Paired Samples t-test
(part 1)

Lecture 12

Emma Ning, M.A.

Things Going Well

Lectures & Slides

“Going back on lecture slides in case I forget anything. I use this for hw assignments.”

“I really like the way the lecture notes are **structured**, as they provide examples of how to do specific types of problems and **explain** what everything means.”

“I love the powerpoint slides. The 'simple english' explanations really make things **intuitive**.”

ICAs

“I think the ICAs are really good at helping **solidify** the material. I like how it relates to the in-class lecture that day so that there **isn't a delay** between learning the material and trying to apply it.”

“I like getting examples from the professor and then **working as a team** to solve a problem with our tables.”

Things Going Well

Homework

“The homework assignments allow me to apply what we've learned in class and actually **practice** it.”

“I also like that the homework is based on the class material questions (like ICAs) as it helps me really understand how to **work through different types of problems** that might pop up on exams.”

Labs

“**Working with others** during the lab helped me talk through to the right answers.”

“I also like the lab sections since we have the opportunity to review questions (one by one) and it **doesn't feel like I'm being rushed**.”

Constructive Feedback

"I feel that I need to do **more practice** problems, reflecting back on exam 1. I ran out of time so I feel that I need to be more efficient and get more concise with my explanations by going through all the concepts and focus on **understanding** the main ideas and topics."

"For now i think i have a much better understanding of the material. But in the future office hours might be a big help for me.."

Yes! Come to office hours. You don't need to have a specific question prepared — sometimes it's just helpful to talk things through. If you've noticed that ICAs or labs are taking you a long time, it's a great time to stop by and make sure you're on the right track before things start to pile up.

Constructive Feedback

"A little more practice with the **interpretations** and **notation** in the tests."

I will try to be more consistent with statistical notations and be mindful of explaining them. If it gets confusing during lecture, please stop me and ask.

"We've gone over the APA format used to summarize certain data and maybe it'd help to see **more examples of that in actual research articles**. What I mean is maybe providing snippets of research articles and have us **interpret specific results focusing on the concepts we just learned**."

I think this is a really great idea! I am starting that today and will try my best to incorporate for every lecture/topic.

Using AI

“Yes, I have had AI explain concepts or problems I don't understand. I found this helpful as it is very **concise** and has good explanations.”

“I find it helpful in explaining concepts I'm stuck on in a **simpler** way. Or sometimes I use it to **elaborate** on steps for a problem I may have not fully understood.”

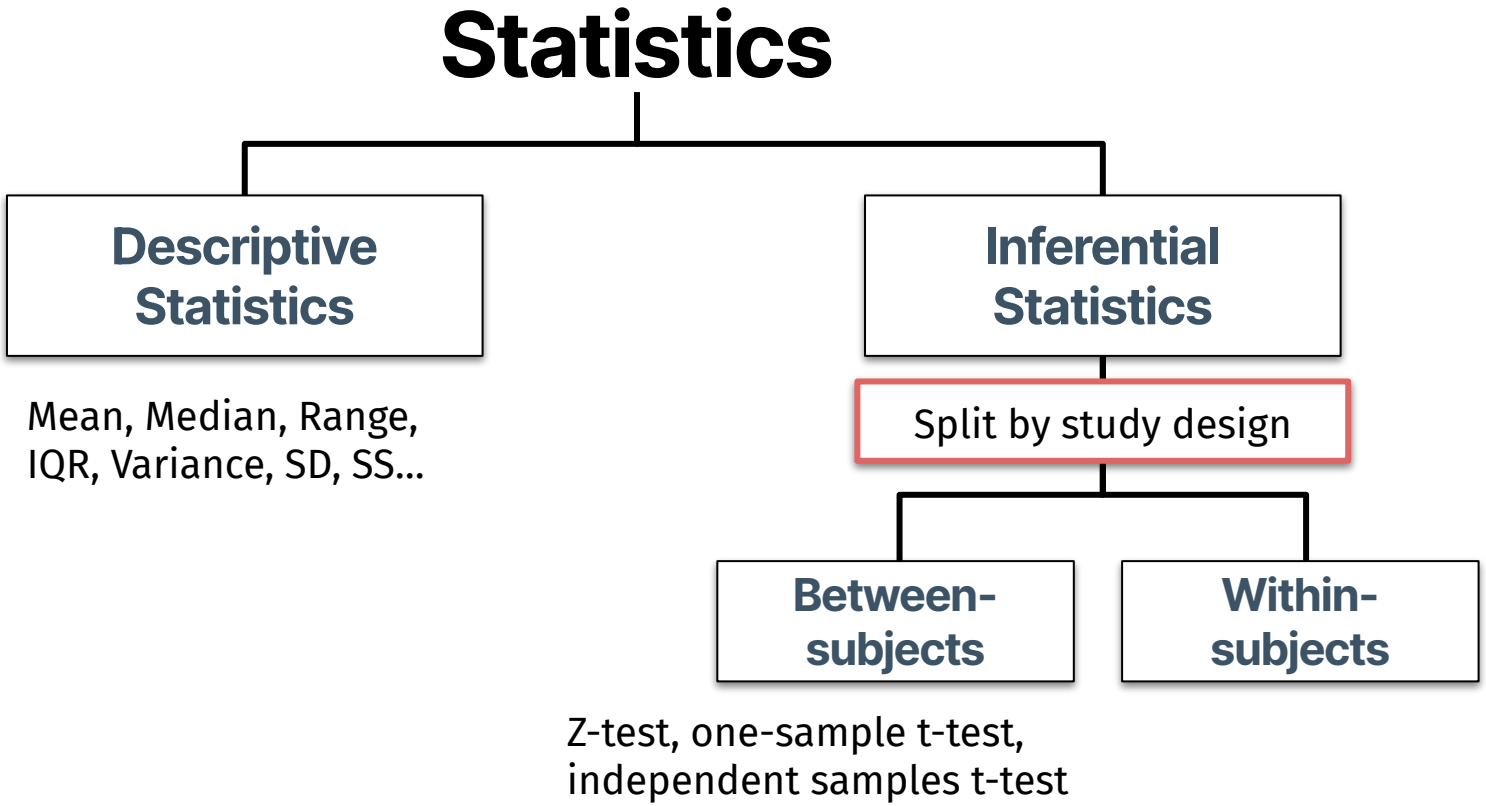
“I’ve used it to help me figure out the **correct structure** for interpretations.”

“No I don’t use AI but I do **google** things when I forget thing such as what’s the difference between t crit or t stats. Things like that that I might forget the reasoning for them or what they are.”

Thank you for your feedback!

It made my day reading them. I’m glad I can make an impact on your learning, and thank you for teaching me how to teach.

What we have learned so far



TODAY'S PLAN

01

Within-Subjects Design & The "Paired" Sample

02

Paired Samples t-test

03

Worked Example

04

Wrap-Up

Learning objectives

- Differentiate between a **within-subjects** design from a **between-subjects** design
- Calculate and interpret the **standard error, t statistic, effect size, and 95% confidence interval** for a **paired sample t test**
- **Conduct** a paired sample t test following the steps of NHST and **interpret** the results to draw a conclusion
- **Write and interpret** the results of a paired sample t-test in APA style

Within-Subjects Design & The “Paired” Sample

Recap: Between vs. Within Subjects Design

Let's go back to our example a few classes ago...
Suppose we are a pharma company, and we want to see if our drug improves sleep meaningfully.

We take our sample and randomly assign participants to either a placebo group or a drug group.




Avg hours slept | 6.50 | 8.50

Recap: Between vs. Within Subjects Design

This would work nicely, right?

Let's go back to our example a few classes ago...

Suppose we are a pharma company, and we want to see if our drug

If the drug group sleeps significantly longer than the placebo group, we can say our drug works!

We take our sample and randomly assign participants to either a placebo group or a drug group.

BUT...

What if the drug group happens to be better sleepers to begin with?

Avg hours slept 6.50 8.50

First week with my Garmin. Am I doing it right?

Wellness & Training Metrics / Features

1.3K 242 Share

Join the conversation

Sort by: Best Search Comments

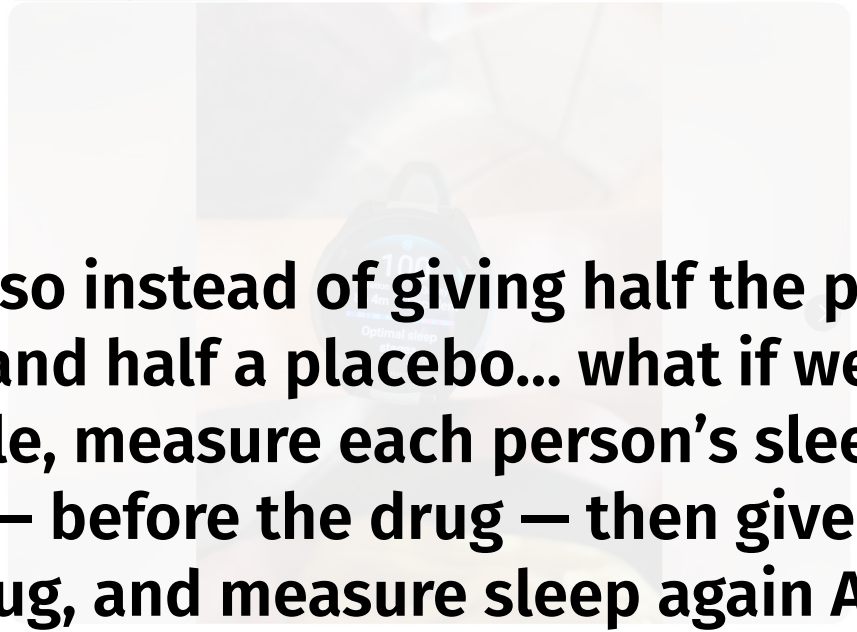
panconpinga_ • 3d ago

congratulations and fuck you

1.8K Reply 2 Share ...

First week with my Garmin. Am I doing it right?

Wellness & Training Metrics / Features



Okay, so instead of giving half the people the drug and half a placebo... what if we take our sample, measure each person's sleep quality FIRST— before the drug — then give them the drug, and measure sleep again AFTER?

1.3K 242 Share

Join the conversation

Sort by: Best Search Comments

panconpinga • 3d ago

congratulations and fuck you

1.8K Reply 2 Share

Recap: Between vs. Within Subjects Design

Between-subjects Design

One group of people experiences one condition, and a **different** group experiences another. Then we compare their outcomes

Examples:
t-test, one-sample t-test, Independent samples t-test

Within-subjects Design

Every person in our sample provides **more than 1 data point** – at different times, in different situations, or after different experiences — so we can track how they **change**.

Examples:
Paired samples t-test

Recap: Between vs. Within Subjects Design

One of the biggest advantages of a within-subjects design is that it's *more powerful*.

We need fewer people to detect an effect if it exists.

This is because each person is compared to themselves. This cancels out individual differences, like whether someone is naturally a good sleeper.

What Paired Samples Are

After - Before

Participant	Sleep *Before*	Sleep *After*	X_D
Josh	5	7	+ 2
Claire	4	9	+ 5
Shruti	5	5	0
Renan	6	5	- 1
Emma	5	9	+ 4
	$M_{pre} = 5$	$M_{post} = 7$	$M_D = 2$

We are examining how the scores **change** for each participant's "pair" of scores.

What Paired Samples Are

Of course, we can measure each person more than 2 times. In that case, we don't call it "paired" anymore.

Instead, we simply refer to that study design as a "within-subjects" design.

But if we measure the same person only twice, we can do a "paired-samples t-test" to see if they changed.

Participant	Sleep *Before*	Sleep *After*	X_D
Josh	5	7	+ 2
Claire	4	9	+ 5
Shruti	5	8	+ 3
Renan	6	5	- 1
Emma	5	9	+ 4
	$M_{pre} = 5$	$M_{post} = 7$	$M_D = 2$

We are examining how the scores in each participant's "pair" of scores

What Paired Samples Are

Participant	Sleep *Before*	Sleep *After*	X_D
Josh	4	8	+ 4
Cla	7	8	+ 1
Shruti	5	5	0
Renan	6	5	- 1
Emma	5	9	+ 4
	$M_{pre} = 5$	$M_{post} = 7$	$M_D = 2$

We are examining how the scores change for each participant's "pair" of scores.

In other words, you do a paired-samples t-test when you have a within-subjects design AND each person is only measured twice. (not more, not less)

A Brief Note: Synonyms

Between-subjects Design

Also known as:

- Independent-measures design
- Independent-measures

Within-subjects Design

Also known as:

- Repeated-measures design
- Repeated-measures
- Dependent samples
- Longitudinal (a lot more data points per person)

THINK - PAIR - SHARE

Which of the following are **within-subjects designs**? Why?

Examining the growth of a child's math abilities from the beginning of the school year to the end.

Examining the difference in participants' cortisol levels when waking up and when going to bed.

Comparing the stress levels between students at UIC, Depaul, and Northwestern.

Comparing student activism levels of undergraduates and graduate students.

Examining whether students read faster in silent environments or with background music. Each participant reads two different passages (one in silence and one with background music).

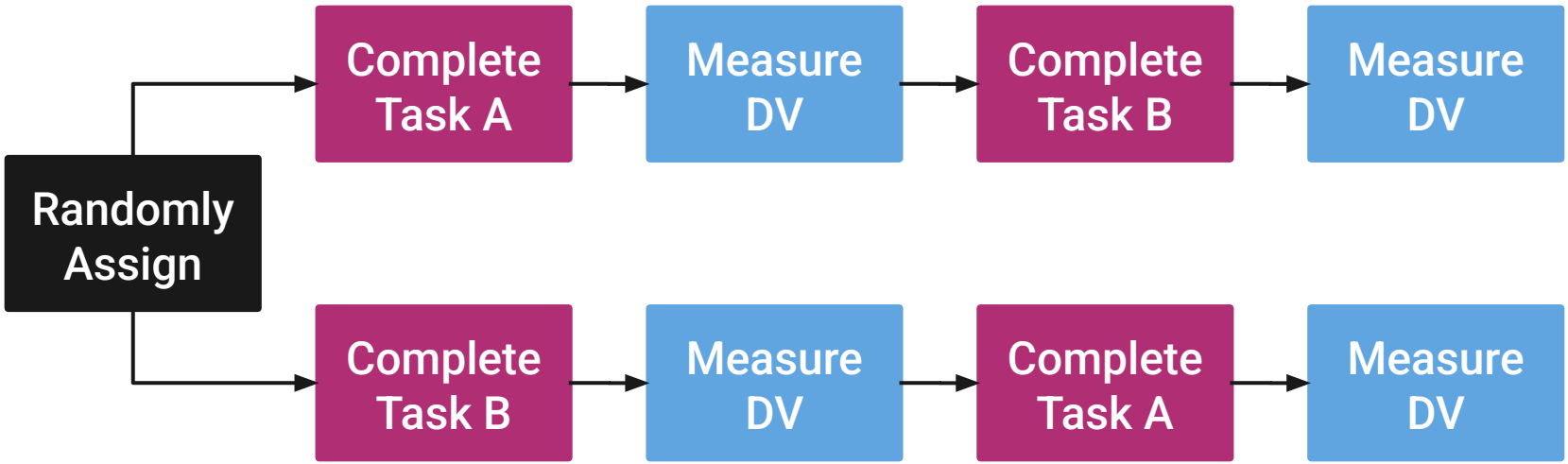
A Disadvantage of Repeated-Measures Designs

Order Effects

Participants might do worse the second time because they are fatigued; or might do better the second time because they have practice.

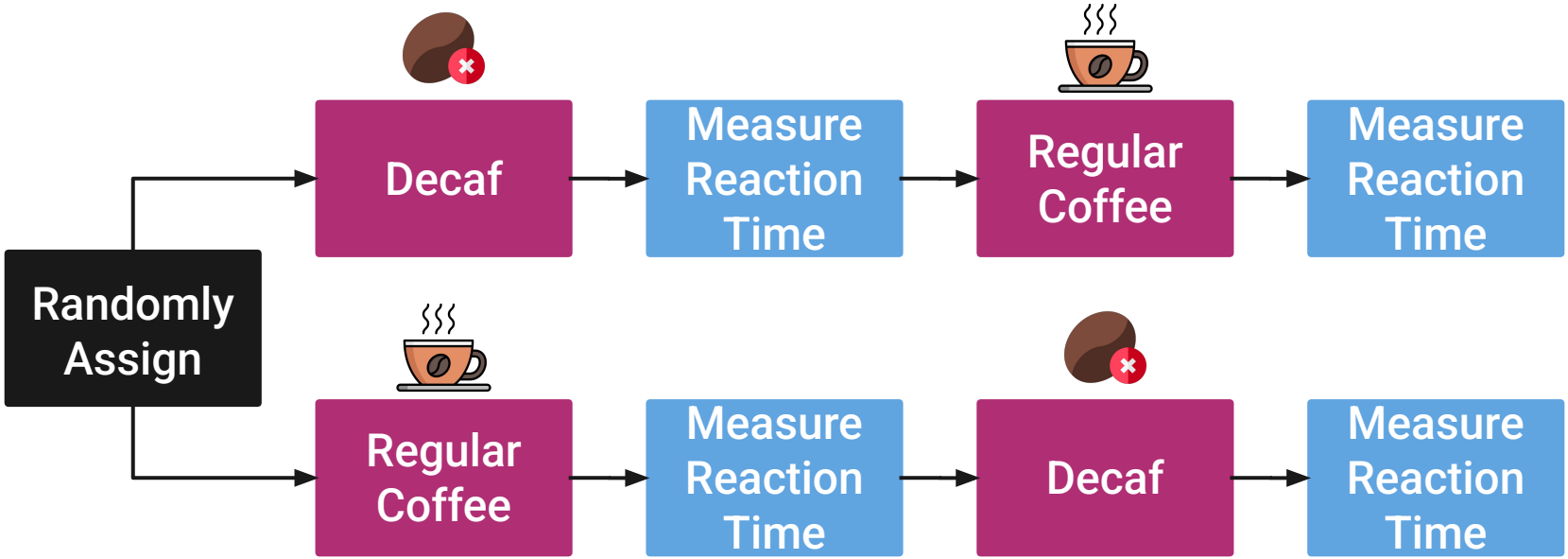
To address order effects, we might choose to counterbalance our tasks.

Counterbalancing Example



When researchers use **counterbalancing**, they present the levels of the independent variable to participants in **different sequences**. Thus, any order effects should cancel each other out.

Does caffeine consumption significantly affect reaction time?



Paired Samples t-test

NHST STEPS

1. We **state our hypotheses** (null and alternative).
2. We find our **cutoff value (t_{crit})** using a table.
3. We **calculate our t-statistic**, **effect size**, and **95% CI**.
4. We **make a decision** (**reject** the null, **fail to reject** the null).
5. We **interpret** and **report** our results (in APA Style).

Not independent anymore

Participant	Sleep *Before*	Sleep *After*	X_D
Josh	5	7	+ 2
Claire	4	9	+ 5
Renan	5	5	0
Shruti	6	5	- 1
Emma	5	9	+ 4
	$M_{pre} = 5$	$M_{post} = 7$	$M_D = 2$

Our Hypothesis - A Slight Change

	Null Hypothesis H_0	**Alternative Hypothesis** H_A
	There is **no difference** in XX between Group A and B	There **is a difference** in XX between Group A and B
Before:	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$
In paired-samples t-test:	$\mu_D = 0$	$\mu_D \neq 0$

Degrees of Freedom

(df)

n - 1

A Familiar Formula

$$t = \frac{\text{difference in means}}{\text{estimated standard error}}$$

$$t = \frac{M_D}{S_{MD}}$$

A Look at “Difference Scores”

Participant	Sleep *Before*	Sleep *After*	X_D	$(X_D - M_D)^2$
Josh	5	7		
Claire	4	9		
Renan	5	5		
Shruti	6	5		
Emma	5	9		

$M_{pre} = 5$ $M_{post} = 7$

A Look at “Difference Scores”

Participant	Sleep *Before*	Sleep *After*	X_D	$(X_D - M_D)^2$
Josh	5	7	+ 2	
Claire	4	9	+ 5	
Renan	5	5	0	
Shruti	6	5	- 1	
Emma	5	9	+ 4	

$M_{pre} = 5$ $M_{post} = 7$ $M_D = 2$

A Look at “Difference Scores”

Participant	Sleep *Before*	Sleep *After*	X_D	$(X_D - M_D)^2$
Josh	5	7	+ 2	$(2 - 2)^2 = 0$
Claire	4	9	+ 5	$(5 - 2)^2 = 9$
Renan	5	5	0	$(0 - 2)^2 = 4$
Shruti	6	5	- 1	$(-1 - 2)^2 = 9$
Emma	5	9	+ 4	$(4 - 2)^2 = 4$
	$M_{pre} = 5$	$M_{post} = 7$	$M_D = 2$	$SS_D = 26$

Standard Deviation & Standard Error

Standard Deviation of Difference Scores

$$S_D = \sqrt{\frac{SS_D}{df}}$$

Standard Error of Difference Scores

$$S_{MD} = \frac{S_D}{\sqrt{n}}$$

For this class, **you will be expected to calculate** S_{MD} when given SS_D for homework and exams.

EFFECT SIZE

tells us the ***magnitude* of the difference** between the means

Cohen's *d*

$$d = \frac{M_D}{s_D}$$

Mean of the difference scores

SD of difference scores

What are the **units**?

CONFIDENCE INTERVAL

To calculate a **95% confidence interval** for a paired samples *t* test, we take the difference in the two means and add/subtract the critical value multiplied by the standard error.

$$CI = M_D \pm t_{crit}(s_{MD})$$

Remember: A confidence interval shows the range of values that are likely to contain *the true population value*, based on your data, with a certain level of confidence (usually 95%).

Worked Example

EXAMPLE

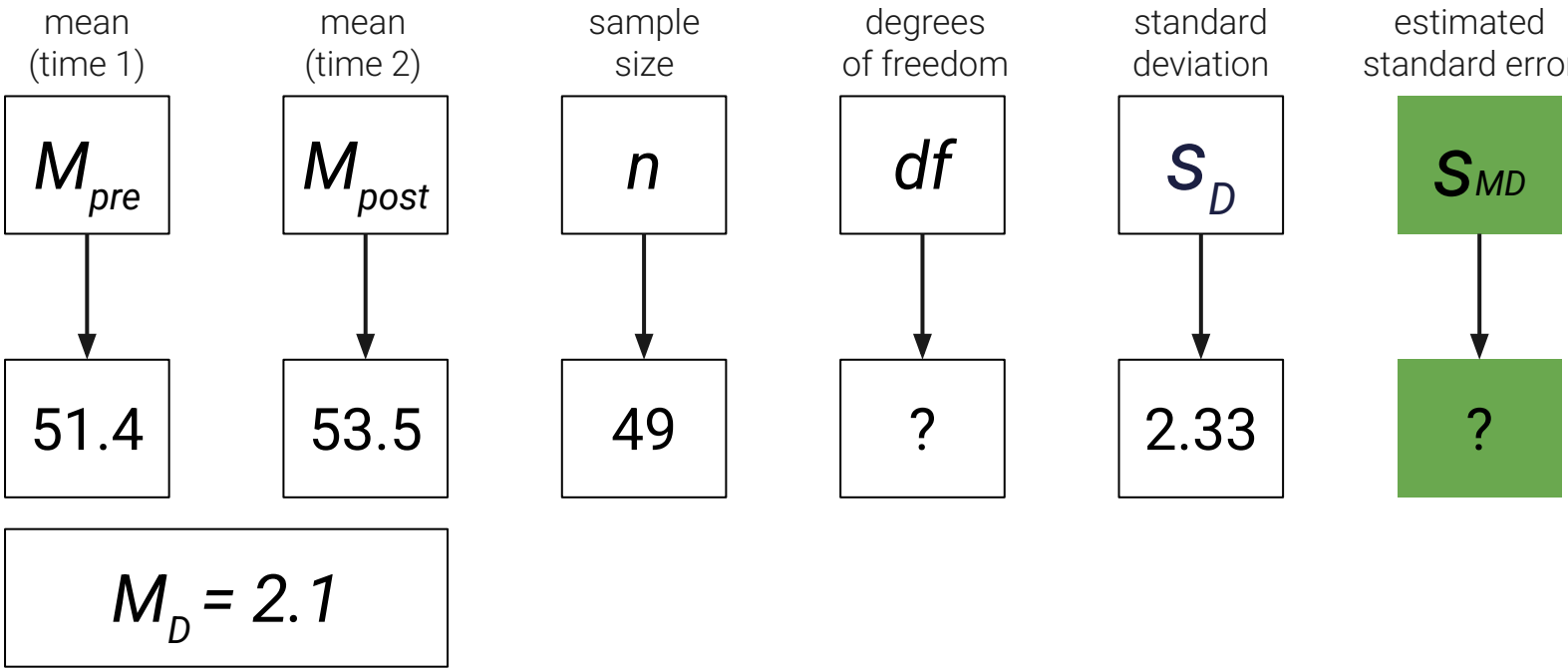
Geier & Morris (2022) conducted a study to investigate if a gratitude intervention (i.e., a weekly gratitude journal) led to higher mental well-being of university students during the COVID-19 pandemic. For the **49** participants in the treatment condition, their pre- and post-intervention scores are below.

Before	**After**
M = 51.4	M = 53.5
S_D = 2.33	

Conduct a paired samples *t* test using the NHST steps (α = 0.05, two-tailed test).

Geier & Morris (2022)

STEP 0: ANNOTATE



NHST STEPS

1. We **state our hypotheses** (null and alternative).
2. We find our **cutoff value (t_{crit})** using a table.
3. We **calculate our t-statistic**, effect size, and 95% CI.
4. We **make a decision** (**reject** the null, **fail to reject** the null).
5. We **interpret** and **report** our results (in APA Style).

STEP 1: State Your Hypotheses

Null Hypothesis

H_0

There is **no difference** in mental well-being before and after the gratitude intervention.

$\mu_D = 0$

Alternative Hypothesis

H_A

There is **a difference** in mental well-being before and after the gratitude intervention.

$\mu_D \neq 0$

NHST STEPS

1. We **state our hypotheses** (null and alternative).
2. We find our **cutoff value (t_{crit})** using a table.
3. We **calculate our t-statistic**, effect size, and 95% CI.
4. We **make a decision** (**reject** the null, **fail to reject** the null).
5. We **interpret** and **report** our results (in APA Style).

STEP 2: Find Your Cutoff (t_{crit})

	Proportion (a) in One tail				
	.05	.025	.01	.005	.0005
	Proportion (a) in Two tails combined				
df	.10	.05	.02	.01	.001
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.093	2.539	2.861	3.883
20	1.725	2.086	2.528	2.845	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.768
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.689
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.660
30	1.697	2.042	2.457	2.750	3.646
40	1.684	2.021	2.423	2.704	3.551
60	1.671	2.000	2.390	2.660	3.460
120	1.658	1.980	2.358	2.617	3.373
∞	1.645	1.960	2.326	2.576	3.290

What we need to know:

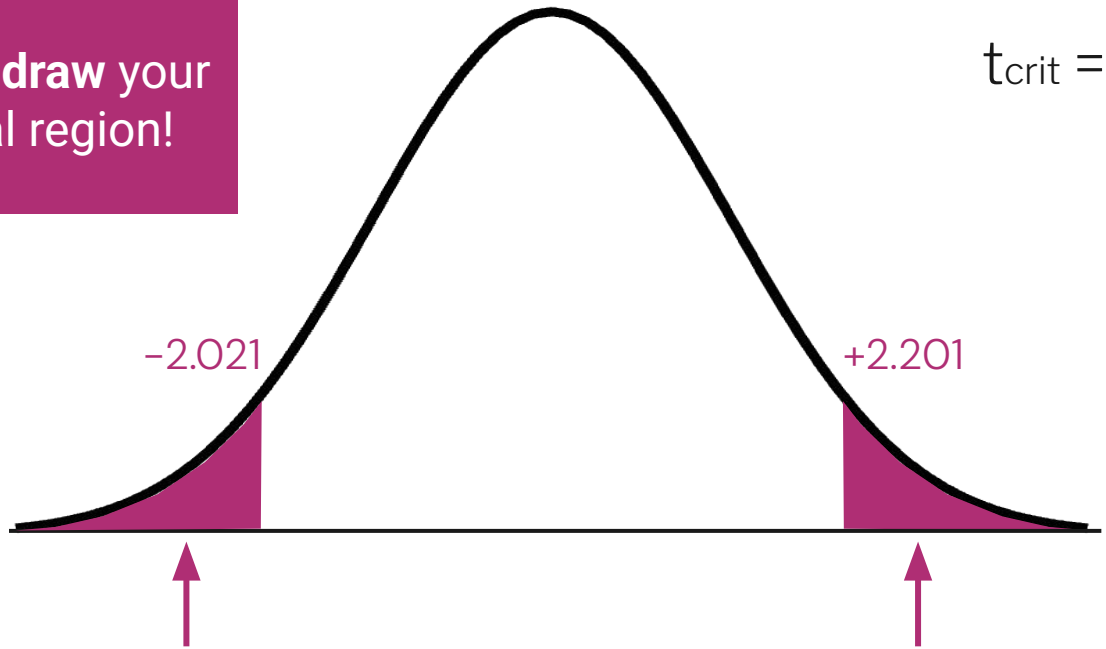
- ***n*** = 49
- ***df*** = 48
- ***α*** = 0.05
- **two**-tailed test

Our t_{crit} is **±2.021**

STEP 2: Find Your Cutoff (t_{crit})

Always **draw** your critical region!

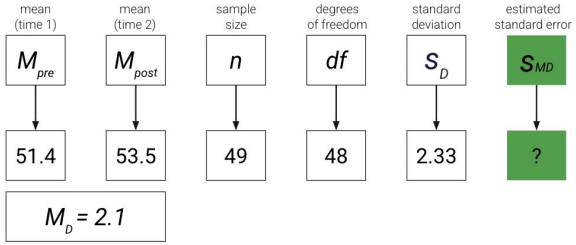
$t_{crit} = \pm 2.021$



NHST STEPS

1. We **state our hypotheses** (null and alternative).
2. We find our **cutoff value (t_{crit})** using a table.
3. We **calculate our t-statistic**, effect size, and 95% CI.
4. We **make a decision** (**reject** the null, **fail to reject** the null).
5. We **interpret** and **report** our results (in APA Style).

STEP 3: Calculate your t, d, and 95% CI.



3A Compute your **standard error**.

$$S_{MD} = \frac{s_D}{\sqrt{n}} = \frac{2.33}{\sqrt{49}} = 0.33$$

3B Compute your **t statistic**.

$$t = \frac{M_D}{s_{MD}} = \frac{2.1}{0.33} = \boxed{6.36}$$

STEP 3: Calculate your t, d, and 95% CI.

3C Compute your ***effect size***.

$$d = \frac{M_D}{s_D} = \frac{2.1}{2.33} = \boxed{0.91}$$

3D Calculate your **95% confidence interval**.

95% CI = $M_D \pm t_{crit}(s_{MD})$
= $2.1 \pm 2.021(0.33)$
= 1.43 & 2.77

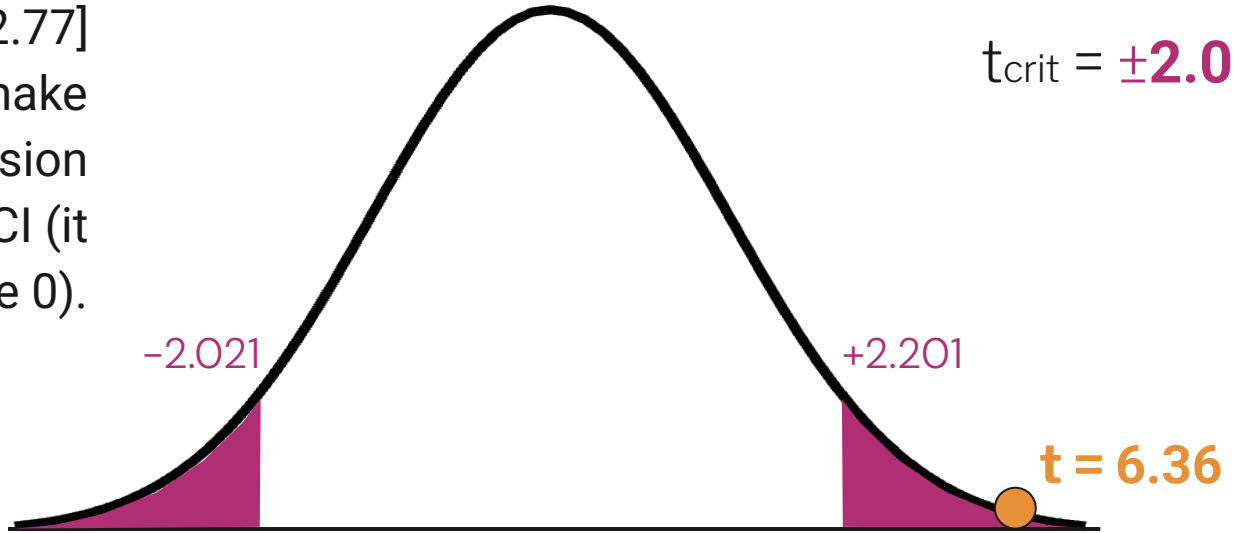
95% CI [1.43, 2.77]

NHST STEPS

1. We **state our hypotheses** (null and alternative).
2. We find our **cutoff value (t_{crit})** using a table.
3. We **calculate our t-statistic**, effect size, and 95% CI.
4. We **make a decision** (**reject** the null, **fail to reject** the null).
5. We **interpret** and **report** our results (in APA Style).

STEP 4: Make Your Decision.

95% CI [1.43, 2.77]
We can also make the same decision based on our CI (it doesn't include 0).



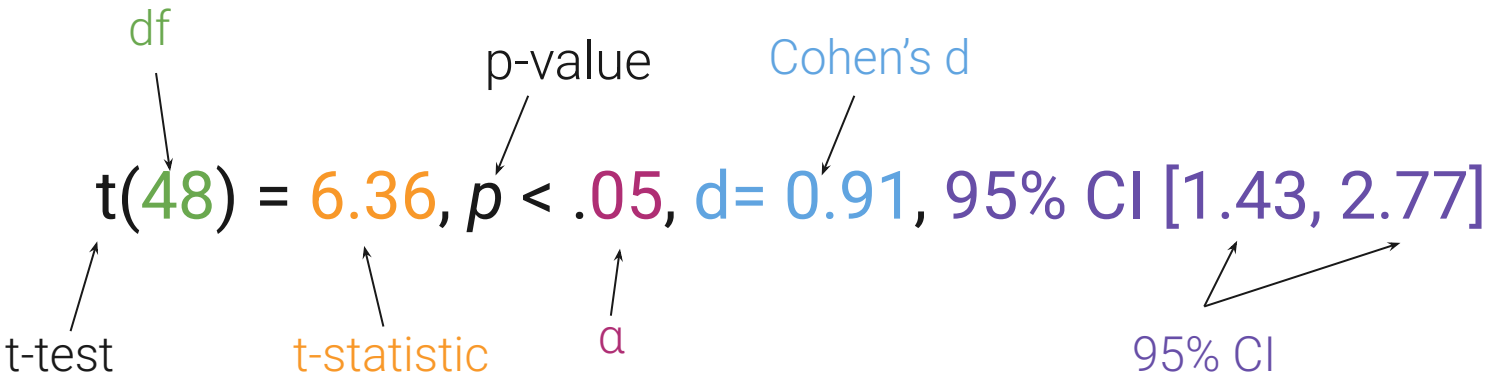
Our t statistic is in the critical region, so we ***reject the null***.

NHST STEPS

1. We **state our hypotheses** (null and alternative).
2. We find our **cutoff value (t_{crit})** using a table.
3. We **calculate our t-statistic**, effect size, and 95% CI.
4. We **make a decision** (**reject** the null, **fail to reject** the null).
5. We **interpret** and **report** our results (in APA Style).

STEP 5: Interpret Results in APA Style.

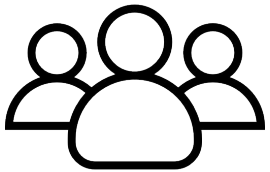
Results from a **paired sample *t* test** indicated that there was a **significant difference (increase)** in **mental well-being** before (*M* = 51.4) and after (*M* = 53.5) the gratitude intervention, ***t*(48) = 6.36, *p* < .05, *d* = 0.91, 95% CI [1.43, 2.77]** with a **large** effect size.



Research example

Nikolin et al., 2019

Examined the effect of transcranial direct current stimulation (tDCS) and cognitive-emotional training (CET)'s effects on treatment-resistant depression.



Received 18 sessions of tDCS and concurrent CET over 6 weeks.

Statistical analyses were performed using R statistical software version 3.5.1 (R Core Team, 2018) and the FieldTrip MATLAB toolbox (Oostenveld et al., 2011). Scripts used for statistical analysis are available at the following link (https://github.com/snikolin/tDCSandCET). Two-tailed paired-samples *t* tests were performed for mood, working memory and EEG neurophysiological outcomes to evaluate changes from BASELINE to POST. Normality of paired differences was tested using the Shapiro–Wilk test. A *p*-value of <.05 was considered non-normally distributed, in which case we additionally performed a non-parametric Wilcoxon signed-rank test. We applied a Bonferroni correction to the six neurophysiological outcomes to reduce the false-positive (Type 1) error associated with multiple comparisons. As such, the threshold for statistical significance for EEG measures was set at $p = .008$ (i.e. 0.05/6). A within-subjects study design of $n = 20$ allows for the detection of moderate–large effect sizes of Cohen's $d \geq 0.65$ using $\alpha = 0.05$ and 80% statistical power. Therefore, lack of statistical significance on a paired-samples *t* test may be indicative of no effect or of insufficient power to detect effects sizes smaller than this threshold. To statistically reject the

Test	Baseline Mean (*SD*)		Post Mean (*SD*)		Post – Baseline Mean (95% CI)		Paired-samples *t* test		
							t_{19}	*p*	Cohen's *d*
Mood outcome									
MADRS	30.0	5.0	19.5	9.5	−10.5	−5.7 to −15.2	−4.61	<.001	1.03
Working memory									
D-prime	1.62	0.76	2.01	0.84	0.39	0.05 to 0.73	2.40	.027	0.54
RT (ms)	796	174	835	235	40	−68 to 148	0.77	.452	0.17
Power spectral density									
Frontal alpha asymmetry	0.09	0.31	0.04	0.25	−0.05	−0.24 to 0.14	−0.55	.591	0.12
Frontal theta	−0.18	1.00	−0.09	1.17	0.09	−0.44 to 0.62	0.35	.729	0.08
Event-related potentials									
P2	1.64	1.27	2.06	1.46	0.42	−0.004 to 0.85	2.07	.052	0.46
P3	1.01	1.65	1.04	1.27	0.03	−0.41 to 0.47	0.31	.888	0.03

I C A 1 2

Let's try to run our own paired-samples t-test:

Test	Baseline Mean (*SD*)		Post Mean (*SD*)		Post – Baseline Mean (95% CI)		Paired-samples *t* test		
							t_{19}	*p*	Cohen's *d*
Mood outcome									
MADRS	30.0	5.0	19.5	9.5	−10.5	−5.7 to −15.2	−4.61	<.001	1.03

$S_D = 9.49, n = 20$

Note: our results will be different from theirs, because they did not tell us what SS_D is, so we cannot do an exact replication of their analysis.

Wrap Up

Key Takeaways

- Paired-samples t-test:
  - What it is
  - Why do we do it, advantage over independent samples-t test
  - Disadvantage
  - Calculation

- Between-subjects vs. within subjects design

- Extracting information from research paper with the knowledge you already have

Extra Credit Opportunity

At the end of the semester, we'll spend two class sessions doing a **hands-on data analysis** workshop, where we'll analyze real datasets together using the tools you've learned this semester.
To make this as relevant and fun as possible, I'd like **you** to help decide what data we'll analyze.

Each student who submits a dataset idea earns **+1 extra credit point** (added to your lowest exam score).
If your dataset is selected for the workshop, you'll receive **+5 extra credit points** added to your lowest exam score.

Due Thursday midnight, October 16th.

More details on Blackboard tab "Syllabus & Other Links"!

Applying to Graduate School Workshop

Tomorrow, October 10th @ 12PM in BSB 1076

Thinking about a graduate degree in psychology but feeling overwhelmed by the application process? Join current UIC Psychology graduate students for a hands-on workshop to learn key strategies for crafting a standout application with step-by-step guidance.

Event Link