Study of a possible association between level of education and job satisfaction

June 2015

Introduction:

Some people believe that a higher academic degree leads to a more satisfying work. A lively discussion is going on concerning this issue. This project puts this belief to a test by examining data from the General Social Survey (GSS).

We try to explore a possible correlation between the respondent's highest educational degree and the level of satisfaction with the job he/she obtained.

Data:

The data represents a random sample of the US population. The cases are randomly selected individuals from the US population.

For the purposes of our study we will examine the following variables: **'satjob'** (On the whole, how satisfied are you with the work you do?) and **'degree'** (degree obtained by the respondents). Both variables are categorical and ordinal. The levels of 'satjob' can be ordered in the following decreasing order (Very Satisfied, Mod. satisfied, A Little Dissatisfied, Very Dissatisfied). The levels of 'degree' can be ordered in the increasing order as follows: Less Than High School, High School, Junior College, Bachelor, Graduate).

It is an observational study. Respondents were randomly selected from the US population. For this reason, the findings **can be generalized to the entire population**. Since this is an observational study and not an experiment (with random assignment to groups), **only a correlation** (association) can be established between the variables, **not a causal link**.

Exploratory data analysis:

I will use the statistical software R to conduct this analysis. However, I will perform a step by step calculation of the statistical test and compare it with the corresponding function provided by R.

Some of the rows in the data set contain missing values (NA - not available).

Number of rows without missing values (NA - not available): **40672** (rows lost: **16389**). I decided to delete the rows of data containing NA values (roughly 25% of data will be lost). I will assume that NA values are randomly distributed and that no bias is introduced. It must be noted that it would be difficult to impute such a large number of missing values without introducing a bias. With this understanding, I decided to procede with a new data set which represents roughly 75% of the original data set and does not contain NA values.

- Contingency table - Total frequency per income

```
##                    Lt High School High School Junior College Bachelor
## Very Satisfied               3349       10005           1201     3106
## Mod. Satisfied               2793        8497            883     2386
## A Little Dissat               821        2281            214      546
## Very Dissatisfied             378         961             69      208
## coltotals                    7341       21744           2367     6246
##                     Graduate rowtotals
## Very Satisfied          1753     19414
## Mod. Satisfied           954     15513
## A Little Dissat          195      4057
## Very Dissatisfied         72      1688
## coltotals               2974     40672
```

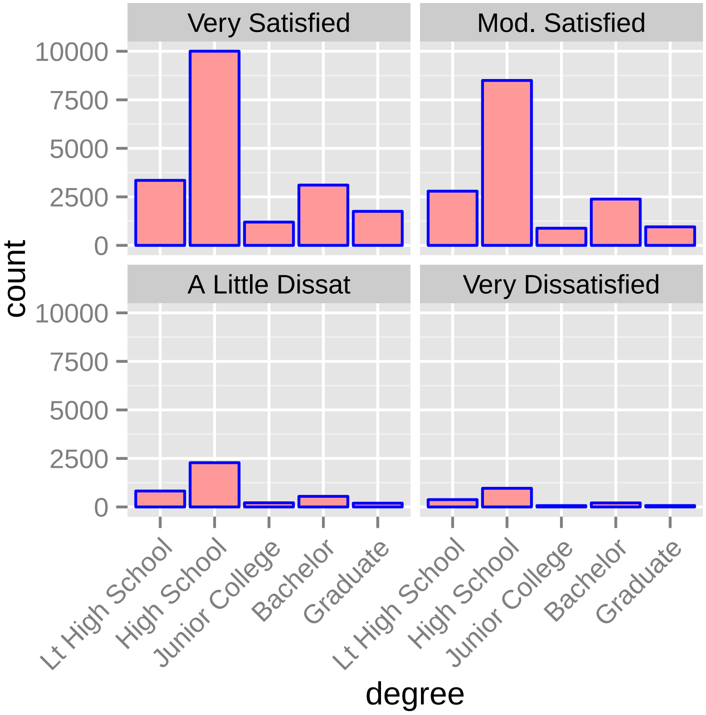
The numbers of respondents belonging to each income group is not equal, hence it makes more sense to consider relative frequencies and not total counts.

- Contingency table - Relative frequency per income

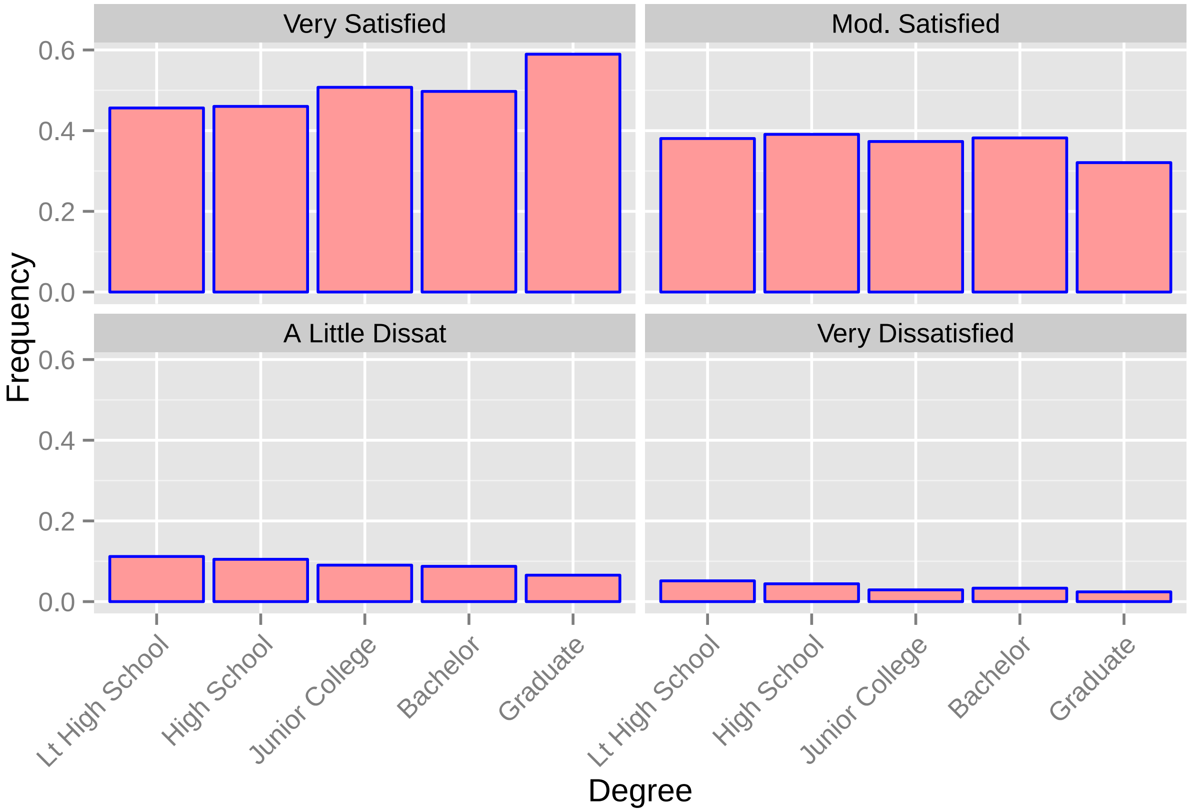
```
##                      y
## x                     Lt High School High School Junior College   Bachelor
##   Very Satisfied          0.45620488  0.46012693     0.50739332 0.49727826
##   Mod. Satisfied          0.38046588  0.39077447     0.37304605 0.38200448
##   A Little Dissat         0.11183762  0.10490250     0.09040980 0.08741595
##   Very Dissatisfied       0.05149162  0.04419610     0.02915082 0.03330131
##                      y
## x                       Graduate
##   Very Satisfied      0.58944183
##   Mod. Satisfied      0.32078009
##   A Little Dissat     0.06556826
##   Very Dissatisfied 0.02420982
```

- Joint count barplot

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```



- Joint relative frequency barplots per level of satisfaction



From the above plots, first of all we conclude that most of the respondents are either very satisfied or moderately satisfied with their work (roughly, 80%). The level of satisfaction grows slightly as the level of degree increases (for example, in the "Very Satisfied" group).

Inference:

Let us see if our observation based on visual examination of data is statistically significant.

We have two categorical variables each of which has more than two levels. In this situation we use a **chi-squared independence test**.

- Null hypothesis - H0 = job satisfaction and degree are independent
- Alternative hypothesis - Ha = job satisfaction and degree are dependent
- Conditions for chi-squared independence test

1. **Independence**. Sampled observations are independent since the sample is random, the sample represents less than 10% of the population and each case only contributes to one cell in the table.
2. **Sample size**. Each particular scenario (cell) has at least 5 expected cases.

First we will perform a chi-squared independence test manually and then apply the function integrated in R. The manual calculation is done purely to review the procedure. - Compute the expected values:

```
##
##                      Lt High School High School Junior College Bachelor
##   Very Satisfied              3504       10379           1130     2981
##   Mod. Satisfied              2800        8294            903     2382
##   A Little Dissat              732        2169            236      623
##   Very Dissatisfied            305         902             98      259
##
##                      Graduate
##   Very Satisfied         1420
##   Mod. Satisfied         1134
##   A Little Dissat         297
##   Very Dissatisfied       123
```

- Compute chi-squared:

```
chi_sq <- sum((tbl-expected)^2/expected); chi_sq #tbl contains observed values
```

```
## [1] 266.438
```

- Degrees of freedom: `(4-1)*(5-1)` = 12, p-value:

```
pchisq(chi_sq, 12, lower.tail = FALSE)
```

```
## [1] 5.056602e-50
```

P-value is close to zero, we reject the null hypothesis

Now we will compare this result with the result of the R's `chisq.test` function.

```
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 267.1376, df = 12, p-value < 2.2e-16
```

The R's chisq.test function produces approximately the same results: p-value is far less than 0.05 at 95% significance level. **We reject the null hypothesis**.

Conclusion:

We **reject the null hypothesis** that the level of job satisfaction and academic degree are independent and **accept the alternative hypothesis** - these variables are dependent. There might be a **positive correlation** between the level of degree and job satisfaction (i.e. respondents with a higher degree seem to be more satisfied with their work). Though, **we cannot conlcude that there is a casual link** since this is an observational study and not an experiment. It might be interesting to conduct a **further study** on other possible factors contributing to the satisfaction at work (for example, level of happiness, optimism, pessimism, etc).

References

- Data citation Smith, Tom W., Michael Hout, and Peter V. Marsden. General Social Survey, 1972-2012 [Cumulative File]. ICPSR34802-v1. Storrs, CT: Roper Center for Public Opinion Research, University of Connecticut /Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 2013-09-11. doi:10.3886/ICPSR34802.v1
- Persistent URL: http://doi.org/10.3886/ICPSR34802.v1
- A modified version of the survey was used for the purposes of this study taken fromthe Data Analysis and Statistical Inference class. Link to download data.