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?
(b) How many 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.

Answers - Smoking habits of UK residents

(a) Each row represents the individuals that took the survey. Participants or Subjects. Basically the Sample.

(b) There were 1691 participants included in the survey.

(c) Below are the types of variables in the dataset.

- sex: Categorical Variable
- age: Numerical Variable - Discrete
- marital: Categorical Variable
- grossIncome: Categorical Variable - Ordinal
- smoke: Categorical Variable
- amtWeekends: Categorical Variable - 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.

(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.

Answers - Cheaters, scope of inference

(a) Populaton is the children ages between 5 and 15. Sample is the 160 children participated in the study.

(b) In order to generalize the results, the sample needs to represent the population. If the selected children are representative of the population which is children ages between 5 and 15, then the result of the study can be generalized. The findings of the study can be used to establish casual relationships, as it is an experiment.

---

[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.

(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?

Answers - Reading the paper

(a)Part of the data was collected as a voluntary exam, which indicates that there might be a Voluntary Response Sample Bias. Another part of the data was collected through health behavior survey specific to the years 1978 to 1985, when the participants where 50-60 years old. The data does not seem to be collected in a random framework of population. There is no notion of implied randomness. Hence, the sample may not be representative of the population and we can not conclude that smoking causes dementia later in life.

(b)Considering the survey was collected through parents on their children, there might be indications of convenience sample bias. The survey also was specific to the students that attended to "The University of Michigan" and was not random and the result might be limited to the sample size. The sample may not be representative of the population and the statement of "The study shows that sleep disorders lead to bullying in school children" is not justified. The response variable in this case is, behavioral concerns such as disruptive behavior and bullying. The explanatory variable is sleeping disorder. The sample may not represents the population and there may not be causal relationship between sleeping disorder and disruptive / bullying, however there might be an association between them.

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?
(b) What are the treatment and control groups in this study?
(c) Does this study make use of blocking? If so, what is the blocking variable?
(d) Does this study make use of blinding?
(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.
(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?

Answers - Exercise and mental health.

(a)The study is experiment. Researchers randomly assign subjects to establish casual releationship between the explanatory and response variable.

(b)The treatment group is the group that exercises twice a week. Control group is the group that does not exercise.

(c)We can suspect that the age group is the known variable that might impact the response variable (mental health exam results in terms of effectiveness of the exercise to mental health). Age group would be the blocking variable.

(d)The subjects are aware of what group they are in (treatment or control). The study does not use blinding.

(e)There is random sampling with no mention of non response, voluntery response or convinience sample examples, this might indicate that sample used in the study can be the representative of the population. As this is an experiment study, it can also be used to establish a casual relationship between exercise and mental health.

(f)In my opinion, the scope of the study needs details around the stratified random sampling and experimental design. For example; does the study includes replicating the experiment by collecting larger sample or re-do the entire experiment?