SOC 4015/5050: PS-04 - Difference of Means Tests

Christopher Prener, Ph.D.

Fall 2018

Directions

Please complete all steps below. Your well-formatted R Notebook source (the `.Rmd` file) and output along with your project folder structure should be uploaded to your GitHub assignment repository by 4:15pm on Monday, October 29th, 2018. You will need to have the package `gapminder` installed to access the data for this assignment.

Analysis Development

Using RStudio and your operating system's file manager, create an R Project in the *existing* directory in your assignments repository named `Lab-08`. Add a `README.md` file, notebook, and all necessary folders before beginning.[1]

[1] This initial section follows the project workflow that is available in the `lecture-03` repo!

Part 1: Data Preparation

1. Using the data table `gapminder` in the `gapminder` package, create a new data frame that has the following characteristics:
   (a) contains a binary variable that is `TRUE` for Asian countries
   (b) contains a binary variable that is `TRUE` for countries in Oceania
   (c) contains only data for the year 2007
   (d) contains only the the variables `country`, `continent`, and `lifeExp` in addition to the two new binary variables created above
2. Using the same source data as above, create a new data frame that has the following characteristics:
   (a) contains only data for the years 1957 and 2007
   (b) contains only the the variables `country`, `year`, and `lifeExp`
3. Reshape the data created in the previous question and store those reshaped data in a new data frame. Export these data as a `.csv` file.

Part 2: Independent T Test

Using the life expectancy data created above in Part 1, question 1, answer the following questions.

4. Calculate the appropriate descriptive statistics for the life expectancy variable.
5. Calculate the appropriate descriptive statistics for the binary variable you created representing Asian countries.
6. Test these data (the life expectancy variable and the Asian countries variable) to see whether they meet the assumptions for the independent t test. Make sure to include a narrative in your `.Rmd` file that evaluates each of the assumptions and makes a concluding statement about how valid our t test results may be given the assumptions testing.
7. Create two plots of the relationship between Asian countries and life expectancy. One of these plots should be ideal for exploratory data analysis and one should be ideal for communicating our findings with an audience that includes members without statistical training. Save both plots as `.png` files at 300 dots per inch.
8. Create a ridge plot using the life expectancy and continent variables. Why do you think no data appears for Oceania?[2]
9. Calculate and interpret a difference of means test of the variation in life expectancy between Asian and non-Asian countries. Include an interpretation of the *effect size* of this relationship in addition to assessing its statistical significance. Export the cleaned test results to a `.csv` file (s).

[2] *Hint:* Use the Oceania binary variable you created to help answer this question.

Part 3: Dependent T Test

Using the life expectancy data created above in Part 1, questions 2 and 3, answer the following questions.

10. Calculate the appropriate descriptive statistics for the 1957 life expectancy data.[3]
11. Calculate the appropriate descriptive statistics for the 2007 life expectancy data.
12. Create a descriptive statistics summary of all variables in these data using `skimr`.

[3] *Hint:* See the lecture-08 page on the website for tips on using variables that have fully numeric names, like `1957`.

13. Test these data (the variation in life expectancy between 1957 and 2007) to see whether they meet the assumptions for the dependent t test. Make sure to include a narrative in your `.Rmd` file that evaluates each of the assumptions and makes a concluding statement about how valid our t test results may be given the assumptions testing.

14. Create two plots showing the variation in life expectancy between 1957 and 2007. One of these plots should be ideal for exploratory data analysis and one should be ideal for communicating our findings with an audience that includes members without statistical training. Save both plots as `.png` files at 300 dots per inch.

15. Calculate and interpret a difference of means test of the variation in life expectancy between 1957 and 2007. Include an interpretation of the *effect size* of this relationship in addition to assessing its statistical significance. Export the cleaned test results to a `.csv` file(s).