MODULE-1 INTRODUCTION TO DATA SCIENCE

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Data Science Lifecycle:

The steps in Data Science includes:

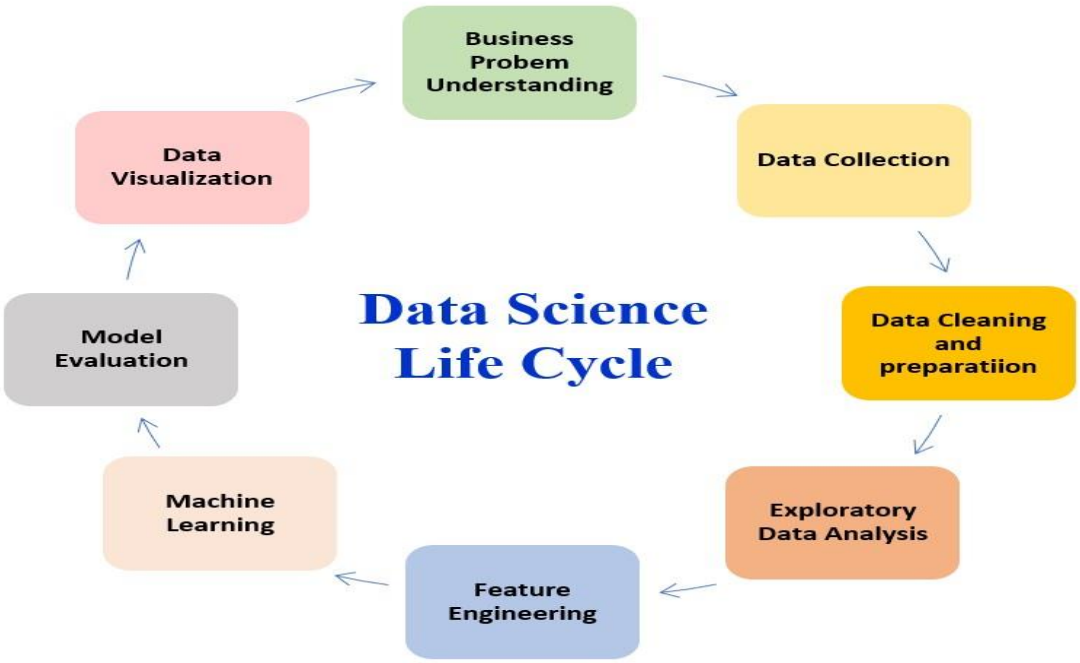
1. Capture: Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.

2. Maintain: Data Warehousing, Data Cleansing, Data Staging, Data Processing, and Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.

3. Process: Data Mining, Clustering/Classification, Data Modelling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.

4. Analyze: Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, and Qualitative Analysis. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.

5. Communicate: Data Reporting, Data Visualization, Business Intelligence, and Decision Making. In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.



- Data Acquisition or Collection: Here, data scientists take data from all its raw sources, such as databases and flat-files. Then, they integrate and transform it into a homogenous format, collecting it into what is known as a "data warehouse," a system by which the data can be used to extract information from easily. Also known as ETL, this step can be done with some tools, such as Talend Studio, DataStage and Informatica.

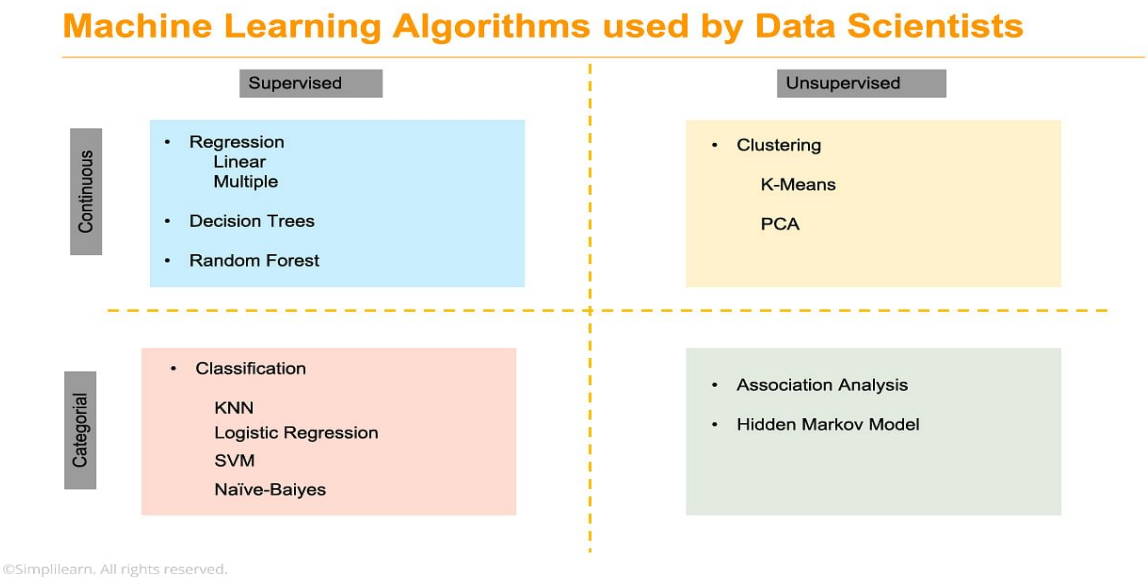
- Data Preparation: This is the most important stage, wherein 60 percent of a data scientist's time is spent because often data is "dirty" or unfit for use and must be scalable, productive and meaningful. In fact, five sub-steps exist here:

- Data Cleaning: Important because bad data can lead to bad models, this step handles missing values and null or void values that might cause the models to fail. Ultimately, it improves business decisions and productivity.

- Data Transformation: Takes raw data and turns it into desired outputs by normalizing it. This step can use, for example, min-max normalization or z-score normalization.

- Handling Outliers: This happens when some data falls outside the scope of the realm of the rest of the data. Using exploratory analysis, a data scientist quickly uses plots and graphs to determine what to do with the outliers and see why they're there. Often, outliers are used for fraud detection.

- Data Integration: Here, the data scientist ensures the data is accurate and reliable.
- Data Reduction: This compiles multiple sources of data into one, increases storage capabilities, reduces costs and eliminates duplicate, redundant data.
- Data Mining: Here, data scientists uncover the data patterns and relationships to take better business decisions. It's a discovery process to get hidden and useful knowledge, commonly known as exploratory data analysis. Data mining is useful for predicting future trends, recognizing customer patterns, helping to make decisions, quickly detecting fraud and choosing the correct algorithms. Tableau works nicely for data mining.
- Model Building: This goes further than simple data mining and requires building a machine learning model. The model is built by selecting a machine learning algorithm that suits the data, problem statement and available resources



- There are two types of machine learning algorithms: Supervised and Unsupervised:

1. Supervised: Supervised learning algorithms are used when the data is labelled. There are two types:

- Regression: When you need to predict continuous values and variables are linearly dependent, algorithms used are linear and multiple regression, decision trees and random forest

- Classification: When you need to predict categorical values, some of the classification algorithms used are KNN, logistic regression, SVM and Naïve-Bayes

2. Unsupervised: Unsupervised learning algorithms are used when the data is unlabelled, there is no labelled data to learn from. There are two types:

- Clustering: This is the method of dividing the objects which are similar between them and dissimilar to others. K-Means and PCA clustering algorithms are commonly used.
- Association-rule analysis: This is used to discover interesting relations between variables, Apriori and Hidden Markov Model algorithm can be used
- Model Maintenance: After gathering data and performing the mining and model building, data scientists must maintain the model accuracy. Thus, they take the following steps:

1. Assess: Running a sample through the data occasionally to make sure it remains accurate
2. Retrain: When the results of the reassessment aren't right, the data scientist must retrain the algorithm to provide the correct results again
3. Rebuild: If retraining fails, rebuilding must occur.

As you can see, data science is a complex process of various steps taking massive effort to achieve continuous, excellent results.

Applications of Data Science

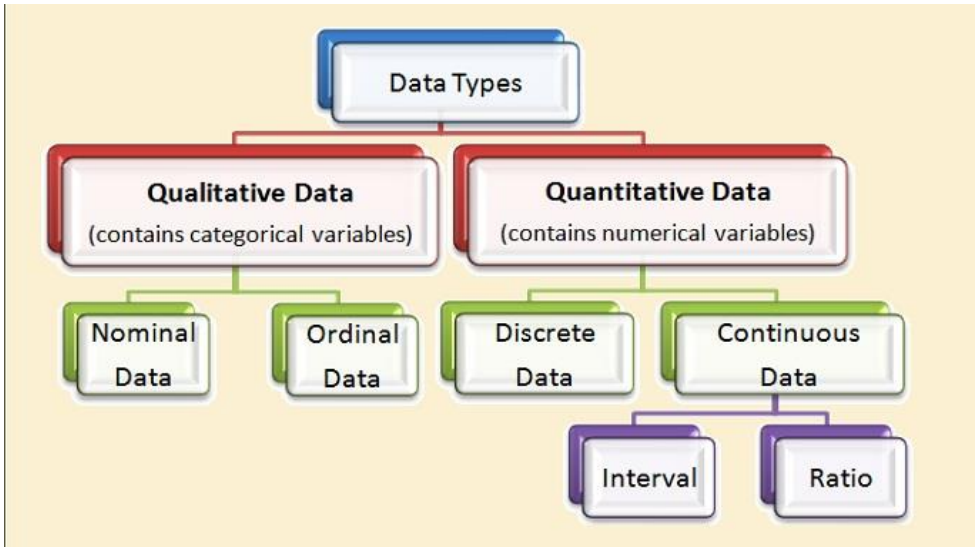
1. Data science detects patterns in seemingly unstructured or unconnected data, allowing conclusions and predictions to be made.
2. Tech businesses that acquire user data can utilise strategies to transform that data into valuable or profitable information.
3. Data Science has also made inroads into the transportation industry, such as with driverless cars. It is simple to lower the number of accidents with the use of driverless cars. For example, with driverless cars, training data is supplied to the algorithm, and the data is examined using data Science approaches, such as the speed limit on the highway, busy streets, etc.

4. Data Science applications provide a better level of beneficial customization through genetics and genomics research.

Structured data stands for information that is highly organized, factual, and to-the-point. It usually comes in the form of letters and numbers that fit nicely into the rows and columns of tables. Structured data commonly exists in tables similar to Excel files and Google Docs spreadsheets.

Unstructured data doesn't have any pre-defined structure to it and comes in all its diversity of forms. The examples of unstructured data vary from imagery and text files like PDF documents to video and audio files, to name a few.

4 Types of Data:



Qualitative Data Type

Qualitative or Categorical Data describes the object under consideration using a finite set of discrete classes. It means that this type of data can't be counted or measured easily using numbers and therefore divided into categories. The gender of a person (male, female, or others) is a good example of this data type.

These are usually extracted from audio, images, or text medium. Another example can be of a smartphone brand that provides information about the current rating, the color of the phone, category of the phone, and so on. All this information can be categorized as Qualitative data. There are two subcategories under this:

Nominal:

These are the set of values that don't possess a natural ordering. Let's understand this with some examples. The color of a smartphone can be considered as a nominal data type as we can't compare one color with others.

It is not possible to state that 'Red' is greater than 'Blue'. The gender of a person is another one where we can't differentiate between male, female, or others. Mobile phone categories whether it is midrange, budget segment, or premium smartphone is also nominal data type.

Ordinal

These types of values have a natural ordering while maintaining their class of values. If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of small < medium < large. The grading system while marking candidates in a test can also be considered as an ordinal data type where A+ is definitely better than B grade.

These categories help us deciding which encoding strategy can be applied to which type of data. Data encoding for Qualitative data is important because machine learning models can't handle these values directly and needed to be converted to numerical types as the models are mathematical in nature.

Quantitative Data Type:

This data type tries to quantify things and it does by considering numerical values that make it countable in nature. The price of a smartphone, discount offered, number of ratings on a product, the frequency of processor of a smartphone, or ram of that particular phone, all these things fall under the category of Quantitative data types.

Discrete

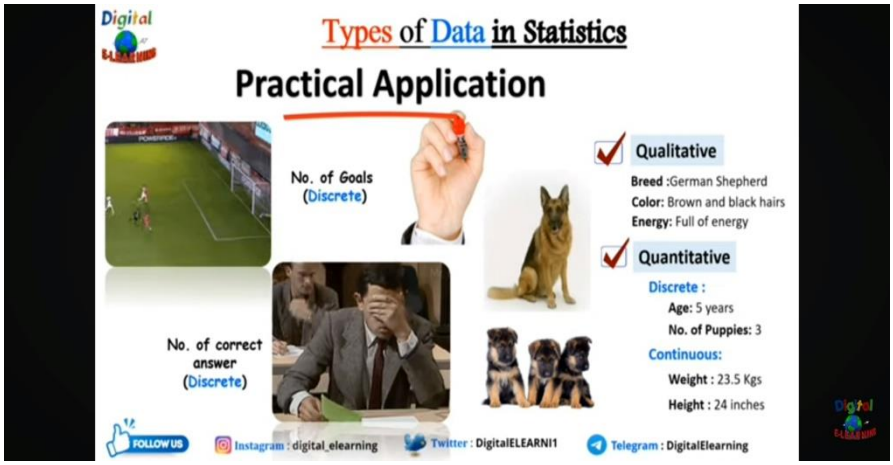
The numerical values which fall under are integers or whole numbers are placed under this category. The number of speakers in the phone, cameras, cores in the processor, the number of sims supported all these are some of the examples of the discrete data type.

Discrete data types in statistics cannot be measured – it can only be counted as the objects included in discrete data have a fixed value. The value can be represented in decimal, but it has to be whole. Discrete data is often identified through charts, including bar charts, pie charts, and tally charts.

Continuous

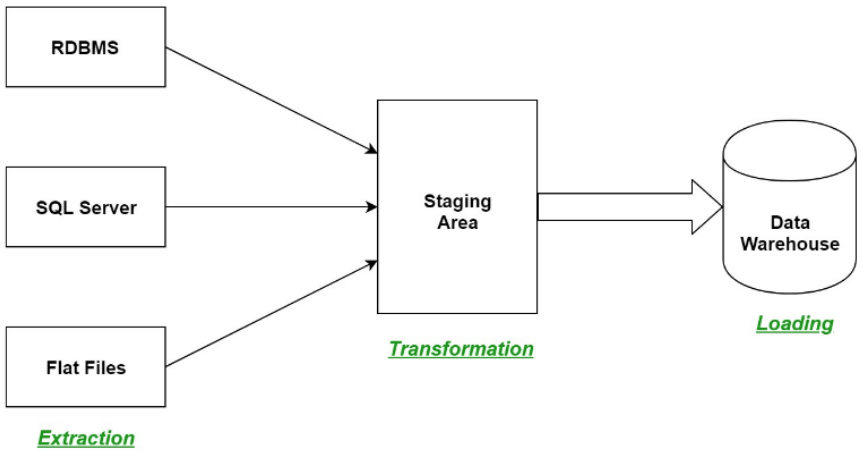
The fractional numbers are considered as continuous values. These can take the form of the operating frequency of the processors, the android version of the phone, wifi frequency, temperature of the cores, and so on.

Unlike discrete data types of data in research, with a whole and fixed value, continuous data can break down into smaller pieces and can take any value. For example, volatile values such as temperature and the weight of a human can be included in the continuous value. Continuous types of statistical data are represented using a graph that easily reflects value fluctuation by the highs and lows of the line through a certain period of time.



Data mining and Data warehousing:

Data warehousing is a method of organizing and compiling data into one database, whereas data mining deals with fetching important data from databases. Data mining attempts to depict meaningful patterns through a dependency on the data that is compiled in the data warehouse.



DATA WAREHOUSE:

A data warehouse is where data can be collected for mining purposes, usually with large storage capacity. Various organizations' systems are in the data warehouse, where it can be fetched as per usage.

Source □ Extract □Transform □ Load □ Target.

(Data warehouse process)

Data warehouses collaborate data from several sources and ensure data accuracy, quality, and consistency. System execution is boosted by differentiating the process of analytics from traditional databases. In a data warehouse, data is sorted into a formatted pattern by type and as needed. The data is examined by query tools using several patterns.

Data warehouses store historical data and handle requests faster, helping in online analytical processing, whereas a database is used to store current transactions in a business process that is called online transaction processing.

FEATURES OF DATA WAREHOUSES:

- Subject Oriented:
  It provides you with important data about a specific subject like suppliers, products, promotion, customers, etc. Data warehousing usually handles the analysis and modeling of data that assist any organization to make data-driven decisions.

- Integrated:
  Different heterogeneous sources are put together to build a data warehouse, such as level documents or social databases.

- Time-Variant:
  The data collected in a data warehouse is identified with a specific period.

- Nonvolatile:
  This means the earlier data is not deleted when new data is added to the data warehouse. The operational database and data warehouse are kept separate and thus continuous changes in the operational database are not shown in the data warehouse.

APPLICATIONS OF DATA WAREHOUSES:

Data warehouses help analysts or senior executives analyze, organize, and use data for decision making.

It is used in the following fields:

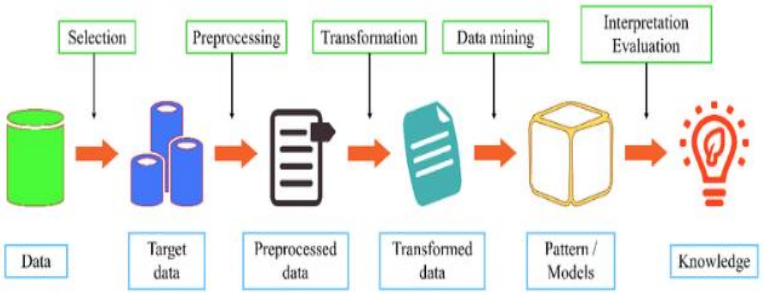
- Consumer goods
- Banking services
- Financial services
- Manufacturing
- Retail sectors

ADVANTAGES OF DATA WAREHOUSING:

- Cost-efficient and provides quality of data
- Performance and productivity are improved
- Accurate data access and consistency.

DATA MINING:

In this process, data is extracted and analysed to fetch useful information. In data mining hidden patterns are researched from the dataset to predict future behavior. Data mining is used to indicate and discover relationships through the data. Data mining uses statistics, artificial intelligence, machine learning systems, and some databases to find hidden patterns in the data. It supports business-related queries that are time-consuming to resolve.



FEATURES OF DATA MINING:

- It is good with large databases and datasets
- It predicts future results
- It creates actionable insights
- It utilizes the automated discovery of patterns

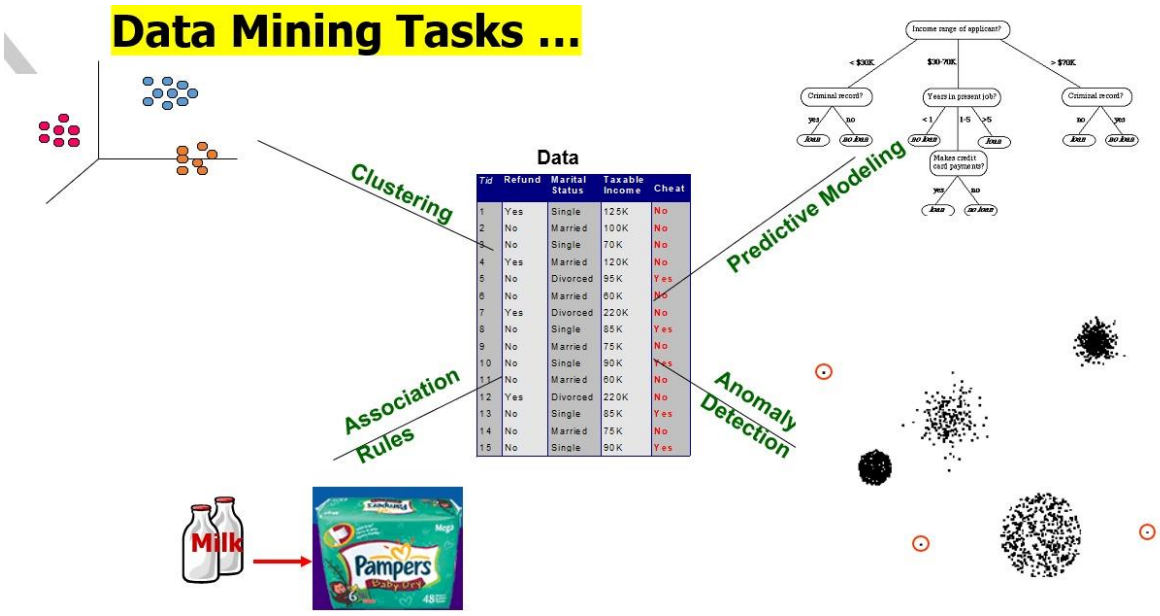
ADVANTAGES OF DATA MINING:

- Fraud Detection:
  It is used to find which insurance claims, phone calls, debit or credit purchases are fraud.

- Trend Analysis:
  Existing marketplace trends are analysed, which provides a strategic benefit as it helps in reduction of costs, as in manufacturing per demand.

- Market Analysis:
  It can predict the market and therefore help to make business decisions. For example: it can identify a target market for a retailer, or certain types of products desired by types of customers.

DATA MINING TECHNIQUES:



- Classification:
  It is used to fetch the appropriate information from the dataset and to segregate different classes that are present in the dataset. Below are the classification models.

1. K-nearest neighbors
2. Support Vector Machine
3. Gaussian Naïve Bayes, etc.

- Clustering:
  It is used to find similarities in data by putting related data together and helping to identify different variations in the dataset. It helps to find hidden patterns. An example of clustering is text mining, medical diagnostics, etc.

- Association Rules:
  They are used to identify a connection of two or more items. For example, if-then scenarios of items that are frequently purchased in tandem in a grocery store can calculate the proportion of items that are bought by customers together. Lift, confidence, and support are techniques used in association rules.

- Outlier Detection:
  It is used to identify patterns that do not match the normal behavior in the data, as the outlier deviates from the rest of the data points. It helps in fraud detection, intrusion, etc. Boxplot and z-score are ways to detect outliers.

Descriptive Analytics:

Descriptive analytics is the interpretation of historical data to better understand changes that have occurred in a business. Descriptive analytics describes the use of a range of historic data to draw comparisons. Most commonly reported financial metrics are a product of descriptive analytics, for example, year-over-year pricing changes, month-over-month sales growth, the number of users, or the total revenue per subscriber. These measures all describe what has occurred in a business during a set period.

- Descriptive analytics is the process of parsing historical data to better understand the changes that have occurred in a business.
- Using a range of historic data and benchmarking, decision-makers obtain a holistic view of performance and trends on which to base business strategy.
- Descriptive analytics can help to identify the areas of strength and weakness in an organization.
- Examples of metrics used in descriptive analytics include year-over-year pricing changes, month-over-month sales growth, the number of users, or the total revenue per subscriber.

- Descriptive analytics is now being used in conjunction with newer analytics, such as predictive and prescriptive analytics.

Working of Descriptive Analytics:

Data aggregation and **data mining** are two techniques used in descriptive analytics to discover historical data. Data is first gathered and sorted by data aggregation in order to make the datasets more manageable by analysts. Data mining describes the next step of the analysis and involves a search of the data to identify patterns and meaning. Identified patterns are analysed to discover the specific ways that learners interacted with the learning content and within the learning environment.

Examples of descriptive analytics:

Many LMS platforms and learning systems offer descriptive analytical reporting with the aim of help businesses and institutions measure learner performance to ensure that training goals and targets are met. The findings from descriptive analytics can quickly identify areas that require improvement - whether that be improving learner engagement or the effectiveness of course delivery.

Here are some examples of how descriptive analytics is being used in the field of learning analytics:

- Tracking course enrollments, course compliance rates, Recording which learning resources are accessed and how often
- Summarizing the number of times a learner posts in a discussion board
- Tracking assignment and assessment grades
- Comparing pre-test and post-test assessments
- Analysing course completion rates by learner or by course
- Collating course survey results

- Identifying length of time that learners took to complete a course.

Advantages of descriptive analytics

When learners engage in online learning, they leave a digital trace behind with every interaction they have in the learning environment.

This means that descriptive analytics in online learning can gain insight into behaviours and performance indicators that would otherwise not be known.

Here are some advantages to utilizing this information:

- Quickly and easily report on the Return on Investment (ROI) by showing how performance achieved business or target goals.
- Identify gaps and performance issues early - before they become problems.
- Identify specific learners who require additional support, regardless of how many students or employees there are.
- Identify successful learners in order to offer positive feedback or additional resources.
- Analyze the value and impact of course design and learning resources.

Probability Theory:

Statistics

Descriptive statistics involves summarizing and organizing the data so they can be easily understood. Descriptive statistics, unlike inferential statistics, seeks to describe the data, but do not attempt to make inferences from the sample to the whole population. Here, we typically describe the data in a sample. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory. Descriptive statistics are broken down into two categories. Measures of central tendency and measures of variability (spread).

Measure of Central Tendency

Central tendency refers to the idea that there is one number that best summarizes the

entireset of measurements, a number that is in some way “central” to the set.

Mean / Median /Mode/ Variance /Standard Deviation are all very basic but very important concept of statistics used in data science. Almost all the machine learning algorithm uses these concepts in data pre-processing steps. These concepts are part of descriptive statistics where we basically used to describe and understand the data for features in Machine learning

Mean :

Mean is also known as average of all the numbers in the data set which is calculated by below equation.

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

Symbolically,

$$\overline{x} = \frac{\sum x}{n}$$

where $\overline{x}$ (read as 'x bar') is the mean of the set of x values,
$\sum x$ is the sum of all the x values, and
n is the number of x values.

Lets say we have below heights of persons.

heights=[168,170,150,160,182,140,175,191,152,150]

Mean of dataset

```
In [41]: import numpy as np
         heights=[168,170,150,160,182,140,175,191,152,150]

In [42]: mean=np.mean(heights)
         mean

Out[42]: 163.8
```

Median :

Median is mid value in this ordered data set.

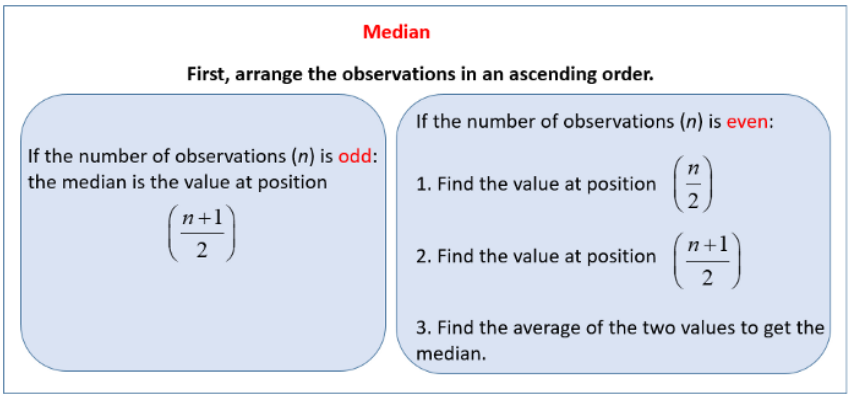


image:source unknown

Arrange the data in the increasing order and then find the mid value.

Sort Data in increasing order

```
In [33]: heights.sort()

In [34]: heights
Out[34]: [140, 150, 150, 152, 160, 168, 170, 175, 182, 191]
```

If we have even number of values in the data set then median is sum of mid two numbers divided by 2

Median of dataset

```
In [37]: median=np.median(heights)

In [38]: median
Out[38]: 164.0
```

In we have odd number in the data set like below we have 9 heights the median will be 5th number value.

```
In [43]: import numpy as np
         heights=[168,170,150,160,182,140,175,191,152]
```

Sort Data in increasing order

```
In [46]: heights.sort()

In [47]: heights
Out[47]: [140, 150, 152, 160, 168, 170, 175, 182, 191]
```

Median of dataset

```
In [44]: median=np.median(heights)

In [45]: median
Out[45]: 168.0
```

Mode :

Mode is the number which occur most often in the data set.Here 150 is occurring twice so this is our mode.

Mode in datset

```
In [49]: import numpy as np
         heights=[168,170,150,160,182,140,175,191,152,150]

In [50]: import statistics as stats
         stats.mode(heights)
Out[50]: 150
```

Variance :

Variance is the numerical values that describe the variability of the observations from its arithmetic mean and denoted by sigma-squared(σ2)

Variance measure how far individuals in the group are spread out, in the set of data from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$$

Where Xi : Elements in the data set mu : the population mean =the population mean

Step 1: This formula says that take each element from dataset(population) and subtract from mean of data set.Later sum all the values.

Step 2: Take the sum in Step 1 and divide by total number of elements.

Square in the above formula will nullify the effect of negative sign(-)

Variance

```
In [51]: import numpy as np
         heights=[168,170,150,160,182,140,175,191,152,150]

In [52]: np.var(heights)
Out[52]: 235.35999999999999
```

Standard Deviation :

It is a measure of dispersion of observation within dataset relative to their mean.It is square root of the variance and denoted by Sigma (σ) .

Standard deviation is expressed in the same unit as the values in the dataset so it measure how much observations of the data set differs from its mean.

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

σ = population standard deviation
N = the size of the population
x_i = each value from the population
μ = the population mean

Standard Deviations

```
In [54]: import numpy as np
         heights=[168,170,150,160,182,140,175,191,152,150]

In [56]: std=np.std(heights)

In [57]: std
Out[57]: 15.341447128612085
```

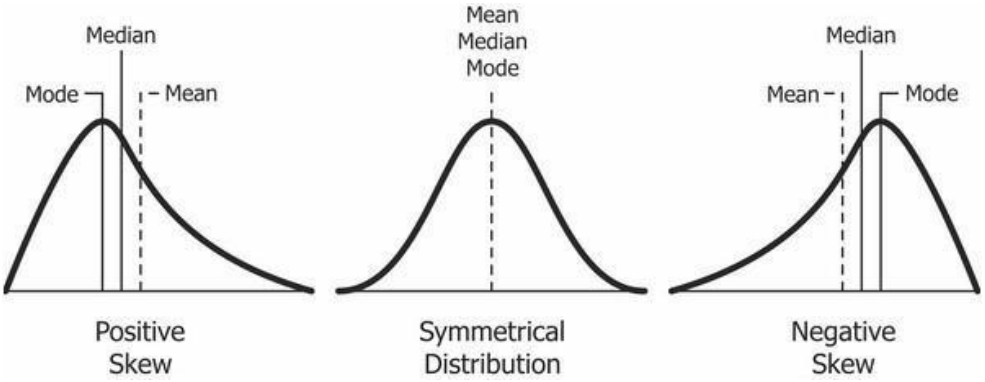
Mean / Median /Mode/ Variance /Standard Deviation are simple yet very important concept in statistics

Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or

undefined. In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other.

When a distribution is skewed to the left, the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative skewness. When a distribution is skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode.



This situation is also called positive skewness.

PROBABILITY DISTRIBUTION:

Probability distribution yields the possible outcomes for any random event. It is also defined based on the underlying sample space as a set of possible outcomes of any random experiment. These settings could be a set of real numbers or a set of vectors or a set of any entities. It is a part of probability and statistics.

Random experiments are defined as the result of an experiment, whose outcome cannot be predicted. Suppose, if we toss a coin, we cannot predict, what outcome it will appear either it will come as Head or as Tail. The possible result of a random experiment is called an outcome. And the set of outcomes is called a sample point. With the help of these experiments or events, we can always create a probability pattern table in terms of variables and probabilities.

Types of Probability Distribution

There are two types of probability distribution which are used for different purposes and various types of the data generation process.

1. Normal or Cumulative Probability Distribution
2. Binomial or Discrete Probability Distribution

Cumulative Probability Distribution

The cumulative probability distribution is also known as a continuous probability distribution. In this distribution, the set of possible outcomes can take on values in a continuous range.

For example, a set of real numbers, is a continuous or normal distribution, as it gives all the possible outcomes of real numbers. Similarly, a set of complex numbers, a set of prime numbers, a set of whole numbers etc. are examples of Normal Probability distribution. Also, in real-life scenarios, the temperature of the day is an example of continuous probability. Based on these outcomes we can create a distribution table. A probability density function describes it. The formula for the normal distribution is;

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where,

-
  -
  - μ = Mean Value
  - σ = Standard Distribution of probability.
  - If mean(μ) = 0 and standard deviation(σ) = 1, then this distribution is known to be normal distribution.
  - x = Normal random variable

Normal Distribution Examples

Since the normal distribution statistics estimates many natural events so well, it has evolved into a standard of recommendation for many probability queries. Some of the examples are:

-
  -

- Height of the Population of the world
- Rolling a dice (once or multiple times)
- To judge the Intelligent Quotient Level of children in this competitive world
- Tossing a coin
- Income distribution in countries economy among poor and rich
- The sizes of females shoes
- Weight of newly born babies range
- Average report of Students based on their performance

Discrete Probability Distribution

A distribution is called a discrete probability distribution, where the set of outcomes are discrete in nature.

For example, if a dice is rolled, then all the possible outcomes are discrete and give a mass of outcomes. It is also known as the probability mass function.

So, the outcomes of binomial distribution consist of n repeated trials and the outcome may or may not occur. The formula for the binomial distribution is;

$$P(x) = \frac{n!}{r!(n-r)!} \cdot p^r (1-p)^{n-r}$$

$$P(x) = C(n, r).p^r (1-p)^{n-r}$$

Where,

-
  -
- n = Total number of events
- r = Total number of successful events.
- p = Success on a single trial probability.
- ${}^nC_r = [n!/r!(n-r)]!$
- 1 – p = Failure Probability

Binomial Distribution Examples

As we already know, binomial distribution gives the possibility of a different set of outcomes. In the real-life, the concept is used for:

- To find the number of used and unused materials while manufacturing a product.
- To take a survey of positive and negative feedback from the people for anything.
- To check if a particular channel is watched by how many viewers by calculating the survey of YES/NO.
- The number of men and women working in a company.
- To count the votes for a candidate in an election and many more.

Confidence Interval:

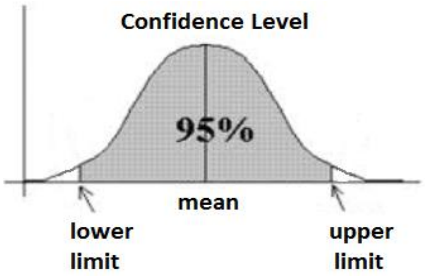
Introduction:

Confidence Interval is a range where we are certain that true value exists. The selection of a confidence level for an interval determines the probability that the confidence interval will contain the true parameter value. The confidence level describes the uncertainty associated with a sampling method.

A confidence interval, in statistics, **refers to the probability that a population parameter will fall between a set of values for a certain proportion of times**. Analysts often use confidence intervals than contain either 95% or 99% of expected observations.

Suppose we used the same sampling method (say sample mean) to compute a different interval estimate for each sample. Some interval estimates would include the true population parameter and some would not.

A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter. A 95% confidence level means that 95% of the intervals would include the population parameter.



Confidence Interval	z-value
90%	1.645
95%	1.960
99%	2.576

Why Are Confidence Intervals Used?

Statisticians use confidence intervals to measure uncertainty in a sample variable. For example, a researcher selects different samples randomly from the same population and computes a confidence interval for each sample to see how it may represent the true value of the population variable. The resulting datasets are all different where some intervals include the true population parameter and others do not

What Is a T-Test?

Confidence intervals are conducted using statistical methods, such as a t-test. A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related to certain features. Calculating a t-test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group.

Confidence Interval for population mean:

How to Construct a Confidence Interval for a Population Mean

Step 1: Identify the sample mean $\bar{x}$, the sample size n, and the sample standard deviation s.

Step 2: Find the degrees of freedom using df=n−1. Then look up the critical value tc from the Student's t-distribution (found here: https://study.com/academy/lesson/critical-values-of-the-t-distribution-statistical-table.html).

Step 3: Calculate the left endpoint of the confidence interval using $a=\bar{x}-tc(s/\sqrt{n})$ and the right endpoint using $b=\bar{x}+tc(s/\sqrt{n})$ **Step 4:** Write the confidence interval (a,b)

What is a Confidence Interval?

Confidence Interval: A confidence interval is a range of estimated values for an unknown population parameter. The confidence interval is calculated using a sample mean and standard deviation. A confidence interval means that you can say with confidence that the actual population parameter value will lie within the interval.

We will use these steps, definitions, and equations to construct a confidence interval for a population mean in the following two examples.

Examples of Constructing a Confidence Interval for a Population Mean

Example 1

A researcher wants to estimate the mean household income in a town of 25,000 households. The researcher takes a simple random sample of 501 households in the town and finds the sample mean household income is $57,250 with a standard deviation of $1,203. Construct a 95% confidence interval for the population mean household income.

Step 1: Identify the sample mean $\bar{x}$, the sample size n, and the sample standard deviation s.

We have:

$\bar{x}=\$57,250$

$n=501$

$s=\$1,203$

Step 2: Find the degrees of freedom using df=n−1. Then look up the critical value tc from the Student's t-distribution (found here: https://study.com/academy/lesson/critical-values-of-the-t-distribution-statistical-table.html).

Calculating the degrees of freedom,

$df=n-1=501-1=500$

According to the table with 500 degrees of freedom and 95% confidence level, we have tc=1.65

Step 3: Calculate the left endpoint of the confidence interval using $a=\bar{x}-tc(s/\sqrt{n})$ and the right endpoint using $b=\bar{x}+tc(s/\sqrt{n})$

Using the formula for the left endpoint:

$a=\bar{x}-tc(s/\sqrt{n})=\$57{,}250-1.65(\$1{,}203\sqrt{501})\approx 57{,}161.32$

Using the formula for the right endpoint:

$b=\bar{x}+tc(s/\sqrt{n})=\$57{,}250+1.65(\$1{,}203\sqrt{501})\approx 57{,}338.68$

Step 4: Write the confidence interval (a,b)

A 95% confidence interval for the population mean is ($57,161.32,$57,338.68)

Example 2

A teacher wants to estimate the mean height of all 400 students at her school. She takes a simple random sample of 30 students and finds the sample mean height is 63 inches with a standard deviation of 1.5 inches. Construct a 90% confidence interval for the population mean height.

Step 1: Identify the sample mean $\bar{x}$ the sample size n, and the sample standard deviation s We have:

$\bar{x}=63$ inches$=63$ inches

$n=30=30$

$s=1.5$ inches$=1.5$ inches

Step 2: Find the degrees of freedom using $df=n-1$. Then look up the critical value tc from the Student's t-distribution (found here: https://study.com/academy/lesson/critical-values-of-the-t-distribution-statistical-table.html).

Calculating the degrees of freedom,

$df=n-1=30-1=29$

According to the table with 29 degrees of freedom and 90% confidence level, we have $tc=1.31=1.31$.

Step 3: Calculate the left endpoint of the confidence interval using $a=\bar{x}-tc(s/\sqrt{n})$ and the right endpoint using $b=\bar{x}+tc(s/\sqrt{n})$

Using the formula for the left endpoint:

$a=\bar{x}-tc(s/\sqrt{n})=63 \text{ inches}-1.31(1.5 \text{ inches}\sqrt{30})\approx 62.64 \text{ inches}$

Using the formula for the right endpoint:

$b=\bar{x}+tc(s/\sqrt{n})=63 \text{ inches}+1.31(1.5 \text{ inches}\sqrt{30})\approx 63.36 \text{ inches}$

Step 4: Write the confidence interval (a,b)

A 90% confidence interval for the population height is (62.64 inches,63.36 inches).

Confidence Interval for population Proportion:

The procedure to find the confidence interval for a population proportion is similar to that for the population mean, but the formulas are a bit different although conceptually identical. While the formulas are different, they are based upon the same mathematical foundation given to us by the Central Limit Theorem. Because of this we will see the same basic format using the same three pieces of information: the sample value of the parameter in question, the standard deviation of the relevant sampling distribution, and the number of standard deviations we need to have the confidence in our estimate that we desire.

The random variable P' (read "P prime") is the sample proportion,

$P'=X/n$

(Sometimes the random variable is denoted as Pˆ, read "P hat".)

p' = the **estimated proportion** of successes or sample proportion of successes (p' is a **point estimate** for p, the true population proportion, and thus q is the probability of a failure in any one trial.)

x = the **number** of successes in the sample

n = the size of the sample

The formula for the confidence interval for a population proportion follows the same format as that for an estimate of a population mean. Remembering the sampling distribution for the proportion from Chapter 7, the standard deviation was found to be:

$$\sigma p'=\sqrt{p(1-p)/n}$$

The confidence interval for a population proportion, therefore, becomes:

$$p=p'\pm[Z(a2)\ \sqrt{p'(1-p')/n}$$

Z(a2) is set according to our desired degree of confidence and $\sqrt{p'(1-p')/n}$ is the standard deviation of the sampling distribution.

The **sample proportions** p' **and** q' **are estimates of the unknown population proportions** p **and** q. The estimated proportions p' and q' are used because p and q are not known.

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones.

1. Construct a 95% confidence interval for the proportion of adult residents of this city who have cell phones.
2. Interpret the confidence interval found in part 1.
3. Is it reasonable to conclude that 85% of the adult residents of this city have cell phones? Explain.

Solution:

1. The sample proportion is p^=421/500=0.842. We need to check n×p^and n×(1−p^)

$$n\times\hat{p}=500\times0.842=421\geq5$$
$$n\times(1-\hat{p})=500\times(1-0.842)=79\geq5$$

Because both $n\times\hat{p}\geq5$and $n\times(1-\hat{p})\geq5$, the sample proportions follow a normal distribution and we can construct the confidence interval.

To find the confidence interval, we need to find the z-score for the 95% confidence interval. This means that we need to find the z-score so that the entire area to the left of z is 0.95+1−0.95/2=0.975.

Function	norm.s.inv	**Answer**
Field 1	0.975	1.9599…

So z=1.9599 For The 95% confidence interval

$$\text{LowerLimit}=\hat{p}-z\times\sqrt{\hat{p}\times(1-\hat{p})n}=0.842-1.9599\times\sqrt{0.842\times(1-0.842)/500}=0.810$$

$$\text{UpperLimit}=\hat{p}+z\times\sqrt{\hat{p}\times(1-\hat{p})n}=0.842+1.9599\times\sqrt{0.842\times(1-0.842)/500}=0.874$$

2. We are 95% confident that the proportion of adult residents of this city who have cell phones is between 81% and 87.4%.
3. It is reasonable to conclude that 85% of the adult residents of this city have cell phones because 85% is inside the confidence interval.

Hypothesis Testing

Setting Up a Hypothesis Test:

Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses that arise from theories.

Steps:

Step 1: State your null and alternate hypothesis

Step 2: Collect data

Step 3: Perform a statistical test

Step 4: Decide whether to reject or fail to reject your null hypothesis

Step 5: Present your findings

Hypothesis testing example

you want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, you restate it as:

H_0: Men are, on average, not taller than women.

H_a: Men are, on average, taller than women.

Step 2: Collect data

For a statistical test to be valid, it is important to perform sampling and collect data in a way that is designed to test your hypothesis. If your data are not representative, then you cannot make statistical inferences about the population you are interested in.

Step 3: Perform a statistical test

There are a variety of statistical tests available, but they are all based on the comparison of **within-group variance** (how spread out the data is within a category) versus **between-group variance** (how different the categories are from one another).

Hypothesis testing example

Based on the type of data you collected, you perform a one-tailed *t*-test to test whether men are in fact taller than women. This test gives you:

an estimate of the difference in average height between the two groups.

a *p*-value showing how likely you are to see this difference if the null hypothesis of no difference is true.

Your *t*-test shows an average height of 175.4 cm for men and an average height of 161.7 cm for women, with an estimate of the true difference ranging from 10.2 cm to infinity. The *p*-value is 0.002.

Step 4: Decide whether to reject or fail to reject your null hypothesis

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

Hypothesis testing example : In your analysis of the difference in average height between men and women, you find that the *p*-value of 0.002 is below your cutoff of 0.05, so you decide to reject your null hypothesis of no difference.

Step 5: Present your findings

The results of hypothesis testing will be presented in the results and discussion sections of your research paper, dissertation or thesis.

In our comparison of mean height between men and women we found an average difference of 13.7 cm and a *p*-value of 0.002; therefore, we can reject the null hypothesis that men are not taller than women and conclude that there is likely a difference in height between men and women.

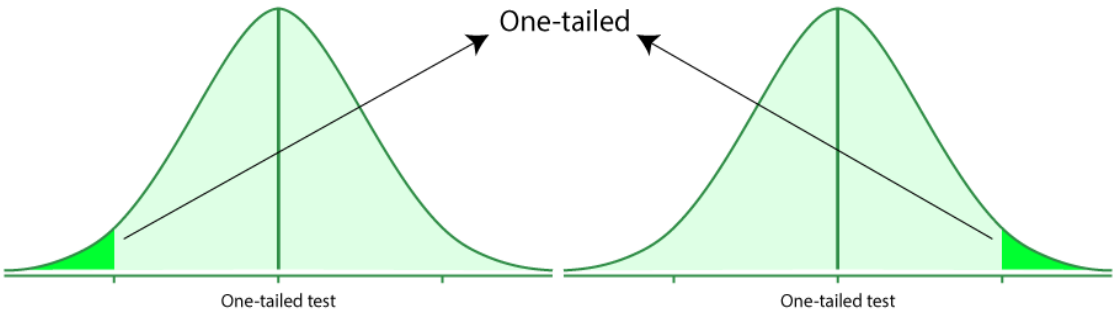
One-tail and two-tail test:

Typically, hypothesis tests take all of the sample data and convert it to a single value, which is known as a test statistic. You're probably already familiar with some test statistics. For example, t-tests calculate t-values. F-tests, such as ANOVA, generate F-values. The chi-square test of independence and some distribution tests produce chi-square values. All of these values are test statistics.

One and Two-Tailed Tests are ways to identify the relationship between the statistical variables. For checking the relationship between variables in a **single direction** (Left or Right direction), we use a one-tailed test. A two-tailed test is used for checking whether the relations between variables are in any direction or not.

One-Tailed Test

A one-tailed test is based on a uni-directional hypothesis where the area of rejection is on only one side of the sampling distribution. It determines whether a particular population parameter is larger or smaller than the predefined parameter. It uses one single critical value to test the data.



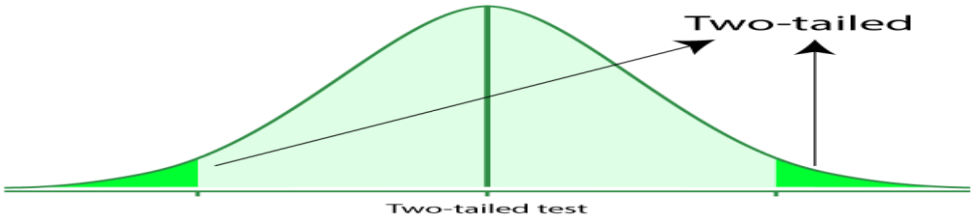
Example: Effect of participants of students in coding competition on their fear level.

- H0: There is no important effect of students in coding competition on their fear level.

The main intention is to check the **decreased** fear level when students participate in a coding competition.

Two-Tailed Test

A two-tailed test is also called a nondirectional hypothesis. For checking whether the sample is greater or less than a range of values, we use the two-tailed. It is used for null hypothesis testing.



Example: Effect of new bill pass on the loan of farmers.

- H0: There is no significant effect of the new bill passed on loans of farmers.

New bill passes can affect in both ways either **increase or decrease** the loan of farmers.

Example:

Suppose H0: mean = 50 and H1: mean not equal to 50

According to the H1, the mean can be greater than or less than 50. This is an example of a Two-tailed test.

In a similar manner, if H0: mean >=50, then H1: mean <50

Here the mean is less than 50. It is called a One-tailed test.

Simple and Composite Hypothesis Testing

Depending on the population distribution, you can classify the statistical hypothesis into two types.

Simple Hypothesis: A simple hypothesis specifies an exact value for the parameter.

Composite Hypothesis: A composite hypothesis specifies a range of values.

Example:

A company is claiming that their average sales for this quarter are 1000 units. This is an example of a simple hypothesis.

Suppose the company claims that the sales are in the range of 900 to 1000 units. Then this is a case of a composite hypothesis.

Type 1 and Type 2 Error:

A hypothesis test can result in two types of errors.

Type 1 Error: A Type-I error occurs when sample results reject the null hypothesis despite being true.

Type 2 Error: A Type-II error occurs when the null hypothesis is not rejected when it is false, unlike a Type-I error.

Example:

Suppose a teacher evaluates the examination paper to decide whether a student passes or fails.

H0: Student has passed

H1: Student has failed

Type I error will be the teacher failing the student [rejects H0] although the student scored the passing marks [H0 was true].

Type II error will be the case where the teacher passes the student [do not reject H0] although the student did not score the passing marks [H1 is true].

Power of The Hypothesis Test:

The power of hypothesis test is a measure of how effective the test is at identifying (say) a difference in populations if such a difference exists. It is the probability of rejecting the null hypothesis when it is false.

The probability of correctly rejecting H_0 when it is false is known as *the power of the test.*

Suppose you want to calculate the power of a hypothesis test on a population mean when the standard deviation is known. Before calculating the power of a test, you need the following:

1. The value of μ, H0= $\mu = \mu 0$
2. The mean of the observed values (X)
3. The population standard deviation(Sigma)
4. The sample size (denoted *n*)

5. The level of significance(Alpha)

To calculate power, you basically work two problems back-to-back. First, find a percentile assuming that H_0 is true. Then, turn it around and find the probability that you'd get that value assuming H_0 is false (and instead H_a is true).

1. Assume that H_0 is true, and

   $\mu = \mu 0$

2. Find the percentile value corresponding to alpha sitting in the tail(s) corresponding to H_a. That is, if

$H_\partial : \mu > \mu_0$ then find *b* where

$$p(X > b) = \alpha$$

If

$H_\partial : \mu < \mu_0$

then find *b* where

$$p(X < b) = \alpha$$

3. Assume that H_0 is false, and instead H_a is true. Since

   $\mu \neq \mu_0$ under this assumption, then let

$\mu = \overline{X}$ in the next step.

4. Find the power by calculating the probability of getting a value more extreme than *b* from Step 2 in the direction of H_a. This process is similar to finding the *p*-value in a test of a single population mean, but instead of using

μ_0 ,you use $\overline{X}$

Example:

Suppose a child psychologist says that the average time that working mothers spend talking to their children is 11 minutes per day. You want to test

$H_0 : \mu = 11$

versus

$$H_a : \mu > 11$$

You conduct a random sample of 100 working mothers and find they spend an average of 11.5 minutes per day talking with their children. Assume prior research suggests the population standard deviation is 2.3 minutes.

When conducting this hypothesis test for a population mean, you find that the *p*-value = 0.015, and with a level of significance of

$$\alpha = 0.05$$

you reject the null hypothesis. But there are a lot of different values of

$$\overline{X}$$

(not just 11.5) that would lead you to reject H_0. So how strong is this specific test? Find the power.

1. Assume that H_0 is true, and
   $$\mu = 11$$
2. Find the percentile value corresponding to
   $$\alpha = 0.05$$
   sitting in the upper tail. If $p(Z > z_b) = 0.05$, then $z_b = 1.645$. Further,
   $$b = \mu + z \cdot \frac{\sigma}{\sqrt{n}} = 11 + 1.645\left(\frac{2.3}{\sqrt{100}}\right) = 11.38$$
3. Assume that H_0 is false, and instead
   $$\mu = 11.5$$
4. Find the power by calculating the probability of getting a value more extreme than *b* from Step 2 in the direction of H_a. Here, you need to find $p(Z > z)$ where
   $$z = \frac{b - \mu}{\sigma/\sqrt{n}} = \frac{11.38 - 11.5}{2.3/\sqrt{100}} = -0.52$$
5. Using the *Z*-table, you find that

6. Hopefully, you were already feeling good about your decision to reject the null hypothesis since the *p*-value of 0.015 was significant at an

   α-level

   of 0.05. Further, you found that Power = 0.6985, meaning that there was nearly a 70 percent chance of correctly rejecting a false null hypothesis.

7. $$\text{Power} = p(Z > -0.52) = 1 - p(Z \leq -0.52) = 1 - 0.3015 = 0.6985$$

Question Bank

1 What is Data Science process?

2 Differentiate Business Intelligence (BI) and Data Science.

3 Compare Data Science and Statistics.

4 Define Data Science.

5 List out the areas in which Data Science can be applied.

6 Who is a Data Scientist?

7.Describe life cycle of Data Science with neat diagram.

8.Explain normal distribution with an example,

9. Justify the need for normal distribution.

10. Describe Normal Distribution in detail.

11. A researcher wants to estimate the mean household income in a town of 25,000 households. The researcher takes a simple random sample of 501 households in the town and finds the sample mean household income is $57,250 with a standard deviation of $1,203. Construct a 95% confidence interval for the population mean household income.

12. In a time use study 20 randomly selected managers were found to spend a mean time of 2.4 hours per day on paperwork. The standard deviation of the 20 scores was 1.30 hours. Construct a 98% confidence interval for the mean time spent on paperwork by all managers.

13. Calculating the confidence interval ln the survey of Americans' and Brits' television watching habits, we can use the sample mean, sample standard deviation, and sample size in place of the population mean, population standard deviation, and population size.calculate the 95% confidence interval.

14.Explain one-tailed and two-tailed hypothesis testing.

15.Explain types of error in hypothesis testing.

16.Explain the power of the hypothesis test.

17. Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes – they own cell phones.

1. **Construct a 95% confidence interval for the proportion of adult residents of this city who have cell phones.**
2. **Interpret the confidence interval found in part 1.**
3. **Is it reasonable to conclude that 85% of the adult residents of this city have cell phones? Explain.**