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

(a) Each row in the data matrix represents a single person who took part in the survey. The information in each row includes their sex, age, marital status, income, whether they smoke, and how much they smoke on weekends and weekdays.

(b) The survey included **1691 participants**, as there are 1691 rows in the data matrix

(c)

Sex

- **Type:** Categorical
- **What it represents:** It tells us whether the person is Male or Female.
- **Order:** No order (it's just a label).

Age

- **Type:** Numerical
- **What it represents:** It shows how old the person is.
- **Order:** Continuous (it can be any number and can be very specific).

Marital

- **Type:** Categorical
- **What it represents:** It tells us if the person is Single or Married.
- **Order:** No order (it's just a label).

Gross Income

- **Type:** Categorical
- **What it represents:** It shows the range of the person's income, like "Under £2,600" or "£10,400 to £15,600."
- **Order:** Yes, there is an order (from lower to higher income).

Smoke

- **Type:** Categorical
- **What it represents:** It tells us whether the person smokes or not (Yes or No).
- **Order:** No order (it's just a label).

AmtWeekends

- **Type:** Categorical
- **What it represents:** It shows how many cigarettes the person smokes each day on weekends.
- **Order:** No order (but if we wanted to, we could turn this into numbers).

AmtWeekdays

- **Type:** Categorical
- **What it represents:** It shows how many cigarettes the person smokes each day on weekdays.
- **Order:** No order (but like AmtWeekends, it could be turned into numbers if needed).

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

a.

Population of Interest: The researchers want to learn about how all children aged 5 to 15 act when it comes to cheating. They aim to understand cheating behaviour in this entire age group.

For Example: The sample consists of 160 children who were part of the study. They were between 5 and 15 years old and were split into two groups. One group was told not to cheat, while the other group received no instructions.

b.

Generalizability:
The results might be useful for other children aged 5 to 15, but it depends on how the 160 kids were chosen. If these kids were picked randomly from different places and backgrounds, the results could apply to all kids in this age group. However, if the kids were all from the same area or had similar backgrounds, the results might not be accurate for other children.

Causal Relationships:
The study can help show if telling kids not to cheat actually changes their behavior. The researchers randomly told some kids not to cheat and didn't give instructions to others. Because of this random assignment, any differences in cheating are likely because of the instructions, not other factors. This way, the study can help us understand if telling kids not to cheat really works.

[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 analysed data from 23,123 health plan members who participated in a voluntary exam and health behaviour 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 behavioural concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behaviour or bullying. The researchers found that children who had behavioural 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

a.

Analysis:

- **Link vs. Cause:** The study finds that smoking is linked to a higher risk of getting dementia later in life. But just because smoking and dementia are connected doesn't mean smoking causes dementia directly. The study shows a relationship, not a cause-and-effect.

- **Adjustments for Other Factors:** The researchers considered other factors that might affect the results, which helps make the findings more reliable. However, there might still be other factors that weren't considered.

- **Reasoning:** The study suggests that smoking is associated with a higher risk of dementia, but it doesn't prove that smoking directly causes dementia. To confirm that smoking causes dementia, more research is needed, such as long-term studies or more evidence about how smoking might lead to dementia.

b.

Analysis:

- **Link vs. Cause:** The study finds a connection between sleep problems and behavioural issues like bullying. However, it doesn't prove that sleep problems directly cause bullying.

- **Incorrect Conclusion:** Saying "the study shows that sleep problems lead to bullying" isn't correct. The study only shows a link, not a cause. Sleep problems and behavioural issues might be related because of another factor, or the relationship might go both ways.

- **Best Description:** The study shows a link between sleep problems and behavioural issues like bullying. This means kids with sleep problems are more likely to have behavioural issues, but it doesn't mean sleep problems cause bullying. More research is needed to understand if there is a direct cause-and-effect relationship.

In summary:

- For the smoking and dementia study, we can infer a strong association but not causation.
- For the sleep disorders and bullying study, we can identify a correlation but cannot conclude causation.

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- presentative 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

a. **Type of Study:** This is an **experiment** where people are randomly assigned to different groups to see the effects of exercise on mental health.

b. **Treatment Group:** The people who are told to exercise twice a week.
Control Group: The people who are told not to exercise at all.

c. **Blocking:** Yes, the study uses blocking by dividing participants into age groups (18-30, 31-40, and 41-55). This ensures that each age group is equally represented in both the exercise and no-exercise groups.

d. **Blinding:** The study doesn't mention blinding, which means participants know if they are exercising or not. They can't be kept unaware of their own activity. However, the people who assess mental health could be kept in the dark about who exercised and who didn't to prevent any bias in their measurements.

e. **Causal Relationship:** Since the study randomly assigns people to either exercise or not, it can show whether exercise directly affects mental health.

Generalizability: The results should apply to people in the same age groups. However, we need to ensure the sample also reflects different genders and backgrounds to confirm that the findings apply to a wider group of people.

f. **Reservations:**

- **Blinding Concerns:** Since participants know if they are exercising, it could influence how they report their mental health. To reduce bias, ensure that the people assessing mental health don't know which group the participants are in.
- **Sample Details:** Make sure the sample includes people from various backgrounds, not just different ages, to get a more accurate result.

- **Exercise Details:** Clearly explain what “exercise twice a week” means so everyone follows the same plan.
- **Study Duration:** Ensure the study lasts long enough to see if exercise really impacts mental health.
- **Measurement Tools:** Use reliable and well-tested methods to measure mental health to ensure accurate results.