Chapter 1 - Introduction to Data

Jeff Shamp - Spring 2020 Data 606

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 represent one observation of a given individual.

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

It looks like total sample of 1691.

(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 and Smoke are both categorical. Neither are ordinal, they are presumably dichotomous. The rest are numerical. Of the numerical variables; age, amtWeekend, amtWeekday are discrete and Income is continuous.

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.

Population of interest are children between 5 and 15, so school=aged children. The samples are the selected children.

(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.

Since to do not know how the children were selected or how children were assigned treatments in this study, we can only conclude that this is a correalated and not a causal statement. Nor dose this study generalize to the population given the explicit information given. If we knew something about assignment or sampling, we might be able to conclude more.

---

[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.

This is observational data based on a voluntary sampling. So this study does not generalize to population and casuation cannot be shown with this data alone. Voluntary data can be biased so that should rule out generalizability and the observational nature means that there was no random assignment of 'treatment'.

(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 study has some serious problems with causation and generalization. It's voluntary data so it will not gereneralize well and we as readers do not know the details on how teacher assessments were made. We also don't know about any kind of treatment from this data so we lose out on assignment information. As such this is correlation data of this sample only.

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 an experiment with random sampling and assignment. It's not clear if treatments were assigned in a blind fashion or not.

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

control appears to be non-exercise and treatment is exercise 2x per week.

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

No. This experiment uses strata to divide the individuals into representative proportions of age groups. One was they could use blocking is to further divide the age groups into blocks of people who exercise at the same frequency already.

(d) Does this study make use of blinding?

It is unclear if the people in the study knew which group they were in of if there were other groups

(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.

This does appear to check all the boxes in experimental design. Random sampling, random assignment within strata. These kinds of studies have many confunding variables, though which inhibits causation.

(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 want to see maybe more detailed samples of people, something more than age group. Also other lifestyle choices could be the source of confunding variables. If someone exercises 5 times a week and then gets the rest group, their base fitness might affect the result of the study. This could also be true of a reduced workout schedule in the treatment group. Also, big things like, is an individual an alcoholic? That might put a damper on the results of the treatment.