Statistical Methods in Education Research
Vanderbilt University
Prof. Sean P. Corcoran

September 5, 2023
42 total points

Problem Set 1 Solutions

---

1. (**23 points**) The pro-school choice organization EdChoice conducts a regular opinion survey on matters related to school enrollment. You can find highlights from the July 2021 edition here: `https://edchoice.morningconsultintelligence.com/assets/127517.pdf` and detailed results at the following website: `https://edchoice.morningconsultintelligence.com/assets/127480.pdf`. Refer to this report to answer the questions below.

   (a) The EdChoice report contains many interesting statistics, including items related to support for school choice, teacher salaries, vaccination, and in-person schooling during COVID-19. Based on a quick skim of this report, would you say these are intended as descriptive or inferential statistics? Briefly explain your answer. (**3 points**)

   > These are intended to be inferential statistics. This is clear since the data were collected via a "national sample of adults" and a margin of error ("measure of precision") was reported. The margin of error is an example of "quantifying the uncertainty" associated with an inferential statistic, as described in class.

   (b) Consider the data the authors used for this report. What was the unit of observation? Was the data a population or a sample? How were the data obtained? How many observations were in the data? (**4 points**)

   > The unit of observation was an adult respondent to a survey conducted online. The data were a *sample* of 2,200 adults and 1,228 school parents.

   (c) Consider the following list of measures used in this report. For each, state whether the measure is: (1) quantitative or categorical; (2) nominal, ordinal, or interval/ratio scale. Base your answer on how they describe the outcome, and provide a brief explanation if necessary. (**6 points**)

   > i. Feelings about whether K-12 education is generally going in the right direction: **categorical, nominal**. Respondents were given the choice between "generally going in the right direction" or "generally gotten off on the wrong track." One could also argue this is an ordinal scale since the aim is to

measure the extent to which individuals thing education is going in the right direction.

ii. How much trust one has in principals to make good decisions about education: **quantitative, ordinal**. Choices were ordinal, including "a lot," "some", "not that much", and "not at all".

iii. If given the option, what type of school parents would select for their child: **categorical, nominal**. Options appear to have included private school, regular school, home school, charter school, and "don't know".

iv. School parents' comfort with their children returning to school in person: **quantitative, ordinal**. Choices were ordinal, including "very comfortable," "somewhat comfortable", "not that comfortable", and "not at all comfortable". (An "I don't know" option was also provided).

v. Parents' opinion on homeschooling since COVID-19 began: **quantitative, ordinal**. Similar Likert scale to the trust and comfort items above.

vi. How much parents would be willing to spend on a monthly basis to participate in a "learning pod": **quantitative, ratio**. Respondents provided a dollar amount.

(d) What is an example of a dichotomous variable used in this report? (**2 points**)

One example is the question about general feelings about K-12 education. There were two possible answers: "generally going in the right direction" or "generally gotten off on the wrong track."

(e) Take a look at Table EC1_1 in the detailed report. Which subgroups had the most positive view of K-12 education? Which subgroups had the most negative view? Do you see any interesting differences in the responses by gender, race, income, or political party? (**4 points**)

Groups that had the most positive view of education nationally included Democrats (61%), Liberals (63%), Labor Union members (64%), Blacks (60%), private school parents (65%), and special education parents (57%). Those with the most negative view included those age 65+ (30%), Republicans (33%), conservatives (33%), high school parents (38%), small town/rural area residents (34-35%).

(f) What is the population of interest in this study, the sampling frame, and sampling method? Briefly summarize. Note: more details on the report's methodology are provided here: `https://edchoice.morningconsultintelligence.com/methodology/` (**4 points**)

> There are several populations of interest in this study, including all U.S. adults, parents of school-aged children, and teachers. An important thing to note about the survey is that it uses *non-probability* sampling. This term means that the sample was not drawn from a well-defined frame from which one might calculate the probability of obtaining various samples. The survey is internet-based and respondents self-select into participating. (Interestingly, Morning Consult—the survey company—bids for a stratified sample of respondents from a large exchange of survey-takers). Weights are used to make the resulting sample "look" like the population of interest.

2. (**6 points**) A sociologist wants to estimate the average age at marriage for women in New England in the early 18th century. She finds within her state archives of marriage records for a large Puritan village for the years 1700-1730. She then takes a sample of those records, noting the age of the bride for each. The average age in these records is 24.1 years. Using a statistical method covered later, the sociologist estimates the average age of bridges at marriage for the population to be between 23.5 and 24.7 years.

> (a) What part of this example is *descriptive?* The descriptive part of this example is the average age of marriage in the sample (24.1).
>
> (b) What part of this example is *inferential?* The inferential part of this example is the estimate of the average age of brides in the population (23.5 to 24.7).
>
> (c) To what population does the inference refer? The population of interest is women who married in New England in the early 18th century.

3. (**3 points**) Identify the following quantitative variables as either discrete or continuous and identify their most likely level of measurement (nominal, ordinal, interval, or ratio).

(a) Number of students enrolled at Vanderbilt University in any particular term: interval/ratio and discrete. While the number of students is definitely a discrete measure, in practice we would analyze it as a continuous variable, since it takes on many values.

(b) Distance an individual can run in five minutes: interval/ratio and continuous.

(c) Hair length: interval/ratio and continuous.

(d) Number of hot dogs sold as baseball games: interval/ratio and discrete. While the number of hot dogs is definitely a discrete measure, in practice we would analyze it as a continuous variable, since it takes on many values.

(e) Self-concept as measured by the degree of agreement with the statement, "I feel good about myself," on a five point (Likert) scale: ordinal. While the underlying construct of "self-concept" is best thought of as continuous, it is operationalized here as a discrete variable, and thus would be analyzed as such.

(f) Lack of coordination as measured by the length of time it takes an individual to assemble a puzzle: interval/ratio and continuous.

4. (**10 points**) Visit the **amazing** website `www.socialexplorer.com`. Create a user account with your Vanderbilt email address so that you can have Basic access to this site. When prompted, you can affiliate with Vanderbilt's institutional subscription to get full functionality. From the main page, select "Explore Maps" and "Explore" under United States. On the next screen, select "Change Data." Use the Social Explorer site to find answers to the following questions.

   (a) In the list of variables for 2021, identify one categorical variable and one quantitative variable. (Consider how the variable was measured in the original data collection, not how it is summarized at a higher level). (**2 points**)

   The scale of the measure differs depending on whether we are thinking of the way the original question was posed to individuals, or how it is reported at aggregate levels of geography. For example, the Census question pertaining to "household type" is categorical, with separate categories for married-couple families, other families, and non-family households. Since these are Census tabulations for different geographic locations, the data we see here are quantitative: the count of households in a geographic area in each category. An example of a quantitative variable

at both levels—the original Census question and the aggregated variable—is household income. (e.g., median household income for the geographic area).

(b) In 2021, what is the median household income in Tennessee? In the Nashville-Davidson-Murfreesboro-Franklin metropolitan statistical area (MSA)? Are these statistics based on a population or sample? Explain how you know. (**4 points**)

Median household income in 2021 is $58,516 in Tennessee, and $72,537 in the Nashville MSA. (I obtained these by selecting median household income from the "change data" option, and selecting the appropriate geographic area). These data come from the 2021 American Community Survey, which is conducted by the U.S. Census. It is a sample of the population. These happen to be 5-year estimates, which means the data were collected over the 5-year period ending in 2021. To get more information about each variable you can click on the 'i' info button and then on 'Open Data Dictionary'. Incidentally, it is possible to download a dataset you have selected. Look for the "Reports" option.

(c) In 2020, what percentage of voters in Tennessee voted for the Democratic candidate (Joseph Biden) in the presidential election? What percentage of voters in Davidson County (Nashville) voted for the Democratic candidate? Are these statistics based on a population or sample? Explain how you know. **(4 points)**

This requires changing the source data, selecting 2020 and 'Presidential Elections'. With "state" selected as the geography level we see that 37.45% of voters in Tennessee voted for Joe Biden. Changing the geography level to county, we see 64.49% of voters in Davidson County supported Joe Biden. (Incidentally, this is up from 59.8% of voters supporting Hillary Clinton in 2016). These data are a *population* since they represent all votes cast in the 2020 election.

In addition to its maps, Social Explorer is a valuable resource for downloading data of all types. For example, from the main page you can select "Tables" and then a table of interest (e.g., American Community Survey 2017-2021). You can select levels of geography (e.g., county), and one or more tables of interest (e.g., school enrollment, educational attainment). After the table is generated you can download it as raw data, an Excel file, etc. Socialexplorer will even generate the Stata .do file to allow you to read the data into Stata.