Live MCQ
সঠিক প্রস্তুতি, সহজ প্রিলি

Live MCQ™

৪৯তম স্পেশাল বিসিএস (শিক্ষা) বিষয়ভিত্তিক প্রস্তুতি

বিষয়: পরিসংখ্যান (৯৮১)

Compendium PDF

(সিলেবাস অনুসারে সর্বাধিক গুরুত্বপূর্ণ টপিক ও এমসিকিউ-এর সমন্বয়ে রচিত)

Mentor:

মোঃ রাকিবুল ইসলাম,

৪১ তম বিসিএস,

সহকারী কর কমিশনার।

সূচিপত্র:

৪৯তম বিসিএস (বিশেষ) পরীক্ষা-২০২৫ এর সিলেবাস

STATISTICS
(POST RELATED)
Subject Code: 981
Total Marks-100

Part-I
Marks - 50

1. Introduction to Statistics: Definition and scope, Scope of Statistics, Classification, Variables.
2. Presentation of Data: Charts or Diagrams, Types of diagrams.
3. Grouping Data: Frequency Distribution, General rules for forming frequency, Graphical presentation of frequency distribution, Relative frequency distribution.
4. Measures of Central Tendency: The Arithmetic mean, the Median, The Mode, The Geometric Mean. The Harmonic Mean, Finding Measures of Central tendency from Grouped data, Graphical determination of Measures of Central tendency, Comparative discussion on measures of central tendency.
5. Measures of Dispersion: Dispersion or variation, Measures of dispersion from grouped data, Interpretation of Standard deviation, Chebyshev rule, Normal rule, Relative dispersion: Co-efficient of Variation.
6. Skewness and Kurtosis: Skewness, Kurtosis, Skewness and kurtosis from graphical displays, Descriptive measures of skewness and kurtosis.
7. Regression and correlation: Simple regression and correlation. Least squares estimates of simple linear regression, regression coefficient and correlation coefficient. Rank correlation, correlation ratio and partial correlation. Multiple regression and multiple correlation coefficient. Coefficient of determination.
8. Demography: Crude birth and death rates, Fertility rate, Age specific and total fertility rates, Population growth in Bangladesh, Migration, Nuptiality.

9. Index Number: Definition, Properties of index numbers, Significance of index numbers, Classification of index numbers, Simple Index Number, Un weighted indices, Simple average of price index, Simple Aggregate Index, Weighted Indices, Laspeyres index, Paasche method, Fisher's Ideal Index, Weighted average of relatives.

10. Time Series Analysis: Components of a time series. Measurement of secular trend, seasonal variations, cyclical variations and measurement of irregular variations.

11. Sampling: Statistical population and sample. Advantages and disadvantages of sampling over census. Sample design. Probability and non-probability sampling. Simple random sampling, stratified random sampling and systematic sampling. Cluster sampling, sampling error and non-sampling error. Determination of sample size.

STATISTICS

Part-II

Marks - 50

1. Concept of probability: Basic Definitions, Approaches of Defining probability, Basic properties of probabilities, Notation and Graphical displays for events.

2. Rules of Probability: Special Addition rule, The complementation Rule, General Addition rule, Bivariate data and Contingency table. Joint and marginal probabilities, Multiplication rules, Conditional probabilities, Concept of Bayes' Theorem.

3. Random Variables and probability Distributions : Random variable, Discrete Probability Distribution, Binomial Probability, Hypergeometric distribution, Poisson distribution, Normal distribution.

4. Sampling Distribution: Sampling distribution of the sample mean for a normally distributed variable, The Central Limit Theorem (CLT), Sampling Distribution of the Sample Mean, Sampling distribution of the sample proportion, Sampling distribution of function of mean and proportion. Confidence interval, Confidence interval of Population mean. Depermination of Sample size, Sampling for estimating mean, Sampling for estimating proportion.

5. Basic Concepts of Hypothesis Testing : Null and Alternative hypothesis, simple and composite hypotheses, Test statistic, acceptance and rejection regions, type I and type II errors, the significance level, one tailed and two tailed tests, general procedure for test of hypothesis. Tests based on normal, student's t, F, and X^2 distribution. The Z- test for two population means. The pooled t-test for two population means. The paired t-test for two population means.

6. Analysis of Variance : Concept of analysis of variance, treatment, response, extraneous variables, One-Way Anova Model, Estimate of The Model Parameters, Hypothesis Testing In Anova. Two-Way Anova, significance of Correlation and rank correlation coefficients. Multiple comparison test. Two-way analysis of variance with and without interaction.

7. Experimental Designs: Basic principles of Experimental Design. Randomization, Replication and Local control. The completely Randomized Design (CRD), Randomized Complete Block Design (RCBD) and Latin square Design.

৪৯তম বিসিএস (বিশেষ) পরীক্ষা-২০২৫ এর সিলেবাস

STATISTICS
(POST RELATED)
Subject Code: 981
Total Marks-100

Part-I
Marks - 50

1. Introduction to Statistics: Definition and scope, Scope of Statistics, Classification, Variables.
2. Presentation of Data: Charts or Diagrams, Types of diagrams.
3. Grouping Data: Frequency Distribution, General rules for forming frequency, Graphical presentation of frequency distribution, Relative frequency distribution.

1) Introduction to Statistics
   Definition & characteristics and limitation (★★)
   Scope of Statistics (applications) (★★)
   Classification of data (qualitative/quantitative; primary/secondary) (★★)
   Variables (types, scales of measurement) (★★★)
2) Presentation of Data
   Charts/Diagrams (histogram, polygon, ogive, bar, pie, scatter) (★★)
   Types of diagrams & when to use them (★★)
3) Grouping Data
   Frequency distribution & class design (bounds, width) (★★)
   General rules for forming frequency (Sturges, practical rules) (★★)
   Graphical presentation of frequency distribution (histogram/ogive) (★★)
   Relative frequency distribution (★)
4. Measures of Central Tendency: The Arithmetic mean, the Median, The Mode, The Geometric Mean. The Harmonic Mean, Finding Measures of Central tendency from Grouped data, Graphical determination of Measures of Central tendency, Comparative discussion on measures of central tendency.

Arithmetic Mean (ungrouped & grouped) (★★)
Median (location, interpolation) (★★)
Mode (grouped data, modal class) (★★)
Geometric Mean (★★★)
Harmonic Mean (★★★)
From grouped data (all three) (★★★)
Relationship among them (★★)

Graphical determination (median/mode via ogives, histogram) (★★)

Comparative discussion (when to use which) (★★★)

5. Measures of Dispersion: Dispersion or variation, Measures of dispersion from grouped data, Interpretation of Standard deviation, Chebyshev rule, Normal rule, Relative dispersion: Co-efficient of Variation.
6. Skewness and Kurtosis: Skewness, Kurtosis, Skewness and kurtosis from graphical displays, Descriptive measures of skewness and kurtosis.
7. Regression and correlation: Simple regression and correlation. Least squares estimates of simple linear regression, regression coefficient and correlation coefficient. Rank correlation, correlation ratio and partial correlation. Multiple regression and multiple correlation coefficient. Coefficient of determination.
8. Demography: Crude birth and death rates, Fertility rate, Age specific and total fertility rates, Population growth in Bangladesh, Migration, Nuptiality.

5) Measures of Dispersion

Concept of dispersion/variation (★)

Measures from grouped data (range, IQR, variance, SD) (★★★)

Interpretation of Standard Deviation (coefficients, units) (★★)

Chebyshev's rule (★★)

Normal rule (empirical 68–95–99.7%) (★★)

Relative dispersion: Coefficient of Variation (★★★)

6) Skewness & Kurtosis

Skewness (Pearson/Bowley) (★★)

Kurtosis (β2, excess kurtosis) (★★)

From graphical displays (shape reading) (★★★)

Descriptive measures (formulas & interpretation) (★★★)

7) Regression & Correlation

Simple regression & correlation (concept + uses) (★★★)

Least squares estimates in simple linear regression (slope/intercept) (★★★)

Regression coefficient & correlation coefficient (r, bxy, byx, relation) (★★★)

Rank correlation (Spearman) (★★)

Correlation ratio (η) (★★)

Partial correlation (★★)

Multiple regression (setup, interpretation) (★★)

Multiple correlation coefficient (R) (★★)

Coefficient of determination (R^2) (★★★)

8) Demography

Crude birth & death rates (★)

Fertility rate (general/total) (★★)

Age-specific & total fertility rates (★★)

GRR, NRR (★★★)

Interpretation of those rates (★★★)

Population growth in Bangladesh (trends/indicators) (★★)
Migration (★)
Nuptiality (★)

9. Index Number: Definition, Properties of index numbers, Significance of index numbers, Classification of index numbers, Simple Index Number, Un weighted indices, Simple average of price index, Simple Aggregate Index, Weighted Indices, Laspeyres index, Paasche method, Fisher's Ideal Index, Weighted average of relatives.
10. Time Series Analysis: Components of a time series. Measurement of secular trend, seasonal variations, cyclical variations and measurement of irregular variations.
11. Sampling: Statistical population and sample. Advantages and disadvantages of sampling over census. Sample design. Probability and non-probability sampling. Simple random sampling, stratified random sampling and systematic sampling. Cluster sampling, sampling error and non-sampling error. Determination of sample size.

9) Index Numbers
Definition (★)
Properties (time reversal, factor reversal) (★★)
Significance/uses (★)
Classification (price, quantity, value; simple/weighted) (★★)
Simple Index Number (unweighted) (★)
Unweighted indices—simple average of price relatives (★)
Simple Aggregate Index (★)
Weighted Indices (general approach) (★★)
Laspeyres Index (★★★)
Paasche Index (★★★)
Fisher's Ideal Index (tests, superiority) (★★★)
CPI, CPI interpretation(★★★)
Weighted average of relatives (★★)
10) Time Series Analysis
Components (trend, seasonal, cyclical, irregular) (★★★)
Measurement of secular trend (moving average, least squares) (★★★)
Measurement of seasonal variation (ratio-to-moving-avg, link relatives) (★★★)
Cyclical variations (identification, limitations) (★★)
Irregular variations (★)
11) Sampling
Statistical population & sample (★★)
Sampling vs census: advantages & disadvantages (★★)
Sample design (steps, frames) (★★)
Probability & non-probability sampling (★★★)
Simple random sampling (methods, estimators) (★★★)
Stratified random sampling (allocation rules) (★★★)
Systematic sampling (interval, start, issues) (★★)

Cluster sampling (when/why; intra-cluster correlation) (★★)
Sampling & non-sampling error (★★)
Determination of sample size (for mean/proportion, precision) (★★★)

STATISTICS
Part-II
Marks - 50

1. Concept of probability: Basic Definitions, Approaches of Defining probability, Basic properties of probabilities, Notation and Graphical displays for events.

2. Rules of Probability: Special Addition rule, The complementation Rule, General Addition rule, Bivariate data and Contingency table. Joint and marginal probabilities, Multiplication rules, Conditional probabilities, Concept of Bayes' Theorem.

3. Random Variables and probability Distributions : Random variable, Discrete Probability Distribution, Binomial Probability, Hypergeometric distribution, Poisson distribution, Normal distribution.

1) Concept of Probability
Basic definitions (sample space, events) (★★)
Approaches (classical, relative frequency, axiomatic) (★★)
Basic properties of probabilities (complements, bounds) (★★)
Notation & graphical displays for events (Venn, set ops) (★)
2) Rules of Probability
Special Addition Rule (mutually exclusive) (★★)
Complementation Rule (★★)
General Addition Rule (non-mutually exclusive) (★★★)
Bivariate data & contingency table (setup, counts) (★★)
Joint & marginal probabilities (★★)
Multiplication rules (independent/dependent) (★★★)
Conditional probability (★★★)
Bayes' Theorem (classification/diagnostic problems) (★★★)
3) Random Variables & Probability Distributions
Random variable (discrete/continuous; pmf/pdf, cdf) (★★)
Discrete distributions overview (★)
Binomial distribution (mean/var; fitting; problems) (★★★)
Hypergeometric distribution (without replacement) (★★)
Poisson distribution (law of rare events; Poisson approx.) (★★★)
Normal distribution (standardization; areas) (★★★)

4. Sampling Distribution: Sampling distribution of the sample mean for a normally distributed variable, The Central Limit Theorem (CLT), Sampling Distribution of the Sample Mean, Sampling distribution of the sample proportion, Sampling distribution of function of mean and proportion. Confidence interval, Confidence interval of Population mean. Depermination of Sample size, Sampling for estimating mean, Sampling for estimating proportion.

Sampling distribution of sample mean (normal population) (★★★)

Central Limit Theorem (intuition & use) (★★★)

Sampling distribution of sample mean (general case via CLT) (★★★)

Sampling distribution of sample proportion (★★★)

Sampling distribution of functions of mean & proportion (ratios/linear comb.) (★★)

Confidence interval (concept & structure) (★★)

CI of population mean (known/unknown σ) (★★★)

Determination of sample size (error/CI-width based) (★★★)

Sampling for estimating mean (design + CI) (★★★)

Sampling for estimating proportion (design + CI) (★★★)

5. Basic Concepts of Hypothesis Testing : Null and Alternative hypothesis, simple and composite hypotheses, Test statistic, acceptance and rejection regions, type I and type II errors, the significance level, one tailed and two tailed tests, general procedure for test of hypothesis. Tests based on normal, student's t, F, and X^2 distribution. The Z- test for two population means. The pooled t-test for two population means. The paired t-test for two population means.

6. Analysis of Variance : Concept of analysis of variance, treatment, response, extraneous variables, One-Way Anova Model, Estimate of The Model Parameters, Hypothesis Testing In Anova. Two-Way Anova, significance of Correlation and rank correlation coefficients. Multiple comparison test. Two-way analysis of variance with and without interaction.

7. Experimental Designs: Basic principles of Experimental Design. Randomization, Replication and Local control. The completely Randomized Design (CRD), Randomized Complete Block Design (RCBD) and Latin square Design.

Basic Concepts of Hypothesis Testing

Null & alternative hypotheses (★★★)

Simple vs composite hypotheses (★★)

Test statistic; acceptance/rejection regions (★★★)

Type I & II errors; significance level; power (★★★)

One-tailed vs two-tailed tests (★★)

General procedure (formulate–assumptions–compute–decide) (★★★)

Tests using normal, Student's t, F, and χ^2 distributions (families & use-cases) (★★★)

Z-test for two population means (assumptions) (★★★)

Pooled t-test for two population means (equal variances) (★★★)

Paired t-test for two population means (dependent samples) (★★★)

Analysis of Variance (ANOVA)

Concept (variance decomposition) (★★★)

Treatment, response, extraneous variables (blocking) (★★)

One-Way ANOVA model & assumptions (★★★)
Estimation of model parameters (SS, MS, σ^2 via MSE) (★★)
Hypothesis testing in ANOVA (F-test) (★★★)
Two-Way ANOVA (with & without interaction) (★★★)
Significance of correlation and rank correlation coefficients (ANOVA/ttest logic) (★★)
Multiple comparison tests (LSD/Tukey outline; when used) (★★)
7) Experimental Designs
Principles: Randomization, Replication, Local Control (★★★)
Completely Randomized Design (CRD) (model, ANOVA table) (★★★)
Randomized Complete Block Design (RCBD) (blocks, efficiency) (★★★)
Latin Square Design (rows, columns, treatments) (★★★)

❒ **Roughly we get-**

(★★★): Means/median/mode; SD/CV; linear regression by least squares; R, r, R^2; index formulas (Laspeyres/Paasche/Fisher); trend/seasonal measures; sampling designs & size; CLT; CIs; z/t/χ^2/F tests; ANOVA; CRD/RCBD/LSD; Binomial/Poisson/Normal; Bayes/conditional rules.

(★★): Chebyshev vs normal rule; skewness/kurtosis; partial/rank/correlation ratio; hypergeometric; multiple comparison; significance of (rank) correlation.

(★): Diagram types; relative frequency; nuptiality/migration; general definitions/notations; irregular/cyclical components.

20 High voltage topic to be covered must

1. Mean formula comparison and limitations and when to use what (Grouped/Ungrouped) ★★★
2. Median (Grouped Data) ★★★
   - Must cover: median formula, interprelation, ogive method.
3. Mode (Grouped Data) ★★★
   - Must cover: formula for modal class, graphical method.
4. Standard Deviation & Variance ★★★
   - Must cover: raw data, grouped data, shortcut formulas.
5. Coefficient of Variation (CV) ★★★
   - Must cover: formula, interpretation, comparison between series.
6. Correlation Coefficient (Pearson's r) ★★★
   - Must cover: formula, interpretation (strength/direction), $-1 \leq r \leq +1$.
7. Regression (Least Squares Line) ★★★
   - Must cover: regression equations (y on x, x on y), slope interpretation.
8. Fisher's Ideal Index ★★★
   - Must cover: formula, time reversal test, factor reversal test.
9. Laspeyres & Paasche Indices ★★★
   - Must cover: formulas, differences, biases.

10. Time Series – Secular Trend ★★★
    - Must cover: moving average, least squares methods.
11. Time Series – Seasonal Variations ★★★
    - Must cover: ratio-to-moving average, link relatives method.
12. Conditional Probability & Bayes' Theorem ★★★
    - Must cover: P(A|B), diagnostic problems, tree diagram use.
13. Binomial Distribution ★★★
    - Must cover: formula, mean, variance, fitting, normal approximation.
14. Poisson Distribution ★★★
    - Must cover: formula, mean = variance property, approximation to Binomial.
15. Normal Distribution ★★★
    - Must cover: properties, standardization (Z), areas under curve.
16. Central Limit Theorem (CLT) ★★★
    - Must cover: statement, role in approximation, importance in inference.
17. Sampling Distribution of Mean & Proportion ★★★
    - Must cover: formulas, standard error, link to CI.
18. Confidence Interval (CI) ★★★
    - Must cover: CI for mean (σ known & unknown), CI for proportion.
19. Hypothesis Testing – z-test & t-test ★★★
    - Must cover: one-sample, two-sample, paired & pooled t-test.
20. One-Way ANOVA (F-test) ★★★
    - Must cover: model setup, assumptions, ANOVA table, interpretation.

1. Paasche=120, Laspeyres=125. Fisher's Index=?

Answer: 122.47 ✓

Explanation: Fisher=√(120×125)=√15000≈122.47.

2. Binomial, n=5, p=0.4, variance=?

Answer: 1.2 ✓

Explanation: Variance=npq=5×0.4×0.6=1.2.

3. In ANOVA, TSS decomposed as:

Answer: BSS+WSS ✓

Explanation: Total=Between+Within variation.

4. The sampling distribution of mean approaches normal if n is large. This is:

Answer: Central Limit Theorem ✓

Explanation: CLT states sample mean is normal for large n.

5. If Y = 10 + 0.8X: What is the interpretation of values a,b.

6. Two events A, B Given. Their values P(A), P(B) are given. P (AUB) also given. Check whether independent or not. Mutually exclusive or not.

7. R^2= value, interpretation.

7.1. beta value interpretation

8. In time series, seasonal variation repeats in:

Answer: Fixed period ✓

Explanation: Seasonal cycles repeat regularly (e.g., yearly).

9.Regression coefficients bxy=0.4, byx=0.5, correlation?

Answer: ± √0.2✓

Explanation: r^2=bxy·byx=0.2 ⇒ r=±√0.2.

10. In a normal distribution, the total area under the curve is:

Answer: 1 ✓

Explanation: Probability distribution integrates to 1.

11 . Which index number satisfies both time reversal and factor reversal tests?

Answer: Fisher's Index ✓

Explanation: Fisher's Ideal Index passes both tests.

12. Which of the following is a non-parametric test?

Answer: Chi-Square Test ✓

Explanation: Chi-square does not depend on population parameters.

13. When Strafied Sampling useful?

Answer: Sub groups are internally homogenous and externally heterogenous.

14. The variance of 5, 7, 9, 11 is:

Answer: 5 ✓

Explanation: Mean=8, deviations squared sum=20, variance=20/4=5.

15. The mean of first 10 natural numbers is:

Answer: 5.5 ✓

Explanation: Mean = (sum of 1 to 10)/10 = 55/10 = 5.5.

16. If median = mean = mode in a distribution, the distribution is:

Answer: Normal Distribution ✓

Explanation: In normal distribution, mean = median = mode.

17. Which of the following is not a measure of central tendency?

Answer: Standard Deviation ✓

Explanation: Standard deviation is a measure of dispersion, not central tendency.

18. The regression line of Y on X passes through:

Answer: (Mean of X, Mean of Y) ✓

Explanation: Regression lines always pass through means of variables.

19. The sum of deviations from mean is always:

Answer: Zero ✓

Explanation: By definition of arithmetic mean.

20. Coefficient of variation is:

Answer: (SD/Mean)×100 ✓

Explanation: Relative measure of dispersion.

পড়াশোনার কৌশল (পরীক্ষার আগে শেষ কয়েকদিনে)
প্রধান অধ্যায় নির্ধারণ করুন → Probability, Hypothesis Testing, t/z/F-test, ANOVA, Regression & Correlation, Time Series, Sampling – এগুলো সাধারণত বেশি আসে।
Formula Sheet বানান → শুধু ফর্মুলা ও সংক্ষিপ্ত নোট লিখে নিন। যেমন:
Mean, Variance, SD, Skewness, Kurtosis
Binomial/Poisson/Normal Distribution এর PMF/PDF, Expectation
t-test, z-test, F-test এর test statistic ফর্মুলা
ANOVA table structure
Correlation (Karl Pearson, Spearman Rank)
MCQ ও Conceptual প্রশ্ন রিভাইস করুন → যেমন Type-I, Type-II error, Degrees of freedom, Assumption of tests, Difference between tests.
পূর্বের প্রশ্নপত্র দেখে নিন → কোন টপিক থেকে বেশি আসে, তার প্রতি বেশি সময় দিন।
সংক্ষিপ্ত প্র্যাকটিস → প্রতিদিন numerical solve করে হাতের অভ্যাস বজায় রাখুন।
⏰ পরীক্ষার আগের দিন ও সকালে
নতুন কিছু শুরু করবেন না, শুধু ফর্মুলা ও ছোট নোট রিভাইস করুন।
দীর্ঘ numerical না, শুধু ছোট উদাহরণ ও স্টেপ মনে করুন।
ঘুম যথেষ্ট নিন (কমপক্ষে ৬–৭ ঘণ্টা)। পানি পান+ নিয়মিত ব্যায়াম করুন।

পরীক্ষার জন্য প্রয়োজনীয় জিনিস (কলম, ক্যালকুলেটর, এডমিট কার্ড) আগে গুছিয়ে নিন।
📝 ৩. পরীক্ষার হলে করণীয়
প্রশ্ন ভালোভাবে পড়ে বুঝুন → Numerical এর data গুলা হাইলাইট করুন।
সহজ প্রশ্ন আগে করুন → এতে সময় বাঁচবে ও আত্মবিশ্বাস বাড়বে।
সময়ের হিসাব রাখুন → শেষ ৫ মিনিট শুধু উত্তর চেক করার জন্য রাখুন। কোনো জানা উত্তর বাদ গেল কিনা+ রোল নম্বর ঠিক লিখলেন কিনা+ সেট নম্বর ঠিক লিখলেন কিনা চেক করুন।
মনে রাখবেন, ফিফটি ফিফটি সিউর থাকলে সেগুলো আন্সার করবেন, একদমই সিউর না হলে আন্সার করবেন না। বড় অঙ্ক আসলে, টাইম নষ্ট করবেন না, ছেড়ে দিন ওটা।
Last Moment Suggestions
Statistics: Definition, Characteristics and limitations
Statistics is specifically concerned with data — how to collect it, organize it, analyze it, interpret results, and present findings for decision-making. In other words- Statistics is the branch of mathematics that deals with the **collection, organization, presentation, analysis, and interpretation of numerical data** to assist in decision-making.
Characteristics of Statistics
-**Deals with data** – Statistics always works on collected data, not imagination.
-**Numerical in nature** – Information must be expressed in numbers for statistical analysis.
-**Aggregates of facts** – Statistics deals with groups or masses, not individual facts. Example: "The average income of 1000 families" (not one person's income).
-**Collected for a purpose** – Data is always gathered with a specific objective.

-Comparable – Data should be homogeneous (similar in type) to make valid comparisons.

-Subject to errors – Since it often involves sampling, some margin of error is always present.

- Statistics are **aggregate of interrelated** information.

Limitations of Statistics

-Does not study individuals – Focuses on groups, not single cases.

-Cannot give exact results – Always has some approximation or error.

-Depends on correct data – Wrong or biased data = wrong conclusions ("garbage in, garbage out").

-Can be misused purposively – Data can be manipulated to mislead people.

-Limited scope in qualitative aspects – Feelings, emotions, morality cannot be measured exactly in numbers.

-Does not prove causation – It shows correlation, not cause-effect directly.

- **Statistics rules are mutable.**

Scales of Measurement

Measurement are classified into **four levels**:

1. **Nominal Scale**
2. **Ordinal Scale**
3. **Interval Scale**
4. **Ratio Scale**

1. Nominal Scale (Classification / Naming Scale)

Meaning:

-It is the lowest level of measurement.

-Data are **classified into categories** without any order or ranking.

-Numbers are only **labels** (not quantities).

Properties:

-Categories are mutually exclusive (a person belongs to only one category).

-No mathematical operations (only counting frequencies, mode, percentages).

Examples:

-Gender: Male = 1, Female = 2, Others = 3.

-Religion: Muslim, Hindu, Christian, Buddhist.

-Blood group: A, B, AB, O.

Allowed Statistics:

Frequency, percentage, mode, chi-square test.

2. Ordinal Scale (Ranking Scale)

Meaning:

-Data can be **ordered or ranked** but the difference between ranks is not meaningful.

-Shows relative position, but not the exact difference.

Properties:

-Higher numbers = higher rank/order.
-Cannot measure how much greater one is than another.

Examples:

-Class position: 1st, 2nd, 3rd.
-Customer satisfaction: Very satisfied (5), Satisfied (4), Neutral (3), Dissatisfied (2), Very dissatisfied (1).
-Socioeconomic status: High, Middle, Low.

Allowed Statistics:

Median, percentile, rank correlation (Spearman's).

3. Interval Scale

Meaning:

-Data are ordered, and **differences between values are meaningful**.
-But there is **no true zero point** (zero does not mean absence).

Properties:

-Equal intervals between values.

-Ratios are meaningless because zero is arbitrary.

Examples:

-Temperature (Celsius, Fahrenheit):
30°C is hotter than 20°C (difference = 10°C meaningful). But 0°C ≠ "no temperature."
-IQ scores.
-Calendar years (2000, 2010, 2020).

Allowed Statistics:

Mean, standard deviation, correlation, regression.

4. Ratio Scale (Highest Level)

Meaning:

-Has all features of interval scale **plus a true zero point**.
-Allows comparison of both differences and ratios.

Properties:

-Equal intervals + meaningful zero.
-Ratios are meaningful (you can say "twice as much").

Examples:

-Height (0 cm = no height).
-Weight (0 kg = no weight).
-Age (0 years = newborn).
-Income (0 taka = no income).

Allowed Statistics:

All mathematical operations: mean, variance, standard deviation, geometric mean, coefficient of variation, etc.

Statistical Graphs & Diagrams

1. Histogram

Use:

-Shows the **frequency distribution of continuous data** (interval/ratio scale).

-Helps to identify distribution shape (normal, skewed, etc.).

-Useful for detecting patterns, spread, and outliers.

-Mode can be calculated from Histogram.

Limitations:

-Only for continuous data, not nominal.

-Choice of class interval (bin width) can change the appearance.

-Exact values are not shown, only ranges.

Comparison:

Similar to a bar chart, but bars are **adjacent (no gaps)** since data is continuous.

2. Frequency Polygon

Use:

-Line graph joining midpoints of histogram bars.

-Shows **shape of distribution** more clearly than histogram.

-Easy to compare two or more distributions on the same graph.

Limitations:

-Not as intuitive as histogram for beginners.

-Exact frequencies are harder to read.

Comparison:

Preferred when comparing multiple distributions, while histogram is better for single distribution.

3. Ogive Curve (Cumulative Frequency Curve)

Use:

-Shows **cumulative frequency** (less than or greater than type).

-Useful to find **median, quartiles, percentiles** graphically.

-Helps in understanding how data accumulates over a range.

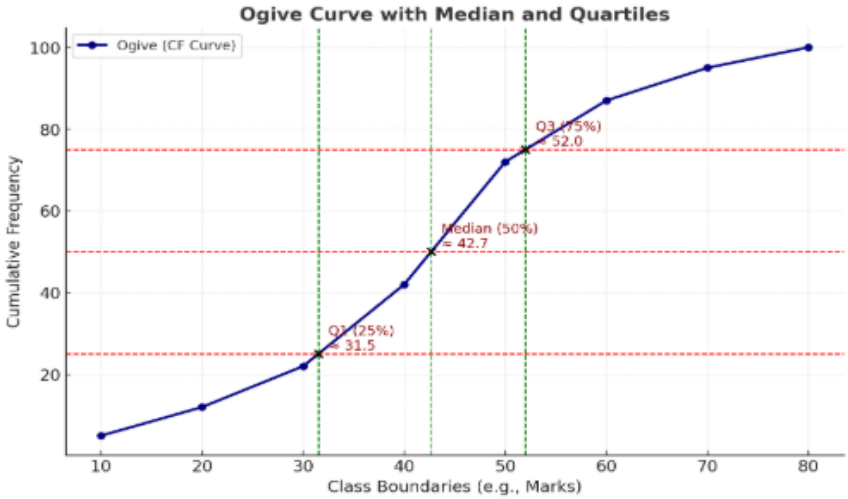
Limitations:

-Cannot display exact frequency of each class.

-Less useful for distribution shape.

Comparison:

Histogram/frequency polygon show distribution shape, while **ogive shows cumulative trend**.



4. Bar Chart

Use:

-Represents **categorical (discrete) data**.

-Each category shown with a bar of proportional length.

-Easy to compare different categories (e.g., sales of products, population by gender).

Limitations:

-Cannot represent continuous data directly.

-Misleading if scales are inconsistent or 3D bars are used.

Comparison:

Histogram vs Bar Chart → Histogram = continuous data (no gaps), Bar Chart = categorical data (gaps).

5. Pie Chart

Use:

-Circular chart divided into sectors (angles proportional to frequencies).

-Useful for showing **relative proportions/percentages** of categories.

-Easy for quick visual impression.

Limitations:

-Hard to compare slices if differences are small.

-Not suitable for large number of categories (>6).

-Does not show trend over time.

Comparison:

Bar chart is better when comparing many categories, pie chart is better for **composition (parts of a whole)**.

6. Boxplot (or Whisker Plot)

Use:

-Shows **five-number summary** (minimum, Q1, median, Q3, maximum).

-Identifies **spread, symmetry, skewness, and outliers**.

-Good for comparing distributions across groups.

Limitations:

-Does not show detailed frequency distribution.

-Not intuitive for non-statistical audiences.

Comparison:

-Histogram shows full distribution shape, Boxplot summarizes data with quartiles and outliers.

-Boxplot is excellent for comparing multiple groups side by side.

Measures of Central Tendency

Central Tendency = A single value that represents the **center** of a dataset (around which other values cluster).

The main measures are:

Mean (Arithmetic, Geometric, Harmonic)

Median

Mode

1. Arithmetic Mean (AM)

- **Definition:** Sum of all values divided by number of observations.
- **Formula:**

$$\bar{X} = \frac{\sum X}{N}$$

- **Strengths:**
  - Uses all data values.
  - Easy to compute.
  - Suitable for further statistical analysis (variance, regression, etc.).
- **Weaknesses:**
  - Affected by extreme values (outliers).
  - Not suitable for qualitative data or skewed distribution.
- **Best Use:** Balanced data without outliers (e.g., average marks, production).

2. Median

Formula:

- For **ungrouped data:**

  Arrange values in ascending order.
  - If N is odd → Median = Middle value.
  - If N is even → Median = Average of two middle values.
- For **grouped data:**

$$\text{Median} = L + \left(\frac{\frac{N}{2} - CF}{f} \right) \times h$$

Where:

- L = lower boundary of median class
- N = total frequency
- CF = cumulative frequency before median class
- f = frequency of median class
- h = class width

Uses:

- Best when data is skewed or has outliers.
- Used in income distribution, property prices, test scores.

Limitations:

- Ignores extreme values and exact magnitudes.
- Not suitable for algebraic treatment.
- Depends on arrangement of data.

🔹 3. Mode

Formula (grouped data):

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where:

- L = lower boundary of modal class
- f_1 = frequency of modal class
- f_0 = frequency of class before modal class
- f_2 = frequency of class after modal class
- h = class width

Uses:

- Shows most common value.
- Useful in business and marketing (e.g., most sold shoe size, most preferred product).
- Works well with nominal data.

Limitations:

- Sometimes no mode, sometimes multiple modes.
- Not suitable for small datasets.
- Ignores extreme values and distribution shape.

4. Geometric Mean (GM)

Formula:

- For **ungrouped data:**

$$GM = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$$

- For **grouped data:**

$$GM = \text{antilog}\left(\frac{\sum f \log X}{\sum f}\right)$$

Uses:

-Used for growth rates (population, interest rate, inflation).

-Suitable when values are percentages, ratios, or indices.

Limitations:

-Cannot be used if data has zero or negative values.

-Calculation is complex (requires log tables/computer).

5. Harmonic Mean (HM)

Formula:

- For **ungrouped data:**

$$HM = \frac{n}{\sum \frac{1}{X_i}}$$

- For **grouped data:**

$$HM = \frac{\sum f}{\sum \frac{f}{X}}$$

Uses:

-Best for rates, speeds, ratios.

Example: Average speed of a car traveling different distances at different speeds.

Limitations:

Strongly affected by small values.

Difficult to calculate manually.

Not widely understood compared to AM or Median.

Comparison among the Measures of Central Tendency

The measures of central tendency—mean, median, mode, geometric mean, and harmonic mean—each describe the "center" of a dataset in their own way, but they differ in

definition, method of calculation, and suitability for various types of data. The **arithmetic mean** is the most common, obtained by dividing the sum of all observations by their number. It is simple, makes use of every data value, and is algebraically convenient for further analysis, but it is highly affected by extreme values and is not suitable for skewed distributions. The **median**, on the other hand, represents the middle value when data are arranged in order. It is not influenced by very large or small values and is therefore useful in skewed or open-ended distributions such as income and wealth data. However, it does not utilize all values and cannot be used in mathematical analysis. The **mode** indicates the most frequently occurring value in a dataset. It is the only measure that can be used with nominal or categorical data, making it valuable in business or marketing to identify the most preferred product or size. Its weakness is that a dataset may have no mode or multiple modes, making it less precise. The **geometric mean** is the nth root of the product of observations, and it is particularly useful in analyzing growth rates, percentages, and index numbers. It is less affected by extreme values than the mean, but it cannot be applied to data containing zero or negative values and is computationally complex. Finally, the **harmonic mean** is the reciprocal of the arithmetic mean of reciprocals, and it is best for averaging rates, speeds, and ratios. It gives greater weight to smaller values, but this also makes it very sensitive to extremely small data points.

In summary, the arithmetic mean is most suitable for general balanced data, the median is best when distributions are skewed, the mode is appropriate for categorical or nominal data, the geometric mean is ideal for growth-related studies, and the harmonic mean is used for rates and ratios. Each measure has its own advantages and limitations, and the choice depends on the nature of the data and the purpose of analysis.

Some important formulas:

- If each values of a series are equal, then, AM=GM=HM

-In case of two non zero and positive values , $AM*HM=(GM)^2$

- For two or more non zero positive values, AM≥GM≥HM

- For first n natural numbers mean= $\frac{n+1}{2}$

Measurement of dispersion

1. Range

Formula:

$$\text{Range} = X_{\max} - X_{\min}$$

where $X_{\max}$ = maximum value, $X_{\min}$ = minimum value.

Use:

- Measures the total spread of the data.
- Easy to calculate and understand.
- Helpful in quick comparison of datasets.

Limitation:

- Depends only on extreme values, ignores rest of data.
- Very sensitive to outliers.
- Not a reliable measure of variability for large datasets.

2. Variance (σ^2 or s^2)

Formula (Population):

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

Formula (Sample):

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1}$$

Use:

- Measures average squared deviation from mean.
- Forms basis for other measures (e.g., SD, ANOVA, regression).
- Useful in comparing variability across datasets.

Limitation:

- Units are squared, so interpretation is not straightforward.
- Sensitive to extreme values.

3. Standard Deviation (SD, σ or s)

Formula:

$$\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}} \quad ; \quad s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Use:

- Square root of variance; same units as data.
- Widely used to measure dispersion.
- Useful in finance, science, social studies, hypothesis testing.

Limitation:

- Still sensitive to extreme values.
- Assumes data is interval/ratio scale.

4. Coefficient of Variation (CV)

Formula:

$$CV = \frac{\text{SD}}{\text{Mean}} \times 100\%$$

Use:

- Standardizes variability, independent of units.
- Useful for comparing variation between datasets with different means/scales.
- Widely used in finance (risk-to-return ratio).

Limitation:

- Not meaningful when mean = 0 or very close to 0.
- Only valid for ratio scale data.

5. Quartiles (Q1, Q2, Q3)

Formula (general position):

Q_k=Value at position $k(N+1)/4$, $k=1,2,3$

Q1: 25th percentile (lower quartile).

Q2: 50th percentile (median).

Q3: 75th percentile (upper quartile).

Use:

-Divide data into four equal parts.

-Helpful in understanding distribution and spread.

-Basis for IQR and boxplots.

Limitation:

-Ignores finer details of distribution.

6. Percentiles

Formula (general position):

P_k=Value at position $k(N+1)/100, k=1,2,...,99$

Use:

-Divide data into 100 equal parts.

-Widely used in standardized tests (e.g., scoring in exams, health data).

-Helps compare individual observation relative to group.

Limitation:

-Sensitive to extreme data values.

7. **Deciles**

Formula:

D_k=Value at position $k(N+1)/10, k=1,2,...,9$

Use:

-Divide data into 10 equal parts.

-Useful in economics (income distribution, wealth inequality).

-Helps summarize large datasets.

Limitation:

-Like quartiles, may ignore finer data details.

-Different formulas exist for position.

8. Interquartile Range (IQR)

Formula:

IQR=Q3-Q1

Use:

-Measures spread of the middle 50% of data.

-Robust against outliers.

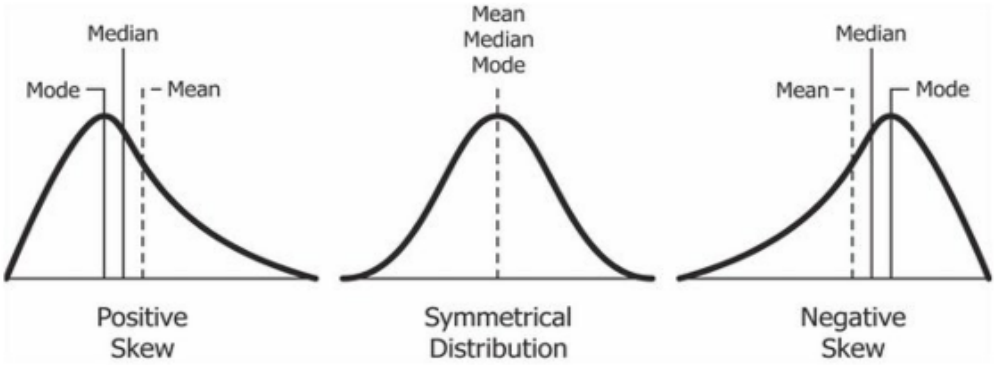
-Common in boxplots and non-parametric statistics.

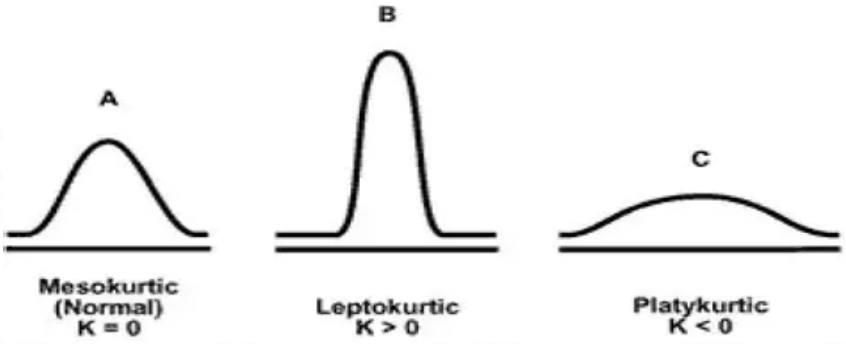
Limitation:

-Ignores data outside Q1 and Q3.

-Less informative if extreme values are important.

Skewness and kurtosis





Moments are numerical values that describe the **shape** of a distribution.

(b) Central Moments (about mean)

$$\mu_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^r$$

- $\mu_1 = 0$ (always zero about mean)
- $\mu_2 =$ variance
- $\mu_3, \mu_4 \rightarrow$ used for skewness & kurtosis.

(c) Skewness (3rd Moment)

It measures asymmetry.

$$\text{Skewness} = \gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

- If **γ1 = 0** → symmetric (normal distribution).
- If **γ1 > 0** → positively skewed (tail on right).
- If **γ1 < 0** → negatively skewed (tail on left).

Alternative (Karl Pearson's Coefficient):

$$Sk = \frac{\bar{x} - \text{Mode}}{SD} \quad \text{or} \quad \frac{3(\bar{x} - \text{Median})}{SD}$$

(d) Kurtosis (4th Moment)

It measures "peakedness" or "flatness".

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\gamma_2 = \beta_2 - 3$$

- **γ2 = 0** → Mesokurtic (Normal curve).
- **γ2 > 0** → Leptokurtic (More peaked).
- **γ2 < 0** → Platykurtic (Flatter).

1. Given Moments (symmetric case)

$$\mu_2 = 16,\ \mu_3 = 0,\ \mu_4 = 768$$

- Skewness:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{0}{16^{3/2}} = 0$$

👉 Distribution is **symmetric**.

- Kurtosis:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{768}{16^2} = \frac{768}{256} = 3$$

$$\gamma_2 = \beta_2 - 3 = 0$$

👉 **Mesokurtic (normal-like).**

Important Formula: β2≥β1+1

Regression, Correlation

In statistics, correlation and regression are two essential tools to study relationships between variables.

- Correlation → Measures how strongly two variables move together.

- Regression → Predicts the value of one variable (dependent) from another (independent).

Example:

Correlation: Height & Weight – do taller students weigh more?

Regression: Predicting a student's weight from their height.

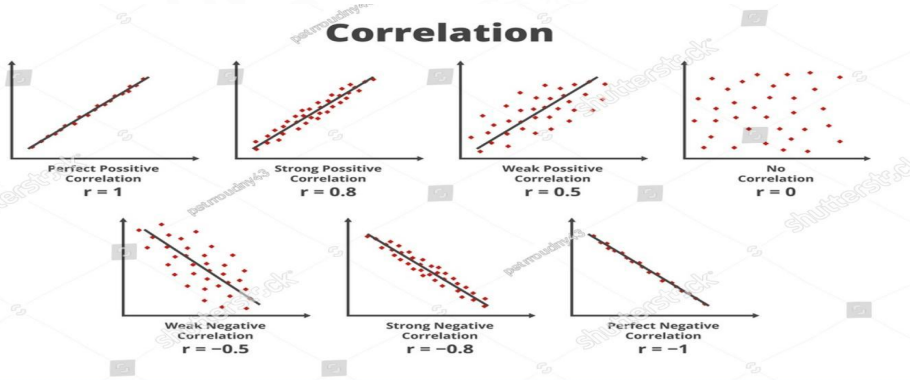
Correlation

Correlation measures strength and direction of a relationship between variables. It is mainly association between variables.

Positive correlation → both increase together.

Negative correlation → one increases, other decreases.

Zero correlation → no predictable change.



Pearson's Correlation Coefficient (r):

Formula:

$r = \Sigma(x - \bar{x})(y - \bar{y}) / \sqrt{(\Sigma(x - \bar{x})^2 * \Sigma(y - \bar{y})^2)}$

Range: $-1 \leq r \leq +1$

Interpretation: r = +1 perfect positive, r = -1 perfect negative, r = 0 no relation.

Rank Correlation (Spearman's ρ):

Formula: $\rho = 1 - (6 \Sigma d^2) / (n(n^2 - 1))$

Partial Correlation:

Measures correlation between first & second variable after removing effect of third variable.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Correlation Ratio (η):

Used for non-linear relationships.

Regression

Regression finds the best-fit equation predicting Y from X.

Simple Linear Regression Equation:

$Y = a + bX$

a = intercept, b = slope.

Least Squares Method:

$b = \Sigma(x - \bar{x})(y - \bar{y}) / \Sigma(x - \bar{x})^2$

$a = \bar{y} - b\,\bar{x}$

Regression Coefficients:

b_{xy}: Y on X slope, b_{yx}: X on Y slope.

Relation: $r^2 = b_{yx} \times b_{xy}$

Multiple Regression

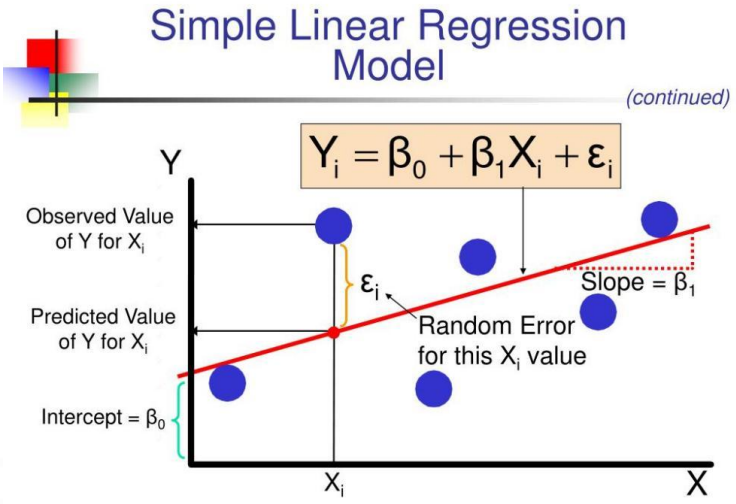
Equation:

$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$

Multiple Correlation Coefficient (R):

Measures correlation between actual Y & predicted $\hat{Y}$.

Coefficient of Determination (R^2):
R^2 = Explained SS / Total SS
Indicates % of variation in Y explained by X variables.
Some Conceptual Questions.



1. What is correlation in statistics?
Correlation measures the degree and direction of a linear (or other) relationship between two variables.
Explanation: It does not imply causation — it only shows association.
2. Name the two main types of correlation.
Positive and negative correlation.
Explanation: Positive means variables move in the same direction; negative means they move in opposite directions.
3. What is the range of Pearson's correlation coefficient?
From **–1 to +1.**
Explanation: –1 = perfect negative linear relation, +1 = perfect positive linear relation, 0 = no linear relation.
4. What does a correlation coefficient of 0.8 indicate?
A strong positive linear relationship between the variables.
Explanation: The closer to ±1, the stronger the linear association.
5. When should rank correlation be used?
When data is ordinal or when precise numerical values are not available.
Explanation: Spearman's ρ works on ranks, not raw values.
8. What is partial correlation?
It measures the correlation between two variables while controlling for the effect of one or more other variables.
Explanation: Useful to isolate the relationship between two variables.

9. What is correlation ratio (η)?

A measure of association used when the relationship may be nonlinear.

Explanation: It's often used when one variable is categorical.

10. Why is correlation not the same as regression?

Correlation measures association, regression models prediction.

Explanation: Regression assumes dependence of Y on X.

11. What is regression analysis used for?

Predicting the value of a dependent variable from one or more independent variables.

Explanation: It also quantifies the relationship between variables.

12. What is simple linear regression?

Regression with one dependent variable and one independent variable.

Explanation: The equation is $Y=a+bX$.

13. What does the least squares method do in regression?

It minimizes the sum of squared residuals between observed and predicted values.

Explanation: This ensures the "best fit" line.

14. What is a regression coefficient?

The value that represents the change in Y for a one-unit change in X.

Explanation: In $Y=a+bX$; b is the regression coefficient.

15. What is the intercept in regression?

The predicted value of Y when X = 0.

16. What is multiple regression?

Regression with one dependent variable and two or more independent variables.

Explanation: Equation: $Y=a+b_1X_1+b_2X_2+...+b_kX_k$.

17. What is the multiple correlation coefficient (R)?

The correlation between the actual Y values and the predicted Y values from multiple regression.

Explanation: R ranges from 0 to 1.

18. What is the coefficient of determination (R^2)?

The proportion of variance in Y explained by the model.

Explanation: R^2 = SSR / SST.

19. What does R^2 = 0.85 mean?

85% of the variation in Y is explained by the independent variables.

Explanation: The remaining 15% is due to other factors or randomness.

20. What is the adjusted R^2 used for?

It adjusts R^2 for the number of predictors to prevent overestimation.

Explanation: Useful in multiple regression when comparing models.

21. Q: X: 1,2,3,4,5; Y: 2,4,6,8,10 — What is r?

Answer: Answer: r = 1. Reason: Y = 2X, perfect positive linear relationship.

22. Pearson's r (Perfect Negative Correlation)

Q: X: 1,2,3,4,5; Y: 10,8,6,4,2 — What is r?

Answer: Answer: r = -1. Reason: perfect negative linear relationship.

23. Regression equation from b and a

Q: If b = 1.1 and a = 2.6, write regression equation.

Answer: Answer: Y = 2.6 + 1.1X.

24. Predict Y using regression

Q: Using Y = 2.6 + 1.1X, find Y when X = 7.

Answer: Answer: Y = 2.6 + 1.1*7=10.3 .

25. Spearman's ρ interpretation

Q: Given ρ = 0.8 — what does it mean?

Answer: Answer: Strong positive rank correlation.

Demography and Index Number

- GFR – General Fertility Rate

Definition:

The General Fertility Rate measures the number of live births per 1,000 women of reproductive age (usually aged 15–49) in a year. It focuses only on the segment of the population capable of giving birth, unlike CBR which considers the entire population.

Formula:

GFR = (Number of live births in a year / Number of women aged 15–49) × 1000

Example:

Live births in a year: 60,000

Women aged 15–49: 1,200,000

GFR = (60,000 / 1,200,000) × 1000 = 50 births per 1,000 women aged 15–49

- ASFR – Age-Specific Fertility Rate

Definition:

ASFR measures the number of live births per 1,000 women in a specific age group (e.g. 15-19, 20–24, 25–29, 30-34, 35-39,40-44,45-49) in a given year. It shows which age group has the highest fertility.

Formula:

ASFR(i) = (Number of live births to women in age group(i) / Number of women in that age group(i))× 1000

Example:

Live births to women aged 20–24: 12,000

Women aged 20–24: 240,000

$ASFR_{(20–24)}$ = (12,000 / 240,000) × 1000 = 50 births per 1,000 women aged 20–24

- TFR – Total Fertility Rate

Definition:

TFR estimates the average number of children a woman would have during her reproductive years (15–49), if she experienced the current ASFRs for each age group throughout her life. It is calculated by summing up ASFRs and adjusting for the width

of the age groups. TFR adjusts for age composition, making it better for comparing across countries or years.

Formula (with 5-year age groups):

TFR = 5Σ(ASFR for each age group)

The "× 5" adjusts for each 5-year interval in the age groups.

- Example:

Suppose we have ASFR data:

15–19: 40

20–24: 120

25–29: 100

30–34: 60

35–39: 30

40–44: 10

45–49: 2

- TFR = ((40 + 120 + 100 + 60 + 30 + 10 + 2) × 5) ÷ 1000 = (362 × 5) ÷ 1000 = 1.81 children per woman
- Gross Reproduction Rate

The Gross Reproduction Rate (GRR) is the average number of daughters a woman would have during her lifetime if she experienced the current age-specific fertility rates (ASFR), assuming no mortality before the end of the reproductive period. It is similar to TFR but considers only female births.

- Net Reproduction Rate

The Net Reproduction Rate (NRR) measures the average number of daughters a woman would have during her lifetime, accounting for mortality. It reflects the extent to which each generation of women is replacing itself.

Formula: NRR = GRR × Survival rate of women to each reproductive age

Or, more formally: NRR = Σ (ASFR_x × Proportion female births × Probability of survival to age group × 5) ÷ 1000

Example:

TFR = 3.0, Female proportion = 0.49, GRR = 1.47, Survival rate to end of reproductive period = 0.90

NRR = 1.47 × 0.90 = 1.323 daughters per woman

Interpretation of GRR and NRR Values

1. GRR (Gross Reproduction Rate)

> 1 → Each woman, on average, has more than 1 daughter → potential for population growth (assuming no female mortality before end of reproductive years).

= 1 → Each woman has exactly 1 daughter → replacement level (population stable if mortality and migration are neutral).

< 1 → Each woman has fewer than 1 daughter → potential population decline over time if conditions remain unchanged.

2. NRR (Net Reproduction Rate)

> 1 → Each generation of women is larger than the previous one (growth after accounting for mortality).
= 1 → Each generation exactly replaces itself → long-term stability.
< 1 → Each generation of women is smaller → declining population in the long run (if migration is absent).
GRR=NRR= All new born girl babies cannot die before their last potent age.

Dependency Ratio in Bangladesh

The dependency ratio measures the proportion of individuals considered economically dependent—typically children (ages 0–14) and older adults (65+)—relative to the working-age population (ages 15–64). It is expressed per 100 working-age individuals.
Formula:
Dependency Ratio = [(Population aged 0–14) + (Population aged 65+)] / (Population aged 15–64) × 100

1. What is an Index Number?

An **index number** is a statistical measure that shows **relative change in a variable or group of related variables over time, across regions, or between different groups.**
It's usually expressed as a percentage relative to a **base period.**

Formula (general):

$$\text{Index Number} = \frac{\text{Value in current period}}{\text{Value in base period}} \times 100$$

Example:

If the price of rice was 50 BDT last year and 55 BDT this year:

$$\text{Price Index} = \frac{55}{50} \times 100 = 110$$

→ Prices increased by 10%.

---Classification of Index Numbers

1. Price Index – measures price changes.
2. Quantity Index – measures change in quantities.
3. Value Index – measures change in total value.
4. Special Purpose Indices – cost of living index, etc.

1. Unweighted Index Numbers

These do not use weights; all items are treated equally.
Examples:
-Simple Price Index (Unweighted) – Measures average price change of a group of commodities.
-Simple Quantity Index (Unweighted) – Measures average change in quantity of items.
-Aggregate Price Index (Unweighted) – (Sum of current prices / sum of base prices) × 100.
-Aggregate Quantity Index (Unweighted) – (Sum of current quantities / sum of base quantities) × 100.
Key point: Easy to calculate, but less accurate when items have different importance.
2. Weighted Index Numbers
These use weights to reflect relative importance of items.

Examples:
-Laspeyres Price Index (Weighted)
Uses base period quantities as weights.
-Paasche Price Index (Weighted)
Uses current period quantities as weights.
-Fisher's Ideal Index (Weighted)
Geometric mean of Laspeyres and Paasche → considered most reliable.
-Marshall-Edgeworth Index (Weighted)
Uses average of base and current period quantities as weights.
-Cost-of-Living Index (Weighted)
Measures changes in expenditure needed to maintain standard of living; uses weights based on consumption.
Consumer Price Index (CPI) – Weighted, based on expenditure patterns of consumers.

4. Key Differences Between Weighted and Unweighted Index Numbers

Feature	Unweighted Index	Weighted Index
Weight	No weight, all items treated equally	Weight assigned based on importance (price, quantity, expenditure)
Accuracy	Less accurate	More accurate reflection of real change
Formula	Simple arithmetic mean	Weighted mean or ratio of weighted sums
Example	Average of price changes of items	CPI uses expenditure weights of goods

Unweighted Index Numbers

1. **Simple Price Index (Laspeyres, Paasche simplified)**
   - **Simple Price Index Formula:**

$$I_p = \frac{\sum P_t}{\sum P_0} \times 100$$

   Where P_t = price in current period, P_0 = price in base period.
2. **Simple Quantity Index**
   Measures change in quantity over time:

$$I_q = \frac{\sum Q_t}{\sum Q_0} \times 100$$

Weighted Index

1. **Laspeyres Price Index** (base period weighting)

$$I_L = \frac{\sum(P_t \cdot Q_0)}{\sum(P_0 \cdot Q_0)} \times 100$$

- P = price, Q_0 = quantity in base period
- Emphasizes **base period importance.**

2. **Paasche Price Index** (current period weighting)

$$I_P = \frac{\sum(P_t \cdot Q_t)}{\sum(P_0 \cdot Q_t)} \times 100$$

- Uses **current period quantities** as weights.

3. **Fisher's Ideal Index**

$$I_F = \sqrt{I_L \cdot I_P}$$

- Geometric mean of Laspeyres and Paasche → considered **most accurate.**

4. Dorbish-Bowley's Index= (L+P)/2

- Arithmetic mean of Laspeyres and Paasche

Laspeyres Price Index (LPI) Bias

Definition: LPI uses base period quantities as weights.

Main Bias: Upward bias (overstatement of inflation)

Reason:

-It assumes people continue to buy the same quantity of goods as in the base period, even if prices rise.

-In reality, consumers substitute expensive items with cheaper alternatives.

-Because LPI ignores substitution, it overstates the cost of living.

Paasche Price Index (PPI) Bias

Definition: PPI uses current period quantities as weights.

Main Bias: Downward bias (understatement of inflation)

Reason:

PPI reflects that people buy less of the more expensive goods in the current period.

This reduces the measured price increase compared to reality.

Therefore, PPI tends to understate the true increase in cost of living.

Key Point:

This is why **Fisher's Ideal Index** is preferred—it **balances the upward bias of LPI and downward bias of PPI.**

Why Fisher Index is Ideal?

-Balances the overstatement of LPI and understatement of PPI.

-Considered most accurate and reliable.

-Satisfies time reversal and factor reversal tests → mathematically consistent.

But It has several limitations:

-More complex to calculate.

-Requires data for both base and current period quantities → sometimes hard to collect.

Index	Strengths	Limitations
Laspeyres	Easy, based on base quantities, widely used	Overstates inflation, ignores substitution effect, outdated quantities
Paasche	Reflects current consumption, more realistic	Harder to collect data, understates inflation, not good for historical comparison
Fisher	Accurate, balances L & P, mathematically consistent	Complex calculation, needs both period data
Durbin	Flexible weighting, accommodates changing quantities	Complex, rare, less intuitive

Data for two commodities:

Commodity	P_0	P_1	Q_0	Q_1
M	10	12	5	6
N	20	25	8	7

Compute **Laspeyres, Paasche, and Fisher indices.**

Step 1: Laspeyres Price Index (LPI)

Formula:

$$L = \frac{\sum(P_1 \cdot Q_0)}{\sum(P_0 \cdot Q_0)} \times 100$$

Numerator:

$$(12 \cdot 5) + (25 \cdot 8) = 60 + 200 = 260$$

Denominator:

$$(10 \cdot 5) + (20 \cdot 8) = 50 + 160 = 210$$

LPI:

$$L = \frac{260}{210} \times 100 \approx 123.81$$

Step 2: Paasche Price Index (PPI)

Formula:

$$P = \frac{\sum(P_1 \cdot Q_1)}{\sum(P_0 \cdot Q_1)} \times 100$$

Numerator:

$$(12 \cdot 6) + (25 \cdot 7) = 72 + 175 = 247$$

Denominator:

$$(10 \cdot 6) + (20 \cdot 7) = 60 + 140 = 200$$

PPI:

$$P = \frac{247}{200} \times 100 = 123.5$$

Step 3: Fisher Ideal Index (F)

Formula:

$$F = \sqrt{L \cdot P}$$

Calculation:

$$F = \sqrt{123.81 \cdot 123.5} = \sqrt{15294} \approx 123.6$$

1. Time Reversal Test

👉 **Meaning:**

An index number should be consistent if we reverse the time periods.

Mathematically,

$$P_{01} \times P_{10} = 1$$

where P_{01} = price index of current year (1) with base year (0),

and P_{10} = price index of base year with current year as base.

👉 **Example:**

Suppose two commodities:

Commodity	Price in 0 (base)	Price in 1 (current)
A	10	20
B	20	25

- Using Fisher's Index:
  $P_{01} = 1.414$ (approx)
  $P_{10} = 0.707$ (approx)
  Multiply: $1.414 \times 0.707 = 1$ ✅ (passes test).

👉 **Conclusion**: Fisher's index passes time reversal, Laspeyres/Paasche do not.

$$P_{01}^F = \sqrt{P_{01}^L \cdot P_{01}^P} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \cdot \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

$$P_{10}^F = \sqrt{P_{10}^L \, P_{10}^P} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \cdot \frac{\sum p_0 q_0}{\sum p_1 q_0}}.$$

We see that their product equals 1.

The intuition is: Let, price of wheat 2003-2004 increased from 1110 to 1480 per quintal, the price in 2004 should be 133.33 percent of price in 2003. And price in 2003 should be 75 percent of price in 2004.
One figure is reciprocal of other. Their product (1.33*.75) =1

2. Factor Reversal Test

👉 **Meaning**:
If we multiply price index (P) and quantity index (Q), we should get the value index (ratio of total values).

$$P_{01} \times Q_{01} = V_{01}$$

where $V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$.

$$P_{01}^F \cdot Q_{01}^F = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \cdot \frac{\sum p_1 q_1}{\sum p_0 q_1}} \cdot \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \cdot \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

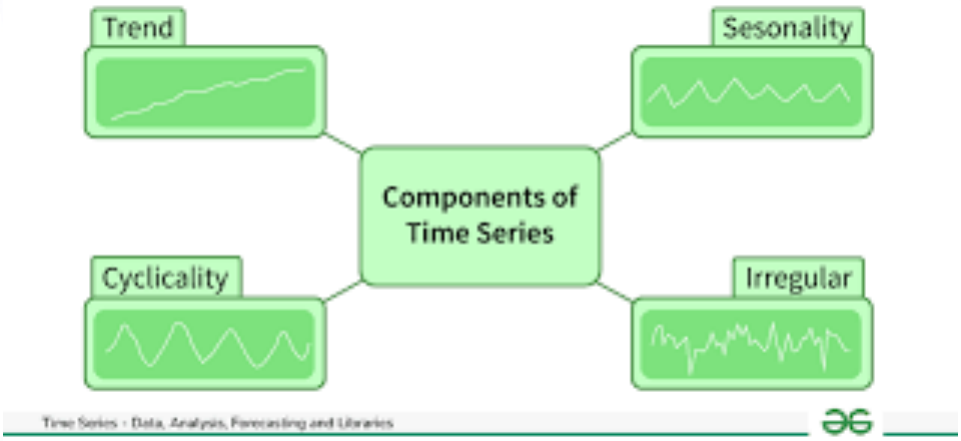
Imagine this:
You run a small grocery shop.
One month, you sell apples and bananas.
Next month, prices go up and maybe you sell more or fewer fruits.
Now you want to measure:
How much prices changed.
How much quantities sold changed.

How much total sales value changed.
The Factor Reversal Test says:
If you measure the price change and the quantity change correctly, then multiply them, you should get exactly the total change in sales value.
Real-life analogy
Price Index = "things got more expensive."
Quantity Index = "you sold more or fewer things."
Multiply them → "total money earned changed exactly as it should."
✓ If your method satisfies this, it passes the Factor Reversal Test.
It's like checking your shopping bill:
If apples are more expensive but you bought fewer, the net effect on the total bill should match what you get by combining price and quantity changes.

Laspeyres Formula for Consumer Price Index(CPI) or Cost of Living Index(CLI)

The Laspeyres Price Index fixes the quantities of the base year and compares how much the same basket costs in the current year.
Formula: CPI = $(\Sigma p_1 q_0 / \Sigma p_0 q_0) \times 100$
Example: Rice: $p_0=30$, $q_0=100$; Wheat: $p_0=20$, $q_0=50$; $p_1=36$, 25.
CPI = (3600 + 1250) / (3000 + 1000) × 100 = 4850 / 4000 × 100 = 121.25
CPI = 121.25 means the price level has increased by 21.25% compared to the base year.

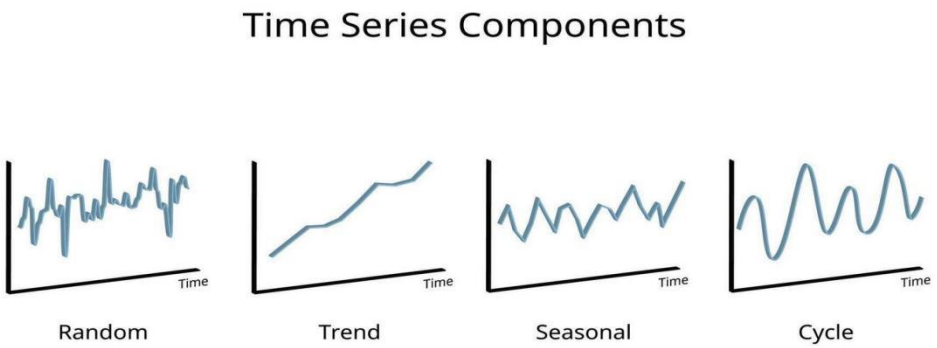
- **Laspeyres → overstates inflation (upward bias)**
- **Paasche → understates inflation (downward bias)**
- **Fisher → balances over- and underestimation**

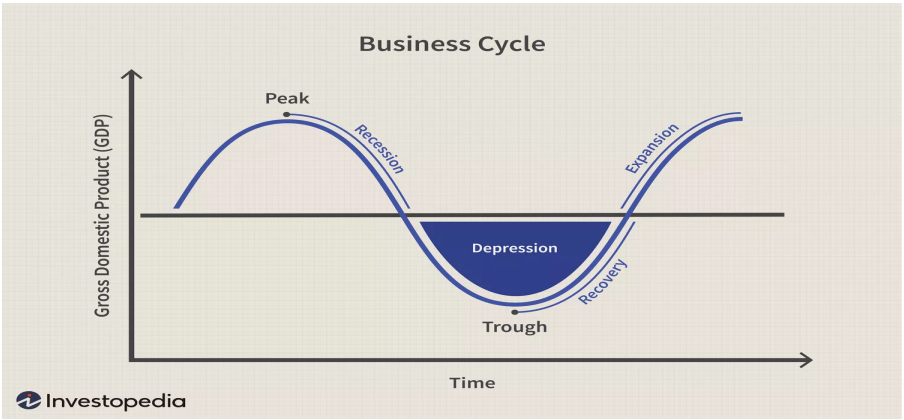


Time Series

Components of Time Series

1. Secular Trend (T): Long-term movement over a long period (e.g., gradual rise in population).
2. Seasonal Variation (S): Regular patterns that repeat within a year (e.g., more ice-cream sales in summer).
3. Cyclical Variation (C): Long-term wave-like movements often linked to business cycles (e.g., recession and boom).
4. Irregular Variation (I): Unpredictable variations caused by unusual events (e.g., flood, war, pandemic).





Formula of Time Series

General model: $Y = T \times S \times C \times I$ (Multiplicative model)

Or, $Y = T + S + C + I$ (Additive model)

Methods of Measuring Straight line Trend

- Graphical/free hand method

-Semi average method

-Method of least squares

Methods of Measuring Non linear Trend

-Freehand or graphical method

-Moving Average method

Methods of Measuring Seasonal Variation

-Method of simple averages

-Ratio to trend method

-Ratio to moving average

-Link relatives method

Methods of Measuring cyclic variation

-Residual method

- Direct method

- Reference cycle analysis method

Semi average method

Description:

Divide the time series into two equal parts, calculate the average of each part, then plot these averages and draw a straight line between them.

Advantages:

Simple → only uses averages.

Gives a straight-line trend.

More objective than graphical method.

Limitations:

Only suitable for linear trends.

If number of observations is odd, technique needs adjustment.

Ignores fluctuations within the parts.

Affect of extreme values

Usefulness:

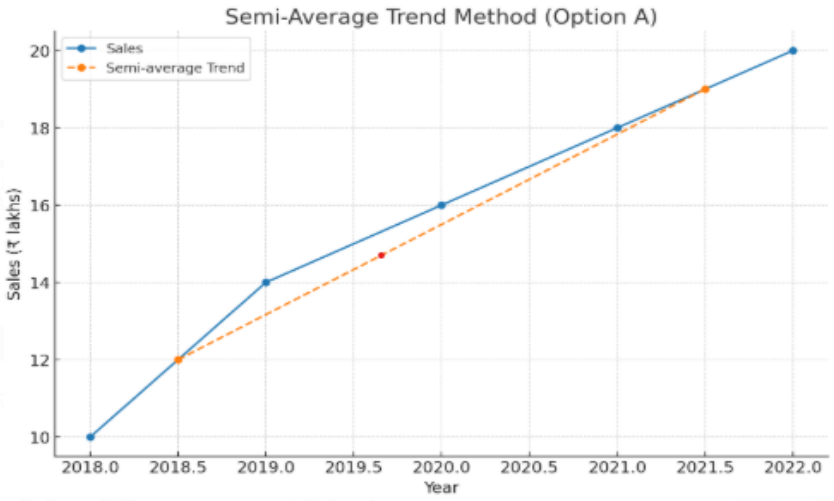
Quick but crude. Works when a broad linear trend is adequate.

Year	Sales
2018	10
2019	14
2020	16
2021	18
2022	20

Drop 2020, so we compare:

1st half: 2018 & 2019: Average is calculated

2nd half: 2021 & 2022: Average is calculated



Moving Average Method

Description:

Calculate arithmetic average of observations within a fixed window (say 3-year, 5-year, etc.), then move the window forward one period at a time.

Features:

Smooths out short-term seasonal & irregular fluctuations.

Flexible window length.

Limitations:

Trend values are not available for all periods (missing at ends).

Choice of period length is arbitrary.

Doesn't give a forecasting equation.

Usefulness:

Excellent at smoothing data when seasonal effects are present. Good for exploratory analysis.

Example: 3-Year Moving Average

Suppose the annual sales (in ₹ lakhs) of a company are:

Year	Sales
2018	10
2019	14
2020	16
2021	18
2022	20

Step-by-Step:

We take 3-year moving averages:

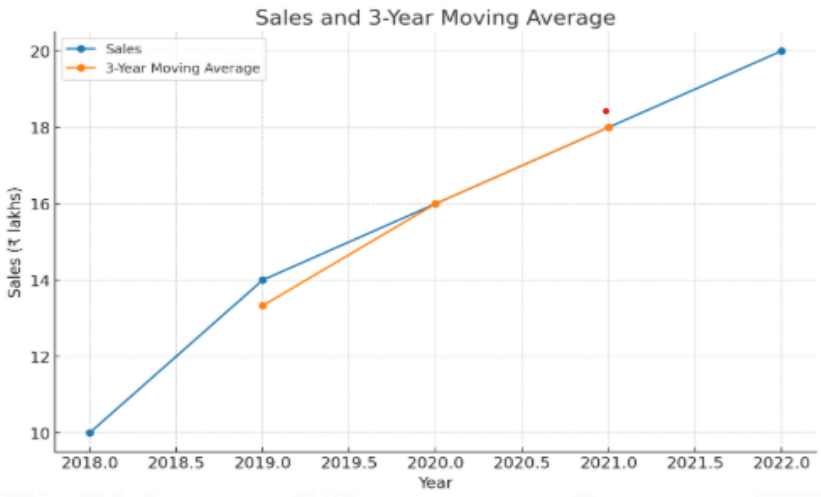
MA1=(10+14+16)/3=40/3=13.33

MA2= (14+16+18)/3=16

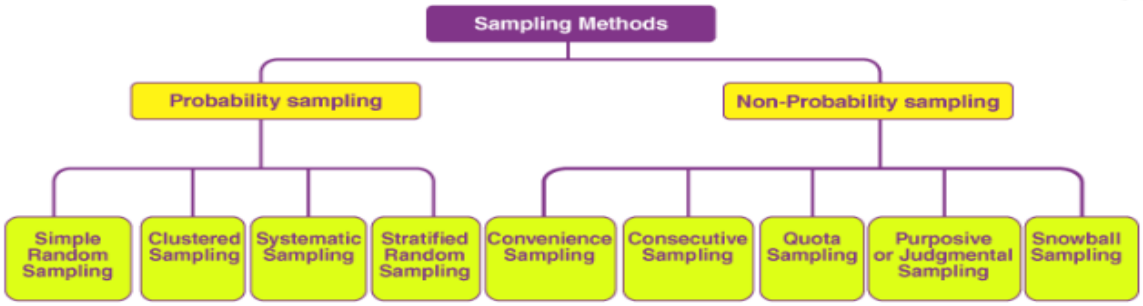
MA3= (16+18+20)/3=18

Trend (Smoothed Values)

Centre Year	3-Year Moving Average
2019	13.33
2020	16.00
2021	18.00



The method of least squares fits a mathematical trend line (usually straight line) that best represents the long-term movement and then subtracts this trend from the original data. **Best fit Line**



1. Probability Sampling

Every unit has a known, non-zero chance of being selected. Reduces bias.

(a) Simple Random Sampling (SRS)

Definition: Each unit has equal chance of selection (lottery/ random number table).

Strengths: Easy, unbiased, statistically sound.

Limitations: Needs complete population list, not practical for large populations.

Use: Election polls, small surveys.

(b) Systematic Sampling

Definition: Select every k-th item from ordered population after a random start.

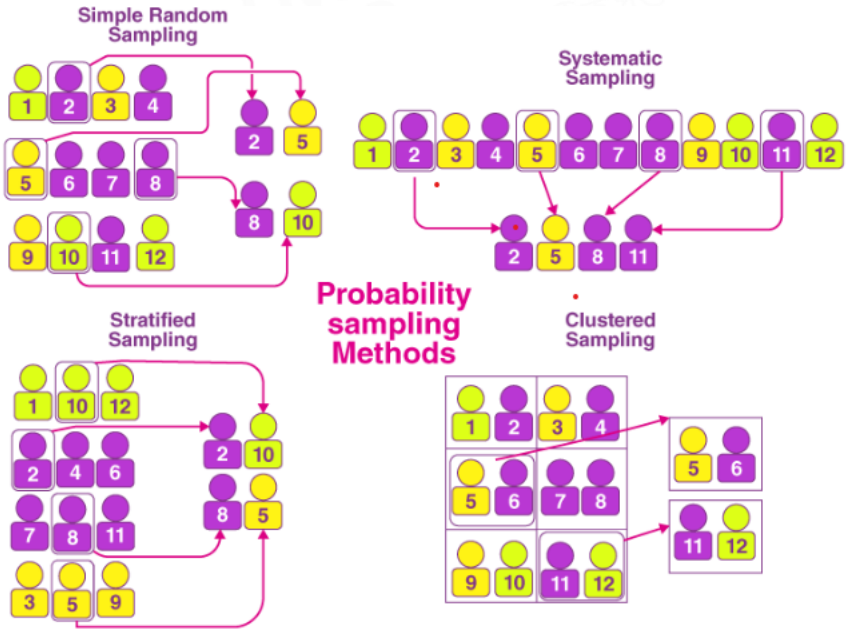
Strengths: Simpler than SRS, evenly spread.

Limitations: Risk of bias if there is periodic pattern in data.

Use: Quality control in manufacturing.

(c) Stratified Sampling

Definition: Divide population into subgroups (strata) and randomly sample from each.
Strengths: Ensures representation of subgroups, more precise estimates.
Limitations: Requires detailed population info, complex.
Use: Population divided by age, income, region.
(d) Cluster Sampling
Definition: Divide population into clusters (groups), randomly select some clusters, and study all/ some units within them.
Strengths: Cost-effective, useful for large/geographically spread populations.
Limitations: Less precise, higher sampling error compared to stratified.
Use: Household surveys in different districts.
(e) Multistage Sampling
Definition: Combination of methods, often cluster + random sampling at multiple stages.
Strengths: Flexible, practical for large populations.
Limitations: Complex design, chance of compounded error.
Use: National-level surveys (e.g., census).
2. Non-Probability Sampling
Selection is subjective; not every unit has known chance of inclusion.
(a) Convenience Sampling
Definition: Choose respondents easiest to access.
Strengths: Fast, cheap, easy.
Limitations: Highly biased, not representative.
Use: Pilot studies, quick feedback.
(b) Judgment / Purposive Sampling
Definition: Researcher selects units based on expertise.
Strengths: Useful for special/rare populations.
Limitations: Subjective, researcher bias.
Use: Expert interviews, case studies.
(c) Quota Sampling
Definition: Population divided into groups, fixed quota taken (non-randomly).
Strengths: Ensures subgroup representation.
Limitations: Not random, may bias within groups.
Use: Market research surveys.
(d) Snowball Sampling
Definition: Existing participants recruit further participants.
Strengths: Good for hard-to-reach groups.
Limitations: Non-random, biased, network dependent.
Use: Drug users, refugees, rare disease patients.



Conceptual Analysis

Population vs Sample

Population:

The entire group of subjects or items you are interested in studying.

Example: All households in Dhaka city, all students in Bangladesh, all mobile phones produced in 2024.

Denoted by: Usually N.

Sample:

A subset of the population chosen for actual study.

Example: 500 households from Dhaka, 200 students from a university, 50 mobile phones tested.

Denoted by: Usually n.

Key Points:

Population data is often impractical or costly to collect.

Sample is used to estimate population characteristics.

Visual analogy: If the population is a big cake, the sample is a slice. You taste the slice to judge the whole cake.

Census Survey vs Sample Survey

Census Survey:

Data collected from the entire population.

Example: Population and Housing Census in Bangladesh.

Advantages:

Complete coverage, no sampling error.

Limitations: Very costly, time-consuming, data may quickly become outdated.
Sample Survey:
Data collected from a subset (sample) of the population.
Example: Household Income and Expenditure Survey (HIES) in Bangladesh.
Advantages: Less expensive, faster, easier to manage.
Limitations: Subject to sampling error; needs careful sample design.
Parameter vs Statistic
Parameter:
Numerical characteristic of a population.
Example: Population mean (μ), population proportion (P), population variance (σ^2).
Usually fixed, unknown unless a census is conducted.
Statistic:
Numerical characteristic of a sample.
Example: Sample mean ($\bar{x}$), sample proportion (p), sample variance (s^2).
Used to estimate population parameters.
Key Difference:
Parameter → describes population
Statistic → describes sample
Statistic is a random variable (changes with each sample), parameter is fixed.
Sampling Error vs Non-Sampling Error
Sampling Error:

Error caused by studying a sample instead of the whole population.

Example: If a sample of students shows average height slightly different from true population height.

Can be reduced by increasing sample size or using better sampling design.

Non-Sampling Error:

Error not related to sampling, occurs even in a census.

Examples: Data entry mistakes, response bias, faulty measuring instruments.

Cannot always be reduced by larger sample size.
sampling error এর প্রাথমিক কারণ হচ্ছে - আমরা যে পপুলেশন না নিয়ে স্যাম্পল নিয়ে কাজ করছি- এটাই। স্যাম্পলিং ম্যাথড যতই ভালো আর ত্রুটিমুক্ত হোক, এরর আসবেই।
Sampling error arises when a sample, rather than the entire population, is studied.

It reflects the natural variability between sample estimates and true population values, even if sampling is random and correctly conducted.

Calculating Sample Size

$$n = \left(\frac{Z \cdot \sigma}{E}\right)^2$$

Where:

n = required sample size

Z = Z-value for chosen confidence level

σ = population standard deviation (or estimated from pilot sample)

E = maximum allowable margin of error (precision you want)

A researcher wants to estimate the **average monthly expenditure** of college students.

- Wants **95% confidence** → $Z = 1.96$
- From a pilot study, **standard deviation** $\sigma = \$50$
- Wants estimate within **±$5** → $E = 5$

$$n = \left(\frac{1.96 \times 50}{5}\right)^2 = \left(\frac{98}{5}\right)^2 = (19.6)^2 = 384.16$$

Probability

Approaches to Defining Probability

There are three classical ways to define and interpret probability:

(a) Classical/Laplacean/Priori Approach

Assumes all outcomes are equally likely.

P(A)= n(A)/n(S); n(A)= favourable outcome, n(S)= total number of outcome

(b)Relative Frequency or Empirical or posterior Approach

Probability of happening an event= no. of times events in past/Total no. of observations

(c)Axiomatic Approach

Based on set theory. Combines both classical+Empirical. There are 3 axoms:

a) $0 \le P(A) \le 1$

b) $P(S)=1$

c) $P(A_1 U A_2 U A_3 .... U A_n) = P(A_1)+P(A_2)+P(A_3)+.....P(A_n)$

Basic Properties of Probability

1. $0 \leq P(A) \leq 1$
2. $P(S)=1$
3. $P(A^c) = 1-P(A)$, where A^c = complement.
4. For mutually exclusive events A and B: $\rightarrow P(A \cup B)=P(A)+P(B)$

Important Rules of Probability

(a) Multiplication Rule

Independent events: $P(A \cap B) = P(A) \times P(B)$

Example: Rolling a 2 on a die and tossing heads = $1/6 \times 1/2 = 1/12$

Dependent events: $P(A \cap B) = P(A|B) \times P(B)$

(b) Addition Rule

For any two non mutually exclusive events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

For any two mutually exclusive events: $P(A \cup B) = P(A) + P(B)$

(c) Complementation Rule

$P(A^c) = 1 - P(A)$

Example: Probability of not rolling a 6 on a die = $1 - 1/6 = 5/6$

Bivariate Data

Bivariate data involves pairs of linked observations, often to explore relationships between two variables.

Example: A researcher collects students' study hours and corresponding exam scores. Each pair (hours, score) is a bivariate observation.

Contingency Table

A grid used to display the frequency distribution of two categorical variables jointly.

Suppose a company randomly selects an employee.

- Let event **M** = selected employee is *Male*
- Event **T** = selected employee is *Trained*
- From records:

	Trained (T)	Not Trained (T')	Total
Male (M)	30	10	40
Female (F)	20	40	60
Total	50	50	**100**

Here we can analyze joint, marginal, and conditional probabilities

Joint. Marginal and Conditional Probability

Joint Probability= $P(A \cap B)$

The probability that two events happen simultaneously.

(a) Joint probability

$P(M \cap T)$ = probability employee is male and trained

= 30 / 100 = 0.30

Marginal Probability:

The probability of a single event occurring, regardless of the outcomes of any other events. These are called marginal probabilities because they are taken from the totals (margins) of the table — not depending on any other variable.

(b) Marginal probability in the previous example-

P(T) = total trained / total employees = 50 / 100 = 0.50

P(M)= ?

Conditional Probability

Probability of event A occurring given that B has already occurred.

P(A|B)=P(A∩B)/P(B)

P(B|A)=P(A∩B)/P(BA)

We find multiplication rule here.

Using the above table:

✅ **Probability that an employee is *trained* given that the employee is *male*.**

$$P(T|M) = \frac{P(M \cap T)}{P(M)} = \frac{0.30}{0.40} = 0.75$$

So there's a **75% chance** a male employee is trained.

Bayes' Theorem: Used to update probabilities based on new evidence.

Random Variables

Random Variable (RV)

A random variable is a rule that assigns a number to each outcome of a random experiment.

☞ In other words, it converts chance outcomes into numerical values so we can calculate things like probability, mean, variance, etc.

Types

1. Discrete RV – takes countable values (like 0,1,2,...).
2. Continuous RV – takes any value in an interval (like height, time, weight).

Expected Value (Mean)

The expected value of a random variable is like its long-run average.

It tells us what we can expect on average if we repeat an experiment many times.

- For a **discrete random variable** X with possible values $x_1, x_2, \ldots, x_n$ and probabilities $P(X = x_i) = p_i$:

$$E[X] = \sum_{i=1}^{n} x_i \, p_i$$

- For a **continuous random variable** with probability density function $f(x)$:

$$E[X] = \int_{-\infty}^{\infty} x f(x) \, dx$$

For a fair 6-sided die, $P(X = i) = \frac{1}{6}$ for $i = 1, \ldots, 6$.

$$E[X] = \sum_{i=1}^{6} i \cdot \frac{1}{6} = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6)$$
$$= \frac{1}{6} \cdot 21 = \frac{21}{6} = 3.5$$

So, the expected value is **3.5**.

1. Properties of Expectation (E[X])

1. **Linearity**

$$E[aX + bY] = aE[X] + bE[Y]$$

(for any constants a, b; true whether X, Y are independent or not)

2. **Constant**

$$E[c] = c$$

(for any constant c)

3. **Multiplying by a constant**

$$E[aX] = aE[X]$$

4. **Additivity**

$$E[X + Y] = E[X] + E[Y]$$

(holds always, not only for independence)

5. **Product of independent random variables**
   If X, Y are independent:

$$E[XY] = E[X] \cdot E[Y]$$

Variance

The variance measures how spread out the values of a random variable are from its expected value.

$$Var(X) = E[(X - E[X])^2]$$

This expands to:

$$Var(X) = E[X^2] - (E[X])^2$$

- **Standard deviation** is the square root of variance:

$$\sigma = \sqrt{Var(X)}$$

2. Properties of Variance (Var(X))

1. **Definition**

$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

2. **Non-negativity**

$$Var(X) \geq 0$$

3. **Variance of a constant**

$$Var(c) = 0$$

4. **Scaling property**

$$Var(aX) = a^2 Var(X)$$

5. **Shifting property**

$$Var(X + c) = Var(X)$$

(adding a constant does not change variance)

6. **Variance of sum of two random variables**

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

7. **If independent**

$$Var(X + Y) = Var(X) + Var(Y)$$

(because $Cov(X, Y) = 0$ when independent)

Math 1:

Let X = outcome of a fair die.

- $E[X] = 3.5$.

Now define $Y = 2X + 3$.

$$E[Y] = E[2X + 3] = 2E[X] + 3 = 2(3.5) + 3 = 10$$

Math 2:

We already know for a die:

$$Var(X) = \frac{35}{12} \approx 2.92$$

Now for $Y = 2X + 3$:

$$Var(Y) = Var(2X + 3) = 2^2 Var(X) = 4 \times \frac{35}{12} = \frac{140}{12} = \frac{35}{3} \approx 11.67$$

Math 3:

Flip 2 fair coins. Let

- X = number of heads on coin 1 (so X is 0 or 1, with $E[X] = 0.5, Var(X) = 0.25$)
- Y = number of heads on coin 2 (same distribution).

Define total heads: $Z = X + Y$.

$$Var(Z) = Var(X + Y) = Var(X) + Var(Y) \quad \text{(independent)}$$

$$= 0.25 + 0.25 = 0.5$$

Probability Mass Function

- Used for **discrete random variables.**
- It gives the probability of **each possible value.**

👉 Properties:

1. $P(X = x) \geq 0$
2. $\sum P(X = x) = 1$

👉 Formula:

$$P(X = x) = f(x)$$

👉 Example: Tossing a fair die.

$$P(X = x) = \frac{1}{6}, \quad x = 1, 2, 3, 4, 5, 6$$

Probability Density Function

- Used for **continuous random variables.**
- Probability at a single point = **0.**
- Instead, probability is calculated over an **interval.**

👉 Properties:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

👉 Formula for probability:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

CDF (Cumulative Distribution Function)

The CDF of a random variable X is defined as:

$$F(x) = P(X \leq x)$$

-Works for both discrete & continuous random variables.

-Non-decreasing function

-Step function

As x increases, F(x) never decreases.

Because probabilities accumulate.

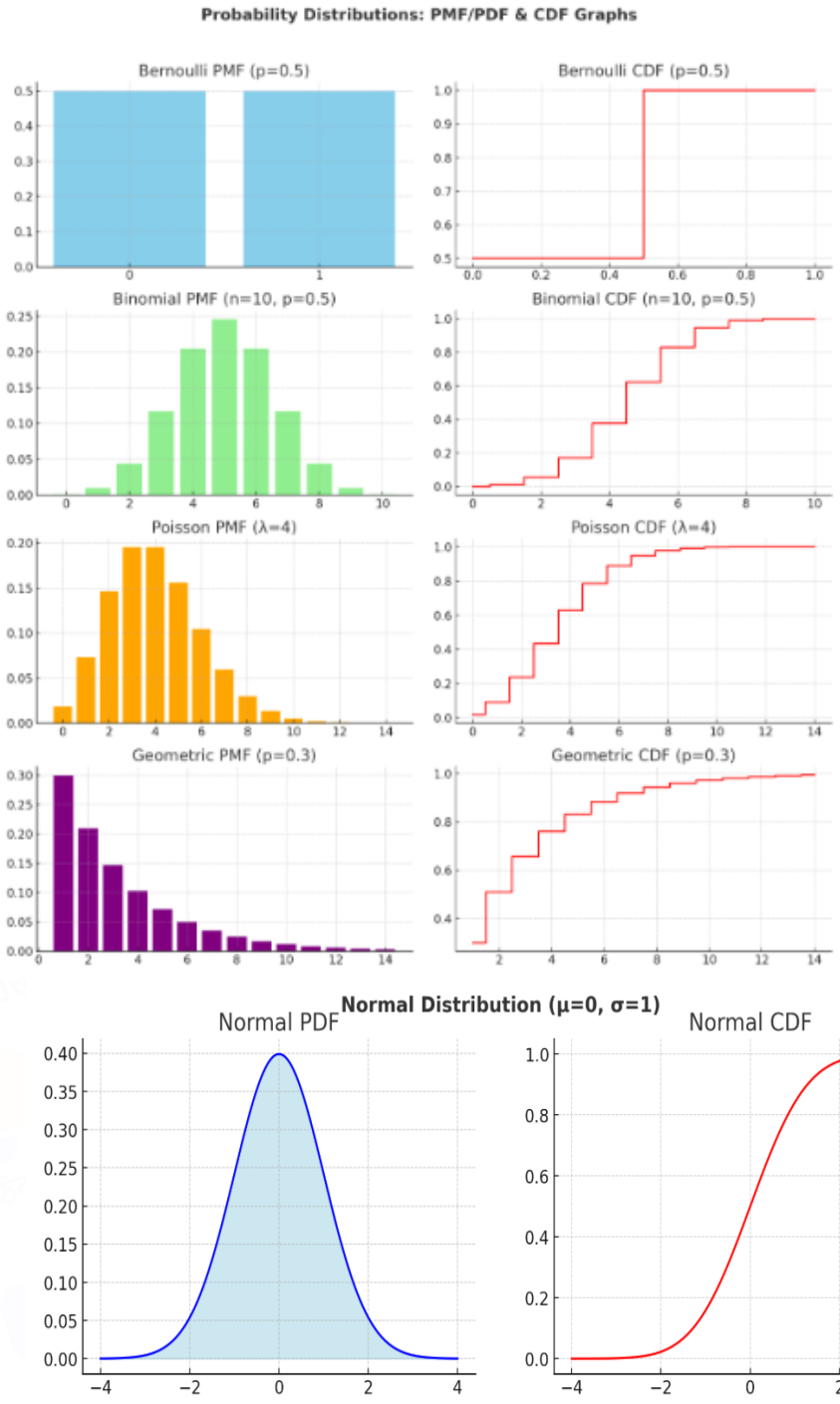
It gives the probability that random variable X takes a value less than or equal to x.

Throwing a die-

$$F(3) = P(X \leq 3) = P(1) + P(2) + P(3) = \frac{3}{6} = 0.5$$

If $a < b$, then

$$F(a) \leq F(b)$$



Here are the **PDF** (bell-shaped curve) and **CDF** (S-shaped curve) for the **Normal Distribution** with mean $\mu = 0$ and standard deviation $\sigma = 1$.

📊 Specific Graph Interpretations

1. **Bernoulli (p=0.5)**
   - PMF: Two bars → 50% at 0, 50% at 1.
   - CDF: Jumps from 0.5 at x=0 to 1 at x=1.
2. **Binomial (n=10, p=0.5)**
   - PMF: Peaks around 5 (most likely ~half heads). Symmetric.
   - CDF: Smooth staircase climbing up to 1.
3. **Poisson (λ=4)**
   - PMF: Skewed bar shape, peaks near 4.
   - CDF: Step curve rising toward 1.
4. **Geometric (p=0.3)**
   - PMF: Bars decrease as trial number grows.
   - CDF: Rises quickly because success is likely within first few trials.

Probability Distribution

A probability distribution tells us how the probabilities are spread out over possible outcomes of a random variable.

Two main types:

Discrete distribution → for outcomes that are countable (like dice, coins).

Continuous distribution → for outcomes over a range (like height, weight, time).

1. Bernoulli Trial

Concept

A single experiment with only two possible outcomes: Success (1) or Failure (0).

Probability of success = p, probability of failure = q = 1 − p.

When to Use

When analyzing binary events: coin toss (H/T), pass/fail, defective/not defective, etc.

Properties

Mean = p

Variance = p(1 − p)

Support: {0,1}

2. Binomial Distribution

Extension of Bernoulli trial.

Number of successes in n independent Bernoulli trials with probability of success = p.

Probability Mass Function (PMF):

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

When to Use

Fixed number of trials, independent, identical probability.

Examples: number of heads in 10 coin tosses, number of defective bulbs in a sample of 20.

Properties

Mean = np

Variance = np(1 – p)

3. Poisson Distribution

Concept

Counts the number of events in a fixed interval of time/space, if events occur independently at a constant rate.

PMF

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

When to Use

Rare events over time/space.

Examples: number of phone calls per hour, accidents per week, typing errors per page.

Properties

Mean = Variance = λ

Approximates Binomial when n large, p small.

4. Geometric Distribution

Concept

Number of trials until the first success in independent Bernoulli trials.

PMF

$$P(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \ldots$$

When to Use

- Waiting time for first success.
- Example: How many coin tosses until the first head?

Properties

- Mean = **1/p**
- Variance = **(1-p)/p²**
- Memoryless property: P(X > m + n | X > m) = P(X > n).

5. Normal Dstribution

Concept

- Continuous distribution, symmetric, bell-shaped curve.
- Defined by mean (μ) and standard deviation (σ).

Probability Density Function (PDF)

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

When to Use

- Data naturally follows bell-shape (heights, IQ, measurement errors).
- Central Limit Theorem: sample means tend to be normal as n grows large.

Properties

- Mean = μ, Variance = σ^2
- Symmetric about μ
- 68-95-99.7% rule.

6. Hypergeometric Dstribution

- Probability of **k successes in n draws without replacement** from a finite population containing successes and failures.

PMF

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where

- N = population size,
- K = number of successes in population,
- n = sample size,
- k = successes in sample.

When to Use

- Sampling **without replacement**.
- Example: Picking red balls from a box without putting them back.

7. Uniform Distribution

Concept

- All outcomes equally likely within interval [a, b].

PDF

$$f(x) = \frac{1}{b-a}, \quad a \le x \le b$$

When to Use

- Random number generation, equal-likelihood situations.

Properties

- Mean = (a + b)/2
- Variance = $(b - a)^2 / 12$

8. Exponential Distribution

Concept

- Models **time between successive events** in a Poisson process.

PDF

$$f(x) = \lambda e^{-\lambda x}, \quad x \ge 0$$

When to Use

- Lifetimes of devices, time between arrivals in a queue, waiting times.

Properties

- Mean = $1/\lambda$
- Variance = $1/\lambda^2$
- Memoryless property: $P(X > s + t \mid X > s) = P(X > t)$.

Comparison & Relations

Bernoulli → Binomial → Poisson: Bernoulli is 1 trial, Binomial is multiple trials, Poisson is limit of Binomial.

Geometric vs Exponential: Both measure "waiting time" – geometric (discrete), exponential (continuous).

Binomial vs Hypergeometric: With replacement vs without replacement.

Normal vs Uniform: Normal bell-shaped, Uniform flat.

Exponential vs Poisson: Poisson counts number of events; exponential models time between events.

Sampling Distribution

1. **Sampling distribution of the mean** is the distribution of means from all possible samples.
2. Its **mean** equals the population mean (μ).
3. Its **variance** equals the population variance divided by sample size (σ^2/n).
4. Standard error decreases as sample size increases → sample mean becomes **more reliable**.
5. **Central Limit Theorem** ensures the sampling distribution is approximately **normal** if n is large.

Scenario:
A factory produces bolts with average length $\mu = 10$ cm and standard deviation $\sigma = 2$ cm.

Step 1: Sample mean distribution for n = 16

- Mean:

$$E(\bar{X}) = \mu = 10 \text{ cm}$$

- Variance:

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{2^2}{16} = \frac{4}{16} = 0.25$$

- Standard error:

$$SE = \sqrt{0.25} = 0.5 \text{ cm}$$

Interpretation:
If we repeatedly take samples of 16 bolts, the sample mean length will typically vary ±0.5 cm around 10 cm.

Key Takeaways

The **sampling distribution of a proportion** is the distribution of sample proportions from repeated sampling.

Its **mean** equals the population proportion (p).

Its **variance** is $\frac{p(1-p)}{n}$ and decreases as sample size increases.

Standard error decreases with larger n, making sample estimates more reliable.

If n is large and p is not too close to 0 or 1, the distribution is approximately **normal**, allowing for confidence intervals and hypothesis tests.

Scenario:

A city survey shows that **40% of households own a car** ($p = 0.4$).

Step 1: Sample proportion distribution for n = 100

- Mean:

$$E(\hat{p}) = p = 0.4$$

- Variance:

$$Var(\hat{p}) = \frac{0.4 \cdot 0.6}{100} = \frac{0.24}{100} = 0.0024$$

- Standard error:

$$SE_{\hat{p}} = \sqrt{0.0024} \approx 0.049$$

Interpretation:

If we repeatedly take samples of 100 house[illegible]s, the **sample proportion** will typically vary ±4.9% around 40%.

Central Limit Theorem (CLT)

The Central Limit Theorem states that, regardless of the population distribution, the sampling distribution of the sample mean approaches a normal distribution as the sample size becomes large (n≥30 is commonly used).

Real-life example:

A hospital records patient waiting times. Even if waiting times are skewed (most patients wait 5–10 minutes, some wait 60+), the average waiting time from repeated samples of 40 patients will form a normal distribution.

Implication: CLT allows us to apply normal probability methods for inference, even if the population is not normally distributed.

A **confidence interval** provides a range of plausible values for a population parameter (mean or proportion) based on sample data.

CI for Population Mean (known σ):

$$\bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- $\bar{X}$ = sample mean
- $Z_{\alpha/2}$ = critical value from the standard normal distribution (e.g., 1.96 for 95% CI)
- $\sigma/\sqrt{n}$ = standard error

Example:

Using the apple weight example, if the sample mean of 25 apples is 152 g and $\sigma = 30$ g:

$$CI = 152 \pm 1.96 \cdot \frac{30}{\sqrt{25}} = 152 \pm 11.76 \implies (140.24, 163.76)$$

We are 95% confident that the true mean weight of apples in the orchard is between 140.24 g and 163.76 g.

CI for Population Proportion:

$$\hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example:

From the car ownership example, sample proportion $\hat{p} = 0.42, n = 100$:

$$CI = 0.42 \pm 1.96 \cdot \sqrt{\frac{0.42(1-0.42)}{100}} = 0.42 \pm 0.098 \implies (0.322, 0.518)$$

So the proportion of households with a car is likely between 32.2% and 51.8%.

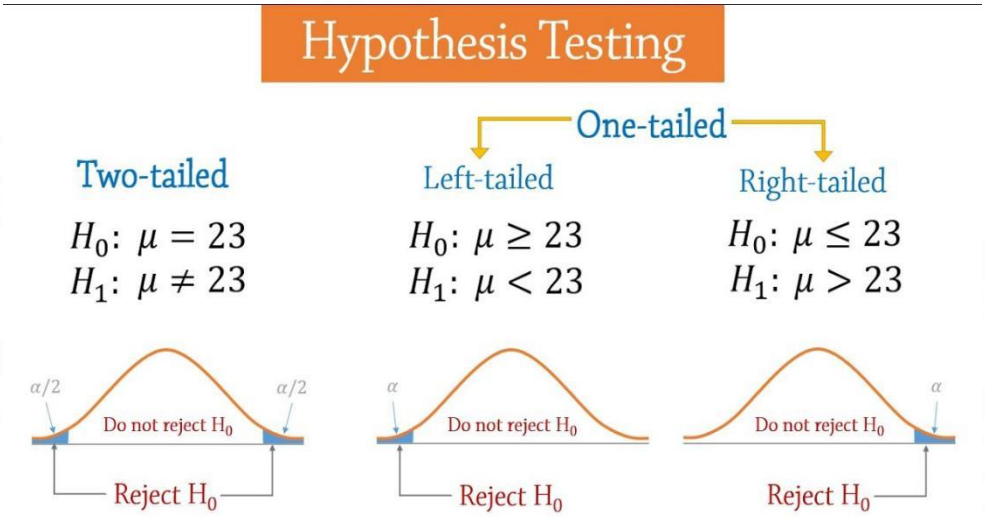
Hypothesis Testing

what hypothesis testing does:

☞ It gives us a rule to decide whether to accept or reject a claim using sample data.

Null Hypothesis (H_0): The claim of "no effect" or "no difference". It assumes the current belief is true.

Alternative Hypothesis (H_1 or H_a): The rival claim that we want to test for (usually indicating a difference or effect).



নমুনা থেকে নির্ণীত মান ও সমগ্রক থেকে নির্ণীত মানের মধ্যে কিছু পার্থক্য থাকবে। এই পার্থক্যের মাত্রা কতটুকু অর্থাৎ তা গ্রহণযোগ্য সীমার মধ্যে অবস্থান করছে কিনা তা পরিমাপ করার জন্য পরিসংখ্যান সংক্রান্ত যে পদ্ধতি অনুসরণ করা হয় তাকেই বলে টেস্ট অব হাইপোথিসিস। (প্রকল্প বিচার)

এক্ষেত্রে নিম্নোক্ত পরীক্ষাগুলো করা হয়:

১. পরিমিত পরীক্ষা (normal test or Z test)

২. t পরীক্ষা (ক্ষুদ্র নমুনার ক্ষেত্রে)

৩. কাই বর্গ পরীক্ষা

8) F test

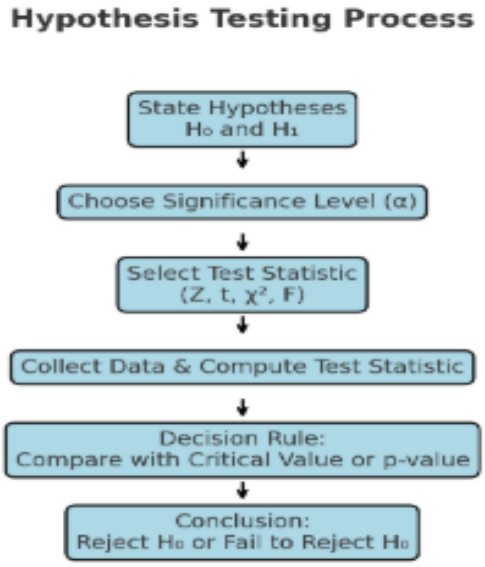
মূখ্য প্রকল্প (null hypothesis)

বিকল্প প্রকল্প (alternative hypothesis)

বাতিল এলাকা (rejection region)

গ্রহণ এলাকা (accepted region)

এক প্রান্তিক দ্বি প্রান্তিক পরীক্ষা (one sided test, two sided test)



-Choose the appropriate statistical test

-Z-test (large samples, known σ)

-t-test (small samples, unknown σ)
-Chi-square test (categorical data)
-F-test (variances comparison)
Now, Calculate the test statistic from sample data.Find the critical value / p-value. Compare the calculated test statistic with the critical value from tables (Z, t, χ^2).

4. Quick Comparison Table

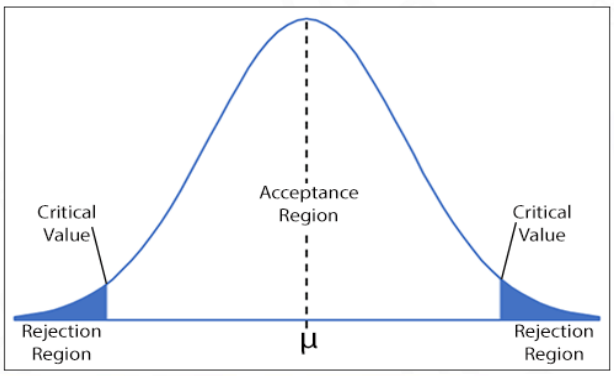
Test	When to Use	Data Type	Example
Z-test	Large n, σ known	Mean/Proportion	Avg. weight ≠ 60 kg
One-sample t	Small n, σ unknown	Mean	Sample mean vs population mean
Pooled t	Two groups, equal variance	Mean	Boys vs Girls test scores
Paired t	Same group before/after	Mean	Before & after training scores
F-test	Compare 2 variances	Variance	Variability of yields between two farms
χ^2 test	Categorical data	Frequencies	Independence of gender & choice of product

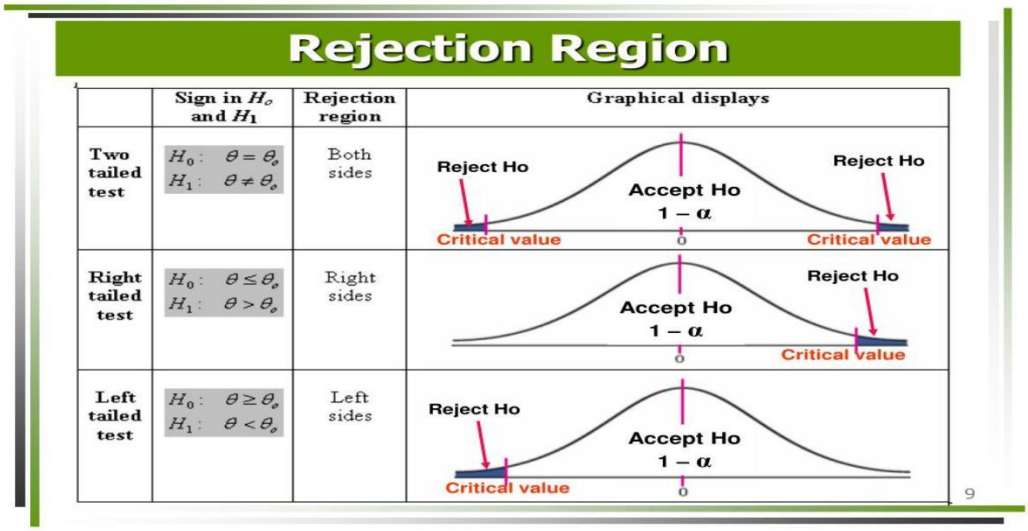
Errors in Hypothesis Testing
Type I Error (α error): Rejecting H_0 when it is true.
Example: Concluding the cement bags are not 50 kg when they actually are.
Type II Error (β error): Failing to reject H_0 when it is false.
Example: Concluding the cement bags are 50 kg when in fact they are lighter.

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

Scribbr





p-value

The smallest level of significance at which H_0 can be rejected

Start with a claim (H_0).

Compare p-value with significance level (α):

If $p \leq \alpha \Rightarrow$ Reject H_0

If $p > \alpha \Rightarrow$ Fail to reject H_0.

ANOVA

Definition: ANOVA is a statistical method used to compare means of more than two groups to see if at least one differs significantly.

Idea: Total variation in the data = Variation between groups + Variation within groups.

Formula for Total Sum of Squares (SST):

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

Where:

- X_{ij} = value of j-th observation in i-th group
- $\bar{X}$ = grand mean

$$SST = SSB + SSW$$

- **SSB (Sum of Squares Between Groups):** variation due to difference in group means.
- **SSW (Sum of Squares Within Groups):** variation due to random error inside groups.

$$SST = SSB + SSW$$

ANOVA checks if **variation between groups** is much larger than **variation within groups.**

$$F = \frac{MSB}{MSW}$$

If F is large → groups differ significantly.

Why not multiple t-tests?

If you compare 4 groups using t-tests → 6 pairwise comparisons. Each has 5% error chance. Overall error risk (Type I error) increases.

ANOVA controls this error by using one test.

ANOVA Table (One-Way Example)

Source of Variation	SS	df	MS	F = MSB/MSW
Between Groups	SSB	k-1	MSB	
Within Groups	SSW	N-k	MSW	
Total	SST	N-1		

Two-Way ANOVA Table Format

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-ratio
Factor A (rows)	SSA	a – 1	MSA = SSA / (a–1)	F = MSA / MSE
Factor B (columns)	SSB	b – 1	MSB = SSB / (b–1)	F = MSB / MSE
Interaction (A×B)	SSAB	(a–1)(b–1)	MSAB = SSAB / (a–1)(b–1)	F = MSAB / MSE
Error (Within)	SSE	ab(n–1)	MSE = SSE / ab(n–1)	–
Total	SST	N – 1	–	–

Design of Experiments

Design of Experiments is the planned way of collecting data so that we can:

-Study the effect of one or more factors (independent variables) on a response (dependent variable).

-Minimize error and maximize reliability of results.

-Use resources (time, money, subjects) efficiently.

In simple words: “Design of experiment is a blueprint that tells us how to collect data, in what order, how many samples, and under what conditions.”

🎯 1. Completely Randomized Design (CRD)

👉 Treatments are assigned **completely at random**.

Example: 3 fertilizers (A, B, C) applied randomly on 9 plots.

Plot	1	2	3	4	5	6	7
Fertilizer	A	B	C	A	C	B	A

✅ Very simple: just random assignment.

🎯 2. Randomized Block Design (RBD)

👉 Units are grouped into **blocks**, then treatments randomized inside each block.

Example: Soil type is blocking factor (2 types: Clay & Sandy). Each block gets all 3 fertilizers.

Block: Clay Soil	**Plot 1**	**Plot 2**	**Plot 3**
Fertilizer	A	B	C

Block: Sandy Soil	**Plot 4**	**Plot 5**	**Plot 6**
Fertilizer	B	C	A

✅ Each block gets all treatments → reduces error from soil differences.

🎯 3. Latin Square Design (LSD)

👉 Used when **two nuisance factors** need control (rows = soil, columns = irrigation).
Treatments arranged so each appears once per row & once per column.

Example: 3 fertilizers (A, B, C), Soil type = rows, Irrigation level = columns.

Soil \ Irrigation	**Low**	**Medium**	**High**
Clay	A	B	C
Sandy	B	C	A
Loamy	C	A	B

✅ Each fertilizer appears once per row and once per column.

Main Types of Experimental Design

Design	Explanation	Real-Life Example	Advantages	Limitations
Completely Randomized Design (CRD)	Treatments are assigned to experimental units *completely at random*.	Testing 3 teaching methods by assigning students randomly to each method.	Simple, easy to analyze.	May be inefficient if units are heterogeneous (e.g., students with very different backgrounds).
Randomized Block Design (RBD)	Units are divided into *blocks* (similar groups) to reduce variability, then treatments are randomly assigned within each block.	Fertilizers tested on different soil types → soil type = block, fertilizer = treatment.	Controls for variability due to blocks, more precise.	More complex than CRD.
Latin Square Design (LSD)	Treatments are arranged so that each appears once in every row and column.	Comparing fertilizers while controlling both **soil type** (rows) and **irrigation method** (columns).	Controls for *two sources of variation* at once.	Needs equal numbers of rows/columns; difficult to arrange.

পরিসংখ্যান একটি বিস্তৃত বিজ্ঞান, যা ডেটা সংগ্রহ, বিশ্লেষণ, ব্যাখ্যা এবং উপস্থাপনের মাধ্যমে আমাদের জীবনের বিভিন্ন ক্ষেত্রে সিদ্ধান্ত নিতে সাহায্য করে।সাধারণ গড়, তরঙ্গ গড় এবং জ্যামিতিক গড়—এগুলো কেন্দ্রীয় প্রবণতার পরিমাপ, কিন্তু প্রত্যেকটির অনন্য বৈশিষ্ট্য রয়েছে যা ডেটার ধরন অনুসারে ব্যবহার হয়। সাধারণ গড়, যা সকল মানের যোগফলকে মানের সংখ্যা দিয়ে ভাগ করে পাওয়া যায়, সবচেয়ে সাধারণ এবং সহজবোধ্য, কিন্তু এটি চরম মান বা আউটলায়ার্সের প্রভাবে সহজেই বিকৃত হয়ে যায়। উদাহরণস্বরূপ, একটি কোম্পানিতে কর্মচারীদের গড় বেতন হিসাব করতে গেলে যদি সিইও-এর বিপুল বেতন থাকে, তাহলে সাধারণ গড় উচ্চ দেখাবে, যা সাধারণ কর্মচারীদের বাস্তবতা প্রতিফলিত করবে না—এজন্য এটি সাধারণত সমান বিতরণের ডেটায় ব্যবহার হয়, যেমন একটি ক্লাসের ছাত্রদের গড় নম্বর। অন্যদিকে, তরঙ্গ গড়, যা মানের সংখ্যাকে প্রত্যেক মানের ব্যস্তকের যোগফল দিয়ে ভাগ করে গণনা করা হয়, গতি বা হার-সম্পর্কিত ডেটায় উপযোগী কারণ এটি ছোট মানগুলোকে বেশি ওজন দেয়; বাস্তব জীবনে, যদি আপনি ঢাকা থেকে চট্টগ্রাম যাওয়ার সময় বিভিন্ন গতিতে গাড়ি চালান, তাহলে গড় গতি হিসাব করতে তরঙ্গ গড় ব্যবহার হয় যাতে দূরত্ব এবং সময়ের অনুপাত সঠিকভাবে ধরা পড়ে, না হলে সাধারণ গড় অতিরিক্ত উচ্চ দেখাতে পারে। জ্যামিতিক গড়, যা সকল মানের গুণফলের n-তম মূল নিয়ে গণনা করা হয় (n হলো মানের সংখ্যা), বৃদ্ধির হার বা অনুপাতিক ডেটায় সবচেয়ে কার্যকর কারণ এটি গুণনীয় সম্পর্ক ধরে রাখে এবং নেগেটিভ মান নিয়ে কাজ করে না; উদাহরণ হিসেবে, বিনিয়োগের বার্ষিক রিটার্ন হার হিসাব করতে এটি ব্যবহার হয়—যেমন যদি একটি স্টকের রিটার্ন প্রথম বছর ১০%, দ্বিতীয় বছর ২০%, তাহলে জ্যামিতিক গড় (প্রায় ১৪.৯%) সঠিকভাবে কম্পাউন্ডিং ইফেক্ট দেখায়, যা সাধারণ গড় (১৫%) দেখাতে পারে না। গভীরভাবে চিন্তা করলে দেখা যায়, এই গড়গুলোর মধ্যে সম্পর্ক রয়েছে—সাধারণত arithmetic mean ≥ geometric mean ≥ harmonic mean (পজিটিভ ডেটায়), এবং ভুল গড় ব্যবহার করলে সিদ্ধান্ত ভুল হয়ে যেতে পারে, যেমন অর্থনৈতিক নীতিতে।

গড়, মধ্যক এবং প্রচুরক—এগুলো কেন্দ্রীয় প্রবণতার তিনটি প্রধান পরিমাপ, যা ডেটার "সাধারণ" মান খুঁজে বের করে কিন্তু ভিন্ন দৃষ্টিকোণ থেকে। গড়, যা ইতিমধ্যে উল্লেখ করা হয়েছে, সকল ডেটার গণিতিক কেন্দ্রবিন্দু এবং পরিসংখ্যানীয় বিশ্লেষণে সবচেয়ে শক্তিশালী কারণ এটি আরো উন্নত গণনায় (যেমন ভ্যারিয়েন্স) ব্যবহার হয়, কিন্তু

আউটলায়ার্সের কারণে এটি স্কিউড ডেটায় অসঠিক হতে পারে; বাস্তবে, একটি দেশের গড় আয় হিসাব করতে গেলে যদি কয়েকজন বিলিয়নিয়ার থাকে, তাহলে গড় উচ্চ দেখাবে যা সাধারণ মানুষের জীবনযাত্রা প্রতিফলিত করবে না। মধ্যক, যা ডেটা সাজিয়ে মাঝের মান (বা জোড় সংখ্যায় দুটির গড়), আউটলায়ার্সের প্রভাব থেকে মুক্ত এবং স্কিউড ডেটায় আরো নির্ভরযোগ্য; উদাহরণ হিসেবে, রিয়েল এস্টেট মার্কেটে বাড়ির গড় দামের পরিবর্তে মধ্যক দাম ব্যবহার হয় কারণ কয়েকটা অত্যধিক দামি বাড়ি গড়কে বাড়িয়ে দিতে পারে, যেমন ঢাকার একটি এলাকায় যদি বেশিরভাগ বাড়ির দাম ৫০ লাখ টাকা হয় কিন্তু একটা ৫ কোটি টাকার, তাহলে মধ্যক সঠিকতর ছবি দেবে। প্রচুরক, যা সবচেয়ে ঘন ঘন ঘটিত মান (একাধিকও হতে পারে), ক্যাটাগরিকাল বা নামিনাল ডেটায় উপযোগী এবং কোনো গণনা ছাড়াই পাওয়া যায়; বাস্তব জীবনে, একটি সুপারমার্কেটে সবচেয়ে বেশি বিক্রিত পণ্য (যেমন কোকাকোলা) খুঁজতে প্রচুরক ব্যবহার হয়, যা মার্কেটিং স্ট্র্যাটেজিতে সাহায্য করে। গভীরভাবে দেখলে, এগুলোর মধ্যে সম্পর্ক রয়েছে—সমান বিতরণে তিনটিই সমান, কিন্তু ডানদিকে স্কিউড ডেটায় mean > median > mode, এবং ভুল পরিমাপ ব্যবহার করলে ডেটা মিসইন্টারপ্রেট হয়, যেমন স্বাস্থ্য গবেষণায়।

পরিসংখ্যানে ব্যবহৃত চার্ট এবং গ্রাফগুলো ডেটাকে চাক্ষুষ করে তোলে, যাতে জটিল তথ্য সহজে বোঝা যায়, কিন্তু প্রত্যেকটির নির্দিষ্ট উদ্দেশ্য রয়েছে যা ডেটার ধরন এবং উদ্দেশ্য অনুসারে নির্বাচন করতে হয়। বার চার্ট, যা উল্লম্ব বা অনুভূমিক বার দিয়ে ক্যাটাগরিকাল ডেটা তুলনা করে, বিভিন্ন গ্রুপের মধ্যে পার্থক্য দেখাতে সাহায্য করে; উদাহরণস্বরূপ, একটি কোম্পানির বিভিন্ন ডিপার্টমেন্টের বিক্রয় দেখাতে এটি ব্যবহার হয়, যাতে সহজেই বোঝা যায় কোন ডিপার্টমেন্ট সবচেয়ে লাভজনক। হিস্টোগ্রাম, যা ফ্রিকোয়েন্সি বিতরণ দেখাতে বার ব্যবহার করে কিন্তু ডেটাকে বিনে ভাগ করে, কনটিনুয়াস ডেটার আকার দেখায়; বাস্তবে, একটি হাসপাতালে রোগীদের বয়সের বিতরণ দেখাতে এটি ব্যবহার হয়, যা দেখিয়ে দেয় যে বেশিরভাগ রোগী ৪০-৬০ বছরের মধ্যে, এবং এটি স্কিউ বা নরমাল ডিস্ট্রিবিউশন চিহ্নিত করে। লাইন গ্রাফ, যা সময়ের সাথে পরিবর্তন দেখাতে পয়েন্ট যুক্ত করে লাইন তৈরি করে, ট্রেন্ড অ্যানালাইসিসে উপযোগী; উদাহরণ হিসেবে, একটি দেশের জিডিপি বৃদ্ধির গ্রাফে এটি ব্যবহার হয় যাতে অর্থনৈতিক মন্দা বা উত্থান স্পষ্ট হয়। পাই চার্ট, যা পুরোটাকে সেক্টরে ভাগ করে অংশের অনুপাত দেখায়, সীমিত ক্যাটাগরিতে ভালো কাজ করে; বাস্তব জীবনে, একটি পরিবারের মাসিক খরচের বিভাজন (যেমন ৪০% খাবারে, ৩০% ভাড়ায়) দেখাতে এটি ব্যবহার হয়, কিন্তু অনেক অংশ থাকলে বিভ্রান্তিকর হয়ে যায়। স্ক্যাটার প্লট, যা দুটি ভ্যারিয়েবলের সম্পর্ক দেখাতে পয়েন্ট ছড়িয়ে দেয়, করেলেশন চিহ্নিত করে; উদাহরণস্বরূপ, শিক্ষা স্তর এবং আয়ের মধ্যে সম্পর্ক দেখাতে এটি ব্যবহার হয়, যেখানে পয়েন্টের ক্লাস্টারিং পজিটিভ করেলেশন নির্দেশ করে। গভীর চিন্তায়, চার্ট নির্বাচন ডেটার ভলিউম এবং অডিয়েন্সের উপর নির্ভর করে—ভুল চার্ট (যেমন পাই চার্টে অনেক অংশ) ডেটা মিসলিড করতে পারে, এবং আধুনিক টুলস যেমন ট্যাবলো বা এক্সেল এগুলোকে ইন্টারেক্টিভ করে তোলে।

ডেমোগ্রাফি বা জনসংখ্যা বিজ্ঞান জনসংখ্যার গতিবিদ্যা অধ্যয়ন করে, যা সরকারি নীতি, অর্থনীতি এবং সমাজের জন্য অপরিহার্য। জনসংখ্যার আকার এবং ঘনত্ব দেখায় কোনো এলাকায় কত লোক বাস করে এবং প্রতি বর্গকিলোমিটারে কতজন, যা শহুরে পরিকল্পনায় সাহায্য করে; উদাহরণস্বরূপ, বাংলাদেশের উচ্চ জনঘনত্ব (প্রায় ১২০০ জন/বর্গকিমি) কারণে ঢাকায় ট্রাফিক এবং আবাসন সমস্যা হয়, যা ডেমোগ্রাফিক ডেটা থেকে অনুমান করা যায়। জন্মহার এবং মৃত্যুহার, যা প্রতি হাজারে জন্ম বা মৃত্যুর সংখ্যা, জনসংখ্যা বৃদ্ধির মূল চালক; বাস্তবে, ভারতে জন্মহার কমার কারণে বয়স্ক জনসংখ্যা বাড়ছে, যা পেনশন সিস্টেমকে চাপে ফেলে। প্রবাস বা মাইগ্রেশন লোকের স্থানান্তর দেখায়, যেমন গ্রাম থেকে শহরে—উদাহরণ হিসেবে, বাংলাদেশ থেকে মধ্যপ্রাচ্যে শ্রমিক প্রবাস

অর্থনীতিকে সাহায্য করে কিন্তু পরিবার বিচ্ছেদ ঘটায়। বয়স-লিঙ্গ কাঠামো, যা পিরামিড চার্টে দেখানো হয়, ভবিষ্যত চাহিদা অনুমান করে; যেমন চীনে এক সন্তান নীতির কারণে বয়স্ক জনসংখ্যা বাড়ায়, যা লেবর ফোর্স কমায়। ফার্টিলিটি (মহিলাদের গড় সন্তান সংখ্যা) এবং মর্টালিটি (মৃত্যুর কারণ) স্বাস্থ্য নীতিতে সাহায্য করে; উদাহরণস্বরূপ, কোভিড-১৯ মহামারিতে মর্টালিটি রেট বিশ্লেষণ করে ভ্যাকসিন বিতরণ করা হয়। গভীরভাবে, ডেমোগ্রাফি জলবায়ু পরিবর্তনের সাথে যুক্ত—উচ্চ জনসংখ্যা সম্পদের চাপ বাড়ায়, এবং ডেটা থেকে ট্রানজিশন মডেল (যেমন ডেমোগ্রাফিক ট্রানজিশন) বোঝা যায়।

নির্ভরণ রেখা বা রিগ্রেশন লাইন এবং সংশ্লেষণ বা করেলেশন দুটি ভ্যারিয়েবলের সম্পর্ক বোঝায়, কিন্তু করেলেশন শুধু সম্পর্কের শক্তি এবং দিক মাপে যখন রিগ্রেশন ভবিষ্যদ্বাণী করে। করেলেশন কোএফিশিয়েন্ট (-১ থেকে +১) পিয়ারসন সূত্র দিয়ে গণনা হয়, যা কভ্যারিয়েন্সকে স্ট্যান্ডার্ড ডেভিয়েশন দিয়ে ভাগ করে পাওয়া যায়; উদাহরণস্বরূপ, ধূমপান এবং ফুসফুস ক্যান্সারের মধ্যে পজিটিভ করেলেশন (প্রায় ০.৮) দেখায় যে একটি বাড়লে অন্যটিও বাড়ে, কিন্তু এটি কজালিটি প্রমাণ করে না—হয়তো অন্য ফ্যাক্টর (যেমন জেনেটিক্স) জড়িত। রিগ্রেশন লাইন, যেমন লিনিয়ার রিগ্রেশন $Y = a + bX$, স্লোপ (b) এবং ইন্টারসেপ্ট (a) দিয়ে একটি থেকে অন্যটি অনুমান করে; বাস্তবে, বিক্রয় ডেটা থেকে ভবিষ্যত লাভ অনুমান করতে এটি ব্যবহার হয়, যেমন একটি কফি শপে বিজ্ঞাপন খরচ বাড়ালে বিক্রয় কত বাড়বে। গভীর চিন্তায়, করেলেশন $\neq$ কজেশন (যেমন আইসক্রিম বিক্রয় এবং ডুবিয়ে মৃত্যু—দুটোই গরমে বাড়ে), এবং মাল্টিপল রিগ্রেশন একাধিক ভ্যারিয়েবল ধরে, যা মেশিন লার্নিং-এর ভিত্তি।

বিভিন্ন নমুনায়ন পদ্ধতি পুরো জনসংখ্যা থেকে প্রতিনিধিত্বমূলক অংশ নেওয়ার উপায়, যা খরচ এবং সময় সাশ্রয় করে কিন্তু bias এড়াতে সতর্কতা লাগে। সরল দৈব নমুনায়ন, যা লটারির মতো প্রত্যেককে সমান সুযোগ দেয়, সবচেয়ে নিরপেক্ষ কিন্তু বড় জনসংখ্যায় কঠিন; উদাহরণস্বরূপ, একটি স্কুলের ৫০০ ছাত্র থেকে ৫০ জনের মতামত নিতে এটি ব্যবহার হয়। সিস্টেম্যাটিক নমুনায়ন, যা একটা নির্দিষ্ট ক্রমে (যেমন প্রতি ১০ম) নির্বাচন করে, সহজ কিন্তু যদি লিস্টে প্যাটার্ন থাকে তাহলে bias হয়; বাস্তবে, কারখানার প্রোডাক্ট চেক করতে প্রতি ২০তম আইটেম নেওয়া হয়। ক্লাস্টার নমুনায়ন, যা জনসংখ্যাকে গ্রুপে ভাগ করে কয়েকটা গ্রুপ নিয়ে, খরচ কমায়; উদাহরণ হিসেবে, দেশের বিভিন্ন জেলা থেকে কয়েকটা জেলা নিয়ে জরিপ করা হয় যাতে ভ্রমণ খরচ কমে। স্ট্র্যাটিফাইড নমুনায়ন, যা স্তরে (যেমন বয়স বা লিঙ্গ) ভাগ করে প্রত্যেক স্তর থেকে নমুনা নেয়, ভিন্নতা ধরে; বাস্তবে, নির্বাচনী জরিপে লিঙ্গ বা আয়ের ভিত্তিতে এটি ব্যবহার হয় যাতে সকল গ্রুপ প্রতিনিধিত্ব পায়। গভীরভাবে, নমুনায়ন error (যেমন স্যাম্পলিং error) কমাতে র‍্যান্ডমাইজেশন লাগে, এবং ভুল পদ্ধতি (যেমন ভলান্টিয়ার স্যাম্পল) bias ঘটায় যা গবেষণা অসঠিক করে।

কালীন সারণী বা টাইম সিরিজ ডেটা সময়ের সাথে পরিবর্তন দেখায়, যা ভবিষ্যদ্বাণী এবং ট্রেন্ড অ্যানালাইসিসে অপরিহার্য। এতে চারটি উপাদান রয়েছে: ট্রেন্ড (দীর্ঘমেয়াদি পরিবর্তন), সিজনালিটি (মৌসুমী পুনরাবৃত্তি), সাইক্লিক (চক্রাকার ওঠানামা) এবং আইরেগুলার (অনিয়মিত ঘটনা); উদাহরণস্বরূপ, একটি দোকানের মাসিক বিক্রয় ডেটায় ট্রেন্ড দেখায় যে বিক্রয় বাড়ছে, সিজনালিটি দেখায় ঈদের সময় উচ্চ, সাইক্লিক অর্থনৈতিক মন্দায় কম, এবং আইরেগুলার কোনো দুর্ঘটনায় প্রভাবিত। বাস্তবে, স্টক মার্কেটের শেয়ার দামের টাইম সিরিজ বিশ্লেষণ করে ইনভেস্টররা ভবিষ্যত অনুমান করে, যেমন ARIMA মডেল ব্যবহার করে। গভীর চিন্তায়, টাইম সিরিজ স্টেশনারিটি (স্থিরতা) চেক করতে হয়, না হলে অনুমান ভুল হয়, এবং এটি মহামারি ট্র্যাকিং-এর মতো রিয়েল-টাইম অ্যাপ্লিকেশনে ব্যবহার হয়।

এনোভা বা অ্যানালাইসিস অফ ভ্যারিয়েন্স গ্রুপের মধ্যে পার্থক্য পরীক্ষা করে, যা t-টেস্টের এক্সটেনশন; এটি F-স্ট্যাটিস্টিক দিয়ে ভ্যারিয়েন্স তুলনা করে দেখায় যে গ্রুপ মিনের পার্থক্য সিগনিফিক্যান্ট কি না। উদাহরণস্বরূপ, তিন ধরনের সারের প্রভাব ফসলে দেখতে এনোভা ব্যবহার হয়—যদি F-ভ্যালু উচ্চ হয়, তাহলে অন্তত একটা সার ভিন্ন। ডিজাইন অফ এক্সপেরিমেন্টস (DOE) পরীক্ষা পরিকল্পনা করে ফ্যাক্টরের প্রভাব মাপে; এতে রেপ্লিকেশন (একই পরীক্ষা পুনরাবৃত্তি) error কমায়, র‍্যান্ডমাইজেশন bias এড়ায়, এবং ব্লকিং অনাকাঙ্ক্ষিত ভ্যারিয়েবল (যেমন মাটির ধরন) নিয়ন্ত্রণ করে। বাস্তবে, ফার্মাসিউটিক্যাল কোম্পানিতে ওষুধ টেস্টিং-এ DOE ব্যবহার হয় যাতে ডোজ এবং বয়সের প্রভাব সঠিকভাবে ধরা পড়ে। গভীরভাবে, এগুলো র‍্যান্ডমাইজড কন্ট্রোল্ড ট্রায়ালের ভিত্তি, এবং ভুল ডিজাইন টাইপ I/II error ঘটায়।

অতিরিক্তভাবে, ডিসপার্শনের পরিমাপ যেমন ভ্যারিয়েন্স এবং স্ট্যান্ডার্ড ডেভিয়েশন ডেটার ছড়ানো মাপে, যা কেন্দ্রীয় প্রবণতার সাথে যুক্ত। ভ্যারিয়েন্স গড় থেকে মানের বর্গকের গড় বিচ্যুতি, এবং স্ট্যান্ডার্ড ডেভিয়েশন তার বর্গমূল; উদাহরণস্বরূপ, স্টক মার্কেটে ভ্যারিয়েন্স উচ্চ হলে রিস্ক বেশি, যা ইনভেস্টরদের সতর্ক করে। সম্ভাব্যতা বিতরণ যেমন নরমাল ডিস্ট্রিবিউশন (বেল কার্ভ, যেখানে ৬৮% ডেটা ১ SD-এর মধ্যে) বা বাইনোমিয়াল (সাকসেস/ফেলিওর ট্রায়াল) ডেটার প্যাটার্ন বর্ণনা করে; বাস্তবে, IQ স্কোর নরমাল ডিস্ট্রিবিউটেড, যা শিক্ষা নীতিতে সাহায্য করে। ইনফারেন্সিয়াল পরিসংখ্যান স্যাম্পল থেকে পপুলেশন অনুমান করে, যেমন কনফিডেন্স ইন্টারভাল (৯৫% কনফিডেন্সে গড় ৫০ ± ৫); উদাহরণস্বরূপ, পোলিং-এ ভোটার মতামত অনুমান। বায়েসিয়ান পরিসংখ্যান প্রায়র প্রবাবিলিটি আপডেট করে, যা মেডিকেল ডায়াগনোসিসে ব্যবহার হয়—যেমন টেস্ট পজিটিভ হলে ডিজিজের প্রবাবিলিটি আপডেট। অবশেষে, পরিসংখ্যান মেশিন লার্নিং-এর ভিত্তি, যেমন রিগ্রেশন ML মডেলে ব্যবহার হয়

1. **Which of the following is not true for Statistics?**

A) Statistics are aggregate of facts.

B) The science of collecting, organizing, analyzing, interpreting, and presenting data.

C) All numerical statements of facts are statistics but all statistics are not numerical statements of facts.

D) Statistics are collected for a predetermined purpose.

Answer: C

Explanation: Statistics is specifically concerned with data — how to collect it, organize it, analyze it, interpret results, and present findings for decision-making. Moreover, single and isolated figures are not statistics for the simple reason that such figures are unrelated and cannot be compared. For example, the age of a student is 20. This is not statistics. But if we consider the age of student1=22, student2=23, student3=21, this is statistics. Besides, the purpose of collecting data must be decided in advance. But, *"C) All numerical statements of facts are statistics but all statistics are not numerical statements of facts"*- is false. **Correct statement should be-** ***"All numerical statements of facts are not statistics, but all statistics are numerical statements of facts".*** For example, "Textile industries are growing" is not statistics.

"Number of Textile industries increased from 250 in 2005 to 350 in 2015" is statistics. So, statistics must be numerical. But isolated numerical number (The age of only 1 student) is not statistics.

Source: Business Statistics, S.P. Gupta, M.P. Gupta.

1. Which is NOT a/an characteristic of statistics?

A) Helps in predictive analysis
B) Summarizes large data
C) Directly measures emotions like happiness
D) Informed decision-making

Answer:C

Explanation: Statistics helps summarize data, make predictions, and support decisions. But it cannot *directly* measure subjective internal states (like "happiness") with absolute precision. You can *approximate* emotions via surveys (Likert scales) or proxy measures (behavioural indicators), but that is indirect and subject to measurement error and interpretation—so C is not a true "importance" of statistics. A, B, and D are genuine applications

Source: a) Business Statistics, Md. Abdul Aziz. b) Live MCQ Class lecture.

3. Which of the following is a limitation of statistics?

A) Purposive misuses.
B) Statistics rules are mutable.
C) Statistics gives only an overall measures
D) All of the above.

Answer:D

Explanation: Statistics helps summarize data, make predictions, and support decisions. But it cannot measure an individual's condition, rather gives a trend. The average mark of a class is 60. This does not tell the condition of an individual student. Statistical numbers can be misused. For previous example, one teacher can say that his class is doing fine. But most of the students might get less than 30 and average number might be 50 because some exceptional students got more than 90. Moreover, statistics fit model and "All models are wrong, some models are useful"- Box. The rules vary based on dataset.

Source: a) Business Statistics, Md. Abdul Aziz. b) George E.P. Box c) Live MCQ Class lecture.

4. Which of the following is a qualitative variable?

A) Height of a person
B) Blood type
C) Number of books in a library
D) Weight of an object

Answer: B

Explanation: Blood type is categorical (qualitative) because it describes a quality or characteristic without numeric measurement.

Source: a) Business Statistics, Md. Abdul Aziz. b) মৌলিক পরিসংখ্যান- ছিদ্দিক উর রহমান

5. Which scale of measurement does "I.Q. score" represent?

A) Nominal B) Ordinal
C) Interval D) Ratio

Answer: C

Explanation: I.Q. score has equal intervals but **no true zero** (Zero I.Q. score does not mean absence of any intelligence), so it's interval scale.

Source: a) Business Statistics, Md. Abdul Aziz. b) মৌলিক পরিসংখ্যান- ছিদ্দিক উর রহমান

6. Which is true for Q1?

A) 10% of data below this point B) It is equal to 25^{th} percentile
C) 50% of data below this point D) It is equal to median.

Answer: B

Explanation: Q1 (lower quartile) is the value below which 25% of data falls. 25^{th} percentile also tells us same thing.

Source: Mathspace, Statistics Libre Texts, Business Statistics, Md. Abdul Aziz.

7. Which measure of central tendency is most affected by extreme values?

A) Mean B) Median
C) Mode D) Quartile

Answer: A

Explanation: The arithmetic mean uses every value in the dataset and sums them. A single very large or very small value (an outlier) can pull the mean substantially. Example: dataset [2, 3, 4, 5, 100] → mean = (2+3+4+5+100)/5 = 114/5 = 22.8, which is far from the "center" of the lower four numbers. The median (middle value) would be 4, which is not affected by the 100 as much.

Source: a) Basic Statistics, B.L. Agarwal b) Live MCQ class lecture

8. In a stem-and-leaf plot, the "stem" represents:

A) The last digit B) The leading digit(s)
C) The frequency D) The midpoint of class intervals

Answer: B

Explanation: A stem-and-leaf plot shows each data value split into a stem (leading part) and leaf (final digit). Example dataset: 12, 14, 17, 21, 23 → Stem | Leaf

1 | 2 4 7
2 | 1 3

Here stems are the tens (1, 2) and leaves are unit digits (2,4,7...). This preserves raw data while showing distribution.

Source: a) Business Statistics, Md. Abdul Aziz.

9. A cyclist travels 40 km at 20 km/h and then another 40 km at 40 km/h. Find the average speed.

A) 30 km/h
B) 26.67 km/h
C) 40km/h
D) None of above

Answer: B

Explanation: Formula of Harmonic mean will be used here because distance is same and speed unequal.

Solution: The **Harmonic Mean (HM)** formula is:

For n numbers $x_1,x_2,x_3,...,x_n$:

HM= $n/(1/x_1+1/x_2...)$

So, HM= $2/(1/20+1/40)=2/(0.05+0.025)=2/0.075=26.67$km/h

10. Find geometric mean of the numbers 2,4,8.

E) 2

F) 4

G) 6

H) None of above

I) Ans: B

Explanation: GM= $(x_1*x_2*x_3....x_n)^{(1/n)}$

Here, GM= $(2*4*8)^{(1/3)}=64^{(1/3)}=4$

11. If AM=15, and GM=12 for two positive numbers, what is the value of HM?

(a) 9.6 (b) 8

(c) 9.4 (d) 9.2

Answer: (a) 9.6

Explanation:

For two positive numbers we know:

$A.M.*H.M.=(G.M)^2$

So, HM= $G.M.^2/A.M.=144/15=9.6$

12. The mean height of 50 men is 175 cm, and the mean height of 70 women is 165 cm. What is the combined mean height?

(a) 169 cm (b) 169.16 cm

(c) 170 cm (d) 170.6 cm

Answer: (b) 169.16 cm

Explanation:

- Combined mean = $(n_1 * \bar{x_1} + n_2 * \bar{x_2}) / (n_1 + n_2)$:
  - n_1: Number of data points in the first group.
  - $\bar{X_1}$: Mean of the first group.
  - n_2: Number of data points in the second group.
  - $\bar{x_2}$: Mean of the second group.

Combined mean= $(50*175+70*165)/(50+70)=20300/120=169.16$

13. Monthly incomes (in $) of 8 workers are:

2000, 2200, 2400, 2600, 2800, 3400,3000, 3200.

Find Q3 using the formula.

A) 4.5

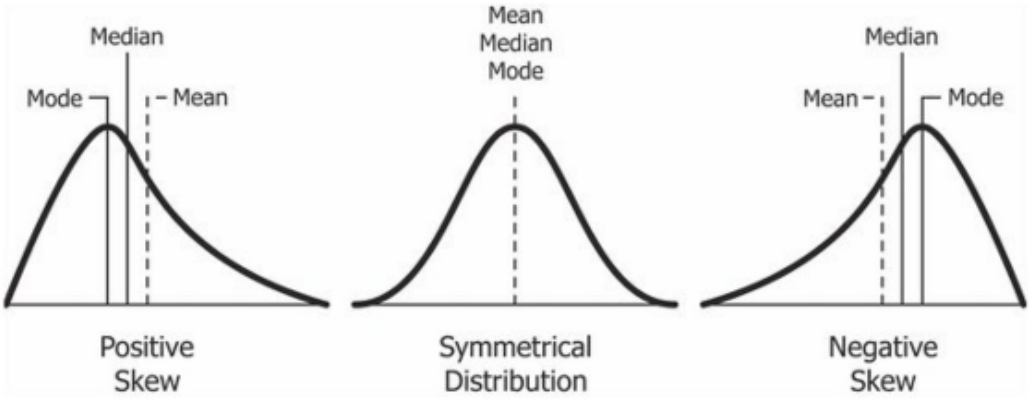
B) 5.5
C) 6.5
D) None of the above
Ans. D.
Explanation:
First, we have to rearrange the number in ascending order. we get,
2000, 2200, 2400, 2600, 2800, 3000, 3200, 3400.
Now, Qk=k(n+1)/4
So, Q3=3(8+1)/4=27/4=6.75. So, Q3 is between 6^{th} and 7^{th} value.
Q3 = 3000 + 0.75 × (3200 − 3000) = 3000 + 150 = **3150**

14. If the coefficient of variation (CV) of dataset A is 15% and of dataset B is 20%, which dataset is more consistent? (যদি ডেটাসেট A এর CV ১৫% এবং ডেটাসেট B এর CV ২০% হয়, তাহলে কোন ডেটাসেটটি বেশি সামঞ্জস্যপূর্ণ?)

A) Dataset A (ডেটাসেট A)
B) Dataset B (ডেটাসেট B)
C) Both are equally consistent (দুটোই সমানভাবে সামঞ্জস্যপূর্ণ)
D) Can't be determined (নির্ধারণ করা যায় না)
Ans. A
Explanation:
The coefficient of variation is used to compare two or more data sets. A lower CV indicates
less relative variability and therefore more consistency.
Source: Investopedia

15. What does a positively skewed distribution indicate?

A) Mean < Median
B) Mean = Median = Mode
C) Mean>Median
D) None of the above.
Ans. C
Explanation:
In a positively skewed distribution, the tail is on the right side, so the mean is greater than the median.
A positively skewed distribution, also known as a right-skewed distribution, is characterized by most of the data being concentrated on the left side, with a long tail extending to the right.
This type of distribution is asymmetrical and often occurs in real-world scenarios such as income levels, stock returns, and hospital stay durations.

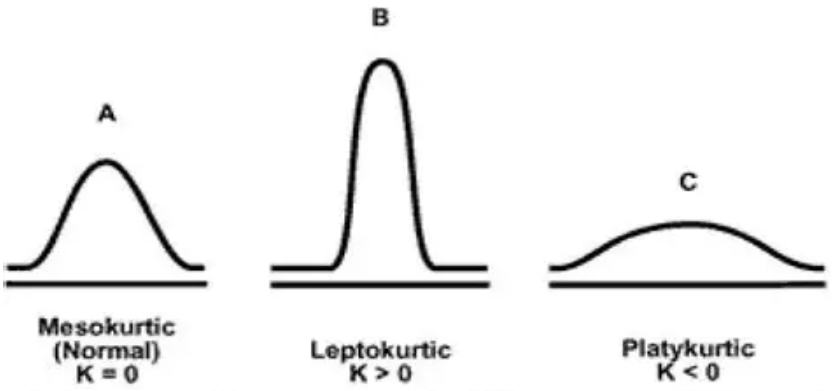


16. What type of kurtosis is indicated by a sharp peak and heavy tails? (sharp peak এবং heavy tails দ্বারা কোন ধরনের কার্টোসিস নির্দেশ করা হয়?)

A) Mesokurtic
B) Platykurtic
C) Leptokurtic
D) Normal kurtosis

Ans. C

Explanation:
Leptokurtic distributions have sharper peaks and heavier tails compared to a normal distribution, indicating more extreme values.
Here are the definitions of leptokurtic, mesokurtic, and platykurtic distributions:
Leptokurtic: Distributions with high kurtosis (fat tails) that are more outlier-prone than a normal distribution.
They have a sharper peak and heavier tails, indicating more values in the tails and closer to the mean.
Mesokurtic: Distributions with medium kurtosis (normal distribution).
They have a moderate peak and tails, similar to the normal distribution.
Platykurtic: Distributions with low kurtosis (thin tails) that are less outlier-prone than a normal distribution.
They have a flatter peak and lighter tails, indicating fewer values in the tails.

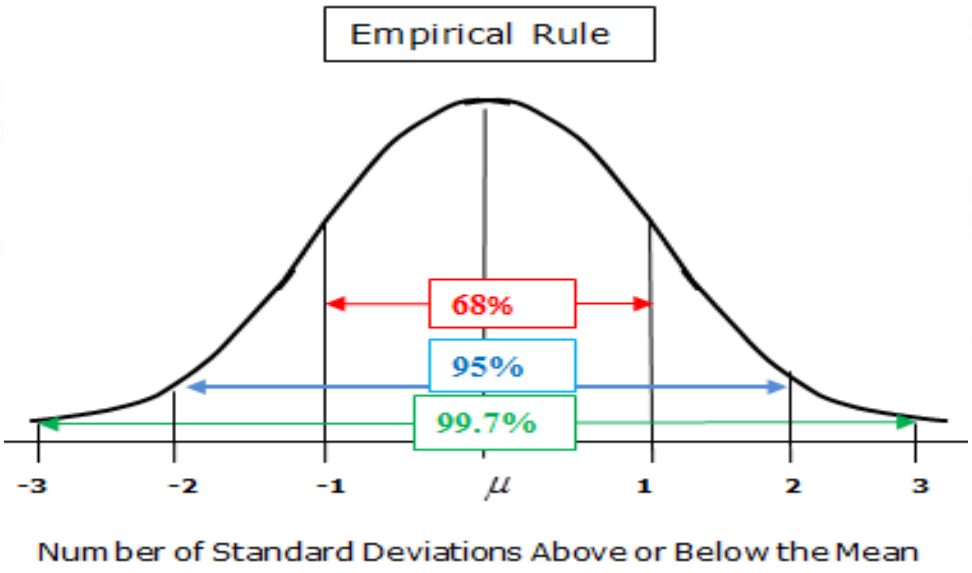


17. A dataset has mean = 50 and standard deviation = 5. What is the coefficient of variation (CV) expressed as a percentage?

A) 5% B) 15%
C) 25% D) None
Ans. D
Explanation:
CV (%) = (SD / Mean) × 100.
SD / Mean = 5 ÷ 50 = 0.1
0.1 × 100 = 10%.

18. For a normal distribution with mean = 100 and standard deviation = 10, approximately what proportion of the data lies between 90 and 110? (গড় = ১০০ এবং বিচ্যুতি = ১০ হলে, প্রায় কত শতাংশ তথ্য ৯০–১১০ এর মধ্যে থাকে?)

ক) 50% খ) 68%
গ) 95% ঘ) 99.7%
Ans. B
Explanation:
In a normal (bell-shaped) distribution, the empirical (68–95–99.7) rule says about 68% of observations lie within ±1 SD of the mean.Here, 90–110 is mean ±10 (±1 SD), so ≈68%.



19. Using Chebyshev's inequality, at least what percentage of observations lie within 1.5 standard deviations of the mean for any distribution? (চেবিসেভের অসমতার ব্যবহার করে, যেকোনো distribution এর জন্য গড় থেকে ১.৫ বিচ্যুতির মধ্যে কত শতাংশ observation থাকে?)

ক) 44.44%
খ) 55.56%
গ) 66.67%
ঘ) 75.00%

Ans. B

Explanation :

Chebyshev rule uses the formula : $1-(1/k^2)$ to find the minimum proportion of data within k standard deviation. (where k must be greater than 1)

Here,

k=1.5

so, $k^2=1.5\times1.5=2.25$.

Then, 1-1/2.25=1-0.444...=0.555...⟶ rounded = 55.56%.

20. When Standard deviation takes the value 0?

(a) When all the observations are non zero

(b) When all the observations are positive

(c) When all the observations are equal

(d) None

Answer: (c)

Explanation: Let three observations 7,7,7. There is no dispersion from mean, hence standard deviation equals 0.

21. Given Central moments: μ2=16, μ3=64, μ4=1536, which of the following is correct?

(a) The distribution is negatively skewed (long left tail) and leptokurtic (sharply peaked).

(b) The distribution is positively skewed (long right tail) and leptokurtic (sharply peaked).

(c) The distribution is positively skewed (long right tail) and mesokurtic (normal like).

(d) Not enough information given to decide shape and peakedness.

Answer: (b)

Explanation:

Given: $\mu_2 = 16,\ \mu_3 = 64,\ \mu_4 = 1536$

- Skewness:

$$\gamma_1 = \frac{64}{16^{3/2}} = \frac{64}{64} = 1$$

- Kurtosis:

$$\beta_2 = \frac{1536}{16^2} = \frac{1536}{256} = 6, \quad \gamma_2 = 3$$

👉 **Description:** The distribution is **positively skewed** (long right tail) and **leptokurtic** (sharply peaked).

22.A dataset has variance μ2= 4 and fourth central moment μ4= 150. Find Pearson's kurtosis and classify the distribution.

A) 9.375 ⟶ Platykurtic

B) 9.375 ⟶ Leptokurtic

C) 4.25 → Mesokurtic

D) 3.75 → Leptokurtic

Ans (B)

Explanation:

$\beta_2 = \mu_4/\mu_2^2$

$= 150/4^2$

$= 150/16$

$= 9.375$

Since β2 > 3, the distribution is Leptokurtic (heavier tails than normal).

Basic definition of Kurtosis states that a distribution is leptokurtic if $\beta_2 > 3$.

23.Given grouped data with mean = 50, median = 48, and standard deviation = 5, calculate Pearson's coefficient of skewness (median-based) [গড় = ৫০, মধ্যক = ৪৮ এবং বিচ্যুতি = ৫, Pearson's coefficient of skewness (median-based গণনা করুন]

(A) 1.20 (B) 0.80

(C) -1.20 (D) -0.80

Explanation:

Formula:

Skewness = 3(Mean-Median)/σ

= 3(50-48)/5 = 1.20

24. The arithmetic mean of 10 numbers is 25. If each number is increased by 5, the new mean will be:

a) 20 b) 25 c) 30 d) 35

Answer: c) 30

Explanation: New mean = old mean + increase = 25 + 5 = 30. Shifting data by a constant increases mean by that constant.

25. Which measure of dispersion is most **sensitive to extreme values**?

a) Interquartile Range b) Range

c) Mean d) Both b and c

Answer: b) Range

Explanation: Range = Max – Min, depends only on extreme values. Variance & SD are influenced but not solely determined by extremes. Mean is a measurement of central tendency, not dispersion.

26. A **histogram** is most suitable for representing:

a) Qualitative data

b) Discrete frequency distribution

c) Continuous frequency distribution

d) Time series data

Answer: c) Continuous frequency distribution

Explanation: Histogram shows frequencies of continuous intervals as adjacent rectangles.

27. Which of the following is a **common error in questionnaire design**?
a) Using mutually exclusive categories
b) Using double-barreled questions
c) Keeping questions short and simple
d) Ensuring anonymity
Answer: b) Using double-barreled questions
Explanation: Double-barreled (two questions in one) confuses respondents and biases results. Another error is leading question.
❌Double-Barreled Question
☞ A single question that actually asks two things at once, but only allows one answer.
Example:
"Do you think the government should increase spending on education and healthcare?"
Problem: Education and healthcare are two separate issues. A respondent might support one but not the other.
"How satisfied are you with your salary and job security?"
Problem: Salary and job security are different things, but only one answer option.
✓ Better version:
"Do you think the government should increase spending on education?"
"Do you think the government should increase spending on healthcare?"
✗ Leading Question
☞ A question that suggests or influences a particular answer, often biased.
Example:
"Don't you agree that our company provides the best customer service?"
Problem: It suggests that the company already provides the best service.
"Since this policy is unfair, do you think it should be changed?"
Problem: The wording "unfair" already pushes the respondent toward "Yes".
✓ Better version:
"How would you rate our company's customer service compared to competitors?"
"What is your opinion about the new policy?"
☞ Double-barreled = two questions in one
☞ Leading = biasing the answer with wording
28. If mean > median > mode, the distribution is:
a) Symmetrical
b) Negatively skewed
c) Positively skewed
d) Leptokurtic
Answer: c) Positively skewed
Explanation: In positive skew, tail on right → Mean > Median > Mode.
29. A distribution with kurtosis $\beta_2 < 3$ is called:
a) Mesokurtic b) Leptokurtic

c) Platykurtic d) Skewed

Answer: c) Platykurtic

Explanation:

$\beta_2 = 3 \rightarrow$ Normal (Mesokurtic)

$\beta_2 > 3 \rightarrow$ Leptokurtic (peaked)

$\beta_2 < 3 \rightarrow$ Platykurtic (flat).

30. In a moderately skewed distribution, which is the best measure of central tendency?

a) Mean b) Median

c) Mode d) Geometric Mean

✓ Answer: b) Median

Mean (average) is heavily affected by extreme values (outliers).

In a skewed distribution, the mean gets pulled in the direction of the skew.

Example: In income data, a few billionaires make the average (mean) much higher than the "typical" income.

Mode is just the most frequent value.

It is not always unique and can be misleading in continuous or irregular distributions.

Geometric Mean is useful mainly for growth rates, ratios, or percentages—not for skewed raw data.

Median (middle value) is not influenced by extremes.

It divides the dataset into two equal halves.

In skewed data, it better represents the "typical" observation.

📌 Rule of thumb:

Symmetrical data → Mean, Median, Mode ≈ same.

Moderately skewed data → Median is preferred.

Heavily skewed data → Median or sometimes Mode (if meaningful).

Example (Skewed Income Data):

Data: 20k, 22k, 25k, 28k, 30k, 500k (an outlier)

Mean ≈ 105,833 → too high, distorted by the outlier.

Median = 26.5k → represents "typical" worker better.

Mode → not very meaningful here.

So, Median = best choice in skewed data.

31. Given data: 5, 10, 15, 20, 25. Find variance.

a) 50 b) 62.5 c) 40 d) 25

Answer: b) 62.5

Explanation:

Mean = 15.

$\Sigma(x-\bar{x})^2 = (100+25+0+25+100)=250.$

Variance = 250/4 = 62.5 (population).

32. Cumulative frequency curve is also known as:

a) Histogram b) Ogive

c) Scatter plot d) Frequency polygon

Answer: b) Ogive

Explanation: Ogive shows cumulative frequencies, useful for medians and percentiles.

An ogive is a cumulative frequency curve.

It plots cumulative frequencies (CF) against class boundaries.

Two types exist:

Less than Ogive → shows cumulative frequency up to a value.

More than Ogive → shows cumulative frequency greater than or equal to a value.

🔻 Why Ogive helps with Median

Median = middle value (50th percentile).

On the ogive graph:

Find N/2 (where N = total frequency).

Draw a horizontal line from N/2 on the Y-axis until it meets the ogive.

Drop a vertical line down to the X-axis.

The corresponding X value is the median.

☞ This works because the ogive shows how many observations are up to a certain point, so the middle point can be located graphically.

🔻 Why Ogive helps with Percentiles / Quartiles

Percentile = value below which a certain % of data lies.

Example: 25th percentile (Q1) = value below which 25% of data lies.

On ogive:

Mark 0.25N, 0.50N, 0.75N on Y-axis (for Q1, Q2, Q3).

Draw horizontals to the ogive, then verticals to X-axis.

The X values are the quartiles/percentiles.

🔻 Quick Example

Suppose we have exam scores of 100 students, and we draw an ogive.

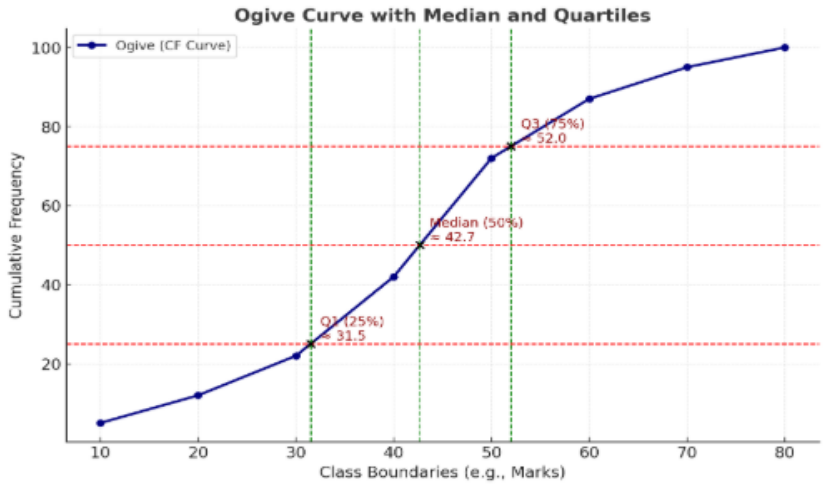
Total frequency = 100

To find Median (50th percentile) → mark 50 on Y-axis → go to curve → drop to X → median score.

To find Q1 (25th percentile) → mark 25 → repeat.

To find Q3 (75th percentile) → mark 75 → repeat.

Thus, ogive converts cumulative frequency into a graphical tool for locating median, quartiles, percentiles.



33. A distribution has Mean = 60, Median = 50, Mode = 40. What type of skewness?

a) Symmetrical b) Negatively skewed

c) Positively skewed d) Normal

Answer: c) Positively skewed

Explanation: Mean > Median > Mode → Positive skew.

34. If the fourth central moment $\mu_4 = 400$ and variance = 25, what is β_2?

a) 2.56 b) 3 c) 2 d) 0.64

Answer: d) 0.64

Explanation: $\beta_2 = \mu_4 / (\mu_2)^2 = 400 / (25^2) = 400 / 625 = 0.64$

Here, Variance=μ_2.

35. Coefficient of variation (CV) is useful for:

a) Comparing variability between datasets with different units

b) Finding absolute dispersion

c) Constructing frequency polygon

d) Detecting skewness

Answer: a) Comparing variability between datasets with different units

Explanation: CV = (SD/Mean)×100 removes units → relative measure.

36. Which mean is **most suitable for averaging rates** (like speed, efficiency)?

a) Arithmetic Mean b) Geometric Mean

c) Harmonic Mean d) Median

Answer: c) Harmonic Mean

Explanation: HM = suitable for rates, ratios.

37. Suppose a histogram of exam scores shows a **U-shaped distribution** (two peaks). Which of the following is **most correct**?

a) Data is positively skewed. b) Data is negatively skewed.

c) Data is bimodal. d) Data is mesokurtic.

Answer: c) Data is bimodal.

Explanation: Two peaks (modes) → bimodal. Skewness describes asymmetry, kurtosis describes peakedness, but "U-shape" suggests bimodal.

38. A question asks:

"How much do you agree that our innovative new product is superior to all competitors?"

This question suffers from:

a) Double-barreled error
b) Leading wording bias
c) Social desirability bias
d) Non-response error

Answer: b) Leading wording bias

Explanation: The phrase *"innovative new product"* and *"superior"* suggest a positive answer. The wording itself biases the response.

39. Two distributions have the **same mean and variance**, but different **coefficients of variation (CV)**. What can you conclude?

a) CV must be the same if variance is same.
b) The distribution with lower mean has higher CV.
c) The distribution with higher mean has higher CV.
d) Nothing can be concluded.

Answer: b) The distribution with lower mean has higher CV.

Explanation: CV = SD/Mean. If SD is same but mean differs → smaller mean → larger CV (greater relative variability).

40. A dataset has **extreme positive outliers**, but the researcher insists on using the **mean** instead of the median. Which **statistical adjustment** would make the mean more reliable?

a) Use geometric mean
b) Winsorizing (trimming extreme values)
c) Replace mean with mode
d) Increase sample size

Answer: b) Winsorizing

Explanation: Winsorizing reduces the effect of outliers by replacing extreme values with the nearest non-extreme values. This makes mean less distorted in skewed distributions.

41. A dataset has: Mean = 50, Median = 40, SD = 20. Using **Pearson's second coefficient of skewness**: Sk=3(Mean-Median)/SD

Find skewness and interpret.

a) +1.5 → moderately positively skewed
b) −1.5 → moderately negatively skewed
c) +0.75 → positively skewed
d) 0 → symmetric

Answer: a) +1.5 → moderately positively skewed

Explanation:

Sk=3(50−40)/20=1.5

Since it's positive → distribution has a long right tail (positively skewed).

42. Which chart is best when there are many categories?

(a) Bar Chart

(b) Pie Chart
(c) Histogram
(d) None

If Categories Are Many:

Bar Chart

Preferred when categories are many.

Each bar represents one category → still readable even if you have 15–20 categories (though better to keep <12 for clarity).

Can be sorted (descending order) to improve readability.

Pie Chart

✗ Not preferred when categories are many (more than ~6–7).

Becomes crowded, slices too small to interpret.

Histogram

Works for continuous data, not categorical.

So not relevant here.

43. Class intervals (Marks): 0–10, 10–20, 20–30, 30–40, 40–50

Frequencies: 5, 12, 18, 14, 6

Calculate Mode and Median from the data.

A) Mode=23 , median= 23.5
B) Mode= 26, Median=27
C) Mode=26 , Median= 25.83
D) None

Ans. C

- **Modal class** = 20–30 (highest frequency = 18)
- $L = 20, f_1 = 18, f_0 = 12, f_2 = 14, h = 10$

$$\text{Mode} = 20 + \frac{(18-12)}{(2(18)-12-14)} \times 10$$

$$= 20 + \frac{6}{36-26} \times 10$$

$$= 20 + \frac{6}{10} \times 10 = 20 + 6 = 26$$

Median: Data (class intervals & frequencies)

0–10 → 5

10–20 → 12

20–30 → 18

30–40 ⟶ 14

40–50 ⟶ 6

☞ Total frequency N=5+12+18+14+6=55.

Step 1: Find cumulative frequencies

0–10 ⟶ 5

10–20 ⟶ 5 + 12 = 17

20–30 ⟶ 17 + 18 = 35

30–40 ⟶ 35 + 14 = 49

40–50 ⟶ 49 + 6 = 55

Step 2: Locate Median class

N/2=55/2=27.5.

The **27.5th value** lies in the **20–30 class** (since CF before it is 17, and CF after it is 35).

So, **Median class = 20–30**.

- **Step 3: Apply Median formula**

$$\text{Median} = L + \left(\frac{\frac{N}{2} - CF_{prev}}{f} \right) \times h$$

Where:

- $L = 20$ (lower boundary of median class)
- $N/2 = 27.5$
- $CF_{prev} = 17$ (cumulative frequency before median class)
- $f = 18$ (frequency of median class)
- $h = 10$ (class width)

- **Step 4: Substitute values**

$$\text{Median} = 20 + \left(\frac{27.5 - 17}{18} \right) \times 10$$

$$= 20 + \left(\frac{10.5}{18} \right) \times 10$$

$$= 20 + 5.83 = 25.83$$

44. **In the equation Y=a+bX, what does b represent?**

(a) Intercept (b) Regression coefficient

(c) Correlation coefficient (d) Error term

Answer: (b)

Explanation: b is the slope, indicating change in Y per unit change in X.

45. **If r = –0.95, then:**

(a) Strong positive correlation (b) Weak negative correlation

(c) Strong negative correlation (d) No correlation

Answer: (c)

Explanation: –0.95 is close to –1, indicating a strong negative linear relation.

46. **Pearson's correlation coefficient measures:**

(a) Association between two variables (b) Causation between two variables

(c) Difference between means (d) Both a and b

Answer: (a)

Explanation: Correlation measures association, not causation.

47. Which method is commonly used to estimate regression coefficients?

(a) Maximum likelihood

(b) Least squares

(c) Chi-square

(d) Z-test

Answer: (b)

Explanation: Least squares minimizes the sum of squared errors.

48. **Which statement is true?**

A) A correlation of 0.8 implies X causes Y.

B) Pearson's r measures monotonic relationships only.

C) A high R^2 always means the model is valid.

D) Pearson's r does not imply causation.

Answer: D

Explanation: Pearson's r does not imply causation, It reflects association.

49. **Partial correlation measures:**

A) Total effect of X on Y

B) Effect of X on Y, controlling for Z

C) Correlation between X and Z

D) None of the above

Answer: B

Partial correlation (r_xy.z) measures association between X and Y **holding Z constant**.

50. Adjusted R^2 is preferred over R^2 when:

(a) The number of predictors increases

(b) We have only one predictor

(c) R^2 is negative

(d) Sample size is zero

Answer: (a)

Explanation: Adjusted R^2 corrects for the inflation in R^2 caused by adding predictors.

51. If the correlation coefficient (r) between two variables is –0.85, which of the following statements is correct? (যদি দুটি চলকের মধ্যে correlation coefficient –0.85 হয়, তবে নিচের কোনটি সঠিক?)

ক) The relationship between the two variables is weak and positive. (দুটি চলকের মধ্যে সম্পর্ক দুর্বল এবং ধনাত্মক)

খ) The relationship between the two variables is strong and positive. (দুটি চলকের মধ্যে সম্পর্ক শক্তিশালী এবং ধনাত্মক)

গ) The relationship between the two variables is strong and negative. (দুটি চলকের মধ্যে সম্পর্ক শক্তিশালী এবং ঋণাত্মক)

ঘ) There is no relationship between the two variables. (দুটি চলকের মধ্যে কোনো সম্পর্ক নেই)

Ans. C

The correlation coefficient (r) ranges from –1 to +1. The closer |r| is to 1, the stronger the relationship. Here, r = –0.85 indicates a strong negative correlation.

Source: Applied General Statistics. Coxton and Crowden.

52. If the correlation coefficient (r) between study hours and exam score is 0, what does this imply? (যদি পড়ার সময় এবং পরীক্ষার নম্বরের মধ্যে correlation coefficient 0 হয়, তাহলে এর দ্বারা কি বুঝায়?)

ক) Studying more hours causes better scores. (বেশি পড়াশোনা করলে ভালো নম্বর আসে)

খ) There is no linear relationship between study hours and exam score. (পড়াশোনার সময় ও পরীক্ষার নম্বরের মধ্যে কোনো linear সম্পর্ক নেই)

গ) There may still have a non-linear relationship. (তাদের মধ্যে non-linear সম্পর্ক থাকতে পারে)

ঘ) Both b,c

Ans. D

When r = 0, there is no linear correlation between the variables, but they may still have a non-linear relationship. Correlation would be called non linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. For example, if we double the amount of rainfall, the production of rice would not necessarily be doubled. It may be pointed out that in most practical cases we find a non linear relationship between the variables.

Source: Business Statistics, SP Gupta, MP Gupta.

53. Two variables X and Y have a correlation coefficient r = 0.9. What percentage of the variation in Y can be explained by X? (যদি X এবং Y চলকের মধ্যে correlation coefficient 0.9 হয়, তবে Y এর কত শতাংশ পরিবর্তন X দ্বারা ব্যাখ্যা করা যায়?)

ক) 81%

খ) 90%

গ) 95%

ঘ) 99%

Ans. A

The coefficient of determination $R^2 = r^2 \times 100\% = (0.9)^2 \times 100\% = 81\%$. This means that 81% of the variation in the dependent variable has been explained by the independent variable. The maximum value r2 can take is 1. Not more than 1. Because r2=1 means all the variation is captured, not more than this variation can be achieved.

Source: Business Statistics, SP Gupta, MP Gupta, Live MCQ class lecture.

54. If $r_{XY} = 0.6$, $\sigma_X = 10$ and $\sigma_Y = 15$, what is Cov(X,Y)? (যদি $r_{XY} = 0.6$, $\sigma_X = 10$ এবং $\sigma_Y = 15$ হয়, তবে Cov(X,Y) কত?)

ক) 90 খ) 60 গ) 150 ঘ) 0.4

Ans. A

Using $r = Cov(X,Y)/(\sigma_X \times \sigma_Y)$,

so, $Cov(X,Y) = r \times (\sigma_X \times \sigma_Y)$

$= 0.6 \times 10 \times 15 = 90$.

55. Which of the following correlation measures is most suitable when the data is ordinal and may have ties? (ডাটা ক্রমিক (ordinal) হলে এবং সমমান (ties) থাকলে কোন correlation measures সবচেয়ে উপযুক্ত?)

ক) Pearson's r (পিয়ারসনের r)

খ) Spearman's ρ (স্পিয়ারম্যানের ρ)

গ) Regression slope (রিগ্রেশন ঢাল)

ঘ) Coefficient of determination (নির্ধারণ সহগ)

Ans. B

Explanation: Spearman's rank correlation uses ranked data, making it suitable for ordinal variables and capable of handling tied ranks.

Why Spearman's ??

ρ: It's a rank-based (non-parametric) measure—replace values with ranks and correlate the ranks. That makes it appropriate for ordinal data and more robust to outliers and certain nonlinear monotonic patterns.

Why not Pearson's ??

r: Pearson assumes roughly interval/ratio scale and linearity; using it on purely ordinal data can be misleading.

56. In, Y = 2.5-0.7X, what does the slope indicate? (Y = 2.5 - 0.7X সমীকরণে ঢালের মান কী বোঝায়?)

ক) A unit increase in X reduces Y by 0.7 (on average) (X এক একক বাড়লে গড়ে Y এর মান 0.7 কমে যায়)

খ) A unit increase in X increases Y by 0.7 (X এক একক বাড়লে Y এর মান 0.7 বেড়ে যায়)

গ) Change in X when Y changes by one (Y এক একক বাড়লে X এর পরিবর্তন কত হয়)

ঘ) The correlation between X and Y (X এবং Y এর মধ্যে correlation)

Ans. A

Interpretation: In the simple linear model
Y = a + bX,
b is the regression coefficient (slope)—the expected change in Y for a one-unit increase in X, holding the model fixed. A negative b means an inverse relationship: as X rises, Y falls on average.

57. A researcher finds a high positive correlation (r = 0.92) between the number of ice creams sold and the number of drowning incidents in a city. Which of the following is the most valid conclusion? (একজন গবেষক একটি শহরে বিক্রি হওয়া আইসক্রিমের সংখ্যা এবং drowning incident এর মধ্যে high positive correlation (r = 0.92) খুঁজে পান। তাহলে নিচের কোনটি সবচেয়ে যুক্তিসঙ্গত ?)

ক) Eating ice cream causes drowning. (আইসক্রিম খাওয়া drowning এর কারণ)

খ) Drowning incidents cause people to buy ice cream (drowning এর ঘটনা মানুষকে আইসক্রিম কিনতে প্ররোচিত করে)

গ) Both are likely influenced by a lurking variable such as temperature (উভয়ই তাপমাত্রার মতো একটি lurking variable দ্বারা প্রভাবিত)

ঘ) The data must be wrong because correlation cannot be that high. (ডাটা ভুল, কারণ সহসম্পর্ক এত বেশি হতে পারে না)

Ans. C
Correlation does not imply causation. Here, both variables are likely related to hot weather — more ice creams are sold, and more people swim, increasing drowning risk. This is a classic example of spurious correlation caused by a third variable.

58. Which of the following is not an assumption of the classical simple linear regression model?
(নিম্নলিখিত কোনটি classical simple linear regression মডেলের অনুমান নয়?)
ক) The relationship between X and Y is linear (X এবং Y এর মধ্যে সম্পর্ক linear)
খ) Residuals have constant variance (homoscedasticity) (অবশিষ্টাংশের constant variance (homoscedasticity) থাকে)
গ) Residuals are normally distributed for all values of X (সব X মানের জন্য অবশিষ্টাংশ স্বাভাবিক বণ্টিত)
ঘ) The correlation between X and Y is exactly 1 (X এবং Y এর মধ্যে সম্পর্কের সহগ ঠিক 1)
Ans. D
Classical regression assumptions include:

Linearity between X and Y
Homoscedasticity of residuals
Independence of residuals
Normality of residuals (for inference)
Perfect correlation (r=±1) is not an assumption; in fact, perfect correlation would be problematic because it would make predictions exact with no error (which rarely occurs in reality).
Source: Regression Analysis, Montogomery.

59. Given the ASFRs below for a population, calculate the TFR: Age Group-(15–19); ASFR=0.05, Age Group-(20–24); ASFR=0.21, Age Group-(25-29); ASFR=0.29, Age Group-(30–34); ASFR=0.19, Age Group-(35–39); ASFR=0.09, Age Group-(40-44); ASFR=0.03, Age Group-(45–49); ASFR=0.01.

A) 2.4 B) 3.8 C) 5.0 D) 4.35

Ans. D

We have to add all ASFRs: 0.05 + 0.21 + 0.29 + 0.19 + 0.09 + 0.03 + 0.01 = 0.87
Multiplying by 5: TFR = 0.87 × 5 = 4.35
Interpretation: Each woman is expected to give birth to around 4 children — indicating a high fertility population.

60. If a country has Crude Birth Rate (CBR) = 30 per 1,000 and Crude Death Rate (CDR) = 10 per 1,000, what is the Natural Growth Rate (NGR)?
(প্রতি 1000 জনে CBR = 30, CDR = 10 হলে, NGR কত?)

ক) 2% খ) 20% গ) 3% ঘ) 4%

Ans. A

Natural Growth Rate = CBR – CDR
Here, Natural Growth Rate = (30/1000) – (10/1000)
= 0.03 - 0.01
= 0.02
= 2%

61. If in a population GRR = NRR, what does it indicate?(যদি GRR = NRR হয়, তবে এর মানে কী?)

A) The average number of children per woman is decreasing (প্রতি নারীর গড়ে সন্তান সংখ্যা কমছে)
B) All women survive through the reproductive age(সব নারী প্রজনন বয়স পার করছে)
C) The fertility rate is very low(জনন হার খুবই কম)
D) Mortality rate is high (মৃত্যুহার বেশি)

Ans. B

GRR (Gross Reproduction Rate): এটা বোঝায়, "একজন মহিলার পুরো জীবনকালে কত মেয়ে সন্তান হবার সম্ভাবনা আছে যদি বর্তমান জন্মহার (ASFR) একই থাকে।" এখানে মৃত্যুর হিসাব (mortality) নেওয়া হয় না।

NRR (Net Reproduction Rate): এটা বোঝায়, "একজন মহিলার পুরো জীবনকালে কত মেয়ে সন্তান বাঁচে এবং প্রজননযোগ্য বয়সে পৌঁছায়।" এখানে মৃত্যুর প্রভাব ধরা হয়।

যদি GRR = NRR হয়, তাহলে এর মানে হলো:

মহিলারা প্রজননযোগ্য বয়স পর্যন্ত সকলেই বেঁচে থাকে, কোনো মৃত্যু হয়নি। অর্থাৎ, নবজাতক কন্যা শিশুদের মধ্যে কেউই সন্তান ধারণে সক্ষম বয়সের শেষ সীমার অর্থাৎ 49 বছরের পূর্ব পর্যন্ত মারা যাবে না।

GRR considers only fertility, while NRR accounts for both fertility and mortality. If GRR = NRR, mortality has no effect, meaning all women survive through their reproductive years.

62. The Crude Birth Rate (CBR) of a town is 25 per 1,000 population. If the total population is 80,000, how many births occurred in that year?

(CBR = প্রতি 1000 জনে 25 জন ; জনসংখ্যা = 80,000 । ওই বছর কত জনের জন্ম হয়েছে?)

ক) 2,000 খ) 2500 গ) 3000 ঘ) 3500

Ans. A

Explanation: CBR = (Births / Total population) × 1,000

→ Births = (25 × 80,000) / 1,000

= 2,000.

63. Which rate considers only live female births expected per woman? (কোন হারটি কেবলমাত্র প্রতি নারীর প্রত্যাশিত জীবিত কন্যা সন্তানের জন্ম বিবেচনা করে?)

ক) TFR খ) CBR গ) GRR ঘ) CDR

Ans. D

Gross Reproduction Rate estimates daughters a woman will have assuming no mortality. It is more refined version than TFR.

64. The annual population growth rate of a country is 2%. Using the Rule of 70, what is the approximate doubling time of the population? (বার্ষিক জনসংখ্যা বৃদ্ধির হার যদি 2% হয়, তবে Rule of 70 অনুযায়ী জনসংখ্যা দ্বিগুণ হতে কত বছর লাগবে?)

A) 35 years (৩৫ বছর) B) 50 years (৫০ বছর)

C) 70 years(৭০ বছর) D) 14 years(১৪ বছর)

Ans. A

Explanation:

Population doubling time = 70 / growth rate (%)

Doubling time = 70/2

= 35 years

65. Which fertility measure is the most refined version among all the measures?(নিম্নলিখিত কোন Fertility Measure সবচেয়ে সূক্ষ্ম বা refined?)

ক) ASFR খ) CBR গ) GRR ঘ) NRR

Ans. D

Each measure is a refined version of the previous one:

CBR → GFR → ASFR → TFR → GRR → NRR; They move from general to specific, becoming more accurate and meaningful for demographic analysis.

66. Which statement is TRUE regarding Age-Specific Fertility Rate (ASFR)? (Age-Specific Fertility Rate (ASFR) সম্পর্কে কোনটি সঠিক?)

ক) It is always less than Gross Reproduction Rate (এটি সবসময় GRR এর চেয়ে কম)

খ) It is unaffected by women's reproductive age(এটি নারীর প্রজনন বয়স দ্বারা প্রভাবিত নয়)

গ) It refers to female births per 1,000 women in a specific age group (এটি নির্দিষ্ট বয়সগোষ্ঠীতে প্রতি ১০০০ নারীর জন্য কন্যা জন্ম নির্দেশ করে)

ঘ) None (কোনোটিই নয়)

Ans. D

ASFR is calculated for specific reproductive 7 age groups (e.g. 15-19, 20–24, 25-29, 30-34, 35-39, 40-44, 45-49).

It refers to live births (both boy and girl) per 1,000 women in a specific age group, not female births only.

It is always more than GRR because GRR only considers female births.

67. If the Net Reproduction Rate (NRR) is equal to 1, then: (NRR যদি 1 হয়, তবে এর মানে কী?)

ক) Population will reduce to half (জনসংখ্যা অর্ধেক হবে)

খ) Each generation of mothers is exactly reproducing itself (প্রতিটি প্রজন্ম নিজেকে সঠিকভাবে প্রতিস্থাপন করছে)

গ) Population will triple in next generation (জনসংখ্যা তিনগুণ হবে)

ঘ) There is zero female birth (কোনো কন্যাশিশু জন্ম নিচ্ছে না)

Ans. B

Explanation: NRR=1 indicates each generation of mothers is exactly reproducing and hence replacing itself. It interprets a population stability in the long run.

NRR<1 interprets a population decline in the long run.

NRR>1 interprets a population growth in the long run.

68. Given: Women aged 20–24 = 5,000; Births to them in a year = 400. Find ASFR (per 1,000 women).(দেওয়া আছে: 20–24 বছর বয়সী নারী = 5,000; জন্ম = 400। ASFR প্রতি 1,000 নারী কত হবে?)

A) 50 B) 80

C) 40 D) None of the above

Ans. B

ASFR considers 7 age structures (15-19,20-24, 25-29, 30-34, 35-39, 40-44, 45-49).

They calculate fertility rates for each age group.

Formula of ASFR (for age group i) = (Births of age group i / Women age group i) × 1,000

= (400/ 5,000) × 1,000

= 80. (Here, i= 20-24)

69. Main purpose of Index Numbers is: (Index Numbers-এর মূল উদ্দেশ্য কী?)

ক) Study changes over time(সময় অনুযায়ী পরিবর্তন অধ্যয়ন)

খ) Predict population(জনসংখ্যা পূর্বাভাস)

গ) Economic forecasting (অর্থনৈতিক পূর্বাভাস)

ঘ) Both a,c (উভয়ই, ক ও গ)

Ans. D

Index numbers are like economic barometer. They measure the change of price, quantity, value etc. over time. They are useful in calculating Consumer price index/cost of living index. They measure relative change in variables and thus give important insights on economy.

70. An NRR of 0.7 implies:

ক) Rapid population growth (দ্রুত জনসংখ্যা বৃদ্ধি)

খ) Each generation is replacing 170% of itself (প্রতিটি প্রজন্ম নিজেদের 170% প্রতিস্থাপন করছে)

গ) Each generation is replacing 70% of itself (প্রতিটি প্রজন্ম নিজেদের 70% প্রতিস্থাপন করছে)

ঘ) Each generation is replacing 1.7 of itself (প্রতিটি প্রজন্ম নিজেদের 1.7 গুণ প্রতিস্থাপন করছে)

Ans.C

Explanation: NRR = Less than 1 means reduction in size over generations.

Here, the replacement level is less than 1, which interprets that, population will decrease.

71. Which of the following best explains Fisher's Ideal Index Number?

(নিচের কোনটি দ্বারা Fisher's Ideal Index Number সবচেয়ে ভালোভাবে বোঝা যায়?)

ক) It is simply the arithmetic mean of Laspeyres and Paasche indices

(এটি ল্যাসপায়রেস ও Paasche সূচকের গাণিতিক গড়)

খ) It is the geometric mean of Laspeyres and Paasche indices

(এটি ল্যাসপায়রেস ও Paasche সূচকের জ্যামিতিক গড়)

গ) It passes both time reversal and factor reversal tests

(time reversal ও factor reversal পরীক্ষায় উত্তীর্ণ হয়)

ঘ) Both b, c (খ ও গ উভয়ই)

Ans. D

Explanation:

Fisher's Ideal Index is called ideal because it overcomes the bias of both Laspeyres' (base year weighted) and Paasche's (current year weighted) indices by taking their geometric mean, satisfying time reversal and factor reversal tests.

Time reversal test- An index number formula should give consistent results over time. If we reverse the time period (swap base year and current year), the product of the two index numbers should be 1 (or 100, depending on scale).

Mathematically:

$P_{01} \times P_{10} = 1$

where, P01= price index from time 0 to 1

P10= price index from time 1 back to 0

Interpretation:

If an index satisfies this test, it means it is time-consistent — the upward change going forward should be exactly offset by the backward change.

Which index passes?

Fisher's Ideal Index satisfies the time reversal test, whereas Laspeyres' and Paasche's indices do not, due to their fixed weighting systems.

Factor reversal test- This test checks whether the price index multiplied by the quantity index equals the value index (i.e., the change in total monetary value from base period to current period).

Mathematically:

$P_{01} \times Q_{01} = V_{01}$

Interpretation:

If a formula passes this test, it means it measures price and quantity changes without leaving any part unexplained, perfectly decomposing the change in total value.

Which index passes?

Again, Fisher's Ideal Index satisfies this test, while Laspeyres' and Paasche's indices do not.

Source: Statistics, HSC, Md. Abdul Aziz.

S.P. Gupta, Statistical Methods, 50th Edition, 2020

M.P. Gupta, Fundamentals of Statistics, 2018

72.Which of the following statements is true about GRR and NRR?

A) NRR is always greater than GRR

B) GRR accounts for mortality, NRR does not

C) NRR is always less than or equal to GRR

D) GRR measures both male and female births

Correct Answer: C

73.An index number of 150 means:

A. Price increased by 50%
B. Price decreased by 50%
C. Price doubled
D. Price stayed the same

Ans. A

Explanation: Index = 150 means prices are 50% higher compared to the base year (base = 100).

74.In Paasche index, quantities are taken from:

A. Base Year
B. Current Year
C. Average of base and current
D. None of these

Ans. B

Explanation: Paasche uses current year quantities as weights

75. Which index number method is the geometric mean of Laspeyres and Paasche?

A. Marshall–Edgeworth

B. Fisher's

C. Bowley's

D. Kelly's

Ans. B

Explanation: Fisher's Ideal Index = √(Laspeyres × Paasche). It is called "ideal" because it balances both methods. It balances upward bias of L and Downward bias of P.

76.Nuptiality refers to: (Nuptiality বলতে বোঝায়-)

ক) The study of fertility and birth rates (জনন ক্ষমতা ও জন্মহার বিষয়ক গবেষণা)

খ) The death rate among infants (শিশুমৃত্যুর হার)

গ) The frequency, characteristics and patterns of marriage in a population (একটি সমাজে বিবাহের হার, বৈশিষ্ট্য ও ধরণ)

ঘ) The migration of unmarried females (অবিবাহিত নারীদের অভিবাসন)

Ans. C

Explanation:

In demography, nuptiality deals with marriage-related aspects such as age at marriage, proportion of married persons, marriage rates, etc. It helps understand the social structure and reproductive potential of a population. In countries like Bangladesh, nuptiality indicators (e.g., average age at first marriage) are crucial for policymaking in health, population control, and women's empowerment.

77.Which of the following statements correctly compares the biases of Laspeyres and Paasche index numbers?

ক) Both Laspeyres and Paasche tend to overstate price rise

খ) Laspeyres tends to overstate, while Paasche tends to understate the price rise

গ) Laspeyres tends to understate, while Paasche tends to overstate the price rise

ঘ) Both always give the same result as Fisher's Ideal Index

Ans. B

Laspeyres index uses base-year quantities, ignoring substitution toward cheaper goods — so it overstates inflation. Paasche uses current-year quantities, incorporating substitution, which tends to understate inflation. Fisher, being geometric mean of the two, typically lies in between. Hence option B is correct.

78. One limitation of the Paasche Price Index is: (Price Index এর সীমাবদ্ধতাঃ)

ক) It overestimates price changes due to base year quantities

খ) It requires current year quantities, which may not always be available

গ) It is a geometric mean of two indices

ঘ) It cannot be used for international comparison

Ans. B

The Paasche Price Index uses current year quantities as weights, which can be hard to obtain or unreliable. Option A describes Laspeyres' bias. Option C is characteristic of Fisher. Option D is not a general limitation; Paasche can be used internationally if data is available.

79.

Commodity	P_0	P_1	Q_1
A	20	25	10
B	30	35	5

Compute **Paasche Price Index.**

A) 112
B) 121.4
C) 109.5
D) None

Ans. B

Solution:

$$P = \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times 100$$

- Numerator: 25×10 + 35×5 = 250 + 175 = 425
- Denominator: 20×10 + 30×5 = 200 + 150 = 350

$$P = \frac{425}{350} \times 100 \approx 121.43$$

80..Which of the following is true about CPI?

A) It measures the average change in quantity of goods consumed
B) It measures changes in prices of a fixed basket of consumer goods
C) It ignores weights of goods
D) It always uses current period quantities as weights

Answer: B

Explanation:

CPI tracks price changes for a **fixed basket of goods**, often weighted according to expenditure shares of households.

81. A country has:

Population aged 0–14 = 20 million
Population aged 65+ = 5 million
Working-age population (15–64) = 50 million
What is the dependency ratio?

A) 50%
B) 50.5%
C) 50.0%
D) 25%

Ans. A

Solution:

Formula:

$$\text{Dependency Ratio (\%)} = \frac{\text{Population aged 0–14} + 65+}{\text{Working age population}} \times 100$$

$$= \frac{20+5}{50} \times 100 = \frac{25}{50} \times 100 = 50\%$$

Answer: A ✅

82.Which statement is correct?

A) CPI is always unweighted

B) CPI uses weighted index numbers because some goods are more important

C) CPI ignores consumption pattern

D) CPI is same as Paasche always

Answer: B

Explanation:

CPI gives **weights based on expenditure patterns** to reflect real consumption impact.

83.Which of the following is a key difference between weighted and unweighted index numbers?

(weighted ও unweighted সূচকের মধ্যে মূল পার্থক্য কোনটি?)

ক) Unweighted index numbers require prices and quantities; weighted do not (unweighted সূচক দাম ও পরিমাণ চায়; Weighted তা চায় না)

খ) Weighted index numbers assign importance to items based on quantities or value, while unweighted treat all items equally

(Weighted সূচক পণ্যের গুরুত্ব নির্ধারণ করে পরিমাণ বা মূল্যের ভিত্তিতে, আর unweighted সব পণ্যকে সমানভাবে দেখে)

গ) Weighted index numbers are used only for quantity indices (Weighted সূচক কেবল quantity সূচকের জন্য ব্যবহৃত হয়)

ঘ) Unweighted index numbers are more accurate because they use weights (unweighted সূচক বেশি নির্ভুল কারণ এগুলো Weight ব্যবহার করে)

Ans. B

In weighted index numbers, each commodity is given a weight denoting its relative importance (generally quantity or value). In unweighted index numbers, all items are given equal importance, regardless of how much is consumed or produced.

84. Which one of the following is an example of an unweighted index number?

(নিচের কোনটি একটি unweighted সূচকের উদাহরণ?)

ক) Laspeyres Price Index (ল্যাসপিয়ার্স মূল্যসূচক)

খ) Paasche Quantity Index (পাশ্চে পরিমাণ সূচক)

গ) Simple Aggregative Price index (সাধারণ সমষ্টিগত মূল্যসূচক)

ঘ) Fisher's Ideal Index (ফিশারের আদর্শ সূচক)

Ans. C

The Simple Aggregative Price Index is calculated by summing current prices and dividing by the sum of base prices — it doesn't use any weights for individual commodities, hence it is unweighted. All the others (Laspeyres, Paasche, Fisher) are weighted index numbers, as they use quantities as weights.

Source: Business Statistics (SP Gupta, MP Gupta.)

85. What is the primary difference between a price index and a quantity index? (একটি মূল্য সূচক এবং একটি পরিমাণ সূচকের মধ্যে প্রধান পার্থক্য কী?)

ক) A price index measures changes in the quantity of goods, while a quantity index measures changes in the price of goods. (একটি মূল্য সূচক পণ্যের পরিমাণের পরিবর্তন পরিমাপ করে, অন্যদিকে একটি পরিমাণ সূচক পণ্যের মূল্যের পরিবর্তন পরিমাপ করে।)

খ) A price index measures changes in the price of goods, while a quantity index measures changes in the quantity of goods. (একটি মূল্য সূচক পণ্যের মূল্যের পরিবর্তন পরিমাপ করে, অন্যদিকে একটি পরিমাণ সূচক পণ্যের পরিমাণের পরিবর্তন পরিমাপ করে।)

গ) A price index uses base-year prices, while a quantity index uses current-year quantities (একটি মূল্য সূচক ভিত্তি বছরের মূল্য ব্যবহার করে, অন্যদিকে একটি পরিমাণ সূচক বর্তমান বছরের পরিমাণ ব্যবহার করে।)

ঘ) A price index uses current-year prices, while a quantity index uses base-year quantities (একটি মূল্য সূচক বর্তমান বছরের মূল্য ব্যবহার করে, অন্যদিকে একটি পরিমাণ সূচক ভিত্তি বছরের পরিমাণ ব্যবহার করে।)

Ans. B

Option b directly defines the core function of each type of index. Hence, this the correct one.

option a তে price index quantity of goods এর change করে বলা হয়েছে, যা ভুল।

option b তে- সঠিক উত্তর।

option c তে- price index, base year ও current year দুটাই ব্যবহার করতে পারে, তাই ভুল।

option d- option c এর মতোই ভুল

86. Suppose the price of 3 commodities in the base year (P_0) are: A = 10, B = 20, C = 30. In the current year (P_1): A = 15, B = 25, C = 45. If we calculate the index number using: (i) Simple Aggregate Method (ii) Simple Average of Price Relatives Method, Which of the following statements is correct?

(ধরা যাক, ভিত্তি বছরে (P_0) তিনটি পণ্যের মূল্য ছিল: A = ১০, B = ২০, C = ৩০। বর্তমান বছরে (P_1): A = ১৫, B = ২৫, C = ৪৫। যদি আমরা সূচক সংখ্যা গণনা করি: (i) সাধারণ সমষ্টি পদ্ধতি এবং (ii) মূল্য আপেক্ষিকের সাধারণ গড় পদ্ধতি ব্যবহার করে, তবে নিম্নলিখিত কোন বিবৃতিটি সঠিক?)

ক) Both methods give the same result (উভয় পদ্ধতি একই ফলাফল দেয়।)

খ) Aggregate method gives a higher index than average relative method (সমষ্টি পদ্ধতি আপেক্ষিক গড় পদ্ধতির চেয়ে উচ্চতর সূচক দেয়।)

গ) Average relative method gives a higher index than aggregate method (আপেক্ষিক গড় পদ্ধতি সমষ্টি পদ্ধতির চেয়ে উচ্চতর সূচক দেয়।)

ঘ) Both methods are biased but give opposite directions (উভয় পদ্ধতিই পক্ষপাতমূলক কিন্তু বিপরীত দিকে ফলাফল দেয়।)

Ans. A

Simple Aggregate Index = $(\Sigma P_1 / \Sigma P_0) \times 100 = (15+25+45) / (10+20+30) \times 100 = 85/60 \times 100 = 141.67$

Simple Average of Price Relatives = $[(15/10\times100 + 25/20\times100 + 45/30\times100)] / 3 = (150 + 125 + 150)/3 = 141.67$

87. A CPI of 140 (base year 2010 = 100) indicates that: (যদি CPI = ১৪০ হয় (ভিত্তি বছর ২০১০ = ১০০), তবে এটি কী নির্দেশ করে?)

ক) The general price level has fallen by 40% since 2010. (২০১০ সাল থেকে সাধারণ মূল্যস্তর ৪০% কমেছে।)

খ) Household consumption has increased by 40% since 2010 (২০১০ সাল থেকে পরিবারের ব্যয় ৪০% বেড়েছে।)

গ) The general price level has increased by 140% since 2010. (২০১০ সাল থেকে সাধারণ মূল্যস্তর ১৪০% বেড়েছে।)

ঘ) None (কোনোটিই নয়।)

Ans. D.

Explanation:

CPI compares the cost of a fixed basket of goods and services relative to the base year. A value of 140 means the basket now costs 40% more than in 2010, reflecting the rise in general prices.

A correct statement for the questions is- " The general price level has increased by 40% since 2010"

88. The long-term upward or downward movement in data is known as (ডেটায় দীর্ঘমেয়াদি ঊর্ধ্বগতি বা নিম্নগতি কী নামে পরিচিত?)

ক) Seasonal Fluctuation (ঋতুগত ভিন্নতা)

খ) Trend (প্রবণতা)

গ) Irregular Variation (অনিয়মিত ভিন্নতা)

ঘ) Cyclic Variation (চক্রাকার ভিন্নতা)

Ans. B

Explanation:

Secular Trend or normally "Trend" reflects persistent long-term growth or decline in a series over several years, such as population growth.

When we talk of trend, we mean smooth, regular, long term movement of the data- sudden and erratic movements either in upwards or in downward direction have nothing to do with the trend.

Source: Business Statistics (SP Gupta, MP Gupta)

89. The method of moving averages is used to

(Moving average পদ্ধতি ব্যবহৃত হয় —)

ক) Measure correlation (সহসম্বন্ধ পরিমাপ করতে)

খ) Remove trend (Trend অপসারণ করতে)
গ) Remove non-linear trend (অরৈখিক Trend মুছে ফেলতে)
ঘ) Measure dispersion (বিচ্যুতি পরিমাপ করতে)
Ans. B
There are broadly two types of trend -Linear trend, Non-linear trend.
For measuring straight line trend, we use Semi average method, Method of least square etc.
On the other hand, Moving average method is used to measure non-linear trend.
The moving average method is a technique used to analyze data trends by calculating the average of a specific set of data points over a defined period, which is then "moved" through the data set.
Moving averages smooth out short-term fluctuations, isolating the long-term trend.

90. Deseasonalizing a time series means —
(টাইম সিরিজের Deseasonalizing বলতে বোঝায় —)
ক) Adding trend (Trend যোগ করা)
খ) Removing cyclic effect (চক্রাকার প্রভাব দূর করা)
গ) Eliminating seasonal effect (ঋতুগত প্রভাব অপসারণ করা)
ঘ) Removing random effect (এলোমেলো প্রভাব দূর করা)
Ans. C
Deseasonalizing time series data involves removing the seasonal component to reveal underlying trends and make it easier to analyze the data.
This process is useful for identifying patterns and making predictions, especially when seasonal variations obscure the true trend.

91. The irregular component in a time series is also called —
(টাইম সিরিজে Irregular component আর কী নামে পরিচিত?)
ক) Random component (এলোমেলো উপাদান)
খ) Seasonal component (ঋতুগত উপাদান)
গ) Trend component (প্রবণতা উপাদান)
ঘ) Cyclic component (চক্রাকার উপাদান)
Ans. A
A time series can be broken down into four key components: trend, seasonality, cyclicality and irregularity. These components help to understand patterns and make predictions about future values in a dataset. **Irregular variation represents unpredictable, random shocks like wars or natural disasters.**

92. Year: Value= (2016,20), (2017, 24), (2018,27), (2019,32), (2020,39), (2021,43).
calculate the trend value using the Semi-Average Method. What is the estimated trend for 2019? (Semi-Average Method ব্যবহার করে 2019 সালের Trend Value কত হবে?)
ক) 29 খ) 30.5 গ) 32.5 ঘ) 33.22
Ans. D
Step-1: Since we have even number (6) of years, split into two equal parts:

First Half (2016-2018):

Values = 20, 24, 27 ⟶ Mean = (20+24+27)/3 = 23.67

Second Half (2019-2021):

Values = 32, 39, 43 ⟶ Mean = (32+39+43)/3 = 38

Step-2: These represent average trend at the "middle year" of each half ⟶

Trend value at 2017 = 23.67

Trend value at 2020 = 38

Step-3: Fit a straight line between (2017,23.67) and (2020,38) and interpolate for 2019.

Slope = (38-23.67)/ (2020-2017)= 14.33/3=4.77

Change from 2017 to 2019 is 2 years ⟶ Trend at 2019 =

23.67+(4.777×2)=23.67+9.554=33.224

93. Which model represents time series as: Y = T × S × C × I?

(কোন মডেলে টাইম সিরিজকে প্রকাশ করা হয়: Y = T × S × C × I ?)

ক) Additive model (যোগফল মডেল)

খ) Multiplicative model (গুণফল মডেল)

গ) Regression model (রিগ্রেশন মডেল)

ঘ) Autoregressive model (অটো-রিগ্রেসিভ মডেল)

Ans. B

In the multiplicative model components multiply; additive means they sum.

94. Eliminating trend from time series data is called-

(টাইম সিরিজ থেকে Trend বাদ দেওয়াকে কী বলে?)

ক) Additivity (যোগফল) খ) Multiplicity (গুণফল)

গ) deseasoning ঘ) Detrending

Ans. D

Eliminating trend from time series data is called detrending.

While eliminating seasonality from time series data is called deseasoning.

Additive model is a model where components add up.

Source: Business Statistics (Md. Abdul Aziz)

95. A regular fluctuation within a year is called —

(এক বছরের ভেতর নিয়মিত ওঠানামা কী নামে পরিচিত?)

ক) Trend (প্রবণতা)

খ) Cyclical variation (চক্রাকার ভিন্নতা)

গ) Seasonal variation (ঋতুগত ভিন্নতা)

ঘ) Irregular variation (অনিয়মিত ভিন্নতা)

Ans. C

Seasonal variation refers to periodic changes repeating within 12 months (e.g. ice cream sales in summer).

96. What is the main purpose of deseasoning a time series?

(টাইম সিরিজ Deseasoning করার মূল উদ্দেশ্য কী?)

ক) To eliminate random errors. (Random Error দূর করা)

খ) To study long-term trend and cyclical variation without seasonal effects. (Seasonal Effect বাদ দিয়ে Long-term Trend ও Cycle অধ্যয়ন করা)

গ) To make data stationary for ARIMA modeling. (ARIMA Model-এর জন্য Data Stationary করা)

ঘ) To reduce data size for easier computation. (Data ছোট করা যাতে সহজে হিসাব হয়)

Ans.B

Deseasoning removes seasonal fluctuations so the underlying trend and cycle can be clearly analyzed.

97. Which of the following statements about the graphical method of trend projection is correct?

(Graphical Method of Trend Projection সম্পর্কিত সঠিক বক্তব্য কোনটি?)

ক) It is the most accurate method of forecasting. (এটি সবচেয়ে নির্ভুল Forecasting Method)

খ) It is subjective and depends on the person drawing the trend line. (এটি Subjective এবং যিনি আঁকছেন তার উপর নির্ভরশীল)

গ) It requires complex statistical computation. (এতে জটিল পরিসংখ্যান দরকার হয়)

ঘ) It automatically adjusts for seasonality. (এটি স্বয়ংক্রিয়ভাবে Seasonality সামলায়)

Ans. B

The graphical method is a visual approach where a trend line is drawn by inspection. Its strength is simplicity, but its limitation is subjectivity — different people may draw slightly different lines, leading to inconsistent forecasts.

98. Trend Can be measured by-

(Trend কোন পদ্ধতিতে মাপা যায়?)

ক) Graphical method

খ) Method of Semi-averages

গ) Method of moving averages

ঘ) All of the above

Ans. D

Trend can be measured by mainly 4 methods---

1. Graphical method,
2. Method of Semi-averages,
3. Method of moving averages,
4. Method of Least square.

Source: Business Statistics (Md. Abdul Aziz)

99. When using the Semi-Average Method for trend estimation, which of the following is a key limitation?

(Semi-Average Method-এ Trend অনুমান করার প্রধান সীমাবদ্ধতা কী?)

ক) It requires a large number of observations to compute the trend. (এতে অনেক Observation প্রয়োজন হয়।)

খ) The method assumes a linear trend and ignores possible curvature in the data. (এটি Linear Trend ধরে নেয় এবং Curvature উপেক্ষা করে)

গ) It can only be used for odd-numbered time-series data. (এটি কেবল Odd সংখ্যক Observation-এ ব্যবহার করা যায়।)

ঘ) It provides exact predictions for future points. (এটি ভবিষ্যতের সঠিক পূর্বাভাস দেয়।)

Ans. B

The Semi-Average Method splits the series into two equal halves, computes the average of each half, and connects these averages with a straight line.

This inherently assumes that the trend is linear, which can lead to inaccurate estimates if the actual trend is non-linear or exponential.

Option A is not strictly true; the method works even for small series.

Option C is false; the method works for both odd and even numbers of observations (splitting may be slightly adjusted for odd-length series).

Option D is incorrect; semi-average provides trend estimates, not exact forecasts.

Source: Business Statistics (SP Gupta, MP Gupta.), Live MCQ class lecture

100. In simple random sampling, the probability of selecting any particular unit is:

a) 0

b) Equal and non-zero

c) Unequal but non-zero

d) Depends on researcher's choice

Answer: b

☞ Each unit has equal chance in SRS.In SRS, every unit of the population has the same probability of being selected. This makes it unbiased. Example: Lottery draw.

101. Which sampling method ensures each subgroup of population is represented proportionally?

a) Systematic b) Stratified

c) Cluster d) Quota

Answer: b

☞ Stratified guarantees proportional subgroup representation.Population is divided into homogeneous subgroups (strata), and random samples are taken from each. This ensures representation of subgroups like gender, income levels.

102. Cluster sampling is most useful when:

a) Population is small

b) Population is geographically scattered

c) Subgroup information is easily available

d) Sampling error must be minimized

Answer: b

☞ Clusters reduce cost for widely spread populations.When a population is too large and spread out, it's easier to divide into clusters (e.g., villages, schools) and randomly pick clusters. Saves time & cost, but less precise.

103. Which is a major limitation of systematic sampling?
a) Needs a population frame
b) High cost
c) Periodic patterns may bias sample
d) Requires subgroup info
Answer: c
If the list has a repeating cycle (e.g., every 10th person is from same group), systematic sampling gives biased results.
104. Which is a non-probability method?
a) Multistage b) Stratified
c) Snowball d) Systematic
Answer: c
Snowball relies on participants referring others; it's non-random. Probability methods must give every unit a known chance. Refugees, drug users, HIV patients are hidden populations. Snowball sampling helps reach them via referrals.
105. The biggest advantage of probability sampling over non-probability is:
a) Saves cost
b) Ensures unbiased representation
c) Faster implementation
d) Requires no population frame
Answer: b
106. The difference between population parameter and sample statistic is:
a) Statistic always equals parameter
b) Statistic is unknown, parameter is known
c) Parameter describes population, statistic describes sample
d) Both are same in large samples
Answer: c
107. Which is an example of non-sampling error?
a) Wrong sample size
b) Biased responses
c) Improper random selection
d) Large variance due to small sample
Answer: b
Non-sampling errors occur due to wrong questionnaire design, recording mistakes, non-response bias, even in census. Not related to sample size.
108. Sampling error can be reduced by:
a) Increasing sample size
b) Using non-probability method
c) Reducing population
d) Ignoring extreme values

Answer: a

Larger samples better approximate population, lowering variability. Example: Averages stabilize as sample size increases.

Even in full census, mistakes in data entry, non-response, or interviewer bias may happen.

Sampling error = 0 in census, but non-sampling error remains.

109. In a 3-year moving average, the value of year t is:

a) Average of year t only

b) Average of years (t, t+1, t+2)

c) Average of (t-1, t, t+1)

d) None

Answer: c

🔎 Explanation

- A **moving average** smooths out fluctuations by averaging consecutive values.
- For a **3-year moving average**:
  - We take **three consecutive years**, centered at year *t*.
  - The formula is:

$$MA_t = \frac{Y_{t-1} + Y_t + Y_{t+1}}{3}$$

- Example: Suppose sales data (in units):
  - 2018: 100, 2019: 120, 2020: 150, 2021: 180, 2022: 210
  - To compute moving average for **2020**:

$$MA_{2020} = \frac{120 + 150 + 180}{3} = \frac{450}{3} = 150$$

- This smooths out random ups and downs, giving a clearer **trend**.

👉 In short, **3-year moving average uses the year before, the year itself, and the year after.**

110. A drawback of moving average method is:

a) Easy to compute

b) Removes short-term fluctuations

c) Loses some original data at ends

d) Suitable for all series

Ans. C

🔎 **Explanation**

- When calculating moving averages, you need values **before and after** the year in question.
- So, at the **start and end** of the data, you cannot compute the moving average.
  - Example: With a **3-year moving average**:
    - For the **first year (2018)**, you don't have (t–1).
    - For the **last year (2022)**, you don't have (t+1).
    - So, you cannot calculate MA for those boundary years.
- This means:
  1. **Some data points are "lost."**
  2. **Original values are replaced with smoothed values**, so short-term variations are hidden.
  3. It **lags behind turning points** (trend reversals may be detected late).

👉 Thus, while moving average is **good for trend detection**, it sacrifices **accuracy at ends** and hides **recent changes.**

111. Which time series component is most difficult to measure?

a) Trend b) Seasonal variation

c) Cyclical variation d) Random variation

Answer: d (Random variation)

☞ Trend and seasonal can be measured by methods. Cyclical requires long data. Random is unpredictable, hence hardest.

112. A researcher uses census tracts as primary sampling units and randomly selects households within them. This is —

(একজন গবেষক Census Tract-কে প্রাথমিক স্যাম্পল ইউনিট হিসেবে নিয়ে তার মধ্যে বাড়ি র‍্যান্ডমভাবে বেছে নেয়। এটি কী?)

ক) Two-stage sampling (দুই-স্তর স্যাম্পলিং)

খ) Stratified sampling (স্তরীভূত স্যাম্পলিং)

গ) Quota sampling (কোটাভিত্তিক স্যাম্পলিং)

ঘ) Snowball sampling (স্নোবল স্যাম্পলিং)

Ans. A

Firstly, the researcher choose cluster (tract), then sample units inside cluster (households). This type of sampling is called two stage sampling.

113. In which situation is stratified sampling more advantageous than simple random sampling?

(Stratified Sampling কখন Simple Random Sampling-এর চেয়ে বেশি উপকারী?)

ক) When the population is perfectly homogeneous (যখন জনসংখ্যা পুরোপুরি সমজাতীয়)

খ) when the population can be divided into distinct, internally homogeneous but mutually heterogeneous subgroups

গ) When there is no sampling frame (যখন স্যাম্পলিং ফ্রেম নেই)

ঘ) When time and cost are unlimited (যখন সময় ও খরচ সীমাহীন)

Ans. B

Stratified sampling is useful when the population can be divided into distinct, internally homogeneous but mutually heterogeneous subgroups called strata (like gender, age groups, income levels).

Sampling is done separately from each stratum to improve precision and ensure representation.

If the population is perfectly homogeneous (A), SRS is sufficient. Lack of sampling frame (C) blocks both methods.

Time/cost (D) is not the key consideration.

114. In which of the following situations is cluster sampling more appropriate and efficient than stratified sampling or simple random sampling?

(নিচের কোন ক্ষেত্রে ক্লাস্টার স্যাম্পলিং Stratified বা Simple Random Sampling-এর তুলনায় বেশি কার্যকর?)

ক) When the population is highly heterogeneous and geographically dispersed, but elements within groups are similar (যখন জনসংখ্যা ভৌগোলিকভাবে ছড়ানো এবং অত্যন্ত বৈচিত্র্যময়, কিন্তু গ্রুপের ভেতরে মিল রয়েছে)

খ) When the population is heterogeneous and geographically dispersed, but elements within groups are also heterogeneous

গ) When the objective is to increase precision by reducing sampling error (যখন লক্ষ্য বেশি নির্ভুলতা বাড়ানো)

ঘ) When the cost of listing and contacting each unit in the population is low (যখন প্রতিটি ইউনিট তালিকাভুক্ত করা ও যোগাযোগের খরচ কম)

Ans. B

Option A is Stratified sampling.

Cluster sampling is most efficient when-

The population is geographically scattered, making it costly and impractical to list or reach every unit individually.

There exist naturally occurring clusters (like villages, schools, hospitals) where units within a cluster are similar, but clusters differ from each other.

Option B correctly reflects this situation and captures the practical motivation: cluster sampling minimizes cost and logistical complexity by sampling groups (clusters) rather than every unit directly.

Option C is more aligned with stratified sampling, which increases precision by controlling variability within strata.

Option D favors SRS, since low cost of contacting units removes the need for clustering.

Cluster = heterogeneous inside (mini population), Stratified = homogeneous inside.

115. Which of the following sampling methods is most likely to introduce selection bias because the choice of units depends on the researcher's judgment or convenience?

(নিচের কোন স্যাম্পলিং পদ্ধতিতে সবচেয়ে বেশি Selection Bias হওয়ার সম্ভাবনা থাকে, কারণ গবেষকের সিদ্ধান্ত বা সুবিধার উপর নির্ভর করে?)

ক) Simple Random Sampling (সাধারণ র‍্যান্ডম স্যাম্পলিং)

খ) Stratified Sampling (স্তরীভূত স্যাম্পলিং)
গ) Cluster Sampling (ক্লাস্টার স্যাম্পলিং)
ঘ) Purposive Sampling (উদ্দেশ্যমূলক স্যাম্পলিং)
Ans. D
Purposive sampling is a type of non-probability sampling where the researcher intentionally selects units based on specific characteristics or their own judgment, believing those units are most representative of the population.
Since selection is subjective and not based on chance, results may suffer from selection bias and are not generalizable to the whole population.

Options A, B, and C are probability sampling methods, where each unit has some known chance of selection — which helps reduce bias.

In contrast, non-probability sampling methods like purposive, convenience, and quota sampling do not rely on randomness, increasing the potential for systematic bias.

116. Which of the following is a non-sampling error?
(নিচের কোনটি Non-sampling Error?)
ক) Selecting an unrepresentative sample (অপ্রতিনিধিত্বমূলক স্যাম্পল নির্বাচন করা)
খ) Wrongly recording survey responses (সার্ভে রেসপন্স ভুলভাবে রেকর্ড করা)
গ) Using multi-stage sampling instead of simple random sampling (Simple Random Sampling-এর বদলে Multi-stage Sampling ব্যবহার করা)
ঘ) Sampling a very small portion of population (জনসংখ্যার খুব ছোট অংশ স্যাম্পল করা)
Ans. B
Non-sampling errors occur regardless of sample size or sampling technique—they relate to defects in data collection process (e.g., data entry error, response bias, measurement mistakes).
Option B clearly reflects that.
Options A, C, and D are flaws related to the sampling design itself, hence are sampling errors.

117. A researcher selects every 10th name from a list of employees to form a sample. This is an example of:
(একজন গবেষক কর্মচারীদের তালিকা থেকে প্রতি ১০ম নাম নির্বাচন করে নমুনা তৈরি করেন। এটি কোনটির উদাহরণ?)
ক) Random sampling (এলোমেলো নমুনা)
খ) Stratified sampling (স্তরবিন্যাস নমুনা)
গ) Systematic sampling (পদ্ধতিগত নমুনা)
ঘ) Cluster sampling (ক্লাস্টার নমুনা)
Ans. C
Systematic sampling selects every k-th element from an ordered list after a random start. Here, every 10th name is chosen.

118. A coin is tossed and a die is rolled. What is the probability of getting a Head and a 6?

(একটি কয়েন টস করা হলো এবং একটি ছক্কা নিক্ষেপ করা হলো। হেড এবং 6 পাওয়ার সম্ভাবনা কত?)

ক) 1/6

খ) 1/12

গ) 1/2

ঘ) 1/36

Ans. B

In case of Tossing a coin: P(Head) = 1/2

In case of Rolling a die: P(6) = 1/6

এখানে, দুটা ঘটনা স্বাধীন।

অর্থাৎ, একটার ঘটার সাথে আরেকটার ঘটার সম্ভাবনা নির্ভর করে না।

Independent event এর ক্ষেত্রে আমরা জানি, $P(A \cap B) = P(A) \times P(B)$

So, Since independent → Multiply: $P(\text{Head} \cap 6)$

$= (1/2) \times (1/6)$

$= 1/12$

119. If two events A and B are independent, then:

(যদি দুটি ঘটনা A এবং B স্বাধীন হয়, তবে:)

ক) $P(A \cap B) = P(A)+P(B)$

খ) $P(A \cap B) = P(A) \cdot P(B)$

গ) $P(A \cup B) = P(A) \cdot P(B)$

ঘ) $P(A \cap B) = 0$

Ans. B

Independence means occurrence of one event does not affect the other, so joint probability is the product of individual probabilities.

Source: Business Statistics, Md. Abdul Aziz.

120. Which of the following is a fundamental property of probability?

(নিচের কোনটি সম্ভাব্যতার মৌলিক বৈশিষ্ট্য?)

ক) $P(S) = 0$ খ) $P(\phi) = 1$

গ) $P(S) = 1$ ঘ) $0 < P(A) < 1$

Ans. C

The probability of the sample space S (sure event) is always 1.

মানে নমুনাক্ষেত্র বা স্যাম্পল স্পেসের যেকোন পয়েন্ট ঘটবেই তার সম্ভাবনা ১.

121. If $P(A)=0.4$ and $P(B)=0.5$, A and B are mutually exclusive, then $P(A \cup B)$ is:

(যদি $P(A) = 0.4$ এবং $P(B) = 0.5$ হয় এবং A ও B পরস্পর বর্জনযোগ্য হয়, তবে $P(A \cup B)$ এর মান কত?)

ক) 0.2 খ) 0.5 গ) 0.7 ঘ) 0.9

Ans. D

For mutually exclusive events: $P(A \cup B) = P(A)+P(B)$

$= 0.4+0.5$

$= 0.9$

122. If events A and B are independent, which is always true?

(যদি A এবং B স্বাধীন ঘটনা হয়, তবে নিচের কোনটি সবসময় সত্য?)

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

খ) $P(A \cap B) = P(A) \times P(B)$

গ) $P(A \cap B) = P(A) + P(B)$

ঘ) None

Ans. B

Independence definition = multiplication rule.

For two independent event->

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

123. The classical definition of probability applies only when:

(ক্লাসিকাল সম্ভাব্যতার সংজ্ঞা কেবল কোন ক্ষেত্রে প্রযোজ্য?)

ক) Events are based on personal judgment (ঘটনাগুলো ব্যক্তিগত মতামতের উপর ভিত্তি করে)

খ) Outcomes are equally likely (ফলাফলগুলো সমসম্ভাব্য হলে)

গ) Events are mutually exclusive (ঘটনাগুলো পরস্পর বর্জনযোগ্য হলে)

ঘ) Experiments cannot be repeated (পরীক্ষা পুনরাবৃত্ত করা যায় না যখন)

Ans. B

124.Which probability approach relies on long-run relative frequencies of events?

(নিচের কোন সম্ভাব্যতার পদ্ধতিতে ঘটনার দীর্ঘমেয়াদী আপেক্ষিক ফ্রিকোয়েন্সির উপর ভিত্তি করা হয়?)

ক) Classical approach (ক্লাসিকাল পদ্ধতি)

খ) Subjective approach (সাবজেক্টিভ পদ্ধতি)

গ) Axiomatic approach (এক্সিওম্যাটিক পদ্ধতি)

ঘ) Empirical approach (এম্পিরিকাল পদ্ধতি)

Ans. D

The empirical approach defines probability as the relative frequency of an event in a large number of trials.

125.Which of the following best explains the use of Bayes' theorem?

(নিচের কোনটি বেইজ উপপাদ্যের ব্যবহার সবচেয়ে ভালোভাবে ব্যাখ্যা করে?)

ক) It helps to compute the probability of independent events (এটি স্বাধীন ঘটনার সম্ভাবনা নির্ণয়ে সাহায্য করে)

খ) It updates prior probabilities when new information is available (নতুন তথ্য পাওয়া গেলে পূর্ব সম্ভাবনা আপডেট করতে এটি ব্যবহৃত হয়)

গ) It calculates probability only when all outcomes are equally likely (এটি কেবল তখনই প্রযোজ্য যখন সব ফলাফল সমসম্ভাব্য)

ঘ) It applies only when two events are mutually exclusive. (এটি কেবল তখনই প্রযোজ্য যখন দুটি ঘটনা পরস্পর বর্জনযোগ্য)

Ans. B

Bayes' theorem is used to revise prior beliefs (prior probabilities) into posterior probabilities when new evidence is observed.

126. Which of the following is NOT a valid probability axiom?

(নিচের কোনটি সম্ভাব্যতার একটি বৈধ স্বতঃসিদ্ধ নয়?)

ক) $0 \leq P(A) \leq 1$

খ) $P(S) = 1$, where S is the sample space

গ) For disjoint events A and B, $P(A \cup B) = P(A) + P(B)$

ঘ) $P(A) = -P(Ac)$

Ans. D

Probability is always non-negative; this is not a valid axiom.

127. Which of the following is a false assumption about independent events? (নিচের কোনটি স্বাধীন ঘটনার ক্ষেত্রে একটি মিথ্যা ধারণা?)

ক) If two events are independent, knowledge of one does not change the probability of the other (যদি দুটি ঘটনা স্বাধীন হয়, তাহলে একটি ঘটনার ঘটার তথ্য অন্য ঘটনার সম্ভাবনাকে পরিবর্তন করে না।)

খ) Two events with nonzero probabilities can be both independent and mutually exclusive (দুটি ইভেন্ট যার সম্ভাবনা শূন্য নয়, তা একসাথে স্বাধীন এবং পরস্পর সংঘর্ষহীন (mutually exclusive) হতে পারে।)

গ) Independence implies $P(A \cap B)=P(A).P(B)$ (স্বাধীন ঘটনা হলে $P(A \cap B)=P(A).P(B)$)

ঘ) Dependence means occurrence of one affects the probability of the other (নির্ভরতা (Dependence) মানে হলো একটি ঘটনার হওয়া অন্য ঘটনার সম্ভাবনাকে প্রভাবিত করে।)

Ans. B

Suppose A and B are mutually exclusive: Then, $P(A \cap B)=0$

But if they were also independent, then:

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

so, একই সাথে mutually exclusive ও independent হলে,

$P(A) \cdot P(B)= 0$ হতে হবে। কিন্তু,

If both P(A), P(B) are greater than 0, then their product is greater than 0, which contradicts mutual exclusivity.

That means at least one of them must be 0. অর্থাৎ সহজ কথায়, দুটা অশূন্য সম্ভাবনা বিশিষ্ট ঘটনা একই সাথে mutually exclusive ও independent হতে পারে না।

128. Which condition defines independent events? (কোন্ শর্ত স্বাধীন ঘটনা নির্দেশ করে?)

ক) $P(A \cup B) = P(A) + P(B)$

খ) $P(A|B) = P(A)$

গ) $P(A|B) = 0$

ঘ) $P(A \cap B) = P(A) + P(B)$

Ans. B

Explanation:

Independence means occurrence of one does not change probability of the other:

If, $P(A|B) = P(A)$

129. Events A1,A2,...,An are said to be collectively exhaustive when: (A1,A2,...,An ঘটনাগুলোকে একত্রে পূর্ণ ঘটনা বলা হবে যখন)

ক) They are mutually exclusive (তারা নিশ্ছেদ)

খ) Their union covers the whole sample space (তাদের ইউনিয়ন সমগ্র নমুনাক্ষেত্রকে কাভার করে)

গ) Their intersection is equal to the sample space (তাদের ছেদ তাদের নমুনা ক্ষেত্রের সমান
ঘ) Each has equal probability (প্রত্যেকটার সম সম্ভাব্যতা আছে)
Ans. B
Explanation:
Collectively exhaustive = at least one of the events must happen (their union = sample space).
So, If A and B are exhaustive event, then AUB= Total Sample space

130. A factory produces bulbs with probability of defect p=0.05. Out of 20 bulbs, what is the probability of finding exactly 2 defective bulbs?
(একটি কারখানা ২০টি বাল্ব উৎপাদন করে যার মধ্যে খারাপ হওয়ার সম্ভাবনা p=0.05। ঠিক ২টি খারাপ বাল্ব পাওয়ার সম্ভাবনা কত?)
ক) 0.1887
খ) 0.2641
গ) 0.3774
ঘ) 0.4112
Ans. A

$$P(X=2) \doteq \binom{20}{2}(0.05)^2(0.95)^{18} = 190 \times 0.0025 \times 0.397 = 0.1887$$

131. Bivariate data refers to:
(Bivariate data কী বোঝায়?)
ক) Data with only one variable (শুধু একটি চলক যুক্ত তথ্য)
খ) Data with two related variables (দুটি সম্পর্কযুক্ত চলক যুক্ত তথ্য)
গ) Data collected from more than two sources (একাধিক উৎস থেকে সংগৃহীত তথ্য)
ঘ) Data that is qualitative only (শুধু গুণগত তথ্য)
Ans. B
Bivariate data involves the analysis of two variables simultaneously to study the relationship between them.
For example, height and weight of students.
This is different from univariate (one variable) or multivariate (more than two variables).

Suppose a 2×2 contingency table shows:

	Likes Tea	Likes Coffee	Total
Male	40	20	60
Female	30	10	40
Total	70	30	100

132. What is the probability that a randomly selected person is a female who likes tea?

(এলোমেলোভাবে নির্বাচিত একজন ব্যক্তির মহিলা এবং চা পছন্দকারী হওয়ার সম্ভাবনা কত?)

ক) 0.30 খ) 0.40 গ) 0.70 ঘ) 0.10

Ans. A

Number of females who like tea = 30

Total number of people = 100

So, Probability = 30/100=0.30

133. If E(X) = 10 and E(Y) = 15, then E(2X + 3Y) = ?

(যদি E(X)=10 এবং E(Y)=15 হয়, তবে E(2X+3Y) কত?)

ক) 20 খ) 25 গ) 65 ঘ) 70

Ans. C

E(2X+3Y) = 2E(X)+3E(Y)

= 2(10)+3(15) = 20+45 = 65

134. If X is a random variable with Var(X) = 9, then Var(2X + 3) = ?

(যদি X একটি র‍্যান্ডম চলক হয় এবং এর প্রসরণ/ভেদাংক Var(X)=9 হয়, তবে Var(2X+3) কত?)

ক) 9 খ) 12 গ) 18 ঘ) 36

Ans. D

Var(aX+b)=a2.Var(X)

Here, a=2,Var(X)=9

Var(2X+3) = 22.9 = 36

135. The Central Limit Theorem (CLT) states:

(সেন্ট্রাল লিমিট থিয়োরেম বলে:)

ক) Any variable is normally distributed if n is large (যেকোনো চলক n বড় হলে নরমাল বিন্যাসিত হয়)

খ) Distribution of sample means approaches normal as n grows (নমুনা গড়ের বিন্যাস n বাড়লে নরমালের কাছাকাছি হয়)

গ) Variance becomes zero as n grows (n বড় হলে প্রসরণ/ভেদাংক শূন্য হয়ে যায়)

ঘ) Variance becomes 1 as n grows (n বড় হলে প্রসরণ/ভেদাংক 1 হয়)

Ans. B

CLT → sample mean of large samples ≈ normal, regardless of parent distribution.

136. Which scenario best fits a hypergeometric model?

(কোন পরিস্থিতি hypergeometric মডেলের জন্য উপযুক্ত?)

ক) Tossing a fair die repeatedly (একটি fair die বারবার নিক্ষেপ করা)

খ) Drawing 3 red balls from a bag of 10 balls without replacement (১০টি বলের ব্যাগ থেকে প্রতিস্থাপন ছাড়া ৩টি লাল বল তোলা)

গ) Number of customers arriving in an hour (এক ঘণ্টায় গ্রাহক আগমনের সংখ্যা)

ঘ) Time until first success in coin tosses (Coin toss-এ প্রথম সফলতা পর্যন্ত সময়)

Ans. B

Hypergeometric = sampling without replacement.

option a-> Binomial distribution

Option c-> poisson distribution

Option d-> Geometric distrbution

137. If the probability of hitting a target is p=0.2, what is the probability that the first hit occurs on the 4th trial?

(লক্ষ্যে আঘাতের সম্ভাবনা p=০.২ হলে, প্রথম আঘাত ৪র্থ পরীক্ষায় হওয়ার সম্ভাবনা কত?)

ক) 0.0512 খ) 0.0819 গ) 0.1024 ঘ) 0.1280

Ans. C

$$P(X = 4) = (1 - p)^3 p = (0.8)^3(0.2) = 0.1024$$

138. A researcher wants to estimate the mean weight of a fruit within ±2 g with 95% confidence. Population SD = 8 g. Find required sample size.

A) 62 B) 45 C) 76 D) 48

Ans.A

Explanation:

Step 1: Formula

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Step 2: Substitute values

$$n = \left(\frac{1.96 \cdot 8}{2}\right)^2 = (7.84)^2 \approx 61.47$$

Answer: n ≈ 62

Explanation: A sample of 62 fruits ensures the mean estimate is within ±2 g at 95% confidence.

139. A sample of 50 apples has mean weight 152 g. Population standard deviation is 30 g. Construct a 95% CI for the population mean.

A) (140.1,152.5)

B) (141.1,152.5)

C) (143.7,160.3)

D) None

Ans. C

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{50}} \approx 4.24$$

$$CI = \bar{X} \pm 1.96 \cdot SE = 152 \pm 1.96 \cdot 4.24 \approx 152 \pm 8.3$$

$$CI \approx (143.7, 160.3)$$

Explanation:

We are 95% confident the true population mean lies in this interval.

140. Two factories produce light bulbs. Factory A: u_1=1000 hr, sigma$_1$=50 hr, n_1=36,

Factory B: u_2=1020 hr, sigma$_2$=60 hr, n_2=49. Find the standard error of the difference of sample means.

A) 11

B) 22

C) 11.96

D) 8.98

Ans. C

Step 1: Formula

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

Step 2: Substitute values

$$SE = \sqrt{\frac{50^2}{36} + \frac{60^2}{49}} = \sqrt{\frac{2500}{36} + \frac{3600}{49}}$$

$$SE = \sqrt{69.44 + 73.47} = \sqrt{142.91} \approx 11.96$$

Answer: SE ≈ 11.96 hr

Explanation: This measures expected fluctuation in the difference of sample means.

141. Factory A: u_1=10 cm, sigma$_1$=1 cm, n_1=25,
Factory B: u_2=10.5 cm, sigma$_2$=1.2 cm, n_2=36. Find the mean and variance of the distribution for difference of two means.

A) -0.5,0.08

B) -0.3,0.08

C) 0.5,0.08

D) 0.5,-0.08

Ans.A

2. Formula for Difference of Two Means

Let:

- μ_1, μ_2 = population means
- σ_1^2, σ_2^2 = population variances
- n_1, n_2 = sample sizes

Then:

1. **Mean of the sampling distribution:**

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

2. **Variance:**

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Scenario:

- Factory A: $\mu_1 = 10$ cm, $\sigma_1 = 1$ cm, $n_1 = 25$
- Factory B: $\mu_2 = 10.5$ cm, $\sigma_2 = 1.2$ cm, $n_2 = 36$

Step 1: Mean of sampling distribution

$$E(\bar{X}_1 - \bar{X}_2) = 10 - 10.5 = -0.5 \text{ cm}$$

Step 2: Variance of sampling distribution

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{1^2}{25} + \frac{1.2^2}{36} = 0.04 + 0.04 = 0.08$$

Step 3: Standard error

$$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{0.08} \approx 0.283$$

Interpretation:

If we repeatedly sample pens from both factories, the **difference in sample means** will typically vary ±0.283 cm around -0.5 cm.

142. If α = 0.01 is chosen, what does it mean?

A) We allow 1% risk of rejecting H_0 when it is true

B) We allow 1% risk of accepting H_0 when it is false

C) We guarantee 99% probability of accepting H_0

D) We guarantee 99% probability of rejecting H_0

Answer: A

Explanation: α (significance level) = probability of committing Type I error.

143. .The level of significance (α) in hypothesis testing is:

A) Probability of Type II error

B) Maximum risk of rejecting H_0 when it is true

C) Always equal to 0.10

D) Probability that H_1 is false

Answer: B) Maximum risk of rejecting H_0 when it is true

Explanation: α is chosen before the test (commonly 0.05 or 0.01) and controls the probability of Type I error.

144. Which test is used to compare the variances of two populations?

A) Z-test

B) t-test

C) F-test

D) χ^2 test

Answer: C) F-test

Explanation: F-test is used to test the equality of two population variances (basis of ANOVA).

145. If we fail to reject H_0, it means:

A) H_0 is proven true

B) We have insufficient evidence to reject H_0

C) H_1 is true

D) A Type I error has occurred

Answer: B) We have insufficient evidence to reject H_0

Explanation: Failing to reject does not prove H_0 is true; it only means the sample does not provide enough evidence against it.

146. Which test is appropriate for checking the independence of two categorical variables?

A) Z-test B) t-test

C) χ^2 test D) F-test

Answer: C) χ^2 test

Explanation: The chi-square test for independence is used in contingency tables (e.g., gender vs. preference).

147. The p-value represents:

A) The probability of rejecting H_0

B) The smallest level of significance at which H_0 can be rejected

C) The probability that H_0 is true

D) The probability of committing a Type II error

Answer: B) The smallest level of significance at which H_0 can be rejected

Explanation: p-value measures the strength of evidence against H_0. Smaller p-values → stronger evidence. If, P<alpha, reject H_0

148. Which of the following is used when sample size is small ($n < 30$) and σ is unknown?

A) Z-test B) t-test

C) χ^2 test D) F-test

Answer: B) t-test

Explanation: Student's t-test is used for small samples when population variance is unknown.

149. The null hypothesis (H_0) usually states:

A) A difference exists in the population

B) No difference or no effect exists

C) The sample mean is always equal to the population mean

D) A hypothesis that is always rejected

Answer: B) No difference or no effect exists

Explanation: H_0 is a conservative assumption that assumes "no change" or "no effect." The test checks if sample evidence is strong enough to reject it.

150. One-Way ANOVA is used when:

A) We compare two sample means

B) We compare more than two sample means

C) We compare variances of two groups

D) We check correlation between variables

✓ Answer: B

☞ ANOVA is mainly used to compare more than two means at once.

151. In one-way ANOVA with 4 groups and total 20 observations, what are the degrees of freedom (between, within)?

A) (3, 16) B) (4, 15)

C) (3, 20) D) (4, 19)

✓ Answer: A

☞ df between = k−1=4−1=3, df within = N−k=20−4=16.

152. In a two-way ANOVA, interaction effect means:

A) Two factors work independently

B) The effect of one factor depends on the level of another factor

C) Factors are not significant

D) Variances are equal

✓ Answer: B

☞ Interaction = when one factor's effect changes depending on the level of another factor.

153. In a Completely Randomized Design (CRD), treatments are assigned:

A) According to blocks B) Completely at random

C) According to rows and columns D) By controlling two nuisance factors

✓ Answer: B

☞ CRD = simplest design, treatments assigned at random.

154. A Randomized Block Design (RBD) is best used when:

A) All experimental units are homogeneous

B) There are two nuisance factors

C) There is one nuisance factor that must be controlled

D) Treatments must be replicated equally

✓ Answer: C

☞ RBD controls for one source of variation (e.g., soil type, gender).

155. Which design controls two nuisance factors at the same time?

A) CRD B) RBD

C) Latin Square Design D) Factorial Design

✓ Answer: C

☞ LSD handles two nuisance factors (rows & columns).

156. In a factorial design with 3 fertilizers and 2 irrigation levels, how many treatment combinations exist?

A) 5 B) 6 C) 3 D) 2

✓ Answer: B

☞ 3×2=6 combinations.

157. Which of the following statements about ANOVA is true?

A) A significant F-test means all group means are different

B) A significant F-test means at least one mean is different

C) ANOVA can only be used with two groups

D) ANOVA does not require assumptions

✓ Answer: B

☞ ANOVA only tells us at least one mean differs; post-hoc tests find which means differ.

158. In ANOVA, the F-ratio is computed as:

A) Variance within groups ÷ Variance between groups

B) Variance between groups ÷ Variance within groups

C) Mean of all group means ÷ Overall variance

D) Total variance ÷ Between-group variance

Answer: B

Explanation:

The **F-statistic** in ANOVA is the ratio of **between-group variance (treatment effect)** to **within-group variance (error effect)**. A higher F-value indicates stronger evidence that group means differ significantly. In ANOVA, the F-test checks whether the variance explained by group differences (between-group variance) is significantly larger than the variance within groups (error variance). It's essentially testing the overall model fit.

159. Which of the following is **NOT** an assumption of one-way ANOVA?

A) Homogeneity of variances

B) Independence of observations

C) Normality of residuals

D) Equal sample sizes across groups

Answer: D

Explanation:

One-way ANOVA requires **homogeneity of variances, independence, and approximate normality**. Equal sample sizes are not required (though they make the test more robust).

160. The Least Significant Difference (LSD) test in ANOVA is mainly used for:

A) Testing whether ANOVA assumptions are satisfied

B) Comparing variances among groups

C) Performing post-hoc pairwise comparisons of group means

D) Estimating the standard error of the mean

Answer: C

Explanation:

The LSD test is a **post-hoc multiple comparison test** applied after ANOVA finds a significant difference. It helps identify **which specific group means differ**.

161. The main reason for introducing a blocking factor in an ANOVA design is:

A) To reduce the total sample size needed

B) To control for variability due to nuisance factors

C) To increase the number of treatments

D) To relax the assumption of normality

Answer: B

Explanation:

A **blocking factor** is used to account for variability caused by nuisance variables (e.g., location, time, batch). This helps reduce **experimental error** and increases the sensitivity of the test.

162. Why is ANOVA preferred over multiple independent t-tests when comparing several group means?

A) ANOVA gives exact values of mean differences

B) ANOVA avoids inflation of Type I error rate

C) ANOVA requires fewer assumptions than t-test

D) ANOVA always requires equal sample sizes

Answer: B

Explanation:

If multiple t-tests are used, the **probability of committing Type I error increases**. ANOVA controls the **family-wise error rate** by testing all group means simultaneously with a single F-test.

163. The probability of rejecting a true null hypothesis is:

A) Type II error B) Power

C) Type I error D) Test statistic

Answer: C

Explanation: Type I error (α) = rejecting H_0 when it is true.

164. The probability of failing to reject a false null hypothesis is:

A) Type I error B) Type II error

C) Confidence level D) Test power

Answer: B

Explanation: Type II error (β) = not detecting a real effect.

165. The power of a test is mathematically equal to:

A) $1 - \alpha$ B) $1 - \beta$

C) $\alpha + \beta$ D) β

Answer: B

Explanation: Power = 1 – Probability of Type II error = $1 - \beta$.

166. A two-tailed Z-test is appropriate when:

A) Testing directional hypotheses

B) Variances are unequal

C) Interest is in both directions of difference

D) Comparing more than two means

Answer: C

Explanation: Two-tailed tests check for deviations on both sides of mean.

167. When population variance is unknown and sample size is small (<30), which test is more appropriate?

A) Z-test B) t-test

C) ANOVA D) Chi-square test

Answer: B

Explanation: t-test is used when variance is unknown and sample size is small.

168. The pooled t-test assumes:

A) Equal variances in two populations B) Unequal variances

C) Dependent samples D) Non-normality

Answer: A

Explanation: Pooled t-test requires homogeneity of variances.

169. Which test is suitable for comparing two independent groups with unequal variances?

A) Pooled t-test B) Paired t-test

C) Welch's t-test D) Z-test

Answer: C

Explanation: Welch's t-test does not assume equal variances.

170. The paired t-test is most appropriate when:

A) Groups are independent

B) Groups are matched or repeated measures on same subjects

C) Variances are unequal

D) Data are nominal

Answer: B

Explanation: Paired t-test is for dependent/matched pairs.

171. The assumption of ANOVA includes:

A) Independence, normality, equal variances

B) Non-parametric distribution

C) Dependent samples

D) Randomness not required

Answer: A

Explanation: These are the standard assumptions of ANOVA.

172. Which of the following is a discrete random variable?

A) Number of students absent in class

B) Height of students

C) Weight of apples

D) Length of road

Answer: A

Explanation: Counted values (like absent students) are discrete.

173. Randomized Block Design (RBD) is used to:

A) Increase variability
B) Control block-to-block variability
C) Eliminate replication
D) Randomize treatments completely

Answer: B

Explanation: RBD controls nuisance variation by blocking.

174. Latin Square Design (LSD) controls for:

A) Only one source of variation
B) Two sources of variation simultaneously
C) Three sources of variation
D) Random errors

Answer: B

Explanation: LSD controls for two blocking factors.

175. Which test statistic is most sensitive to unequal sample sizes and unequal variances?

A) Z-test
B) Pooled t-test
C) Welch's t-test
D) Paired t-test

Answer: C

Explanation: Welch's test adjusts for unequal sample sizes and variances.

176. The critical region in hypothesis testing represents:

A) Values of test statistic that lead to rejecting H_0
B) Region of acceptance of H_0
C) Always at center of distribution
D) Values only greater than mean

Answer: A

Explanation: Critical region = rejection region.

177. The main advantage of blocking in RBD is:

A) Reducing within-block variation
B) Increasing randomization
C) Eliminating treatments
D) Avoiding replication

Answer: A

Explanation: Blocks reduce error due to nuisance variables.

178. The treatment degrees of freedom in one-way ANOVA with k groups is:

A) k
B) k – 1
C) n – k
D) n – 1

Answer: B

Explanation: Treatment df = k – 1.

179. In a completely randomized design (CRD), treatments are:

A) Assigned randomly to all experimental units
B) Assigned systematically
C) Blocked by factors
D) Restricted by square arrangement

Answer: A

Explanation: CRD relies on full randomization.

180. Which of the following increases the probability of a Type II error?

A) Large sample size B) Small sample size

C) High α level D) Strong effect size

Answer: B

Explanation: Small samples reduce test power, increasing β.

181. Reducing α (significance level) will generally:

A) Reduce Type II error B) Increase Type II error

C) Leave Type II error unaffected D) Eliminate both errors

Answer: B

Explanation: Lower α reduces Type I error but increases β.

182. A one-tailed test is more powerful than a two-tailed test when:

A) Effect can occur in either direction

B) Effect is expected in only one direction

C) Variances are equal

D) Population is normal

Answer: B

Explanation: Power increases if direction is specified.

183. The paired t-test uses differences because:

A) It reduces variability due to paired units

B) It increases degrees of freedom

C) It avoids hypothesis testing

D) It ignores dependence

Answer: A

Explanation: Taking differences cancels subject-specific effects.

184. . Welch's t-test is often preferred over pooled t-test when:

A) Variances are equal

B) Sample sizes are equal

C) Variances are unequal and sample sizes differ

D) Data are paired

Answer: C

Explanation: Welch adjusts df under unequal variance/sample sizes.

185. In experimental designs, replication is used to:

A) Reduce bias of treatment effects

B) Control block effects

C) Estimate experimental error

D) Avoid randomization

Answer: C

Explanation: Replication provides error estimation.

186. The F-distribution is:

A) Symmetric
B) Always positive and skewed right
C) Negative skewed
D) Discrete

Answer: B

Explanation: F is right-skewed and non-negative.

187. In CRD, experimental error tends to be:

A) Larger than in RBD
B) Smaller than in RBD
C) Same as in LSD
D) Eliminated

Answer: A

Explanation: CRD ignores blocking, so error variance can be larger.

188. The test used when comparing two means with dependent samples is:

A) Z-test
B) Pooled t-test
C) Paired t-test
D) Welch test

Answer: C

Explanation: Paired t-test is for matched/dependent samples.

189. Which of the following tests requires homogeneity of variances assumption?

A) Welch's t-test
B) Pooled t-test
C) Paired t-test
D) Non-parametric tests

Answer: B

Explanation: Pooled t-test needs equal variances.

190. In Latin Square Design, the number of rows, columns, and treatments must be:

A) Equal
B) Different
C) Random
D) Proportional only

Answer: A

Explanation: LSD requires equal rows, columns, and treatments.

191. Which of the following is NOT an assumption of CRD?

A) Homogeneity of experimental units
B) Independence of observations
C) Random assignment of treatments
D) Control of two blocking factors

Answer: D

Explanation: Controlling two blocking factors is for LSD, not CRD.

192. The correct sequence of hypothesis testing procedure is:

A) Select α → Collect data → State hypotheses → Compute statistic → Conclude
B) State hypotheses → Select α → Collect data → Compute statistic → Conclude
C) Collect data → State hypotheses → Select α → Compute statistic → Conclude
D) Compute statistic → State hypotheses → Select α → Collect data → Conclude

Answer: B

Explanation: Standard order: Hypotheses → α → Data → Statistic → Decision.

193. In stratified random sampling, compared to simple random sampling (SRS) with the same sample size, the variance of the estimator of the population mean is:

A. Always greater. B. Always equal.

C. Always smaller. D. Smaller if strata are internally homogeneous.

Answer: D.

Explanation: Stratification reduces variance if within-stratum units are homogeneous relative to the population. But if strata are not well-chosen, variance might not improve. Thus the correct subtle answer is "smaller if strata are internally homogeneous."

194. In a multiplicative time series model $Yt=Tt\times St\times Ct\times It$, where Tt=trend, St =seasonal, Ct=cyclical, It=irregular, the seasonal effect is interpreted as:

A. An additive constant repeating each season.

B. A percentage/ratio effect applied to the trend-cycle component.

C. Random noise unrelated to time.

D. The difference between actual and trend values.

Answer: B.

Explanation: In a multiplicative model, seasonal effect is multiplicative — e.g., "sales are 20% higher in December." Additive seasonal effects (constant increases/decreases) occur in an additive model. So B is correct.

195. If A and B are independent events, and C is any event with $P(C)>0$, then in general:

A. A and B remain independent given C.

B. A and B are mutually exclusive given C.

C. A and B may or may not remain independent given C.

D. Independence of A,B is equivalent to independence given any C.

Answer: C.

Explanation: Independence does not automatically hold under conditioning. For example, even if A,BA,BA,B are independent, conditioning on a third event CCC can introduce dependence. Thus the safe statement is that they *may or may not* remain independent.

196. You run a randomized block design (RBD) with $b=4b=4b=4$ blocks and $t=5t=5t=5$ treatments (one observation per treatment in each block). What is the correct residual (error) degrees of freedom?

A. bt−1 B. (b−1)(t−1)

C. t−1

D. b−1

Answer: B

197. Under the classical linear regression assumptions (linearity in parameters, zero conditional mean of errors, homoscedasticity, no perfect multicollinearity), the OLS estimator of coefficients is:

A. The unique unbiased estimator.
B. The BLUE (Best Linear Unbiased Estimator).
C. Maximum likelihood estimator even if errors are non-normal.
D. Biased but consistent.
Answer: B.
Explanation: The Gauss–Markov theorem states OLS is BLUE: among linear unbiased estimators it has minimum variance. It's not necessarily the unique unbiased estimator (A is false in wording). OLS is MLE only if errors are normal (so C is false if non-normal). OLS is unbiased (under assumptions), so D is false.

198. Which scenario requires a *paired* t-test rather than an independent two-sample t-test?
A. Comparing test scores from two randomly selected classes.
B. Measuring blood pressure of patients before and after a single treatment.
C. Comparing heights of men vs women measured independently.
D. Comparing two machines' product dimensions where each machine produced different independent items.
Answer: B.
Explanation: A paired t-test is used when observations are naturally paired/dependent (e.g., before–after on same subjects). A, C, and D describe independent samples.

199. You want to compare means of two independent groups. Which test is most robust for controlling Type I error when the two populations have unequal variances and potentially unequal sample sizes?
A. Two-sample pooled t-test (assumes equal variances).
B. Two-sample Welch t-test (does not assume equal variances).
C. Paired t-test.
D. Mann–Whitney U test (always better).
Answer: B.
Explanation: Welch's t-test adjusts degrees of freedom for unequal variances and unequal sample sizes and maintains nominal Type I error better than the pooled t when variances differ. Pooled t (A) is invalid if variances differ. C is for dependent pairs. D is nonparametric and useful sometimes, but it tests a different hypothesis (shift in distributions) and is not "always better."

200. You draw a random sample of size n=25 from a population with unknown non-normal distribution but finite mean μ and finite variance σ2. Which statement is most appropriate about the sampling distribution of the sample mean $\bar{X}$?
A. Exactly normal with mean μ and variance σ2/n.
B. Approximately normal with mean μ and variance σ2/n (by the CLT).
C. Has the same shape as the population distribution.
D. Approximately normal only if the population is symmetric.
Answer: B.

Explanation: The Central Limit Theorem implies $\bar{X}$ is approximately $N(\mu,\sigma 2/n)$ for sufficiently large n. n=25 is often large enough for many practical populations, but "exactly" normal (A) is false unless the population is normal. C and D are incorrect because CLT does not require symmetry (D is too restrictive) and the sample mean's shape tends to normality even if the original shape differs.

201.

Which scenario requires a *paired* t-test rather than an independent two-sample t-test?

A. Comparing test scores from two randomly selected classes.

B. Measuring blood pressure of patients before and after a single treatment.

C. Comparing heights of men vs women measured independently.

D. Comparing two machines' product dimensions where each machine produced different independent items.

Answer: B.

Explanation: A paired t-test is used when observations are naturally paired/dependent (e.g., before–after on same subjects). A, C, and D describe independent samples.

202. Under the classical linear regression assumptions (linearity in parameters, zero conditional mean of errors, homoscedasticity, no perfect multicollinearity), the OLS estimator of coefficients is:

A. The unique unbiased estimator.

B. The BLUE (Best Linear Unbiased Estimator).

C. Maximum likelihood estimator even if errors are non-normal.

D. Biased but consistent.

Answer: B.

Explanation: The Gauss–Markov theorem states OLS is BLUE: among linear unbiased estimators it has minimum variance. It's not necessarily the unique unbiased estimator (A is false in wording). OLS is MLE only if errors are normal (so C is false if non-normal). OLS is unbiased (under assumptions), so D is false.

203.

The sample autocorrelation function (ACF) at lag 0 is always:

A. Between -1 and +1.

B. Exactly 1.

C. Equal to the variance.

D. Dependent on the process being stationary.

Answer: B.

Explanation: By definition, the correlation of a variable with itself is 1. So sample ACF at lag 0 = 1. It is not just "between -1 and +1." Variance corresponds to autocovariance at lag 0, not autocorrelation.