pratice

December 2, 2021

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]: da['DMDMARTL'].value_counts()
```

```
Out[2]: 1.0    2780
        5.0    1004
        3.0     579
        6.0     527
        2.0     396
        4.0     186
```

```
        77.0       2
        Name: DMDMARTL, dtype: int64

In [3]: da.columns

Out[3]: Index(['SEQN', 'ALQ101', 'ALQ110', 'ALQ130', 'SMQ020', 'RIAGENDR', 'RIDAGEYR',
               'RIDRETH1', 'DMDCITZN', 'DMDEDUC2', 'DMDMARTL', 'DMDHHSIZ', 'WTINT2YR',
               'SDMVPSU', 'SDMVSTRA', 'INDFMPIR', 'BPXSY1', 'BPXDI1', 'BPXSY2',
               'BPXDI2', 'BMXWT', 'BMXHT', 'BMXBMI', 'BMXLEG', 'BMXARML', 'BMXARMC',
               'BMXWAIST', 'HIQ210'],
              dtype='object')

In [4]: da['RIAGENDR'] = da['RIAGENDR'].replace({1:"Male", 2:"Female"})

In [5]: da['DMDMARTL_LABEL'] = da['DMDMARTL'].replace({1:"Married", 2:"Widowed", 3:"Divorced",
                                6:"Living with partner", 77:"Refused",99:"Dont Know"})

In [6]: da['DMDMARTL_LABEL'].value_counts()

Out[6]: Married                2780
        Never married          1004
        Divorced                579
        Living with partner     527
        Widowed                 396
        Separated               186
        Refused                   2
        Name: DMDMARTL_LABEL, dtype: int64

In [7]: da_by_gender = da.groupby(['RIAGENDR'])['DMDMARTL_LABEL'].value_counts()
        da_by_gender

Out[7]: RIAGENDR  DMDMARTL_LABEL
        Female    Married                1303
                  Never married           520
                  Divorced                350
                  Widowed                 296
                  Living with partner     262
                  Separated               118
                  Refused                   1
        Male      Married                1477
                  Never married           484
                  Living with partner     265
                  Divorced                229
                  Widowed                 100
                  Separated                68
                  Refused                   1
        Name: DMDMARTL_LABEL, dtype: int64
```

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. I cannot see to much differences

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.

```
In [17]: da['agegr'] = pd.cut(da.RIDAGEYR,[10,20,30,40,50,60,70,80])
         da_female_age = da.groupby(['RIAGENDR','agegr'])['DMDMARTL_LABEL'].value_counts()
         da_female_age
```

```
Out[17]: RIAGENDR  agegr     DMDMARTL_LABEL
         Female    (10, 20]  Never married           30
                             Living with partner      8
                             Married                  1
                   (20, 30]  Never married          229
                             Married                157
                             Living with partner    106
                             Divorced                11
                             Separated               11
                   (30, 40]  Married                258
                             Never married           97
                             Living with partner     57
                             Divorced                43
                             Separated               17
                             Widowed                  2
                   (40, 50]  Married                288
                             Divorced                69
                             Never married           63
                             Living with partner     37
                             Separated               33
                             Widowed                 12
                   (50, 60]  Married                257
                             Divorced                83
                             Never married           42
                             Living with partner     32
                             Widowed                 28
                             Separated               27
                             Refused                  1
                   (60, 70]  Married                212
                             Divorced                85
                             Widowed                 65
                                                    ...
         Male      (30, 40]  Never married           89
                             Living with partner     72
                             Divorced                24
                             Separated               12
                             Widowed                  2
                             Refused                  1
                   (40, 50]  Married                282
                             Never married           39
```

```
                    Divorced                34
                    Living with partner     33
                    Separated               11
                    Widowed                  2
          (50, 60]  Married                296
                    Divorced                57
                    Never married           47
                    Living with partner     34
                    Separated               10
                    Widowed                 10
          (60, 70]  Married                291
                    Divorced                55
                    Never married           38
                    Living with partner     22
                    Widowed                 17
                    Separated               14
          (70, 80]  Married                246
                    Widowed                 67
                    Divorced                57
                    Separated               14
                    Living with partner      9
                    Never married            9
Name: DMDMARTL_LABEL, Length: 79, dtype: int64
```

In general, most women 30-40 are married or never got married **Q1c.** Repeat part b for the men. Most men in 30-40 never got married or just live with a partner

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 [26]: da_female_age_unstack = da_female_age.unstack(fill_value=0) # Restructure the results
         da_female_age_unstack
```

Out[26]:

DMDMARTL_LABEL		Divorced	Living with partner	Married	Never married	\
RIAGENDR	agegr					
Female	(10, 20]	0	8	1	30	
	(20, 30]	11	106	157	229	
	(30, 40]	43	57	258	97	
	(40, 50]	69	37	288	63	
	(50, 60]	83	32	257	42	
	(60, 70]	85	19	212	38	
	(70, 80]	59	3	130	21	
Male	(10, 20]	0	3	1	36	
	(20, 30]	2	92	103	226	
	(30, 40]	24	72	258	89	
	(40, 50]	34	33	282	39	

```
                (50, 60]        57                  34      296             47
                (60, 70]        55                  22      291             38
                (70, 80]        57                   9      246              9

DMDMARTL_LABEL     Refused  Separated  Widowed
RIAGENDR agegr
Female   (10, 20]        0          0        0
         (20, 30]        0         11        0
         (30, 40]        0         17        2
         (40, 50]        0         33       12
         (50, 60]        1         27       28
         (60, 70]        0         22       65
         (70, 80]        0          8      189
Male     (10, 20]        0          0        0
         (20, 30]        0          7        2
         (30, 40]        1         12        2
         (40, 50]        0         11        2
         (50, 60]        0         10       10
         (60, 70]        0         14       17
         (70, 80]        0         14       67
```

```
In [27]: da_female_age_unstack = da_female_age_unstack.apply(lambda x: x/x.sum(), axis=1) # No
         print(da_female_age_unstack.to_string(float_format="%.3f"))  # Limit display to 3 dec
```

```
DMDMARTL_LABEL     Divorced  Living with partner  Married  Never married  Refused  Separated  W
RIAGENDR agegr
Female   (10, 20]     0.000                0.205    0.026          0.769    0.000      0.000
         (20, 30]     0.021                0.206    0.305          0.446    0.000      0.021
         (30, 40]     0.091                0.120    0.544          0.205    0.000      0.036
         (40, 50]     0.137                0.074    0.574          0.125    0.000      0.066
         (50, 60]     0.177                0.068    0.547          0.089    0.002      0.057
         (60, 70]     0.193                0.043    0.481          0.086    0.000      0.050
         (70, 80]     0.144                0.007    0.317          0.051    0.000      0.020
Male     (10, 20]     0.000                0.075    0.025          0.900    0.000      0.000
         (20, 30]     0.005                0.213    0.238          0.523    0.000      0.016
         (30, 40]     0.052                0.157    0.563          0.194    0.002      0.026
         (40, 50]     0.085                0.082    0.703          0.097    0.000      0.027
         (50, 60]     0.126                0.075    0.652          0.104    0.000      0.022
         (60, 70]     0.126                0.050    0.666          0.087    0.000      0.032
         (70, 80]     0.142                0.022    0.612          0.022    0.000      0.035
```

Q2a. Comment on the trends that you see in this series of marginal distributions.
Q2b. Repeat the construction for males.

```
In [4]: # 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 [5]: # 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 [6]: # 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 [7]: # insert your code here
```

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

```
In [8]: # insert your code here
```

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

```
In [9]: # 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 [10]: # 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 [11]: # 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 [12]: # 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 [13]: # insert your code here
```

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