Inferential Analysis

Anova Analysis

With

Post Hoc Tukey

And

Python

AUTHOR:

LUIS ORELLANA ALTAMIRANO

Anova Analysis

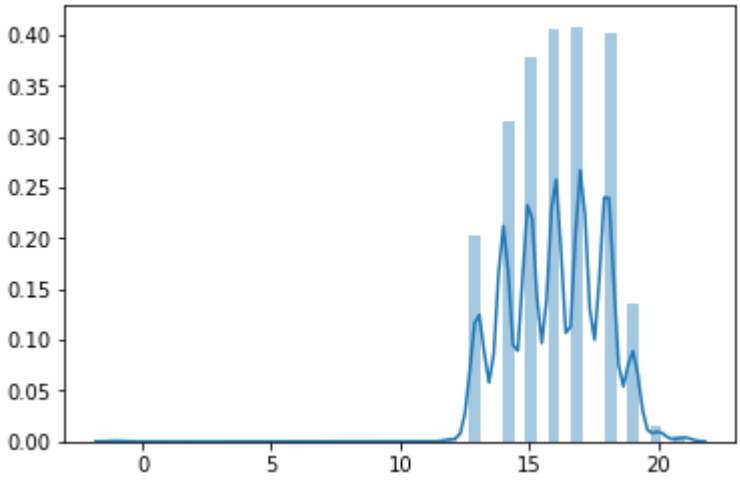
The dataset that was taken for this exercise is “Add Health”[1]. It corresponds to a survey which was focused to adolescents. Something interesting to analyze is the relation between if they enjoy life and their ages. In order to accomplish it, two fields were used, “You enjoyed life” (H1GH11) and the age. The last field that was mentioned, is calculated through this approached:

Age = (Year the survey was conducted) - (year of birth)

The scale of life’s enjoyment is as next:

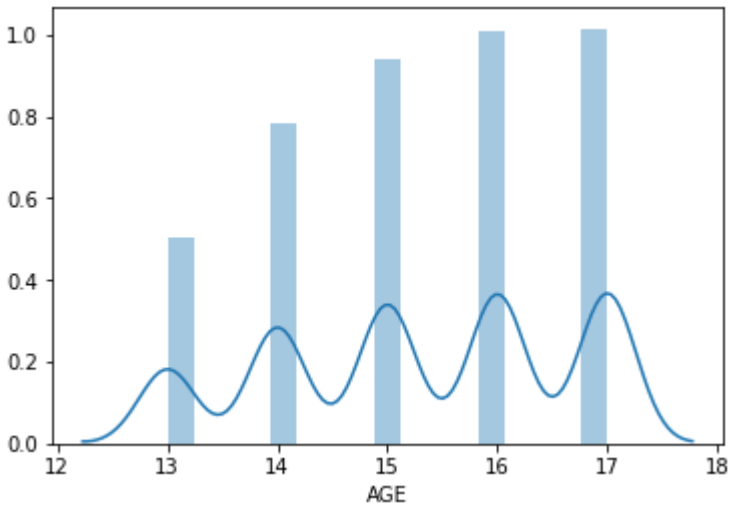
15. You enjoyed life.		
255	0	never or rarely
1043	1	sometimes
2047	2	a lot of the time
3141	3	most of the time or all of the time
8	6	refused
10	8	don’t know

The distribution of age is as below:



However, there is a merely few people who are older than 17 years old. Therefore, they could be taken as outliers. That’s why it was considered the next distribution:

[1] Source: http://www.cpc.unc.edu/projects/addhealth



Due that the explanatory variable (age) has seven levels, post hoc tukey was used. The outcome was:

group1	group2	meandiff	lower	upper	reject
0	1	0.1963	-0.1062	0.4989	False
0	2	0.1303	-0.1562	0.4167	False
0	3	-0.0928	-0.3729	0.1872	False
0	6	0.3963	-1.7981	2.5906	False
0	8	-0.0204	-1.9255	1.8847	False
1	2	-0.0661	-0.2346	0.1025	False
1	3	-0.2892	-0.4465	-0.1318	True
1	6	0.1999	-1.9822	2.382	False
1	8	-0.2168	-2.1078	1.6742	False
2	3	-0.2231	-0.3468	-0.0993	True
2	6	0.266	-1.9139	2.4459	False
2	8	-0.1507	-2.0392	1.7378	False
3	6	0.4891	-1.69	2.6682	False
3	8	0.0724	-1.8151	1.9599	False
6	8	-0.4167	-3.2976	2.4642	False

It’s possible to see, the group 1 is the level of enjoyment and group 2 is the age. In order to check the group that report the strongest feeling of happiness, We will focus only in the level 2,and 3. And the observation indicates that people who are 13 and 14 years old reported have been happier than other people in this survey.

And the means are:

means for age by happiness status

H1FS15	AGE
0	15.270408
1	15.466755
2	15.400667
3	15.177605
6	15.666667
8	15.250000

The standard deviations are:

standard deviations for age by happiness status

H1FS15	AGE
0	1.258139
1	1.281419
2	1.301254
3	1.354955
6	0.577350
8	0.500000

It’s possible to see the means are different for the group 3 and the standard deviation is above 1.

Appendix

Finally, the Python code used through this exercise is:

```
import numpy
import pandas
import statsmodels.formula.api as smf
import statsmodels.stats.multicomp as multi
import matplotlib.pyplot as plt
import seaborn as sns

data = pandas.read_csv(‘addhealth_pds.csv’, low_memory=False)

data = data[['IYEAR’, 'H1GI1Y’, 'H1FS15’]].dropna()

data['AGE’]=data['IYEAR’] - data['H1GI1Y’]

data['AGE’]= data['AGE’].convert_objects(convert_numeric=True)

sub1 = data[['AGE’, 'H1FS15’]].dropna()
sub1=sub1[(sub1['AGE’] >= 13)&(sub1['AGE’] <= 17)]

print ('means for AGE by feeling tired status’)
m1= sub1.groupby('H1FS15’).mean()
print (m1)

print ('standard deviations for AGE by feeling tired status’)
sd1 = sub1.groupby('H1FS15’).std()
print (sd1)

mc1 = multi.MultiComparison(sub1['AGE’], sub1['H1FS15’])
res1 = mc1.tukeyhsd()
print(res1.summary())

sns.distplot(sub1['AGE’]);
```