SOCI 420–903
Advanced Methods of Social Research
Spring 2023

ASSIGNMENT 2
Due by April 11, 2023 (Tuesday) at 11:59pm
Percent of final grade: 20%

Instructor information

Ernesto F. L. Amaral, Associate Professor, Department of Sociology
Office: Liberal Arts Social Sciences Building (LASB) 320
Phone: (979)845–9706
Email: amaral@tamu.edu
Course website: http://www.ernestoamaral.com/soci420-23spring.html

Submission

Assignment should be submitted through Turnitin within Canvas. 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 should develop this assignment **individually**.

Purpose

The purposes of this assignment are for students to investigate microdata from the General Social Survey with the statistical package Stata, select and transform variables, and examine these variables with statistical methods discussed in the course.

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 2018 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. You should generate new variables to recode original ones if appropriate, as we performed in class.

A. Estimating means (chapter 7)
Include at least 1 table.

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 one variable** that fit this description, which will also be used as a dependent variable later in the assignment.

2. **Estimate means, standard errors, sample size, and construct 95% confidence intervals for the mean of your variable 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 your 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 1 table.

1. **Choose one nominal- or ordinal-level variables**. This variable should be different than the variable used in Part A. This variable will also be used as a dependent variable later in the assignment.

2. **Estimate proportions, standard errors, sample size, and construct 95% confidence intervals for the proportions of each category of your variable 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 your 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 1 table.

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

2. **Choose one independent variable** that might logically be a cause of your dependent variable. This independent variable should have only two categories. You can still use an independent variable with more than two categories by collapsing their scores with the "generate" and "replace" commands in Stata.

3. **Estimate two-sample *t*-tests with equal variances between your dependent and independent 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 your 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. Two-sample test of proportions (chapter 9)
Include at least 1 table.

1. **Choose one dependent variable with only two categories (dummy variable)**. You can use the dependent variable from part B of this assignment. You can transform a variable with more than two categories into two categories by collapsing their scores with the "generate" and "replace" commands in Stata.

2. **Choose one independent variable** that might logically be a cause of your dependent variable. This independent variable should have only two categories. You can transform a variable with more than two categories into two categories by collapsing their scores with the "generate" and "replace" commands in Stata. This variable should be different than the independent variable used in Part C.

3. **Estimate two-sample test of proportions between your dependent and independent variables for each year**. You do not have to consider the complex survey design or the sampling weight for these estimations. Use the command "prtest" with your dependent variable and the option "by" to indicate the independent variable. For example, prtest proimmig, by(democrat).

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.

E. Analysis of variance (chapter 10)
Include at least 1 table.

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

2. **Choose one independent variable** that might logically be a cause of your dependent variable and that have between three and five categories. This variable should be different than the independent variables used in Parts C and D.

3. **Estimate one-way analysis of variance between your dependent and independent 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.

F. Chi square (chapter 11)
Include at least 1 table.

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

2. **Choose one independent variable** that might logically be a cause of your dependent variable. 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. This variable should be different than the independent variables used in Parts C, D, and E.

3. **Estimate the chi square tests between your dependent and independent 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) ".n" No answer (NA); (3) ".d" Do not know/Cannot choose (DK); and (4) ".s" Skipped on web. 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/soci420-23spring/Examples_tab_fig.pdf). There are also some papers (http://www.ernestoamaral.com/papers.html) and drafts (http://www.ernestoamaral.com/drafts.html) on my website, which 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/soci420-23spring/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://doi.org/10.17605/OSF.IO/8FRJ4). 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://doi.org/10.17605/OSF.IO/QFHXE). 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://doi.org/10.17605/OSF.IO/C5P2A).

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) Up to GSS 2018, you should use the weight variable "wtssall," which applies an adult weight to years before 2004. For GSS 2021, you should use the weight variable "wtssnrps," since the variable "wtssall" is not available.

11) You can simply use the survey weight if you are estimating only frequency distributions and measures of central tendency (e.g., mean, median). However, you need to utilize the complex survey design ("svyset" and "svy") if you are estimating measures of dispersion (e.g., standard deviation, standard error), margins of error, confidence intervals, and statistical significance (e.g., *t*-test, *p*-value).

12) 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 deviation. 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 considering 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/soci420-23spring/Modules.zip).