S371: Statistics for Sociology

**Final Project**

Due Thursday, May 7, 12:30pm via Canvas

100 points possible

This project provides an opportunity for you to review and apply the skills you learned throughout the course. You will analyze data from the General Social Survey (GSS) and summarize and interpret your results in a report.

Data

This wave of the GSS was collected in 2018 from a probability sample of 2,348 of American adults aged 18 and above. The data, “project.RData,” and accompanying codebook are available on Canvas. Additional information on the questionnaires and data can be found at gss.norc.org.

This dataset includes 56 variables. The data has been minimally preprocessed to simplify analysis. Most variables are not available for all 2,348 individuals though. There are some cases missing for most variables. Missing values, refused questions, and inapplicable questions have all been marked with NA.

Analysis

You will choose and analyze data for four variables from the dataset. You must choose one of each of the following variable types:

1. quantitative dependent (response) variable
2. categorical dependent (response) variable
3. quantitative independent (explanatory) variable
4. categorical independent (explanatory) variable <

You will examine univariate distributions and bivariate relationships. The bivariate relationships you examine will be between the dependent and independent variables. So, you will have four variable pairs:

|  |  |  |  |
| --- | --- | --- | --- |
| Dependent variable (response) |  | Independent variable  (explanatory) | |
|  | Quantitative | Categorical |
| Quantitative | pair 1 | pair 2 |
| Categorical | pair 3 | pair 4 |

There are few true quantitative variables available in the data set. You may treat *ordinal variables with five or more categories* as quantitative (e.g. POLVIEWS).

If you choose, you can create new secondary variable(s) based on one or more of the variables in the dataset. For example, I recommend limiting categorical variables to between two and five categories. You should also collapse categories with a small number of individuals (i.e. less than 5% of your sample).

You will complete the analysis in R and report and summarize your results in a Word document. Results must be presented in tables and graphs in the body of the report as instructed below. Graphs should be made in R. Raw R output should not be included in the body of the report. Instead, a final R script of your analysis will be pasted into an appendix.

Sections

1. Background (10 pts)
   1. Data: Briefly describe the original source of your data, the 2016 General Social Survey. Identify your sample, describing who is in the sample and the sample size. (This may just be all adults 18+ without missing data or you may further limit the sample in some way to match your target population of interest.)
   2. Variables: Identify your four variables, describing first your dependent variables and then your independent variables. Say what concept they measure (e.g. socio-economic status), say how they measure it (describe response categories or units), and note whether each is quantitative or categorical.
   3. Hypotheses: Briefly explain why and how you expect each independent variable to be related to each of the dependent variables for each of your four pairs. You do not have to include formal hypotheses (although you can), but you should comprehensively describe the relationships you expect to find for each pair of independent and dependent variables. (You should draft this before you do the analysis so that you are not influenced by the results.)
2. Univariate Distributions (30 pts)

For each of your four variables, describe and interpret the univariate distributions. Each variable’s univariate distribution should also be presented in a table and graph as described below:

* Two quantitative variables - create a summary table that includes:
  1. Measures of center (mean and median)
  2. Measures of variation (standard deviation and five number summary)
  3. Lower and upper bounds of 95% confidence intervals for the means
* Two categorical variables - create a summary table that includes:
  1. Categories for each categorical variable
  2. Percentages of the sample in each category
  3. Lower and upper bounds of 95% confidence intervals for the percentage in each category
* All four variables – create univariate graphs (one for each variable). Provide an appropriate graph showing the distribution of each variable. These graphs may be histograms, boxplots, or bar graphs and should be made in R.

1. Bivariate Relationships (30 pts)

For each of the four pairs of dependent and independent variables, report and interpret your analysis of the relationship between the two variables. This text will be based on describing the analyses below. The analyses will differ depending on the variable types, for:

* Pair 1 – both dependent and independent variables are quantitative
  + Scatterplot (include this graph)
  + Correlation (can be reported in text only)
  + Regression (can report slope and r2 in text only)
* Pair 4 – both dependent and independent variables are categorical
  + Two-way table with column percentages (include this as a table)
  + Chi-square test (if one variable has three or more categories) or two-sample Z test of proportions (if both variables have two categories)
* Pairs 2 and 3 – one variable is categorical and the other is quantitative
  + If categorical variable has two categories: two-sample t-test of means (include the means and test results in a table)
  + If categorical variable has three or more categories: Choose one category as a “reference” category and perform a two-sample t-test of means for each other category and this reference category.

1. Conclusion (10 pts) – Provide a summary and reflections on your findings. Return to your hypotheses from the background section and describe whether your results met those expectations. For areas in which your results did not meet your expectations, speculate on why you may have gotten the unexpected results. Also, if you thought the variables would be associated due to a causal relationship state whether or not this provides definitive evidence of such a causal relationship and explain why.
2. R appendix (20 pts) – Paste an R script from your final analyses. In the course of your analysis, you will perform many analyses that are not part of the final report. These extra analyses include, for example, mistakes and additional exploration. Do NOT include the output from the R script.

Format

Text

* Main text
  + double-spaced
  + 12 pt font size
  + one-inch margins
* R script appendix
  + single-spaced
  + Courier New font
  + size 9
  + one-inch margins
* There are no length requirements; take the space you need

Tables and Figures

* Must have numbered titles that include the sample size
  + Example: Table 1. Descriptive statistics for variable 1 and variable 2 (n = 1,256).
* Should be placed near the text that they go with
* Figures (graphs) should be made directly in R
* Tables should be “pretty” aka report ready. Don’t copy raw R output as your tables.

Notes and tips:

1. Use the word significant as appropriate. Always use a significance level (α) of 0.05 with two-tailed tests.
2. Look back at the tables used to summarize R output in the problem sets for ideas on how to format tables.
3. As described on the GSS website, not all individuals had the same probability of being selected into the samples. Thus, to make the sample truly representative of the population, sample weights should be used in the analyses. We did not cover the use of sample weights in our course though. So, act as if each individual did have the same probability of being selected into the sample and that the sample is representative of the broader population (without the use of weights).