USC Price
Sol Price School of Public Policy

Professor John Romley
PPD303 Statistics for Policy, Planning and Development
Fall 2020

Problem Set 3

Sampling: Determine whether the sample is a simple random sample. Give a brief explanation of your choice.

1. (10 points) In the last general election, 132,312 adults voted in Dutchess County, New York. You plan to conduct a post-election survey of 500 of those voters. After obtaining a list of those who voted, you number the list from 1 to 132,312. Then, you use a computer to randomly generate 500 numbers between 1 and 132,312. Your sample consists of the voters corresponding to the selected numbers. Determine whether the sample is based on simple random sampling. Give a brief explanation of your choice.

Identify which of the following types of sampling is used: random or convenience.

2. (5 points) In a clinical trial of the cholesterol drug Lipitor, subjects were partitioned into groups which were given a placebo or Lipitor doses of 10 mg, 20 mg, 40 mg, or 80 mg. The subjects were randomly assigned to the different treatment groups (based on data from Pfizer, Inc.). Identify which of these types of sampling is used: random or convenience.

3. (5 points) In 1936, *Literary Digest* magazine mailed questionnaires to 10 million people and obtained 2,266,566 responses. The responses indicated that Alf Landon would win the presidential election, but Franklin D. Roosevelt actually won the election. Identify which of these types of sampling is used: random or convenience.

4. (15 points) Consider a study to evaluate the effect of *access to* dorm room Internet connections on the grades of college students in general. In a large representative dorm, half the rooms are randomly wired for high-speed Internet connections (the treatment group), and final course grades are collected for all residents. Which of the following pose threats to the validity of the study, and why? (Hint: There can be multiple threats. An explanation is required for each threat.)

Note: Causality can be established when there are no other sources of bias such as confoundedness due to lurking variables or sampling biases (voluntary response, attrition, mortality bias, etc.).

a. Midway through the year all the male athletes move into a fraternity and drop out of the study (their final grades are not observed).

b. Engineering students assigned to the control group put together a local area network so that they can share a private wireless Internet connection that they pay for jointly.

c. The art majors in the treatment group never learn how to access their Internet accounts.

d. The economics majors in the treatment group provide access to their Internet connection to those in the control group, for a fee.

5. (10 points) When it was found that hydroxyurea reduced the symptoms of sickle cell anemia, the National Institutes of Health released a medical bulletin. The bulletin said, "These findings are the results of data analyzed from the Multicenter Study of Hydroxyurea in Sickle Cell Anemia (MSH), which was a double-blind, placebo-controlled trial in which half of the patients received hydroxyurea and half received a placebo capsule." Explain to someone who knows no statistics what the terms (1) "placebo-controlled" and (2) "double-blind" mean here.

Confounding

6. (35 points) To study the effect of neighborhood on academic performance, 1000 families were given federal housing vouchers to move out of their low-income neighborhoods. No statistically significant improvement in the academic performance of the children in the families was found one year after the move.

a. What are the explanatory and response variables?

b. What are the subjects, factor(s), and treatment?

c. What does no significant difference mean in describing the outcome of this study?

d. Explain clearly why the lack of improvement in academic performance after one year does not necessarily mean that neighborhood does not affect academic performance.

e. In particular, identify some lurking variables whose effect on academic performance may be confounded with the effect of the neighborhood.

f. Use a figure like Figure 9.1 from the textbook to illustrate your explanation.

g. Suppose instead that families could apply for vouchers, but that a random group of applicants would receive them. Use a diagram like Figure 9.3 from the textbook to display this study design.

h. Suppose that 1376 families applied. Use Table B from the textbook, starting at line 128, to assign the first five families to the treatment (voucher) group.

Data Ethics

Most of the exercises pose issues for discussion. There are no right or wrong answers, but there are more and less thoughtful answers.

7. (10 points) Texas A&M, like many universities, offers free screening for HIV, the virus that causes AIDS. An announcement regarding the screening says, "Persons who sign up for the HIV Screening will be assigned a number so that they do not have to give their name." They can learn the results of the test by telephone, still without giving their name. Does this practice offer (1) *anonymity*, (2) *confidentiality*, *or (3) both*? Explain.

8. (10 points) The presidential election campaign is in full swing, and the candidates have hired polling organizations to take regular polls to find out what the voters think about the issues. *What information should the pollsters be required to give out?* Please answer the following questions:

(a) What does the standard of *informed consent* require the pollsters to tell potential respondents?

(b) The standards accepted by polling organizations also require giving respondents the *name and address of the organization that carries out the poll*. Why do you think this is required?

(c) The polling organization usually has a professional name such as "Samples Incorporated" in order NOT to reveal to the respondents whether the poll isbeing conducted by/financed by a political party or candidate. Would *revealing the sponsor* to respondents bias the poll? Or should the sponsor always be announced whenever poll results are made public?