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 reflects data on one sample resident of the UK. A row is referred to as an observational unit or case

The columns for sex, age, marital, grossIncome are demographical, and can be used for sampling strategies, experiment strategies, or data analysis.

The columns smoke, amtWeekends, amtWeekdays constitute the extent of the habits we are trying to analyze.

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

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.

Table 1: Variable Types

variable	type	subtype
sex	categorical	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 is children between 5 and 15. Its not clear if it was meant to be culturally independent, i.e. to study children in general, or if its understood to be children of a certain country or city

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.

I can imagine certain sampling biases can exist to render the sample and target populations to be incompatible.

I dont know the researchers nor have I seen how the study is being implemented.

Its possible children in rural areas tend to be more honest or more compliant than children raised in an urban environment.

Also, I would have to assume that the instructions actually carried enought weight that a reasonable person would be effected, and that they would impress equally to each age/sex/cultures.

[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.

Intuitively I would think that dementia might be influenced by a lack of physical or mental exercise.

I would think that the study needs controls to measure dementia levels across samples that have the same baselines of stress levels, nutrition, exercise, or anything else that we think might be a factor.

I am not a doctor but I would think smoking effects the lungs and heart foremost.

(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?

Again, my intuition tells me that factors other than lack of sleep could lead to bullying.

Perhaps the kids have a lot of stress due to issues in their social or family environemnt.

However, unlike the question about smoking, where I didnt think there would be a link between smoking and dementia, it does make sense to me that a lack of sleep would increase the general irritability in any person.

The study needs to be implemented in a manner that respects the reality several factors can increase bullying.

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?

Its a stratified study, with each strata divided into treatment/control groups.

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

The Treatment group *is the group receiving some sort of stimulus, in this case it is exercise*

The Control Group *is meant to demonstrate the effect of the treatment, in this case it is the subjects who do not exercise*

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

Blocking infers grouping by an attribute suspected of being confounding, I dont see that here.

(d) Does this study make use of blinding?

No. The patients seem to be aware of their 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.

I would suspect that exercise improves mental health. I also would think the association wouldnt be confounding by many other factors, that it would be true for anyone in any culture or demographic.

(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 support the study as I think its important for people to understand the positive benefits of exercise. If the study demonstrates a positive association between exercise and happiness (like I think it would) it would be good to use in public service advertising.