Hypothesis Testing I

- Housekeeping:
  - HW grades to be posted soon
  - Next lecture we'll do a bit more on hypothesis testing and then start exam review
  - I'll be making this year's exam over the weekend and will review everything on it
- Today:
  - Recap a bit from last lecture
  - A different approach to hypothesis testing → Signal Detection Theory

Sampling Distributions

- Why We Test: From Uncertainty to Decision
  - We’ve spent time learning:
    - How data vary (probability).
    - How to summarize uncertainty (inference, confidence intervals).
    - How to simulate what “random” looks like (permutation tests).
    - How important is the difference?
  - Today marks a shift: inference becomes decision-making.
  - Framing question: When the world is noisy, how do we decide if something is “real”?

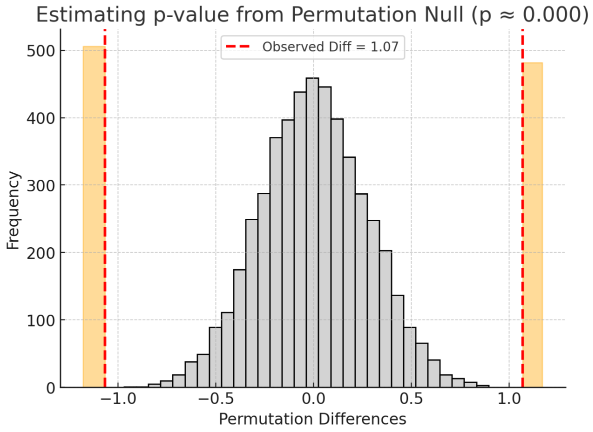
Summary

Concept	Question it answers	DS connection
p-value	"Is there signal beyond chance?"	Detects presence
Effect size	"How strong is the signal?"	Measures predictability
CI	"How stable is that estimate?"	Quantifies reliability

- Make sure the relationships you expect are real and stable
- Make sure you have enough data
- Make sure the effect is big enough to be meaningful
- Can this be answered with a classical statistical test, is machine learning even necessary?

Recap from Last Lecture

- Inference answers: What range of outcomes are plausible if nothing special is happening?
- Confidence intervals tell us how much wiggle room noise can explain.
- Permutation tests showed: random shuffling generates an expected “null” world.
- Effect size tells us *how big* the observed difference is — the *practical* impact beyond mere statistical detectability.
- **Hypothesis testing formalizes that process into a repeatable, communicable decision rule.**



Description → Inference → Action

Stage	Question	Data Science Example
Description (EDA)	What do we see?	Mean CTR = 3.2 %
Inference	What range is plausible?	3.2 % ± 0.4 % margin of error
Decision (Testing)	Should we act?	Launch new design?

- EDA → explore; Inference → quantify; Testing → decide.
- Testing becomes crucial when decisions have cost or risk.
- This mirrors the DS workflow: exploration → model → evaluation → deployment.

Why Formal Testing Exists

- Simulation builds intuition; formal testing provides standards (α levels, p-values).
- Science, healthcare, industry need a common protocol for uncertainty.
- Core ingredients we'll formalize next:
  - Competing claims → H_0 and H_1
  - Evidence → test statistic
  - Benchmark → sampling distribution
  - Decision rule → compare to α
  - Goal = defensible decisions that balance error.

Hypothesis Tests Have the Same Core

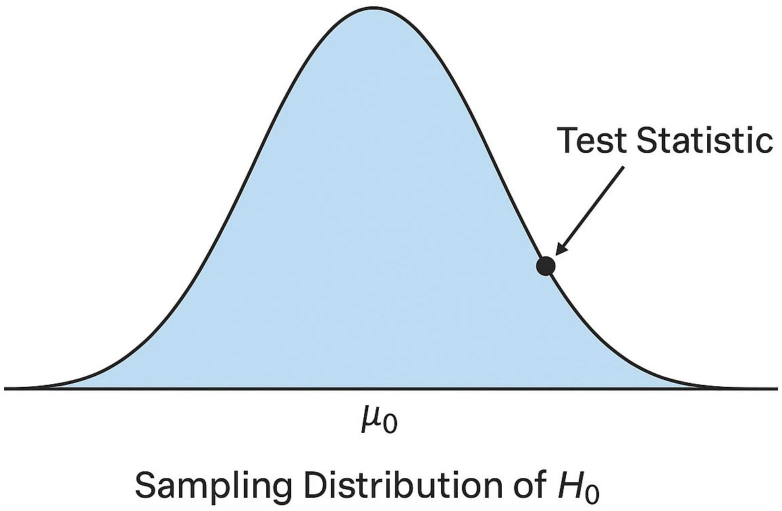
- For any test → Define two competing claims
  - H_0 ("null"): no real effect, pattern due to noise
  - H_1 ("alternative"): a real effect exists
- Collect data → compute a test statistic
- Assume H_0 is true → know what results are typical by chance
- Compare our observed statistic to that "null" world
- Decide: is this result too extreme to attribute to noise?

Competing Explanations

- H_0: status quo, “nothing special happening”
- H_1: there’s an effect, difference, or association
- Example:
  - A/B test: H_0 = mean CTR_A = mean CTR_B
  - H_1 = mean CTR_A ≠ mean CTR_B
- Always phrase in terms of population parameters, not samples.
- You can test
  - directional (one-tailed) → greater than | less than
  - non-directional (two-tailed) claims → not equal to

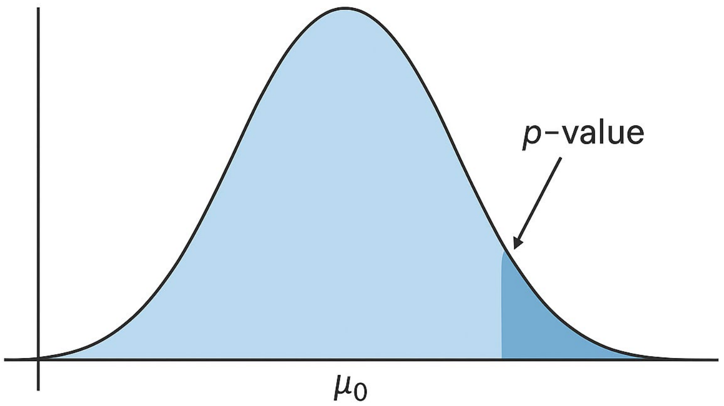
How Far Do Our Data Deviate from H_0

- Sampling distribution = distribution of the test statistic if we repeated sampling infinitely under H_0.
  - Permutation tests showed this empirically — now we model it analytically.
- Tails of the distribution = “rare” outcomes under H_0.
- The test statistic summarizes the evidence against H_0.
- It could be a mean difference, a correlation, a count ratio, etc.
- The statistic converts data → single number reflecting deviation.
- Large |statistic| ⇒ data are far from what H_0 predicts.
- Each test type defines its own statistic (t, z, χ^2, F…).



How Surprising is Our Result?

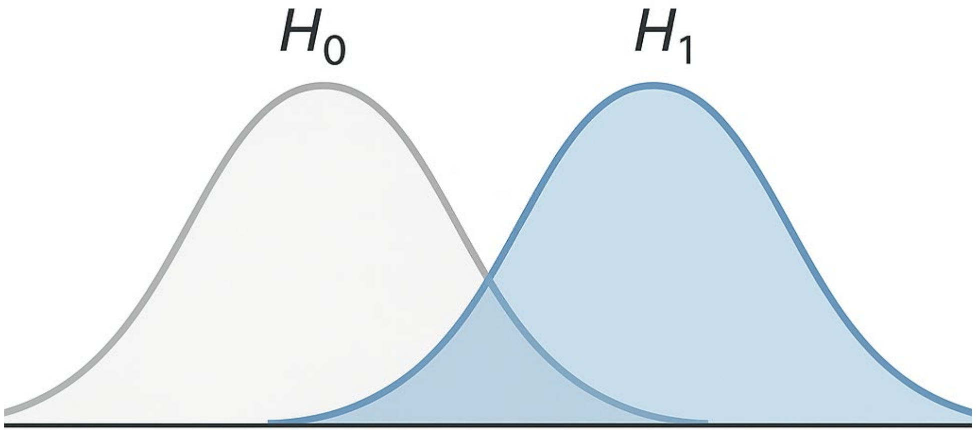
- The p-value = probability of observing a test statistic as extreme (or more) than ours if H_0 were true.
- Small p → result is rare under H_0 → stronger evidence against H_0.
- Important: $p \neq$ "probability H_0 is true."
- Convention: $\alpha = 0.05$ (arbitrary, not sacred).
- Decision rule:
  - If $p < \alpha$ → reject H_0 (evidence for effect).
  - If $p \geq \alpha$ → fail to reject H_0 (insufficient evidence).
- p-values quantify *surprise*, not *truth.* Even a small p doesn't make H_0 impossible — just implausible



Sampling Distribution of H_0

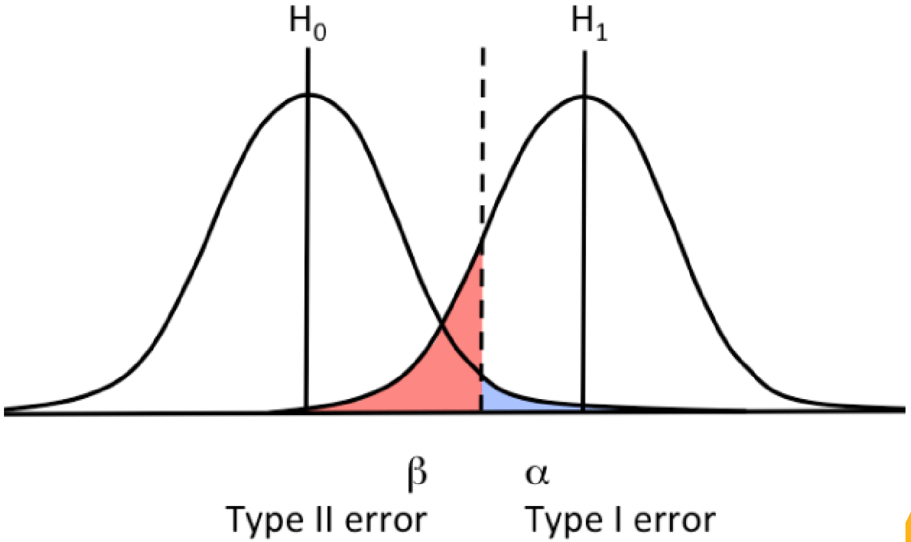
The Decision Framework Intro

- Each curve (H_0 and H_1) shows what your test statistic (like a difference in means) would look like if you ran the experiment many times.
- The x-axis is the value of that test statistic (e.g., the difference between conversion rates in Group A and Group B).
- The y-axis is the probability density of getting that value, assuming H_0 or H_1 is true.
- They are distributions of possible sample outcomes, *not* a histogram of your actual data



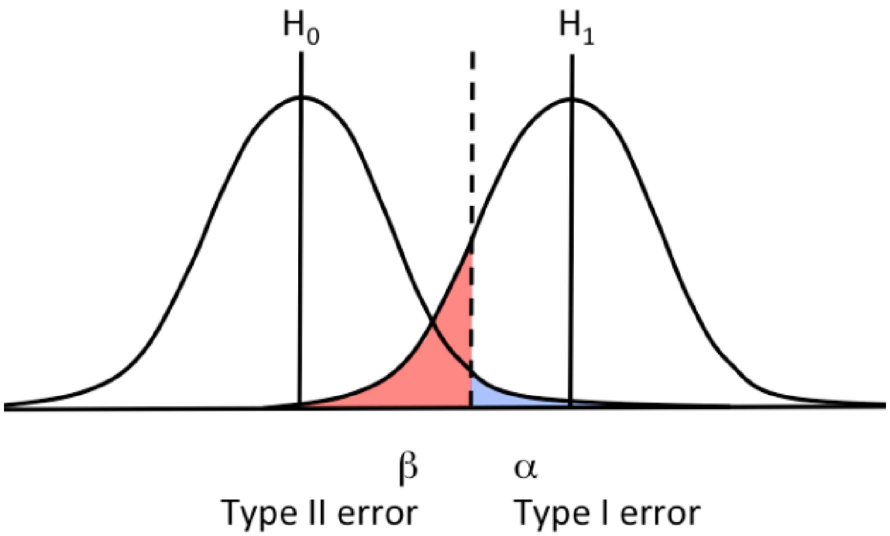
Threshold and alpha

- A decision threshold (vertical line) is the value of the test statistic that marks the cutoff for rejecting H_0.
  - Everything to the right of the threshold = “significant” (reject H_0).
  - Everything to the left = “not significant” (fail to reject H_0).
- Alpha (α) – **Type I Error**: Probability of rejecting H_0 when it’s true.
  - Interpretation: **False positive** — seeing a “significant” effect that’s really just noise.
  - In the plot:
    - The blue (or right) shaded tail under the H_0 curve beyond the decision threshold.
    - Represents results so extreme that you’d call them “significant,” even though H_0 was true.
  - Typical value: $\alpha = 0.05$ → willing to be wrong 5% of the time when no real effect exists.
  - Analogy: You raise a false alarm — “There’s a fire!” when there isn’t.



Beta, power

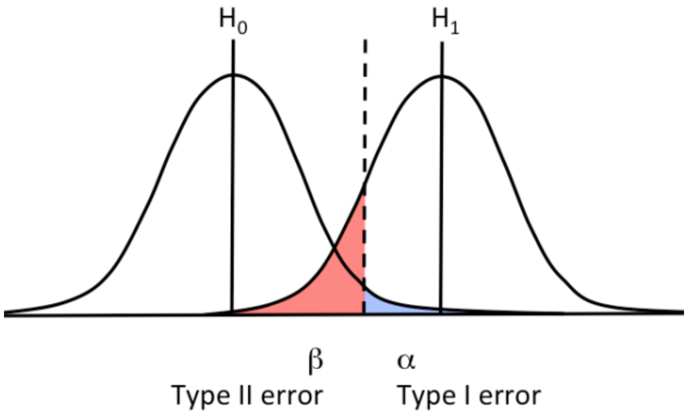
- Beta (β) – **Type II Error**
- Definition: Probability of failing to reject H_0 when H_1 is actually true.
- Interpretation: **False negative** — missing a real effect.
- In the plot:
  - The red shaded area under the H_1 curve to the left of the decision threshold.
  - Represents cases where the true effect exists, but your data don't look extreme enough to detect it.
- Analogy: You miss a real signal — "No fire," when smoke detectors were right.
- Power = 1 – β
  - The proportion of the H_1 curve to the right of the threshold (correct rejections of H_0).
  - Represents your ability to detect a true effect when it exists.
  - Increases when:
    - The true effect size is larger (H_1 curve moves farther from H_0).
    - Sample size is larger (distributions get narrower).
    - You relax α (allow more false positives).



α and β trade off: reducing false alarms (smaller α) increases missed detections (larger β).

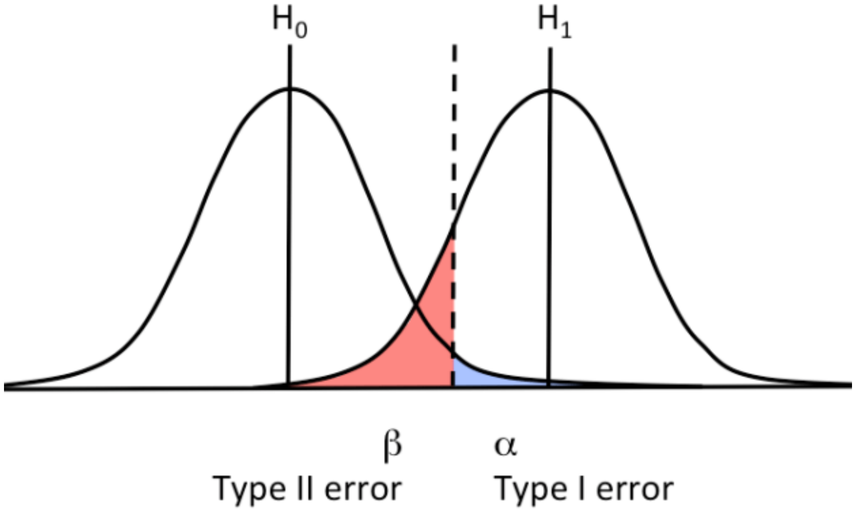
Summary

Region	True World	Decision	Probability	Meaning
Left of threshold under H_0	H_0 true	Fail to reject H_0	**$1 - \alpha$**	Correct "no effect" call
Right tail under H_0	H_0 true	Reject H_0	α	False positive
Left tail under H_1	H_1 true	Fail to reject H_0	β	Missed detection
Right of threshold under H_1	H_1 true	Reject H_0	**$1 - \beta$**	Correct detection (power)



Every Decision Has a Cost

- Hypothesis testing forces explicit trade-offs:
  - Type I error (α): false alarm → acting on noise.
  - Type II error (β): missed detection → ignoring a real effect.
- Each context values these errors differently.
- No test eliminates error — we choose how much risk to tolerate.



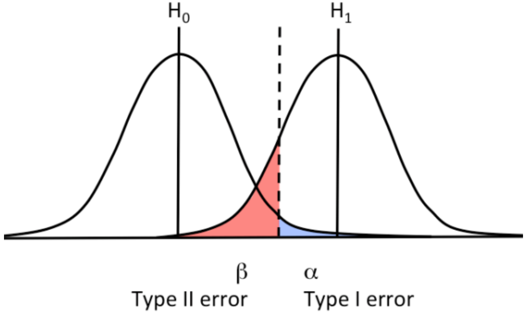
Hypothesis Tests & Medical Tests

- Statistical world: H_0 vs H_1.
- Diagnostic world: “no disease” vs “disease present.”

Reality	Decision: Positive (Reject H_0)	Decision: Negative (Fail to Reject H_0)
No disease (H_0)	Type I error (α) → False Positive	$1 - \alpha$ → True Negative
Disease present (H_1)	$1 - \beta$ → True Positive	Type II error (β) → False Negative

Quantifying Our Decisions

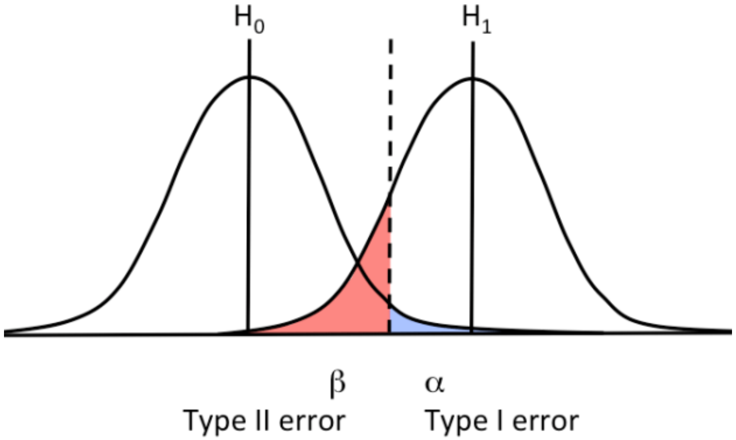
		Predicted	
		Positive	Negative
Predictive	Positive	**True Positive** TP	**False Negative** FN
	Negative	**False Positive** FP	**True Negative** TN



- Sensitivity = 1 – β = ability to catch true positives.
  - In hypothesis testing, this is power.
- Specificity = 1 – α = ability to correctly reject false alarms.
- False Positive Rate (FPR) = α.
- False Negative Rate (FNR) = β.

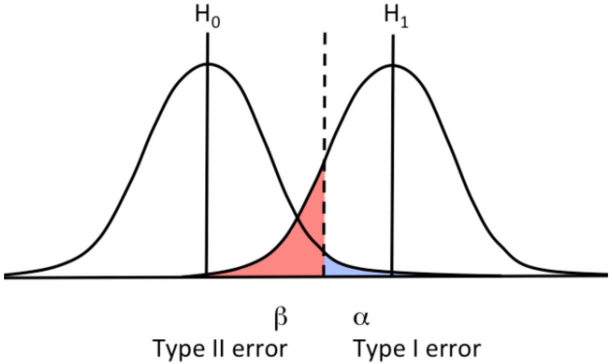
Threshold Shift the Balance

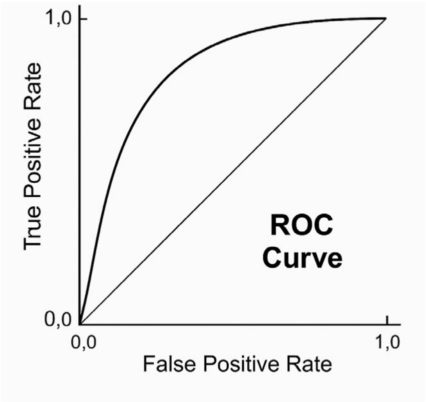
- The decision threshold (critical value) controls α and β.
- Move it right → fewer false positives (↓α) but more misses (↑β).
- Move it left → catch more signals (↓β) but risk more false alarms (↑α).
- In A/B tests, fraud detection, or classification, this is your cutoff.
- Choosing it = choosing what kind of mistake you can live with.



ROC: Seeing All Thresholds at Once

- ROC (Receiver Operating Characteristic) curve = plots all thresholds.
- x-axis: False Positive Rate (α).
- y-axis: True Positive Rate (1−β) = Sensitivity.
- Each point = one possible decision threshold.
  - Area Under Curve (AUC): overall ability to separate signal from noise.
- Hypothesis testing → picks one α; ROC → shows all α–β trade-offs.





Choosing the Right Balance

- Upper-left corner = perfect classifier (FPR = 0, TPR = 1).
- Diagonal line = random guessing (AUC = 0.5).
- Higher curve → better separation between H_0 and H_1.
- In practice, the optimal threshold depends on:
  - Cost of false alarms vs misses.
  - Base rates (how common the positive class is).
- Same logic as choosing α and β in testing.

UMBC

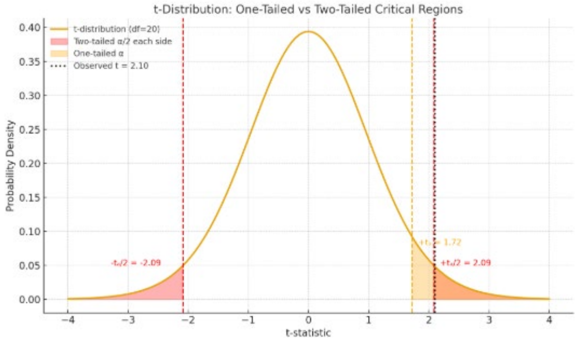
Summary

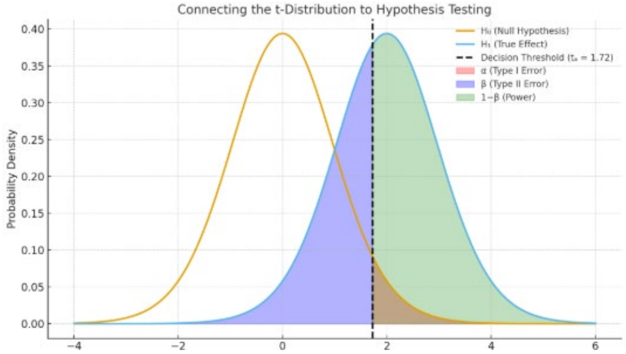
- α, β, power, sensitivity, and specificity all describe uncertainty under noise.
- Thresholds are how we turn uncertainty into action.
- ROC curves generalize hypothesis testing to many thresholds.
- Next: applying these ideas to real-world models and interpreting statistical significance vs practical value.

The t-Test: Hypothesis Testing

- The t-test is just one implementation of our hypothesis-testing logic.
- Example: compare two sample means (A vs. B).
- Null (H_0): no difference → $\mu_a = \mu_\beta$.
- Alternative (H_1): difference exists → $\mu_a \neq \mu_\beta$.
- Compute test statistic:

$$t = \frac{\bar{x}_A - \bar{x}_B}{SE_{\text{diff}}}$$





Each Test Is a Binary Classifier

- Each hypothesis test is a binary decision: “signal” or “no signal.”
- “Reject H_0” = predict “signal present.”
- “Fail to reject H_0” = predict “no signal.”

Reality	Decision: Reject H_0	Decision: Fail to Reject H_0
H_0 true	Type I (α)	Correct ($1-\alpha$)
H_1 true	Correct ($1-\beta$)	Type II (β)

- Think of every t-test as a one-point classifier on the ROC curve — one threshold, one trade-off

Changing α = Changing the Threshold

- t-tests “reject” when |t| exceeds the critical value.
- Smaller α → threshold moves further into the tail → fewer false positives, more misses.
- Larger α → threshold moves closer to center → catch more signals, risk more false alarms.
- Same logic applies in all domains (disease detection, fraud alerts, etc.).
- Setting $\alpha = 0.05$ is arbitrary — you’re just choosing a point on a trade-off curve

What if We Varied α ?

- You could, in theory, move the decision threshold (t-critical) around (α):
  - If you set α = 0.10 → smaller threshold → more rejections (higher power, more false alarms).
  - If you set α = 0.01 → larger threshold → fewer rejections (lower power, fewer false alarms).
- In practice, a t-test gives you one decision at one false-positive tolerance — α = 0.05
- But mathematically, we can imagine sliding that α threshold up and down, just like changing a classifier threshold.
- Each setting gives us a different trade-off between false alarms (α) and missed detections (β).
  The curve that traces all those trade-offs is the ROC curve — it's the t-test generalized.
- So power analysis is just ROC analysis in disguise — we're studying how well our test separates signal from noise."