Chapter 1 - Introduction to Data

Gabriel Campos

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?
ANSWER: The rows represent a UK resident and participant in the study.

```
glimpse(smoking)
```

```
## Rows: 1,691
## Columns: 12
## $ gender                <fct> Male, Female, Male, Female, Female, Female, M...
## $ age                   <int> 38, 42, 40, 40, 39, 37, 53, 44, 40, 41, 72, 4...
## $ marital_status        <fct> Divorced, Single, Married, Married, Married, ...
## $ highest_qualification <fct> No Qualification, No Qualification, Degree, D...
## $ nationality           <fct> British, British, English, English, British, ...
## $ ethnicity             <fct> White, White, White, White, White, White, Whi...
## $ gross_income          <fct> "2,600 to 5,200", "Under 2,600", "28,600 to 3...
## $ region                <fct> The North, The North, The North, The North, T...
## $ smoke                 <fct> No, Yes, No, No, No, No, Yes, No, Yes, Yes, N...
## $ amt_weekends          <int> NA, 12, NA, NA, NA, NA, 6, NA, 8, 15, NA, NA,...
## $ amt_weekdays          <int> NA, 12, NA, NA, NA, NA, 6, NA, 8, 12, NA, NA,...
## $ type                  <fct> , Packets, , , , , Packets, , Hand-Rolled, Pa...
```

(b) How many participants were included in the survey?
ANSWER: 1,693

(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.
ANSWER:
Sex: Categorical, Not Ordinal
Age: Numerical, Continuous
'Marital Status': Categorical, Not Ordinal
`Highest Qualification`: Categorical, Ordinal
Nationality: Categorical, Not Ordinal
Ethnicity: Categorical, Not Ordinal
`Gross Income`: Numerical, Continuous

Region: Categorical, Not Ordinal
`Smoke?`: Categorical, Not Ordinal
`Amount Weekends`: Numerical, Discrete
`Amount Weekdays`: Numerical, Discrete
Type: Categorical, 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.
ANSWER:
The **population** of interest are all children between 5 and 15 years old.

(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. *The study splits the group into 2 groups (instructions and no instructions) and results were recorded. This qualifies the case study as experimental and not observational.*
ii. *The explanatory variable is children being instructed not to cheat and the response variable is cheating rate.*
ANSWER:
Noting the above points, the findings can be used to establish a causal relationship and be generalized to the population. However, the results could potentially be skewed, if the ages for each group are disproportional (e.g. younger children in instucted group and older in non instructed, vice versa).

---

[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.
ANSWER::
The case study was observational, not experimental. We should take caution to not generalize these results, because they involve volunteers with individual characteristics. These characteristic might not be representative of all smokers and people with with dementia. Subsequently, we cannot conclude that smoking causes dementia later in life.

(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?
Association != causation
i.This study was observational, not experimental.
ii. Causation can only be inferred from a randomized **experiment.**
iii. Bullying may be a **confounding variable** because it may be correlated with both the explanatory and response variable.
** **In other words, we can adversely conclude bullying and/or behavioral issues lead to sleep disorders.****
v.Ultimately other confounding variables may exist that should be considered.
vi. Bullying cannot be quantitatively examined or measured.
ANSWER:
The conclusion "sleep disorders lead to bullying in school children" is **not** justified. At best the conclusion to be made is that "children with behavioral issues and those **identified** as bullies are twice as **likely** to have shown symptoms of sleep disorders."

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?
ANSWER:
The study is experimental.

(b) What are the treatment and control groups in this study?
ANSWER:
i)**Treatment Group**: Subjects assigned to exercise twice a week.
ii)**Control Group** Subjects instructed not to exercise. The control group provides a reference to compare and measure the impact made from exercising.

(c) Does this study make use of blocking? If so, what is the blocking variable?
ANSWER:
Yes the studying makes use of blocking by grouping individuals base on age variables:
18-30,31-40 and 41-55 year old from the population.

(d) Does this study make use of blinding?
ANSWER:
No the study does not make use of blinding, because the researchers to not keep the subjects uninformed about treatment (e.g. instructions to exercise or 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.
i)The study is experimental, using random sampling.
ii) Explanatory variable is instructions to exercise/not exercise and response variable is results of mental health exam.
ANSWER:
Noting the above points, the study can establish a causal relationship between exercise and health and can be generalized for the at large ages 18-55. The possible discrepancy could come from the subjects mental and physical health before participating in the study.

(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?
ANSWER:
My concerns would be having an otherwise healthy participant become unhealthy by not exercising. Likewise the risk that a participant with poor mental hygiene may get worse from not exercising exist. I would require heavy oversite to ensure safety of those involved and would ask participant to lessen excercising, rather than discontinue altogether.