Data Analysis Tools

**Assignment – Week 1**

**Running the Analysis of Variance**

**ANOVA F test**

**By mapolarbear@gmail.com**

Started new course Data Analysis Tools, the first week was very interesting, lots of conceptual knowledge related to how to test and define Zero Hypothesis (H0), what role it plays, how to reject or accept H0 and why and how to accept Alternate Hypothesis (Ha). It was also interesting to go through ANOVA procedure, how to better understand it’s output.

**Research question:** As long as the main study of my research is female and male suicide ratings and what possible relationship they have to employment/unemployment and other ratings. To complete my assignment and extend my research, I decided to add some flavor to research, to check if **fsuicides2004** and **msuicides2004** are correlating to **region** to which countries belong. Another words, if, let say suicides rating in North-American countries are different from Oceania, or from Asia or Eastern Europe. If there is a dependency. And, again if there is a difference between female and male statistics.

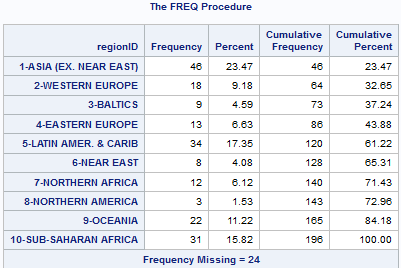
**Zero Hypothesis (H0)**: There is no correlation between suicide ratings and region of country.

**Alternate Hypothesis (Ha):** There is a correlation between suicide ratings and region of country.

**Variables:**

**fsuicides2004, msuicides2004 -** Female and male suicide ratings

**regionID -** Categorical variable created for the purpose of my research question and represents countries grouped into 10 regions, based on World Back classification. Each region has its unique numeric ID, with ranging values from 1 to 10. Here is the distribution output of FREQ procedure:



**Missing values -** Some countries didn’t get corresponding region ID (24 missing values), just because of different spelling or country names. All countries with missing **regionID** excluded from study, because majority of them do not have suicide rating values anyway.

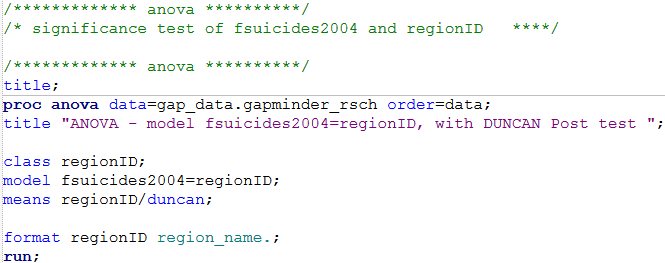
**Statistical procedures** **and methods**

**ANOVA** - A one-way analysis of variance considers one treatment factor with two or more treatment levels. The goal of the analysis is to test for differences among the means of the levels and to quantify these differences. If there are two treatment levels, this analysis is equivalent to a T test comparing two group means.

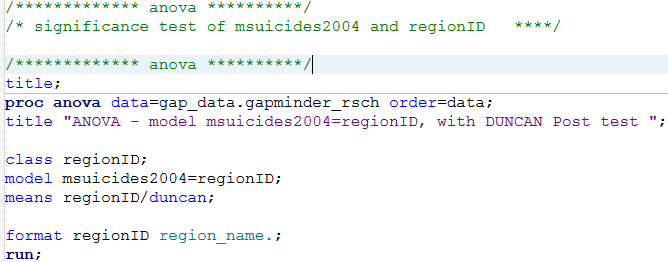
**DUNCAN** test for posttest check.

**Program code:**

1. **ANOVA - model fsuicides2004=regionID, with DUNCAN Post test**



1. **ANOVA - model msuicides2004=regionID, with DUNCAN Post te**st



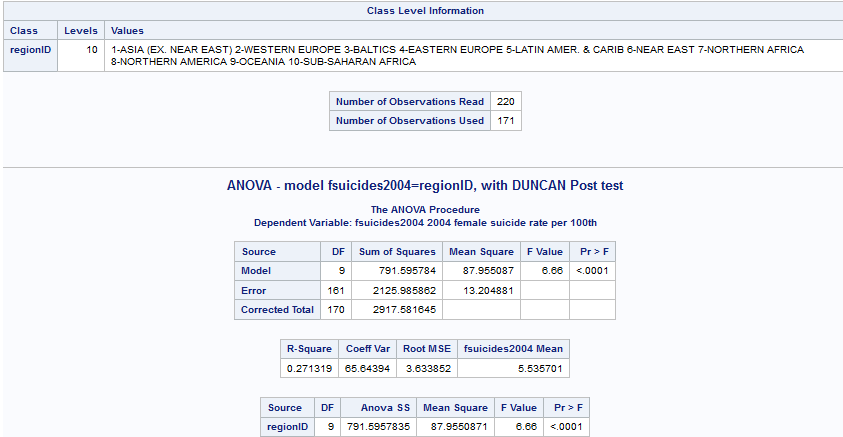
**Interpretation of ANOVA output results**

* **Model fsuicides2004=regionID**

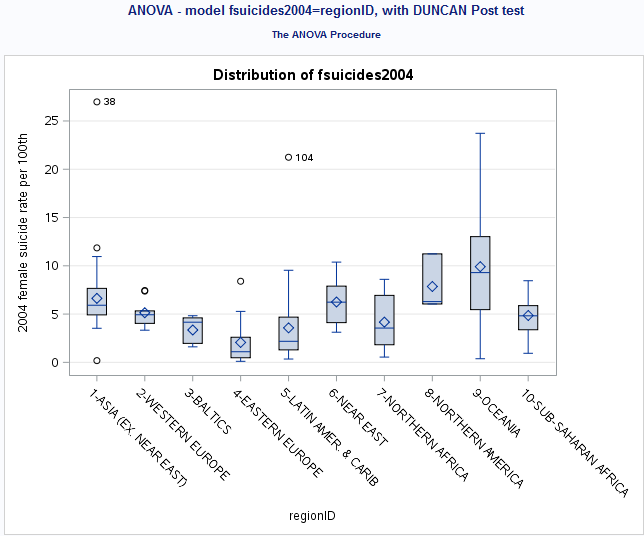
The "Class Level Information" table shows list the variables that appear in CLASS variable **regionID**, their levels, and the number of observations in the data set.

The degrees of freedom (DF) column we are using to check the analysis results is 9. The model degrees of freedom for a one-way analysis of variance are the number of levels minus 1; in this case, 10 – 1 = 9. The Corrected Total degrees of freedom is 170; in this case 171 – 1 = 170. The sum of Model and Error degrees of freedom equal the Corrected Total.

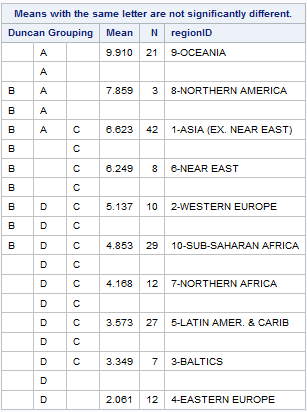
The **overall F test is significant** (**F=6.66, P < .0001**), indicating that the model as a whole accounts for a significant portion of the variability in the dependent variable. The **F test for regionID is significant**, indicating that some contrast between the means for the different **regionID** is different from zero. I think, preliminary, we can reject **Zero Hypothesis** **(H0)** and check DUNCAN test to make sure that we can accept **Alternate Hypothesis (Ha)**



ANOVA procedure output also includes a box plot of the dependent variable values within each classification level of the independent variable. Based on box plot graphics we can check how wide **fsuicides2004** values are in each of 10 country regions, including position **MAX** values, **MEDIAN** (horizontal line) and **MEAN** (little diamonds) values.



Results of **DUNKAN**’s procedure with examples of implications of the multiple comparisons results are as follows:



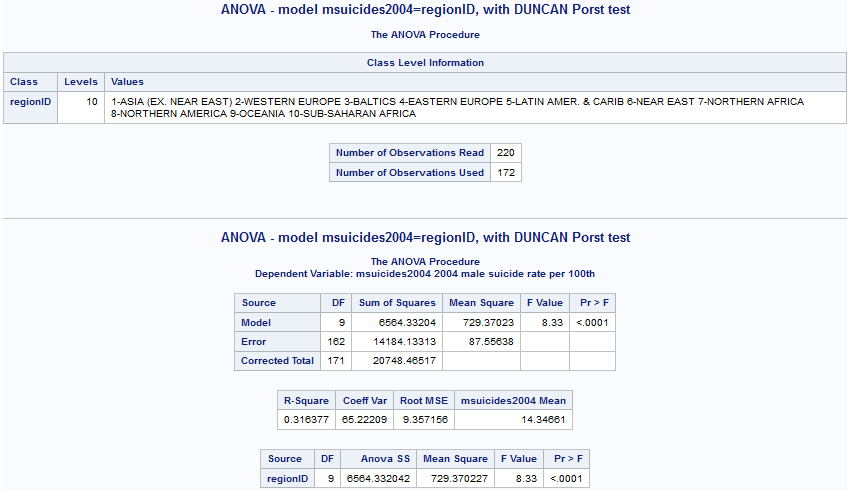
* **Group A** - 9-OCENIA has significantly higher value of 9.910 among other regions in the group.
* **Group B** – 8-NOTHERN AMERICA has significantly higher value of 7.859 among other regions in the group with lowest of 4.853 for 10-SUN-SAHARAN AFRICA.
* **Group C** – 1-ASIA (EX. NEAR EAST) has significantly higher value of 6.623 among in the group with lowest of 3.349 for 3-BALTICS.
* **Group D** – 2-WESTERN EUROPE with value of 6.623 among is not significantly higher than other regions in the group, but the lowest value of 2.061 for 4-EASTERN EUROPE, which is significantly lower than 6.623.

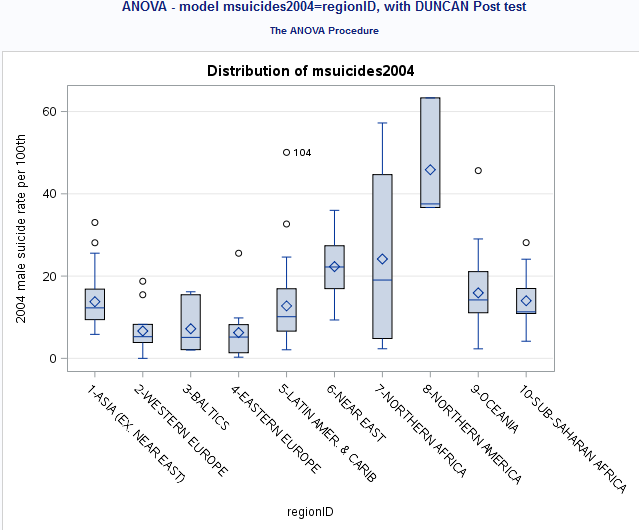
**Conclusion:**

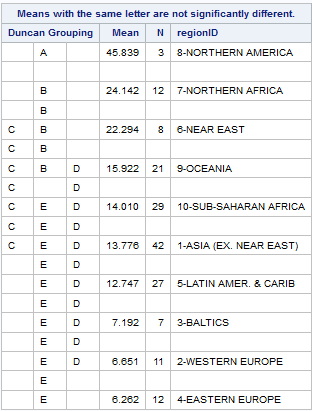
**I think we can accept Alternate Hypothesis (Ha) and make a conclusion that rating of suicides (fsuicides2004 in current study) is dependent on the countries region and has significant relationship.**

**Although the experiment has succeeded in separating the regions with highest female suicide ratings from the lowest.**

* In addition, here is ANOVA statistics for **model msuicides2004=regionID**
* The "Class Level Information" table shows list the variables that appear in CLASS variable **regionID**, their levels, and the number of observations in the data set.
* The degrees of freedom (DF) column we are using to check the analysis results is 9. The model degrees of freedom for a one-way analysis of variance are the number of levels minus 1; in this case, 10 – 1 = 9. The Corrected Total degrees of freedom is 170; in this case 171 – 1 = 170. The sum of Model and Error degrees of freedom equal the Corrected Total.
* The **overall F test is significant** (**F=8.33, P < .0001**), indicating that the model as a whole accounts for a significant portion of the variability in the dependent variable. The **F test for regionID is significant**, indicating that some contrast between the means for the different **regionID** is different from zero. I think, preliminary, we can reject **Zero Hypothesis** **(H0)** and check DUNCAN test to make sure that we can accept **Alternate Hypothesis (Ha)**







\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* End of week 1 assignment \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

<http://coursera-sas-tools-a-week-1.tumblr.com/post/152079337232/data-analysis-tools-assignment-week-1-running>

<http://coursera-sas-dm-week4.tumblr.com/post/151798525721/data-management-and-visualization-assignment>