

Statistics For Data Science

What is Data?

According to the Oxford “Data is distinct pieces of information, usually formatted in a special way”.

Data is measured, collected and reported, and analysed, whereupon it is often visualized using graphs, images or other analysis tools. Raw data may be a collection of numbers or characters before it's been cleaned and corrected by researchers.

What is data science?

Data science is the study of data. It involves developing methods of recording, storing, and analysing data to effectively extract useful information. The goal of data science is to gain insights and knowledge from any type of data - both structured and unstructured.

Its simple definition is Collect the information, make an analysis and take the Decision for the growth of the business.

Why Data Scientist?

Data scientists straddle the world of both business and IT and possess unique skill sets. Their role has assumed significance thanks to how businesses today think of big data. Business wants to make use of the unstructured data which can boost their revenue. Data scientists analyse this information to make sense of it and bring out business insights that will aid in the growth of the business.

Analytics Project Life Cycle:

The following are the broadly divided steps which are usually performed in any Predictive analytics / Machine Learning problem.

1. Understanding the problem
2. Collecting and Reading data
3. Exploratory Data Analysis
4. Data cleaning/Data pre-processing
5. Data transformation
6. Data partition
7. Selecting few models
8. Cross validation for all chosen models
9. Evaluation of all models and selecting the best model
10. Predictions on unknown or unseen data

1. Understanding the Problem:

Before you solve a problem, you have to define the problem. You'll often get ambiguous inputs from the people who are your clients. You'll have to develop the intuition to translate scarce inputs in to actionable outputs and to ask the questions that nobody else is asking.

If you are solving a problem for a VP of sales in a company. You should start understanding by their goals and you need to work with them to clearly define the problem.

To define the problem, you need to ask the right questions.

1. Who are the customers?
2. Why are they buying our products?
3. How do we predict whether a customer is going to buy our product or not?
4. What is different from segments who are performing well and those that are performing below expectations?
5. How much money we may lose if we don't actively sell the products to these groups?

2. Collecting & Reading data:

The first step is to get the data. Once you get the data you need to read the data into Machine Learning tools like R, python and so on. One should check the format of the data before reading. There can be many formats of the data. Commonly used and easiest format of data is csv format. Sometimes you need to do ETL (Extract, Transform and Load).

Simplest way to begin Machine Learning is to get the data in csv format. Data can be collected from repositories which are available free and publicly.

3. Exploratory Data Analysis:

Before proceeding towards doing anything related to data, one should clearly and precisely know about the problem and the questions which are required to be answered through Machine Learning. Only then one can be certain about the results which the Machine Learning algorithm is going to give. In the csv format, data is in the form of tables which have rows and columns. One row belongs to one observation or record & one column belongs to one variable. Variable can be independent or dependent variable. So, one of the columns belongs to dependent variable, also known as target variable. One should check the meanings of each of the variables before going ahead.

One should explore the data. There are many ways to explore the data including data visualization. This will help you get more insights on the problem and also it will help you to get intuition on how you can get better results from Machine Learning. It can tell you which variables are important and it can also tell you which data columns or rows have missing values. One can also find the patterns, if any in the data.

4. Data Cleaning / Data pre-processing:

We need to find the missing values like NA, NAN, blanks, etc and then impute (or fill) them with something like average of non-missing values in the columns. We

also need to remove the unnecessary columns and/or rows. One should do data exploration before doing any data cleaning blindly.

5. Data transformation:

For numerical data we may require centering, scaling or normalization like log-normalization, etc in order to avoid issues like overfitting. We may also require dimensionality reduction techniques like Principal component analysis to remove dimensionality issues. We may require one-hot encoding if we have categorical data.

6. Data partition:

We need to split the dataset into a training (known) set and testing (unknown) dataset. We need “test data set” in order to validate the model or check the model performance on unseen or test dataset.

7. Selecting few models:

Based on your intuition and your experience you can choose few models from list of machine learning models. They may or may not work so you need to choose different model then or you may need to tune the model. There are several models for different needs. For classification problems we have models like logistic regression, decision tree, random forest, etc and for regression problems we have models like linear regression, Gradient Descent, neural networks, lasso, etc.

8. Cross validation for all chosen models:

Model is fitted on training or known data. One must do the cross-validation & model tuning before making any conclusions about the results. Cross-validation is done to issues like over fitting and model tuning is done to get the best model parameters which can give best required results. Once you have chosen the models, then you can perform model tuning and cross-validation for each of the chosen models. Cross-validation is like repeatedly checking the model performance on unknown dataset and thereby increasing the assurance of the model performance on any data set which will be fed into this model in future.

9. Evaluation of all models and selecting the best model:

Once the model is fitted on the training data, it is used to predict the target or dependent variable for the test data. The predicted value of the target is then compared with the actual target values of the test data set. The accuracy of the model is the percentage of correct predictions which are made. There are several evaluation metrics like R², RMSE, Accuracy score, F1 score, AUC, log loss and so on. Depending on the requirement you can choose evaluation metric and then calculate it for each of the models.

Then you choose the model which has performed best in the evaluation. With this chosen model, you can then train this model on the training data set again.

10. Predictions on new data by deployment:

And now you ready to get the final predictions for the data which is unseen data.

Types of Data:

Generally, data can be classified into two parts:

1. Categorical Data:

Categorical data refers to a data type that can be stored and identified based on the names or labels given to them. A process called matching is done, to draw out the similarities or relations between the data and then they are grouped accordingly.

The data collected in the categorical form is also known as qualitative data. Each dataset can be grouped and labelled depending on their matching qualities, under only one category

Example: Marital Status, Political Party, Eye colour

2. Numerical Data:

Numerical data can further be classified into two categories:

Discrete Data:

Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers.

We speak of discrete data if the data can only take on certain values. This type of data **can't be measured but it can be counted**. It basically represents information that can be categorized into a classification.

Example: Number of students in the class, Mobiles, Brothers, etc.

Continuous Data:

Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range.

Continuous Data represents measurements and therefore their values **can't be counted but they can be measured**.

Example: Temperature, Time, Age, Water, Currency, etc.

At advanced level, we can further classify the data into four parts

1. Nominal Data:

Nominal data are used to label variables where there is no quantitative value and has no order. So, if you change the order of the value then the meaning will remain the same.

Thus, nominal data are observed but not measured, are unordered but non-equidistant, and have no meaningful zero.

The only numerical activities you can perform on nominal data is to state that perception is (or isn't) equivalent to another (equity or inequity), and you can use this data to amass them. You can't organize nominal data, so you can't sort them.

Neither would you be able to do any numerical tasks as they are saved for numerical data. With nominal data, you can calculate frequencies, proportions, percentages, and central points.

Examples: Language, Gender, Location, Education, Yes/No, etc.

2. Ordinal Data:

Ordinal data is almost the same as nominal data but not in the case of order as their categories can be ordered like 1st, 2nd, etc. However, there is no continuity in the relative distances between adjacent categories.

Ordinal Data is observed but not measured, is ordered but non-equidistant, and has no meaningful zero. Ordinal scales are always used for measuring happiness, satisfaction, etc.

As ordinal data are ordered, they can be arranged by making basic comparisons between the categories, for example, greater or less than, higher or lower, and so on. You can't do any numerical activities with ordinal data, however, as they are numerical data. Examples: Excellent to poor, Grades, Opinion, etc.

3. Interval Data:

Interval values represent **ordered units that have the same difference**. Therefore, we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values.

The problem with interval values data is that they **don't have a true zero**. That means in regards to our example, that there is no such thing as no temperature. With interval data, we can add and subtract, but we cannot multiply, divide or calculate ratios. Because there is no true zero, a lot of descriptive and inferential statistics can't be applied. **Example:** Body temperature.

4. Ratio Data:

Ratio values are also ordered units that have the same difference. Ratio values are **the same as interval values, with the difference that they do have an absolute zero**. Examples: height, weight, length etc.

5. Special Numeric Data:

Phone number, Pin codes, Aadhar number, Emp id, Transaction number, Order number.

Statistics:

Statistics is a branch of mathematics that deals with the study of collecting, analysing, interpreting, presenting, and organizing data in a particular manner. Statistics is defined as the process of collection of data, classifying data, representing the data for easy interpretation, and further analysis of data. Statistics also is referred to as arriving at conclusions from the sample data that is collected using surveys or experiments. Different sectors such as psychology, sociology, geology, probability, and so on also use statistics to function.

Mathematical Statistics:

Statistics is used mainly to gain an understanding of the data and focus on various applications. Statistics is the process of collecting data, evaluating data, and summarizing it into a mathematical form. Initially, statistics were related to the science of the state where it was used in the collection and analysis of facts and data about a country such as its economy, population, etc. Mathematical statistics applies mathematical techniques like linear algebra, differential equations, mathematical analysis, and theories of probability.

There are two methods of analysing data in mathematical statistics that are used on a large scale:

Descriptive Statistics:

The descriptive method of statistics is used to describe the data collected and summarize the data and its properties using the measures of central tendencies and the measures of dispersion.

Inferential Statistics:

This method of statistics is used to draw conclusions from the data. Inferential statistics requires statistical tests performed on samples, and it draws conclusions by identifying the differences between the 2 groups. Tests calculate the p-value that is compared with the probability of chance(α) = 0.05. If the p-value is less than α , then it is concluded that the p-value is statistically significant.

| Descriptive Statistics | Inferential Statistics |
|---|---|
| Descriptive statistics are used to quantify the characteristics of the data. | Inferential statistics are used to make conclusions about the population by using analytical tools on the sample data |
| Measures of central tendency and measures of dispersion are the important tools used. | Hypothesis testing and regression analysis are the analytical tools used. |
| It is used to describe the characteristics of a known sample or population. | It is used to make inferences about an unknown population |
| Measures of descriptive statistics are variance, range, mean, median, etc. | Measures of inferential statistics are t-test, z test, linear regression, etc. |

Descriptive Statistics:

The study of numerical and graphical ways to describe and display your data is called descriptive statistics. It describes the data and helps us understand the features of the data by summarizing the given sample set or population of data. In descriptive statistics, we usually take the sample into account.

Statisticians use graphical representation of data to get a clear picture of the data. Business trends can be analysed easily with these representations. visual representation is more effective than presenting huge numbers.

We can describe these data in various dimensions. Various dimensions of describing data are

1. Central Tendency of Data
2. Dispersion of Data
3. Shape of the Data

1. Central Tendency of Data:

This is the centre of the distribution of data. It describes the location of data and concentrates where the data is located.

The three most widely used measures of the “centre” of the data are

1.1 Mean

The “Mean” is the average of the data.

Average can be identified by summing up all the numbers and then dividing them by the number of observations.

Mean = $X_1 + X_2 + X_3 + \dots + X_n / n$

Example: Data – 10,20,30,40,50 and Number of observations = 5

Mean = $[10+20+30+40+50] / 5$

Mean = 30

Outliers influence the central tendency of the data.

1.2 Median

Median is the 50%th percentile of the data. It is exactly the centre point of the data.

Median can be identified by ordering the data and splits the data into two equal parts and find the number. It is the best way to find the centre of the data.

Because the central tendency of the data is not affected by outliers. Outliers don't influence the data.

Example: Odd number of Data – 10,20,30,40,50

Median is 30.

Even number of data – 10,20,30,40,50,60

Find the middle 2 data and take the mean of those two values.

Here 30 and 40 are middle values.

$= 30+40 / 2$

$= 35$. Median is 35

1.3 Mode

Mode is frequently occurring data or elements.

If an element occurs the highest number of times, it is the mode of that data. If no number in the data is repeated, then there is no mode for that data. There can be more than one mode in a dataset if two values have the same frequency and also the highest frequency.

Outliers don't influence the data.

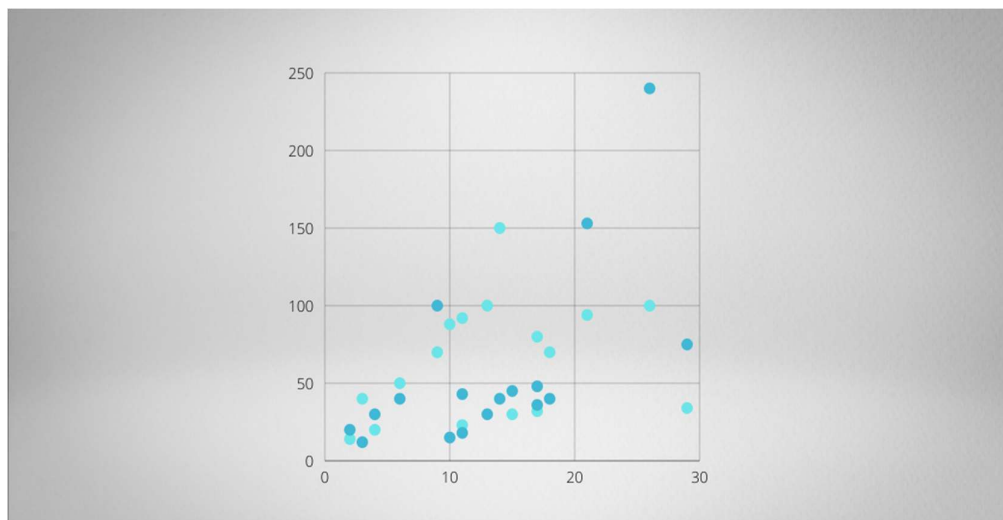
The mode can be calculated for both quantitative and qualitative data.

Example: Data – 1,3,4,6,7,3,3,5,10, 3

Mode is 3

because 3 has the highest frequency (4 times)

2. Dispersion of Data:



The dispersion is the “Spread of the data”. It measures how far the data is spread. In most of the dataset, the data values are closely located near the mean. On some other dataset, the values are widely spread out of the mean. These dispersions of data can be measured by

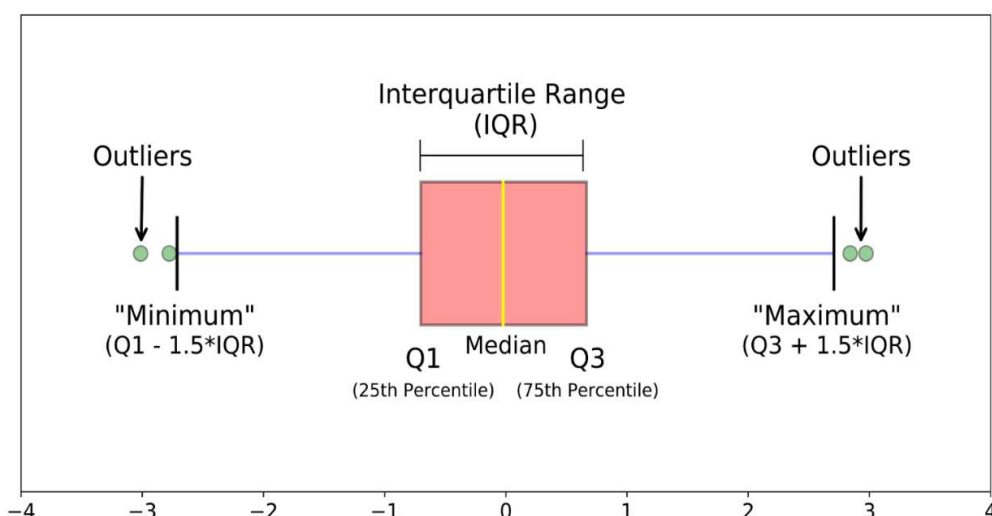
2.1 Inter Quartile Range (IQR)

Quartiles are special percentiles.

1st Quartile Q1 is the same as the 25th percentile.

2nd Quartile Q2 is the same as 50th percentile.

3rd Quartile Q3 is same as 75th percentile



Steps to find quartile and percentile:

- The data should sort and ordered from the smallest to the largest.
- For Quartiles, ordered data is divided into 4 equal parts.
- For Percentiles, ordered data is divided into 100 equal parts.

Inter Quartile Range is the difference between the third quartile(Q3) and the first Quartile (Q1)

$$\text{IQR} = Q3 - Q1$$

Inter Quartile range

It is the spread of the middle half (50%) of the data

2.2 Range

The range is the difference between the largest and the smallest value in the data.

$$\text{Max} - \text{Min} = \text{Range}$$

2.3 Standard Deviation

The most common measure of spread is the standard deviation.

The Standard deviation is the measure of how far the data deviates from the mean value.

The standard deviation formula varies for population and sample. Both formulas are similar, but not the same.

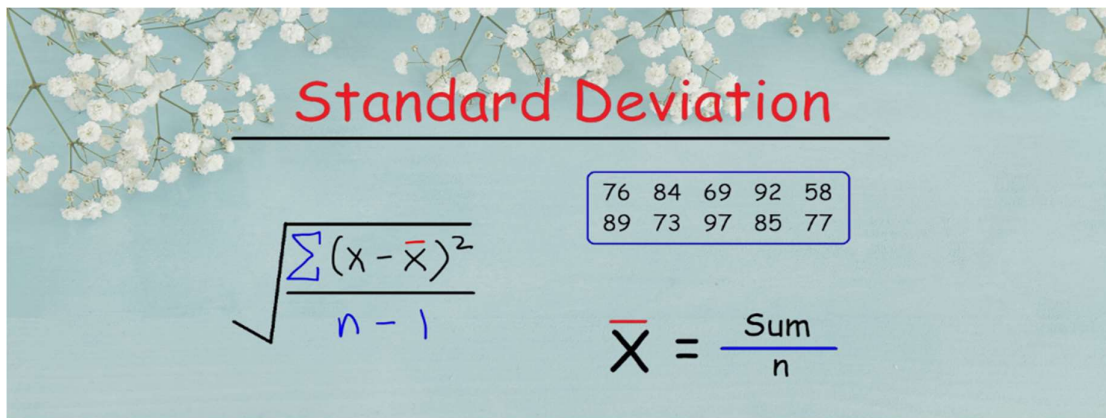
- Symbol used for Sample Standard Deviation – “s” (lowercase)
- Symbol used for Population Standard Deviation – “ σ ” (sigma, lower case)

Steps to find Standard deviation:

If x is a number, then the difference “x – mean” is its deviation. The deviations are used to calculate the standard deviation.

Sample Standard Deviation, s = Square root of sample variance

Sample Standard Deviation, s = Square root of $[\sum (x - \bar{x})^2 / n - 1]$ where \bar{x} is average and n is no. of samples



Standard Deviation

$$\sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

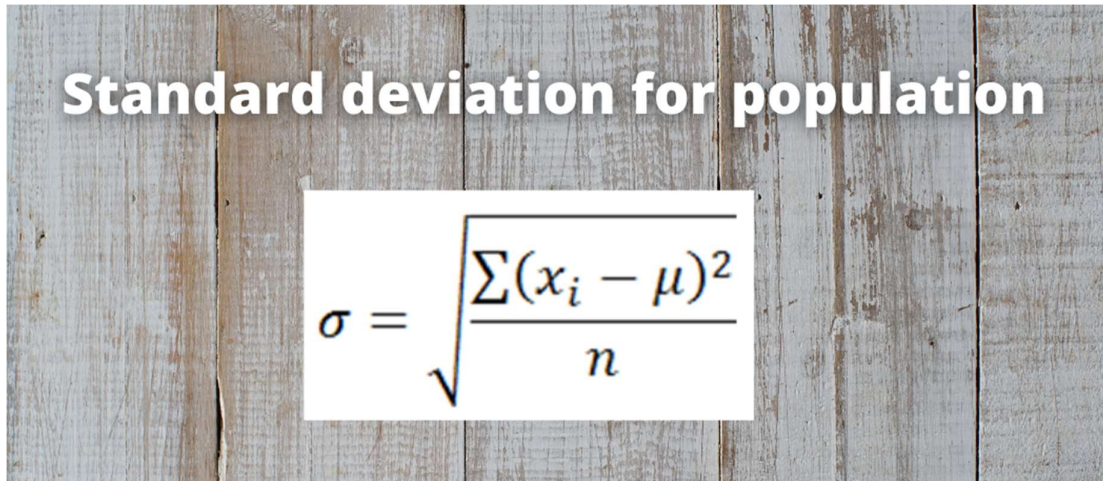
| | | | | |
|----|----|----|----|----|
| 76 | 84 | 69 | 92 | 58 |
| 89 | 73 | 97 | 85 | 77 |

$$\bar{x} = \frac{\text{Sum}}{n}$$

Standard Deviation for sample

Population Standard Deviation, σ = Square root of population variance

Population Standard Deviation, σ = Square root of $[\sum (x - \mu)^2 / N]$ where μ is Mean and N is no. of population.



Standard deviation for population

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

The standard deviation for population

The standard deviation is always positive or zero. It will be large when the data values are spread out from the mean.

2.4 Variance

The variance is a measure of variability. It is the average squared deviation from the mean.

The symbol σ^2 represents the population variance and the symbol for s^2 represents sample variance.

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

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

3. Shape of the Data:

The shape describes the type of the graph.

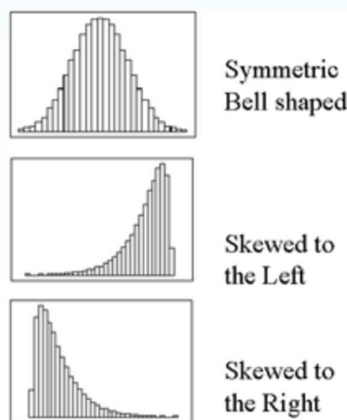
The shape of the data is important because making a decision about the probability of data is based on its shape.

The shape of the data can be measured by two methodologies.

3.1 Symmetric:

In the symmetric shape of the graph, the data is distributed the same on both sides.

In symmetric data, the mean and median are located close together.



The curve formed by this symmetric graph is called a normal curve.

3.2 Skewness:

Skewness is the measure of the asymmetry of the distribution of data.

The data is not symmetrical (i.e.) it is skewed towards one side.

Skewness is classified into two types.

1. Positively skewed

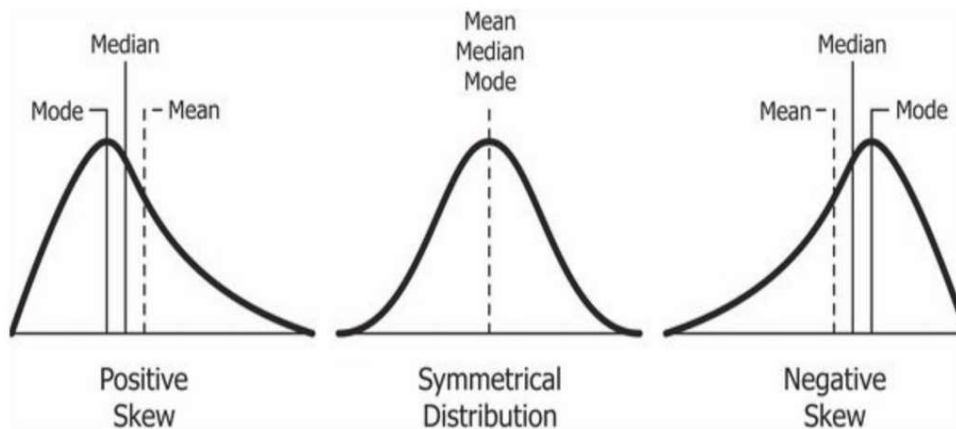
In a Positively skewed distribution, the data values are clustered around the left side of the distribution and the right side is longer.

The mean and median will be greater than the mode in the positive skew.

2. Negatively skewed

In a Negatively skewed distribution, the data values are clustered around the right side of the distribution and the left side is longer.

The mean and median will be less than the mode in the negative skew.



3.3 Kurtosis

Kurtosis is the measure of describing the distribution of data.

This data is distributed in different ways. They are,

1. Platykurtic

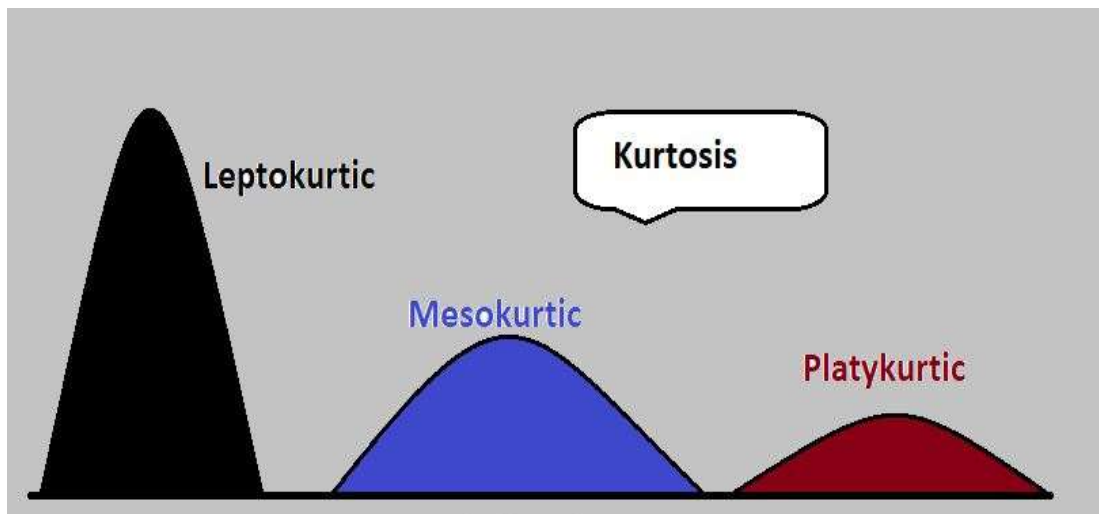
The platykurtic shows a distribution with flat tails. Here the data is distributed flatly. The flat tails indicated the small outliers in the distribution.

2. Mesokurtic

In Mesokurtic, the data is widely distributed. It is normally distributed and it also matches normal distribution.

3. Leptokurtic

In leptokurtic, the data is very closely distributed. The height of the peak is greater than width of the peak.



Scatter Plot:

A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

Scatter plots' primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Scatter plot Correlation:

We know that the correlation is a statistical measure of the relationship between the two variables' relative movements. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the closer the points will touch the line. This cause examination tool is considered as one of the seven essential quality tools.

Types of correlation:

The scatter plot explains the correlation between two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

Positive Correlation:

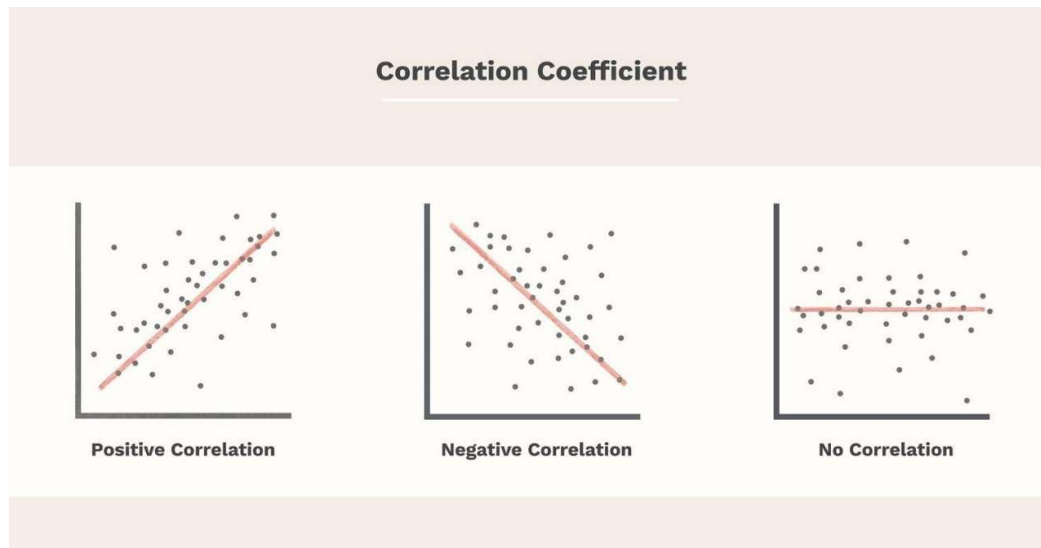
When the points in the graph are rising, moving from left to right, then the scatter plot shows a positive correlation. It means the values of one variable are increasing with respect to another.

Negative Correlation:

When the points in the scatter graph fall while moving left to right, then it is called a negative correlation. It means the values of one variable are decreasing with respect to another.

No Correlation:

When the points are scattered all over the graph and it is difficult to conclude whether the values are increasing or decreasing, then there is no correlation between the variables.



Inferential Statistics:

Inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data. Apart from inferential statistics, descriptive statistics forms another branch of statistics. Inferential statistics help to draw conclusions about the population while descriptive statistics summarizes the features of the data set.

There are two main types of inferential statistics - hypothesis testing and regression analysis. The samples chosen in inferential statistics need to be representative of the entire population. In this article, we will learn more about inferential statistics, its types, examples, and see the important formulas.

What are Inferential Statistics?

Inferential statistics helps to develop a good understanding of the population data by analysing the samples obtained from it. It helps in making generalizations about the population by using various analytical tests and tools. In order to pick out random samples that will represent the population accurately many sampling techniques are used. Some of the important methods are simple random sampling, stratified sampling, cluster sampling, and systematic sampling techniques.

Inferential Statistics Definition:

Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. The goal of inferential statistics is to make generalizations about a population. In inferential statistics, a statistic is taken from the sample data (e.g., the sample mean) that used to make inferences about the population parameter (e.g., the population mean).

Inferential Statistics Examples:

Inferential statistics is very useful and cost-effective as it can make inferences about the population without collecting the complete data. Some inferential statistics examples are given below:

- Suppose the mean marks of 100 students in a particular country are known. Using this sample information, the mean marks of students in the country can be approximated using inferential statistics.
- Suppose a coach wants to find out how many average cartwheels sophomores at his college can do without stopping. A sample of a few students will be asked to perform cartwheels and the average will be calculated. Inferential statistics will use this data to make a conclusion regarding how many cartwheel sophomores can perform on average.

Why do we need Inferential Statistics?

In contrast to Descriptive Statistics, rather than having access to the whole population, we often have a limited amount of data.

In such cases, Inferential Statistics come into action. For example, we might be interested in finding the average of the entire school's exam marks. It is not reasonable because we might find it impracticable to get the data we need. So, rather than getting

the entire school's exam marks, we measure a smaller sample of students (for example, a sample of 50 students). This sample of 50 students will now describe the complete population of all students of that school.

Simply put, Inferential Statistics make predictions about a population based on a sample of data taken from that population.

The technique of Inferential Statistics involves the following steps:

- First, take some samples and try to find one that represents the entire population accurately.
- Next, test the sample and use it to draw generalizations about the whole population.

Types of Inferential Statistics:

1. Estimating parameters:

We take a statistic from the collected data, such as the standard deviation, and use it to define a more general parameter, such as the standard deviation of the complete population.

2. Hypothesis testing:

Very beneficial when we are looking to gather data on something that can only be given to a very confined population, such as a new drug. If we want to know whether this drug will work for all patients ("complete population"), we can use the data collected to predict this (often by calculating a z-score).

3. Confidence Interval:

The confidence interval is the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or re-sample the population in the same way.

The confidence level is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of the confidence interval, and is set by the alpha value.

Confidence interval for the mean of normally-distributed data:

Normally-distributed data forms a bell shape when plotted on a graph, with the sample mean in the middle and the rest of the data distributed fairly evenly on either side of the mean.

The confidence interval for data which follows a standard normal distribution is:

$$\text{Confidence Interval} = \bar{X} \pm Z \times \frac{\sigma}{\sqrt{n}}$$

Where:

- CI = the confidence interval
- \bar{X} = the population mean
- Z = the critical value of the z-distribution
- σ = the population standard deviation
- \sqrt{n} = the square root of the population size

What exactly is a confidence interval?

A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence.

Confidence, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

Your desired confidence level is usually one minus the alpha (α) value you used in your statistical test

$$\text{Confidence level} = 1 - \alpha$$

So, if you use an alpha value of $p < 0.05$ for statistical significance, then your confidence level would be $1 - 0.05 = 0.95$, or 95%.

When do you use confidence intervals?

You can calculate confidence intervals for many kinds of statistical estimates, including:

- Proportions
- Population means
- Differences between population means or proportions
- Estimates of variation among groups

These are all point estimates, and don't give any information about the variation around the number. Confidence intervals are useful for communicating the variation around a point estimate.

Example: Variation around an estimate

You survey 100 Brits and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week.

However, the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.

Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.

Finding the standard deviation:

Most statistical software will have a built-in function to calculate your standard deviation, but to find it by hand you can first find your sample variance, then take the square root to get the standard deviation.

1. Find the sample variance

Sample variance is defined as the sum of squared differences from the mean, also known as the mean-squared-error (MSE):

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n - 1}$$

To find the MSE, subtract your sample mean from each value in the dataset, square the resulting number, and divide that number by $n - 1$

Then add up all of these numbers to get your total sample variance (s^2). For larger sample sets, it's easiest to do this in Excel.

2. Find the standard deviation.

The standard deviation of your estimate (s) is equal to the square root of the sample variance/sample error (s^2):

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

Example:

In the television-watching survey, the variance in the GB estimate is 100, while the variance in the USA estimate is 25. Taking the square root of the variance gives us a sample standard deviation (s) of:

- 10 for the GB estimate.
- 5 for the USA estimate.

Central Limit Theorem:

The central limit theorem is the basis for how normal distributions work in statistics.

In research, to get a good idea of a population mean, ideally you'd collect data from multiple random samples within the population. A **sampling distribution of the mean** is the distribution of the means of these different samples.

The central limit theorem shows the following:

- Law of Large Numbers: As you increase sample size (or the number of samples), then the sample mean will approach the population mean.
- With multiple large samples, the sampling distribution of the mean is normally distributed, even if your original variable is not normally distributed.

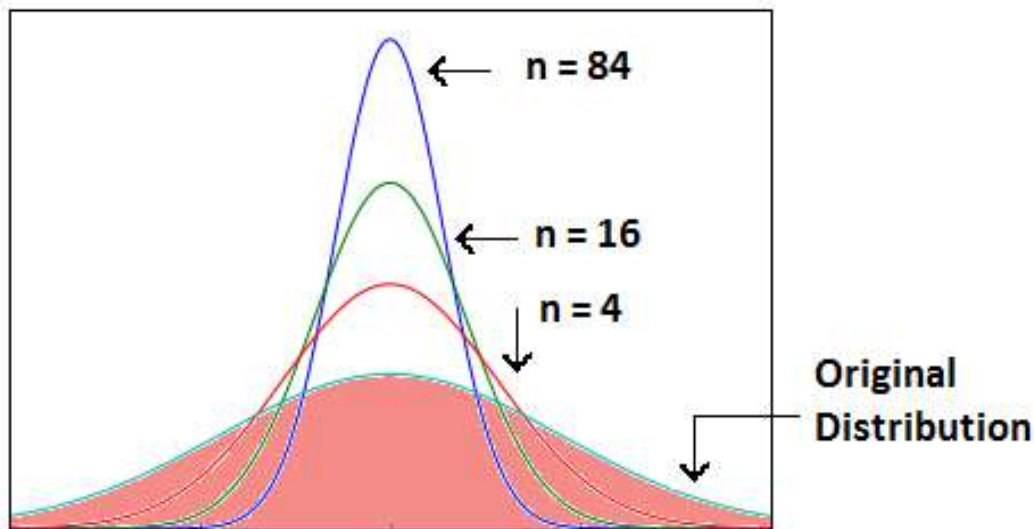
Parametric statistical tests typically assume that samples come from normally distributed populations, but the central limit theorem means that this assumption isn't necessary to meet when you have a large enough sample.

You can use parametric tests for large samples from populations with any kind of distribution as long as other important assumptions are met. A sample size of 30 or more is generally considered large.

For small samples, the assumption of normality is important because the sampling distribution of the mean isn't known. For accurate results, you have to be sure that the population is normally distributed before you can use parametric tests with small samples.

Why Is the Central Limit Theorem Useful?

The central limit theorem is useful when analysing 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.



Hypothesis Testing:

Hypothesis testing is a part of statistics in which we make assumptions about the population parameter. So, hypothesis testing mentions a proper procedure by analysing a random sample of the population to accept or reject the assumption.

Hypothesis testing is the way of trying to make sense of assumptions by looking at the sample data.

Type of Hypothesis:

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses.

- **Null Hypothesis.** The null hypothesis, denoted by **H₀**, is usually the hypothesis that sample observations result purely from chance.
- **Alternative Hypothesis.** The alternative hypothesis, denoted by **H₁ or H_a**, is the hypothesis that sample observations are influenced by some non-random cause.

Steps of Hypothesis Testing:

The process to determine whether to reject a null hypothesis or to fail to reject the null hypothesis, based on sample data is called hypothesis testing. It consists of four steps:

1. Define the null and alternate hypothesis
2. Define an analysis plan to find how to use sample data to estimate the null hypothesis
3. Do some analysis on the sample data to create a single number called '**test statistic**'
4. Understand the result by applying the decision rule to check whether the Null hypothesis is true or not

If the value of t-stat is less than the significance level we will reject the null hypothesis, otherwise, we will fail to reject the null hypothesis.

Technically, we never accept the null hypothesis, we say that either we fail to reject or we reject the null hypothesis.

Errors in hypothesis testing:

We have explained what is hypothesis testing and the steps to do the testing. Now, while performing the hypothesis testing, there might be some errors.

- **Type I error.** A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called **alpha** and is often denoted by α .
- **Type II error.** A Type II error occurs when the researcher fails to reject a false null hypothesis. The probability of committing a Type II error is called **beta** and is often denoted by β . The probability of *not* committing a Type II error is called the **Power** of the test.

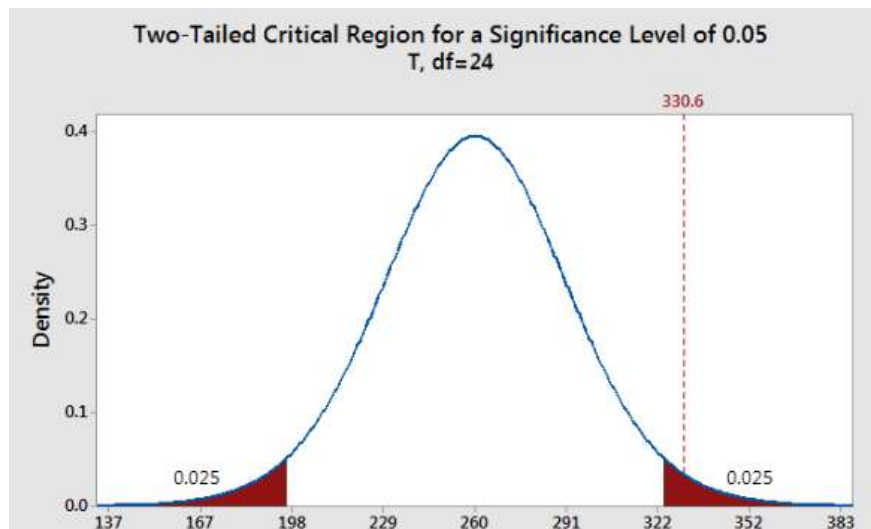
| Decision ----- Actual | Reject the null hypothesis | Fail to reject the null hypothesis |
|------------------------------|----------------------------|------------------------------------|
| Null Hypothesis is True | Type-1 Error | Decision is correct |
| Alternate hypothesis is true | Decision is correct | Type-2 Error |

Terms in Hypothesis testing:

Significance level:

The significance level is defined as the probability of the case when we reject the null hypothesis, but in actuality, it is true. For example, a 0.05 significance level indicates that there is a 5% risk in assuming that there is some difference when, in actuality, there is no difference. It is denoted by alpha (α).

The below figure shows that the two shaded regions are equidistant from the null hypothesis, each having a probability of 0.025 and a total of 0.05, which is our significance level. The shaded region in case of a two-tailed test is called the critical region.



P-value:

The p-value is defined as the probability of seeing a **t-statistic** as extreme as the calculated value if the null hypothesis value is true. A low enough p-value is the ground for rejecting the null hypothesis. We reject the null hypothesis if the p-value is less than the significance level.

Z-test:

A z test is used on data that follows a normal distribution and has a sample size greater than or equal to 30. It is used to test if the means of the sample and population are equal when the population variance is known. The right tailed hypothesis can be set up as follows:

Null Hypothesis: $H_0: \mu = \mu_0$

Alternate Hypothesis: $H_1: \mu > \mu_0$

Decision Criteria: If the z statistic $>$ z critical value then rejects the null hypothesis.

We find the Z-statistic of the sample means and calculate the z-score. Z-score is given by the formula,

$$z = \frac{x - \mu}{\sigma}$$

Z-test is mainly used when the population mean and standard deviation are given.

T-test:

A t test is used when the data follows a student t distribution and the sample size is lesser than 30. It is used to compare the sample and population mean when the population variance is unknown. The hypothesis test for inferential statistics is given as follows:

Null Hypothesis: $H_0: \mu = \mu_0$

Alternate Hypothesis: $H_1: \mu > \mu_0$

Decision Criteria: If the t statistic > t critical value then rejects the null hypothesis.

The Sample Standard Deviation is given as:

$$S = \frac{\sqrt{\sum (x - \bar{x})^2}}{(n-1)}$$

where $n-1$ is Bessel's correction for estimating the population parameter.

Another difference between z-scores and t-values is that t-values are dependent on the Degree of Freedom of a sample. Let us define what degree of freedom is for a sample.

The Degree of Freedom:

It is the number of variables that have the choice of having more than one arbitrary value. For example, in a sample of size 10 with a mean of 10, 9 values can be arbitrary, but the 10th value is forced by the sample mean.

Points to note about the t-tests:

1. The greater the difference between the sample mean and the population mean, the greater the chance of rejecting the Null Hypothesis.
2. Greater the sample size, the greater the chance of rejection of the Null Hypothesis.

Different types of T-tests:

1. One Sample T-test:

The one-sample t-test compares the mean of sample data to a known value. So, if we have to compare the mean of sample data to the population mean, we use the One-Sample T-test.

We can run a one-sample T-test when we do not have the population S.D., or we have a sample of size less than 30.

t-statistic is given by:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

where, \bar{X} is the sample mean, μ the population mean, s the sample standard deviation, and N the sample size.

2. Two sample T-test:

We use a two-sample T-test when we want to evaluate whether the mean of the two samples is different or not. In a two-sample T-test, we have another two categories:

- **Independent Sample T-test:**

Independent sample means that the two different samples should be selected from two completely different populations. In other words, we can say that one population should not be dependent on the other population.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

$\bar{x}_1 - \bar{x}_2$ is the difference between the sample means

$\mu_1 - \mu_2$ is the difference between the hypothesized population means

$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is the standard error of the difference between the sample means

- **Paired T-test:**

If our samples are connected in some way, we have to use the paired t-test. Here, 'connecting' means that the samples are connected as we are collecting data from the same group two times, e.g., blood tests of patients of a hospital before and after medication.

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where, \bar{d} is mean of the case wise difference between before and after case,

s_d = standard deviation of the difference

n = sample size.

Chi-Square test:

The Chi-square test is used in the case when we have to compare categorical data.

The Chi-square test is of two types. Both use chi-square statistics and distribution for different purposes.

- **The goodness of fit:** It determines if sample data of categorical variables match with population or not.
- **Test of Independence:** It compares two categorical variables to find whether they are related to each other or not.

Chi-square statistic is given by:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

ANOVA (Analysis of variance):

ANOVA (Analysis of Variance) is used to check if at least one of two or more groups have statistically different means. Now, the question arises — Why do we need another test for checking the difference of means between independent groups? Why can we not use multiple t-tests to check for the difference in means?

The answer is simple. Multiple t-tests will have a compound effect on the error rate of the result. Performing a t-test thrice will give an error rate of ~15%, which is too high, whereas ANOVA keeps it at 5% for a 95% confidence interval.

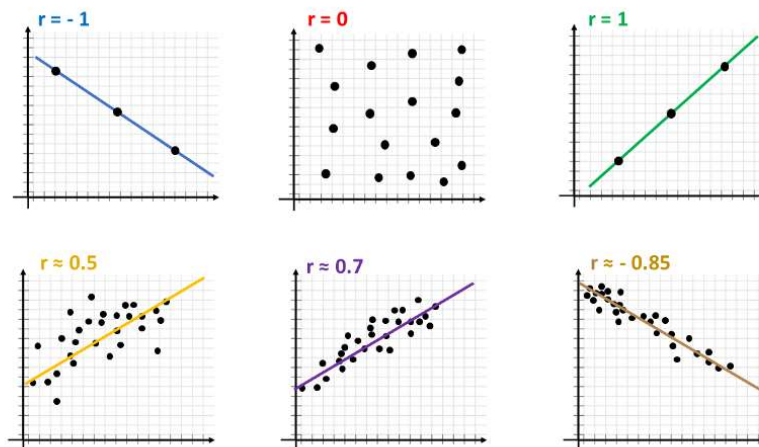
To perform an ANOVA, you must have a continuous response variable and at least one categorical factor with two or more levels. ANOVA requires data from approximately normally distributed populations with equal variances between factor levels.

There are two types of ANOVA test:

1. **One-way ANOVA:** when only 1 independent variable is considered.
2. **Two-way ANOVA:** when 2 independent variables are considered.
3. **N-way ANOVA:** when N number of independent variables are considered.

Correlation Coefficient (R or r):

It is used to measure the strength between two variables. It is simply the square root of the coefficient of Determination and ranges from -1 to 1 where 0 represents no correlation, and 1 represents positive strong correlation while -1 represents negative strong correlation.



Carsten Grube - dataZ4s.com - Statistical Data Analysis

Difference between Z-Test and T-Test:

| Basis | Z Test | T-Test |
|--|--|--|
| Basic Definition | Z-test is a kind of hypothesis test which ascertains if the averages of the 2 datasets are different from each other when standard deviation or variance is given. | The t-test can be referred to as a kind of parametric test that is applied to an identity, how the averages of 2 sets of data differ from each other when the standard deviation or variance is not given. |
| Population Variance | The Population variance or standard deviation is known here. | The Population variance or standard deviation is unknown here. |
| Sample Size | The Sample size is large. | Here the Sample Size is small. |
| Key Assumptions | All data points are independent. Normal Distribution for Z, with an average zero and variance = 1. | All data points are not dependent. Sample values are to be recorded and taken accurately. |
| Based upon (a type of distribution) | Based on Normal distribution. | Based on Student-t distribution. |

Difference between One Tailed Test and Two Tailed Test:

| BASIS OF COMPARISON | ONE-TAILED TEST | TWO-TAILED TEST |
|--------------------------------|--|--|
| Meaning | A statistical hypothesis test in which alternative hypothesis has only one end, is known as one tailed test. | A significance test in which alternative hypothesis has two ends, is called two-tailed test. |
| Hypothesis | Directional | Non-directional |
| Region of rejection | Either left or right | Both left and right |
| Determines | If there is a relationship between variables in single direction. | If there is a relationship between variables in either direction. |
| Result | Greater or less than certain value. | Greater or less than certain range of values. |
| Sign in alternative hypothesis | $>$ or $<$ | \neq |

Sample Vs Population:

Sample:

A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

You should calculate the sample standard deviation when the dataset you're working with represents a sample taken from a larger population of interest. The formula to calculate a sample standard deviation, denoted as s .

Whatever statistical measures I can calculate Here we called them as **“Descriptive statistics.”**

Population:

A population is the entire group that you want to draw conclusions about.

You should calculate the population standard deviation when the dataset you're working with represents an entire population, i.e., every value that you're interested in. The formula to calculate a population standard deviation, denoted as σ .

Whatever the sample data, I have I will apply some additional theory and I will estimate on population. **‘Inferential statistics’**

| Sample | Population |
|---|--|
| $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$ | $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ |

| | |
|---|---|
| Σ : A symbol that means “sum” x_i : The i th value in a dataset \bar{x} : The sample mean n : The sample size | Σ : A symbol that means “sum” x_i : The i th value in a dataset μ : The population mean N : The population size |
|---|---|

Brute Force method:

When we try to get all the data and compute it to make a statement, i.e., when we reach out to the whole population. But it is difficult to compute and get each and every individual, that why we are getting into Sample.

Reasons for sampling:

- **Necessity:**
Sometimes it's simply not possible to study the whole population due to its size or inaccessibility.
- **Practicality:**
It's easier and more efficient to collect data from a sample.
- **Cost-effectiveness:**
There are fewer participant, laboratory, equipment, and researcher costs involved.
- **Manageability:**
Storing and running statistical analyses on smaller datasets is easier and reliable.

Practice Problem 1: Height

Suppose a gym teacher wants to summarize the mean and standard deviation of heights of students in his class.

When calculating the standard deviation of height, should he use the population or sample standard deviation formula?

Answer: He should use the **population standard deviation** because he is only interested in the height of students in this one particular class.

Practice Problem 2: Manufacturing

Suppose an inspector wants to summarize the mean and standard deviation of the weight of tires produced at a certain factory. He decides to collect a simple random sample of 40 tires from the factory and weighs each of them.

When calculating the standard deviation of weights, should he use the population or sample standard deviation formula?

Answer: He should use the **sample standard deviation** because he is interested in the weights of all tires produced at this factory, not just the weights of the tires in his sample.

Practice Problem 3: Biology

Suppose a biologist wants to summarize the mean and standard deviation of the weight of a particular species of turtles. She decides to go out and collect a simple random sample of 20 turtles from the population.

When calculating the standard deviation of weights, should she use the population or sample standard deviation formula?

Answer: She should use the **sample standard deviation** because she is interested in the weights of the entire population of turtles, not just the weights of the turtles in her sample.

Population and Sampling methods:

The population contains all the data points from a set of data, while a sample consists of few observations selected from the population. The sample from the population should be selected such that it has all the properties that a population has. Population's measurable properties such as mean, standard deviation, etc., are called parameters, while Sample's measurable property is known as a statistic.

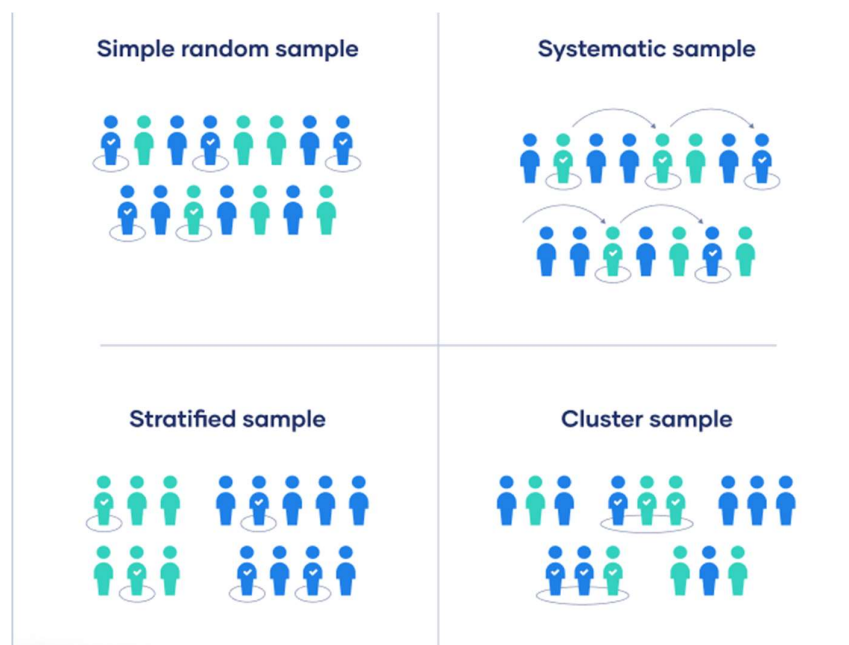
To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. There are two types of sampling methods:

- **Probability sampling** involves random selection, allowing you to make strong statistical inferences about the whole group.
- **Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Probability Sampling: (Unbiased Sample)

Probability sampling means that every member of the population has a chance of being selected. It is mainly used in 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:

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: You want to select a simple random sample of 100 employees of Company X. 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:

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: All 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:

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, 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 to select a sample from each subgroup.

Example: The 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:

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.

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.

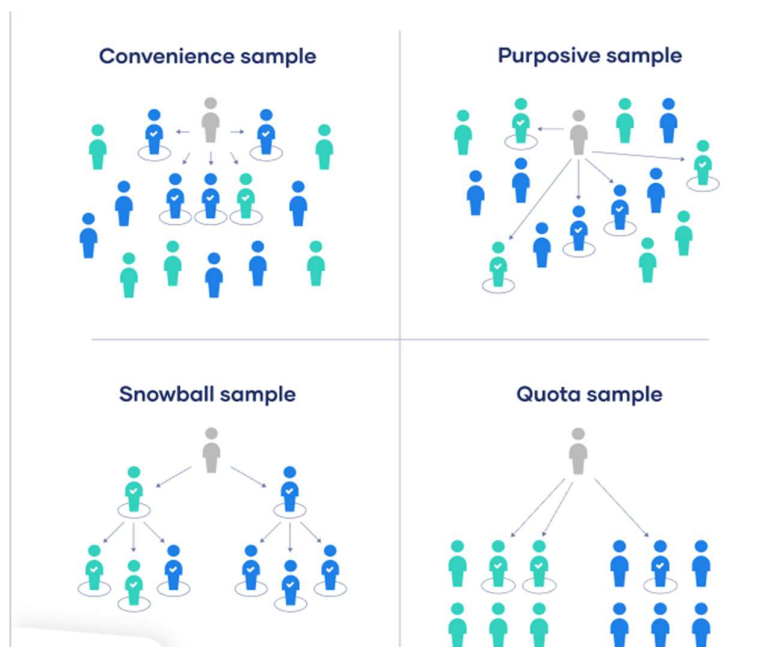
Example: The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Non-probability sampling: (Biased Sample)

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. 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 and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.



1. 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 results.

Example: You 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 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, as some people will inherently be more likely to volunteer than others.

Example: You 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:

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, 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.

Example: You 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

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.

Example: You 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.

Random Variable and Probability Distributions:

Probability is defined as the likeliness of something to occur or happen and probability distributions are functions that give the relation between all the outcomes of a random variable in any random experiment and its probable values.

These distribution functions are used in predicting the stock prices, weather prediction.

What is Random Variable?

Set of all possible values from a Random Experiment is called Random Variable.

It is represented by X .

Example: Outcome of coin toss.

Types of Random Variable:

- Discrete Random Variable:

X is a discrete because it has a countable value between two numbers

Example: number of balls in a bag, number of tails in tossing coin

- Continuous Random Variable:

X is a continuous because it has an infinite number of values between two values

Example: distance travelled, Height of students

What is Probability Distribution?

A Probability Distribution of a random variable is a list of all possible outcomes with corresponding probability values.

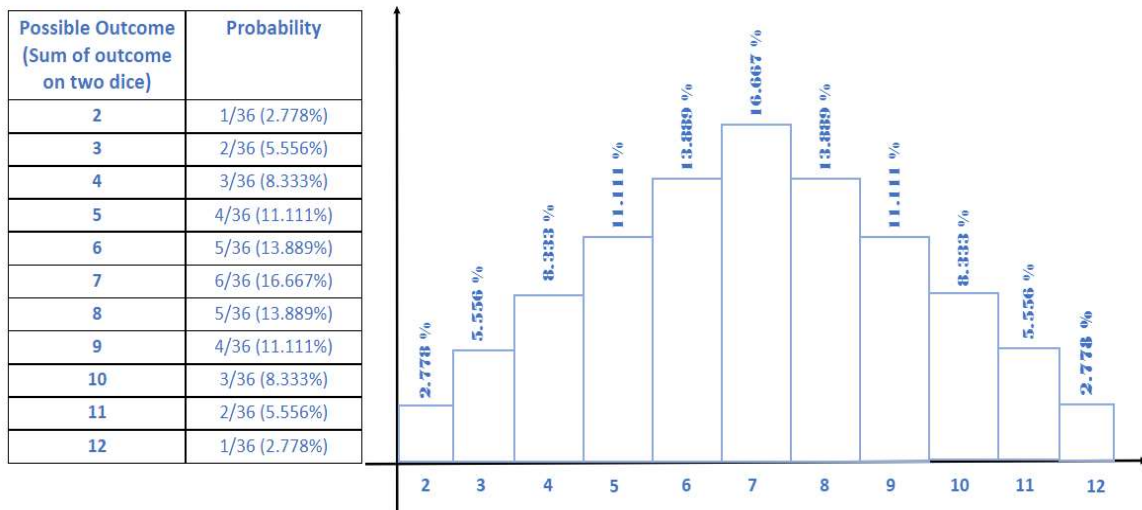
Note: The value of the probability always lies between 0 to 1.

| Outcome of die roll | 1 | 2 | 3 | 4 | 5 | 6 |
|---------------------|-----|-----|-----|-----|-----|-----|
| Probability | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

The probability distribution for a fair six-sided die

What is an example of Probability Distribution?

When two dice are rolled with six sided dots, let the possible probability of rolling is as follows:



- If a random variable is a discrete variable, its probability distribution is called discrete probability distribution.
 - Example: Flipping of two coins
 - Functions that represent a discrete probability distribution is known as **Probability Mass Function**.
- If a random variable is a continuous variable, its probability distribution is called continuous probability distribution.
 - Example: Measuring temperature over a period of time
 - Functions that represent a continuous probability distribution is known as **Probability Density Function**.

Types of Probability Distributions:

Uniform Distribution:

Probability distribution in which all the outcome has equal probability is known as Uniform Distribution.

Example: Perfect Random Generator

Consider an experiment of tossing a single coin:



- Probability of getting Head = 0.5
- Probability of getting Tail = 0.5

Bernoulli Distribution:

A discrete probability distribution for a random experiment that has only two possible outcomes (Bernoulli trials) is known Bernoulli Distribution.

Example: India will win cricket world cup or not

It has only two possible outcomes

- Success (1)
- Failure (0)

Consider an experiment of Shooting of Basketball

- Shoots the Ball ($n = 1$) = p
- Doesn't Shoots the Ball ($n = 0$) = $q = 1 - p$



Binomial Distribution:

A discrete probability distribution that gives only two possible outcomes in n independent trials is known as Binomial Distribution.

Example: Yes/No survey

- Extension of Bernoulli Distribution
- Represent the number of success and failure into n independent trials
- The probability of success and failure is the same for all independent and identical trials.

Let's understand the Binomial Distribution by an example,

Consider the experiment of Picking Balls

Problem Statement:

Let there are 8 white balls and 2 black balls, then the probability of drawing 3 white balls, if the probability of selecting white ball is 0.6.

$$n = 8 + 2 = 10$$

$$p = 0.6$$

$$P(X = 3) = \frac{10!}{3!7!}(0.6)^3(1 - 0.6)^7 = 0.04247$$



| Bernoulli | Binomial |
|---------------------------------------|--|
| Deals with the single trial event | Deals with the outcome of Multiple trials of the single events |
| Has only two possible outcome 0 and 1 | Sum of identically and independent distributed Bernoulli Random Variable |

Poisson Distribution:

A discrete probability distribution that measures the probability of a random variable over a specific period of time is known as Poisson Distribution.

Example: Probability of Asteroid collision over a selected year of period.

- Used to predict probability of number of successful events.
- Random variable X is Poisson distributed if the distribution function is given by:

Note: In case of Poisson Distribution **Mean = Variance**

Let's understand the Poisson Distribution by an example,

Consider the experiment of Number of patients visiting in a hospital

Problem Statement:

Let in a hospital patient arriving in a hospital at expected value is 6, then what is the probability of five patients will visit the hospital in that day?

- Patients arriving at expected value = 6
- $P(\text{Five patients will visit the hospital}) = P(X=5)$

$$P(X=5) = \frac{6^5 e^{-6}}{5!} = 0.1606$$

| Poisson | Binomial |
|---------------------------------------|---|
| Number of trials are infinite | Number of trials are fixed |
| Unlimited number of possible outcomes | Only two possible outcomes (Success or Failure) |
| Mean = Variance | Mean > Variance |

Normal Distribution:

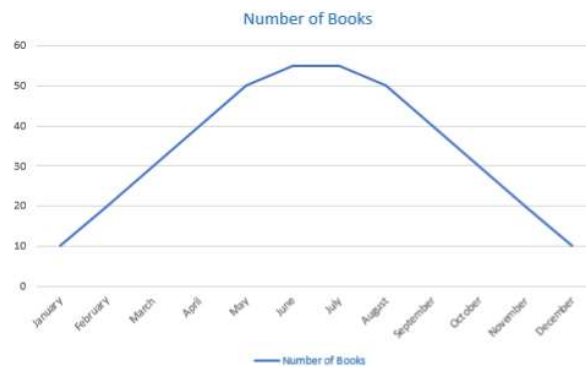
(Gaussian Distribution, Bell Curve, Symmetric Around Mean):

A continuous probability distribution, which is symmetric about its mean value (i.e., data near the mean are more frequency in occurrence) is known as Normal Distribution.

Let's understand the Normal Distribution by an example,
Consider the experiment of Number of books read by students in a school

Number of Books Read by Students

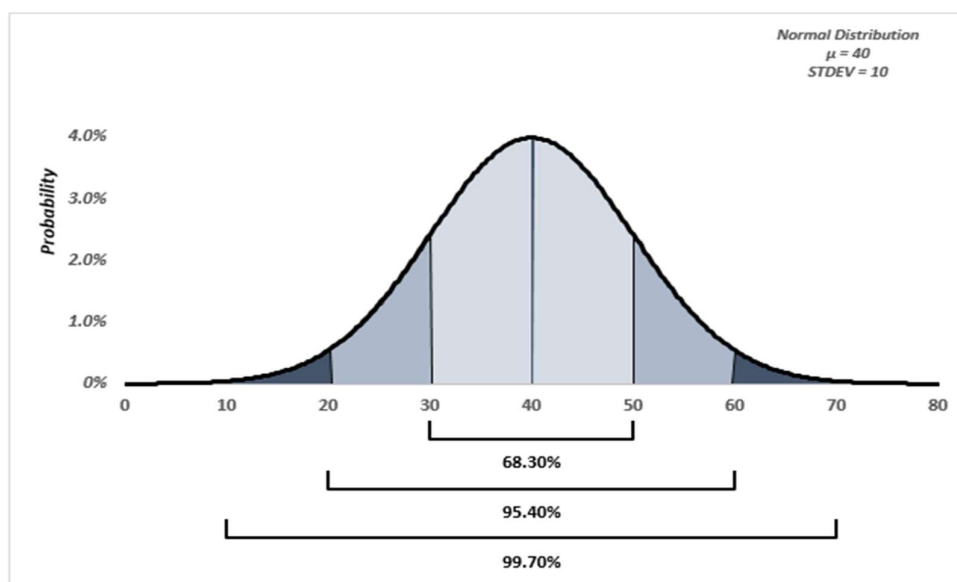
| Months | Number of Books |
|-----------|-----------------|
| January | 10 |
| February | 20 |
| March | 30 |
| April | 40 |
| May | 50 |
| June | 55 |
| July | 55 |
| August | 50 |
| September | 40 |
| October | 30 |
| November | 20 |
| December | 10 |



Empirical Rule:

Empirical Rule is often called the **68 – 95 – 99.7** rule or **Three Sigma Rule**. It states that on a Normal Distribution:

- 68% of the data will be within one Standard Deviation of the Mean
- 95% of the data will be within two Standard Deviations of the Mean
- 99.7 of the data will be within three Standard Deviations of the Mean



A Normal distribution curve has the following properties:

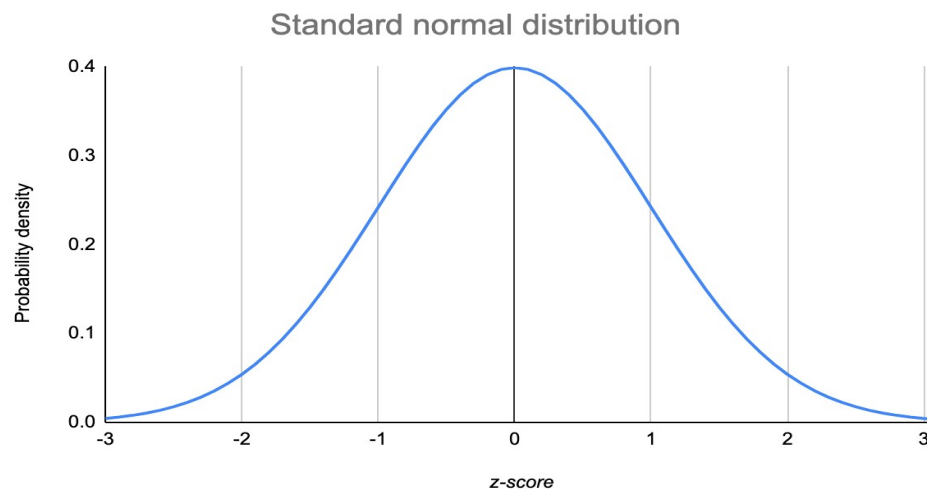
- Symmetrical around its mean value
- Mean = Median = Mode
- Total area under the curve is 1
- Curve of the distribution is bell curve
- The curve is symmetric, with half of the values on the left and half of the values on the right.

Difference between Poisson and Normal Distribution:

| Poisson | Normal |
|-----------------------------------|----------------------------|
| Use Discrete Data | Use Continuous Data |
| Distribution varies on mean value | Symmetric about mean value |
| Mean = Variance | Mean = Median = Mode |

Standard normal distribution:

- Normal distribution with mean = 0 and standard deviation = 1.



Probability distributions are not a Graph.
A graph is just a visual representation.

The End