Chapter 1 - Introduction to Data

Adam Gersowitz

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 the data matrix represents the results of the survey for an individual.

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

There were 1691 participants included in the survey.

(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: catigorical, non-ordinal

age: numerical, continuous

martial: categorical, non-ordinal

grossIncome categorical, ordinal

smoke: categorical, non-ordinal

amtWeekend: numerical, descrete

amtWeekdays: numerical, discrete

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.

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

The sample are the 160 children involved in the study.

(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 results of the study can be generalized to the population as long as there was no bias in which children were selected. Assuming the children were randomly selected the study can be genrealized.

The findings can be used to establish causal relationships because the study was experimental not observational.

[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.

Based on this study we cannot conclude that smoking causes dimentia later in life. Due to the fact this study is observational we cannot use it to conclude causal relationships. Additonally, since the sample was self selected selected through a voluntary exam there could be additional factors that could be correlated with both dimentia and smoking that were not included or controlled for on the exam.

(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?

This statement is not justified. This study was observational not a randomized experiment, so it can not be used to conclude a causal relationships. Additionally, since the study asked the parents and teachers opinions of the children's sleep sleep habits and behavioral concerns the results could be skewed due to their relationship with child. What can be concluded from this study is that the parental/teacher assessment that a child is having problems with disruptive behvior and bullying is correlated with the prental assessment of sleep habits which indicate the child exhibited symptoms of sleep disorders.

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 rep- resentative 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 a randomized experiment.

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

The treatment group of this study is the group that exercised twice a week while the control group were the ones instructed not to exercise.

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

This study does make use of blocking. The blocking variable is the age and age range of participants.

(d) Does this study make use of blinding?

There is no blinding in this study. The patients in the second group know they are in the control group because they are not receiving any type of treatment.

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

The results of this study can be used to establish causal relationship between exercising twice a week and mental health because it was an experimental study that utilized random sampling. The conclusions can be generalized to the public at large because the experiment used random sampling.

(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?

I would have reservations about the limitations of the study regarding the instruction to exercise only twice a week. I would want to see what the relationship between exercising 5 times a week is with mental health as compared with 2 or even 10 times. By only having people exercising twice it makes it more difficult to extablish a broad causation between exercising and mental health.