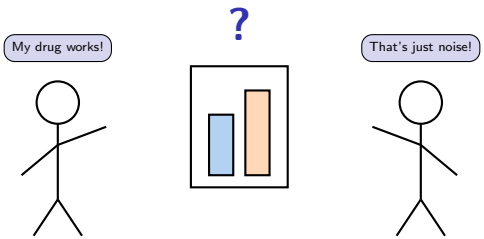
Hypothesis Testing: A Visual Guide

A Formula-Free Introduction

Statistical Data Analysis Course

March 13, 2026

Can We Settle This Argument?



When two scientists disagree about data, who wins?

Statistics provides the referee

What Will You Learn Today?

1. **Decide if a difference is real or just luck**
2. **Pick the right test for 2 vs. 3+ groups**
3. **Avoid the most common testing mistakes**

No formulas needed — just concepts and visual intuition

Why Can't We Just Eyeball It?

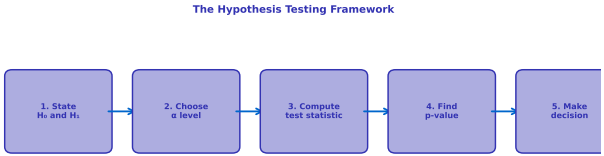
Story: The Tempting A/B Test

An A/B test shows **52% vs. 48%** conversion rate. Is that a real difference?

- Looks different? Maybe — but the sample was only 200 people
- Could easily be coin-flip noise: flip 100 coins twice and you'll rarely get exactly 50/50
- The naive approach: "Looks different, ship it!" — this leads to costly mistakes

Human intuition fails when differences are small and data is noisy

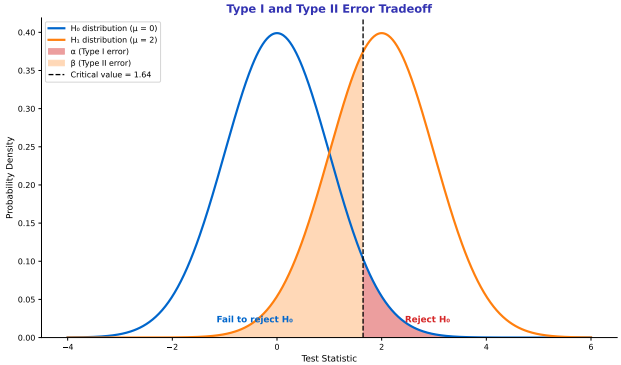
How Do We Structure the Question?



- State what you expect (H_1) and what "no effect" looks like (H_0)
- Collect data and measure how surprising it is
- Make a decision based on the evidence

Every statistical test follows these five steps

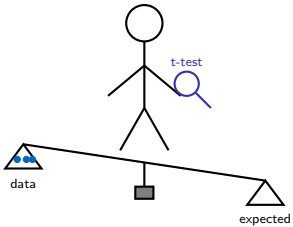
What Could Go Wrong?



- **False alarm** (Type I) — you see an effect that isn't there
- **Missed finding** (Type II) — you miss a real effect
- The trade-off: reducing one increases the other

Setting alpha to 0.05 means accepting a 5 percent false alarm rate

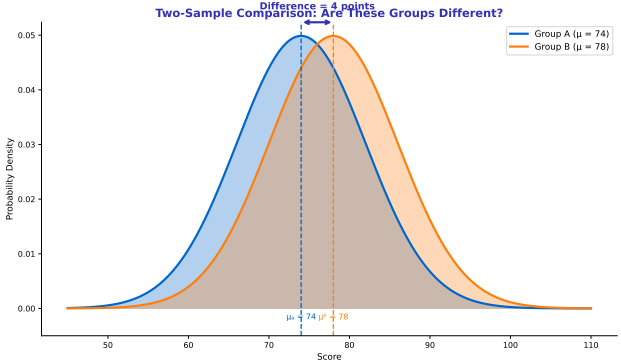
How Does a T-Test Weigh Evidence?



- A t-test compares your data to what you'd expect if nothing happened
- The bigger the tilt, the stronger the evidence
- If the tilt is extreme enough, we reject "no effect"

The t-test is a precision instrument for comparing two things

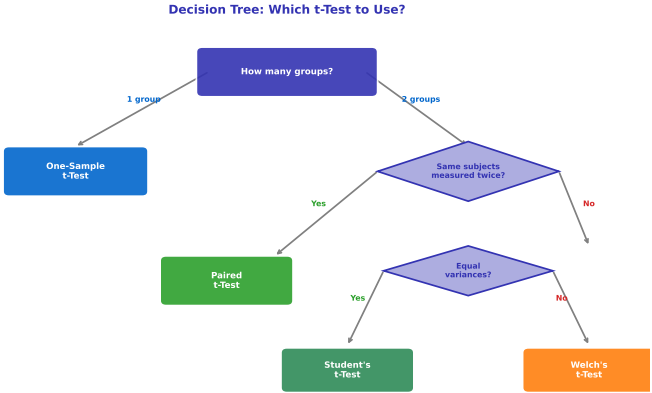
Do These Two Groups Really Differ?



- Large overlap means hard to distinguish groups
- Small overlap means a clear difference
- The t-test quantifies exactly how much overlap matters

Visual overlap is intuitive — the t-test makes it rigorous

Which T-Test Should You Pick?



- One group vs. a known value? **One-sample**
- Same subjects measured twice? **Paired**
- Two different groups? **Two-sample**

The decision depends on how your data was collected, not what answer you want

What Does the Code Look Like?

R

```r
# One-sample
t.test(data, mu = 100)

# Paired
t.test(before, after,
       paired = TRUE)

# Two-sample
t.test(group_a, group_b)
```

Python

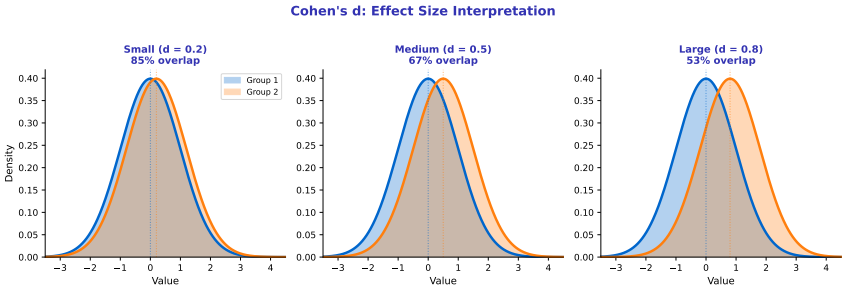
```python
# One-sample
ttest_1samp(data, 100)

# Paired
ttest_rel(before, after)

# Two-sample
ttest_ind(group_a, group_b)
```

Three lines of code per test — the computer handles the math

Is a Significant Difference Actually Big?



- A tiny p-value with a tiny effect = statistically real but practically useless
- Cohen's d measures how **big** the difference is
- Small (0.2), Medium (0.5), Large (0.8) — always report alongside p

Statistical significance is not the same as practical importance

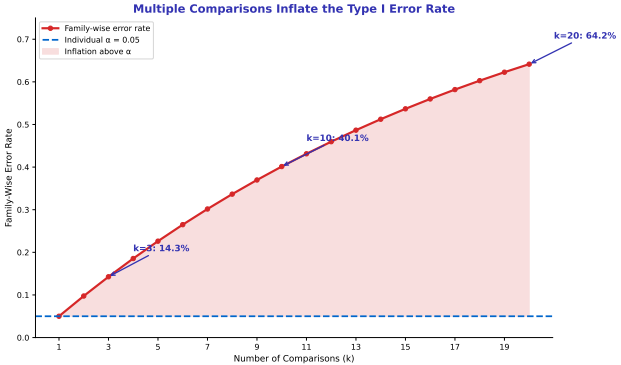
When Do T-Tests Break Down?

- T-tests only handle **2 groups** at a time
- Running multiple t-tests **inflates your false alarm rate**
- With 10 groups you'd need 45 t-tests — and a **40% chance** of at least one false alarm!

$\Longrightarrow$ **We need a better tool...**

The more tests you run, the more likely you are to fool yourself

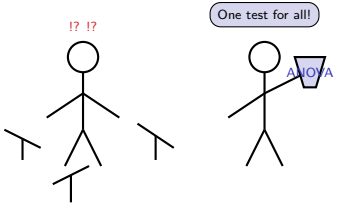
How Fast Do False Alarms Pile Up?



- With 3 groups and 3 tests, false alarm rate jumps to **14%**
- With 10 groups and 45 tests, it reaches **40%**
- This is why we need ANOVA — one test for all groups

Multiple testing is the silent killer of scientific credibility

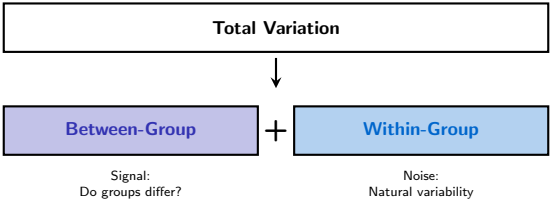
Can One Test Handle All Groups?



- Instead of comparing every pair separately, ANOVA asks one question: "Do **any** of these groups differ?"
- If the answer is yes, **then** we investigate which pairs

ANOVA replaces many fragile tests with one robust test

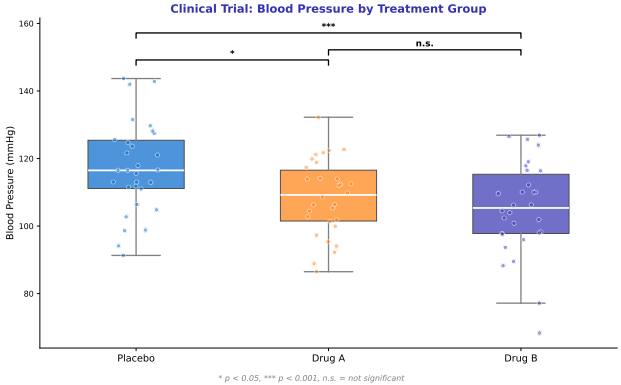
What Does ANOVA Actually Compare?



- If between-group variation is large relative to within-group, groups differ
- If within-group noise dominates, groups look the same
- The F-statistic is simply the ratio: between divided by within

ANOVA decomposes total variation to find the signal hidden in the noise

See ANOVA in Action!



- Three treatments tested on blood pressure
- Boxplots show clear separation between groups
- ANOVA confirms: at least one treatment works differently

ANOVA tells you something differs — post-hoc tests tell you what

Which Pairs Are Actually Different?

The Post-Hoc Problem

ANOVA says "yes, groups differ" — but **which** ones?

- **Tukey HSD** compares every pair while controlling false alarms
- It adjusts p-values so the overall error rate stays at 5%
- Code: `TukeyHSD(aov(score ~ group, data = df))`

Tukey HSD is like running protected t-tests that won't inflate your error rate

Watch Out for These ANOVA Gotchas!

Assumption	What to Check
Normality	Q-Q plot or Shapiro-Wilk test
Equal variances	Levene's test
Independence	Study design (not a statistical test)

- If assumptions fail, use **Kruskal-Wallis** (non-parametric alternative)

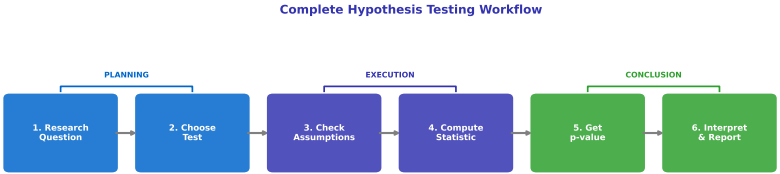
Assumptions matter — violating them can invalidate your conclusion

T-Test or ANOVA — How Do You Decide?

	T-Test	**ANOVA**
Groups	2	3 or more
Output	t-statistic	F-statistic
Follow-up	None needed	Post-hoc tests
Non-parametric	Mann-Whitney	Kruskal-Wallis

When in doubt, start with ANOVA — it works for 2 groups too

What Is the Complete Testing Workflow?



- This flowchart covers every decision from research question to final report
- Print it, pin it above your desk, and follow it every time

A systematic workflow prevents the most common statistical mistakes

What About Non-Normal Data?

- Real data is often skewed (financial returns, reaction times, counts)
- Non-parametric tests make fewer assumptions about distribution shape
- **Mann-Whitney** replaces the t-test; **Kruskal-Wallis** replaces ANOVA

Non-parametric tests trade some power for robustness to non-normality

How Do You Report Results?

Template Sentence

"A one-way ANOVA revealed a significant effect of treatment, $F(2, 87) = 4.32$, $p = .016$, eta-squared $= .09$ (medium effect)."

Key reporting checklist:

- Name the test and its statistics
- Report the *p*-value **and** the effect size
- State the practical conclusion in plain English

A well-reported result tells the reader what happened, how strong it was, and why it matters

Remember the Three Big Takeaways!

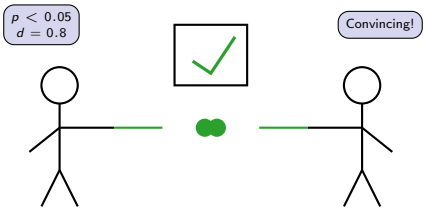
1. A p-value is **not** the probability your hypothesis is true — it measures how surprising your data would be if nothing happened

2. Always report **effect size** alongside p-value — significance without magnitude is meaningless

3. Use **ANOVA** for 3+ groups — never chain t-tests

These three rules will prevent 90 percent of common statistical mistakes

Can We Settle Arguments Now?



The argument from Slide 2 is settled — with evidence, not opinion.

Statistics does not prove truth — it quantifies evidence against reasonable doubt

Where Do You Go From Here?

- For the full technical details, see the main lecture (`hypothesis_testing.tex`)
- Practice with **11 real datasets** in the `datasets/` folder
- Test yourself with the **20-question quiz**

"Statistics: the science of being precisely uncertain."

The best way to learn statistics is to practice on real data