Lab 3

DATA1220-55

2024-11-13

Research Question: Do babies born to mothers who smoke have lower birth weights than babies born to non-smokers?

1. Start a new project in RStudio.
2. Start a new R script and save it as `lab03.R`.
3. You will need to use the `dplyr` library for this lab. Load it now using the `library()` function.
4. Download the Excel spreadsheet `births14.xlsx`.
    a. Go to the URL https://github.com/sarah-grabinski/data1220-55_fall2024/raw/refs/heads/main/slides/lab01/births14.xlsx and manually download the file.
    b. Modify the code below to download the file directly to your project folder using the URL https://github.com/sarah-grabinski/data1220-55_fall2024/raw/refs/heads/main/slides/lab01/births14.xlsx.

```
download.file("http://www.website.com/excel_spreadsheet.xlsx",
              destfile = "excel_spreadsheet.xlsx",
              mode = "wb")
```

5. Use the "Import Dataset" tool in RStudio to load the file `births14.xlsx` into RStudio as the dataframe `births14`.
6. Modify the function below to inspect your data using the `summary()` function.
    a. What is the value of the the 3rd quartile of the variable `weight`?
    b. Compare the mean and median of the variable `weight`. Is it reasonable to assume normality here?

```
summary(dataframe)
```

7. Generate summary statistics $\bar{x}$ for the mean birth weight `weight` of babies by smoking habit `habit` using the `summarize()` function from the `dplyr` library and the `mean()` function. You will also need to estimate the standard deviation s with the `sd()` function and get the sample size n with the `n()` function for each group. Modify the example below.

```
dataframe |>
  summarize(newcolumn1 = function1(),
            newcolumn2 = function2(variable1),
            newcolumn3 = function3(variable1),
            .by = "variable2")
```

8. Calculate a point estimate for the difference $\bar{x}_1 - \bar{x}_2$ between the mean birth weight of babies born to smoking moms and the mean birth weight of babies born to non-smoking moms. Store the result as a variable called `mean_diff`.

```
new_variable <- 1 + 2
```

9. Calculate the standard error SE for the difference between the 2 mean birth weights $\bar{x}_1 - \bar{x}_2$. Store the result as a variable called `se_diff`.

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

```
new_variable <- sqrt((s1^2)/n1 + (s2^2)/n2)
```

10. Construct a 95% confidence interval using the Student's t distribution for the difference in average birth weights by smoking habit.

    a. Calculate the degrees of freedom for this t distribution and store it as the variable `dof`.

    $$\text{degrees of freedom} = \min(n_1, n_2) - 1$$

    b. Find the value of alpha α for a 95% confidence interval and store it as the variable `alpha`.

    $$\alpha = 1 - \text{confidence}$$

    c. Find the critical value T^* for a 95% confidence interval using the `qt(probability, df)` function. Use the probability $\alpha/2$ or $1 - \alpha/2$ and the degrees of freedom `dof` that you just calculated. Store it as the variable `t_star`.

d. Calculate the upper and lower bounds of your confidence interval as point estimate+ $T^* \times$ SE and point estimate $- T^* \times$ SE.

e. Can you interpret your confidence interval in a complete sentence?

f. Based on the boundaries of your confidence interval, can you make any early guesses about whether your hypothesis test will be significant?

11. Perform a 2-sample t-test to test the hypothesis that there is a difference in average birth weights between babies born to smoking and non-smoking mothers.

$$H_0 : \mu_{\text{smoker}} - \mu_{\text{nonsmoker}} = 0$$
$$H_A : \mu_{\text{smoker}} - \mu_{\text{nonsmoker}} < 0$$

a. Calculate the test statistic T for your observed difference under the null hypothesis that there is no difference $H_0 : \mu_{\text{smoker}} - \mu_{\text{nonsmoker}} = 0$. Save the result as the variable `test_statistic`.

$$`T = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

b. Get the p-value for the one-sided hypothesis test (left-tailed) from your `test_statistic` T using the `pt()` function and your degrees of freedom `dof`. Depending on the value of your test statistic T, you may need to change the `lower.tail` parameter to false `F` to get the proper probability.

::: {.cell}

```
pt(statistic, df, lower.tail = T)
```

:::

c. Using the significance level `alpha` or α that you calculated previously, would you reject the null hypothesis? Why or why not?