

Chapter 7: Comparing a Numerical Variable Across Two Groups

In this chapter, we will consider methods that allow us to make comparisons on numerical variables between two different groups. In general, these methods should be used to address research questions involving a categorical predictor variable (with two categories) and a numerical response variable.

✓ explanatory

Two different cases will be considered:

- (1) When the samples are independent, and
- (2) when the samples are dependent (i.e., matched or paired).

COMPARING TWO POPULATION MEANS: INDEPENDENT SAMPLES (two-sample independent t-test)

The hypothesis testing procedures presented in this section should be used when the observations from the two groups being compared are independent. Whether or not the observations are dependent is determined by how the data are collected. To see this, consider the following example.

Example 7.1: Nicotine Withdrawal and Perceived Elapsed Time

Recall the study conducted by researchers at Pennsylvania State University which investigated whether time perception, a simple indication of a person's ability to concentrate, is impaired during nicotine withdrawal. This study was discussed in Chapter 6. Twenty smokers were put through a 24-hour smoking abstinence and were asked to estimate how much time had passed during a 45-second period. Another 22 individuals who are non-smokers were recruited; they were also asked to estimate how much time had passed during a 45-second period. Suppose the resulting data on perceived elapsed time (in seconds) were analyzed as shown below (these results are artificial but are similar to the actual findings).

```
head(time_elapsed)

# A tibble: 6 x 4
  status     sex   age time_passed
  <chr>     <chr> <dbl>      <dbl>
1 non-smoker male    42       42.5
2 non-smoker male    40       41.2
3 non-smoker male    45       51.7
4 daily smoker male   31       59.9
5 daily smoker female 42       52.6
6 daily smoker male   32       42.8
```

Research Question: Do those smokers suffering from nicotine withdrawal tend to believe that more time has elapsed than non-smokers?

1. What is the observational unit for this study?

A person/subject

2. What are the variables assessed in this study? What are their roles (explanatory / response) and data types?

- Explanatory: status (smoker/non-smoker) -categorical
- Response: time passed (seconds) -numerical

Note that when the response variable is numerical and the explanatory variable is categorical (with two categories), a two-sample independent t-test is appropriate for testing for differences across the two groups.

3. What are the parameters of interest for this study?

$\mu_{\text{non-smokers}}$: The mean perceived time elapsed for all non-smokers.

$\mu_{\text{daily smokers}}$: The mean perceived time elapsed for all daily smokers suffering from nicotine withdrawal.

Similar to summarizing a single numeric variable, we can compare the two groups with summary statistics from the `favstats()` function.

favstats(time_passed ~ status, data = time_elapsed)										
	status	min	Q1	median	Q3	max	mean	sd	n	missing
1	daily smoker	39.12	48.7450	53.260	56.6575	69.37	52.6535	7.268202	20	0
2	non-smoker	30.86	41.2375	45.865	49.4625	61.21	45.2200	6.826466	22	0

4. Report the observed mean perceived time passed for study participants suffering from nicotine withdrawl. How about for non-smokers?

$$\bar{x}_{\text{smokers}} = 52.65 \text{ seconds}$$

$$\bar{x}_{\text{non-smokers}} = 45.22 \text{ seconds}$$

5. Calculate the observed difference in the mean perceived time elapsed between daily smokers and non-smokers.

$$\bar{x}_{\text{smokers}} - \bar{x}_{\text{non-smokers}} = 52.65 - 45.22 = 7.43 \text{ seconds}$$

↳ Note: this is estimating $M_{\text{smokers}} - M_{\text{non-smokers}}$

6. Let's refresh ourselves, what are the three types of plots used to plot a single numerical variable?

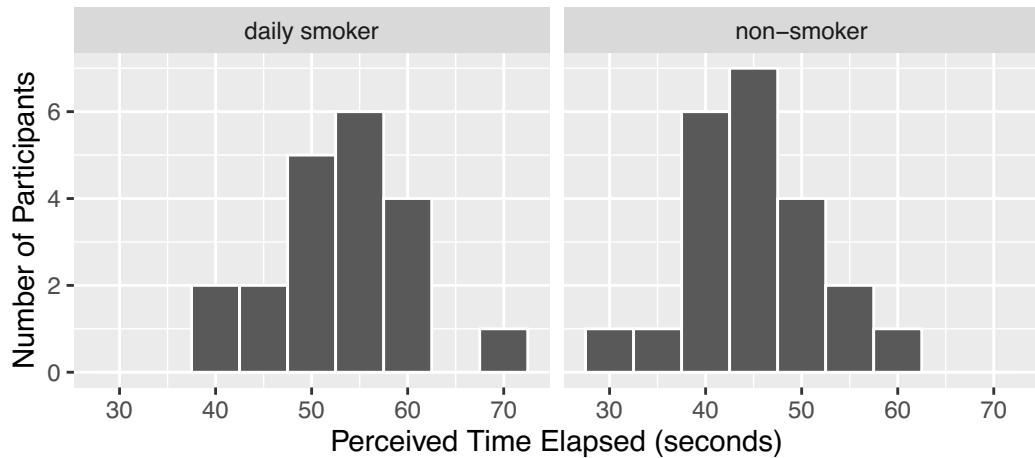
Histograms, dotplots, Boxplots.

Faceted Histograms

When we want to add a categorical variable (like ~~dict~~) to a histogram, we create separate plots for each level of the categorical variable. These separate plots are called facets. We are comparing the ~~cholesterol~~ levels for ~~corn flake~~ and ~~oat bran~~ diets, so we will have two facets, one per ~~dict~~ ~~time passed~~ ~~smokers~~ ~~non-smokers~~
~~status~~

```
ggplot(data = time_elapsed,
       mapping = aes(x = time_passed)) +
  geom_histogram(binwidth = 5, color = "white") +
  labs(x = "Perceived Time Elapsed (seconds)",
       y = "Number of Participants") +
```

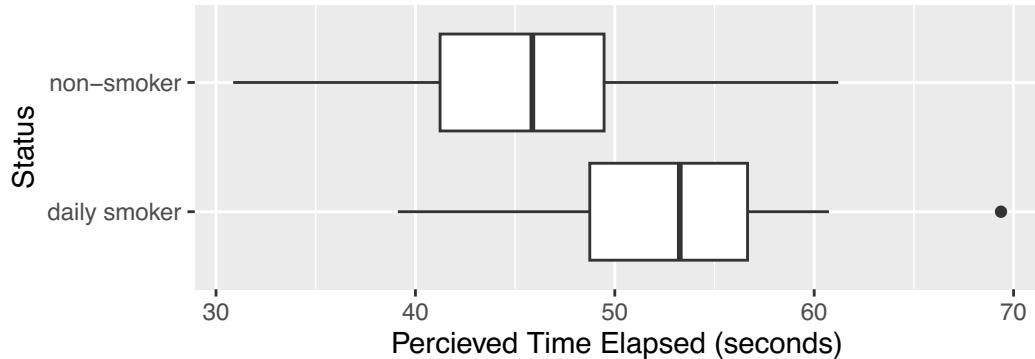
```
facet_wrap(~ status)
```



Side-by-Side Box Plots

Another way we can incorporate a categorical into our plots is to plot our boxplots for each group side-by-side. As opposed to faceting, these boxplots will be on the **same** plot. We only need to add one extra piece to our code from the previous chapter: a categorical variable.

```
ggplot(data = time_elapsed,
       mapping = aes(x = time_passed,
                      y = status)) +
  geom_boxplot() +
  labs(x = "Percieved Time Elapsed (seconds)",
       y = "Status")
```



Now that we have explored our data with summary statistics and visualizations, we want to use our data to draw inferences and make claims about the larger population.

Since the difference will change from sample to sample, in order to make valid inferences about the true population difference, we must first understand how the difference in sample means is expected to change from sample to sample. That is, we must determine what differences in means are likely to happen by chance when taking random samples from populations with the same mean.

14. Set up the null and alternative hypotheses (in words and symbols).

Null: The difference in mean perceived time elapsed between all daily smokers suffering from nicotine withdrawal and all nonsmokers is equal to 0.

$$H_0: \mu_s - \mu_{ns} = 0 \rightarrow \mu_s = \mu_{ns}$$

Alt: " is greater than.

$$H_A: \mu_s - \mu_{ns} > 0$$

In order to test our research question, we could conduct a simulation similar to what we did with two categorical variables (yawn experiment). Recall, we:

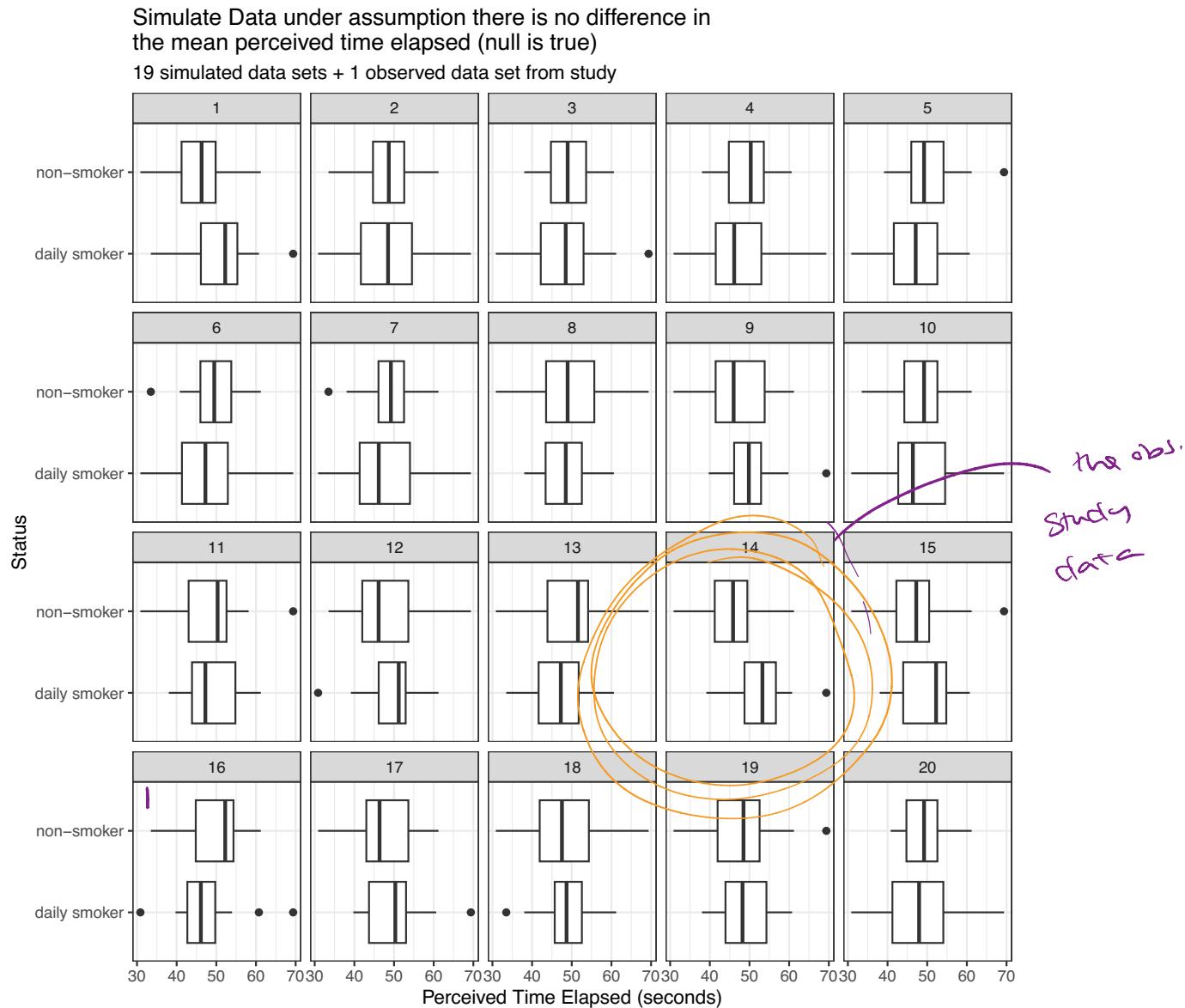
- Step 1: Write the status and time passed on 42 cards.
(explanatory) (response)
- Step 2: Simulate what could have happened if the null was true and ripped apart.
- Step 3: Generate a new data set by shuffling & randomly matching pairs of status & time passed.
- Step 4: Calculate the simulated difference in means for the new simulated data set and add it to the dot plot.

We would then repeat this process 100 or 1000 times to get an idea of what the sampling distribution of the difference in means looks like.

15. Assuming there is no difference in mean perceived time elapsed between the smokers and non-smokers, where do you expect the distribution to be centered? Explain.

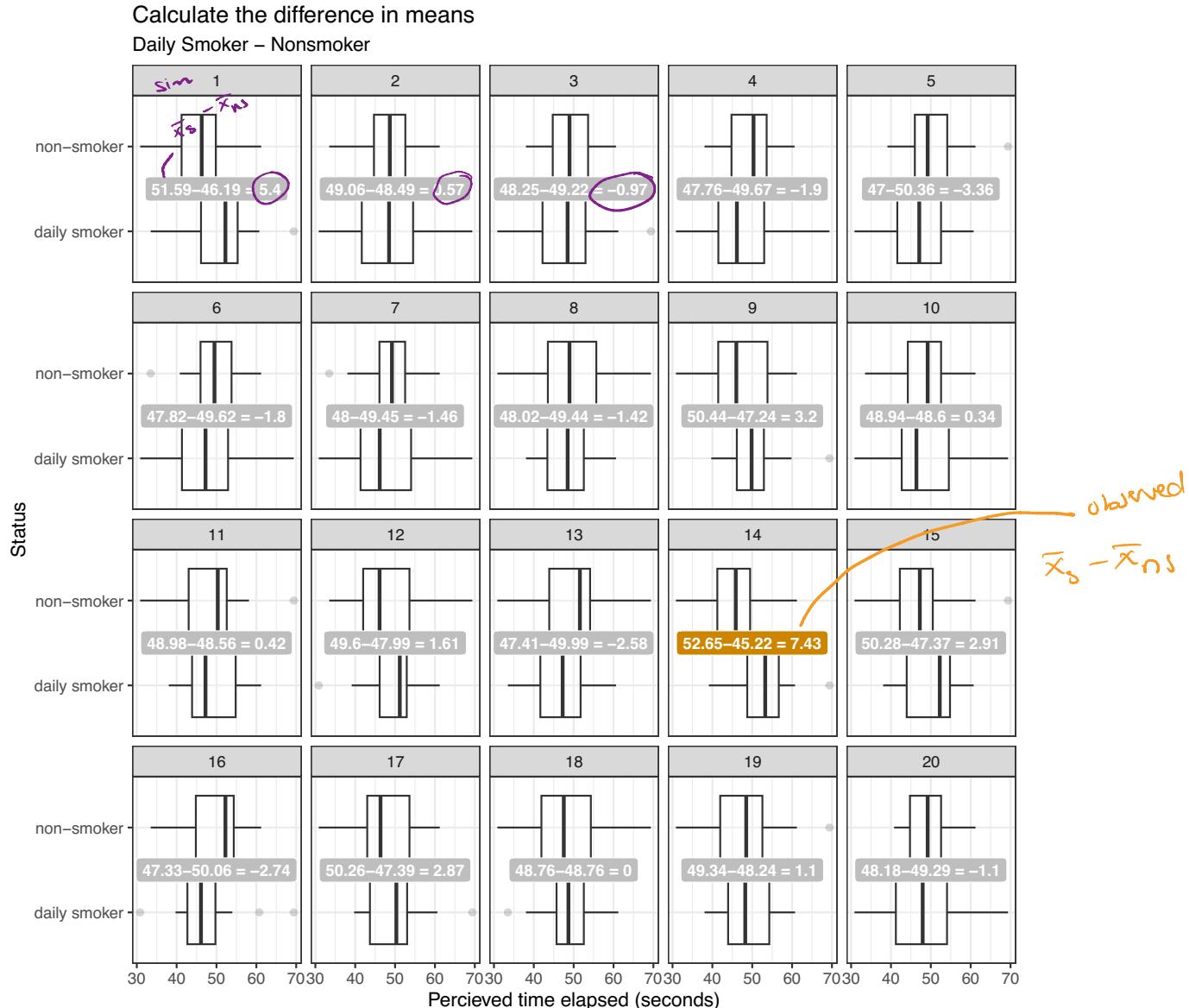
0, if no difference / relationship
(the null is true)

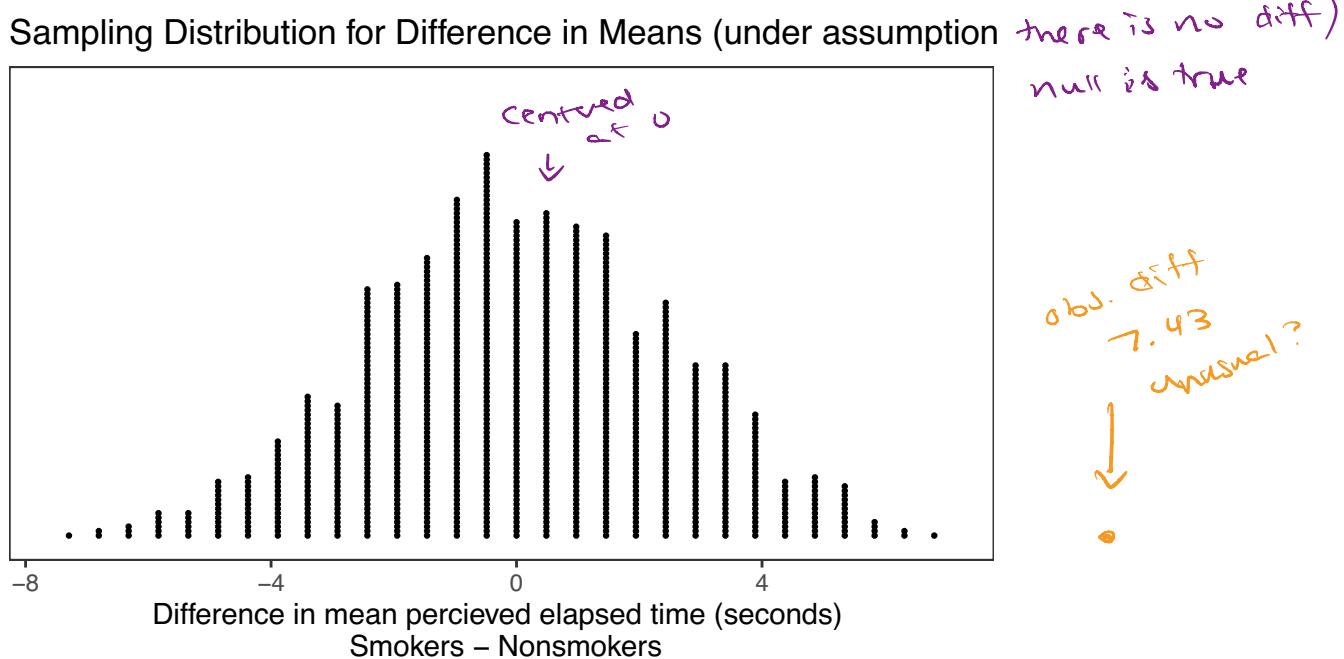
The plot below contains 19 panels of “new simulated” data sets (under the assumption there is no difference in mean perceived time elapsed between the two groups, following the process above) and one panel of the actual data observed from the study.



16. Which panel contains the actual data observed in the study? Was it hard to pick out? Remember what it was like trying to pick this out.

We could take these and calculate the observed difference in cholesterol levels for each panel and plot this to begin creating the distribution of our sampling difference in means.





17. Does our observed data appear to be unusual compared to the data simulated under the assumption there is no difference?

yes, our data seem unusual compared to the simulated data assuming no difference.

It turns out that these sample means vary in a predictable way, and we can use what we know about this predictable pattern to determine what outcomes are likely (or not likely) to happen by chance. This predictable pattern is called the **sampling distribution for the difference in means**.

i Characteristics for Sampling Distribution for Difference in Means

When the samples are **independent**, the sampling distribution for the difference in means can be described as follows.

1. mean = $\mu_1 - \mu_2$. *← from null (usually 0)*

2. standard deviation of the sampling distribution (i.e., standard error) =

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Previously
center: μ
sd: σ/\sqrt{n}

Now

Note: we are assuming the variances are unequal in this calculation, there are other formulas for assuming equal variances.

3. The shape of the sampling distribution is approximately normal if both sample sizes are **sufficiently large** (> 30 in each group) OR if both **original populations** are normally distributed.

Given these characteristics, the test statistic we will use when testing for differences in two population means for independent samples is as follows:

$$T = \frac{\text{observed value} - \text{null value}}{\text{standard error}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{typically 0}$$

These test statistics both follow a t-distribution, and the degrees of freedom are given by $(n_1 - 1) + (n_2 - 1)$.

Note: there are more complicated ways to adjust for these degrees of freedom, but that is beyond the scope of this class. You may notice that R outputs a different number of degrees of freedom than what this formula gives you.

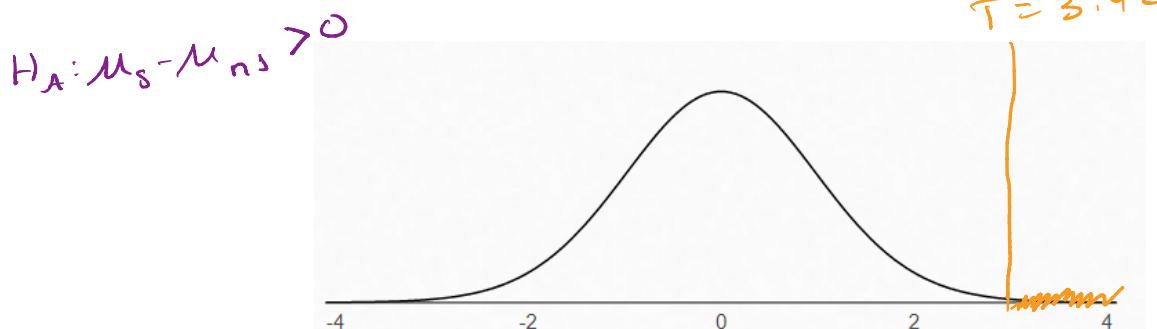
18. Calculate the observed T test-statistic for the study data.

$$T = \frac{(52.65 - 45.22) - 0}{\sqrt{\frac{7.26^2}{20} + \frac{6.82^2}{22}}} = 3.42$$

we assume no diff under H_0

T observed test stat

19. Using the t-distribution provided below, show how you would calculate/estimate the p-value.



Additionally, we can use R to conduct our two-sample independent t-test to answer our research question.

```
t_test(x = time_elapsed,
       explanatory = status,
       response = time_passed,
       mu = 0,
       alternative = "greater",
       conf_int = FALSE)

# A tibble: 1 x 5
  statistic  t_df  p_value alternative estimate
  <dbl>    <dbl>    <dbl>    <chr>      <dbl>
1      3.41    39.0  0.000768 greater     7.43
```

$T \uparrow$

$p < 0.001 \Rightarrow \text{strong evidence}$

20. Write the conclusion in context of the problem.

We have strong evidence to conclude the difference in mean perceived time elapsed between all daily smokers suffering from nicotine withdrawal and all nonsmokers is greater than 0 seconds. ($T = 3.41$, $dF = 39$, $p\text{value} < 0.001$).

i Conditions for two-sample independent t-test

1. Are the two groups independent?
2. Are both sample sizes sufficiently large? If not, is it reasonable to assume that both populations are normally distributed?

21. Check the conditions for using the two-sample independent t-test.

✓ **Independent Groups:** The perceived time passed for subjects in the smoking group doesn't tell us anything about the time passed for subjects in the non-smoking group.

✓ **Normality/Large Sample Size:** $n_g = 20 < 30$ & $n_ns = 22 < 30$ X
look at faceted histograms \Rightarrow approximately normal ✓ okay

22. Can we say that smoking causes a longer perceived elapsed time? Explain.

No, subjects were not randomly assigned to smoking status \Rightarrow association only!

Example 7.2: Diet and Cholesterol (Two-sample Independent t-test)

Researchers investigated whether eating corn flakes compared to oat bran had an effect on serum cholesterol levels. Twenty-eight (28) adults were randomly assigned a diet that included either corn flakes (14 individuals) or oat bran (14 individuals). After two weeks, cholesterol levels (mmol/L) of the participant were recorded.

```
head(cholesterol_data_long)
```

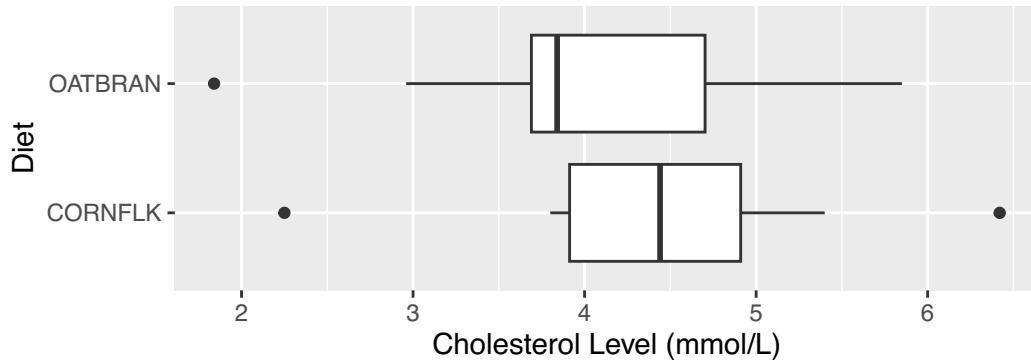
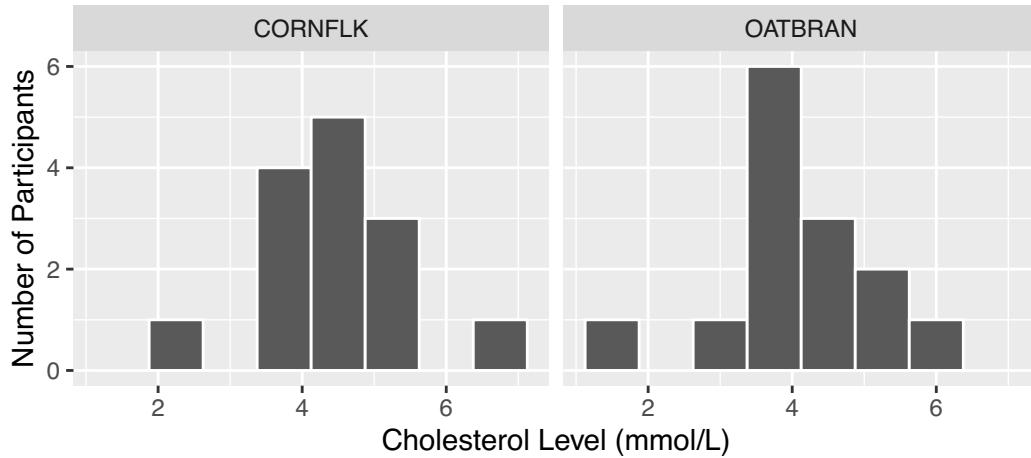
```
# A tibble: 6 x 2
  Diet      Cholesterol
  <chr>      <dbl>
1 CORNFLK      4.61
2 OATBRAN      3.84
3 CORNFLK      6.42
4 OATBRAN      5.57
5 CORNFLK      5.4 
6 OATBRAN      5.85
```

Research Question: Does eating corn flakes compared to oat bran have an effect on serum cholesterol levels? In other words, is there a difference in mean cholesterol levels between all adults on a corn-flake diet compared to all adults on an oat bran diet?

The summary statistics and visualizations for the study are provided below:

```
favstats(Cholesterol ~ Diet, data = cholesterol_data_long)
```

Diet	min	Q1	median	Q3	max	mean	sd	n	missing
1 CORNFLK	2.25	3.9125	4.44	4.9100	6.42	4.443571	0.9688344	14	0
2 OATBRAN	1.84	3.6900	3.84	4.7025	5.85	4.080714	1.0569802	14	0



1. What is the observational unit for this study?

An adult

2. What are the variables assessed in this study? What are their roles (explanatory / response) and data types?

- Explanatory: Diet (cornflake/oat bran) - categorical
- Response: Cholesterol (mmol/L) - numerical

3. What are the parameters of interest for this study?

$\mu_{\text{cornflakes}}$: The mean cholesterol for all adults on a cornflake diet.

$\mu_{\text{oat bran}}$: The mean cholesterol for all adults on an oat bran diet.

4. Report the observed mean cholesterol level for study participants on the corn flake diet. How about on the oat bran diet?

$$\bar{x}_{\text{CF}} = 4.44 \text{ mmol/L}$$

$$\bar{x}_{\text{OR}} = 4.08 \text{ mmol/L}$$

5. Calculate the observed difference in the mean cholesterol level between participants on the corn flake diet and participants on the oat bran diet.

$$\bar{x}_{\text{CF}} - \bar{x}_{\text{OR}} = 4.44 - 4.08 = 0.36 \text{ mmol/L}$$

6. Set up the null and alternative hypotheses, in symbols.

$$H_0: \mu_{\text{CF}} - \mu_{\text{OR}} = 0 \quad \mu_{\text{CF}} = \mu_{\text{OR}}$$

$$H_A: \mu_{\text{CF}} - \mu_{\text{OR}} \neq 0 \quad \mu_{\text{CF}} \neq \mu_{\text{OR}}$$

7. Check the conditions for using the two-sample independent t-test.

- Independent groups: The cholesterol for an adult on the LF diet does not tell us about the cholesterol for an adult on the oat bran.
- Normality assumption: $n_{LF} = 14 < 30 \Rightarrow$ check histograms of observed data for $n_{OB} = 14 < 30$ normality.

8. What do we know about the sampling distribution of the difference in means? From CLT

- Shape: \approx Normal

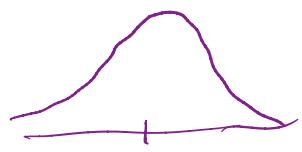
↑ under assumption null is true
(no diff).

- Mean: 0

- Standard Deviation (i.e., Standard Error):

$$\sqrt{\frac{0.97^2}{14} + \frac{1.05^2}{14}} = 0.38 \text{ Vmmol/L}$$

↑ SE

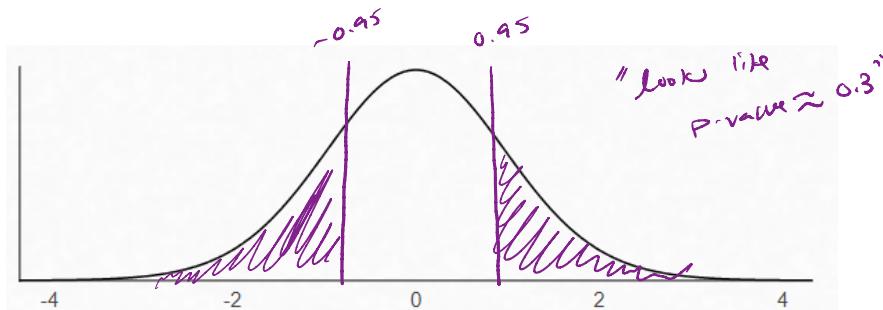


9. Calculate the observed T test-statistic for the data in the study.

$$\frac{0.38 - 0}{0.38} = 0.95$$

$$T = \frac{\text{obs} - \text{null}}{\text{SE}}$$

10. Using the t-distribution provided below, show how you would calculate/estimate the p-value.



Alternatively, we could conduct the two-sample independent t-test via R.

```
t_test(x = cholesterol_data_long,
       explanatory = Diet,
       response = Cholesterol,
       mu = 0,
       alternative = "two-sided",
```

```

conf_int = FALSE)

# A tibble: 1 x 5
statistic t_df p_value alternative estimate
<dbl> <dbl> <dbl> <chr> <dbl>
1 0.947 25.8 0.352 two.sided 0.363
    
```

11. Using the output above, write a conclusion in context of the study.

we do not have evidence to conclude the difference in mean cholesterol levels for all adults on the corn flake diet compared to the oat bran diet is different from 0 mmol/L ($t = 0.947$, $df = 25.8$, $p = 0.352$).

12. If we had found evidence of a difference in cholesterol levels between the two diets, could we say that cereal diet *causes* a change in cholesterol levels? Explain.

yes, diet was randomly assigned to subjects.

i Constructing a Confidence Interval for the Difference in Means

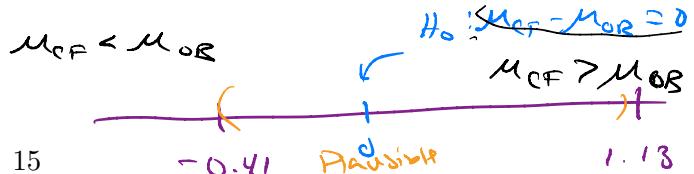
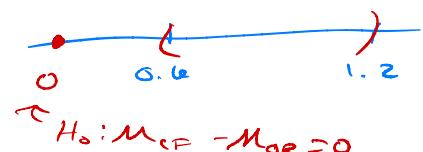
$$(\bar{x}_1 - \bar{x}_2) \pm t\text{-quantile} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{point estimate} \pm \text{mult} \times \text{SE}$$

13. Calculate the 95% confidence interval for the study using the t-quantile provided below:

$qt(0.975, df = 40)$

[1] 2.021075

- Lower Endpoint = $0.36 \pm 2.02 (0.38) = [-0.4076, 1.1276]$



$$M_{CF} - M_{OB}$$

- Upper Endpoint =

We are 95% confident the mean cholesterol level for all adults on the cornflake diet is between 0.41 mmol/L lower and 1.13 mmol/L higher than the mean cholesterol level for all adults on the oat bran diet.

14. Interpret the meaning of this 95% confidence interval in context of the study.

We are 95% confident the difference in mean cholesterol levels for all adults on the cornflake diet compared to oat bran diet is between -0.41 and 1.13 mmol/L .

15. What does it mean for 0 to fall within the confidence interval?

0 is a plausible difference in cholesterol levels (i.e. it is plausible the cholesterol levels are the same). Recall, "no evidence"

COMPARING TWO POPULATION MEANS: DEPENDENT SAMPLES (paired t-test)

The hypothesis testing procedures presented in this section should be used when the observations from the two groups being compared are dependent. Whether or not the observations are dependent is determined by how the data are collected. To see this, consider the following example.

Example 7.3: Diet and Cholesterol (Paired t-test)

Recall, in **Example 7.2**, researchers investigated whether eating corn flakes compared to oat bran had an effect on serum cholesterol levels. Twenty-eight (28) individuals were randomly assigned a diet that included either corn flakes (14 individuals) or oat bran (14 individuals). After two weeks, cholesterol levels (mmol/L) of the participant were recorded.

But actually what happened was...

Fourteen (14) individuals were randomly assigned a diet that included either oat bran or corn flakes. After two weeks on the initial diet, serum cholesterol were measured and the participants were then crossed-over to the alternate diet. After two-weeks on the second diet, cholesterol levels were once again recorded.

1. How was the data collected differently from the previous activities?

Each subject was on each diet.

2. Are the cholesterol levels independent across the 28 measurements? Explain.

*No, there were only 14 subjects and
one person's cholesterol on a cornflake diet
will probably be related to the same person's
cholesterol on the oat bran diet.*

Since we can match or pair the cholesterol level measurements between the treatment groups, we call this a **paired study design**. These could be paired up by taking measurements on the same observational unit twice or taking measurements on similar observational units (e.g. identical twins, similar fields, etc.). When analyzing the data and making conclusions, we must take this design into consideration or we are violating a key assumption (independence).

The data set below shows the data for each participant ID. The CORNFK column indicates the cholesterol level for that individual on the corn flake diet and the OATBRAN column indicates the cholesterol level for that individual on the oat bran diet.

```
head(cholesterol_data)
```

ID	CORNFLK	OATBRAN	CholesterolDiff
			<dbl>
1 1	4.61	3.84	0.770
2 2	6.42	5.57	0.85
3 3	5.4	5.85	-0.450
4 4	4.54	4.8	-0.260
5 5	3.98	3.68	0.300
6 6	3.82	2.96	0.86

3. What can be said about the Cholesterol of Subject ID #1 as compared to Subject ID #2, for example, regardless of which diet the Cholesterol levels were measured after?

Subject #1 has lower cholesterol levels overall.

Big Idea

For these data, the first cholesterol level on the cornflake diet is related to the first cholesterol level on the oat bran diet (the two measurements were made on the same person). Thus, these two samples are **dependent**.

In other words, much of the variability in the observations is due to differences between people. So, to control for this variability in weights from person to person (which will help us isolate the effect of diet), we will work with the **DIFFERENCES** on each subject, instead. This will remove the structure of dependence between the cornflake and oat bran groups and will control for the fact that some people, in general, tend to have higher (or lower) cholesterol levels than others.

4. What does the `CholesterolDiff` column represent? Why does this calculation make sense?

The change in cholesterol between the two diets for each subject.

5. What does a positive difference of 0.77 indicate?

The subject's cholesterol was higher on the cornflake diet.

6. What does a negative difference of -0.45 indicate?

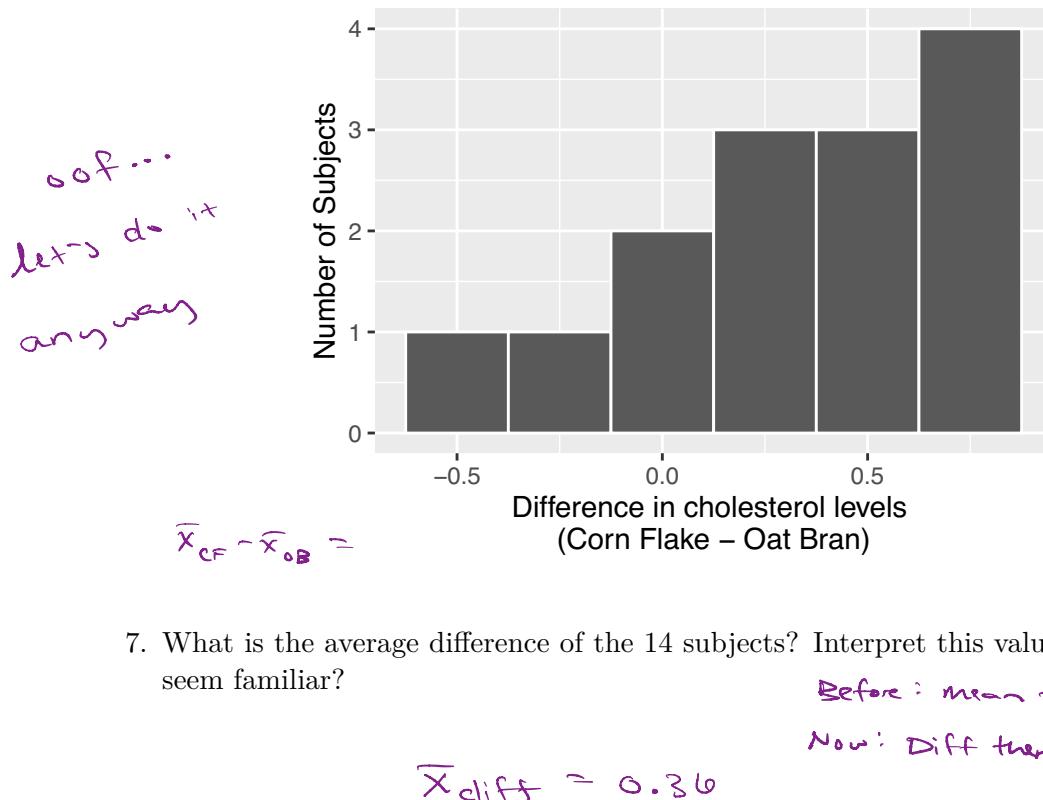
The subject's cholesterol was higher on the oat bran diet.

A histogram of the differences in cholesterol levels for the 14 individuals and a table of summary statistics are shown below.

Summary statistics for the difference in cholesterol levels between the cornflake diet and oat bran diet are given below:

```
favstats(~ CholesterolDiff, data = cholesterol_data)
```

min	Q1	median	Q3	max	mean	sd	n	missing
-0.45	0.115	0.36	0.695	0.86	0.3628571	0.4059638	14	0
					\bar{x}_{diff}	s_{diff}	n	



7. What is the average difference of the 14 subjects? Interpret this value. Does this value seem familiar?

Before: mean then diff

Now: Diff then mean

$$\bar{x}_{\text{diff}} = 0.36$$

8. If cereal diet had no effect on cholesterol levels, what would you expect these differences to be, on average?

0

Big Idea: paired t-test

Note that these differences are represented by a single column of data. So, instead of viewing this as a problem involving a categorical predictor and a numerical response, you could view this as a problem involving a single numerical variable – the differences! Therefore, the hypothesis testing procedure is exactly the same as the procedure for testing a **single population mean** we discussed in Chapter 6.

That is, the parameter of interest is the true population mean of the differences which we will represent by $\mu_{\text{difference}}$.

- Our best estimate for this parameter is the sample mean of the observed differences. We'll call this quantity $\bar{x}_{\text{difference}}$.
- The sample standard deviation of the differences will be denoted by $s_{\text{difference}}$.

We can carry out the hypothesis test as follows to determine whether cholesterol levels of all adults tend to *differ* between cereal diets.

Research Question Researchers are still interested in whether eating corn flakes compared to oat bran had an effect on serum cholesterol levels.

9. Write the parameter of interest in words.

The **mean difference in cholesterol levels for all adults on the cornflake diet compared to the oat bran diet.** (μ_{diff})

10. Set up the null and alternative hypotheses in words and symbols.

Before: mean then diff

$$H_0: \underline{\mu_{\text{diff}} = 0}$$

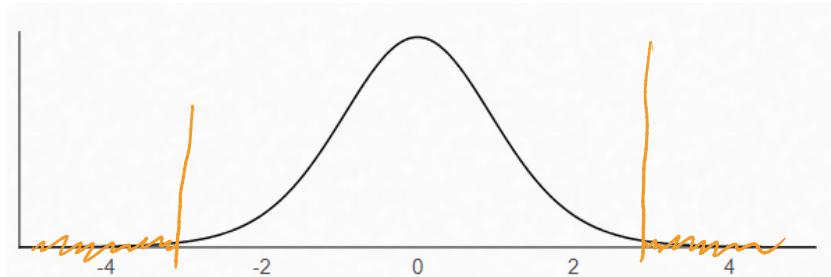
Now: diff then mean

$$H_A: \underline{\mu_{\text{diff}} \neq 0}$$

11. Calculate the T test-statistic:

$$T = \frac{\bar{x}_{\text{difference}} - \mu_{\text{difference}}}{s_{\text{difference}}/\sqrt{n}} = \frac{(0.34 - 0)}{0.404/\sqrt{14}} = 3.31$$

12. Using the t-distribution provided below, show how you would calculate/estimate the p-value.



Similar to Chapter 6, the R code below can be used to conduct the paired t-test:

```
t_test(x = cholesterol_data,
       response = CholesterolDiff,
       mu = 0,
       alternative = "two-sided")

# A tibble: 1 x 7
  statistic  t_df p_value alternative estimate lower_ci upper_ci
  <dbl> <dbl> <dbl> <chr>      <dbl>    <dbl>    <dbl>
1     3.34     13 0.00528 two.sided     0.363    0.128    0.597
```

13. State a conclusion in context of the study.

We have strong evidence the mean difference in cholesterol levels between the cornflake diet & oat bran diet is greater than 0 mmol/L for all subjects.

14. How much evidence does this provide for a change in cholesterol level due to diet? How does this differ from **Example 7.2** when we assumed independent samples?

Recall 7.2: $t = 0.947$, $p = 0.352 \Rightarrow$ no evidence

We found strong evidence in the paired analysis with the same data analyzed before. It is important to analyze data in the way it is collected.

15. Can we conclude that the diet caused the change? Explain.

yes, they randomly assigned the order of diet to subjects

i Checking the Normality Assumption for a Paired t-test

Recall that for the t-test to be valid, at least one of the following conditions must be met:

- Either the sample size is sufficiently large (greater than 30 or so), or
- The distribution of the observed data is approximately normal (which would indicate that the population is normally distributed so that the Central Limit Theorem would apply even with a small sample size)

16. Does the t-test appear to be a valid approach for testing this research question? Justify your reasoning.

17. Using the formula for confidence intervals for single numerical variables from Chapter 6 and the t-quantile provided below, calculate the 90% confidence interval.

```
qt(0.95, 13)
```

```
[1] 1.770933
```

$$\bar{x}_{\text{difference}} \pm \text{t-quantile}\left(\frac{s_{\text{difference}}}{\sqrt{n}}\right) =$$

18. Where does 0 fall within the confidence interval? Why does this make sense?