Data Analysis Tools

**Assignment – Week 2**

**Running a Chi-Square Test of Independence**

**By mapolarbear@gmail.com**

This assignment is dedicated to **CHI Square Independence test**, **to determine level of independence/dependence between two variables**. In theory, in case if **p** value is greater than 0.05, means that variables are more likely to be independent and we cannot safely reject **Zero Hypothesis (H0)**, that variables are not dependent on each other. On a contrary, if our study provides opposite result, **p value is less <= 0.05** and result is significant, we may accept **Alternate Hypothesis (Ha)**, that variables are dependent on each other (when value of one variable changes, the value of another variable also changing ) and we also may have to conduct post-hoc test to make sure that we can safely reject H0 hypothesis.

**Research question:** If changes in in 2004 suicide rates of females and males are related to changes in corresponding 2004 gender employment rate statistics and if we can safely reject Zero Hypothesis(H0) or not, and we can accept Alternative Hypothesis(Ha).

This research study will be little bit tricky, because GAPMINDER doesn’t has categorical variable, and for purpose of my study I have to create my own categorical variables, grouping **fsucides2004, msuicides2004** and **employrate\_f2004, employrate\_m2004** respectively.

My research question in this case will be if newly created categorized variables dependent on each other:

* **fsucidesAVG** dependent on **employrate\_f2004CAT** and vice versa
* **fsucidesAVG** dependent on **femployrateAVG**  and vice versa
* **msuicidesAVG** dependent on **employrate\_m2004CAT** and vice versa
* **msucidesAVG** dependent on **memployrateAVG**  and vice versa

**Zero Hypothesis (H0)**: Suicide ratings **are not dependent** on employment rating for each gender.

**Alternate Hypothesis (Ha):** Suicide ratings **are dependent** on employment rating for each gender.

**Variables:**

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

**fsuicidesAVG, msuicidesAVG -** New, categorized variables, representing suicide ratings ABOVE and BELOW rating **means**.

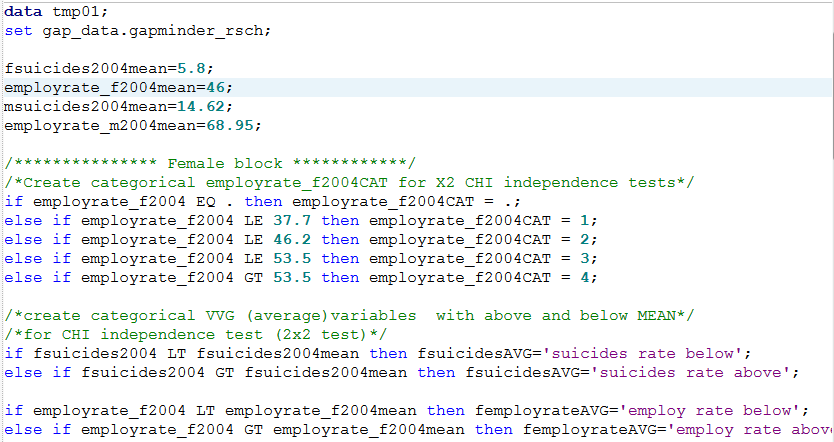
**femployrateAVG, memployrateAVG -** New, categorized variables, representing suicide ratings ABOVE and BELOW rating **means**.

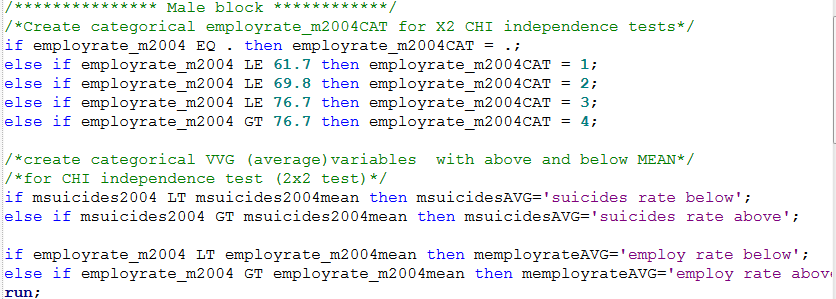
**employrate\_f2004CAT, employrate\_m2004CAT -** New, categorized variables into 4 groups each, representing 2004 employment rate for corresponding gender.

Variables for **C 🡪 C** (2x2) CHI Square test: **fsuicidesAVG, msuicidesAVG** versus **femployrateAVG, memployrateAVG**

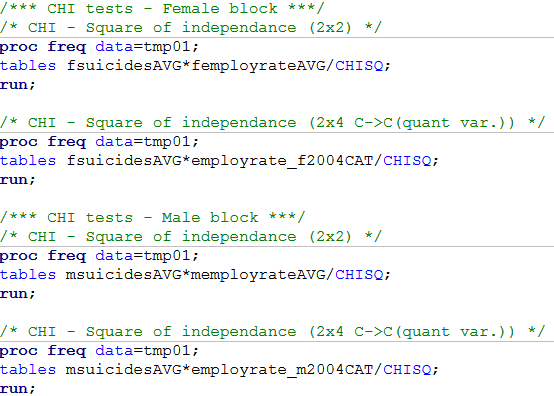
Variables for **C 🡪 Q** (2x4) X2 CHI Square test**: fsuicidesAVG, msuicidesAVG** versus categorized into 4 groupseach **employrate\_f2004CAT, employrate\_m2004CAT**

**Code to create categorical variables:**





**CHI Square tests of Independence:**

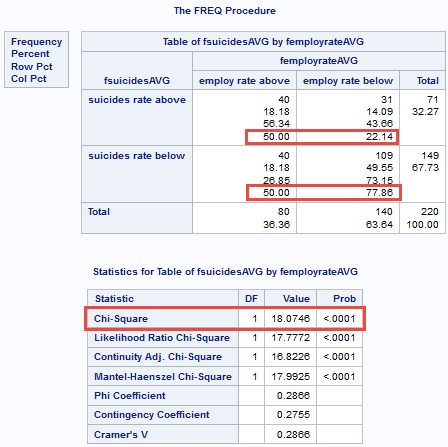


**CHI Square tests of Independence – Sample Results for FEMALE portion only, MALE portion failed test completely and Null Hypothesis was accepted:**

Variables for **C 🡪 C** (2x2) CHI Square test: variables **femployrateAVG** and **fsuicidesAVG**

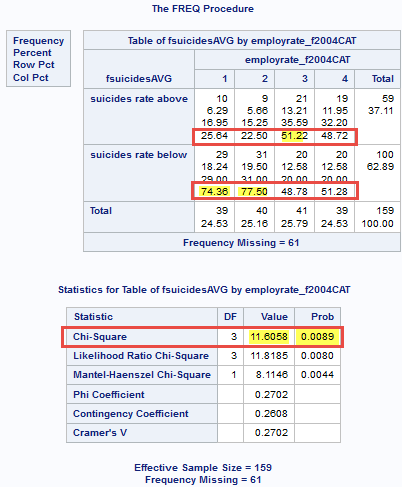
Based on results, we can conclude that column percent of ‘suicides rate above’ and **‘employ rate above**’ is **50%** and **significantly higher**, about 30% than corresponding ‘employ rate below’, which is **22.14**%. In addition, ‘**suicides rate below’** and ‘**employ rate above**’ is **50%** and significantly lower that corresponding ‘**employ rate below**’, which is **77.86**. In addition, **Chi-Square value has value of 18.0746, and Prob (p-value) of <.0001**, which is very significant and we can safely reject **Null Hypothesis(H0)** and accept **Alternate Hypothesis(Ha)** and accept the fact that **femployrateAVG** and **fsuicidesAVG** are dependent variables. In addition please note that ‘**employ rate above**’ and **‘..Below’** with corresponding **‘below**’ and ‘**above**’ suicides are very well distributed, with 50% each.

This findings confirms our previous tests, that suicide rate is related to employment rate and it looks like the higher female employment rate, the higher suicide rate and in opposite side, the lower female employment rate the lower suicide rate. This is quite interesting findings, but again, these categorical variables are sort of ‘made up’, ‘artificial, build out of normal quantitative variables.



Variables for **C 🡪 Q (categorized exploratory QUANT variable)** (2x4) X2 CHI Square test**:** variables **fsuicidesAVG** and **employrate\_f2004CAT**

Based test results you can CHI Square Independence test between variables, we can see that column percent of **‘suicides rate above’ for group 3 (46.2 – 53.5) of employrate\_f2004CAT is significantly high: 51.22**, than **25.64** and **22.50** for **groups 1 (0-37.7) and 2 (37.7-46.2)** and for group **4(53.5)** somewhere in between lowest (22.50) and highest (51.22) and it looks like we cannot safely reject NULL Hypothesis(H0), because only group 3 and 4 have significant difference. However, Chi Square statistics showing that there is overall **significant difference because of 11.6058** value **and p-value significantly below 0.05 and is 0.0089**. At the same time we can see that column percentage for ‘suicides bellow’ rates are significantly different for corresponding groups, but values between employrate\_f2004CAT groups 1 and 2 are not significantly different and not significantly different between groups 3 and 4.

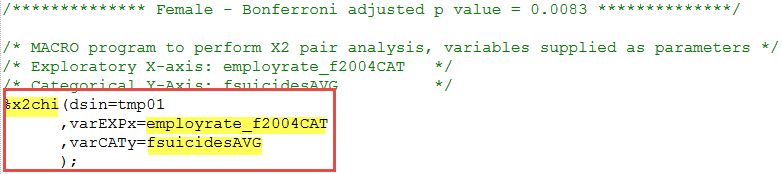


Based on observed results we can make a conclusion that we cannot reject Null Hypothesis(H0) and cannot accept Alternate Hypothesis(Ha), we need to conduct post-hoc test between pairs of groups for all possible combinations, total of 6, and use **Bonferroni adjusted p-value**(0.05/4=**0.0083**) to determine if we can safely reject H0 hypothesis and accept Ha hypothesis and conclude that variables **fsuicidesAVG** and **employrate\_f2004CAT** are dependent to each other.

For the purpose of current test and future **Chi Square X2 independence tests**, I’ve the **macro program** x2chi, which determines how many groups exploratory variable has and runs CHISQ for each pair. Here is the link to x2chi code:

<https://github.com/mapolarbear/Data-Analysis-Interpretation-Wesleyan-SAS/blob/master/x2chi>

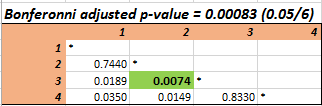
To run program, following macro execution added into my open code program:

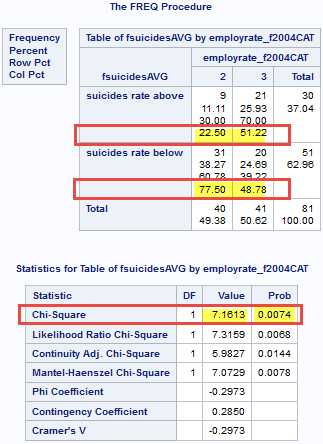


**Here is output for all pairs:**

For only one pair, (2 and 3) we can safely **reject H0 hypothesis** and **accept Ha alternate hypothesis:**

Pair 2 and 3 - There is significant difference (7.1613), p-value = 0.0074, there is a difference in column percentage values, p-value = 0.0074 is below **Bonferroni adjusted p-value of 0.00083 is little**, we **can safely reject H0 hypothesis or accept Ha hypothesis**





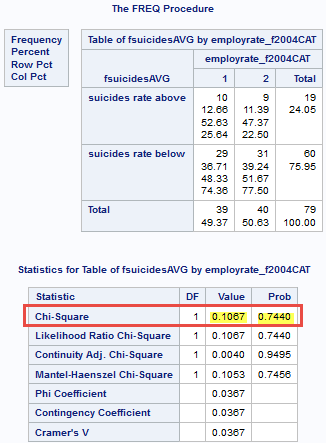
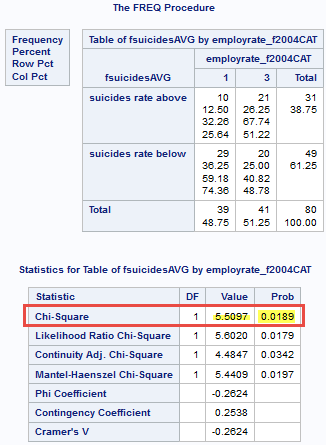
Pair 1 and 2 - No significant difference (1.1067), p-value = 0.7440, taking in consideration column percentage values, and way **above Bonferroni adjusted p-value of 0.00083**, result shows that there is no relationship between variables and we **cannot safely reject H0 hypothesis**

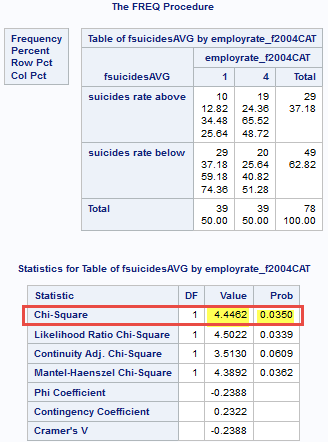
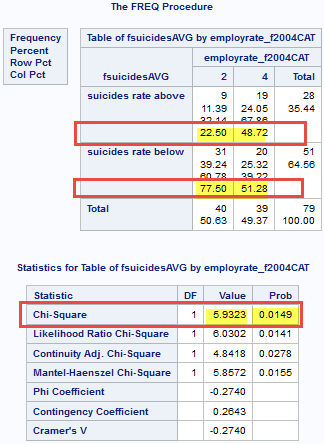
Pair 1 and 3 - There is significant difference (5.5097), p-value = 0.0189, there is a difference in column percentage values, but p-value = 0.0189 is little bit **Bonferroni adjusted p-value of 0.00083 is little**, we **cannot safely reject H0 hypothesis or accept Ha hypothesis**

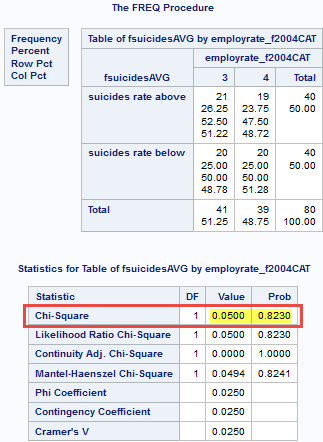
Pair 1 and 4 - There is significant difference (4.462), p-value = 0.0350, there is a difference in column percentage values, but p-value = 0.0189 is higher **Bonferroni adjusted p-value of 0.00083 is little**, we **cannot safely reject H0 hypothesis or accept Ha hypothesis**

Pair 2 and 4 - There is significant difference (5.9323), p-value = 0.0149, there is a difference in column percentage values, but p-value = 0.0149 is little bit **Bonferroni adjusted p-value of 0.00083 is little**, we **cannot safely reject H0 hypothesis or accept Ha hypothesis**

Pair 3 and 4 - No significant difference (0.0500), p-value = 0.8330, taking in consideration column percentage values, and way **above Bonferroni adjusted p-value of 0.00083**, result shows that there is no relationship between variables and we **cannot safely reject H0 hypothesis**



**Conclusion:**

**I think that studied data does not fully fit Chi Square Test of independence and as a conclusion of our test, taking in consideration artificial nature of our exploratory (categorized quantitative variable) and male portion of test ( where test failed and Alternate hypothesis was rejected), we have to accept Null Hypothesis (H0) and reject Alternate Hypothesis (Ha).**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* End of week 2 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>