



Interview Bank: Maths and Statistics

AI course, interview bank | About 35 interview questions with model answers and what the interviewer listens for: probability, Bayes, hypothesis tests and A/B tests, estimation, linear algebra, calculus and optimisation

#Interviews #Statistics #Probability #LinearAlgebra

What these questions test

A criminal trial	A hypothesis test
Presumed innocent	Null hypothesis H0: no effect
Evidence against the accused	Test statistic from the data
How surprising the evidence is if innocent	p-value
Standard of proof	Significance level α , e.g. 0.05
Convicting an innocent person	Type I error (false positive)
Acquitting a guilty person	Type II error (missed effect)
Chance of convicting the guilty	Power, $1 - \beta$

WHERE THE ANALOGY BREAKS 'Not guilty' is not 'innocent': failing to reject H0 does not prove no effect, it may mean too little data.

The plain-language picture interviewers want behind every testing answer: a trial. Use it when explaining to a non-statistician

Prerequisites and scope

Taught in **maths 08-11** (probability, distributions and Bayes, statistics, information theory), **maths 02-07** (linear algebra, calculus, optimisation) and **machine-learning 16** (A/B testing). Data-scientist loops weigh this sheet most; ML-engineer loops ask the starred basics and the derivations. Levels: [J] junior, [M] mid, [S] senior.

Name	Formula	Use it for
Bayes	$P(H E) = \frac{P(E H) P(H)}{P(E)}$	posterior from prior and evidence
Standard error of a mean	$SE = \sigma / \sqrt{n}$	confidence intervals, sample size
Variance of a sum	$Var(X + Y) = VarX + VarY + 2Cov(X, Y)$	portfolio, correlated errors
Expected value	$E[X] = \sum_x x P(X = x)$	dice, games, cost of a policy

where H , E = hypothesis and evidence (events) (-); σ = standard deviation of one observation (metric unit); n = sample size (observations); X , Y = random variables; Cov their covariance (metric unit); x = a value X can take, with probability $P(X = x)$ (metric unit)

Four formulas interviewers expect you to write without notes

Deep answers: A/B tests and likelihood

MID

Q. An A/B test on checkout shows +1.2 % conversion with $p = 0.04$. Explain it to the product manager and say whether to ship.

Structure

1. Restate what was tested and the primary metric

2. What $p = 0.04$ means, and what it does not mean

3. Give the effect with its confidence interval

4. Check validity: planned duration, sample ratio, peeking, number of metrics

5. Recommend, with guardrails and the cost of being wrong

Model answer

If the change had no effect, a gap at least this large would appear about 4 % of the time, so chance alone is an unlikely explanation. It is not a 96 % chance that the variant is better. The decision rests on the interval: if the 95 % CI of the relative lift is +0.1 % to +2.3 %, the effect is probably positive but may be tiny. Before shipping I check that the test ran its planned length without early stopping, that the group sizes match the planned split, and that conversion was the primary metric chosen in advance. If guardrails (revenue per user, refunds, latency) are flat, ship and keep monitoring. Taught in machine-learning 16 and maths 10.

Follow-ups

Q: We tracked 20 metrics and only this one is significant. Ship?

A: At $\alpha = 0.05$, 20 null metrics give about one false positive on average; apply Bonferroni ($0.05 / 20 = 0.0025$) or Benjamini-Hochberg, or confirm with a new test.

Q: The PM wants to stop as soon as p drops below 0.05.

A: Repeated looks inflate the false-positive rate far above 5 %; fix the horizon in advance or use a sequential test built for continuous monitoring.

Q: And if p had been 0.06?

A: Not a different world from 0.04: report the interval; decide on the cost of a wrong ship versus a missed gain and on the power the test had.

Red flags

- '96 % chance the variant wins'
- no effect size or interval
- ignores peeking and multiple metrics

t = 60 · w / r: speaking time t (s) of the model answer, w: its words, r: speaking rate (130 words/min). Aim for 1-2 min, then stop and invite follow-ups.

MID Q. Show that minimising log loss is maximum likelihood for a binary classifier.

Structure

- 1 Model each label as Bernoulli with $p = \sigma(w \cdot x)$
- 2 Write the likelihood of the data set
- 3 Take the log, negate, average
- 4 Recognise binary cross-entropy
- 5 Generalise: other noise models, other losses

Model answer

Each label $y_i \in \{0, 1\}$ has $P(y_i) = p_i^{y_i} (1 - p_i)^{1 - y_i}$ with $p_i = \sigma(w \cdot x_i)$. For independent examples the likelihood is the product over i . The log turns the product into a sum and keeps the maximiser; negating and dividing by n gives $-(1/n) \sum [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$, the binary cross-entropy. So minimising log loss is maximising likelihood. The same argument gives mean squared error for Gaussian noise of fixed variance and mean absolute error for Laplace noise. Taught in maths 10-11 and machine-learning 03.

Follow-ups

Q: Why maximise the log of the likelihood?
 A: Products of many small numbers underflow; log is monotone so the optimum is the same, and sums differentiate term by term.

Q: Add a Gaussian prior $N(0, \tau^2)$ on the weights. What changes?
 A: You get the MAP estimate: the negative log prior adds $\|w\|^2 / (2\tau^2)$, an L2 penalty (ridge, weight decay).

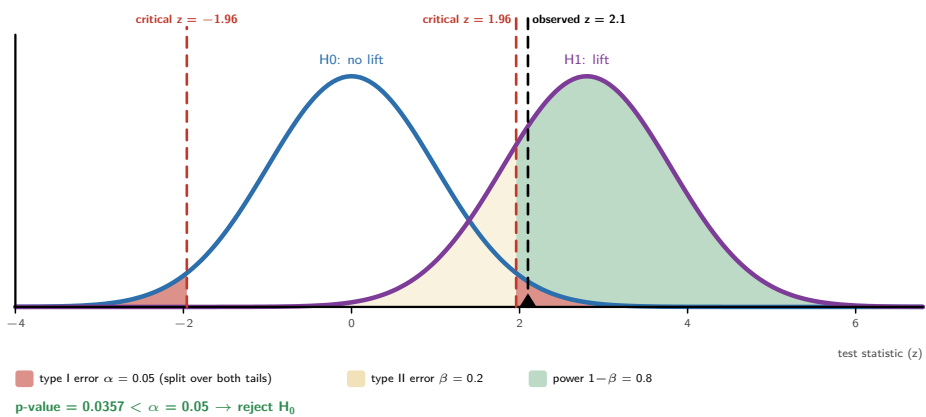
Q: Gradient of the loss with respect to the logit $z = w \cdot x$?
 A: $p - y$: prediction minus label, bounded and well scaled, one reason this pairing trains well.

Red flags

- calls the likelihood 'the probability of the parameters'
- cannot say where the log comes from

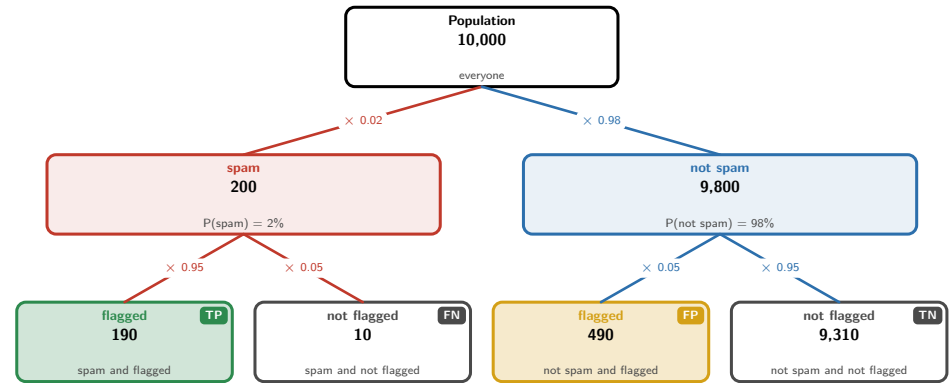
96 words · ≈ 44 s spoken

$t = 60 - w / r$: speaking time t (s) of the model answer, w : its words, r : speaking rate (130 words/min). Aim for 1-2 min, then stop and invite follow-ups.

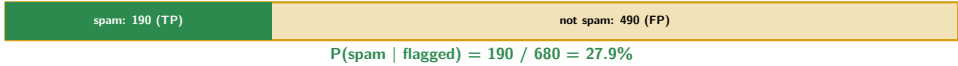


[M] 'With a true effect of 2.8 standard errors and a two-sided α of 0.05, what is the power? And for an observed z of 2.1?'
 Answer: power ≈ 0.80 ($2.8 = 1.96 + 0.84$), $p \approx 0.036$, reject H_0 . They look for: power depends on effect size and n , not on the observed p (maths 10)

Probability and testing, computed



Evidence "flagged": $TP + FP = 190 + 490 = 680$



Natural frequencies: $P(H|E) = TP / (TP + FP) = 190 / (190 + 490) = 0.279$ (27.9%)

Bayes' rule: $P(H|E) = P(E|H) \cdot P(H) / [P(E|H) \cdot P(H) + P(E|\neg H) \cdot P(\neg H)] = 0.95 \cdot 0.02 / (0.95 \cdot 0.02 + 0.05 \cdot 0.98) = 0.279$

Prior 2% $\rightarrow$ posterior 27.9%: the evidence multiplies the odds by the likelihood ratio $LR = P(E|H) / P(E|\neg H) = 0.95 / 0.05 = 19$

Without the evidence: $P(H|\neg E) = FN / (FN + TN) = 10 / (10 + 9,310) = 0.11\%$

Where: H : spam (the hypothesis); E : flagged (the evidence); $\neg$: not; $P(H)$ = prior, the base rate before seeing E ; $P(E|H)$ = sensitivity (true-positive rate); $P(E|\neg H)$ = false-positive rate; $P(H|E)$ = posterior; TP, FN, FP, TN: counts of true / false positives / negatives among $N = 10,000$ people (or items)

[J] '2 % of emails are spam. The filter flags 95 % of spam and 5 % of legitimate mail. What share of flagged mail is spam?'
 Answer: about 28 %. They look for: base rates handled with counts, not intuition (maths 09)

Sample size per group, $\alpha = 0.05$ two-sided, power 0.8 (Lehr's rule)

FORMULA $n = 16 \times V / D^2$

WHERE

- n = users needed in each group (users)
- V = variance of the metric per user; for a conversion rate p , $V = p(1 - p)$ (unitless)
- D = minimum detectable effect, absolute difference (same unit as the metric)

WORKED

$V = 0.09$
 $D = 0.01$
 $n = 16 \times 0.09 / 0.01^2$
 $n = 14\,400$ users per group

RESULT $n = 14\,400$ users per group

READ IT $16 \approx 2 \times (1.96 + 0.84)^2$: the z for $\alpha/2 = 0.025$ plus the z for power 0.8, squared, times two groups' variance

MISTAKES

- using a relative effect (10 %) where an absolute one (1 point) is meant
- forgetting it is per group: the total is $2n$

[M] 'Baseline conversion 10 %. How many users per group to detect +1 point?' Answer: about 14,400 per arm with Lehr's rule. They look for: variance of a proportion $p(1 - p)$, the square on the effect, and that halving the effect quadruples n (machine-learning 16)

Question

20 INDEPENDENT ? = 0.05 MOST ACCURATE

[M] A team tests 20 independent features at $\alpha = 0.05$ and finds one significant. Which statement is MOST accurate?

TESTS Hypothesis testing and multiple comparisons (maths 10)

✓ A. Even if no feature has an effect, about one significant result is expected

$20 \times 0.05 = 1$ expected false positive; $P(\text{at least one}) = 1 - 0.95^{20} \approx 0.64$

✗ B. The significant feature has a 95 % chance of being real

Confuses the p -value with the probability the effect is real

✗ C. Bonferroni would set the per-test threshold to 0.01

$0.05 / 20 = 0.0025$

✗ D. Multiple testing matters only when the tests are dependent

Independence makes the inflation worse, not absent

Strategy: Count the tests, multiply by α for the expected false positives, then name a correction. Interviewers want the number 0.64 or 'about one'.

Quick questions: probability and statistics

Statement	True	False
1. Zero correlation means X and Y are independent. → Only no linear relation: $Y = X^2$ with X symmetric around 0 has zero correlation (maths 08).	☐	☒
2. The sum of two independent normal variables is normal. → Means add and variances add (maths 09).	☒	☐
3. With 10 million users, almost any non-zero difference is significant. → SE shrinks as $1/\sqrt{n}$: ask whether the effect is large enough to matter (maths 10).	☒	☐
4. A 95 % confidence interval contains 95 % of the data. → It is about the parameter; the spread of data is a prediction or reference interval (maths 10).	☐	☒
5. The maximum-likelihood variance of a Gaussian divides by n and is biased. → Dividing by $n - 1$ (Bessel) removes the bias (maths 10).	☒	☐
6. The central limit theorem needs normally distributed data. → Means of i.i.d. data with finite variance become normal whatever the shape (maths 09).	☐	☒
7. More data always removes bias from an estimate. → It cuts variance; a biased sample (selection, survivorship) stays biased (data 02).	☐	☒

[J/M] Seven statements interviewers use to probe misconceptions

Probability

parameters fixed, data vary; sums to 1 over outcomes

Likelihood

data fixed, a function of the parameters; need not sum to 1

vs

Tell: What is held fixed?

Confidence interval

frequentist: the procedure covers the parameter 95 % of the time

Credible interval

Bayesian: 95 % posterior probability the parameter is inside

vs

Tell: Is the parameter random?

Type I error

false positive, rate α

Type II error

missed effect, rate β

vs

Tell: Was H_0 true?

MLE

maximises the likelihood

MAP

maximises likelihood $\times$ prior; prior = regulariser

vs

Tell: Is there a prior?

Standard deviation

spread of individual observations

Standard error

spread of an estimate, $\sigma/\sqrt{n}$

vs

Tell: Data or estimate?

Covariance

in units of $X \times Y$

Correlation

covariance / ($\sigma_x \sigma_y$), between -1 and 1

vs

Tell: Is it scale-free?

[M] 'What is the difference between ...?' The one question that separates each pair is the answer interviewers want

Q1 [M] 8 heads in 10 tosses. Is the coin unfair at $\alpha = 0.05$?

A $P(\geq 8 \text{ heads}) = (45 + 10 + 1) / 1024 \approx 0.055$; two-sided $p \approx 0.11$: not significant. Looks for: an exact binomial count, two-sided reasoning (maths 10).

Q2 [M] Expected value of the larger of two fair dice?

A $E = \sum P(\max \geq k) = 6 - (0 + 1 + 4 + 9 + 16 + 25)/36 = 161/36 \approx 4.47$. Looks for: a clean method, not enumeration by luck (maths 08).

Q3 [J] Expected number of rolls to get a six?

A Geometric with $p = 1/6$: $1/p = 6$ rolls. Looks for: naming the distribution (maths 09).

Q4 [S] A 50/50 test shows 50,800 vs 49,200 users. Concern?

A Sample ratio mismatch: SD of a group count = $\sqrt{(100,000 \times 0.25)} \approx 158$, so 800 is about 5 SD: an assignment or logging bug; do not read the result (machine-learning 16).

Q5 [M] What does the bootstrap do and when do you use it?

A Resample the data with replacement many times, recompute the statistic, read the CI from its percentiles; for medians, ratios or metrics with no closed-form SE (maths 10).

Q6 [M] Is KL divergence symmetric? Why care?

A No: $KL(p||q) \neq KL(q||p)$. Minimising $KL(p||q)$ over q covers all modes of p ; the reverse picks one mode, as in variational inference (maths 11).

Q7 [J] When does a Poisson model fit counts?

A Independent rare events at a constant rate per interval; mean = variance = λ . Variance far above the mean: use a negative binomial (maths 09).

Q8 [M] What is Simpson's paradox?

A A trend in every subgroup reverses when pooled, because a confounder changes the group mix; report by segment (metrics-and-kpis 11).

Q9 [J] Mean or median for salaries? Why?

A Median: skewed data with a long tail pulls the mean up. Looks for: robustness to outliers (maths 10).

Q10 [M] Entropy of a fair 8-sided die in bits?

A $\log_2 8 = 3$ bits: the number of yes/no questions needed on average (maths 11).

☀ **Quick questions: linear algebra, calculus, optimisation**

Question	Model answer	They listen for	Read
What is an eigenvector? [J]	A direction a matrix only stretches: $Av = \lambda v$, λ the stretch factor	the geometric picture	maths 04
How does PCA relate to the SVD? [M]	Principal axes are the right singular vectors of the centred data; variance of PC $i = s_i^2 / (n - 1)$	centring first	maths 04
What does a zero determinant mean? [M]	The map squashes space into a lower dimension: singular, not invertible	volume picture	maths 03
$X^T X$ is singular in least squares. Fix? [M]	Features are collinear or $p > n$: drop or combine features, or add a ridge term λI	rank, regularisation	maths 03
Why step against the gradient? [J]	The gradient is the direction of steepest increase; its negative decreases f fastest locally	directional derivative	maths 06
Shape of the Jacobian of f from R^n to R^m ? [M]	$m \times n$: one row per output, one column per input	shapes, not formulas	maths 06

Why does convexity matter? [M]

Every local minimum is global, so gradient descent finds it; neural nets are not convex

global vs local

maths 07

Minimise f subject to $g(x) = 0$? [S]

Lagrange: $\nabla f = \lambda \nabla g$ at the optimum; KKT conditions for inequalities, as in the SVM dual

multipliers, KKT

maths 17

Eight questions with the model answer, what the interviewer listens for and the course sheet

⊕ **[S] 'Why does the learning rate have an upper limit on a quadratic?'**

Model answer: on $f(x) = (a/2)x^2$, one step gives $x \leftarrow (1 - \eta a)x$, which shrinks only if $|1 - \eta a| < 1$, that is $\eta < 2/a$. For a full model, a is the largest Hessian eigenvalue (the sharpest direction), so the step size is capped by curvature. They look for: a one-line derivation and the link to curvature (maths 07, maths 17).

★ **Traps and recap**

MYTH OR FACT?

"A non-significant result proves there is no effect"

MYTH: It shows the data cannot rule out zero
Low power is the usual reason; report the interval.

MYTH OR FACT?

" $p < 0.05$ means the effect is large"

MYTH: Significance is about noise, not size
Large samples make tiny effects significant.

MYTH OR FACT?

"Standardising features makes them normally distributed"

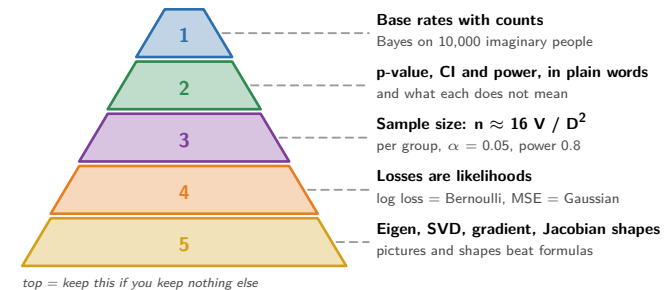
MYTH: It only shifts and rescales
The shape, skew included, is unchanged.

MYTH OR FACT?

"A correlation of 0.9 means X causes Y"

MYTH: Correlation alone cannot give direction or rule out confounders
Causal claims need an experiment or a causal design.

Four statements candidates say confidently. Front: the claim; back: the verdict



What to have ready, most asked first