Chapter 1 - Introduction to Data

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 in the data matrix represents one participant (one UK resident) from the survey conducted. Each respondent is an observational unit.

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

According to the view of the data matrix above, it appears that 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: categorical variable, nominal
-age: numerical, discrete
-marital: categorical, nominal
-grossIncome: categorical, ordinal
-smoke: categorical, nominal
-amtWeekends: numerical, discrete
-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 in this study are children between the ages of 5 and 15. The sample of the study are 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.

The explanation of the study above doesn't give much insight into the sampling method used in the experiment. I also couldn't access the full paper from the link provided below, and the abstract doesn't give additional information about the sampling method. Therefore, we do not know enough to determine whether or not the sample population of 160 children would allow for findings to be generalizable. Since the sample population of 160 children could all be from the same geographic location, from similar lived experiences, etc., selection bias could affect generalizeable statments about differences between the two treatment groups. We'd have to confirm that the sample population was determined through random sampling before being able to make generalizable assumptions.

The study is an observational study and does not explicitly say that the groups were subjected to random assignment. Therefore, causal relationships cannot be established.

[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 conlcude that smoking causes dementia later in life. Since the 23,123 members that participated in the exam did so voluntarily, the sample is not a random sample. It is also an observational study, with no designed experimental protocol - no treatment and control groups were established.

(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. There is no designed experimental protocol in this study either, with no distinct treatment and control groups. We can say that children with sleep disorders and tendencies to be identified as bullies by parents and teachers are assoicated.

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 groups are half the subjects from each age group that are randomly assigned to exercise twice a week. The control groups are the other half of the subjects from each age group that are randomly assigned to not exercise.

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

The study does make use of blocking. The blocking variable is a categorical age variable (18-30, 31-40, and 41-55)

(d) Does this study make use of blinding?

This study does not make use of blinding, since the control group is told explicitly to not exercise.

(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 of this study can be used to establish a causal relationship between exercise and mental health and the conclusions can be generalized to the population at large since there was use of stratified random sampling in the experiment, and there is an explicit experimental protocol established (it is not an observational study). The only reservation I'd have is the number of participants in the study. Establishing causal relationships and generalizable conclusions from this study would hold more weight with a larger sample size.

(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 a few reservations about the study proposal. I would like to know how many study participants there will be in both the treatment and control groups. Any differences exhibited between the two groups would be better justified with higher Ns in the proposed study. Additionally, some may argue the ethics of instructing the control group to not exercise, since it could be restricting this group from mental health gains.