Chapter 1 - Introduction to Data

Tyler Frankenberg

2021-02-07

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 represents the survey data gathered for a single respondent. There is one row included for each respondent.

(b) How many participants were included in the survey?
1691 participants were 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.
Age - numerical, discrete
Marital - categorical, nominal
grossIncome - categorical, ordinal
smoke - categorical, nominal
amtWeekends - categorical, ordinal
amtWeekdays - categorical, ordinal

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: all 5-15 year olds. The sample is 160 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.
We don't know enough about how the sample was selected, or about how representative it was, to confidently generalize findings to the population at large. We cannot establish causal relationships for this reason because we don't know if other characteristics (like age, for example) are confounding variables. Because there doesn't appear to have been any blinding used in the study, we must also be cognizant of bias.

---

[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.
The study provides compelling evidence to suggest smoking causes or at least contributes to the development of dementia. We should not conclude a causal relationship, however, without seeing the results replicated in additional studies that all control for confounding variables. Even then, we'd want to consider the possibility that tendency to smoke and dementia are both caused by a third variable, such as stress level.

(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?

While the study results suggest a link between sleep disorders and behavioral concerns or bullying, the study as summarized here does not prove causality. It is based on multiple surveys completed by more than one third party for each subject. There is no control for the point in time at which disciplinary issues were experienced, versus when the student experienced sleep issues. Therefore, it could be just as likely that the students develop sleep disorders due to anxiety related to their disciplinary issues. There could also be other variables. For example, does the study take into account whether the student's home life is influenced by poverty or by a parent's alcoholism or domestic violence? The presence of any of these variables, or many other more innocuous ones, could hypothetically cause both sleep disorder and influence the student's likelihood of experiencing behavioral issues.

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, controlled experiment.

(b) What are the treatment and control groups in this study?
The treatment group is the group given instruction to exercise, and the control group is the group instructed not to exercise.

(c) Does this study make use of blocking? If so, what is the blocking variable?
Yes. The blocking variable is age group.

(d) Does this study make use of blinding?
This study does not seem to make use of blinding.

(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.
The results should be interpreted with caution. They may suggest a relationship between exercise and mental health, but it should not be assumed to be causal without further replication. Furthermore, we should be concerned about confounding variables because the study only blocks for age group and does not seem to control for other demographics or for other contributors to mental health like diet, medications, employment type and work schedule, relationship status and other social factors.

(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'd tell them to come back to me when they've addressed my concerns in question (e) above.