1. What is Analytics?

Analytics is a broad term that encompasses the processes, technologies, frameworks, and algorithms to extract meaningful insights from data. Raw data in itself does not have a meaning until it is contextualized and processed into useful information. Analytics is this process of extracting and creating information from raw data by filtering, processing, categorizing, condensing, and contextualizing the data. This information obtained is then organized and structured to infer knowledge about the system and/or its users, its environment, and its operations and progress towards its objectives, thus making the systems smarter and more efficient. The choice of the technologies, algorithms, and frameworks for analytics is driven by the analytics goals of the application. For example, the goals of the analytics task may be: (1) to predict something (for example whether a transaction is a fraud or not, whether it will rain on a particular day, or whether a tumour is benign or malignant), (2) to find patterns in the data (for example, finding the top 10 coldest days in the year, finding which pages are visited the most on a particular website, or finding the most searched celebrity in a particular year), (3) finding relationships in the data (for example, finding similar news articles, finding similar patients in an electronic health record system, finding related products on an eCommerce website, finding similar images, or finding correlation between news items and stock prices).

1.1 Types of Analytics

1.1.1 Descriptive Analytics Descriptive analytics comprises analyzing past data to present it in a summarized form which can be easily interpreted. Descriptive analytics aims to answer - What has happened? A major portion of analytics done today is descriptive analytics through use of statistics functions such as counts, maximum, minimum, mean, top-N, percentage, for instance. These statistics help in describing patterns in the data and present the data in a summarized form. For example, computing the total number of likes for a particular post, computing the average monthly rainfall or finding the average number of visitors per month on a website. Descriptive analytics is useful to summarize the data.

1.1.2 Diagnostic Analytics Diagnostic analytics comprises analysis of past data to diagnose the reasons as to why certain events happened. Diagnostic analytics aims to answer - Why did it happen? Let us consider an example of a system that collects and analyzes sensor data from machines for monitoring their health and predicting failures. While descriptive analytics can be useful for summarizing the data by computing various statistics (such as mean, minimum, maximum, variance, or top-N), diagnostic analytics can provide more insights into why certain a fault has occurred based on the patterns in the sensor data for previous faults. Among the seven computational tasks, the computational tasks such as Linear Algebraic Computations, General N-Body Problems, and Graph-theoretic Computations can be used for diagnostic analytics.

1.1.3 Predictive Analytics Predictive analytics comprises predicting the occurrence of an event or the likely outcome of an event or forecasting the future values using prediction models. Predictive analytics aims to answer - What is likely to happen? For example, predictive analytics can be used for predicting when a fault will occur in a machine, predicting whether a tumour is benign or malignant, predicting the occurrence of natural emergency (events such as forest fires or river floods) or forecasting the pollution levels. Predictive Analytics is done using predictive models which are trained by existing data. These models learn patterns and trends from the existing data and predict the occurrence of an event or the likely outcome of an event (classification models) or forecast numbers (regression models). The accuracy of prediction models depends on the quality and volume of the existing data available for training the models, such that all the patterns and trends in the existing data can be learned accurately. Before a model is used for prediction, it must be validated with existing data. The typical approach adopted while developing prediction models is to divide the existing data into training and test data sets (for example 75% of the data is used for training and 25% data is used for testing the prediction model).

Tasks such as Linear Algebraic Computations, General N-Body Problems, Graph-theoretic Computations, Integration and Alignment Problems can be used for predictive analytics.

1.1.4 Prescriptive Analytics While predictive analytics uses prediction models to predict the likely outcome of an event, prescriptive analytics uses multiple prediction models to predict various outcomes and the best course of action for each outcome. Prescriptive analytics aims to answer - What can we do to make it happen? Prescriptive Analytics can predict the possible outcomes based on the current choice of actions. We can consider prescriptive analytics as a type of analytics that uses different prediction models for different inputs. Prescriptive analytics prescribes actions or the best option to follow from the available options. For example, prescriptive analytics can be used to prescribe the best medicine for treatment of a patient based on the outcomes of various medicines for similar patients. Another example of prescriptive analytics would be to suggest the best mobile data plan for a customer based on the customer's browsing patterns. Tasks such as General N-Body Problems, Graph theoretic Computations, Optimization and Alignment Problems can be used for prescriptive analytics.

2. What is Big Data?

Big data is defined as collections of datasets whose volume, velocity or variety is so large that it is difficult to store, manage, process, and analyze the data using traditional databases and data processing tools. In the recent years, there has been an exponential growth in both structured and unstructured data generated by information technology, industrial, healthcare, Internet of Things, and other systems Big Data has the potential to power next generation of smart applications that will leverage the power of the data to make the applications intelligent. Applications of big data span a wide range of domains such as web, retail and marketing, banking and financial, industrial, healthcare, environmental, Internet of Things and cyber-physical systems. Big Data analytics deals with collection, storage, processing, and analysis of this massive-scale data.

Specialized tools and frameworks are required for big data analysis when:

(1) the volume of data involved is so large that it is difficult to store, process and analyze data on a single machine,

(2) the velocity of data is very high, and the data needs to be analyzed in real-time,

(3) there is variety of data involved, which can be structured, unstructured or semi-structured, and is collected from multiple data sources,

(4) various types of analytics need to be performed to extract value from the data such as descriptive, diagnostic, predictive and prescriptive analytics.

Big Data tools and frameworks have distributed and parallel processing architectures and can leverage the storage and computational resources of a large cluster of machines. Big data analytics involves several steps starting from data cleansing, data munging (or wrangling), data processing and visualization. Big data analytics lifecycle starts from the collection of data from multiple data sources. Specialized tools and frameworks are required to ingest the data from different sources into the dig data analytics backend. The data is stored in specialized storage solutions (such as distributed filesystems and non-relational databases) which are designed to scale. Based on the analysis requirements (batch or real-time), and type of analysis to be performed (descriptive, diagnostic, predictive, or predictive) specialized frameworks are used. Big data analytics is enabled by several technologies such as cloud computing, distributed and parallel processing frameworks, non-relational databases, in-memory computing, for instance.

Some examples of big data are listed as follows:

- Data generated by social networks including text, images, audio and video data

- Click-stream data generated by web applications such as e-Commerce to analyze user behaviour

- Machine sensor data collected from sensors embedded in industrial and energy systems for monitoring their health and detecting failures

- Healthcare data collected in electronic health record (EHR) systems

- Logs generated by web applications

- Stock markets data

- Transactional data generated by banking and financial applications

2.1 Characteristics of Big Data (Vs of Big Data)

The underlying characteristics of big data include:

2.1.1 Volume - Big data is a form of data whose volume is so large that it would not fit on a single machine therefore specialized tools and frameworks are required to store process and analyze such data. For example, social media applications process billions of messages every day, industrial and energy systems can generate terabytes of sensor data every day, cab aggregation applications can process millions of transactions in a day, etc. The volumes of data generated by modern IT, industrial, healthcare, Internet of Things, and other systems is growing exponentially driven by the lowering costs of data storage and processing architectures and the need to extract valuable insights from the data to improve business processes, efficiency and service to consumers. Though there is no fixed threshold for the volume of data to be considered as big data, however, typically, the term big data is used for massive scale data that is difficult to store, manage and process using traditional databases and data processing architectures.

2.1.2 Velocity - Velocity of data refers to how fast the data is generated. Data generated by certain sources can arrive at very high velocities, for example, social media data or sensor data. Velocity is another important characteristic of big data and the primary reason for the exponential growth of data. High velocity of data results in the volume of data accumulated to become very large, in short span of time. Some applications can have strict deadlines for data analysis (such as trading or online fraud detection) and the data needs to be analyzed in real-time. Specialized tools are required to ingest such high velocity data into the big data infrastructure and analyze the data in real-time.

2.1.3 Variety - Variety refers to the forms of the data. Big data comes in different forms such as structured, unstructured or semi-structured, including text data, image, audio, video and sensor data. Big data systems need to be flexible enough to handle such variety of data

2.1.4 Veracity - Veracity refers to how accurate is the data. To extract value from the data, the data needs to be cleaned to remove noise. Data-driven applications can reap the benefits of big data only when the data is meaningful and accurate. Therefore, cleansing of data is important so that incorrect and faulty data can be filtered out.

2.1.5 Value - Value of data refers to the usefulness of data for the intended purpose. The end goal of any big data analytics system is to extract value from the data. The value of the data is also related to the veracity or accuracy of the data. For some applications value also depends on how fast we are able to process the data.

2.2 Types of Big Data

Data can be broadly classified as being structured, unstructured, or semi-structured. Although these distinctions have always existed, the classification of data into these categories has become more prominent with the advent of big data.

2.2.1 Structured data - Structured data, as the name implies, indicates datasets that have a defined organizational structure such as Microsoft Excel or CSV files. In pure database terms, the data should be representable using a schema. As an example, the following table representing the top five happiest countries in the world published by the United Nations in its 2017 World Happiness Index ranking would be an atypical representation of structured data. We can clearly define the data types of the columns--Rank, Score, GDP per capita, social support, Healthy life expectancy, Trust, Generosity, and Dystopia are numerical columns, whereas Country is represented using letters, or more specifically, strings. Commercial databases such as Teradata, Greenplum as well as Redis, Cassandra, and Hive in the open-source domain are examples of technologies that provide the ability to manage and query structured data.

2.2.2 Unstructured data - Unstructured data consists of any dataset that does not have a predefined organizational schema as in the table in the prior section. Spoken words, music, videos, and even books, including this one, would be considered unstructured. This by no means implies that the content doesn't have organization. Indeed, a book has a table of contents, chapters, subchapters, and an index--in that sense, it follows a definite organization. However, it would be futile to represent every word and sentence as being part of a strict set of rules. A sentence can consist of words, numbers, punctuation marks, and so on and does not have a predefined data type as spreadsheets do. To be structured, the book would need to have an exact set of characteristics in every sentence, which would be both unreasonable and impractical

Unstructured data can be stored in various formats. They can be Blobs or, in the case of textual data, freeform text held in a data storage medium. For textual data, technologies such as Lucene/Solr, Elasticsearch, and others are generally used to query, index, and other operations.

2.2.3 Semi-structured data - Semi-structured data refers to data that has both the elements of an organizational schema as well as aspects that are arbitrary. A personal phone diary (increasingly rare these days!) with columns for name, address, phone number, and notes could be considered a semi-structured dataset. The user might not be aware of the addresses of all individuals and hence some of the entries may have just a phone number and vice versa. Similarly, the column for notes may contain additional descriptive information (such as a facsimile number, name of a relative associated with the individual, and so on). It is an arbitrary field that allows the user to add complementary information. The columns for name, address, and phone number can thus be considered structured in the sense that they can be presented in a tabular format, whereas the notes section is unstructured in the sense that it may contain an arbitrary set of descriptive information that cannot be represented in the other columns in the diary. In computing, semi-structured data is usually represented by formats, such as JSON, that can encapsulate both structured as well as schema less or arbitrary associations, generally using key-value pairs. A more common example could be emailing messages, which have both a structured part, such as name of the sender, time when the message was received, and so on, that is common to all email messages and an unstructured portion represented by the body or content of the email. Platforms such as Mongo and CouchDB are generally used to store and query semi□structured datasets.

2.3 Sources of big data

Technology today allows us to collect data at an astounding rate--both in terms of volume and variety. There are various sources that generate data, but in the context of big data, the primary sources are as follows:

Social networks: Arguably, the primary source of all big data that we know of today is the social networks that have proliferated over the past 5-10 years. This is by and large unstructured data that is represented by millions of social media postings and other data that is generated on a second-by-second basis through user interactions on the web across the world. Increase in access to the internet across the world has been a self-fulfilling act for the growth of data in social networks.

Media: Largely a result of the growth of social networks, media represents the millions, if not billions, of audio and visual uploads that take place on a daily basis. Videos uploaded on YouTube, music recordings on SoundCloud, and pictures posted on Instagram are prime examples of media, whose volume continues to grow in an unrestrained manner.

Data warehouses: Companies have long invested in specialized data storage facilities commonly known as data warehouses. A DW is essentially collections of historical data that companies wish to maintain and catalogue for easy retrieval, whether for internal use or regulatory purposes. As industries gradually shift toward the practice of storing data in platforms such as Hadoop and NoSQL, more and more companies are moving data from their pre-existing data warehouses to some of the newer technologies. Company emails, accounting records, databases, and internal documents are some examples of DW data that is now being offloaded onto Hadoop or Hadoop-like platforms that leverage multiple nodes to provide a highly available and fault-tolerant platform.

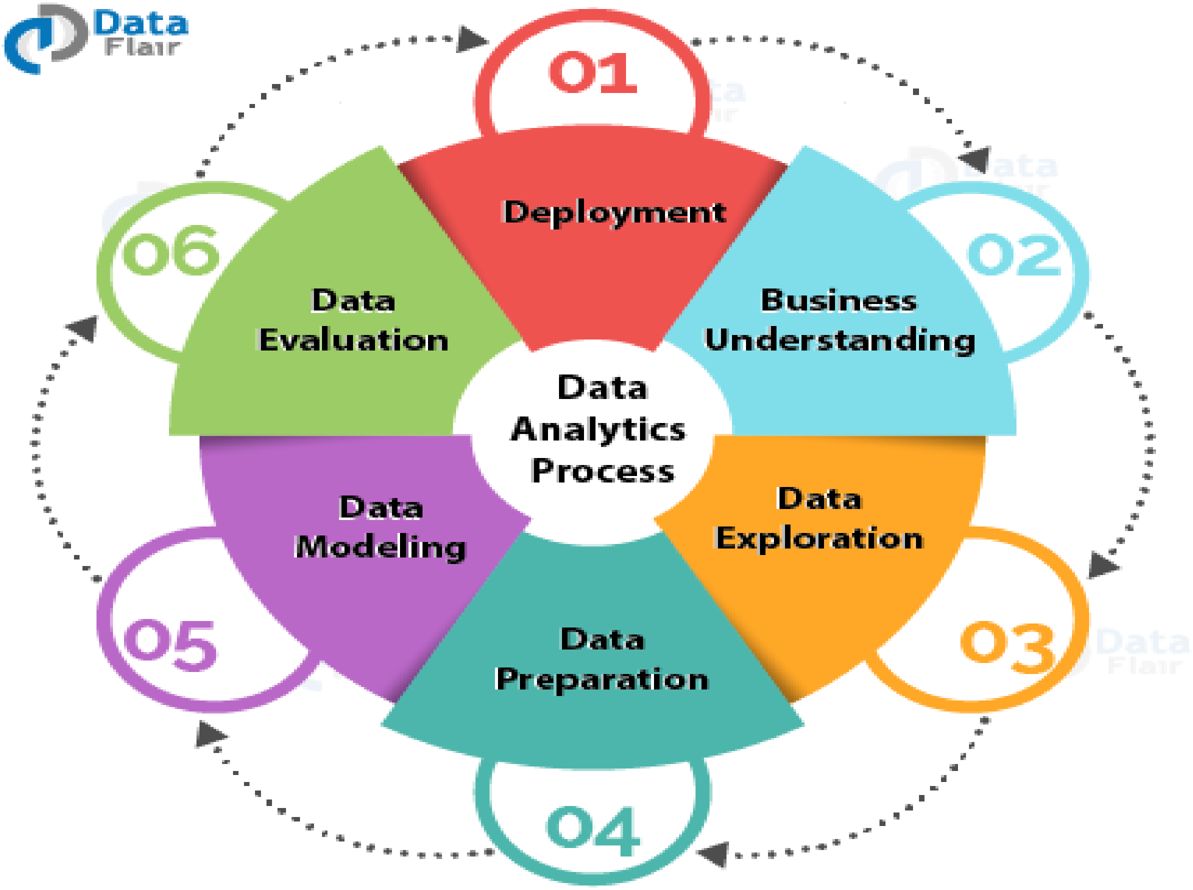
Sensors: A more recent phenomenon in the space of big data has been the collection of data from sensor devices. While sensors have always existed and industries such as oil and gas have been using drilling sensors for measurements at oil rigs for many decades, the advent of wearable devices, also known as the Internet Of Things, such as Fitbit and Apple Watch, meant that now each individual could stream data at the same rate at which a few oil rigs used to do just 10 years back.

Wearable devices can collect hundreds of measurements from an individual at any given point in time. While not yet a big data problem as such, as the industry keeps evolving, sensor-related data is likely to become more akin to the kind of spontaneous data that is generated on the web through social network activities.

3. Analytic Processes and Tools

There are 6 analytic processes:

1. Deployment
2. Business Understanding
3. Data Exploration
4. Data Preparation
5. Data Modelling
6. Data Evaluation



Step 1: Deployment

- Here we need to:
  - plan the deployment and monitoring and maintenance,
  - we need to produce a final report and review the project.
  - In this phase,
- we deploy the results of the analysis.
- This is also known as reviewing the project.

Step 2: Business Understanding

- Business Understanding
  - The very first step consists of business understanding.
  - Whenever any requirement occurs, firstly we need to determine the business objective,
  - assess the situation,
  - determine data mining goals and then
  - produce the project plan as per the requirement.
- Business objectives are defined in this phase.

Step 3: Data Exploration

- The second step consists of Data understanding.
  - For the further process, we need to gather initial data, describe and explore the data and verify data quality to ensure it contains the data we require.
  - Data collected from the various sources is described in terms of its application and the need for the project in this phase.
  - This is also known as data exploration.
- This is necessary to verify the quality of data collected.

Step 4: Data Preparation

- From the data collected in the last step,
    - we need to select data as per the need, clean it, construct it to get useful information and
    - then integrate it all.
- Finally, we need to format the data to get the appropriate data.
- Data is selected, cleaned, and integrated into the format finalized for the analysis in this phase.

Step 5: Data Modelling

- we need to
  - select a modelling technique, generate test design, build a model and assess the model built.
- The data model is built to
  - analyze relationships between various selected objects in the data,
  - test cases are built for assessing the model and model is tested and implemented on the data in this phase.

4. ANALYSIS AND REPORTING

What is Analysis?

- The process of exploring data and reports
  - in order to extract meaningful insights,
  - which can be used to better understand and improve business performance.
- What is Reporting?
- Reporting is
  - "the process of organizing data into informational summaries in order to monitor how different areas of a business are performing."

4.1 COMPARING ANALYSIS WITH REPORTING

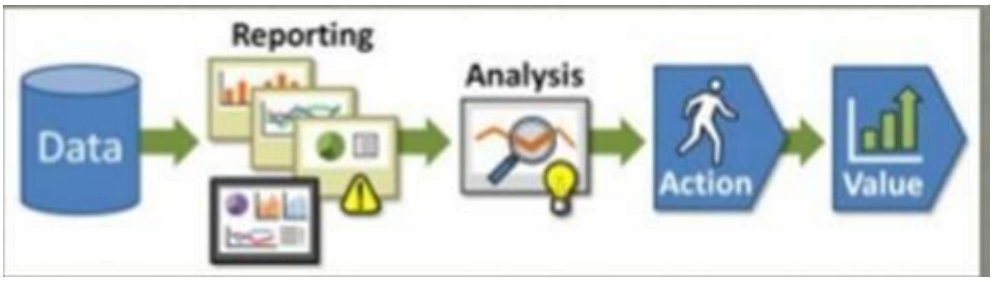
- Measuring core metrics and presenting them — whether in an email, a slide deck, or online dashboard — falls under Reporting.

- **Analytics is** "the process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance."

- **Reporting** helps companies to monitor their online business and be alerted to when data falls outside of expected ranges.

- **Good reporting** should raise questions about the business from its end users.

- **The goal of analysis** is to answer questions by interpreting the data at a deeper level and providing actionable recommendations.



4.2 CONTRAST BETWEEN ANALYSIS AND REPORTING

The basic differences between Analysis and Reporting are as follows:

Analysis	**Reporting**
Provides what is needed	Provides what is asked for
Is typically customized	Is Typically standardized
Involves a person	Does not involve a person
Is extremely flexible	Is Inflexible

- Reporting translates raw data into information, analysis transforms data and information into insights.

- Reporting shows you what is happening while analysis focuses on explaining why it is happening and what you can do about it.

- Reporting and analysis can go hand-in-hand:

- Reporting provides no limited context about what is happening in the data. Context is critical to good analysis.

- Reporting usually raises a question – What is happening?

- Analysis transforms the data into insights - Why is it happening? What can you do about it?

5. Sampling Fundamentals

Sampling may be defined as the selection of some part of an aggregate or totality on the basis of which a judgement or inference about the aggregate or totality is made. In other words, it is the process of obtaining information about an entire population by examining only a part of it. In most of the research work and surveys, the usual approach happens to be to make generalisations or to draw inferences based on samples about the parameters of population from which the samples are taken. The researcher quite often selects only a few items from the universe for his study purposes. All this is done on the assumption that the sample data will enable him to estimate the population parameters. The items so selected constitute what is technically called a sample, their selection process or technique is called sample design and the survey conducted on the basis of sample is described as sample survey. Sample should be truly representative of population characteristics without any bias so that it may result in valid and reliable conclusions.

5.1 NEED FOR SAMPLING

Sampling is used in practice for a variety of reasons such as:

1. Sampling can save time and money. A sample study is usually less expensive than a census study and produces results at a relatively faster speed.

2. Sampling may enable more accurate measurements for a sample study is generally conducted by trained and experienced investigators.

3. Sampling remains the only way when population contains infinitely many members.

4. Sampling remains the only choice when a test involves the destruction of the item under study.

5. Sampling usually enables to estimate the sampling errors and, thus, assists in obtaining information concerning some characteristic of the population.

5.2 SOME FUNDAMENTAL DEFINITIONS

1. Universe/Population: From a statistical point of view, the term ‘Universe ‘refers to the total of the items or units in any field of inquiry, whereas the term ‘population’ refers to the total of items about which information is desired. The attributes that are the object of study are referred to as characteristics and the units possessing them are called as elementary units. The aggregate of such units is generally described as population. Thus, all units in any field of inquiry constitute universe and all elementary units (on the basis of one characteristic or more) constitute population. Quit often, we do not find any difference between population and universe, and as such the two terms are taken as interchangeable. However, a researcher must necessarily define these terms precisely.

2. Sampling frame: The elementary units or the group or cluster of such units may form the basis of sampling process in which case they are called as sampling units. A list containing all such sampling units is known as sampling frame. Thus, sampling frame consists of a list of items from which the sample is to be drawn. If the population is finite and the time frame is in the present or past, then it is possible for the frame to be identical with the population. In most cases they are not identical because it is often impossible to draw a sample directly from population. As such this frame is either constructed by a researcher for the purpose of his study or may consist of some existing list of the population. For instance, one can use telephone directory as a frame for conducting opinion survey in a city. Whatever the frame may be, it should be a good representative of the population.

3. Sampling design: A sample design is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or the procedure the researcher would adopt in selecting some sampling units from which inferences about the population is drawn. Sampling design is determined before any data are collected.

4. Statistic(s) and parameter(s): A statistic is a characteristic of a sample, whereas a parameter is a characteristic of a population. Thus, when we work out certain measures such as mean, median, mode or the like ones from samples, then they are called statistic(s) for they describe the characteristics of a sample. But when such measures describe the characteristics of a population, they are known as parameter(s). For instance, the population mean b g μ is a parameter, whereas the sample mean (X) is a statistic. To obtain the estimate of a parameter from a statistic constitutes the prime objective of sampling analysis.

5. Sampling error: Sample surveys do imply the study of a small portion of the population and as such there would naturally be a certain amount of inaccuracy in the information collected. This inaccuracy may be termed as sampling error or error variance. In other words, sampling errors are those errors which arise on account of sampling, and they generally happen to be random variations (in case of random sampling) in the sample estimates around the true population values.

$$\text{Sampling error} = \text{Frame error} + \text{Chance error} + \text{Response error}$$

6. Precision: Precision is the range within which the population average (or other parameter) will lie in accordance with the reliability specified in the confidence level as a percentage of the estimate ± or as a numerical quantity. For instance, if the estimate is Rs 4000 and the precision desired is ± 4%, then the true value will be no less than Rs 3840 and no more than Rs 4160. This is the range (Rs 3840 to Rs 4160) within which the true answer should lie. But if we desire that the estimate should not deviate from the actual value by more than Rs 200 in either direction, in that case the range would be Rs 3800 to Rs 4200.

7. Confidence level and significance level: The confidence level or reliability is the expected percentage of times that the actual value will fall within the stated precision limits. Thus, if we take a confidence level of 95%, then we mean that there are 95 chances in 100 (or .95 in 1) that the sample results represent the true condition of the population within a specified precision range against 5 chances in 100 (or .05 in 1) that it does not. Precision is the range within which the answer may vary and still be acceptable; confidence level indicates the likelihood that the answer will fall within that range, and the significance level indicates the likelihood that the answer will fall outside that range. We can always remember that if the confidence level is 95%, then the significance level will be (100 – 95) i.e., 5%; if the confidence level is 99%, the significance level is (100 – 99) i.e., 1%, and so on. We should also remember that the area of normal curve within precision limits for the specified confidence level constitute the acceptance region and the area of the curve outside these limits in either direction constitutes the rejection regions.

8. Sampling distribution: We are often concerned with sampling distribution in sampling analysis. If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the particular statistic, say mean. Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure. It may be noted that each item in a sampling distribution is a particular statistic of a sample. The sampling distribution tends quite closer to the normal distribution if the number of samples is large. The significance of sampling distribution follows from the fact that the mean of a sampling distribution is the same as the mean of the universe. Thus, the mean of the sampling distribution can be taken as the mean of the universe.

5.3 IMPORTANT SAMPLING DISTRIBUTIONS:

Some important sampling distributions, which are commonly used, are:

(1) sampling distribution of mean.

(2) sampling distribution of proportion.

(3) student's 't' distribution.

(4) F distribution; and

(5) Chi-square distribution.

A brief mention of each one of these sampling distributions will be helpful.

1. Sampling distribution of mean: Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population. If samples are taken from a normal population, N(μ, σ_p), the sampling distribution of mean would also be normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation = $\sigma p\sqrt{n}$, where μ is the mean of the population, σ p is the standard deviation of the population and n means the number of items in a sample. But when sampling is from a population which is not normal (may be positively or negatively skewed), even then, as per the central limit theorem, the sampling distribution of mean tends quite closer to the normal distribution, provided the number of sample items is large i.e., more than 30. In case we want to reduce the sampling distribution of mean to unit normal distribution i.e., N (0,1), we can write the normal variate $z = \frac{\bar{x}-\mu}{\sigma_p/\sqrt{n}}$ for the sampling distribution of mean. This characteristic of the sampling distribution of mean is very useful in several decision situations for accepting or rejection of hypotheses.

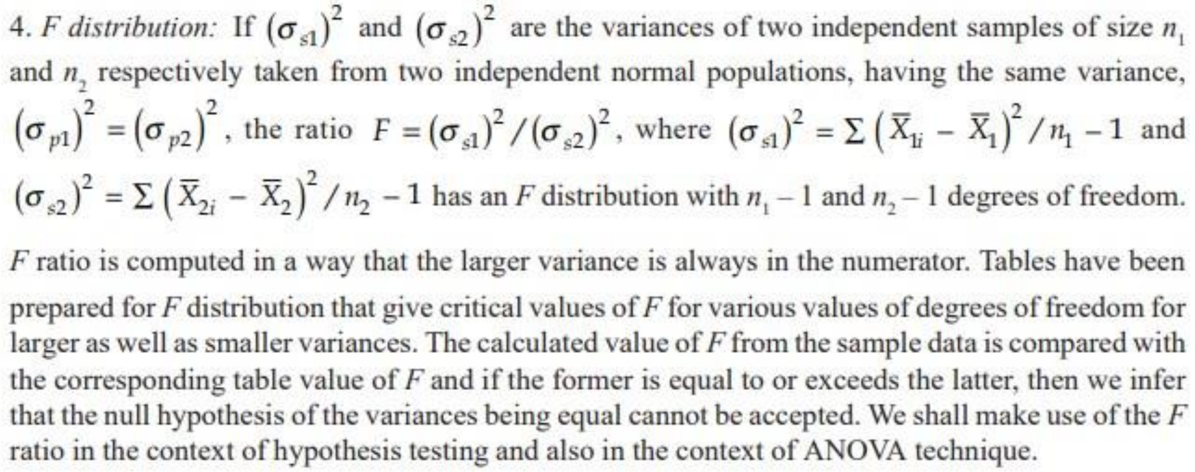
2. Sampling distribution of proportion: Like sampling distribution of mean, we can as well have a sampling distribution of proportion. This happens in case of statistics of attributes. Assume that we have worked out the proportion of defective parts in large number of samples, each with say 100 items, that have been taken from an infinite population and plot a probability distribution of the said proportions, we obtain what is known as the sampling distribution of the said proportions, we obtain what is known as the sampling distribution of proportion. Usually, the statistics of attributes correspond to the conditions of a binomial distribution that tends to become normal distribution as n becomes larger and larger. If p represents the proportion of defectives i.e., of successes and q the proportion of non-defectives i.e., of failures (or q = 1 – p) and if p is treated as a random variable, then the sampling distribution of proportion of successes has a mean = p with standard deviation = $\sqrt{\frac{p \cdot q}{n}}$, where n is the sample size. Presuming the binomial distribution approximating the normal distribution for large n, the normal variate of the sampling distribution of proportion $z = \frac{\hat{p}-p}{\sqrt{(p \cdot q)/n}}$, where $\hat{p}$ (pronounced as p-hat) is the sample proportion of successes, can be used for testing of hypotheses.

3. Student's t-distribution: When population standard deviation $(\sigma_p)$is not known and the sample is of a small size i.e., $n \leq 30$, we use t distribution for the sampling distribution of mean and workout t variable as:

$$t = (\bar{X} - \mu) / (\sigma_s / \sqrt{n})$$

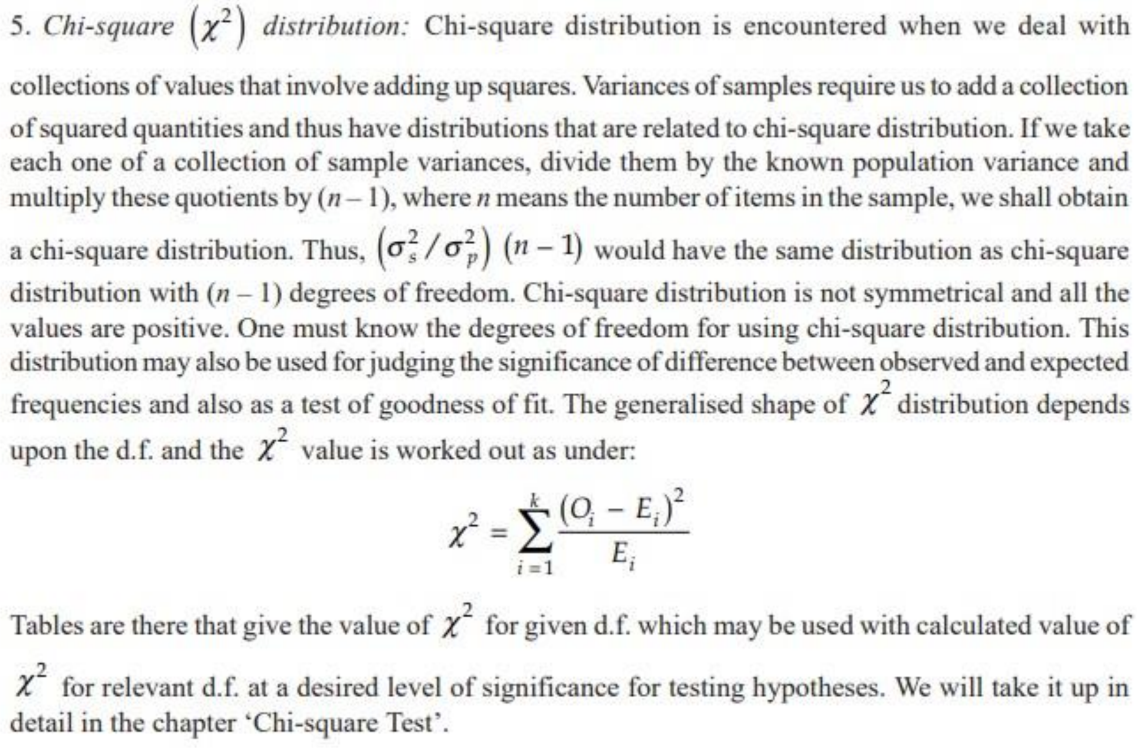
Where $\sigma_s = \sqrt{\frac{\Sigma(X_i-\bar{X})^2}{n} - 1}$

i.e., the sample standard deviation. t-distribution is also symmetrical and is very close to the distribution of standard normal variate, z, except for small values of n. The variable t differs from z in the sense that we use sample standard deviation σ_S in the calculation of t, whereas we use standard deviation of population (σ_p) in the calculation of z. There is a different t distribution for every possible sample size i.e., for different degrees of freedom. The degrees of freedom for a sample of size n is *n – 1*. As the sample size gets larger, the shape of the t distribution becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the t distribution is so close to the normal distribution that we can use the normal to approximate the t-distribution. But when n is small, the t-distribution is far from normal but when $n \rightarrow \alpha$, t-distribution is identical with normal distribution. The t-distribution tables are available which give the critical values of t for different degrees of freedom at various levels of significance. The table value of t for given degrees of freedom at a certain level of significance is compared with the calculated value of t from the sample data, and if the latter is either equal to or exceeds, we infer that the null hypothesis cannot be accepted.



4. *F distribution:* If $(\sigma_{s1})^2$ and $(\sigma_{s2})^2$ are the variances of two independent samples of size n_1 and n_2 respectively taken from two independent normal populations, having the same variance, $(\sigma_{p1})^2 = (\sigma_{p2})^2$, the ratio $F = (\sigma_{s1})^2 / (\sigma_{s2})^2$, where $(\sigma_{s1})^2 = \Sigma\left(\overline{X}_{1i} - \overline{X}_1\right)^2 / n_1 - 1$ and $(\sigma_{s2})^2 = \Sigma\left(\overline{X}_{2i} - \overline{X}_2\right)^2 / n_2 - 1$ has an *F* distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

F ratio is computed in a way that the larger variance is always in the numerator. Tables have been prepared for *F* distribution that give critical values of *F* for various values of degrees of freedom for larger as well as smaller variances. The calculated value of *F* from the sample data is compared with the corresponding table value of *F* and if the former is equal to or exceeds the latter, then we infer that the null hypothesis of the variances being equal cannot be accepted. We shall make use of the *F* ratio in the context of hypothesis testing and also in the context of ANOVA technique.



5. *Chi-square* (χ^2) *distribution:* Chi-square distribution is encountered when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and thus have distributions that are related to chi-square distribution. If we take each one of a collection of sample variances, divide them by the known population variance and multiply these quotients by $(n - 1)$, where *n* means the number of items in the sample, we shall obtain a chi-square distribution. Thus, $\left(\sigma_s^2 / \sigma_p^2\right)(n-1)$ would have the same distribution as chi-square distribution with $(n - 1)$ degrees of freedom. Chi-square distribution is not symmetrical and all the values are positive. One must know the degrees of freedom for using chi-square distribution. This distribution may also be used for judging the significance of difference between observed and expected frequencies and also as a test of goodness of fit. The generalised shape of χ^2 distribution depends upon the d.f. and the χ^2 value is worked out as under:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Tables are there that give the value of χ^2 for given d.f. which may be used with calculated value of χ^2 for relevant d.f. at a desired level of significance for testing hypotheses. We will take it up in detail in the chapter 'Chi-square Test'.

6. Re-Sampling

Resampling is the method that consists of drawing repeated samples from the original data samples. The method of Resampling is a nonparametric method of statistical inference. In other words, the method of resampling does not involve the utilization of the generic distribution tables (for example, normal distribution tables) in order to compute approximate p probability values.

Resampling involves the selection of randomized cases with replacement from the original data sample in such a manner that each number of the sample drawn has a number of cases that are similar to the original data sample. Due to replacement, the drawn number of samples that are used by the method of resampling consists of repetitive cases.

Resampling generates a unique sampling distribution on the basis of the actual data. The method of resampling uses experimental methods, rather than analytical methods, to generate the unique sampling distribution. The method of resampling yields unbiased estimates as it is based on the unbiased samples of all the possible results of the data studied by the researcher.

Specific Resampling Techniques

The main techniques are:

1. Bootstrapping and Normal resampling (sampling from a normal distribution).

2. Permutation Resampling (also called Rearrangements or Rerandomization),

3. Cross Validation.

1. Bootstrapping and Normal Resampling

Bootstrapping is a type of resampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample. Normal resampling is very similar to bootstrapping as it is a special case of the normal shift model—one of the assumptions for bootstrapping. Both bootstrapping and normal resampling both assume that samples are drawn from an actual population (either a real one or a theoretical one). Another similarity is that both techniques use sampling with replacement.

Ideally, you would want to draw large, non-repeated, samples from a population in order to create a sampling distribution for a statistic. However, limited resources may prevent you from getting the ideal statistic. Resampling means that you can draw small samples over and over again from the same population. As well as saving time and money, the samples can be quite good approximations for population parameters.

2. Permutation Resampling

Unlike bootstrapping, permutation resampling doesn't need any "population"; resampling is dependent only on the assignment of units to treatment groups. The fact that you're dealing with actual samples, instead of populations, is one reason why it's sometimes referred to as the Gold standard bootstrapping technique. Another important difference is that permutation resampling is a without replacement sampling technique.

3. Cross Validation

Cross-validation is a way to validate a predictive model. Subsets of the data are removed to be used as a validating set; the remaining data is used to form a training set, which is used to predict the validation set.

7. Statistical Inference

Using data analysis and statistics to make conclusions about a population is called statistical inference.

The main types of statistical inference are:

1. Estimation

2. Hypothesis testing

1. Estimation

Statistics from a sample are used to estimate population parameters.

- The most likely value is called a point estimate.
- There is always uncertainty when estimating.
- The uncertainty is often expressed as confidence intervals defined by a likely lowest and highest value for the parameter.
- An example could be a confidence interval for the number of bicycles a Dutch person owns:

> "The average number of bikes a Dutch person owns is between 3.5 and 6."

2. Hypothesis Testing

- Hypothesis testing is a method to check if a claim about a population is true. More precisely, it checks how likely it is that a hypothesis is true is based on the sample data.
- There are different types of hypothesis testing.
- The step of the test depends on:
    - Type of data (categorical or numerical)
    - If you are looking at:
        - A single group
        - Comparing one group to another
        - Comparing the same group before and after a change

Some examples of claims or questions that can be checked with hypothesis testing:

- 90% of Australians are left-handed.
- Is the average weight of dogs more than 40kg?
- Do doctors make more money than lawyers?

8. Prediction Error

Prediction error quantifies one of two things:

- In regression analysis, it's a measure of how well the model predicts the response variable.
- In classification (machine learning), it's a measure of how well samples are classified to the correct category.

Sometimes, the term is used informally to mean exactly what it means in plain English (you've made some predictions, and there are some errors). In regression, the term "prediction error" and "Residuals" are sometimes used synonymously. Therefore, check the author's intent before assuming they mean something specific (like the mean squared prediction error).

Mean Squared Prediction Error (MSPE)

MSPE summarizes the predictive ability of a model. Ideally, this value should be close to zero, which means that your predictor is close to the true value. The concept is similar to Mean Squared Error (MSE), which is a measure of the how well an estimator measures a parameter (or how close a regression line is to a set of points). The difference is that while MSE measures of an estimator's fit, the MSPE is a measure of a predictor's fit— or how well it predicts the true value.

Quantifying Prediction Errors

Prediction error can be quantified in several ways, depending on where you're using it. In general, you can analyze the behaviour of prediction error with bias and variance.

In statistics, the root-mean-square error (RMSE) aggregates the magnitudes of prediction errors. The Rao-Blackwell theory can estimate prediction error as well as improve the efficiency of initial estimators.

In machine learning, Cross-validation (CV) assesses prediction error and trains the prediction rule. A second method, the bootstrap, begins by estimating the prediction rule's sampling distribution (or the sampling distribution's parameters); It can also quantify prediction error and other aspects of the prediction rule.