STT 5811 HW01: Data Basics and Study Design

SOLUTIONS

Monday, January 26, 2026 @ 08:25 PM

Contents

Instructions

Do exercises 1, 2, 3, 4, and 8 in **IMS2E Section 1.4** and all even-numbered exercises in **IMS2E Section 2.5**. Use complete sentences and explain your answers for those questions where you are asked to do so.

Section 1.4 Exercises

Exercise 1

The dataset has 23 observations and 7 variables.

Exercise 2

The dataset has 19,961 observations and 9 variables.

Exercise 4

a. Based on the study findings in (c), the research question would be something like: "Does explicitly telling children not to cheat affect whether or not they cheat?" This is because the instructions seem to be the explanatory variable in the experiment. However, using only the initial description at the top of the exercise, it would be: "Do characteristics like age, sex, or only-child status impact honesty/cheating." However, those are not variables the researchers can randomize in an experiment, and are more like confounding variables. (Variations on this answer are possible/acceptable.)

b. The subjects are 160 children between the ages of 5 and 15.

c. There are five variables recorded in this study: (1) age (continuous numerical), (2) sex (categorical), (3) only child or not (categorical), (4) cheated or not (categorical), and (5) type of instructions given (categorical). Note: Type of instructions is the explanatory variable; cheated or not is the response.

Exercise 6

a. The research question is: "Is there a difference in unethical behavior between people who self-identify as having low social class versus high social class?"

b. The cases are 129 University of California at Berkeley undergraduates.

c. Two variables are: (1) low or high social class (categorical) and (2) number of candies the participant took (discrete numerical). The research scenario mentions money, education level and job respectability, but these were not variables; they were questions used to determine social class.

Exercise 8

a. The percent of treatment group participants who reported improvement is $66/85 = 0.776 = 77.6\%$.

b. The percent of control group participants who reported improvement is $65/81 = 0.802 = 80.2\%$.

c. There was a (slightly) higher percentage of improvement in the control group.

d. Even if the two treatments were in fact equally effective, it is unlikely that we would observe *exactly* the same improvement rates in the two experimental groups. The difference of 2.6%, which equates to only about one or two participants, seems like it could be due to just random chance.

e. The explanatory variable is the type of treatment the participant received (antibiotics versus placebo). The response variable is improvement in sinusitis symptoms.

Section 2.5 Exercises

Exercise 2

The population mean is 5.5 hours and the sample mean is 6.25 hours.

Exercise 4

a. Based on the scenario description, the population of interest is presumably all children aged 5 and 15. The sample is 160 children in this age group.

b. It is unlikely that the participants are a random (or representative) sample from the population, so the results are probably not generalizable. This study is an experiment, so if participants were randomly assigned to treatment groups (no cheating versus no instructions), the findings can be used to establish causal relationships, since that helps to eliminate confounding variables like age, sex, or singleton status. Refer back to Figure 2.8 in **Section 2.4**.

Exercise 6

a. The population of interest is probably people in general. However, since the sample consists of 129 UC Berkeley undergraduates, the target population is limited to UC Berkeley undergraduates.

b. If the students in this sample (who are likely not randomly sampled) can be considered representative of UC Berkeley undergraduates, then the results are generalizable to the UC Berkeley undergraduate population. If students from this school can be considered representative of all college students, or all people, we could generalize to these populations as well. However, being representative of all people is quite unlikely. The study is observational (participants were not randomly assigned to lower and upper class groups), so the findings cannot be used to establish causal relationships.

Exercise 8

a. The percentage of all videos on YouTube that are cat videos is the population parameter.

b. The value 2% is a sample statistic.

c. A video in the sample is an observation.

d. Whether or not a video is a cat video is a variable.

Exercise 10

a. This is an observational study.

b. To ensure that students from each year are fairly represented, we could use a stratified random sample. For example, we could randomly choose a fixed number of students from each part of campus (east and west), since different class years reside in different parts. We could also select a fixed number of students from each of the four class years, but this might be more challenging.

Exercise 12

a. This is an observational study.

b. Since the study is observational, we cannot infer causation.

c. Caffeine and lack of sleep are potential confounding variables that might explain the observed relationship between stress and muscle cramps. Students could be getting muscle cramps from increased caffeine consumption and/or lack of sleep, rather than directly from stress.

Exercise 14

Sampling from the phone book would introduce bias since it would not include unlisted phone numbers or cell phone numbers. If these numbers are missing at random, it might not be too much of a problem. However, but if people who choose to not list their numbers share a certain characteristic, the sample would not fairly capture such people and therefore would not be representative of the population.

Exercise 16

a. This is an observational study.

b. The friend's statement is not justified because it implies a causal association between sleep disorders and bullying, which we cannot infer from an observational study. However, it is reasonable to say that the variables are related. A better statement might be: "School children identified as bullies are more likely to be found to suffer from sleep disorders than children not identified as bullies."

Exercise 18

The estimate will be biased and tend to overestimate true family size. Families without children have zero chance of being sampled. A family with two children is twice as likely as a family with one, since there are twice as many kids in school. Even bigger families will have a proportionally larger chance.

Exercise 20

a. This is an experiment, since the researchers randomly assigned different treatments to the participants.

b. The explanatory variable is the Vitamin C treatment, which has four levels: 1g, 3g, 3g with additives, or placebo. The response variable is the duration of the cold.

c. The participants were blinded, since they did not know which treatment they received.

d. The study was double-blind with respect to the researchers who were assessing the participants, but the nurses who interacted with participants during the distribution of the medication were not blinded.

e. Participants were randomly assigned and blinded, so we would expect a similar number in each group to not to take their pills. Thus, non-adherence would not be a confounding variable in the study.

Exercise 22

Randomly assign participants to treatment groups: no music, instrumental, or music with lyrics. Have each person study a new concept. Give them the same quiz to assess what they learned. Compare the number of questions they get correct (on average) across the groups. (This is only one example. The important aspects are that the treatment groups are the music, participants learn the same thing and are assessed as the response variable. Random selection is not feasible, but random assignment should be used.)

Exercise 24

a. This is an experiment.

b. The treatment group is exercise twice a week and the control group is no exercise.

c. Yes, the study uses blocking. The blocking variable is age group (18-30, 31-40, 41-55).

d. The study is not blinded. Patients definitely will know whether or not they are exercising.

e. This is an experiment, so we can make causal inferences from the data. Because the sample is random, cause-and-effect can be generalized (cautiously, due to lack of blinding) to the population.

f. We cannot ethically require randomly selected people to be participants in a study. It also might not be ethical to tell people not to exercise, or impossible due to jobs or life circumstances.

Exercise 26

a. This is a simple random sample (SRS). Generally, the SRS method is effective in achieving a sample that is representative of the population. However, given the broad diversity (variability) in neighborhoods, a SRS might miss some neighborhoods that have distinct characteristics.

b. This is a stratified random sample. It would be an effective method, given the neighborhood variability. The sample will include representation from each neighborhood. However, it requires more work.

c. This is a cluster sample. It would *not* be an effective method, given the stated broad variability. Clusters should be similar to one another and the population as a whole. The sample definitely would exclude many neighborhoods with distinct characteristics, and thus not be representative.

d. This is a multi-stage sample. It starts with a cluster sample, and therefore it has the same flaws.

e. This is a convenience sample. It is not an effective method. Houses that are not close to the city council offices have zero probability of being chosen. It likely would represent only one neighborhood.

Exercise 28

a. The explanatory variable is the percentage of the population with a bachelor's degree. The response variable is per capita income (in thousands). In a scatterplot, we usually think of a response variable (Y) as being a function of an explanatory variable (X).

b. There is a strong positive linear relationship between the variables. As the percent of the population with a bachelor's degree increases, the per capita income increases. There are very few counties where more than 60% of the population have a bachelor's degree and very few countries that have a more than $50,000 in per capita income (upper right of the plot).

c. This is an observational study, so we cannot make a causal statement using these data. However, we can reasonably say that having a higher percent of the population with bachelor's degree is associated with having a higher per capita income (positive association). For example, in some cases, having a higher income or coming from a family with a higher income facilitates getting a bachelor's degree, which is the reverse of the initially suggested causation.

Exercise 30

a. This is an observational study.

b. The explanatory variables are screen time, the child's sex and age, and the mother's education, ethnicity, psychological distress, and employment.

c. The response variable is psychological well-being.

d. The data come from a "nationally representative" sample, so results can be generalized to the population from which these samples were drawn.

e. Because this is an observational study, the results cannot be used to establish any causal relationships.