Probability & Statistics — 100 Q&A Interview Guide (Thorough Answers)

Section 1 — Probability Fundamentals (Q1–Q20)

Q1. What is probability?
A1. Probability quantifies uncertainty: it assigns a number between 0 and 1 to an event to express how likely the event is to occur. Formally, under Kolmogorov's axioms, probability P is a function on a sigma-algebra of events such that (1) $P(A) \geq 0$, (2) $P(\Omega) = 1$ for the sample space Ω, and (3) for countable disjoint events A_i, $P(\cup A_i) = \Sigma P(A_i)$. Intuitively, probability can be interpreted as frequency (long-run relative frequency), degree-of-belief (Bayesian), or propensity depending on context.

Q2. What is an event, sample space, and sigma-algebra?
A2. The **sample space (Ω)** is the set of all possible outcomes of an experiment (e.g., $\Omega = \{1,\ldots,6\}$ for a die). An **event** is a subset of Ω (e.g., {2,4,6} = "even roll"). A **sigma-algebra** (σ-algebra) is a collection of events closed under countable unions/complements; it ensures we can assign probabilities consistently even for infinite outcomes. These form the formal structure on which probability measures are defined.

Q3. State the addition rule for probabilities.
A3. For two events A and B, the general addition rule is: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. For mutually exclusive (disjoint) events where $A \cap B = \emptyset$, it reduces to $P(A \cup B) = P(A) + P(B)$. The rule generalizes to any finite number of events with alternating inclusion-exclusion terms to correct for overlaps.

Q4. What is conditional probability and its basic formula?
A4. The conditional probability of A given B ($P(A|B)$) is the probability of A occurring assuming B has occurred: $P(A|B) = P(A \cap B) / P(B)$, provided $P(B) > 0$. It underpins sequential and dependent events and is the foundation for Bayes' theorem and many inference techniques.

Q5. What is Bayes' theorem and give an example of its use.
A5. Bayes' theorem relates $P(A|B)$ to $P(B|A)$: $P(A|B) = [P(B|A)\,P(A)] / P(B)$, where $P(B) = \Sigma_i P(B|A_i)\,P(A_i)$ in a partition. Practically, it updates prior beliefs $P(A)$ after observing data B. Example: in medical testing, A = disease present, B = positive test. Bayes yields the post-test probability of disease, showing how prevalence and false positives affect results.

Q6. Define independence of events.
A6. Events A and B are independent if $P(A \cap B) = P(A)\,P(B)$, equivalently $P(A|B) = P(A)$. Independence implies knowledge of one event gives no probabilistic information about the other. For more than two events, mutual independence requires the property to hold for every subset, not just pairwise.

Q7. What is the law of total probability?
A7. If {B_i} is a partition of the sample space (disjoint, union = Ω), then for any event A: P(A) = Σ_i P(A|B_i) P(B_i). This is used to compute marginal probabilities by averaging conditional probabilities over a partition, and it is central to hierarchical modeling and Bayesian marginal likelihood computations.

Q8. What is combinatorics — permutations and combinations — and when do you use each?
A8. Permutations count ordered arrangements; for n distinct objects taken k at a time: P(n,k) = n!/(n−k)!. **Combinations** count unordered selections: C(n,k) = n!/[k!(n−k)!]. Use permutations when order matters (passwords), combinations when it does not (committee selection). These counts feed into classical probability calculations.

Q9. What is a probability mass function (PMF) and cumulative distribution function (CDF) for discrete variables?
A9. A PMF p(x) = P(X = x) gives probabilities at points for a discrete random variable. The CDF F(x) = P(X ≤ x) accumulates these probabilities: F(x) = Σ_{t ≤ x} p(t). The PMF summarizes point probabilities; the CDF provides ordering and is right-continuous for discrete variables.

Q10. What is a probability density function (PDF) and PDF vs. PMF?
A10. A PDF f(x) for a continuous random variable satisfies P(a ≤ X ≤ b) = ∫*a^b f(x) dx. Unlike PMF values, f(x) is not a probability but a density; probabilities come from integrals. The CDF for continuous X is F(x) =* ∫{−∞}^x f(t) dt. PDFs must be nonnegative and integrate to 1.

Q11. Define expectation (mean) and variance; give formulas.
A11. Expectation E[X] is the weighted average of outcomes: discrete: E[X] = Σ x p(x); continuous: E[X] = ∫ x f(x) dx. Variance Var(X) measures dispersion: Var(X) = E[(X − E[X])^2] = E[X^2] − (E[X])^2. Standard deviation is √Var(X). Expectation is linear: E[aX + b] = a E[X] + b.

Q12. What is covariance and correlation? Interpret them.
A12. Covariance Cov(X,Y) = E[(X−E[X])(Y−E[Y])] measures joint variability; positive covariance means X and Y move together. Correlation ρ(X,Y) = Cov(X,Y)/[σ_X σ_Y] normalizes covariance to [−1,1], facilitating comparison across scales. Correlation = 0 does not imply independence unless variables are jointly normally distributed.

Q13. What is the moment generating function (MGF) and why is it useful?
A13. M_X(t) = E[e^{tX}] is the MGF. If it exists in a neighborhood of t=0, derivatives at zero give moments: M_X^{(k)}(0) = E[X^k]. MGFs uniquely characterize distributions (when they exist), and they simplify derivation of sums of independent random variables because M_{X+Y}(t) = M_X(t) M_Y(t).

Q14. Explain conditional expectation and its properties.
A14. The conditional expectation E[X|Y] is a random variable giving the mean of X given Y. Properties: linearity E[aX+b|Y] = a E[X|Y] + b, tower property E[E[X|Y]] = E[X], and if X independent of Y then E[X|Y] = E[X]. Conditional expectation underlies regression, prediction, and many probabilistic decompositions (e.g., Var(X) = E[Var(X|Y)] + Var(E[X|Y])).

Q15. What is the difference between frequentist and Bayesian probability?
A15. Frequentist probability interprets probability as long-run relative frequency; parameters are fixed unknowns, inference uses sampling distributions. Bayesian probability treats probability as degree-of-belief; parameters are random with prior distributions updated by data via Bayes' theorem to produce posterior distributions. Bayesian methods explicitly quantify parameter uncertainty; frequentist methods rely on estimators and tests based on repeated sampling properties.

Q16. Define convergence in probability and almost sure convergence.
A16. A sequence X_n converges **in probability** to X if, for any ε>0, P(|X_n − X| > ε) → 0 as n→∞. **Almost sure (a.s.) convergence** means P(lim_{n→∞} X_n = X) = 1. Almost sure convergence implies convergence in probability; the converse is not true. These concepts are foundational in asymptotic theory.

Q17. What is the law of large numbers (LLN)?
A17. LLN states that sample averages converge to the expected value as sample size increases. The **weak LLN**: $\bar{X}_n \to \mu$ in probability. The **strong LLN**: $\bar{X}_n \to \mu$ almost surely. LLN justifies using sample means to estimate population means in practice.

Q18. State the central limit theorem (CLT) and why it matters.
A18. For i.i.d. variables X_i with mean μ and variance σ^2, the standardized sum (Σ X_i − nμ)/(σ√n) converges in distribution to a standard normal as n→∞. CLT explains why many sampling distributions are approximately normal even if the underlying data aren't, underpinning confidence intervals and hypothesis tests.

Q19. What is a stochastic process? Give examples.
A19. A stochastic process is a collection {X(t): t∈T} of random variables indexed by time or space. Examples: random walk, Poisson process (counts over time), Brownian motion (continuous-time limit of random walks), and Markov chains. Processes model temporal/spatial randomness in finance, queuing, and physics.

Q20. What is a Poisson process and key properties?
A20. A Poisson process counts occurrences over time with independent increments and stationary increment rates. If events occur at constant rate λ, then N(t) ~ Poisson(λ t). Interarrival times are i.i.d. exponential with mean 1/λ. This process models rare events like call arrivals or decay events.

Section 2 — Discrete Distributions & Tools (Q21–Q40)

Q21. What is the Bernoulli distribution?
A21. Bernoulli(p) models a single binary trial: P(X=1)=p (success), P(X=0)=1−p. E[X]=p, Var(X)=p(1−p). It is the building block for binomial and many binary models in logistic regression.

Q22. Describe the Binomial distribution and its uses.
A22. Binomial(n,p) is the sum of n independent Bernoulli(p) trials: P(X=k)=C(n,k) p^k (1−p)^{n−k}. Expectation np and variance np(1−p). It models counts of successes in fixed trials (A/B testing, quality control). Approximate by Poisson (when n large, p small, np moderate) or by normal (when np and n(1−p) both large) for inference.

Q23. What is the geometric distribution? Two parameterizations and interpretations.
A23. Geometric models #trials until first success. Two parameterizations: support on {1,2,...} (P(X=k)=(1−p)^{k−1} p: counts trials including success) and support on {0,1,2,...} (counts failures before first success). E[X] = 1/p (trials version), Var(X) = (1−p)/p^2. Memoryless property: P(X>m+n | X>m) = P(X>n), unique among discrete distributions.

Q24. What is the Negative Binomial distribution?
A24. Negative binomial models #trials to achieve r successes (or number of failures before r successes). PMF: P(X=k) = C(k+r−1, k) (1−p)^k p^r. E[X] = r(1−p)/p (failure-count form) and Var[X] = r(1−p)/p^2. It generalizes geometric and is useful for overdispersed count data relative to Poisson.

Q25. Explain the Poisson distribution, mean = variance property, and where it's applied.
A25. Poisson(λ) has PMF P(X=k)=e^{−λ} λ^k / k!, mean=Var=λ. It models counts of independent rare events per unit (time/area). The equal mean-variance property can be violated in real data (overdispersion), necessitating negative binomial or mixture models.

Q26. When is the hypergeometric distribution used?
A26. Hypergeometric models draws without replacement from a finite population: population size N with K successes; drawing n items, the count X of successes follows P(X=k)= [C(K,k) C(N−K, n−k)]/C(N,n). Use when sampling without replacement (quality inspection from a lot), where binomial (replacement) approximation may be adequate for large N.

Q27. What is the Multinomial distribution?
A27. Multinomial generalizes binomial to more categories: for n independent trials, each outcome falls into one of k categories with probabilities p_1…p_k, and counts (X_1…X_k) ~ Multinomial(n, p). Marginally X_i ~ Binomial(n, p_i). Covariances are negative between categories: Cov(X_i,X_j) = −n p_i p_j for i≠j.

Q28. How do you compute expectation and variance for discrete distributions?
A28. For discrete X with PMF p(x): E[X] = Σ x p(x), E[g(X)] = Σ g(x) p(x). Var(X) = E[X^2] − (E[X])^2. Use linearity for sums of r.v.s: E[ΣX_i] = ΣE[X_i]. For independent sums, Var(ΣX_i) = ΣVar(X_i); if dependent, include covariances.

Q29. What is overdispersion and how to detect and handle it in count data?
A29. Overdispersion occurs when observed variance of counts exceeds the variance implied by the model (e.g., Var > mean for Poisson). Detect via comparing sample variance to mean or using dispersion tests. Address by using negative binomial, quasi-Poisson, zero-inflated models, or random effects/mixed models.

Q30. What is the probability generating function (PGF) and its use?
A30. PGF G_X(s) = E[s^X] for discrete nonnegative integer-valued X. Derivatives at s=1 give moments: G'_X(1) = E[X], etc. PGFs compactly encode distributions and make computation of sums of independent integer-valued variables easier (product of PGFs).

Q31. Explain the relation between Binomial, Poisson, and Normal approximations.
A31. For Binomial(n,p): when n large and p small with λ = n p moderate, Binomial ≈ Poisson(λ). When n large and both np and n(1−p) are large, a normal approx with continuity correction works: Binomial ≈ Normal(mean=np, var=np(1−p)). Central limit theorem underlies the normal approximation.

Q32. Define the Dirichlet-multinomial and when it is used.
A32. Dirichlet-multinomial is a compound model: categories' probabilities p come from a Dirichlet prior; conditional on p, counts follow multinomial. It yields overdispersion relative to multinomial and is used when trial-to-trial category probabilities vary (e.g., ecological counts).

Q33. How do you test whether two categorical variables are independent?
A33. Use the chi-square test of independence on a contingency table: compute expected counts under independence, Σ (O−E)^2/E approximates χ^2 with (r−1)(c−1) df if counts are large. For small expected counts, use Fisher's exact test. Also consider effect size measures like Cramér's V.

Q34. What is a zero-inflated model?
A34. Zero-inflated models handle excess zeros by mixing a point mass at zero with a count distribution (e.g., Poisson, NB). Structure: with prob π an observation is an "excess zero"; otherwise it follows the count model. Useful for data with more zeros than standard distributions predict (insurance claims, defect counts).

Q35. What is a discrete-time Markov chain?
A35. A sequence X_n where P(X_{n+1}|X_n,…,X_0) = P(X_{n+1}|X_n) — the Markov property. Characterized by a transition probability matrix P with rows summing to 1. Key topics: n-step

transition probabilities, stationary (invariant) distribution π solving πP = π, and classification of states (irreducible, recurrent, transient).

Q36. How do you compute conditional probabilities for discrete joint distributions?
A36. For joint PMF p_{X,Y}(x,y), conditional PMF p_{X|Y}(x|y) = p_{X,Y}(x,y) / p_Y(y) where p_Y(y) = Σ_x p_{X,Y}(x,y). This decomposes joint behavior into conditional and marginal components and is central to Bayesian and hierarchical models.

Q37. What is empirical distribution and empirical CDF?
A37. Given data x_1…x_n, the empirical distribution places mass 1/n at each observed value. The empirical CDF F_n(x) = (1/n) Σ I(x_i ≤ x) estimates the true CDF; by Glivenko–Cantelli theorem, sup_x |F_n(x) − F(x)| → 0 a.s. This underlies nonparametric inference (e.g., bootstrap).

Q38. Describe coupon collector and occupancy problems briefly.
A38. Coupon collector: how many samples until you collect all k types? Expected time ≈ k H_k (harmonic sum). Occupancy: distributing n balls into k boxes; questions include distribution of counts, empty boxes, and collisions. These combinatorial problems have applications in hashing and biodiversity measurements.

Q39. What is the role of indicator variables in probability calculations?
A39. Indicators I_A equal 1 if event A occurs, 0 otherwise. They simplify proofs using linearity: E[I_A]=P(A); Var(I_A)=P(A)(1−P(A)); sums of indicators count occurrences. They are useful to derive expectations and variances in combinatorial contexts.

Q40. How does conditional independence differ from marginal independence?
A40. X and Y are marginally independent if P(X,Y)=P(X)P(Y). They can be conditionally independent given Z if P(X,Y|Z)=P(X|Z)P(Y|Z) for all Z values. Two variables might be dependent marginally but independent conditional on a third (e.g., confounding), so conditional independence is crucial in graphical models and causal inference.

Section 3 — Continuous Distributions & Multivariate (Q41–Q60)

Q41. Explain the Uniform distribution (continuous) and its properties.
A41. Continuous Uniform(a,b) has PDF f(x)=1/(b−a) for x∈[a,b]. E[X]=(a+b)/2, Var(X)=(b−a)^2/12. It represents maximum ignorance subject to range constraints and is used for simulations and as a base for inverse transform sampling.

Q42. Define the Normal distribution and central properties.
A42. Normal(μ,σ^2) has PDF (1/(σ√(2π))) exp(−(x−μ)^2/(2σ^2)). It is fully characterized by mean μ and variance σ^2, is stable under linear combinations of independent normals, and the CLT makes it

ubiquitous for approximations. Z = (X−μ)/σ is standard normal; many inference procedures rely on normality or approximate normality.

Q43. What is the exponential distribution and the memoryless property?
A43. Exponential(λ) has PDF λ e^{−λ x} for x≥0, mean 1/λ, Var 1/λ^2. It is continuous analogue to geometric with the **memoryless** property: P(X>t+s | X>t) = P(X>s). It models waiting times for Poisson processes and is used in reliability and survival analysis.

Q44. Describe the Gamma distribution and its relation to Exponential and Chi-square.
A44. Gamma(α, β) (shape α, rate β) generalizes exponential: sum of α independent exponential(β) variables if α integer. Mean α/β, Var α/β^2. Chi-square(ν) is Gamma(ν/2, 1/2). Gamma is flexible for modeling skewed positive data and arises in Bayesian conjugacy with Poisson counts.

Q45. Explain the Beta distribution and applications.
A45. Beta(α,β) is defined on [0,1] with density proportional to x^{α−1} (1−x)^{β−1}. It models probabilities and proportions; it's conjugate prior for binomial likelihood: posterior Beta(α+successes, β+failures). Mean α/(α+β), variance αβ/[(α+β)^2(α+β+1)].

Q46. What is the Log-normal distribution? When does it appear?
A46. If Y ~ Normal(μ,σ^2), then X = e^Y is log-normal. It models positive-skewed multiplicative phenomena (income, stock prices, biological sizes). E[X] = exp(μ + σ^2/2), Var(X) = exp(2μ + σ^2)(exp(σ^2)−1).

Q47. What is the Student's t-distribution and why is it used?
A47. The t-distribution with ν degrees of freedom arises when estimating a mean with unknown variance: t = ($\bar{X}$ − μ)/(S/√n) follows t_ν under normality. It has heavier tails than normal, protecting against small-sample variance uncertainty. As ν→∞ it approaches the normal distribution.

Q48. Define Chi-square and F distributions and their uses.
A48. Chi-square(ν) is the sum of squares of ν independent standard normals; key in variance inference. F distribution is ratio (U1/ν1)/(U2/ν2) where U1~χ^2_{ν1}, U2~χ^2_{ν2}; used in ANOVA and testing equality of variances. Both are central to classical hypothesis testing frameworks.

Q49. What is joint PDF, marginal, and conditional PDF for continuous variables?
A49. For joint density f_{X,Y}(x,y), the marginal f_X(x)=∫ f_{X,Y}(x,y) dy. Conditional density f_{X|Y}(x|y) = f_{X,Y}(x,y)/f_Y(y) (if f_Y(y)>0). These relationships parallel discrete cases and allow decomposition of multivariate behavior.

Q50. What does it mean for continuous variables to be independent?
A50. X and Y are independent if f_{X,Y}(x,y) = f_X(x) f_Y(y) for all x,y. Independence implies that conditional density equals marginal: f_{X|Y}(x|y) = f_X(x). For joint normal vectors, zero covariance implies independence, but not generally for arbitrary distributions.

Q51. What is a multivariate normal distribution? Key properties.
A51. Multivariate normal (μ, Σ) has density proportional to exp(−½ (x−μ)' Σ^{-1} (x−μ)). Any linear combination of components is normal; marginals and conditionals are normal; mean vector μ and covariance matrix Σ fully characterize it. Useful in multivariate regression, PCA, and inference.

Q52. How do you compute the distribution of a function of random variables (change of variables)?
A52. For monotonic transformations Y = g(X), PDF f_Y(y) = f_X(g^{-1}(y)) |d/dy g^{-1}(y)|. For multivariate transforms, use Jacobian determinant: f_Y(y) = f_X(x) |det(J)|^{-1} where x = g^{-1}(y). This is key in deriving distributions of sums, products, and order statistics.

Q53. What are order statistics and their importance?
A53. Order statistics X_{(1)} ≤ … ≤ X_{(n)} are sorted sample values. They are used for extremes (min/max), quantile estimation, and nonparametric inference. The joint distribution of order stats has known forms; the k-th order stat's pdf relates to Beta distributions when sampling from uniform.

Q54. Explain joint moment generating functions and independence.
A54. Joint MGF M_{X,Y}(s,t)=E[e^{sX + tY}]. If X and Y independent then M_{X,Y}(s,t)=M_X(s) M_Y(t). Joint MGFs encode full joint distribution (if they exist) and facilitate finding distributions of linear combinations and moments through differentiation.

Q55. What is the multivariate central limit theorem?
A55. For i.i.d. random vectors X_i with mean μ and covariance Σ, √n ($\bar{X}$_n − μ) converges in distribution to multivariate normal N(0, Σ). This generalizes CLT to vector-valued data, underpinning asymptotic inference for multivariate estimators.

Q56. What is conditional density for continuous distributions and an example?
A56. Conditional density f_{X|Y}(x|y) = f_{X,Y}(x,y)/f_Y(y). Example: If (X,Y) jointly normal, the conditional distribution X|Y=y is normal with mean μ_X + Σ_{XY} Σ_{YY}^{−1} (y−μ_Y) and reduced variance. Conditional densities are used for prediction and Bayesian updates.

Q57. How do you handle measurement error in continuous variables?
A57. Model measurement error explicitly: observed Z = X + ε, where ε is error. Methods: errors-in-variables regression, SIMEX, instrumental variables, or hierarchical modeling to separate signal from noise. Accounting for measurement error avoids biased parameter estimates.

Q58. What is a Gaussian mixture model (GMM)?
A58. GMM is a weighted sum of Gaussian components: $f(x)=\Sigma\ \pi_k\ N(x|\mu_k, \Sigma_k)$. It models multimodal continuous data, performs soft clustering, and is fitted via EM algorithm (expectation-maximization). Identifiability and choice of K are practical considerations.

Q59. Describe techniques for multivariate density estimation.
A59. Techniques include parametric models (multivariate normals, mixtures), kernel density estimation (KDE) with multivariate kernels (needs careful bandwidth selection), and copulas to model dependence structure separately from marginals. High dimensionality poses challenges (curse of dimensionality).

Q60. What is the Mahalanobis distance and when do you use it?
A60. $D^2 = (x-\mu)' \Sigma^{-1} (x-\mu)$ measures distance accounting for covariance structure. It's scale-invariant and used for outlier detection, classification (e.g., discriminant analysis), and multivariate anomaly detection where Euclidean distance is inadequate.

Section 4 — Statistical Inference & Estimation (Q61–Q80)

Q61. What is a point estimator and desirable properties (bias, consistency, efficiency)?
A61. A point estimator $\hat{\theta}$ estimates parameter θ. **Bias**: $E[\hat{\theta}]-\theta$ (zero ⇒ unbiased). **Consistency**: $\hat{\theta} \to \theta$ (in probability) as $n\to\infty$. **Efficiency**: among unbiased estimators, one with smallest variance is efficient. Mean squared error MSE = Var + Bias^2 trades off bias and variance.

Q62. What is Maximum Likelihood Estimation (MLE)?
A62. MLE chooses $\hat{\theta}$ to maximize the likelihood $L(\theta) = \Pi\ f(x_i|\theta)$. Under regularity, MLE is consistent, asymptotically normal ($\hat{\theta} \sim N(\theta, I(\theta)^{-1}/n)$ where I is Fisher information), and asymptotically efficient. Compute via calculus or numerical optimization; watch for boundary issues and multimodality.

Q63. What is method of moments?
A63. Method of moments equates sample moments (e.g., sample mean, variance) to theoretical moments to solve for parameters. It's simple and often closed-form, but may be less efficient than MLE and produce estimates outside parameter space. Useful for initial estimates or complex likelihoods.

Q64. Define Fisher information and its role.
A64. Fisher information $I(\theta) = E[\ (\partial/\partial\theta \log f(X|\theta))^2\ ]$ measures information data provide about θ. In regular models, $Var(\hat{\theta}_{MLE}) \approx 1/[n\ I(\theta)]$, providing Cramér–Rao lower bound: any unbiased estimator must have $Var \geq 1/[n\ I(\theta)]$. It quantifies estimator precision asymptotically.

Q65. What is the Cramér–Rao lower bound?
A65. The CR bound states $\mathrm{Var}(\hat{\theta}) \geq [I(\theta)]^{-1}/n$ for unbiased estimators, where $I(\theta)$ is Fisher information. If an unbiased estimator reaches the bound, it is efficient. It provides a benchmark for the best possible variance.

Q66. Explain confidence intervals and interpretation.
A66. A 95% confidence interval for θ is an interval computed from data that covers the true θ in 95% of repeated samples. It is a procedure-based statement (random interval, fixed θ), not "probability θ lies in this interval" (Bayesian). Constructed via pivot quantities or asymptotic normality: $\hat{\theta} \pm z_{0.975} SE$.

Q67. What is hypothesis testing and p-value interpretation?
A67. Hypothesis testing evaluates H0 vs H1. A **p-value** is the probability, under H0, of observing data at least as extreme as observed. Small p-value provides evidence against H0. Misinterpretations: p-value ≠ probability H0 true; it depends on sample size and test choice.

Q68. Type I and Type II errors, and statistical power.
A68. Type I error (α): rejecting H0 when true (false positive). **Type II error (β)**: failing to reject H0 when false (false negative). **Power = 1−β** is the probability the test detects a true alternative. Power depends on effect size, variance, α, and sample size; design studies to achieve desired power.

Q69. What is the Likelihood Ratio Test (LRT)?
A69. LRT compares nested models H0: $\theta \in \Theta 0$ vs H1: $\theta \in \Theta$. Statistic $\Lambda = -2 \log [\sup_{\Theta 0} L / \sup_{\Theta} L]$, which under regularity asymptotically follows χ^2 with df = $\dim(\Theta) - \dim(\Theta 0)$. LRT is broadly applicable and forms the basis for many model comparison tests.

Q70. What are nonparametric tests and when are they used?
A70. Nonparametric tests (e.g., Wilcoxon, Mann–Whitney, Kruskal–Wallis) make fewer distributional assumptions and are used when normality fails or data are ordinal. They often test medians/ranks rather than means and are robust but may be less powerful if parametric assumptions hold.

Q71. Explain bootstrap and its applications.
A71. Bootstrap resampling approximates sampling distributions by repeatedly resampling (with replacement) from observed data and recalculating statistics. It yields empirical SEs, bias estimates, and confidence intervals (percentile, BCa) without strong parametric assumptions. It's powerful for complex estimators or small-sample inference, though dependent on data representativeness.

Q72. What is cross-validation and why is it useful?
A72. Cross-validation partitions data into train/test folds (k-fold, leave-one-out) to assess model generalization. It estimates out-of-sample error and aids model selection/tuning (e.g., hyperparameters). It reduces overfitting risk compared to in-sample metrics, but must respect data dependencies (time series, grouped data).

Q73. What is the difference between parametric and nonparametric estimation?
A73. Parametric assumes a specific family (e.g., normal) with finite parameters; estimation targets parameters (MLE). **Nonparametric** makes minimal assumptions about functional form (e.g., KDE, splines); it provides flexible fits but often needs more data and careful smoothing/regularization.

Q74. Describe Bayesian inference and conjugate priors.
A74. Bayesian inference combines a prior $p(\theta)$ with likelihood $p(x|\theta)$ to yield posterior $p(\theta|x) \propto p(x|\theta)\, p(\theta)$. **Conjugate priors** yield posterior in same family as prior (e.g., Beta prior & binomial likelihood → Beta posterior), simplifying closed-form updates and interpretability. Bayesian outputs include posterior means, credible intervals, and predictive distributions.

Q75. What is a credible interval and how does it differ from a confidence interval?
A75. A 95% credible interval from the posterior is an interval $[a,b]$ such that $P(\theta \in [a,b] \mid \text{data}) = 0.95$ — a direct probability statement about θ given data. Unlike frequentist CIs, credible intervals depend on the prior and are interpreted probabilistically about θ.

Q76. Explain the concept of multiple testing and control of false discoveries.
A76. Multiple hypothesis tests inflate Type I error across tests. Family-wise error rate (FWER) controls (e.g., Bonferroni) control probability of any false positives, often conservative. False discovery rate (FDR, Benjamini–Hochberg) controls expected proportion of false positives among rejected hypotheses and is more powerful in high-dimensional testing (genomics).

Q77. What is an ANOVA and when is it used?
A77. ANOVA (analysis of variance) tests differences among group means by partitioning variance into between-group and within-group components and comparing their ratio (F-statistic). One-way ANOVA tests one factor; multi-way ANOVA includes multiple factors and interactions. Assumptions include independence, normality, and homogeneity of variances.

Q78. How do you test normality of data?
A78. Graphical checks: Q-Q plots, histograms. Formal tests: Shapiro–Wilk, Anderson–Darling, Kolmogorov–Smirnov (with caveats). Tests can be sensitive to large samples (small deviations become significant); combine tests with visual assessment and consider robust or nonparametric methods if normality is questionable.

Q79. What is asymptotic normality and its role in inference?
A79. Many estimators (e.g., MLE) are asymptotically normal: √n(θ̂ − θ) → N(0, Σ). This justifies approximate confidence intervals and tests using normal theory even when finite-sample distributions are unknown, provided sample size is large and regularity conditions hold.

Q80. What is statistical power analysis and how do you compute sample size?
A80. Power analysis determines required sample size to detect an effect of interest with desired power (1−β) at significance level α. For simple cases, use analytic formulas: n = [(z_{1−α/2} + z_{1−β}) σ / δ]^2 for mean difference with effect size δ. For complex models, simulate data to estimate required n under realistic assumptions.

Section 5 — Advanced & Applied Topics (Q81–Q100)

Q81. Explain linear regression assumptions and Gauss–Markov theorem.
A81. Linear regression y = Xβ + ε assumes (1) linearity in parameters, (2) exogeneity E[ε|X] = 0, (3) homoscedasticity Var(ε|X)=σ^2, (4) no perfect multicollinearity in X, and often (5) normality of errors for small-sample inference. Gauss–Markov theorem: under assumptions 1–4, OLS estimator β̂ is the Best Linear Unbiased Estimator (BLUE) — it has minimum variance among linear unbiased estimators.

Q82. How do you detect and handle multicollinearity?
A82. Detect via variance inflation factor (VIF), large standard errors, or unstable coefficients under small model changes. Handle by removing or combining collinear predictors, centering variables, using PCA, or applying regularization (ridge regression) which shrinks coefficients and reduces variance.

Q83. What is regularization (Ridge and Lasso) and when to use them?
A83. Regularization penalizes large coefficients to reduce overfitting. **Ridge** adds λ Σ β_j^2 (L2) and shrinks coefficients; good when predictors are many and correlated. **Lasso** adds λ Σ |β_j| (L1) and can produce sparse models (variable selection). Choose via cross-validation to balance bias-variance tradeoff.

Q84. Explain logistic regression and interpretation of coefficients.
A84. Logistic regression models binary outcomes via log-odds: logit P(Y=1|X) = Xβ. Coefficient β_j represents change in log-odds per unit change in X_j; exp(β_j) is the odds ratio. Estimation via MLE; evaluate model with ROC/AUC, calibration plots, confusion matrices, and consider class imbalance adjustments.

Q85. What is ROC curve and AUC?
A85. ROC (receiver operating characteristic) plots true positive rate vs false positive rate across thresholds. AUC (area under ROC) measures discrimination: 0.5 is random, 1 is perfect. ROC is threshold-independent; for imbalanced data, precision-recall curves may be more informative.

Q86. Describe principal component analysis (PCA) and when to use it.
A86. PCA is a dimensionality reduction technique that finds orthogonal directions (principal components) maximizing variance. It transforms correlated variables into uncorrelated components; first few components capture most variance. Use for visualization, noise reduction, and pre-processing before modeling; beware of interpretability loss and scaling effects.

Q87. What is the Expectation-Maximization (EM) algorithm?
A87. EM iteratively computes maximum likelihood estimates for models with latent variables. **E-step**: compute expected complete-data log-likelihood given current parameters; **M-step**: maximize this expectation to update parameters. Converges to local maxima; common in mixture models, missing data, and latent variable models.

Q88. Explain random effects vs fixed effects in mixed models.
A88. Fixed effects estimate population-level parameters (effects of predictors). Random effects model subject-specific deviations drawn from distributions (e.g., intercepts ~ N(0, σ^2)). Mixed models capture hierarchical or clustered data (repeated measures) and partition variability between and within clusters.

Q89. What is time series stationarity and why does it matter?
A89. Stationarity means statistical properties (mean, variance, auto-covariance) are time-invariant. Many time-series methods (ARIMA) assume stationarity for parameter estimation and forecasting. Nonstationary series (trend, seasonality) often require differencing, detrending, or seasonal adjustment to meet assumptions.

Q90. Describe ARIMA modeling basics.
A90. ARIMA(p,d,q) combines autoregression (AR p terms), differencing (d) to induce stationarity, and moving average (q terms). Fit by identifying p and q via ACF/PACF plots, estimate parameters (MLE or least squares), and validate residuals (white noise). Extend with seasonal terms (SARIMA) for seasonal data.

Q91. What is survival analysis; define hazard and survival functions.
A91. Survival analysis studies time-to-event data. **Survival function S(t)=P(T>t)** is probability of surviving past time t. **Hazard function λ(t)=f(t)/S(t)** is instantaneous event rate at time t given survival to t. Models: Kaplan–Meier (nonparametric), Cox proportional hazards (semi-parametric), and parametric models (Weibull, exponential).

Q92. Explain the Cox proportional hazards model.
A92. Cox model: hazard $\lambda(t|X)=\lambda_0(t)\exp(X\beta)$, where $\lambda_0(t)$ is baseline hazard and β are coefficients. It's semi-parametric: no need to specify $\lambda_0(t)$. Interpretation: $\exp(\beta_j)$ is hazard ratio per unit

increase in X_j. Assumes proportional hazards (constant hazard ratios over time); test via Schoenfeld residuals.

Q93. What is Monte Carlo simulation and variance reduction techniques?
A93. Monte Carlo simulates random draws to approximate integrals or distributions (expectations, probabilities). Variance reduction methods include antithetic variates, control variates, importance sampling, and stratified sampling, which increase estimator efficiency for a given computational budget.

Q94. Describe causal inference basics: confounding, randomization, and instrumental variables.
A94. Causal inference seeks to estimate effect of treatment on outcomes. **Confounding** occurs when variables affect both treatment and outcome, biasing estimates. **Randomization** removes confounding by design. **Instrumental variables (IV)** allow identification when unmeasured confounding exists: an IV affects treatment but only affects outcome via treatment (exclusion restriction). Other tools: propensity scores, difference-in-differences, regression discontinuity.

Q95. What is propensity score matching and when to use it?
A95. Propensity score is P(Treatment|X). Matching treated to control units with similar scores aims to balance covariates and reduce confounding in observational studies. Use when randomization is unavailable; ensure common support and check balance diagnostics post-matching.

Q96. Explain dimensionality curse and strategies to mitigate it.
A96. The curse of dimensionality refers to exponential growth in volume as feature count increases, making density estimation, distance metrics, and sampling sparse. Mitigate by feature selection, dimensionality reduction (PCA, manifold learning), regularization, and using models that scale well (tree-based methods) or incorporate structure (sparsity).

Q97. How do you handle missing data types: MCAR, MAR, MNAR?
A97. MCAR (missing completely at random): missingness independent of data; listwise deletion unbiased. MAR (missing at random): missingness depends on observed data; use multiple imputation or model-based methods. MNAR (missing not at random): missingness depends on unobserved values; requires explicit modeling of missingness mechanism or sensitivity analyses.

Q98. What is multiple imputation?
A98. Multiple imputation fills missing values multiple times (m datasets) using a model that accounts for uncertainty, analyzes each dataset, and pools results (Rubin's rules). It provides valid inference under MAR and preserves variability introduced by imputation compared to single imputation.

Q99. What are information criteria (AIC/BIC) and model selection trade-offs?
A99. AIC = −2 log L + 2k favors goodness-of-fit with penalty 2k (k parameters). BIC = −2 log L + k log n

penalizes complexity more strongly as n grows. AIC targets predictive accuracy; BIC emphasizes model parsimony and consistency under true model in candidate set. Use cross-validation as complementary approach.

Q100. Describe reproducible statistical analysis best practices.
A100. Reproducible analysis includes: clear data provenance and versioning, code in scripts/notebooks with comments, use of package/environment managers (conda, renv), seed control for randomness, automated pipelines, documentation, unit tests for analysis code, and sharing artifacts (data or sanitized versions) plus scripts so others can rerun results. Reproducibility improves credibility and collaboration.
