Data Analysis Assignment 1

**Part 1: Introduction**

**What is your research question? Why do you care? Why should others care?**

The analysis investigates the relationship between the educational attainment and the one's perception on his/her financial status during last few years. Most people grew up hearing the lectures from their parents and teachers how important to go to school. The prospect of the US economy continues to look grim while the price tag on the higher education is on the rise. It is a high time to ask the questions: Is there a return on investment in monetary sense? How much is enough to get one equipped for what is called knowledge economy? Is more the better?

**Part 2: Data (10 points)**

**Write about the data from your proposal in text (not bullet-point) form. Address these points:**

The General Social Survey (GSS) data used in this analysis was obtained from the web site of the National Opinion Research Center. [1] The original dataset had been modified by facilitators of Statistical Analysis and Inference class on Coursera. [2]

**Data collection: Describe how the data were collected.**

The General Social Survey (GSS) was conducted by the National Opinion Research Center. The survey collects demographic data such as the respondent's age, gender, and ethnicity as well as religious affiliations along with his/her opinion on social and political issues in the contemporary America.

The data set that will be used in the analysis is a cumulative data from 1972 to 2012. The total number of the observation is 57061 with 144 variables. The interviews were conducted by one of these methods: in-person, telephone, computer-aided.

**Cases: What are the cases? (Remember: case = units of observation or units of experiment)**

**Variables: What are the two variables you will be studying? State the type of each variable.**

In this analysis, a case is defined as an observed response on one’s highest degree completed and his/her perception about change in his/her financial situation in the last few years. The variables of the interest are only two out of 144 collected by the GSS, namely degree and finalter. The independent variable is degree. This indicates the highest degree completed by the respondent. This is an ordinal categorical variable with five levels: Less than High School (coded as Lt High School), High School, Junior College, Bachelor, and Graduate. The dependent variable is finalter. This is the respondent’s opinion on of whether his/her financial situation has gotten better, stayed the same, or worth in the last few years.  This is an ordinal categorical variable with three levels: Worse, Stayed Same, and Better.

**Study: What is the type of study? Is it an observational study or an experiment? Explain how you've arrived at your conclusion using information on the sampling and/or experimental design.**

This is an observational study. The data was collected by conducting a survey and were not results from a designed experiment where subjects are randomly assigned to a control group and a treatment group.

**Scope of inference - generalizability: Identify the population of interest, and whether the findings from this analysis can be generalized to that population, or, if not, a subsection of that population. Explain why or why not. Also discuss any potential sources of bias that might prevent generalizability.**

**Scope of inference - causality: Can these data be used to establish causal links between the variables of interest? Explain why or why not.**

The population of the concern is males and females whose age range from 18 and 89 who are living in the United States. The data was collected from the people with various educational levels, financial status, ethnic backgrounds, marital status, and regions of the United States. The intention is to be representative of the current US population. Therefore the analysis can be generalize to adults who are living in the United States.

The survey data was collected from the people who agreed to participated in the survey and were reachable by the means of a telephone line, a computer network, or in-person. This could be a source of a bias.

Since this is an observational study, this data cannot be used to establish causality between the educational attainment and the financial situation. Instead, the analysis seeks correlations between two variables.

**Part 3: Exploratory data analysis (10 points)**

**Calculate and discuss relevant descriptive statistics, including summary statistics and visualizations of the data. Also address what the exploratory data analysis suggests about your research question.**

**Part 4: Inference (20 points)**

**State hypotheses**

**The null hypothesis is that there is no difference in perceptions on their financial situations in recent years between groups separated by their educational attainment.**

**Check conditions**

**State the method(s) to be used and why and how**

Perform inference

Interpret results

If applicable, state whether results from various methods agree

It is your responsibility to figure out the appropriate methodology. What techniques you use to conduct inference will depend on the type of data you're using, and your sample size. All of you should conduct at least a hypothesis test, and report the associated p-value and the conclusion. Those of you comparing two means, two medians, or two proportions should also calculate a confidence interval for the parameter of interest. Those of you working with categorical variables with more than two levels will need to use methods like ANOVA and chi-square testing for which there is no associated confidence interval, and that's ok. If your data fails some conditions and you can't use a theoretical method, then you should use an appropriate simulation based method.

If you can use both theoretical and simulation based methods, then choose one and stick with it. You don't have to do both. However if you can't use both, then you need to decide which is appropriate.

If you can do both a hypothesis test and a confidence interval, do both, and comment on agreement of the results from the two methods. However if your variables do not lend themselves to a confidence interval, that's ok.

It's essential to make sure the method you're using is appropriate for the dataset and the research question you're working with.

- Two categorical variables (either one or both with more than 2 levels): hypothesis test only

compare proportions across several groups

no defined parameter of interest, Chi-square test only (theoretical if expected sample size condition met, simulation if not)

Part 5: Conclusion (5 points)

Write a brief summary of your findings without repeating your statements from earlier. Include a discussion of what you have learned about your research question and the data you collected, and include ideas for possible future research.

# References (1 point)

Include a citation for your data, and if your data set is online, provide a link to the source. Also list other references (if any). Please upload your file containing all of your responses here.

The data set had been modified for Data Analysis and Statistical Inference on Coursera.

<https://d396qusza40orc.cloudfront.net/statistics%2Fproject%2Fgss1.html#finalter_toc>

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.v1http://doi.org/10.3886/ICPSR34802.v1

R-Tutorial: Chi-Squared Test of Independence, accessed on March, 19, 2014

http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence