1. **What is central limit theorem and why is it important?**

The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all [samples](https://www.simplilearn.com/tutorials/machine-learning-tutorial/population-vs-sample) from that population will be roughly equal to the population mean.

The central limit theorem doesn't have its own formula, but it relies on sample mean and standard deviation. As sample means are gathered from the population, standard deviation is used to distribute the data across a probability distribution curve.

Importance of Central limit theorem:

* The central limit theorem is useful when analyzing large data sets because it allows one to assume that the sampling distribution of the mean will be normally-distributed in most cases. This allows for easier statistical analysis and inference. For example, investors can use central limit theorem to aggregate individual security performance data and generate distribution of sample means that represent a larger population distribution for security returns over a period of time.

1. **What is sampling? How many sampling methods do you know?**

When you conduct research about a group of people, it’s rarely possible to collect data from every person in that group. Instead, you select a **sample**. The sample is the group of individuals who will actually participate in the research.

To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. This is called a **sampling method**. There are two primary types of sampling methods that you can use in your research:

* [**Probability sampling**](https://www.scribbr.com/methodology/probability-sampling/) involves random selection, allowing you to make strong statistical inferences about the whole group.
* [**Non-probability sampling**](https://www.scribbr.com/methodology/non-probability-sampling/) involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

You should clearly explain how you selected your sample in the [methodology](https://www.scribbr.com/dissertation/methodology/) section of your paper or thesis, as well as how you approached minimizing [research bias](https://www.scribbr.com/faq-category/research-bias/) in your work.

**Probability sampling methods**

[Probability sampling](https://www.scribbr.com/methodology/probability-sampling/) means that every member of the population has a chance of being selected. It is mainly used in [quantitative research](https://www.scribbr.com/methodology/quantitative-research/). If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

There are four main types of probability sample.

**1.**[**Simple random sampling**](https://www.scribbr.com/methodology/simple-random-sampling/)

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example: Simple random samplingYou want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

**2.**[**Systematic sampling**](https://www.scribbr.com/methodology/systematic-sampling/)

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example: Systematic samplingAll employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

**3.**[**Stratified sampling**](https://www.scribbr.com/methodology/stratified-sampling/)

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or [systematic sampling](https://www.scribbr.com/methodology/systematic-sampling/) to select a sample from each subgroup.

Example: Stratified samplingThe company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

**4.**[**Cluster sampling**](https://www.scribbr.com/methodology/cluster-sampling/)

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called [multistage sampling](https://www.scribbr.com/methodology/multistage-sampling/).

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It’s difficult to guarantee that the sampled clusters are really representative of the whole population.

Example: Cluster samplingThe company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don’t have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

**Non-probability sampling methods**

In a non-probability sample, individuals are selected based on non-random criteria, and not every individual has a chance of being included.

This type of sample is easier and cheaper to access, but it has a higher risk of [sampling bias](https://www.scribbr.com/research-bias/sampling-bias/). That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Non-probability sampling techniques are often used in [exploratory](https://www.scribbr.com/methodology/exploratory-research/) and [qualitative research](https://www.scribbr.com/methodology/qualitative-research/). In these types of research, the aim is not to test a[hypothesis](https://www.scribbr.nl/methodology/hypotheses/) about a broad population, but to develop an initial understanding of a small or under-researched population.

**1.**[**Convenience sampling**](https://www.scribbr.com/methodology/convenience-sampling/)

A convenience sample simply includes the individuals who happen to be most accessible to the researcher.

This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can’t produce [generalizable](https://www.scribbr.com/research-bias/generalizability/) results. Convenience samples are at risk for both [sampling bias](https://www.scribbr.com/research-bias/sampling-bias/) and [selection bias](https://www.scribbr.com/research-bias/selection-bias/).

Example: Convenience samplingYou are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a[survey](https://www.scribbr.com/methodology/survey-research/) on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

**2. Voluntary response sampling**

Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).

Voluntary response samples are always at least somewhat [biased](https://www.scribbr.com/faq-category/research-bias/), as some people will inherently be more likely to volunteer than others, leading to [self-selection bias](https://www.scribbr.com/research-bias/self-selection-bias/).

Example: Voluntary response samplingYou send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can’t be sure that their opinions are representative of all students.

**3.**[**Purposive sampling**](https://www.scribbr.com/methodology/purposive-sampling/)

This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

It is often used in [qualitative research](https://www.scribbr.com/methodology/qualitative-research/), where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your [inclusion and exclusion criteria](https://www.scribbr.com/methodology/inclusion-exclusion-criteria/) and beware of [observer bias](https://www.scribbr.com/research-bias/observer-bias/) affecting your arguments.

Example: Purposive samplingYou want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

**4.**[**Snowball sampling**](https://www.scribbr.com/methodology/snowball-sampling/)

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to “snowballs” as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to [sampling bias](https://www.scribbr.com/research-bias/sampling-bias/).

Example: Snowball samplingYou are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn’t possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

1. **What is the difference between type1 and typeII error?**

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

Examle: Type I error (false positive): the test result says you have coronavirus, but you actually don't. Type II error (false negative): the test result says you don't have coronavirus, but you actually do

1. **What do you understand by the term Normal distribution?**

The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in [statistics](https://statisticsbyjim.com/glossary/statistics/) for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

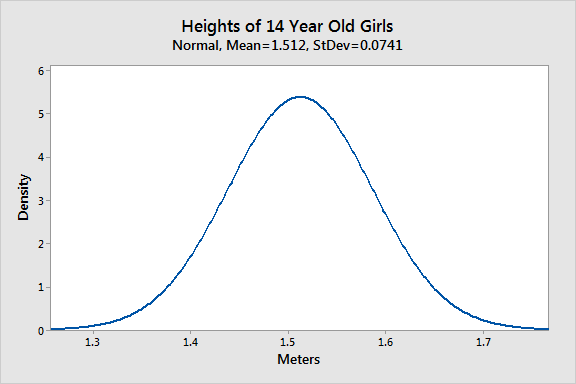
The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal. For example, the Student’s t, Cauchy, and logistic distributions are symmetric.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

In this blog post, learn how to use the normal distribution, about its [parameters](https://statisticsbyjim.com/glossary/parameter/), the Empirical Rule, and how to calculate Z-scores to standardize your data and find probabilities.

Example of Normally Distributed Data: Heights

Height data are normally distributed. The distribution in this example fits real data that I collected from 14-year-old girls during a study. The graph below displays the probability distribution function for this normal distribution. Learn more about [Probability Density Functions](https://statisticsbyjim.com/probability/probability-density-function/).



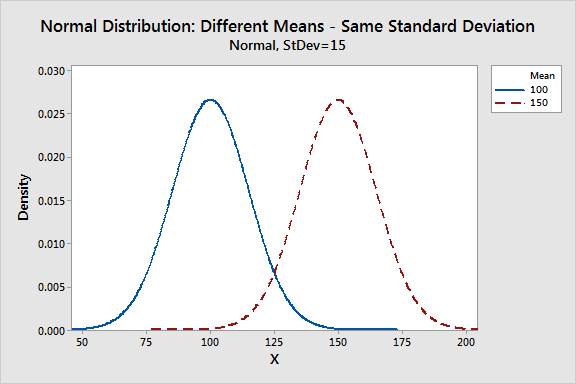
As you can see, the distribution of heights follows the typical bell curve pattern for all normal distributions. Most girls are close to the average (1.512 meters). Small differences between an individual’s height and the mean occur more frequently than substantial deviations from the mean. The standard deviation is 0.0741m, which indicates the typical distance that individual girls tend to fall from mean height.

The distribution is symmetric. The number of girls shorter than average equals the number of girls taller than average. In both tails of the distribution, extremely short girls occur as infrequently as extremely tall girls.

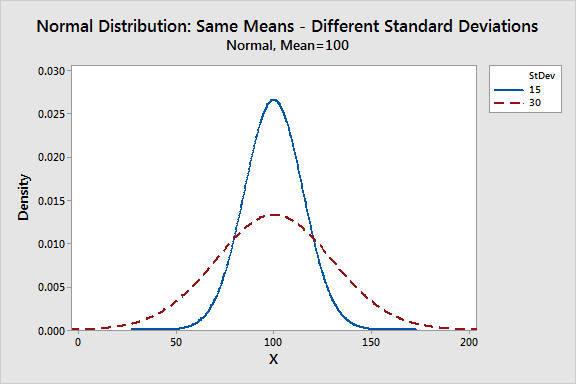
Parameters of the Normal Distribution

As with any probability distribution, the parameters for the normal distribution define its shape and probabilities entirely. The normal distribution has two parameters, the mean and standard deviation. The Gaussian distribution does not have just one form. Instead, the shape changes based on the parameter values, as shown in the graphs below.

Mean

The mean is the central tendency of the normal distribution. It defines the location of the peak for the bell curve. Most values cluster around the mean. On a graph, changing the mean shifts the entire curve left or right on the X-axis.

Standard deviation

The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far away from the mean the values tend to fall. It represents the typical distance between the observations and the average.

On a graph, changing the standard deviation either tightens or spreads out the width of the distribution along the X-axis. Larger standard deviations produce wider distributions.

When you have narrow distributions, the probabilities are higher that values won’t fall far from the mean. As you increase the spread of the bell curve, the likelihood that observations will be further away from the mean also increases.

Population parameters versus sample estimates

The mean and standard deviation are parameter values that apply to entire populations. For the Gaussian distribution, statisticians signify the parameters by using the Greek symbol μ (mu) for the [population](https://statisticsbyjim.com/glossary/population/) mean and σ (sigma) for the population standard deviation.

Unfortunately, population parameters are usually unknown because it’s generally impossible to measure an entire population. However, you can use random samples to calculate [estimates](https://statisticsbyjim.com/glossary/estimator/) of these parameters. Statisticians represent [sample](https://statisticsbyjim.com/glossary/sample/) estimates of these parameters using x̅ for the sample mean and s for the sample standard deviation.

Common Properties for All Forms of the Normal Distribution

Despite the different shapes, all forms of the normal distribution have the following characteristic properties.

* They’re all symmetric bell curves. The Gaussian distribution cannot model skewed distributions.
* The mean, median, and [mode](https://statisticsbyjim.com/glossary/mode/) are all equal.
* Half of the population is less than the mean and half is greater than the mean.
* The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean. More on this below!

While the normal distribution is essential in statistics, it is just one of many probability distributions, and it does not fit all populations. To learn how to determine whether the normal distribution provides the best fit to your sample data, read my posts about [How to Identify the Distribution of Your Data](https://statisticsbyjim.com/hypothesis-testing/identify-distribution-data/) and [Assessing Normality: Histograms vs. Normal Probability Plots](https://statisticsbyjim.com/basics/assessing-normality-histograms-probability-plots/).

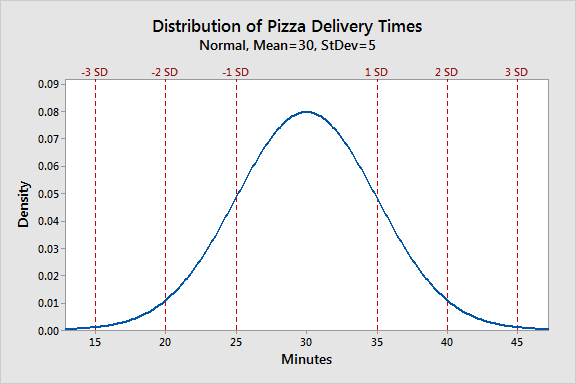
The [uniform distribution](https://statisticsbyjim.com/probability/uniform-distribution/) also models symmetric, [continuous data](https://statisticsbyjim.com/glossary/continuous-variables/), but all equal-sized ranges in this distribution have the same probability, which differs from the normal distribution.

If you have continuous data that are skewed, you’ll need to use a different distribution, such as the [Weibull](https://statisticsbyjim.com/probability/weibull-distribution/" \t "_blank), [lognormal](https://statisticsbyjim.com/probability/lognormal-distribution/), [exponential](https://statisticsbyjim.com/probability/exponential-distribution/), or [gamma](https://statisticsbyjim.com/probability/gamma-distribution/) distribution.

The Empirical Rule for the Normal Distribution

When you have normally distributed data, the standard deviation becomes particularly valuable. You can use it to determine the proportion of the values that fall within a specified number of standard deviations from the mean. For example, in a normal distribution, 68% of the observations fall within +/- 1 standard deviation from the mean. This property is part of the Empirical Rule, which describes the percentage of the data that fall within specific numbers of standard deviations from the mean for bell-shaped curves.

|  |  |
| --- | --- |
| **Mean +/- standard deviations** | **Percentage of data contained** |
| 1 | 68% |
| 2 | 95% |
| 3 | 99.7% |

Let’s look at a pizza delivery example. Assume that a pizza restaurant has a mean delivery time of 30 minutes and a standard deviation of 5 minutes. Using the Empirical Rule, we can determine that 68% of the delivery times are between 25-35 minutes (30 +/- 5), 95% are between 20-40 minutes (30 +/- 2\*5), and 99.7% are between 15-45 minutes (30 +/-3\*5). The chart below illustrates this property graphically.

If your data do not follow the Gaussian distribution and you want an easy method to determine proportions for various standard deviations, use [Chebyshev’s Theorem](https://statisticsbyjim.com/basics/chebyshevs-theorem-in-statistics/" \t "_blank)! That method provides a similar type of result as the Empirical Rule but for non-normal data.

1. **What is correlation and covariance in statistics?**

**Covariance:**

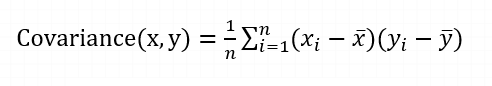
Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.

The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.

The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number.

A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables. Covariance is great for defining the type of relationship, but it's terrible for interpreting the magnitude.

Let Σ(X) and Σ(Y) be the expected values of the variables, the covariance formula can be represented as:



Where,

* xi = data value of x
* yi = data value of y
* x̄ = mean of x
* ȳ = mean of y
* N = number of data values.

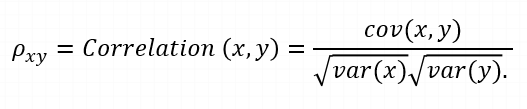
## Application Of Covariance

The following are the most common applications of Covariance:

* Simulating systems with multiple correlated variables is done using Cholesky decomposition. A covariance matrix helps determine the Cholesky decomposition because it is positive semi-definite. The matrix is decomposed by the product of the lower matrix and its transpose.
* To reduce the dimensions of large data sets, principal component analysis is used. To perform principal component analysis, an eigen decomposition is applied to the covariance matrix

**Correlation:**

In statistics, correlation is a measure that determines the degree to which two or more random variables move in sequence. When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated. The formula for correlation is:



where,

var(X) = standard deviation of X

var(Y) = standard deviation of Y

Positive correlation occurs when two variables move in the same direction. When variables move in the opposite direction, they are said to be negatively correlated.

Correlation is of three types:

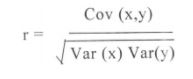
1. **Simple Correlation:** In simple correlation, a single number expresses the degree to which two variables are related.
2. **Partial Correlation:** When one variable's effects are removed, the correlation between two variables is revealed in partial correlation.
3. **Multiple correlation:** A statistical technique that uses two or more variables to predict the value of one variable.

## Methods of Calculating Correlation

There are a number of methods to calculate correlation coefficient. Here are some of the most common ones:

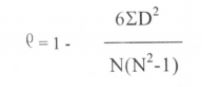
### Coefficient of correlation

This is the most common method of determining the correlation coefficient of two variables. It is obtained by dividing the covariance of two variables with the product of their [standard deviations.](https://www.simplilearn.com/tutorials/statistics-tutorial/what-is-normal-distribution)



### Rank Correlation Coefficient

A [rank correlation coefficient](https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation) measures the degree of similarity between two variables, and can be used to assess the significance of the relation between them. It measures the extent to which, as one variable increases, the other decreases.



where,

ρ = coefficient of rank relation

D = difference between paired ranks

N = number of items ranked

### Coefficient of Concurrent Deviations

Coefficient of concurrent deviations is used when you want to study the correlation in a very casual manner and there is not much need to attain precision.



where,

rc = coefficient of concurrent deviations

n = number of pairs of deviations

We will continue our learning of the covariance vs correlation differences with these applications of the correlation matrix.

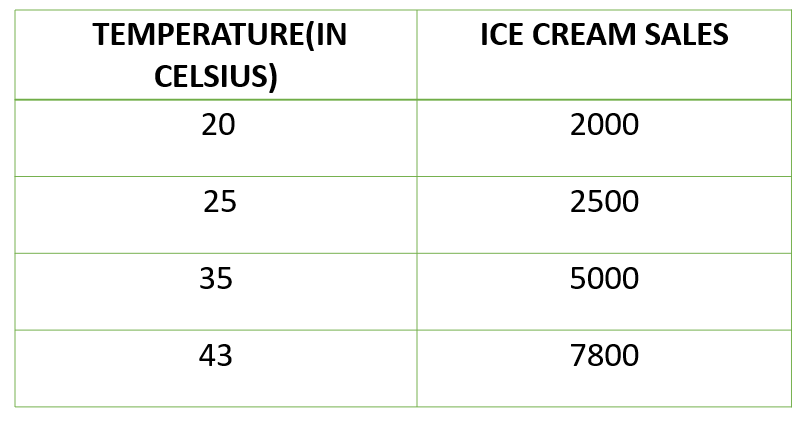
## Applications of Correlation

A correlation matrix is computed for three main reasons:

* The goal when dealing with large amounts of data is to find patterns. As a result, a correlation matrix is used to look for a pattern in the data and determine whether the variables are highly correlated.
* For use in other analyses. When excluding missing values pairwise, correlation matrices are commonly used as inputs for exploratory factor analysis, confirmatory factor analysis, structural equation models, and [linear regression.](https://www.simplilearn.com/tutorials/machine-learning-tutorial/linear-regression-in-python)
* When checking other analyses, as a diagnostic. When it comes to linear regression, for example, a large number of correlations indicate that the linear regression estimates will be unreliable.

1. **Differentiate between univariate ,Biavariate,and multivariate analysis.**
   1. **Univariate data –**  
      This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

Suppose that the heights of seven students of a class is recorded(figure 1),there is only one variable that is height and it is not dealing with any cause or relationship. The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

* 1. ** Bivariate data –**  
     This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.Example of bivariate data can be temperature and ice cream sales in summer season.

Suppose the temperature and ice cream sales are the two variables of a bivariate data(figure 2). Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase. Thus bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

* 1. **Multivariate data –**  
     When the data involves **three or more variables**, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved.Some of the techniques are regression analysis,path analysis,factor analysis and multivariate analysis of variance (MANOVA).

1. **What do you understand by sensitivity and how would you calculate it?**

The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as **sensitive analysis**. It’s usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond’s price.

It is also known as the what – if analysis. Sensitivity analysis can be used for any activity or system. All from planning a family vacation with the variables in mind to the decisions at corporate levels can be done through sensitivity analysis.**Apply sensitivity analysis in investments**

Sensitivity analysis works on the simple principle: **Change the model and observe the behavior.**

**The parameters that one needs to note while doing the above are:**

***A) Experimental design:*** It includes combination of parameters that are to be varied. This includes a check on which and how many parameters need to vary at a given point in time, assigning values (maximum and minimum levels) before the experiment, study the correlations: positive or negative and accordingly assign values for the combination.

***B) What tfo vary:***The different parameters that can be chosen to vary in the model could be:  
a) the number of activities  
b) the objective in relation to the risk assumed and the profits expected  
c) technical parameters  
d) number of constraints and its limits

***C) What to observe:***  
a) the value of the objective as per the strategy  
b) value of the decision variables  
c) value of the objective function between two strategies adopted

Measurement of sensitivity analysis

Below are mentioned the steps used to conduct sensitivity analysis:

1. Firstly the base case output is defined; say the NPV at a particular base case input value (V1) for which the sensitivity is to be measured. All the other inputs of the model  are kept constant.
2. Then the value of the output at a new value of the input (V2) while keeping other inputs constant is calculated.
3. Find the percentage change in the output and the percentage change in the input.
4. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

This process of testing sensitivity for another input (say cash flows growth rate) while keeping the rest of inputs constant is repeated until the sensitivity figure for each of the inputs is obtained. The conclusion would be that the higher the sensitivity figure, the more sensitive the output is to any change in that input and vice versa.

Methods of Sensitivity Analysis

There are different methods to carry out the sensitivity analysis:

* Modeling and simulation techniques
* Scenario management tools through Microsoft excel

There are mainly two approaches to analyzing sensitivity:

* Local Sensitivity Analysis
* Global Sensitivity Analysis

**Local sensitivity analysis** is derivative based (numerical or analytical). The term local indicates that the derivatives are taken at a single point. This method is apt for simple cost functions, but not feasible for complex models, like models with discontinuities do not always have derivatives.

Mathematically, the sensitivity of the cost function with respect to certain parameters is equal to the partial derivative of the cost function with respect to those parameters.

Local sensitivity analysis is a *one-at-a-time* (OAT) technique that analyzes the impact of one parameter on the cost function at a time, keeping the other parameters fixed.

**Global sensitivity analysis** is the second approach to sensitivity analysis, often implemented using Monte Carlo techniques. This approach uses a global set of samples to explore the design space.

**The various techniques widely applied include:**

* ***Differential sensitivity analysis:*** It is also referred to the direct method. It involves solving simple partial derivatives to temporal sensitivity analysis. Although this method is computationally efficient, solving equations is intensive task to handle.
* ***One at a time sensitivity measures:*** It is the most fundamental method with partial differentiation, in which varying parameters values are taken one at a time. It is also called as local analysis as it is an indicator only for the addressed point estimates and not the entire distribution.
* ***Factorial Analysis:*** It involves the selection of given number of samples for a specific parameter and then running the model for the combinations. The outcome is then used to carry out parameter sensitivity.

Through the sensitivity index one can calculate the output % difference when one input parameter varies from minimum to maximum value.

* ***Correlation analysis*** helps in defining the relation between independent and dependent variables.
* ***Regression analysis*** is a comprehensive method used to get responses for complex models.
* ***Subjective sensitivity analysis:*** In this method the individual parameters are analyzed. This is a subjective method, simple, qualitative and an easy method to rule out input parameters.

1. **What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?**

Hypothesis testing in statistics refers to analyzing an assumption about a population parameter. It is used to make an educated guess about an assumption using statistics. With the use of sample data, hypothesis testing makes an assumption about how true the assumption is for the entire population from where the sample is being taken.

Any hypothetical statement we make may or may not be valid, and it is then our responsibility to provide evidence for its possibility. To approach any hypothesis, we follow these four simple steps that test its validity.

1. First, we formulate two hypothetical statements such that only one of them is true. By doing so, we can check the validity of our own hypothesis.
2. The next step is to formulate the statistical analysis to be followed based upon the data points.
3. Then we analyze the given data using our methodology.
4. The final step is to analyze the result and judge whether the null hypothesis will be rejected or is true.

### Let’s look at several hypothesis testing examples:

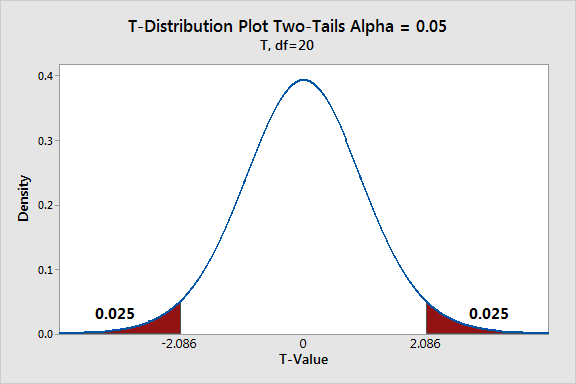
* It is observed that the average recovery time for a knee-surgery patient is 8 weeks. A physician believes that after successful knee surgery if the patient goes for physical therapy twice a week rather than thrice a week, the recovery period will be longer. Conduct hypothesis for this statement.
* David is a ten-year-old who finishes a 25-yard freestyle in the meantime of 16.43 seconds. David’s father bought goggles for his son, believing that it would help him to reduce his time. He then recorded a total of fifteen 25-yard freestyle for David, and the average time came out to be 16 seconds. Conduct a hypothesis.
* A tire company claims their A-segment of tires have a running life of 50,000 miles before they need to be replaced, and previous studies show a standard deviation of 8,000 miles. After surveying a total of 28 tires, the mean run time came to be 46,500 miles with a standard deviation of 9800 miles. Is the claim made by the tire company consistent with the given data? Conduct hypothesis testing.

H0 and H1:

Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. In our case it will be the average height of students in the batch is 100. Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed.

Two-Tailed Hypothesis Tests

Two-tailed hypothesis tests are also known as nondirectional and two-sided tests because you can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution. In the example below, I use an alpha of 5% and the distribution has two shaded regions of 2.5% (2 \* 2.5% = 5%).



When a test statistic falls in either critical region, your sample data are sufficiently incompatible with the null hypothesis that you can reject it for the population.

In a two-tailed test, the generic null and alternative hypotheses are the following:

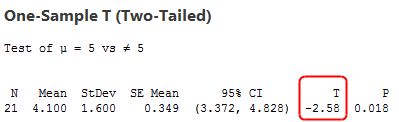
* **Null**: The effect equals zero.
* **Alternative**:  The effect does not equal zero.

The specifics of the hypotheses depend on the type of test you perform because you might be assessing means, proportions, or rates.

Example of a two-tailed 1-sample t-test

Suppose we perform a two-sided 1-sample t-test where we compare the mean strength (4.1) of parts from a supplier to a target value (5). We use a two-tailed test because we care whether the mean is greater than or less than the target value.

To interpret the results, simply compare the [p-value](https://statisticsbyjim.com/glossary/p-value/) to your significance level. If the p-value is less than the significance level, you know that the test statistic fell into one of the critical regions, but which one? Just look at the estimated effect. In the output below, the t-value is negative, so we know that the test statistic fell in the critical region in the left tail of the distribution, indicating the mean is less than the target value. Now we know this difference is statistically significant.



We can conclude that the population mean for part strength is less than the target value. However, the test had the capacity to detect a positive difference as well. You can also assess the confidence interval. With a two-tailed hypothesis test, you’ll obtain a two-sided confidence interval. The confidence interval tells us that the population mean is likely to fall between 3.372 and 4.828. This range excludes the target value (5), which is another indicator of significance.

Advantages of two-tailed hypothesis tests

You can detect both positive and negative effects. Two-tailed tests are standard in scientific research where discovering any type of effect is usually of interest to researchers.

1. **What is quantitative data and qualitative data?**

Quantitative Data:

Quantitative data is **data expressing a certain quantity, amount or range**. Usually, there are measurement units associated with the data, e.g. metres, in the case of the height of a person. It makes sense to set boundary limits to such data, and it is also meaningful to apply arithmetic operations to the data.

Qualitative Data:

What is Qualitative Data? Qualitative data is the **descriptive and conceptual findings collected through questionnaires, interviews, or observation**. Analyzing qualitative data allows us to explore ideas and further explain quantitative results.

1. **How to calculate range and interquartile range?**

To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). The range only takes into account these two values and ignore the data points between the two extremities of the distribution. It's used as a supplement to other measures, but it is rarely used as the sole measure of dispersion because it’s sensitive to extreme values.

The interquartile range and semi-interquartile range give a better idea of the dispersion of data. To calculate these two measures, you need to know the values of the lower and upper quartiles. The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range.

When the data set is small, it is simple to identify the values of quartiles. Let’s look at an example.

**Example 1 – Range and interquartile range of a data set**

Find the quartiles of this data set: 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, 36.

You first need to arrange the data points in increasing order. As you do so, you can give them a rank to indicate their position in the data set. Rank 1 is the data point with the smallest value, rank 2 is the data point with the second-lowest value, etc.

|  |  |
| --- | --- |
| **Table 4.5.1.1 Rank of data points Table summary This table displays the results of Rank of data points. The information is grouped by Rank (appearing as row headers), Value (appearing as column headers).** | |
| **Rank** | **Value** |
| 1 | 6 |
| 2 | 7 |
| 3 | 15 |
| 4 | 36 |
| 5 | 39 |
| 6 | 41 |
| 7 | 41 |
| 8 | 43 |
| 9 | 43 |
| 10 | 47 |
| 11 | 49 |

Then you need to find the rank of the median to split the data set in two. As we have seen in the section on the median, if the number of data points is an uneven value, the rank of the [median](https://www150.statcan.gc.ca/edu/power-pouvoir/ch11/median-mediane/5214872-eng.htm) will be

(n + 1) ÷ 2 = (11 + 1) ÷ 2 = 6

The rank of the median is 6, which means there are five points on each side.

Then you need to split the lower half of the data in two again to find the lower quartile. The lower quartile will be the point of rank (5 + 1) ÷ 2 = 3. The result is Q1 = 15. The second half must also be split in two to find the value of the upper quartile. The rank of the upper quartile will be 6 + 3 = 9. So Q3 = 43.

Once you have the quartiles, you can easily measure the spread. The interquartile range will be Q3 - Q1, which gives 28 (43-15). The semi-interquartile range is 14 (28 ÷ 2) and the range is 43 (49-6).

For larger data sets, you can use the cumulative relative frequency distribution to help identify the quartiles or, even better, the basic statistics functions available in a spreadsheet or statistical software that give results more easily.

What happens when the data set includes a data point whose value is considered extreme compared to the rest of the distribution?

**Example 2 – Range and interquartile range in presence of an extreme value**

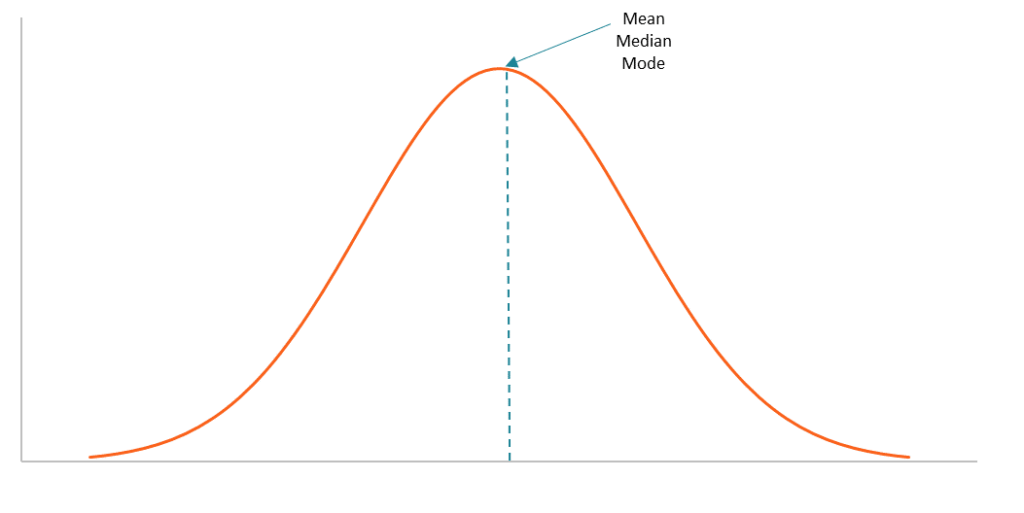
Find the range and interquartile range of the data set of example 1, to which a data point of value 75 was added.

The range would now be 69 (75-6). The median would be the mean of the values of the data point of rank 12 ÷ 2 = 6 and the data point of rank (12 ÷ 2) + 1 = 7. Because it falls between ranks 6 and 7, there are six data points on each side of the median. The lower quartile is the mean of the values of the data point of rank 6 ÷ 2 = 3 and the data points of rank (6 ÷ 2) + 1 = 4. The result is (15 + 36) ÷ 2 = 25.5. The upper quartile is the mean of the values of data point of rank 6 + 3 = 9 and the data point of rank 6 + 4 = 10, which is (43 + 47) ÷ 2 = 45. The interquartile range is 45 - 25.5 = 19.5.

In summary, the range went from 43 to 69, an increase of 26 compared to example 1, just because of a single extreme value. The more robust interquartile range went from 28 to 19.5, a decrease of only 8.5.

The second example demonstrated that the interquartile range is more robust than the range when the data set includes a value considered extreme. It’s not a perfect measure, though. In this example, we might have expected that when adding an extreme value, the measure of dispersion would increase, but the opposite happened because there was a great difference between the values of data points of ranks 3 and 4.

1. **What do you understand by bell curve distribution ?**



A bell curve is the informal name of a graph that depicts a normal probability distribution. The term obtained its name due to the bell-shaped curve of the normal probability distribution graph.

However, the term is not quite correct because the normal probability distribution is not the only probability distribution whose graph shows a bell-shaped curve. For example, the graphs of the Cauchy and logistic distributions also demonstrate a bell-shaped curve.

**Characteristics of a Bell Curve**

The bell curve is perfectly symmetrical. It is concentrated around the peak and decreases on either side. In a bell curve, the peak represents the most probable event in the dataset while the other events are equally distributed around the peak. The peak of the curve corresponds to the mean of the dataset (note that the mean in a normal probability distribution also equals the [median](https://corporatefinanceinstitute.com/resources/knowledge/other/median/) and the mode).

The dispersion of the data on the bell curve is measured by the [standard deviation](https://corporatefinanceinstitute.com/resources/knowledge/standard-deviation/). The probabilities of the bell curve and the standard deviation share a few important relationships, including:

* Around 68% of the data lies within 1 standard deviation.
* Around 95% of the data lies within 2 standard deviations.
* Around 99.7% of the data lies within 3 standard deviations.

The relationships described above are known as the 68-95-99.7 rule or the empirical rule. The empirical rule is primarily used to calculate the [confidence interval](https://corporatefinanceinstitute.com/resources/knowledge/finance/confidence-interval/) of a normal probability distribution.

The concept is extremely important in [statistics](https://www.originlab.com/index.aspx?go=Products/Origin/Statistics) due to the wide applications of the normal probability distribution. For instance, the normal probability distribution is used as a representation of the distribution of random variables whose real distribution is unknown.

1. **Mention one method to find outliers**

Outliers are values at the extreme ends of a dataset.

Some outliers represent true values from natural variation in the population. Other outliers may result from incorrect data entry, equipment malfunctions, or other [measurement errors](https://www.scribbr.com/methodology/random-vs-systematic-error/).

An outlier isn’t always a form of dirty or incorrect data, so you have to be careful with them in [data cleansing](https://www.scribbr.com/methodology/data-cleansing/). What you should do with an outlier depends on its most likely cause.

You can choose from several methods to detect outliers depending on your time and resources.

**Sorting method**

You can **sort** [quantitative variables](https://www.scribbr.com/methodology/types-of-variables/#quantitative-variables) from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.

This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

Example: Sorting methodYour dataset for a pilot experiment consists of 8 values.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 180 | 156 | 9 | 176 | 163 | 1827 | 166 | 171 |

You sort the values from low to high and scan for extreme values.

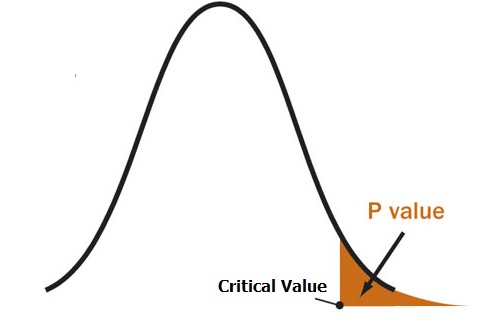
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **9** | 156 | 163 | 166 | 171 | 176 | 180 | **1872** |

1. **What is p-value in hypothesis testing?**

A p value is used in [hypothesis testing](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/) to help you [support or reject the null hypothesis](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/support-or-reject-null-hypothesis/). The p value is the evidence **against** a [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/). The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you [convert them to a percentage](http://wagner.nyu.edu/files/students/Math_Review_-_Review_topics_-_Percents.pdf). For example, a p value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That’s pretty tiny. On the other hand, a large p-value of .9(90%) means your results have a 90% probability of being completely random and not due to anything in your experiment. Therefore, the smaller the p-value, the more important (“[significant](https://www.statisticshowto.com/what-is-statistical-significance/)“) your results.

When you run a [hypothesis test](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/), you compare the p value from your test to the [alpha level](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-an-alpha-level/)you selected when you ran the test. Alpha levels can also be written as percentages.

[](https://www.statisticshowto.com/wp-content/uploads/2014/01/p-value1.jpg)Graphically, the p value is the area in the **tail** of a [probability distribution](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/probability-distribution/). It’s calculated when you run hypothesis test and is the area to the right of the [test statistic](https://www.statisticshowto.com/test-statistic/) (if you’re running a [two-tailed test](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/), it’s the area to the left and to the right).

1. **What is the Binomial Probability Formula?**

The binomial distribution formula is for any random variable X, given by;  P(x:n,p) = nC*x* px(1-p)n-x **Or** P(x:n,p) = nCx px (q)n-x

where,

* n = the number of experiments
* x = 0, 1, 2, 3, 4, …
* p = Probability of success in a single experiment
* q = Probability of failure in a single experiment (= 1 – p)

The binomial distribution formula is also written in the form of n-Bernoulli trials, where nCx = n!/x!(n-x)!. Hence, P(x:n,p) = n!/[x!(n-x)!].px.(q)n-x

1. **Explain ANOVA and it’s applications.**

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

## The Formula for ANOVA is:

​ **F=**

**where:**

F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error​

## Example of How to Use ANOVA

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample [t-tests](https://www.investopedia.com/terms/t/t-test.asp). However, it results in fewer [type I errors](https://www.investopedia.com/terms/t/type_1_error.asp) and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.