# 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?

Each row represents an individual person, or a participant.

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

There were 1,691 participants 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.

See the following table where the first column "Variable" is the variable name, the second column "Identity" identifies the variable as categorical or numerical, and the last column "Type" describe the categorical variables as ordinal or nominal and numerical variables as discrete or continuous.

| *Variable* | *Identity* | *Type* |
|---|---|---|
| Sex | Categorical | Nominal |
| Age | Numerical | Discrete |
| Marital | Categorical | Nominal |
| GrossIncome | Categorical | Ordinal |
| Smoke | Categorical | Numerical |
| amtWeekends | Numerical | Discrete |
| amtWeekdays | Numerical | Discrete |

The variables "GrossIncome," "amtWeekends," and "amtWeekdays" were interesting to identify. In the data

GrossIncome appears to be organized into economic brackets similar to classifying income into low, middle, and high categories. For this reason, I believe it must be considered a ordinal categorical value because there is a natural order to low, middle, and high income brackets. Variables "amtWeekends" and "amtWeekdays" were interesting because there is the potential for a measurement of less than one cigarette. However, the data only displayed positive whole values and it would be quite difficult to measure a less than a whole cigarette. I would also expect the result to have jumps from one value to the next, regardless of how it was measured. For this reason, I believe both are discrete numerical values.

---

**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 is 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.

Assuming the kids were a representative sample of the population then the results of this study could be applied to the general population of children aged 5 to 15. Causal relationships can be inferred from randomized experiments, such as this one.

Their use of two groups of randomly selected children, one with a treatment and one without, introduced a control group that was used to compare and validate the results of the treatment. This control group allowed the researchers to isolate the treatment and make a reasonable conclusion that girls, especially as they age, cheat less than boys when they are explicitly told not to cheat. Meanwhile, the rate of cheating for boys at all ages did not vary. The response in this experiment indicates that the likelihood for girls aged 5 to 15 to cheat for a reward is dependent on whether on or not they are explicitly told not to cheat while it does not vary for boys of the same age group.

---

[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.

Since the researchers are not directly interfering with the data and are using medical records to show only a correlation in the data, this is an observational study using observational data. Thus, we cannot conclude that smoking causes dementia later in life because observational studies merely observe the relationship between variables. This is correlation without causation.

(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 is another example of an observational study and as such, it can only indicate a correlation, not causation. There is no treatment to imply causation is occurring. My friend's statement is not justified because he is implying causation from a naturally occurring association between two variables without conducting an experiment.

---

**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 experimental study with stratified random sampling.

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

The treatment groups are the individuals in each age group that were instructed to exercise twice a week. Exercise is the treatment. In this case, the control groups are the other halves of each age group that were instructed not to exercise.

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

Based on the need for blocking to have a variable to separate indiviudals into distinct groups, I would say this study does not make use of blocking. Stratified random sampling is not equivalent to blocking as the former separates the population to create samples of the population called strata. This is more of a divide and conquer method to sample the population at large using similar characteristics. The latter, blocking, pulls from the sample any groups that might effect the experiment and divides them into randomized treatment groups to ensure equal representation in the study.

(d) Does this study make use of blinding?

This study does not make use of blinding. Both researcher and the subjects are aware of the treatment and control groups.

(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.

Provided there is a large enough sample to be representative of the population, individuals are randomly selected (not volunteers), and there is a best attempt at controlling variables to ensure any response in the treatment group comes from the treatment only, then the results of this study could be used to establish a causal relationship between exercise and mental health. It is an experimental study with stratified random sampling and as such can establish a causal relationship that 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?

Ideally, this experiment would be double-blind to ensure both researcher and subjects are not inadvertently introducing thier own biases. Without the use of blinding this method could cause subjects' mental health to be influenced by those in the treatment group, control group, or from the researcher administering the mental health exam. This shows a lack of controlling in their variables and it will make it more difficult to generalize to the population. The researcher should check for any other underlying variables that could influence the results (such as pre-existing mental health conditions and confounding variables) after controlling for the placebo effect. Another reservation in determining if this proposed study should get funding is that I am also left to assume they know the sample size necessary to produce repeatable results.