Seung Yun Choi

72929094

SOCI 328 Social Statistics

Professor Dr. Gerry Veenstra

November 15, 2017

Assignment 1

Part I  
We expect that the two variables be related since the different classes in the National Occupational Classification correspond to different levels of skill. For example, we expect a higher level of income for Management Occupations (01) category compared with the Sales and Services Occupations (07). We expect higher skill occupations to have higher income level. We suspect that the two variables are strongly related. We also suspect that the skill level (classification) results in higher income and therefore it would be the cause or influence. It would therefore make sense to designate the NOC as the independent variable and Income as dependent.

Part II

The variable **NOC2006\_C10** measures the occupational class of the respondents. It is a nominal variable with 10 categories for each occupational class. The variable **incmhsd** measures which income interval the respondent lies in. Therefore, it is an ordinal variable with categories defined as different income ranges. We remove invalid data from our dataset by running the following command:

*drop if (incmhsd>= 90) | (NOCS2006\_C10 >= 90)*

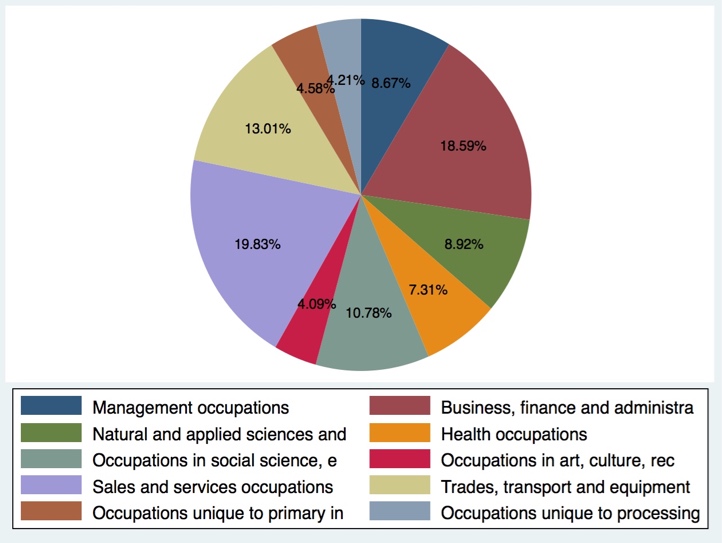
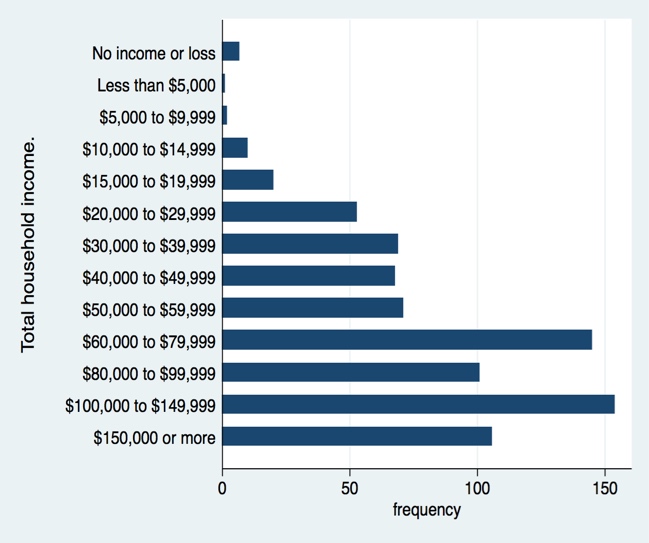
To see how the distribution of different categories looks like, we use a bar chart for the incmhsd variable. This is suitable since bar charts also show order. For NOCS2006\_C10, we use a pie chart since there is no order:  


Figure 1. Bar and Pie charts for Income and NOC respectively

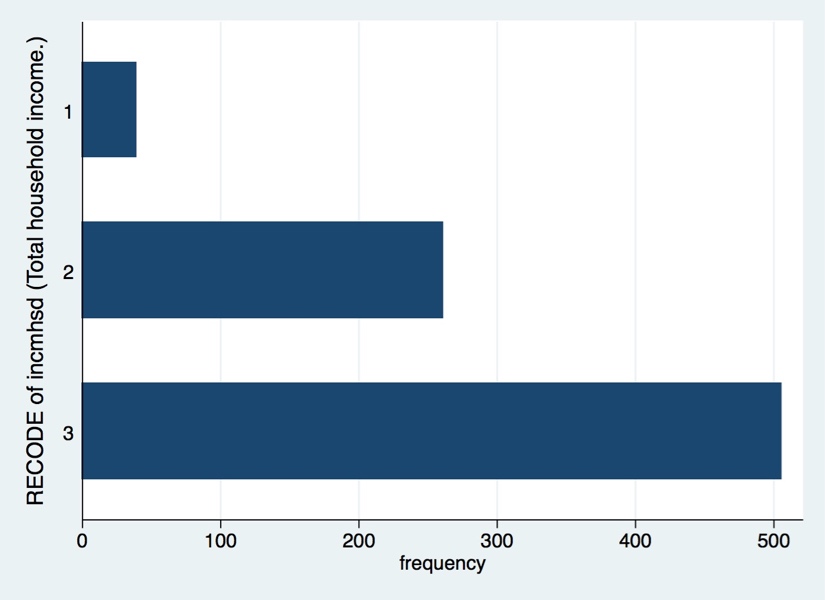
Part III  
We recode the income variable to make any possible relationship more obvious. For this purpose, we combine the categories into 3 bigger categories of low (1 to 5), medium (6 to 9) and high (10 to 13) income:  
*recode incmhsd (1/5 = 1) (6/9 = 2) (10/13 = 3), gen(incmhsd\_basic)*  


Figure 2. Bar chart for recoded income

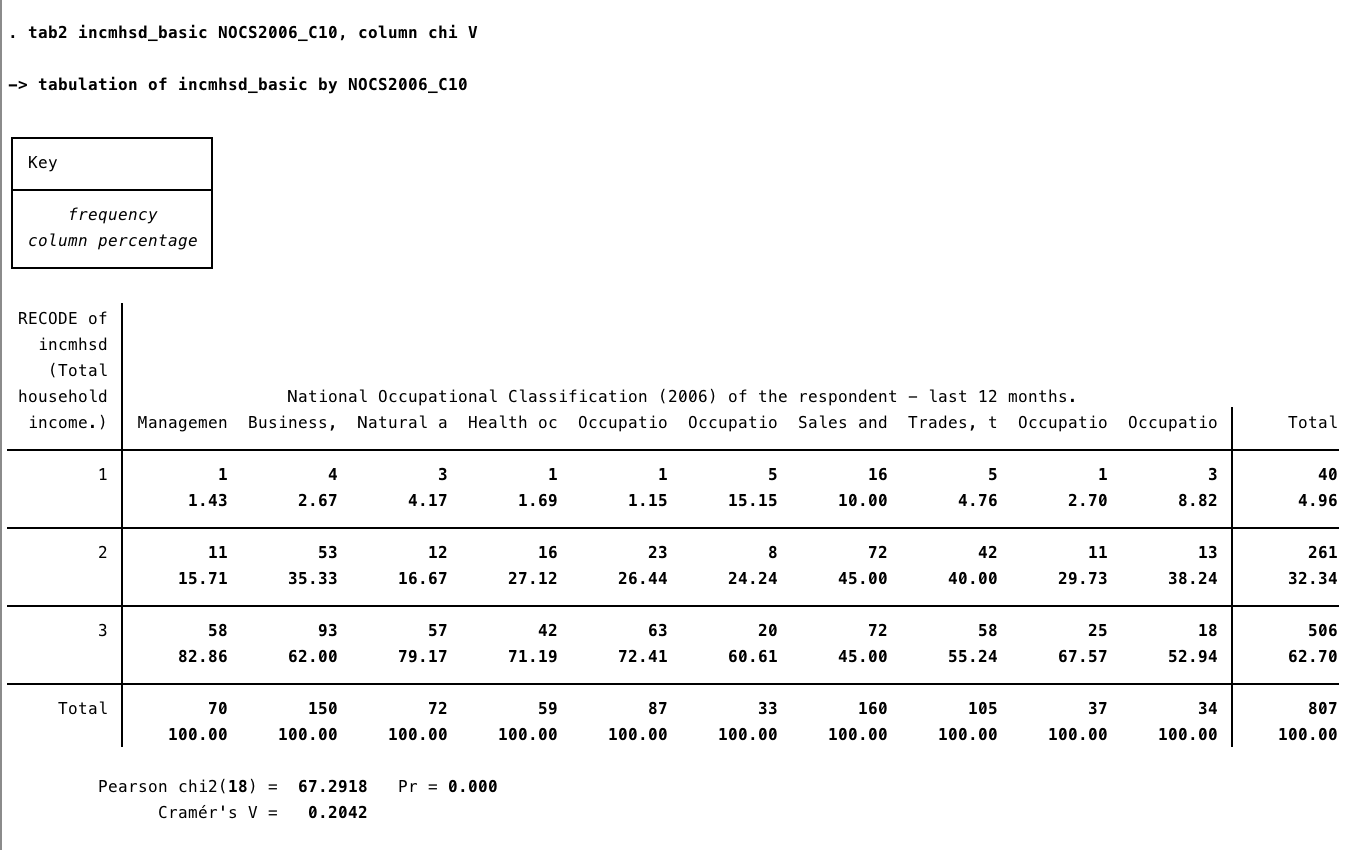
We do not recode the classification variable since there is no obvious way to combine different classes. Since we are analyzing the association between two categorical variables, we have to use either the Cramer’s V measure or the Kendall’s Tau-b (for two ordinal variables). Since NOC is not an ordinal variable we will use Cramer’s V to measure the level of association:  
*tab2 incmhsd\_basic NOCS2006\_C10, column chi V*  


Figure 3. Statistical Analysis of the Association Between NOC and Income

We see that for some classes the marginal frequency percentages are not the same as the total marginal frequency percentages. This means that the percentages are not what we expect them to be. For example for Management class we see that the percentages are greatly skewed towards higher income brackets whereas for sales the opposite is true and the percentages are skewed towards lower income brackets. To quantitatively confirm the suspected relationship between classification and income we use the Cramer’s V statistic. We find the Cramer’s V to be 0.2042. From the lecture note, we can see that this value corresponds to a modest strength of association.  
Part IV  
Overall, the Cramer’s V measure shows a modest (neither weak nor strong) relationship and therefore confirms our hypothesis that these two variables are related. However, the strength of the relationship is not strong, contrary to what we hypothesized. Throughout the investigation of the variables NOC2006\_C10 (National Occupational Classification of the respondent) and incmhsd (household income) I hypothesized that their relationship would be strong because in Korea, the higher the occupation, the income is higher.