Lesson 12: Inference for mean difference from two-sample dependent/paired data

TB sections 5.2

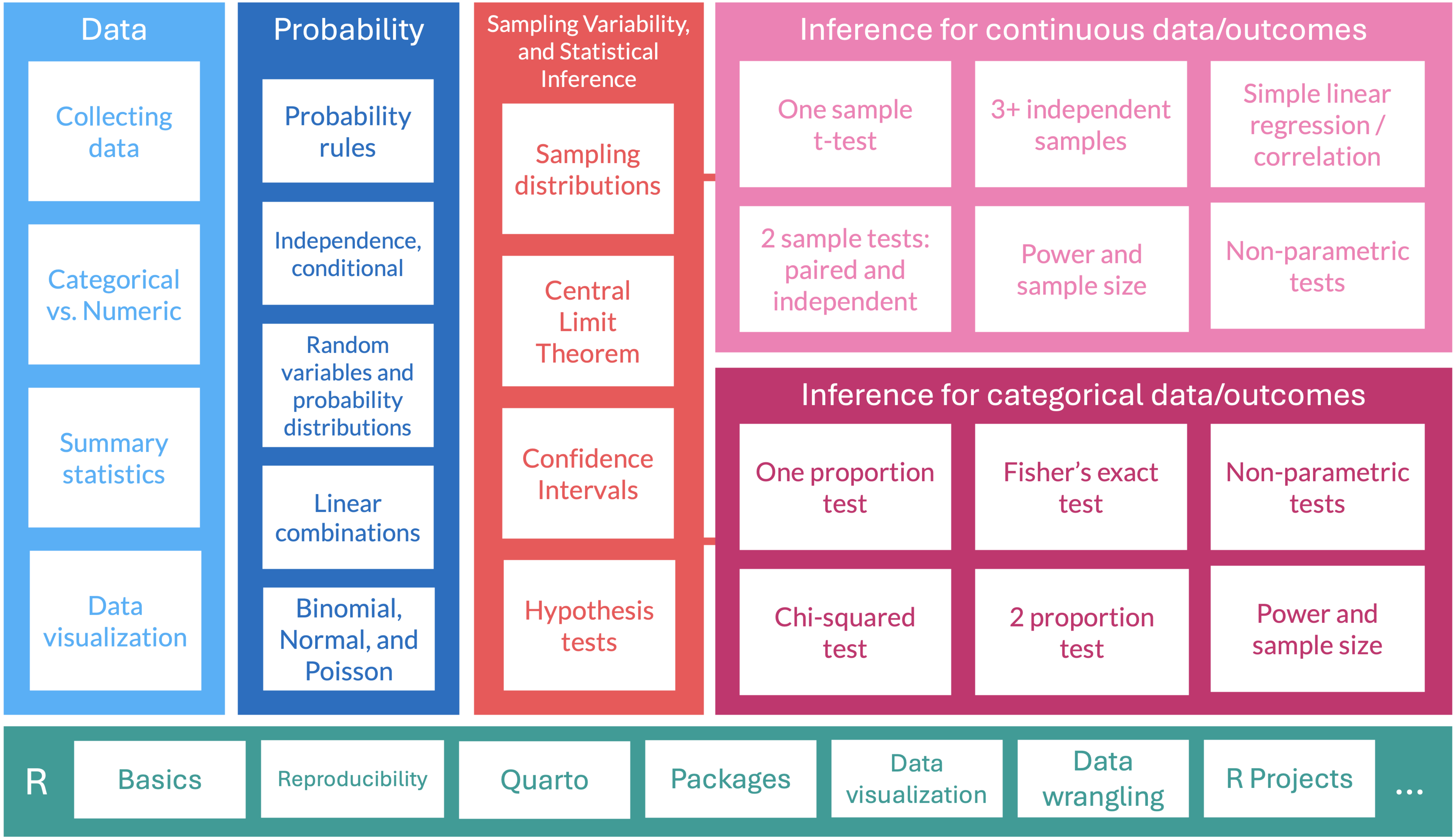
Meike Niederhausen and Nicky Wakim

2024-11-11

Learning Objectives

1. Define paired data and explain how it differs from independent samples in the context of statistical analysis.
2. Construct confidence intervals for the mean difference in paired data and interpret these intervals in the context of the research question.
3. Perform the appropriate hypothesis test for paired 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

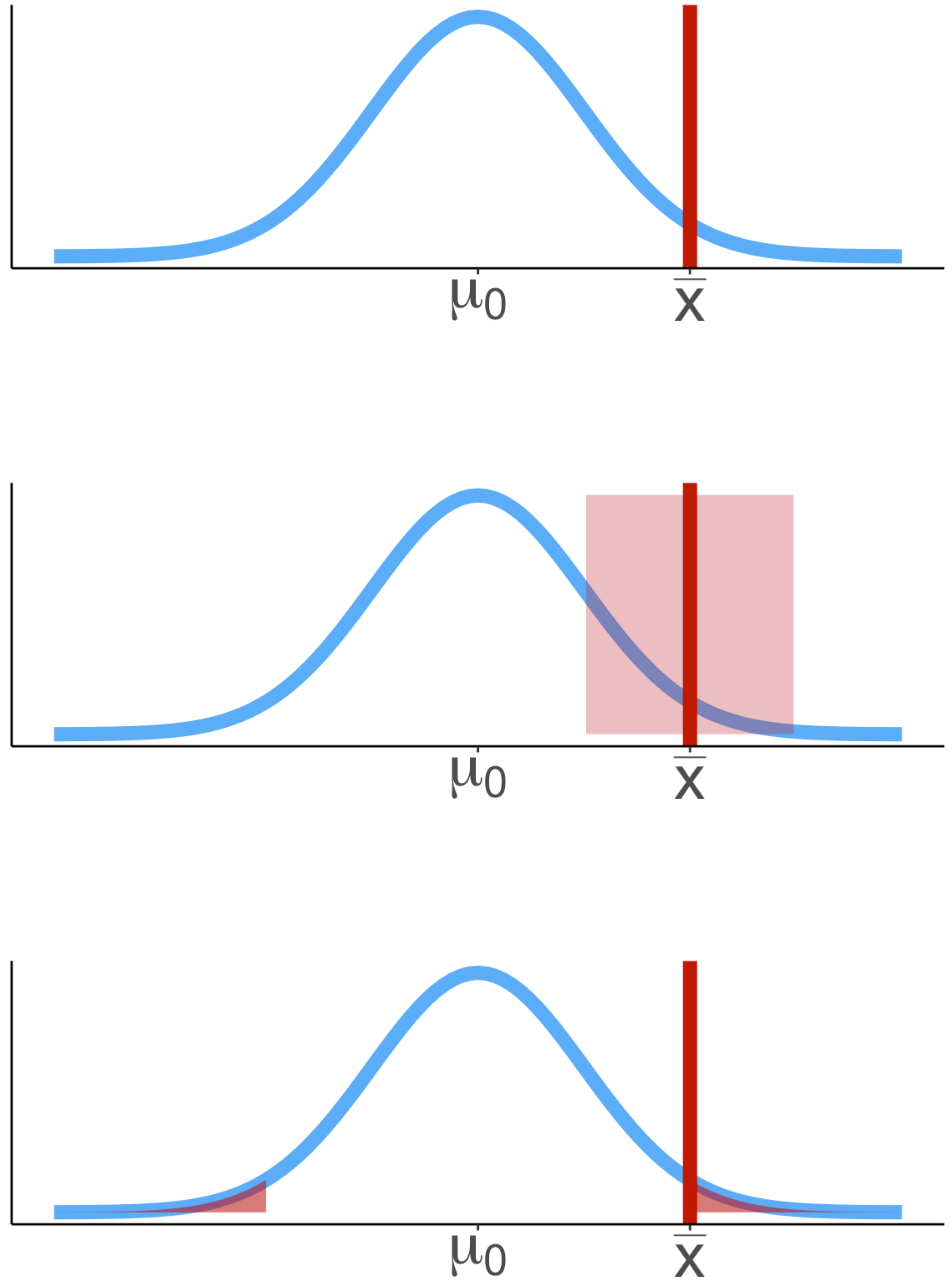
Inference for categorical data/outcomes
- One proportion test
- Fisher’s exact test
- Non-parametric tests
- Chi-squared test
- 2 proportion test
- Power and sample size

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

Last 2 times: Inference for a single-sample mean

- Inference for a single-sample mean includes:
  - Confidence intervals (Lesson 10)
  - Hypothesis testing (Lesson 11)

Single-sample mean:



Last time: example of a hypothesis test for a single-sample mean

Example of hypothesis test based on the 1992 JAMA data

Is there evidence to support that the population mean body temperature is different from 98.6°F?

1. **Assumptions:** The individual observations are independent and the number of individuals in our sample is 130. Thus, we can use CLT to approximate the sampling distribution.

2. Set $\alpha = 0.05$

3. **Hypothesis:**

$$\begin{aligned} H_0 &: \mu = 98.6 \\ \text{vs. } H_A &: \mu \neq 98.6 \end{aligned}$$

4-5.

```r
temps_ttest <- t.test(x = BodyTemps$Temperature, mu = 98.6)
tidy(temps_ttest) %>% gt() %>% tab_options(table.font.size = 36)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
98.24923	-5.454823	2.410632e-07	129	98.122	98.37646	One Sample t-test	two.sided

6. **Conclusion:** We reject the null hypothesis. The average body temperature in the sample was 98.25°F (95% CI 98.12, 98.38°F), which is discernibly different from 98.6°F (p-value < 0.001).

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	$\overline{x}_d$	???
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 Objectives

1. Define paired data and explain how it differs from independent samples in the context of statistical analysis.
2. Construct confidence intervals for the mean difference in paired data and interpret these intervals in the context of the research question.
3. Perform the appropriate hypothesis test for paired data and interpret the results.

What are paired data?

- **Paired data:** two sets of observations are uniquely paired so that an observation in one set matches an observation in the other

- Examples
  - Enroll pairs of identical twins to study a disease
  - Enroll people and collect data before & after an intervention (longitudinal data)
  - Textbook example: Compare maximal speed of competitive swimmers wearing a wetsuit vs. wearing a regular swimsuit

- Paired data result in a **natural measure of difference**
  - Example: Enroll parent and child pairs to study cholesterol levels
    - We can look at the difference in cholesterol levels between parent and child

For paired data: Population parameters vs. sample statistics

Population parameter

- Mean difference: δ (“delta”, lowercase)
- Standard deviation: σ_d (“sigma”)
- Variance: σ_d^2

Sample statistic (point estimate)

- Sample mean difference: $\overline{x}_d$
- Sample standard deviation: s_d
- Sample variance: s_d^2

- Using d helps us distinguish between a single sample and paired data

Can a vegetarian diet change cholesterol levels?

- We will illustrate how to perform a hypothesis test and calculate a confidence interval for paired data as we work through this example
- **Scenario**:
  - 43 non-vegetarian people were enrolled in a study and were instructed to adopt a vegetarian diet
  - Cholesterol levels were measured before and after the vegetarian diet

Question: Is there evidence to support that cholesterol levels changed after the vegetarian diet?

- How do we answer this question?
  - First, calculate changes (differences) in cholesterol levels
    - We usually do after - before if the data are longitudinal
  - Then find CI or perform hypothesis test

EDA: Explore the cholesterol data

- Read in the data with `read.csv()`

```r
chol <- read.csv(here::here("data", "chol213_n40.csv"))
```

- Take a look at the variables with `glimpse()`

```r
glimpse(chol)
```

```
Rows: 43
Columns: 2
$ Before <int> 195, 145, 205, 159, 244, 166, 250, 236, 192, 224, 238, 197, 169…
$ After  <int> 146, 155, 178, 146, 208, 147, 202, 215, 184, 208, 206, 169, 182…
```

- Get summary statistics with `get_summary_stats()`

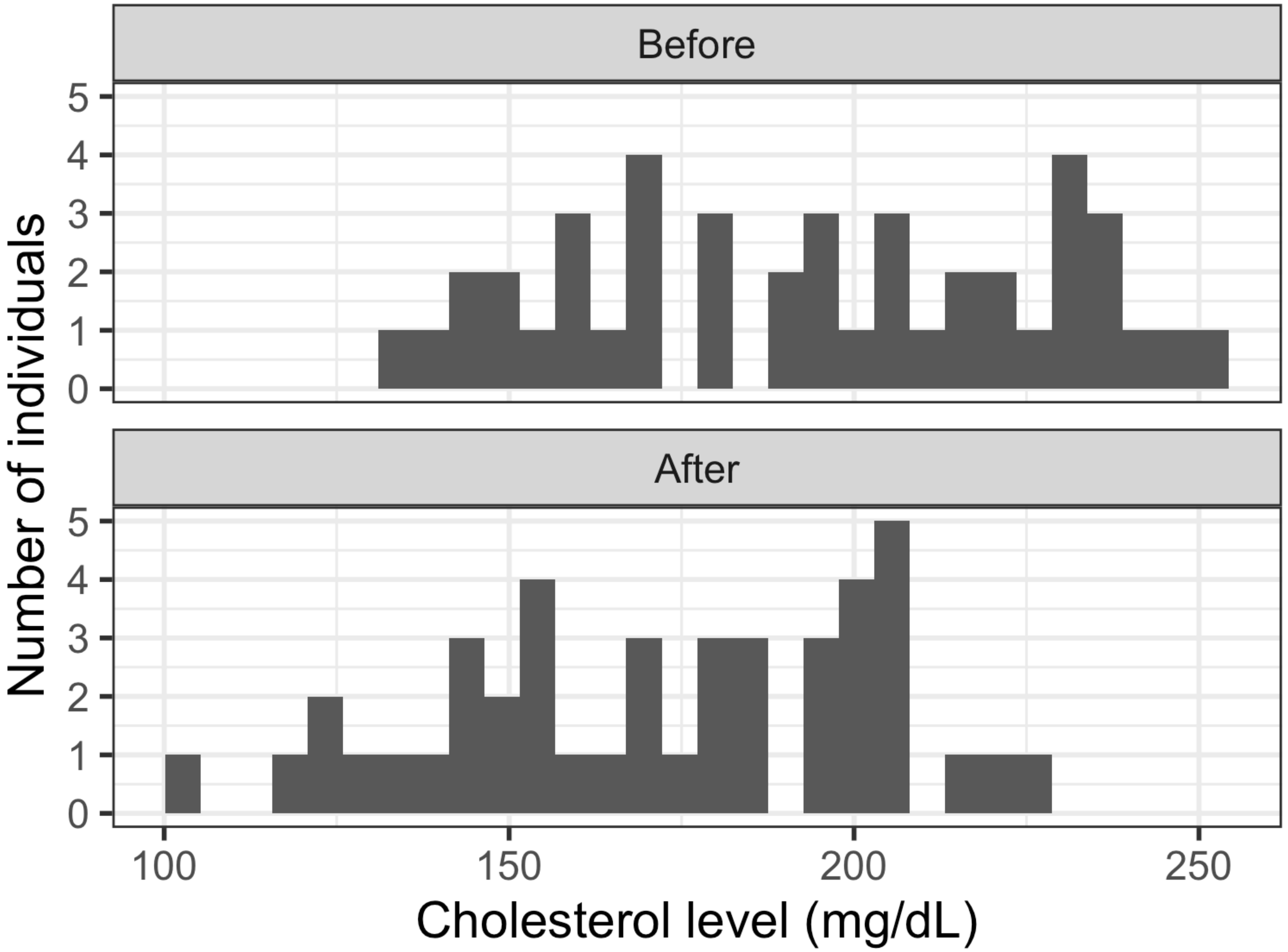
```r
chol %>% get_summary_stats(type = "common") %>%
  gt() %>% tab_options(table.font.size = 40)
```

variable	n	min	max	median	iqr	mean	sd	se	ci
Before	43	132	250	197	56.5	193.977	34.098	5.200	10.494
After	43	101	227	176	50.5	172.209	31.112	4.744	9.575

EDA: Cholesterol levels before and after vegetarian diet

- Behind the scenes: I changed the data from wide to long format to make this plot (to be covered in R08)

```r
ggplot(chol_long, aes(x = Cholesterol)) + geom_histogram() +
  facet_wrap(~ Time, ncol = 1) +
  labs(y = "Number of individuals", x = "Cholesterol level (mg/dL)")
```



EDA: Differences in cholesterol levels: After - Before diet

- How do we calculate the difference in cholesterol levels?
- I can create a new variable called "DiffChol" using the `mutate()` function (look more closely at this in R08)

```r
chol <- chol %>%
  mutate(DiffChol = After - Before)
glimpse(chol)
```

```
Rows: 43
Columns: 3
$ Before   <int> 195, 145, 205, 159, 244, 166, 250, 236, 192, 224, 238, 197, 1…
$ After    <int> 146, 155, 178, 146, 208, 147, 202, 215, 184, 208, 206, 169, 1…
$ DiffChol <int> -49, 10, -27, -13, -36, -19, -48, -21, -8, -16, -32, -28, 13,…
```

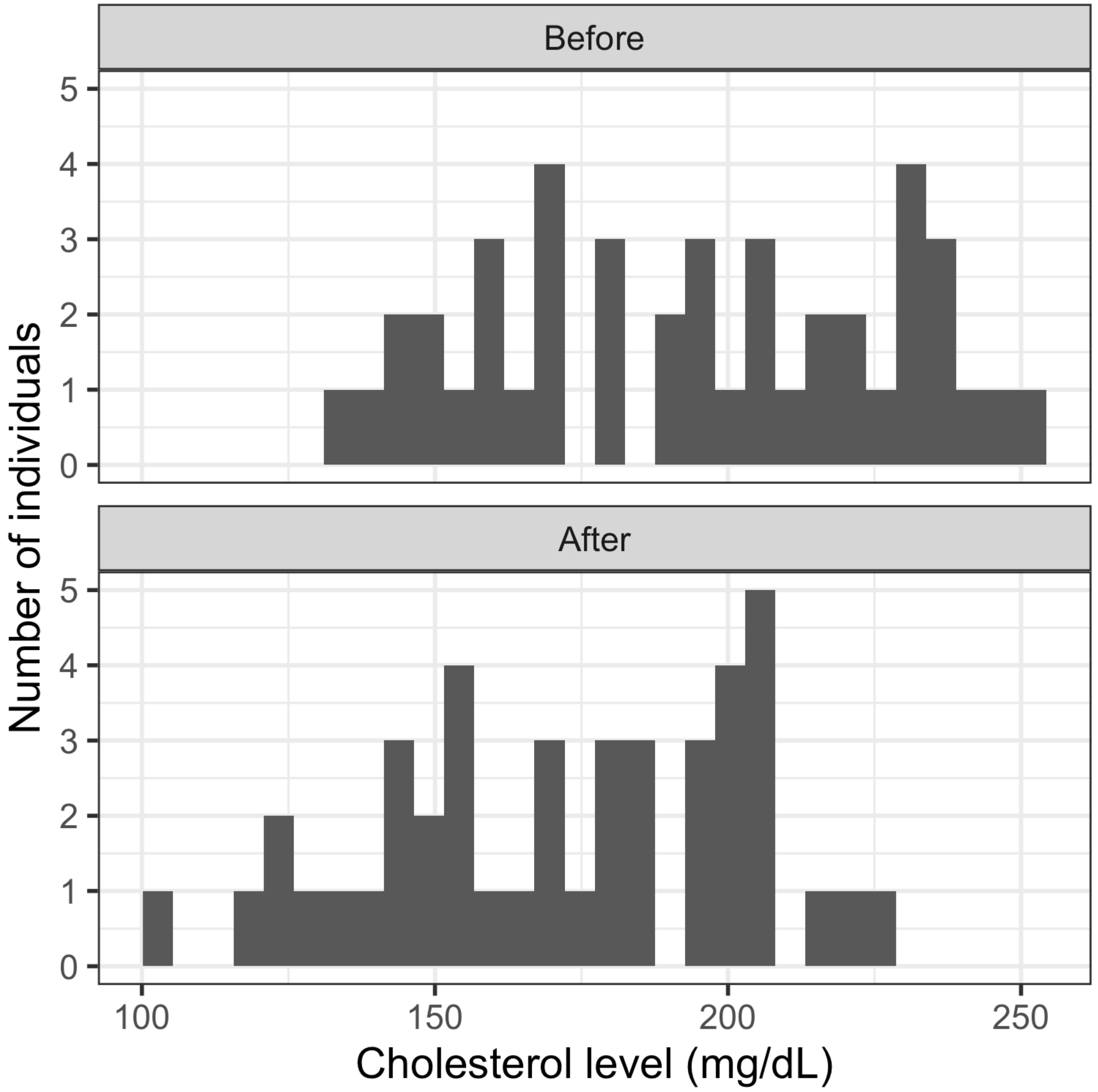
Poll Everywhere Question 1

Summary stats including difference in cholesterol:

variable	n	min	max	median	iqr	mean	sd	se	ci
Before	43	132	250	197	56.5	193.977	34.098	5.200	10.494
After	43	101	227	176	50.5	172.209	31.112	4.744	9.575
DiffChol	43	-49	13	-23	16.0	-21.767	13.890	2.118	4.275

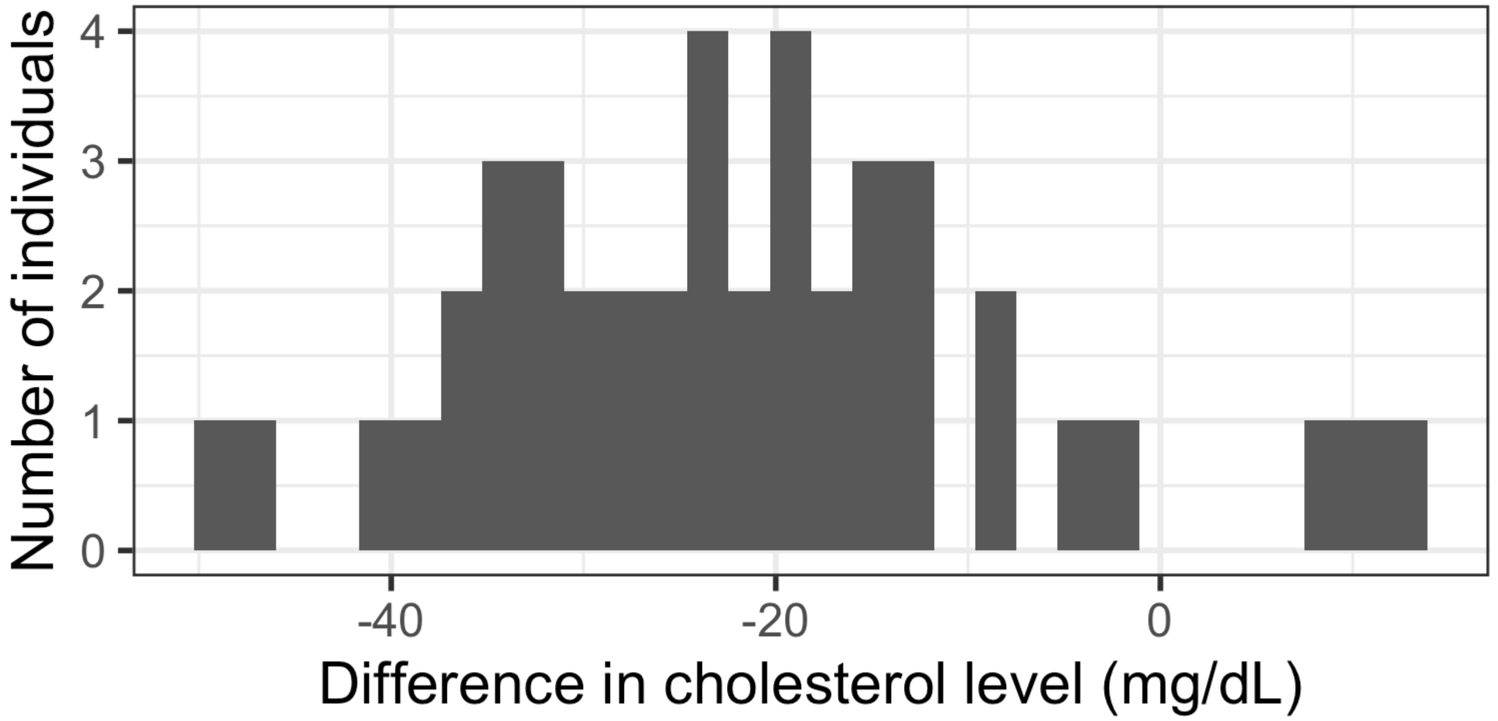
EDA: Differences in cholesterol levels: After - Before diet

- Code for below plot



Difference:

```r
ggplot(chol, aes(x=DiffChol)) +
  geom_histogram() +
  labs(y = "Number of individuals",
       x = "Difference in cholesterol l
```



Same distribution: single-sample mean & paired mean difference (1/2)

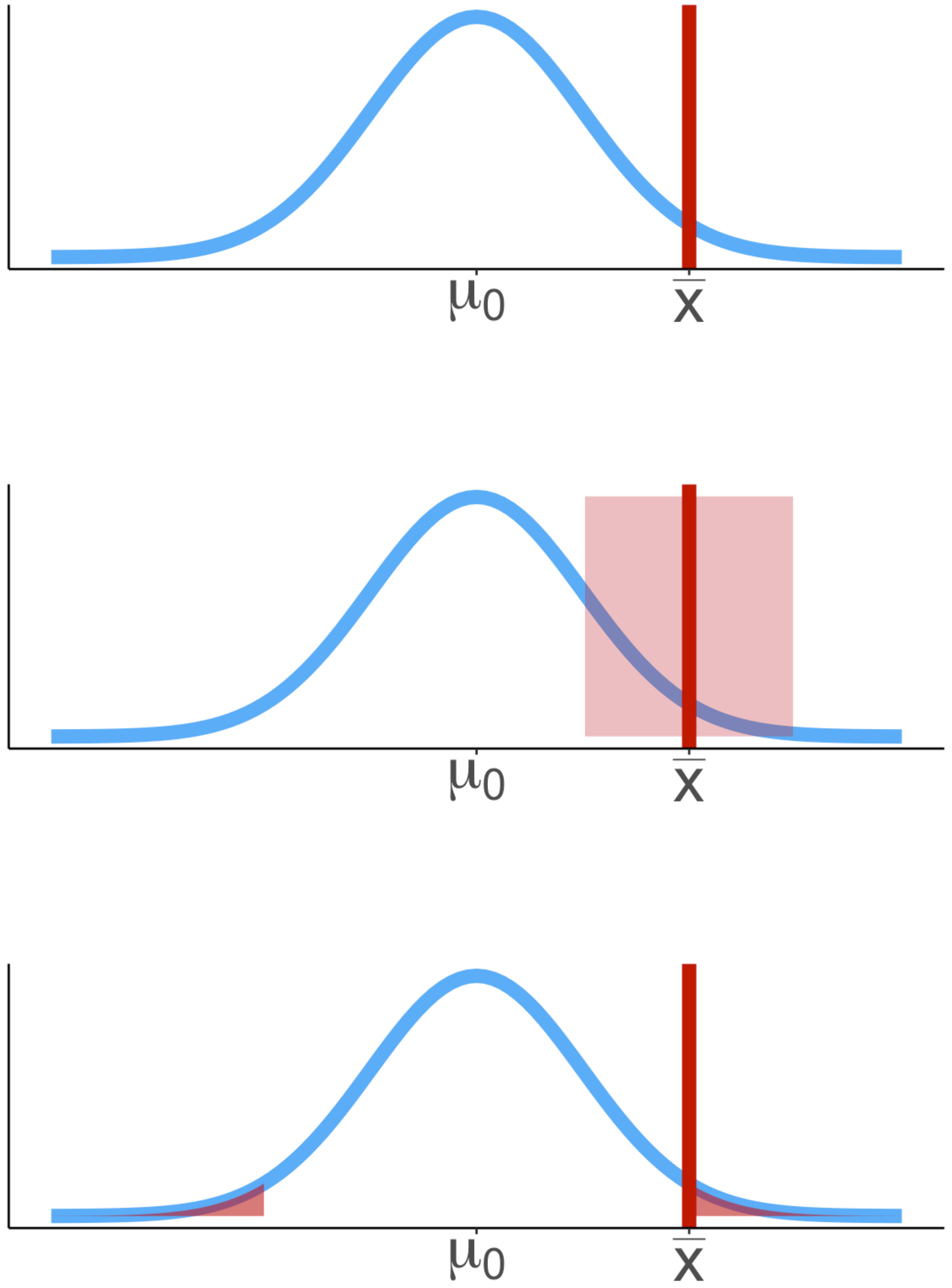
- Even though we are looking at a difference, we have a single sample mean to represent the difference
  - Before, we had single sample mean $\overline{x}$
  - Now we have a sample mean difference $\overline{x}_d$

- Distribution for the mean difference for paired data is the same as the distribution for a single mean
  - Use the t-distribution to build our inference

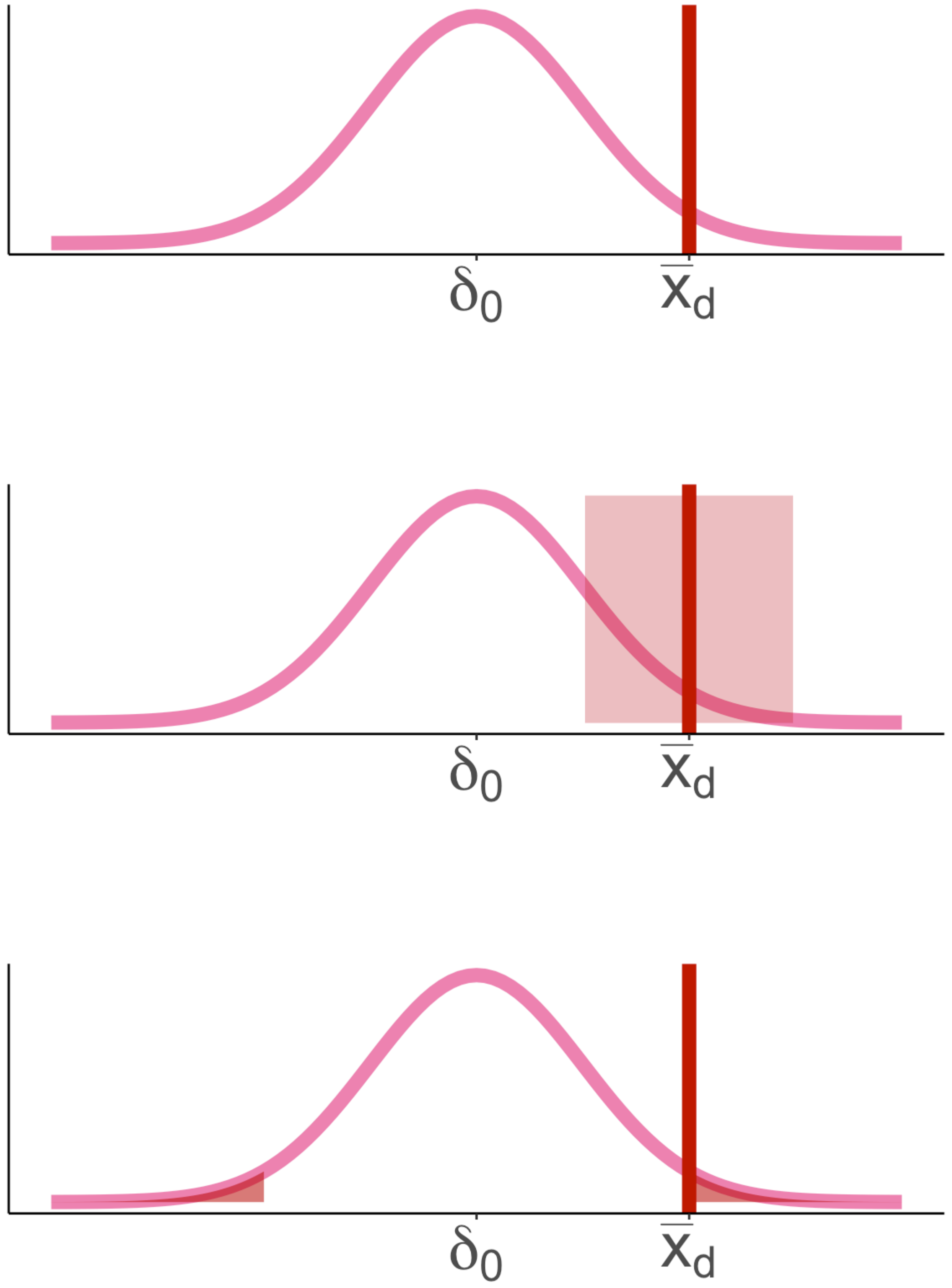
- We can use the same procedure for confidence intervals and hypothesis testing as we did for the single-sample mean

Same distribution: single-sample mean & paired mean difference (2/2)

Single-sample mean:



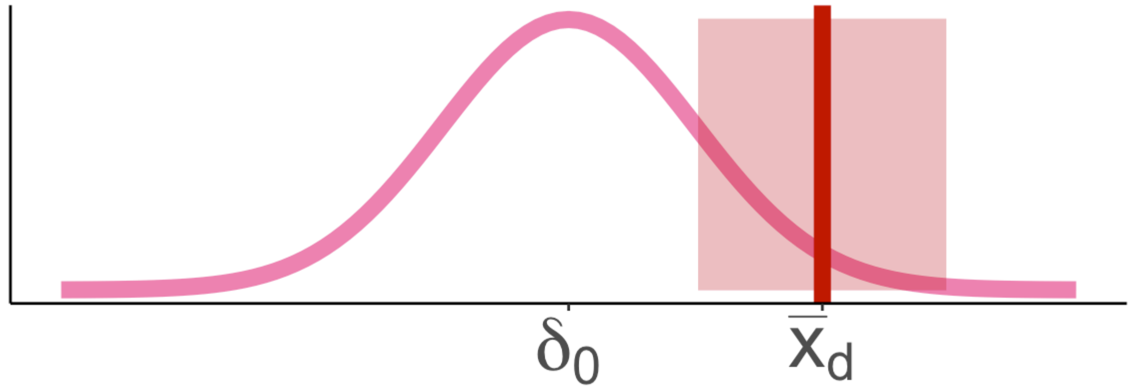
Paired mean difference:



Approaches to answer a research question

- **Research question is a generic form for paired data:** Is there evidence to support that the population mean difference is different than δ_0?

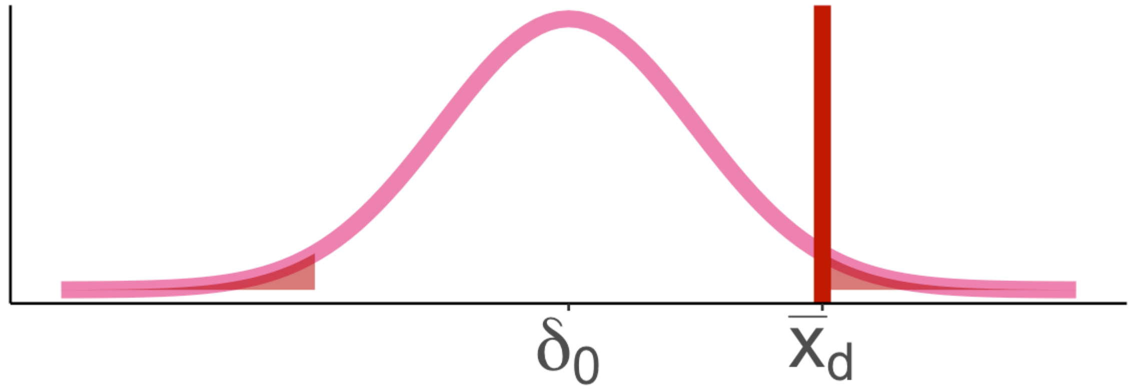
Calculate **CI for the mean difference** δ:



$$\overline{x}_d \pm t^* \cdot \frac{s_d}{\sqrt{n}}$$

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

Run a **hypothesis test:**



Hypotheses

$$H_0 : \delta = \delta_0$$
$$H_A : \delta \neq \delta_0$$
$$(or \quad <, >)$$

Test statistic

$$t_{\overline{x}_d} = \frac{\overline{x}_d - \delta_0}{\frac{s_d}{\sqrt{n}}}$$

Learning Objectives

1. Define paired data and explain how it differs from independent samples in the context of statistical analysis.
2. Construct confidence intervals for the mean difference in paired data and interpret these intervals in the context of the research question.
3. Perform the appropriate hypothesis test for paired data and interpret the results.

95% CI for the mean difference in cholesterol levels

```r
chol %>%
  select(DiffChol) %>%
  get_summary_stats(type = "common") %>%
  gt() %>% tab_options(table.font.size = 40)
```

variable	n	min	max	median	iqr	mean	sd	se	ci
DiffChol	43	-49	13	-23	16	-21.767	13.89	2.118	4.275

95% CI for population mean difference δ:

$$\bar{x}_d \pm t^* \cdot \frac{s_d}{\sqrt{n}}$$

$$-21.767 \pm 2.018 \cdot \frac{13.89}{\sqrt{43}}$$

$$-21.767 \pm 2.018 \cdot 2.118$$

$$-21.767 \pm 4.275$$

$$(-26.042, -17.493)$$

Used t^* = `qt(0.975, df=42)` = 2.018

Conclusion:
We are 95% confident that the (population) mean difference in cholesterol levels after a vegetarian diet is between -26.042 mg/dL and -17.493 mg/dL.

95% CI for the mean difference in cholesterol levels (using R)

- We can use R to get those same values

```r
t.test(x = chol$DiffChol, mu = 0)
```

```

	One Sample t-test

data:  chol$DiffChol
t = -10.276, df = 42, p-value = 4.946e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -26.04229 -17.49259
sample estimates:
mean of x 
-21.76744 
```

▶ We can tidy the output

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-21.76744	-10.27603	4.945625e-13	42	-26.04229	-17.49259	One Sample t-test	two.sided

Conclusion:
We are 95% confident that the (population) mean difference in cholesterol levels after a vegetarian diet is between -26.042 mg/dL and -17.493 mg/dL.

Poll Everywhere Question 2

Learning Objectives

1. Define paired data and explain how it differs from independent samples in the context of statistical analysis.
2. Construct confidence intervals for the mean difference in paired data and interpret these intervals in the context of the research question.
3. Perform the appropriate hypothesis test for paired data and interpret the results.

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 pairs**: Each pair is independent from all other pairs,
  - **Approximately normal sample or big n**: the distribution of the sample should be approximately normal, *or* the sample size 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 pairs of observations are independent from each other and the number of pairs in our sample is 43. Thus, we can use CLT to approximate the sampling distribution.

Step 2: Set the level of significance α

- **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 (1/2)

In statistics, a **hypothesis** is a statement about the value of an **unknown population parameter**.

A **hypothesis test** consists of a test between two competing hypotheses:

1. a **null** hypothesis H_0 (pronounced "H-naught") vs.
2. an **alternative** hypothesis H_A (also denoted H_1)

Example of hypotheses in words:

$$\begin{aligned} H_0 :& \text{The population mean difference in cholesterol levels after a vegetarian diet is zero} \\ \text{vs. } H_A :& \text{The population mean difference in cholesterol levels after a vegetarian diet is} \\ & \text{different than zero} \end{aligned}$$

1. H_0 is a claim that there is "no effect" or "no difference of interest."
2. H_A is the claim a researcher wants to establish or find evidence to support. It is viewed as a "challenger" hypothesis to the null hypothesis H_0

Step 3: Null & Alternative Hypotheses (2/2)

Notation for hypotheses (for paired data)

$$H_0 : \delta = \delta_0$$
$$\text{vs. } H_A : \delta \neq, <, \text{or}, > \delta_0$$

Hypotheses test for example

$$H_0 : \delta = 0$$
$$\text{vs. } H_A : \delta \neq 0$$

We call δ_0 the *null value* (hypothesized population mean difference from H_0)

$H_A : \delta \neq \delta_0$

- not choosing a priori whether we believe the population mean difference is greater or less than the null value δ_0

$H_A : \delta < \delta_0$

- believe the population mean difference is **less** than the null value δ_0

$H_A : \delta > \delta_0$

- believe the population mean difference is **greater** than the null value δ_0

- $H_A : \delta \neq \delta_0$ is the most common option, since it's the most conservative

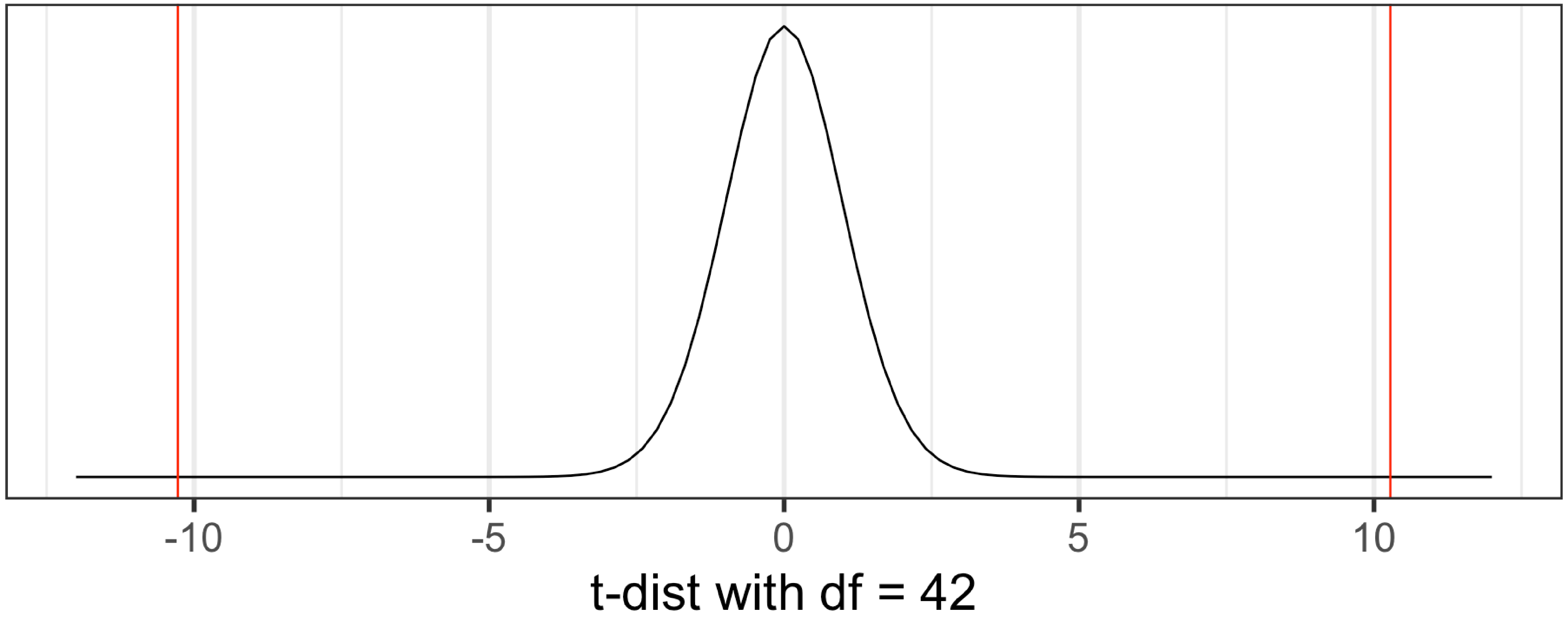
Step 4: Test statistic (where we do not know population sd)

From our example: Recall that $\overline{x}_d = -21.767$, $s_d = 13.89$, and $n = 43$

The test statistic is:

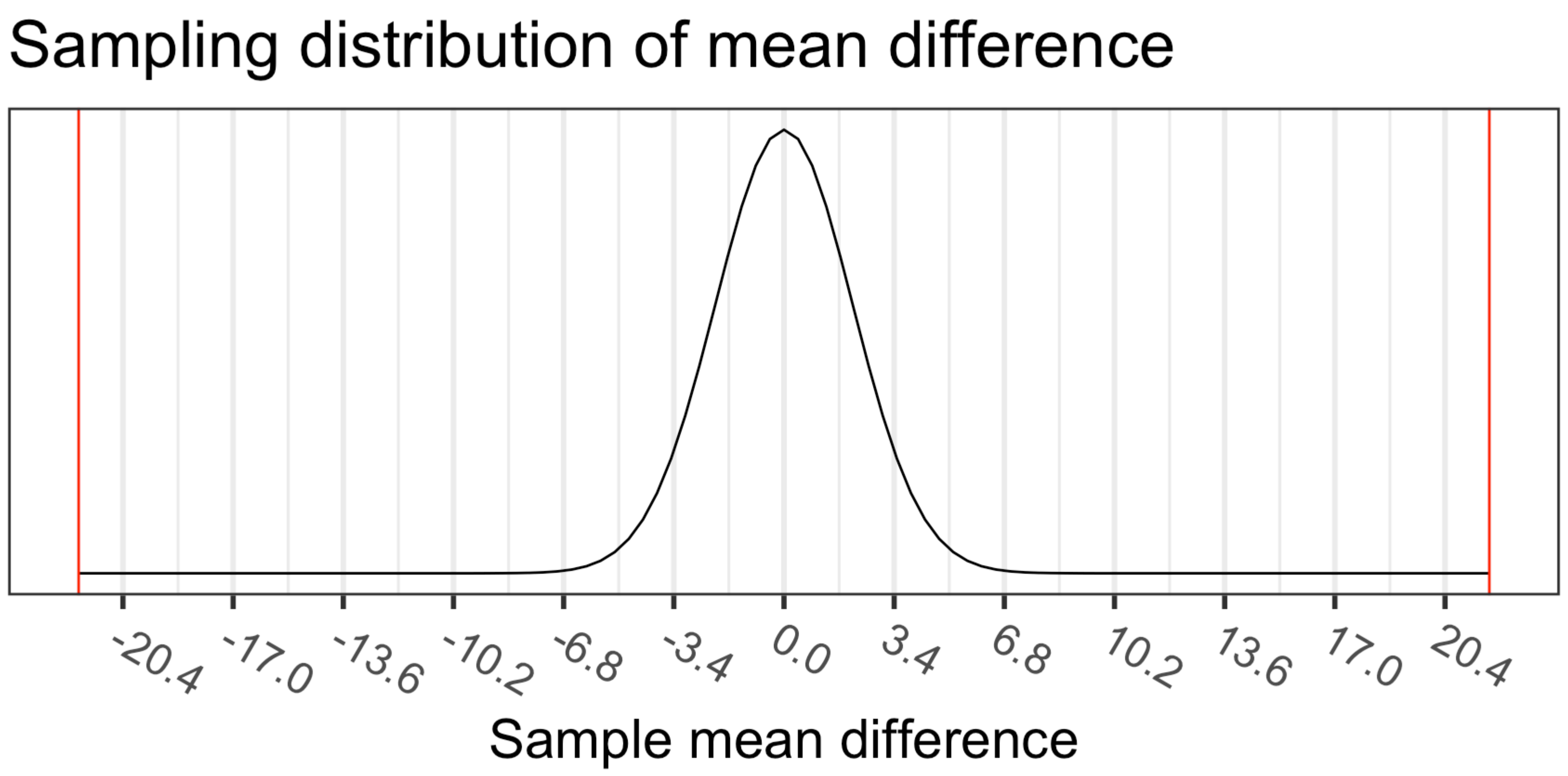
$$t_{\overline{x}_d} = \frac{\overline{x}_d - \delta_0}{\frac{s_d}{\sqrt{n}}} = \frac{-21.767 - 0}{\frac{13.89}{\sqrt{43}}} = -10.276$$

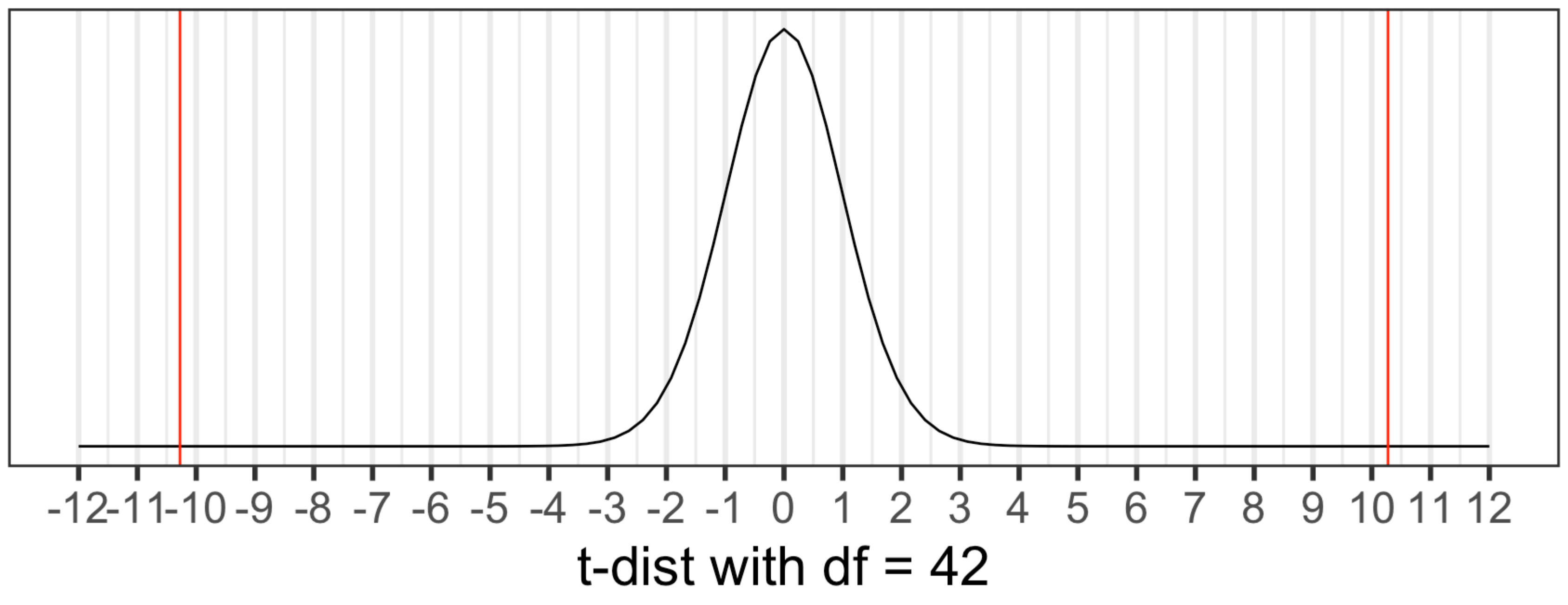
- Statistical theory tells us that $t_{\overline{x}}$ follows a **Student's t-distribution** with $df = n - 1 = 42$



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 $df = n - 1 = 43 - 1 = 42$:

$$\begin{aligned}\text{p-value} &= P(T \leq -10.276) + P(T \geq 10.276)\\ &= 4.946032 \times 10^{-13} < 0.001\end{aligned}$$

```r
2*pt(-10.276, df = 43-1,
      lower.tail = TRUE)
```

```
[1] 4.946032e-13
```

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

```r
t.test(x = chol$DiffChol, mu = 0)
```

```

	One Sample t-test

data:  chol$DiffChol
t = -10.276, df = 42, p-value = 4.946e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -26.04229 -17.49259
sample estimates:
mean of x 
-21.76744 
```

- We can "tidy" the results

```r
t.test(x = chol$DiffChol, mu = 0) %>% tidy() %>% gt() %>%
  tab_options(table.font.size = 40) # use a different size in your HW
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-21.76744	-10.27603	4.945625e-13	42	-26.04229	-17.49259	One Sample t-test	two.sided

Poll Everywhere Question 3

Step 6: Conclusion to hypothesis test

$$H_0 : \delta = \delta_0$$
$$\text{vs. } H_A : \delta \neq \delta_0$$

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

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

- There is sufficient evidence that the (population) mean difference is discernibly different from δ_0 (p-value = __)
- The mean difference (insert measure) in the sample was $\overline{x}_d$ (95% CI ___,___), which is discernibly different from δ_0 (p-value = ___).

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

- There is insufficient evidence that the (population) mean difference of (insert measure) is discernibly different from δ_0 (p-value = __)
- The average (insert measure) in the sample was $\overline{x}_d$ (95% CI ___,___), which is not discernibly different from δ_0 (p-value = ___).

Step 6: Conclusion to hypothesis test

$$\begin{aligned} H_0 &: \delta = 0 \\ \text{vs. } H_A &: \delta \neq 0 \end{aligned}$$

- Recall the p-value = $4.9456253 \times 10^{-13}$
- Use α = 0.05.
- Do we reject or fail to reject H_0?

Conclusion statement:

- Stats class conclusion (and good enough for our class!)
  - There is sufficient evidence that the (population) mean difference in cholesterol levels after a vegetarian diet is different from 0 mg/dL (p-value < 0.001).
- More realistic manuscript conclusion:
  - After a vegetarian diet, cholesterol levels decreased by on average 21.77 mg/dL (95% CI: 17.49, 26.04), which is discernably different than 0 (p-value < 0.001).

What if we wanted to test whether the diet *decreased* cholesterol levels?

Example of hypothesis test

Is there evidence to support that cholesterol levels decreased after the vegetarian diet?

1. **Assumptions:** The pairs of observations are independent from each other and the number of pairs in our sample is 43. Thus, we can use CLT to approximate the sampling distribution.

2. Set $\alpha = 0.05$

3. **Hypothesis:**

$$\begin{aligned} H_0 &: \delta = 0 \\ \text{vs. } H_A &: \delta < 0 \end{aligned}$$

4-5.

```r
chol_ttest <- t.test(x = chol$DiffChol, mu = 0, alternative = "less")
tidy(chol_ttest) %>% gt() %>% tab_options(table.font.size = 36)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-21.76744	-10.27603	2.472813e-13	42	-Inf	-18.20461	One Sample t-test	less

6. **Conclusion:** We reject the null hypothesis. There is sufficient evidence that cholesterol levels decreased with the vegetarian diet (p-value < 0.001).

Reference: Ways to run a paired t-test in R

R option 1: Run a 1-sample `t.test` using the paired differences

$H_A : \delta \neq 0$

```r
t.test(x = chol$DiffChol, mu = 0)
```

```

	One Sample t-test

data:  chol$DiffChol
t = -10.276, df = 42, p-value = 4.946e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -26.04229 -17.49259
sample estimates:
mean of x 
-21.76744 
```

Run the code without `mu = 0`. Do the results change? Why or why not?

R option 2: Run a 2-sample `t.test` with `paired = TRUE` option

$H_A : \delta \neq 0$

- For a 2-sample t-test we specify both `x=` and `y=`
- Note: `mu = 0` is the default value and doesn't need to be specified

```r
t.test(x = chol$Before, y = chol$After, mu = 0, paired = TRUE)
```

```

	Paired t-test

data:  chol$Before and chol$After
t = 10.276, df = 42, p-value = 4.946e-13
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 17.49259 26.04229
sample estimates:
mean difference 
       21.76744 
```

What is different in the output compared to option 1?