Homework 4

Due: June 20, 2018 @ 11:59pm

Instructions

For this homework assignment, you will use statistical inference to answer a question about the National Survey of Family Growth, Cycle 6 dataset published by the National Center for Health Statistics.

Obtain the Github repository you will use to complete homework 4 that contains a starter RMarkdown file named `homework_4.Rmd`, which you will use to do your work and write-up when completing the questions below. Remember to fill in your name at the top of the RMarkdown document and be sure to save, commit, and push (upload) frequently to Github so that you have incremental snapshots of your work. When you're done, follow the How to submit section below to setup a Pull Request, which will be used for feedback.

About the dataset

This homework uses the *National Survey of Family Growth, Cycle 6* dataset in the file `2002FemPreg.rds`, published by the National Center for Health Statistics. Complete descriptions of all the variables can be found in the NSFG Cycle 6: Female Pregnancy File Codebook. Below are selected descriptions of variables that will be used for this homework assignment:

Variable	Description
`caseid`	integer ID of the respondent
`prglngth`	integer duration of the pregnancy in weeks
`outcome`	integer code for the outcome of the pregnancy, with a 1 indicating a live birth
`birthord`	serial number for live births; the code for a respondent's first child is 1, and so on. For outcomes other than live birth, this field is blank

Questions

This homework assignment revolves around answering the following question using this dataset:

> Do first born children either arrive early or late when compared with non-first-borns?

The questions in this assignment will guide you through the process of answering this question using statistical inference.

1. Addressing the question *"do first born children either arrive early or late when compared with non-first-borns?"* means that we should only consider live births in the dataset. Filter the dataset so that it only contains outcomes with live births and assign this result to the variable `live_births`.

   Next, we need to label all births in `live_births` as either "first" or "other" so that we can easily find the children that are first borns and the ones that are not. There are a couple of different ways that you can label the births:

   - Split the dataset into two parts, a `first_births` dataset and an `other_births` dataset. Do this by applying a filter to extract the first births, then use `mutate()` to create a new column called `birth_order` that labels these rows as "first". Then repeat this process, except apply a filter to extract all other births and label those as "other" in `birth_order`. To recombine the datasets into one, use `bind_rows()`.
   - Use `if_else()` with `mutate()` to create the `birth_order` column and the "first" and "other" labels. You can also try using `case_when()` instead of `if_else()` to accomplish this. If you do it this way, you won't need to use `bind_rows()`.

After labeling the births, remove the extraneous variables from the data frame leaving just the `prglngth` and `birth_order` columns. Assign the resulting data frame to a variable named `pregnancy_length`.

2. Take the `pregnancy_length` dataset and plot a probability mass function (PMF) histogram of the pregnancy length in weeks that shows "first" births and "other" births on the same plot. Choose a reasonable binwidth for the histogram and add `coord_cartesian(xlim = c(27, 46))` to your plot so that the window focuses where most of the data is. **After creating the plot, describe the shape, center, and spread of the two distributions**. Based on the visualization, do you think the data looks like it supports the statement that "first born children either arrive early or arrive late when compared with non-first-borns"?

3. Group the variable `prglngth` into "first" and "other" births and compute the summary statistics (mean, median, standard deviation, inter-quartile range, minimum, maximum) for each group. How do the different summary statistics compare between the two distributions? Does it look like there may be a notable difference between the two distributions? Explain.

4. Plot the cumulative distribution functions (CDFs) of the pregnancy lengths for "first" and "other" births. The CDF for both distributions should be on the same figure so that we can directly compare them (**hint:** they should be two different colors and partially transparent). How do the distributions compare? Does it look like there is there a meaningful difference between the two distributions?

5. If we want to determine whether or not the difference between two distributions is statistically significant, we need to run a hypothesis test. Formalize your analysis by writing down the null hypothesis for the question of whether first babies arrive early or arrive late when compared with non-first-borns. It should be clear from how you write your null hypothesis whether you're conducting a **one-sided** or **two-sided** hypothesis test. You should also restate what the **observed value** is (this will be a number you compute using the data in `prglngth`) that you will be comparing against the null distribution.

6. Use a simulation to generate the null distribution so that you can perform the hypothesis test. Do this using the functions provided in the `infer` package, `specify()`, `hypothesize()`, `generate()`, and `calculate()`. To collect enough statistics, you should set the simulation to repeat 10,000 times.

   Once you've obtained the null distribution, use it to compute the p-value for your hypothesis test. Assuming a significance level of $\alpha = 0.05$, state whether we can we reject the null hypothesis.

   Finally, visualize the simulated null distribution as a histogram and use `geom_vline()` to show where the **observed value** sits relative to it.

7. Use a bootstrap simulation to calculate the 95% confidence interval for your **observed value**. Do this using the following functions from the `infer` package, `specify()`, ‘`generate()`, and `calculate()`. To collect enough statistics, you should set the bootstrap simulation to repeat 10,000 times.

   Once you've obtained the bootstrap distribution, use the method demonstrated in the class 17 slides to find the upper and lower bounds of the 95% confidence interval. Does the value corresponding to the null result fall within the range of the 95% confidence interval?

   Finally, visualize the bootstrap distribution as a histogram and use two `geom_vline()`s to show the region corresponding to the 95% confidence interval.

8. In addition to hypothesis tests and confidence intervals, we should also consider the **effect size**, which measures the relative difference between two distributions. The effect size helps us better know how important a given result actually is, not just whether or not we can reject the null hypothesis. One measure of the effect size is called Cohen's d (https://en.wikipedia.org/wiki/Effect_size#Cohen.27s_d), which we will use to compute the effect size between the pregnancy lengths for "first" and "other" births. The different ranges of d can be interpreted using the following table:

Effect size	d
Very small	0.01
Small	0.20
Medium	0.50

Effect size	d
Large	0.80
Very large	1.20
Huge	2.00

The following set of functions should also be preloaded for you: `cohens_d_bootstrap()`, `bootstrap_report()`, and `plot_ci()`. These functions will use bootstrap simulations to compute the confidence interval for the Cohen's *d* parameter. Run the bootstrap simulation as follows (you can just copy and paste this code):

```
cohens_d_bootstrap(data = pregnancy_length, model = prglngth ~ birth_order)
```

Be sure to assign the results to a variable, for example `bootstrap_results`.

To print a report for the bootstrap simulation, run:

```
bootstrap_report(bootstrap_results)
```

To visualize the bootstrap distribution and confidence interval, run:

```
plot_ci(bootstrap_results)
```

Using the provided table, report how large the effect size is for the difference in pregnancy lengths for "first" and "other" births.

How to submit

When you are ready to submit, be sure to save, commit, and push your final result so that everything is synchronized to Github. Then, navigate to **your copy** of the Github repository you used for this assignment. You should see your repository, along with the updated files that you just synchronized to Github. Confirm that your files are up-to-date, and then do the following steps:

1. Click the *Pull Requests* tab near the top of the page.
2. Click the green button that says "New pull request".
3. Click the dropdown menu button labeled "base:", and select the option `starting`.
4. Confirm that the dropdown menu button labeled "compare:" is set to `master`.
5. Click the green button that says "Create pull request".
6. Give the *pull request* the following title: `Submission: Homework 4, FirstName LastName`, replacing `FirstName` and `LastName` with your actual first and last name.
7. In the messagebox, write: `My homework submission is ready for grading @jkglasbrenner`.
8. Click "Create pull request" to lock in your submission.

Cheatsheets

You are encouraged to review and keep the following cheatsheets handy while working on this assignment:

- RStudio cheatsheet
- RMarkdown cheatsheet
- RMarkdown reference
- `ggplot2` cheatsheet
- Data transformation cheatsheet
- Data import cheatsheet