
Q1. What is an outlier?
A1. An outlier is an observation that lies significantly far from other data points in a dataset. Outliers can occur due to natural variability, measurement errors, data entry mistakes, or exceptional events.

- **Impact:** Outliers can skew mean, variance, correlation, and regression models, leading to incorrect conclusions.
- **Detection Methods:**
    - **Visualization:** Boxplots, scatterplots, histograms.
    - **Statistical Methods:**
        - **Z-score:** Observations with $|Z| > 3$ may be outliers.
        - **IQR Rule:** Values below Q1 − 1.5×IQR or above Q3 + 1.5×IQR.
- **Treatment Approaches:**
    - **Remove:** If clearly erroneous.
    - **Transform:** Log or square root to reduce effect.
    - **Impute/Adjust:** Replace with mean/median or cap extreme values.

Example: In a salary dataset, most salaries range from 30k–80k, but a CEO earns 500k. This 500k is an outlier that could distort the average.

Q2. What is the measure of central tendency?
A2. Measures of central tendency summarize a dataset by identifying a central or typical value:

- **Mean:** Sum of all values ÷ number of values. Sensitive to outliers.
- **Median:** Middle value when data is sorted. Robust to skewed distributions.
- **Mode:** Most frequently occurring value. Useful for categorical data.

Example: For salaries [30k, 35k, 40k, 500k],

- Mean = 151.25k (skewed by 500k)
- Median = 37.5k (better central measure)

Q3. Significance level vs confidence level
A3.

- **Significance Level (α):** Probability of rejecting the null hypothesis when it is actually true (Type I error). Common values: 0.05, 0.01.
- **Confidence Level (1 − α):** Probability that the calculated confidence interval contains the true population parameter if the study is repeated multiple times.

Example: $\alpha = 0.05 \rightarrow$ 95% confidence that the interval contains the true mean.

Q4. Bias vs variance
A4.

- **Bias:** Error due to assumptions in the model; leads to **underfitting**.
- **Variance:** Error due to sensitivity to training data; leads to **overfitting**.
- **Trade-off:** Reducing bias may increase variance and vice versa. Total error = Bias² + Variance + irreducible error.

Example: Linear regression on nonlinear data → high bias. Complex decision tree → high variance.

Q5. What is the sampling method? List different types of sampling methods.
A5. Sampling selects a subset of a population to make inferences. Types:

1. **Simple Random Sampling:** Every element has equal probability.
2. **Stratified Sampling:** Population divided into strata; sample drawn proportionally from each.
3. **Cluster Sampling:** Randomly select clusters and include all elements.
4. **Systematic Sampling:** Every k-th element selected after random start.
5. **Convenience Sampling:** Non-random, based on ease.

Example: To survey students in a university: stratified sampling by year ensures representation from freshmen to seniors.

Q6. What is the correlation coefficient? Range?
A6. Pearson correlation coefficient (r) quantifies **linear relationship** between two variables.

- **Range:** $-1 \leq r \leq +1$
    - $r \approx +1$ → strong positive correlation
    - $r \approx -1$ → strong negative correlation
    - $r \approx 0$ → no linear correlation

Example: Height vs weight in adults → $r \approx +0.8$ (strong positive correlation).

Q7. What is A/B testing?
A7. A/B testing evaluates two variants of a product/feature:

- **A:** Control (current version)
- **B:** Treatment (new version)
- **Process:** Randomly split users → measure key metric (conversion, CTR) → statistical test (t-test, chi-square) to determine significance.

Example: Test two website button colors. If B increases clicks significantly → deploy B.

Q8. Difference between sample and population
A8.

- **Population:** Entire set of interest (e.g., all customers of a bank).
- **Sample:** Subset used for analysis.
- **Importance:** Samples reduce cost/time; must be representative to avoid bias.

Q9. Difference between Descriptive and Inferential Statistics
A9.

- **Descriptive Statistics:** Summarizes observed data using mean, median, variance, graphs.
- **Inferential Statistics:** Draws conclusions or predictions about a population from a sample (hypothesis tests, CI, regression).

Example: Survey 1000 users (sample) → infer preferences for 1 million users (population).

Q10. Descriptive, Predictive, Prescriptive Analytics
A10.

- **Descriptive:** What happened? Reports, dashboards.
- **Predictive:** What will happen? Forecasts using ML/statistical models.
- **Prescriptive:** What should we do? Optimizations and simulations to guide decisions.

Example: Retail:

- Descriptive → last month's sales report
- Predictive → next month's sales forecast
- Prescriptive → recommended stock reorder quantity

Q11. Handling missing values in a dataset
A11.

1. **Deletion:** Drop rows/columns if missingness is low.
2. **Imputation:** Replace with mean, median, mode, forward/backward fill.
3. **Model-based:** Regression, kNN imputation.
4. **Treat as category:** Encode missing values separately.

Consideration: Depends on missingness type: MCAR, MAR, MNAR.

Q12. Example of root cause analysis
A12.

Example: High manufacturing defect rate → **5 Whys Analysis:**

- Why defects? → Machine miscalibrated
- Why miscalibrated? → No regular maintenance
- Solution: Implement preventive maintenance → reduce defects

Focus: Identify underlying cause, not just symptoms.

Q13. Probability of sum = 5 and 8 with two dice
A13.

- Total outcomes = 6 × 6 = 36
- Sum = 5 → (1,4),(2,3),(3,2),(4,1) → 4 outcomes → 4/36 = 1/9
- Sum = 8 → (2,6),(3,5),(4,4),(5,3),(6,2) → 5 outcomes → 5/36

Q14. Quantitative vs Qualitative Data
A14.

- **Quantitative:** Numeric, measurable.
    - Discrete → counts (e.g., number of cars)
    - Continuous → measurements (e.g., height, weight)
- **Qualitative:** Categorical, descriptive
    - Nominal → unordered (e.g., gender, color)
    - Ordinal → ordered (e.g., rating 1–5)

Q15. Meaning of KPI & examples from personal projects
A15.

- **KPI (Key Performance Indicator):** A measurable metric that reflects progress toward objectives. SMART KPIs are Specific, Measurable, Achievable, Relevant, Time-bound.

Examples:

- **Power BI project:** Dashboard refresh time, user adoption rate
- **ML project:** Accuracy, F1-score, ROC-AUC
- **Agile project:** Sprint velocity, cycle time, defect rate

---