Lecture 9: Hypothesis Testing

p-Values · p-Hacking · Power Analysis · Permutation Tests · Multiple Testing

Previously, on Lecture 8...

Confidence interval: A random interval $[L, U]$ with $P(L \leq \theta \leq U) = 1 - \alpha$. The *procedure* is 95%.

Wald CI: $\hat{\theta} \pm z_{\alpha/2} \cdot \text{SE}$. Works when CLT kicks in. Use Wilson for proportions.

t-interval: Use when σ unknown and n small. Heavier tails $\Rightarrow$ wider CI.

Bootstrap: Resample with replacement $\Rightarrow$ SE and CI without formulas. BCa is the gold standard.

Two paths: Analytical when formulas exist; bootstrap when they don't. Same goal.

Today: We can estimate and quantify uncertainty. Now: is the effect **real** or **noise**?

Part I: The Logic of Testing

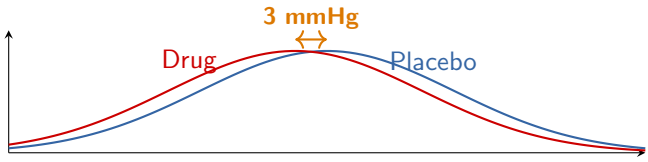
Is 3 mmHg a real effect or just noise?

The Drug Trial

Scenario: A pharmaceutical company tests a new blood pressure drug.

- $n = 100$ patients per group (drug vs. placebo)
- Placebo group mean: $\bar{X}_P = 140$ mmHg
- Drug group mean: $\bar{X}_D = 137$ mmHg
- Pooled SD: $S = 12$ mmHg

Question: Is this 3 mmHg reduction *real*, or could random chance produce it?



Innocent Until Proven Guilty

Courtroom

1. **Assume** defendant is innocent
2. **Examine** the evidence
3. **Convict** only if evidence is overwhelming

"Innocent until proven guilty"

Verdict: guilty / *not proven guilty*
(never "proven innocent")

Statistics

1. **Assume** H_0: no effect
2. **Compute** how surprising the data is
3. **Reject** H_0 only if the data is extremely unlikely under H_0

"No effect until proven otherwise"

Verdict: reject H_0 / *fail to reject* H_0
(never "H_0 is true")

H_0 and H_1

H_0 (null hypothesis)

"Nothing is happening"

The default / boring explanation

Examples:
Drug has no effect: $\mu_D = \mu_P$
The coin is fair: $p = 0.5$
Treatment A = Treatment B

H_1 (alternative)

"Something is happening"

What we want evidence *for*

One-sided: $\mu_D < \mu_P$
(drug lowers BP)
Two-sided: $\mu_D \neq \mu_P$
(drug changes BP)

Key: We never "prove" H_0. We either **reject** it (strong evidence against) or **fail to reject** it (insufficient evidence). Absence of evidence $\neq$ evidence of absence.

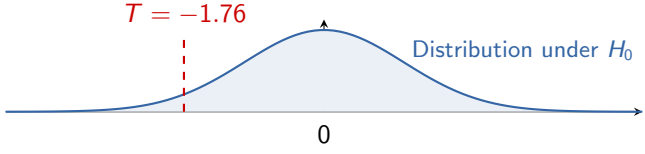
The Test Statistic

A **test statistic** T compresses the data into one number measuring "distance from H_0".

$$T = \frac{\bar{X}_D - \bar{X}_P}{\text{SE}} = \frac{\bar{X}_D - \bar{X}_P}{S\sqrt{2/n}}$$

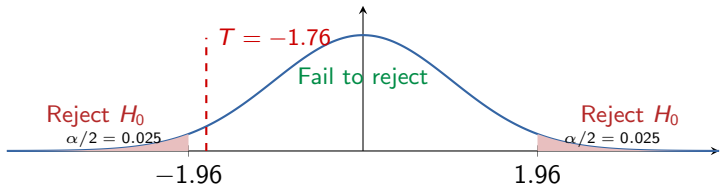
Drug trial: $$T = \frac{137 - 140}{12\sqrt{2/100}} = \frac{-3}{1.70} = -1.76$$

Under H_0: $T \sim N(0,1)$ approximately (by CLT, Lecture 7). Large $|T|$ = data far from H_0 = evidence against H_0.



The Rejection Region

Choose a **significance level** α (typically 0.05) *before* looking at the data.



Drug trial: $|T| = 1.76 < 1.96 \Rightarrow$ **fail to reject** H_0 at $\alpha = 0.05$.

The 3 mmHg reduction is *not statistically significant* at the 5% level.

The Hypothesis Testing Recipe

Step 1: State H_0 and H_1 *before* looking at the data.

Steps 1–2 must happen **before** you see the data. Otherwise you risk p-hacking (more on this later).

The Hypothesis Testing Recipe

Step 1: State H_0 and H_1 *before* looking at the data.

Step 2: Choose α (typically 0.05). This is your tolerance for false alarms.

Steps 1–2 must happen **before** you see the data. Otherwise you risk *p*-hacking (more on this later).

The Hypothesis Testing Recipe

Step 1: State H_0 and H_1 *before* looking at the data.

Step 2: Choose α (typically 0.05). This is your tolerance for false alarms.

Step 3: Compute the test statistic $T = (\hat{\theta} - \theta_0)/\text{SE}$.

Steps 1–2 must happen **before** you see the data. Otherwise you risk *p*-hacking (more on this later).

The Hypothesis Testing Recipe

Step 1: State H_0 and H_1 *before* looking at the data.

Step 2: Choose α (typically 0.05). This is your tolerance for false alarms.

Step 3: Compute the test statistic $T = (\hat{\theta} - \theta_0)/\text{SE}$.

Step 4: Find the p-value: $p = P(|T| \geq |T_{\text{obs}}| \mid H_0)$.

Steps 1–2 must happen **before** you see the data. Otherwise you risk *p*-hacking (more on this later).

The Hypothesis Testing Recipe

Step 1: State H_0 and H_1 *before* looking at the data.

Step 2: Choose α (typically 0.05). This is your tolerance for false alarms.

Step 3: Compute the test statistic $T = (\hat{\theta} - \theta_0)/\text{SE}$.

Step 4: Find the p-value: $p = P(|T| \geq |T_{\text{obs}}| \mid H_0)$.

Step 5: Decide: if $p \leq \alpha$, reject H_0. Otherwise, fail to reject.

Steps 1–2 must happen **before** you see the data. Otherwise you risk *p*-hacking (more on this later).

The Hypothesis Testing Recipe

Step 1: State H_0 and H_1 *before* looking at the data.

↓

Step 2: Choose α (typically 0.05). This is your tolerance for false alarms.

↓

Step 3: Compute the test statistic $T = (\hat{\theta} - \theta_0)/\text{SE}$.

↓

Step 4: Find the p-value: $p = P(|T| \geq |T_{\text{obs}}| \mid H_0)$.

↓

Step 5: Decide: if $p \leq \alpha$, reject H_0. Otherwise, fail to reject.

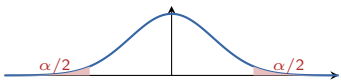
↓

Step 6: Report **effect size** + **confidence interval** + p-value. Never just "$p < 0.05$".

Steps 1–2 must happen **before** you see the data. Otherwise you risk *p*-hacking (more on this later).

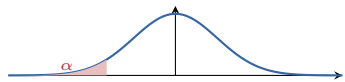
One-Sided vs Two-Sided Tests

Two-sided: $H_1 : \mu \neq \mu_0$



Reject if $|T| > z_{\alpha/2}$

One-sided: $H_1 : \mu < \mu_0$



Reject if $T < -z_\alpha$

When to use which?

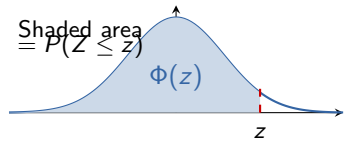
Two-sided: "Is there any difference?" Default choice. Used when the effect could go either way.

One-sided: "Is the drug *better*?" Only when direction is specified *before* seeing data.

One-sided has more power (lower threshold) but only detects effects in one direction.

Reading the z-Table: How It Works

Standard Normal cumulative distribution: $\Phi(z) = P(Z \le z)$



The table gives $\Phi(z)$: the area to the **left** of z. Row = first digit, column = second decimal.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812

Row 1.7 + Col .06

Read: $z = 1.76 \Rightarrow$ row **1.7**, column **.06** $\Rightarrow \Phi(1.76) = 0.9608$.

Using the z-Table: Two Key Tasks

Task 1: Find a p-value

Given $T = 1.76$, find $P(Z > 1.76)$.

Step 1: Look up $\Phi(1.76) = 0.9608$

Step 2: Right tail:
$P(Z > 1.76) = 1 - 0.9608 = 0.0392$

Step 3: For two-sided test:
$p = 2 \times 0.0392 = \mathbf{0.0784}$

(Multiply by 2 because extreme values in *either* tail count as evidence.)

Task 2: Find a critical value

For $\alpha = 0.05$ (two-sided), find $z_{\alpha/2}$.

Step 1: We need $P(Z > z) = 0.025$,
i.e. $\Phi(z) = 1 - 0.025 = 0.975$

Step 2: Scan the table for 0.9750

Step 3: Found at row **1.9**, col **.06**:
$\Phi(1.96) = 0.9750$
$\Rightarrow z_{0.025} = \mathbf{1.96}$

(This is where the famous 1.96 comes from!)

Common critical values: $z_{0.05} = 1.645$ (90% CI) $z_{0.025} = 1.960$ (95% CI) $z_{0.005} = 2.576$ (99% CI)

The Student's t-Distribution

William Sealy Gosset, head brewer at Guinness, Dublin, 1908

The problem

Gosset tested barley and hops with **tiny samples** ($n = 3$–4).

He noticed: using S instead of σ in $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ made the distribution **wider** than $N(0,1)$.

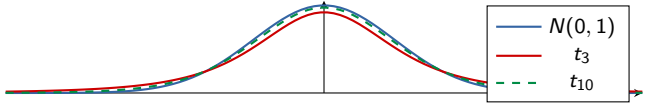
With small n, S is a noisy estimate of σ — extra uncertainty!

The solution

He derived the exact distribution: the t-distribution with df $= n - 1$.

Why "Student"?
Guinness banned employees from publishing. Gosset published under the pseudonym **"Student"** in *Biometrika* (1908).

The name stuck forever.



z vs t: When to Use Which

When σ is **known**: $$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \Rightarrow \text{use } z\text{-table}$$

When σ is **unknown**: $$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \Rightarrow \text{use } t\text{-table}$$

z vs t: When to Use Which

When σ is **known**: $$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \Rightarrow \text{use } z\text{-table}$$

When σ is **unknown**: $$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \Rightarrow \text{use } t\text{-table}$$

In practice:

σ is almost never known!

So we almost always use t.

The drug trial used z because $n = 100$ is large enough that $t_{99} \approx N(0,1)$.

Rule of thumb:

$n \geq 30$: t and z give similar results

$n < 30$: t gives wider intervals (correctly accounts for uncertainty)

$n < 10$: t is essential!

Reading the t-Table

Organized differently from z-table: rows = df, columns = tail probabilities

	Upper-tail probability $P(T > t)$				
df	0.100	0.050	0.025	0.010	0.005
3	1.638	2.353	3.182	4.541	5.841
10	1.372	1.812	2.228	2.764	3.169
24	1.318	1.711	2.064	2.492	2.797
30	1.310	1.697	2.042	2.457	2.750
60	1.296	1.671	2.000	2.390	2.660
∞	1.282	1.645	1.960	2.326	2.576

Row 24, Col 0.025

Coffee shop: $n = 25$, **df** $= 24$

For 95% two-sided: $\alpha/2 = 0.025$ column

$\Rightarrow t_{0.025,24} =$ **2.064** (compare $z_{0.025} = 1.960$, **5% larger**)

Key observations

df $\uparrow \Rightarrow$ critical value $\downarrow$ (more data = less uncertainty)

df $= 3$: $t_{0.025} = 3.182$ (huge!) df $= \infty$: $= 1.960 = z_{0.025}$

Bottom row = the z-table!

Part II: Two Ways to Be Wrong

False alarms and missed discoveries

Type I and Type II Errors

	Fail to reject H_0	**Reject H_0**
H_0 true	**Correct** True negative $P = 1 - \alpha$	**Type I Error** False alarm $P = \alpha$
H_0 false	**Type II Error** Missed detection $P = \beta$	**Correct** Power $= 1 - \beta$

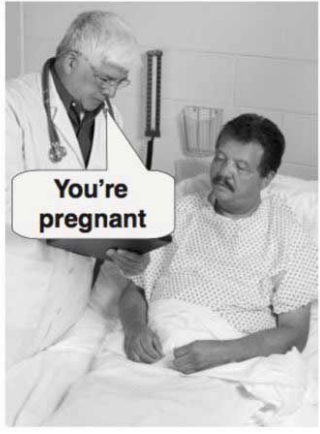
Type I and Type II Errors

	Fail to reject H_0	**Reject H_0**
H_0 true	**Correct** True negative $P = 1 - \alpha$	**Type I Error** False alarm $P = \alpha$
H_0 false	**Type II Error** Missed detection $P = \beta$	**Correct** Power $= 1 - \beta$

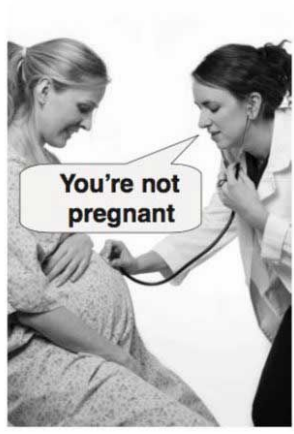
Type I (false positive):
Telling a man he's pregnant.
Convict an innocent person.
Approve a useless drug.

Type II (false negative):
Telling a pregnant woman she's not.
Free a guilty person.
Miss an effective drug.

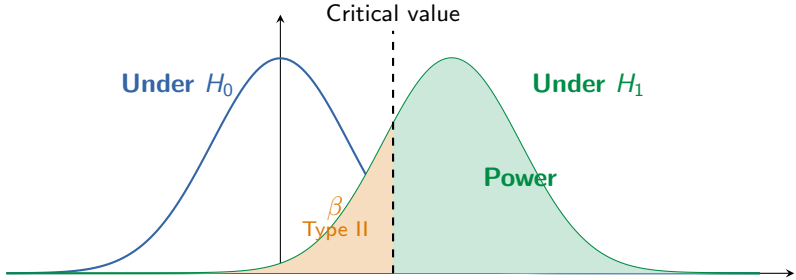
Type I error
(false positive)



Type II error
(false negative)



Visualizing the Two Errors



Lowering α (moving the cutoff right) $\Rightarrow$ fewer false alarms but more missed effects (β grows).
There's always a **trade-off** between the two errors.

Part III: The p-Value

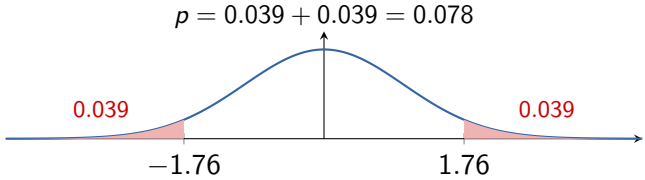
The most used — and most abused — number in science

The p-Value: Definition

The **p-value** is the probability of observing data *this extreme or more*,

assuming H_0 is true:

$$p \quad = \quad P\big(\,|T| \quad \geq \quad |T_{\text{obs}}| \quad \big| \quad H_0\big)$$



Drug trial: $p = 0.078 > 0.05 = \alpha \;\Rightarrow$ fail to reject H_0.

Rule: Reject H_0 if and only if $p \leq \alpha$.

What the p-Value Is NOT

✗ "$p = 0.03$ means there's a 3% chance H_0 is true."

Reality: $p = P(\text{data} \mid H_0)$, not $P(H_0 \mid \text{data})$. The transpose fallacy!

What the p-Value Is NOT

✗ "$p = 0.03$ means there's a 3% chance H_0 is true."
Reality: $p = P(\text{data} \mid H_0)$, not $P(H_0 \mid \text{data})$. The transpose fallacy!

✗ "$1 - p$ = probability the effect is real."
Reality: p says nothing about $P(H_1)$. For that you need Bayes.

What the p-Value Is NOT

✗ "$p = 0.03$ means there's a 3% chance H_0 is true."
Reality: $p = P(\text{data} \mid H_0)$, not $P(H_0 \mid \text{data})$. The transpose fallacy!

✗ "$1 - p$ = probability the effect is real."
Reality: p says nothing about $P(H_1)$. For that you need Bayes.

✗ "$1 - p$ = probability the result will replicate."
Reality: Replication depends on effect size, n, and design. Not on p alone.

What the p-Value Is NOT

✗ "$p = 0.03$ means there's a 3% chance H_0 is true."
Reality: $p = P(\text{data} \mid H_0)$, not $P(H_0 \mid \text{data})$. The transpose fallacy!

✗ "$1 - p$ = probability the effect is real."
Reality: p says nothing about $P(H_1)$. For that you need Bayes.

✗ "$1 - p$ = probability the result will replicate."
Reality: Replication depends on effect size, n, and design. Not on p alone.

✗ "Smaller p = bigger effect."
Reality: p depends on *both* effect size and n. With $n = 10^6$, any tiny effect gives $p \approx 0$.

Simply Statistics

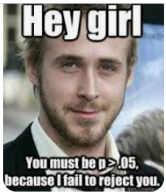
Statistics meme: Sad …

Reddit

Sigh... p-values... : r/m…

Reddit

probability of the null…

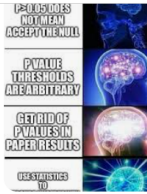


Medium

Easy p-value interp…

x.com

Customary p-value …



x.com

The holiday p valu…

JacqueLENS PhD

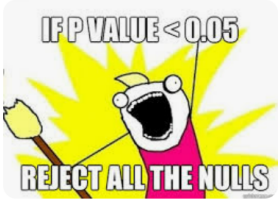
JacqueLENS PhD - Perspectives on the …

LinkedIn

hypothesis probabilit…

Facebook

A short video about p-valu…



Medium

P-value — A measure of surprise. I…

Arbor

p-value

Statistical vs Practical Significance

With enough data, *any* nonzero effect becomes "statistically significant."

Scenario A

$n = 100{,}000$ per group
Drug lowers BP by **0.1 mmHg**
$p < 0.001$

Statistically significant
Clinically meaningless

Would you take a pill every day for 0.1 mmHg?

Scenario B

$n = 15$ per group
Drug lowers BP by **8 mmHg**
$p = 0.12$

Not significant
Clinically important

Study was underpowered, not the drug's fault.

Always report: **(1) effect size** (e.g., Cohen's $d = \frac{\mu_1 - \mu_2}{\sigma}$, the difference in SD units), **(2) CI**, **(3) p-value**.
Ask "**How big** is the effect?", not just "Is it nonzero?"

Tests and CIs Are Two Sides of the Same Coin

Key duality: A two-sided test at level α and a $(1-\alpha)$ CI always give the same answer:

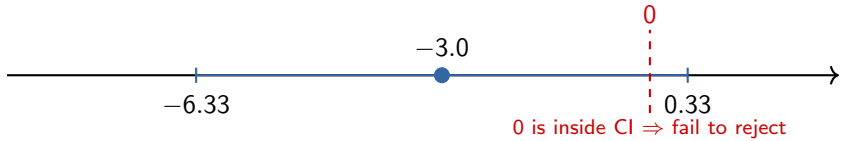
Reject $H_0 : \theta = \theta_0$ at level α $\iff$ θ_0 lies **outside** the $(1-\alpha)$ CI

Tests and CIs Are Two Sides of the Same Coin

Key duality: A two-sided test at level α and a $(1-\alpha)$ CI always give the same answer:

Reject $H_0 : \theta = \theta_0$ at level α $\iff$ θ_0 lies **outside** the $(1-\alpha)$ CI

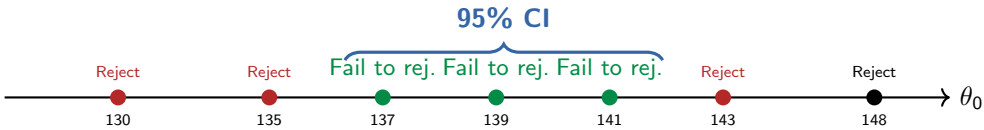
Drug trial check: $p = 0.078 > 0.05 \Rightarrow$ fail to reject.
95% CI for $\bar{X}_D - \bar{X}_P$: $-3 \pm 1.96 \times 1.70 = [-6.33,\ 0.33]$ — **contains 0**. Same conclusion!



CIs are more informative than tests: they tell you both *whether* to reject and *where* θ plausibly lies.

CI by Inversion: Collect All Non-Rejected θ_0

A CI can be built by "inverting" the test: try every possible θ_0 and keep those you **can't reject**.

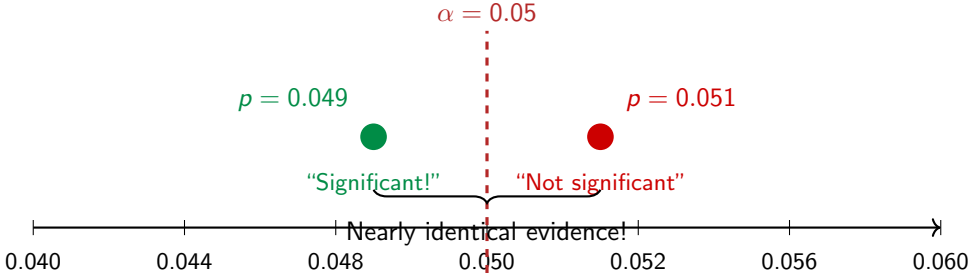


$$\text{CI} = \{\theta_0 : \text{we fail to reject } H_0 : \theta = \theta_0 \text{ at level } \alpha\}$$

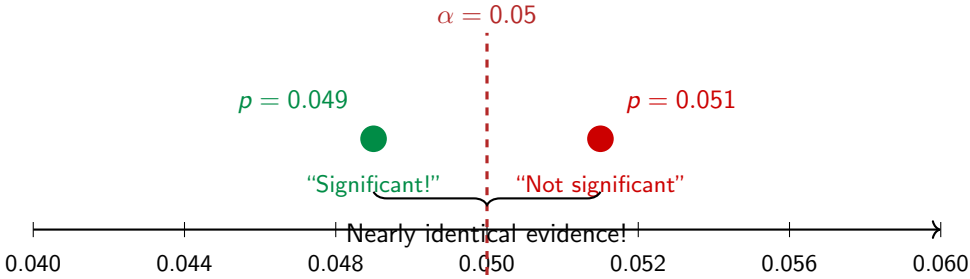
The CI is exactly the set of "plausible" parameter values — those consistent with the data.

For the Wald test, this recovers $\hat{\theta} \pm z_{\alpha/2} \cdot \text{SE}$. For other tests (LRT, score), it gives different — sometimes better — CIs.

$p = 0.049$ vs $p = 0.051$



$p = 0.049$ vs $p = 0.051$



The 0.05 threshold is a **convention**, not a law of nature.

ASA (American Statistical Association) **(2016):** "Scientific conclusions should not be based only on whether a p-value passes a specific threshold."

ASA (2019): "We recommend that declarations of 'statistical significance' be **abandoned**."

The Replication Crisis: Why This Matters

The problem

In 2015, only **36%** of 100 psychology studies replicated.

Similar failures in cancer biology, economics, and social science.

Cause: *p*-hacking, underpowered studies, publication bias.

The response

Pre-registration: commit to your analysis plan *before* seeing data.

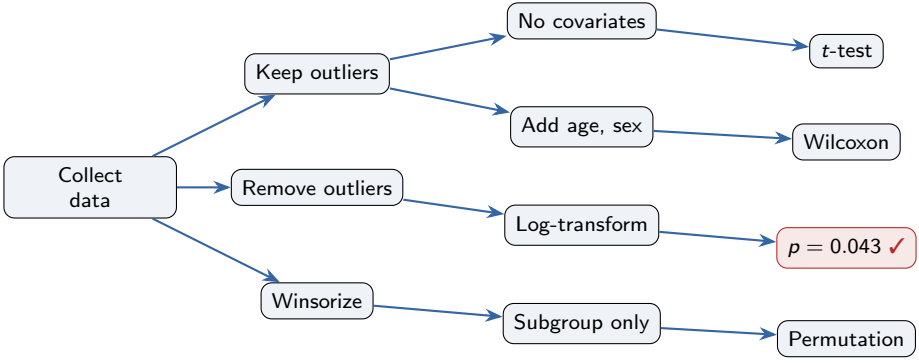
Report effect sizes & CIs, not just *p*-values.

Journals now require it. Science is self-correcting.

The biggest culprit? A practice called ***p*-hacking**. Let's see how it works...

p-Hacking: The Garden of Forking Paths

Definition: Trying multiple analyses until you find $p < 0.05$, then reporting only that one.

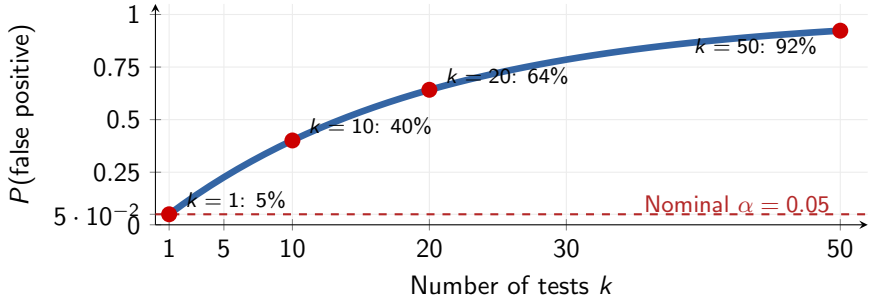


With enough "researcher degrees of freedom," you can *almost always* find $p < 0.05$ —
even when there is **no real effect**. This is the **garden of forking paths** (Gelman & Loken, 2013).

p-Hacking: How Bad Can It Get?

Thought experiment: The null is *true* ($\mu = 0$). A researcher tries k independent tests.

$$P(\text{at least one } p < 0.05 \text{ in } k \text{ tests}) = 1 - (1 - 0.05)^k = 1 - 0.95^k$$



With 20 tests, there's a **64% chance** of a false "discovery" — even with no real effect.

Real-World *p*-Hacking: FiveThirtyEight Demo

Hack Your Way to Scientific Glory (FiveThirtyEight, 2015): an interactive tool where you manipulate researcher choices to make *anything* "significant."

The setup:

Does the US economy do better under Democrats or Republicans?

Choose: which party, which metric, which years, which controls...

The antidote:

1. **Pre-register** your analysis plan before seeing data
2. **Report all analyses** you ran, not just the "significant" ones
3. **Correct for multiple comparisons** (Bonferroni, BH — covered later)
4. **Focus on effect sizes and CIs**, not just $p < 0.05$

Real-World p-Hacking: FiveThirtyEight Demo

Hack Your Way to Scientific Glory (FiveThirtyEight, 2015): an interactive tool where you manipulate researcher choices to make *anything* "significant."

The setup:

Does the US economy do better under Democrats or Republicans?

Choose: which party, which metric, which years, which controls...

cherry-pick →

The result:

You can "prove" *either* party is better, with $p < 0.05$, by choosing different combinations of the same data.

Same data, opposite conclusions!

Real-World p-Hacking: FiveThirtyEight Demo

Hack Your Way to Scientific Glory (FiveThirtyEight, 2015): an interactive tool where you manipulate researcher choices to make *anything* "significant."

The setup:

Does the US economy do better under Democrats or Republicans?

Choose: which party, which metric, which years, which controls...

cherry-pick →

The result:

You can "prove" *either* party is better, with $p < 0.05$, by choosing different combinations of the same data.

Same data, opposite conclusions!

The antidote:

1. **Pre-register** your analysis plan before seeing data
2. **Report all analyses** you ran, not just the "significant" ones
3. **Correct for multiple comparisons** (Bonferroni, BH — covered later)
4. **Focus on effect sizes and CIs**, not just $p < 0.05$

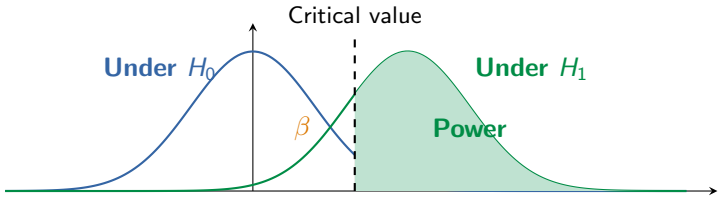
Part IV: Power Analysis

How to design a study that can actually find what you're looking for

Power: The Ability to Detect Real Effects

$$\text{Power} = P(\text{reject } H_0 \mid H_1 \text{ is true}) = 1 - \beta$$

Convention: aim for power $\geq$ 0.80 (80%).



What Determines Power?

1. Effect size $\delta \uparrow$

Bigger real effect
$\Rightarrow$ easier to detect
$\Rightarrow$ higher power

2. Sample size $n \uparrow$

More data $\Rightarrow$ narrower SE
$\Rightarrow$ distributions separate
$\Rightarrow$ higher power

3. Significance level $\alpha \uparrow$

Larger $\alpha \Rightarrow$ bigger rejection region
$\Rightarrow$ higher power (but more Type I)

4. Variability $\sigma \downarrow$

Less noise $\Rightarrow$ clearer signal
$\Rightarrow$ narrower distributions
$\Rightarrow$ higher power

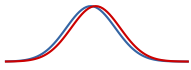
You control n and α. Nature controls δ and σ.
Power analysis = choosing n so that you can detect a meaningful δ.

Effect Size: Cohen's d

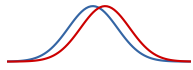
$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

Effect in units of standard deviations.

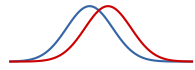
$d = 0.2$ (small)



$d = 0.5$ (medium)



$d = 0.8$ (large)



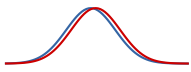
Drug trial: $d = 3/12 = 0.25$ (small).
Effect size is **independent of sample size** — it measures the *phenomenon*, not the study.

Effect Size: Cohen's d

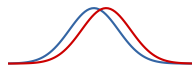
$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

Effect in units of standard deviations.

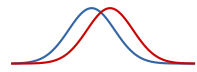
$d = 0.2$ (small)



$d = 0.5$ (medium)



$d = 0.8$ (large)



Drug trial: $d = 3/12 = 0.25$ (small).

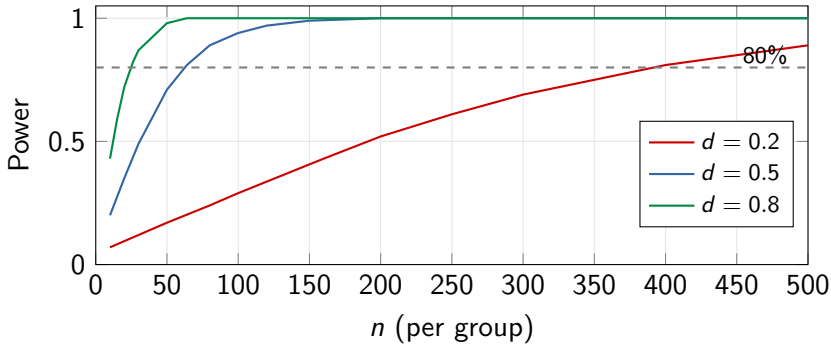
Effect size is **independent of sample size** — it measures the *phenomenon*, not the study.

Don't blindly trust 0.2 / 0.5 / 0.8! Cohen himself called them a "last resort." **In medicine:** $d = 0.1$ across millions saves thousands of lives. **In education:** $d = 0.8$ is extraordinarily rare — $d = 0.3$ is already large. Always define the **SESOI** (smallest effect size of interest) from domain knowledge.

Other effect sizes: Pearson's r, odds ratio, η^2 (ANOVA). Each context has its own.

Power Curves: Planning Your Study

Two-sample z-test, $\alpha = 0.05$, two-sided



Required n per group for 80% power: $d = 0.8$: $n \approx 26$ $d = 0.5$: $n \approx 64$ $d = 0.2$: $n \approx 394$

Sample Size Formula

$$n_{\text{per group}} = \frac{2\left(z_{\alpha/2} + z_{\beta}\right)^2}{d^2}$$

where $d = \delta/\sigma$ and $z_\beta = 0.84$ for 80% power.

Large effect
$d = 0.8$
$n = 26$

Medium effect
$d = 0.5$
$n = 64$

Small effect
$d = 0.2$
$n = 394$

An underpowered study wastes resources. If you can only recruit 50 patients,
you can only detect $d \geq 0.57$. Know this *before* you start collecting data.

Warning: Post-Hoc Power Is Meaningless

The mistake

Study finds $p = 0.12$.
Reviewer asks: "What was the power?"
Researcher computes power using the *observed* effect size.

Gets: "Power was 43%."
Concludes: "Study was underpowered."

Why it's circular

Observed power is just a 1-to-1 transform of the p-value!

$p > 0.05 \Leftrightarrow$ low observed power
$p < 0.05 \Leftrightarrow$ high observed power

It **adds no new information** beyond the p-value itself.

Power is a planning tool. Compute it *before* the study, using the **smallest effect size of interest** — not *after*, using the observed effect.

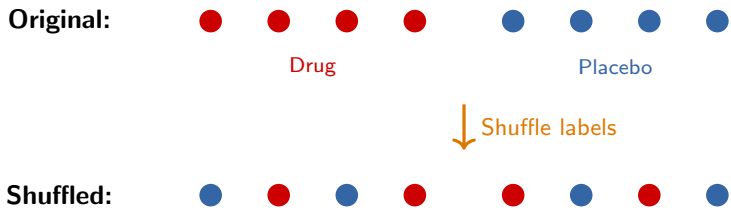
Part V: Permutation Tests

No distributional assumptions — let the data speak

If H_0 Is True, Labels Don't Matter

Under H_0: drug and placebo groups come from the *same* distribution.
So the group labels are **arbitrary** — shuffling them shouldn't change anything.

Intuition: Under H_0, swapping a patient between drug and placebo groups shouldn't change anything — the groups are interchangeable. This is called **exchangeability**.



Permutation Test Algorithm

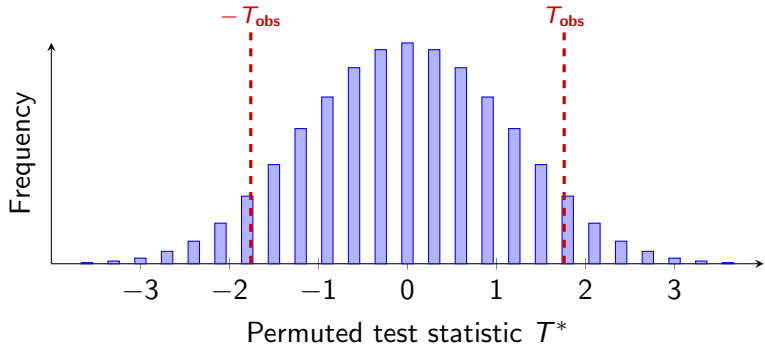
Algorithm:

1. Compute T_{obs} from the original data
2. Shuffle the group labels randomly (keeping group sizes fixed)
3. Compute T^* on the shuffled data
4. Repeat B times (e.g., $B = 10{,}000$)
5. p-value $= \dfrac{\#\{|T^*| \geq |T_{\text{obs}}|\} + 1}{B + 1}$

The "+1" in numerator and denominator ensures the p-value is never exactly 0
and accounts for the observed test statistic itself.

Permutation Distribution

Drug trial: $T_{\text{obs}} = -1.76$, $B = 10{,}000$ permutations



Count: ≈ 780 of 10,000 permutations have $|T^*| \geq 1.76$.

Permutation p-value: ≈ 0.078 — nearly identical to the z-test $p = 0.078$.

Permutation Tests: Pros and Cons

Advantages

+ **Distribution-free** (but not assumption-free!)
+ Exact p-values (not approximate)
+ Works for *any* test statistic (median, ratio, custom metric)
+ Easy to understand and implement

Limitations

– Computationally intensive (but cheap on modern machines)
– Tests the "sharp null" only (H_0: identical distributions)
– Cannot easily give CIs (use bootstrap for that)
– Needs exchangeability under H_0

When Is Exchangeability Violated?

Permutation tests assume: under H_0, swapping labels doesn't change the distribution. This fails when:

1. Paired / matched data

Before-after measurements on the *same* patients. Shuffling breaks the pairing. **Fix:** permute the *sign* of differences instead.

2. Unequal variances

Groups have same mean under H_0 but different spreads. Shuffling mixes the variances.
Fix: use a studentized statistic.

3. Dependent observations — Time series, spatial data, clustered data. Observations within a group are correlated; shuffling individuals destroys the dependence structure. **Fix:** permute *blocks* or *clusters*, not individual observations.

Permutation Test in Python

```
import numpy as np

# Data: drug (n=100) and placebo (n=100)
drug = np.random.normal(137, 12, 100)
placebo = np.random.normal(140, 12, 100)
t_obs = drug.mean() - placebo.mean()

# Permutation test
combined = np.concatenate([drug, placebo])
B = 10_000
t_perm = np.zeros(B)
for i in range(B):
  np.random.shuffle(combined)
  t_perm[i] = combined[:100].mean() - combined[100:].mean()

# Two-sided p-value
p_perm = (np.sum(np.abs(t_perm) >= np.abs(t_obs)) + 1) / (B + 1)
```

Part VI: Multiple Testing

Test 20 hypotheses, and one will be "significant" by chance

(See xkcd.com/882 — "Significant" — for the classic illustration)

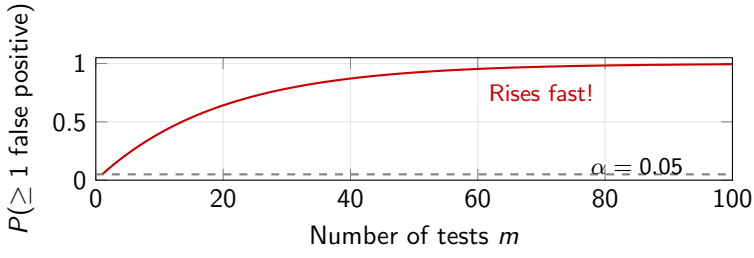
The Multiple Testing Problem

Examples: 20,000 genes, 50 A/B test variants, multiple clinical endpoints

If you test $m = 20$ independent true nulls at $\alpha = 0.05$:

$$P(\text{at least one false positive}) = 1 - (1 - \alpha)^m = 1 - 0.95^{20} = \mathbf{0.64}$$

64% chance of a false discovery — even when nothing is real!



Correcting for Multiple Tests

Bonferroni / Holm (FWER)

Bonferroni: reject if $p_i \leq \alpha/m$

Holm (step-down): sort p-values, compare $p_{(k)}$ to $\alpha/(m-k+1)$

Controls: $P(\text{any false positive}) \leq \alpha$

Holm is **always** $\geq$ as powerful as Bonferroni. No reason to use Bonferroni alone.

Benjamini–Hochberg (FDR)

Sort p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$

Find largest k: $p_{(k)} \leq \frac{k}{m}\alpha$

Reject $H_{(1)}, \ldots, H_{(k)}$

Controls: expected *fraction* of false discoveries $\leq \alpha$

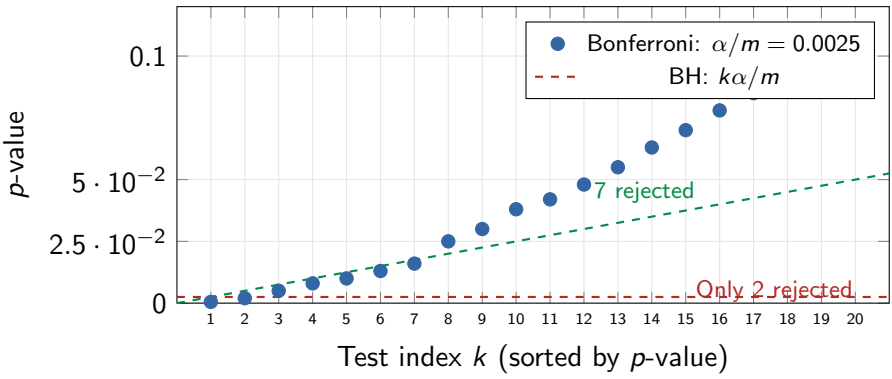
$+$ More powerful — the modern standard

FWER (Bonferroni): "Every single rejection must be real." Use for small m.

FDR (BH): "Some false discoveries OK if I find more true ones." Use for large m.

BH in Action: The Step-Up Procedure

$m = 20$ tests, $\alpha = 0.05$



Dots below a line = rejected by that method.
BH rejects 7 hypotheses; Bonferroni only 2. BH finds more while controlling the *fraction* of false discoveries.

Summary: Hypothesis Testing

Logic: Assume H_0 (nothing happening). Ask: how surprising is my data under H_0?

Test statistic: Compresses data into one number measuring distance from H_0.

p-value: $P(\text{this extreme or more} \mid H_0)$. **NOT** $P(H_0 \mid \text{data})$.

Type I (α): False alarm. **Type II (β):** Missed detection. **Power** $= 1 - \beta$.

Effect size: Cohen's d measures "how big?" independently of sample size.

Power analysis: Plan n before collecting data. Underpowered studies waste resources.

Permutation tests: Shuffle labels under H_0. No distributional assumptions needed.

Multiple testing: Many tests $\Rightarrow$ many false positives. Bonferroni (strict) or BH-FDR (modern).

Replication crisis: Pre-register, report effect sizes + CIs, don't p-hack. Science is self-correcting.

Practical: Hypothesis Testing & Power

1. ***z*-test by hand:**
   - Drug trial: $n = 100$, $\bar{X}_D = 137$, $\bar{X}_P = 140$, $S = 12$
   - Compute T, find p-value, decide at $\alpha = 0.05$
2. **Power simulation:**
   - Under H_1: $\delta = 3$, $\sigma = 12$. Simulate 10,000 experiments
   - Try $n = 100$, 250, 500. What fraction reject H_0?
3. **Permutation test:**
   - With the drug trial data, implement a permutation test ($B = 10{,}000$)
   - Compare p_{perm} with $p_{\text{z-test}}$
4. **Multiple testing:**
   - Simulate 1,000 null experiments ($\mu_1 = \mu_2$). How many give $p < 0.05$?
   - Apply Bonferroni and BH. How many survive each correction?

Homework

1. A coffee shop claims cups contain 350 mL. You measure $n = 25$: $\bar{X} = 345$, $S = 10$.
   Test $H_0 : \mu = 350$ vs $H_1 : \mu \neq 350$ at $\alpha = 0.05$.
   Compute T, p-value, and 95% CI. Does the CI agree with the test?

2. A researcher wants to detect $d = 0.3$ with 80% power at $\alpha = 0.05$.
   (a) How large a sample per group?
   (b) If budget allows only $n = 50$ per group, what is the minimum detectable d?

3. You test 100 genes for disease association. 7 give $p < 0.05$.
   (a) How many false positives do you expect if all nulls are true?
   (b) Apply Bonferroni: how many survive?
   (c) Apply BH at $q = 0.10$: how many survive?

4. Two teaching methods. A scores: 78, 85, 91, 72, 88. B scores: 65, 70, 82, 68, 74.
   Implement a permutation test for the difference in means ($B = 10{,}000$).
   Is the difference significant at $\alpha = 0.05$?

Recommended Visualizations & Resources

Interactive: Seeing Theory — Frequentist Inference (Brown)

`seeing-theory.brown.edu/frequentist-inference` — drag sliders for rejection regions, Type I/II errors, power.

Interactive: R Psychologist — Understanding Statistical Power

`rpsychologist.com/d3/nhst/` — the best interactive power visualization. Drag effect size and n.

Video: StatQuest — p-Values, Clearly Explained

The clearest short explanation of what p-values are and are not. Also: "Hypothesis Testing and p-Values."

Reading: ASA Statement on p-Values (2016) + 2019 Follow-Up

2016: Six principles every scientist should know. 2019: "Moving to a World Beyond $p < 0.05$" — abandon "stat. significant."

Reading: Wasserman, "All of Statistics" — Ch. 10

Hypothesis testing, Neyman–Pearson, power. Concise and rigorous.

Questions?

Next: Lecture 10 — Classical tests & the LRT framework

Preview: Is there a *best* test? The Neyman–Pearson lemma says yes — the likelihood ratio test. One principle, many tests.