Chapter 1 - Introduction to Data

Amber Ferger

Problem 1

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.

Load smoking data

```
library(openintro)

## Please visit openintro.org for free statistics materials

##
## Attaching package: 'openintro'

## The following objects are masked from 'package:datasets':
##
##     cars, trees

smokingData <- openintro::smoking
```

(a) What does each row of the data matrix represent? **Each row represents an individual in the dataset.**

(b) How many participants were included in the survey? **There are 1691 participants in the survey.**

```
nrow(smokingData)

## [1] 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.

```
str(smokingData)

## 'data.frame':    1691 obs. of  12 variables:
##  $ gender              : Factor w/ 2 levels "Female","Male": 2 1 2 1 1 1 2
2 2 1 ...
##  $ age                 : int  38 42 40 40 39 37 53 44 40 41 ...
##  $ maritalStatus       : Factor w/ 5 levels "Divorced","Married",..: 1 4 2
2 2 2 2 4 4 2 ...
##  $ highestQualification: Factor w/ 8 levels "A Levels","Degree",..: 6 6 2
2 4 4 2 2 3 6 ...
##  $ nationality         : Factor w/ 8 levels "British","English",..: 1 1 2
```

```
2 1 1 1 2 2 2 ...
##  $ ethnicity           : Factor w/ 7 levels "Asian","Black",..: 7 7 7 7 7
7 7 7 7 7 ...
##  $ grossIncome         : Factor w/ 10 levels "10,400 to 15,600",..: 3 9 5
1 3 2 7 1 3 6 ...
##  $ region              : Factor w/ 7 levels "London","Midlands & East
Anglia",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ smoke               : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 2
2 ...
##  $ amtWeekends         : int  NA 12 NA NA NA NA 6 NA 8 15 ...
##  $ amtWeekdays         : int  NA 12 NA NA NA NA 6 NA 8 12 ...
##  $ type                : Factor w/ 5 levels "","Both/Mainly Hand-
Rolled",..: 1 5 1 1 1 1 5 1 4 5 ...
```

- **gender**: categorical, nominal
- **age**: numerical, discrete
- **maritalStatus**: categorical, nominal
- **highestQualification**: categorical, ordinal
- **nationality**: categorical, nominal
- **ethnicity**: categorical, nominal
- **grossIncome**: categorical, ordinal
- **region**: categorical, nominal
- **smoke**: categorical, nominal
- **amtWeekends**: numerical,discrete
- **amtWeekdays**: numerical, discrete
- **type**: categorical, nominal

Problem 2

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. **Because there is no specification of a particular type of individual, we can assume that the**

[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

population is all of humankind. The sample is the 160 selected 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 results of this study should not be generalized to the population because there are differences across the children's characteristics within each group. If one group appears to be less honest than the other, it could be a result of these unaccounted (confounding) factors instead of the explicit instruction. In order for the results of this experiment to be more generalizable, it would be appropriate to account for these factors within the study design. Although this is an experiment and not solely observational (because the researchers are collecting data utilizing an explanatory and response variable), we should not use the findings to establish a casual relationship because of the reasons listed above. If the experiment were altered to account for the differences in children's characteristics, then yes, we could use it to establish a causal relationship.**

Problem 3

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.

No, we cannot conclude that smoking causes dementia later in life. This is because this is an observational study, where the researchers collected data in a way that does not directly interfere with how the data arises. In observational studies, we typically cannot assume a causal relationship.

(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 cannot be justified for a few reasons. First, as discussed above, observational studies can only provide evidence of an association between things, not a causal relationship. Second, it is entirely possible that the sleep disorder is a symptom of something else entirely, and this may be the actual root of the bullying. Third, as with any survey, we could potentially be dealing with classifications that are subjective. For example, one teacher might view a behavior as indicative of bullying and another does not. Despite these limitations, we can conclude with reasonable certainty that there is an association between sleep disorders and bullying. Further research taking additional factors into account might be able to provide some additional insight.

Problem 4

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 an experiment using stratified sampling.**

(b) What are the treatment and control groups in this study? **The control group is the one instructed to rest and not exercise. The treatment group is the one instructed to exercise twice a week.**

(c) Does this study make use of blocking? If so, what is the blocking variable? **No, this study does not make use of blocking. However, both groups are split into strata and randomly assigned so that the sample is more adequately representative of the population.**

(d) Does this study make use of blinding? **This study does not make use of blinding because the groups know that they are either exercising or not exercising.**

(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. **In general, yes, this study can be used to establish a causal relationship between exercise and mental health and also generalised to the population at large. Because the study made**

sure to stratify the population, it is more representative of the population at large. However, because there was no blocking (ex: individuals that currently have lower self esteem should be split evenly between the control and treatment groups), we might have an unbalanced representative of each between groups.

(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? **The biggest reservation I would have is that there is no mention of requiring the participants to document their exercises. It's entirely possible that the treatment group lies and says that they did exercise when they actually didn't. I would also be concerned that the individuals in the treatment group are not performing the same level of exercise. Is it adequate to compare someone that works out at a high level intensity vs someone that works out at a low level?**