PS 211: Introduction to Experimental Design

Fall 2025 · Section C1

Lecture 13: Reporting results in APA style, One-way ANOVA

Updates & Reminders

- Homework 3 is due on **Friday, October 31** at 11:59 PM.
  - Submit via Blackboard as a PDF file.
  - Late submissions will incur a penalty of 3% per day.
  - We are going to try to post grades and answers by Monday night.
- Exam 3 will be on **Thursday, Nov. 6th** during our regular class time.
  - It will focus on material from Lectures 10-13 (today).
  - It will build on material from earlier in the course, so be sure to review those topics as well.
  - The format will be the same as Exam 1 (33 multiple choice questions).
  - A review sheet will be posted by the end of the day tomorrow.

Review: Confidence Intervals

- Remember, a *z* or *t* statistic corresponds to a percentile in the sampling distribution.
- For example, if we have $t = 2.45$ with df = 28, we can look up the corresponding p-value: $p = .022$.
- This means that if the null hypothesis were true, we would expect to see a *t* value this extreme (or more extreme) only 2.2% of the time.
- Confidence intervals rely on this logic **in reverse**.
- Instead of asking "Given the null hypothesis, what is the probability of observing this data?",
- We ask "Given the observed data, what range of population values are plausible?"

Review: Confidence Intervals (Continued)

- Instead of starting with the *null* distribution, we start with the *sampling* distribution.
- We assume our sample mean was drawn from a population with some unknown mean μ.
- Our best estimate of μ is the sample mean *M*.
- But because of sampling variability, we know that *M* might not equal μ exactly.
- So we construct a range of values around *M* that are plausible values for μ.
- This range is called a **confidence interval (CI)**.

Review: Confidence Intervals (Continued)

- To find a **95% confidence interval**, we identify the range of values that would fall within the middle 95% of the sampling distribution.
- This corresponds to the values between the 2.5th and 97.5th percentiles.
- We can use the critical *t* value for our sample size to find these cutoffs.
- Our critical *t* value tells us how far from *M* we need to go to capture 95% of the sampling distribution.

Review: Confidence Intervals (Continued)

- Remember, our *t* value is in units of standard errors (SE).
- A critical *t* of 2.57 means that 95% of the distribution falls within 2.57 SEs of the mean.

Review: Confidence Intervals (Continued)

- Remember, our *t* value is in units of standard errors (SE).
- A critical *t* of 2.57 means that 95% of the distribution falls within 2.57 SEs of the mean.
- We can use this to calculate the CI around our sample mean:

$$\mathrm{CI} = M \pm (t_{critical} \times SE)$$

Where:

- *M* = sample mean
- $t_{critical}$ = critical *t* value for desired confidence level
- *SE* = standard error of the mean

Review: Confidence Intervals (Continued)

One more time...

- Our best estimate of μ is the sample mean *M*.
- But because of sampling variability, we know that *M* might not equal μ exactly.
- So we construct a range of values around *M* that are plausible values for μ.
- We do this by determining where 95% (or 99% or 99.9%, etc.) of the sampling distribution lies.
- We compute *this* by finding the critical *t* value and multiplying it by the standard error.
- This range is called a **confidence interval (CI)**.
- Lower standard error -> narrower CI.
- Higher standard error -> wider CI.

Review: Confidence Intervals (Continued)

Practice question:

How does increasing the sample size affect the width of a confidence interval?

A. Increases the width because there is more data.

B. Increases the width because the critical t-value gets larger.

C. Decreases the width because the standard error is smaller and the estimate is more precise.

D. No effect on the width because n does not factor into CI calculations.

Answer: C. Decreases the width because the standard error is smaller and the estimate is more precise.

Review: Confidence Intervals (Continued)

Practice question:

How does increasing the estimate of the sample mean affect the width of a confidence interval?

A. Increases the width because the values are higher.

B. Increases the width because the critical t-value gets larger.

C. Decreases the width because the standard error will also increase.

D. No effect on the width because the sample mean does not factor into CI calculations.

Answer: D. No effect on the width because the sample mean does not factor into CI calculations.

Recap: Worksheet example

- Let’s go through the worksheet together.
- I will attempt to fill it out in real time.
- Please correct me if I make a mistake!

Reporting Results in APA Style

APA style tells us **how to clearly report statistics** so others can understand and replicate our work.
It ensures clarity and consistency across psychology and related sciences.

What is APA Style?

- **APA** = *American Psychological Association*
- A standardized format for writing and reporting scientific results
- Ensures clarity, transparency, and consistency across research papers

> **APA style guides how we:**
> • write numbers and statistics
> • report results of tests (t, z, F, etc.)
> • format p-values, CIs, and effect sizes

Why Use APA Style?

- Helps readers quickly interpret results.
- Makes research **replicable** and **comparable**.
- Prevents ambiguity — everyone reports in the same format.
- Required by most psychology journals and conferences.

Think of APA as the "grammar rules" of scientific writing.

Always Include These Elements

Element	Description	Example
Test statistic	Which test you ran (*t*, *z*, *F*)	*t*
Degrees of freedom (df)	Indicates sample size info	*t*(28)
Test value	The computed statistic	*t*(28) = 2.45
p value	Probability under H_0	*p* = .004 or *p* < .001
Descriptive stats	Means, SDs for each group	*M* = 5.3, *SD* = 1.2

✅ These should appear in **every** statistical report.

Technically optional (but strongly recommended)

Element	Why Include It	Example
Effect size	Shows *magnitude* of difference	d = 0.68
Confidence interval (CI)	Range of plausible values	95% CI 0.12, 1.24
Exact p value	More informative than p < .05	p = .037

- Including these helps readers understand *how strong* and *how precise* your findings are.
- Many journals now require effect sizes, CIs, and exact p-values.
- There is no reason *not* to include them!

Reporting a *z* Test

Template

$$M = X,\ SD = Y,\ z = Z,\ p = P,\ d = D,\ 95\%\ \text{CI}\ [\text{LL, UL}]$$

Example

Students' average rent (*M* = \$920, *SD* = \$192) was significantly higher than the population mean of \$500, *z* = 9.39, *p* < .001, *d* = 4.20, 95% CI [\$832, \$1008].

✅ No *degrees of freedom* are reported for *z* tests. *Z* tests are based on known population parameters, not sample estimates, so df are not applicable.

Reporting a One-Sample *t* Test

Template

$$t(df) = X.XX,\ p = P,\ d = D,\ 95\%\ \mathrm{CI}\ [\mathrm{LL},\ \mathrm{UL}]$$

Example

The sample’s mean reaction time (*M* = 312 ms, *SD* = 28) was significantly faster than the population mean of 350 ms, *t*(14) = –5.68, *p* < .001, *d* = 1.47, 95% CI [–50.1, –21.9].

✅ State the **known or hypothesized** population mean and the sample mean and SD.

Reporting a Paired-Samples *t* Test

Template

$$t(df) = X.XX,\ p = P,\ d = D,\ 95\%\ \mathrm{CI}\ [\mathrm{LL},\ \mathrm{UL}]$$

Example

Participants recalled more words after caffeine (*M* = 12.3, *SD* = 2.1) than without caffeine (*M* = 9.8, *SD* = 2.5), *t*(19) = 3.42, *p* = .003, *d* = 0.76, 95% CI [0.92, 4.11].

✅ State the means and SDs for **both** conditions.

Reporting an Independent-Samples *t* Test

Template

$$t(df) = X.XX,\ p = P,\ d = D,\ 95\%\ \text{CI}\ [\text{LL, UL}]$$

Example

Participants in the study-group condition had higher quiz scores (*M* = 84.6, *SD* = 4.2) than participants in the individual condition (*M* = 78.3, *SD* = 5.0), *t*(38) = 4.11, *p* < .001, *d* = 1.30, 95% CI [3.21, 9.59].

✅ State the means and SDs for **both** groups.

Practice question: How many participants were in this study?

A. 38

B. 40

C. 38 per group (76 total)

D. 36

Formatting Rules (APA 7th Edition)

- Italicize *t*, *z*, *p*, *M*, *SD*, *F*, *r*
- Use **two decimal places** for statistics (except *p*)
- Report *p* values as:
  - *p* = .042 (no leading zero)
  - *p* < .001 (for very small values)
- Match decimal precision across CIs and descriptive stats

Putting It All Together

Always Include

- Test type and statistic (*t*, *z*, etc.)
- df (for *t* tests)
- *p* value
- Descriptive stats (*M*, *SD*)

Usually Include

- Effect size (*d*, *r*, etc.)
- Confidence interval
- Directional phrasing ("significantly greater than…")

APA Reporting: Practice 🧠

Rewrite each in APA style:

1. The Caffeine group got an average of 14.2 right answers (standard deviation = 3), and the Placebo group got 11.1 right answers (SD 2.8). The difference was statistically significant with t = 3.2 and p=0.004, df = 28. The Effect size was D= 0.79 and 95 percent CI = (0.50 TO 4.80).

2. Participants slept on average 7.6 Hours (s.d.=0.8) which was not significantly different from the national Mean of 7.5. The T test was not significant (P = .66; t = .45; DF=19). Cohen’s D = 0.06 and the 95% confidence interval was (-0.42 , 0.62).

Practice: Answers

Below are the **APA-corrected versions** of those two messy examples.

☕ Correct Example 1 — Independent-Samples *t*

> The caffeine group (*M* = 14.2, *SD* = 3.0) scored higher than the placebo group (*M* = 11.1, *SD* = 2.8), *t*(28) = 3.20, *p* = .004, *d* = 0.79, 95% CI [0.50, 4.80].

What was fixed?

- Italicize statistical symbols (*t*, *p*, *d*, *M*, *SD*).
- Report *t*(df) = value, *p*, *d*, and 95% CI in this order.
- Use brackets for the CI and **no leading zeros** for *p* values less than 1.

Practice: Answers (Continued)

😴 Correct Example 2 — One-Sample *t*

> Participants slept an average of 7.6 hours ($SD = 0.8$), which did not differ significantly from the national mean of 7.5, $t(19) = 0.45$, $p = .66$, $d = 0.06$, 95% CI [−0.42, 0.62].

What was fixed?

- Write numbers and units clearly (“7.6 hours”).
- Keep $p = .66$ (not $P = 0.66$).
- Report CI in brackets.

Moving to our next topic...

One-Way ANOVA

Are We Worse Drivers When on the Phone?

A simple *t*-test compares driving ability when talking vs. not talking on the phone.

But what if we wanted to compare **more than two conditions**?

💭 Possible conditions:
– Driving alone
– With a passenger
– On a cell phone
– On a video call

Can I Use a *t* Test to Compare >2 Groups?

If you used a *t*-test for every possible combination, you'd run many tests!

- Driving alone vs. passenger
- Driving alone vs. video call
- Driving alone vs. cell phone
- Passenger vs. video call
- Passenger vs. cell phone
- Video call vs. cell phone

That's **6 *t*-tests!**

Is there any problem with that? 🤔
Yes — it **increases the chance of a Type I error** (false positive).

Why You Can't Do Many *t* Tests

- Each test carries a 5% chance (α = .05) of a **Type I error** — falsely rejecting the null.
  - There is a 5% chance of concluding there is a difference when there really isn't one.
  - There is a 95% chance of correctly *failing* to reject the null.
- The more *t*-tests you do, the higher your overall risk of error.
- This problem is called **alpha inflation**.

With one test → 95% chance of correctly retaining the null & 5% chance of a false positive

With two tests -> 95% chance of correctly retaining the null **twice** → $(0.95)^2 = 0.903$ → 10% chance of error

With three tests → 95% chance of correctly retaining the null **three** times → $(0.95)^3 = 0.857$ → 14% chance of error

Why You Can't Do Many *t* Tests (continued)

So if you did 6 *t*-tests (like in our example), you'd have **a 54% chance of a Type I error!**

> **Conclusion:** Multiple t-tests inflate the error rate. We need a single test for 3+ groups.

The Solution: Using the *F* Statistic

When we want to compare **3 or more means**, we use the **F distribution** — the foundation of ANOVA.

Like *z* and *t* tests, the *F* statistic is a **ratio**:

$$z = \frac{\text{Difference Between Means}}{\text{Standard Error}}$$

$$t = \frac{\text{Difference Between Means}}{\text{Standard Error}}$$

$$F = \frac{\text{Between-Groups Variance}}{\text{Within-Groups Variance}}$$

The Solution: Using the *F* Statistic

$$F = \frac{\text{Between-Groups Variance}}{\text{Within-Groups Variance}}$$

- The **numerator** captures how far apart the group means are.
- The **denominator** captures how much variability exists within each group.

The bigger the ratio, the more evidence that not all groups come from the same population.

Intuition for the *F* Statistic

- Think of the *F* statistic as an expansion of the *z* and *t* statistics:
  - *z* → can do one thing
  - *t* → can do a few things
  - *F* → can do lots of things
- Each builds on the same idea: comparing variability due to chance vs. systematic differences.

> The *F* statistic captures both the **differences between groups** and the **noise within them**.

Which part of this equation captures systematic differences between groups?

Which *part captures "noise" within groups?*

$$F = \frac{\text{Between-Groups Variance}}{\text{Within-Groups Variance}}$$

Characteristics of the *F* Distribution

- *F* is a **ratio of two variances** (between-groups / within-groups).
  - Variances are based on sums of squares (SS), so *F* is always **positive**.
- The *F* distribution is a *series* of distributions (like the *t* distribution)
  - There are **two values for degrees of freedom** in every *F* test:
    - One for the numerator (df_1 = between-groups)
    - One for the denominator (df_2 = within-groups)

Like *t*, the *F* distribution changes shape depending on the sample size and number of groups.

To use *F*, your data must be on a **numeric (interval or ratio)** scale.

The *F* Table and Degrees of Freedom

- The *F* table expands the *t* table by adding **another dimension** for the number of groups.
- There's an *F* distribution for every combination of:
  - Sample size (→ df_1, numerator)
  - Number of groups (→ df_2, denominator)

F(df_1, df_2) helps us decide whether group differences are larger than expected by chance.

In practice, most people don't use the F table directly — statistical software calculates the exact p-value for you!

The important thing to understand is that your *p* value depends on both degrees of freedom, which are determined by:

1. your sample size and
2. your number of groups.

ANOVA Overview

An **ANOVA** (*Analysis of Variance*) compares **differences between 3+ groups using one test.**

One-way between-groups ANOVA

- Hypothesis test used to compare means across **more than two independent groups**:
  - Groups are defined by a single independent variable (IV) with 3+ levels.
    - These levels are **categorical** (nominal/ordinal).
    - A between-groups ANOVA is used for a design where each participant appears in only one group.
  - The dependent variable (DV) is **numeric** (interval/ratio).

Example: Comparing mean driving performance across four phone-use conditions.

Phone-use conditions: Categorical IV with 4 levels

Driving performance: Numeric DV

ANOVA: What Are We Analyzing?

Even though the test is called an **analysis of variance**, what we're really doing is comparing **means**.

We ask:
Are the differences among group means larger than we would expect from random chance?

If the groups come from the same population, their means should be similar.

If the means are *very different*, that suggests at least one group differs systematically.

The Two Sources of Variability

1️⃣ Between-Groups Variance

- How far apart are the **group means**?
- Captures **systematic differences** due to the independent variable.

2️⃣ Within-Groups Variance

- How spread out are scores **inside each group**?
- Captures **unsystematic noise** — individual differences or measurement error.

> When between-group variability ≫ within-group variability, the *F* ratio becomes large → **evidence against the null**.

How Do We Compute Between vs. Within Variability?

ANOVA divides the **total variability** in a dataset into two parts:

$$SS_{\text{Total}} = SS_{\text{Between}} + SS_{\text{Within}}$$

Recap: Sum of Squares (SS) Refresher

To calculate variance, we compute the **sum of squared deviations** from the mean.

1. Subtract the mean from each score.
2. Square each deviation.
3. Add them up.

$$SS = \sum (X - \bar{X})^2$$

- This is the **Sum of Squares (SS)**.

1️⃣ Between-Groups Variability ($SS_{Between}$)

- Start with the **mean of each group** (M_1, M_2, M_3, …).
- Compute how far each group mean is from the **grand mean** (M_G, the mean of all scores).
- Weight each squared deviation by the **group size** (*n*).
- Add them up.

$$SS_{\text{Between}} = \sum n_i(M_i - M_G)^2$$

This reflects the variability **explained by group membership**.

2️⃣ Within-Groups Variability (SS_{Within})

- For each group, measure how much each individual score deviates from its group mean.
- Square those deviations and sum them across all groups.

$$SS_{\text{Within}} = \sum\sum(X_{ij} - M_i)^2$$

This captures **unexplained variability** — random noise and individual differences.

Why are there two summation symbols?

- The inner Σ says: For a given group i, sum the squared deviations of each person's score (j) from that group's mean
- The outer Σ says: Now repeat that process for each group, and add them all up.

Converting Sums of Squares to the *F* Statistic

We then divide each sum of squares by its degrees of freedom to obtain **mean squares (MS):**

$$MS_{\text{Between}} = \frac{SS_{\text{Between}}}{df_{\text{Between}}}, \quad MS_{\text{Within}} = \frac{SS_{\text{Within}}}{df_{\text{Within}}}$$

Where:

- $df_{\text{Between}} = k - 1$ (k = number of groups)
- $df_{\text{Within}} = N - k$ (N = total sample size)

Finally, we compare the two:

$$F = \frac{MS_{\text{Between}}}{MS_{\text{Within}}}$$

A large *F* value indicates that between-group variability is much greater than within-group variability → evidence against the null hypothesis.

The Logic of the *F* Statistic

At its core, ANOVA is a **signal-to-noise ratio**:

$$F = \frac{\text{Systematic variance (between groups)}}{\text{Random variance (within groups)}}$$

- **Signal** → variability explained by your independent variable
- **Noise** → variability due to chance

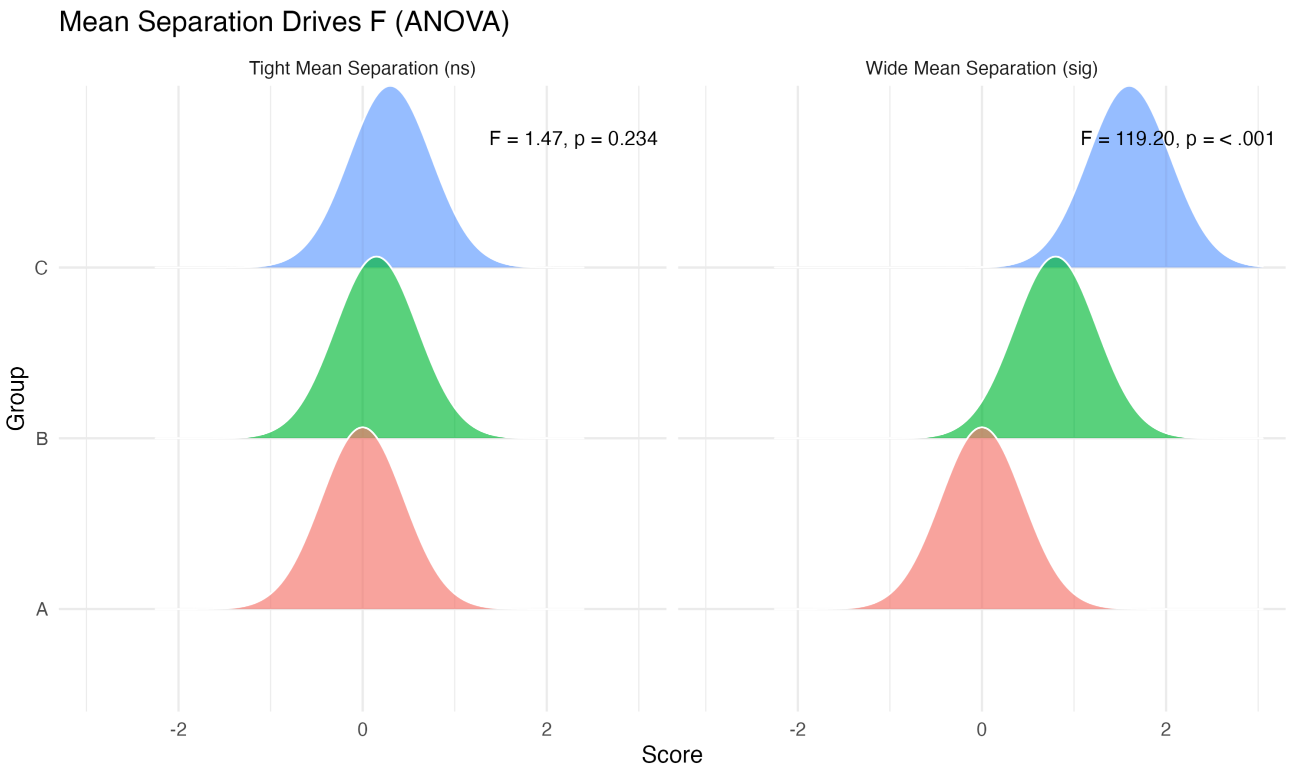
If *F* ≈ 1 → group differences are about what chance would produce.
If *F* >> 1 → the manipulation explains meaningful variance.

Interpreting *F*

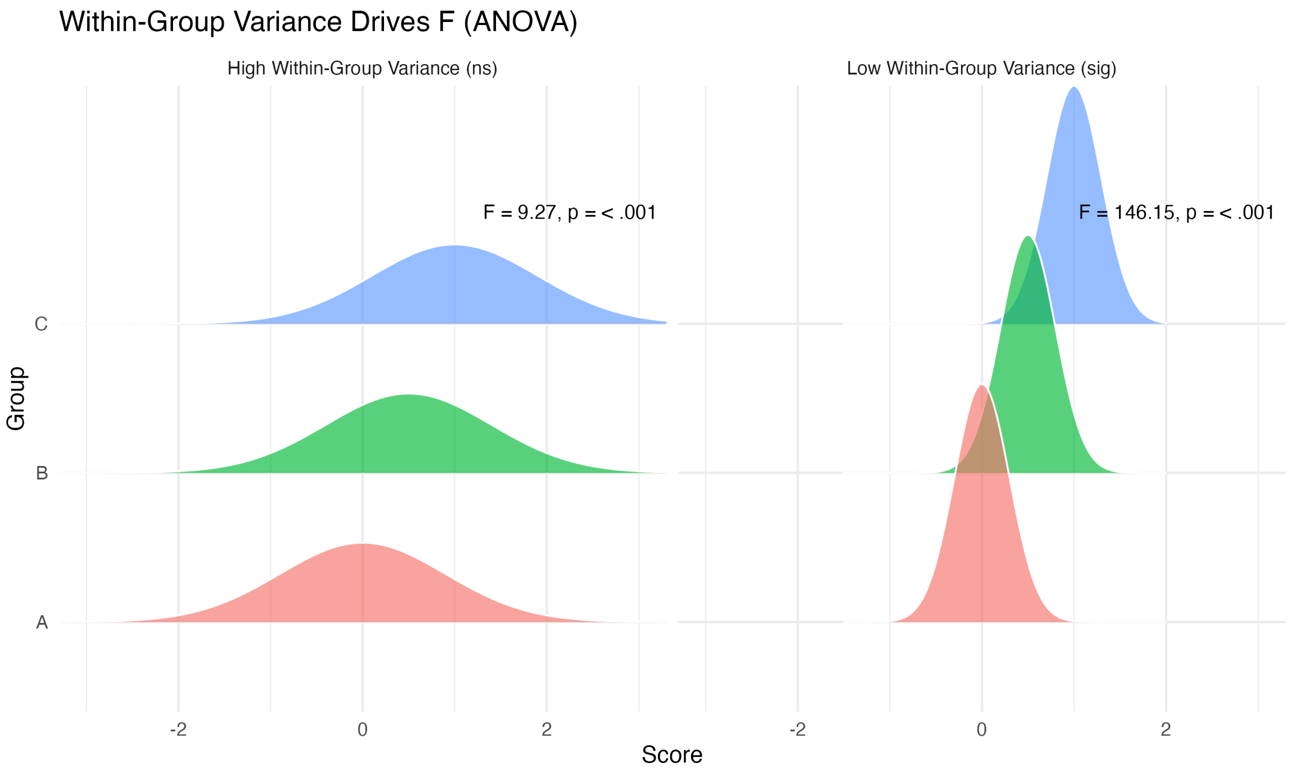
- *F* values can't be negative.
- The **larger the *F***, the stronger the evidence that not all group means are equal.
- Whether an *F* is "large enough" depends on its **critical value** from the *F* distribution (df_1, df_2).

If your observed *F* > critical *F*, you reject H_0 and conclude that at least one mean differs from the others.

The F Statistic Visualized: Between-Group Variance



The F Statistic Visualized: Within-Group Variance



Setting Up the Hypotheses

Null hypothesis (H_0):
All population means are equal.

$$\mu_1 = \mu_2 = \mu_3 = \dots$$

Research hypothesis (H_1):
At least one population mean differs.

> Unlike a *t*-test, ANOVA does not tell us *which* means differ —
> only that **not all of them are the same**.

Why Use ANOVA Instead of Multiple *t* Tests?

- Keeps the **Type I error rate** at 5%.
- Tests all group differences **in one analysis**.
- Provides the foundation for **multi-factor and repeated-measures** designs.
  - We will talk about repeated-measures designs soon.

ANOVA is the statistically correct, efficient way to ask:

"Are these group means different enough that it's unlikely to be chance?"

Assumptions of One-Way ANOVA

Every ANOVA relies on three key assumptions, which represent the ideal conditions for valid results:

1. **Random selection** of participants

2. **Normality** — each group's population is roughly normal

3. **Homogeneity of variance** — groups have similar spread

- Homoscedastic populations are those that have the same variance.
- Heteroscedastic populations are those that have different variances.

When these hold, *F* is robust and trustworthy.
If they're violated, be cautious in interpreting results, or may need to use non-parametric alternatives.

Example: One-Way ANOVA (Source Table)

Imagine we conduct a one-way ANOVA comparing driving performance across four driving conditions: 1. No phone, 2. Passenger, 3. Cell phone, and 4. Video call.

In total, we have 40 participants divided among the four groups (n = 10 per group).

The between-groups and within-groups sums of squares are:

Source	SS	df	MS	F
Between Groups	450.0			
Within Groups	100.0			
Total	550.0			

What should go in the DF column?

Example: One-Way ANOVA (Source Table)

The between-groups and within-groups sums of squares are:

Source	SS	df	MS	F
Between Groups	450.0	K-1 = 3		
Within Groups	100.0	N-K = 36		
Total	550.0	39		

What should go in the MS column?

Example: One-Way ANOVA (Source Table)

The between-groups and within-groups sums of squares are:

Source	SS	df	MS	F
Between Groups	450.0	K-1 = 3	SS/df = 150.0	
Within Groups	100.0	N-K = 36	SS/df = 2.78	
Total	550.0	39		

What should go in the F column?

Example: One-Way ANOVA (Source Table)

The between-groups and within-groups sums of squares are:

Source	SS	df	MS	F
Between Groups	450.0	K-1 = 3	SS/df = 150.0	MS_Between / MS_Within = 54.0
Within Groups	100.0	N-K = 36	SS/df = 2.78	
Total	550.0	39		

Final ANOVA result: F(3, 36) = 54.0

We need to look up the critical F value for (3, 36) df to determine significance, or use software to get the exact p-value.

Beyond Significance: Measuring Effect Size

A significant *F* tells us that not all group means are equal —
but **how big** is that difference in practical terms?

That's where **effect size** comes in.
For ANOVA, we most often report **η^2 (eta squared)** or **R^2**. (In a one-way ANOVA, they are equivalent.)

$$\eta^2 = \frac{SS_{\text{Between}}}{SS_{\text{Total}}}$$

It represents the **proportion of total variance** in the dependent variable that is explained by the independent variable.

Understanding η² and R² Conceptually

- Think of η^2 like a “percentage of variance explained.”
- $\eta^2 = .20$ means 20 % of the variability in your data is accounted for by your experimental manipulation (group membership).
- Larger η^2 → stronger effect.

🧠 Quick interpretation guide (Cohen, 1988):
• Small ≈ .01 • Medium ≈ .06 • Large ≈ .14

These cutoffs are rough — context matters.

Example: Reporting Effect Size

Imagine we find:

> $F(2, 36) = 5.47$, $p = .008$, $\eta^2 = .23$

That means **23 %** of the variance in scores is explained by our independent variable.

Including effect sizes helps readers judge:

- whether the difference is **meaningful**, not just statistically significant,
- and whether findings might replicate with new samples.

When *F* Is Significant: What’s Next?

- A significant ANOVA tells us that **at least one group differs**, but not *which* ones.
- To find out *which* groups differ, we use **follow-up tests** called *post-hoc comparisons*.

Why Not Just Run More *t*-Tests?

Because each extra *t*-test increases the risk of **Type I error** (false positives).

Post-hoc procedures control that risk by adjusting the critical value or significance threshold.

Planned vs. Post-Hoc Comparisons

Planned (a priori)

- Decided *before* data collection
- Test specific hypotheses
- Fewer tests → no need for strong correction

Post-hoc

- Done *after* seeing results
- Explore where the differences lie
- Need corrections to keep α = .05 overall

> Think of post-hoc tests as “honest ways to peek under the hood” after finding a significant overall *F*.
>
> In general, with post-hoc tests, we adjust the size of the effect needed to declare significance to control for multiple comparisons.

Common Post-Hoc Tests

Test	Key Idea	Conservative?	Typical Use
Scheffé	Controls α for all possible contrasts	Most conservative	Exploratory analyses
Tukey HSD	Tests all pairwise mean comparisons equally	Moderate	Balanced designs
Bonferroni	Divides α by number of comparisons	Flexible	Small number of tests

All aim to protect against **false positives** while allowing fair comparisons among multiple groups.

Conceptual Takeaway

- **Scheffé:** “Play it safe” — harder to reach significance
- **Tukey HSD:** “Middle ground” — good balance of safety and power
- **Bonferroni:** “Divide and conquer” — simple, but can be overly strict when many tests

Different journals or software packages may default to different methods — always report which you used.

Bonferroni example: You run an ANOVA with 4 groups (6 pairwise comparisons) and find a significant effect. You now want to know *which* groups differ. To keep overall α = .05, each test must meet $p < .0083$ (0.05/6) to be significant.

That’s all for today!

Tuesday: Exam 3 Review Session featuring lots of practice questions