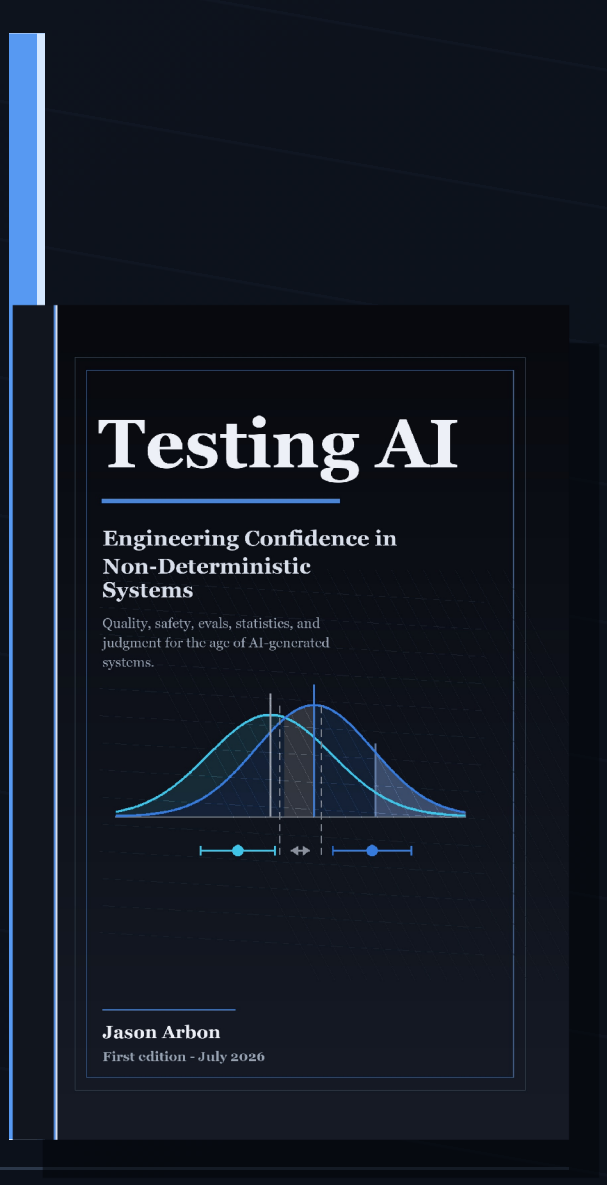
STATISTICAL TESTS FOR AI QUALITY

Statistics make uncertainty visible.

They do not decide whether a product is good enough to ship. They help a team decide honestly.

Swipe for a practical way to compare AI versions.



COMPARING VERSIONS FAIRLY

Run the same cases through both versions.

Version A		Version B	
Hard refund	6	Hard refund	8
Simple return	9	Simple return	9
Policy edge case	5	Policy edge case	6

A paired comparison asks about the **per-case difference**, so easy and difficult cases do not get mistaken for a model improvement.

A paired t-test is a useful starting point for numeric scores on the same eval cases.

NULL HYPOTHESIS

Start with the boring explanation.

Default claim

H0

The new version did not create a meaningful difference. The apparent lift may be ordinary sampling noise.

Evidence challenge

H1

The observed results are hard to explain as noise alone, under the stated plan and assumptions.

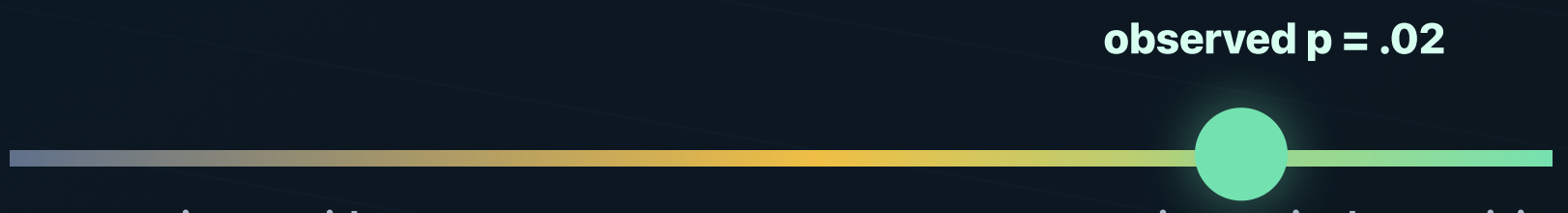
Define the population, metric, slices, and stopping rule before the results arrive.

A null hypothesis does not prove two versions are equal. It prevents a lucky story from becoming a release story.

P-VALUES

A p-value is evidence, not permission.

How surprising would this result be if there were no real difference?



P-values are continuous. Their meaning still depends on the test plan, assumptions, sample, effect size, repeated looks, and how many comparisons the team inspected.

Do not turn one number into a ship button. Read it beside the actual size and risk of the change.

PRACTICAL SIGNIFICANCE

Real is not always worth it.

+0.02

Credible, but tiny

A huge sample may show a real average lift. It may still be invisible to users, slower to serve, or not worth migration risk.

6% → 2%

Small average, big consequence

A reduction in serious policy-boundary failures can be worth further validation even when the global score barely moves.

Statistical significance asks whether a measured difference is likely real. Practical significance asks whether it matters.

POWER AND MINIMUM DETECTABLE EFFECT

Decide the smallest win that matters.

Too small to act on

10% → 9%

Maybe better. Probably not enough benefit to justify a disruptive model migration.

→

Worth detecting

10% → 5%

A drop in serious failures that could change the product decision. Plan enough sample to see it.

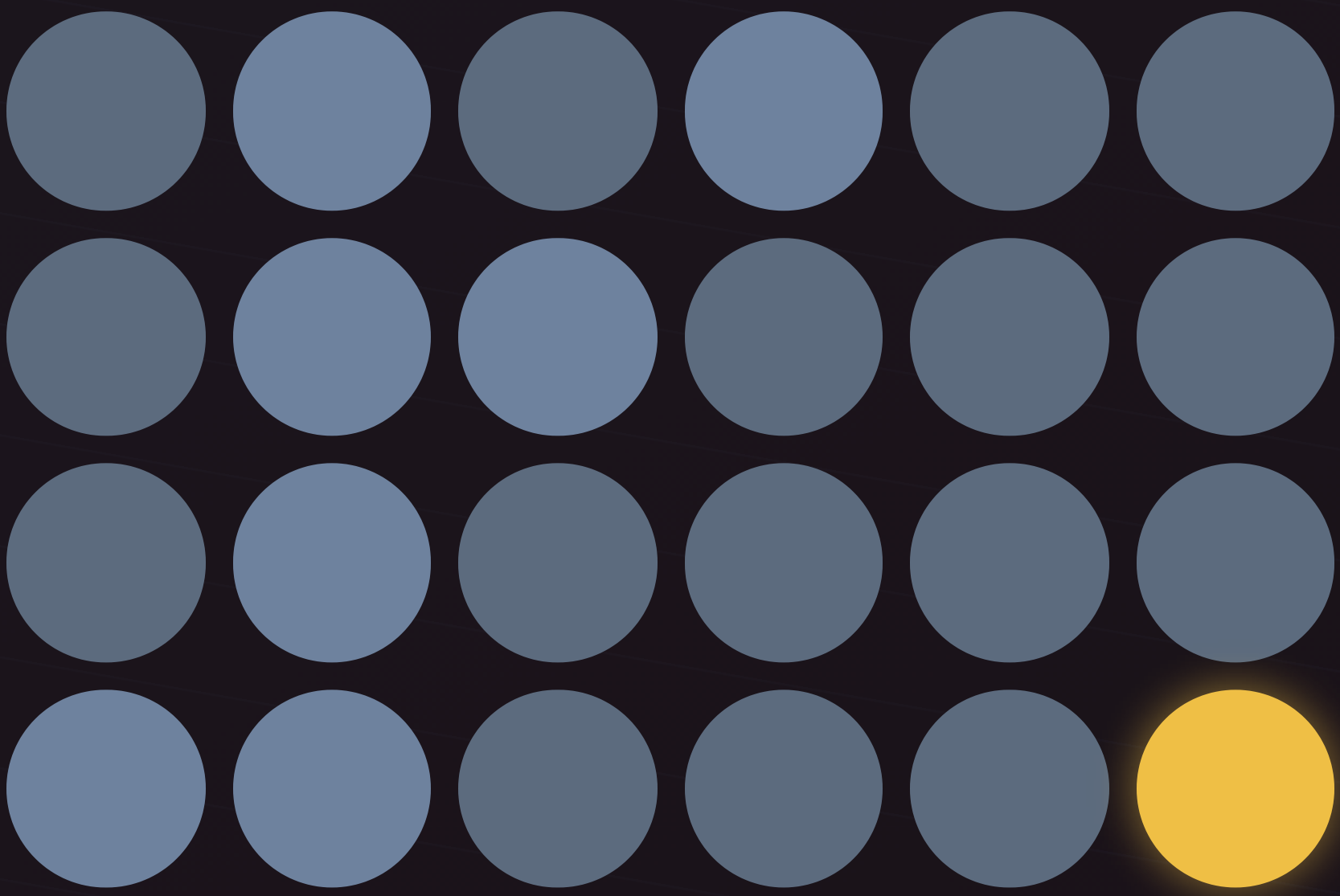
Power asks whether the eval has enough evidence to notice the change the team actually cares about.

A small eval that finds nothing may be underpowered, not proof that nothing changed.

MULTIPLE COMPARISONS

After 200 tries, a shiny win may be luck.

Prompt, model, tool, temperature, retry, judge, slice...



One winner is a lead, not a discovery. Freeze it and rerun it on fresh holdout cases the team did not tune against.

Exploration is useful. Treating its best score as proof is the trap.

MATCH THE QUESTION TO THE DATA

There is no universal test.

Same cases, numeric score

Paired comparison
Inspect each A-to-B difference.

Different groups, numeric score

Independent comparison
Random assignment and variance matter.

Pass / fail or outcome buckets

Categorical comparison
Ask whether the pattern of outcomes shifted.

Repeated users, topics, or runs

Cluster-aware analysis
Do not count dependent rows as new evidence.

Start with the data-generating process: assignment unit, measurement unit, and real source of dependence.

PRECISION, RECALL, AND F-SCORES

Pick the score for the cost of being wrong.

F0.5

Precision first

False alarms are expensive. Flag fewer things, but be right when you do.

F1

Balanced

False alarms and missed issues have roughly similar costs.

F2

Recall first

Missing a real safety or security problem is worse than extra review.

> Report **precision and recall beside the F-score.** A combined number should not hide which mistake the product has chosen to tolerate.

Metrics are product values expressed as math.

RELEASE EVIDENCE

The decision still belongs to the builder.

Effect size

Did the improvement move enough to matter?

Uncertainty

How wide is the plausible range around it?

Risk slices

Did a rare but important category regress?

Operational cost

Can the system be served, monitored, and rolled back?

Statistics are one part of a release decision. Quality, safety, cost, and reversibility still get a vote.

The goal is not certainty. It is evidence strong enough for the risk.

IcebergQA

AI-FIRST QUALITY WORK

PUT IT INTO PRACTICE

Build release evidence your product can actually use.

AI quality work is nuanced: the experiment, the data, the risks, and the decision all have to fit together. IcebergQA helps teams make that transition.

Talk to IcebergQA

Testing AI by Jason Arbon · Chapter 4: Statistical Tests for AI Quality