

STATISTICS & PROBABILITY

for Data Science / Analytics Interviews



- 1. Mean, Median, Mode, Variance, Std Dev
- 2. Probability Distributions — Normal, Binomial, Poisson
- 3. Central Limit Theorem
- 4. Hypothesis Testing — Null/Alt, Type I/II Errors
- 5. p-value & Confidence Interval — Interpretation
- 6. Correlation vs Causation
- 7. A/B Testing Framework
- 8. Bayes Theorem — Worked Example

Phase 3 • Intuition-First, Interview-Focused • Not Derivation-Heavy

Mean, Median, Mode, Variance, Std Dev

Central Tendency

Measure	What it is	When it breaks / misleads
Mean	Sum / count — the balance point	Heavily pulled by outliers / skew
Median	Middle value when sorted	Robust to outliers — preferred for skewed data (e.g. salary, house prices)
Mode	Most frequent value	Only useful measure for categorical data

Rule of thumb: if mean $\neq$ median by a lot, your data is skewed — always plot a histogram before trusting the mean.

Spread

Variance = average of squared deviations from the mean $\rightarrow$ tells you how spread out the data is, but in squared units (hard to interpret directly).

Standard Deviation = $\sqrt{\text{variance}}$ $\rightarrow$ brings it back to the original units, so it's directly comparable to the mean. This is why std is reported far more often than variance.

Variance = avg squared distance from mean

Std Dev = same idea, real-world units

🕒 Common Interview Questions

- When would you report median instead of mean? $\rightarrow$ Skewed data with outliers — e.g. average salary is misleading if a few execs earn 50x the median employee; median better represents the 'typical' person.
- Why do we square the deviations in variance instead of just taking absolute value? $\rightarrow$ Squaring penalizes larger deviations more, and makes the math differentiable/tractable for further statistical methods (unlike absolute value).
- If std dev is 0, what does that tell you? $\rightarrow$ Every value in the dataset is identical — zero spread.
- How would you explain standard deviation to a non-technical stakeholder? $\rightarrow$ 'On average, how far a typical data point is from the overall average' — a measure of how spread out or consistent the data is.

Probability Distributions

Normal (Gaussian) Distribution

Symmetric bell curve. 68% of data within 1 std dev of mean, 95% within 2, 99.7% within 3 ("68-95-99.7 rule").

DS example: heights, exam scores, measurement errors, and — thanks to CLT — sample means of almost anything.

Binomial Distribution

Number of "successes" in a fixed number of independent yes/no trials, each with the same success probability p .

DS example: out of 100 emails sent, how many get opened (each email = open or not, fixed n , fixed p).

Poisson Distribution

Number of events happening in a fixed time/space window, when events happen independently at a known average rate.

DS example: number of website visits per minute, number of support tickets per day, number of defects per batch.

► Binomial vs Poisson — how to tell them apart

Aspect	Binomial	Poisson
Setup	Fixed number of trials (n)	Fixed interval of time/space, no fixed " n "
Question	"Out of n tries, how many succeed?"	"How many events occur in this window?"
Example	Out of 50 users, how many convert?	How many users arrive per hour?

Common Interview Questions

- Why does the Central Limit Theorem make the Normal distribution so important even if raw data isn't normal? → Because sample means/sums of almost any distribution tend toward Normal as sample size grows — this is what justifies using Normal-based tests (t-test, z-test) so widely.
- Give a real DS scenario each for Binomial and Poisson. → Binomial: A/B test conversions out of a fixed number of visitors. Poisson: number of server errors per hour (rate-based, no fixed ' n ').
- What connects Binomial and Poisson? → Poisson is the limiting case of Binomial when n is very large and p is very small, but $n \times p$ (the rate) stays constant.

Central Limit Theorem (CLT)

CLT: If you repeatedly take random samples from ANY population (regardless of its shape) and compute the sample mean each time, the distribution of those sample means will approach a Normal distribution as sample size grows ($n \geq \sim 30$ is the common rule of thumb).

Population — any shape (even skewed)

Take many samples, compute mean each time

Distribution of sample means → Normal

► Why interviewers care about this

CLT is the entire justification for using Normal-distribution-based tools (confidence intervals, t-tests, z-tests) on real-world business data that is almost never perfectly normal itself. Without CLT, most classical hypothesis testing wouldn't be valid.

🕒 Common Interview Questions

- Explain CLT like you would to a product manager. → 'Even if individual customer purchase amounts are unpredictable and skewed, if you repeatedly look at the AVERAGE purchase of groups of customers, those averages will form a nice predictable bell curve — that's what lets us build reliable confidence intervals around an average.'
- Why is the $n \geq 30$ rule of thumb used? → Empirically, sample means tend to look 'Normal enough' by then for most underlying distributions — though more skewed populations may need a larger n .
- Does CLT say the original data becomes normal? → No — it says the distribution of SAMPLE MEANS becomes normal, not the raw underlying data.

Hypothesis Testing — Null/Alt, Type I/II Errors

H_0 (Null) — "no effect / no difference" **vs** H_1 (Alt) — "there IS an effect / difference"

We always assume H_0 is true by default and look for enough evidence in the data to reject it — like "innocent until proven guilty" in court.

► Type I vs Type II Error

Error	What happened	Court analogy	Business example
Type I (α)	Rejected H_0 when it was actually TRUE — false positive	Convicting an innocent person	Launching a "winning" feature that actually did nothing
Type II (β)	Failed to reject H_0 when it was actually FALSE — false negative	Letting a guilty person go free	Missing a feature that actually would have worked

There's a trade-off: lowering Type I error (stricter threshold) tends to raise Type II error, and vice versa — you can't minimize both freely with a fixed sample size.

Quick scipy.stats snippets

```
from scipy import stats

# Independent samples t-test: do two groups have different means?
t_stat, p_value = stats.ttest_ind(group_a_scores, group_b_scores)

# Chi-square test: are two categorical variables associated?
chi2, p_value, dof, expected = stats.chi2_contingency(contingency_table)
```

🔗 Common Interview Questions

- Explain Type I vs Type II error using a real business scenario. → Type I: rolling out a new UI because the test said it improved conversion, when really it didn't (wasted engineering effort). Type II: shelving a genuinely better UI because the test failed to detect the improvement (missed opportunity).
- Which error is 'worse' — doesn't it depend on context? → Yes — in medical testing, Type II (missing a real disease) is often considered worse than Type I (false alarm leading to extra tests); in spam filtering, Type I (blocking a real email) may be worse than Type II (letting spam through).
- How do you choose between a t-test and a chi-square test? → t-test compares MEANS of a numeric variable between groups; chi-square tests association between two CATEGORICAL variables.

p-value & Confidence Interval — Interpretation

"Explain p-value to a non-technical person"

"If there were truly NO real effect/difference, the p-value is the probability we'd still see data this extreme (or more) just by random chance. A small p-value means our result would be a very unlikely coincidence if nothing real was going on — so we start doubting the 'nothing going on' story."

p-value is NOT the probability that H_0 is true, and it's NOT the probability the result happened by chance — it's the probability of seeing data this extreme, assuming H_0 is true. This distinction is one of the most commonly tested interview traps.

► The $p < 0.05$ Convention

0.05 is just a widely-used convention, not a law of nature — it means we accept a 5% chance of a Type I error (false positive) as "acceptable" by community standard. Some fields (e.g. physics) use much stricter thresholds.

► Confidence Interval

"Explain a 95% CI to a non-technical person"

"If we repeated this study 100 times and built a confidence interval each time, about 95 of those intervals would contain the true population value." It is NOT "there's a 95% chance the true value is in THIS specific interval" — the true value either is or isn't in it; the 95% describes the reliability of the METHOD across repeats.

🕒 Common Interview Questions

- What's the most common misinterpretation of a p-value? → Thinking it's 'the probability the null hypothesis is true' or 'the probability the result is due to chance' — it's actually the probability of the observed data (or more extreme) GIVEN that H_0 is true.
- A p-value of 0.04 vs 0.001 — does the smaller one mean a 'bigger' effect? → Not necessarily — a very small p-value can come from a large sample size detecting a tiny, practically meaningless effect. Always check effect size alongside p-value.
- Narrow vs wide confidence interval — what does each tell you? → Narrow CI = more precise estimate (usually from larger sample size); wide CI = more uncertainty about the true value.

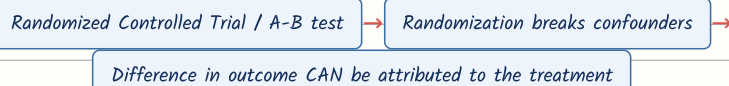
Correlation vs Causation

Correlation = two variables move together. Causation = one variable's change directly PRODUCES a change in the other. Correlation is necessary but nowhere near sufficient to claim causation.

► Why variables can be correlated without causation

Reason	Classic example
Confounding variable (hidden 3rd cause)	Ice cream sales & drowning deaths both rise in summer (heat is the real cause of both)
Reverse causation	Companies with more revenue hire more marketers — does marketing cause revenue, or does revenue enable more marketing hires?
Pure coincidence	Spurious correlations (e.g. Nicolas Cage films vs pool drownings — famous internet example)

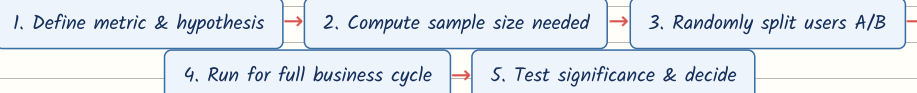
► How to move toward causation



🕒 Common Interview Questions

- Give an example from your own project where you had to be careful about correlation vs causation. → (Prepare a real example: e.g. a feature correlated with churn might just be correlated with tenure, not actually cause churn.)
- If A/B testing isn't possible, how else might you argue for causation? → Natural experiments, instrumental variables, difference-in-differences, controlling for known confounders in regression — but always with a weaker causal claim than a true RCT.
- Why is correlation alone dangerous for business decisions? → Acting on a correlated-but-not-causal signal wastes resources on a lever that won't actually move the outcome (e.g. optimizing a metric that doesn't cause revenue growth).

A/B Testing Framework



Step-by-step, what actually matters in an interview

- Define success metric upfront — e.g. click-through rate, conversion rate — and avoid "peeking" and switching metrics mid-test (p-hacking).
- Sample size / power analysis — determines how long to run the test; too short = underpowered, can't detect real effects (Type II error risk).
- Randomization — must be truly random assignment to A/B to eliminate confounders and enable causal claims.
- Run for a full cycle — at least one full week (captures weekday/weekend behavior), avoid stopping early just because results look significant ("peeking problem").
- Statistical significance $\neq$ practical significance — a statistically significant 0.01% lift may not be worth shipping given engineering cost.

Common pitfall interviewers probe: stopping a test early the moment $p < 0.05$ appears — this inflates the false positive rate because you're effectively running multiple implicit tests (the "peeking problem").

🔗 Common Interview Questions

- How would you determine how long to run an A/B test? → Power analysis based on baseline conversion rate, minimum detectable effect you care about, and desired significance/power levels (commonly 80% power, 5% significance) — determines required sample size, then converted to duration using expected traffic.
- What's the 'peeking problem' and how do you avoid it? → Repeatedly checking significance mid-test and stopping as soon as $p < 0.05$ appears inflates false positives; fix with sequential testing methods or committing to a pre-determined sample size/duration.
- Metric shows statistical significance but stakeholders are unimpressed — why? → The effect size might be tiny and not practically meaningful even if 'real' — always report effect size and confidence interval, not just the p-value.
- How do you handle novelty effects in A/B tests? → Run the test long enough that the initial excitement/curiosity about a new feature wears off, so you're measuring steady-state behavior, not a temporary spike.

Bayes Theorem — Worked Example

Bayes Theorem lets you UPDATE a belief (probability) after seeing new evidence. It's the mathematical version of "how surprising is this evidence, and how should it change what I believed before?"

Prior belief + New evidence → Updated (posterior) belief

► Worked Example — The Classic Disease Test Problem

Setup: A disease affects 1% of a population. A test correctly detects the disease 99% of the time (true positive rate), and correctly identifies healthy people 95% of the time (so 5% false positive rate). A random person tests positive — what's the probability they actually have the disease?

$$P(\text{Disease}) = 0.01 \quad P(\text{No Disease}) = 0.99$$

$$P(\text{Positive} | \text{Disease}) = 0.99 \quad P(\text{Positive} | \text{No Disease}) = 0.05$$

$$P(\text{Disease} | \text{Positive}) = [P(\text{Positive} | \text{Disease}) \times P(\text{Disease})] / P(\text{Positive})$$

$$P(\text{Positive}) = (0.99 \times 0.01) + (0.05 \times 0.99) = 0.0099 + 0.0495 = 0.0594$$

$$P(\text{Disease} | \text{Positive}) = 0.0099 / 0.0594 \approx 16.7\%$$

Even with a "99% accurate" test, a positive result only means a ~17% real chance of disease — because the disease is RARE (low prior), so false positives from the large healthy population outnumber true positives.

💡 Why this trips people up (and why interviewers love it)

Our intuition anchors on the test's 99% accuracy and ignores how rare the disease is (the prior). Bayes forces you to combine BOTH — this exact pattern shows up in spam filters, fraud detection, and medical ML models.

🔗 Common Interview Questions

- Why doesn't a '99% accurate' test mean a positive result is 99% reliable? → Because accuracy alone ignores the base rate (prior probability) of the condition — with a rare condition, false positives from the much larger healthy population can outnumber true positives.
- Where does Bayes' theorem show up in real ML systems? → Naive Bayes classifiers, spam filters, fraud detection scoring, medical diagnosis models, and recommendation systems updating beliefs about user preference as new interactions come in.
- How would you explain 'prior' and 'posterior' in plain English? → Prior = what you believed before seeing this evidence; Posterior = what you believe AFTER updating on the new evidence.