UNIT 9 · BLOCK 3

⚖ Hypothesis Testing

The courtroom logic behind every "significant" result in psychology.

Why this matters

Every study in psychology ends with the same nagging question — *is this effect real, or did we just get lucky?* This unit hands you the formal procedure researchers use to answer it, called **null-hypothesis significance testing (NHST)**. The move is sneaky and brilliant: you start by assuming **nothing** is going on, then ask how *surprising* your data would be if that were true. Master this logic and you can read a results section, smell a bad claim, and never again be fooled by the word "significant." The arithmetic is a single formula. The reasoning is the prize.

By the end you can…

- ✓ State a **null hypothesis** (H0) and an **alternative hypothesis** (H1) in symbols and in plain English.
- ✓ Explain the **p-value** correctly — and spot the misinterpretation almost everyone makes.
- ✓ Walk the **four steps of NHST** and decide to **reject** or **fail to reject** H0 (never "accept").
- ✓

Tell a **Type I error** from a **Type II error**, choose **one- vs two-tailed**, and separate **statistical** significance from **practical** significance.

1. The big idea — assume nothing, then look for surprise

NHST runs on a maneuver borrowed straight from a courtroom. A jury doesn't start by asking "is this person guilty?" It starts by *assuming* the defendant is innocent and then asks whether the evidence is too strong to keep believing that. Statistics does the identical thing. We assume there is **no effect** — the dullest possible state of the world — and then ask: *if that were true, how unlikely would data like ours be?* If our data would be a freak accident under "no effect," we stop believing "no effect."

That's the whole engine, and it fits in four moves:

- **Start skeptical.** Assume the **null hypothesis (H0)** is true. H0 always says: *nothing is going on — no difference, no effect, no relationship.*
- **Name the rival.** The **alternative hypothesis (H1)** says: *something IS going on — a real difference or effect exists.* It's what the researcher usually hopes to support.
- **Measure the surprise.** Collect data and ask one question: *if H0 were true, how often would we see results this extreme just by chance?* That surprise gets a number — the **p-value**.
- **Decide.** If the data are too surprising to chalk up to luck (a small p), we **reject** H0. If they aren't, we **fail to reject** it.

Why we test the boring claim

You can never directly *prove* there's an effect — there are too many ways to be wrong. But you *can* show that "no effect" makes the data look absurd. So we set up the boring claim (H0) as a target and try to knock it down. It's proof by contradiction, dressed in numbers.

2. The two hypotheses — H0 and H1

Every test names two rival claims about a **population** (never just the sample). They are mutually exclusive and they cover all the possibilities between them.

Hypothesis	Symbols (one-sample mean)	Plain English
H0 (null)	$\mu = 18$	Nothing's going on. No difference from the known value. The skeptical default.
H1 (alternative)	$\mu \neq 18$	Something's going on. The true mean really differs.

Notice the hypotheses are about $\boldsymbol{\mu}$ (the population mean), not $\bar{\mathbf{x}}$ (our sample mean). We already *know* our sample mean — the question is what it implies about the population that produced it. H0 gets the equals sign; H1 gets "different from" (and as you'll see in section 5, "different" can mean ≠, >, or <).

Set them BEFORE you peek

You write H0 and H1 — and pick your cutoff — *before* looking at the data. Choosing your hypothesis after seeing which way the numbers fell is a cardinal sin (it's called HARKing), because it lets you "win" no matter what happens.

CHECK **A researcher claims SJNY students sleep a *different* amount than the national norm of 7 hours. Which is the correct null hypothesis?**

μ = 7. The null is always the boring "no difference" claim and always carries the equals sign. The researcher's hunch — that they *differ* — is the *alternative* (H1: μ ≠ 7).

3. The p-value and alpha — stated carefully

The **p-value** is the heart of the test, and it is also the single most misstated number in all of science. Here is the only correct definition:

The p-value, said right

The p-value is the probability of obtaining data **at least as extreme as ours**, **assuming the null hypothesis is true.** In symbols: p = P(data this extreme | H0 true). A small p means: "*data like this would almost never happen by chance — so doubting H0 makes sense.*"

Read that conditional carefully — the order matters enormously. The p-value assumes H0 and asks about the *data*. It does **not** assume the data and tell you about H0. So a p of .03 does **NOT** mean "there's a 3% chance H0 is true" or "a 97% chance the effect is real." Those are flipped conditionals, and the flip is wrong. The p-value can only ever speak about how the data behave *in a world where nothing is going on.*

Before we collect a single data point, we choose a threshold for "surprising enough." It's called **alpha (α)**, and by long convention in psychology **α = .05**. Crucially, alpha is not just a cutoff — it *is* the rate at which we'll falsely cry "effect!" when H0 is actually true. Set α = .05

and you've agreed to be fooled by random noise about 5% of the time. That's the price of admission, and it's exactly why we keep α small.

CHECK **A study reports *p* = .02. Which statement is correct?**

The second. The p-value is the probability of the *data* (this extreme or more) *given* H0 — not the probability of H0 given the data. "A 2% chance H0 is true" flips the conditional and is the classic misinterpretation. p never tells you the probability that any hypothesis is true.

4. The decision rule — reject or fail to reject

Now the payoff. Compare your p-value to alpha and the verdict is automatic:

The decision, in two lines

p < α → the data are too surprising to credit to luck. We **REJECT H0**. The result is **statistically significant**.
p ≥ α → the data aren't surprising enough. We **FAIL TO REJECT H0**.

You'll often see the same decision made with a **test statistic** instead of a p-value: compute something like a *z* or *t*, then compare it to a **critical value** that marks the edge of the rejection region. If the statistic lands past the critical value (out in the tail), p is automatically below α and you reject. The two routes — "p vs α" and "statistic vs critical value" — always agree; they're the same boundary described two ways.

Never say "accept H0"

When p ≥ α we say **"fail to reject H0,"** *never* "accept H0" or "prove H0 true." Failing to reject only means we didn't gather enough evidence — like a jury returning "not guilty," which is not the same as "proven innocent." Absence of evidence isn't evidence of absence. A bigger or better study might still find the effect.

5. One-tailed vs. two-tailed tests

Before collecting data you also decide *which direction* counts as surprising. This sets where the rejection region lives on the null distribution.

- **Two-tailed (non-directional).** H1 says only that the mean is *different* ($\mu \neq 18$). You'd be surprised by an extreme result on *either* side, so α is split into *both* tails (.025 in each for $\alpha = .05$). This is the cautious default and what you should use unless you have strong prior reason not to.
- **One-tailed (directional).** H1 names a direction in advance — $\mu > 18$ *or* $\mu < 18$ — so the entire $\alpha = .05$ sits in one tail. That makes it easier to reach significance *in the predicted direction*, but you forfeit any ability to detect a surprise in the other direction.

Here is the picture for a **two-tailed** test — the null distribution with both rejection regions shaded and our observed statistic marked out in the right tail:

CHECK **You split α = .05 so that .025 sits in each tail of the null distribution. What kind of test is this?**

Two-tailed. Splitting alpha into both tails means you're alert to a surprising result in *either* direction, so H1 is the non-directional μ ≠ value. A one-tailed test would put all .05 in a single tail.

6. Two ways to be wrong — Type I and Type II

Even with flawless arithmetic, a decision can be wrong, because we only ever see a sample, not the whole truth. Reality is either "H0 true" or "H0 false," and we either reject or don't — so there are four outcomes. Two are correct; two are mistakes. This 2×2 table is the one to commit to memory:

	H0 is really TRUE (no effect exists)	**H0 is really FALSE** (an effect exists)
We reject H0	❌ **Type I error** false positive · false alarm · rate = **α**	Correct decision (a "hit" — rate = **power, 1 − β**)
We fail to reject H0	Correct decision (rate = **1 − α**)	❌ **Type II error** false negative · a miss · rate = **β**

The two mistakes, in one breath

Type I error = a false positive. We yelled "effect!" when there was none. Its rate is **alpha (α)** — which is precisely why we keep α small (usually .05). **Type II error = a false negative.** A real effect was there and we walked right past it. Its rate is called **beta (β)**, and **power = 1 − β** is your chance of catching a real effect when one truly exists (we aim for power ≥ .80).

The two errors trade off against each other. Shrink α to almost zero and you'll almost never raise a false alarm (low Type I) — but you'll also miss more real effects (higher Type II). The biggest lever you actually control is **sample size**: a larger *n* shrinks the standard error, sharpens the test, and lowers β *without* touching α. That's why bigger studies have more power.

Stat in the Wild — the courtroom

A criminal trial runs on this exact logic. The jury starts with **H0 = "innocent"** — the default, the boring assumption. To **reject** it and convict, the prosecution needs **strong evidence**, beyond a reasonable doubt. Weak evidence? You don't convict; you "fail to reject" — which is *not* the same as declaring the defendant truly innocent. So **Type I error = convicting an innocent person** (a false alarm — we rejected a true null), and **Type II error = letting a guilty person walk** (a miss — we failed to reject a false null).

Alpha is how much false-conviction risk the system tolerates. Setting **α = .05** is like saying, "*We'll convict only when the evidence is so strong it would fool us less than 5% of the time.*" We keep α small because convicting the innocent is the scarier mistake — just as a false positive is usually the scarier mistake in science.

CHECK

A new therapy truly has no effect, but a study's data happen to look extreme and the researchers reject H0. What error is this?

Type I error. H0 was actually true (no effect), yet we rejected it — a false alarm. Its long-run rate equals α. A Type II error is the opposite: a real effect exists but we fail to reject H0.

7. Statistical significance is NOT practical significance

Here is the distinction that separates a careful reader from a gullible one. "Statistically significant" means only one thing: *the effect is probably not zero.* It says nothing about whether the effect is **big enough to matter.** Those are two different questions, and the p-value answers only the first.

Why does this matter so much? Because the p-value depends heavily on **sample size**. The standard error shrinks as *n* grows, so with a large enough sample *any* difference from H0 — no matter how trivial — will eventually cross $p < .05$. Pour in 50,000 participants and a difference of half a point on a 100-point scale can be "highly significant" and utterly meaningless. Significance tells you an effect is *real*; it does not tell you it is *large*.

> **Always report an effect size**
>
> That's what **effect size** is for — a number that captures *how big* the effect is, independent of sample size (e.g., **Cohen's d** for mean differences, where ≈ 0.2 is small, 0.5 medium, 0.8 large; or **r** for correlations). Best practice is to report *both*: the p-value answers "is it real?" and the effect size answers "is it big enough to care about?" A complete result needs both.

The flip side is just as important. A **non**-significant result with a tiny sample might be hiding a real, even large, effect — the study simply lacked the **power** to detect it. So "$p > .05$" never means "no effect"; it means "this study couldn't rule chance out." Sample size cuts both ways.

CHECK **A study of 80,000 people finds that a supplement raises mood by 0.1 points on a 50-point scale, *p* < .001. What's the best takeaway?**

The second. With a huge *n*, even a microscopic difference gets a tiny p-value. Statistical significance ≠ practical significance — you must check the effect size (here, a 0.1-point bump is negligible) before deciding the result matters.

Worked example — the four steps on OUR data

The question: Are SJNY students more stressed than the national norm? The national **Perceived-Stress** norm is μ = 18, with a known population σ = 6. Our sample of **n = 200** students has a mean stress of **20.09** (column J). Because we know σ, we can use a **z-test**, where the test statistic is $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$. Walk the four steps — predict each before you reveal it.

NHST in four steps — the stress z-test

The same four moves work for every significance test you'll ever run. Reveal one step at a time.

STEP 1 — STATE H0 AND H1

H0: $\mu = 18$ (SJNY students are just like the nation). **H1:** $\mu \neq 18$ (they differ). We have no strong directional prediction, so this is **two-tailed**.

STEP 2 — SET A AND THE TAILS

Use the convention $\alpha = .05$, two-tailed. That puts the critical values at $z = \pm 1.96$ — reject if our statistic lands beyond either one (equivalently, if $p < .05$).

STEP 3 — COMPUTE THE STATISTIC AND COMPARE

Standard error = 6 / √200 = 0.424. Then z = (20.09 − 18) / 0.424 = 2.09 / 0.424 = $+4.91$. Since +4.91 is far beyond the critical +1.96, it sits deep in the rejection region; the corresponding p-value is $< .001$ — well under α.

STEP 4 — DECIDE AND INTERPRET IN PLAIN WORDS

Because $p < \alpha$ (and z is past the critical value), we **reject H0**. *In plain English:* SJNY students score significantly higher on perceived stress than the national norm; a sample mean this far above 18 would almost never occur if the true mean were really 18. (We'd still report an effect size to say *how much* higher.)

Read the verdict carefully

"Reject H0" here means the data are incompatible with "no difference" — *not* that we've proven the exact size of the gap, and *not* that there's only a 0.1% chance H0 is true. The p-value (< .001) is the probability of data this extreme *if* μ really were 18.

In Google Sheets

You won't compute z by hand on exams — Sheets does it. Open the **SJNY Student Life** dataset and type these in any empty cell (perceived stress is column J):

```
=(AVERAGE(J2:J201) - 18) / (6 / SQRT(COUNT(J2:J201)))     → the z-score (≈ 4.91)
=2*(1 - NORM.S.DIST(ABS(z), TRUE))                        → the two-tailed p-value (<
```

In the second formula, put your z-value where it says `z`. The `2*` at the front makes it **two-tailed** (it counts both tails); drop the `2*` for a one-tailed p. A handy gut check: a z beyond about ±1.96 always yields a two-tailed p below .05 — that's the same ±1.96 boundary drawn in the figure above.

Watch out for

"Fail to reject" is not "prove H0 true" — it only means *not enough evidence*; absence of proof isn't proof of absence, so never write "accept H0." **The p-value is P(data | H0), not P(H0 | data)** — a small p does *not* mean H0 has a small chance of being true. **Significant ≠ large:** a huge sample can make a trivial difference "significant," so always pair p with an effect size. And pick **alpha and your tails before** you see the data — never after, to make a result "work."

Key terms

- Null hypothesis (H0)
- Alternative hypothesis (H1)
- NHST
- p-value
- Alpha (α)
- Critical value
- Rejection region
- Reject / fail to reject
- One- vs two-tailed
- Statistically significant
- z-test
- Standard error
- Type I error (α)
- Type II error (β)
- Power (1 − β)
- Effect size (Cohen's d)
- Practical significance

PRACTICE — NEVER GRADED

Check yourself

Ten quick true/false items with instant explanations. (Different from your Canvas quiz — real practice.)

Sigma says: When a p-value confuses you, swap in one word — *surprise.* A small p means "this data would be shocking if nothing were going on," so we stop believing nothing is going on. Big p means "eh, not surprising," so we shrug. Small p, reject. Big p, fail to reject. That's the whole game — just never let "significant" trick you into thinking "big."