nhanes_hypothesis_test_practice

February 26, 2022

1 Practice notebook for hypothesis tests using NHANES data

This notebook will give you the opportunity to perform some hypothesis tests with the NHANES data that are similar to what was done in the week 3 case study notebook.

You can enter your code into the cells that say "enter your code here", and you can type responses to the questions into the cells that say "Type Markdown and Latex".

Note that most of the code that you will need to write below is very similar to code that appears in the case study notebook. You will need to edit code from that notebook in small ways to adapt it to the prompts below.

To get started, we will use the same module imports and read the data in the same way as we did in the case study:

```
In [1]: %matplotlib inline
        import matplotlib.pyplot as plt
        import seaborn as sns
        import pandas as pd
        import statsmodels.api as sm
        import numpy as np

        da = pd.read_csv("nhanes_2015_2016.csv")
```

1.1 Question 1

Conduct a hypothesis test (at the 0.05 level) for the null hypothesis that the proportion of women who smoke is equal to the proportion of men who smoke.

```
In [1]: # insert your code here
```

Q1a. Write 1-2 sentences explaining the substance of your findings to someone who does not know anything about statistical hypothesis tests.

Q1b. Construct three 95% confidence intervals: one for the proportion of women who smoke, one for the proportion of men who smoke, and one for the difference in the rates of smoking between women and men.

```
In [ ]: # insert your code here
```

Q1c. Comment on any ways in which the confidence intervals that you found in part b reinforce, contradict, or add support to the hypothesis test conducted in part a.

1.2 Question 2

Partition the population into two groups based on whether a person has graduated college or not, using the educational attainment variable DMDEDUC2. Then conduct a test of the null hypothesis that the average heights (in centimeters) of the two groups are equal. Next, convert the heights from centimeters to inches, and conduct a test of the null hypothesis that the average heights (in inches) of the two groups are equal.

```
In [ ]: # insert your code here
```

Q2a. Based on the analysis performed here, are you confident that people who graduated from college have a different average height compared to people who did not graduate from college?

Q2b: How do the results obtained using the heights expressed in inches compare to the results obtained using the heights expressed in centimeters?

1.3 Question 3

Conduct a hypothesis test of the null hypothesis that the average BMI for men between 30 and 40 is equal to the average BMI for men between 50 and 60. Then carry out this test again after log transforming the BMI values.

```
In [ ]: # insert your code here
```

Q3a. How would you characterize the evidence that mean BMI differs between these age bands, and how would you characterize the evidence that mean log BMI differs between these age bands?

1.4 Question 4

Suppose we wish to compare the mean BMI between college graduates and people who have not graduated from college, focusing on women between the ages of 30 and 40. First, consider the variance of BMI within each of these subpopulations using graphical techniques, and through the estimated subpopulation variances. Then, calculate pooled and unpooled estimates of the standard error for the difference between the mean BMI in the two populations being compared. Finally, test the null hypothesis that the two population means are equal, using each of the two different standard errors.

```
In [ ]: # insert your code here
```

Q4a. Comment on the strength of evidence against the null hypothesis that these two populations have equal mean BMI.

Q4b. Comment on the degree to which the two populations have different variances, and on the extent to which the results using different approaches to estimating the standard error of the mean difference give divergent results.

1.5 Question 5

Conduct a test of the null hypothesis that the first and second diastolic blood pressure measurements within a subject have the same mean values.

```
In [ ]: # insert your code here
```

Q5a. Briefly describe your findings for an audience that is not familiar with statistical hypothesis testing.

Q5b. Pretend that the first and second diastolic blood pressure measurements were taken on different people. Modfify the analysis above as appropriate for this setting.

```
In [ ]: # insert your code here
```

Q5c. Briefly describe how the approaches used and the results obtained in the preceeding two parts of the question differ.