nhanes_univariate_practice

August 9, 2020

1 Practice notebook for univariate analysis using NHANES data

This notebook will give you the opportunity to perform some univariate analyses on your own using the NHANES. These analyses are similar to what was done in the week 2 NHANES 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

Relabel the marital status variable DMDMARTL to have brief but informative character labels. Then construct a frequency table of these values for all people, then for women only, and for men only. Then construct these three frequency tables using only people whose age is between 30 and 40.

```
In [2]: # insert your code here
        da["Marital_status"]=da["DMDMARTL"]
        da["Marital_status"]
        da["Marital_status"].value_counts()
        male_counts=da[da["RIAGENDR"]==1]["Marital_status"].value_counts()
        female_counts=da[da["RIAGENDR"]==2]["Marital_status"].value_counts()
```

```
#da[(da["RIAGENDR"]==2) & (da["RIDAGEYR"]>30) & (da["RIDAGEYR"] <= 40) ]["Marital_stat
```

Q1a. Briefly comment on some of the differences that you observe between the distribution of marital status between women and men, for people of all ages. da["DMDHRAGE"] **Q1b.** Briefly comment on the differences that you observe between the distribution of marital status states for women between the overall population, and for women between the ages of 30 and 40.

Q1c. Repeat part b for the men.

1.2 Question 2

Restricting to the female population, stratify the subjects into age bands no wider than ten years, and construct the distribution of marital status within each age band. Within each age band, present the distribution in terms of proportions that must sum to 1.

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

Q2a. Comment on the trends that you see in this series of marginal distributions.

Q2b. Repeat the construction for males.

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

Q2c. Comment on any notable differences that you see when comparing these results for females and for males.

1.3 Question 3

Construct a histogram of the distribution of heights using the BMXHT variable in the NHANES sample.

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

Q3a. Use the `bins` argument to distplot to produce histograms with different numbers of bins. Assess whether the default value for this argument gives a meaningful result, and comment on what happens as the number of bins grows excessively large or excessively small.

Q3b. Make separate histograms for the heights of women and men, then make a side-by-side boxplot showing the heights of women and men.

```
In [3]: # insert your code here
```

Q3c. Comment on what features, if any are not represented clearly in the boxplots, and what features, if any, are easier to see in the boxplots than in the histograms.

1.4 Question 4

Make a boxplot showing the distribution of within-subject differences between the first and second systolic blood pressure measurents (BPXSY1 and BPXSY2).

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

Q4a. What proportion of the subjects have a lower SBP on the second reading compared to the first?

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

Q4b. Make side-by-side boxplots of the two systolic blood pressure variables.

```
In [4]: # insert your code here
```

Q4c. Comment on the variation within either the first or second systolic blood pressure measurements, and the variation in the within-subject differences between the first and second systolic blood pressure measurements.

1.5 Question 5

Construct a frequency table of household sizes for people within each educational attainment category (the relevant variable is DMDEDUC2). Convert the frequencies to proportions.

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

Q5a. Comment on any major differences among the distributions.

Q5b. Restrict the sample to people between 30 and 40 years of age. Then calculate the median household size for women and men within each level of educational attainment.

```
In [7]: # insert your code here
```

1.6 Question 6

The participants can be clustered into "maked variance units" (MVU) based on every combination of the variables SDMVSTRA and SDMVPSU. Calculate the mean age (RIDAGEYR), height (BMXHT), and BMI (BMXBMI) for each gender (RIAGENDR), within each MVU, and report the ratio between the largest and smallest mean (e.g. for height) across the MVUs.

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

Q6a. Comment on the extent to which mean age, height, and BMI vary among the MVUs.

Q6b. Calculate the inter-quartile range (IQR) for age, height, and BMI for each gender and each MVU. Report the ratio between the largest and smalles IQR across the MVUs.

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

Q6c. Comment on the extent to which the IQR for age, height, and BMI vary among the MVUs.