Lectures 10–11: Classical Tests, ANOVA & A/B Testing

A tour of the named tests · then one principle that unifies them

Previously, on Lecture 9...

Logic: Assume H_0 (nothing happening). Ask: how surprising is my data under H_0?

p-value: $P(\text{this extreme or more} \mid H_0)$. **NOT** $P(H_0 \mid \text{data})$.

Errors: Type I (α) = false alarm. Type II (β) = missed detection. Power $= 1 - \beta$.

Tables: z-table when σ known. t-table when σ unknown (df $= n - 1$).

Tools: Permutation tests (no assumptions). Multiple testing: Holm/BF (FWER) or BH (FDR).

Today: The framework is clear. But which **specific test** do I actually use?

Today's Plan: A Tour, Then a Unification

We'll meet the named tests one by one, like tools on a shelf. At the end, we'll pull back and see they're all the same machine.

Part I Tests for means
one-sample, paired, two-sample t

Part II Proportions & counts
z-test, χ^2 goodness-of-fit, independence

Part III When assumptions fail
Mann–Whitney, Wilcoxon

Part IV Decision flowchart + Python

Part V The reveal: "wait, these are all the same test" — the Likelihood Ratio framework.

Parts VI–IX Then: 3+ groups (ANOVA), A/B testing, pitfalls, big picture.

Part I: Tests for Means

One-sample, paired, two-sample

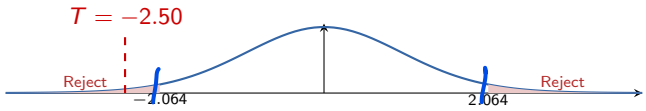
One-Sample t-Test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \text{under } H_0$$

Example: Coffee shop claims cups contain $\mu_0 = 350$ mL. We measure $n = 25$: $\bar{X} = 345$, $S = 10$.

$$T = \frac{345 - 350}{10/\sqrt{25}} = \frac{-5}{2} = -2.50, \qquad p = 2 \cdot P(t_{24} < -2.50) \approx 0.020$$



$p = 0.020 < 0.05$: **reject** H_0. The cups are underfilled.

Assumptions: (1) Approx. Normal (or $n \geq 30$), (2) independent. Non-Normal + small n: Wilcoxon (Part III).

One-sided: $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ — one tail only.

Paired t-Test: Before vs After

If you have paired observations (X_i, Y_i), compute differences $D_i = X_i - Y_i$ and apply a one-sample t-test on D_i.

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_1 : \mu_D \neq 0$$

$$T = \frac{\bar{D}}{S_D/\sqrt{n}} \sim t_{n-1}$$

Example: Blood pressure before/after drug, $n = 8$ patients.

Before	148	142	136	134	138	140	132	144
After	140	138	132	130	135	137	130	140
D_i	8	4	4	4	3	3	2	4

$\bar{D} = 4.0$, $S_D = 1.77$, $T = \dfrac{4.0}{1.77/\sqrt{8}} = 6.39$, $p \approx 0.0004$.

$p \ll 0.05$: **reject** H_0. The drug works. Cohen's $d = \bar{D}/S_D = 4.0/1.77 = 2.26$ (very large).

Why Pairing Matters: Same Data, Two Stories

Run the BP example as if it were unpaired — watch the verdict change

The paired analysis on the previous slide gave $T = 6.39$, $p \approx 0.0004$.
What if we (wrongly) treat "before" and "after" as two independent groups?

Paired (correct)

Compute differences first: D_i.

$\bar{D} = 4.0, \;\; S_D = 1.77$

$$T = \frac{4.0}{1.77/\sqrt{8}} = \mathbf{6.39}$$

$p \approx 0.0004 \Rightarrow$ **reject**: drug works.

Unpaired (wrong here)

Treat as two groups. Compute group SDs:

$\bar{X}_B = 139.25, \; S_B = 5.34$
$\bar{X}_A = 135.25, \; S_A = 4.17$

$$T = \frac{4.0}{\sqrt{5.34^2/8 + 4.17^2/8}} = \mathbf{1.67}$$

$p \approx 0.12 \Rightarrow$ **fail to reject**.

Why so different? The patient-to-patient BP swing ($\sim$ 16 mmHg, hence $S_B \approx 5$) dominates the drug effect ($\sim$ 4 mmHg). Unpaired treats that swing as "noise". Paired *cancels it out*: each patient is their own baseline, leaving only the within-patient change ($S_D = 1.77$, roughly 3$\times$ smaller SD, $\sim 9\times$ less variance).

Two-Sample t-Test (Welch's): Comparing Two Groups

Two independent groups, $(\bar{X}_1, S_1, n_1)$ and $(\bar{X}_2, S_2, n_2)$

Building the statistic: the numerator is the observed difference of means $\bar{X}_1 - \bar{X}_2$. The denominator is its standard error.

Since the samples are *independent*: $\mathrm{Var}(\bar{X}_1 - \bar{X}_2) = \mathrm{Var}(\bar{X}_1) + \mathrm{Var}(\bar{X}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$. Plug in S_i^2 for σ_i^2:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \qquad \mathrm{df} = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

Why the awkward df? It's a weighted average of $n_1 - 1$ and $n_2 - 1$ that adjusts for the two unequal variance estimates. Usually a non-integer; `scipy` rounds and uses tables.

Default to Welch's. R and Python use it by default. The pooled t-test assumes $\sigma_1 = \sigma_2$; tiny extra power, not worth the assumption.

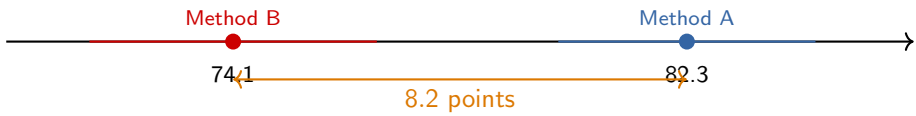
Two-Sample Worked Example: Two Teaching Methods

Method A ($n_1 = 10$): $\bar{X}_1 = 82.3$, $S_1 = 8.5$. Method B ($n_2 = 12$): $\bar{X}_2 = 74.1$, $S_2 = 10.2$.

Welch's t-test:

$$T = \frac{82.3 - 74.1}{\sqrt{8.5^2/10 + 10.2^2/12}} = \frac{8.2}{\sqrt{7.225 + 8.67}} = \frac{8.2}{3.99} = 2.06$$

df ≈ 19.8, $t_{19.8,\,0.025} \approx 2.09$, $p \approx 0.053$.



$|T| = 2.06 < 2.09$: **fail to reject** at $\alpha = 0.05$ (just barely!). Cohen's $d \approx 0.87$ (large effect) — the data *look* like a real gap, but n is too small to call it. Underpowered, not no-effect.

Part II: Proportions and Counts

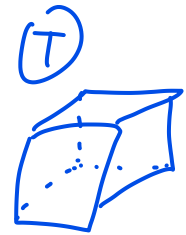
z-test for one proportion, χ^2 for tables of counts

Test for a Proportion

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p \neq p_0$$

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$

(Use p_0 in SE, not $\hat{p}$!)



Example: Is a coin fair? $n = 200$ flips, $k = 115$ heads, $\hat{p} = 0.575$.

$$Z = \frac{0.575 - 0.5}{\sqrt{0.5 \times 0.5/200}} = \frac{0.075}{0.0354} = 2.12, \qquad p = 2(1 - \Phi(2.12)) = 0.034$$

$p < 0.05$: **reject** H_0. Evidence the coin is biased.

Rule of thumb: use this test when $np_0 \geq 10$ and $n(1-p_0) \geq 10$.
For two proportions: $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ where $\hat{p} = \frac{k_1+k_2}{n_1+n_2}$
(pooled under $H_0 : p_1 = p_2$).

What About Many Categories?

The z-test handled 2 outcomes. What about 6 (a die) or a 2×3 table?

For each of k categories we have an **observed** count O_i and an **expected** count E_i (under H_0). We want one number that summarises "how off" the data are.

Naive idea: $\sum(O_i - E_i)$**.** But positives cancel negatives: a die that came up $25, 20, 15, 25, 20, 15$ would give 0, even though it's clearly not flat. Useless.

Better: square each deviation, then standardise by E_i:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Why divide by E_i? A count under H_0 is roughly Poisson with variance $\approx E_i$, so $(O_i - E_i)/\sqrt{E_i}$ is approximately a standardised Z. Squaring and summing k such Z's gives...

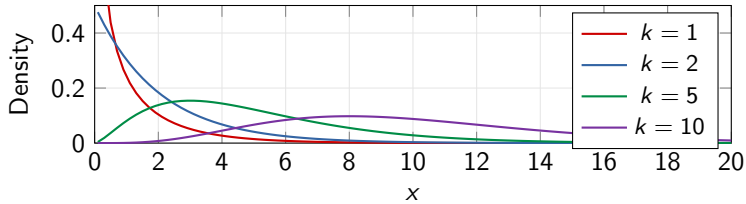
... a χ^2 **distribution.** (Defined on the next slide.)

The χ^2 Distribution

Sum of squared independent standard normals

Definition: if $Z_1, \ldots, Z_k$ are independent $N(0,1)$, then

$$Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi^2_k \quad (\text{"chi-squared with } k \text{ degrees of freedom"}).$$

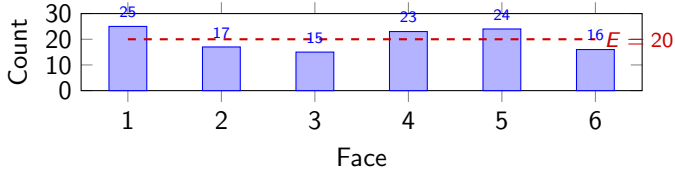


Key facts: mean $= k$, variance $= 2k$, always ≥ 0, skewed right (symmetric as k grows).

Where k comes from (in tests): roughly the number of "free squared deviations" in X^2. We'll see it shake out for each test.

Goodness-of-Fit: Is a Die Loaded?

Story: Someone hands you a die. You suspect it's loaded. You roll it $n = 120$ times.



Question: could a fair die have produced this much wobble around the expected 20 per face?

Under H_0 : $p_i = 1/6$, expected $E_i = np_i = 20$. Apply the recipe:

$$X^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \;\dot{\sim}\; \chi^2_{k-1} \quad (k - 1 = 5 \text{ df})$$

$X^2 = \frac{25+9+25+9+16+16}{20} = 5.0$. Critical: $\chi^2_{5,\,0.05} = 11.07$. $5.0 < 11.07$: **fail to reject**. The wobble is consistent with chance — no evidence the die is loaded.

Why $k-1$, not k? The six counts sum to n, so once five are known the sixth is fixed. Only 5 are free to deviate.

Test of Independence: Two Categorical Variables

Setup: Contingency table with r rows and c columns. H_0: variables are independent.

$$\text{Under } H_0\text{: } \quad E_{ij} \;=\; \frac{(\text{row } i \text{ total}) \times (\text{col } j \text{ total})}{n}$$

$$X^2 \;=\; \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \;\dot{\sim}\; \chi^2_{(r-1)(c-1)}$$

Test of Independence: Two Categorical Variables

Setup: Contingency table with r rows and c columns. H_0: variables are independent.

$$\text{Under } H_0: \quad E_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{col } j \text{ total})}{n}$$

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \;\dot{\sim}\; \chi^2_{(r-1)(c-1)}$$

Example: Smoking and lung cancer ($n = 200$).

	Cancer	No cancer	Total
Smoker	**40** (E=25)	**60** (E=75)	100
Non-smoker	**10** (E=25)	**90** (E=75)	100
Total	50	150	200

$X^2 = \frac{(40-25)^2}{25} + \frac{(60-75)^2}{75} + \frac{(10-25)^2}{25} + \frac{(90-75)^2}{75} = 9 + 3 + 9 + 3 = \mathbf{24.0}$

df $= (2-1)(2-1) = 1$. $\chi^2_{1,\,0.05} = 3.84$. $24.0 \gg 3.84$: **reject**. Strong association.

χ^2 **detects association, not causation** — could be a hidden confounder (see L14).

Part III: When Assumptions Fail

Nonparametric alternatives: no Normal required

Mann–Whitney U Test

Comparing two independent groups — nonparametric alternative to two-sample t

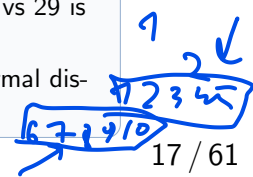
Use case: you A/B-test two checkout flows and record *completion time*. Most users finish in ~ 30 s, but a few hit the bathroom mid-flow and take 20 minutes. A t-test gets dragged around by those outliers; the means become useless. **Mann–Whitney** replaces every value with its *rank* — the slow outlier becomes "last," nothing more.

Idea: pool all values, rank them $1, 2, \ldots, n_1 + n_2$, sum each group's ranks.

Value (sorted)	12	24	36	41	50	6	7	8	9	10	sum
Rank	1	2	3	4	5	6	7	8	9	10	—
Group A		A	A		A		A		A		26
Group B	B			B		B		B		B	29

A's values were $\{3, 5, 7, 9, 2\}$, occupying rank positions $2, 3, 5, 7, 9$ (sum $= 26$). If groups are identical, A and B should split the ranks evenly — here 26 vs 29 is close, no evidence of a difference.

When to use: skewed data, outliers, ordinal data, small n with non-Normal distributions.



Wilcoxon Signed-Rank Test (1/2): The Idea

Paired nonparametric alternative to the paired t-test

Setup: same as paired t — n paired observations, compute differences D_i. Difference: instead of using $\bar{D}$ and S_D, we use *ranks* of $|D_i|$.

1. Compute differences D_i (could be positive, negative, or zero — drop zeros).

2. Take absolute values $|D_i|$ and *rank* them ($1 =$ smallest absolute change, $n =$ largest).

3. Sum the ranks where the original D_i was positive: W^+. Same for negative: W^-.

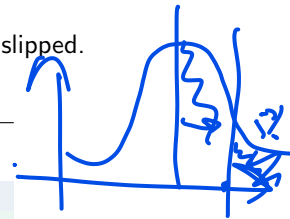
4. The test statistic is $W = \min(W^+, W^-)$. Small $W \Rightarrow$ one sign dominates $\Rightarrow$ reject H_0.

Intuition: if the treatment has no effect, half the differences should be positive, half negative, with similar magnitudes. So the rank sums should split roughly evenly.

Wilcoxon Signed-Rank Test (2/2): Worked Example

Data: 7 students, scores before vs after a workshop. Mostly improved, two slipped.

Subject	1	2	3	4	5	6	7
D_i (after−before)	$+5$	$+3$	-1	$+7$	$+4$	-2	$+6$
$\lvert D_i \rvert$	5	3	1	7	4	2	6
Rank of $\lvert D_i \rvert$	5	3	1	7	4	2	6
Sign	$+$	$+$	$-$	$+$	$+$	$-$	$+$



Compute: $W^+ = 5+3+7+4+6 = 25$, $W^- = 1+2 = 3$, $W = \min(25, 3) = 3$.

Decide: at $n = 7$, two-sided $\alpha = 0.05$ rejects when $W \le 2$. Here $W = 3$: **fail to reject** ($p \approx 0.063$ via normal approximation). Suggestive but not conclusive.

When to use: paired data that's not Normal; ordinal scales (e.g. Likert 1–5); small samples with outliers.

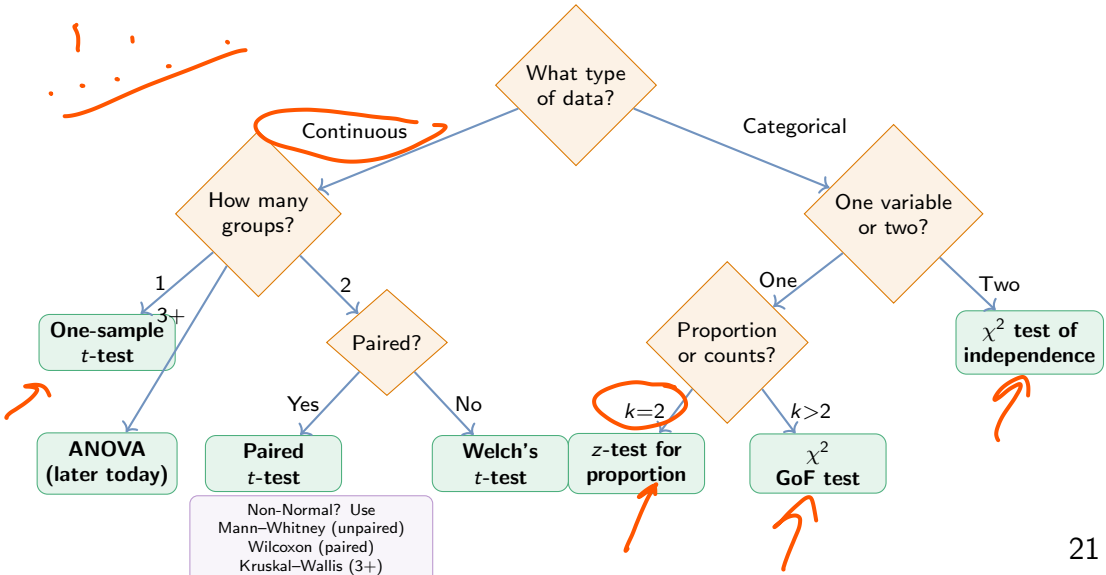
Trade-off: $\sim 5\%$ less power than paired t when data *is* Normal — a small price for safety.

Three or more groups, non-Normal: Kruskal–Wallis (covered later).

Part IV: Putting It Together

Decision flowchart and Python

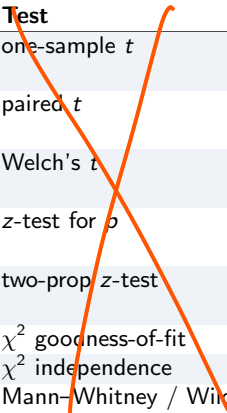
Which Test Do I Use?



Scenario Cheat Sheet

Real-world question → test

What you want to do	Test
"Are these cups really 350 mL on average?" (one mean vs claim)	one-sample t
"Did the drug change blood pressure?" (same patients before/after)	paired t
"Is method A better than method B?" (two independent groups)	Welch's t
"Is this coin / button / outcome at the claimed rate?" (one proportion)	z-test for p
"Is page A's conversion different from page B's?" (two proportions)	two-prop z-test
"Is this die fair?" (one categorical variable, k buckets)	χ^2 goodness-of-fit
"Are smoking and cancer related?" (two categorical variables)	χ^2 independence
"Same as above but with extreme outliers / ordinal data"	Mann–Whitney / Wilcoxon
"Are 3+ teaching methods different?" (3+ group means)	ANOVA + Tukey (later)
"Did effect 1 happen *or* effect 2 happen *or* . . . " (many tests)	multiple-test correction (L9)



First question to ask: "what kind of data do I have?" — then "one group, two, or many?". The flowchart and the table both lead there.

Classical Tests in Python

```python
from scipy import stats

# One-sample t-test: H0: mu = 350
stats.ttest_1samp(data, popmean=350)

# Paired t-test
stats.ttest_rel(before, after)

# Two-sample Welch's t-test (default)
stats.ttest_ind(group_a, group_b, equal_var=False)

# Chi-squared test of independence
stats.chi2_contingency([[40,60],[10,90]])

# Nonparametric alternatives
stats.mannwhitneyu(group_a, group_b)  # unpaired
stats.wilcoxon(before - after)        # paired
```

Each function returns `(statistic, p_value)`. Always check which alternative hypothesis is default.

Part V: The Unifying Picture

Wait — those tests were all the same test in disguise.

The Likelihood Ratio framework

Notice the Pattern

Look back at every test we just saw. They share the same shape:

H_0 *pins down* a parameter (or a relationship).
H_1 *lets it be free.*

The test asks: **does pinning it down hurt the fit to the data?**

One-sample t: H_0 pins $\mu = \mu_0$. H_1: μ is free. Test: does μ_0 fit as well as $\bar{X}$?

Two-sample t: H_0 pins one shared μ. H_1: two free μ's. Test: is one mean as good as two?

z-test for proportion: H_0 pins $p = p_0$. H_1: p is free. Test: does p_0 fit the count?

χ^2 GoF: H_0 pins all $p_i = p_i^0$. H_1: p_i free. Test: does the fixed distribution fit the counts?

χ^2 independence: H_0 pins joint = product of marginals. H_1: joint is free.

Restricted vs Full Models

Putting names on the pattern

The pattern has names. The "pinned" model is **restricted**; the "free" model is **full**.

H_0: Restricted

Some parameters *fixed* by the hypothesis.

Fewer **free parameters** (= unknowns to estimate)

$H_0 \subset H_1$

H_1: Full

All parameters *free*.

More free parameters (more unknowns to estimate from data)

Coin fair? $p = 0.5$ (0 free) vs $p \in [0,1]$ (1 free) $\Rightarrow k = 1$ free param difference

Same mean? $\mu_1 = \mu_2$ (1 shared) vs μ_1, μ_2 separate (2) $\Rightarrow k = 1$

Independent? joint = product of marginals vs arbitrary joint $\Rightarrow k = (r-1)(c-1)$

The number k is the count of constraints H_0 imposes. It will determine the test's degrees of freedom.

The Likelihood Ratio Statistic Λ (1/2)

Definition and interpretation

Recipe: fit both models using MLE (Lecture 5), then compare how well they fit.

$$\Lambda = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} = \frac{\text{best likelihood under } H_0}{\text{best likelihood overall}}$$

Numerator: best fit *with* the H_0 constraint. **Denominator:** best fit *without* any constraint.

Denominator $\geq$ numerator (more freedom $\Rightarrow$ better fit), so $0 \leq \Lambda \leq 1$.

Λ close to 1

H_0 fits almost as well as H_1.
Restricting barely hurts.
$\Rightarrow$ fail to reject.

Λ close to 0

H_0 fits much worse than H_1.
The restriction really hurts.
$\Rightarrow$ reject H_0.

The Likelihood Ratio Statistic Λ (2/2)

Why we transform to $-2\log\Lambda$

Λ is awkward to work with directly. We use $-2\log\Lambda$ instead. Three reasons:

1. Likelihoods are tiny. For the coin: $L(0.5) = 0.5^{100} \approx 10^{-31}$. Ratios of tiny numbers are unstable. log converts product to sum, and $\log(10^{-31}) = -71$ is a manageable number.

2. "Bigger = reject" is more natural. Λ is small when we reject (confusing). The minus sign flips it: $\Lambda \to 0$ becomes $-\log\Lambda \to \infty$. Now "larger statistic = stronger evidence," like every other test.

3. The factor of 2 matches χ^2. The constant is chosen so that $-2\log\Lambda$ converges to a clean, tabulated distribution: the χ^2_k. That's Wilks' theorem (next slide).

Bottom line: $-2\log\Lambda \geq 0$. Reject when it's large.

Worked LRT: The Coin, End to End

Data: 100 flips, 62 heads. $H_0 : p = 0.5$ vs $H_1 : p \in [0, 1]$ ($k = 1$ free-param diff).

1. Likelihoods. $L(0.5) = 0.5^{100} \approx 7.9 \times 10^{-31}$. $L(\hat{p}) = L(0.62) = 0.62^{62} \cdot 0.38^{38} \approx 1.45 \times 10^{-29}$.

2. Ratio. $\Lambda = L(0.5)/L(0.62) \approx 0.055$. $-2 \log \Lambda \approx 5.82$.

3. Reference distribution. Under H_0, $-2 \log \Lambda \dot{\sim} \chi^2_1$ (Wilks, $k = 1$ constraint).

4. Decide. Critical value $\chi^2_{1,\,0.05} = 3.84$. $5.82 > 3.84 \Rightarrow$ **reject** H_0. $p = P(\chi^2_1 > 5.82) \approx 0.016$.

Sanity check (same data): z-test gives $Z = 2.40$, $Z^2 = 5.76$, $p \approx 0.016$. Same answer — the z-test **is** this LRT in disguise.

Wilks' Theorem: Why It All Works

The reason every test today produced a valid p-value is one theorem. Under H_0, as $n \to \infty$:

$$-2 \log \Lambda \quad \xrightarrow{d} \quad \chi^2_k$$

where k = (free params in H_1) − (free params in H_0) = constraints.

The tests we already saw, retroactively explained:

One-sample t
$k = 1$
$T^2 \dot{\sim} \chi^2_1$

Two-sample / Welch
$k = 1$
$T^2 \dot{\sim} \chi^2_1$

χ^2 independence
$k = (r-1)(c-1)$
already χ^2

The takeaway: every test today is an LRT on nested models, and Wilks tells you what distribution to compare against. **Caveat:** asymptotic. For small n, prefer exact tests (Fisher, binomial).

Recap: Classical Tests & LRT

One-sample t: $T = (\bar{X} - \mu_0)/(S/\sqrt{n}) \sim t_{n-1}$. The workhorse.

Paired t: differences first, then one-sample t. More powerful than unpaired.

Welch's t: two independent means. No equal-variance assumption. Default choice.

Proportions: z-test with SE $= \sqrt{p_0(1-p_0)/n}$. Use when $np \geq 10$.

χ^2 tests: GoF, independence. Statistic is sum of squared standardized errors.

Nonparametric: Mann–Whitney (unpaired), Wilcoxon (paired) when normality fails.

LRT: all of the above are special cases. $-2\log\Lambda \dot{\sim} \chi^2_k$ (Wilks).

Effect size matters: a small p on a tiny effect is real but unimportant.

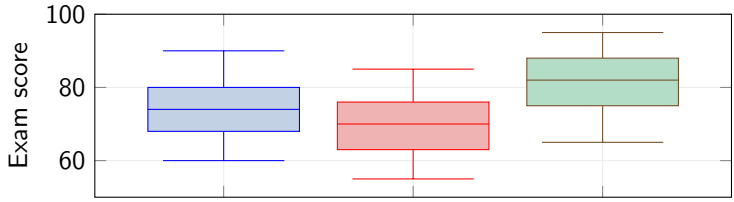
Up next: What if we have **three or more** groups? And how does industry actually run experiments?

Part VI: From 2 to k Groups

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?



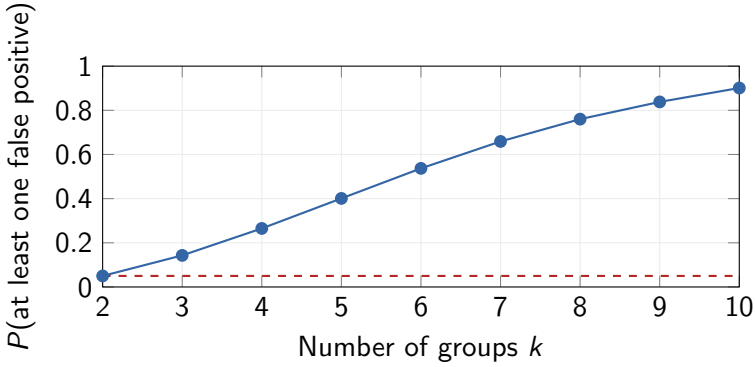
Naïve approach: Run three pairwise t-tests: A vs B, A vs C, B vs C.

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

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

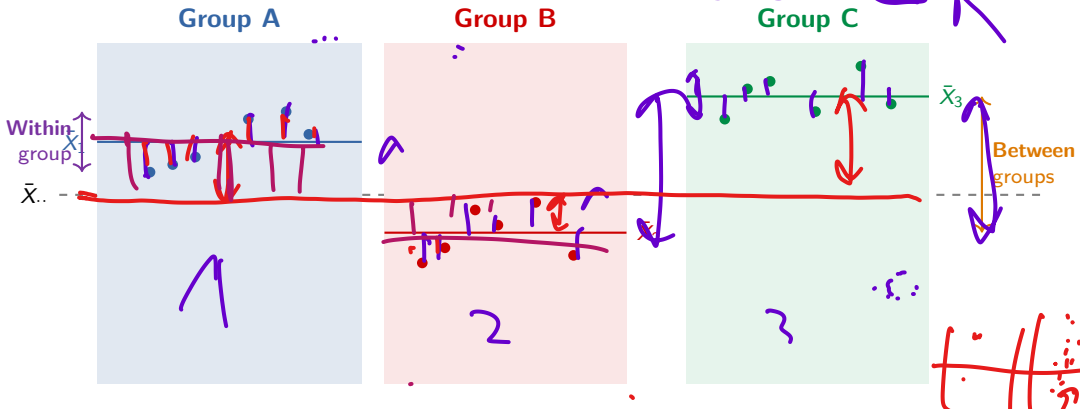
With 10 groups: $\binom{10}{2} = 45$ tests $\Rightarrow 1 - 0.95^{45} = 0.90$. Almost certain to find "something."

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).

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

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

SS_{Total}
Total spread of all obs 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
$\text{df} = N - k$

Concrete example (3 methods, $N = 30$, used on the next slide):
$\text{SS}_{\text{Total}} = 1800$, $\text{SS}_{\text{Between}} = 720$ (df $= 2$), $\text{SS}_{\text{Within}} = 1080$ (df $= 27$).

The F-Statistic

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

under H_0 : $\mu_1 = \mu_2 = \cdots = \mu_k$

$F \approx 1$: No evidence

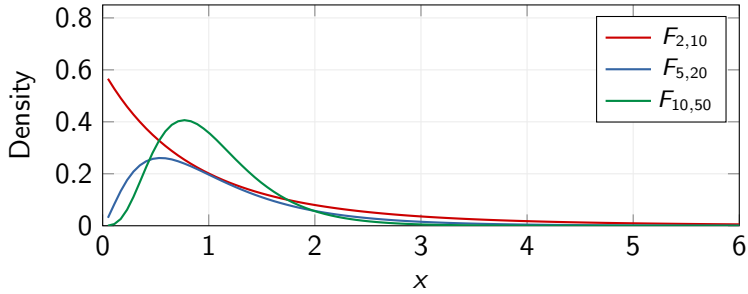
Between $\approx$ within variance.
Group means are similar.
$\Rightarrow$ Fail to reject H_0.

$F \gg 1$: Evidence

Between $\gg$ within 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: the three teaching methods, $n_i = 10$ each ($N = 30$), with the SS values from two slides ago.

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): $\sigma_1^2 = \cdots = \sigma_k^2$.
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).

Post-Hoc: Tukey's HSD vs Bonferroni

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

ANOVA told us *some* group means differ. Now we ask: *which pairs?* Two standard methods:

Tukey's HSD

Honestly **S**ignificant **D**ifference.

Designed for *all pairs* of k groups. Controls family-wise error rate at α. Uses the *studentized range* distribution q (designed for max-of-k comparisons).

Best for: all pairs, equal n_i.

Bonferroni

Run pairwise t-tests, but compare each p-value to α/m instead of α ($m = \#$ tests).

More conservative than Tukey when comparing all pairs (Tukey is sharper here).

Works for *any* set of comparisons, not just pairs.

Best for: few planned comparisons.

Tukey's HSD: Worked Example

Three teaching methods (means $74, 70, 82$, $\text{MS}_\text{W} = 40$, $n_i = 10$, $\text{df} = 27$)

Step 1. Compute the HSD threshold. Look up $q_{\alpha,\,k,\,\text{df}}$ for $\alpha = 0.05$, $k = 3$ groups, $\text{df} = 27$:

$$q_{0.05,\,3,\,27} \approx 3.51 \quad \text{(studentized range table)}$$

Step 2. Convert to a difference-of-means threshold:

$$\text{HSD} = q \cdot \sqrt{\frac{\text{MS}_\text{W}}{n}} = 3.51 \cdot \sqrt{\frac{40}{10}} = 3.51 \cdot 2 = \mathbf{7.02}$$

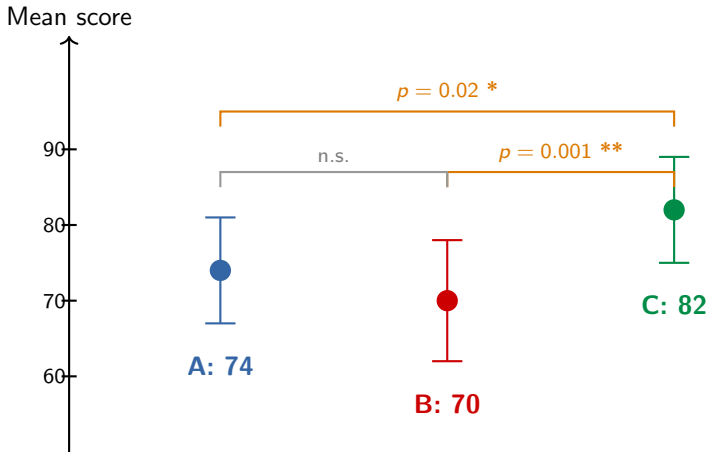
Step 3. Any pair of group means whose absolute difference exceeds 7.02 is significantly different.

Pair	$\lvert\bar{X}_i - \bar{X}_j\rvert$	vs HSD $= 7.02$	**Verdict**
C (82) vs A (74)	8	> 7.02	**significant** ($p \approx 0.02$)
C (82) vs B (70)	12	> 7.02	**significant** ($p \approx 0.001$)
A (74) vs B (70)	4	< 7.02	not significant ($p \approx 0.38$)

Conclusion: Method C beats both A and B. A and B are statistically indistinguishable.

In Python: `from statsmodels.stats.multicomp import pairwise_tukeyhsd; pairwise_tukeyhsd(scores, group_labels)`.

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.

Effect Size: η^2 (Eta-Squared)

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

Proportion of total variance explained by group membership.

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

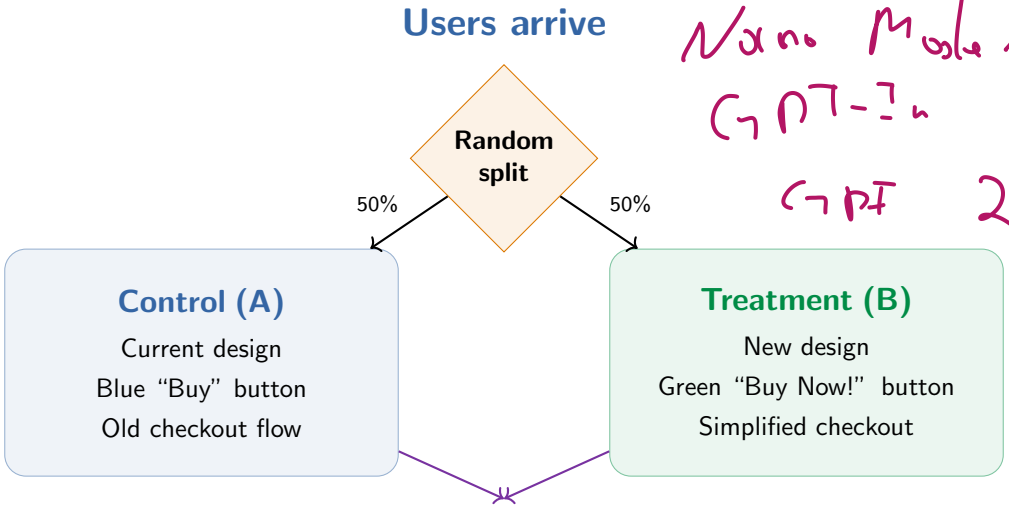
Teaching example: $\eta^2 = 720/1800 = 0.40$. **Very large** — 40% of score variation is explained by teaching method.

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.

Part VII: 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.

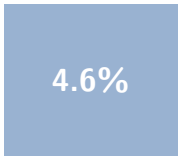
A/B Test Example: Button Color

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

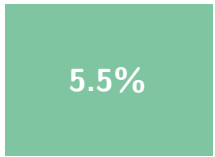
Pooled proportion: $\hat{p} = (230 + 275)/(5000 + 5000) = 0.0505$.

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

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



Control (Blue)



Treatment (Green)

+19.6% relative lift
$p = 0.040$

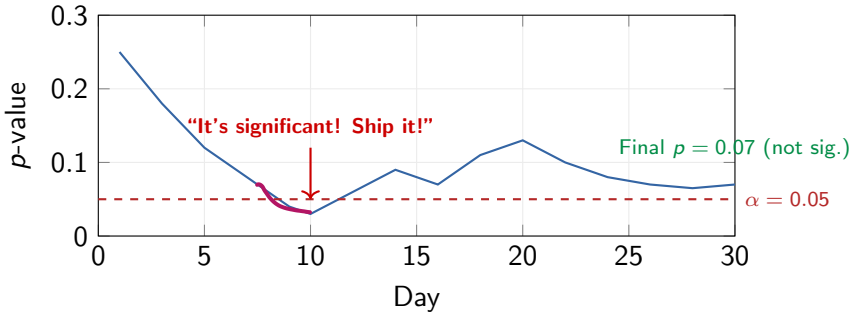
Decision: Statistically significant. +19.6% relative lift is practically meaningful. Ship the green button.

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 VIII: A/B Testing Pitfalls

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

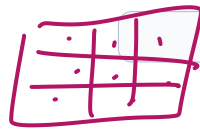
Pitfall 1: Peeking (The Most Common Mistake)



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.

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).

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 categorical variables, association: χ^2 test of independence.

Summary: Lectures 10–11

t-tests: one-sample, paired, Welch's two-sample. Workhorses of inference.

χ^2 tests: GoF and independence. Statistic = sum of squared standardized errors.

Nonparametric: Mann–Whitney, Wilcoxon when normality fails.

LRT: all of the above are special cases. $-2\log\Lambda \dot{\sim} \chi^2_k$ (Wilks).

ANOVA: $F = \mathrm{MS_B}/\mathrm{MS_W}$. Single test for k group means.

Post-hoc: Tukey HSD or Bonferroni once ANOVA rejects.

Effect size: η^2, Cohen's d. "Real" $\neq$ "meaningful".

A/B testing: same machinery, plus engineering: randomization, sample size, deployment.

Pitfalls: peeking, too many metrics, novelty, Simpson's, interference.

Practical: Classical Tests

1. **One-sample t-test:**
   - Coffee shop data (n=25, $\bar{X}$=345, S=10, μ_0=350)
   - Compute T, find p from t-table, verify with `scipy.stats.ttest_1samp`
2. **Paired t-test:**
   - Blood pressure before/after. Compute by hand and with `ttest_rel`
3. **Welch's t-test:**
   - Teaching method data. Use `ttest_ind(equal_var=False)`
   - Compare p with the permutation test from L9
4. **χ^2 independence test:**
   - Smoking/cancer 2×2 table. Compute by hand
   - Verify with `scipy.stats.chi2_contingency`

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. A manufacturer claims mean weight is 500 g. Sample $n = 16$: $\bar{X} = 495$, $S = 8$. Test at $\alpha = 0.05$. Compute the 95% CI. Does it contain 500?

2. 12 runners' times (s) before and after training:
   Before: 58.2, 57.1, 60.3, 55.8, 59.6, 62.1, 56.4, 61.2, 58.7, 57.5, 60.8, 59.3.
   After: 56.1, 55.8, 58.7, 54.2, 57.3, 59.8, 55.1, 59.0, 56.4, 56.0, 58.2, 57.1.
   (a) Paired t-test. (b) Unpaired t-test on the same data. Why does the p-value change?

3. Survey of 300 people on preferred transport: bus 90, metro 120, car 60, bicycle 30. Test whether preferences are equally distributed (χ^2 goodness-of-fit).

4. Study method (group/solo) vs exam result (pass/fail): Group+pass 70, Group+fail 30, Solo+pass 55, Solo+fail 45. Test independence at $\alpha = 0.05$.

5. Startup salaries (thousands): Team A: 120, 45, 38, 200, 52. Team B: 55, 62, 48, 70, 58.
   (a) Why is a t-test problematic here? (b) Use Mann–Whitney U (`scipy.stats.mannwhitneyu`).

6. 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 $\alpha = 0.05$ ($F_{2,42,0.05} = 3.22$).

Recommended Resources

Interactive: Seeing Theory — Brown University

`seeing-theory.brown.edu/frequentist-inference` — t-tests, χ^2, regression with sliders.

Video: StatQuest

"t-Tests Explained", "Chi-Square Tests", "ANOVA", "A/B Testing". Clear and concise.

Reading: Wasserman, "All of Statistics" — Ch. 10–11

Hypothesis testing, LRT framework, t-tests, χ^2 tests. Concise and rigorous.

Reading: Kohavi, Tang & Xu — "Trustworthy Online Controlled Experiments"

The definitive book on A/B testing from Microsoft researchers.

Blog: Evan Miller — "How Not to Run an A/B Test"

Classic post on the peeking problem. Required reading for anyone running A/B tests.

Python: `scipy.stats` + `statsmodels`

`ttest_*`, `chi2_contingency`, `mannwhitneyu`, `wilcoxon`, `f_oneway`, `kruskal`, `levene`, `pairwise_tukeyhsd`, `proportions_ztest`.

Questions?

Next: Lecture 12 — Regression inference