Power and Sample Size

Textbook Section 5.4

Emile Latour, Nicky Wakim, Meike Niederhausen

February 18, 2026

Learning Objectives

By the end of today's lecture, you will be able to:

1. Understand the four components in equilibrium in a hypothesis test
2. Define and interpret Type I and Type II errors
3. Define power and understand its role in study design
4. Calculate power and sample size using R for one-sample t-tests
5. Calculate power and sample size using R for two-sample t-tests

Roadmap for Today

Part 1: The Four Components

- What affects our ability to detect effects?
- Significance level (α)
- Sample size (n)
- Effect size
- Power ($1 - \beta$)

Part 2: Errors in Hypothesis Testing

- Type I errors (α)
- Type II errors (β)
- Power as "correct detection"
- Visualizing errors with distributions

Part 3: Calculating Power and Sample Size

- Introduction to Cohen's *d*
- Using the `pwr` package in R
- One-sample t-test examples
- Two-sample t-test examples

Part 4: Study Design Applications

- Planning studies with power in mind
- Interpreting existing study results
- Trade-offs and practical considerations

Connecting to What We Know

Where we've been: Hypothesis testing

Over the past few lectures, we've learned to:

- Use **confidence intervals** to estimate population parameters
- Conduct **hypothesis tests** to evaluate claims about parameters
- Work with three types of t-tests:
  - One-sample (compare to known value)
  - Paired (compare before/after)
  - Two independent samples (compare groups)

In each case, we:

1. Collected data
2. Calculated a test statistic
3. Got a p-value
4. Made a conclusion (reject H_0 or fail to reject)

Quick reference: Everything we've covered

	One-sample	Independent two-sample	Paired sample
Example	Body temp: Population mean is 98.6°F, is sample different?	Caffeine: taps/min between caffeine and non-caffeine group	Vegetarian diet: cholesterol before and after
Sample statistic	$\bar{x}$	$\bar{x}_1 - \bar{x}_2$	$\bar{x}_d$
Population parameter	μ	$\mu_1 - \mu_2$	μ_d or δ
Possible hypothesis tests	$H_0 : \mu = \mu_0$	$H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$	$H_0 : \mu_d = 0$ or $\delta = 0$
Standard error	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$SE = \frac{s_d}{\sqrt{n}}$
Test statistic	$t = \frac{\bar{x} - \mu_0}{SE}$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$	$t = \frac{\bar{x}_d - 0}{SE}$
Confidence intervals	$\bar{x} \pm t^* \times SE$	$(\bar{x}_1 - \bar{x}_2) \pm t^* \times SE$	$\bar{x}_d \pm t^* \times SE$

This summarizes all three test types - we'll use these concepts as we think about power

The question we haven't answered yet

Scenario: You're planning a new study

Questions you need to answer:

- How many participants do I need?
- Can I detect the effect I'm looking for?
- What if I can only recruit 20 people?
- Is my study worth running?

This is about study design:

- *Before* collecting data
- *Before* spending time/money
- *Before* asking participants to volunteer
- Making sure your study can answer your question

Today: We learn how to design studies with adequate **power** to detect real effects

Motivating example: The caffeine study

Research Question: Does caffeine increase finger tapping speed?

Study Design:

- 70 college students trained to tap fingers rapidly
- Randomly assigned to two groups:
  - **Control group:** Decaffeinated coffee (n=35)
  - **Caffeine group:** Coffee with 200mg caffeine (n=35)
- After 2 hours, measure taps per minute

Results: The caffeine group had significantly higher taps/min ($p < 0.001$)

The questions we SHOULD have asked first

But what if we had asked BEFORE running the study:

- “We can only recruit 20 per group - is that enough?”
- “What difference can we actually detect with this sample size?”
- “How likely are we to find an effect if it’s really there?”

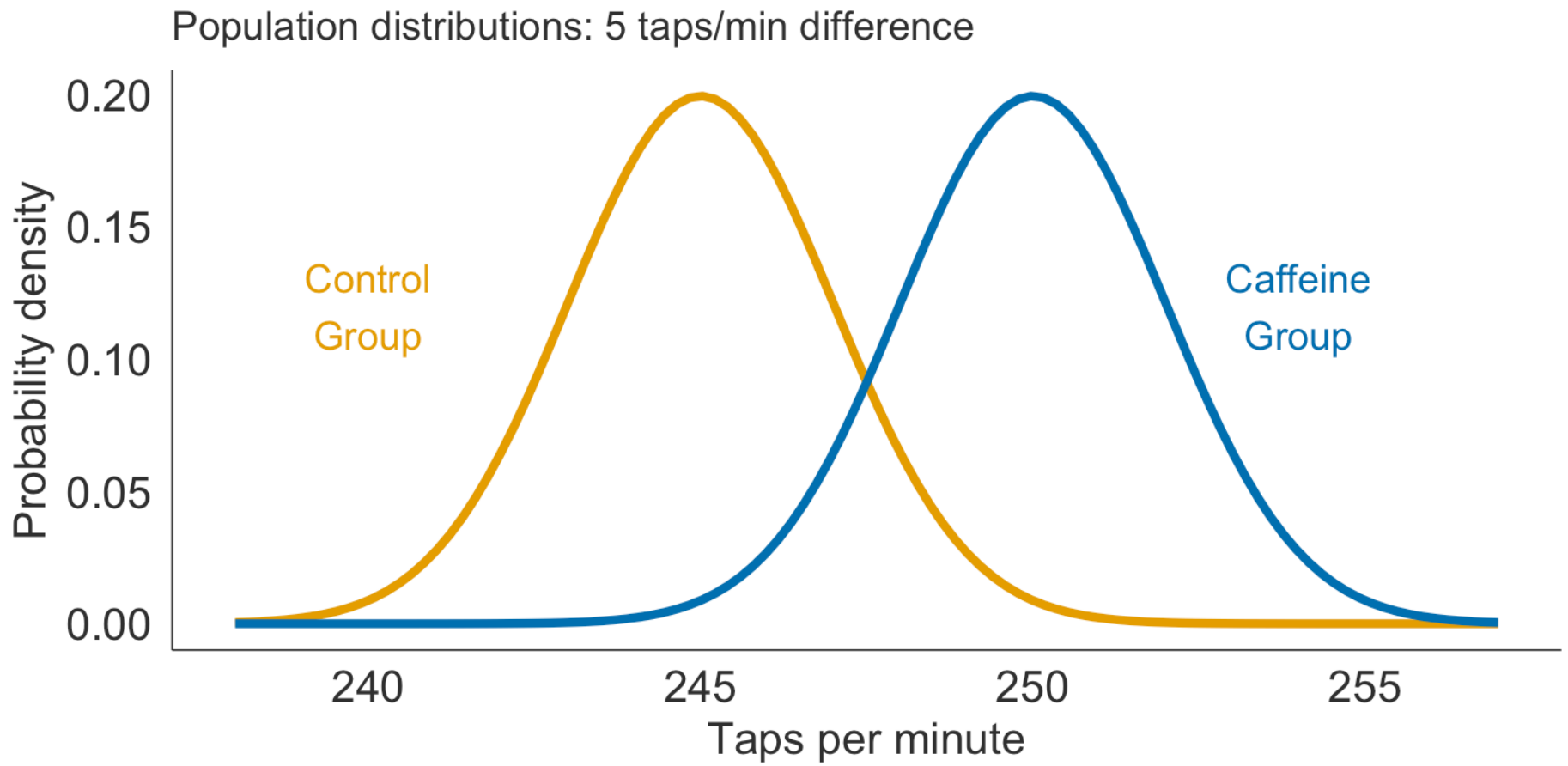
These questions are about POWER - and that’s what we’ll learn today

> **Note**
>
> Power analysis happens *before* data collection, not after!

Part 1: The Four Components

What affects our ability to detect an effect?

Scenario: Imagine two populations that differ in their mean tapping speed

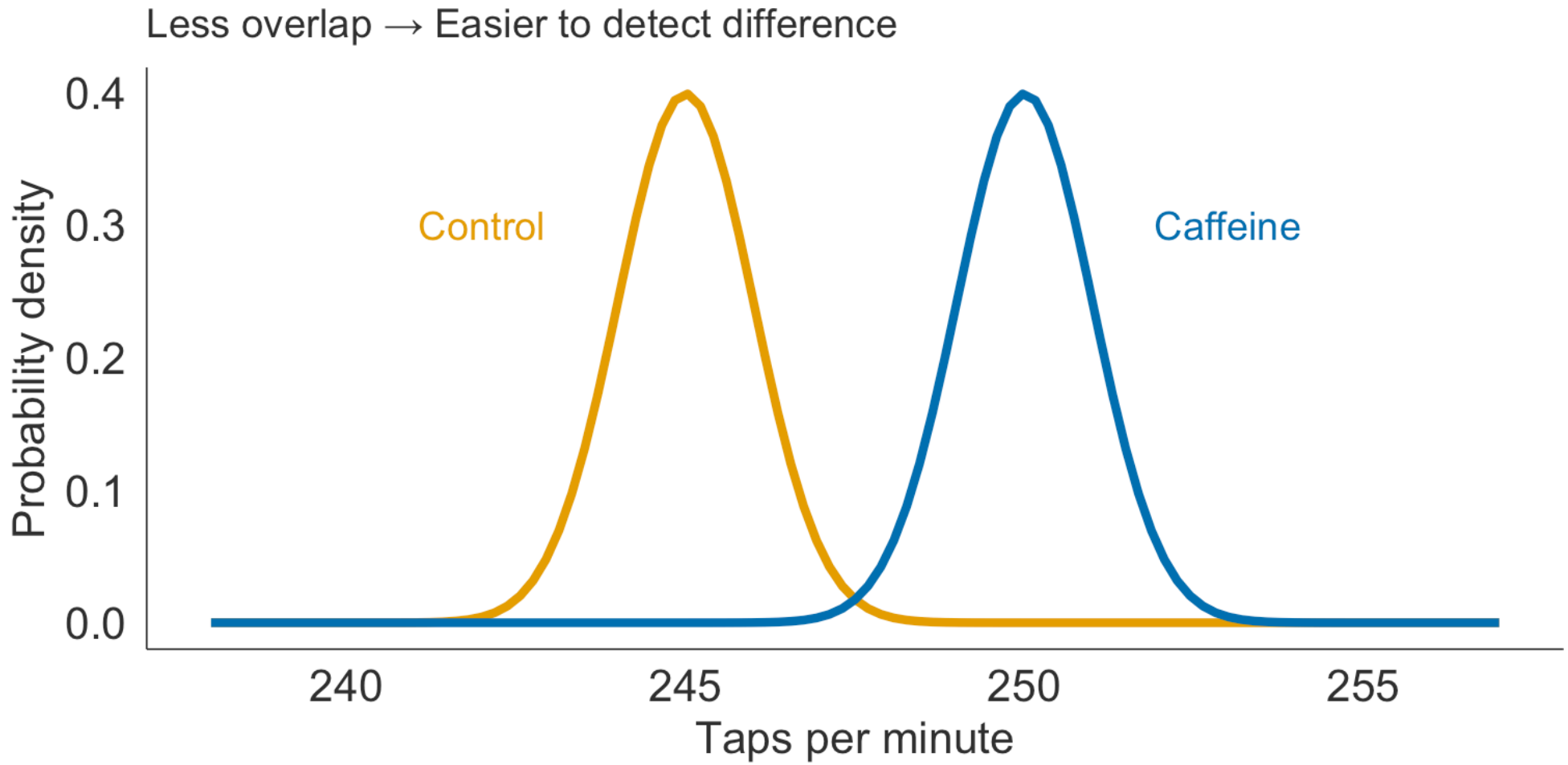


Question: When we take samples from these groups, will it be easy to tell them apart?

Answer: It depends on several factors...

Scenario 1: Small standard deviation

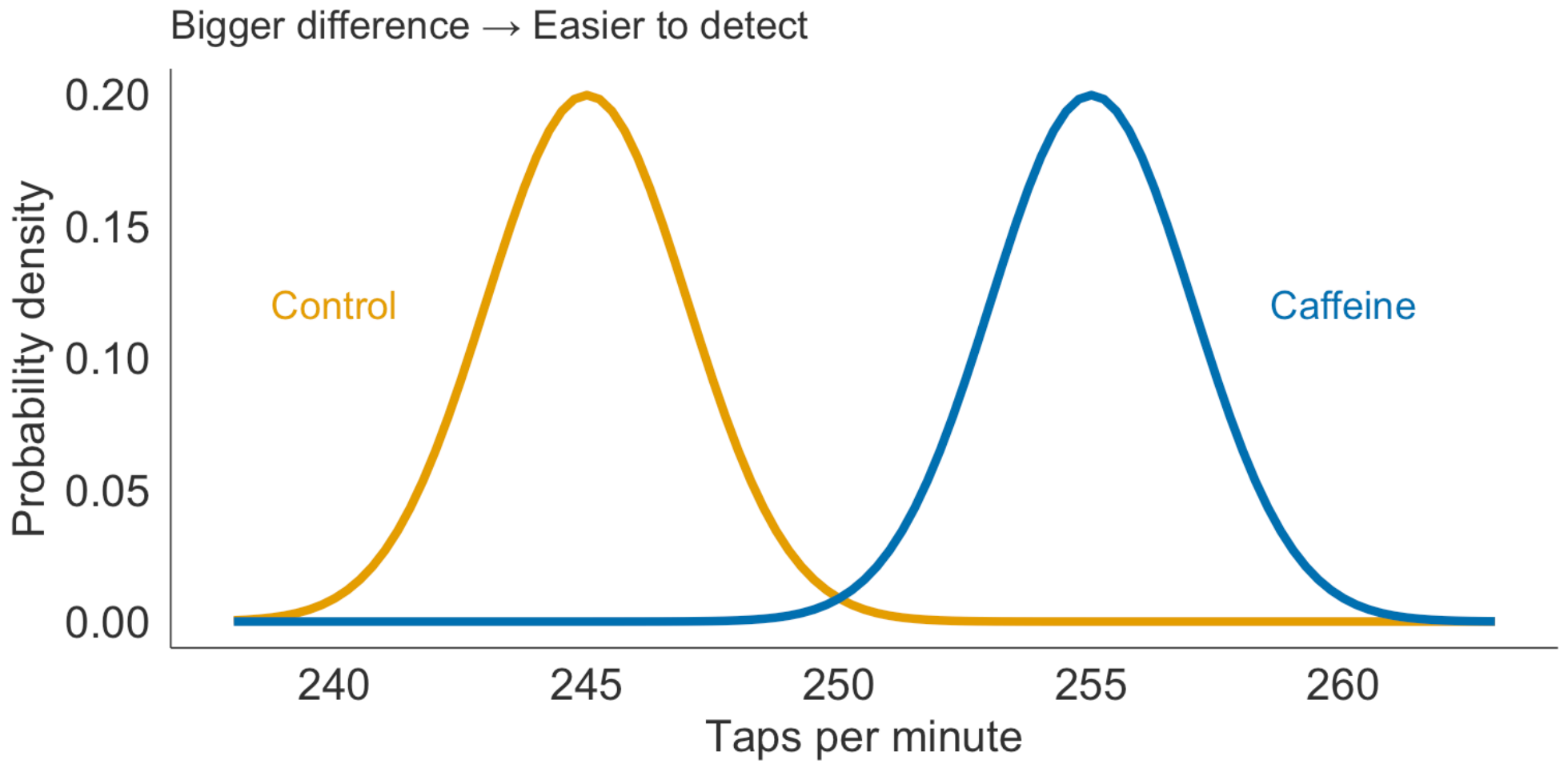
Same difference (5 taps/min), but less variability (SD = 1 instead of 2)



Observation: Less overlap between distributions makes the difference easier to detect!

Scenario 2: Larger difference

Larger difference (10 taps/min instead of 5), same variability (SD = 2)



Observation: A larger true difference is easier to detect!

What did we just observe?

From our visual examples, we noticed:

1. **Less variability** (smaller SD) → distributions pull apart → easier to detect difference
2. **Bigger true difference** → distributions pull apart → easier to detect difference
3. These affect how much **overlap** there is between groups

But we're missing something important:

- What about sample size? More data should help us detect differences
- What about our significance level? Being more strict (lower α) changes our threshold

Let's formalize these ideas...

Statisticians recognize **four components** that work together in any hypothesis test. Change one, and you affect the others!

The four components in equilibrium

In any hypothesis test, four quantities are mathematically related:

1. Significance level (α)

- Probability of Type I error
- Usually set at 0.05
- Set *before* collecting data
- Determines rejection region

3. Sample size (n)

- Number of observations
- Larger n → smaller SE
- Often what we're trying to determine
- Constrained by time, cost, ethics

2. Power ($1 - \beta$)

- Probability of detecting a real effect
- Typically want 80% or 90%
- What we calculate given the others
- The probability of "correctly rejecting"

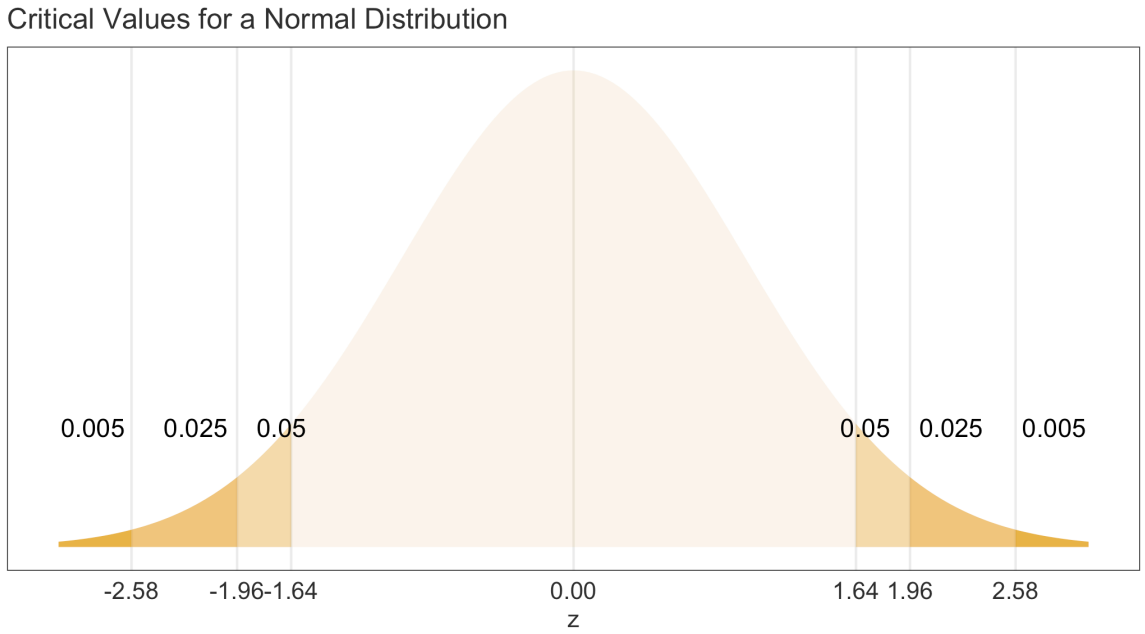
4. Effect size

- Difference we want to detect
- Relative to variability
- Often measured by Cohen's *d*
- Based on pilot data or literature

Key insight: Given any 3 pieces, we can solve for the 4th.

Part 2: Errors in Hypothesis Testing

Significance levels and critical values



What are critical values?

Cutoff points that determine when to reject H_0

Determined by:

- Significance level (α)
- One- vs two-sided test
- Distribution type

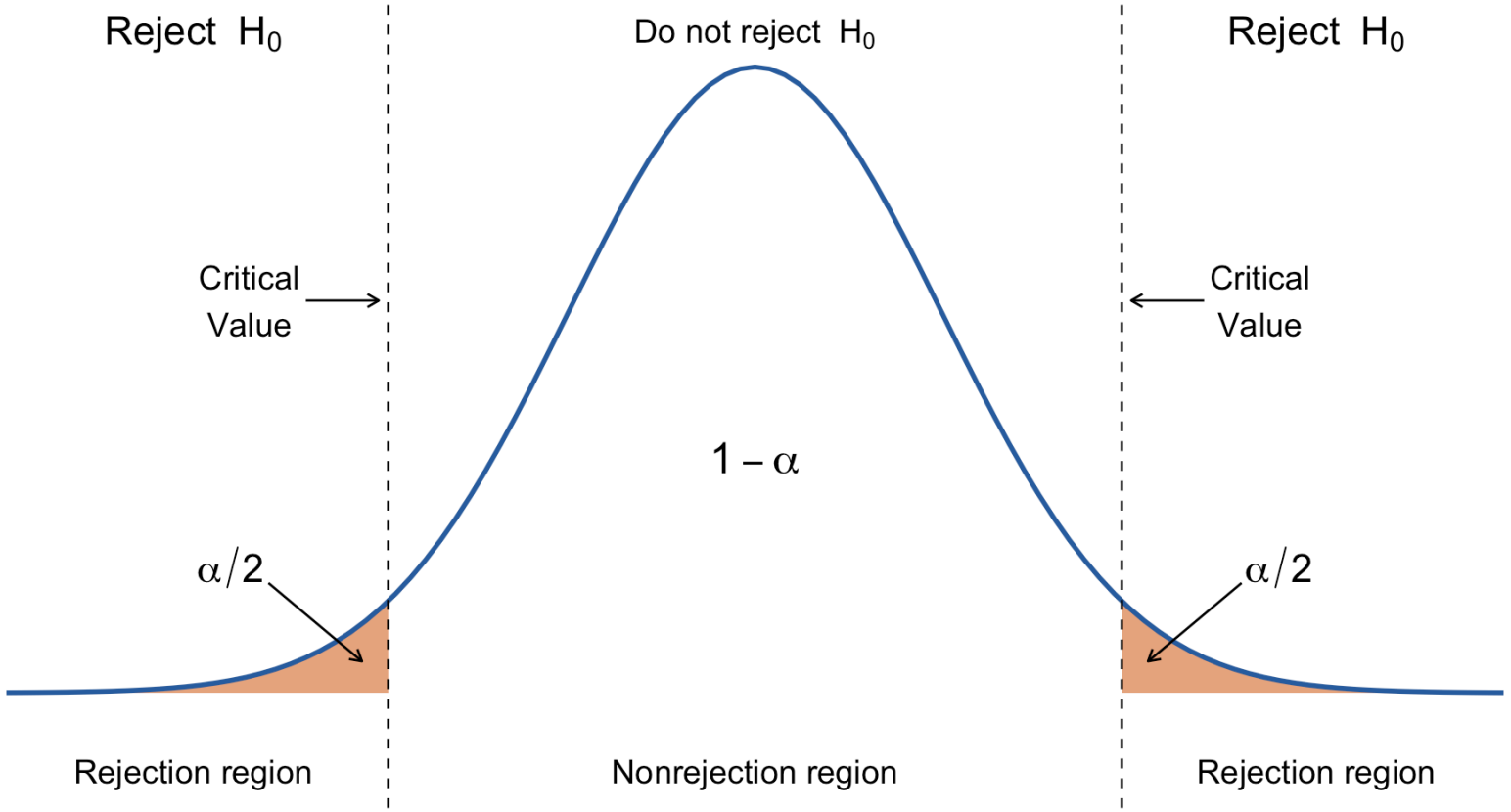
For our tests:

- We've been using t^* from t-distribution (figure shows z-distribution)
- Typically $\alpha = 0.05$
- Two-sided tests (most common)

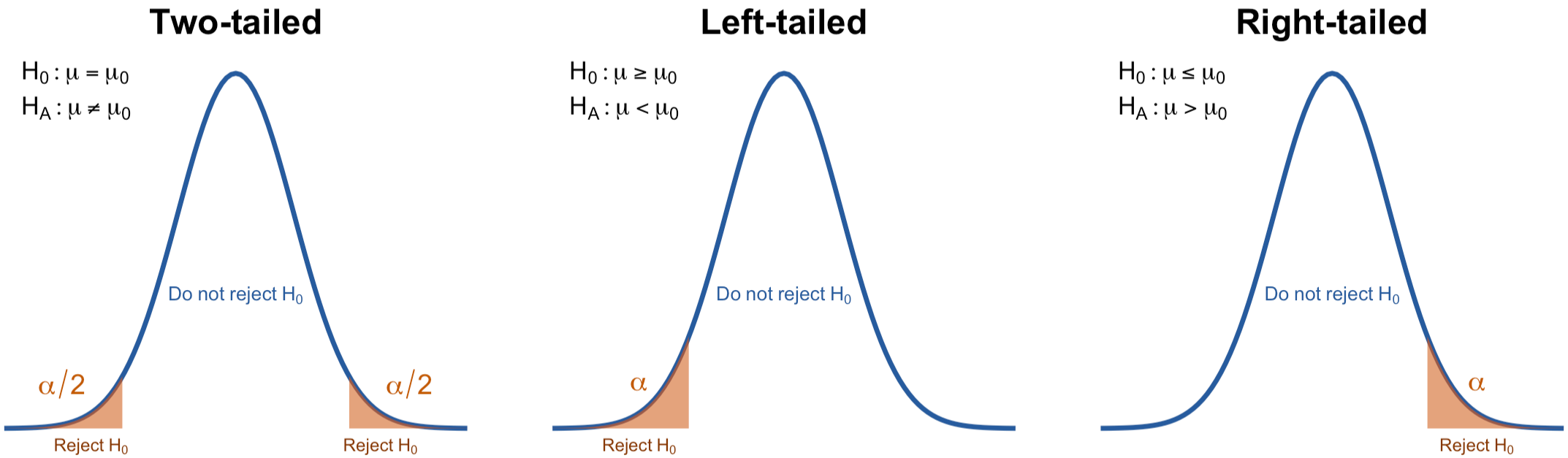
If |test statistic| > critical value → reject H_0

Rejection regions

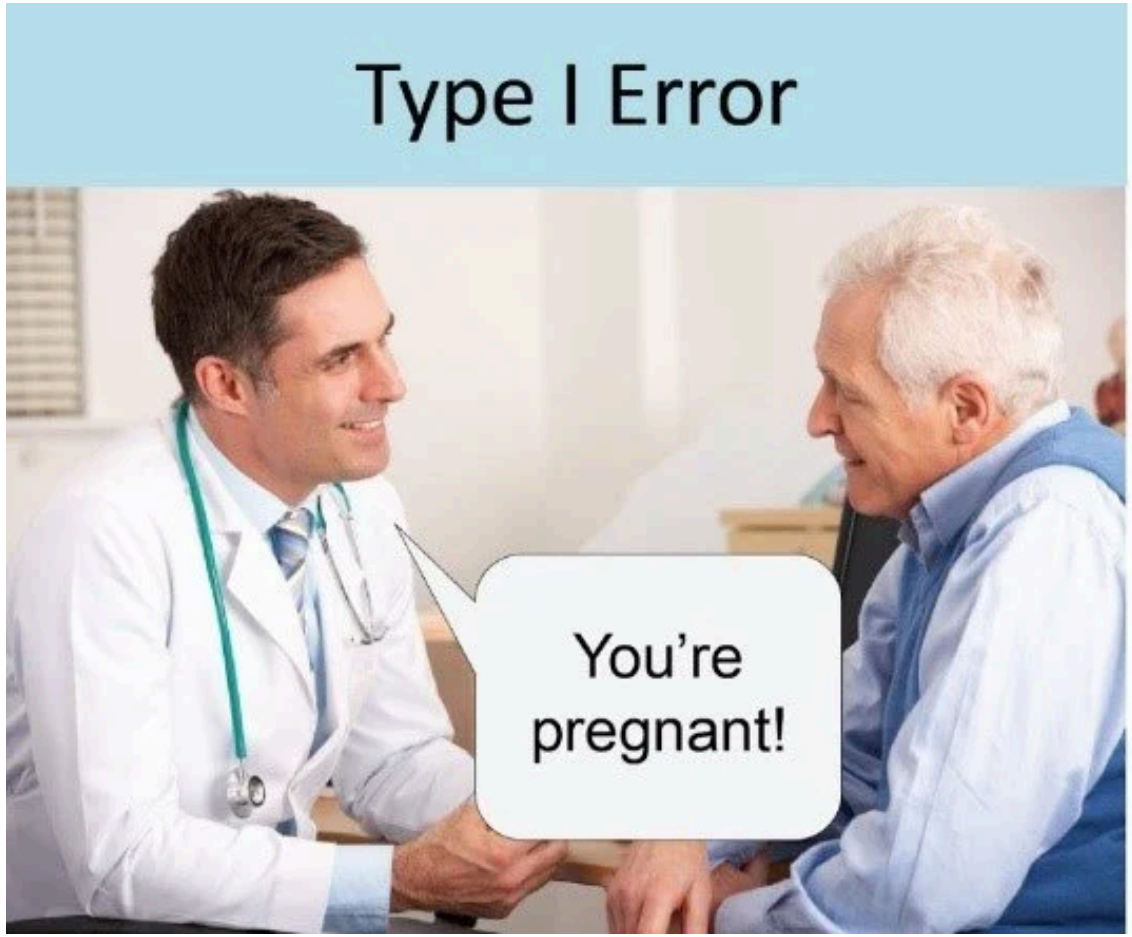
- If the absolute value of the test statistic is greater than the critical value, we reject H_0
  - In this case the test statistic is in the **rejection region**.
  - Otherwise it's in the non-rejection region.

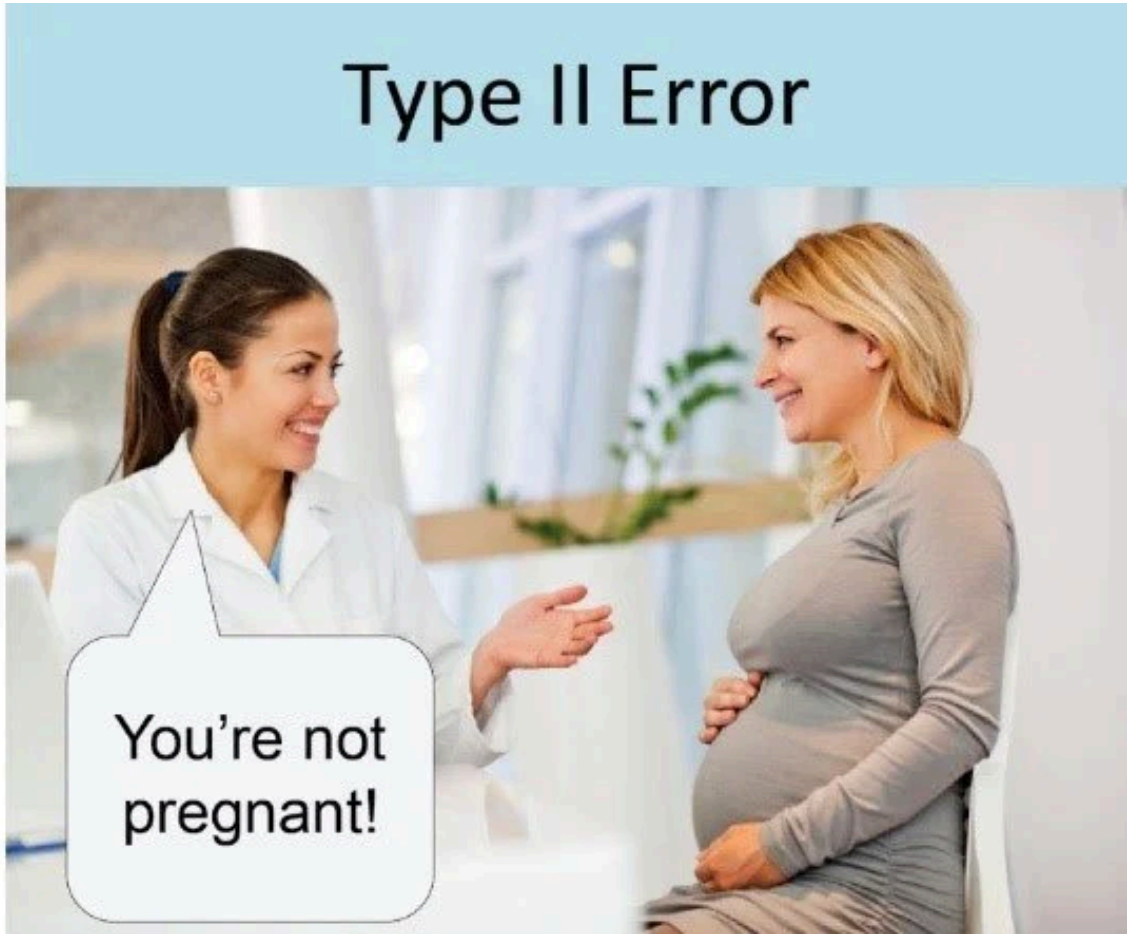


Visual: One-sided vs. two-sided



Hypothesis testing errors





StatisticsSolutions

- Null hypothesis: The patient is not pregnant.
- Alternative hypothesis: The patient is pregnant.

Justice system analogy

Justice System - Trial

	Defendant Innocent	Defendant Guilty
Reject Presumption of Innocence (Guilty Verdict)	**Type I Error**	**Correct**
Fail to Reject Presumption of Innocence (Not Guilty Verdict)	**Correct**	**Type II Error**

Statistics - Hypothesis Test

	Null Hypothesis True	Null Hypothesis False
Reject Null Hypothesis	**Type I Error**	**Correct**
Fail to Reject Null Hypothesis	**Correct**	**Type II Error**

Type I and Type II Errors - Making Mistakes in the Justice System

Type I and Type II errors

Remember: Hypothesis tests can make mistakes in two ways

Type I Error (False Positive)

Definition: Reject H_0 when it's actually true

Notation: Probability of making a Type I error = α

Example: Conclude caffeine increases tapping when it really doesn't

Control: Set α before study (usually 0.05)

Also called: False positive, α error

Type II Error (False Negative)

Definition: Fail to reject H_0 when it's false

Notation: Probability of making a Type II error = β

Example: Fail to detect caffeine effect when it really exists

Control: Increase sample size, decrease variability

Also called: False negative, β error

Trade-off: Decreasing one type of error often increases the other!

Power: The probability of correctly detecting an effect

Power Definition

Power = $1 - \beta$ = Probability of correctly rejecting H_0 when it's false

- The probability of detecting an effect that actually exists
- Typically want power ≥ 0.80 (80%)
- Some fields require 0.90 (90%)

Why 80%? This is a convention balancing:

- Reasonable chance of detecting real effects
- Practical constraints on sample size
- Cost and feasibility

Think of it this way:

- High power = good "detector" - likely to find effect if it exists
- Low power = bad "detector" - might miss real effects

The complete picture

	Reality	
Test decision	H_0 **True**	H_A **True**
Reject H_0	Type I Error (α)	**Power** ($1 - \beta$) ✓
Fail to reject H_0	Correct ($1 - \alpha$) ✓	Type II Error (β)

Correct decisions:

- $1 - \alpha$: Confidence (correctly fail to reject when H_0 is true)
- $1 - \beta$: Power (correctly reject H_0 when H_A is true)

Errors:

- α: Type I error rate (false positive)
- β: Type II error rate (false negative)

What we control in study design: We typically fix α at 0.05 and choose n to achieve desired power

Visualizing Type I error, Type II error, and Power

Two possible realities:

- Dashed grey: H_0 is true (no effect)
- Solid blue: H_A is true (real effect exists)

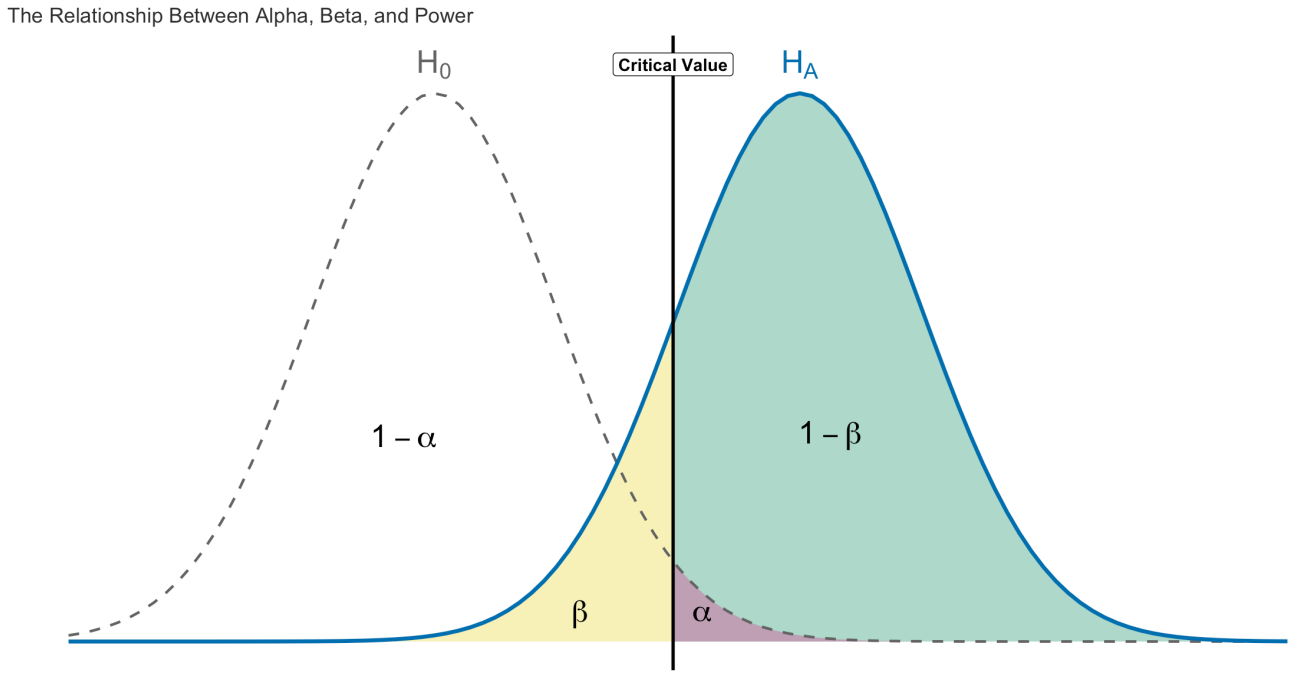
Our decision boundary: The critical value separates "reject" from "don't reject"

Under H_0 (grey curve):

- Pink region (α) = Type I error when we wrongly reject

Under H_A (blue curve):

- Yellow (β) = Type II error - we miss the real effect
- Green (1-β) = **Power** - we correctly detect the effect!



The key insight: These are all connected - changing one affects the others!

What increases power?

Power increases when:

1. Larger sample size (n)

- Reduces standard error
- Narrows sampling distributions
- Makes distributions more separated

2. Larger effect size

- Greater true difference
- More separation between null and alternative
- Easier to distinguish groups

3. Less variability

- Smaller population SD (σ)
- Tighter distributions
- Less overlap

4. Higher significance level

- Larger α (but increases Type I error!)
- Usually not recommended
- Trade-off between errors

Practical implications:

- We usually have most control over sample size
- Can sometimes reduce variability through better measurement
- Effect size is a property of reality — we estimate it, we don't choose it
- Significance level is conventionally fixed at 0.05

Part 3: Calculating Power and Sample Size

Why do power calculations?

Before collecting data (prospective):

- Determine required sample size for adequate power
- Justify sample size in grant proposals/protocols
- Avoid underpowered studies that waste resources
- Ensure ethical use of participants

After collecting data (post-hoc):

- Understand power of completed study
- Plan follow-up studies

> ⚠ **Important Note**
>
> Post-hoc power calculations for non-significant results can be misleading! Better to report confidence intervals showing uncertainty.

Cohen's d: Standardized effect size

Problem: Effect sizes are in original units (e.g., taps/min, mmHg, °F)

- Hard to compare across studies
- Can't have general guidelines

Solution: Cohen's *d* standardizes the effect size

Cohen's d: Standardized effect size

Cohen's d Formulas

One-sample test (or paired):

$$d = \frac{\bar{x} - \mu_0}{s} \quad \text{or} \quad d = \frac{\bar{x}_d - \delta_0}{s_d}$$

Two-sample test:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}}$$

where $s_{\text{pooled}} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

For two-sample d, we often use pooled SD when group variances are similar; there are variants when they are not.

Interpreting Cohen's d

Cohen's guidelines for effect size:

Effect Size	d value	Interpretation
Small	0.2	Difficult to detect, subtle effect
Medium	0.5	Moderate effect, visible to careful observer
Large	0.8	Large effect, obvious to casual observer

Example interpretations:

- $d = 0.2$: Treatment shifts mean by 0.2 standard deviations
- $d = 0.5$: Treatment shifts mean by half a standard deviation
- $d = 0.8$: Treatment shifts mean by 0.8 standard deviations

Important: These are just guidelines! What's "small" or "large" depends on context

- In medicine: Small effects can be clinically important
- In psychology: Large effects might indicate measurement problems

The pwr package in R

We'll use the `pwr` package for power calculations

```r
library(pwr)
```

Key function: `pwr.t.test()`

- Works for one-sample, two-sample, and paired t-tests
- Specify all parameters except one
- Returns the missing parameter

Function structure:

```r
pwr.t.test(n = NULL,                    # Sample size per group
           d = NULL,                    # Cohen's d effect size
           sig.level = 0.05,            # Significance level (α)
           power = NULL,                # Power (1-β)
           type = "two.sample",         # or "one.sample", "paired"
           alternative = "two.sided")   # or "less", "greater"
```

Leave out the parameter you want to calculate!

Example 1: One-sample test - Finding sample size

Scenario: Body temperature study

- We believe true mean is 98.25°F (vs. claimed 98.6°F)
- Pilot data suggests SD ≈ 0.73°F
- Want 80% power with $\alpha = 0.05$
- **Question:** How many people do we need?

Step 1: Calculate Cohen's d

```
# Effect size
mu0 <- 98.6          # Null value
mu_true <- 98.25     # What we believe is true
s <- 0.73            # Standard deviation from pilot data

d <- (mu_true - mu0) / s
d
```

```
[1] -0.4794521
```

Interpretation: The true mean differs from null by 0.48 standard deviations

Example 1: One-sample test - Finding sample size (cont.)

Step 2: Use `pwr.t.test()` to find required sample size

```r
# Specify all parameters except for sample size, n

result <- pwr.t.test(
  d = (98.6 - 98.25) / 0.73,   # Cohen's d
  sig.level = 0.05,            # α = 0.05
  power = 0.80,                # Want 80% power
  type = "one.sample",         # One-sample test
  alternative = "two.sided"    # Two-sided test
)

result
```

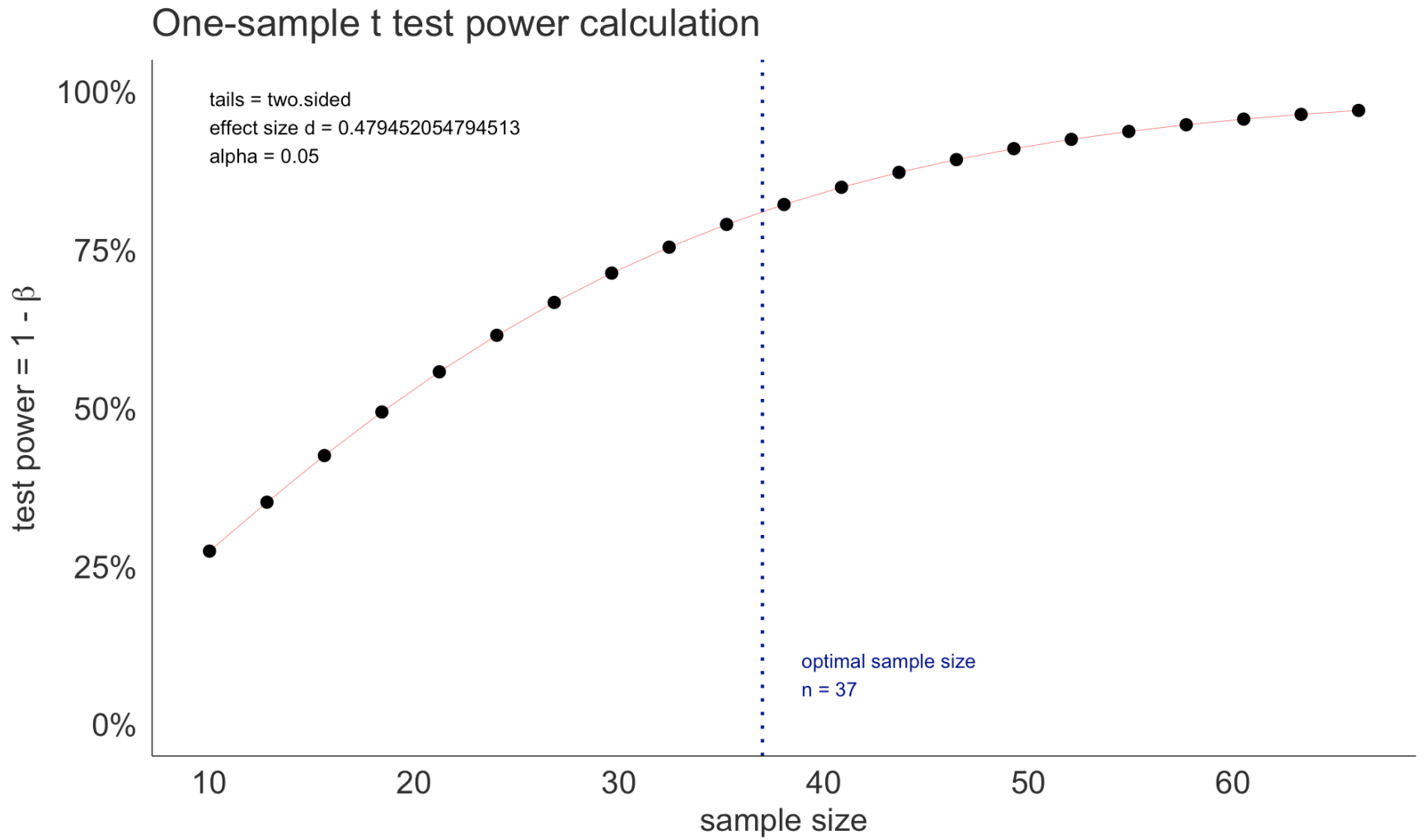
```
     One-sample t test power calculation

              n = 36.11196
              d = 0.4794521
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Conclusion: We need **37 participants** to have 80% power to detect this difference at $\alpha = 0.05$

Visualizing the power curve

```
1 plot(result)
```



The curve shows: As sample size increases, power increases (holding other factors constant)

Example 2: One-sample test - Calculating power

Scenario: Same body temperature study, but the sample size is fixed

- n = 130 participants
- Effect size: d = 0.479
- $\alpha = 0.05$
- **Question:** What power do we have given the sample size we have?

```r
# Specify all parameters except for power

result_power <- pwr.t.test(
  n = 130,                      # Sample size we have
  d = (98.6 - 98.25) / 0.73,    # Cohen's d
  sig.level = 0.05,             # α = 0.05
  type = "one.sample",
  alternative = "two.sided"
)
```

Example 2: One-sample test - Calculating power (cont.)

```
1 result_power
```

```
     One-sample t test power calculation 

              n = 130
              d = 0.4794521
      sig.level = 0.05
          power = 0.9997354
    alternative = two.sided
```

Conclusion: With n=130, we had **100%** power! (Very high - almost certain to detect the effect if it exists)

Example 3: Two-sample test - Finding sample size

Scenario: Caffeine tapping study

- Want to detect 2 taps/min difference between groups
- Expect SD ≈ 2.6 taps/min in each group
- Want 80% power with $\alpha = 0.05$
- **Question:** How many participants per group do we need?

Step 1: Calculate Cohen's d

```
diff <- 2           # Difference we want to detect
sd_pooled <- 2.6    # Expected SD in each group

d <- diff / sd_pooled
d
```

```
[1] 0.7692308
```

Example 3: Two-sample test - Finding sample size (cont.)

Step 2: Calculate required sample size

```r
# Specify all parameters except for sample size, n

result_caff <- pwr.t.test(
  d = d,
  sig.level = 0.05,
  power = 0.80,
  type = "two.sample",          # Two independent groups
  alternative = "two.sided"
)

result_caff
```

```
     Two-sample t test power calculation

              n = 27.52331
              d = 0.7692308
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Conclusion: Need **28 participants per group** (total n = 56)

Example 4: Two-sample test - Calculating power

Scenario: We have 35 subjects per group in the Caffeine study

- Effect size: $d = 0.77$
  - difference of 2 points between the two groups
  - assuming $SD = 2.6$ in both groups
- $\alpha = 0.05$
- **Question:** What power do we have to detect this difference?

```r
# Specify all parameters except for power

result_caff_power <- pwr.t.test(
  n = 35,       # Sample size per group
  d = 2 / 2.6,
  sig.level = 0.05,
  type = "two.sample",
  alternative = "two.sided"
)
```

Example 4: Two-sample test - Calculating power (cont.)

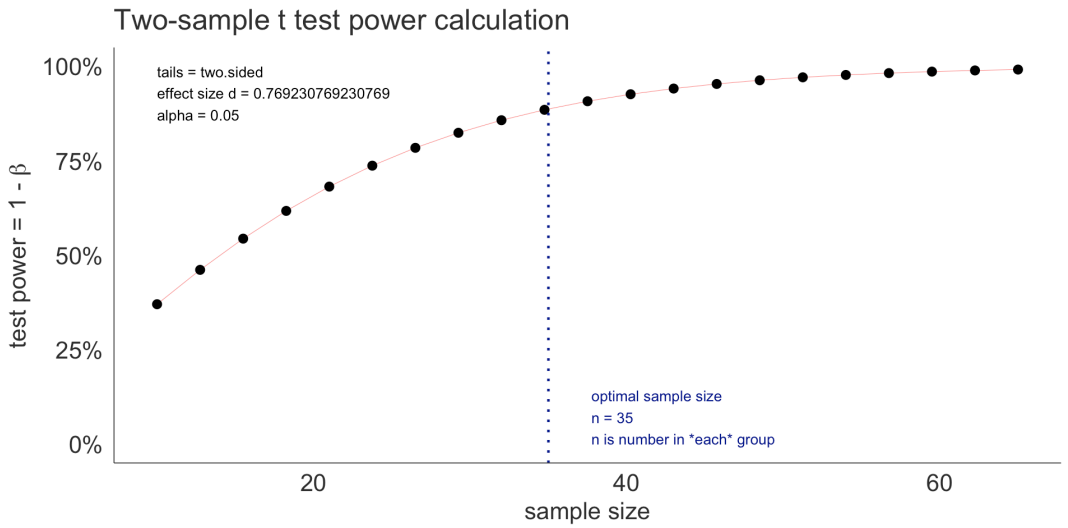
```
1 result_caff_power
```

```
     Two-sample t test power calculation

              n = 35
              d = 0.7692308
      sig.level = 0.05
          power = 0.8872805
    alternative = two.sided

NOTE: n is number in *each* group
```

```
1 plot(result_caff_power)
```



Conclusion: With n=35 per group, we have **88.7%** power (excellent!)

One-sided vs two-sided tests and power

Two-sided test:

$H_A : \mu \neq \mu_0$ (most common)

```r
# Two-sided test
two_sided <- pwr.t.test(
  d = 0.5,
  sig.level = 0.05,
  power = 0.80,
  type = "two.sample",
  alternative = "two.sided"
)

two_sided
```

```
     Two-sample t test power calculation

              n = 63.76561
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

One-sided test:

$H_A : \mu > \mu_0$ or $H_A : \mu < \mu_0$

```r
# One-sided test
one_sided <- pwr.t.test(
  d = 0.5,
  sig.level = 0.05,
  power = 0.80,
  type = "two.sample",
  alternative = "greater"
)

one_sided
```

```
     Two-sample t test power calculation

              n = 50.1508
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = greater

NOTE: n is number in *each* group
```

One-sided vs two-sided tests and power (cont.)

Key point: One-sided tests require smaller sample sizes BUT:

- Only justified when direction is known *a priori*
- Can't detect effects in "wrong" direction
- Generally not recommended unless strong scientific justification
- Reviewers/collaborators often require lower α (e.g., 0.025) for one-sided tests

Part 4: Study Design Applications

The power-sample size trade-off

Increasing sample size:

✓ Increases power

✓ More likely to detect real effects

✓ More precise estimates

✗ More expensive

✗ Takes longer

✗ May not be feasible

Practical constraints:

- Budget limitations
- Time constraints
- Available participants
- Ethical considerations
- Feasibility

The balance: Choose smallest n that gives adequate power (usually 80-90%)

Common mistakes in power analysis

⚠ Mistake 1: Post-hoc power for non-significant results

Don't do this: "Our result was non-significant (p=0.12). Let me calculate power..."

Why it's wrong: Post-hoc power is mathematically tied to your p-value — it simply restates your result in different units and adds no new information

Do instead: Report confidence interval showing uncertainty

⚠ Mistake 2: Using observed effect size for post-hoc power

Don't do this: Calculate power using your observed effect size to evaluate your own completed study

Why it's wrong: Post-hoc power will almost always be low when $p > 0.05$ — it tells you nothing beyond what the p-value already told you

Note: Your observed effect *can* inform planning a future study, but treat it as a noisy estimate and consider a range of plausible values

Do instead: Use effect size from pilot data, literature, or smallest clinically meaningful effect

Common mistakes continued

⚠ Mistake 3: Ignoring practical significance

Don't do this: Design study to detect any statistically significant difference

Why it's wrong: Tiny, clinically meaningless effects can be significant with large n

Do instead: Base power on *clinically/scientifically meaningful* effect size

⚠ Mistake 4: Not considering variability

Don't do this: Assume optimistic (low) SD when planning sample size

Why it's wrong: Real data often more variable → study underpowered

Do instead: Use conservative (higher) SD estimate, or add 10-20% to planned n as buffer

Real-world example: A cautionary tale

Study: New drug to lower blood pressure

- Powered to detect 5 mmHg reduction
- Assumed SD = 10 mmHg (from literature)
- Calculated need for n = 64 per group

- Only recruited n = 50 per group due to budget constraints
- Actual SD in study was 12 mmHg (more variability than expected)

Result:

- Observed reduction: 4 mmHg (close to target!)
- *p*-value = 0.090 (not significant at 0.05 level)
- Actual power was only 38% (not the planned 80%)
- Study deemed “negative” despite clinically meaningful effect

Lessons:

1. Small deviations from plan can substantially reduce power
2. Under-recruitment is very common - build in buffer
3. Conservative SD estimates are wise

When is a study “underpowered”?

Conventionally:

- Power < 50%: Severely underpowered
- Power 50-70%: Underpowered
- Power 70-80%: Marginally adequate
- Power 80-90%: Good
- Power > 90%: Excellent (sometimes wasteful)

Practical considerations:

- **80% power** is standard for many studies
- **90% power** for important decisions (e.g., FDA approval)
- **Lower power acceptable** for:
  - Pilot/feasibility studies
  - Exploratory research
  - Studies where negative result is informative

Tip

Always report your power analysis in papers/grants! Shows thoughtful study design.

Using power analysis in different scenarios

Planning a new study:

1. Review literature for expected effect size
2. Use conservative estimates
3. Choose a target power (usually 80–90%) and calculate required sample size
4. Add 10-20% buffer for dropout
5. Check if feasible

Evaluating a completed study:

1. Focus on effect estimates and confidence intervals
2. If non-significant: Is CI compatible with meaningful effect?
3. Don't calculate "achieved power" (mathematically determined by the *p*-value)
4. Use observed effect size & SD to plan follow-up studies

Reviewing others' work:

1. Check if power analysis reported
2. Evaluate if assumptions reasonable
3. Consider if study adequately powered
4. Be skeptical of underpowered studies

Grant writing:

1. Justify sample size with power analysis
2. Show you calculated it prospectively
3. Document all assumptions
4. Include sensitivity analyses

Wrap-up and Key Takeaways

Summary: The big picture

Four quantities in equilibrium:

1. **Significance level** (α) - usually fixed at 0.05
2. **Effect size** - a property of reality (we estimate it, we do not choose it)
3. **Sample size** (n) - what we typically calculate
4. **Power** ($1 - \beta$) - probability of detecting real effect

Key concepts:

- Power = Probability of correctly detecting an effect when it exists
- Type I error (α) = False positive
- Type II error (β) = False negative
- Cohen's *d* = Standardized effect size
- Typical target: 80-90% power

Key R functions

Main function: `pwr.t.test()`

```r
library(pwr)

# Calculate sample size (leave n = NULL)
pwr.t.test(d = 0.5, sig.level = 0.05, power = 0.80,
           type = "two.sample", alternative = "two.sided")

# Calculate power (leave power = NULL)
pwr.t.test(n = 50, d = 0.5, sig.level = 0.05,
           type = "two.sample", alternative = "two.sided")

# Calculate detectable effect size (leave d = NULL)
pwr.t.test(n = 50, sig.level = 0.05, power = 0.80,
           type = "two.sample", alternative = "two.sided")
```

Types: `"one.sample"`, `"two.sample"`, `"paired"`

Alternatives: `"two.sided"` (most common), `"less"`, `"greater"`

Best practices for power analysis

1. **Plan prospectively** - before collecting data
2. **Use realistic effect sizes** - from literature or pilot data
3. **Be conservative** - overestimate SD, add buffer to sample size
4. **Consider practical significance** - not just statistical
5. **Report your analysis** - document all assumptions
6. **Don't do post-hoc power** for non-significant results
7. **Use confidence intervals** to show uncertainty
8. **Consider feasibility** - balance power with resources

Resources for power and sample size calculations

More software for power and sample size calculations: PASS

- PASS is a very powerful (& expensive) software that does power and sample size calculations for many advanced statistical modeling techniques.
  - Even if you don't have access to PASS, their **documentation** is very good and free online.
  - Documentation includes formulas and references.
  - PASS documentation for powering **means**
    - One mean, paired means, two independent means
- One-sample t-test documentation: **https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/PASS/One-Sample_T-Tests.pdf**

OCTRI-BERD power & sample size presentations

- **Power and Sample Size 101**
  - Presented by Meike Niederhausen; April 13, 2023
  - Slides: http://bit.ly/PSS101-BERD-April2023
  - Recording
- **Power and Sample Size for Clinical Trials: An Introduction**
  - Presented by Yiyi Chen; Feb 18, 2021
  - Slides: http://bit.ly/PSS-ClinicalTrials
  - Recording
- **Planning a Study with Power and Sample Size Considerations in Mind**
  - Presented by David Yanez; May 29, 2019
  - Slides
  - Recording
- **Power and Sample Size Simulations in R**
  - Presented by Robin Baudier; Sept 21, 2023
  - Slides
  - Recording

Additional resources

Good paper

- Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies (Serdar et al.)

Interactive tools:

- Understanding Statistical Power and Significance Testing

Free software:

- Sample size calculators from UCSF
- CRAB (Cancer Research and Biostatistics) Statistical Tools
- G*Power - free, open source power analysis software