UNIT 13 · BLOCK 4

One-Way ANOVA

One honest test to compare three or more groups at once.

Why this matters

You've made it to the big one. All year a *t*-test has compared **two** groups. But real questions usually have **three or more** — omnivores vs. vegetarians vs. vegans, three teaching methods, four age groups. You *could* run a whole pile of *t*-tests… but that quietly stacks up your chance of a false alarm. **ANOVA** compares every group in **one** honest test. The name (Analysis of Variance) looks scary, but the engine is one tidy idea — split the spread in the data into two piles and take their ratio.

By the end you can…

- Explain **why running many *t*-tests is a problem** — the familywise false-alarm rate piles up.
- Describe ANOVA as **partitioning variance** into a **between-group** and a **within-group** part.
- Compute and read an ***F* ratio** ($F = \text{MS}_{\text{between}} \div \text{MS}_{\text{within}}$) with its two ***df*** and a ***p*-value**.
- Fill in an **ANOVA source table** (SS MS *F*) and interpret the result.
-

Follow a significant ANOVA with **post-hoc tests** — compute a **Tukey HSD** comparison with the ***q* statistic**, and know when **Bonferroni** is the honest alternative.

Check ANOVA's **three assumptions**, and **report the full result in APA style** — *F*, *df*, *p*, η^2, and the pairwise follow-ups.

1. Why not just run a bunch of *t*-tests?

Say you have three groups. To compare them with *t*-tests you'd need three separate runs: A-vs-B, A-vs-C, B-vs-C. Four groups? That jumps to **six** runs. Five groups? Ten. The comparisons multiply fast.

Here's the trap. **Every *t*-test carries a small risk of a false alarm** — calling a difference "real" when it's actually just chance. That's a **Type I error**, and we usually set the risk at **5%** per test ($\alpha = .05$). Run one test, you live with 5%. But run *many* tests and those little risks **pile up**. The chance of *at least one* false alarm across the whole family of tests — the **familywise error rate** — climbs well past 5%.

The arithmetic is unforgiving. With three independent tests each at 5%, the chance of avoiding every false alarm is $0.95 \times 0.95 \times 0.95 \approx 0.857$, so the chance of *at least one* false positive is about **14%** — nearly triple what you signed up for. With ten comparisons it's about **40%**. Run enough *t*-tests and you're almost *guaranteed* to "find" something that isn't there.

The fix in one line

ANOVA tests all the groups together in one shot — holding the familywise false-alarm rate at a single honest 5%. That's the whole reason it exists: more than two groups, one omnibus test.

CHECK **A student wants to compare four groups, so she runs all six pairwise *t*-tests at α = .05 and figures her false-alarm risk is still just 5%. Is she right?**

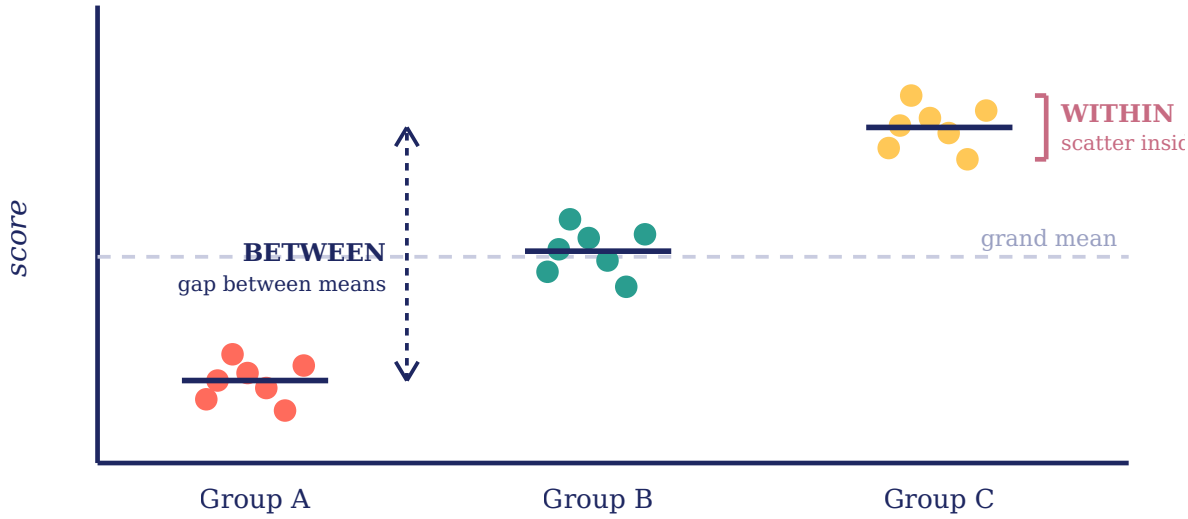
No. Each test carries its own 5% risk, and across six tests the chance of *at least one* false alarm balloons toward ~25%. That pile-up — the inflated familywise error rate — is exactly why ANOVA exists: it compares all four groups in one test that keeps the overall risk at 5%.

2. How ANOVA actually works — partitioning the spread

The deep idea behind ANOVA is hiding in its name: **Analysis of Variance**. We take the total spread in the whole dataset and **split it into two piles** — two sources that, added together, account for all the variation. This split is called **partitioning the variance**, and it's the heart of the method.

- **Between-group variability** — how far apart the **group means** sit from each other (and from the grand mean of everyone combined). If the groups truly differ, their averages land far apart. *(big = interesting — this is the signal)*
- **Within-group variability** — the everyday **noise inside** each group. People in the same group still aren't identical; they scatter around their own group's mean. *(this is just static — the noise)*

Picture three groups as three clouds of dots. **Between**-group spread is the distance between the *centers* of the clouds. **Within**-group spread is how puffy each cloud is on its own. ANOVA's whole job is to ask: are the clouds farther apart than they are puffy?



Between vs. within. The three group means (navy bars) climb from A to C — that vertical gap is **between-group** spread, the signal. The puff of dots around each bar is **within-group** spread, the noise. ANOVA divides the first by the second.

The whole picture in one sentence

***F* asks: is the gap *between* the groups bigger than the noise *within* them?** If yes (big *F*), something real is going on. If no (*F* near 1), it's all static.

3. The *F* ratio — signal over noise

To turn the two piles of spread into one number, ANOVA first expresses each as a **variance** — in ANOVA-speak, a **mean square (MS)**. (A mean square is just a variance: a sum of squared deviations divided by its degrees of freedom — same recipe you met back in Unit 4.) There's a mean square for each pile:

- $\mathbf{MS_{between}}$ — the variance *among the group means*. The signal.
- $\mathbf{MS_{within}}$ — the average variance *inside* the groups. The noise. (You'll also see it called $\mathbf{MS_{error}}$.)

Then the test statistic is simply one divided by the other:

```
F = MS_between ÷ MS_within
  = (variance BETWEEN groups) ÷ (variance WITHIN groups)
  = signal ÷ noise
```

A **big *F*** means the gap between groups is much larger than the random noise — so the difference is probably real. An *F* **near 1** means the groups are no more different than ordinary within-group noise would produce all by itself. *F* can't go below 0 (it's a ratio of two variances, which are never negative), so the whole action is on the high side: the bigger *F* climbs above 1, the stronger the evidence.

> **Why we never "accept"**
>
> Like every test this year, ANOVA can only **reject** or **fail to reject** the null. A big *F* lets you **reject** "all the means are equal." A small *F* means you **fail to reject** it — *not* that you've proven the groups identical. We never "accept the null"; absence of evidence isn't evidence of absence.

4. Degrees of freedom — the two numbers riding with *F*

F always travels with **two** degrees-of-freedom values, one for each pile of spread. With **k** = number of groups and ***N*** = total people across all groups:

df	Formula	Why
Between-groups *df*	$k - 1$	You have *k* group means; once the grand mean is fixed, only $k - 1$ are free to vary.
Within-groups *df*	$N - k$	Each group spends one *df* on its own mean, so you lose one per group — *k* in all.
Total *df*	$N - 1$	The two above add up to this: $(k - 1) + (N - k) = N - 1$. The partition is exact.

So three diet groups (k = 3) with 200 students total (N = 200) give between-df = 3 − 1 = **2** and within-df = 200 − 3 = **197**. Reported in full, the result reads:

```
F(2, 197) = 32.2, p < .001
```

The two numbers in parentheses are **always** the between-df then the within-df, in that order — and notice they encode your design: the first tells a reader how many groups you had (k − 1), the second how big your sample was (N − k). The decision rule is the one you already know: **if *p* < .05, reject the null** — at least one group mean stands apart from the others.

CHECK **A study compares 5 teaching methods with 85 students total. What are the two degrees of freedom for the *F* ratio?**

***df* = 4 and 80. Between-df = k − 1 = 5 − 1 = 4. Within-df = *N* − k = 85 − 5 = 80. (Check: 4 + 80 = 84 = *N* − 1.) It would report as *F*(4, 80).**

5. The ANOVA source table

Every ANOVA bookkeeps itself in a **source table** — the standard layout SPSS, Sheets templates, and journals all use. It walks left to right from raw spread (**SS**, sums of squares) through the *df*, to the variances (**MS**), and finally the ***F***. Here's the skeleton, with the formula in each cell:

Source	SS	*df*	MS	*F*
Between groups	SS_B	$k - 1$	$SS_B \div df_B$	$MS_B \div MS_W$
Within groups	SS_W	$N - k$	$SS_W \div df_W$	—
Total	$SS_B + SS_W$	$N - 1$	—	—

Read it as a recipe: divide each **SS** by its own ***df*** to get an **MS**, then divide the between-MS by the within-MS to get ***F***. The "Total" row is just a checksum — the two SS rows add to it, and the two *df* rows add to it, proving the variance was partitioned exactly. Let's fill one in with real numbers.

Worked example — fill in the source table and get *F*

Three groups (k = 3), $N = 30$ people total. The two sums of squares are already computed: $SS_{between} = 60$, $SS_{within} = 135$. Predict each step before revealing it.

STEP 1 — DEGREES OF FREEDOM

$df_{between} = k - 1 = 3 - 1 =$ `2`. $df_{within} = N - k = 30 - 3 =$ `27`. (Check: 2 + 27 = 29 = $N - 1$.)

STEP 2 — MS BETWEEN

Divide its SS by its *df*: $MS_{between} = 60 \div 2 =$ `30`. This is the variance among the three group means — the signal.

STEP 3 — MS WITHIN

Same recipe: $MS_{within} = 135 \div 27 =$ `5`. This is the average noise inside the groups.

STEP 4 — THE *F* RATIO

$F = MS_{between} \div MS_{within} = 30 \div 5 =$ `6.00`. The between-group spread is six times the within-group noise.

STEP 5 — COMPARE TO CRITICAL *F*

With $df = (2, 27)$ the critical F at $\alpha = .05$ is about `3.35`. Our $F = 6.00$ sits well above it (equivalently, $p < .05$).

STEP 6 — DECIDE AND INTERPRET

6.00 > 3.35, so we `reject` the null. At least one group mean differs from the others — write it as $F(2, 27) = 6.00$, $p < .05$. (Which groups? That's the post-hoc step, next.)

Your completed table

Between: SS = 60, *df* = 2, MS = 30. Within: SS = 135, *df* = 27, MS = 5. ***F* = 30 ÷ 5 = 6.00.** Total: SS = 195, *df* = 29. Every number in the table came from just two operations — divide SS by *df*, then divide MS by MS.

The fine print — ANOVA's three assumptions

Like the *t*-tests it grew out of, ANOVA leans on a few conditions. They don't have to hold perfectly — the test is famously sturdy — but you should be able to name them and know which one actually bites:

- **Independent observations.** Each score comes from a separate, unrelated case. Same non-negotiable rule as Units 10–12 — no formula rescues linked observations. (Same people measured in every condition? That's a *repeated-measures* design, and it needs a different ANOVA.)
- **Roughly normal populations.** Each group's scores should come from an approximately normal population. Thanks to the Central Limit Theorem, with decent group sizes ANOVA barely notices moderate skew — worry only when groups are small *and* obviously lopsided.
- **Homogeneity of variance.** The groups should have roughly *similar spreads* — one puffy cloud and two tight ones make MS_{within} a misleading average of the noise. Rule of thumb: trouble starts when one group's SD is more than about twice another's, *especially* when the group sizes are also unequal.

Which one bites?

Non-independence is fatal, always. Normality is the forgiving one. Homogeneity of variance sits in between: with equal-ish group sizes ANOVA shrugs it off, but with unequal *n*s *and* unequal spreads (our diet groups run 136 / 46 / 18 — worth a glance!) the *F* can drift. The standard check is eyeballing the group SDs side by side; the standard fix is a sturdier variant (Welch's ANOVA — one click away in SPSS).

6. A significant *F* run post-hoc tests

ANOVA has one honest limit. A significant result is an **omnibus** finding — it says "*somebody is different*" — but **not who**. Was it Vegans beating both? Or Vegetarians edging out Omnivores only? The *F* ratio can't tell you; it lumps all the groups together.

To find out, you run **post-hoc tests** — small follow-up comparisons *after* a significant ANOVA that pinpoint **which** pairs of groups actually differ, while still protecting the familywise error rate (common ones: Tukey's HSD, Bonferroni, Scheffé). Think of ANOVA as the smoke alarm and post-hoc tests as walking room to room to find the fire. The order is fixed: significant ANOVA first, then post-hoc — never the reverse.

The two-step idea

ANOVA says *whether* any group differs. Post-hoc tests say *which* ones. You only walk the rooms after the alarm goes off — post-hoc tests come *after* a significant *F*, never before. And if the omnibus *F* is *not* significant, you generally stop: there's no overall difference to chase down.

CHECK

A one-way ANOVA on four groups comes back $F(3, 96) = 9.4$, $p < .001$. What has it told you?

That at least one group differs — but not which. A significant omnibus *F* only says *somebody* stands apart. To learn *which* pairs differ, you follow up with post-hoc tests. The *F* alone never identifies the culprit.

Tukey's HSD — the workhorse

The most-used post-hoc test is **Tukey's HSD** — *Honestly Significant Difference*. The "honestly" is the whole sales pitch: it compares every pair of means while keeping the *familywise* false-alarm rate at 5% across the whole set of comparisons. Its engine is a cousin of *t* called the **studentized range statistic, *q***. For any two group means:

```
q = (larger mean − smaller mean) ÷ SE_pair

where  SE_pair = √[ (MS_within / 2) × (1/n₁ + 1/n₂) ]
```

Look at what's inside: the top is the familiar **gap** between two means, and the bottom recycles $\mathbf{MS_{within}}$ — the noise estimate you already computed for the ANOVA — sized to the two groups being compared. (With equal group sizes this simplifies to the classic $\sqrt{(MS_{within}/n)}$; the version above handles unequal *n*s, which our diet groups certainly are.) Then the decision rule is the same rhythm as every test this year: compare your *q* to a **critical *q***, which depends on **k** (how many groups — more groups, higher bar) and $\boldsymbol{df_{within}}$. For our design — k = 3 groups, $df_{within} = 197$ — the critical *q* at α = .05 is about **3.34**. Beat it and that pair "honestly" differs.

Why not just three t-tests *after* the ANOVA?

Because that would sneak the familywise problem right back in through the side door — three tests, three stacked 5% risks. Tukey's critical q is deliberately *higher* than the matching critical t: each pair pays a little extra so that the *family* of comparisons stays at an honest 5%. Same logic as the ANOVA itself, applied to the follow-ups.

Worked example — Tukey's HSD on OUR diet data

From the ANOVA below: means Omnivore 3.91 (n = 136), Vegetarian 4.92 (n = 46), Vegan 5.67 (n = 18); MS_{within} = 1.12; critical $q(3, 197) \approx 3.34$. Three pairs to check — predict each verdict before revealing it.

STEP 1 — OMNIVORE VS. VEGETARIAN

Gap = 4.92 − 3.91 = 1.01. SE = √[(1.12/2)(1/136 + 1/46)] ≈ 0.13. q = 1.01 ÷ 0.13 ≈ `7.9`. Way past 3.34 `significant` (p < .001). Vegetarians really do score higher than omnivores.

STEP 2 — OMNIVORE VS. VEGAN

Gap = 5.67 − 3.91 = 1.76. SE = √[(1.12/2)(1/136 + 1/18)] ≈ 0.19. q = 1.76 ÷ 0.19 ≈ `9.4`. The biggest gap and the biggest q `significant` (p < .001).

STEP 3 — VEGETARIAN VS. VEGAN

Gap = 5.67 − 4.92 = 0.74. SE = √[(1.12/2)(1/46 + 1/18)] ≈ 0.21. q = 0.74 ÷ 0.21 ≈ `3.6`. Just clears 3.34 `significant` (p = .03) — the closest call of the three.

STEP 4 — READ THE PATTERN

All three pairs differ: the staircase is real at `every step`. Omnivore < Vegetarian < Vegan, each jump bigger than chance. Notice the small-n price: the vegan group (n = 18) makes its comparisons noisier — bigger SE, so a bigger gap is needed to clear the same bar.

Tukey vs. Bonferroni — picking your protection

The other name you'll meet constantly is the **Bonferroni correction** — less a test than a blunt, honest rule: run your pairwise t-tests, but **divide α by the number of comparisons**. Three comparisons? Each

must beat α = .05 ÷ 3 ≈ **.017** to count. (Equivalently: multiply each *p* by 3 and see if it's still under .05.)

	Tukey's HSD	Bonferroni
What it is	A purpose-built test (the *q* statistic)	A penalty applied to ordinary *t*-tests
Best when	Comparing *all* pairs after an ANOVA	Testing a *few* chosen comparisons — or when no Tukey is on the menu (e.g., Sheets)
Personality	Balanced — built for exactly this job	Simple but conservative: with many comparisons the bar gets brutally high, and real effects can slip away

The one-line summary

Comparing **every pair after a significant *F***? Reach for **Tukey**. Only have plain *t*-tests available, or just a couple of planned comparisons? **Bonferroni** your α and be honest about it. Both exist for one reason: to keep the familywise false-alarm rate at 5% while you hunt for *who* differs.

CHECK

After a significant ANOVA on four groups, a researcher Bonferroni-corrects her six pairwise *t*-tests. What α must each comparison now beat?

.05 ÷ 6 ≈ .008. Bonferroni divides the familywise α across the comparisons: 6 tests each at .008 keeps the whole family at ~5%. (The flip side: that's a high bar — with many comparisons Bonferroni gets conservative, which is why Tukey's HSD is usually the better tool for all-pairs testing.)

7. Try it yourself — the One-Way ANOVA calculator

Here's the real thing, embedded right on the page. Hit **“Load class data (compassion by diet)”** to drop in our three diet groups, then **Run** and read the plain-English result. When the *F* comes back significant, the calculator now walks the rooms for you — it prints the full **Tukey HSD pairwise table** underneath, exactly the follow-ups from §6. This is the exact analysis from our worked example below.

Worked example — on OUR data

Open the **SJNY Student Life** dataset. We're looking at one of your instructor's own research questions: *do people who eat differently also **feel** differently about animals?* Each of our 200 students reported their **diet** and filled out the **Compassion for Animals scale** (1 = low, 7 = high):

- **Omnivore** — eats meat
- **Vegetarian** — no meat, but eats some animal products
- **Vegan** — no animal products at all

This is exactly an ANOVA design: **three** groups (k = 3), one quantitative outcome. The three group means tell the story — they climb steadily:

Group	*n*	Mean compassion (1–7)
Omnivore	136	3.91
Vegetarian	46	4.92
Vegan	18	5.67

In Sheets (column **X** = diet, column **Y** = animal_compassion), grab each mean with one function:

```
=AVERAGEIF(X2:X201, "Omnivore",   Y2:Y201)    3.91
=AVERAGEIF(X2:X201, "Vegetarian", Y2:Y201)    4.92
=AVERAGEIF(X2:X201, "Vegan",      Y2:Y201)    5.67
```

Then make a **bar chart** of the three means. The bars climb steadily — a clear step up at each level, with Vegan towering over Omnivore. Now feed those numbers into the calculator above and you get the full result, laid out in a source table:

Source	SS	*df*	MS	*F*
Between groups	72.1	2	36.1	32.2
Within groups	220.7	197	1.12	—
Total	292.8	199	—	—

The result, read aloud

***F*(2, 197) = 32.2, *p* < .001, η² = .25.** Trace it in the table: $MS_{between}$ = 72.1 ÷ 2 = 36.1; MS_{within} = 220.7 ÷ 197 ≈ 1.12; *F* = 36.1 ÷ 1.12 ≈ 32.2. That big *F* and tiny *p* mean the gap between the groups is *way* bigger than chance noise — so we **reject** the null that all three means are equal. The **η² = .25** (eta-squared, an effect size) says diet accounts for about **25% of the variation** in compassion scores — a large effect. A significant *F* is the green light to ask *which* groups differ; the Tukey follow-ups in §6 confirmed each step of the staircase: all three pairs differ. The students draw their own conclusions from there.

Stat in the Wild — a real finding, in our own data

This isn't a textbook toy — it's a live line of research. Across studies in the **United States and the United Kingdom**, vegans report markedly more compassion for animals than omnivores do, and the gap is **large** (about $d \approx 1.0$ — the average vegan sits a full standard deviation above the average omnivore). Vegetarians land in between. That steady climb — Omnivore 3.91 < Vegetarian 4.92 < Vegan 5.67 — is exactly the pattern those studies keep finding.

The takeaway: notice the design — **three** groups, not two — exactly when you reach for **ANOVA** instead of a *t*-test. One ANOVA keeps the false-alarm rate honest, and a significant *F* is your green light to ask *which* groups differ. *(Magee, M. W. (2026). Compassion for animals and self-compassion among vegans and omnivores in the US and UK [Manuscript in preparation].)*

Report it in APA style

Every stats result you'll ever publish, present, or hand in gets written in the same standardized sentence — the APA (7th ed.) results format. It packs the test, the *df*, the statistic, the *p*, and the effect size into one line a reader anywhere can decode. The template for a one-way ANOVA:

```
F(df_between, df_within) = value, p = value, η² = value
```

The house rules, once, since they apply to every test from here on: statistical letters (*F*, *t*, *p*, *M*, *SD*) are **italicized**; statistics get **two decimals**; *p*-values get two or three decimals with **no leading zero** (.032, not 0.032) because *p* can't exceed 1; and *p* below .001 is written ***p* < .001**, never "= .000". Report the **means and SDs** so readers see the actual pattern, and **always include the effect size**. Our full result, written for a journal:

The write-up

"A one-way ANOVA revealed a significant effect of diet on animal compassion, $F(2, 197) = 32.20$, $p < .001$, $\eta^2 = .25$. Tukey HSD comparisons showed that vegans ($M = 5.67$) reported more animal compassion than both vegetarians ($M = 4.92$, $p = .032$) and omnivores ($M = 3.91$, $p < .001$), and vegetarians in turn exceeded omnivores ($p < .001$)."

Read the anatomy of that sentence: **what test** ("a one-way ANOVA"), **the omnibus verdict** with its full statistics, then **the post-hoc story** — which pairs, in which direction, with the group means anchoring each claim. That sentence structure — omnibus first, follow-ups second — mirrors the order of the analysis itself. To the extent that a reader only skims one sentence of your results, this is the sentence.

CHECK **Which write-up is in correct APA style for a significant one-way ANOVA?**

The second. Three fixes in the first version: *p* is never "= .000" (write $p < .001$); *p*-values drop the leading zero; and the effect size (η^2) should ride along. Statistics also get two decimals — 32.20, not 32.2 — and the letters are italicized.

In Google Sheets

Sheets has **no one-click one-way-ANOVA button** in the base app, so for the full source table you have two honest routes:

- **The add-on route.** Install the free **XLMiner Analysis ToolPak** add-on, then **Extensions ▸ Add-ons ▸** open it ▸ choose **Anova: Single Factor**, highlight your columns of group data, and it prints the whole source table (SS, *df*, MS, *F*, *p*) — the same one SPSS

gives. (In Excel proper it lives under **Data ▸ Data Analysis ▸ Anova: Single Factor**.)

- **The shortcut formula.** If you only need the *p*-value for a quick check, `=FTEST(range1, range2)` compares the variances of two groups — useful, but it's a two-group tool, not the full multi-group source table. For a real three-group ANOVA, use the add-on above (or the calculator on this page).

Either way, Sheets nails the part that builds your intuition — the group means and the chart:

```
=AVERAGEIF(X2:X201, "Omnivore",   Y2:Y201)    each group mean
=COUNTIF(X2:X201, "Vegan")                     each group's n
```

Highlight your little table of three means, then **Insert ▸ Chart** and pick a column chart. You'll see the staircase climb with your own eyes — then read the ***F*(2, 197) = 32.2, *p* < .001** from the add-on (or this page's calculator) to confirm the steps are real, not noise.

Post-hoc follow-ups in Sheets. Sheets has no Tukey button, so the honest route is **Bonferroni** — run the three pairwise *t*-tests with `T.TEST` and hold each to α = .05 ÷ 3 ≈ .017:

```
=T.TEST(omnivore_range,   vegetarian_range, 2, 3)    compare to .017
=T.TEST(omnivore_range,   vegan_range,      2, 3)    compare to .017
=T.TEST(vegetarian_range, vegan_range,      2, 3)    compare to .017
```

(The final `3` = unequal-variances type, the safe default when group sizes differ this much.) For real Tukey values, use this page's calculator — after a significant *F* it prints the full pairwise table — or SPSS's one-click **Post Hoc ▸ Tukey**.

Watch out for

3+ groups ANOVA. Exactly 2 groups *t*-test. That's the whole decision. Don't run a stack of *t*-tests to dodge ANOVA — the familywise false-alarm risk piles up fast. A significant ANOVA doesn't say *which* group differs; that's what **post-hoc tests** are for, and they come *after* a significant *F*. $\boldsymbol{F = MS_{between} \div MS_{within}}$: big *F* + small *p* = real difference; *F* near 1 = just noise. Remember the *df* pair — between = k − 1, within = *N* − k. And say "**fail to reject**," never "accept" — a small *F* doesn't prove the groups are identical.

Key terms

ANOVA (Analysis of Variance) · **Partitioning variance** · **Between-group variability** · **Within-group variability** · **Sum of squares (SS)** · **Mean square (MS)** · ***F* ratio** · **Source table** · **Degrees of freedom (*df*)** · **Omnibus test** · ***p*-value** · **Eta-squared (η²)** · **Post-hoc test** · **Familywise error rate** · **Type I error (false alarm)** · **Tukey's HSD** · **Studentized range (*q*)** · **Bonferroni correction** · **Homogeneity of variance** · **APA style (results)**

PRACTICE — NEVER GRADED

Check yourself

Twelve quick true/false items with instant explanations. Real practice — different from your Canvas quiz.

Sigma says: Two numbers, one question. $F = MS_{between} \div MS_{within}$ — the gap-between-groups divided by the noise-within-groups. When that ratio is big, the groups really differ. Count your groups first: more than two means ANOVA, every time — and a significant *F* sends you to post-hoc tests to find *who*.