Lecture 11: ANOVA & A/B Testing

Comparing Many Groups · Experimenting at Scale

Previously, on Lecture 10...

LRT: Compare restricted vs full model. $-2\log\Lambda \sim \chi^2_k$ (Wilks' theorem).

t-tests: One-sample, paired, Welch's two-sample. The workhorses of inference.

χ^2 tests: Goodness-of-fit (one variable) and independence (two categorical variables).

Nonparametric: Mann–Whitney and Wilcoxon when normality fails.

Flowchart: Data type $\rightarrow$ number of groups $\rightarrow$ paired? $\rightarrow$ specific test.

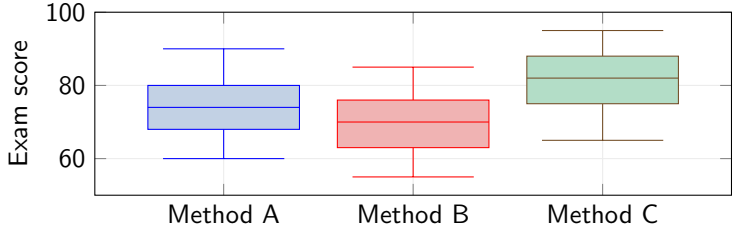
Today: What if we have **three or more** groups? And how does industry actually run experiments?

Part I: The Problem

Three groups, three t-tests — what could go wrong?

Comparing More Than Two Groups

Scenario: A teacher tries three teaching methods (A, B, C) on different classes. Are exam scores different?



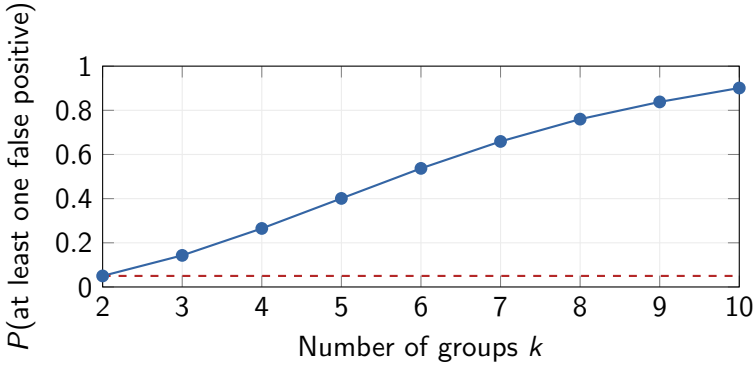
Naïve approach: Run three pairwise t-tests: A vs B, A vs C, B vs C.

Problem: Each test has $\alpha = 0.05$ false positive rate. Three tests?

$$P(\text{at least one false positive}) = 1 - (1 - 0.05)^3 = 0.143$$

With 10 groups: $\binom{10}{2} = 45$ tests $\Rightarrow 1 - 0.95^{45} = 0.90$. Almost certain to find "something."

The Multiple Comparisons Explosion

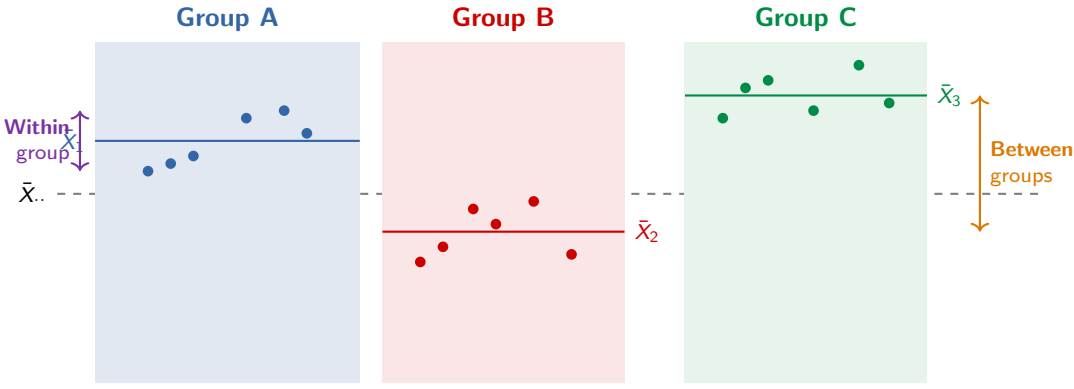


We need a single test that asks: "Are *any* of these group means different?"
while controlling the overall Type I error rate at α.
That test is **ANOVA** (Analysis of Variance).

Part II: One-Way ANOVA

One factor, k groups — split the total variability

The Core Idea: Between vs Within



If group means are truly equal, the "between-group" variability should be no larger than the "within-group" variability (just random noise).

Sum of Squares Decomposition

$$\underbrace{\sum_{i,j}(X_{ij} - \bar{X}_{..})^2}_{\text{SS}_{\text{Total}}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2}_{\text{SS}_{\text{Between}}} + \underbrace{\sum_{i,j}(X_{ij} - \bar{X}_{i.})^2}_{\text{SS}_{\text{Within}}}$$

SS_{Total}
Total spread of all observations around $\bar{X}_{..}$
$\text{df} = N - 1$

$=$

$\text{SS}_{\text{Between}}$
Spread of *group means* around $\bar{X}_{..}$
$\text{df} = k - 1$

$+$

$\text{SS}_{\text{Within}}$
Spread around *own* group mean
$\text{df} = N - k$

The F-Statistic

$$F = \frac{\text{MS}_{\text{Between}}}{\text{MS}_{\text{Within}}} = \frac{\text{SS}_{\text{Between}}/(k-1)}{\text{SS}_{\text{Within}}/(N-k)} \sim F_{k-1,\,N-k}$$

$$\text{under } H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$F \approx 1$: No evidence

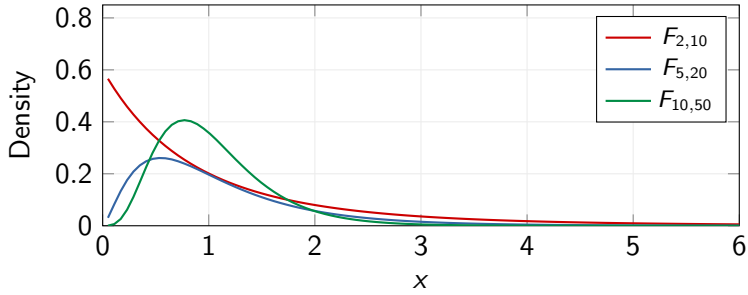
Between-group variance $\approx$ within-group variance.
Group means are similar.
$\Rightarrow$ Fail to reject H_0.

$F \gg 1$: Evidence

Between-group variance $\gg$ within-group variance.
Group means differ.
$\Rightarrow$ Reject H_0.

F is always ≥ 0. We only reject in the **right tail** (one-sided).

The F-Distribution



Key facts: Ratio of two independent χ^2 variables, each divided by their df. Right-skewed; mean ≈ 1 when H_0 is true (for large denominator df).
Special case: $F_{1,\,n-2} = t^2_{n-2}$. The t-test is a special case of ANOVA with $k = 2$.

The ANOVA Table

Source	SS	df	MS	F
Between (Treatment)	SS_B	$k-1$	$SS_B/(k-1)$	MS_B/MS_W
Within (Error)	SS_W	$N-k$	$SS_W/(N-k)$	
Total	SS_T	$N-1$		

Example: Three methods, $n_i = 10$ each ($N = 30$).

Source	SS	df	MS	F
Between	720	2	360	$360/40 = \mathbf{9.0}$
Within	1080	27	40	
Total	1800	29		

$F_{2,\,27,\,0.05} = 3.35$. Since $F = 9.0 > 3.35$: **reject** H_0.
At least one teaching method gives different results. But which one?

ANOVA: Assumptions

1. **Independence:** Observations within and between groups are independent.
   Violated by: repeated measures, clustered data, time series.

2. **Normality:** Each group is (approximately) normally distributed.
   Robust to violations when $n_i \geq 20$ (CLT). Check with QQ plots.

3. **Equal variances (homoscedasticity):** $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.
   Rule of thumb: OK if largest S_i / smallest $S_i < 2$.
   Test with Levene's test. Alternative: Welch's ANOVA.

If equal-variance assumption fails: use **Welch's ANOVA** (`pingouin.welch_anova`).
Note: `scipy.stats.f_oneway` does *standard* ANOVA (assumes equal variances).
Post-hoc for unequal variances: Games–Howell instead of Tukey HSD.
If normality fails: use **Kruskal–Wallis test** (nonparametric ANOVA).

Part III: Post-Hoc Tests

ANOVA says "something differs." Post-hoc says *what.*

Post-Hoc: Which Pairs Are Different?

Only run post-hoc tests *after* ANOVA rejects H_0

Tukey's HSD

Honestly Significant Difference.

Compares *all* pairs.

Controls **family-wise** error rate.

Uses studentized range (q).

Best for: equal n_i, all pairs.

Bonferroni

Divide α by number of tests.

Works with any comparisons.

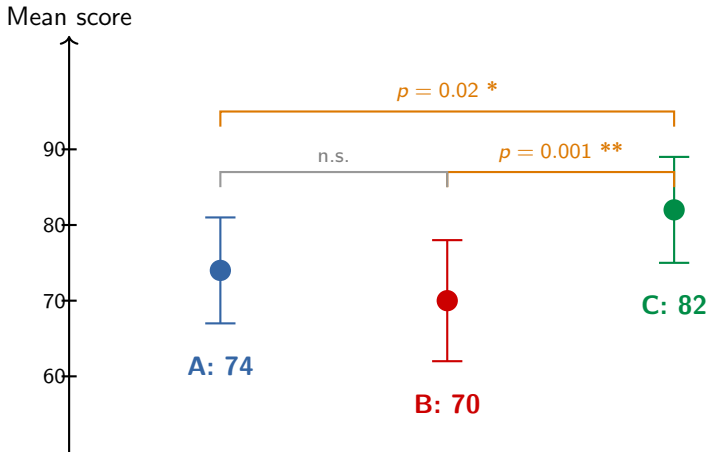
Conservative (wider CIs).

We already know this from L9!

Best for: few planned comparisons.

Teaching methods example: Tukey's HSD finds that C > A ($p = 0.02$) and C > B ($p = 0.001$),
but A vs B is not significant ($p = 0.38$). Method C is the winner.

Visualizing Post-Hoc Results



Error bars:
95% Tukey
HSD intervals

If two intervals **don't overlap**, the difference is significant. Method C is significantly better.

Part IV: Beyond the Basics

Effect sizes and two-factor designs

Effect Size: η^2 (Eta-Squared)

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

Proportion of total variance explained by group membership.

Small	**Medium**	**Large**
$\eta^2 = 0.01$	$\eta^2 = 0.06$	$\eta^2 = 0.14$

Teaching example: $\eta^2 = 720/1800 = 0.40$. **Very large** — 40% of score variation is explained by teaching method.

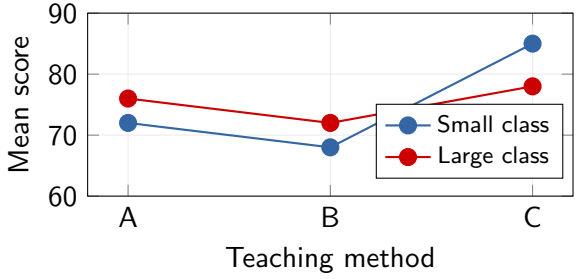
Like Cohen's *d* for *t*-tests, η^2 answers: how **big** is the effect?
A small *p*-value with tiny η^2 means: real but unimportant.
Note: η^2 is biased upward for small samples. Use ω^2 (omega-squared) for less bias.

Two-Way ANOVA: Two Factors at Once

Example: Teaching method (A/B/C) × class size (small/large)

$$SS_{Total} = SS_{Method} + SS_{Size} + SS_{Interaction} + SS_{Within}$$

Now we get **three** F-tests: one for each factor, one for their interaction.

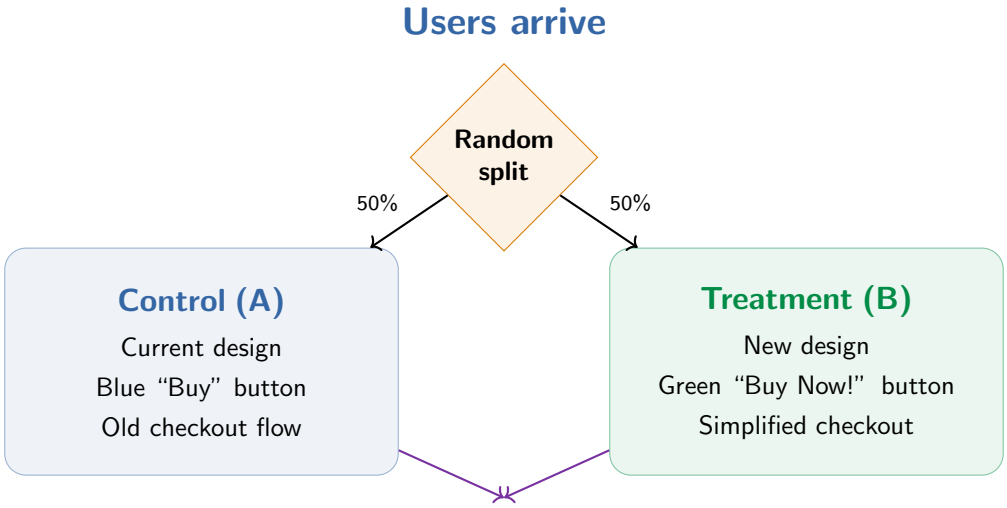


Interaction: Method C shines in small classes but not in large ones.
Lines are **not parallel** ⇒ significant interaction. The effect of method *depends on* class size.

Part V: A/B Testing

Hypothesis testing meets the real world

What Is A/B Testing?



A/B Testing Is Just Hypothesis Testing

Step 1 — Hypothesis: $H_0 : p_B = p_A$ (no difference). $H_1 : p_B \neq p_A$ (or $p_B > p_A$).

Step 2 — Sample size: Choose n *before* the test using power analysis (L9). Fix α, β, MDE.

Step 3 — Randomize: Randomly assign users to A or B. Run until planned n is reached.

Step 4 — Analyze: Two-proportion z-test or Welch's t-test depending on metric.

Step 5 — Decide: If $p < \alpha$ **and** effect is practically significant $\Rightarrow$ ship B.

MDE = Minimum Detectable Effect.
The smallest improvement worth detecting. Business decides this, not statistics.

Sample Size Planning for A/B Tests

For comparing two proportions (p_A vs p_B)

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2 \cdot (p_A(1-p_A) + p_B(1-p_B))}{(p_B - p_A)^2}$$

n = per group. This is the two-proportion version from L9's power formula.

Example: Baseline conversion $p_A = 5\%$. Want to detect $p_B = 6\%$ (MDE = 1 pp). $\alpha = 0.05$, power = 80% ($z_{0.025} = 1.96$, $z_{0.20} = 0.84$).

$$n = \frac{(1.96 + 0.84)^2 \times (0.05 \times 0.95 + 0.06 \times 0.94)}{(0.01)^2} = \frac{7.84 \times 0.1039}{0.0001} \approx \mathbf{8{,}146}$$

Need ~**8,000 users per group** (16,000 total) to detect a 1% lift in conversion.
Small effects need big samples. This is why A/B tests take weeks.

A/B Test Example: Button Color

Setup: $n_A = 5{,}000$ (blue button), $n_B = 5{,}000$ (green button).
Results: $k_A = 230$ conversions ($\hat{p}_A = 4.6\%$), $k_B = 275$ ($\hat{p}_B = 5.5\%$).

Pooled proportion: $\hat{p} = (230 + 275)/(5000 + 5000) = 0.0505$.

$$Z = \frac{0.055 - 0.046}{\sqrt{0.0505 \times 0.9495 \times (1/5000 + 1/5000)}} = \frac{0.009}{0.0044} = 2.05$$

p-value $= 2(1 - \Phi(2.05)) = 0.040 < 0.05$: **reject** H_0.

+19.6% relative lift
$p = 0.040$

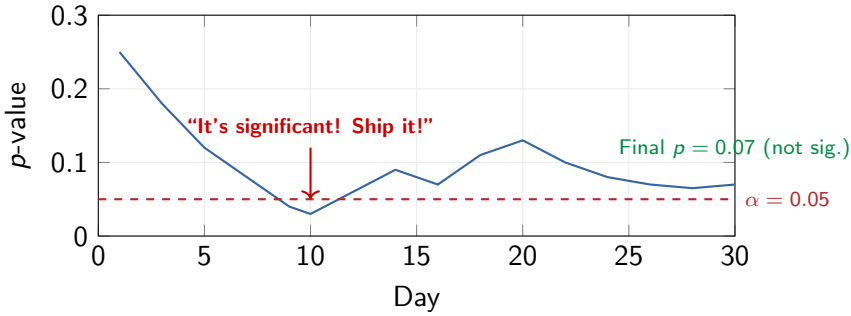
Decision: Statistically significant. +19.6% relative lift is practically meaningful. Ship the green button.

Absolute vs relative: MDE and sample size formulas use *absolute* differences (0.9 pp here). Lift is usually reported *relative* ($0.009/0.046 = 19.6\%$). Don't confuse the two!

Part VI: A/B Testing Pitfalls

The theory is simple. The practice is full of traps.

Pitfall 1: Peeking (The Most Common Mistake)



If you check daily and stop when $p < 0.05$, the true false positive rate can be **26%** or higher, not 5%. The p-value fluctuates — early dips are noise.

Fix: Decide sample size *before* the test. Don't peek.

Pitfall 2: Too Many Metrics

Conversion rate

Revenue per user

Time on page

Bounce rate

Pages per session

Cart abandonment

Newsletter signup

Customer satisfaction

Test 8 metrics at $\alpha = 0.05$: $P(\text{at least one false positive}) = 1 - 0.95^8 = 0.34$.

Fix: Pick **one primary metric** before the test. Others are exploratory.
If you must test many, apply Bonferroni or BH correction (L9).

More Pitfalls

3. Novelty & primacy effects: Users click the new thing *because* it's new, not because it's better. Effect fades after 1–2 weeks. **Fix:** run test long enough (2+ weeks).

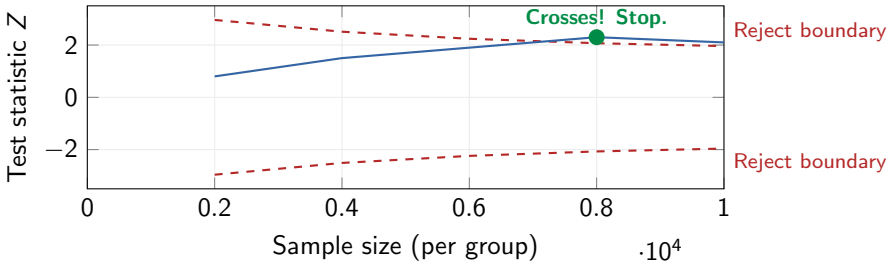
4. Simpson's paradox: Overall B wins, but A wins in *every* segment (mobile, desktop, tablet). Caused by unequal traffic mix. **Fix:** stratified randomization.

5. Interference / network effects: If user A's experience affects user B (social networks, marketplaces), standard randomization breaks. **Fix:** cluster randomization (by region, by network cluster).

6. Survivorship bias: Only analyzing users who *completed* the funnel. Ignoring those who dropped off. **Fix:** intent-to-treat analysis.

Sequential Testing: A Principled Way to Peek

Sometimes you *need* to check results early (e.g., detecting harm).



O'Brien–Fleming boundaries: tighter at early looks (harder to reject), relaxing over time.

Overall Type I error is still $\alpha = 0.05$, but you can stop early if the effect is large.

Common myth: "Bayesian A/B testing solves the peeking problem." It doesn't — naive Bayesian stopping still inflates errors.

Python: `statsmodels.stats.GroupSequential`

ANOVA and A/B Testing: Same Core, Different Worlds

	ANOVA	**A/B Testing**
Setting	Science, medicine, education	Tech, product, marketing
Groups	2+ (often 3–5)	Usually 2 (A vs B)
Design	Controlled experiment	Online randomized experiment
Metric	Continuous (scores, times)	Often proportions (conversion)
Analysis	F-test + post-hoc	z-test or t-test
Pitfalls	Assumptions, multiple comp.	Peeking, novelty, interference
Effect size	η^2	Relative lift (%)

Both are hypothesis tests. ANOVA generalizes to k groups. A/B testing adds the engineering of randomization, sample size, and deployment.
An A/B test with 3+ variants is literally ANOVA (sometimes called A/B/n testing).

When To Use What

2 groups, continuous metric: Welch's t-test (or Mann–Whitney if non-Normal). From L10.

2 groups, proportions: Two-proportion z-test. The classic A/B test.

$k \geq 3$ **groups, continuous:** One-way ANOVA $\rightarrow$ post-hoc (Tukey). Or Kruskal–Wallis if non-Normal.

$k \geq 3$ **groups, proportions:** χ^2 test of homogeneity ($= \chi^2$ independence test on proportions).

Two factors (e.g., method $\times$ class size): Two-way ANOVA. Look at main effects and interaction.

Need to peek at results early: Sequential testing (O'Brien–Fleming). Pre-plan the number of looks.

Summary: ANOVA & A/B Testing

Problem: Multiple t-tests inflate false positives. Need a single test for k groups.

ANOVA: $F = \mathrm{MS}_{\mathrm{Between}}/\mathrm{MS}_{\mathrm{Within}}$. Large $F \Rightarrow$ group means differ.

Post-hoc: Tukey HSD or Bonferroni to find *which* pairs differ.

Effect size: $\eta^2 = \mathrm{SS}_{\mathrm{B}}/\mathrm{SS}_{\mathrm{T}}$. How much variance is explained by groups.

A/B testing: Hypothesis testing in tech. Same math, different engineering.

Sample size: Small effects need big samples. Plan n before running the test.

Pitfalls: Peeking, too many metrics, novelty effects, interference.

Sequential: O'Brien–Fleming lets you peek with controlled error rate.

Practical: ANOVA & A/B Testing in Python

1. **One-way ANOVA:**
   - Generate 3 groups with `np.random.normal` (same vs different means)
   - Run `scipy.stats.f_oneway`. Check F and p
   - Post-hoc with `statsmodels.stats.multicomp.pairwise_tukeyhsd`
2. **Assumptions check:**
   - Levene's test: `scipy.stats.levene`
   - Kruskal–Wallis: `scipy.stats.kruskal`
3. **A/B test simulation:**
   - Simulate n Bernoulli trials for A and B. Run z-test
   - Repeat 10,000 times. Check: is the rejection rate $\approx \alpha$?
4. **Peeking simulation:**
   - Simulate with $p_A = p_B$ (no effect). Check daily
   - Count how often $p < 0.05$ at *any* day. Compare to 5%

Homework

1. Three fertilizers tested on plant height ($n = 15$ per group):
   A: $\bar{X}_1 = 22.1$, $S_1 = 3.2$. B: $\bar{X}_2 = 25.4$, $S_2 = 3.5$. C: $\bar{X}_3 = 24.8$, $S_3 = 2.9$.
   (a) Compute SS_B, SS_W, F. (b) Test at $\alpha = 0.05$ ($F_{2,42,0.05} = 3.22$).
   (c) Compute η^2. (d) Which fertilizer is best? (Use Bonferroni with $\alpha/3$.)
   Hint: $SS_W = \sum(n_i - 1)S_i^2$. $\bar{X}_{\cdot\cdot} = (\bar{X}_1 + \bar{X}_2 + \bar{X}_3)/3$ when n_i are equal.

2. An e-commerce site tests two checkout flows ($n = 3{,}000$ per group):
   Control: 156 conversions. Treatment: 189 conversions.
   (a) Compute $\hat{p}_A$, $\hat{p}_B$, Z, p-value. (b) Is the result significant at $\alpha = 0.05$?
   (c) What is the relative lift? (d) How many users would you need to detect a 0.5% absolute lift?

3. **Peeking simulation** (Python): Simulate 1,000 A/B tests where $p_A = p_B = 0.05$ (no true effect), $n = 5{,}000$ per group. For each test, compute the p-value after every 500 users. What fraction of tests show $p < 0.05$ at *any* checkpoint? Compare to $\alpha = 0.05$.

4. A 3-way A/B/C test: $n_A = 1000$, $k_A = 48$; $n_B = 1000$, $k_B = 55$; $n_C = 1000$, $k_C = 62$. Test using χ^2 test of homogeneity. Then do pairwise z-tests with Bonferroni.

Recommended Resources

Interactive: Seeing Theory — ANOVA (Brown University)
`seeing-theory.brown.edu/regression-analysis` — visual F-test with adjustable groups.

Video: StatQuest — ANOVA
Clear walkthrough of one-way ANOVA, F-statistic, and post-hoc tests. Also: "A/B Testing."

Reading: Kohavi, Tang & Xu — "Trustworthy Online Controlled Experiments"
The definitive book on A/B testing from Microsoft researchers. Practical and rigorous.

Blog: Evan Miller — "How Not to Run an A/B Test"
Classic blog post on the peeking problem. Required reading for anyone running A/B tests.

Python: `scipy.stats` + `statsmodels`
`f_oneway`, `kruskal`, `levene`, `pairwise_tukeyhsd`, `proportions_ztest`.

Questions?

Next: Lecture 12 — Regression inference