Chapter 1 - Introduction to Data

Zachary Safir

Smoking habits of UK residents. (1.10, p. 20) A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

(a) What does each row of the data matrix represent?

Each row of a data matrix corresponds to a unique case (observational unit), and each column corresponds to a variable. In the case of this data, it appears that every row corresponds to a respondent.

(b) How many participants were included in the survey?

According to textbook, this portion of the data has 1691 participants, the data loaded into R however has 1693 observations.

(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

Sex is nominal, age is discrete, marital status is nominal, gross income is ordinal, smoker is nominal, amount on weekends is ordinal or discrete (if we cut off the text), amount on weekdays is ordinal or discrete (if we cut off the text).

Cheaters, scope of inference. (1.14, p. 29) Exercise 1.5 introduces a study where researchers studying the relationship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15[1]. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group.

(a) Identify the population of interest and the sample in this study.

Our population of interest is children between the ages of 5 and 15.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

The study needs to be repeated more times before it can be generalized to the population. The researchers also need to make the sampling method they used to pick the participants clear. It's possible the method they used did not create a sample group that represents the larger population.

---

[1] Alessandro Bucciol and Marco Piovesan. "Luck or cheating? A field experiment on honesty with children". In: Journal of Economic Psychology 32.1 (2011), pp. 73-78. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1307694

Reading the paper. (1.28, p. 31) Below are excerpts from two articles published in the NY Times:

(a) An article titled Risks: Smokers Found More Prone to Dementia states the following:

"Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a- day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

Association does not mean causation. While it's clear that smoking may increase your risk by a statistically significant amount, it would still be incorrect to conclude that smoking causes dementia from this information alone. We also cannot because this was an observational study.

(b) Another article titled The School Bully Is Sleepy states the following:

"The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders."

A friend of yours who read the article says, "The study shows that sleep disorders lead to bullying in school children." Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

As mentioned above, association does not mean causation and this is another observational study. While it may be true that "bullies were twice as likely to have shown symptoms of sleep disorders" it does not mean that having a sleep disorder alone can lead to bullying. There are likely many other factors that are important to this discussion that are not being observed. It's also simply not clear if the sleep disorders are a cause for the bullying, or symptom of a larger problem.

Exercise and mental health. (1.34, p. 35) A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41-55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

(a) What type of study is this?

This is an experiment.

(b) What are the treatment and control groups in this study?

The group that exercises two times a week is the treatment group and the control group the is one that was instructed to not exercise.

(c) Does this study make use of blocking? If so, what is the blocking variable?

This study uses age as a blocking variable. The blocks used are ages 18-30, 31-40, and 41-55.

(d) Does this study make use of blinding?

It does not. The participants know which groups they are in.

(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.

Repeated trials of the experiment need to be done. However, given that they properly conducted stratified random sampling, a casual relationship can begun to be formed from the experiment. More control over the participants health prior to starting the trial needs to be in order to get proper results.

(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

As mentioned, there are further improvements that need to be made in order for this study to work. Another important piece that I did not mention would be diet. The diet of each participant will have a huge effect on the results. Having more control over that will be crucial to improving this study.