ASSIGNMENT 1
Due by September 19, 2023 (Tuesday) at 11:59pm
Percent of final grade: 10%

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/soci600-23fall.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 American Community Survey with the statistical package Stata, select and transform variables, and examine these variables with the initial 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 Texas population with the American Community Survey (ACS).

The document should be on US Letter paper size, one-inch margins, Arial font, size 11, 1.5 line spacing, and a **maximum of 1,000 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 **one nominal-level variable, one ordinal-level variable, and one interval-ratio-level variable from the 2021 American Community Survey (ACS)**. Write a sentence or two to explain each variable, being careful to include a description of the overall shape of the distribution (**chapter 2**), the central tendency (**chapter 3**), and the dispersion (**chapter 4**). For nominal- and ordinal-level variables, be sure to explain any arbitrary numerical codes. For example, in the variable for employment status (“empstat”) in ACS, a 1 is coded as “employed,” a 2 indicates “unemployed,” and a 3 is “not in labor force.” **This exercise will generate at least 2 figures and 2 tables in your assignment.**

Generate **one figure** for the nominal-level variable informing percentage distributions. Report total number of cases and missing cases in the footnote of the figure (see further information below).

Generate **one table** for the ordinal-level variable, including percentages, cumulative percentages, and total number of cases. Report missing cases in a row below the total (see further information below).

Generate **one table** for the interval-ratio-level variable including measures of central tendency (median, mean), measures of dispersion (lowest score, highest score, 25th percentile, 75th percentile, range, interquartile range, standard deviation), and number of cases. Check for skew both by comparing the mean and median. Report missing cases in a row below the total (see further information below).

Generate **one boxplot** for the same interval-ratio-level variable selected above. Report total number of cases and missing cases in the footnote of the figure (see further information below).

For this assignment, you are performing **univariate descriptive statistics**, thus you do not have to generate cross tabulations among variables (bivariate or multivariate descriptive statistics). It would be good to select variables that relate to each other in some way, so you can write a “coherent story about a subject.” Basically, you will write a report that explains the different distributions of your variables. At the end of the report (paragraph with final considerations), you could mention that a next step in the analysis would be to provide bivariate and multivariate descriptive statistics to explore correlations among the selected variables.

See the ACS codebook in the IPUMS website (https://usa.ipums.org/usa-action/variables/group) for a list of available variables. You should generate new variables to recode original ones if appropriate, as we performed in class.

Other considerations

1) The American Community Survey (ACS) microdata is available on the course website, as well as from the IPUMS website (https://usa.ipums.org/usa-action/samples).

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 ACS. For instance, there are respondents who are not asked to answer a specific question, so the variable for those cases are assigned as not applicable (N/A) or observations not in universe (NIU). In other cases, respondents might not have provided information, so the variable has a missing case. 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/soci600-23fall/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/soci600-23fall/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) You should use the person weight (“perwt”) on your analysis.

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 ACS 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/soci600-23fall/Modules.zip).