UNIT 12 · BLOCK 3

Paired-Samples t-Test

Measuring the SAME people twice to see what changed.

Why this matters

Does the medicine lower blood pressure? Does the tutoring raise the score? Does the gratitude exercise lift the mood? To answer questions like these, researchers measure the same people **before** and **after**. The paired *t*-test tells you whether the change is real. And the trick behind it is gorgeous: it quietly turns each person's two scores into **one difference score**, then asks a question you already know how to answer — is that average difference different from zero?

By the end you can…

- Recognize a **paired** (within-subjects) design — and tell it apart from two independent groups.
- Reduce each pair to a **difference score** *d* and explain why that one move *is* the whole test.
- Compute $t = \bar{d} \div (s_d/\sqrt{n})$ with $df = n_{pairs} - 1$ and decide whether to reject the null.
- Say **why pairing is so powerful** — it removes between-person variability, so each person is their own control.

Run it in Google Sheets and read the ***t*** and ***p***.

1. The big idea

Last unit, the **independent** *t*-test compared **two separate groups** of people — coffee drinkers versus non-drinkers, one mean against another. This unit, the **paired** (also called **dependent** or **within-subjects**) *t*-test compares **two scores from the SAME people** — a *before* and an *after*, or a matched pair. The two columns aren't strangers; they're tied together, row by row, because *the same person produced both*.

Test	What it compares	Example
Independent	Two *different* groups of people	Coffee drinkers vs. non-drinkers
Paired	The *same* people, measured twice	Everyone's mood *before* vs. *after*

The rule of thumb: **same people measured twice = paired.** Two separate, unconnected groups = independent. That single distinction decides which test you reach for — and it's set by your *design*, long before you touch a number.

2. Recognizing a paired (within-subjects) design

The paired test fits whenever your two scores are **linked in pairs**. There are two flavors, and both count:

- **Same people, twice (repeated measures).** You measure each person before and after something — a treatment, a lesson, an exercise. In our dataset, `mood_pre` and `mood_post` are exactly this:

every student gave a mood rating before a gratitude exercise and again after.

- **Matched pairs.** Different people, but deliberately paired off so each pair is as alike as possible — twins, married couples, or two participants matched on age and baseline score, one assigned to each condition. The pairing makes their scores belong together just like a before/after pair.

The tell is always the same: **can you line the two columns up row by row, with a meaningful reason each row's two values belong together?** If yes — same person, or a matched partner — it's paired. If the two columns are just two piles of unrelated people, it's independent.

CHECK

A sleep researcher records each participant's reaction time *after a full night's sleep* and again *after staying up all night* — every person contributes two scores. Which design is this?

Paired (within-subjects). The *same* people are measured twice — rested and sleep-deprived — so the two columns line up person by person. That linkage is the signature of a paired design, and it calls for a paired *t*-test.

3. The trick — it works on the differences

Here's the clever part, and it's the single most important idea in the unit. The paired *t*-test doesn't really juggle two columns at all. **First it does subtraction.** For **each person** it collapses the two scores into one **difference score**, usually written *d*:

The one formula

***d* = after − before.** If a student's mood went from 4 to 6, their difference is $d = +2$. If it went from 7 to 5, their difference is $d = -2$. Now every person is a single number — and a column of difference scores is all the test ever looks at.

That subtraction is the whole magic move. Once you have a column of *d*'s, the paired test is **nothing more than a one-sample *t*-test on those differences**, with a test value of **zero**. You already learned that test in Unit 10 — so you already know how to run a paired *t*-test. It just asks one simple question:

Is the average difference ($\bar{d}$) different from zero?

- If nothing really changed, the ups and downs cancel and the differences average out to about **0**.
- If something genuinely happened, the differences pile up on one side — and their average sits clearly **above** (or below) zero, far enough that chance can't comfortably explain it.

That's the entire test. Turn two scores into one difference, then check whether the mean difference is far enough from zero to believe. Two columns in; one column of *d*'s; a one-sample *t* against 0 comes out.

4. The formula — a one-sample *t* in disguise

Because the paired test is just a one-sample *t* on the difference scores, the formula is the one you already know, dressed in *d*'s:

The paired-t formula

$$t = \bar{d} \div (s_d / \sqrt{n})$$

where $\bar{d}$ is the mean of the difference scores, s_d is their standard deviation, n is the number of *pairs*, and $s_d/\sqrt{n}$ is the standard error of the differences. Degrees of freedom: $df = n_{pairs} - 1$.

Read it in plain English: the top is "how big is the average change?" and the bottom is "how much do the changes wobble around from person to person?" A big, consistent change over a small wobble makes *t* large — large enough to clear the critical value and reject the null. Watch it work on five clean numbers:

Worked example — a paired *t* from scratch

Five students rate their mood before and after a gratitude exercise (1–9 scale). Predict each step before revealing it.

THE DATA

Before: 4, 6, 5, 3, 7. After: 6, 7, 8, 5, 9. Five people, two scores each — a paired design, $n = 5$ pairs.

STEP 1 — DIFFERENCE SCORE FOR EACH PERSON (*D* = AFTER − BEFORE)

6−4= $+2$; 7−6=+1; 8−5=+3; 5−3=+2; 9−7=+2. Difference column: **2, 1, 3, 2, 2**. From here on we forget the original columns entirely.

STEP 2 — MEAN OF THE DIFFERENCES ($\bar{D}$)

Add them and divide by n: (2 + 1 + 3 + 2 + 2) ÷ 5 = 10 ÷ 5 = $\bar{d} = 2.0$. On average mood rose two points.

STEP 3 — STANDARD DEVIATION OF THE DIFFERENCES (S_D)

Deviations from 2.0: 0, −1, +1, 0, 0. Square them: 0, 1, 1, 0, 0 sum of squares = 2. Divide by $n - 1 = 4$: variance = 0.5.

Square-root: $s_d \approx 0.71$.

STEP 4 — STANDARD ERROR ($S_D/\sqrt{N}$)

0.71 ÷ √5 = 0.71 ÷ 2.236 = ≈ 0.32. This is the typical wobble of the average difference.

STEP 5 — THE T-STATISTIC

$t = \bar{d} \div SE = 2.0 \div 0.32 =$ $t \approx 6.32$, with $\boldsymbol{df = n - 1 = 4}$.

STEP 6 — DECIDE

The critical *t* at *df* = 4, two-tailed α = .05, is about 2.78. Our 6.32 is way past it, so **we reject the null** ($p < .05$): the mood lift is bigger and more consistent than chance would produce.

> We *do not* "accept" anything — we simply reject the no-change hypothesis.

CHECK **In a paired *t*-test with $n = 25$ pairs of scores, the degrees of freedom are…**

24. For a paired test, $df = n_{pairs} - 1$, because the test runs on the *one* column of difference scores. With 25 difference scores, $df = 25 - 1 = 24$ — exactly like a one-sample *t* on those differences. (It is *not* based on the total of 50 individual measurements.)

5. Why pairing is so powerful

Everyone starts at a different baseline. One student is a naturally cheerful 8; another is a quiet 3. In the **independent** *t*-test, those huge person-to-person gaps become **noise** in the denominator that swamps the real effect — the spread *between* people drowns out the signal. Pairing cancels that noise. Because each person is compared **to themselves**, their personal baseline subtracts away in the very first step (the difference score). The cheerful student and the quiet student can *both* show a clean +1 jump, and pairing sees both jumps with no baseline static in the way.

Statistically, this is the payoff: subtracting each person's own baseline **removes the between-person variability**, which shrinks the standard error in the denominator. A smaller denominator makes *t* larger for the same average change — so a real effect is easier to detect. That is what we mean by **more statistical power**.

Quick test

Think of it this way: **each person serves as their own control.** Remove the person-to-person noise, and a real change becomes much easier to spot. That's why the paired test is often *more powerful* than the independent test for the same number of people — you get more signal for the same sample size, simply by pairing.

6. The headline study — a gratitude lift

We ran a quick **gratitude exercise** in class: students listed a few things they were grateful for. We measured **mood before** (column `mood_pre`) and **mood after** (column `mood_post`) for all **200 students**. Reduce each student to a difference score, average them, and run a one-sample *t* against zero — exactly the recipe above, just with 200 pairs instead of 5.

When	Average mood
Before (mood_pre)	5.36
After (mood_post)	6.47
Average difference ($\bar{d}$)	**+1.12**

$t(199) = 11.9, p < .001.$

Because ***p* is far below .05**, we **reject the null** hypothesis of no change: the lift is real, not a fluke. The gratitude exercise **significantly raised mood** — about a full point on the scale, and not by chance. Our result echoes a famous study: Emmons and McCullough (2003) had people keep weekly lists of things they were grateful for and found the gratitude group reported higher well-being than a comparison group. A tiny exercise, a real and reliable lift — exactly what we saw in our own data.

CHECK **A paired *t*-test returns $t(199) = 11.9$, $p < .001$. The right thing to conclude is that we…**

Reject the null hypothesis. A *p* below .05 means the mean difference is too far from zero to be chance, so we reject "no change." We never *accept* the null (and never *prove* the alternative either) — in null-hypothesis testing we only ever reject or fail to reject. Here, with $p < .001$, we reject.

Try it yourself

Here's the live **Paired-Samples t-Test** calculator. Hit **Load class data** to drop in the mood pre/post numbers, then **Run** and read the plain-English result. Behind the scenes it does precisely the six steps above — difference scores, their mean and SD, then $t = \bar{d} \div (s_d/\sqrt{n})$.

How big? Cohen's *d* for paired designs

Same lesson as the last two units: $t(199) = 11.9$ says the lift is **real**, not how **big** it is. For a paired design, Cohen's *d* takes the cleanest possible form — the average change divided by the SD of the change scores, both of which you already computed on the way to *t*:

```
d = d̄ ÷ s_d = 1.12 ÷ 1.33 = 0.84
```

Read it plainly

The average mood lift is **0.84 standard deviations** of the change scores — by the usual rule of thumb (0.2 small, 0.5 medium, 0.8 large), a **large effect**. Notice the family resemblance: *t* divides $\bar{d}$ by $\mathbf{s_d/\sqrt{n}}$ (the standard error), *d* divides the same $\bar{d}$ by $\mathbf{s_d}$ (the raw spread). That's why 200 pairs can make *t* a whopping 11.9 while *d* stays a more honest 0.84 — *t* grows with the sample, *d* doesn't. The calculator above now reports *d* alongside *t* and *p*.

Report it in APA style

The APA (7th ed.) sentence for a paired *t* gives both time-points' means, then the statistics — same house rules as always (italicized letters, two decimals, no leading zero on *p*, "*p* < .001" never "= .000", effect size included):

The write-up

"Mood was significantly higher after the gratitude exercise (*M* = 6.47) than before it (*M* = 5.36), *t*(199) = 11.88, *p* < .001, *d* = 0.84."

Direction in plain words, both means, then the full inferential sentence. Note the *df* quietly telling the reader this was 200 *pairs* — in a paired design the write-up describes the same people at two moments, not two separate groups.

In Google Sheets

Both scores live in their own columns — one for *before*, one for *after*. There are two equally good ways to do this, and they give the same answer.

The one-formula way. In any empty cell:

```
=T.TEST(V2:V201, U2:U201, 2, 1)
```

Reading the arguments left to right:

- `V2:V201` = mood *after*
- `U2:U201` = mood *before*
- **2** = two-tailed (we'd accept a change in either direction)
- **1** = **PAIRED** — this last digit (the `type` argument) is the switch that makes it a paired test

That single **1** is the whole difference from last unit. A `2` there would run an independent test instead. One digit changes everything. `T.TEST` hands you the *p-value* directly.

The build-the-difference-column way (and the one that shows you what's really happening). In a new column, subtract row by row, then run a **one-sample test on that single column against 0**:

```
W2:  =V2 - U2            drag down to W201: a difference score per person
=AVERAGE(W2:W201)        d̄  (the mean difference)
=STDEV(W2:W201)          s_d (SD of the differences)
=AVERAGE(W2:W201) / (STDEV(W2:W201) / SQRT(200))    t
df = 200 - 1 = 199
```

The two recipes agree because they are the same test. The first hides the difference column; the second shows it. Either way you're testing whether $\bar{d}$ is reliably different from zero.

Before and after, on the farm question

Before-and-after designs are **everywhere**, and they're a natural fit for studying how attitudes shift. Imagine showing students a short documentary about modern animal agriculture, then measuring their **animal-compassion score** both before and after. Same people, two scores, one difference apiece — a textbook paired design, scored with exactly the formula above.

Pairing is precisely the right tool here, because people walk in at wildly different starting points: some already deeply concerned, some never having thought about it. An independent comparison would drown the real shift in all that baseline variation — that between-person spread would inflate the standard error and bury the signal. Pairing strips it out — each person is compared to *their own* earlier score — so even a modest, genuine change in compassion becomes easy to detect. The same logic that caught a one-point mood lift would catch an attitude shift after a single film. Matter-of-fact, and entirely testable.

Stat in the Wild — the before-and-after superpower

Before-and-after designs show up across all of science. Does the blood-pressure drug actually lower BP? Does the tutoring program raise test scores? Does therapy reduce anxiety? In every case, researchers measure the **same people twice** and pair the scores.

The lesson: pairing is a superpower because it strips out the one thing that usually hides an effect — the fact that everyone is different to begin with. Measure the same person before and after, and you see *their* change, cleanly. When a design lets you pair, pair: you get more signal for the same number of people.

Watch out for

Same people twice (or matched pairs) = paired. Two separate, unconnected groups = independent — pick the test that matches the design. ***df* = n_{pairs} − 1**, based on the count of difference scores, not the total number of measurements. **Order of the columns sets the sign:** `after − before` gives a positive *t* for an increase; flip them and the sign flips, but the size and *p*-value don't change. **The last argument is the switch** — a `1` makes it paired, a `2` makes it independent. And a small *p* means *real*, not *huge*: $p < .001$ says the change is reliable; the **+1.12** tells you how big it is — always report both. Finally, when you decide, say "**reject**" or "**fail to reject**" the null — never "accept" it.

Key terms

Paired (dependent) *t*-test · **Within-subjects** · **Repeated measures** · **Matched pairs** · **Difference score (*d*)** · **Mean difference ($\bar{d}$)** · **SD of differences (s_d)** · **Baseline** · **Own control** · **Statistical power** · ***df* = *n* − 1** · **Two-tailed** · **Statistically significant** · **Fail to reject**

PRACTICE — NEVER GRADED

Check yourself

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

Sigma says: When you see "before and after on the same people," your hand should reach for the paired test automatically. The secret is that it quietly turns two columns into one column of differences — then just checks if that average is far from zero. *df* = number of difference scores − 1.