
Q. What is a normal (bell curve) distribution?
Ans: A **normal distribution** is a symmetric, bell-shaped probability distribution where data clusters around the mean.

- **Symmetry:** Mean = Median = Mode.
- **Bell-shape:** High at center, tapers at tails.
- **Parameters:** Mean (μ) = center, Standard deviation (σ) = spread.
- **Empirical rule:** ~68% ($\pm1\sigma$), 95% ($\pm2\sigma$), 99.7% ($\pm3\sigma$).
- **Continuous distribution** (infinite values).

Q. How do you calculate the needed sample size?
Ans: Depends on:

1. **Confidence level (CL):** e.g., 95%.
2. **Margin of error (MOE):** max acceptable difference.
3. **Population variability:** σ (or conservative 0.5 for binary).
4. **Law of large numbers:** $n \geq 30$ ensures reliability.
   → Plug values into formula to compute.

Q. One-tailed vs two-tailed hypothesis testing?
Ans:

- **One-tailed:** Used when effect expected in one direction.
  - H1: $\mu > \mu_0$ or $\mu < \mu_0$.
  - Critical region = one tail.
- **Two-tailed:** Detects any difference.
  - H1: $\mu \neq \mu_0$.
  - Critical region = both tails.

Q. Type I vs Type II errors?
Ans:

- **Type I error (α, false positive):** Reject true null. → Conclude effect exists when it doesn't.
  - Example: Diagnosing healthy person as diseased.
- **Type II error (β, false negative):** Fail to reject false null. → Miss real effect.
  - Example: Failing to diagnose actual disease.

Q. When use t-test vs z-test?
Ans:

- **t-test:** Population σ unknown, sample size small ($n < 30$), random sample, data ~normal.
- **z-test:** Population σ known, large sample ($n \geq 30$), comparing sample mean with population mean.

Q. Difference between F-test and ANOVA?
Ans:

- **F-test:** Compares **variances** of 2+ samples.
  - Two-sample F-test → check homogeneity of variances.
  - Output: F-statistic, p-value.
- **ANOVA:** Compares **means** of 3+ groups.
  - One-way = one factor, Two-way = two factors.
  - Output: F-statistic, p-value → if significant, run post-hoc tests (Tukey, Bonferroni).

Q. What is the effect of skewed data in query processing?
Ans: Skewed data leads to **imbalanced partitions** where some servers process far more data than others. This causes **slow joins, poor parallelism, and longer query times**.

Example: In e-commerce, if 70% of orders come from one city, the server handling that partition gets overloaded.

Q. How do you calculate Z and T scores?
Ans:
- **Z-score:** (Value minus population mean) divided by population standard deviation. Used when population parameters are known.
- **T-score:** (Value minus sample mean) divided by (sample standard deviation divided by square root of sample size). Used when population standard deviation is unknown or sample size is small.
  Example: Standardizing exam scores to compare across different tests.

Q. How does statistical hypothesis testing work?
Ans:
1. State null hypothesis (no effect) and alternative hypothesis (some effect).
2. Choose significance level (like 5%).
3. Calculate test statistic from sample.
4. Compare with critical value or p-value.
5. Reject or fail to reject null hypothesis.
   Example: Testing if a new ad campaign increases sales compared to the old one.

Q. How do you visualize multidimensional data?
Ans: Use scatterplot matrices, heatmaps, PCA or t-SNE projections, parallel coordinates, or 3D plots.
Example: In finance, PCA is used to reduce many stock features to 2D for clustering.

Q. What is the difference between DataFrame and RDD?
Ans:
- **RDD:** Low-level distributed collection, less optimized.
- **DataFrame:** Schema-based, Catalyst-optimized, allows SQL-like queries, faster for structured data.
  Example: Analyzing transactions → DataFrame is preferred for SQL-style queries.

Q. How do you find mean, median, and mode of a gradient distribution?
Ans:
- **Mean:** Sum of all gradient values divided by total count.
- **Median:** Middle gradient value when sorted.
- **Mode:** Most frequent gradient value.
  Example: Used in neural networks to summarize weight updates during training.

Q. What is the expected value of a binomial distribution?
Ans: Number of trials multiplied by probability of success.
Example: In 10 coin flips with probability 0.5, expected heads = 10 × 0.5 = 5.

Q. Describe normal and lognormal distribution.
Ans:
- **Normal:** Symmetric bell curve; mean, median, and mode are equal. *Example:* Heights of people.
- **Lognormal:** Distribution of a variable whose logarithm is normal; skewed right, only positive values. *Example:* Stock prices or income.

Q. Why is MSE used as loss function in Linear Regression?
Ans: Mean Squared Error penalizes large errors more, is smooth for optimization, and has a closed-form least squares solution.
Example: Predicting house prices where larger errors should hurt more.

Q. What is correlation coefficient, and what should its value be?

Ans: Correlation coefficient measures strength and direction of linear relationship, ranging from negative one to positive one.

- Value near zero → weak relation.
- Value near ±1 → strong relation.

Example: Height and weight usually have positive correlation.

Q. How do you do A/B testing?

Ans: Split users into control group A and treatment group B, measure performance metric, run hypothesis test to check significance.

Example: Comparing old vs. new website design click-through rate.

Q. Some packages give a% broken, others b% broken. How do you decide?

Ans: Perform proportion test (z-test for proportions) or compare confidence intervals to check if difference is statistically significant.

Example: Deciding between two suppliers based on defective rate.

Q. What does beta tell us about systematic risk?

Ans: Beta measures stock volatility relative to the market.

- Beta equals one → same risk as market.
- Greater than one → more volatile.
- Less than one → less volatile.

Example: Tech stocks usually have beta > 1, utilities often < 1.

Q. What is F1 score, and how is it calculated?

Ans: F1 is harmonic mean of precision and recall. Formula=2* (precision times recall) divided /(precision + recall).

Example: Used in spam detection to balance false positives and false negatives.

Q. Can we model nonlinear relationships with linear regression?

Ans: Yes, by adding polynomial terms, logarithmic or exponential transformations, and interaction effects while keeping coefficients linear.

Example: Modeling price vs. square footage squared in real estate.

Q. When do you use mean vs median?

Ans:

- **Mean:** For symmetric distributions without outliers.
- **Median:** For skewed data or with extreme values.

Example: Median income better reflects population than mean income, since a few billionaires distort the average.

Q. What happens if sampling distribution is altered (e.g., remove values below mean)?

Ans: Distribution becomes biased and skewed, violating assumptions → increases chance of false positives (Type I error).

Example: Ignoring low-spending customers inflates average purchase estimates.

Q. Difference between linear regression and t-test?

Ans:

- **Linear regression:** Estimates relationship between dependent and independent variables.
- **T-test:** Compares means of groups (special case of regression with categorical predictor).

Example: Regression predicts sales from ad spend; t-test compares average sales between two ad campaigns.

Q. Explain Central Limit Theorem (CLT).

Ans: With large enough sample size, distribution of sample means approaches normal distribution, regardless of population distribution, with mean equal to population mean and standard error equal to population standard deviation divided by square root of sample size.
Example: Average daily sales across 100 stores will be approximately normal even if individual store sales are skewed.

Q. Explain right-skewed and left-skewed distributions.
Ans:

- **Right-skewed:** Long tail on right, mean greater than median. *Example:* Income distribution.
- **Left-skewed:** Long tail on left, mean less than median. *Example:* Age at retirement.
- **Normal:** Symmetric with equal mean and median.

Q. What is Linear Discriminant Analysis (LDA)?
Ans: LDA is a supervised method for classification and dimensionality reduction; it projects data onto new axes that maximize class separation.
Example: Used in face recognition for separating classes by features.

Q. Techniques to assess multicollinearity?
Ans: Correlation matrix, Variance Inflation Factor (VIF), condition index, and eigenvalue analysis.
Example: Checking if "years of education" and "age" are too correlated in a salary regression model.

Q. Methods to detect outliers in a dataset?
Ans:

- **Statistical:** Z-scores, Interquartile Range rule.
- **Visual:** Boxplots, scatterplots.
- **Model-based:** Isolation forest, DBSCAN, Mahalanobis distance.
  Example: Detecting fraudulent transactions that deviate strongly from normal spending patterns.

Q. Difference between standard normal distribution (Z-distribution) and Student's t-distribution?
Ans:

- **Z-distribution:** Mean = 0, standard deviation = 1. Used when population standard deviation is known or sample size is large.
- **t-distribution:** Looks similar but has heavier tails to account for extra uncertainty when standard deviation is estimated from the sample. Approaches Z-distribution as sample size increases.
  Example: Testing average height of 1,000 students (use Z); testing with only 20 students and unknown σ (use t).

Q. How can normal distribution approximate binomial probabilities?
Ans: For large number of trials (n) and probability not near 0 or 1, a binomial distribution can be approximated by a normal distribution with mean = $n \times p$ and standard deviation = square root of $[n \times p \times (1 - p)]$. A continuity correction (± 0.5) is applied.
Example: In 100 coin flips with $p=0.5$, probability of 45–55 heads can be estimated using normal approximation.

Q. What is the standard error of the mean (SEM)?
Ans: SEM = population standard deviation divided by square root of sample size. It shows how much the sample mean is expected to vary from the true population mean.
Example: If $\sigma=10$ and $n=100$, SEM = $10/10 = 1$.

Q. What best describes the central tendency of a dataset?

Ans: Central tendency is summarized using:

- **Mean:** Arithmetic average.
- **Median:** Middle value in ordered data.
- **Mode:** Most frequent value.
  Choice depends on data distribution.
  Example: For incomes, the median is better since a few billionaires distort the mean.

Q. What is Maximum Likelihood Estimation (MLE)?

Ans: MLE finds parameter values that maximize the likelihood of observing the given data. It is widely used for estimating parameters in probability distributions and machine learning models.
Example: Estimating the mean of normally distributed exam scores by choosing the mean that makes observed scores most probable.

Q. What statistical methods compare two populations/groups?

Ans:

- **t-test:** Compares means.
- **z-test:** Compares proportions.
- **Chi-square test:** Compares categorical frequencies.
- **Mann-Whitney U:** Non-parametric test for medians.
- **ANOVA:** Extends to compare more than two groups.
  Example: Comparing average sales between two marketing campaigns (t-test).

Q. How to plot ROC curve from a confusion matrix?

Ans:

1. Vary classification threshold.
2. For each threshold, calculate True Positive Rate (Recall = TP / [TP+FN]) and False Positive Rate (FP / [FP+TN]).
3. Plot FPR on X-axis and TPR on Y-axis.
4. The curve shows trade-off; Area Under Curve (AUC) closer to 1 indicates better model.
   Example: In fraud detection, ROC curve helps compare logistic regression vs decision tree performance.

Q. Methods to analyze skewed data?

Ans: Use log or square-root transformations, Box-Cox transformation, robust measures like median and IQR, or non-parametric tests that don't assume normality.
Example: Transforming highly skewed income data with log scale.

Q. Formulas for precision and recall?

Ans:

- **Precision:** True Positives / (True Positives + False Positives). It answers "Of all predicted positives, how many were correct?"
- **Recall (Sensitivity):** True Positives /(True Positives + False Negatives). It answers "Of all actual positives, how many did we detect?"
  Example: In spam detection, high recall catches most spam but may lower precision.

Q. Differences between panel data and cross-sectional data?

Ans:

- **Cross-sectional:** Snapshot of many subjects at one point in time.
- **Panel data:** Same subjects followed across multiple time periods, combining cross-sectional and time series.
  Example: Surveying 1,000 people once = cross-sectional; surveying same 1,000 people each year = panel data.

Q. When does coefficient of determination (R^2) take negative values?

Ans: R^2 becomes negative when the model fits worse than simply predicting the mean for all observations, meaning residual errors are larger than total variance.
Example: Using a wrong predictor variable to explain sales can give negative R^2.

Q. What is the bias-variance tradeoff?
Ans:

- **Bias:** Error due to overly simple models that underfit data.
- **Variance:** Error due to overly complex models that overfit training data but fail on new data. Goal is to find balance that minimizes total error.
Example: Linear regression on a curved dataset (high bias); deep tree on small dataset (high variance).

Q. Assumptions of Ordinary Least Squares (OLS)?
Ans:

1. Relationship between variables is linear.
2. Errors are independent.
3. Errors have constant variance (homoscedasticity).
4. No perfect multicollinearity among predictors.
5. Errors are normally distributed (important for hypothesis testing).
Example: Predicting house prices with OLS requires that predictor variables like square footage and number of bedrooms are not perfectly correlated.

Q. What is the effect of skewed data in query processing?
Ans: Skewed data creates imbalanced partitions, which slows down joins and reduces parallel efficiency. In distributed systems, one machine may process far more data than others, becoming a bottleneck. For example, if most orders in a sales dataset come from one country, queries involving region-based joins may slow down.

Q. How do you calculate Z and T scores?
Ans: A Z-score is the difference between a value and the mean, divided by the standard deviation of the population. A T-score is similar but uses the sample standard deviation, making it suitable when population parameters are unknown and sample size is small. For instance, comparing a student's test score to the class average would use a T-score if the class size is only 20.

Q. Why do we use ANOVA instead of multiple t-tests?
Ans: ANOVA (Analysis of Variance) checks whether three or more groups differ significantly by comparing their variances. Using multiple t-tests increases the chance of Type I error (false positives), whereas ANOVA controls this risk. For example, testing whether sales differ across four store branches is better done with ANOVA.

Q. How do you identify outliers in data?
Ans: Outliers are values that deviate significantly from the majority. They can be detected using statistical rules such as values beyond 1.5 times the interquartile range, or Z-scores greater than 3 in absolute value. For example, in employee salary data, a CEO's pay may appear as an outlier.

Q. What is the difference between correlation and causation?
Ans: Correlation measures the strength of association between two variables but does not imply that one causes the other. Causation means one variable directly influences another. For example, ice cream sales and drowning incidents are correlated, but both are caused by hot weather, not each other.

Q. What is p-value in hypothesis testing?
Ans: The p-value is the probability of observing results at least as extreme as the sample data,

assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against the null. For instance, if a new drug yields a p-value of 0.01, it suggests the improvement is unlikely due to chance.

Q. What is the difference between Type I and Type II error?

Ans: Type I error occurs when we reject a true null hypothesis (false alarm), while Type II error occurs when we fail to reject a false null hypothesis (missed detection). For example, wrongly approving a defective batch of medicines is a Type II error, while wrongly rejecting a safe batch is Type I.

Q. What is multicollinearity and why is it a problem in regression?

Ans: Multicollinearity occurs when independent variables in regression are highly correlated, making it difficult to separate their effects. This leads to unstable coefficient estimates and reduced interpretability. For example, predicting house price with both “size in square feet” and “number of rooms” may create multicollinearity.

Q. How do you handle missing values in data?

Ans: Missing values can be handled by deleting rows, imputing with mean/median/mode, using predictive models, or applying domain-specific rules. The choice depends on data size and importance of the variable. For instance, missing customer ages in survey data might be filled using the median age.

Q. What is the difference between population and sample?

Ans: A population is the entire group we are studying, while a sample is a subset of that group used for analysis. Sampling is used because studying the full population is often impractical. For example, instead of surveying all citizens of a country, polling agencies survey a few thousand people.

Q. What is the Law of Large Numbers?

Ans: The Law of Large Numbers states that as sample size increases, the sample mean gets closer to the population mean. It explains why larger samples provide more accurate estimates. For example, flipping a coin 10 times may give 70% heads, but 10,000 flips will get close to 50%.

Q. What is heteroscedasticity in regression?

Ans: Heteroscedasticity means that the variance of residuals is not constant across all levels of an independent variable. It violates regression assumptions and can bias inference. For instance, predicting household spending may show higher error variance for wealthy families compared to low-income ones.

Q. What is cross-validation and why is it important?

Ans: Cross-validation splits data into training and testing parts multiple times to ensure model performance is not dependent on one split. It provides a reliable estimate of generalization ability. For example, in fraud detection, cross-validation helps confirm that accuracy is not just due to one dataset partition.

Q. What is overfitting and underfitting in models?

Ans: Overfitting occurs when a model learns noise along with patterns, performing well on training data but poorly on new data. Underfitting occurs when a model is too simple to capture relationships. For example, a complex decision tree that memorizes training transactions may overfit fraud detection data.

Q. What is the difference between parametric and non-parametric tests?

Ans: Parametric tests assume data follows a known distribution (like normal), while non-parametric tests do not rely on such assumptions. Parametric tests are more powerful if

assumptions hold. For example, a t-test is parametric, while a Mann-Whitney test is non-parametric.

Q. What is standard error and how is it different from standard deviation?

Ans: Standard deviation measures variability in individual data points, while standard error measures variability in sample means. Standard error decreases with larger samples. For instance, individual student scores vary a lot (SD), but the average score across many classrooms varies less (SE).

Q. What is the difference between descriptive and inferential statistics?

Ans: Descriptive statistics summarize data using measures like mean and variance, while inferential statistics use samples to draw conclusions about populations. For example, average customer age is descriptive, but predicting future customer age trends is inferential.

Q. What is a confidence interval?

Ans: A confidence interval is a range around a sample estimate that likely contains the true population parameter, with a specified confidence level (like 95%). It reflects uncertainty in estimates. For instance, estimating average delivery time as 3–5 days with 95% confidence means the true average is very likely within that range.

Q. What is hypothesis testing?

Ans: Hypothesis testing is a statistical method to decide if there is enough evidence to reject a null hypothesis in favor of an alternative. It relies on p-values or test statistics. For example, a company may test if a new ad campaign increases sales compared to the old one.

Q. What is the difference between one-tailed and two-tailed tests?

Ans: A one-tailed test checks for effect in only one direction, while a two-tailed test checks for effect in both directions. Choice depends on the research question. For example, testing if a new fertilizer increases crop yield is one-tailed, but testing if it changes yield in any way is two-tailed.

Q. What is the difference between variance and standard deviation?

Ans: Variance measures average squared deviation from the mean, while standard deviation is the square root of variance, keeping the unit same as data. Standard deviation is easier to interpret. For instance, in exam scores, a higher standard deviation means more variation in student performance.

Q. What is the difference between a normal distribution and a skewed distribution?

Ans: Normal distribution is symmetric around the mean, while skewed distributions lean to one side. Right skew has a long tail on the right, left skew on the left. For example, salaries in a company are usually right-skewed because a few employees earn much more than the majority.

Q. What is the difference between probability and odds?

Ans: Probability is the chance of an event happening out of all possibilities, while odds compare the chance of success to failure. For instance, if a team has a 0.25 probability of winning, its odds are 1 to 3.

Q. What is the difference between discrete and continuous variables?

Ans: Discrete variables take countable values (like number of children), while continuous variables can take any value within a range (like height). For example, temperature is continuous, while number of cars in a parking lot is discrete.

Q. What is sampling bias?

Ans: Sampling bias occurs when some members of the population are more likely to be

included in the sample than others, making the sample unrepresentative. For example, surveying only urban customers about product preferences may misrepresent rural views.

Q. What is statistical significance?

Ans: Statistical significance means the observed effect in data is unlikely to have occurred by chance, based on a threshold like 5%. It indicates reliability of findings. For example, if a new drug reduces recovery time with $p < 0.05$, the effect is considered statistically significant.

Q. Customer spend data is highly skewed with a few very large purchases. Which measure of central tendency would you report — mean or median, and why?

Ans: I'd report the **median**, because it's not influenced by extreme purchases the way the mean is.

Explanation to interviewer: In skewed datasets, the mean can give a misleading picture of "typical" spending. For example, if most customers spend ₹500 but a few spend ₹1,00,000, the mean shoots up, but the median still represents the center of most customers' behavior.

Q. Sales data has extreme outliers (holiday season). How would you treat them before reporting?

Ans: I'd either separate seasonal effects, apply log transformation, or winsorize extreme values.

Explanation: Outliers like holiday spikes are real business effects, so instead of blindly removing them, I'd compare regular vs. seasonal sales. This way, the report is accurate for both daily stability and special occasions.

Q. What does standard deviation tell you about sales per store?

Ans: It shows how spread out sales are from the average.

Explanation: If Store A and Store B both average ₹50,000 sales but A has an SD of ₹10,000 and B has ₹50,000, it means A is stable while B is volatile. That informs resource planning.

Q. How would you detect if customer satisfaction scores are normally distributed?

Ans: I'd use histograms, QQ-plots, and normality tests like Shapiro-Wilk.

Explanation: Many statistical tests assume normality. If scores are skewed (say customers mostly rate 4–5 stars), I may use non-parametric tests instead.

Q. What is the probability that it rains on the 4th day given probabilities for previous days?

Ans: If independent, it's just the probability of rain on day 4. If dependent, I'd use Bayes' theorem.

Explanation: In business, independence vs. dependence matters. For example, repeat purchases are not independent — yesterday's purchase affects today's.

Q. Probability that product X is available in warehouse A = 0.6, in B = 0.8. What's the probability it's available on the website?

Ans: Assuming independence, $0.6 + 0.8 - (0.6×0.8) = 0.92$.

Explanation: The website shows "available" if at least one warehouse has stock, so it's the union probability.

Q. Which distribution models the number of customer complaints per week? Why?

Ans: Poisson distribution.

Explanation: Poisson models counts of events in fixed time intervals. Complaints are discrete, random, and occur at some average rate.

Q. Why is the Central Limit Theorem so important in analytics?

Ans: It allows us to assume normality for sample means.

Explanation: Even if spending isn't normally distributed, averages across large samples will be. That makes confidence intervals and hypothesis testing valid.

Q. What is the difference between Binomial and Poisson distributions in business use cases?
Ans: Binomial → fixed trials with success/failure (e.g., email clicks). Poisson → count of rare events over time (e.g., server failures).
Explanation: The binomial deals with bounded trials, while Poisson assumes theoretically infinite opportunities.

Q. Explain p-value in the context of credit risk analysis.
Ans: The p-value tells us how likely it is that the observed risk difference happened by chance if there's no real difference.
Explanation: A low p-value (e.g., 0.01) means there's strong evidence that one group is actually riskier. If it's high (e.g., 0.3), we can't conclude a real difference.

Q. You test if Prime customers spend more than non-Prime customers. Which statistical test would you use?
Ans: Two-sample t-test (or Mann-Whitney if not normal).
Explanation: We're comparing averages between two independent groups. If distributions are skewed, I'd use a non-parametric version.

Q. In an A/B test for a new recommendation algorithm, Group B shows 2% higher engagement but p-value = 0.07. What's your conclusion?
Ans: Statistically not significant at 5% level, but potentially business-relevant.
Explanation: It means the observed lift could be due to chance. I'd recommend running a longer test or considering business trade-offs — sometimes a 2% lift is huge financially.

Q. What's the difference between statistical significance and business significance?
Ans: Statistical significance asks "is this real?" Business significance asks "does it matter?"
Explanation: For example, a 0.1% increase in click-through might be statistically significant, but financially meaningless.

Q. If you conduct 100 hypothesis tests at $\alpha = 0.05$, how many false positives do you expect?
Ans: Around 5.
Explanation: At 5% significance level, you allow 1 in 20 tests to be falsely positive.

Q. How would you design a sample to estimate average spend per store?
Ans: Use stratified sampling across store size or region.
Explanation: Random sampling may overrepresent small or large stores. Stratification ensures fairness and accuracy.

Q. What is stratified sampling and when would you use it in a customer survey?
Ans: Split population into subgroups (e.g., age, gender), then sample proportionally.
Explanation: Useful when subgroups behave differently and we want them represented in estimates.

Q. How would you test if two datasets come from the same distribution?
Ans: Use KS test, Chi-square test, or Mann-Whitney.
Explanation: These don't just compare means, but overall distribution shapes.

Q. If margin of error is 3%, how many more samples do you need to reduce it to 0.3%?
Ans: 100× more samples.
Explanation: Margin of error decreases with square root of n. So a 10× reduction requires 100× sample.

Q. We want to understand the impact of discount % on sales. Which regression model would you use and why?

Ans: Linear regression.
Explanation: Sales = continuous dependent variable. Predictor is discount %. It directly shows how each % discount translates to sales lift.

Q. How do you check for multicollinearity in regression models?
Ans: Use VIF or correlation matrix.
Explanation: If discount % and advertising spend are highly correlated, coefficients become unstable. VIF > 10 is a red flag.

Q. How would you measure correlation between time spent watching and churn rate?
Ans: Use point-biserial correlation or logistic regression.
Explanation: Churn is binary, so logistic regression gives probability estimates based on watch time.

Q. What is the difference between linear regression and logistic regression in business analytics?
Ans: Linear predicts continuous outcomes (sales). Logistic predicts probabilities for binary outcomes (churn/no churn).
Explanation: Logistic regression outputs between 0–1, making it ideal for classification.

Q. What does R^2 mean in business forecasting?
Ans: It's the % of variance in sales explained by the model.
Explanation: An R^2 of 0.8 means 80% of fluctuations are explained by predictors. But I'd caution that high $R^2 \neq$ causation.

Q. Customer age data has outliers (age = 150). How would you handle it?
Ans: Validate, cap, or remove.
Explanation: Likely a data entry error. If it's valid (like an error in system ID), I'd investigate source; if not, remove.

Q. Product reviews dataset has 20% missing values in ratings. What would you do?
Ans: If missing randomly, impute with median/regression. If not random, analyze pattern.
Explanation: If unhappy customers skip ratings, imputing may bias results.

Q. What are MCAR, MAR, MNAR in missing data?
Ans:

- MCAR: missing unrelated to anything.
- MAR: missing depends on observed variables.
- MNAR: missing depends on the value itself.

Explanation: MNAR is hardest — e.g., high-income people skipping salary disclosure.

Q. When would you drop missing values instead of imputing?
Ans: If proportion is small (<5%) and random.
Explanation: Safer than introducing artificial data.

Q. How would you design an experiment to test if new thumbnails improve click-through rate?
Ans: Randomly split users into control and treatment groups, compare CTR.
Explanation: Randomization ensures unbiased comparison.

Q. How would you calculate required sample size for an A/B test?
Ans: Use power analysis with expected effect size, α, and power.
Explanation: This avoids underpowered tests that give inconclusive results.

Q. What statistical test would you use if conversion rate is binary? What if it's average order value (continuous)?
Ans: Binary → proportion z-test or Chi-square. Continuous → t-test.
Explanation: Choice depends on data type.

Q. What are common pitfalls in A/B testing (e.g., peeking early, multiple comparisons)?
Ans: Stopping test too soon, testing too many variants, seasonal effects.
Explanation: These inflate false positives and mislead conclusions.

Q. Are we overestimating or underestimating true population credit scores using a fixed cutoff?
Ans: Likely biased because we're truncating distribution.
Explanation: Scores below cutoff are excluded → underestimates variability and mean.

Q. How would you use confidence intervals to report average basket size across stores?
Ans: Report mean with CI, e.g., ₹800 (95% CI: ₹780–₹820).
Explanation: It shows both estimate and uncertainty, not just a point number.

Q. How do you measure customer lifetime value statistically?
Ans: Use discounted cash flow models, combining purchase frequency with survival models.
Explanation: CLV predicts expected revenue per customer, useful for acquisition budgets.

Q. How would you evaluate if personalized recommendations increase retention?
Ans: Run an A/B test comparing retention; supplement with survival analysis.
Explanation: If treatment group stays longer, it's evidence of success.

Q. How would you explain the result of a hypothesis test to a non-technical executive?
Ans: "We're confident the new strategy improves retention, with less than 5% chance this is random."
Explanation: Translate p-values and confidence into plain business terms.

Q. How do you identify whether customer transaction amounts follow a normal distribution?
Answer: Use visuals (**histogram, KDE, QQ-plot**) and tests (**Shapiro–Wilk, Anderson–Darling**); also check **skewness/kurtosis**.
Explain: I first eyeball the shape via histogram/QQ-plot. If points deviate systematically from the QQ line (especially in tails), it's not normal. Then I confirm with Shapiro–Wilk. For heavy skew (common in spend), I try a **log transform** and re-check.

Q. Employee salaries are right-skewed. What measure would you use to describe average salary?
Answer: Median for central tendency; report **IQR** and also provide **mean** separately for finance.
Explain: Right-skew means a few high earners pull up the mean. The median reflects the "typical" employee. I include mean for budgeting and show both to be transparent.

Q. How would you compare variability of sales across two different regions?
Answer: Use **standard deviation** and **coefficient of variation (CV = SD/mean)**; Levene's test for equality of variances.
Explain: If means differ a lot, CV is better because it's scale-free. I'd test variance equality (Levene) before comparing means to choose the right inference method.

Q. How would you explain variance and standard deviation to a stakeholder?
Answer: Variance is the average squared distance from the mean; **SD** is its square root (in original units).

Explain: If SD of weekly sales is ₹10k, most weeks are roughly within ±₹10k of average. Larger SD = more unpredictability → affects staffing/inventory.

Q. Call center averages 10 calls/hour. Probability of >12 calls in an hour?
Answer: Model **Poisson(λ=10)**; compute $P(X>12)=1-\sum_{k=0}^{12} e^{-10}10^k/k!$.
Explain: Calls are counts in fixed time — classic Poisson. I'd give the formula or use software to get the tail probability; also mention using **normal approximation** when λ is large.

Q. Defect rate 1%. Probability at least 1 defective in 200?
Answer: $1-(0.99)^{200}$ ≈ $1-e^{-2}$ ≈ **0.865**.
Explain: "At least one" is **1 − none defective**. The exponential approximation $(1-p)^n \approx e^{-np}$ is handy.

Q. Probability of drawing two red cards consecutively without replacement?
Answer: $(26/52) \times (25/51) = 25/102 \approx 0.245$.
Explain: First red: 26/52. Then reds left: 25/51. Multiply since sequential and dependent.

Q. How would you simulate customer arrivals?
Answer: Use a **Poisson process**: inter-arrival times are **Exponential(λ)**; simulate with $T_i = -\ln(U_i)/\lambda$.
Explain: If arrivals per time are Poisson, the gaps between arrivals are exponential. In Python/R, sample exponential inter-arrivals and cumulate.

Q. What distribution models time until a customer unsubscribes?
Answer: Exponential (memoryless) or **Weibull** (flexible hazard); for covariates, **Cox proportional hazards**.
Explain: If churn risk changes over customer age/tenure, Weibull/Cox are better than exponential.

Q. How would you test if conversion rates differ between two campaigns?
Answer: Two-proportion z-test or **Chi-square**, with **power analysis** beforehand.
Explain: Binary outcome (convert/not). I ensure randomization, adequate sample, and check for multiple testing if many segments.

Q. What statistical test would you use to check if male and female employees have different average salaries?
Answer: Welch's two-sample t-test (unequal variances) + **covariate-adjusted OLS**.
Explain: Start with t-test; then run regression controlling for role, tenure, location to avoid confounding.

Q. Do churn rates differ across three regions?
Answer: Chi-square test of independence on contingency table; or **logistic regression with region dummies**.
Explain: Chi-square tests association; logistic adjusts for other drivers (price, plan, tenure).

Q. Two models have similar accuracy — which is better statistically?
Answer: McNemar's test for paired classifications; **DeLong's test** for AUC; **paired t-test** over **CV folds** on a chosen metric.
Explain: Because predictions are paired on the same cases, use paired tests; I compare across folds to avoid single-split variance.

Q. How would you test whether website load time impacts purchase probability?
Answer: Logistic regression of purchase ~ load time (+ controls), test coefficient; or **A/B** by throttling load time (ethically) / using **causal IV/RDD**.

Explain: Correlation ≠ causation; if randomization is hard, use IV (e.g., exogenous network congestion) to identify causal effect.

Q. How would you design a sampling plan to estimate average daily spend across 5,000 stores?
Answer: Stratified sampling by region/store size; compute **sample size** for target MOE; apply **finite population correction (FPC)**.
Explain: Heterogeneous stores → stratify for precision. I'd oversample small or high-variance strata (Neyman allocation) and weight back.

Q. Customer groups very different in size — which sampling method would you use?
Answer: Stratified sampling with **proportional or Neyman allocation**.
Explain: Ensures small but important groups aren't missed while optimizing precision by variance and cost.

Q. How would you ensure a survey sample represents the population?
Answer: Probability sampling, enforce **coverage**, use **quotas**, and apply **post-stratification/weighting** to match population margins.
Explain: I compare sample demographics to census/CRM benchmarks and reweight to correct drift.

Q. How would you measure non-response bias?
Answer: Compare respondents vs frame on known vars, run **follow-ups/incentives**, model **response propensity**, **reweight**.
Explain: If late responders resemble non-responders, follow-up differences estimate bias.

Q. When would you use random sampling vs stratified sampling in customer surveys?
Answer: Stratified when subgroups differ materially or must be represented; **simple random** when population is homogeneous and cheap.
Explain: Stratification reduces variance and ensures coverage of small but critical segments.

Q. How would you prove advertising causes sales (not just correlated)?
Answer: Randomized geo-experiments, **difference-in-differences**, **instrumental variables**, or **causal forests**; control for seasonality and trends.
Explain: I'd randomize spend across matched geos; if not possible, DiD with untreated controls or IV (e.g., auction shocks) to identify causal lift.

Q. Regression shows a negative coefficient for "loyalty score" on churn. How do you interpret it?
Answer: Higher loyalty score → **lower** churn odds, holding other factors constant.
Explain: In logistic regression, a negative coefficient reduces log-odds. I might convert to **odds ratio** (e.g., OR = e^β) to translate impact.

Q. How would you decide whether to use linear regression or a non-linear model?
Answer: Inspect **residuals**, test **interactions/polynomials/splines**, compare via **cross-validated error** and **business interpretability**.
Explain: If effects bend (saturation), a GAM or tree model may fit better. I pick the simplest model that hits accuracy targets.

Q. Residuals are not normally distributed — what do you do?
Answer: Use **robust/HC standard errors**, **transform the target** (log), consider **GLMs** or **quantile regression**, and **bootstrap** CIs.
Explain: Non-normal residuals mainly affect inference; robust SEs and bootstrapping protect p-values/CIs.

Q. How would you explain multicollinearity to an executive (with example)?

Answer: "Two inputs move together so much that the model can't tell which one drives the result."

Explain: Example: **discount %** and **coupon usage** always rise together; coefficients swing wildly and signs flip with small data changes. Fixes: combine features, drop one, or **regularize (Ridge/Lasso)** to stabilize.

Hypothesis Testing Flowchart (Full Version with Decisions, Choices, and When to Choose)

1. Start

---

2. Define Research Question

- **Decision:** What are you testing?
- **When to choose:** At the very beginning, before any data collection.
- **Choices / How to choose:**
  - Choose a **clear, measurable variable**.
  - Decide the **population** and **effect you expect**.
  - Example: "Does new teaching method improve test scores?"

---

3. Identify Population and Sample

- **Decision:** Who or what will be studied?
- **When to choose:** After defining research question.
- **Choices / How to choose:**
  - Choose a **representative sample** from the population.
  - Consider **sample size** (large enough for power, small enough for feasibility).
  - Decide **sampling method**: random, stratified, cluster, convenience.

---

4. Determine Type of Data

- **Decision:** What is the nature of your data?
- **When to choose:** After collecting or identifying variables.
- **Choices / How to choose:**
  - **Quantitative (numeric)** → continuous or discrete
    - Use for t-tests, ANOVA, correlation, regression.
  - **Qualitative / Categorical** → nominal or ordinal
    - Use for chi-square, Fisher's exact, proportion tests.

---

5. Formulate Hypotheses

- **Decision:** What is H0 and H1?
- **When to choose:** Before analysis.
- **Choices / How to choose:**
  - **Null Hypothesis (H0):** assumes no effect, difference, or relationship.
  - **Alternative Hypothesis (H1):** assumes effect exists.
  - **Directional (one-tailed) or Non-directional (two-tailed)**:
    - Choose one-tailed if prior research predicts **specific direction**.
    - Choose two-tailed if **any difference matters**.

---

6. Choose Significance Level (α)

- **Decision:** How strict is the evidence needed to reject H0?
- **When to choose:** Before conducting test.
- **Choices / How to choose:**
  - Common $\alpha = 0.05$ → 5% risk of Type I error
  - $\alpha = 0.01$ → more conservative, less risk of false positive
  - Choose **smaller α** when stakes are high or false positives are costly.

---

7. Select Type of Test

- **Decision:** Which statistical test fits the data & hypothesis?
- **When to choose:** After data type and hypotheses are known.
- **Choices / How to choose:**

A. Parametric vs Non-parametric

- **Parametric** → Use when:
  - Data is **numeric/continuous**
  - **Normality assumption** holds
  - **Equal variances** (for two-sample tests)
  - Independent observations
- **Non-parametric** → Use when:
  - Data is **ordinal**, skewed, or violates parametric assumptions
  - Sample size is small

B. Specific Test Selection Based on Scenario

Scenario	Test	When to choose
1 sample, numeric	One-sample t-test	Compare sample mean to known value
2 independent samples	Independent t-test	Compare means of two groups
2 paired samples	Paired t-test	Pre-post or matched pairs
More than 2 groups	ANOVA	Compare means across multiple groups
Ordinal / non-normal	Mann-Whitney, Wilcoxon	Compare medians instead of means
Categorical	Chi-square, Fisher exact	Compare proportions or counts
Correlation	Pearson / Spearman	Relationship between variables

---

8. Check Assumptions

- **Decision:** Are assumptions of chosen test valid?
- **When to choose:** Before calculating test statistic.
- **Choices / How to choose:**
  - Normality: use Shapiro-Wilk, Q-Q plot
  - Equal variances: use Levene's test
  - Independence: check study design
  - If assumptions **violate**, choose alternative test or transform data

---

9. Calculate Test Statistic

- **Decision:** Compute statistic for hypothesis testing.
- **When to choose:** Once test type is finalized and data assumptions are valid.
- **Choices / How to choose:**
  - Use formula for chosen test (t, z, chi-square, ANOVA F, Mann-Whitney U).

---

10. Compare Test Statistic / Calculate p-value

- **Decision:** Is the evidence strong enough to reject H0?
- **When to choose:** After test statistic is calculated.
- **Choices / How to choose:**
  - **p-value ≤ α** → reject H0
  - **p-value > α** → fail to reject H0
  - Optional: compare test statistic to **critical value**

---

11. Decision on Hypothesis

- **Decision:** What conclusion to draw?
- **When to choose:** After p-value comparison.
- **Choices / How to choose:**
  - Reject H0 → statistically significant effect

- Fail to reject H0 → no statistically significant effect

---

12. Effect Size and Confidence Interval
- **Decision:** How large is the effect and how precise?
- **When to choose:** After hypothesis decision.
- **Choices / How to choose:**
  - Effect size metrics: Cohen's d, odds ratio, mean difference
  - Confidence interval: 95% or 99%
  - Helps interpret **practical significance**

---

13. Interpret Results
- **Decision:** Is the effect meaningful in real-world context?
- **When to choose:** After calculating effect size & CI.
- **Choices / How to choose:**
  - Statistically significant but small effect → may not be practically relevant
  - Consider context, cost, or impact

---

14. Report Results
- **Decision:** How to summarize findings?
- **When to choose:** After analysis is complete.
- **Choices / How to choose:**
  - Include: test type, assumptions, sample size, test statistic, p-value, confidence interval, effect size, conclusion.

Absolutely! Let's create a comprehensive A/B testing flowchart in text, following the same structure as the hypothesis testing one, including all decisions, choices, and when/how to make them.

A/B Testing Flowchart (Full Version)
1. Start

---

2. Define Objective
- Decision: What do you want to test?
- When to choose: Before designing the experiment.
- Choices / How to choose:
  - Conversion rate, click-through rate, engagement, revenue, or other KPIs.
  - Ensure metric is measurable and meaningful.

---

3. Identify Population
- Decision: Who will see the A/B test?
- When to choose: Before sampling or randomization.
- Choices / How to choose:
  - Define target audience or user segment.
  - Consider inclusion/exclusion criteria.

---

4. Determine Sample Size
- Decision: How many users do you need?
- When to choose: Before running the test.
- Choices / How to choose:
  - Use power analysis to detect a minimum effect size with desired significance (α) and power (1-β).
  - Larger sample → more accurate results, slower test.
  - Smaller sample → faster test, higher risk of Type II error.

---

5. Randomization / Split

- Decision: How to assign users to groups?
- When to choose: Before experiment begins.
- Choices / How to choose:
    - Randomly assign users to Group A (control) and Group B (variant).
    - Ensure equal probability, avoid bias.
    - Optional: stratified randomization for key user segments.

---

6. Design Experiment
- Decision: How long should the test run, and what variations to show?
- When to choose: Before launch.
- Choices / How to choose:
    - Duration: Long enough to capture sufficient conversions, avoid daily/weekly bias.
    - Variation: Only change one element at a time to isolate effect.
    - Ensure consistency in tracking.

---

7. Collect Data
- Decision: Are data collection methods reliable?
- When to choose: During the experiment.
- Choices / How to choose:
    - Track clicks, conversions, engagement accurately.
    - Ensure no data loss, duplication, or misattribution.

---

8. Check Assumptions
- Decision: Are the standard A/B test assumptions valid?
- When to choose: Before statistical analysis.
- Choices / How to choose:
    - Independence: Each user contributes only once.
    - Random assignment: No bias in group allocation.
    - Sufficient sample size for approximation to normal distribution (Central Limit Theorem).
    - If assumptions violated → consider non-parametric tests or blocking techniques.

---

9. Choose Statistical Test
- Decision: Which test is appropriate for the metric?
- When to choose: Before analyzing data.
- Choices / How to choose:
    - Binary outcomes (conversion yes/no) → z-test for proportions, chi-square test.
    - Continuous outcomes (revenue, time) → t-test (independent samples).
    - Non-normal data or small sample → Mann-Whitney U test.
    - Multiple variants (>2) → ANOVA or chi-square for proportions.

---

10. Calculate Test Statistic
- Decision: How to measure difference between A and B?
- When to choose: After selecting the test.
- Choices / How to choose:
    - Compute difference in means or proportions.
    - Calculate standard error and z/t statistic.

---

11. Compare with Critical Value / Calculate p-value
- Decision: Is the difference statistically significant?
- When to choose: After test statistic calculation.
- Choices / How to choose:
    - p-value $\leq \alpha$ → reject null → significant difference.
    - p-value $> \alpha$ → fail to reject null → no significant difference.
    - Optional: use confidence interval to check effect range.

12. Decision on Variant
- Decision: Which variant performs better?
- When to choose: After significance check.
- Choices / How to choose:
    - Variant B better → consider rolling out.
    - No significant difference → retain control or redesign test.
    - Check effect size and business impact before decision.

---

13. Calculate Effect Size and Confidence Interval
- Decision: How meaningful is the difference?
- When to choose: After statistical decision.
- Choices / How to choose:
    - Effect size: difference in conversion rates, mean difference, odds ratio.
    - Confidence interval: 95% or 99% to understand precision.
    - Assess practical significance for business decision.

---

14. Interpret Results
- Decision: Should the change be implemented?
- When to choose: After analysis and effect size evaluation.
- Choices / How to choose:
    - Statistically significant and practically meaningful → implement variant.
    - Not significant or effect too small → do not implement, redesign test.

---

15. Report Results
- Decision: How to document findings?
- When to choose: At the conclusion of test.
- Choices / How to choose:
    - Include: objective, metric, groups, sample size, statistical test, p-value, confidence interval, effect size, conclusion, and recommendation.

---

16. End
