assignment2

April 29, 2022

1 Assignment 2

For this assignment you'll be looking at 2017 data on immunizations from the CDC. Your datafile for this assignment is in `assets/NISPUF17.csv`. A data users guide for this, which you'll need to map the variables in the data to the questions being asked, is available at `assets/NIS-PUF17-DUG.pdf`. **Note: you may have to go to your Jupyter tree (click on the Coursera image) and navigate to the assignment 2 assets folder to see this PDF file).**

1.1 Question 1

Write a function called `proportion_of_education` which returns the proportion of children in the dataset who had a mother with the education levels equal to less than high school (<12), high school (12), more than high school but not a college graduate (>12) and college degree.

This function should return a dictionary in the form of (use the correct numbers, do not round numbers):

```
{"less than high school":0.2,
"high school":0.4,
"more than high school but not college":0.2,
"college":0.2}
```

[18]:
```
import pandas as pd
df= pd.read_csv('assets/NISPUF17.csv', index_col=0)
ed=df['EDUC1']

def proportion_of_education():

    NumMothers=len(ed)
    lessth=[]
    eq=[]
    higherth=[]
    colleg=[]
    for i in ed:
        if   i == 1:
            lessth.append(i)
            portionless=len(lessth)/NumMothers
        elif i == 2:
            eq.append(i)
```

```
            portioneq=len(eq)/NumMothers
        elif i == 3:
            higherth.append(i)
            portionhigher=len(higherth)/NumMothers
        else :
            colleg.append(i)
            portioncolleg=len(colleg)/NumMothers
    proportion_of_education= {"less than high school":portionless,
                              "high school":portioneq,
                              "more than high school but not college":
 ↪portionhigher,
                              "college":portioncolleg}
    return proportion_of_education
```

[19]:
```
assert type(proportion_of_education())==type({}), "You must return a dictionary.
 ↪"
assert len(proportion_of_education()) == 4, "You have not returned a dictionary␣
 ↪with four items in it."
assert "less than high school" in proportion_of_education().keys(), "You have␣
 ↪not returned a dictionary with the correct keys."
assert "high school" in proportion_of_education().keys(), "You have not␣
 ↪returned a dictionary with the correct keys."
assert "more than high school but not college" in proportion_of_education().
 ↪keys(), "You have not returned a dictionary with the correct keys."
assert "college" in proportion_of_education().keys(), "You have not returned a␣
 ↪dictionary with the correct keys."
```

1.2 Question 2

Let's explore the relationship between being fed breastmilk as a child and getting a seasonal influenza vaccine from a healthcare provider. Return a tuple of the average number of influenza vaccines for those children we know received breastmilk as a child and those who know did not.

This function should return a tuple in the form (use the correct numbers:

```
(2.5, 0.1)
```

[4]:
```
import pandas as pd
df= pd.read_csv('assets/NISPUF17.csv', index_col=0)

def average_influenza_doses():
    copia= df.drop(df[df['CBF_01']>2].index)        #aplico filtro 1. CBF=1 o 2.
    columnas=copia[["CBF_01","P_NUMFLU"]].dropna() #aplico filtro2. P_NUMFLU␣
 ↪sin datos Nan
    fedyes=columnas[columnas['CBF_01']==1]          #con esto saco valores de␣
 ↪P_NUMFLU cuando CBF=1
    promfedyes=fedyes['P_NUMFLU'].mean()            #promedio valores de␣
 ↪P_NUMFLU cuando CBF=1
```

```
    fedNo=columnas[columnas['CBF_01']==2]          #con esto saco valores de␣
↪P_NUMFLU cuando CBF=2
    promfedNo=fedNo['P_NUMFLU'].mean()             #promedio valores de␣
↪P_NUMFLU cuando CBF=2
    average_influenza_doses=(promfedyes,promfedNo)
    return average_influenza_doses
```

[5]:
```
assert len(average_influenza_doses())==2, "Return two values in a tuple, the␣
↪first for yes and the second for no."
```

1.3 Question 3

It would be interesting to see if there is any evidence of a link between vaccine effectiveness and sex of the child. Calculate the ratio of the number of children who contracted chickenpox but were vaccinated against it (at least one varicella dose) versus those who were vaccinated but did not contract chicken pox. Return results by sex.

This function should return a dictionary in the form of (use the correct numbers):

```
{"male":0.2,
"female":0.4}
```

Note: To aid in verification, the `chickenpox_by_sex()['female']` value the autograder is looking for starts with the digits `0.0077`.

[8]:
```
import pandas as pd
df=pd.read_csv('assets/NISPUF17.csv', index_col=0)
def chickenpox_by_sex():
    copy= df
    columnas=copy[["HAD_CPOX","P_NUMVRC", "SEX"]]
    print(columnas)
    print('\n')
    #------------------------------ male -----------------------------------
    mask1=(columnas['SEX']==1) & (columnas['HAD_CPOX']==1) &␣
↪(columnas['P_NUMVRC']>=1)
    hvc=columnas.where(mask1).dropna()             #Dataframe de hombres con␣
↪vacuna contagiados
    print(hvc)
    thc=len(hvc)                                   #total hombres contagiados␣
↪con al menos una vacuna
    print('Hombres contagiados con al menos una vacuna : ', thc)

    mask2=(columnas['SEX']==1) & (columnas['HAD_CPOX']==2) &␣
↪(columnas['P_NUMVRC']>=1)
    hvnc=columnas.where(mask2).dropna()            #Dataframe de hombres con␣
↪vacuna no contagiados
    print('\n',hvnc)
    thnc=len(hvnc)                                 #total hombres no contagiados␣
↪con al menos una vacuna
```

```
    print('Hombres no contagiados con al menos una vacuna : ', thnc)

    maleratio=thc/thnc                                  #Hombres contagiados vs No␣
↪contagiados, todos con vacuna
    print('\nMale Ratio', maleratio)
    #------------------------------ female ----------------------------------
    mask3=(columnas['SEX']==2) & (columnas['HAD_CPOX']==1) &␣
↪(columnas['P_NUMVRC']>=1)
    mvc=columnas.where(mask3).dropna()                  #Dataframe de mujeres con␣
↪vacuna contagiadas
    print('\n',mvc)
    tmc=len(mvc)                                        #total mujeres contagiadass␣
↪con al menos una vacuna
    print('\nMujeres contagiadas con al menos una vacuna : ', tmc)

    mask4=(columnas['SEX']==2) & (columnas['HAD_CPOX']==2) &␣
↪(columnas['P_NUMVRC']>=1)
    mvnc=columnas.where(mask4).dropna()                 #Dataframe de mujeres con␣
↪vacuna no contagiadas
    print('\n',mvnc)
    tmnc=len(mvnc)                                      #total mujeres no contagiados␣
↪con al menos una vacuna
    print('\nMujeres no contagiadas con al menos una vacuna : ', tmnc)

    femaleratio=tmc/tmnc                                  #Hombres contagiados vs No␣
↪contagiados, todos con vacuna
    print('\nFemale Ratio', femaleratio)

    chickenpox_by_sex={"male":maleratio, "female":femaleratio}
    print('\n',chickenpox_by_sex)
    return chickenpox_by_sex
```

[9]:
```
assert len(chickenpox_by_sex())==2, "Return a dictionary with two items, the␣
↪first for males and the second for females."
```

```
       HAD_CPOX  P_NUMVRC  SEX
1             2       NaN    1
2             2       NaN    1
3             2       NaN    2
4             2       1.0    2
5             2       0.0    2
...         ...       ...  ...
28461         2       NaN    2
28462         2       NaN    2
28463         2       NaN    2
28464         2       NaN    2
28465         2       NaN    2
```

```
[28465 rows x 3 columns]


       HAD_CPOX  P_NUMVRC  SEX
462         1.0       1.0  1.0
688         1.0       1.0  1.0
864         1.0       1.0  1.0
1480        1.0       1.0  1.0
1907        1.0       1.0  1.0
...         ...       ...  ...
27046       1.0       1.0  1.0
27721       1.0       1.0  1.0
28107       1.0       1.0  1.0
28108       1.0       1.0  1.0
28405       1.0       1.0  1.0

[68 rows x 3 columns]
Hombres contagiados con al menos una vacuna :  68

        HAD_CPOX  P_NUMVRC  SEX
11          2.0       1.0  1.0
13          2.0       1.0  1.0
17          2.0       2.0  1.0
20          2.0       1.0  1.0
22          2.0       1.0  1.0
...         ...       ...  ...
28436       2.0       1.0  1.0
28437       2.0       1.0  1.0
28439       2.0       1.0  1.0
28448       2.0       1.0  1.0
28450       2.0       1.0  1.0

[7028 rows x 3 columns]
Hombres no contagiados con al menos una vacuna :  7028

Male Ratio 0.009675583380762664

        HAD_CPOX  P_NUMVRC  SEX
722         1.0       1.0  2.0
1325        1.0       1.0  2.0
1572        1.0       1.0  2.0
2125        1.0       1.0  2.0
2628        1.0       1.0  2.0
3107        1.0       1.0  2.0
3207        1.0       1.0  2.0
3573        1.0       1.0  2.0
3820        1.0       1.0  2.0
```

```
4205        1.0       1.0  2.0
4470        1.0       1.0  2.0
4561        1.0       1.0  2.0
4763        1.0       1.0  2.0
4778        1.0       1.0  2.0
6907        1.0       1.0  2.0
7316        1.0       1.0  2.0
7606        1.0       1.0  2.0
8987        1.0       1.0  2.0
9829        1.0       1.0  2.0
10925       1.0       1.0  2.0
11426       1.0       1.0  2.0
11489       1.0       1.0  2.0
11710       1.0       1.0  2.0
11979       1.0       1.0  2.0
13401       1.0       1.0  2.0
13535       1.0       1.0  2.0
14157       1.0       2.0  2.0
15720       1.0       1.0  2.0
16036       1.0       1.0  2.0
16037       1.0       1.0  2.0
16230       1.0       1.0  2.0
16279       1.0       1.0  2.0
16799       1.0       1.0  2.0
16952       1.0       1.0  2.0
17452       1.0       1.0  2.0
18347       1.0       1.0  2.0
18821       1.0       1.0  2.0
19036       1.0       1.0  2.0
19362       1.0       1.0  2.0
19412       1.0       1.0  2.0
19826       1.0       1.0  2.0
20150       1.0       1.0  2.0
20825       1.0       1.0  2.0
21222       1.0       1.0  2.0
22491       1.0       1.0  2.0
22492       1.0       1.0  2.0
23009       1.0       1.0  2.0
23056       1.0       1.0  2.0
25024       1.0       1.0  2.0
25142       1.0       1.0  2.0
26740       1.0       1.0  2.0
26967       1.0       1.0  2.0
28315       1.0       1.0  2.0

Mujeres contagiadas con al menos una vacuna :  53

        HAD_CPOX  P_NUMVRC  SEX
```

```
4           2.0       1.0  2.0
7           2.0       1.0  2.0
21          2.0       1.0  2.0
23          2.0       1.0  2.0
25          2.0       1.0  2.0
...         ...       ...  ...
28438       2.0       1.0  2.0
28441       2.0       1.0  2.0
28447       2.0       1.0  2.0
28453       2.0       1.0  2.0
28455       2.0       1.0  2.0

[6802 rows x 3 columns]

Mujeres no contagiadas con al menos una vacuna :  6802

Female Ratio 0.0077918259335489565

 {'male': 0.009675583380762664, 'female': 0.0077918259335489565}
```

1.4 Question 4

A correlation is a statistical relationship between two variables. If we wanted to know if vaccines work, we might look at the correlation between the use of the vaccine and whether it results in prevention of the infection or disease [1]. In this question, you are to see if there is a correlation between having had the chicken pox and the number of chickenpox vaccine doses given (varicella).

Some notes on interpreting the answer. The `had_chickenpox_column` is either `1` (for yes) or `2` (for no), and the `num_chickenpox_vaccine_column` is the number of doses a child has been given of the varicella vaccine. A positive correlation (e.g., `corr > 0`) means that an increase in `had_chickenpox_column` (which means more no's) would also increase the values of `num_chickenpox_vaccine_column` (which means more doses of vaccine). If there is a negative correlation (e.g., `corr < 0`), it indicates that having had chickenpox is related to an increase in the number of vaccine doses.

Also, `pval` is the probability that we observe a correlation between `had_chickenpox_column` and `num_chickenpox_vaccine_column` which is greater than or equal to a particular value occurred by chance. A small `pval` means that the observed correlation is highly unlikely to occur by chance. In this case, `pval` should be very small (will end in `e-18` indicating a very small number).

[1] This isn't really the full picture, since we are not looking at when the dose was given. It's possible that children had chickenpox and then their parents went to get them the vaccine. Does this dataset have the data we would need to investigate the timing of the dose?

[30]:
```
import scipy.stats as stats
import pandas as pd
df=pd.read_csv('assets/NISPUF17.csv', index_col=0)
def corr_chickenpox():
    copy= df
    columnas=copy[["HAD_CPOX","P_NUMVRC"]]
    mask4=(columnas["HAD_CPOX"].lt(3))
```

```
    res=columnas.where(mask4).dropna()
    res.sort_index(inplace=True)
    #print(res)

    #print(res["HAD_CPOX"].unique())
    #print(res["P_NUMVRC"].unique())


    # here is some stub code to actually run the correlation
    corr, pval=stats.pearsonr(res["HAD_CPOX"],res["P_NUMVRC"])
    #print('Correlacion: ',corr)
    #print('pval:          ', pval)
    return corr
corr_chickenpox()
```

[30]:
```
0.07044873460147986
```

[31]:
```
assert -1<=corr_chickenpox()<=1, "You must return a float number between -1.0␣
↪and 1.0."
```

[ ]:
```

```