SOCI 420
Advanced Methods of Social Research
Spring 2018

ASSIGNMENT 2
Due by April 2, 2018 (Monday) at 8:00am
Percent of final grade: 25%

Course information

Course website: http://www.ernestoamaral.com/soci420-18spring.html

Meeting times: Lecture, Monday and Wednesday, 1:50–2:40pm
Lab, Friday, 1:50–2:40pm

Instructor information

Ernesto F. L. Amaral, Assistant Professor, Department of Sociology
Office location: Academic Building 415
Phone: (979)845–9706
Email: amaral@tamu.edu

Submission

Assignment should be submitted through Turnitin within eCampus. Turnitin is an online database system designed to help instructors **detect plagiarism**, track citations, facilitate peer reviews, and provide paperless grading markup in written assignments. Students can develop this assignment **individually or in pairs**. Students doing the assignment in pairs must individually submit the same version of the document to their own eCampus account.

General information

This assignment is based on the ideas for research projects included in Appendix E of the course textbook (Healey, Joseph F. 2015. Statistics: A Tool for Social Research. Stamford: Cengage Learning. 10th edition).

Your grade for this assignment will be determined by the use of **several statistical tools** with a focus on the **quality of your analysis**, and the elaboration of **coherent interpretations**. The accuracy of the formatting of your tables and graphs will also be evaluated. The Stata codes used for this assignment (do-file) should be included at the end of the document as an appendix.

When interpreting tables and graphs, write in plain English, as if you were reporting results in a newspaper. You should have an introductory paragraph explaining the main purpose of your analysis, another paragraph briefly explaining your data and methods, a few paragraphs with the analyses of the tables and graphs, and a concluding paragraph with final considerations. This assignment should be seen as a document that tells a **coherent story about a subject**. Thus, it is important to think wisely about selecting variables for your analysis. You should also make clear that you are estimating characteristics of the adult population of the entire United States.

The document should be on US Letter paper size, one-inch margins, Arial font, size 11, 1.5 line spacing, and a **maximum of 2,000 words** (excluding tables, figures, and Stata do-file). Font size within tables can have a smaller size, such as size 9 for numbers and text within tables and size 8 for table footnotes.

Students should take advantage of **regular classes** and **office hours** to clarify any questions with the professor and/or the teaching assistant. The days and time of office hours are listed in the syllabus and course website.

Exercise

Select **variables that are available in the 2004, 2010, and 2016 General Social Survey (GSS)** to estimate characteristics of the U.S. adult population. You will use Stata to estimate and analyze sample statistics and confidence intervals. These variables can be the same as those used in the classroom or in the previous assignment. See the GSS codebook or the GSS Data Explorer website (https://gssdataexplorer.norc.org) for a list of variables available in GSS.

A. Estimating means (chapter 7)
Include at least 2 tables (1 for each variable).

1. There are relatively few interval-ratio variables in GSS, and for this part of the project you may use ordinal variables that have at least three categories or scores. **Choose two variables** that fit this description.

2. **Estimate means, standard errors, sample size, and construct 95% confidence intervals for the mean of each of your variables for each year.** When computing the standard error, consider the effect of clustering and stratification, as well as the effect of sampling weights (i.e. complex survey design), using the "svyset" command. The command "svy: mean" provides an estimate of the population mean and an estimate of its standard error. For example, svy: mean conrinc.

3. For each variable, **report and explain the results**, the confidence interval, the confidence level, and the sample size.

4. Include in your analysis a brief explanation of the role of these concepts and terms in the estimation: sample, population, statistic, parameter, equal probability of selection method (EPSEM), representative, and confidence level. This can be part of the overall data and methods paragraph.

B. Estimating proportions (chapter 7)
Include at least 2 tables (1 for each variable).

1. **Choose two nominal- or ordinal-level variables**. These variables should be different than the variables used for estimating means.

2. **Estimate proportions, standard errors, sample size, and construct 95% confidence intervals for the proportions of each category of your variables for each year.** When computing the standard error, consider the effect of clustering and stratification, as well as the effect of sampling weights (i.e. complex survey design), using the "svyset" command. The command "svy: prop" provides an estimate of the population proportion and an estimate of its standard error. For example, svy: prop letin1.

3. For each variable, **report and explain the results**, the confidence interval, the confidence level, and the sample size.

4. Include in your analysis a brief explanation of the role of these concepts and terms in the estimation: sample, population, statistic, parameter, equal probability of selection method (EPSEM), representative, and confidence level. This can be part of the overall data and methods paragraph.

C. Two-sample *t*-test (chapter 9)
Include at least 2 tables (1 for each dependent variable).

1. **Use the two variables from part A of this assignment (interval-ratio or ordinal-level variables that have three or more scores)**. They will be the dependent variables.

2. **Choose independent variables** that might logically be a cause of your dependent variables. These independent variables should have only two categories. You can still use independent variables with more than two categories by collapsing their scores with the "generate" and "replace" commands in Stata. Independent variables can be at any level of measurement. You may use the same independent variable for both dependent variables.

3. **Estimate two-sample *t*-tests with equal variances for each of your dependent variables for each year**. You do not have to consider the complex survey design or the sampling weight for these estimations. Use the command "ttest" with a specific dependent variable and the option "by" to indicate the independent variable. For example, ttest conrinc, by(sex).

4. **Report and explain the results** of the tests. At a minimum, your analysis should clearly identify the independent and dependent variables, the sample statistics, the value of the test statistic, the results of the test, and the confidence level.

D. Analysis of variance (chapter 10)
Include at least 2 tables (1 for each dependent variable).

1. **Use the two variables from part A of this assignment (interval-ratio or ordinal-level variables that have three or more scores)**. They will be the dependent variables.

2. **Choose independent variables** that might logically be a cause of your dependent variables and that have between three and five categories. You may use the same independent variable for both dependent variables.

3. **Estimate one-way analysis of variance for each of your dependent variables for each year**. You should consider the effect of sampling weights, using the "aweight" command. The "oneway" command reports one-way analysis-of-variance (ANOVA) models. For example, oneway conrinc raceeth [aweight=wtssall].

4. **Report and explain the results** of the tests. At a minimum, your analysis should clearly identify the independent and dependent variables, the sample statistics, the value of the test statistic, the results of the test, the degrees of freedom, and the confidence level.

E. Chi square (chapter 11)
Include at least 2 tables (1 for each dependent variable).

1. **Use the two variables from part B of this assignment (nominal- or ordinal-level variables).** They will be the dependent variables.

2. **Choose independent variables** that might logically be a cause of your dependent variables. Independent variables can be at any level of measurement as long as they have five or fewer (preferably two or three) categories. Output will be easier to analyze if you use variables with fewer categories. You may use the same independent variable for both dependent variables.

3. **Estimate the chi square tests for each of your dependent variables for each year**. You do not have to consider the complex survey design or the sampling weight for these estimations. Use the command "tab" with the option "chi" to indicate the Pearson's chi-square test. It is almost always desirable to report the column percentages as well. The dependent variable should have the categories listed on the rows and the independent variable on the columns. For example, tab letin1 sex, chi col.

4. **Report and explain the results** of the tests. At a minimum, your analysis should clearly identify the independent and dependent variables, the sample statistics, the value of the test statistic, the results of the test, the degrees of freedom, and the confidence level.

Other considerations

1) The General Social Survey (GSS) microdata is available on the course website, as well as from the NORC website (http://gss.norc.org).

2) You should avoid including tables and figures in your assignment that do not enhance (or are not related to) your analyses. You should analyze all tables and figures included in your assignment.

3) If reporting missing cases, do not include them in the total of the tables. Preferably, report missing cases in a row below the total. There are three different types of missing values in GSS: (1) ".i" Inapplicable (IAP). Respondents who are not asked to answer a specific question are assigned to IAP; (2) ".d" Don't know (DK); and (3) ".n" No answer (NA). In most cases, it would be better to differentiate between these types of missing cases by including one row for each of them at the bottom of the table.

4) You should utilize appropriate formatting for your tables and graphs. This file has several examples of how to correctly format tables and graphs (http://www.ernestoamaral.com/docs/soci420-18spring/Examples_tab_fig.pdf). There are also some papers on my website (http://www.ernestoamaral.com/papers.html) that can help you with the correct format for tables and graphs.

5) You can copy tables from Stata to Word (highlight table, right click, and select "Copy table as HTML") in order to format them. You can also copy tables from Stata to Excel (highlight table, right click, and select "Copy table" or "Copy table as HTML"), format them, and copy to Word. I suggest copying tables from Excel to Word in an editable format, instead of pasting as figures.

6) If it is complicated to generate all graphs in Stata, you can copy tables from Stata to Excel to generate graphs. There are several examples of how to generate graphs in Excel on the course website (http://www.ernestoamaral.com/docs/soci420-18spring/Excel_charts.zip).

7) Several variables and a large amount of information can be organized in a single table in a clear and objective manner. For example, look at Table 1 (frequency distributions), Table 2 (percentage of one variable by categories of other variables), and Tables 3, 4, and 5 (statistical regressions) in the paper about characterization of fertility levels in Brazil (http://www.ernestoamaral.com/docs/papers/ES2015.pdf). You can also see Table 1 (frequency distributions), Table 2 (rates of one variable by categories of other variables), and Tables 3 and 4 (statistical regressions) in the paper about rising cesarean section rates in Brazil (http://www.ernestoamaral.com/docs/papers/BIRTH2014.pdf). There are also other papers on my website that provide additional examples.

8) You can illustrate descriptive statistics using graphs, instead of tables. For example, look at Figures 2 and 3 in the paper about the growth of Protestantism in Brazil (http://www.ernestoamaral.com/docs/papers/SF2014.pdf).

9) You should perform the data analysis with the statistical software Stata. The codes generated in this software (do-file) must be included at the end of the assignment as an appendix.

10) Descriptive tables should be generated taking into account the GSS complex survey design.

11) The command "summarize" provides descriptive statistics for the sample. It does not provide inferential statistics for the population. You would have to indicate the complex survey design with the command "svyset" to get the standard error of the estimate of the population mean. The command "svy: mean" (followed by "estat sd") provides an estimate of the population mean and an estimate of its standard error. When computing the standard error, consider the effect of clustering and stratification, as well as the effect of sampling weights (i.e. complex survey design). However, the clustering and stratification do not affect the point estimate of the mean. Thus, if you are interested only in the point estimate (e.g. mean, median), you can use "summarize" with "aweights" since it gives the same weighted mean as "svy: mean." For quantiles, "summarize" with "aweights," as well as "pctile" with "aweights" or "pweights," all give the same answers. If you use "summarize" with "aweight" (not considering the complex survey design), this strategy assumes a simple random sample, in which: (1) an estimate of the population mean is the sample mean; and (2) an estimate of the population standard deviation is the sample standard deviation. By not indicating complex survey design variables, Stata will assume a simple random sample and underestimate standard errors. You should explain this limitation in interpreting the weighted standard deviation, when not indicating the complex survey design (https://www.stata.com/support/faqs/statistics/weights-and-summary-statistics).

Considerations for regression models (not applicable for all assignments)

1) Regression models should be estimated taking into account the GSS complex survey design.

2) It is possible to illustrate several regression models in a single table. Remember to include estimated coefficients, robust standard errors (between parentheses), and statistical significance (with asterisks). You can also illustrate the standardized regression coefficients in separated columns. Use the command “outreg2” to transfer the regression models from Stata to Word. If the “outreg2” command is not available in your Stata software, you can install it from this file (http://www.ernestoamaral.com/docs/soci420-17fall/Modules.zip).