Activity 6B: Cholesterol III
Inference for Paired Data

Learning outcomes

- Given a research question involving paired differences, construct the null and alternative hypotheses in words and using appropriate statistical symbols.

- Describe and perform a simulation-based hypothesis test for a paired mean difference.

- Interpret and evaluate a p-value for a simulation-based hypothesis test for a paired mean difference.

- Use bootstrapping to find a confidence interval for a paired mean difference.

- Interpret a confidence interval for a paired mean difference.

- Use a confidence interval to determine the conclusion of a hypothesis test.

Review from Activity 6A: Cholesterol II

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.

Vocabulary Review

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

The data were collected differently because information was measured from each individual twice (once on the corn flake diet and once on the oat bran diet). Previously, the two diet groups were separate.

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

No, the cholesterol levels are not independent across the 28 measurements because only 14 individuals were used, meaning the cholesterol level of that individual on the corn flake diet likely tells us something about their cholesterol level 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 `CORNLK` 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.

ID	CORNFLK	OATBRAN	Cholesterol_Diff
1	4.61	3.84	0.77
2	6.42	5.57	0.85
3	5.40	5.85	-0.45
4	4.54	4.80	-0.26
5	3.98	3.68	0.30
6	3.82	2.96	0.86

3. What does the `Cholesterol_Diff` column represent? Why does this calculation make sense?

The `Cholesterol_Diff` column is the difference in cholesterol levels between corn flake and oat bran for each individual. This makes sense because we are interested in how diet changes an individual's cholesterol level.

Ask a Research Question

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

4. Write the null and alternative hypothesis in **words**.

Null: The true mean difference in cholesterol level between corn flake and oat bran is equal to 0.

Alternative: The true mean difference in cholesterol level between corn flake and oat bran is not equal to 0.

5. Write the null and alternative hypothesis in **notation**.

$H_O : \mu_{\text{CORNFLK - OATBRAN}} = 0$

$H_A : \mu_{\text{CORNFLK - OATBRAN}} \neq 0$

Summarize and Visualize the Data

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

```
ggplot(data = cholesterol_data,
       mapping = aes(x = Cholesterol_Diff)) +
  geom_histogram(binwidth = 0.5) +
  labs(x = "Difference in Cholesterol Levels (Cornflake - Oatbran)")
```
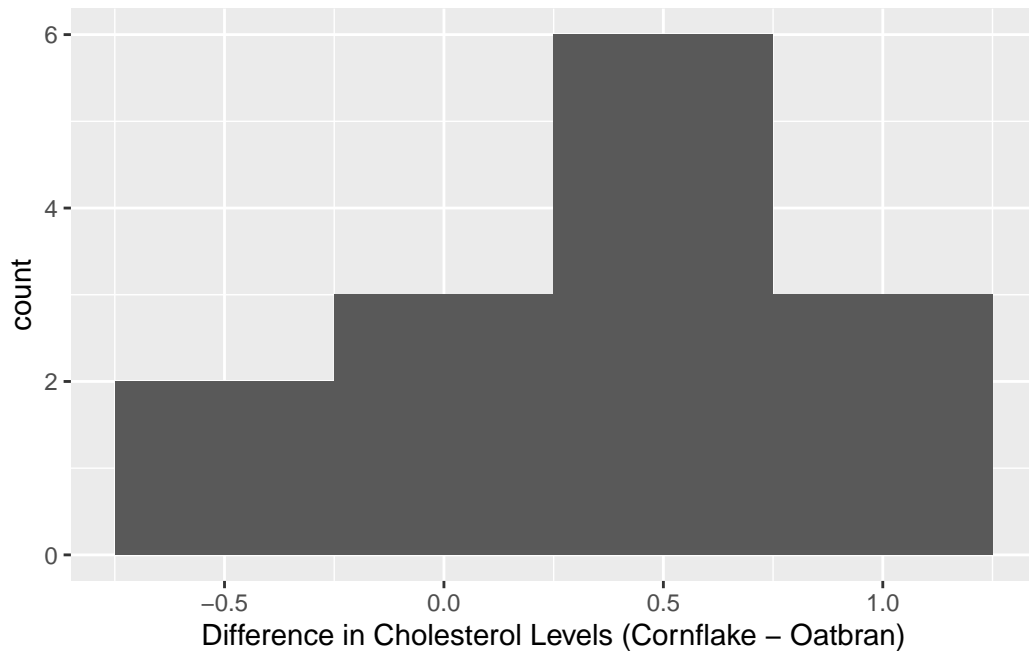
Summary statistics for the cholesterol levels from the corn flake diet `CORNFLK`, cholesterol levels from the oat bran diet `OATBRAN`, and the differences in cholesterol levels between the corn flake and oat bran diets `CholesterolDiff`.

```
        Diet   min     Q1 median     Q3  max      mean        sd  n missing
...1 CORNFLK  2.25 3.9125   4.44 4.9100 6.42 4.4435714 0.9688344 14       0
...2 OATBRAN  1.84 3.6900   3.84 4.7025 5.85 4.0807143 1.0569802 14       0
...3    DIFF -0.45 0.1150   0.36 0.6950 0.86 0.3628571 0.4059638 14       0
```

8. Report the observed statistic of interest (mean difference) for the data. Use notation to assign a symbol to this.

$\bar{x}_{\text{CORNFLK - OATBRAN}} = 0.362$

Use Statistical Methods to Draw Inferences from the Data

Hypothesis Test

To simulate the null distribution of paired sample mean differences we need to randomly assign the cholesterol level for each participant to be either the corn flake or the oat bran cholesterol level.

Take for example participant ID 1. On the corn flake diet, their cholesterol level was 4.61 mmol/L and on the oat bran diet, it was 3.84 mmol/L. If we assume the null is true, that the mean of the differences is 0, then there is no relationship between participant 1's cholesterol level and the diet they were on. That means, Participant 1 would have been just as likely to see a cholesterol level of 3.84 mmol/L on the corn flake diet.

9. How can we use a coin to decide which values for each participant are randomly assigned to the corn flake diet?

Flip the coin for each individual to determine which measurement should be assigned to which diet. For example: heads = corn flake and tails = oat bran; we could flip the coin for the first measurement and assign it to that diet. For Participant ID 1, we could flip a coin and if it lands tails we assign 4.61 mmol/L to the oat bran diet and 3.84 mmol/L to the corn flake diet.

A simulated null distribution is shown below.

```
null_dist <- cholesterol_data %>%
  specify(response = Cholesterol_Diff) %>%
  hypothesise(null = "point", mu = 0) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

null_dist %>%
  visualise() +
  labs(title = "",
       x = "Simulated Mean Difference (Cornflake - Oatbran)")
```
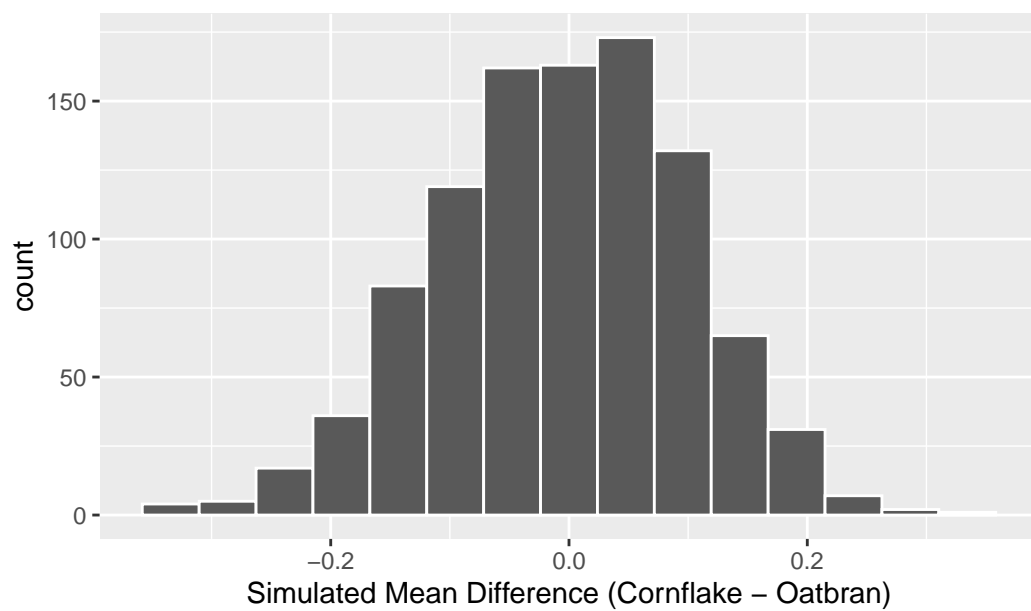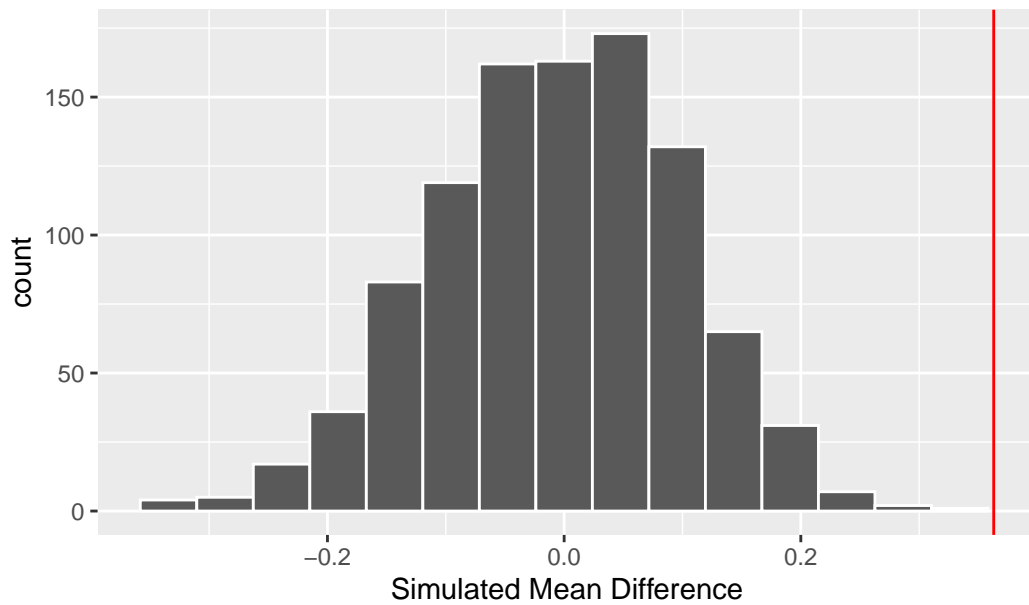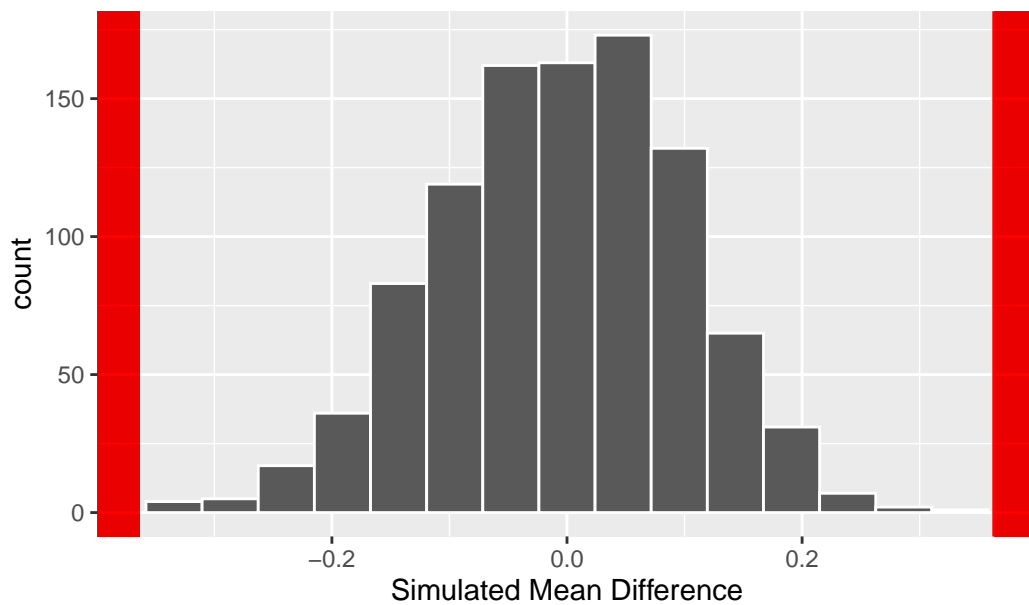
10. Explain why the null distribution is centered at zero.

11. Find the observed **mean difference** on the distribution. *Hint: Look at question 8.*

$\bar{x}_{\text{CORNFLK - OATBRAN}} = 0.363$

12. Shade the area of the distribution you will use to calculate the p-value and estimate the p-value.

15. How much evidence does this provide for a change in cholesterol level due to diet? How does this differ from the previous activities?

Our shaded region is on the tail ends of the simulated null distribution, indicating that our observed mean difference is *very* unlikely if the null was true. Thus, we have strong evidence against the null hypothesis and conclude there is a change in cholesterol level due to diet. In the previous activities we found no evidence of a change in cholesterol levels due to diet (when we analyzed these as two separate groups)! This difference is because of the smaller standard error of the mean difference ($\frac{s}{\sqrt{n}}$) compared to a difference in means ($\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$).

16. Can we conclude that the diet **caused** the change? Explain.

Yes, the order of the diet was randomly assigned to participants in the study. Therefore, we can determine causation!

Confidence Interval

The goal of a confidence interval is to estimate a plausible range of values for the population parameter.

17. What is the population parameter in this study?

$\mu_{CORNFLK-OATBRAN}$: The true mean difference in cholesterol levels between a corn flake and an oat bran diet for all individuals.

To create a bootstrap distribution, we **do not** assume the null hypothesis is true. Instead, we assume that the difference in cholesterol levels for these 14 individuals are representative of the differences in cholesterol levels for other individuals. So, we will randomly sample, with replacement, from our original sample to obtain a bootstrap sample.

I've created a bootstrap distribution below, using 1000 reps.

```
boot_dist <- cholesterol_data %>%
  specify(response = Cholesterol_Diff) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

boot_dist %>%
  visualise() +
  labs(title = "",
       x = "Bootstrap Mean Difference")
```
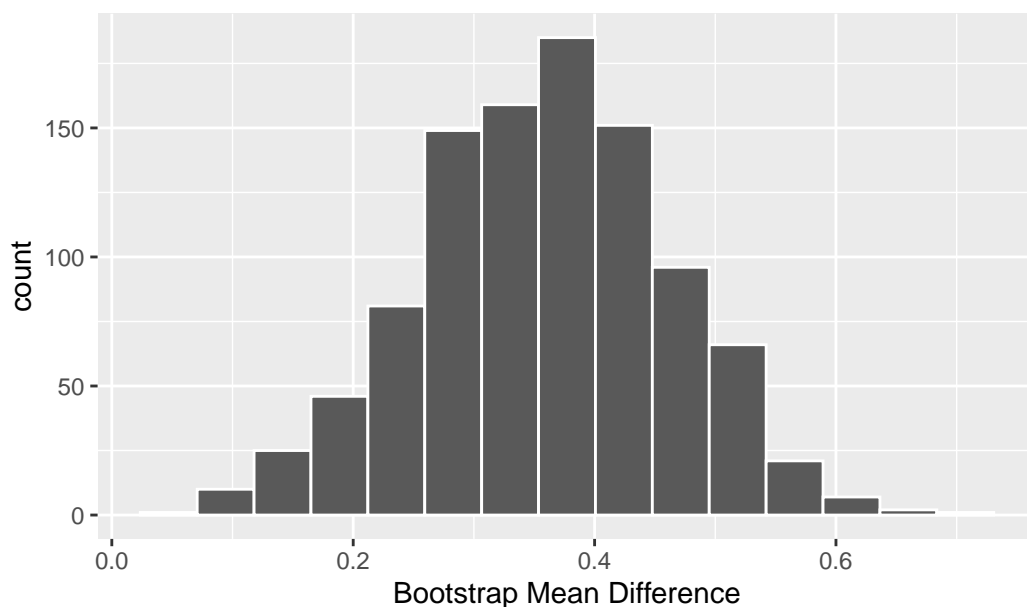
18. Where is the bootstrap distribution be centered? Why is it centered there?

The bootstrap distribution is centered at the original observed mean difference $\bar{x}_{CORNFLK-OATBRAN} = 0.363$. This makes sense because this value is our point estimate / best guess for our parameter.

19. Use the table below to find a 99% confidence interval for μ_{diff}.

To construct a 99% confidence interval, we select our 0.5th and 99.5th quantile values so there is 0.5% in both tails and 99% in the middle of our distribution: (0.184, 0.616).

Quantile	Value
0.5%	0.184
1%	0.118
2.5%	0.150
5%	0.184
90%	0.494
95%	0.526
97.5%	0.549
99.5%	0.616

Communicate the Results and Address the Research Question

20. Interpret the 99% confidence interval in the context of the problem.

We are 99% confident, the true difference in mean cholesterol levels between the corn flake diet and the oat bran diet for all individuals is between 0.108 mmol/L and 0.616 mmol/L.

21. Do the results of your confidence interval and hypothesis test agree? What does each tell you about the null hypothesis?

Yes! The results of the confidence interval agree with the hypothesis test since the null value of 0 is not included in the interval. So, we would reject the null hypothesis based on this confidence interval, which is what we did in #15.