

17.871, Political Science Lab

Spring 2004

Problem set # 1: Research design, data, measurement, and using STATA

Handed out: Feb. 12, 2004

Due: Feb. 26, 2003, *at the beginning of class.*

All answers should be typed. No handwriting, please, unless you want to annotate a printout of an answer beyond what a computer printer can do for you.

I. Research Design

Write out your answers using complete sentences.

- I.1. The following are five relationships that social scientists and other researchers have discovered between two variables. For each relationship, do the following:
- (1) Indicate which is the *dependent variable* and which is the *independent variable* implied by the relationship *as stated*.
 - (2) Indicate whether the relationship *as stated* is likely to be infected by a “lurking variable” (i.e., is “spurious”) or some sort of non-random selection effect.
 - (3) If the relationship is spurious or due to a non-random selection effect, what is the common causal variable that creates the spuriousness? (Use your best judgement here.)
- I.1.a. Frequent smokers are more likely to die from lung cancer than non-smokers.
- I.1.b. Members of the U.S. House of Representatives who receive campaign contributions from the National Rifle Association are more likely to vote against gun control than those who don't receive those contribution.
- I.1.c. Countries that are recipients of large amounts of U.S. foreign aid are more likely to support the U.S. position on Iraq than countries that aren't aid recipients.
- I.1.d. Children who grow up in homes with lots of books perform better in school than children who don't have books at home.
- I.1.e. Children who grow up in homes with lots of televisions perform better in school than children who don't have televisions.
- I.2. Comment on the research designs of the following two studies. Discuss whether they are designed in a way that would allow the researcher to draw the conclusion that is drawn. Comment on the sampling and whether the sampling method may introduce biases into the analysis. State what are the dependent and independent variables in these designs, and what any confounding variables might be. If the research design was insufficient, write a short paragraph indicating why, and what should have been done to improve the design.
- I.2.a. Researchers are interested in determining whether postcards sent to registered voters encouraging them to vote actually worked. The researchers took the list of

registered voters in a town and randomly drew two samples— T , a sample of voters who were sent the get-out-the-vote postcard, and C , a sample of voters who were not sent the get-out-the-vote postcard. After the election, the researchers went to the town clerk to see who voted. They discovered that 85% of the T group voted, whereas 50% of the C group voted. The researchers' best estimate is that the causal effect of sending the post cards is to increase turnout by $85\% - 50\% = 35\%$ points.

- I.2.b. MIT faculty members are interested in determining whether ending spring-term freshman Pass/No Record has been a success. They decide to answer this question by comparing the GPA of spring-term freshmen before and after the change in Pass/No Record grading has taken effect. The average freshman GPA in the spring of 2002 is 4.0; the average freshman GPA in the spring of 2003 is 4.4. The faculty conclude that the change was a success. (Note the obvious: these are made-up data.)

II. Data “Scavenger Hunt”

Find the following facts, using data sources that are available in the Dewey Library, at the Harvard-MIT Data Center, and the web. For your answer, report the result *and* the source of the result. After each fact, I've given you a parenthetical hint about where I'm expecting you to find this information.

Note: The purpose of this set of question is *not* to see how well you do Google searches to pull up random facts, but rather to make sure you can find data that would be good to use in statistical analysis. Each of the answers to these questions will be found in a *bona fide* data source, whether that be a printed volume or an on-line data set. Your answer should yield a source that not only answers the question posed, but also should be able to guide you in doing the more general research that might be used by that data. For instance, if the question were “what was the average verbal SAT I score for college-bound high school seniors in 1980 and 2000?,” an incorrect answer would cite a newspaper article that simply reports the average. The correct answer would cite something like the College Board report entitled “2003 College-Bound Seniors: A Profile of SAT Program Test Takers,” which can be found at the following URL: <http://web.sfgov.org/site/uploadedfiles/election/docs/BT1PARTY.pdf>. The answer, by the way, is 502 (1980) and 505 (2000).

- II.1. What was the life expectancy at birth in the Dominican Republic for the period 1995–2000? (Library reference book)
- II.2. How many Democrats were elected to the U.S. House of Representatives in 1994? (Library reference book or government publication that may, or may not, be available online)
- II.3. What percentage of earned doctorates in the United States in engineering were earned by foreign nationals in 1977 and in 1999? (Government publication that

- may, or may not, be online. The government agency is one that's likely to be interested in the number of scientists and engineers in the U.S.)
- II.4. What were the "support scores" given by the Americans for Democratic Action to Sens. Helms, Durbin, and Clinton in 2002? (Web site)
 - II.5. What percentage of liberals who responded to the 2000 American National Election Study were Republicans? What was the corresponding percentage in 1972? (Web site)
 - II.6. What percentage of respondents to the 1984 American National Election Study, when asked whether the word "religious" described Ronald Reagan, replied "extremely well." (Confine yourself to those who had an opinion; you can use the Havard-MIT Data Center web site to get your answer.)
 - II.7. What is the current percentage of respondents to the Gallup Poll who approve of the way George Bush is handling his job as president? (Web site)
 - II.8. What was the value of the Consumer Price Index in November 2000? (Make sure you indicate the base period.) (Government web site or monthly publication by an agency of the federal government)
 - II.9. How many articles in major national newspapers mentioned the words "John Kerry" and "botox" in the same story during the weeks of January 25, 2004 and February 1, 2004. (I.e., give two numbers, one for each week.) (Online newspaper database, *not* the web site of a single newspaper.)

III. Measurement

- III.1. The *Statistical Abstract of the United States* is the most basic statistical reference for measuring all sorts of political, economic, and social factors in the United States. Using the 2002 *Statistical Abstract*, describe how you would create an indirect indicator to answer the following questions. (Please indicate the table in the *Statistical Abstract* you would use and the measure, or measures, you would use from that table. You may combine measures from multiple tables. You may consult the *Statistical Abstract* at Dewey or online at <http://www.census.gov/prod/www/statistical-abstract-02.html>.)
 - III.1a Which state is the most successful in educating its children?
 - III.1b Which state provides the most medical insurance coverage to its residents?
 - III.1c In what year (in the past twenty years) did the federal government spend the most on domestic programs?
- III.2 Assess the following claim: I have more faith in how well the inflation rate is measured, because it is based on real dollar transactions in the economy, than in the measure of presidential approval reported by the Gallup Poll, because it is based on people's opinions in a survey.
- III.3 What are the most important possible sources of bias in the following measures?

- III.3.a. The number of wars in the world on any given day, based on articles in the *New York Times*.
- III.3.b. The gross domestic product in developing countries.
- III.3.c. Educational achievement in states based on average SAT scores.

IV. Using STATA

- IV.1 Using a text editor such as EMACS, type the text from Exhibit 1 in the document “How to Use the *STATA infile* and *infix* Commands” into Server and save it in a file named scores.dat on your home directory. Write a “do” file that will create a STATA data set from this raw data and save it as a file called “scores.dta”. Turn in a “log” file that documents the STATA commands you issued to read in the data and save it.
- IV.2 The purpose of this question is for you to create two data sets and then merge them together using a common identifier.

Find two tables that interest you in the *Statistical Abstract of the United States* that meet the following criteria: (1) they have between 25 and 52 observations and (2) they have the same units of analysis (e.g., states, years, nations). You may use tables that you referred to in Part III above.

- A. Call these two tables Table A and Table B. Create separate STATA data sets for these two tables, naming the saved file tablea.dta and tableb.dta. Merge the two data sets. Save the merged data set, calling the saved file tablea+b.dta.
- B. Turn in the following:
 - i. The “do file” that shows how you create the data sets and merged them.
 - ii. A printout of the data.
 - iii. A short (one paragraph, 2 or 3 sentences) description of the tables you got your data from.