KIN 610: Quantitative Methods in Kinesiology

Chapter 10: Hypothesis Testing

Ovande Furtado Jr., PhD.

2026-02-11

1 FYI

This presentation is based on the following books. The references are coming from these books unless otherwise specified.

Main sources:

- Moore, D. S., Notz, W. I., & Fligner, M. (2021). *The basic practice of statistics* (9th ed.). W.H. Freeman.
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.
- Furtado, O., Jr. (2026). *Statistics for movement science: A hands-on guide with SPSS* (1st ed.). https://drfurtado.github.io/sms

ClassShare App

You may be asked in class to go to the ClassShare App to answer questions.

- https://classsharedrfurtado.netlify.app/

SPSS Tutorial

- SPSS Tutorial: Hypothesis Testing

2 Intro Question

- Suppose a coach claims that a new training program improves vertical jump height. You test 30 athletes before and after and find the mean improvement is 2.1 cm. Is this a real effect, or could it be due to chance?

Click to reveal answer

We can't just look at the number — we need a formal procedure to determine whether 2.1 cm is large enough relative to the expected variability to conclude it's a **real effect** rather than **sampling error**. This is what hypothesis testing provides.

- **Hypothesis testing** gives us a formal, objective framework for deciding whether observed differences are real or simply the result of random sampling variability. In this chapter, we'll learn the logic, steps, and interpretation of hypothesis tests.

3 Learning Objectives

By the end of this chapter, you should be able to:

- State the null and alternative hypotheses for a given research question
- Explain the logic of hypothesis testing as an indirect proof
- Define and distinguish between Type I and Type II errors
- Calculate and interpret p-values correctly
- Define statistical power and identify factors that influence it
- Conduct and interpret one-sample, independent, and paired t-tests
- Distinguish between statistical significance and practical significance
- Make correct decisions using hypothesis testing procedures

4 Symbols

Symbol	Name	Pronunciation	Definition
H_0	Null hypothesis	"H naught"	Statement of no effect or no difference
H_1 or H_a	Alternative hypothesis	"H one" or "H sub a"	Statement of an effect or difference
α	Significance level	"alpha"	Probability of Type I error (typically 0.05)
β	Type II error rate	"beta"	Probability of failing to detect a real effect
$1 - \beta$	Statistical power	"one minus beta"	Probability of correctly detecting a real effect
p	P-value	"p value"	Probability of data as extreme as observed, if H_0 is true

Symbol	Name	Pronunciation	Definition
t	t-statistic	"t"	Test statistic for comparing means
df	Degrees of freedom	"d.f."	Number of independent pieces of information
d	Cohen's d	"Cohen's d"	Effect size (standardized mean difference)

5 The Logic of Hypothesis Testing

Hypothesis testing uses **indirect reasoning** — similar to a courtroom trial[1,2].

Courtroom analogy:

Courtroom Trial	Hypothesis Testing
Presumption of innocence	Assume H_0 is true (no effect)
Prosecution presents evidence	Calculate test statistic from data
Jury evaluates evidence	Compare p-value to α
Verdict: "Guilty"	**Reject** H_0 (Significant evidence)
Verdict: "Not Guilty"	**Fail to reject** H_0 (Insufficient evidence)

> **!** Important
>
> "Not guilty" "innocent" — just as "fail to reject H_0" "H_0 is true"

- The courtroom analogy helps students understand the logic: We start by assuming no effect, then see if the data provide enough evidence to change our mind.
- Key point: We never "accept" H — we either reject it or fail to reject it.

6 Null and Alternative Hypotheses

Null hypothesis (H_0): Statement of **no effect**, no difference, or no relationship

- $H_0 : \mu = \mu_0$ (population mean equals specified value)
- $H_0 : \mu_1 = \mu_2$ (two population means are equal)
- $H_0 : \mu_D = 0$ (mean difference equals zero)

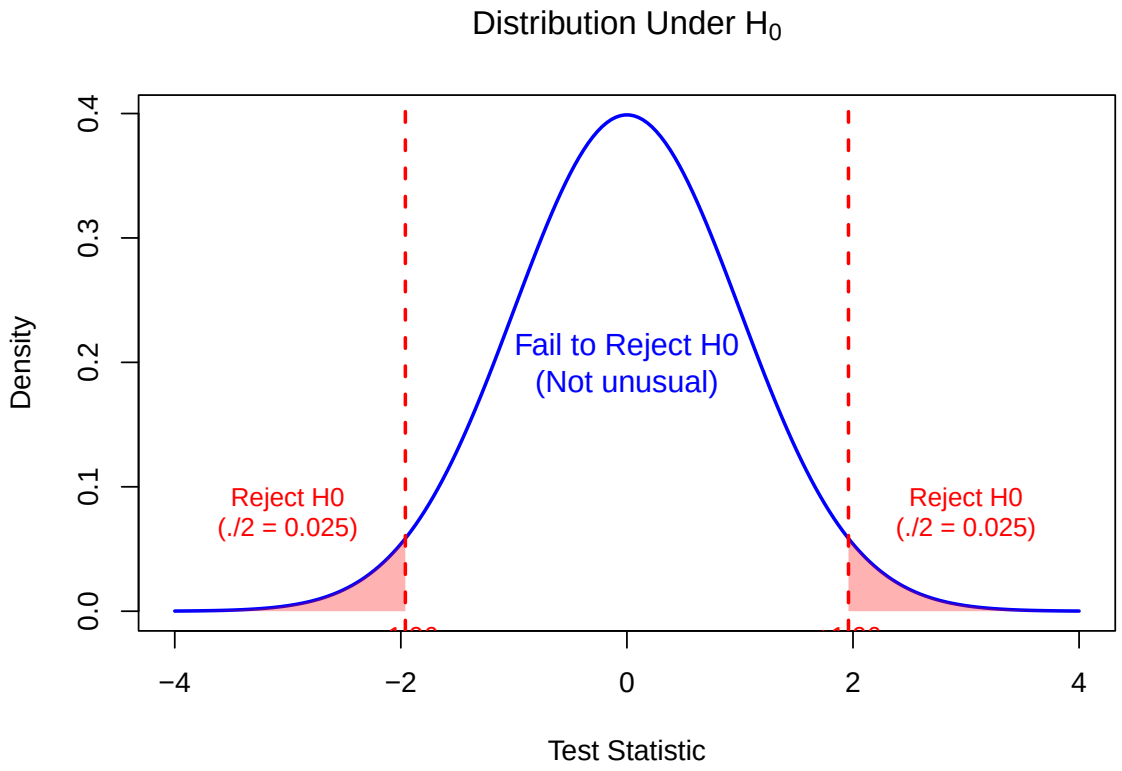

Figure 1: The logic of hypothesis testing: If the observed result is unlikely under H , we reject H

Alternative hypothesis (H_1): Statement that **contradicts** H_0

- **Two-tailed**: $H_1 : \mu \neq \mu_0$ (any direction)
- **One-tailed (right)**: $H_1 : \mu > \mu_0$ (greater than)
- **One-tailed (left)**: $H_1 : \mu < \mu_0$ (less than)

Movement Science examples:

Research Question	H_0	H_1
Does training improve jump height?	$\mu_{\text{post}} = \mu_{\text{pre}}$	$\mu_{\text{post}} > \mu_{\text{pre}}$
Is there a difference in balance between groups?	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$
Does reaction time differ from the norm?	$\mu = 200$ ms	$\mu \neq 200$ ms

Default choice

Use a **two-tailed test** unless you have a strong, pre-specified theoretical reason for a directional hypothesis[1].
Example: If testing a new supplement, use a two-tailed test ($H_1 : \mu \neq 0$) to detect if performance improves OR gets worse. A one-tailed test ($H_1 : \mu > 0$) would ignore the possibility that the supplement actually harms performance.

- Emphasize: H_0 always contains the equality sign (=, ,).
- Common mistake: Students often confuse which is H_0 and which is H_1.
- Rule: H_1 is the research hypothesis — what you're trying to find evidence for.

7 The P-Value

The **p-value** is the probability of observing data as extreme as (or more extreme than) what we actually observed, **assuming H_0 is true**[1,2].

> **In other words:** If there were truly no effect, how surprising would these results be? A small p-value means the results are very surprising (rare), suggesting the "no effect" assumption (H_0) might be wrong.

Interpretation:

- Small p-value ($p < 0.05$): Data are **unlikely under H_0** → Reject H_0
- Large p-value (p 0.05): Data are **compatible with H_0** → Fail to reject H_0

What the p-value is NOT:

The probability that H_0 is true
The probability the results are due to chance
The probability of making an error
The size or importance of the effect

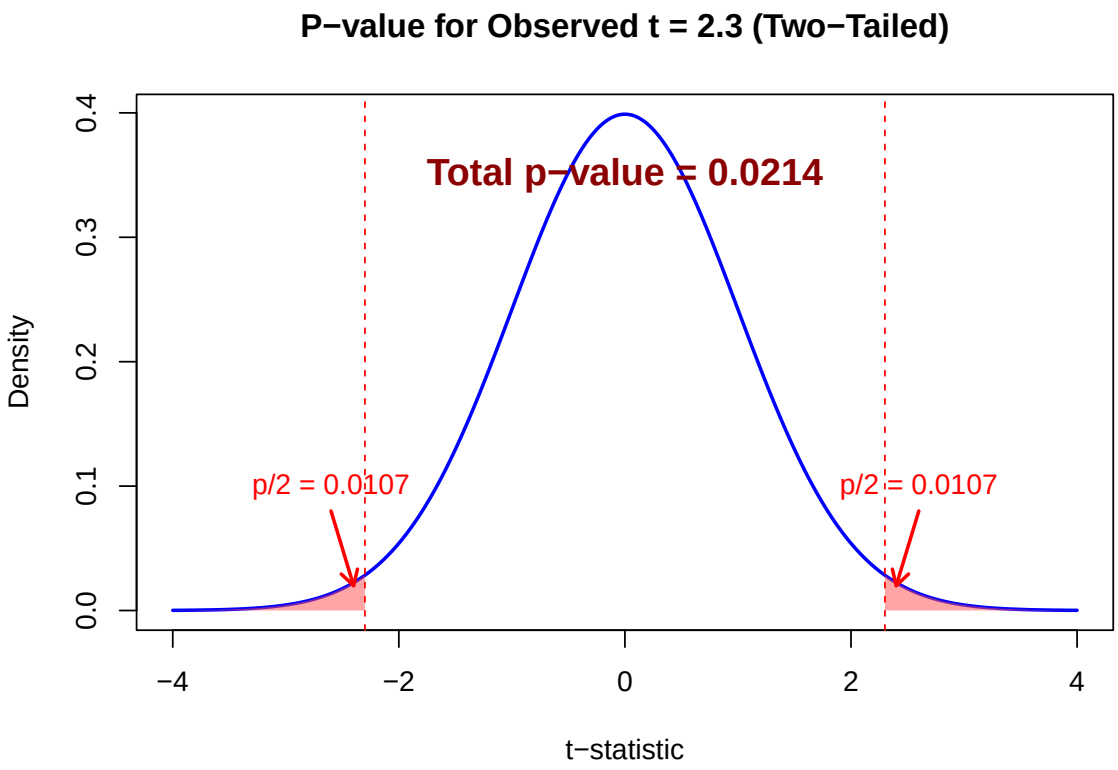

Figure 2: P-value: The probability of observing data this extreme if H_0 is true

> **!** Important
>
> A p-value of `0.021` means: **"If there were truly no effect (H_0 true) and we repeated this study many times, we would obtain results this extreme only 2.1% of the time."** Since $2.1\% < 5\%$ (our threshold), we reject H_0.

- The p-value is the most commonly misinterpreted statistic in research.
- Spend extra time clarifying what p-values are and what they are NOT.
- Key phrase: "The p-value is computed ASSUMING H_0 is true."

8 Type I and Type II Errors

Every hypothesis test can result in one of four outcomes[1,2]:

	H_0 is True	H_0 is False
Reject H_0	Type I Error (α)	Correct Decision (Power)
Fail to reject H_0	Correct Decision	Type II Error (β)

Type I error (False Positive): Concluding there IS an effect when there isn't one

- Probability = α (typically 0.05)
- Example: Concluding a training program works when it doesn't

Type II error (False Negative): Concluding there is NO effect when there is one

- Probability = β
- Example: Concluding a training program doesn't work when it actually does

> ⚠ The trade-off
>
> Reducing Type I error (lowering) **increases** Type II error ( ) — and vice versa. The only way to reduce both is to **increase sample size**.

- Use the fire alarm analogy: Type I = alarm goes off but there's no fire; Type II = there's a fire but no alarm.
- Students should understand the trade-off between the two types of errors.

9 Statistical Power

Statistical power is the probability of correctly rejecting a false null hypothesis — the ability to detect a real effect when one exists[1,2].

$$\text{Power} = 1 - \beta$$

Equation 1: Statistical power formula

Factors affecting power:

1. **Sample size (n)**: Larger → more power

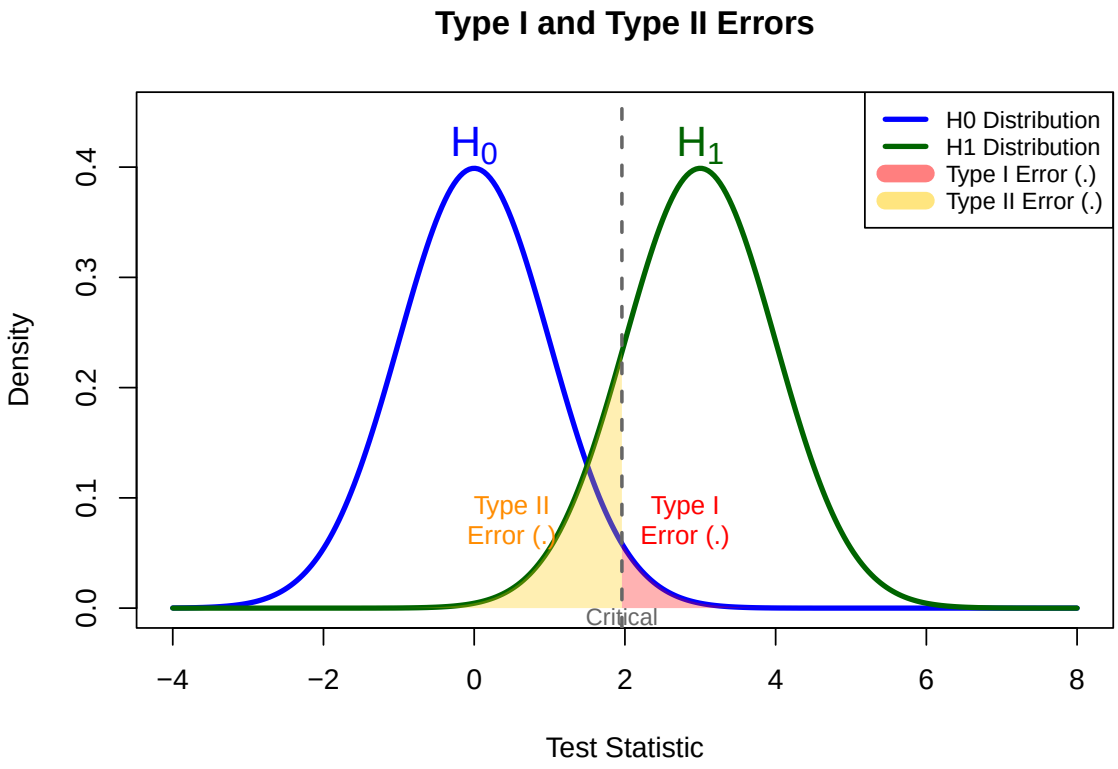

Figure 3: Type I and Type II errors illustrated with two overlapping distributions

2. **Effect size**: Larger effects → more power
3. **Significance level (α)**: Larger → more power (but more Type I errors)
4. **Variability (σ)**: Lower variability → more power

Recommended minimum: Power 0.80 (80%)

This means we want at least an 80% chance of detecting a real effect if one exists.

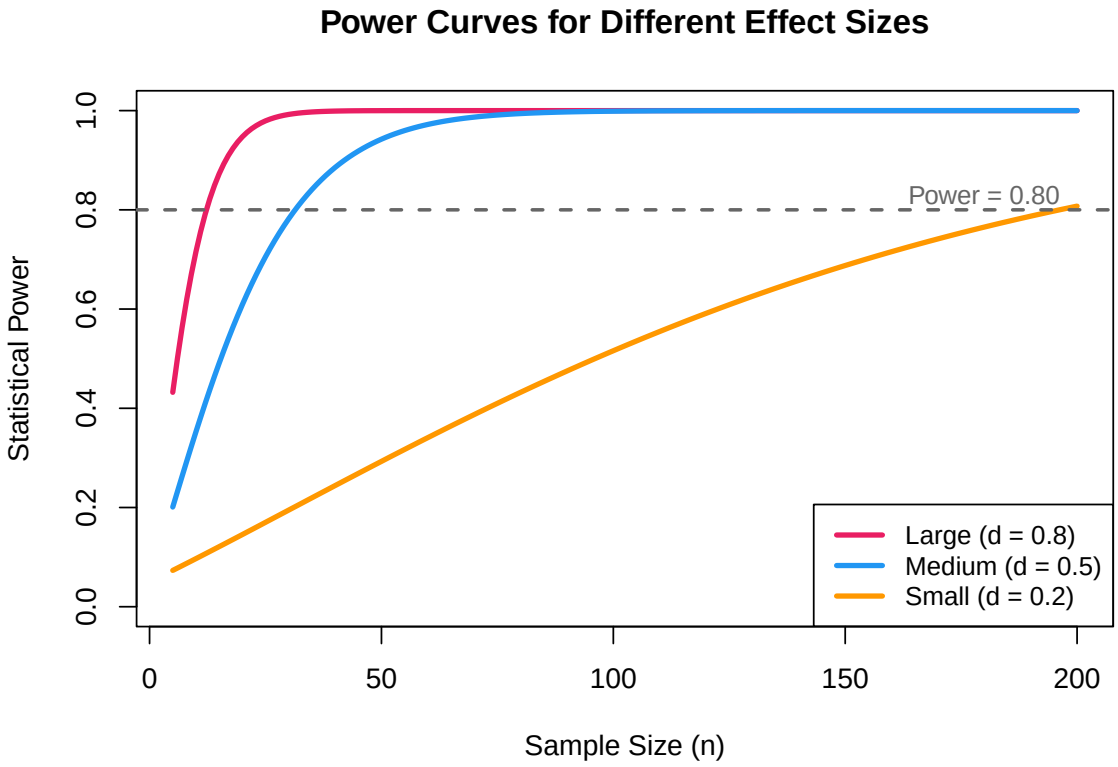

Figure 4: Power increases with sample size

> **!** Important
>
> **Why power matters**: A study with low power has a high probability of missing real effects (Type II error). Always conduct a **power analysis** before collecting data to determine the minimum sample size needed.

- Teaching tip: "If your study has 50% power, it's like flipping a coin to decide if you'll find the effect."

- Practical implication: Many published studies are underpowered, leading to non-replication.

10 Effect Size: Cohen's d

Effect size measures the **magnitude** of the difference — how large the effect is in practical terms, independent of sample size[3].

Formula:

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

Cohen's d formula

Equation 2: **Note:** Use s_{pooled} when comparing two **independent** groups. For paired samples or one-sample tests, use the standard deviation of the difference (s_D) or the sample standard deviation (s), respectively.

Benchmarks[3]:

d Value	Interpretation	Example (Jump Height)
0.2	Small	~1.5 cm improvement
0.5	Medium	~3.7 cm improvement
0.8	Large	~6.0 cm improvement

> **!** Important
>
> **Always report effect sizes alongside p-values.** A result can be statistically significant ($p < .05$) but practically meaningless (tiny d), or statistically non-significant but practically important (large d with small sample).

- Effect sizes tell us HOW BIG the effect is, while p-values tell us whether the effect is detectable.
- In movement science, a 1 cm difference in jump height may be statistically significant with a large enough sample, but practically irrelevant.

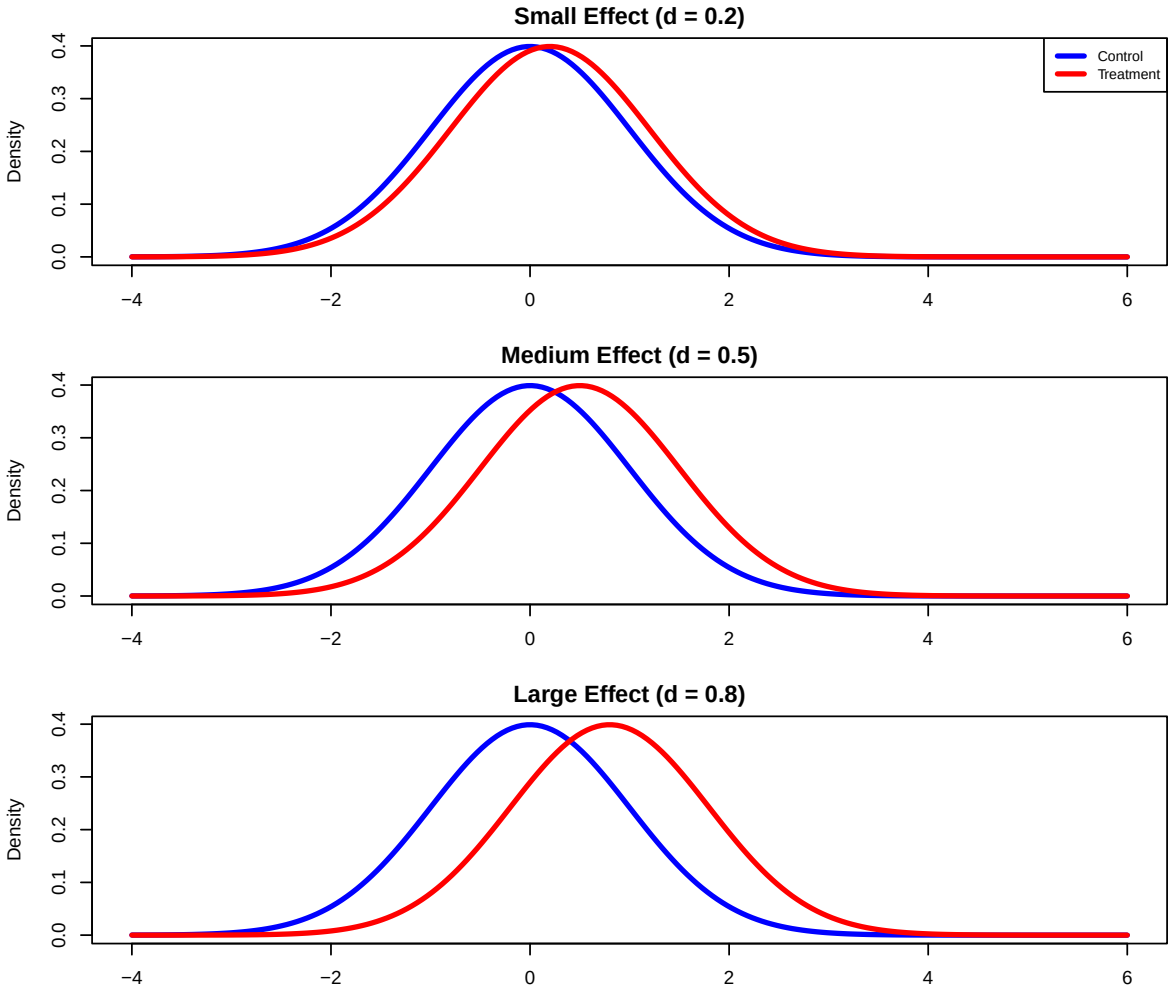

Figure 5: Visual comparison of small, medium, and large effect sizes

11 Statistical vs. Practical Significance

11.1 Statistical Significance

- Based on **p-value** relative to α
- Tells us: "Is the effect **detectable**?"
- Influenced heavily by **sample size**
- Large n can make tiny effects significant

11.2 Practical Significance

- Based on **effect size** and context
- Tells us: "Is the effect **meaningful**?"
- Independent of sample size
- Requires domain knowledge to interpret

Examples in Movement Science:

Scenario	p	d	Interpretation
New shoe improves sprint by 0.001 s	.02	0.05	Stat. sig. but **trivial**
Training improves jump by 8 cm	.08	0.90	Not stat. sig. but **large effect** (underpowered?)
Rehab reduces pain by 2 points	.01	0.65	Stat. sig. AND **meaningful**

⚠ The "significance fallacy"

A "significant" result is not necessarily important, and a "non-significant" result does not mean the effect is zero. Always examine effect sizes and confidence intervals alongside p-values.

- This slide is critical for developing statistical literacy.
- Emphasize: "Significant" in statistics "significant" in everyday language.

12 Steps of Hypothesis Testing: Summary

Follow these five steps for every hypothesis test[1]:

Step	Action	Example
1. **State hypotheses**	Write H_0 and H_1 based on research question	$H_0 : \mu = 50$; $H_1 : \mu \neq 50$
2. **Set criteria**	Choose α (typically 0.05) and test direction	Two-tailed, $\alpha = .05$
3. **Calculate statistic**	Compute test statistic from data	$t = 2.13$
4. **Make decision**	Compare p-value to α	$p = .043 < .05 \rightarrow$ Reject H_0
5. **State conclusion**	Interpret in context, report effect size	"Jump height significantly exceeds 50 cm, $t(24) = 2.13$, $p = .043$, $d = 0.43$"

APA reporting format

"A one-sample t-test revealed that mean vertical jump height ($M = 53.2$, $SD = 7.5$) was significantly greater than the hypothesized mean of 50 cm, $t(24) = 2.13$, $p = .043$, $d = 0.43$."

- Students should memorize this workflow.
- The APA reporting format is essential for writing results sections.

13 Types of t-Tests

The **t-test** is used to compare means when the population standard deviation is unknown[1,2].

13.1 One-Sample t-Test

Purpose: Compare a sample mean to a known or hypothesized value

Hypotheses:

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$

Formula:

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

Example: Is mean jump height different from 50 cm?

13.2 Independent t-Test

Purpose: Compare means between two independent groups

Hypotheses:

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

Formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\text{pooled}}}$$

Example: Do trained vs. untrained athletes differ in balance?

13.3 Paired t-Test

Purpose: Compare means from two related measurements

Hypotheses:

- $H_0 : \mu_D = 0$
- $H_1 : \mu_D \neq 0$

Formula:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}$$

Example: Pre-test vs. post-test strength after training

> Choosing the right t-test
>
> Ask: "Are the two groups **independent** (different people) or **related** (same people measured twice)?" This determines whether to use an independent or paired t-test.

- The t-test replaces the z-test when is unknown (which is almost always the case in real research).
- Students should be able to identify which t-test to use based on the research design.

14 One-Sample t-Test: Worked Example

Research question: Is the mean vertical jump height of kinesiology students different from the national average of 50 cm?

Data: $n = 25$, $\bar{x} = 53.2$ cm, $s = 7.5$ cm

Step 1: State hypotheses

- $H_0 : \mu = 50$ cm
- $H_1 : \mu \neq 50$ cm (two-tailed)

Step 2: Calculate test statistic

$$t = \frac{53.2 - 50}{7.5/\sqrt{25}} = \frac{3.2}{1.5} = 2.13$$

Step 3: Find p-value

- $df = n - 1 = 24$
- From t-table: $p = .043$ (two-tailed)

Step 4: Make decision

- $p = .043 < \alpha = .05 \rightarrow$ **Reject** H_0

Step 5: State conclusion

> "The mean vertical jump height of kinesiology students ($\bar{x} = 53.2$ cm, $s = 7.5$) was significantly greater than the national average of 50 cm, $t(24) = 2.13$, $p = .043$."

Effect size:

$$d = \frac{53.2 - 50}{7.5} = 0.43 \text{ (medium)}$$

> **i** Cohen's d benchmarks
>
> Small = 0.2, Medium = 0.5, Large = 0.8[3]

- Walk through each step carefully — this is the workflow students will follow.
- Emphasize the APA-style conclusion format: what was measured, the statistic, df, and p-value.

15 Common Misconceptions

Misconception 1

"p = 0.03 means there is a 3% chance that H_0 is true."
Correct: p = 0.03 means there is a 3% chance of observing data this extreme **if H_0 were true**.

Misconception 2

"Failing to reject H_0 proves there is no effect."
Correct: Failing to reject means **insufficient evidence** against H_0. The study may lack power.

Misconception 3

"A significant result means the effect is large and important."
Correct: Statistical significance depends on sample size. **Always check effect size.**

Misconception 4

"p = 0.001 is 'more significant' than p = 0.04."
Correct: Both are significant at = 0.05. A smaller p-value indicates **stronger evidence** against H_0, but does not indicate a larger or more important effect.

Misconception 5

"If I run enough tests, I'll eventually find a significant result."
Correct: Multiple testing inflates Type I error. Running 20 tests at = 0.05 will produce about 1 false positive by chance alone. Use corrections (Bonferroni, FDR).

- These are the most common mistakes students (and researchers!) make.
- Spend time on each one — understanding what p-values are NOT is as important as knowing what they are.

16 Summary: Key Takeaways

1. **Hypothesis testing** is a formal framework for deciding whether observed effects are real or due to chance
2. H_0 (no effect) is the default assumption; H_1 is what we're testing for
3. **P-value** = probability of the data (or more extreme) if H_0 is true — NOT the probability H_0 is true
4. **Type I error** (α) = false positive; **Type II error** (β) = false negative
5. **Power** ($1 - \beta$) should be 0.80; depends on sample size, effect size, and variability
6. **t-tests** compare means: one-sample, independent, or paired
7. **Effect sizes** (Cohen's d) tell us how large the effect is — always report alongside p-values
8. **Statistical significance practical significance**: a result can be statistically significant but trivially small

> **!** Important
>
> The goal of hypothesis testing is to make **informed decisions** under uncertainty. Always report effect sizes, confidence intervals, and p-values together for a complete picture.

17 Practice Questions

1. What is the difference between H_0 and H_1? Which contains the equality sign?
2. If p = 0.08 and = 0.05, what is your decision?
3. Explain Type I and Type II errors using a fire alarm analogy.
4. What is statistical power, and why should it be at least 80%?
5. When would you use a paired t-test instead of an independent t-test?
6. If d = 0.15 and p = 0.001, what would you conclude about the effect?
7. Why can't we "accept" or "prove" the null hypothesis?
8. What factors can increase statistical power without changing ?

18 References

1. Moore, D. S., McCabe, G. P., & Craig, B. A. (2021). *Introduction to the practice of statistics* (10th ed.). W. H. Freeman; Company.
2. Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.
3. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
4. Furtado, O., Jr. (2026). *Statistics for movement science: A hands-on guide with SPSS* (1st ed.). https://drfurtado.github.io/sms/