Assignment 3

Due: November 29, 2017 @ 11:59pm

Instructions

In the Github Classroom repository that you copied, you should have a template file named `assignment_3.Rmd`. Use this file to answer the questions below. Also fill out any missing information at the top of the RMarkdown document. Once you're done, be sure to save, commit, and push to Github. Then, submit a pull request.

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
`totalwgt_lb`	the birth weight of the baby in pounds
`postsmks`	integer code about whether or not respondent smoked a cigarette at some point during the pregnancy `value` `label` 1 YES 5 NO

Questions

This homework assignment revolves around answering two questions using this dataset:

- Do first born children either arrive early or late when compared with subsequently born children?
- Do children born to mothers that reported smoking during their pregnancy either weigh more or weigh less than children born to non-smoking mothers?

The majority of the questions will step you through the procedure of answering the first question using inference methods. You will then use this procedure you stepped through as a template for answering the second question on your own.

Question 1

To get started with with answering the question of whether or not *first born children either arrive early or late*, we need to do some basic data wrangling to clean up the dataset. Begin by filtering the dataset so that the dataset only contains outcomes with live births. Assign this result to the variable `live_births`. Then, continuing with `live_births`, filter again and create two additional variables:

- Apply a filter to extract all first births. Then, use `select()` to grab the `prglngth` column and discard all other columns. Pipe this into `mutate(birth_order = "first")`, and then assign the output to the variable `first_births`.

- Apply a filter to extract all other births except first births. Then, use `select()` to grab the `prglngth` column and discard all other columns. Pipe this into `mutate(birth_order = "other")`, and then assign the output to the variable `other_births`.

Use `bind_rows()` to combine the results in `first_births` and `other_births` into a single, tidy `tibble`. Pipe this tibble into `remove_missing()` to remove any rows containing NA entries and assign the result to the variable name `pregnancy_length`.

Question 2

Take `pregnancy_length` 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 choice for the histogram and add `coord_cartesian(xlim = c(27, 46))` to your plot so that the window focuses on 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"?

Question 3

Using `group_by()` and `summarize()`, compute the different summary statistics (mean, median, standard deviation, inter-quartile range) of the variable `prglngth` for `first` and `other` births in `pregnancy_length`. How do the different summary statistics compare between the two distributions? Based on the initial summary statistics, does it look like there may be a statistically significant difference between the two distributions? If so, why?

Question 4

Determine whether the distribution of pregnancy lengths follows a nearly normal distribution. Do this by creating two separate Q-Q plots for the variable `prglngth` in `pregnancy_length`. The first plot is for when `birth_order` equals `"first"` and the second plot is for when `birth_order` equals `"second"`.

To get started, filter `pregnancy_length` so that it only contains `first` births and assign it to the variable named `first_births`. Similarly, filter `pregnancy_length` to only contain `other` births and assign it to the variable named `other_births`.

After filtering, adapt the following code to create a Q-Q plot for `first_births` and also for `other_births`. The additional code helps to plot the "ideal" reference line to show any deviations from normality:

```
qq_x <- qnorm(p = c(0.25, 0.75))
qq_y <- quantile(x = dataset$variable, probs = c(0.25, 0.75), type = 1)
qq_slope <- diff(qq_y) / diff(qq_x)
qq_int <- qq_y[1] - qq_slope * qq_x[1]
ggplot(data = dataset) +
  geom_qq(mapping = aes(sample = variable)) +
  geom_abline(slope = qq_slope, intercept = qq_int, size = 0.75)
```

Please note that you will have to change the `dataset` and `variable` parameters. Based on your plots, does it appear that the pregnancy length distribution is nearly normal for both `first` and `other` births? Why or why not?

Question 5

Returning back to the question of whether or not "first babies arrive early or arrive late", let's plot the cumulative distribution functions (CDFs) of the pregnancy lengths for `first` and `other` births. Plot the CDF for both distributions

on the same figure so that we can directly compare them. You can either use the procedure outlined in *Cumulative distribution functions* from reading assignment 15 or use `stat_ecdf` to do this. How do the distributions compare? Does it look like there is there a meaningful difference between the two distributions?

Question 6

If we want to determine whether or not the difference between two distributions is statistically significant, we need to run a hypothesis test. **Before going further, for the question of "do first babies arrive early or arrive late", formalize your analysis by writing down the null hypothesis**.

Next, use the pre-loaded `inference()` function to run a hypothesis test for the difference in means of the `prglnth` variable between the `first_births` and `other_births` datasets:

```
inference(y = prglngth, x = birth_order, data = pregnancy_length, type = "ht",
          statistic = "mean", null = 0, alternative = "twosided",
          order = c("first", "other"), method = "simulation",
          show_eda_plot = FALSE)
```

Assume that we set our significance level to $\alpha = 0.05$. Based on the outputted p-value, can we reject the null hypothesis?

For completeness, also compute the 95% confidence interval for the difference in means:

```
inference(y = prglngth, x = birth_order, data = pregnancy_length,
          type = "ci", statistic = "mean", null = 0, method = "simulation",
          order = c("first", "other"), boot_method = "perc",
          show_eda_plot = FALSE)
```

Does the confidence interval overlap with the null value?

Question 7

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, which we will use to compute the effect size between the pregnancy lengths for first births 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
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:

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

Be sure to assign the results to a variable, for example `bootstrap_results`. Note that the input `model` specifies how to split the data into categories. You put the response variable (`prglngth`) on left side of a tilde ~, and the categorical explanatory variable (`birth_order`) on the right side.

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)
```

Interpret the outputs from the Cohen's *d* bootstrap simulation. How large is the effect size between the pregnancy length of `first` births and `other` births? Based on this and the previous hypothesis test, provide an answer the question "do first born children arrive early or late compared to other children?"

Question 8

Use the inference tools and procedures that you practiced in the previous exercises to obtain an answer to the question "Do children born to mothers that reported smoking during their pregnancy either weigh more or weigh less than children born to non-smoking mothers?" To answer this question, you will need to start from the dataset stored in `live_births` and work with the columns `totalwgt_lb` and `postsmks`. Please review the About the dataset section for information on what the values in those columns mean.

In order to properly answer the question, you will need to include:

- A visual comparison of the two data distributions
- Use the `inference()` function to run a hypothesis test and to compute the confidence interval
- Use `cohens_d_bootstrap()` to compute and check the *effect size*.

At the end, write a paragraph that summarizes your results and provide an answer as to whether there is a connection between birth weights and smoking. Write the summary as if you are an academic or scientific journalist, focusing on how you can answer the question clearly, precisely, and honestly.