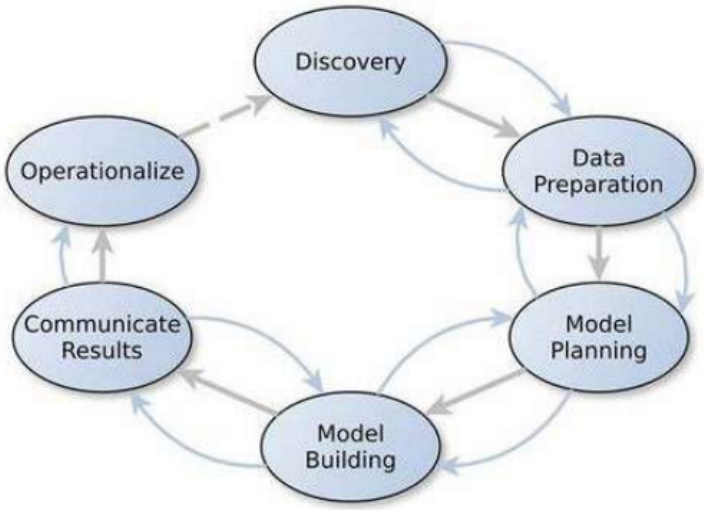
ADVANCED ANALYTICS USING STATISTICS

What is Data?

Data is a collection of facts which is usually obtained as the result of experiences, observations or experiments. It may consist of numbers, words or images.
Data is the lowest level of abstraction in analytics which is the source of information and knowledge.

Data Analytics Life Cycle



Data Discovery: To understand business objectives and data requirements. Tools used are descriptive statistics and data visualisation. For example, exploring healthy drink data to identify trends and patterns.

Data preparation: To clean and prepare data for analysis. The tasks include ETL (Extract, transform and load), data cleaning, preprocessing and feature engineering. Its importance is for data quality and consistency. The tools used are correlation, covariance and inferential statistics. For example, cleaning and transforming raw stock data for investment analysis.

Model planning: To define analytical approach and methods. The tasks include defining problem statements and objectives, selecting relevant variables and features and choosing suitable algorithms and techniques. The considerations include model complexity, interpretability and scalability. For example, planning a machine learning model for customer churn prediction.

Model building: Its purpose is to develop and train predictive models. The tasks include splitting data into training and testing sets, building and training models, and fine-tuning model

parameters. Its importance is for validation and for evaluation metrics. For example, building a neural network for image recognition.
Communicate Results: To document and communicate findings. The tasks include creating reports, dashboards and visualisations and documenting insights and recommendations. The main importance is emphasised on clear and effective communication to stakeholders. For example, presenting data analysis results to company executives.

Operationalize: To implement models into the production environment. The tasks include monitoring model performance and addressing ethical and regulatory considerations. The main importance is emphasised on continuous quality assurance and improvement. The examples include deploying a fraud detection model in banking systems.

What is Business Analytics?

Business Analytics (BA) involves using statistical analysis, data mining, predictive modelling, and other quantitative methods to analyse business data and make informed decisions. BA aims to leverage data to understand and improve business performance.

Probability:

The branch of mathematics that deals with the likelihood of different outcomes is called probability.

Probability = Favourable Outcomes / Total no. of Outcomes

If A and B denote two events,

Joint Probability $P(A \cap B)$: Probability of both events A and B occurring together.

Marginal Probability $P(A)$: Probability of event A occurring, irrespective of B.

Conditional Probability $P(A \mid B)$: Probability of event A occurring given that B has occurred.

Bayes Theorem

It is used to update the probability of a hypothesis based on new evidence. The formula is given by:

$P(A \mid B) = P(B)/(P(B \mid A) \cdot P(A))$

where P(A) and P(B) are probabilities of events A and B respectively and P(B/A) and P(A/B) is the probability of A after B occurred and B after A occurred respectively.

Naive Bayes Classifier

A probabilistic machine learning model used for classification tasks. It is based on Bayes' Theorem with the "naive" assumption of conditional independence between every pair of features given the class variable.

Correlation: It measures the strength and direction of a linear relationship between two variables. It ranges from -1 to +1.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Covariance: It indicates the direction of the linear relationship between two variables. It ranges from -1 to +1.

$$\sigma_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

There are different types of correlations - Pearson's Correlation and Spearman's Correlation.

Pearson's Correlation reflects the noisiness and direction of the linear relationship but neither the slope of the linear relationship nor the aspects of non-linear relationships. Spearman's Correlation is not sensitive to non-linear relationships.

Outliers: it is the data object that significantly deviates from the rest of the data objects and behaves in a different manner. It can be detected using box plots and z-scores.

Probability Distributions

There are two types of distributions -

Continuous Distributions: The distributions are different at infinite sample space. For example Normal Distribution, where a symmetrical, bell-shaped curve where most values cluster around the mean characterised by mean (μ) and standard deviation (σ).

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

Discrete Distributions: The distributions are different at finite sample space. For example, Binomial and Poisson Distribution.

Binomial Distribution: Describes the number of successes in a fixed number of independent Bernoulli trials. Characterised by parameters n (number of trials) and p (probability of success).

$$P(x) = {}^{n}C_{x} p^{x} q^{n-x} = \frac{n!}{(n-x)!x!} p^{x} q^{n-x}$$

$$F(r) = \sum_{x=0}^{r} \binom{n}{x} p^{x} \, q^{(n-x)}$$

Poisson Distribution: Describes the number of events occurring in a fixed interval of time or space. Characterised by the parameter λ (average rate of occurrence).

$$\mathbf{P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}}$$

Descriptive Statistical Measures

There are two types of measures in descriptive statistics. They are as follows -

1. Central tendency:

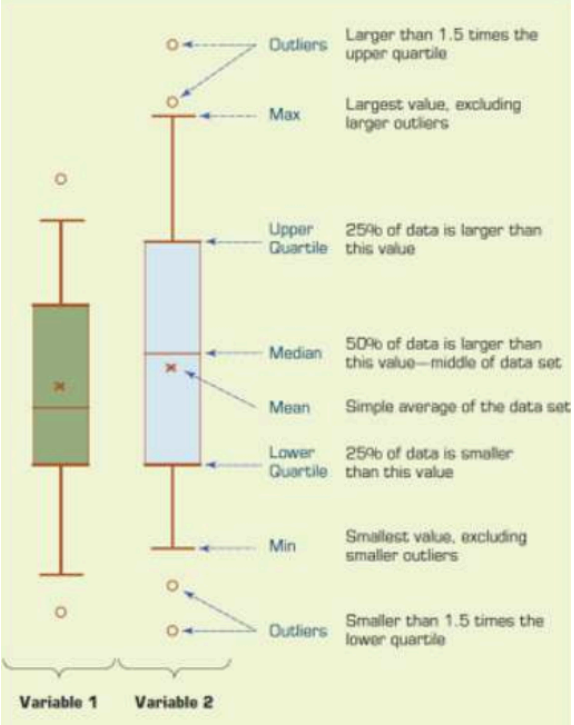
Mean: Average of the data points.
Median: Middle Value in a sorted data.
Mode: Most frequently occurring object in a data.

2. Dispersion:

Range: Difference between highest and lowest values.
Variance: Average of squared deviations from the mean.
Standard Deviation: Measure of the amount of variation from the mean.

Interquartile Range: Difference between first and third quartiles.
First Quartile: The value where 25% of the complete data is below it.
Third Quartile: The value where 25% of the complete data is above it.

The Range, Interquartile Range, Median and Outliers can be described using the diagram shown below.



Inferential Statistics

It is a process of drawing conclusions about a population based on a sample data. The characteristics of the population like mean, standard deviation etc. are called parameters.

The amount of error in the estimation of population parameters that is based on sample statistics is called sampling error.

Central Limit Theorem

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough.

Hypothesis Testing

It is a method of making statistical decisions using experimental data. Its objective is to retain or reject the null hypothesis using data. It consists of two complementary statements.

1. Null Hypothesis: An existing belief.
2. Alternate Hypothesis: What we intend to establish with new evidence.

There are two types of Hypothesis Testing. They are:

1. Parametric Tests: They make use of population parameters like mean, standard deviation. Eg. Z-test, T-test, F-test etc.
2. Non-Parametric Tests: They make use of data distribution to comment on the claim. Eg. Chi-Square etc.

The steps used for performing hypothesis testing is as follows:

1. Formulate Hypotheses: Null (H0) and alternative (H1) hypotheses.
2. Select Significance Level (α): Commonly used values are 0.05 or 0.01.
3. Calculate test statistic Based on the sample data.
4. Compare test statistic to critical value or p-value and make a decision accordingly.
5. If p-value is less than significance level, then reject the null hypothesis otherwise accept the null hypothesis.

Different types of tests are there for different situations. They are as follows:

Z-test

A z test is conducted on a population that follows a normal distribution with independent data points and has a sample size that is greater than or equal to 30. It is used to check whether the means of two populations are equal to each other when the population variance is known.

The formula is given by:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\overline{X} =$ *Sample mean*

$\mu =$ *Population mean*

$\sigma =$ *population standard deviation*

$n =$ *sample size*

Chi-Square Test

Chi-square test is a statistical test for categorical data. It is used to determine whether your data are significantly different from what you expected. Chi-square is often written as X2 and is pronounced “kai-square” (rhymes with “eye-square”). It is also called chi-squared.

A chi-square test (a chi-square goodness of fit test) can test whether these observed frequencies are significantly different from what was expected, such as equal frequencies.

A chi-square test (a test of independence) can test whether these observed frequencies are significantly different from the frequencies expected if handedness is unrelated to nationality.

Both of Pearson’s chi-square tests use the same formula to calculate the test statistic, chi-square (X2):

$$X^2 = \sum \frac{(O - E)^2}{E}$$

where:

- X2 is the chi-square test statistic
- Σ is the summation operator (it means “take the sum of”)
- *O* is the observed frequency
- *E* is the expected frequency

The larger the difference between the observations and the expectations (*O* − *E* in the equation), the bigger the chi-square will be. To decide whether the difference is big enough to be statistically significant, you compare the chi-square value to a critical value.

F-Test

The F test is a statistical technique that determines if the variances of two samples or populations are equal using the F test statistic. Both the samples and the populations need to be independent and fit into an F-distribution.

We can use F-test when:

- The population is normally distributed.
- The samples are taken at random and are independent samples.

The formula for F-test is as follows:

$F_{calc} = \sigma_1^2 / \sigma_2^2$

Where:

Fcalc = critical value.
σ_1^2 and σ_2^2 are the variance of the two samples.

T-Test

Statistical method for the comparison of the mean of the two groups of the normally distributed sample(s).

It is used when:

- Population parameter (mean and standard deviation) is not known.
- Sample size (number of observations) < 30.

The T-test is classified into three parts:

1. **One Sample**

Here, we compare the sample mean with the population mean.

$$t = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\overline{X} = Sample\ mean$
$\mu = Population\ mean$
$\sigma = sample\ standard\ deviation$
$n = sample\ size$

2. **Two Sample**

Here, we compare the means of two different samples.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

$\overline{X}_1, \overline{X}_2$: $Sample\ Mean$
n_1, n_2: $Sample\ Size$
s^2: $estimator\ of\ common\ variance\ such\ that$
$s^2 = \frac{\Sigma(x - \overline{X}_1)^2 + \Sigma(x - \overline{X}_2)^2}{(n_1 - 1) + (n_2 - 1)}$, where
$(n_1 - 1) + (n_2 - 1)$: degree of freedom

3. **Paired Sample**

In this test, we compare the means of two related or same group at two different time.

$$t = \frac{m}{\frac{s}{\sqrt{n}}}$$

m: *mean of difference between each pair of values*
s: *standard deviation of difference between each pair of values*
n: *sample size*

Predictive Modelling

Identifying Informative Attributes:

- Key attributes significantly impact predictions.
- Techniques include statistical measures (e.g., correlation, mutual information) to identify these attributes.

Segmenting Data By Progressive Attributive:

- Data is progressively split based on attribute values to enhance prediction accuracy.
- Segmentation helps in managing large datasets by focusing on more homogeneous subsets.

Models, Induction And Prediction:

- Models are mathematical frameworks that represent the relationship between input features and target variables.
- Induction involves learning the model from historical data.
- Prediction applies the learned model to new data to forecast outcomes.

Supervised Segmentation

A process where the dataset is divided into segments based on supervised learning techniques. Aim is to maximise predictive accuracy by leveraging labelled data.

Visualising Segmentations:

- Tools like decision trees or clustering algorithms are used for visualisation.
- Helps in understanding how data is partitioned and how decisions are made.

Trees As Set Of Rules:

- Decision trees break down data into branches, each representing a decision rule based on attribute values.
- Trees can be converted into a set of if-then rules for easy interpretation.

Probability Estimation:

- Models often output probabilities, providing a measure of confidence in predictions.
- Common methods include logistic regression and Bayesian methods.

Decision Analytics

It involves measuring the performance of predictive models using metrics like accuracy, precision, recall, F1 score, ROC-AUC, etc.It is important for comparing different models.

Analytical Framework:
- Structured approach for analysing data, building models, and interpreting results.
- Includes steps like data preprocessing, model selection, validation, and deployment.

Evaluation:
- Continuous process of assessing model performance to ensure accuracy and reliability.
- Utilises cross-validation, holdout samples, and other techniques.

Baseline:

Simple model or heuristic used as a reference point to compare the performance of more complex models.

Performance And Implications For Investments In Data:
- Strong model performance can justify investments in data collection, storage, and processing.
- Poor performance indicates the need for model refinement or better data quality.

Explicit Evidence Combination With Bayes Rule:
- Bayes' rule provides a formal mechanism to update probabilities based on new evidence.
- Important for probabilistic reasoning and decision-making.

Probabilistic Reasoning:
- Process of drawing conclusions based on probabilities rather than certainties.
- Used extensively in predictive modelling to handle uncertainty.

Business Strategy

Achieving Competitive Advantages:
- Data and predictive models can create competitive advantages by enabling better decision-making, improving efficiency, and enhancing customer experience.
- Examples include personalised marketing and optimised supply chains.

Sustaining Competitive Advantages:
- Continuous innovation, data quality improvement, and model refinement are crucial.
- Maintaining a culture of data-driven decision-making and investing in advanced analytics tools and talent are key strategies.