Statistics learnings experience

The blog about going through the Coursera classes for the data analysis.

Assignment: Running a Chi-Square Test of Independence

Null hypothesis: there is no relation between income and depression.
Alternative hypotheris: there is correlation between income and depression.

Modified code from the task:

```
import pandas
import numpy
import scipy.stats
import seaborn
import matplotlib.pyplot as plt

nesarc_data = pandas.read_csv(‘nesarc_pds.csv’, low_memory=False)

“”“
Null hypothesis: there is no relation between income and depression.
Alternative hypotheris: there is correlation between income and depression
”“”

“”“
206-207 S1Q12B TOTAL HOUSEHOLD INCOME IN LAST 12 MONTHS: CATEGORY
———————————————————————
1531 1. Less than $5,000
2212 2. $5,000 to $7,999
```

1304 3. $8,000 to $9,999
2437 4. $10,000 to $12,999
1288 5. $13,000 to $14,999
3232 6. $15,000 to $19,999
3326 7. $20,000 to $24,999
2961 8. $25,000 to $29,999
3050 9. $30,000 to $34,999
2605 10. $35,000 to $39,999
4407 11. $40,000 to $49,999
3552 12. $50,000 to $59,999
2729 13. $60,000 to $69,999
2084 14. $70,000 to $79,999
1430 15. $80,000 to $89,999
1011 16. $90,000 to $99,999
1171 17. $100,000 to $109,999
451 18. $110,000 to $119,999
939 19. $120,000 to $149,999
745 20. $150,000 to 199,999
628 21. $200,000 or more
———————————————————

2506-2506 S4AQ10BR ONLY/ALL EPISODE(S) LASTED FOR AT LEAST 2 MONTHS (BASED ON S4AQ9E IF ONLY
1 EPISODE)
—————————————————————————-

5008 1. Yes
3636 2. No
173 9. Unknown
34276 BL. NA, worst period did not meet symptom criteria for major depression
"""

new code setting variables you will be working with to numeric
nesarc_data['S1Q12B'] = pandas.to_numeric(nesarc_data['S1Q12B'], errors='coerce')
nesarc_data['S4AQ1'] = pandas.to_numeric(nesarc_data['S4AQ1'], errors='coerce')
nesarc_data['AGE'] = pandas.to_numeric(nesarc_data['AGE'], errors='coerce')

#subset data to young adults age 18 to 25 who have smoked in the past 12 months
#sub1=nesarc_data[(nesarc_data['AGE']>=18) & (nesarc_data['AGE']<=25) & (nesarc_data['CHECK321']==1)]

#make a copy of my new subsetted data
nesarc_subset = nesarc_data.copy()

```
# recode missing values to python missing (NaN)
nesarc_subset['S4AQ1’]=nesarc_subset['S4AQ1’].replace(9, numpy.nan)
nesarc_subset['S4AQ1’]=nesarc_subset['S4AQ1’].replace(1, 'depressed’)
nesarc_subset['S4AQ1’]=nesarc_subset['S4AQ1’].replace(2, 'not depressed’)

#recoding values for S3AQ3B1 into a new variable, S1Q12B
#recode1 = {1: 30, 2: 22, 3: 14, 4: 6, 5: 2.5, 6: 1}
#nesarc_subset['S1Q12B’]= nesarc_subset['S3AQ3B1’].map(recode1)

# contingency table of observed counts
cross_tab = pandas.crosstab(nesarc_subset['S4AQ1’], nesarc_subset['S1Q12B’])
print (cross_tab)

# column percentages
colsum=cross_tab.sum(axis=0)
colpct=cross_tab/colsum
print(colpct)

# chi-square
print ('chi-square value, p value, expected counts’)
chi_square = scipy.stats.chi2_contingency(cross_tab)
print (chi_square)

# set variable types
nesarc_subset[“S1Q12B”] = nesarc_subset[“S1Q12B”].astype('category’)
nesarc_subset[“S4AQ1”] = nesarc_subset[“S4AQ1”].astype('category’)

# graph percent with nicotine dependence within each smoking frequency group
#seaborn.factorplot(x=“S1Q12B”, y=“S4AQ1”, data=nesarc_subset, kind=“bar”, ci=None)
#plt.xlabel('Income’)
#plt.ylabel('Depression’)

# compare all pairs, we are looking for p = 5 % / (21 + 20 / 2) = 0.024 % or 0.00024
for category_a in xrange(1,21):
 for category_b in xrange(1,21):
   recode_ab = {category_a: category_a, category_b: category_b}
   nesarc_subset_ab= nesarc_subset['S1Q12B’].map(recode_ab)
   ct=pandas.crosstab(nesarc_subset['S4AQ1’], nesarc_subset_ab)
   colsum=ct.sum(axis=0)
   colpct=ct/colsum
   cs = scipy.stats.chi2_contingency(ct)
   print str(category_a) + “ vs ” + str(category_b)
   print (cs)
```

Result:

The test shows significance of the alternative hypothesis and p value is 4.7264859104118858e-13 that is smaller than 5%.

Pair analysis displays that large income correlates with smaller percentage of people that had a depression.

S1Q12B 1 2 3 4 5 6 7 8 9 10 … \

S4AQ1 …

depressed 503 761 452 778 390 974 967 897 936 742 …

not depressed 987 1388 815 1595 875 2177 2293 2009 2047 1817 …

S1Q12B 12 13 14 15 16 17 18 19 20 21

S4AQ1

depressed 1007 780 592 414 264 296 124 241 213 169

not depressed 2487 1903 1458 1001 731 852 321 680 522 440

[2 rows x 21 columns]

S1Q12B 1 2 3 4 5 6 \

S4AQ1

depressed 0.337584 0.354118 0.356748 0.327855 0.3083 0.309108

not depressed 0.662416 0.645882 0.643252 0.672145 0.6917 0.690892

S1Q12B 7 8 9 10 … 12 \

S4AQ1 …

depressed 0.296626 0.308672 0.313778 0.289957 … 0.288208

not depressed 0.703374 0.691328 0.686222 0.710043 … 0.711792

S1Q12B 13 14 15 16 17 18 \

S4AQ1

depressed 0.290719 0.28878 0.29258 0.265327 0.25784 0.278652

not depressed 0.709281 0.71122 0.70742 0.734673 0.74216 0.721348

S1Q12B 19 20 21

S4AQ1

depressed 0.261672 0.289796 0.277504

not depressed 0.738328 0.710204 0.722496

[2 rows x 21 columns]

chi-square value, p value, expected counts

(102.37294021704328, 4.7264859104118858e-13, 20, array([[451.40281036, 651.0500936 , 383.84386626, 718.91199261,

383.23795645, 954.61090969, 987.63299448, 880.38695766,

903.71448544, 775.26160518, 1303.61496173, 1058.52444255,

812.82801355, 621.05755788, 428.68119239, 301.44013175,

347.79223241, 134.8149333 , 279.02146869, 222.67185612,

184.49953793],

[1038.59718964, 1497.9499064 , 883.15613374, 1654.08800739,

881.76204355, 2196.38909031, 2272.36700552, 2025.61304234,

2079.28551456, 1783.73839482, 2999.38503827, 2435.47555745,

1870.17198645, 1428.94244212, 986.31880761, 693.55986825,

800.20776759, 310.1850667 , 641.97853131, 512.32814388,

424.50046207]]))

COMP1v15 1 15

S4AQ1

depressed 503 414

not depressed 987 1001

COMP1v15 1 15

S4AQ1

depressed 0.337584 0.29258

```
not depressed  0.662416  0.70742

chi-square value, p value, expected counts

(6.5980203670618227, 0.0102092216197002, 1, array([[  470.3373494,   446.6626506],
       [ 1019.6626506,   968.3373494]]))

COMP1v17        1    17

S4AQ1

depressed      503  296

not depressed  987  852

COMP1v17            1        17

S4AQ1

depressed      0.337584  0.25784

not depressed  0.662416  0.74216

chi-square value, p value, expected counts

(19.152804709353539, 1.2066007461010751e-05, 1, array([[  451.29264594,   347.70735406],
       [ 1038.70735406,   800.29264594]]))

COMP1v21        1     14

S4AQ1

depressed      503   592

not depressed  987  1458

COMP1v21            1        14

S4AQ1

depressed      0.337584  0.28878

not depressed  0.662416  0.71122

chi-square value, p value, expected counts

(9.3923797517081447, 0.0021788918981249464, 1, array([[  460.88983051,   634.11016949],
       [ 1029.11016949,  1415.88983051]]))
```

```
1 vs 1

(0.0, 1.0, 0, array([[ 503.],

    [ 987.]]))

1 vs 2

(0.98944086025962941, 0.31987907153101047, 1, array([[  517.54877714,   746.45122286],

    [  972.45122286,  1402.54877714]]))

1 vs 3

(1.027748122579103, 0.31068815634078129, 1, array([[ 516.12259703,  438.87740297],

    [ 973.87740297,  828.12259703]]))

1 vs 4

(0.34820341989938663, 0.55513189655000028, 1, array([[  494.09526275,   786.90473725],

    [  995.90473725,  1586.09526275]]))

1 vs 5

(2.5460135620036612, 0.110572933003142, 1, array([[  482.96551724,   410.03448276],

    [ 1007.03448276,   854.96551724]]))

1 vs 6

(3.6506387202353463, 0.056047713428024691, 1, array([[  474.19306184,  1002.80693816],

    [ 1015.80693816,  2148.19306184]]))

1 vs 7

(7.8370953629686371, 0.0051184786427701098, 1, array([[  461.11578947,  1008.88421053],

    [ 1028.88421053,  2251.11578947]]))

1 vs 8

(3.6613863970142435, 0.055687276973314236, 1, array([[  474.52229299,   925.47770701],

    [ 1015.47770701,  1980.52229299]]))

1 vs 9

(2.4727028220024185, 0.11583857597284593, 1, array([[  479.34495864,   959.65504136],
```

```
       [ 1010.65504136,  2023.34495864]]))
```

1 vs 10

```
(9.8089725590682644, 0.0017366249925246449, 1, array([[  458.15016053,   786.84983947],
       [ 1031.84983947,  1772.15016053]]))
```

1 vs 11

```
(7.6892879501455713, 0.0055549539583963454, 1, array([[  459.88606939,  1328.11393061],
       [ 1030.11393061,  2974.88606939]]))
```

1 vs 12

```
(11.826123984059663, 0.0005840544721474588, 1, array([[  451.42455859,  1058.57544141],
       [ 1038.57544141,  2435.42455859]]))
```

1 vs 13

```
(9.6625551355525214, 0.0018806187031832511, 1, array([[  458.10448119,   824.89551881],
       [ 1031.89551881,  1858.10448119]]))
```

1 vs 14

```
(9.3923797517081447, 0.0021788918981249464, 1, array([[  460.88983051,   634.11016949],
       [ 1029.11016949,  1415.88983051]]))
```

1 vs 15

```
(6.5980203670618227, 0.0102092216197002, 1, array([[  470.3373494,   446.6626506],
       [ 1019.6626506,   968.3373494]]))
```

1 vs 16

```
(14.260898678275877, 0.00015913661565415731, 1, array([[  459.89134809,   307.10865191],
       [ 1030.10865191,   687.89134809]]))
```

1 vs 17

```
(19.152804709353539, 1.2066007461010751e-05, 1, array([[  451.29264594,   347.70735406],
       [ 1038.70735406,   800.29264594]]))
```

#coursera

Feb 21st, 2016

MORE YOU MIGHT LIKE

Data Analysis Tools. Week 1. Hypothesis Testing and ANOVA.

Instructions

The assignments for this course start where the Data Management and Visualization course assignments left off. Now that you have selected a data set and research question, managed your variables of interest and visualized their relationship graphically, we are ready to test those relationships statistically. We have included the codebooks and data sets from Data Management and Visualization for your convenience. The first assignment deals with analysis of variance. Analysis of variance assesses whether the means of two or more groups are statistically different from each other. This analysis is appropriate whenever you want to compare the means (quantitative variables) of groups (categorical variables). The null hypothesis is that there is no difference in the mean of the quantitative variable across groups (categorical variable), while the alternative is that there is a difference. Note that if your research question does not include one quantitative variable, you can use one from your data set just to get some practice with the tool. If your research question does not include a categorical variable, you can categorize one that is quantitative.

Instructionsless Run an analysis of variance.

You will need to analyze and interpret post hoc paired comparisons in instances where your original statistical test was significant, and you were examining more than two groups (i.e. more than two levels of a categorical, explanatory variable).

WHAT TO SUBMIT:Following completion of the steps described above, create a blog entry where you submit syntax used to run an ANOVA (copied and pasted from your program) along with corresponding output and a few sentences of interpretation.

Review Criterialess Your assessment will be based on the evidence you provide that you have completed all of the steps. In all cases, consider that the peer assessing your work is likely not an expert in the field you are analyzing.