Chapter 1 - Introduction to Data

Joshua Registe

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?

Row 1 "sex" represents the gender of the UK resident that was survey. Row 2 "age" represents the age of the UK resident that was surveyed. Row 3 "marital" represents the marital status of the individual that was surveyed. Row 4 "grossIncome" represents an income bracket that the individual fell in to. Row 5 "smoke" represents whether or not this individual was a smoker. Row 6 "amtWeekends" represents the frequency of cigerattes smoked per day on a weekend. Row 7 "amtWeekdays" represents the frequency of cigerattes smoked per day on a weekday.

(b) How many participants were included in the survey?

The table indicates that 1691 people were included in this 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.

Row 1 "Sex" represents a categorical nominal variable. Row 2 "age" represents a numerical discreet variable. Row 3 "marital" represent a categorical nominal variable. Row 4 "gross income" represents a categorical ordinal variable. Row 5 "Smoke" represents a categorical nominal variable. Row 6 "amtWeekends" represents a categorical ordinal value. Row 7 "amtWeekdays" represents a categorical ordinal value.

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.

The population of interest in this study are Children between the age 5 and 15. and the sample is 160 children.

(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.

Results of the study can be generalized to the population only if the results can be replicated. Currently the results are only applicable to this sample, but if a hypothesis testing was performed and experimental replications showed a distribution of results that implies significance within our experiment, then the results can be generalized to the population, blocking may also be considered the experiment maybe by gender or seeing if the differences accross children characteristics can be grouped. The findings of this study can be used to establish causal relationships if the experiemental design was based on random assignment, causal relationship can be assigned to the sample if random assignment but no random sampling was done, and causal relationship can be assigned to the population if both random sampling and random assignment was done.

---

[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.

Based on the results of this survey, We cannot conclude that smoking causes dementia later in life, there are likely people in that group who smoked 1+ packs of cigerattes a day but did not develop an memory issues. I agree that there may be a correlation between smoking and dementia but memory loss can potentially be simply a product of genetics or something else. That said, the likelihood of dementia likely increases the more you smoke.

(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?

The study states that children who had behavioral issues were twice as likely to have shown symptoms of sleep disorders. I would emphasize the statement "twice as likely". a term like this does not imply causality, it simply states that there is a greater probability that that studet may have sleeping disorders, but that means theres still exists a probability that the student does not show symptoms of sleeping disorders and so you cannot conclude that sleeping disorders lead to bullying. The hypothesis here is also that bullies may have sleeping disorders and not that sleeping disorders result in 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 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 a stratified sampling experiement with random assignment.

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

The treatment group are those that were randomly selected from each age group and told to exercise. the control groups are the rest of the population who were told not to exercise.

(c) Does this study make use of blocking? If so, what is the blocking variable? This study makes use of blocking by stratifying the sample group by age groups. This eliminates sampling biases of younger vs older individual exercising habits.

(d) Does this study make use of blinding? This study does not make use of blinding as all individuals are aware of whether they have been told to exercise or told not to exercise.

(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.

The results of the study may be used to establish a causal relationship as both random sampling and random assignment were employed. this also allows the conclusions to 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?

Some reservations i would have when considering this study is how large our sample size is (large enough to be significant?), how well can we control for certain variables such as: Location - Temperatures may affect how often people exercise, Gender, demographic, etc. Then it would be necessary to quantify just how accesible this data will be to collect and if this may result in biases from convenience sampling.