Chapter 3: Producing Data

Introduction

When we do exploratory data analysis, the graphical representation of distributions is important. From these graphs, we look for patterns that help us interpret our data. Exploratory data analysis is not enough though; patterns in the data can have many different causes, as it is our job to find out what those causes are. The validity of the conclusions that we draw out of data analysis depend not only on using appropriate analytical methods, but also on the quality of the data.

3.1: Sources of Data

Anecdotal Data and Available Data

Anecdotal data is data that is based on our own experiences. This kind of data often represents shocking events or events out of the norm. For example, a psychologist might have one case in which a clinically depressed patient instantly feels better after eating a chocolate bar. While it is tempting to draw conclusions from this singular event, the data was not obtained through experimental methods and we cannot conclude anything from this one case. In short - the plural of anecdote is not data.

Sometimes, it is possible to use data from one experiment to draw conclusions about other research questions. This data is called *available data*.

Samples

Often times, researchers are interested in how the population looks at certain things. For example, what Americans think about abortion, or how people feel about the packaging of a particular product. In these cases, *sample* surveys are given to a random group of people. *Sampling* means that we study a part of a population to draw conclusions about the entire population.

A *census,* on the other hand, is when researchers try to poll every individual in the population of interest. Investigators prefer samples above a census because a census is not efficient. It has also been shown that a well-executed sampling procedure produces more precise results than a census. This is due to the fact that it is tedious to collect data from a large number of people, and it is often easier to make mistakes when forced to do a tedious task for a long period of time.

Studying samples is one type of *observational study*. Observational studies are studies in which individuals are observed and variables are measured. There is no intervention and the experimenter does not have an effect on the reactions of the individuals. In contrast, an experiment is a study in which an intervention is carried out intentionally in order to see how people respond. Experiments are often preferred to observational studies, because we have more control over the variables in experiments.

3.2 Experimental Design

Important Terms

- The individuals that we use for an experiment are called *experimental units*.
- When these units are people, we call them *subjects*.
- A specific experimental condition that is applied on the experimental units is called a *treatment.*
- The distinction between explanatory and response variables for experimentation is important because we want to establish causality. Often, this will succeed only with real experiments. The explanatory variables are called *factors*. Often times, studies look at the combined influence of several factors. In such an experiment, each treatment is formed by combining specific values or quantities of the factors. These specific values are referred to as *levels*.

Comparative Experiments

In many laboratory experiments in science and engineering, only one intervention is carried out at a time. This intervention is then applied to all experimental units. Such a set-up is summarized as follows:

Treatment → Observed response.

In the case of experiments done on living organisms, more complex designs are more convenient. This is to ensure that the observed responses are the result of the treatment and does not, for example, a lurking variable.

- In medical experiments, the *placebo effect* plays an important role in the validity of the experiments. Simply taking a pill, even if the pill does not contain any of the active ingredients being researched, often influences the behaviours of the test subjects in the *placebo group*.
- A control can be used to see if an *intervention* actually leads to specific results. The control group will receive no intervention, while other groups will. If it turns out that the groups who receive an intervention (intervention group) have different scores than those who don't (control group,) then the observed

difference in effect is most likely due to the intervention itself and not by other variables.

- A study is *biased* if it systematically favours particular outcomes.

Randomization

The design of an experiment describes the response variable, the factors (independent variables) and how the experiment was set up. Making comparisons between groups (and discovering differences) is the most important part for researchers. A second aspect of an experiment is about how participants are assigned to conditions. This can be done in various manners, for example, splitting subjects into groups based on sex, age, health, etc. to match with each other. However, this is not entirely adequate because the researcher cannot examine everyone in advance and may miss certain lurking variables that may affect the results of the study. Therefore, subjects are often randomly assigned to participant groups, so that the research can really reflect the results of the intervention, rather than the differences displayed within the groups before the study. *Randomization* is the use of coincidence in order to share in experimental units in groups.

Principles of Experimental Designs

The main principles of experimental designs are:

- Comparison: Compare two or more treatments with one another. This ensures that the effects of lurking variables are kept under control.
- Randomization: Use chance to assign experimental units to treatment groups.
- Repetition: The repetition of each treatment on many different experimental units limits the variation in the study results.

How to Randomise

Often times, researchers will use software to randomise, but it is also possible to do it manually. To do this, a table of random numbers is used. A table of random numbers is a list of the digits 0,1,2,3,4,5,6,7,8,9. These numbers:

- All have the same chance of being selected.
- Are independently of each other, because one value has no effect on a different value.

An experimental design is completely randomized if all experimental units are randomly assigned to all conditions. With such a design, quite a few treatments can becompared with each other.

Double-blind Design

A study is *double-blind* if both the subjects themselves and researchers do not know which treatments have been assigned to whom. Such a design ensures that the expectations of researchers have no influence on their interpretations of the results, and that the researcher examines each subject in the same way. A disadvantage of experiments is the lack of realism. In that case, the test subjects, the handling or the setting of an experiment are not good representations of the conditions that researchers actually want to examine. Many researchers want to generalize their findings to a setting that is different from the setting in which the experiment is done, in order to make it more applicable tot he real world. It is important to remember that statistical analysis of an experiment cannot tell us how well the study results generalise to other settings.

Matched Pair Design

In this design, two treatments are compared in subjects that are matched based on particular characteristics. For example, if you wanted to investigate the effect a new package colour has on consumers, you might match subjects that have the same income, sex, and shopping habits. This way, subjects in each pair are similar to each other than unmatched subjects. The differences in their responses can then be observed and recorded and further analysed.

Block Design

In this design, researchers make use of so-called *blocks*. A block is a group of experimental units or subjects that are similar to each other. In a block design, the random assignment of experimental units of treatments done separately for each block. You can, for example, split men and women into two blocks. Next, within each block, different treatments can be assigned. The results within each group can then be compared

3.3: Sampling Design

Important Terms

- The whole group of individuals we want to know about is called a *population*.
- A *sample* is a part of the population. This is the part we investigate to gather information. We can use this information to draw conclusions about the population as a whole.
- The proportion of the population who give us useable data is called the *response rate.*
- A *voluntary response sample* consists of people who choose to participate in a

survey. These kinds of samples are biased because people with strong opinions tend to respond more frequently.

Sampling Designs

In order to draw correct conclusions, it is important to apply randomisation techniques in the selection of samples.

A *probability sample* is a sample that is selected on the basis of random phenomena. We need to know which samples are possible and what chance each sample is associated with. A probability sample can be *simple random* or *stratified*.
A simple random sample (SRS) is a sample where study participants have an equal chance of being selected from the population.

A stratified random sample is often used when there is an investigation of a large population. SRS is often not adequate enough. In order to attract a stratified random sampling the population must first be divided into groups of similar individuals. These groups are called *strata*. Then, separately for each *stratum* a SRS is done. The sum of the SRSs make up the full stratified random sample.

A *multistage random sample* is selected in stages. This design is often used for national surveys of people or households. This process involves three stages:

- Divide the total population into groups based on a specific criterion. These groups are called *primary sampling units*. Select a random sample of PSU's.
- Divide each PSU further into *blocks.* Use stratified sampling within each block to get a stratified sample from each block.
- Sort the individuals in each block into smaller groups called *clusters* and do a probability sample on each of these clusters.

Things to Look Out For in Sampling

- *Undercoverage* when some groups of the population are not systematically involved in a sample. One example is that a person carries out an investigation by calling people. In America, however, 6% of people do not have a phone. Such research can also lead to misleading results.
- *Nonresponse* is when an individual who is selected for a sample does not complete the experiment or is unable to be contacted
- *Response bias* is when participants provide false information about sensitive issues, such as drug use or stealing. This is often because no one wants to admit undesirable behavior.

- The *wording* of certain questions can also have a large effect on the responses from a sample. Confusing questions, for example, are huge sources of bias because participants may not be answering what the question intended to ask.

3.4: Statistical Inference

Statistical inference is using facts about a sample to draw conclusions or make predictions about a population.

- A *parameter* is a number that describes the population. A parameter is a set number, but we do not actually know its value. For example, it is impossible to know exactly how many Dutch people are against abortion.
- A *statistic* is a number that describes a sample. The value of a statistic known after we have selected a sample, but this value can differ per sample as well. We often use a statistic to estimate an unknown parameter.

Sampling Variability

Sampling variability means that the value of a statistic per sample will be different. Random samples remove bias by selecting a sample based on randomness; however, it appears that the selection of many random samples (of the same size and from the same population) the variation between samples follow a predictable pattern. Statistical inferences are based on the idea that the reliability of sampling depends on the repetition of procedures.

In practice, it is too expensive to run an unlimited number of trials; but we are able to imitate taking many random samples by using *simulation.*

Sampling Distribution

The *sampling distribution* of a statistic is the distribution of all the values that adopt the statistic in all possible samples of the same size and from the same population. If this distribution is plotted on a histogram, it appears that:

- The histogram has a normal distribution
- The histogram also shows that the means and medians are roughly the same.
- In practice, it appears that values from samples of considerable size (e.g.,> 2500) have much less spread than the values of smaller samples (e.g., having a size of 100). This is because larger samples are more representative of the population than smaller samples.

These three facts are true whenever we use random sampling.

Bias and Variability

- A statistic that describes a parameter is *unbiased* if the average of the corresponding samples of distribution is equal to the true value of the estimated parameter. Bias is reduced by using random sampling techniques.
- The *variability* of a statistic is described by the spread of its sampling distribution. This spread is determined by the design and the sample size of the sample (n). Variability is reduced by having larger sample sizes.
- The *margin of error* is a measure of the spread of a sampling distribution
- Low bias can coexist with great variability and low variability may be associated with a lot of bias. In a good research, there is little spread and minimal bias.

3.5: Ethics

Researchers can be confronted with ethical dilemmas when trying to collect data. This is especially common in experiments because it is always accompanied by an intervention. The following are some basic rules for conducting ethical research:

- *Institutional review board*s should be involved in approving studies before data collection begins
- All participants who want to participate in a study must give *informed consent.* This means that participants are fully informed about the procedures that will happen during the experiment, and consent must be obtained before the study starts.
- Individual data must remain *confidential*. Only statistical information about groups or individuals may be disclosed.