Hypothesis Testing

Statistical Methods for Data Analysis

Statistical Data Analysis Course

March 14, 2026

Lecture Outline

Learning Objectives

After completing this lesson, you will be able to:

1. Formulate and test hypotheses using t-tests
2. Compare group means with ANOVA
3. Apply post-hoc corrections for multiple comparisons
4. Calculate power and effect sizes for study planning

These four skills form the foundation of statistical hypothesis testing

Hypothesis Testing Fundamentals

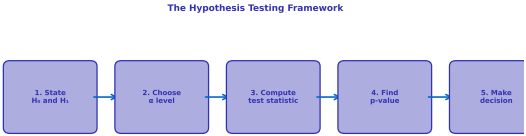
Can We Trust This Result?

The Drug Trial Question

A pharmaceutical company claims their new drug reduces blood pressure.

- 100 patients enrolled in a trial
- Blood pressure drops by 5 mmHg on average
- But patients vary naturally day to day
- Could this drop be pure chance?

Feelings are not evidence—we need a **formal framework** to separate real effects from random noise.



Hypothesis testing provides a principled way to decide whether observed differences are real or due to chance

Worked Example: Drug Trial Intuition

The Setup

- 100 patients split into two groups ($n = 50$ each)
- Placebo group mean: $\bar{x}_P = 130\,\text{mmHg}$
- Drug group mean: $\bar{x}_D = 125\,\text{mmHg}$
- Both groups: $s = 15\,\text{mmHg}$

Building Intuition

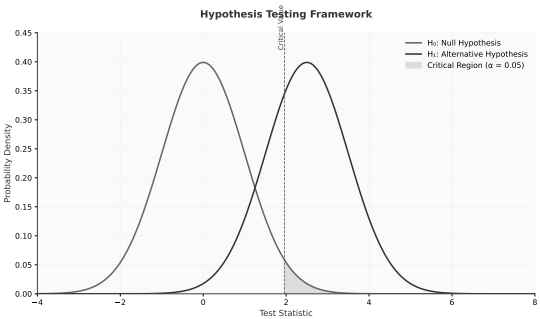
Observed difference: $130 - 125 = 5\,\text{mmHg}$.
Is 5 large relative to the variability?

$$t = \frac{5}{15 / \sqrt{50}} = \frac{5}{2.12} = 2.36$$

$p = 0.02$: only a 2% chance of seeing a difference this large if the drug had *no* effect.

The t-statistic standardises the difference by the noise level—larger t means stronger evidence

The Five-Step Framework



Every Hypothesis Test

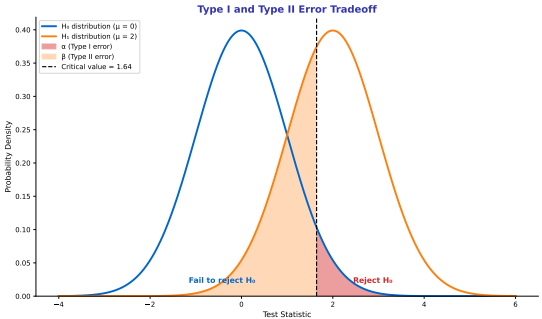
1. **State hypotheses**: H_0 (no effect) vs H_1 (effect exists)
2. **Choose** significance level α (typically 0.05)
3. **Compute** the test statistic from your data
4. **Find** the p-value (or compare to critical value)

Step 5: Decision

If $p < \alpha$: reject H_0 (evidence for an effect).
If $p \geq \alpha$: fail to reject H_0 (insufficient evidence).

This five-step framework applies to every hypothesis test in this course

Type I and Type II Errors



Two Ways to Be Wrong

- **Type I** (α): Reject a true H_0—false positive
- **Type II** (β): Fail to reject a false H_0—false negative
- α is set by the researcher (typically 0.05)
- **Power** $= 1 - \beta$: probability of catching a real effect

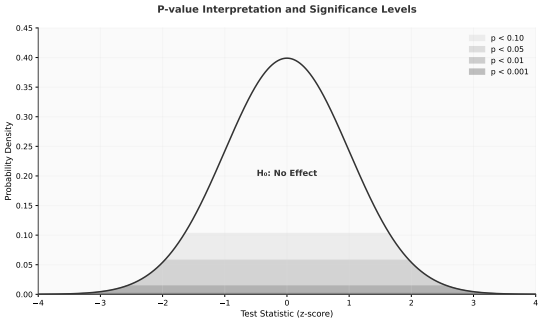
Lowering α reduces false positives but *increases* false negatives—there is always a trade-off.

Minimising Type I error increases Type II error—balance is key

What p-values Really Mean

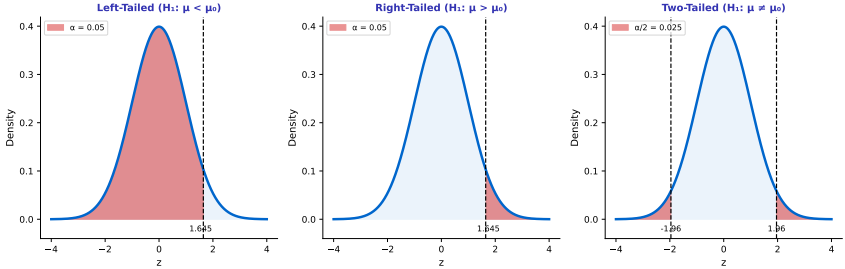
Correct Interpretation

- The p-value is the probability of observing data at least as extreme as yours, *assuming H_0 is true*
- A small p-value means the data are *unlikely* under H_0
- It is **not** the probability that H_0 is false



A small p-value means the data are unlikely under H0—it does not tell you the probability H0 is false

One-Tailed vs Two-Tailed Tests



- **Left-tailed**: H_1: $\mu < \mu_0$ (test for decrease only)
- **Right-tailed**: H_1: $\mu > \mu_0$ (test for increase only)
- **Two-tailed**: H_1: $\mu \neq \mu_0$ (test for any difference—the safe default)

Use two-tailed tests unless you have a strong directional hypothesis stated before data collection

One-Sample t-Test

The Coffee Shop Problem

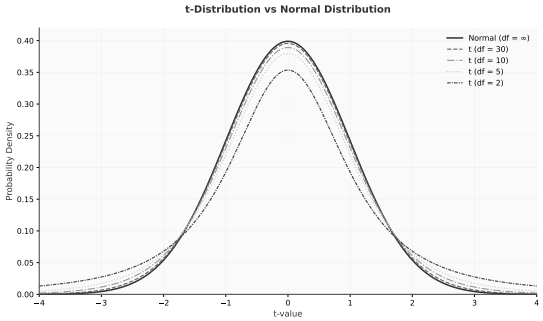
Scenario

A coffee shop advertises 350 ml per cup. You suspect they underfill.

- You measure $n = 10$ cups
- Sample mean: $\bar{x} = 342$ ml
- Sample SD: $s = 8$ ml
- Is this evidence of underfilling?

Hypotheses

H_0: $\mu = 350$ vs H_1: $\mu < 350$



The t-distribution accounts for the extra uncertainty when we estimate σ from a small sample.

Small samples require the t-distribution because we estimate the population SD from the data

The t-Statistic: Measuring Surprise

Formula

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

- $\bar{x}$: sample mean
- μ_0: hypothesised population mean
- s: sample standard deviation
- n: sample size; $df = n - 1$

Coffee Cup Calculation

$$t = \frac{342 - 350}{8 / \sqrt{10}} = \frac{-8}{2.53} = -3.16$$

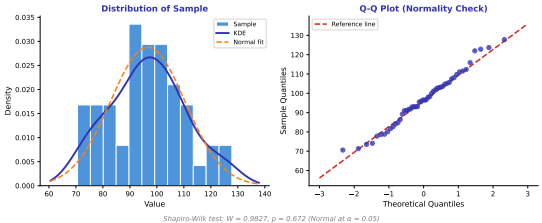
- $df = 9$
- $p = 0.012$ (one-tailed)
- The sample mean is 3.16 standard errors below 350
- At $\alpha = 0.05$: reject H_0—evidence of underfilling

The t-statistic measures how many standard errors the sample mean lies from the hypothesised value

Assumptions: When Can We Use This?

Three Key Assumptions

1. **Random sample**: observations drawn independently from the population
2. **Continuous data**: the outcome variable is measured on a continuous scale
3. **Approximate normality**: population is roughly normal, or $n \geq 30$ (CLT)



Check normality with Q-Q plots and the Shapiro-Wilk test. For $n \geq 30$, mild departures are tolerable.

With n ¿= 30 the Central Limit Theorem provides robustness to non-normality

Worked Example: Coffee Cups in R

```r
# Measured cup volumes (ml)
cups <- c(338, 345, 340, 342,
          348, 335, 344, 339,
          346, 341)

# One-sample t-test
# H0: mu = 350, H1: mu < 350
t.test(cups, mu = 350,
       alternative = "less")

# Output:
# t = -3.16, df = 9
# p-value = 0.006
# 95% CI: (-Inf, 345.7)
# Sample mean: 341.8
```

Interpretation

- $t = -3.16$, $df = 9$, $p = 0.006$
- Reject H_0 at $\alpha = 0.05$: evidence cups are underfilled
- One-sided 95% CI upper bound: $345.7 < 350$
- Average cup contains ≈ 8 ml less than advertised

The one-sided CI confirms the mean is reliably below the advertised volume

Reading the Output

R Output Fields

Field	Meaning
`t`	Test statistic value
`df`	Degrees of freedom
`p-value`	Probability under H_0
`95% CI`	Confidence interval for μ
`mean of x`	Sample mean $\bar{x}$

How to Report

- State test and direction: "one-sample t-test, one-tailed"
- Give key numbers: $t(9) = -3.16$, $p = 0.006$
- State decision: "reject H_0 at $\alpha = 0.05$"
- Interpret practically: "cups underfilled by ≈ 8 ml"

Always report the test statistic, degrees of freedom, exact p-value, and a practical interpretation

Worked Example: Body Temperature

Scenario

The long-accepted normal body temperature is 98.6°F. A researcher suspects it may be lower.

- $n = 25$ healthy adults
- $\bar{x} = 98.2$°F
- $s = 0.7$°F

H_0: $\mu = 98.6$ vs H_1: $\mu < 98.6$

Calculation

$$t = \frac{98.2 - 98.6}{0.7 / \sqrt{25}} = \frac{-0.4}{0.14} = -2.86$$

- $df = 24$
- $p = 0.009$ (one-tailed)
- Reject H_0 at $\alpha = 0.05$
- Evidence the true mean is below 98.6°F

Recent studies confirm average body temperature has declined—this is now a classic textbook example

Your Turn: Practice

Scenario

A battery manufacturer claims their batteries last 500 hours on average.

- You test $n = 15$ batteries
- Sample mean: $\bar{x} = 485$ hours
- Sample SD: $s = 20$ hours

Your Tasks

1. What are H_0 and H_1?
2. Calculate the t-statistic
3. Find the degrees of freedom
4. What is your conclusion at $\alpha = 0.05$?

Try the calculation before looking at the solution—practice builds fluency

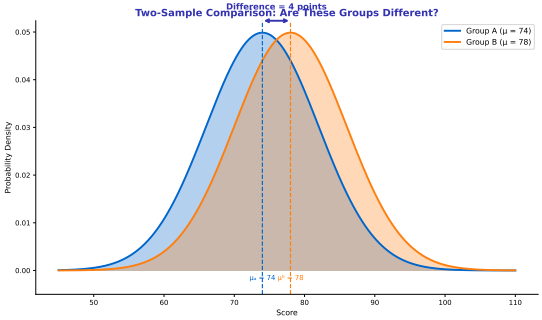
Two-Sample t-Test

Two Groups, One Question

Scenario

Two statistics classes take the same exam:

- Class A ($n = 25$): $\bar{x}_A = 74$, $s_A = 7$
- Class B ($n = 25$): $\bar{x}_B = 78$, $s_B = 6$
- Difference: 4 points
- Is this a real difference or just noise?



The two-sample t-test asks whether two groups were drawn from populations with the same mean

The Pooled t-Statistic

Test Statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Pooled Standard Deviation

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

Worked Example

$$s_p = \sqrt{\frac{24 \cdot 49 + 24 \cdot 36}{48}} = \sqrt{42.5} = 6.52$$

$$t = \frac{74 - 78}{6.52\sqrt{1/25 + 1/25}} = \frac{-4}{1.84} = -2.18$$

- $df = 25 + 25 - 2 = 48$
- $p = 0.034$ (two-tailed)
- Reject H_0: Class B scored significantly higher

The pooled SD combines variability from both groups into a single variance estimate

Effect Size: Cohen's d

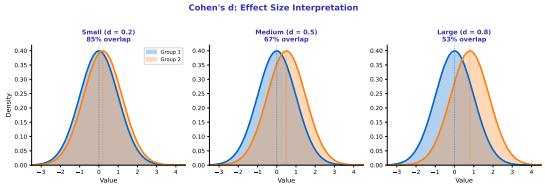
Formula

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

Interpretation Guidelines

Cohen's d	Effect Size
0.2	Small
0.5	Medium
0.8	Large

Example: $d = 4/6.52 = 0.61$ (medium effect)



Cohen's d expresses the difference in standard-deviation units—it is comparable across studies

Equal vs Unequal Variances

Student's t-test

- ✓ Assumes $\sigma_1^2 = \sigma_2^2$
- ✓ Uses pooled s_p and $df = n_1 + n_2 - 2$
- × Verify first with F-test or Levene's test
- × Biased if variances truly differ

Welch's t-test

- ✓ No equal-variance assumption
- ✓ Uses Satterthwaite approximation for df
- ✓ Default in R's `t.test()`
- ✓ The safer, more robust choice

When in doubt, use Welch's test—it is the default in R and performs well even with equal variances

Worked Example: Exam Scores in R

```r
# Class A and Class B scores
classA <- c(72, 68, 74, 80, 77,
            71, 76, 75, 70, 73,
            82, 69, 78, 74, 76,
            72, 75, 71, 80, 74,
            73, 77, 70, 76, 71)
classB <- c(78, 82, 76, 80, 74,
            81, 77, 79, 83, 75,
            80, 78, 76, 82, 79,
            74, 81, 77, 75, 80,
            83, 76, 78, 81, 77)

# Pooled t-test
t.test(classA, classB,
       var.equal = TRUE)
# t = -2.18, df = 48, p = 0.034
```

Interpretation

- $t(48) = -2.18$, $p = 0.034$
- Reject H_0 at $\alpha = 0.05$
- Class B scored significantly higher ($\bar{x}_B = 78$ vs $\bar{x}_A = 74$)
- Cohen's $d = 0.61$: medium effect size

Practical Meaning

A 4-point gap on a 100-point exam is meaningful—roughly half a letter grade.

Always pair the p-value with an effect size to convey practical significance

The Paired t-Test

When to Use

- Same subjects measured twice (before/after)
- Matched pairs (twins, left/right)
- Repeated measures on the same unit

Formula

Compute differences $d_i = x_{1i} - x_{2i}$, then:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}, \quad df = n - 1$$

Example Data

Subject	Before	After	d_i
1	85	82	3
2	92	88	4
3	78	76	2
4	88	83	5
5	76	74	2

$\bar{d} = 3.2\,\text{kg}$, $s_d = 1.30\,\text{kg}$

Pairing removes between-subject variability, increasing statistical power for detecting within-subject changes

Worked Example: Weight Loss in R

```r
# Weight (kg) before and after
before <- c(85, 92, 78,
            88, 76)
after  <- c(82, 88, 76,
            83, 74)

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

# Output:
# t = 3.42, df = 4
# p-value = 0.027
# Mean difference: 3.2 kg

# Effect size
d <- before - after
cohen_d <- mean(d) / sd(d)
# d = 0.68 (medium-large)
```

Interpretation

- $t(4) = 3.42$, $p = 0.027$
- Reject H_0: significant weight loss
- Average loss: 3.2 kg
- Cohen's $d = 0.68$: medium–large effect

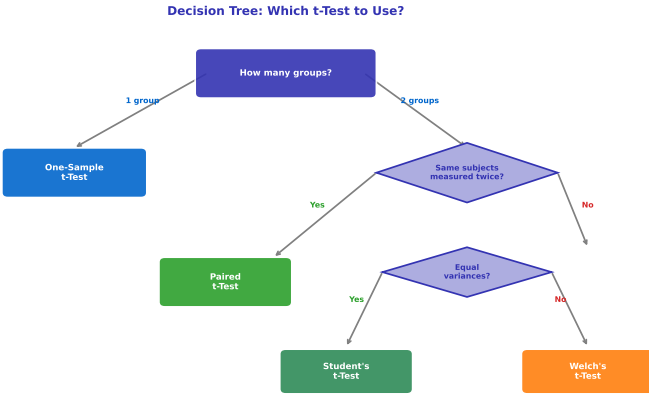
Why Paired?

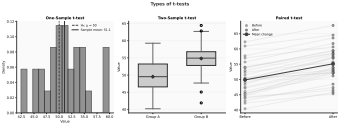
Each subject serves as their own control—individual differences (metabolism, age, starting weight) cancel out.

The paired t-test reduces to a one-sample t-test on the differences

Choosing the Right Test

Decision Flowchart





Three t-test variants share the same logic but differ in how the standard error is computed.

Match your test to your study design: one group vs a value, two independent groups, or paired observations

Worked Example: Classify These Scenarios

Scenario 1
Compare a new drug to placebo in *different* patients.
→ **Two-sample** (independent groups)

Scenario 2
Test whether the average salary at a company equals \$50 000.
→ **One-sample** (single group vs known value)

Scenario 3
Measure anxiety scores before and after therapy in the *same* patients.
→ **Paired** (repeated measures)

Decision Rule

1. How many groups? (one → one-sample)
2. Same subjects? (yes → paired)
3. Different subjects? (→ two-sample)

The study design—not the data—determines which test is appropriate

Common Mistakes

Mistakes

- × Using paired test when groups are independent
- × Ignoring unequal variances (Student's when Welch's is needed)
- × Using one-tailed test without pre-specified direction
- × Interpreting $p > 0.05$ as "no effect exists"

Corrections

- ✓ Check study design first—same subjects or different?
- ✓ Default to Welch's t-test (R's default)
- ✓ Use two-tailed unless you have a prior directional hypothesis
- ✓ "Fail to reject" $\neq$ "accept H_0"

The most common mistake is choosing a test based on what gives the smallest p-value

Analysis of Variance (ANOVA)

Why Not Multiple t-Tests?

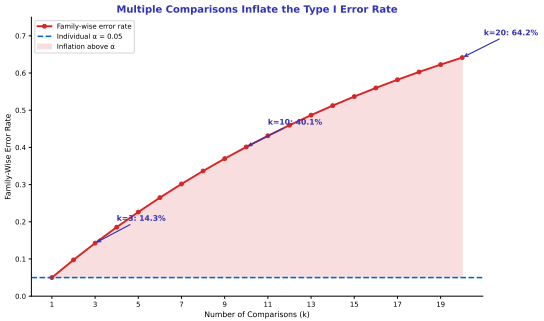
The Inflation Problem

Three fertilisers tested → 3 pairwise *t*-tests.

$$\text{FWER} = 1 - (1 - \alpha)^k$$

- 3 groups → 3 tests: FWER = 14%
- 5 groups → 10 tests: FWER = 40%
- 10 groups → 45 tests: FWER = 90%

ANOVA solves this with a *single* omnibus test.



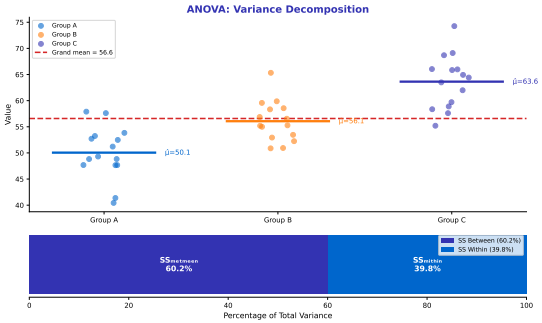
Multiple t-tests inflate the false-positive rate exponentially—ANOVA keeps it at alpha

Variance Decomposition

Partitioning Variance

$$SS_T = SS_B + SS_W$$

- $SS_B = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$ (between groups)
- $SS_W = \sum_{j=1}^{k} \sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$ (within groups)
- $MS_B = SS_B \,/\, (k-1)$
- $MS_W = SS_W \,/\, (N-k)$



ANOVA asks whether between-group variance is large relative to within-group variance

The F-Statistic

Definition

$$F = \frac{MS_B}{MS_W}$$

- Large F: group means differ more than expected by chance
- $F \approx 1$: between-group variance $\approx$ within-group variance
- Always positive; right-skewed distribution

Worked Example

Three teaching methods, $k = 3$, $N = 30$.

Quantity	Value
SS_B	420
SS_W	1200
$MS_B = 420/2$	210
$MS_W = 1200/27$	44.4
$F = 210/44.4$	4.73
df	(2, 27)
p	0.017

An F-ratio much larger than 1 indicates the group means differ more than expected by within-group noise

Worked Example: Teaching Methods in R

```r
# Three teaching methods
scores <- c(75, 82, 78, 80, 77,
            71, 79, 74, 76, 73,
            85, 88, 83, 87, 84,
            90, 86, 82, 88, 85,
            72, 74, 70, 73, 71,
            68, 75, 69, 71, 74)
method <- factor(rep(
  c("A", "B", "C"), each = 10))

# One-way ANOVA
model <- aov(scores ~ method)
summary(model)
```

ANOVA Table

Source	df	SS	MS	F
Method	2	420.0	210.0	4.73
Residual	27	1200.0	44.4	

Interpretation

- $F(2, 27) = 4.73$, $p = 0.017$
- Reject H_0: at least one method differs
- Means: A $= 76.5$, B $= 85.8$, C $= 71.7$
- Post-hoc tests needed to locate differences

A significant ANOVA F-test must be followed by post-hoc tests to identify which groups differ

Assumptions and Alternatives

ANOVA Assumptions

- **Independence**: observations are independent (study design)
- **Normality**: each group is approximately normally distributed
- **Homogeneity**: equal variances across groups (Levene's test)

When Assumptions Fail

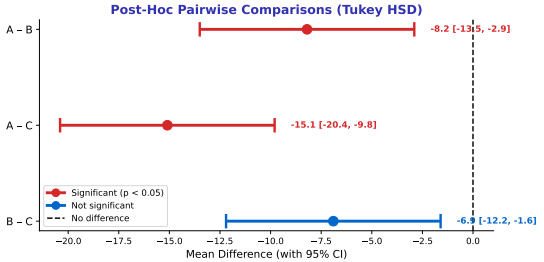
- Unequal variances → **Welch's ANOVA** (`oneway.test()`)
- Non-normality → **Kruskal-Wallis** rank-based test
- Use Shapiro-Wilk per group to check normality
- Use Levene's test (`car::leveneTest()`) to check homogeneity

Welch's ANOVA and Kruskal-Wallis are robust alternatives when classical assumptions are violated

Post-Hoc Tests

Three Main Methods

- **Tukey HSD**: all pairwise comparisons with simultaneous CIs; controls FWER exactly
- **Bonferroni**: adjusts α/k; simple but conservative
- **Scheffé**: tests all possible contrasts; most conservative



Tukey HSD is the standard choice for all pairwise comparisons after a significant ANOVA

Worked Example: Tukey HSD in R

```r
# Tukey HSD post-hoc test
TukeyHSD(model)

# Output (abbreviated):
#       diff     p adj
# B-A    5.2    0.030
# C-A   -7.8    0.001
# C-B  -13.0    0.000

# Alternative: Bonferroni
pairwise.t.test(scores, method,
                p.adjust = "bonf")
```

Tukey HSD Results

Comparison	Diff	95% CI	p_{adj}
B − A	5.2	[0.5, 9.9]	0.030
C − A	−7.8	[−12.5, −3.1]	0.001
C − B	−13.0	[−17.7, −8.3]	<0.001

Interpretation

- B > A: diff = 5.2, $p = 0.03$ (significant)
- C < A: diff = −7.8, $p = 0.001$ (significant)
- Ranking: B > A > C

Post-hoc adjusted p-values already control the family-wise error rate—compare them directly to alpha

Power and Effect Sizes

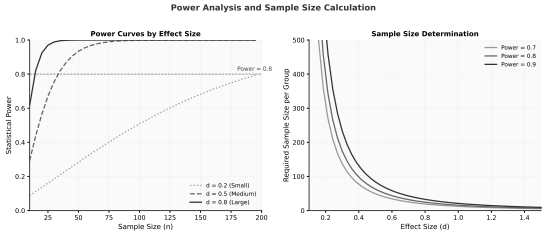
The Underpowered Study Problem

Scenario

A researcher tests a promising drug with only $n = 10$ per group.

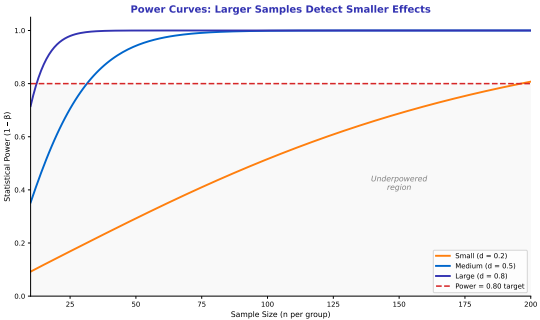
- Result: $p = 0.08$—not significant
- Journal rejects the paper
- But the drug *does* work (true $d = 0.5$)
- Was the study doomed from the start?

With $n = 10$ and $d = 0.5$, power is only **18%**—an 82% chance of missing a real effect.



An underpowered study wastes resources—plan your sample size before collecting data

Statistical Power



Definition

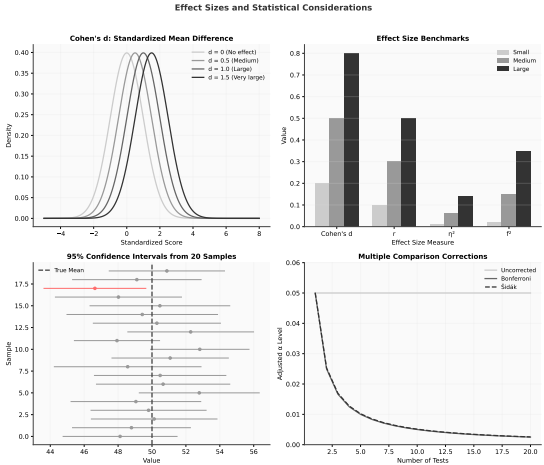
Power $= P(\text{reject } H_0 \mid H_1 \text{ true}) = 1 - \beta$

- Convention: aim for power ≥ 0.80
- Four factors determine power:

Factor	Effect on Power
Effect size ↑	Power ↑
Sample size ↑	Power ↑
α ↑	Power ↑
Variability ↓	Power ↑

Power of 0.80 means an 80% chance of detecting a true effect—one in five real effects will be missed

Effect Sizes: Beyond Significance



Why Report Effect Sizes?

- **Cohen's** d for t-tests: small $= 0.2$, medium $= 0.5$, large $= 0.8$
- η^2 for ANOVA: small $= 0.01$, medium $= 0.06$, large $= 0.14$
- Always report alongside p-values

A significant p-value with a tiny effect size may have no practical importance. Effect sizes answer "how big?"

Statistical significance tells you IF an effect exists; effect size tells you HOW BIG it is

Worked Example: Computing Effect Sizes

Cohen's d (t-test)

From the exam scores example:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} = \frac{74 - 78}{6.52} = 0.61$$

- $d = 0.61 \rightarrow$ medium effect
- 61% of a standard deviation separates the groups

η^2 (ANOVA)

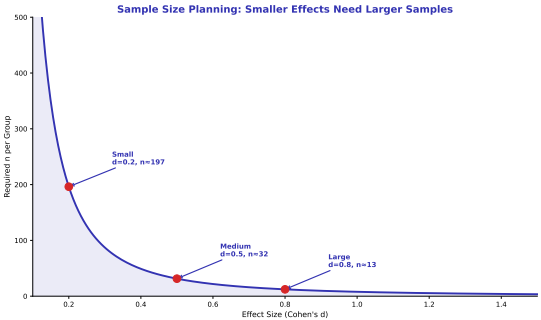
From the teaching methods example:

$$\eta^2 = \frac{SS_B}{SS_T} = \frac{420}{420 + 1200} = \frac{420}{1620} = 0.26$$

- $\eta^2 = 0.26 \rightarrow$ large effect
- Teaching method explains 26% of the variance in scores

Effect sizes are unit-free and allow direct comparison across different studies and scales

Sample Size Planning



Required n per Group

(Two-sample t-test, power $= 0.80$, $\alpha = 0.05$)

Effect Size d	n **per group**
0.2 (small)	393
0.5 (medium)	64
0.8 (large)	26

- Smaller effects need much larger samples
- Doubling n does *not* double power
- Plan *before* data collection

Detecting a small effect requires 15 times more participants than detecting a large effect

Power Analysis in R

```r
library(pwr)

# Two-sample t-test
# How many per group for d=0.5?
pwr.t.test(d = 0.5,
           sig.level = 0.05,
           power = 0.80,
           type = "two.sample")
# n = 63.77 -> need 64 per group

# ANOVA with 3 groups
# Cohen's f = 0.25 (medium)
pwr.anova.test(k = 3,
               f = 0.25,
               sig.level = 0.05,
               power = 0.80)
# n = 52 per group
```

Interpretation

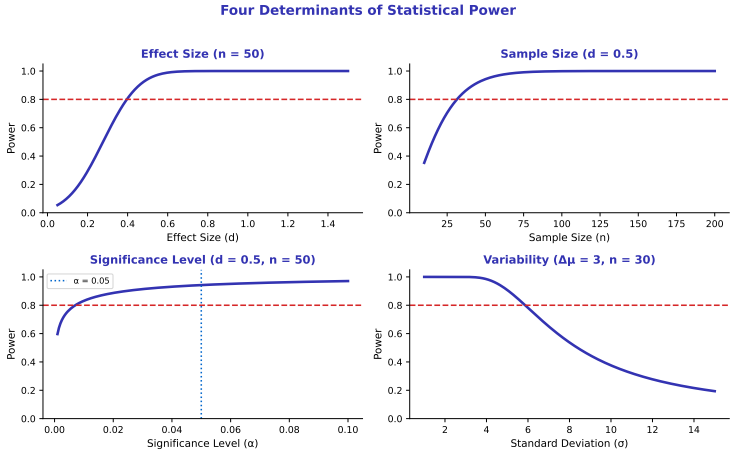
- For $d = 0.5$: need 64 per group (128 total)
- For ANOVA with $f = 0.25$: need 52 per group (156 total)
- Always round up—you cannot have fractional participants

Converting Effect Sizes

$f = \frac{d}{2}$ for two groups, so $d = 0.5 \leftrightarrow f = 0.25$

Run power analysis before your study—it determines whether your budget allows a conclusive answer

The Power Quartet



- Increase n or effect size d to boost power
- Relaxing α boosts power but increases Type I risk—rarely advisable
- Reducing measurement variability (better instruments, controlled conditions) improves power for free

The four levers of power: sample size, effect size, alpha level, and measurement precision

Practice: Plan Your Study

Scenario

You are designing a study to detect a *small* effect ($d = 0.3$) with 80% power at $\alpha = 0.05$.

- Two independent groups (treatment vs control)
- Outcome: continuous test score
- Budget: maximum 100 participants total

Questions to Answer

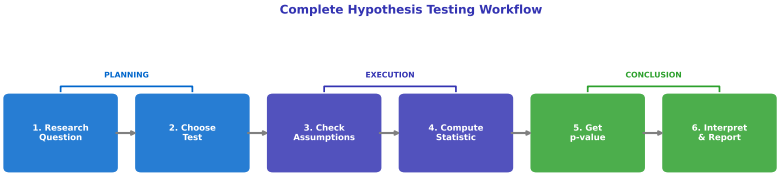
1. How many participants per group do you need?
2. Can you run this study within your budget of $n = 100$?
3. If not, what is the minimum d detectable with $n = 50$ per group?

Use `pwr.t.test()` in R to find the answers.

Planning forces you to confront whether your study can succeed before investing resources

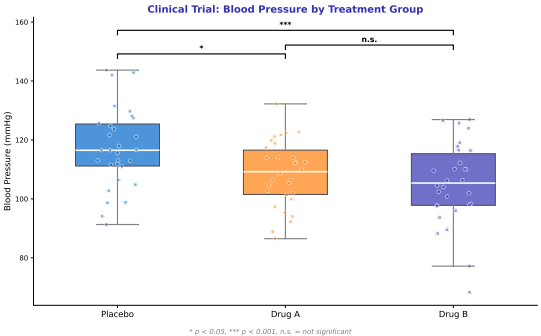
Putting It All Together

Complete Workflow



Follow this workflow for every analysis: design, check assumptions, test, interpret, and report

Case Study: Clinical Trial



Results

- Three treatment groups: Placebo, Drug A, Drug B
- ANOVA: $F(2, 57) = 8.42$, $p = 0.001$
- Tukey HSD: Drug B significantly better than placebo ($p = 0.001$); Drug A vs placebo marginal ($p = 0.07$)
- Effect size: $\eta^2 = 0.16$ (large)

A complete analysis reports hypotheses, assumptions, test statistics, p-values, effect sizes, and practical conclusions

Full Analysis Walkthrough

Dataset

- Three teaching methods (A, B, C)
- Test scores, $n = 10$ per group
- Research question: Do methods differ?

Step 1: Choose Test

Three independent groups $\rightarrow$ one-way ANOVA.

Step 2: Check Assumptions

- Shapiro-Wilk per group: all $p > 0.05$
- Levene's test: $p > 0.05$

Step 3: Run and Interpret

- $F(2, 57) = 4.73$, $p = 0.017$
- Reject H_0: methods differ
- Tukey: B > A ($p = 0.03$), B > C ($p < 0.001$)
- $\eta^2 = 0.14$ (large effect)

A structured walkthrough ensures no step is skipped—from design through interpretation

Best Practices and Pitfalls

Do

- ✓ Report effect sizes alongside p-values
- ✓ Plan sample size *a priori*
- ✓ Use two-tailed tests as the default
- ✓ Check assumptions before running tests

Don't

- × p-hack by running many tests until one is significant
- × Confuse statistical significance with practical importance
- × Ignore violated assumptions
- × Report only "significant" results (publication bias)

Good practice: pre-register hypotheses, report all results, and always include effect sizes

Key Takeaways

1. **Hypothesis testing is a structured framework** for evidence-based decisions—follow the five steps every time
2. **Choose the right test** by matching your study design: one-sample, paired, two-sample, or ANOVA
3. **Always report effect sizes** alongside p-values—statistical significance alone does not imply practical importance
4. **Plan your sample size before collecting data**—an underpowered study wastes resources and risks missing real effects

Hypothesis testing is a tool for evidence-based decision making, not a guarantee of truth

Appendix: Extended Statistical Tests

Beyond t-tests and ANOVA — a visual tour of the statistician's toolbox

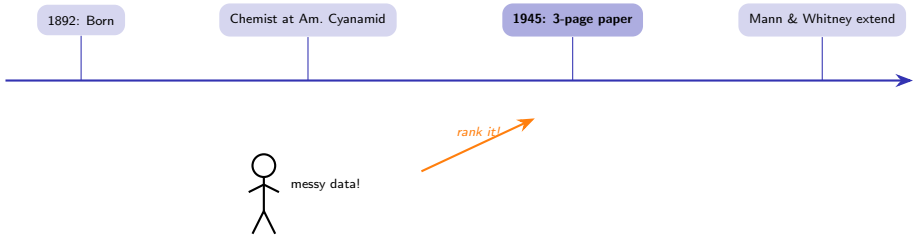
Appendix Overview

1. **Non-Parametric Tests** — Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman
2. **Chi-Squared Tests** — Goodness of fit, independence, McNemar, Fisher's exact
3. **Correlation Tests** — Pearson, Spearman, Kendall, Anscombe's quartet
4. **Normality Tests** — Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling

5. **Variance Tests** — Levene, Brown-Forsythe, Bartlett, F-test
6. **Proportion Tests** — Z-test for proportions, Fisher's exact, A/B testing
7. **Advanced ANOVA** — Two-way, repeated measures, MANOVA, ANCOVA
8. **Multiple Testing & Modern Methods** — Bonferroni, FDR, bootstrap, permutation

Each test group includes history, visual intuition, and R code

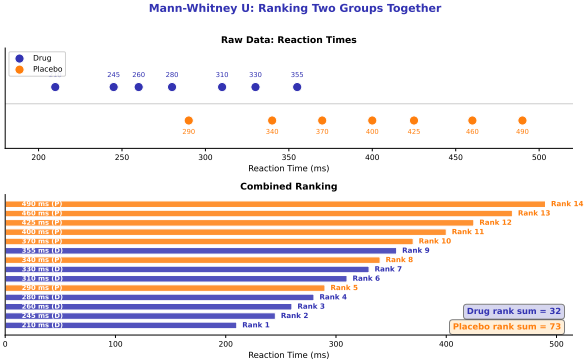
The Chemist Who Ranked Everything



- Frank Wilcoxon, a chemist, found his pesticide data never followed normal distributions
- His 1945 paper was only **3 pages** but revolutionised statistics for messy datasets
- Key idea: replace raw values with **ranks** — eliminates the normality assumption

Wilcoxon showed that replacing values with ranks eliminates the need for normality

Mann-Whitney U Test: Visual Intuition



- Rank **all** values from both groups together
- Sum ranks within each group separately
- A large difference in rank sums $\Rightarrow$ groups differ

Test statistic U counts how often a value from group A exceeds a value from group B.

Mann-Whitney U is the non-parametric alternative to the independent two-sample t-test

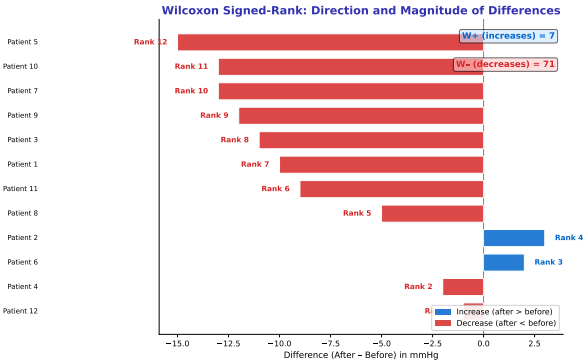
Mann-Whitney U: R Example

```r
# Reaction times: Drug vs Placebo
# (skewed data -- ranks are safer)
drug    <- c(12, 15, 18, 22, 45)
placebo <- c(25, 30, 35, 40, 60)

wilcox.test(drug, placebo)
# W = 3, p-value = 0.032
```

- `wilcox.test()` performs both Mann-Whitney U and Wilcoxon rank-sum
- $p = 0.032 < 0.05$: significant difference in reaction times
- No assumption of normality required

Use when data are ordinal, skewed, or sample size is too small to verify normality

Wilcoxon Signed-Rank Test

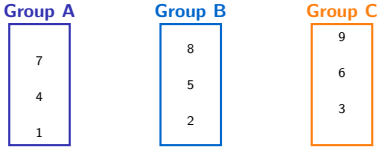


- Compute paired differences $d_i = x_{i,\text{after}} - x_{i,\text{before}}$
- Rank the $|d_i|$ values, ignoring zeros
- Attach the original sign to each rank; sum positive and negative ranks separately

The signed-rank test is the non-parametric counterpart to the paired t-test

Kruskal-Wallis and Friedman Tests

Kruskal-Wallis



Independent groups, all values ranked together

Friedman

Subject	T1	T2	T3
S1	3	3	3
S2	2	3	1
S3	3	1	1

Ranks computed within each subject

- Kruskal-Wallis: ranks across k independent groups (extends Mann-Whitney)
- Friedman: ranks within each subject across k conditions (extends Wilcoxon signed-rank)

Kruskal-Wallis extends Mann-Whitney to 3+ groups; Friedman extends Wilcoxon signed-rank

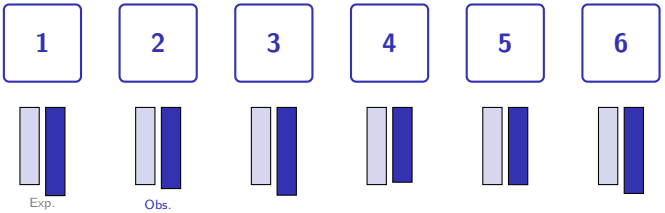
Non-Parametric Tests: When and Why

Parametric Test	Non-Parametric Alternative	When to Switch
One-sample t	Wilcoxon signed-rank	Skewed, ordinal data
Paired t	Wilcoxon signed-rank	Non-normal differences
Two-sample t	Mann-Whitney U	Unequal variances, skew
One-way ANOVA	Kruskal-Wallis	$n < 20$ per group, non-normal
Repeated measures	Friedman	Ordinal or ranked outcomes

- Non-parametric tests trade a small amount of power for freedom from distributional assumptions
- With large n, parametric and non-parametric results usually agree

Rule of thumb: use non-parametric tests when $n < 20$ and normality is doubtful

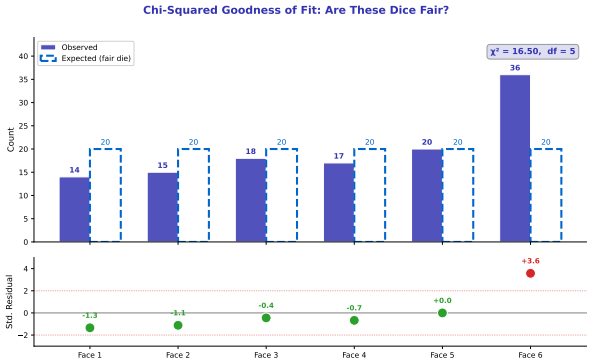
Pearson's Dice Problem



- Karl Pearson (1900) tested whether Monte Carlo dice were loaded
- He compared **observed** frequencies to **expected** frequencies under fair dice
- $\chi^2 = \sum(O_i - E_i)^2/E_i$ — large values indicate poor fit

Pearson asked: "Does the pattern of observed counts deviate too far from what chance predicts?"

Chi-Squared Goodness of Fit



- Tests whether a single categorical variable matches a theoretical distribution
- Requires all expected counts $E_i \geq 5$
- Degrees of freedom: $df = k - 1$ where k is the number of categories

Goodness of fit tests whether a single categorical variable matches a theoretical distribution

Chi-Squared Test of Independence

	Low	Med	High	Row total
Group A	30	15	5	**50**
Group B	10	25	15	**50**
	40	**40**	**20**	**100**

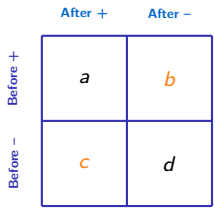
Column totals

- Expected count: $E_{ij} = (\text{row total} \times \text{col total}) \,/\, n$
- Tests whether two categorical variables are **independent**
- $df = (r-1)(c-1)$ where r = rows, c = columns

Independence test asks: "Are two categorical variables related, or is the association due to chance?"

McNemar's Test and Fisher's Exact Test

McNemar's Test



Only b and c matter

- Paired categorical data (before/after)
- $\chi^2 = (b-c)^2/(b+c)$

Fisher's Exact Test

3	1
1	3

Small counts: use exact p

- For small samples where χ^2 approximation fails
- Calculates exact p-value using the hypergeometric distribution

McNemar handles paired categorical data; Fisher's exact test works when expected counts are below 5

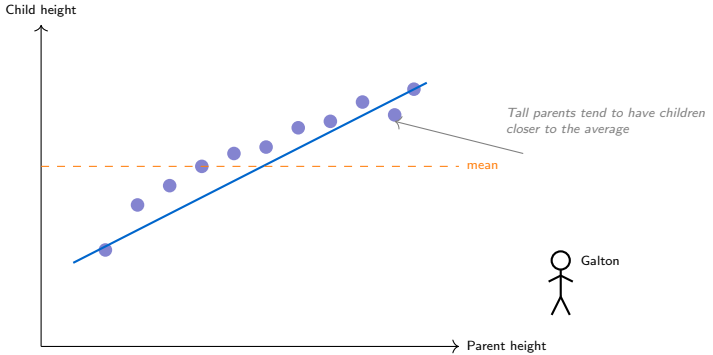
Chi-Squared Tests: R Code

```r
# Contingency table
tbl <- matrix(c(30, 10, 15, 25),
              nrow = 2)

# Chi-squared test
chisq.test(tbl)

# Fisher's exact (small samples)
fisher.test(tbl)

# McNemar's (paired data)
mcnemar.test(tbl)
```

- `chisq.test()` reports χ^2 and p-value
- Check expected counts with `chisq.test()$expected`
- Use `fisher.test()` when any expected count < 5
- `mcnemar.test()` requires paired before/after design

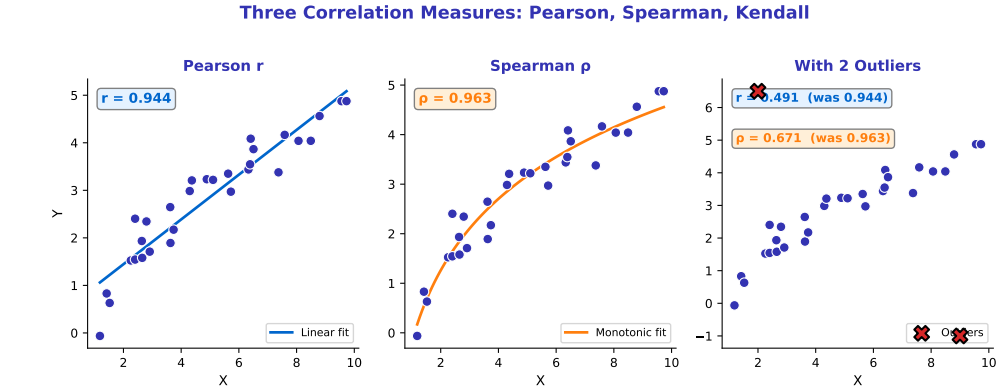
R reports both the chi-squared statistic and p-value; check expected counts with chisq.test()$expected

From Peas to Pearson's r



- Francis Galton studied heredity: parents' height vs children's height
- He discovered **regression to the mean** — extreme parents have less extreme children
- His student Karl Pearson formalised the correlation coefficient r

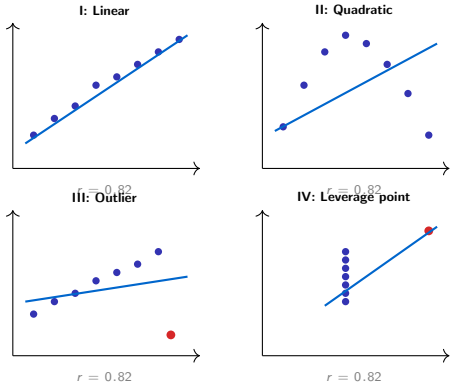
Galton discovered correlation while studying sweet peas — he called it "co-relation"



- **Pearson's** r: measures *linear* association; sensitive to outliers
- **Spearman's** ρ: rank-based, measures *monotonic* association
- **Kendall's** τ: counts concordant vs discordant pairs; robust with small n

Pearson measures linear association; Spearman and Kendall measure monotonic association

When Pearson Fails: Anscombe's Quartet



All four have the same Pearson's r!

Same correlation coefficient can hide very different relationships — always plot your data first

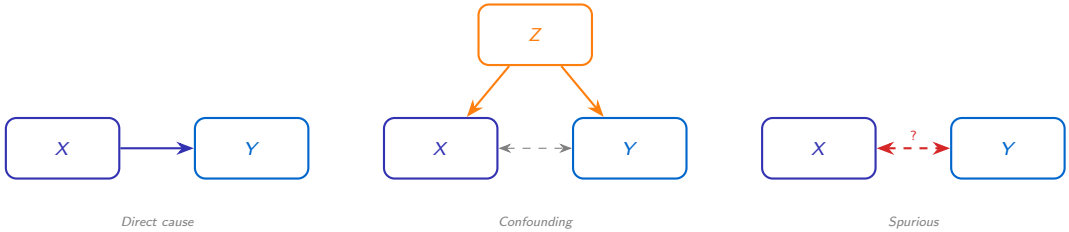
Correlation Tests: R Code

```r
# Pearson (linear)
cor.test(x, y, method = "pearson")

# Spearman (monotonic, rank-based)
cor.test(x, y, method = "spearman")

# Kendall (robust, small samples)
cor.test(x, y, method = "kendall")
```

- All three return a correlation estimate and p-value for H_0: $\rho = 0$
- Pearson assumes bivariate normality
- Spearman and Kendall work on ordinal or skewed data

cor.test() returns both the correlation coefficient and a p-value for testing H_0: $\rho = 0$

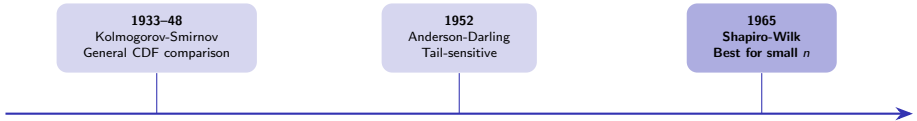
Correlation is Not Causation



- **Direct cause**: X genuinely influences Y (e.g., dosage $\rightarrow$ response)
- **Confounding**: a hidden variable Z drives both (e.g., temperature $\rightarrow$ ice cream and drowning)
- **Spurious**: pure coincidence with no mechanism (e.g., pirate count vs global warming)

A strong correlation between ice cream sales and drowning does not mean ice cream causes drowning

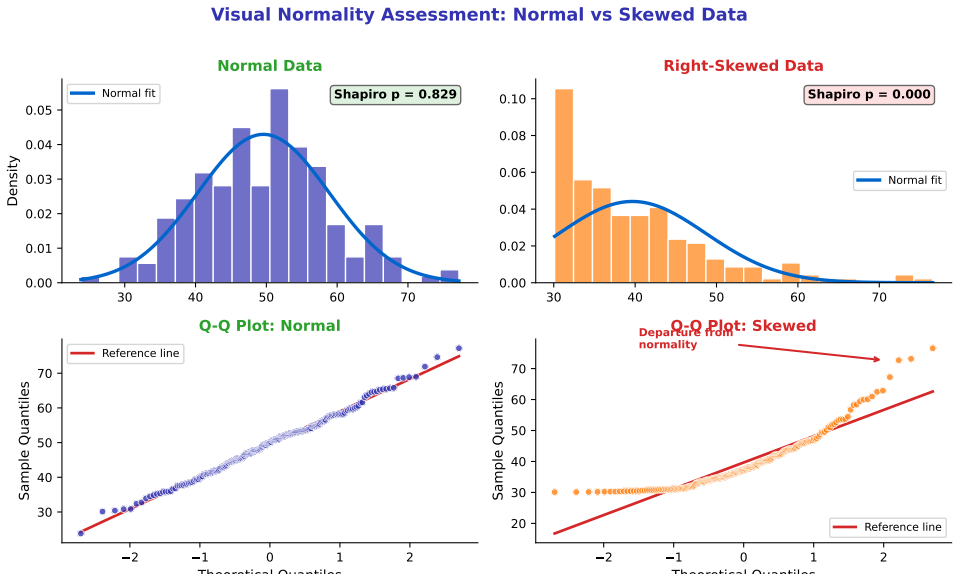
Testing the Bell Curve



- **Kolmogorov-Smirnov**: compares empirical CDF to theoretical CDF; general but less powerful
- **Anderson-Darling**: weights tails more heavily; detects tail deviations
- **Shapiro-Wilk**: designed specifically for small samples ($n < 50$); most powerful for normality

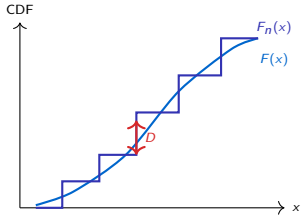
Normality tests answer: "Is it reasonable to assume these data came from a normal distribution?"

Visual Normality Assessment



Shapiro-Wilk vs KS vs Anderson-Darling

	S-W	K-S	A-D
Best for	$n < 50$	General	Tails
Power	High	Low	Medium
Scope	Normal only	Any dist.	Any dist.



- KS measures the maximum distance D between empirical and theoretical CDFs
- Shapiro-Wilk is most powerful for the specific question "is this normal?"

Shapiro-Wilk is most powerful for small samples; KS is more general but less powerful for normality

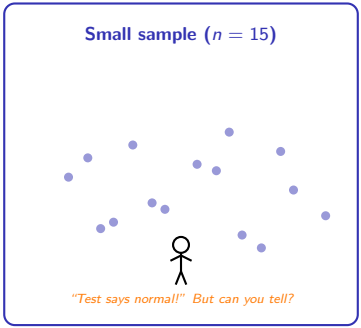
Normality Tests: R Code

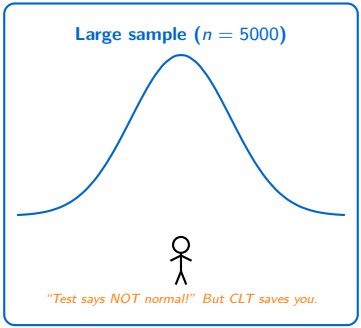
```r
# Shapiro-Wilk (best for n < 50)
shapiro.test(x)

# Kolmogorov-Smirnov (general)
ks.test(x, "pnorm",
        mean(x), sd(x))

# Anderson-Darling (tail-sensitive)
# library(nortest)
# ad.test(x)
```

- Shapiro-Wilk: $p > 0.05 \Rightarrow$ fail to reject normality
- KS: compare to *any* theoretical distribution, not just normal
- Always pair with Q-Q plots for visual confirmation

Normality tests are sensitive to sample size — with $n > 100$, even minor deviations are flagged as significant

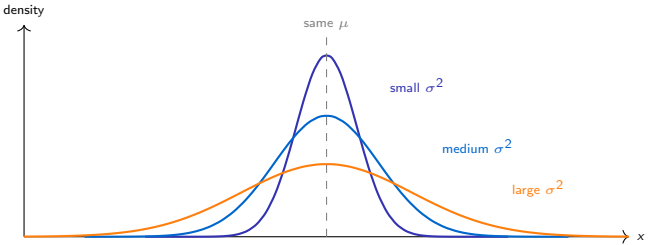
The Normality Paradox





Paradox: when your sample is small enough to need the test, it lacks power; when it is large, the CLT makes it unnecessary

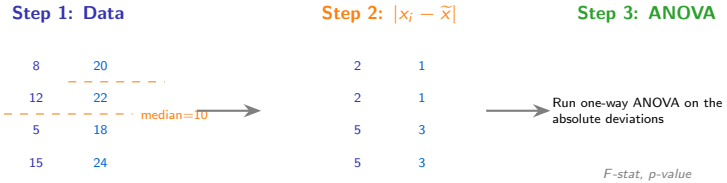
Testing Spread, Not Center



- Levene (1960): robust test using absolute deviations from group medians
- Brown-Forsythe (1974): improved Levene by using the median — robust to skew
- Bartlett (1937): elegant but extremely sensitive to non-normality

Variance equality (homoscedasticity) is a key assumption for t-tests and ANOVA

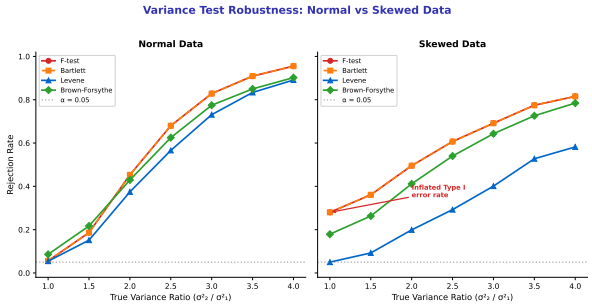
How Levene's Test Works



- Replace each observation with its absolute distance from the group median
- A significant ANOVA on these distances ⇒ unequal variances

Levene's test transforms the variance question into a means question that ANOVA can handle

Comparing Variance Tests



Test	Robust?	Best when
F-test	No	Normal only
Bartlett	No	Normal, $k \geq 2$
Levene	Yes	General use
B-F	Yes	Skewed data

- Brown-Forsythe is the safest default choice
- Bartlett is most powerful under perfect normality

Brown-Forsythe is the safest default; Bartlett is most powerful under normality but fragile otherwise

Variance Tests: R Code

```r
# Levene's test (requires car)
# library(car)
leveneTest(score ~ group,
           data = df)

# Bartlett's test
bartlett.test(score ~ group,
              data = df)

# F-test for 2 groups
var.test(x, y)
```

- `leveneTest()` uses median by default (Brown-Forsythe variant)
- `bartlett.test()` is built-in but fragile with non-normal data
- `var.test()` is the simple F-test for exactly two groups

Always test variance equality before running a pooled t-test or one-way ANOVA

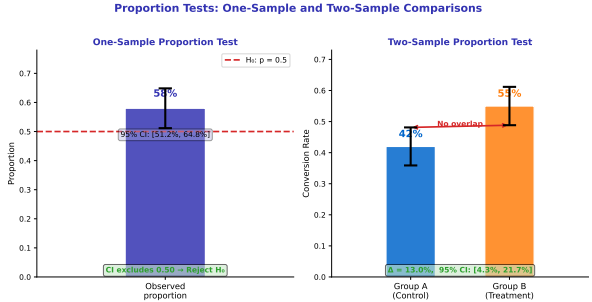
The Lady Tasting Tea

"I can tell whether milk was added before or after the tea."

- Fisher (1920s): colleague Muriel Bristol claimed she could taste the difference
- Fisher designed the **exact test**: how likely is her result if she is just guessing?
- 8 cups, 4 of each type — $\binom{8}{4} = 70$ possible arrangements

Fisher designed the exact test for this experiment — it remains the gold standard for small 2×2 tables

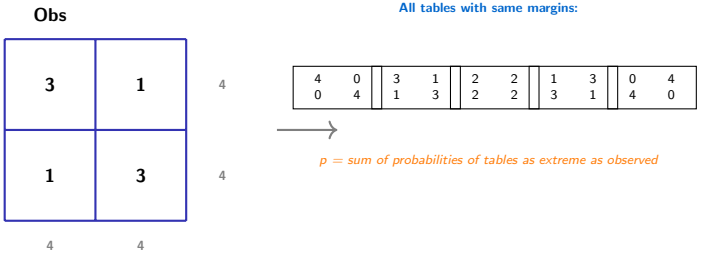
Z-Test for Proportions



- Tests whether an observed proportion $\hat{p}$ differs from a hypothesised value p_0
- Two-proportion version compares $\hat{p}_1$ vs $\hat{p}_2$
- Uses normal approximation to the binomial (requires $np \geq 10$)

The Z-test for proportions uses the normal approximation to the binomial distribution

Fisher's Exact Test: How It Works



- Fix the row and column totals (margins)
- Enumerate all possible tables with those margins
- p-value = probability of observing a table this extreme or more extreme

Fisher's exact test calculates the exact probability rather than relying on a chi-squared approximation

One-Proportion vs Two-Proportion

	One-Proportion	**Two-Proportion**
Question	Is $\hat{p}$ different from p_0?	Is $\hat{p}_1$ different from $\hat{p}_2$?
Example	Is our pass rate $\neq 50\%$?	Is Group A's rate $\neq$ Group B's?
H_0	$p = p_0$	$p_1 = p_2$
Use case	Quality control, surveys	A/B testing, clinical trials

- A/B testing in tech is simply a **two-proportion Z-test**
- Requires $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ for the normal approximation
- For small n: use Fisher's exact test instead

A/B testing in tech is simply a two-proportion Z-test — comparing conversion rates between variants

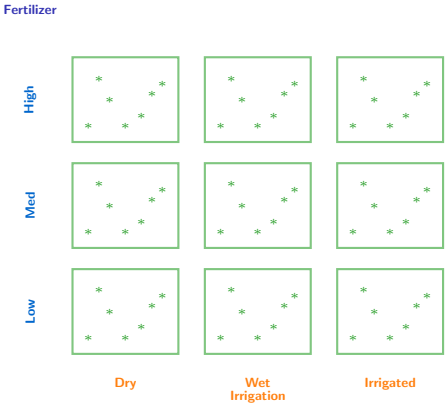
Proportion Tests: R Code

```r
# One proportion: 45/100 vs 50%
prop.test(45, 100, p = 0.5)

# Two proportions: A vs B
prop.test(c(45, 60), c(100, 120))

# Fisher's exact (small samples)
fisher.test(matrix(c(3,1,1,3),
                   nrow = 2))
```

- `prop.test()` applies continuity correction by default
- Set `correct=FALSE` for the standard Z-test
- `fisher.test()` is exact and works for any sample size

prop.test() uses continuity correction by default; set correct=FALSE for the standard Z-test

Fisher's Fields: ANOVA Beyond One Factor

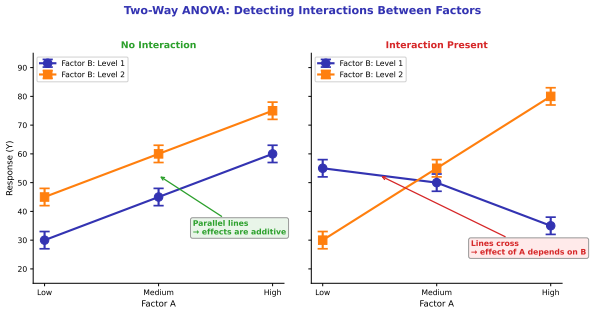


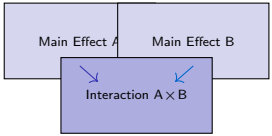
Fisher realised he could study **both factors** *and their* **interaction** *simultaneously*

- Rothamsted Experimental Station, 1920s: Fisher analysed crop yields with multiple factors
- Factorial designs let you study main effects **and** interactions in one experiment
- Split-plot designs arose from practical farming constraints

Fisher's agricultural experiments at Rothamsted gave us factorial designs and the analysis of variance

Two-Way ANOVA: Main Effects and Interaction





- Parallel lines ⇒ no interaction
- Crossing lines ⇒ interaction present
- Always check interaction *before* interpreting main effects

Interaction means the effect of one factor depends on the level of the other — parallel lines indicate no interaction

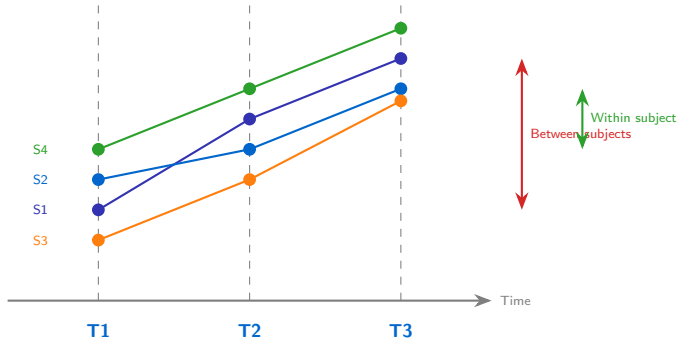
Two-Way ANOVA: R Code

```r
# Two-way ANOVA with interaction
model <- aov(yield ~ fertilizer *
               irrigation,
             data = crops)
summary(model)

# Interaction plot
interaction.plot(
  crops$fertilizer,
  crops$irrigation,
  crops$yield)
```

- The $*$ operator includes both main effects and the interaction: `A + B + A:B`
- Check the interaction p-value first
- If interaction is significant, main effects alone are misleading

The * operator in R formulas includes both main effects and the interaction: A + B + A:B

Repeated Measures ANOVA

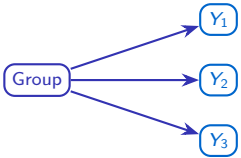


- Each subject is measured at multiple time points — lines connect same subject
- Accounts for within-subject correlation (each person is their own control)
- Mauchly's test checks the **sphericity** assumption (equal variances of differences)

Repeated measures ANOVA accounts for correlation within subjects — each person serves as their own control

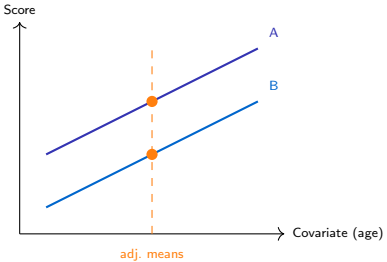
MANOVA and ANCOVA

MANOVA



Multiple DVs simultaneously

ANCOVA



Adjusts group means for a continuous covariate

- MANOVA tests whether groups differ on **multiple** outcomes jointly
- ANCOVA adjusts for a continuous confounder (e.g., age, baseline score)

MANOVA handles multiple outcomes at once; ANCOVA adjusts for continuous confounders

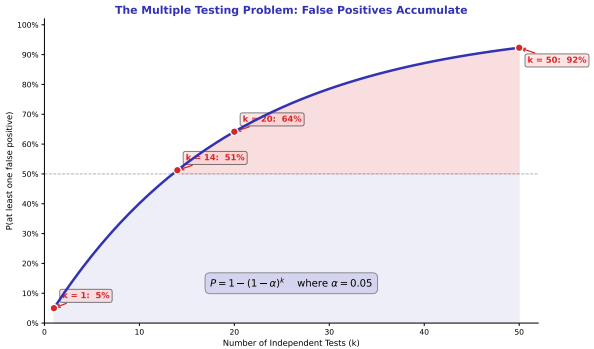
Advanced ANOVA: R Code

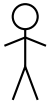
```r
# Repeated measures ANOVA
aov(score ~ time +
    Error(subject/time),
    data = df)

# ANCOVA (adjust for age)
aov(score ~ group + age,
    data = df)

# MANOVA (multiple outcomes)
manova(cbind(y1, y2) ~ group,
       data = df)
```

- `Error(subject/time)`: specifies the within-subject factor structure
- ANCOVA: the covariate must be continuous and linearly related to Y
- MANOVA: `cbind()` bundles multiple dependent variables

For repeated measures in R, the Error() term specifies the within-subject factor structure

The Multiple Testing Crisis





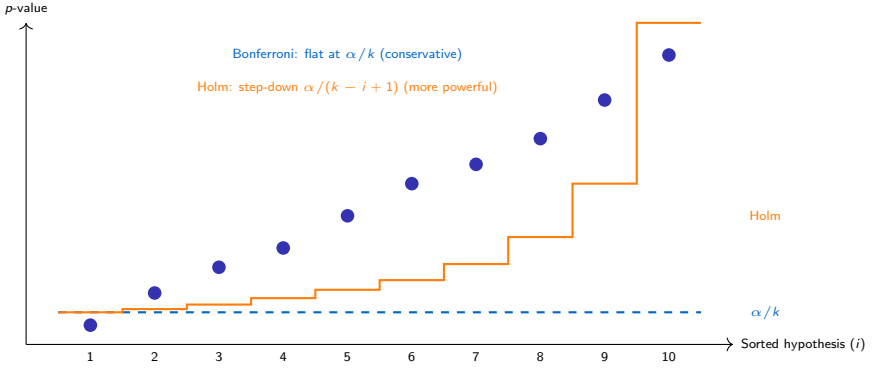
Test 1: $p > 0.05$
Test 2: $p > 0.05$
Test 3: $p > 0.05$
Test 4: $p > 0.05$
Test 5: $p > 0.05$
Test 6: $p = 0.04$ **!!**
"Eureka!"

(false positive)

- With 20 tests at $\alpha = 0.05$: $P(\text{at least 1 false positive}) = 1 - 0.95^{20} = 64\%$
- The more tests you run, the more likely you "find" something by chance

With 20 tests at $\alpha = 0.05$, you have a 64% chance of at least one false positive

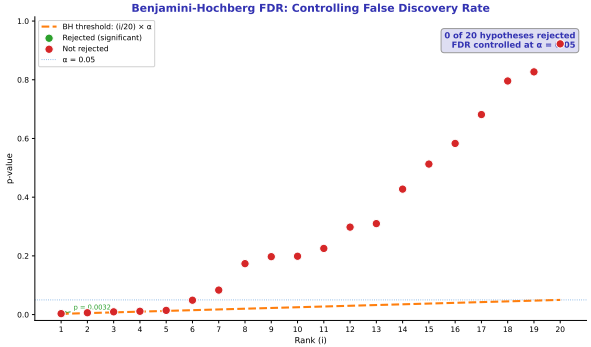
Bonferroni and Holm Corrections



- **Bonferroni**: reject if $p_i < \alpha/k$ — simple but very conservative
- **Holm**: step-down procedure — uniformly more powerful, same Type I control

Holm's method is uniformly more powerful than Bonferroni and should be preferred in practice

False Discovery Rate (BH)



- Sort p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
- Draw threshold line: $p_{(i)} \leq \frac{i}{m} \cdot \alpha$
- Reject all hypotheses below the line
- Controls the **expected proportion** of false discoveries

FDR controls the expected proportion of false discoveries, not the probability of any single error

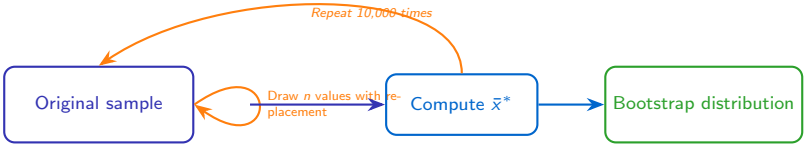
Multiple Testing: R Code

```r
# Raw p-values from 6 tests
p_values <- c(0.001, 0.008,
              0.039, 0.041,
              0.23,  0.99)

# Bonferroni correction
p.adjust(p_values,
         method = "bonferroni")

# Holm step-down
p.adjust(p_values,
         method = "holm")

# Benjamini-Hochberg FDR
p.adjust(p_values,
         method = "BH")
```

- `p.adjust()` handles all major correction methods in base R
- Bonferroni: multiply each p by k (number of tests)
- BH (Benjamini-Hochberg): recommended default for exploratory analyses

In R, p.adjust() handles all major correction methods — BH is the recommended default

Bootstrap Tests: The Computer-Age Revolution

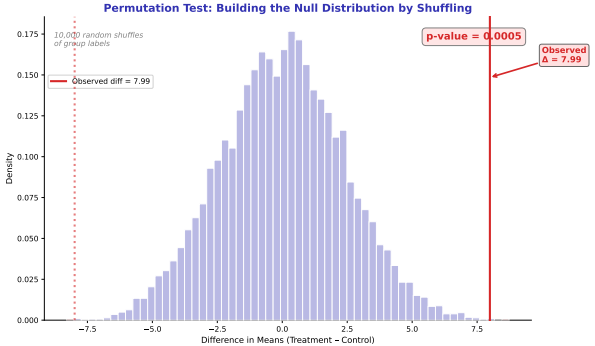


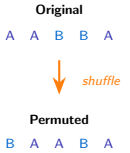
Bradley Efron, Stanford, 1979: "The most important idea in statistics in the last 50 years"

- No distributional assumptions — let the data speak for themselves
- Resample with replacement to build the sampling distribution computationally
- Works for *any* statistic: means, medians, regression coefficients, etc.

Bootstrap replaces mathematical assumptions with computational power — resample to build the sampling distribution

Permutation Tests





- Shuffle group labels randomly
- Recompute the test statistic
- Repeat thousands of times

Permutation tests make no distributional assumptions — they test whether the grouping label matters

Modern Methods: R Code

```r
# Bootstrap confidence interval
library(boot)
boot_fn <- function(data, i)
  mean(data[i])
b <- boot(x, boot_fn, R = 10000)
boot.ci(b, type = "perc")

# Permutation test
library(coin)
independence_test(y ~ group,
                  data = df)
```

- `boot()`: pass data, statistic function, and number of resamples R
- `boot.ci()`: extracts percentile, BCa, or normal CIs
- `coin::independence_test()` provides an exact permutation p-value

Bootstrap and permutation tests are exact for any sample size and make no parametric assumptions

Choosing Among All Tests: Master Decision Tree