Times asked: 6 times
5 times
4 times
3 times
2 times
1 time
# indicates 5-mark question

QA Theory Question bank

1. Introduction to Statistics

1. Define "Statistics". Explain its uses and limitations.# Discuss its use in business and trade.
2. Write short note on: Pie chart and its advantages and disadvantages. #
3. What is diagrammatic representation of data? Explain its advantages. #
4. Justify or contradict – 'Charts or graphs are more effective in attracting attention than any other method of presenting data'. #
5. Write a short note on: Meaning and importance of Tabulation. #
6. Define and explain the following terms with an example:
   Grouped data, class interval, class limits, class boundaries, class mark, inclusive and exclusive series, frequency and tally marks.

2. Data Collection and Sampling methods

1. Explain primary data and secondary data in detail.(Or What are the various methods of collecting statistical data?) Which of these is most reliable and why?
2. Distinguish between primary data and secondary data. #
3. What precautions should be taken in the use of secondary data. #
4. Explain Census method. Its merits and demerits. #
5. Why are personal interviews preferred to questionnaires? Under what conditions may a questionnaire prove as a personal interview? #
6. What do you mean by a questionnaire? What is the difference between a questionnaire and a schedule? State the essential points to be remembered in drafting a questionnaire.
7. Explain sampling and purpose of sampling. #
8. Differentiate between Probability sampling and non-probability sampling. #
9. Explain Simple Random Sampling. #
10. What is Stratified sampling? Explain the merits and demerits of Stratified sampling. #

3. Introduction to Regression

1. Explain the following methods to check the performance of regression model: #
   a) MAE
   b) MAPE
2. Justify or contradict 'b_{xy} and b_{yx}, must be either positive or negative'. #

3. Explain regression and its types. Also explain regression analysis and discuss its applications. How does it differ from correlation.
4. Write a detailed note on least square regression.

4. Introduction to Multiple Linear Regression

1. What do you mean by Partial correlation coefficients? Explain in detail.
2. Write a short note on: Significance of Overall fit of regression model. #
3. What are assumptions of Multiple Linear Regression. #
4. Write short note on: Multiple Regression. #

5. Statistical inference

1. Explain Point Estimation with characteristics.
2. Explain the following point Estimation Properties with Example:
   a) Consistency
   b) Unbiasedness
3. Explain the method of maximum likelihood estimation.
4. Show that sample variance (S^2) is an unbiased estimator of population variance (σ^2). Also illustrate with an example. #
5. Write short note on: Method of moments. #
6. Define a random variable and its mathematical expectation.

6. Test of hypothesis

1. What is Hypothesis testing? Explain
   a) Z-test for single Mean
   b) Z-test for difference of mean
2. Explain in detail Neyman Pearson lemma.
3. Write a short note on: MP and UMP tests. #
4. Differentiate between Null Hypothesis and Alternative Hypothesis. #
5. Differentiate between Critical Region and Region of acceptance.
6. What are the test of skewness.
7. Explain Type I and Type II error in detail.

	1	2	3	4	5	6
2024 Dec	20	20	25	25	15	15
2024 May	15	15	25	20	10	35
2023 Dec	25	10	20	15	20	30
2023 May	20	15	25	25	25	15
2022 Dec	25	10	25	10	20	15
Last 5 Avg	**20**	**15**	**25**	**20**	**20**	**15-25**
***2022 May**	15	15	15	15	15	15
Total	120	85	135	110	100	125

QA Theory Answer bank

multiple times asked questions highlighted
question asked once with red font
# indicates 5-mark question

1. Introduction to Statistics

1. Define “Statistics”. Explain its uses and limitations.# **Discuss its use in business and trade.**

Definition:
Statistics is a branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of numerical data. It helps in drawing meaningful conclusions from data.

Uses of statistics:

1. **Data Analysis:** Statistics helps organize and interpret data to find meaningful patterns.
2. **Decision Making:** It provides data-driven insights to support smart choices.
3. **Planning:** Statistics predicts future trends to assist in strategic planning.
4. **Economics & Finance:** Statistics helps evaluate market trends and financial performance.
5. **Machine Learning & AI:** It supports data modelling, pattern recognition, and result validation.

Limitations of Statistics:

1. **Cannot Replace Common Sense:** Blind use may lead to incorrect conclusions.
2. **Data May Be Misleading:** Biased or inaccurate data can distort analysis.
3. **Cannot Deal with Qualitative Aspects:** Emotions, attitudes, and opinions can't be directly measured statistically.
4. **Requires Skilled Interpretation:** Misinterpretation of data may result in wrong decisions.

Uses in Business and Trade:

1. **Decision Making:** Assists managers in making informed decisions through data analysis.
2. **Market Analysis:** Helps in analysing market trends, consumer preferences, and demand forecasting.
3. **Quality Control:** Used in maintaining and improving product quality through statistical quality control techniques.
4. **Financial Analysis:** Supports budgeting, cost control, and profit analysis.
5. **Performance Evaluation:** Aids in evaluating employee and departmental performance based on statistical indicators.

2. **Write short note on: Pie chart and its advantages and disadvantages.** #

Pie Chart:

A pie chart is a circular diagram divided into slices to illustrate numerical proportions. Each slice represents a category and is proportional to its share of the total.

Advantages:

- **Simple and visually appealing**
- Good for showing **percentage or part-whole relationships**
- Easy to compare a few categories at a glance
- Useful for **non-technical audiences**

Disadvantages:

- Not effective for **large number of categories**
- Difficult to interpret **precise values**
- Can be misleading if **not drawn to scale**
- **Harder to read** compared to bar charts when data is similar in size

3. **What is diagrammatic representation of data? Explain its advantages.** #

Definition:

Diagrammatic representation of data refers to presenting numerical data visually using graphs or diagrams such as bar charts, pie charts, line graphs, etc., to make the information more understandable and interpretable.

Types of Diagrams:

- **Bar Diagram**
- **Pie Chart**
- **Line Graph**
- **Histogram**
- **Pictogram**

Advantages:

1. **Easy to Understand:**
   Visuals simplify complex numerical data.
2. **Quick Comparison:**
   Diagrams make it easy to compare values and trends.
3. **Attractive Presentation:**
   Graphs and diagrams are more engaging than raw tables.
4. **Better Retention:**
   Information is easier to remember when presented visually.
5. **Useful for Analysis:**
   Helps in spotting trends, patterns, and outliers quickly.
6. **Saves Time:**
   Quick overview without reading detailed statistics.

4. **Justify or contradict – 'Charts or graphs are more effective in attracting attention than any other method of presenting data'.** #

Justification –
Charts or graphs are visual representations of data that help simplify complex information, making it easier to understand and interpret.

Reasons why they are more effective:

1. **Visual Clarity** – They present data in a visually appealing way, which grabs attention quickly.
2. **Quick Understanding** – Patterns, trends, and comparisons are easier to grasp at a glance.
3. **Engagement** – Visuals are more engaging than plain text or tables, especially for general audiences.
4. **Simplification** – They break down complex numerical data into easy-to-read formats.
5. **Memory Retention** – People tend to remember visuals better than raw data or text.

5. Write a short note on: Meaning and importance of Tabulation.

Tabulation is the process of arranging raw data systematically in rows and columns to form a table. It organizes information clearly so that it can be easily read, understood, and analyzed.

Importance of Tabulation:

- Simplifies large and complex data for easy understanding.
- Helps in quick comparison and interpretation of information.
- Facilitates efficient data analysis and decision-making.
- Saves time by presenting data clearly and concisely.
- Provides a neat and organized way to display statistical information.

6. **Define and explain the following terms with an example:**
   Grouped data, class interval, class limits, class boundaries, class mark, inclusive and exclusive series, frequency and tally marks.

1. Grouped Data

Data that is organized into groups or classes instead of individual values to make analysis easier.
Example: Marks of students grouped as 0–10, 11–20, 21–30, etc.

2. Class Interval

A range of values that form one group or class in grouped data.
Example: 10–20 is a class interval representing values from 10 up to 20.

3. Class Limits

The smallest and largest values that can belong to a class interval.
Example: For the class interval 10–20, 10 is the lower-class limit and 20 is the upper-class limit.

4. Class Boundaries

The actual limits that separate classes without gaps, usually halfway between class limits.
Example: For classes 10–20 and 21–30, class boundaries could be 9.5–20.5 and 20.5–30.5.

5. Class Mark

The midpoint of a class interval, calculated as (Lower class limit + Upper class limit) ÷ 2.
Example: For 10–20, class mark = (10 + 20) ÷ 2 = 15.

6. Inclusive Series

A grouped data series where class intervals include both their lower and upper limits.
Example: 10–20, 21–30, 31–40 (both 20 and 21 are included in respective classes).

7. Exclusive Series

A grouped data series where the upper limit of one class is the lower limit of the next class but not included in that class.
Example: 10–19, 20–29, 30–39 (19 is included in first class but 20 is excluded, belonging to the second).

8. Frequency

The number of data points or observations within a particular class interval.
Example: If 5 students scored between 10–20, frequency of 10–20 = 5.

9. Tally Marks

A simple way to record frequency by marking groups of five with four vertical lines and a diagonal line across them.
Example: Frequency 7 is represented as |||| \ || (five marks + two marks).

2. Data Collection and Sampling methods

7. Explain primary data and secondary data in detail. (Or What are the various methods of collecting statistical data?) **Which of these is most reliable and why?**

1. Primary Data

- **Definition:**
  Primary data is the data collected directly from the original source for a specific research purpose. It is firsthand information gathered specifically to answer a research question.
- **How it is collected:**
  - **Surveys and Questionnaires** – Structured sets of questions filled by respondents.
  - **Interviews** – Face-to-face, telephonic, or virtual conversations to gather detailed insights.
  - **Observations** – Recording behaviours or events as they occur naturally.
  - **Experiments** – Controlled studies to test hypotheses and collect data.
- **Characteristics:**
  - Original and raw data
  - Highly reliable for the specific purpose it was collected for
  - More time-consuming and costly to collect
  - Tailored to meet the exact needs of the research
- **Examples:**
  - Conducting a survey to know customer satisfaction

2. Secondary Data

- **Definition:**
  Secondary data is data that has already been collected, recorded, and published by someone else for a purpose other than the current research.
- **Sources of Secondary Data:**
  - **Books, Journals, Newspapers** – Contain summarized or analyzed information.
  - **Government Reports and Statistics** – Official data on population, health, economy, etc.
  - **Online Databases and Websites** – Digital repositories of published data and research.
  - **Previous Research Studies –** Includes data from earlier research such as academic projects, case studies.

- **Characteristics:**
    - Easily accessible and inexpensive
    - May not be perfectly suited to the current research needs
    - Can save time and effort in data collection
    - Risk of outdated or biased data
- **Examples:**
    - Census data from government websites

Most Reliable Method:

Primary data collection, especially through direct methods like interviews and experiments, is considered the most reliable.

Why?

- It is specifically tailored to the current research objective.
- Offers high accuracy and relevance.
- Provides firsthand, up-to-date information.

8. Distinguish between primary data and secondary data. #

Basis	Primary Data	Secondary Data
Definition	Data collected firsthand for a specific purpose	Data already collected by someone for another purpose
Source	Original source (e.g., surveys, experiments)	Published sources (e.g., books, journals, reports)
Time & Cost	More time-consuming and expensive	Less time and cost involved
Accuracy	Generally more accurate and reliable	May be less accurate due to unknown collection methods
Suitability	Specifically designed to meet research objectives	May not perfectly fit current research needs
Control	Researcher has full control over data collection process	No control over how or when the data was collected
Up-to-date	Usually up-to-date as collected for current study	May be outdated depending on when it was collected
Example	Conducting a market survey	Using government census data

9. What precautions should be taken in the use of secondary data. #

Precautions in Using Secondary Data:

1. **Check Authenticity** – Ensure the data comes from reliable and credible sources.
2. **Verify Accuracy** – Cross-check with other trusted sources for consistency.
3. **Check Timeliness** – Make sure the data is current and relevant to your study.
4. **Understand Methodology** – Know how the data was collected to judge its reliability.
5. **Check for Bias** – Be cautious of any possible bias in the data or its presentation.

10. Explain Census method. Its merits and demerits.

Census Method:
The Census method involves collecting data from every unit of the population. It is a complete enumeration where information is gathered from each individual or item, leaving no one out.

Merits of Census Method:

1. **Complete Accuracy** – Since every unit is studied, the data is highly accurate.
2. **Detailed Information** – Provides comprehensive and in-depth data.
3. **Reliable for Policy Making** – Useful for government planning and decision-making.
4. **No Sampling Error** – Entire population is covered, so sampling bias is avoided.

Demerits of Census Method:

1. **Time-Consuming** – Collecting data from all units takes a long time.
2. **Costly** – Requires more money, manpower, and resources.
3. **Difficult to Manage** – Managing large-scale data collection is complex.
4. **Not Practical for Frequent Use** – Hard to use repeatedly for regular studies.

11. Why are personal interviews preferred to questionnaires? Under what conditions may a questionnaire prove as a personal interview? #

Why Personal Interviews Are Preferred to Questionnaires:

1. **Better Clarity** – Interviewers can explain questions if the respondent doesn't understand.
2. **Higher Response Rate** – People are more likely to respond in person than return a questionnaire.
3. **More Accurate Data** – Interviewers can ensure accurate and complete answers.
4. **Suitability for Illiterate Respondents** – Interviews can be used even when the respondent cannot read or write.

A questionnaire may be as effective as a personal interview when:

1. **Respondents Are Educated and Willing** – They can understand and complete the form accurately.
2. **Questions Are Simple and Clear** – No need for explanation or clarification.
3. **Large-Scale Surveys** – Where personal interviews are not practical due to cost or distance.
4. **Respondents Prefer Privacy** – They may give more honest answers in written form.

12. What do you mean by a questionnaire? What is the difference between a questionnaire and a schedule? State the essential points to be remembered in drafting a questionnaire.

Questionnaire: A questionnaire is a structured set of written questions used to collect data from respondents. It is usually filled out by the respondent themselves and is widely used in surveys and research.

Difference between Questionnaire and Schedule:

Basis	Questionnaire	Schedule
Filling Method	Filled by the respondent	Filled by the investigator
Cost	Low (no need for trained staff)	High (requires trained enumerators)
Suitability	Literate population	Both literate and illiterate respondents
Non-response rate	Higher	Lower
Control	Less control over responses	Greater control and clarity

Essential Points in Drafting a Good Questionnaire:

1. **Clear Objectives:**
   The purpose of the survey must guide the selection of questions.
2. **Simple and Clear Language:**
   Avoid jargon, technical terms, and complex sentences.
3. **Logical Order:**
   Start with easy and non-sensitive questions; group related questions together.
4. **Avoid Leading Questions:**
   Do not suggest a particular answer in the question.
5. **Use Close-ended and Open-ended Questions Wisely:**
   Close-ended are easier to analyse, while open-ended give more detail.
6. **Avoid Double-Barrelled Questions:**
   Each question should ask only one thing.
7. **Provide Clear Instructions:**
   Indicate how to answer (tick, circle, write) and what to do in case of confusion.
8. **Test the Questionnaire (Pilot Survey):**
   Always pre-test to find errors or confusing items.
9. **Keep it Short and Relevant:**
   Long questionnaires reduce response rate and accuracy.
10. **Ensure Anonymity and Confidentiality (if required):**
    Increases honesty and participation.

13. Explain sampling and purpose of sampling.

Sampling:
Sampling is the process of selecting a small group of individuals or items (called a sample) from a larger group (population) to study and draw conclusions about the entire population. Instead of collecting data from every individual or unit (as in the census method), sampling focuses on a representative subset, making the study more practical and efficient.

Types of sampling methods:

- Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling

Purpose of Sampling:

1. **Saves Time and Cost** – Studying a sample is quicker and less expensive than studying the whole population.
2. **Convenient and Practical** – Easier to manage, especially when the population is large or spread out.
3. **Allows Detailed Study** – Enables in-depth analysis of a manageable group.
4. **Useful When Census is Not Possible** – Ideal when it's not feasible to collect data from everyone.

14. Differentiate between Probability sampling and non-probability sampling.

Basis	Probability Sampling	Non-Probability Sampling
Definition	Every unit in the population has a known, non-zero chance of selection.	Not every unit has a known or equal chance of being selected.
Selection Method	Random and systematic methods are used.	Selection is based on judgment, convenience, or purpose.
Bias	Less prone to bias.	More prone to selection bias.
Representativeness	Usually more representative of the population.	May not represent the population accurately.
Usefulness	Suitable for large-scale and scientific studies.	Suitable for exploratory or preliminary research.
Examples	Simple random sampling, stratified sampling.	Convenience sampling, quota sampling.

15. Explain Simple Random Sampling.

Definition:
Simple Random Sampling is a method where each and every unit in the population has an equal and independent chance of being selected in the sample.

- Selection is completely random, like drawing names from a hat.
- There is no bias in the selection process.
- Can be done using methods like lottery or random number tables.

Merits:

1. Easy to understand and use.
2. Each unit has an equal chance of selection.
3. Reduces bias in sample selection.
4. Results can be generalized to the whole population (if sample is large enough).

Demerits:

1. Not suitable for large or scattered populations.
2. Requires a complete list of the population (sampling frame).
3. Random selection may not represent all sub-groups accurately.

16. What is Stratified sampling? Explain the merits and demerits of Stratified sampling.#

Definition:

Stratified sampling is a sampling method in which the population is divided into homogeneous subgroups or strata based on a specific characteristic (like age, income, gender, etc.), and then random samples are taken from each stratum proportionally or equally.

Example:

If a college has 1000 students: 600 in Engineering, 300 in Commerce, and 100 in Arts, then we divide them into these three strata and take random samples from each group accordingly.

Merits of Stratified Sampling:

1. **More Representative:**
   Ensures that every subgroup is adequately represented in the sample.
2. **Increased Accuracy:**
   Reduces sampling error as variation within strata is lower.
3. **Useful for Comparisons:**
   Facilitates analysis between different strata (e.g., male vs female, rural vs urban).
4. **Improves Reliability:**
   Provides better estimates of the population parameters.

Limitations of Stratified Sampling:

1. **Requires Prior Information:**
   You must know the population structure to form accurate strata.
2. **Complex and Time-Consuming:**
   Dividing population into strata and sampling from each group takes more effort.
3. **Difficulty in Stratification:**
   Not always clear how to divide population; improper stratification can affect results.
4. **Risk of Overlapping:**
   If strata are not mutually exclusive, data may be duplicated or misclassified.

3. Introduction to Regression

17. Explain the following methods to check the performance of regression model: #
a) MAE
b) MAPE

a) MAE (Mean Absolute Error):

Definition:

MAE is the average of the absolute differences between actual and predicted values.

Formula:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where:

- y_i = Actual value
- $\hat{y}_i$ = Predicted value
- n = Number of observations

Interpretation:

Lowe MAE means better model performance. MAE gives equal weight to all errors.

b) MAPE (Mean Absolute Percentage Error):

Definition:

MAPE is the average of the absolute percentage errors between actual and predicted values.

Formula:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Interpretation:

Expresses error as a percentage.

MAPE helps compare model performance across different datasets.

Lower MAPE indicates better accuracy.

Note: MAPE can be misleading when actual values are very close to zero (denominator issue).

18. Justify or contradict 'b_{xy} and b_{yx}, must be either positive or negative'. #

Justification:

- b_{xy} is the regression coefficient of Y on X.
- b_{yx} is the regression coefficient of X on Y.

These coefficients represent the slope in the regression equations and indicate the direction of the relationship between X and Y.

The **signs of b_{xy} and b_{yx} are always the same** because they depend on the sign of the correlation coefficient r.

Since $b_{xy} = r\frac{s_y}{s_x}$ and $b_{yx} = r\frac{s_x}{s_y}$, both depend on r (correlation), which determines the sign.

Conclusion:

The statement is **correct** because b_{xy} and b_{yx} must both be either positive or negative, reflecting the direction of the linear relationship between the variables.

19. Explain regression and its types. Also explain regression analysis and discuss its applications. How does it differ from correlation.

Regression is a statistical method used to model and analyze the relationship between a dependent variable (target) and one or more independent variables (predictors).

Types of Regression

1. Simple Linear Regression

- Involves one independent variable and one dependent variable.
- Model: $Y = a+bX$
- Example: Predicting sales (Y) based on advertising cost (X).

2. Multiple Linear Regression

- Involves two or more independent variables.
- Model: $Y = a+b_1X_1 + b_2X_2 +...+b_nX_nY$
- Example: Predicting house price based on size, location, and number of bedrooms.

Regression Analysis:

Regression analysis is a statistical technique used to examine the relationship between a dependent variable and one or more independent variables.

Example:

$Y = a + bX + \varepsilon$

Where:

- Y: Dependent variable
- X: Independent variable
- a: Intercept
- b: Slope (effect of X on Y)
- ε: Random error

Uses of regression analysis:

1. **Predicting future outcomes** (e.g., sales forecasting).
2. **Analyzing relationships** between dependent and independent variables.
3. **Identifying trends** and patterns in time-series data.
4. **Assessing risk factors** in fields like finance and insurance.
5. **Supporting business decisions** through data-driven insights.

Difference Between Regression and Correlation:

Aspect	Regression	Correlation
Definition	Measures the effect of independent variables on dependent	Measures strength and direction of association
Direction	One-way relationship (X affects Y)	Mutual association (no cause-effect implied)
Use	Prediction and modelling	Interpretation and comparison
Output	Regression equation (Y on X)	Correlation coefficient (r)
Range	Output can be any real number	r lies between -1 and +1

20. Write a detailed note on least square regression.

Definition:

The Least Squares Regression method is a statistical technique used to determine the best-fitting line through a set of data points by minimizing the sum of the squares of the vertical deviations (residuals) between observed values and the estimated values predicted by the line.

Simple Linear Regression Equation:

$$Y = a + bX$$

Where:

- Y = Dependent variable
- X = Independent variable
- a = Intercept of the regression line
- b = Slope of the regression line

Formulas for 'a' and 'b':

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{\sum Y - b\sum X}{n}$$

Steps in Least Squares Method:

1. Collect paired data values of X and Y.
2. Compute $\sum X$, $\sum Y$, $\sum X^2$, $\sum XY$.
3. Apply formulas to calculate slope b and intercept a.
4. Form the regression equation Y = a+bX
5. Use the equation to predict values of Y for given X.

Uses:

- Predicting future trends (e.g., sales forecasting).
- Analyzing relationships between variables.
- Finding line of best fit in data analysis.

4. Introduction to Multiple Linear Regression

21. What do you mean by Partial correlation coefficients? Explain in detail.

Definition:

Partial correlation coefficient measures the strength and direction of a linear relationship between two variables while controlling or removing the effect of one or more other variables.

Example Situation:

Suppose you want to understand the relationship between:

- **X** = Hours Studied
- **Y** = Exam Score
- **Z** = Sleep Hours

There might be a correlation between hours studied and exam score, but sleep hours (Z) could also influence exam score (Y). So, to isolate the direct relationship between X and Y, remove the influence of Z. This is what partial correlation does.

Formula:

For three variables X, Y and Z, the partial correlation coefficient between X and Y controlling for Z is:

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} \cdot r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

Where:

- r_{XY} : Pearson correlation between X and Y
- r_{XZ}: Pearson correlation between X and Z
- r_{YZ} :Pearson correlation between Y and Z

Example:

Assume:

- $r_{XY} = 0.3$
- $r_{XZ} = 0.1$
- $r_{YZ} = 0.2$

Calculate the partial correlation between X and Y controlling for Z, i.e., $r_{XY \cdot Z}$. Substituting:

$$r_{XY \cdot Z} = \frac{0.3 - (0.1 \cdot 0.2)}{\sqrt{(1 - 0.1^2)(1 - 0.2^2)}} = \frac{0.3 - 0.02}{\sqrt{(1 - 0.01)(1 - 0.04)}} = \frac{0.28}{\sqrt{0.99 \cdot 0.96}} = \frac{0.28}{\sqrt{0.9504}} = \frac{0.28}{0.975} \approx 0.287$$

The partial correlation rXY·Z=0.287 indicates a positive but weak relationship between hours studied (X) and exam score (Y), after removing the effect of sleep hours (Z).

22. Write a short note on: Significance of Overall fit of regression model.

The overall fit of a regression model shows how well the model explains the relationship between the independent variables and the dependent variable. It indicates how closely the predicted values match the actual data.

A good overall fit means the model reliably predicts outcomes and captures the true pattern in the data, helping in accurate forecasting and decision-making.

Statistical measures like R-squared and F-test are used to check this fit. A high R-squared and a significant F-test suggest that the model explains a large portion of the variation in the dependent variable, making the model meaningful and useful.

23. What are assumptions of Multiple Linear Regression.

Multiple Linear Regression (MLR) assumes a linear relationship between a dependent variable and two or more independent variables.

Assumptions:

1. **Linearity:**
   The relationship between the dependent and each independent variable is linear.
2. **Independence:**
   Observations are independent of each other.
3. **Homoscedasticity:**
   Constant variance of errors across all levels of the independent variables.
4. **No Multicollinearity:**
   Independent variables are not highly correlated with each other.
5. **Normality of Errors:**
   The residuals (errors) are normally distributed.
6. **No Autocorrelation:**
   Errors are not correlated with one another (important in time series data).
7. **Correct Model Specification:**
   All relevant variables are included; irrelevant ones are excluded.

24. Write short note on: Multiple Regression.

Definition:

Multiple regression is a statistical technique used to model the relationship between one dependent variable and two or more independent variables.

Multiple Regression Equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Where:

- Y = Dependent variable (outcome)
- $X_1, X_2, \ldots, X_k$ = Independent variables (predictors)
- β_0 = Intercept (constant term)
- $\beta_1, \beta_2, \ldots, \beta_k$ = Regression coefficients (effect of each predictor)
- ε = Error term (captures unexplained variation)

Uses:

- Forecasting and prediction
- Understanding impact of multiple factors on an outcome
- Used in economics, marketing, healthcare, etc.

Advantages:

- Can analyze complex relationships
- More realistic models compared to simple regression

Limitations:

- Sensitive to multicollinearity (correlation among predictors)
- Requires large sample sizes for accuracy
- Assumes linear relationship among variables

5. Statistical inference

25. Explain Point Estimation with characteristics.

Point Estimation

Definition:
Point estimation is a statistical technique used to provide a single best guess or value (called a *point estimator*) of an unknown population parameter (like mean, proportion, or variance) based on sample data.

Example:
If we take a sample of students and calculate the average height, that sample mean is a point estimate of the average height of the entire student population.

Characteristics of a Good Point Estimator:

1. **Unbiasedness**
    - The expected value of the estimator should be equal to the true value of the population parameter.
    - *Example:* The sample mean is an unbiased estimator of the population mean.
2. **Consistency**
    - As the sample size increases, the estimator should get closer to the actual population value.
    - *In short:* Larger samples → better estimates.
3. **Efficiency**
    - Among all unbiased estimators, the one with the smallest variance is the most efficient.
    - A more efficient estimator gives more precise results.
4. **Sufficiency**
    - The estimator should use all the relevant information in the sample.
    - A sufficient estimator captures everything the sample tells us about the parameter.

26. Explain the following point Estimation Properties with Example:
a) Consistency
b) Unbiasedness

a) Consistency

Definition:
An estimator is said to be consistent if, as the sample size increases, the estimate converges in probability to the true value of the population parameter.

Explanation:
In simple terms, a consistent estimator gives values that become more accurate as we collect more data. This means the estimation error decreases as sample size increases.

Mathematically:

Let $\hat{\theta}_n$ be the estimator of parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is consistent if:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1 \quad \text{for every } \epsilon > 0$$

Example:
Let $\bar{x}$ be the sample mean used to estimate the population mean μ.

- If we take a small sample, the value of $\bar{x}$ may not be close to μ.
- But as we increase the sample size, $\bar{x}$ becomes closer to μ.

So, the sample mean $\bar{x}$ is a consistent estimator of the population mean μ.

b) Unbiasedness

Definition:
An estimator is unbiased if its expected value is equal to the true value of the population parameter.

Explanation:
This means that, on average, the estimator neither overestimates nor underestimates the actual parameter. Unbiasedness ensures that there is no systematic error in estimation.

Mathematically:

An estimator $\hat{\theta}$ is unbiased for a parameter θ if:

$$E(\hat{\theta}) = \theta$$

Example:
Again, consider the sample mean $\bar{x}$ as an estimator of the population mean μ.

- The expected value of $\bar{x}$ is equal to μ, i.e., $E(\bar{x}) = \mu$.
  Therefore, $\bar{x}$ is an unbiased estimator of μ.

In contrast, if an estimator consistently gives a value higher or lower than the actual parameter, it is biased.

27. Explain the method of maximum likelihood estimation.

Definition:

The **Maximum Likelihood Estimation (MLE)** method is a statistical technique used to estimate the parameters of a probability distribution or a statistical model. It finds the parameter values that maximize the likelihood that the observed data occurred under the given model.

Steps of MLE:

1. **Assume a probability distribution** for the data, with an unknown parameter θ.
2. **Write the likelihood function** $L(\theta)$, which is the joint probability (or density) of the observed data given θ:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

3. **Take the log** of the likelihood function to get the **log-likelihood**:

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

4. **Differentiate** the log-likelihood with respect to θ, set the derivative to **zero**, and solve:

$$\frac{d}{d\theta} \log L(\theta) = 0$$

5. **Verify** that it gives a maximum by checking the second derivative or analyzing behavior.

Advantages of MLE:

1. **Efficiency:**
   MLE provides estimators that are consistent and asymptotically efficient.
2. **Flexibility:**
   Can be applied to a wide range of distributions and models.
3. **Large Sample Properties:**
   Estimators become normally distributed as sample size increases.
4. **Estimates Variance:**
   MLE helps in estimating not only the parameters but also their standard errors.

Disadvantages of MLE:

1. **Computationally Intensive:**
   Involves complex calculus and optimization techniques.

2. **Sensitive to Outliers:**
   Extreme data points can significantly affect the estimation.
3. **Requires Distributional Assumptions:**
   Results are reliable only if the assumed model is correct.
4. **May Fail for Small Samples:**
   Estimates can be biased or unstable when sample size is low.

28. Show that sample variance (S^2) is an unbiased estimator of population variance (σ^2). Also illustrate with an example. #

Population Variance (σ^2):

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

Sample Variance (S^2):

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

Where:

- $\bar{x}$ = sample mean
- μ = population mean
- n = sample size

Goal: Show that

$$E(S^2) = \sigma^2$$

We start from:

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

Use identity:

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

Take expectation:

$$E[S^2] = \frac{1}{n-1}\left[\sum_{i=1}^{n}E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2]\right]$$

Now,

- $E[(X_i - \mu)^2] = \sigma^2$
- $E[(\bar{X} - \mu)^2] = \frac{\sigma^2}{n}$

So:

$$E[S^2] = \frac{1}{n-1}\left[n\sigma^2 - n\left(\frac{\sigma^2}{n}\right)\right] = \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

Hence, $E[S^2] = \sigma^2 \Rightarrow$ sample variance is **unbiased**.

Example:

Suppose we have a sample of size 3:

$$X = \{4, 6, 8\}$$

Step 1: Calculate sample mean

$$\bar{X} = \frac{4 + 6 + 8}{3} = 6$$

Step 2: Calculate sample variance S^2

$$S^2 = \frac{1}{3 - 1} \left[(4 - 6)^2 + (6 - 6)^2 + (8 - 6)^2 \right] = \frac{1}{2}(4 + 0 + 4) = \frac{8}{2} = 4$$

Now, this $S^2 = 4$ is an unbiased estimate of the **true population variance** σ^2 if this sample came from a larger population.

29. Write short note on: Method of moments.

Definition:

The method of moments is a technique for estimating population parameters by equating the sample moments (like mean, variance) to the theoretical moments of a probability distribution.

Formula:

$$\text{Sample moment:} \quad m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$\text{Theoretical moment:} \quad \mu'_k = E(X^k)$$

$$\text{Equating moments:} \quad m_k = \mu'_k \quad \text{for } k = 1, 2, 3, \ldots$$

The unknown parameters can be determined by solving these equations.

Advantages:

- Simple and easy to apply.
- Does not require complex optimization.

Disadvantages:

- Less efficient than Maximum Likelihood Estimation.
- Can give biased estimates in some cases.

30. Define a random variable and its mathematical expectation.

Random Variable:
A random variable is a numerical outcome of a random experiment. It is a function that assigns a real number to each outcome in a sample space. It can be:

- **Discrete:** Takes countable values (e.g., number of heads in coin tosses).
- **Continuous:** Takes an infinite number of values within a range (e.g., height, weight).

Mathematical Expectation (Expected Value):

The mathematical expectation of a random variable is the weighted average of all possible values it can take, where each value is multiplied by its corresponding probability.

For a **discrete random variable** X with possible values $x_1, x_2, \ldots, x_n$ and probabilities $p_1, p_2, \ldots, p_n$, the expectation is given by:

$$E(X) = \sum_{i=1}^{n} x_i \cdot p_i$$

It represents the long-run average value of the variable if the experiment is repeated many times.

6. Test of hypothesis

31. What is Hypothesis testing? Explain
a) Z-test for single Mean
b) Z-test for difference of mean

Hypothesis testing is a statistical method used to make decisions or inferences about population parameters based on sample data.
It involves:

- **Null Hypothesis (H_0):** A statement of no effect or no difference.
- **Alternative Hypothesis (H_1):** What we want to prove or support.

We collect data and use a test statistic to decide whether to reject or fail to reject H_0.

a) Z-test for Single Mean

Definition:
Z-test for a single mean is used to test whether the sample mean significantly differs from a known or hypothesized population mean when the population standard deviation is known and the sample size is large (n ≥ 30).

Hypotheses:

- Null Hypothesis $H_0: \mu = \mu_0$
- Alternative Hypothesis $H_1: \mu \neq \mu_0$ (or $<$ or $>$, depending on the test)

Test Statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Where:

- $\bar{X}$ = Sample mean
- μ_0 = Hypothesized population mean
- σ = Population standard deviation
- n = Sample size

Decision Rule:

- Compare calculated Z value with critical Z value from standard normal table.
- If $|Z| > Z_{\alpha/2}$, reject H_0.

b) Z-test for Difference of Means

Definition:
Z-test for difference of means is used to test whether the means of two independent samples differ significantly when the population standard deviations are known and both sample sizes are large.

Hypotheses:

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$ (or $<$, $>$, depending on the context)

Test Statistic:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where:

- $\bar{X}_1, \bar{X}_2$ = Sample means
- σ_1, σ_2 = Population standard deviations
- n_1, n_2 = Sample sizes

Decision Rule:

- Reject H_0 if calculated $|Z| > Z_{\alpha/2}$.

32. Explain in detail Neyman Pearson lemma.

Definition:

The **Neyman-Pearson Lemma** is a fundamental result in hypothesis testing that provides the most powerful test for deciding between two simple hypotheses.

It states that:

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with probability density (or mass) function $f(x;\theta)$

We want to test:

- Null Hypothesis: $H_0 : \theta = \theta_0$
- Alternative Hypothesis: $H_1 : \theta = \theta_1$

Both are **simple hypotheses** (i.e., the parameter values are fully specified).

Then, the **most powerful test** at a given significance level α (i.e., highest chance of rejecting H_0 when H_1 is true) is the one that rejects H_0 when:

$$\frac{f(x;\theta_1)}{f(x;\theta_0)} > k$$

Where:

- $\frac{f(x;\theta_1)}{f(x;\theta_0)}$ is the **likelihood ratio**
- k is a constant chosen so that the test has the desired level α

Applications:

- Used in designing tests like the **Z-test**, **t-test**, and **likelihood ratio tests**
- Especially relevant in **binary classification** and **signal detection**

Q3. Write a short note on: MP and UMP tests. #

1. Most Powerful (MP) Test:

An **MP test** is the best test for a given significance level α\alphaα when testing a **simple null hypothesis** against a **simple alternative hypothesis**.

Form:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

Definition:

A **Most Powerful test** of level α\alphaα is a test that **maximizes the power** (i.e., the probability of rejecting H_0 when H_1 is true), among all tests of the same size α.

2. Uniformly Most Powerful (UMP) Test:

An **UMP test** is the best test for a given significance level α\alphaα when testing a **simple null hypothesis** against a **composite alternative hypothesis**.

Form:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

Definition:

A **Uniformly Most Powerful test** of level α is a test that **maximizes power not just at one point**, but **for all values of θ in the alternative hypothesis**, among all tests of the same size α.

Type	Applies to	Compares	Power Condition
MP Test	Simple vs Simple	Specific H_0 vs Specific H_1	Most powerful at a fixed θ_1
UMP Test	Simple vs Composite	Specific H_0 vs Range of H_1	Most powerful for all $\theta \in H_1$

34. Differentiate between Null Hypothesis and Alternative Hypothesis. #

Basis	**Null Hypothesis (H_0)**	**Alternative Hypothesis (H_1)**
Definition	Assumes no effect, no difference, or status quo.	Assumes the presence of an effect or difference.
Nature	Conservative or default assumption	Competing claim tested against the null
Purpose	To be tested and possibly rejected	To be accepted if the null is rejected
Symbolically	$H_0 : \mu=\mu_0$	$H_1 : \mu\neq\mu_0$ or >>, <<
Decision	Retain H_0 if no sufficient evidence	Accept H_1 if evidence supports it
Example	A new drug has no effect: $H_0 : \mu=\mu_0$	The drug is effective: $H_1 : \mu\neq\mu_0$

35. Differentiate between Critical Region and Region of acceptance.

Aspect	Critical Region	Region of Acceptance
Definition	The set of values that lead to rejection of H_0	The set of values where we do not reject H_0
Relation to α	Corresponds to the significance level (α)	Complement of significance level (1−α)
Decision	If test statistic falls here, reject H_0	If test statistic falls here, retain H_0
Other Names	Rejection region	Non-rejection region
Example (two-tailed)	For α=0.05, $z < -1.96$ or $z > 1.96$	$-1.96 < z < 1.96$

36. What are the test of skewness.

Skewness:
Skewness is a measure of the asymmetry of a distribution. It indicates whether the data is skewed to the left (negative skew) or to the right (positive skew).

Tests of Skewness:

1. Karl Pearson's Coefficient of Skewness:

$$\text{Sk} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

If mode is not defined:

$$\text{Sk} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

2. Bowley's Coefficient of Skewness (based on quartiles):

$$\text{Sk} = \frac{Q_3 + Q_1 - 2 \cdot \text{Median}}{Q_3 - Q_1}$$

3. Kelly's Coefficient of Skewness (based on percentiles or deciles):

$$\text{Sk} = \frac{P_{90} + P_{10} - 2 \cdot P_{50}}{P_{90} - P_{10}}$$

Interpretation:

- Sk = 0 → Symmetrical distribution
- Sk > 0 → Positively skewed
- Sk < 0 → Negatively skewed

37. Explain Type I and Type II error in detail.

Type I Error (False Positive)

- **Definition:** Rejecting the null hypothesis **when it is actually true**.
- **What it means:** You conclude there is an effect or difference, but in reality, there isn't.
- **Probability:** Denoted by **α (alpha)**, also called the significance level of the test.
- **Example:** A medical test indicates a disease is present when the patient is actually healthy.
- **Consequence:** You may take unnecessary actions based on incorrect rejection.

Type II Error (False Negative)

- **Definition:** Failing to reject the null hypothesis **when it is actually false**.
- **What it means:** You miss detecting an effect or difference that actually exists.
- **Probability:** Denoted by **β (beta)**.
- **Power of test:** Equal to 1−β, which is the probability of correctly rejecting a false null hypothesis.
- **Example:** A medical test fails to detect a disease when the patient actually has it.
- **Consequence:** You may miss important findings or fail to act when necessary.