Lesson 13: Inference for difference in means from two independent samples

TB sections 5.3

Meike Niederhausen and Nicky Wakim

2024-11-13

Learning Objectives

1. Identify when a research question or dataset requires two independent sample inference.
2. Construct and interpret confidence intervals for difference in means of two independent samples.
3. Run a hypothesis test for two sample independent data and interpret the results.

Where are we?

Data

- Collecting data
- Categorical vs. Numeric
- Summary statistics
- Data visualization

Probability

- Probability rules
- Independence, conditional
- Random variables and probability distributions
- Linear combinations
- Binomial, Normal, and Poisson

Sampling Variability, and Statistical Inference

- Sampling distributions
- Central Limit Theorem
- Confidence Intervals
- Hypothesis tests

Inference for continuous data/outcomes

- One sample t-test
- 3+ independent samples
- Simple linear regression / correlation
- 2 sample tests: paired and independent
- Power and sample size
- Non-parametric tests

Inference for categorical data/outcomes

- One proportion test
- Fisher's exact test
- Non-parametric tests
- Chi-squared test
- 2 proportion test
- Power and sample size

R

- Basics
- Reproducibility
- Quarto
- Packages
- Data visualization
- Data wrangling
- R Projects
- …

Different types of inference based on different data types

Lesson	Section	Population parameter	Symbol (pop)	Point estimate	Symbol (sample)	SE
11	5.1	Pop mean	μ	Sample mean	$\overline{x}$	$\frac{s}{\sqrt{n}}$
12	5.2	Pop mean of paired diff	μ_d or δ	Sample mean of paired diff	$\overline{x}_d$	$\frac{s_d}{\sqrt{n}}$
13	5.3	Diff in pop means	$\mu_1 - \mu_2$	Diff in sample means	$\overline{x}_1 - \overline{x}_2$	????
15	8.1	Pop proportion	p	Sample prop	$\widehat{p}$	
15	8.2	Diff in pop prop's	$p_1 - p_2$	Diff in sample prop's	$\widehat{p}_1 - \widehat{p}_2$	

Learning Objective

→ ① independ, 2 samples

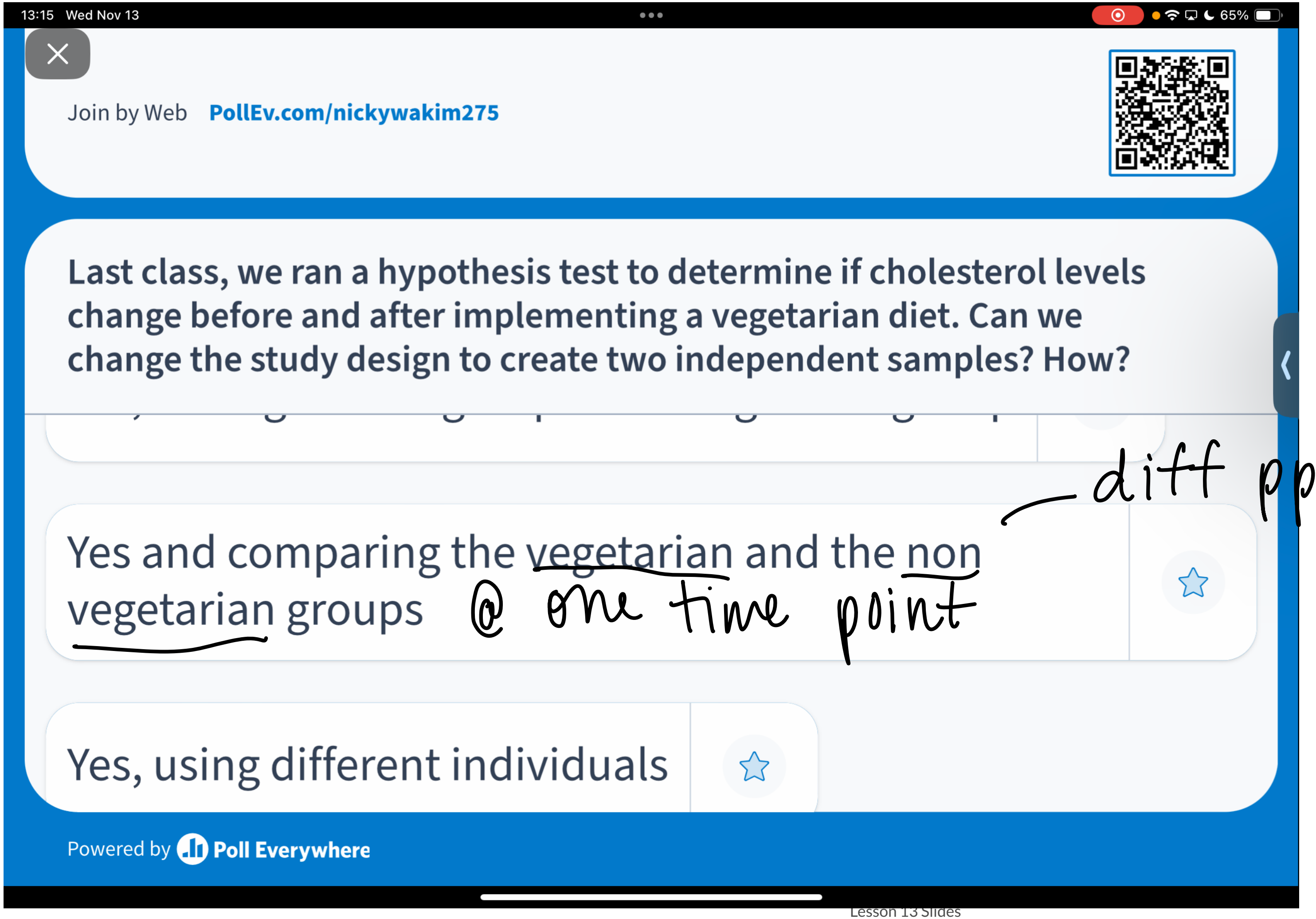
What are data from two independent sample?

- **Two independent samples:** Individuals between and within samples are independent
    - Typically: measure the same outcome for each sample, but typically the two samples differ based on a single variable

- Examples
    - Any study where participants are randomized to a control and treatment group
    - Study with two groups based on their exposure to some condition (can be observational)
    - Book: “Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack?”
    - Book: “Is there evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who do not smoke?”

- Pairing (like comparing before and after) may not be feasible

Poll Everywhere Question 1



PollEv.com/nickywakim275

Last class, we ran a hypothesis test to determine if cholesterol levels change before and after implementing a vegetarian diet. Can we change the study design to create two independent samples? How?

Yes and comparing the vegetarian and the non vegetarian groups @ one time point

diff ppl in each group

Yes, using different individuals

For two independent samples: Population parameters vs. sample statistics

Population parameter

- Population 1 mean: μ_1
- Population 2 mean: μ_2

- Difference in means: $\mu_1 - \mu_2$

- Population 1 standard deviation: σ_1
- Population 2 standard deviation: σ_2

Sample statistic (point estimate)

- Sample 1 mean: $\overline{x}_1$
- Sample 2 mean: $\overline{x}_2$

- Difference in sample means: $\overline{x}_1 - \overline{x}_2$

- Sample 1 standard deviation: s_1
- Sample 2 standard deviation: s_2

Does caffeine increase finger taps/min (on average)?

- Use this example to illustrate how to calculate a confidence interval and perform a hypothesis test for two independent samples

Study Design:[1]

- 70 college students students were trained to tap their fingers at a rapid rate
- Each then drank 2 cups of coffee (double-blind)
  - **Control** group: decaf
  - **Caffeine** group: ~ 200 mg caffeine
- After 2 hours, students were tested.
- **Taps/minute** recorded

Does caffeine increase finger taps/min (on average)?

- Load the data from the csv file `CaffeineTaps`~~`.csv`~~ _n35.csv
- The code below is for when the data file is in a folder called `data` that is in your R project folder (your working directory)

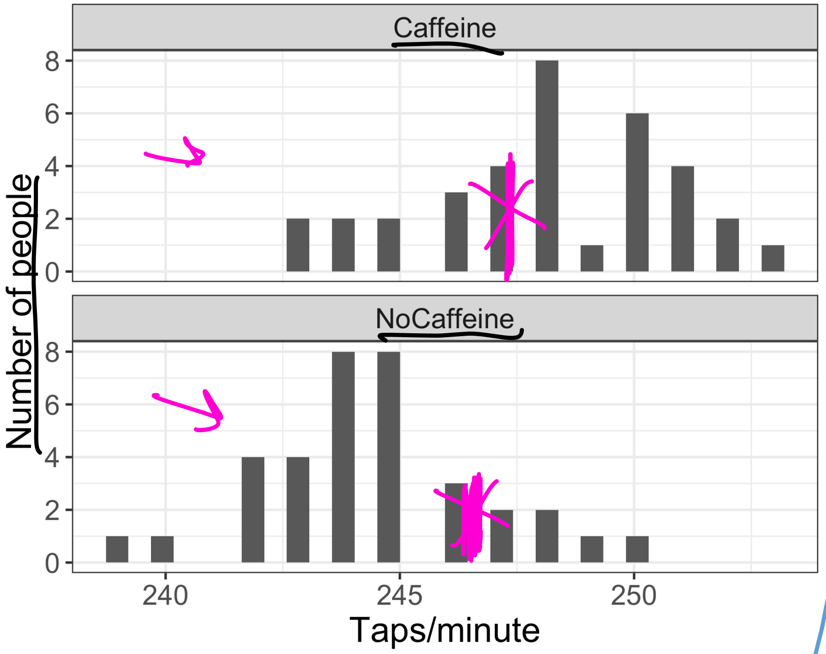
```r
CaffTaps <- read.csv(here::here("data", "CaffeineTaps_n35.csv"))

glimpse(CaffTaps)
```

```
Rows: 70
Columns: 2
$ Taps  <int> 246, 248, 250, 252, 248, 250, 246, 248, 245, 250, 242, 245, 244,…
$ Group <chr> "Caffeine", "Caffeine", "Caffeine", "Caffeine", "Caffeine", "Caf…
```

ind	Taps	Group
1	246	Caffeine
2		
3		
		Non caffeine

EDA: Explore the finger taps data

- Code to make these histograms



- Summary statistics stratified by group

Group	variable	n	mean	sd
Caffeine	Taps	35	248.114	2.621
NoCaffeine	Taps	35	244.514	2.318

Then calculate the difference between the means:

```
diff(sumstats$mean)
```

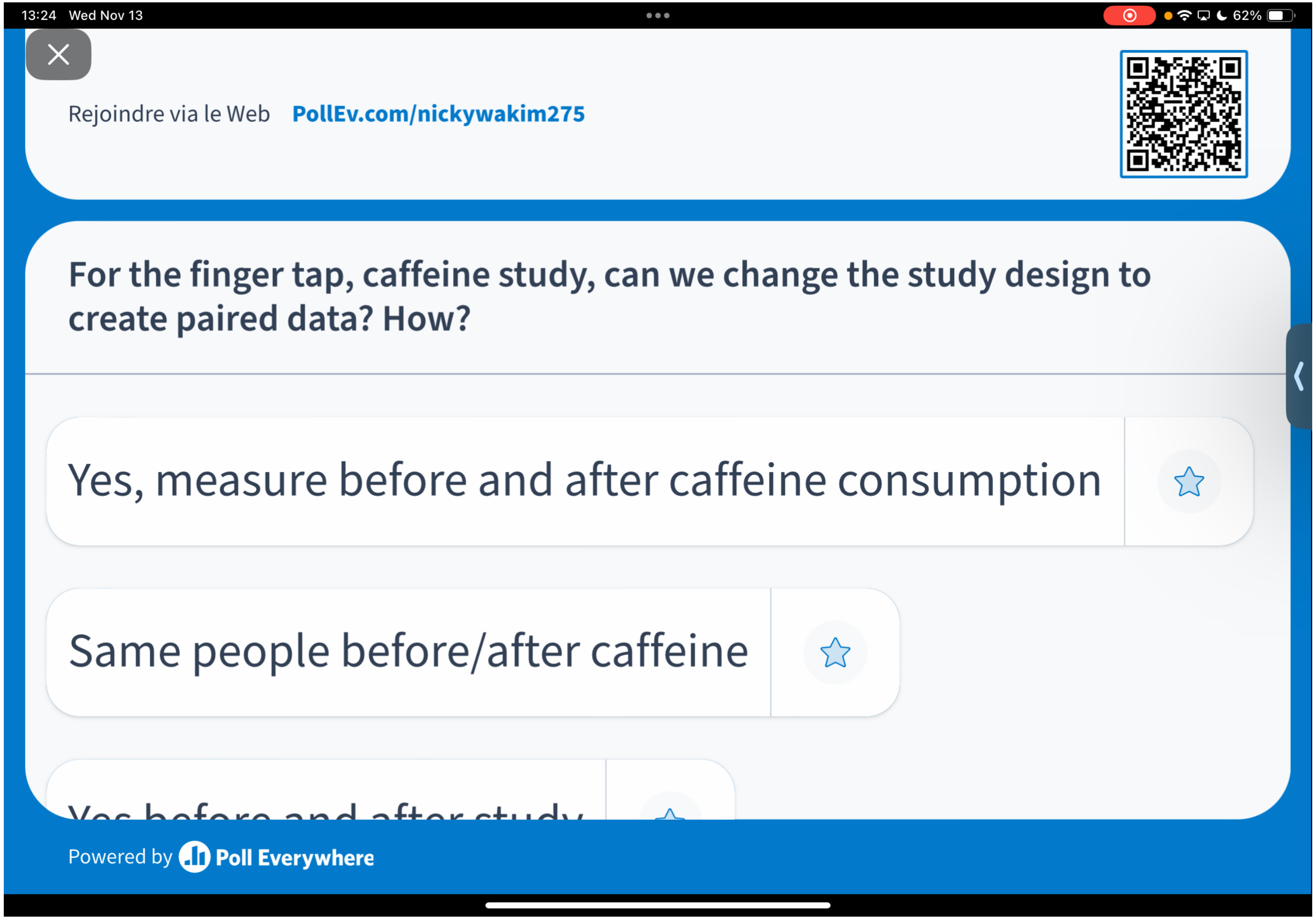
```
[1] -3.6
```

244.514 - 248.114

difference of means

- Note that we cannot calculate 35 differences in taps because these data are not paired!!
- Different individuals receive caffeine vs. do not receive caffeine

Poll Everywhere Question 2



Rejoindre via le Web **PollEv.com/nickywakim275**

For the finger tap, caffeine study, can we change the study design to create paired data? How?

Yes, measure before and after caffeine consumption ☆

Same people before/after caffeine ☆

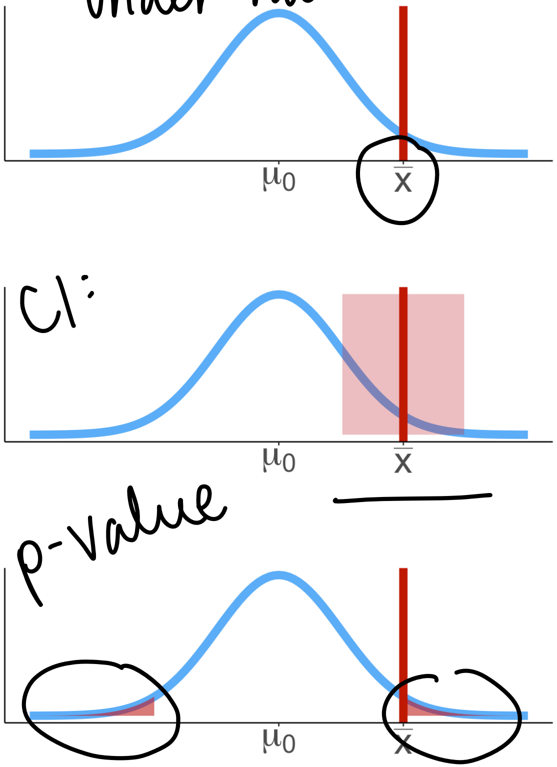
Yes before and after study

Powered by **Poll Everywhere**

What would the distribution look like for 2 independent samples?

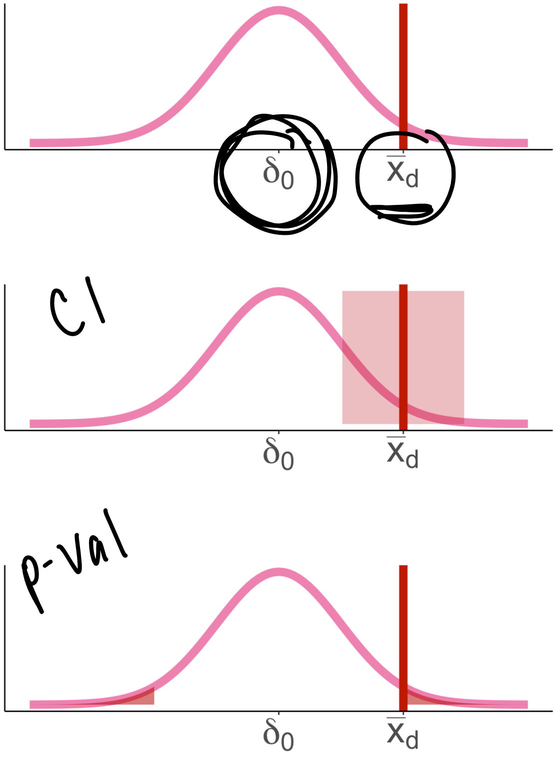
Single-sample mean:

under null



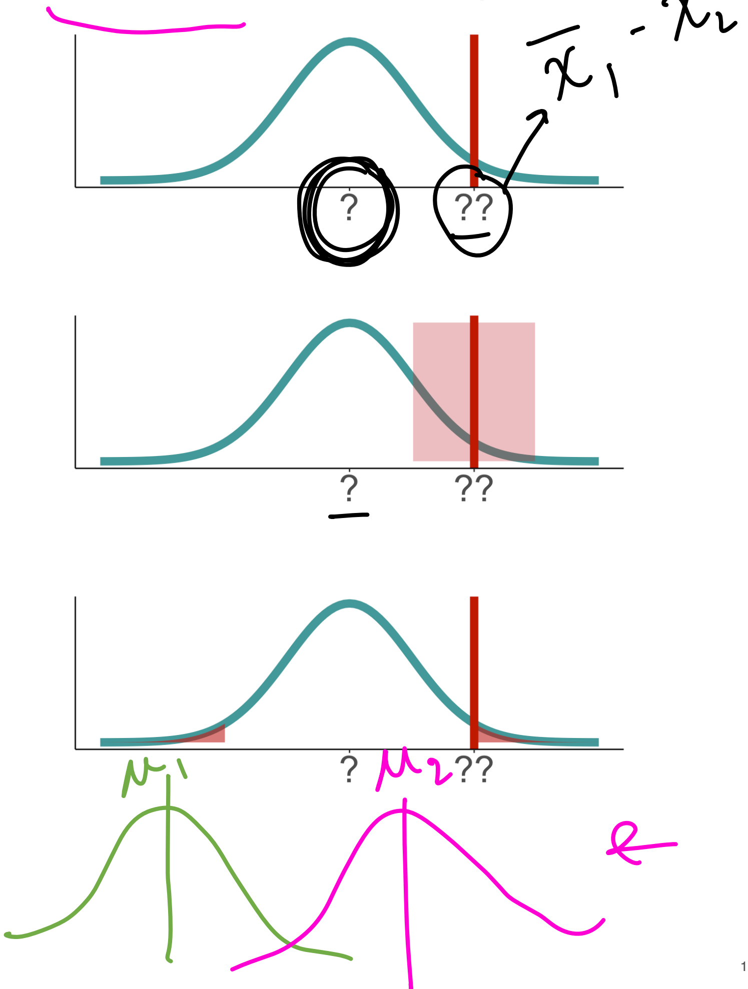
under null

Paired mean difference:



under null

Diff in means of 2 ind samples:



What distribution does $\overline{X}_1 - \overline{X}_2$ have? (when we know pop sd's)

- Let $\overline{X}_1$ and $\overline{X}_2$ be the means of random samples from two independent groups, with parameters shown in table:

	Gp 1	Gp 2
sample size	n_1	n_2
pop mean	μ_1	μ_2
pop sd	σ_1	σ_2

- Some theoretical statistics:
  - If $\overline{X}_1$ and $\overline{X}_2$ are independent normal RVs, then $\overline{X}_1 - \overline{X}_2$ is also normal

    "sum of normals is normal)

  - What is the mean of $\overline{X}_1 - \overline{X}_2$?

mean

$$E[\overline{X}_1 - \overline{X}_2] = E[\overline{X}_1] - E[\overline{X}_2] = \mu_1 - \mu_2$$

  - What is the standard deviation of $\overline{X}_1 - \overline{X}_2$? * independent

$$Var(\overline{X}_1 - \overline{X}_2) = Var(\overline{X}_1) + Var(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Lesson

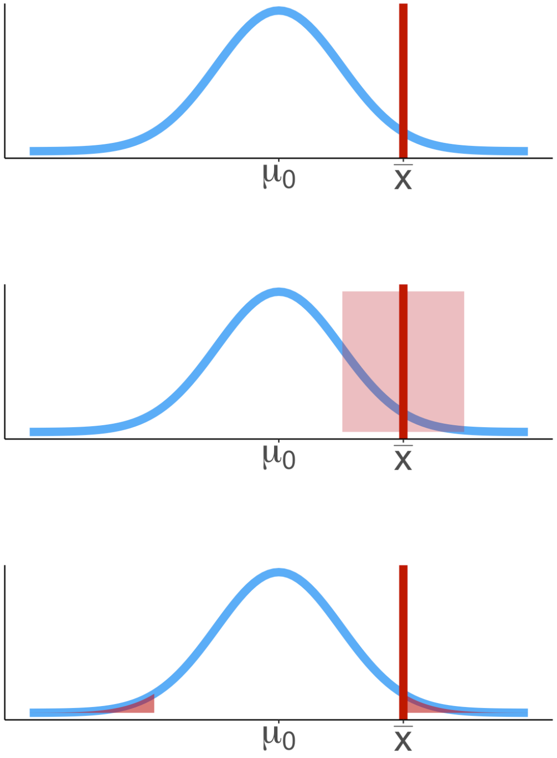
$$SD(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

distrib as

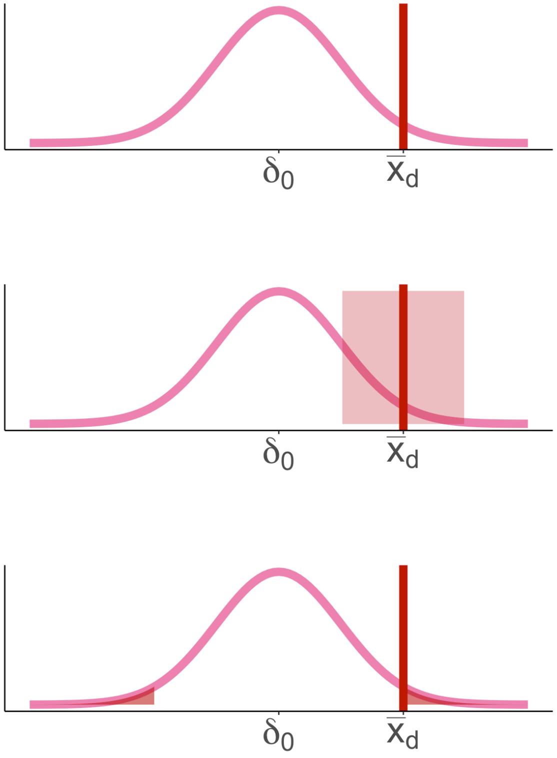
$$\overline{X}_1 - \overline{X}_2 \sim \text{Normal}\left(\mu_1 - \mu_1, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

What would the distribution look like for 2 independent samples?

Single-sample mean:

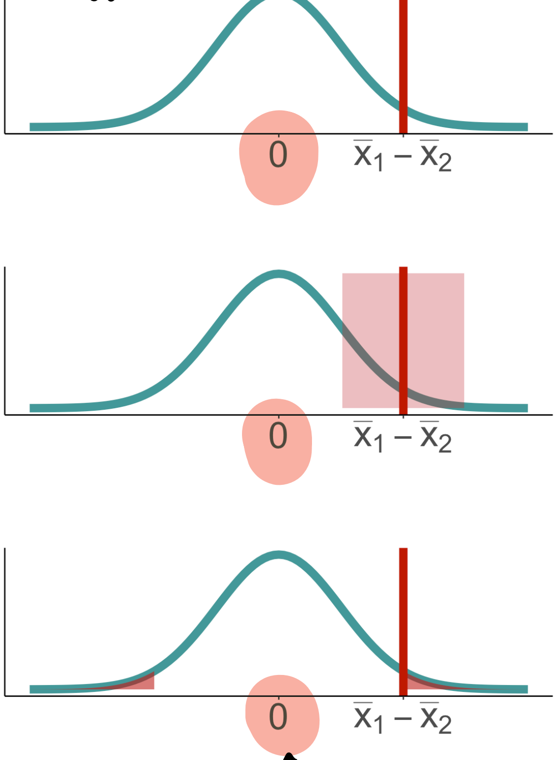


Paired mean difference:



Diff in means of 2 ind samples:

under null: no diff



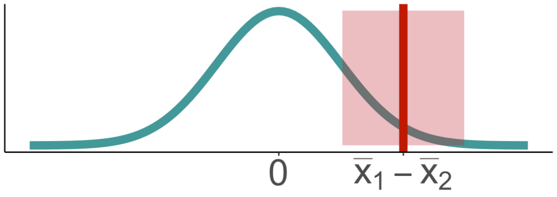
$\uparrow$

$\mu_1 - \mu_2 = 0$

Approaches to answer a research question

- **Research question is a generic form for 2 independent samples:** Is there evidence to support that the population means are different from each other?

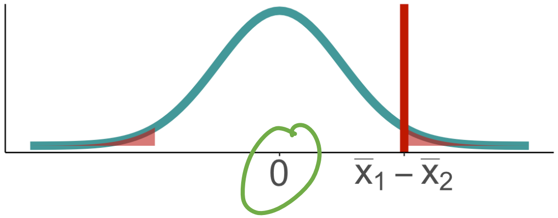
Calculate **CI for the mean difference** δ:



$$\overline{x}_1 - \overline{x}_2 \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- with t^* = t-score that aligns with specific confidence interval

Run a **hypothesis test:**



Hypotheses

$$H_0: \mu_1 = \mu_2$$
$$H_A: \mu_1 \neq \mu_2$$
$$(or <, >)$$

Test statistic

$$t_{\overline{x}_1 - \overline{x}_2} = \frac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\mu_1 = \mu_2$$
$$\rightarrow \mu_1 - \mu_2 = 0$$

② Confidence Interval

Learning Objectives

95% CI for the difference in population mean taps $\mu_1 - \mu_2$

Confidence interval for $\mu_1 - \mu_2$

$$\underline{\bar{x}_1 - \bar{x}_2} \pm t^* \times \text{SE}$$

- with $\text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ if population sd is not known

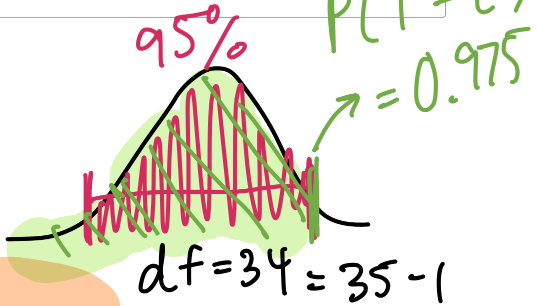
- t^* depends on the confidence level and degrees of freedom
  - degrees of freedom (df) is: $df = n - 1$
  - n is minimum between n_1 and n_2

$\Big\}$ for hand calculating

95% CI for the difference in population mean taps

```r
CaffTaps %>% group_by(Group) %>% get_summary_stats(type = "mean_sd") %>%
  gt() %>% tab_options(table.font.size = 40)
```

Group	variable	n	mean	sd
Caffeine	Taps	35	248.114	2.621
NoCaffeine	Taps	35	244.514	2.318



95% CI for $\mu_{caff} - \mu_{ctrl}$:

$\overline{x}_{ctrl} - \overline{x}_{caff}$

$$\bar{x}_{\text{caff}} - \bar{x}_{\text{ctrl}} \pm t^* \cdot \overbrace{\sqrt{\frac{s^2_{\text{caff}}}{n_{\text{caff}}} + \frac{s^2_{\text{ctrl}}}{n_{\text{ctrl}}}}}^{SE}$$

$$248.114 - 244.514 \pm 2.032 \cdot \sqrt{\frac{2.621^2}{35} + \frac{2.318^2}{35}}$$

$$3.6 \pm 2.032 \cdot \sqrt{0.196 + 0.154}$$

$$(2.398, 4.802)$$

Used t^* = `qt(0.975, df=34)` = 2.032

Conclusion:
We are 95% confident that the difference in (population) mean finger taps/min between the caffeine and control groups is between 2.398 mg/dL and 4.802 mg/dL.

95% CI for the difference in population mean taps (using R)

```
t.test(formula = Taps ~ Group, data = CaffTaps)
```

~

caff no caff

Taps Group
1
2
3

```
     Welch Two Sample t-test

data:  Taps by Group
t = 6.0867, df = 67.002, p-value = 6.266e-08
alternative hypothesis: true difference in means between group Caffeine and group NoCaffeine is
not equal to 0
95 percent confidence interval:
 2.41945 4.78055
sample estimates:
  mean in group Caffeine mean in group NoCaffeine
                248.1143                 244.5143
```

(2.398, 4.802) ← wider

(2.419, 4.781) ← narrower

outcome measurement ~ pop grps

- We can tidy the output

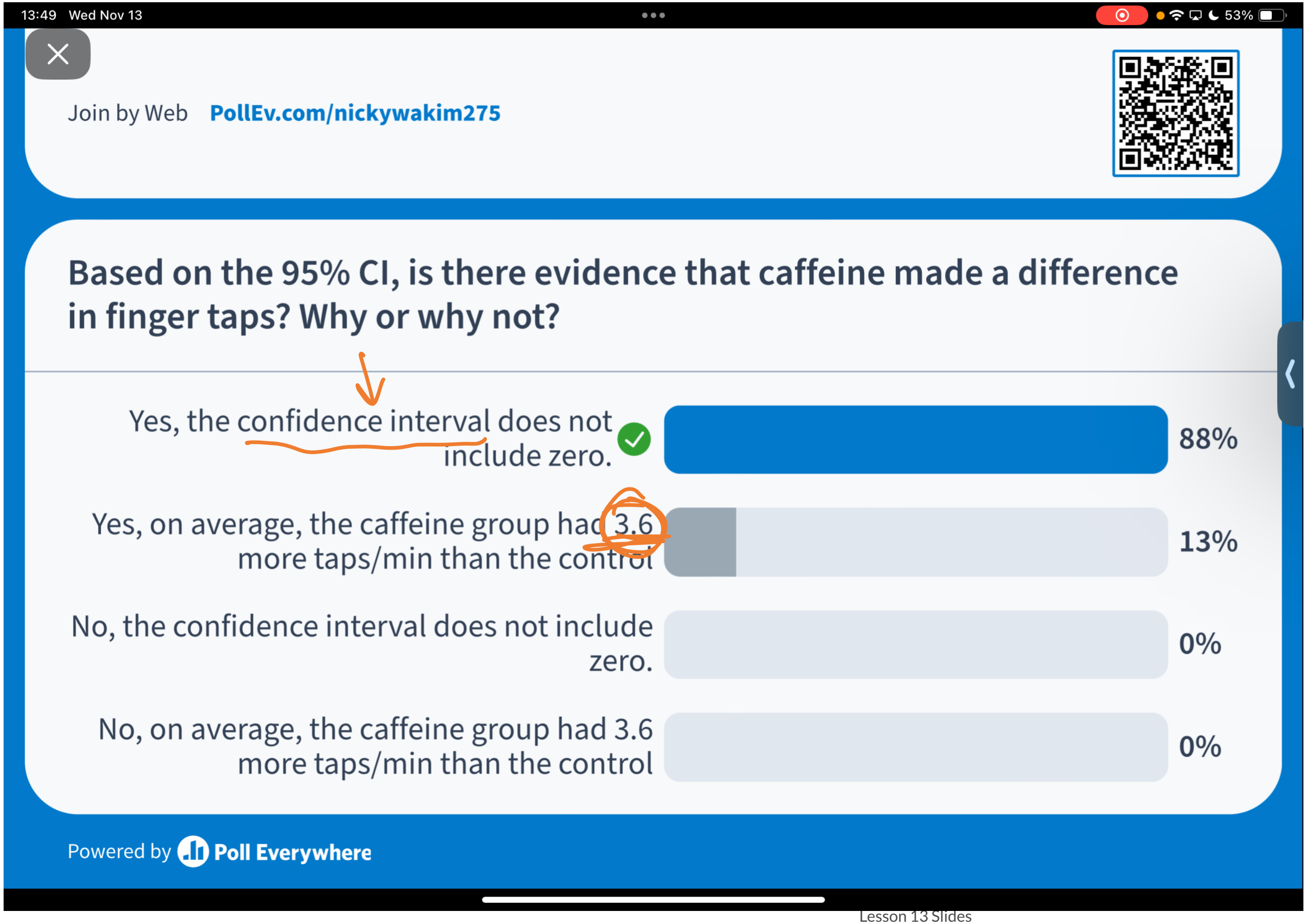
estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
3.6	248.1143	244.5143	6.086677	6.265631e-08	67.00222	2.41945	4.78055	Welch Two Sample t-test	two.sided

Conclusion:

We are 95% confident that the difference in (population) mean finger taps/min between the caffeine and control groups is between ~~2.398~~ mg/dL and ~~4.802~~ mg/dL.

2.419 4.781

Poll Everywhere Question 3



Join by Web PollEv.com/nickywakim275

Based on the 95% CI, is there evidence that caffeine made a difference in finger taps? Why or why not?

Yes, the confidence interval does not include zero. ✓	88%
Yes, on average, the caffeine group had 3.6 more taps/min than the control	13%
No, the confidence interval does not include zero.	0%
No, on average, the caffeine group had 3.6 more taps/min than the control	0%

Learning Objectives

③ hyp testing

Reference: Steps in a Hypothesis Test

1. Check the **assumptions**
2. Set the **level of significance** α
3. Specify the **null** (H_0) and **alternative** (H_A) **hypotheses**
    1. In symbols
    2. In words
    3. Alternative: one- or two-sided?
4. Calculate the **test statistic**.
5. Calculate the **p-value** based on the observed test statistic and its sampling distribution
6. Write a **conclusion** to the hypothesis test
    1. Do we reject or fail to reject H_0?
    2. Write a conclusion in the context of the problem

Step 1: Check the assumptions

- The assumptions to run a hypothesis test on a sample are:
  - **Independent observations**: Each observation from both samples is independent from all other observations
  - **Approximately normal sample or big n**: the distribution of *each sample* should be approximately normal, *or* the sample size of *each sample* should be at least 30

- These are the criteria for the Central Limit Theorem in Lesson 09: Variability in estimates

- In our example, we would check the assumptions with a statement:
  - The observations are independent from each other. Each caffeine group (aka sample) has 35 individuals. Thus, we can use CLT to approximate the sampling distribution for each sample.

Step 2: Set the level of significance

- **Before doing a hypothesis test**, we set a cut-off for how small the p-value should be in order to reject H_0.
- Typically choose $\alpha = 0.05$

- See Lesson 11: Hypothesis Testing 1: Single-sample mean

Step 3: Null & Alternative Hypotheses

Notation for hypotheses (for two ind samples)

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ \text{vs. } H_A &: \mu_1 \neq, <, \text{or}, > \mu_2 \end{aligned}$$

Hypotheses test for example

$$\begin{aligned} H_0 &: \mu_{caff} = \mu_{ctrl} \\ \text{vs. } H_A &: \mu_{caff} > \mu_{ctrl} \end{aligned}$$

- Under the null hypothesis: $\mu_1 = \mu_2$, so the difference in the means is $\mu_1 - \mu_2 = 0$

$H_A : \mu_1 \neq \mu_2$

- not choosing a priori whether we believe the population mean of group 1 is different than the population mean of group 2

$H_A : \mu_1 < \mu_2$

- believe that population mean of group 1 is ~~greater~~ than population mean of group 2

less

$H_A : \mu_1 > \mu_2$

- believe that population mean of group 1 is ~~less~~ than population mean of group 2

greater

- $H_A : \mu_1 \neq \mu_2$ is the most common option, since it's the most conservative

Step 3: Null & Alternative Hypotheses: another way to write it

- Under the null hypothesis: $\mu_1 = \mu_2$, so the difference in the means is $\mu_1 - \mu_2 = 0$

$H_A : \mu_1 \neq \mu_2$

- not choosing a priori whether we believe the population mean of group 1 is different than the population mean of group 2

$H_A : \underline{\mu_1 - \mu_2} \neq 0$

- not choosing a priori whether we believe the difference in population means is greater or less than 0

same

$H_A : \mu_1 > \mu_2$

- believe that population mean of group 1 is greater than population mean of group 2

$H_A : \underline{\mu_1 - \mu_2} > 0$

- believe that difference in population means (mean 1 - mean 2) is greater than 0

same

$H_A : \mu_1 < \mu_2$

- believe that population mean of group 1 is less than population mean of group 2

$H_A : \underline{\mu_1 - \mu_2} \underline{<} 0$

- believe that difference in population means (mean 1 - mean 2) is less than 0

same

Step 3: Null & Alternative Hypotheses

- **Question**: Is there evidence to support that drinking caffeine increases the number of finger taps/min?

Null and alternative hypotheses in **words**

- H_0: The population difference in mean finger taps/min between the caffeine and control groups is 0
- H_A: The population difference in mean finger taps/min between the caffeine and control groups is greater than 0

Null and alternative hypotheses in **symbols**

$$H_0 : \mu_{caff} - \mu_{ctrl} = 0$$
$$H_A : \mu_{caff} - \mu_{ctrl} > 0$$

Step 4: Test statistic

Recall, for a two sample independent means test, we have the following test statistic:

$$t_{\bar{x}_1-\bar{x}_2} = \frac{\boxed{\bar{x}_1 - \bar{x}_2} - \boxed{0}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\mu_1 - \mu_2 = 0$
under
null

- $\bar{x}_1, \bar{x}_2$ are the sample means
- $\mu_0 = 0$ is the mean value specified in H_0
- s_1, s_2 are the sample SD's
- n_1, n_2 are the sample sizes

- Statistical theory tells us that $t_{\bar{x}_1-\bar{x}_2}$ follows a **student's t-distribution** with
  - $\underline{df \approx}$ smaller of $n_1 - 1$ and $n_2 - 1$
  - this is a conservative estimate (smaller than actual df)

Step 4: Test statistic (where we do not know population sd)

From our example: Recall that $\underline{\overline{x}_1} = 248.114$, $\underline{s_1} = 2.621$, $\underline{n_1} = 35$, $\underline{\overline{x}_2} = 244.514$, $\underline{s_2} = 2.318$, and $\underline{n_2} = 35$:

The test statistic is:

→ > 0 alternative

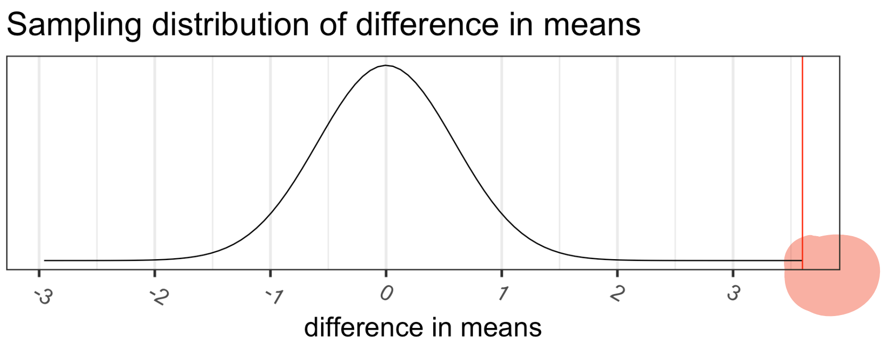
$$\text{test statistic} = t_{\overline{x}_1-\overline{x}_2} = \frac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{248.114 - 244.514 - 0}{\sqrt{\frac{2.621^2}{35} + \frac{2.318^2}{35}}} = \underline{6.0869}$$

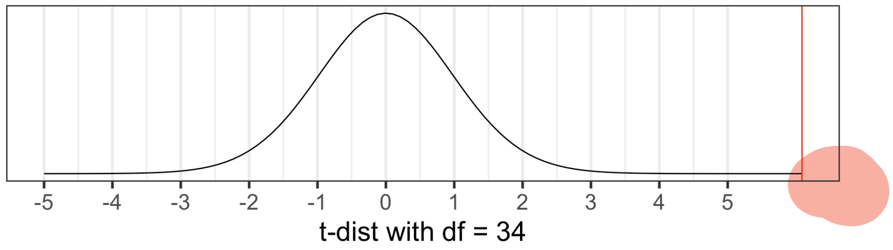
- Statistical theory tells us that $t_{\overline{x}}$ follows a **Student's t-distribution** with $df = n - 1 = 34$

$P(T \geq t^*)$

Step 5: p-value

The **p-value** is the **probability** of obtaining a test statistic *just as extreme or more extreme* than the observed test statistic assuming the null hypothesis H_0 is true.





Calculate the *p*-value using the **Student's t-distribution** with $df = n - 1 = 35 - 1 = 34$:

$$\begin{aligned} \text{p-value} &= P(T > 6.08691) \\ &= 3.3 \times 10^{-7} < 0.001 \end{aligned}$$

```r
pt(tstat,
   df = min(n1 - 1, n2 - 1),
   lower.tail = FALSE)
```

```
[1] 3.321969e-07
```

Step 4-5: test statistic and p-value together using `t.test()`

- I will have reference slides at the end of this lesson to show other options and how to “tidy” the results

```
t.test(formula = Taps ~ Group, alternative = "greater", data = CaffTaps)
```

$H_A: \mu_{caff} - \mu_{ctrl} > 0$

```

	Welch Two Sample t-test

data:  Taps by Group
t = 6.0867, df = 67.002, p-value = 3.133e-08
alternative hypothesis: true difference in means between group Caffeine and group NoCaffeine is
greater than 0
95 percent confidence interval:
 2.613502      Inf
sample estimates:
  mean in group Caffeine mean in group NoCaffeine
                248.1143                 244.5143
```

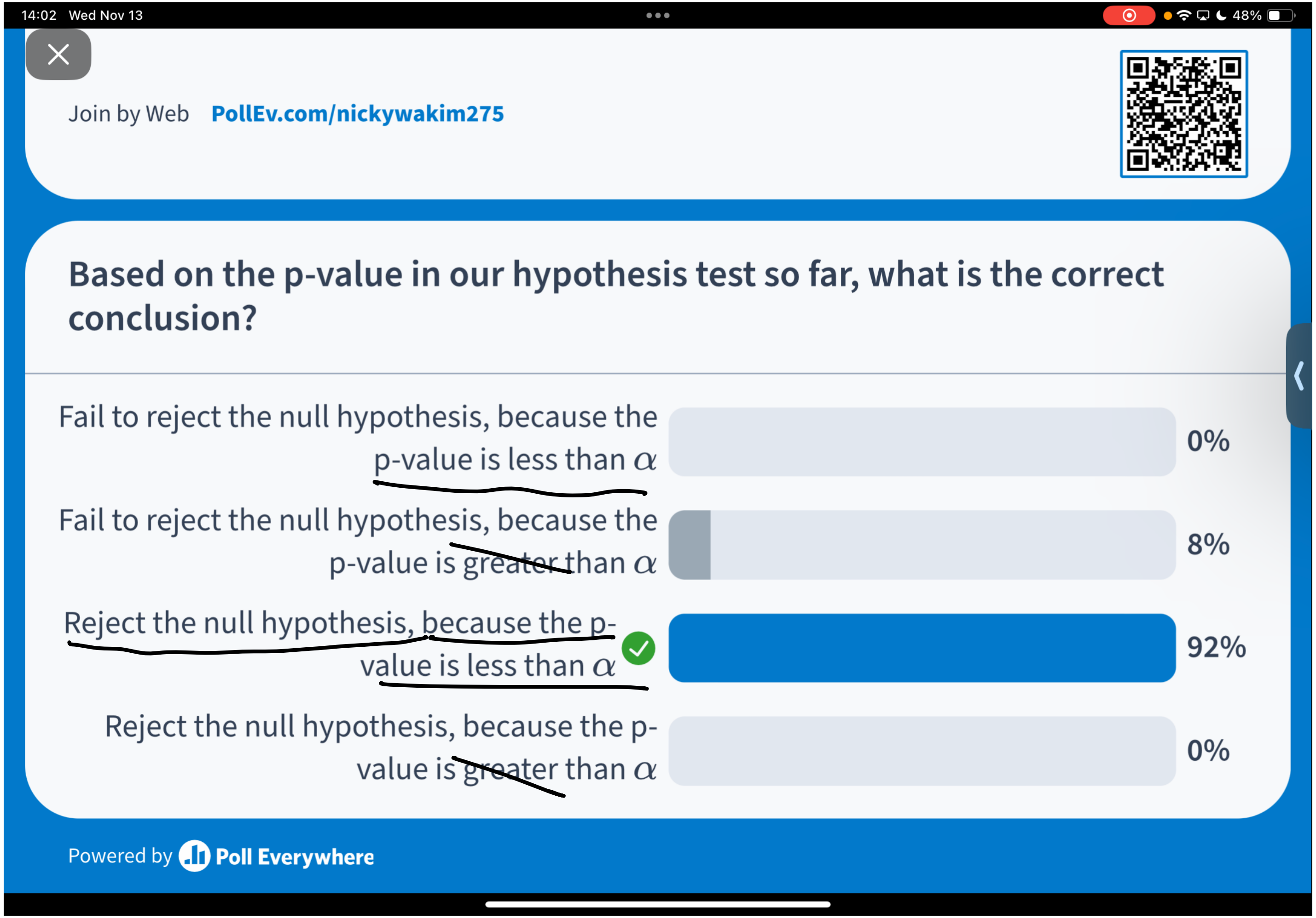
- Why are the degrees of freedom different? (see Slide Section 5.4)
  - Degrees of freedom in R is more accurate
  - Using our approximation in our calculation is okay, but conservative

p-value hand calc

>

[p-value from t.test()]

Poll Everywhere Question 4



Join by Web PollEv.com/nickywakim275

Based on the p-value in our hypothesis test so far, what is the correct conclusion?

Fail to reject the null hypothesis, because the p-value is less than α 0%

Fail to reject the null hypothesis, because the p-value is ~~greater~~ than α 8%

Reject the null hypothesis, because the p-value is less than α ✓ 92%

Reject the null hypothesis, because the p-value is ~~greater~~ than α 0%

p-value 3.3×10^{-8}

< 0.001

$\alpha = 0.05$

p-value < 0.001

< 0.05

Step 6: Conclusion to hypothesis test

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ \text{vs. } H_A &: \mu_1 \underset{\neq}{>} \mu_2 \end{aligned}$$

- Need to compare p-value to our selected $\alpha = 0.05$
- Do we reject or fail to reject H_0?

If p-value $< \alpha$, reject the null hypothesis

- There is sufficient evidence that the difference in population means is discernibly greater than 0 (p-value = __)

discernibly diff
than 0

If p-value $\geq \alpha$, fail to reject the null hypothesis

- There is insufficient evidence that the difference in population means is discernibly greater than 0 (p-value = __)

diff
than 0

Step 6: Conclusion to hypothesis test

$$H_0 : \mu_{caff} - \mu_{ctrl} = 0$$
$$H_A : \mu_{caff} - \mu_{ctrl} > 0$$

- Recall the p-value = 3×10^{-8}
- Use α = 0.05.
- Do we reject or fail to reject H_0?

Conclusion statement:

- Stats class conclusion
  - There is sufficient evidence that the (population) difference in mean finger taps/min with vs. without caffeine is greater than 0 (p-value < 0.001).
- More realistic manuscript conclusion:
  - The mean finger taps/min were 248.114 (SD = 2.621) and 244.514 (SD = 2.318) for the control and caffeine groups, and the increase of 3.6 taps/min was statistically discrenible (p-value < ~~0~~).

0.001

Reference: Ways to run a 2-sample t-test in R

R: 2-sample t-test (with long data)

- The `CaffTaps` data are in a *long* format, meaning that
  - all of the outcome values are in one column and
  - another column indicates which group the values are from
- This is a common format for data from multiple samples, especially if the sample sizes are different.

```r
(Taps_2ttest <- t.test(formula = Taps ~ Group,
                       alternative = "greater",
                       data = CaffTaps))
```

```
	Welch Two Sample t-test

data:  Taps by Group
t = 6.0867, df = 67.002, p-value = 3.133e-08
alternative hypothesis: true difference in means between group Caffeine and group NoCaffeine is
greater than 0
95 percent confidence interval:
 2.613502      Inf
sample estimates:
  mean in group Caffeine mean in group NoCaffeine 
              248.1143                 244.5143 
```

`tidy` the `t.test` output

```r
# use tidy command from broom package for briefer output that's a tibble
tidy(Taps_2ttest) %>% gt() %>% tab_options(table.font.size = 40)
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
3.6	248.1143	244.5143	6.086677	3.132816e-08	67.00222	2.613502	Inf	Welch Two Sample t-test	greater

- Pull the p-value:

```r
tidy(Taps_2ttest)$p.value  # we can pull specific values from the tidy output
```

```
[1] 3.132816e-08
```

R: 2-sample t-test (with wide data)

```r
# make CaffTaps data wide: pivot_wider needs an ID column so that it
# knows how to "match" values from the Caffeine and NoCaffeine groups
CaffTaps_wide <- CaffTaps %>%
  mutate(id = c(rep(1:10, 2), rep(11:35, 2))) %>% #  "fake" IDs for pivot_wider ste
  pivot_wider(names_from = "Group",
              values_from = "Taps")

glimpse(CaffTaps_wide)
```

caff 246 nocaff 242

```
Rows: 35
Columns: 3
$ id         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, …
$ Caffeine   <int> 246, 248, 250, 252, 248, 250, 246, 248, 245, 250, 251, 251,…
$ NoCaffeine <int> 242, 245, 244, 248, 247, 248, 242, 244, 246, 242, 244, 245,…
```

```r
t.test(x = CaffTaps_wide$Caffeine, y = CaffTaps_wide$NoCaffeine, alternative = "gre
  tidy() %>% gt() %>% tab_options(table.font.size = 40)
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
3.6	248.1143	244.5143	6.086677	3.132816e-08	67.00222	2.613502	Inf	Welch Two Sample t-test	greater

Why are the df’s in the R output different?

From many slides ago:

- Statistical theory tells us that $t_{\overline{x}_1-\overline{x}_2}$ follows a **student’s t-distribution** with
    - $df \approx$ smaller of $n_1 - 1$ and $n_2 - 1$
    - this is a **conservative** estimate (smaller than actual df)

The actual degrees of freedom are calculated using Satterthwaite’s method:

$$\nu = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} = \frac{[SE_1^2 + SE_2^2]^2}{SE_1^4/df_1 + SE_2^4/df_2}$$

Verify the *p*-value in the R output using ν = 17.89012:

```r
pt(3.3942, df = 17.89012, lower.tail = FALSE)
```

```
[1] 0.001627588
```