Business Statistics And Analysis

LEARNING WITH DR. ANAND

ANTIM PRAHAR

The Most Important Questions

ACCORDING TO NEW UPDATED SYLLABUS

By

Dr. Anand Vyas

1 Fisher Index Number

- Fisher's method is a way of combining the information in the p-values from different statistical tests so as to form a single overall test: this method requires that the individual test statistics (or, more immediately, their resulting p-values) should be statistically independent.
- Fisher's method is considered the most ideal because it uses both prices and quantities of base and current period and is based on geometric mean.

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 \quad or \quad P_{01} = \sqrt{L \times P}$$

2 Mean Deviation & Standard Deviation

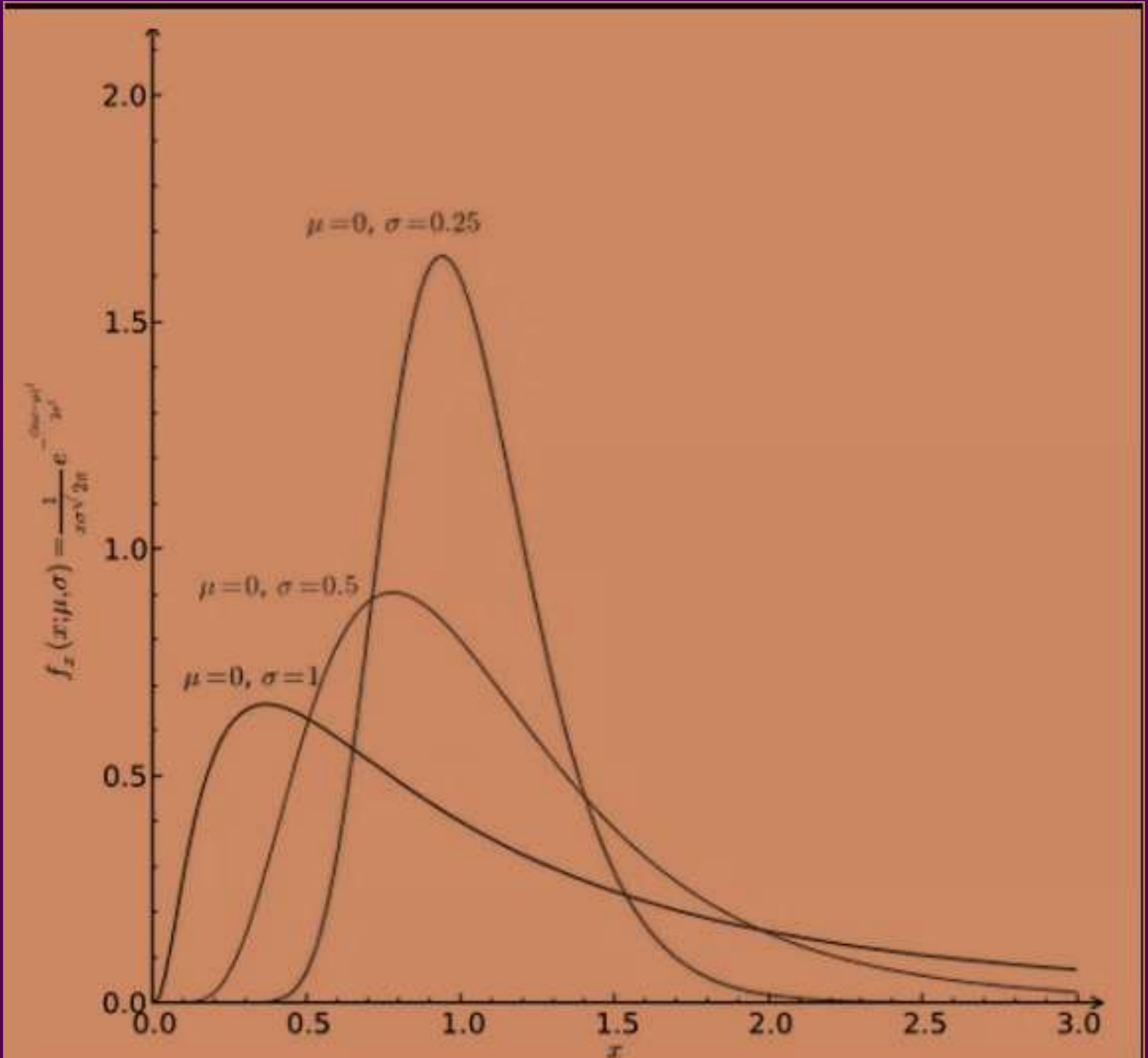
- Mean Deviation
- To understand the dispersion of data from a measure of central tendency, we can use mean deviation. It comes as an improvement over the range. It basically measures the deviations from a value. This value is generally mean or median. Hence although mean deviation about mode can be calculated, mean deviation about mean and median are frequently used.

- Standard Deviation
- As the name suggests, this quantity is a standard measure of the deviation of the entire data in any distribution. Usually represented by s or σ. It uses the arithmetic mean of the distribution as the reference point and normalizes the deviation of all the data values from this mean.

3 Skewness Meaning and Types

- Skewness, in statistics, is the degree of distortion from the symmetrical bell curve, or normal distribution, in a set of data. Skewness can be negative, positive, zero or undefined. A normal distribution has a skew of zero, while a lognormal distribution, for example, would exhibit some degree of right-skew.

•The three probability distributions depicted below depict increasing levels of right (or positive) skewness. Distributions can also be left (negative) skewed. Skewness is used along with kurtosis to better judge the likelihood of events falling in the tails of a probability distribution.


$\mu=0,\ \sigma=0.25$
$\mu=0,\ \sigma=0.5$
$\mu=0,\ \sigma=1$
$f_x(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x

4 Probability (Card & Ball Question)

- In our day to day life the “probability” or “chance” is very commonly used term. Sometimes, we use to say “Probably it may rain tomorrow”, “Probably Mr. X may come for taking his class today”, “Probably you are right”. All these terms, possibility and probability convey the same meaning. But in statistics probability has certain special connotation unlike in Layman’s view.

- Addition and Multiplication Laws

5 Meaning, Scope, functions and Limitations of Statistics

- **Meaning of Statistics:** Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of numerical data. It provides methods and techniques for making inferences and decisions in the face of uncertainty. Statistics is widely used in various fields such as economics, business, medicine, social sciences, engineering, and more.
- **Scope of Statistics:**
- Descriptive Statistics: Involves the organization, summarization, and presentation of data using measures such as mean, median, mode, standard deviation, etc.
- Inferential Statistics: Concerned with making predictions, generalizations, or inferences about a population based on a sample of data.
- Probability: Statistics often involves concepts from probability theory, especially in inferential statistics where uncertainty plays a significant role.
- Statistical Methods: Includes various techniques such as hypothesis testing, regression analysis, analysis of variance (ANOVA), and more, used for data analysis and interpretation.

- **Functions of Statistics:**
- Data Collection: Statistics helps in collecting relevant data through various methods such as surveys, experiments, and observations.
- Data Analysis: It provides techniques for organizing, summarizing, and analyzing data to extract meaningful insights.
- Inference: Statistics enables making predictions and drawing conclusions about populations based on sample data.
- Presentation: It helps in presenting data in a meaningful and comprehensible manner through tables, charts, graphs, and summary statistics.
- Decision Making: Statistics assists in making informed decisions by providing quantitative information about the likelihood and impact of different alternatives.

- **Limitations of Statistics:**
- Sampling Errors: Results obtained from a sample may not accurately represent the entire population due to sampling errors.
- Measurement Errors: Inaccuracies in data collection methods or instruments can lead to measurement errors, affecting the reliability of statistical analysis.
- Assumptions: Statistical methods often rely on certain assumptions about the data, and violations of these assumptions can lead to biased or misleading results.
- Misinterpretation: Statistics can be misinterpreted or manipulated to serve specific agendas, leading to incorrect conclusions.
- Limited Scope: Statistics may not capture all relevant aspects of a phenomenon, especially qualitative or non-quantifiable factors.

6 What do you mean by central tendency? Describe the methods of measuring the central tendency.

- Central tendency refers to a statistical measure that identifies the central or typical value within a dataset. It provides a summary of the central position or the most representative value around which the data points tend to cluster. Central tendency is essential for understanding the distribution of data and making comparisons between different sets of data.
- There are several methods for measuring central tendency, each offering different insights into the characteristics of the data. The most commonly used measures of central tendency include the mean, median, and mode.

- **Mean:** The mean, also known as the average, is calculated by summing up all the values in the dataset and dividing by the total number of values. Mathematically,
- The mean is sensitive to extreme values, also known as outliers, which can skew its value, especially in datasets with asymmetrical distributions.
- **Median:** The median is the middle value in a dataset when the values are arranged in ascending or descending order. If the dataset has an odd number of values, the median is the middle value. If the dataset has an even number of values, the median is the average of the two middle values.
- The median is less affected by extreme values compared to the mean, making it a robust measure of central tendency, especially in skewed distributions.

- **Mode:** The mode is the value that appears most frequently in a dataset. A dataset may have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal). If all values occur with the same frequency, the dataset is said to have no mode.
- The mode is useful for categorical or nominal data but may not be suitable for continuous data with many unique values.

7 Time Series Analysis: Concept, Additive and Multiplicative Models

- Time series analysis is a statistical technique used to analyze and interpret data points collected over a period of time. It involves studying the patterns, trends, and behaviors exhibited by the data over time. Time series data can be found in various fields such as economics, finance, weather forecasting, and engineering, among others.
- **Concept of Time Series Analysis:** Time series analysis involves several steps:
- **Data Collection:** Collecting data points at regular intervals over time.
- **Visualization:** Plotting the data to observe any trends, patterns, or seasonal variations.
- **Modeling:** Fitting mathematical models to the data to describe and predict its behavior.
- **Analysis:** Analyzing the model to understand the underlying patterns and make forecasts or predictions.

- Time series analysis can help in understanding historical trends, forecasting future values, identifying anomalies, and making data-driven decisions.
- **Additive Model:** An additive model decomposes a time series into several components, typically:
- **Trend:** Represents the long-term movement or directionality of the data. It captures the overall increasing or decreasing pattern over time.
- **Seasonality:** Represents the periodic fluctuations or variations that occur at regular intervals within the data, such as daily, weekly, or monthly patterns.
- **Residuals (Error):** Represents the random fluctuations or noise in the data that cannot be explained by the trend or seasonality.
- Mathematically, an additive model can be represented as: $Yt=Tt+St+\varepsilon t$ where:
- Yt represents the observed value at time t,
- Tt represents the trend component at time t,
- St represents the seasonal component at time t,
- εt represents the error term or residuals at time t.
- In an additive model, the components are added together to reconstruct the original time series.

- **Multiplicative Model:** A multiplicative model also decomposes a time series into components, but the components are multiplied together instead of being added: $Yt=Tt\times St\times \varepsilon t$ where the components have the same meanings as in the additive model.
- In a multiplicative model, the components are multiplied together to reconstruct the original time series.

- **Differences between Additive and Multiplicative Models:**
- **Nature of Components:** In additive models, the components are added together, while in multiplicative models, the components are multiplied together.
- **Relative Magnitudes:** In additive models, the magnitudes of the components remain constant over time, while in multiplicative models, the magnitudes of the components may vary proportionally with the level of the series.
- **Application:** Additive models are typically used when the magnitude of seasonality does not depend on the level of the series, while multiplicative models are used when the magnitude of seasonality increases or decreases with the level of the series.

8 What is probability? Explain the calculation of probability under the classical approach. What is Poisson Distribution? How it is calculated? Define its characteristics

- **Probability:** Probability is a measure quantifying the likelihood of an event occurring. It is expressed as a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty. Probability theory provides a mathematical framework for analyzing uncertain events and making predictions based on available information.
- **Calculation of Probability under the Classical Approach:** The classical approach to probability is based on the assumption that all outcomes in a sample space are equally likely. In this approach, the probability of an event *E* occurring is calculated by dividing the number of favorable outcomes for event *E* by the total number of possible outcomes in the sample space.
- Mathematically, the probability of an event *E* is given by: =Number of favorable outcomes for Total number of possible outcomes *P*(*E*)=Total number of possible outcomes Number of favorable outcomes for *E*
- For example, consider rolling a fair six-sided die. The sample space *S* consists of the numbers {1, 2, 3, 4, 5, 6}. To calculate the probability of rolling a 3, since there is only one favorable outcome (rolling a 3) and six possible outcomes, the probability is: (rolling a 3)=16*P*(rolling a 3)=61

- Similarly, for other events, you would count the favorable outcomes and divide by the total number of outcomes.
- **Poisson Distribution:** The Poisson distribution is a probability distribution that describes the number of events that occur in a fixed interval of time or space, given the average rate of occurrence and under certain assumptions. It is often used to model rare events where the probability of occurrence is small but the number of possible occurrences is large.
- **Calculation of Poisson Distribution:** The probability mass function (PMF) of the Poisson distribution is given by:!$P(X=k)=k!e-\lambda\cdot\lambda k$
- Where:
-)$P(X=k)$ is the probability of observing k events,
- e is the base of the natural logarithm (approximately 2.71828),
- λ is the average rate of occurrence of events in the given interval,
- k is the number of events observed,
- !k! denotes the factorial of k.

- **Characteristics of Poisson Distribution:**
- **Discreteness:** The Poisson distribution is a discrete probability distribution, meaning it describes events that occur in distinct, separate units.
- **Non-negative Values:** The number of events described by the Poisson distribution is non-negative (0, 1, 2, ...).
- **Independence:** The occurrence of events in different intervals is assumed to be independent of each other.
- **Memorylessness:** The Poisson distribution has the memoryless property, meaning that the probability of an event occurring in the future is independent of the past.
- **Mean and Variance:** The mean and variance of a Poisson distribution are both equal to the parameter λ, which represents the average rate of occurrence.
-

9 Baye’s Theorem, Addition and Multiplication Laws

- **Bayes' Theorem:**
- **Bayes' Theorem is a fundamental concept in probability theory that describes how to update the probability of a hypothesis based on new evidence or information. It is named after the Reverend Thomas Bayes, an 18th-century British statistician and theologian.**
- **Bayes' Theorem is expressed mathematically as follows:**
- **\[P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \]**
- **Where:**
- **- \(P(A|B) \) is the conditional probability of event \(A \) given event \(B \) has occurred.**
- **- \(P(B|A) \) is the conditional probability of event \(B \) given event \(A \) has occurred.**
- **- \(P(A) \) and \(P(B) \) are the probabilities of events \(A \) and \(B \) occurring, respectively.**

- **Bayes' Theorem allows us to calculate the probability of an event \(A \) given evidence \(B \), based on our prior knowledge of the probability of \(A \) and \(B \), and the probability of observing evidence \(B \) given that \(A \) is true.**
- **Addition Law:**
- **The Addition Law in probability refers to two distinct concepts:**
- **1. Addition Rule for Disjoint Events (Mutually Exclusive Events):**
- **If events \(A \) and \(B \) are disjoint (mutually exclusive), meaning they cannot both occur simultaneously, then the probability of either event occurring is the sum of their individual probabilities:**
- **\[P(A \cup B) = P(A) + P(B) \]**

2. Addition Rule for Non-Disjoint Events:

If events A and B are not mutually exclusive, then the probability of either event occurring is the sum of their individual probabilities minus the probability of their intersection (the event where both A and B occur):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Multiplication Law:

Similar to addition, there are two concepts related to multiplication in probability:

1. Multiplication Rule for Independent Events:

If events A and B are independent, meaning the occurrence of one event does not affect the occurrence of the other, then the probability of both events occurring is the product of their individual probabilities:

$$P(A \cap B) = P(A) \times P(B)$$

2. Multiplication Rule for Dependent Events:

If events A and B are dependent, meaning the occurrence of one event affects the occurrence of the other, then the probability of both events occurring is the product of the probability of the first event and the conditional probability of the second event given that the first event has occurred:

$$P(A \cap B) = P(A) \times P(B|A)$$

These laws and Bayes' Theorem provide powerful tools for calculating probabilities and making inferences in various situations involving uncertainty and random events.

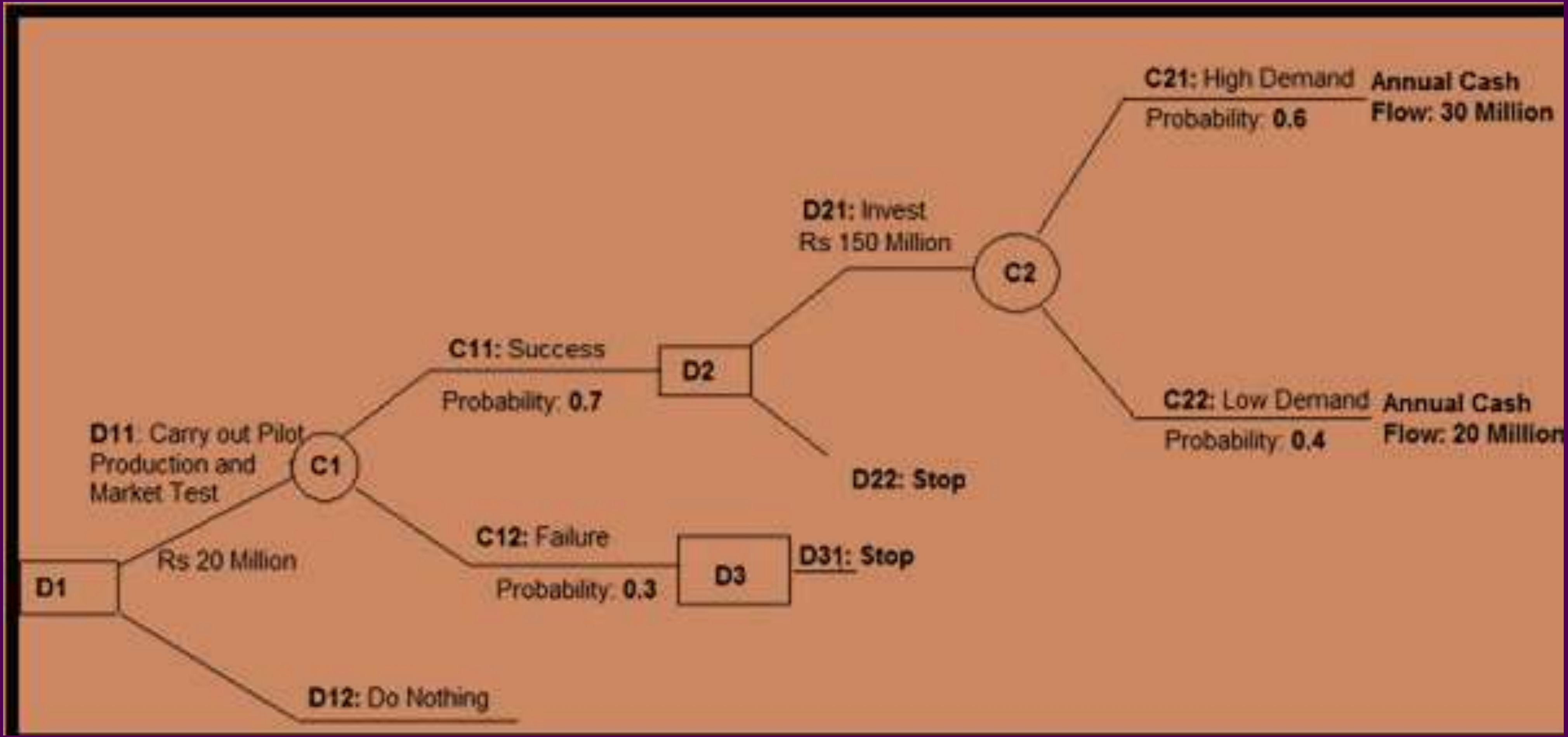
10 What are decision tree? Explain the decision tress with the help of any example. **Decision Tree approach and its Applications**

- The decision tree approach is a predictive modeling technique used in data mining and machine learning for both classification and regression tasks. It is a graphical representation of possible solutions to a decision-making problem, where each node represents a decision, each branch represents an outcome of that decision, and each leaf node represents a final decision or outcome.
- **Components of a Decision Tree:**
- **Root Node:** The topmost node in the tree, representing the initial decision or starting point.
- **Decision Nodes:** Intermediate nodes in the tree where decisions are made based on input features or attributes.
- **Branches:** Arrows connecting nodes, representing possible outcomes of decisions.
- **Leaf Nodes:** Terminal nodes in the tree where final decisions or outcomes are reached.

Decision Tree

- Decision Tree may be understood as the logical tree, is a range of conditions (premises) and actions (conclusions), which are depicted as nodes and the branches of the tree which link the premises with conclusions. It is a decision support tool, having a tree-like representation of decisions and the consequences thereof. It uses ‘AND’ and ‘OR’ operators, to recreate the structure of if-then rules.

- Decision Node: Represented as square, wherein different courses of action arise from decision node in main branches.
- Chance Node: Symbolised as a circle, at the terminal point of decision node, the chance node is present, where they emerge as sub-branches. These depict probabilities and outcomes.



C21: High Demand

Probability: 0.6

Annual Cash Flow: 30 Million

D21: Invest Rs 150 Million

C2

C11: Success

Probability: 0.7

D2

D22: Stop

C22: Low Demand

Probability: 0.4

Annual Cash Flow: 20 Million

D11: Carry out Pilot Production and Market Test

C1

Rs 20 Million

C12: Failure

Probability: 0.3

D3

D31: Stop

D1

D12: Do Nothing

- **Decision Tree Construction:** The construction of a decision tree involves selecting the best attribute at each decision node based on certain criteria, such as information gain (for classification problems) or reduction in variance (for regression problems). This process continues recursively until a stopping criterion is met, such as reaching a maximum tree depth or when further splitting does not provide significant improvement in prediction accuracy.
- **Applications of Decision Trees:**
- **Classification:** Decision trees are commonly used for classification tasks, where the goal is to assign a class label to input data. Examples include:
  - Predicting whether an email is spam or not spam based on its features.
  - Classifying customers into different segments based on their purchasing behavior.
  - Diagnosing medical conditions based on symptoms and test results.

- **Regression:** Decision trees can also be used for regression tasks, where the goal is to predict a continuous numeric value. Examples include:
  - Predicting house prices based on features such as location, size, and amenities.
  - Estimating the demand for a product based on various factors such as price, advertising expenditure, and seasonality.
- **Anomaly Detection:** Decision trees can be used for anomaly detection, where the goal is to identify unusual patterns or outliers in data.
- **Feature Selection:** Decision trees can help identify the most important features or attributes in a dataset, which can be useful for feature selection in other machine learning models.
- **Risk Assessment:** Decision trees can be used to assess risk in various domains, such as finance, insurance, and healthcare, by modeling the likelihood of different outcomes based on input variables.
- **Customer Relationship Management:** Decision trees can be used to optimize marketing strategies by identifying the most effective channels and messages for different customer segments.

11Relationship between Regression and Correlation

- Regression and correlation are both statistical techniques used to analyze the relationship between variables. While they are related, they serve slightly different purposes and provide complementary information about the association between variables.
- **1. Relationship:**
- **Regression:** Regression analysis is used to model the relationship between a dependent variable (response variable) and one or more independent variables (predictor variables). It seeks to estimate the functional form of the relationship and predict the value of the dependent variable based on the values of the independent variables.
- **Correlation:** Correlation analysis measures the strength and direction of the linear relationship between two continuous variables. It quantifies the degree to which changes in one variable are associated with changes in another variable. Correlation does not imply causation but indicates the presence and strength of a relationship between variables.

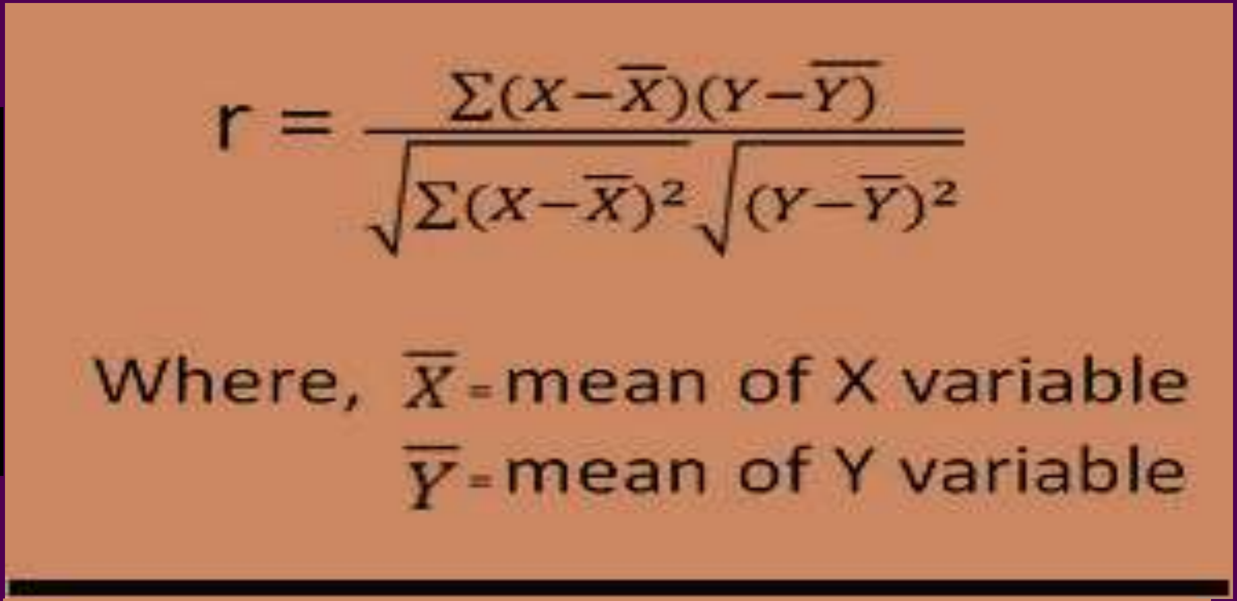
- **2. Purpose:**
- **Regression:** Regression analysis is used for prediction, estimation, and inference. It helps in understanding how changes in one or more independent variables affect the dependent variable. Regression models can be used for forecasting and hypothesis testing.
- **Correlation:** Correlation analysis is used to assess the degree of association between variables. It helps in identifying patterns and trends in data and can be used to screen for potential relationships before conducting regression analysis. Correlation is also useful for variable selection and identifying multicollinearity in regression models.
- **3. Output:**
- **Regression:** The output of regression analysis includes the regression equation, which describes the relationship between variables, as well as estimates of the regression coefficients and their significance. Regression also provides measures of goodness of fit, such as R-squared, which indicates the proportion of variance in the dependent variable explained by the independent variables.
- **Correlation:** The output of correlation analysis includes the correlation coefficient, typically denoted by �*r*, which ranges from -1 to 1. A correlation coefficient of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Correlation analysis also provides a scatter plot to visualize the relationship between variables.
- **4. Assumptions:**
- **Regression:** Regression analysis assumes a causal relationship between the independent and dependent variables. It also assumes linearity, homoscedasticity (constant variance of errors), independence of errors, and normality of errors.
- **Correlation:** Correlation analysis assumes linearity and bivariate normality (normal distribution of each variable and their joint distribution). It is less restrictive than regression analysis and does not imply causality.

12 Correlation Karl Pearson's and Spearman's Rank Correlation

Karl Pearson’s Coefficient of Correlation is widely used mathematical method wherein the numerical expression is used to calculate the degree and direction of the relationship between linear related variables.

Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. The Spearman rank correlation test does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.

Karl Pearson's Correlation

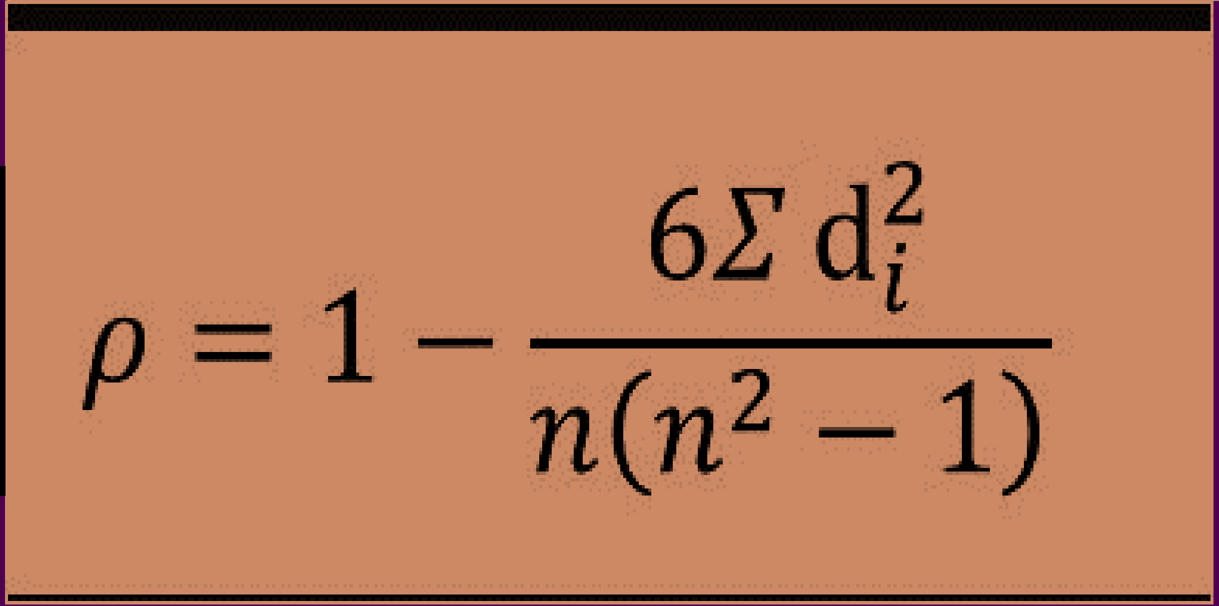


$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}}$$

Where, $\bar{X}$ = mean of X variable

$\bar{Y}$ = mean of Y variable

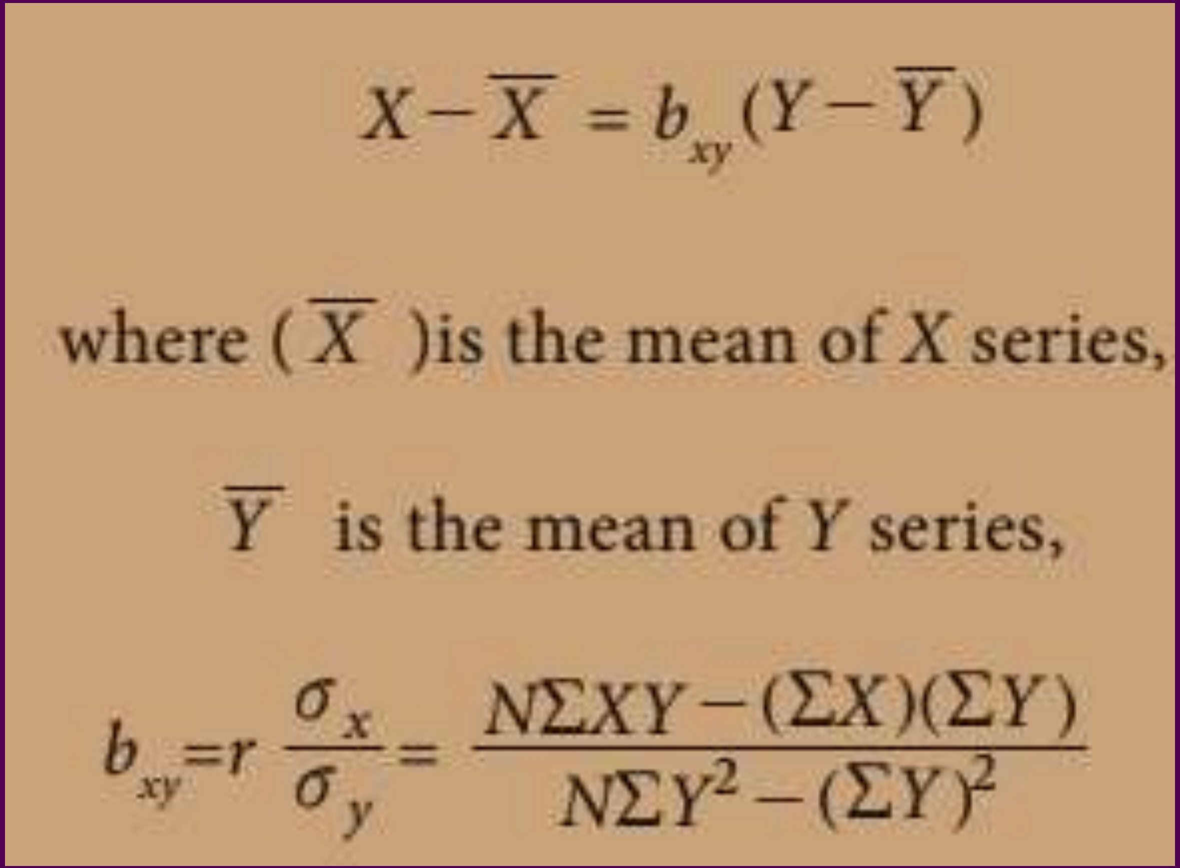
Spearman's Rank Correlation



$$\rho = 1 - \frac{6\Sigma\, \mathrm{d}_i^2}{n(n^2-1)}$$

13 Regression

- Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).



$$X-\overline{X} = b_{xy}(Y-\overline{Y})$$

where ($\overline{X}$)is the mean of X series,

$\overline{Y}$ is the mean of Y series,

$$b_{xy}=r\,\frac{\sigma_x}{\sigma_y}=\frac{N\Sigma XY-(\Sigma X)(\Sigma Y)}{N\Sigma Y^2-(\Sigma Y)^2}$$

14 Mean, Median Mode

- MEAN (ARITHMETIC)
- The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x1, x2, …, xn, the sample mean, usually denoted by (pronounced x bar), is:

- MEDIAN
- The median is the middle score for a set of data that has been arranged in order of magnitude.

- MODE
- The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option