Activity 6A: Cholesterol II

Confidence Intervals for Two Independent Means

Learning outcomes

Part 1: Confidence Interval for Two Independent Means

- Use bootstrapping to find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.
- Use a confidence interval to determine the conclusion of a hypothesis test.

Review from last week

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.

Last week we used cards to simulate what differences in mean cholesterol levels we might have seen if the null hypothesis was true. However, today we're not interested in deciding if we believe the means of these two groups are similar or different.

Today, we are interested in estimating what range of values the **true** difference in means might take on.

1. Based on the p-value you obtained from *Activity 5: Cholesterol I*, do you believe 0 is a plausible value for $\mu_{\text{CORNFLK}} - \mu_{\text{OATBRAN}}$?

Confidence interval

A **confidence interval** represents a range of plausible values for a population parameter. In this case, our population parameter is $\mu_{\text{CORNFLK}} - \mu_{\text{OATBRAN}}$, or the true difference in mean cholesterol levels between corn flake and oatbran diets.

The best way to estimate what range of values a parameter might have is to go out and collect more samples. However, that is often not feasible. So, instead we mimic this process by *resampling with replacement* from our original sample. Remember, this process has a name—**bootstrapping**.

Bootstrapping cholesterol & diets

When bootstrapping with two groups, we're assuming that the sample within each group is *representative* of other possible values in the population. Here, we are assuming that the participants included in our sample are representative of the cholesterol of other individuals who might have been on these diets.

Because we **are not** assuming the null is true (that there is no difference in the means of these two groups), we **do not** combine the groups together. Rather, we keep the groups separate and sample from each group separately.

2. Let's walk through how we would carry out this process:

Step 1:

Step 2:

Step 3:

Step 4:

3. What statistic do we have after step 4?

4. Once we create a bootstrap distribution of 1000 simulations, at what value do you expect the distribution to be centered? Explain your reasoning.

Creating a bootstrap distribution in R

We will use the **infer** package (again) to make our bootstrap distribution. The process we used for this situation will look very similar to before, since all we are changing is the statistic we calculate!

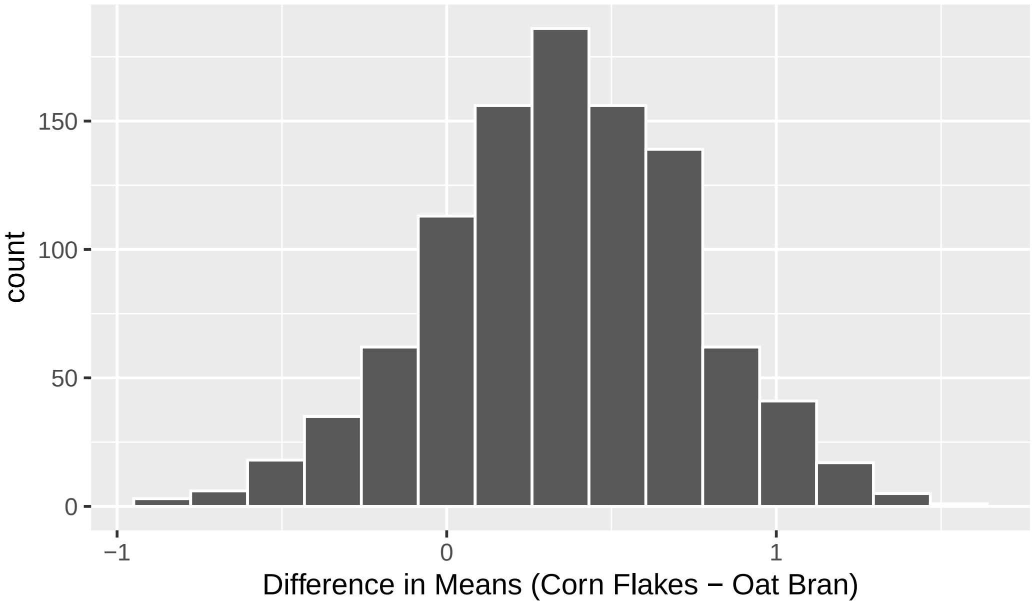
5. Fill in the blanks for the code below.

```
cholesterol_data_long %>%

  specify(response = __________________, explanatory = _______________) %>%

  generate(reps = _______________, type = _______________) %>%

  calculate(stat = "diff in means",
            order = c("CORNFLK", "OATBRAN")
            )
```

6. What is the difference between this code and the code to generate a null distribution (what we did in *Activity 5: Cholesterol I*)?

Obtaining a confidence interval

A bootstrap distribution from 1000 reps is plotted below.



7. Where is this distribution centered? Why does this make sense? How does it compare to our Null Distribution from the previous activity?

8. What are the two ways we could use this distribution to obtain a confidence interval?

Percentile method

I've provided a table of different percentiles to help you create your confidence interval.

Quantile	Value
0.5%	-0.298
1%	-0.572
2.5%	-0.446
5%	-0.298
90%	0.824
95%	0.992
97.5%	1.097
99.5%	0.992

9. Suppose we are interested in constructing a 95% confidence interval. Using the table above, report the end points of this confidence interval.

10. Interpret the confidence interval in the context of this investigation.

SE method

A percentile confidence interval uses **only** the bootstrap distribution. The SE method, on the other hand, uses information from **both** the bootstrap distribution and the *t*-distribution.

Because this method uses a *t*-distribution it should only be used **if the bootstrap distribution is bell-shaped and symmetric**.

11. Do you believe this condition is violated?

Alright, let's see how this confidence interval works. Our formula looks like this:

$$(\bar{x}_{\text{CORNFLK}} - \bar{x}_{\text{OATBRAN}}) \pm t^*_{df} \times SE_{boot}$$

There are three pieces to the interval:

- the observed statistic $(\bar{x}_{\text{CORNFLK}} - \bar{x}_{\text{OATBRAN}})$
- the t-distribution multiplier (t^*_{df})
- the standard error from the bootstrap distribution (SE_{boot})

12. What is the observed statistic (aka point estimate) for this investigation?

13. Using the table below, what is the standard deviation for the bootstrap distribution (aka the estimated standard error for the difference in mean cholesterol levels)?

```r
favstats(~stat, data = bootstrap_dist)
```

```
       min       Q1    median       Q3      max      mean        sd    n
-0.8588718 0.103817 0.3650861 0.618274 1.557857 0.3538702 0.3847116 1000
missing
      0
```

14. Using the table below, circle the correct multiplier we should use to make our interval.

`R` code	Value
`qt(0.90, df = 12)`	1.3562173
`qt(0.90, df = 13)`	1.3501713
`qt(0.90, df = 26)`	1.3149719
`qt(0.95, df = 12)`	1.7822876
`qt(0.95, df = 13)`	1.7709334
`qt(0.95, df = 26)`	1.7056179
`qt(0.975, df = 12)`	2.1788128
`qt(0.975, df = 13)`	2.1603687
`qt(0.975, df = 26)`	2.0555294

R code	Value
qt(0.995, df = 12)	3.0545396
qt(0.995, df = 13)	3.0122758
qt(0.995, df = 26)	2.7787145

15. Using your answers to questions 12, 13, and 14, create a 95% confidence interval for the difference in mean cholesterol levels between the corn flake and oat bran diet.

16. What value do check to see if this interval contains? Does our interval contain this value?

Using the t-distribution to create a confidence interval

So far we've found a confidence interval using the percentile and SE methods. Both of these used some aspect of the bootstrap distribution. One final option is to use **only** the t-distribution only to create our confidence interval.

17. What distribution does a bootstrap distribution approximate?

18. If we wanted to use a t-distribution to approximate this distribution, what conditions do we need to check?