Data606 Homework 1

Joseph Simone

September 1, 2019

Smoking Habits of UK Residents (1.10, p.20)

A. What does each row of the data matrix represent ?

```
##   gender age maritalStatus highestQualification nationality ethnicity
## 1   Male  38      Divorced     No Qualification     British     White
## 2 Female  42        Single     No Qualification     British     White
## 3   Male  40       Married               Degree     English     White
## 4 Female  40       Married               Degree     English     White
## 5 Female  39       Married         GCSE/O Level     British     White
## 6 Female  37       Married         GCSE/O Level     British     White
##        grossIncome    region smoke amtWeekends amtWeekdays    type
## 1   2,600 to 5,200 The North    No          NA          NA
## 2      Under 2,600 The North   Yes          12          12 Packets
## 3 28,600 to 36,400 The North    No          NA          NA
## 4 10,400 to 15,600 The North    No          NA          NA
## 5   2,600 to 5,200 The North    No          NA          NA
## 6 15,600 to 20,800 The North    No          NA          NA
```

Solution: Each row of the matrix represents one observation. In this dataframe, this is a single resident of the UK.

B. How many participants were included in the survey ?

```
## [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.

```
## '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 ...
```

Solution: 1691 particants in this survey.

Cheater, scope of inference (1.14, p.29)

Exercise 1.5 introduces a study where researchers studying the rela- tionship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15. 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.

Solution: The population of interest are children between the ages of 5 to 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.

Solution: This is an experiment, the findings of this experiment are used to establish causal relationships.

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:

"25 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.

Solution: There is no treatment or control group, therefore this is an observational study. Casual relationships between variables cannot be defined in an observational study. Even though, the numbers show a correlation between smoking and dementia, we are not defining all the external factors can can be contributing to these cases.

(b) Another article titled The School Bully Is Sleepy states the following:

"26 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 identifed 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 identi

ed 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 justifed? If not, how best can you describe the conclusion that can be drawn from this study?

Solution: This is also an observational study. We cannot say for certain if the sleep disorders in these cases would lead to 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 stratifed random sampling to ensure representative 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?

Solution: This is a Prospective experimental Study

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

Solution: Treatment Group: exercise 2x's a week Control Group: no exercise

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

Solution: This study does make use of blinding, the age of the groups is the blocking variable.

(d) Does this study make use of blinding?

Solution: This study does not make use of blinding, patients and doctors are completely aware of the groups.

(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.

Solution: Conclusions can be drawn from the population since this experiment was randomized.

(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?

Solution: I would not fun this project on the basis that cluster sampling would be more benefical to the study.