**Week 1 ANOVA testing**

This week we run a simple ANOVA test on European countries to see how income per person is affected my NATO and EU membership. All code is available here: <https://github.com/brennap3/Gapminder/blob/master/Gapminder_Analysis_2015.py>.

The test determines whether there is a difference or not between the means of the different groups or whether the difference is owing to some random variation. The test involves the following steps:

* Compute the group means and standard deviation
* Calculate the within group variance
* Calculate the between group variance
* Compute the F statistic by utilizing the within and between group variance
* Check the significance of the F-statistic

The different groups means, variances and standard deviations tested are:

NATO\_EU\_MEMBERSHIP

Nato\_And\_EU 16145.634068

Nato\_Not\_In\_EU 20295.315233

Not\_In\_Nato\_In\_EU 23345.322253

Not\_In\_Nato\_Not\_In\_EU 19767.767749

The different groups variances are tested are:

print (varianceincomeperpersn)

incomeperperson

NATO\_EU\_MEMBERSHIP

Nato\_And\_EU 1.586113e+08

Nato\_Not\_In\_EU 3.782496e+08

Not\_In\_Nato\_In\_EU 6.775285e+07

Not\_In\_Nato\_Not\_In\_EU 1.058797e+09

The different groups standard deviations tested are:

incomeperperson

NATO\_EU\_MEMBERSHIP

Nato\_And\_EU 12594.097210

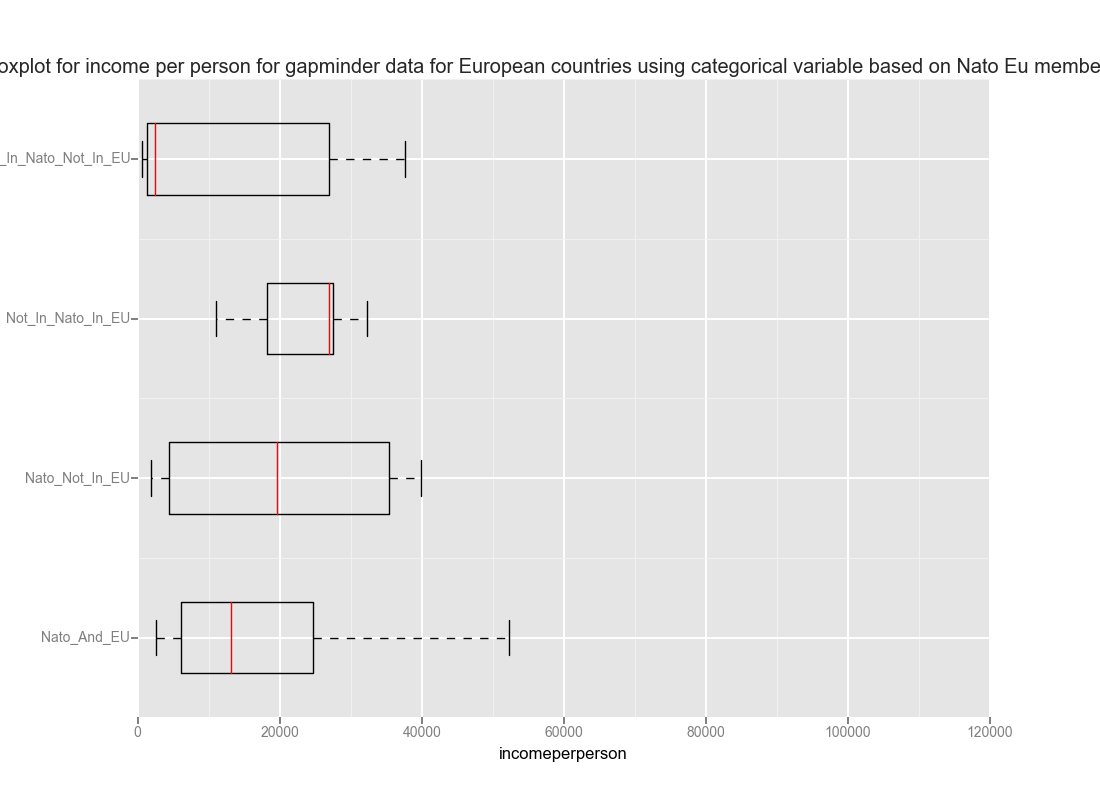
Nato\_Not\_In\_EU 19448.639876

Not\_In\_Nato\_In\_EU 8231.212114

Not\_In\_Nato\_Not\_In\_EU 32539.161967

Comparing the different groups

Comparing the different groups using multiple boxlpots



Using a boxplot we can see the distribution of the income’s per person for each country in Europe based on the following classes:

1. Nato\_And\_EU
2. Nato\_Not\_In\_EU
3. Not\_In\_Nato\_In\_EU
4. Not\_In\_Nato\_Not\_In\_EU

The difference in the distribution can be seen from the above plot, witht Not\_in\_Nato\_In EU showing the narrowest variation and the highest median values.

**Running the ANOVA analysis**

The analysis of variance (and standard deviation) within groups shows that equality of homogeneity assumption maybe violated. Running the ANOVA analysis using the general linear model function we can see that there is evidence obtained that there are differences amongst the different groups. The ANOVA analysis gave an F-statistic of 8.026 and a p value of 4.66e-05 (Prob (F-statistic): 4.66e-05), which is less than our critical value of 0.05. This would indicate that there is a difference in means of the different groups. The model and results are run using the following code:

model1 = smf.ols(formula='incomeperperson ~ C(NATO\_EU\_MEMBERSHIP)', data=data)

results1 = model1.fit()

print(results1.summary())

##F-statistic: 8.026

##Prob (F-statistic): 4.66e-05

To check the pairwise comparisons, a between the different groups Tukey’s honest significant difference test is run in combination with the ANOVA as a post hoc test to show the pairs of groups that the means that are significantly different.

Between the pairs we can find no evidence of statistically significant difference in means (from the summary output we fail to reject the null hypothesis that there is no difference in means).

import statsmodels.stats.multicomp as multi

mc1 = multi.MultiComparison(dataanovatestdf['incomeperperson'],dataanovatestdf['NATO\_EU\_MEMBERSHIP'])

res1 = mc1.tukeyhsd() ##tkeys honestly different test

print(res1.summary())

Multiple Comparison of Means - Tukey HSD,FWER=0.05

================================================================================

group1 group2 meandiff lower upper reject

--------------------------------------------------------------------------------

Nato\_And\_EU Nato\_Not\_In\_EU 4149.6812 -27680.6426 35980.0049 False

Nato\_And\_EU Not\_In\_Nato\_In\_EU 7199.6882 -19850.8854 34250.2618 False

Nato\_And\_EU Not\_In\_Nato\_Not\_In\_EU 3622.1337 -16227.5606 23471.828 False

Nato\_Not\_In\_EU Not\_In\_Nato\_In\_EU 3050.007 -34462.3892 40562.4033 False

Nato\_Not\_In\_EU Not\_In\_Nato\_Not\_In\_EU -527.5475 -33230.0964 32175.0014 False

Not\_In\_Nato\_In\_EU Not\_In\_Nato\_Not\_In\_EU -3577.5545 -31649.2614 24494.1524 False

--------------------------------------------------------------------------------

**Appendix 1: Code to run analysis all code available here**

##week 1

##filter european countries

##filter out NA's

##select columns

##run ANOVA

##

##checking coloumns

##data.columns.values

##checking values

##data['European']

dataanovatestdf=data[['country','incomeperperson','NATO\_EU\_MEMBERSHIP']][data.European=='Europe']

##dataanovatestdf

model1 = smf.ols(formula='incomeperperson ~ C(NATO\_EU\_MEMBERSHIP)', data=data)

results1 = model1.fit()

print(results1.summary())

##F-statistic: 8.026

##Prob (F-statistic): 4.66e-05

## p value less than 0.05 a difference in variance in different groups

print ('means for incomme per person for different groups')

meansincomeperpersn= dataanovatestdf.groupby('NATO\_EU\_MEMBERSHIP').mean()

print (meansincomeperpersn)

##lets display it using boxplots

print ('variances for incomme per person for different groups')

varianceincomeperpersn= dataanovatestdf.groupby('NATO\_EU\_MEMBERSHIP').var()

print (varianceincomeperpersn)

##

print ('standard deviations for incomme per person for different groups')

standarddeviationincomeperpersn= dataanovatestdf.groupby('NATO\_EU\_MEMBERSHIP').std()

print (standarddeviationincomeperpersn)

##boxplot

ggplot(dataanovatestdf, aes(x='incomeperperson', y='NATO\_EU\_MEMBERSHIP')) + geom\_boxplot() +\

xlab("incomeperperson") + ylab(" Nato EU membership status") + ggtitle("Boxplot for income per person for gapminder data for European countries using categorical variable based on Nato Eu membership")

##

##run your post hoc tests

import statsmodels.stats.multicomp as multi

mc1 = multi.MultiComparison(dataanovatestdf['incomeperperson'], dataanovatestdf['NATO\_EU\_MEMBERSHIP'])

res1 = mc1.tukeyhsd() ##tkeys honestly different test

print(res1.summary())

##must be performed after

##cant run pairwise