EASYDEEPLEARN

Statistics Fundamentals

Interview cheat sheet

215 questions · Probability, hypothesis testing, confidence intervals, and the stats DS interviewers ask about.

#bayesian #hypothesis-testing #regression #estimation #probability #causal-inference #ab-testing #theory #eda #descriptive

Probability foundations

State the Central Limit Theorem in one sentence.

For independent, identically distributed samples with finite mean mu and variance σ^2, the sampling distribution of the mean approaches a normal distribution $N(\mu,\ \sigma^2\ /\ n)$ as n grows, regardless of the underlying distribution.

Bayesian vs frequentist — what's the core difference?

Frequentists treat parameters as fixed unknowns and estimate them via long-run frequency properties (unbiased estimators, confidence intervals, p-values). Bayesians treat parameters as random variables with a prior distribution, update with data via Bayes' rule to get a posterior, and summarize with credible intervals.

State the Law of Large Numbers.

As the sample size grows, the sample mean converges to the true expected value. Weak LLN: convergence in probability.

Independence vs uncorrelatedness — what's the difference?

Independent means P(A, B) = P(A) P(B): no relationship at all. Uncorrelated means covariance = 0: no linear relationship.

State Bayes' theorem and one intuitive use.

$P(A\ |\ B)\ =\ P(B\ |\ A)$ P(A) / P(B). Intuition: update a prior belief P(A) with new evidence B to get a posterior belief $P(A\ |\ B)$.

State Kolmogorov's three probability axioms.

For a sample space Ω and events A: (1) P(A) ≥ 0 (non-negativity). (2) P(Ω) = 1 (normalization — some outcome must occur).

Define conditional probability and prove Bayes' rule from it.

$P(A\ |\ B)\ =\ P(A\ B)\ /\ P(B)$ when P(B) > 0. Symmetrically, $P(B\ |\ A)\ =\ P(A\ B)\ /\ P(A) \rightarrow P(A\ B)\ =\ P(B\ |\ A)\ \cdot\ P(A)$.

State the law of total probability.

If $\{B_1, ..., B_n\}$ partition the sample space (mutually exclusive, jointly exhaustive), then P(A) = Σ $P(A\ |\ B_i)\ \cdot\ P(B_i)$. Used constantly: decompose a complex probability by conditioning on a partitioning variable (age, disease status, region).

Bernoulli distribution: parameters, PMF, mean, variance.

X ~ Bernoulli(p): single trial with success prob p. PMF: P(X=1) = p, P(X=0) = 1-p.

Binomial distribution and when to use it.

X ~ Binomial(n, p): number of successes in n independent Bernoulli(p) trials. PMF: $P(X=k)\ =\ C(n,k)\ \cdot\ p^k\ \cdot\ (1-p)$^(n-k).

Poisson distribution and its typical use cases.

X ~ Poisson(λ): count of events in a fixed interval / area, when events happen at constant rate λ independently. PMF: $P(X=k)\ =\ \lambda^k\ \cdot\ e$ ^(-λ) / k!.

Geometric distribution: setup and mean.

X ~ Geometric(p): number of trials until the first success in i.i.d. Bernoulli(p).

Uniform distribution: continuous vs discrete.

Discrete: U{a, ..., b} — each of (b-a+1) integers has probability 1/(b-a+1). Continuous: U(a, b) — density 1/(b-a) on [a,b], zero elsewhere.

Exponential distribution: setup, memorylessness, use cases.

X ~ Exp(λ): time between events in a Poisson process. PDF: f(x) = λ * e^(-λx) for x ≥ 0.

Normal distribution: PDF and key properties.

X ~ $N(\mu,\ \sigma^2)$: $f(x)\ =\ (1\ /\ \mathrm{sqrt}(2\pi\sigma^2))\ \cdot\ \exp(-(x-\mu)^2\ /\ (2\sigma^2))$. Bell-shaped, symmetric around μ. ~68% of mass within 1σ, ~95% within 2σ, ~99.7% within 3σ.

Multivariate normal: parameters and key properties.

X ~ N(μ, Σ), where μ is a length-d mean vector and Σ is a d×d covariance matrix (symmetric positive semi-definite). PDF: |2πΣ|^ $(-1/2)\ \cdot\ \exp(-0.5\ \cdot\ (x-\mu)\ \Sigma^{-1}\ (x-\mu))$.

What is a covariance matrix and its key properties?

For random vector X ∈ R^d: $\mathrm{Cov}(X)\ =\ E[(X-\mu)(X-\mu)]$, a d×d matrix. Diagonal: variances $\mathrm{Var}(X_i)$.

What is linearity of expectation?

For any random variables X, Y and constants a, b: E[aX + bY] = a*E[X] + b*E[Y]. Holds regardless of dependence between X and Y.

Variance of a sum: $\mathrm{Var}(X\ +\ Y)$ = ?

$\mathrm{Var}(X\ +\ Y)\ =\ \mathrm{Var}(X)\ +\ \mathrm{Var}(Y)\ +\ 2\ \cdot\ \mathrm{Cov}(X,\ Y)$. Only when X and Y are uncorrelated (Cov = 0) does the covariance term vanish and $\mathrm{Var}(X+Y)\ =\ \mathrm{Var}(X)\ +\ \mathrm{Var}(Y)$.

Write covariance and correlation formulas.

$\mathrm{Cov}(X,\ Y)\ =\ E[(X\ -\ \mu_X)(Y\ -\ \mu_Y)]\ =\ E[\mathrm{XY}]\ -\ E[X]\mathrm{E[Y]}$. $\mathrm{Corr}(X,\ Y)\ =\ \rho\ =\ \mathrm{Cov}(X,\ Y)\ /\ (\sigma_X\ \sigma_Y)$, in [-1, 1].

What is $E[X\ |\ Y]$ and its Law of Total Expectation?

$E[X\ |\ Y\ =\ y]$ is the expected value of X given that Y = y. As a function of Y, $E[X\ |\ Y]$ is itself a random variable.

Law of Total Variance — $\mathrm{Var}(X)$ = ?

$\mathrm{Var}(X)\ =\ E[\mathrm{Var}(X\ |\ Y)]\ +\ \mathrm{Var}(E[X\ |\ Y])$. 'Within-group' variance (average of conditional variances) plus 'between-group' variance (variance of conditional means).

Why does the Cauchy distribution have no mean?

Cauchy(0,1) PDF: $1\ /\ (\pi(1+x^2))$. Its tails decay only as $1/x^2$ — the integral ∫x * PDF diverges → mean is undefined.

State Jensen's inequality and give an ML example.

For a convex function g: E[g(X)] ≥ g(E[X]). Concave: E[g(X)] ≤ g(E[X]).

What is a moment generating function and why care?

$M_X(t) = E[e(\text{tX})]$. When it exists in an open interval around 0, it uniquely determines the distribution — two RVs with the same MGF are identically distributed.

State Chebyshev's inequality.

$P(|X - \mu| \geq k\sigma) \leq 1/k^2$. Distribution-free bound on tail probabilities — at most $1/k^2$ of the mass is more than k SDs from the mean, for any distribution with finite variance.

State Markov's inequality.

For a non-negative random variable X and a > 0: P(X ≥ a) ≤ E[X] / a. Even weaker than Chebyshev but requires only non-negativity and a finite mean.

What is Hoeffding's inequality?

For bounded i.i.d. X_i ∈ [a, b] with mean μ: $P(|Xn - \mu| \geq t) \leq 2 \cdot \exp(-2n \cdot t^2 / (b-a)^2)$.

How does Monte Carlo estimation work and its convergence rate?

To estimate E[f(X)] where X has a known distribution: draw i.i.d. samples $X_1, ..., X_n$, average $f(X_i)$. By LLN, the average converges to the expectation.

What statistical properties make maximum likelihood the default estimator?

Given data X and a parametric family p(x; θ), MLE picks θ̂ = argmax_θ Σ log $p(x_i\ \theta)$. Equivalently: minimize negative log-likelihood.

What is Fisher information?

I(θ) = $-E[\partial^2 \log p(X\ \theta) / \partial\theta^2] = E[(\partial \log p(X\ \theta)/\partial\theta)^2]$. Measures how much information the observations carry about θ.

Bias, variance, consistency of estimators — define.

Bias: E[θ̂] - θ. Variance: $\text{Var}(\theta)$.

MLE for a Bernoulli(p) — derive.

Likelihood: L(p) = p^Σx_i · $(1-p)^{(n - \Sigma x_i)}$. Log-lik: ℓ $(p) = \Sigma x_i \cdot \log p + (n - \Sigma x_i) \cdot \log(1-p)$. ∂ℓ/∂ $p = \Sigma x_i / p - (n - \Sigma x_i)/(1-p) = 0$ → p̂ = Σx_i / $n = X$.

MLE for $\text{Normal}(\mu, \sigma^2)$ — result.

μ̂ = X̄ (unbiased, minimum variance). σ̂² = $(1/n) \cdot \Sigma(X_i - X)^2$ — the biased MLE. Bias-corrected version divides by n-1 (sample variance).

When would you use Hoeffding's inequality for a CI?

Distribution-free CI for the mean of a bounded random variable in [a, b]: $P(|X - \mu| > t) \leq 2 \cdot \exp(-2\text{nt}^2 / (b-a)^2)$. Gives ε for target coverage: ε = (b-a) * √(log(2/α) / (2n)).

Cramér-Rao lower bound — state it.

For any unbiased estimator θ̂ of θ under regularity conditions: $\text{Var}(\theta)$ ≥ 1 / (n * I(θ)). Fisher information sets the floor for estimator precision.

What is the influence function?

For a functional statistic T(F): $\text{IF}(x\ T,\ F) = \lim_{\varepsilon \to 0}$ ($T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)$) / ε. Measures how much T changes if a single point x is added.

Give the OLS closed-form and its variance.

β̂ = $(XX)^{-1}$ X'y. $\text{Var}(\beta \mid X) = \sigma^2 \cdot (XX)^{-1}$.

State Bayes' theorem and what each term means.

$P(\theta \mid D) = P(D \mid \theta) \cdot P(\theta) / P(D)$. Posterior = Likelihood * Prior / Evidence.

Frequentist vs Bayesian — key philosophical difference.

Frequentist: parameters are fixed but unknown; probability = long-run frequency; CIs, p-values, MLE. Bayesian: parameters are random with a distribution encoding belief; probability = degree of belief; posterior, credible intervals, MAP.

Potential outcomes framework — Rubin causal model.

For each unit i and treatment T, define $Y_i(0)$ (outcome if untreated) and $Y_i(1)$ (outcome if treated). Individual causal effect: $Y_i(1) - Y_i(0)$ — fundamentally unobservable (fundamental problem of causal inference).

Common distributions

Log-normal distribution and its uses.

X ~ $\text{LogNormal}(\mu, \sigma^2)$ if $\log(X) \sim N(\mu, \sigma^2)$. Right-skewed, always positive.

What is a power-law distribution and how do you spot one?

P(X > x) ∝ x^(-α) for large x. Very heavy right tail — a few extreme observations dominate the mean.

Beta distribution: setup and use case.

X ~ Beta(α, β): support [0, 1], flexible shape controlled by α, β. E[X] = α / (α+β); $\text{Var} = \alpha\beta / ((\alpha+\beta)^2(\alpha+\beta+1))$.

Gamma distribution and when it appears.

X ~ Gamma(k, θ): support (0, ∞). PDF ∝ x^(k-1) * e^(-x/θ).

Dirichlet distribution — where does it show up in ML?

$\text{Dirichlet}(\alpha_1, , \alpha_K)$: distribution over probability vectors on the (K-1)-simplex (positive components summing to 1). Multivariate generalization of Beta.

Expectation & variance

Why do we use standard deviation instead of variance in interpretation?

Variance has squared units (dollars^2, meters^2), which are unintuitive. Standard deviation is the square root of variance and shares the units of the data, so it can be plotted alongside the mean and interpreted directly.

Skewness and kurtosis — what do they measure?

$\text{Skewness} = E[((X - \mu)/\sigma)^3]$ — asymmetry of the distribution. Positive: long right tail (log-normal, income).

Why divide by n-1 in the sample variance?

Sample variance uses estimated mean X̄ instead of true μ. Since X̄ is closer to the data than μ (it minimizes SSD by definition), $\Sigma(X_i - X)^2$ is systematically smaller than $\Sigma(X_i - \mu)^2$.

Descriptive statistics & EDA

Mean vs median vs mode — when do you prefer each?

Mean: sensitive to outliers, but efficient for symmetric distributions. Median: robust to outliers, right choice for skewed data (income, response times).

What are quantiles and quartiles?

The q-quantile is the value below which fraction q of the data falls. Median = 0.5-quantile.

How do you read a boxplot?

Box: from Q1 to Q3 (middle 50% of data). Line in box: median.

How do you choose bin width for a histogram?

Rules of thumb: Sturges' (log2(n)+1, small n), Scott's $(3.5\sigma \ / \ n(1/3), \ \text{assumes near} - \text{normal})$, Freedman-Diaconis $(2 \cdot \text{IQR} \ / \ n(1/3), \ \text{robust to skew})$. Modern default in most libraries: Freedman-Diaconis.

What is kernel density estimation?

Non-parametric density estimate: KDE(x) = (1/(nh)) Σ $K((x \ - \ x_i)/h)$, where K is a kernel (usually Gaussian) and h is bandwidth. Bandwidth choice is the key knob: too small → wiggly, too large → over-smoothed.

How do you define an outlier in practice?

No single definition — depends on the model and question. Common rules: (1) Tukey: outside Q1 - 1.5·IQR or Q3 + 1.5·IQR.

What is MAD and why is it robust?

Median Absolute Deviation: $\text{MAD}(X) \ = \ \text{median}(| \ X_i \ - \ \text{median}(X) \ |)$. Robust measure of dispersion — breakdown point 50% (half the data can be arbitrarily corrupted before MAD explodes; SD breaks down at 0%).

Describe a good EDA workflow for a new dataset.

(1) Shape: rows × columns, dtypes, missing values. (2) Univariate: histograms / boxplots / value counts per column.

MCAR vs MAR vs MNAR — what's the difference?

MCAR (Missing Completely At Random): missingness independent of both observed and unobserved data. Can drop rows without bias.

What imputation methods should you consider?

(1) Drop rows: fine if MCAR and few missing. (2) Mean/median/mode imputation: baseline, distorts variance.

When and why do you log-transform a variable?

(1) Right-skewed positive variables (income, prices, response times): log transform pulls in the tail, makes distribution more symmetric, often more Gaussian. (2) Multiplicative relationships become additive after log — foundation of Box-Cox / Yeo-Johnson and log-linear models.

What is the Box-Cox transformation?

$\text{Box} - \text{Cox}(y, \ \lambda) \ = \ (y\lambda \ - \ 1) \ / \ \lambda$ if λ ≠ 0, $\log(y)$ if λ = 0. Choose λ (typically via MLE) to make the transformed variable closest to normal.

Pearson vs Spearman vs Kendall correlation — when do you use each?

Pearson r: measures linear relationship, assumes approximately normal data, sensitive to outliers. Spearman rho: rank-based Pearson — captures any monotonic relationship, robust to outliers.

What is an ECDF and why is it useful?

Empirical Cumulative Distribution Function: $F_n(x) \ = \ (1/n) \ \cdot \ \Sigma \ 1(X_i \ \leq \ x)$. Plots the fraction of data ≤ x for every x.

How do you read a Q-Q plot?

Q-Q (quantile-quantile) plot: plot empirical quantiles vs theoretical quantiles (typically normal). Perfect fit → points on the y=x line.

Standardization vs normalization — what's the difference?

Standardization (Z-score): (x - μ) / σ → mean 0, SD 1. Preserves shape, robust when features have similar scales after transform.

What multivariate EDA plots are most useful?

(1) Correlation heatmap: quick view of all pairwise linear relationships. (2) Scatter plot matrix (pairs plot): all bivariate scatters, diagonal shows univariate — best for small-to-medium feature sets.

How does EDA help catch target leakage?

Look for features with suspicious high correlation to target: (1) any feature perfectly predictive by itself is suspect; (2) features generated after the target event should not exist at prediction time; (3) IDs / timestamps sometimes carry target information indirectly. Cross-checks: check feature availability at inference time (would you have this feature when predicting?); check for target-derived aggregations (e.g., 'average order value' when predicting a churn outcome computed after the churn).

Why should you always do group-wise EDA?

Whole-population statistics can hide subgroup patterns that matter for the model: (1) Simpson's paradox — trends flip when you disaggregate. (2) Fairness — model performance may differ by demographic group.

What time-series-specific EDA should you do?

(1) Plot the raw series and its rolling mean / median at multiple window sizes. (2) Decompose into trend + seasonality + residual (STL).

What do you look for in residual plots?

Residuals vs fitted: no pattern = OK; funnel shape = heteroscedasticity; curvature = missed non-linearity. Q-Q plot: check normality of residuals (banana → skew, sigmoid → heavy tails).

When should you log-transform the response?

(1) Right-skewed positive outcomes (income, prices, spend, time-on-page). (2) Multiplicative errors → additive on log scale.

Overall conversion went down but improved in every country. How is that possible?

The traffic mix changed. Simpson's paradox appears when the groups have different baseline rates and their relative share shifts, so an aggregate can move opposite to every subgroup.

How do you spot selection bias in a dataset someone hands you?

Ask how a row came to exist, which is a different question from what the row contains. Every dataset is the output of a process that decided what to record, and that process is often correlated with the outcome.

Sampling & central limit theorem

What is sampling bias and how do you mitigate it?

Sampling bias occurs when the sample is not representative of the target population — some groups over- or under-represented. Causes: convenience sampling, self-selection, survivorship bias, non-response.

Your revenue-per-user metric is extremely right-skewed. Can you still use a t-test?

Usually yes, but for a reason worth stating precisely: the t-test requires the sampling distribution of the mean to be approximately normal, not the data, and the central limit theorem delivers that at large sample sizes even for skewed data. The practical caveat is that convergence is slower the heavier the tail, so with a few hundred observations and extreme outliers the test can still be unreliable.

Hypothesis testing

What is a p-value, precisely?

The probability, assuming the null hypothesis is true, of observing a test statistic at least as extreme as the one obtained. It is NOT the probability that the null is true, nor the probability of a mistake.

Type I vs Type II error — what's the difference?

Type I error (false positive): rejecting H0 when it's actually true. Its probability is alpha (significance level, often 0.05).

Why is multiple testing a problem and how do you correct for it?

Running many independent tests at $\alpha = 0.05$ will produce many false positives by chance alone (5% per test). Corrections: Bonferroni (divide alpha by number of tests — very conservative), Holm-Bonferroni (stepwise, better power), or control the False Discovery Rate with Benjamini-Hochberg (best for discovery-style work with many tests).

How do you formulate null and alternative hypotheses?

Null H0: the 'no effect' / 'no difference' hypothesis ($\mu_A = \mu_B$, $\beta = 0$, $\text{treatment has no impact}$). Alternative H1: what we suspect ($\mu_A \neq \mu_B \text{ two} - \text{sided}$, or $\mu_A > \mu_B \text{ one} - \text{sided}$).

One-sample t-test: setup and assumptions.

H0: $\mu = \mu_0$. Statistic: $t = (X - \mu_0) / (s / \sqrt{n})$, where s is sample SD.

Two-sample t-test: pooled vs Welch — when to use each?

Compare two group means. Pooled (Student's): assumes equal variances → uses pooled SD, $\text{df} = n_1 + n_2 - 2$.

When do you use a paired t-test?

Same subjects measured under two conditions (before / after treatment), or naturally paired observations (twins, matched pairs). Compute differences $d_i = X_A i - X_B i$, then one-sample t on d.

z-test vs t-test — when to pick each?

z-test: known population variance (rare in practice), or very large n (>~100) where t-distribution ≈ normal. t-test: unknown population variance (usual case). For proportions with large n: z-test for a proportion or two-proportion z-test.

Chi-square test of independence — what does it do?

Tests whether two categorical variables are independent, given an r × c contingency table. Statistic: $\chi^2 = \Sigma\,(O_{\text{ij}} - E_{\text{ij}})^2 / E_{\text{ij}}$, where $E_{\text{ij}} = \text{row}_i \cdot \text{col}_j / N$ under independence. df = (r-1)(c-1).

Chi-square goodness-of-fit test — setup.

Tests whether observed frequencies match an expected distribution: $\chi^2 = \Sigma\,(O_i - E_i)^2 / E_i$, df = k - 1 - (parameters estimated). Uses: is a die fair?

What does one-way ANOVA test?

H0: $\mu_1 = \mu_2 = ... = \mu_k$ (all group means equal). Statistic: F = between-group variance / within-group variance ~ $F_{k-1,\,N-k}$ under H0.

Two-way ANOVA and what interactions mean.

Tests three effects at once: main effect of factor A, main effect of factor B, and A×B interaction. Interaction: the effect of A depends on the level of B.

Mann-Whitney U (Wilcoxon rank-sum) — when and why?

Non-parametric test comparing two independent groups. Combines all data, ranks it, compares rank sums between groups.

Wilcoxon signed-rank test — setup.

Non-parametric alternative to the paired t-test. Compute differences d_i, $\text{rank}\ |d_i|$, sum ranks separately for positive and negative differences.

Kruskal-Wallis — non-parametric ANOVA analog.

Extends Mann-Whitney to k > 2 groups. Rank all data across groups, compute H = 12/(N(N+1)) * Σ $(R_i^2 / n_i) - 3(N+1) \sim \chi^2 k - 1$ approximately.

Friedman test and its use case.

Non-parametric alternative to repeated-measures ANOVA. Same subjects rated on k treatments; rank within each subject, sum ranks per treatment.

Kolmogorov-Smirnov test — one- and two-sample.

Statistic: $D = \max_x |F_1(x) - F_2(x)|$ — max vertical distance between two ECDFs (two-sample) or between empirical and theoretical CDF (one-sample). Distribution-free, works for continuous data, sensitive to shifts in location, scale, or shape.

Shapiro-Wilk normality test — when useful?

Tests whether data comes from a normal distribution. Very sensitive for n < 50; over-powered for large n (reports significant non-normality for trivial deviations).

How do you test equality of variances?

Bartlett: assumes normality; sensitive to non-normality. Levene: uses absolute deviations from mean or median (Brown-Forsythe uses median); more robust.

Why report effect size alongside p-values?

P-values conflate signal size with sample size — huge n makes trivial effects significant. Effect size quantifies practical importance regardless of n.

What is power analysis and how do you use it?

$\text{Power} = P(\text{reject H0} \mid \text{H1 true}) = 1 - \beta$. Power analysis: given effect size, α, and power (typically 0.8), solve for required sample size.

What is MDE (minimum detectable effect)?

Smallest effect size the experiment can detect with a target power (usually 0.8) given the sample size and α. Computed via power formula.

Bonferroni correction — how does it work?

For k independent tests at family-wise error rate α: use $\alpha_{\text{adjusted}} = \alpha / k$ per test. Guarantees P(any false rejection) ≤ α under independence, ≤ α always (union bound).

How does the Benjamini-Hochberg procedure work?

Controls the FDR (expected fraction of false positives among rejections) at level q. Steps: (1) Sort p-values p_(1) ≤ ... ≤ p_(m).

What is a permutation test?

Non-parametric test based on the exchangeability principle: shuffle group labels, recompute the test statistic, build a null distribution empirically. Compute p-value = fraction of permutations with statistic ≥ observed.

What does McNemar's test do?

Paired categorical (usually binary) test. Compares proportions in matched pairs.

When to use Fisher's exact test?

2×2 (or r×c) contingency table with small expected counts (E < 5) where chi-square approximation fails. Computes an exact p-value from the hypergeometric distribution.

What is p-hacking and how do you prevent it?

Running many analyses (test variants, subgroups, transformations) until you find p < 0.05, then reporting only the significant results. Inflates false-positive rate dramatically.

FWER vs FDR — which do you control when?

FWER (Family-Wise Error Rate): P(any false rejection). Controlled by Bonferroni / Holm — safe when even a single false positive is costly (safety, regulatory).

How is a CI related to a hypothesis test?

A 95% CI for a parameter includes all null values that would NOT be rejected at α = 0.05 in a two-sided test. Equivalently: if 0 is outside the 95% CI of the difference, then reject H0: no difference at α = 0.05.

Why is peeking at running A/B tests dangerous?

Peeking = repeatedly running a test and stopping when p < 0.05 → hugely inflates false-positive rate (up to ~40% for one-sided at α=0.05). Fixes: (1) commit to sample size ex ante and don't peek.

What is alpha-spending in sequential testing?

Divide the total α (0.05) across scheduled interim looks so cumulative false-positive rate stays capped. Pocock: uniform per look (conservative).

A product manager asks what a p-value of 0.03 means. What do you say?

Say that if the change truly had no effect, you would see a difference this large or larger about 3% of the time by chance alone. Then say what it does not mean, because that is where decisions go wrong: it is not the probability that the change works, and it is not the probability that the null hypothesis is true.

Non-parametric tests

How many bootstrap resamples do you need, and which statistics does it handle?

Given data $X_1,\ldots,X_n$, sample n observations with replacement to form X*, compute the statistic θ*, repeat B ≈ 1000-10000 times to build an empirical distribution of θ̂. Use it to compute SEs, CIs, and bias.

Power, effect size & multiple testing

What techniques reduce variance in A/B tests?

(1) CUPED: covariate adjustment using a pre-experiment baseline metric — 30-70% variance reduction on retention metrics. (2) Stratification: guarantee balanced enrollment by strata (region, device).

How does CUPED reduce variance in A/B tests?

CUPED (Controlled-experiment Using Pre-Experiment Data, Deng et al. 2013): use a pre-experiment covariate X (e.g. same metric, 30 days pre-treatment) to explain part of Y's variance. Adjusted metric: Y' = Y - θ * (X - E[X]), where θ = Cov(Y, X)/Var(X).

Why is checking an A/B test daily and stopping when it turns significant wrong?

Because the false positive rate is not the nominal 5% under repeated looks; every additional peek is another chance for random fluctuation to cross the threshold, and with daily checks over a few weeks the real error rate rises to something like 20 to 30%. Fixed-horizon tests assume exactly one analysis at a predetermined sample size, and stopping on the first significant result systematically selects the fluctuations that happened to be extreme.

What do you need to know to compute a sample size for an experiment?

Four things, and the hard one is not statistical. The baseline rate of the metric, its variance, which for a proportion follows from the baseline but for a revenue metric must be estimated and is usually much larger than people expect.

Your experiment tracks 20 metrics and one is significant. What now?

Treat it as expected noise until proven otherwise, because with 20 independent tests at the 5% level you expect about one false positive even when nothing changed. The structural fix is to declare a single primary metric before the experiment, with the rest labelled as secondary or guardrail, which turns a fishing expedition into a hypothesis test.

Estimation: MLE, MoM, MAP

What does a 95% confidence interval mean?

If we repeated the sampling and interval construction procedure many times, about 95% of the resulting intervals would contain the true parameter. It's NOT 'there is a 95% probability the parameter is in this specific interval' — the parameter is fixed and the interval is random.

What is the bootstrap and when do you use it?

Bootstrap resamples the data with replacement to approximate the sampling distribution of an estimator. You get standard errors, confidence intervals, and bias estimates without strong parametric assumptions.

What is Maximum Likelihood Estimation?

Choose parameter values that make the observed data most probable — i.e., maximize the likelihood function $L(\theta) = P(\text{data} \mid \theta)$. In practice, maximize log L (numerically stable, turns products into sums).

What is the Method of Moments (MoM)?

Set sample moments equal to theoretical moments and solve for parameters. Example: Gamma(α, β) with mean α/β and variance α/β^2 → set sample mean = α/β and sample $\text{var} = \alpha/\beta^2$, solve for $\hat{\alpha}$ and $\hat{\beta}$.

MAP vs MLE — key difference.

MLE: $\hat{\theta}$ = argmax $p(D \mid \theta)$. MAP: $\hat{\theta}$ = argmax $p(\theta \mid D) = \text{argmax}$ $p(D \mid \theta) \cdot p(\theta)$.

Explain the EM algorithm.

For MLE with latent variables (Z): (E) compute $q(Z) = p(Z \mid X, \theta_{\text{old}})$; (M) update $\theta_{\text{new}} = \text{argmax\_}\theta\ E_q[\log\ p(X, Z \mid \theta)]$. Guaranteed to increase the log-likelihood at each iteration, converges to a local max.

Standard error vs standard deviation — the difference.

SD: spread of the data (√variance) — describes the population. SE: spread of a statistic across hypothetical repeated samples — describes the estimator's precision.

What is the delta method?

For a differentiable function g and asymptotically normal $\hat{\theta}$: g($\hat{\theta}$) ~ $N(g(\theta),\ g(\theta)^2 \cdot \text{Var}(\theta))$ approximately. Use to derive SE of transformations (ratios, log-odds).

What is the jackknife?

Leave-one-out precursor of the bootstrap. For each i, compute $\hat{\theta}_{(-i)}$ = statistic on data without observation i.

What is a conjugate prior?

A prior p(θ) is conjugate to a likelihood $p(x \mid \theta)$ if the posterior $p(\theta \mid x)$ belongs to the same family. Examples: Beta prior + Bernoulli likelihood → Beta posterior; Gamma prior + Poisson → Gamma; Normal prior + Normal (known σ) → Normal; Dirichlet + categorical → Dirichlet.

Beta-Bernoulli conjugate update — derive.

Prior: p ~ Beta(α, β). Data: n Bernoulli trials with k successes.

What is an uninformative prior?

Prior meant to encode minimal prior belief. Options: (1) Flat / uniform (Beta(1,1)) — feels natural but is not invariant to reparameterization.

What is Bayesian shrinkage?

Posterior estimates are pulled ('shrunk') from noisy per-group MLEs toward the prior / group mean. Effect: reduces variance at the cost of small bias — Stein's paradox shows shrinkage dominates MLE in ≥ 3 dimensions.

What is empirical Bayes?

Estimate the hyperparameters of the prior from the data itself (usually by marginal maximum likelihood), then use that prior for the posterior. Cheap approximation to a full hierarchical Bayesian model — avoids MCMC on hyperpriors.

How does Thompson sampling work?

For each arm, maintain posterior over its reward. To choose an action, draw $\theta_j \sim p(\theta_j \mid \text{data})$ for each arm, pick arm with highest sampled θ_j.

Why do conjugate priors matter in production?

Closed-form posterior updates → no MCMC needed, O(1) update per new observation → scales to streaming / online settings. Examples: (1) Beta-Bernoulli for CTR bandits, (2) Gamma-Poisson for arrival-rate estimation, (3) Normal-Normal for personalization.

What does a 95% confidence interval actually assert?

That the procedure generating it captures the true parameter in 95% of hypothetical repetitions of the study. The confidence is a property of the method across repetitions, not of the particular interval you computed, which either contains the parameter or does not.

When is a Bayesian analysis genuinely worth the extra effort?

When you have real prior information, when the sample is small, or when you need the probability statement itself. With little data, a weakly informative prior stabilizes estimates that maximum likelihood pushes to absurd values, which is why hierarchical models are standard for many small groups such as per-store or per-user effects: partial pooling shrinks noisy groups toward the overall mean by exactly as much as the data warrants.

Confidence intervals & bootstrap

95% CI for a mean — formula and interpretation.

CI = $\bar{X} \pm t_{n-1,\ 0.975} \cdot (s\ /\ \sqrt{n})$. Interpretation: if we repeated the experiment many times, 95% of the CIs would contain the true μ.

What CI methods exist for a proportion?

Wald (naive): $\hat{p} \pm z * \sqrt{\hat{p}(1-\hat{p})/n}$ — bad when $\hat{p}$ is near 0/1 or n small (can even exceed [0,1]). Wilson score: preferred default; more accurate coverage.

Bootstrap CI methods — percentile vs BCa vs studentized.

Percentile: use 2.5th and 97.5th quantiles of bootstrap distribution — simple but biased when the sampling distribution is skewed. BCa (bias-corrected & accelerated): corrects bias and skewness → generally best default.

Why do you need block bootstrap for time series?

Standard bootstrap resamples individual points, destroying temporal dependence and grossly under-estimating variance for autocorrelated data. Block bootstrap resamples contiguous blocks (moving or stationary) to preserve short-range dependence.

Prediction interval vs confidence interval — the difference.

CI: uncertainty in a parameter (μ, β, p) — shrinks with n. PI: uncertainty in a new individual observation $Y_{\text{new}} = \mu + \varepsilon$ — includes both parameter uncertainty AND irreducible noise σ, so wider than CI and doesn't shrink to zero (bounded by σ).

Credible interval vs confidence interval.

CI (frequentist): 'if we repeated the experiment many times, 95% of these intervals would contain θ' — probability statement about the procedure. Credible interval (Bayesian): 'given the data + prior, $P(\theta \in [a, b] \mid \text{data}) = 95\%$' — probability statement about θ.

Why do ratio metrics need the delta method?

Ratios like CTR = clicks / impressions have per-user numerators AND denominators → sample-averaged ratio has non-trivial variance. Naive user-level SE is wrong (ignores denominator variability).

When does the bootstrap give you the wrong answer?

It fails where the statistic depends on the extreme tail of the distribution, because resampling can never produce a value larger than the largest observation, so the maximum and near-extreme quantiles are badly estimated. It fails with small samples, since the empirical distribution is a poor stand-in for the population and the intervals come out too narrow.

Regression assumptions

What are the Gauss-Markov assumptions for OLS?

(1) Linearity: $E[Y \mid X] = X\beta$. (2) Strict exogeneity: $E[\varepsilon \mid X] = 0$ (no omitted variable bias).

Heteroscedasticity — what breaks and how do you fix it?

$\text{Var}(\varepsilon \mid X)$ depends on X. OLS β̂ stays unbiased and consistent, but SEs and t-tests are wrong → invalid inference.

How do you detect and handle multicollinearity?

Symptoms: unstable β̂ across subsets, huge SEs, opposite-sign coefficients from univariate expectation, high pairwise correlation. Diagnostics: $\text{VIF} = 1 / (1 - R_j^2)$ — VIF > 5-10 is concerning.

R^2 vs adjusted R^2 — when do you use each?

$R^2 = 1 - \text{RSS}/\text{TSS}$: fraction of variance explained; always increases when you add features (even garbage ones). Adjusted $R^2 = 1 - (1 - R^2) \cdot (n-1)/(n-p-1)$: penalizes model complexity; only increases if the new feature improves fit beyond noise.

Cook's distance, leverage, DFBETAS — define.

Leverage h_{ii}: diagonal of hat matrix $H = X(X'X)^{-1}X'$; measures how extreme x_i is in X-space. Cook's distance: how much β̂ changes if point i is removed; > 4/n is a flag.

Standardized vs studentized residuals — the difference.

Standardized: $r_i / (s \cdot \sqrt{(1 - h_{ii})})$ — divides raw residual by its SE. Studentized (internally): same but using s (uses point i to estimate σ).

Logistic regression: why MLE and not OLS?

Y ∈ {0, 1} → OLS predicts outside [0, 1] and residuals are heteroscedastic (Var = p(1-p)). Logistic: p = σ(x'β), maximize Σ $y_i \log p_i + (1 - y_i) \log(1 - p_i)$.

Odds ratio — interpret and derive.

Odds = p / (1 - p). Odds ratio $(\text{OR}) = \text{odds}_A / \text{odds}_B$.

GLM: link function and exponential family — connection.

GLM: Y ~ exponential family with mean $\mu = E[Y \mid X]$, and link g(μ) = x'β. Canonical links: Gaussian → identity, Bernoulli → logit, Poisson → log, Gamma → inverse.

Poisson regression: setup and pitfalls.

Y ~ Poisson(λ), log λ = x'β. Mean = variance = λ (equidispersion).

Negative binomial regression — why use it?

Adds dispersion parameter θ: $\text{Var} = \mu + \mu^2/\theta$. Handles over-dispersion (Var > Mean) which is the norm for real counts (customer visits, defect counts, rare events).

What is quasi-likelihood?

Rather than a full parametric distribution, specify only a mean-variance relationship: E[Y] = μ, $\text{Var}(Y) = \varphi \cdot V(\mu)$. Quasi-scores solve exact same equations as MLE for the mean, but SEs use estimated dispersion φ.

What is a Generalized Additive Model (GAM)?

$g(E[Y]) = \beta_0 + f_1(x_1) + f_2(x_2)$ + ..., where each f_j is a smooth (usually spline). Extends GLM by letting each feature enter through a non-parametric smooth.

Splines: knots, degrees of freedom, regularization.

Spline = piecewise polynomial with continuity at knots. Cubic splines: 3rd-degree pieces, continuous through 2nd derivative.

How do you include and interpret interactions in regression?

Include $x_1 \cdot x_2$ (product term) plus main effects. Interpretation: coefficient of x_1 depends on level of x_2.

What is a mixed-effects model?

y = Xβ + Zu + ε, where β = fixed effects (population-level slopes), u = random effects (subject / cluster deviations), u ~ N(0, G), ε ~ $N(0, \sigma^2 I)$. Uses: repeated measures on subjects, hierarchical data (students in schools, users in accounts), longitudinal panels.

Random intercepts vs random slopes — the difference.

Random intercept: each cluster has its own baseline level around the global β_0. Random slope: each cluster's slope for a predictor varies around the global β.

Cluster-robust standard errors — when to use?

Observations within clusters (schools, accounts, sessions) are correlated → default OLS SEs are too small → false positives. Cluster-robust ('Liang-Zeger', 'sandwich') SEs correct for within-cluster correlation.

Mixed-effects vs cluster-robust SEs — which do you pick?

Both handle cluster correlation, but differently. Mixed-effects: fully specifies covariance structure → efficient if the model is correct, but misspecification biases estimates.

In practice, how do you handle correlated features in regression?

(1) Domain knowledge: keep the more meaningful one (age vs birthyear). (2) Combine: sum, difference, or PCA / PLS.

Durbin-Watson test — what does it test?

Tests for first-order autocorrelation in regression residuals: $\text{DW} = \Sigma(e_t - e_{t-1})^2 / \Sigma e_t^2$. Range ~ [0, 4].

Two-stage least squares (2SLS) — the recipe.

Stage 1: regress T on instrument Z (and controls X) → predicted T̂. Stage 2: regress Y on T̂ (and X) → coefficient on T̂ is IV estimate.

Your model reports $R^2 = 0.92$. What can go wrong with celebrating that?

Plenty. R^2 rises mechanically with every predictor you add, whether or not it helps, so a high value on training data says nothing about generalization and adjusted R^2 only partially compensates.

Two predictors correlate at 0.95. Does it matter?

It depends entirely on what you want from the model. For prediction, collinearity is largely harmless: the fitted values and out-of-sample error are barely affected, because the pair jointly carries the information regardless of how the coefficient is split between them.

You have 50,000 measurements from 200 patients. What is your sample size?

Closer to 200 than to 50,000, because observations within a patient are correlated and each additional measurement from the same patient carries much less new information than a measurement from a new patient. Treating all 50,000 as independent shrinks the standard errors dramatically and produces confidence intervals far too narrow, which is how clustered data generates confident nonsense.

Mixed effects & clustering

What is a hierarchical Bayesian model?

Multi-level model: parameters θ_j vary across groups j and are themselves drawn from a shared 'population' distribution $\theta_j \sim N(\mu, \tau^2)$. Partial pooling: group estimates borrow strength from each other via the shared prior — automatic shrinkage.

Bayesian methods

What is the posterior predictive distribution?

$p(y_{\text{new}} \mid D) = \int p(y_{\text{new}} \mid \theta)\, p(\theta \mid D)$ dθ. Predicts new data marginalizing over posterior uncertainty.

What is the marginal likelihood (evidence) and why is it hard?

p(D) = ∫ $p(D \mid \theta)$ p(θ) dθ. Normalizing constant for the posterior.

How do you interpret a Bayes factor?

$\text{BF}_{12} = p(D \mid M_1) \,/\, p(D \mid M_2)$ — ratio of marginal likelihoods → posterior odds = BF × prior odds. Jeffreys scale: BF > 10 strong evidence for M_1, > 100 decisive; BF < 1/10 flip.

How does Bayesian A/B testing work?

Model: p_A, p_B ~ Beta prior; observe successes / failures → Beta posteriors. Metric: $P(p_B > p_A \mid \text{data})$ via Monte Carlo (sample from both posteriors, count fraction).

How does Metropolis-Hastings work?

MCMC: build a Markov chain whose stationary distribution is the posterior. Given current θ, propose θ' ~ $q(\theta \mid \theta)$.

What is Gibbs sampling?

Special MCMC: sample each parameter (or block) from its full conditional $p(\theta_j \mid \theta_{-j}, D)$, one at a time. Guaranteed acceptance (α = 1) when full conditionals are known.

Hamiltonian Monte Carlo — intuition.

Uses gradients of log-posterior to propose long, informed steps → efficient in high dimensions. Introduces auxiliary momentum p, runs Hamiltonian dynamics (leapfrog) to propose new (θ, p), MH-accepts.

How do you diagnose MCMC convergence?

(1) R̂ (Gelman-Rubin) < 1.01 across multiple chains → chains agree. (2) ESS (effective sample size) ≥ 400 per parameter.

Burn-in and thinning — why?

Burn-in: discard initial iterations before the chain reaches stationarity. Typical: 10-50% of chain.

Variational inference vs MCMC.

VI: approximate posterior $p(\theta \mid D)$ with a simpler family q_φ(θ) (e.g. mean-field Gaussian) by maximizing $\text{ELBO} = E_q[\log p(D, \theta)] - E_q[\log q(\theta)]$. Advantages: much faster, scales to big data (SVI, mini-batches), differentiable → fits on GPU.

Derive the ELBO for VI.

log p(D) = log ∫ p(D, θ) dθ = log ∫ q(θ) * p(D, θ)/q(θ) dθ ≥ $E_q[\log p(D, \theta)/q(\theta)]$ (Jensen). ELBO = E_q[log p(D, θ)] - E_q[log q(θ)] = E_q[log p(D | θ)] - KL(q || p(θ)). log $p(D) - \text{ELBO} = \text{KL}(q \mid\mid p(\theta \mid D)) \geq 0$ → maximizing ELBO minimizes KL to the posterior.

Informative vs non-informative priors — the tradeoff.

Non-informative (flat, Jeffreys): 'let the data speak' — safe when you have plenty of data, but can be worse than an informative prior at small n. Informative (based on historical data, domain knowledge, similar experiments): dramatically improves inference in small-sample settings, but can also inject bias if wrong.

WAIC and LOO-CV — Bayesian model comparison.

Both estimate out-of-sample predictive performance. WAIC = $-2 \cdot (\text{lppd} - p_{\text{WAIC}})$ where lppd = expected log posterior predictive on training data, $p_{\text{WAIC}} =$ effective number of parameters via posterior variance.

What is a prior predictive check?

Simulate data from the prior alone: θ ~ p(θ), y ~ $p(y \mid \theta)$. Check whether generated y is plausible given domain knowledge (e.g. simulated coin flip probabilities in [0, 1]; simulated log-revenues aren't in the trillions).

Posterior predictive check — how do you use it?

Simulate replicated datasets y_{rep} from posterior predictive; compare their summary statistics (mean, quantiles, distribution shape) to observed data. If observed data looks like a plausible draw from y_{rep}, the model captures the data — otherwise, refine.

When should you use a bandit instead of an A/B test?

Bandit (Thompson sampling / UCB) automatically shifts traffic toward the winning arm during the test → minimizes regret. Use when: (1) short-lived items (news, promotions) where you can't afford to send 50% traffic to the loser; (2) many arms (>10) with fast feedback; (3) time-decaying decisions.

Causal inference basics

Why doesn't correlation imply causation?

Two variables can move together because A causes B, B causes A, both are caused by a third variable Z (confounder), or by pure coincidence. Establishing causation requires either a randomized experiment (breaks confounding by design) or careful causal inference methods (DAGs, matching, instrumental variables, difference-in-differences, regression discontinuity, do-calculus).

What is Simpson's paradox?

A trend that appears in several subgroups can reverse when the groups are combined. Classic example: a treatment looks better in each subgroup but worse overall, because group sizes and baseline rates vary.

Users who adopt feature X churn less. Can you say the feature reduces churn?

Not from that association alone, because four mechanisms produce it without the feature doing anything: (1) confounding — engaged users both adopt features and stay, so engagement causes both, (2) reverse causation — users who were already going to stay are the ones who explore features, (3) selection — the population you measured was filtered in a way that creates the link, (4) coincidence, which matters when the sample is small. To make the causal claim you need randomization, meaning an experiment that offers the feature to a random subset, or a credible identification strategy: instrumental variables, regression discontinuity, difference-in-differences, or matching under an explicitly stated ignorability assumption.

What is SUTVA?

Stable Unit Treatment Value Assumption: (1) no interference between units — one unit's treatment doesn't affect another's outcome (breaks in networks, marketplaces, viral products), (2) no hidden treatment variants — one 'treatment' is well-defined. Violated in social networks (spillovers), two-sided marketplaces (Uber: driver-side changes affect rider outcomes), auctions (competition), any system with global equilibrium effects.

How do you handle confounders in observational studies?

(1) Include them in regression (X in $E[Y \mid T,\ X]$) if measured — assumes correct functional form. (2) Matching / propensity score matching — trims to overlap region.

An aggregate trend reverses once you stratify. Which conditioning set is correct?

An association observed at the aggregate level reverses when data is split by a confounder. Classic example: Berkeley admissions — men admitted at higher aggregate rate but women admitted at higher rate within every department (women applied to more competitive departments).

What is a collider and why is conditioning on it bad?

In a DAG A → C ← B: C is a collider (both A and B point into it). A and B are marginally independent, but conditioning on C creates a spurious association.

What is Pearl's backdoor criterion?

To identify $P(Y \mid \mathrm{do}(T))$ from observational data, find a set of variables Z such that: (1) Z blocks all backdoor paths from T to Y (paths starting with an arrow into T), (2) Z contains no descendant of T. Then $P(Y \mid \mathrm{do}(T)) = \Sigma_z\, P(Y \mid T,\ Z = z)$ P(Z=z).

When is the frontdoor criterion useful?

When unmeasured confounders make backdoor impossible, but a fully-mediating variable M exists. E.g.

What is a propensity score?

$e(x) = P(T = 1 \mid X = x)$. Rosenbaum-Rubin: under unconfoundedness, adjusting for e(x) alone is sufficient (dimension reduction from $\mid X \mid$ to 1).

Inverse Probability Weighting — how does it work?

Weight each treated unit by 1/e(x) and each control by 1/(1-e(x)) → weighted population is balanced on X. Marginal ATE estimator: Σ $(T_i\, Y_i\, /\, e_i)\ -\ \Sigma\, ((1 - T_i)\, Y_i\, /\, (1 - e_i))\, /\, n$.

Why is a doubly-robust estimator useful?

Combines outcome model μ̂(x) and propensity model ê(x). Formula: ATE = E[μ̂(1, X) - μ̂(0, X) + T(Y - μ̂(1, X))/ê(x) - (1-T)(Y - μ̂(0, X))/(1-ê(x))].

What makes a valid instrumental variable?

Z is a valid IV for T → Y if: (1) Relevance: Z affects T (Cov(Z, T) ≠ 0). (2) Exclusion: Z affects Y only through T (no direct Z → Y).

Difference-in-differences — setup and assumption.

Compare change in outcome for treated group vs change for control group before and after intervention. Estimand: $(Y_{\text{treated}}, \text{post}\ -\ Y_{\text{treated}}, \text{pre})\ -\ (Y_{\text{control}}, \text{post}\ -\ Y_{\text{control}}, \text{pre})$.

How do you defend the parallel-trends assumption?

(1) Plot pre-treatment trends of treated and control — visually parallel? (2) Event study: interact treatment indicator with lead / lag indicators; pre-treatment lead coefficients ≈ 0.

Regression discontinuity design — how does it work?

Treatment assigned by a threshold on a continuous running variable X (e.g. score ≥ 60 → scholarship). Compare Y just above vs just below cutoff — near cutoff, individuals are 'as-if random'.

What is synthetic control?

Construct a weighted combination of untreated units that reproduces the pre-treatment trajectory of the treated unit. Weights ≥ 0, sum to 1, chosen to minimize pre-treatment fit error on X and Y.

Mediation analysis — direct vs indirect effects.

Path: T → M → Y (with possible direct T → Y). Decomposition: total effect = direct effect (T → Y not through M) + indirect effect (T → M → Y).

How do you estimate heterogeneous treatment effects (HTE)?

$\mathrm{CATE}(x) = E[Y(1)\ -\ Y(0) \mid X = x]$ varies with covariates. Methods: (1) interaction terms in regression (Y ~ T + T×X); (2) causal forests / GRF (Wager & Athey); (3) meta-learners: T-learner (separate models), S-learner (single joint), X-learner (better in imbalanced treatment), R-learner; (4) doubly-robust CATE.

How does Simpson's paradox strike A/B tests?

Test wins overall but loses in every user segment (or vice versa) → aggregate confounded by mid-experiment composition change. Common causes: (1) SRM by segment (imbalanced assignment across segments over time), (2) traffic mix drift while experiment runs, (3) new-vs-returning composition shift.

How do you test in a marketplace / network with SUTVA violations?

(1) Cluster randomization (randomize at community / graph-community level so spillovers stay within cluster). (2) Switchback experiments (turn treatment on/off over time, randomize periods).

Quantile treatment effect (QTE) vs ATE.

ATE captures mean effect. $\mathrm{QTE}(\tau) = F_Y(1)\text{^(-1)}(\tau)\ -\ F_Y(0)\text{^(-1)}(\tau)$ captures effect on the τ-th quantile — useful when the mean is misleading (heavy tails, revenue metrics dominated by top 1%).

Real interview: your model performs great in A/B but flops post-launch. Why?

Common reasons: (1) Selection bias — the A/B population is not the launch population (early adopters, engaged users). (2) SUTVA violation — 50% traffic doesn't scale to 100% (marketplace saturation, ad auction dynamics).

You cannot randomize. What is the strongest causal claim you can still make?

One conditional on an assumption you state explicitly, which is the honest form of every observational causal claim. Begin by drawing the assumed causal structure, because which variables to adjust for is a question about that structure, not about which improve fit; conditioning on a collider or a mediator introduces bias rather than removing it.

Experimentation & A/B testing

What are the pillars of a solid A/B test design?

(1) Clear primary metric + guardrail metrics chosen in advance. (2) Sample size / MDE via power analysis.

What is Sample Ratio Mismatch (SRM) and why check it?

Chi-square test that observed enrollment ratio (say 50/50) matches expected ratio. $p < 0.005$ → strong evidence something's wrong with the randomizer (bot traffic, ID mismatch, bug in assignment logic, biased filtering). Any downstream analysis with SRM is invalid — you can't trust the effect estimate.

What are guardrail metrics?

Secondary metrics you commit to monitor to catch bad tradeoffs: latency, error rate, retention, revenue-per-user, complaint rate, security signals. Test can 'win' on the primary metric but must not degrade guardrails beyond a preset threshold.

OEC — Overall Evaluation Criterion — what is it?

A single scalar composite of relevant metrics chosen ex ante as the experiment's success criterion. Simplifies decision-making (one number).

Novelty vs primacy effects in long experiments.

Novelty: users react positively to any change initially, then revert → early effect inflated, long-run effect smaller. Primacy: existing users hate change initially, then adapt → early effect deflated, long-run effect larger.

Switchback experiments — when and how?

Alternate treatment on/off over time windows across the whole market → each window randomly assigned. Estimator: difference in outcomes between treated and control windows, with cluster-robust or time-block SEs.

What is the winner's curse in A/B testing?

Because we only 'ship' experiments that cross the significance / effect threshold, the observed effect on the winners is a biased over-estimate of the true effect (selection on the outcome). Empirical rule at Microsoft / Bing: shipped effects shrink 10-40% on re-measurement.

How do you measure long-term effects when A/B tests are short?

(1) Long holdout experiment: keep 1% of traffic unchanged for months → measure long-term differential. (2) Surrogate metrics: proxy predicting long-term outcomes (predicted LTV, 28-day retention as surrogate for annual retention).

How does a company scale to running 1000+ experiments concurrently?

(1) Layered assignment (users get one exposure per orthogonal layer) — Google's classic approach. (2) Mutually exclusive groups when interactions are expected.