diet have a mean glucose level of 140. Test the hypothesis that the raw corn starch had an effect.

Step 1: [State the null hypothesis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/#state): H0:μ=100

Step 2: State the [alternate hypothesis](https://www.statisticshowto.datasciencecentral.com/what-is-an-alternate-hypothesis/): H1:≠100

Step 3: State your[level of significance .](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/) We’ll use 0.05 for this example. As this is a two-tailed test, split the alpha into two.  
0.05/2=0.025

Step 4: Find the [z-score](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) associated with your [level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/) of significance. You’re looking for the area in one tail only. A z-score for 0.95 (1-0.025=0.975) is 1.96. As this is a two-tailed test, you would also be considering the left tail (z=1.96)  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/02/z-score-formula.jpg)

Step 5: Find the [test statistic](https://www.statisticshowto.datasciencecentral.com/test-statistic/) using this formula:    
z=(140-100)/(15/√30)=14.60.

Step 6: If Step 5 is less than -1.96 or greater than 1.96 (Step 3), [reject the null hypothesis](https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/). In this case, it is greater, so you can reject the null.

**T Test**

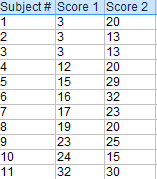
T test is performed when sample size < 30 & Population not known

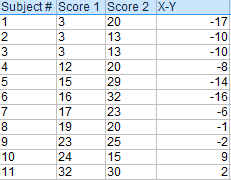
When Sample size increases 🡪 T test becomes Z test

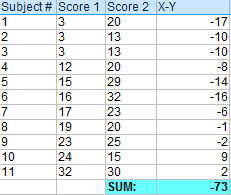
If 1 variable is not available we can find that from population mean.

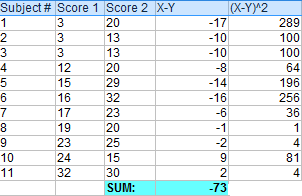
**Degree Of freedom**

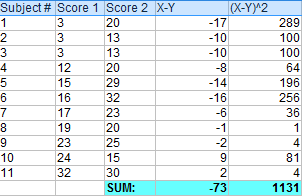
Degrees of freedom are the number of independent values that a statistical analysis can estimate. Typically, the degrees of freedom equal your sample size minus the number of parameters you need to calculate during an analysis.

Example: Calculate a paired t test by hand for the following data:  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-2.png)

Step 1: Subtract each Y score from each X score.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-3.png)

Step 2: Add up all of the values from Step 1.  
Set this number aside for a moment.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-4.png)

Step 3: Square the differences from Step 1.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-5.png)

Step 4: Add up all of the squared differences from Step 3.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-6.png)

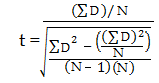
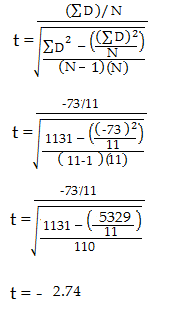
To find 1 variable (Z) = (∑(x-y)^2)/(n-1)

= 1131/10

= 113.1

T = (z-µ)/(s/√n)

= 113.1-

Step 5: Use the following formula to calculate the t-score:  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example.png)  
  
  
ΣD: Sum of the differences (Sum of X-Y from Step 2)  
ΣD2: Sum of the squared differences (from Step 4)  
(ΣD)2: Sum of the differences (from Step 2), squared.  
[](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2014/12/paired-t-test-example-7v3.png)

Step 6: Subtract 1 from the sample size to get the degrees of freedom. We have 11 items, so 11-1 = 10.

Step 7: Find the [p-value](https://www.statisticshowto.datasciencecentral.com/p-value/) in the [t-table](https://www.statisticshowto.datasciencecentral.com/tables/t-distribution-table/), using the [degrees of freedom](https://www.statisticshowto.datasciencecentral.com/degrees-of-freedom/) in Step 6. If you don’t have a specified [alpha level](https://www.statisticshowto.datasciencecentral.com/what-is-an-alpha-level/), use 0.05 (5%). For this sample problem, with df=10, the t-value is 2.228.

Step 8: Compare your t-table value from Step 7 (2.228) to your calculated t-value (-2.74). The calculated t-value is greater than the table value at an alpha level of .05. The p-value is less than the alpha level: p <.05. We can reject the null hypothesis that there is no difference between means.

Note: You can ignore the minus sign when comparing the two t-values, as ± indicates the direction; the p-value remains the same for both directions.