DATA 606 Homework 1

(Chapter 1 - Introduction to Data)

Donny Lofland

8/30/2019

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?

A single *observation* or in this case, an individual (participant) included in the study.

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

Assuming the ID column (index column to left) is autoincremented with no missing values (e.g. removed participants), then there are **1691 participants** in the study.

(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.

- *index*: **numerical**
- *sex*: **categorical** (values: *Female*, *Male*, though there could be others not visible in the table)
- *age*: **numerical, discrete**
- *marital*: **categorical**
- *grossIncome*: **categorical, ordinal**
- *smoke*: **categorical**
- *amtWeekends*: **categorical, ordinal** (as presented in chart, but I would consider converting to numerical)
- *amtWeekdays*: **categorical, ordinal** (as presented in chart, but I would consider converting to numerical)

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. 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 is **children in general** while the sample is **children ages 5-15**.

(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 to generalize the sample back to population. We aren't told if the sample was a truely random sample from kids all over the world, from one geographical area, of a predominant ethonuc group, etc. We also cannot determine a causal relationship ... did some childern understand or interpret the instructions differently? If this experiemnt were run with more affluent children vs poorer children, this might affect outcomes.

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.

> We cannot establish a causal relationship ... there could be a coorelation, but only within the sample. We cannot generalize the sample to the population:
>
> - To establish a causal relationship, we'd ideally need to create random samples, then make one group smoke and the other not smoke. This would be difficult and clearly have ethical concerns.
> - The study was limited to looking at ages 50+ ... there could be other factors earlier in life that led to smoking and dementia and both smoking and dementia were triggered by a 3rd cause.
> - The study only included "Health Plan Members". "non-Health Plan Members" might have different outcomes.
> - The study only included volunteers - this may have created a sample bias.

(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?

> The statement is **not** justified. At best we might be able to say there may be a coorelation between sleep patterns as identified by parents and disuptive behavior. Some possibilties, questions and concerns before trying to draw further inferences from the study:
>
> 1. It's posiible disruptive behavior lead to poor sleep
> 2. It's possible poor sleep led to disruptive behavior
> 3. It's possilbe some other factor led to both poor sleep and disurptive behavior (e.g. stress, diet, family dynamics, etc)
> 4. It's possible parents are not good judges of "poor sleep" and introduced measurement bias.
> 5. "poor sleep" and "disruptive behavior" are both very subjective features which raises concerns about interpretation.
> 6. Did the researchers account for gender, socioeconomic, age, demographic and/or other factors?

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?

Experimental

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

The treatment is **exercise** where the *treatment* group will be exercising and the *control* group will not exercise.

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

The experimenter is blocking for **age**

(d) Does this study make use of blinding?

No, both the researchers and participants know which group they are in.

(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.

While the researchers are certainly proposing a stronger study, at best, they might draw causal relationships within the study group only. As I mention below in (f), there are confounding factors and sufficient questions that would prevent me from accepting conclusions generalized to the population at large. Some of the concerns get at the credibility of within-sample causal conclusions. If researchered measured a large difference in outcomes, that would give more evidence of within-sample causal findings, but should outcome differences be minor, there is a greater chance we aren't seeing a causal relationship even within-sample.

(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?

- Will particiapnts be sampled from different parts of the country? Will ethnic or demographic factors be considered and/or possibly blocked for?
- How are we controlling for and/or tracking the type and duration of exercise? Different participant might define "exercise twice a week" in different ways. A better study might ask particiants to come to a gym and exercise with a set routine under guidance of a personal trainer helping with the study.
- How are we controlling for the control group not accidentally getting exercise? There are lots of other activities that might provide similar physical characteristics of "exercise". For example, gardening, on-the-job exertion, hobbies, sports, etc. Possibly we need to instruct participants to spend fixed timeslots "sitting and doing nothing". Though this could confound as possibly "sitting and doing nothing" leads to changes in mental health.

- Does a fixed exercise routine lead to stress for some participants which might offset hypothesized benefits from exercies? How would this be measured or accounted for? For example, if a segment of the exercise group has physical concerns that make exercise problematic, the researchers might need to further block.