7. Inference from a sample

A common type of argument is one where the conclusion is a generalization based on a *sample* of cases. Here are some examples:

> A random sample of 50 grommets was selected from the production line of this factory and ten of them were found to be faulty. So it is likely that about 20% of the grommets made at this factory are faulty.
>
> The issues that are most important to the electorate have changed since the last election. Previously the most important issue was keeping living costs down, while today the most important issue is honesty in government. (Results based on a survey of 308 eligible voters).
>
> Nobody has ever discovered an insect species that lives on the open ocean. So probably, insects are unable to live in sea water.
>
> Nguyen is not a very sociable person. Every time I've met him, he barely spoke to me at all.

As these examples suggest, arguments of this type are used in a wide variety of contexts, from quality control, opinion polls and scientific evidence, to everyday reasoning about a person's personality. Given their ubiquity, it is worth spending some time discussing how to evaluate these arguments. That is the topic of this chapter.

Sample, Population and Target Property

We begin by introducing three terms that are very useful for describing and evaluating arguments based on samples.

SAMPLE. The group of individuals or cases actually examined is called the **sample.**

POPULATION. The entire group of individuals or cases the argument attempts to draw a conclusion about is the **population**.

TARGET PROPERTY. The **target property** is the feature or property of the individuals or cases which the argument attempts to generalise about.

In the first example argument above, the sample consists of the 50 grommets that were randomly selected, the population consists of all the grommets made at that particular factory (that's what the argument draws a conclusion about) and the target property is whether a grommet is faulty or not.

In the second argument, the *sample* consists of the 308 voters who were surveyed, but what is the population? The conclusion of the argument is that there has been a change in the opinions of *the electorate*. The electorate consists of all current eligible voters in the country, so that is the population in this case. The target property for this argument would be the issues which are most important to a voter in the election, since this is the feature of the individuals in the population the argument is attempting to quantify.

The third argument draws a conclusion about *insects in general* – that they cannot live in sea water. So the population for this argument would consist of all the insects. The sample consists of all the insects ever observed or identified. The target property is 'ability to live in sea water'. The argument tells us that no insects in the sample have this property, and generalises from this to the conclusion that no insects at all have that property.

In the final example, the sample consists of all the occasions on which the speaker has interacted with Nguyen. The population might be thought of as all social situations involving Nguyen while the target property would be sociability.

A general pattern for inference from a sample

All arguments which generalize from a sample can be thought of as fitting the following pattern:

1. X% of cases in a sample have the target property.
2. The sample is representative of the population.
Therefore:
C. It is likely that X% of the population have the target property.

A special case is where X = 100. In that case the argument becomes 'All the cases in the sample have the target property; the sample is representative of the population, therefore, it is likely that *all* members of the population have that property'. You might be familiar with examples such as 'Every swan observed so far has white feathers. Therefore, *all* swans are white'.

This example is often used to make the point that arguments which generalise from a sample are not deductively valid. No matter how many swans you have examined, it is always *possible* that the next one will be black or red, not white. Nevertheless, a representative sample from a population can provided *very strong inductive support* for a generalisation about the whole population.

The second premise – that the sample is *representative* of the population – is often left unstated in arguments like this (as it is in all the examples we began with), but clearly, if the sample is *not* representative of the population, the conclusion will not be supported. To evaluate an argument of this kind then, a key question will be whether the sample is representative or 'typical' of the relevant population.

For example, in the second example argument, we should ask whether the 308 voters surveyed were typical of the population of voters as a whole. Did the sample consist of roughly equal numbers of men and women? Was there a representative spread of incomes and age groups? In the final example, we might wonder if the occasions on which the speaker encountered Nguyen were normal or typical. Perhaps on those occasions Nguyen was behaving out of character for some reason; perhaps he was unusually tired or anxious.

How can we ensure that samples are representative or typical? How can samples fail to be representative? It is to this fascinating and important topic that we now turn.

Sample size

Everyone knows (or should know) that generally speaking, a large sample of individuals is better than a small sample if you want to draw accurate conclusions. Consider the following examples:

> Cycling can't be very effective exercise: my friend Gabriel rides his bike all the time and he is quite overweight.
>
> My grandfather smoked all his life and never had a days' illness. So cigarettes can't be that bad for you.

> Don’t marry her – she’s a brain surgeon! I knew a brain surgeon once and she was one of the most rude and obnoxious people I ever met.

In all these cases, a general conclusion is derived on the basis of just a *single* example – a sample size of one. It ought to go without saying that samples like this (or samples consisting of just two or three cases) are almost always too small to be representative. And yet, it is surprising just how common arguments like this are. Even people well trained in thinking about evidence are prone to this kind of reasoning, though it is often not explicitly stated. We will see a possible psychological explanation for this tendency to generalise from a few striking examples below when we discuss the concept of ‘availability’.

The sample size principle

Everyone knows that a large sample is better than a small sample, but *why* are larger samples better? A useful way to think about this is in terms of the following principle:

The sample size principle

The smaller the sample, the **more** likely it is that you get an atypical or unrepresentative result.

A larger sample is **less** likely to give you an atypical or unrepresentative result.

Suppose you have a very large barrel containing chocolates; some are milk chocolate and some are dark chocolate. You would like to know what proportion of each kind you have in your barrel without having to count them all, so you plan to scoop out a sample of chocolates from the barrel and count them. Let’s assume that unknown to you, exactly half the chocolates in the barrel are milk chocolate and half are dark chocolate.

Suppose you have scoop that always gives you a sample of just four chocolates. Imagine yourself taking *repeated samples* from your barrel. You scoop out four chocolates, count them, then tip them back in. Then you repeat the process again and again, each time scooping out just four chocolates. Obviously you will not always get the same number of milk chocolates in your sample. You could get any number from zero to four. A sample containing four milk chocolates would *not* be representative (remember we’re assuming that as a matter of fact, half the chocolates in the barrel are milk chocolates) but it will *sometimes* happen.

Now imagine, you have a larger scoop; one that always gives you a sample of eight chocolates. And again imagine taking repeated samples from the barrel using this scoop. Again, you will not always get the same number of milk chocolates. This time you might get anything from zero to eight. A sample containing eight milk chocolates would not be representative, but it will *sometimes* happen.

The key point is this: an atypical or unrepresentative sample containing all milk chocolates and no dark chocolates will occur *less* often with the larger scoop. That is what it means to say that a larger sample is ‘better’ or ‘more accurate’. If you took repeated samples, a larger sample will give you atypical results less often.

Of course, this applies not just to the atypical result of getting all milk chocolates in your sample. It also applies to the atypical result of getting *no* milk chocolates, or just one. Those atypical results will also occur less often with the sample size of eight than with the sample size of four. And a sample size of 10 or 20 will give atypical results even less often.

If you're not convinced, try the experiment for yourself with your own bag of equal numbers of two different kinds of chocolate. Then eat the chocolates while repeating to yourself the following mantra: "In a small sample, atypical results will occur more often".

The sample size principle in action

Awareness of the sample size principle can be very useful when critically evaluating claims and evidence of many different kinds. Consider the following story.

There are two hospitals, one a large city hospital and the other a small hospital in a rural area. Both hospitals perform a certain surgical operation, which has a national average success rate of 80%. At the end of every month, each hospital records their proportion of successful operations of this kind. Obviously the exact proportion of successful operations each hospital reports will vary from month to month, but policies dictate that if the proportion of successful operations in a given month is less than 50% (considerably less than the national average) the hospital must report it. Over the last six months, the small rural hospital filed *four* such reports, while the larger hospital filed none. Health officials conclude that there must be a problem with the way the rural hospital is conducting this operation and launch an investigation.

Knowing about the sample size principle, you are not so sure. The smaller hospital will carry out fewer operations each month than the larger hopsital. So the smaller hospital's monthly success rate is based on a smaller sample of operations. *Atypical results will occur more frequently with a smaller sample.* So a success rate of less than 50% will occur more often for the smaller hospital than for the larger one. In other words, the four reports filed by the smaller hospital are what you might expect to see just due to random variation and the smaller monthly sample.

A question to think about: which hospital will more frequently have a monthly success rate of 100% for the operation?

Random samples and margin of error

The above remarks about the sample size principle are only really valid when the sample consists of an *unbiased* or *random* selection from the population. If you had a very special scoop that could only pick up milk chocolates for example, then the atypical result of all milk chocolates in the sample would occur all the time, no matter how large your sample.

A *random* sample is one in which every member of the population has an equal chance of getting into the sample. In many cases of course, it is not practical to arrange things so that every member of the population is equally likely to be sampled. Nonetheless, if the procedure for selecting the sample is carefully designed it is possible for samples to approximate the ideal of a truly random sample and therefore give fairly reliable results when generalising from the sample.

If the sample is random, or close to random, it is possible to give a quite specific answer to the question 'how large does the sample need to be for reliable estimation'? If the sample is random, you can use statistical theory to calculate a **margin of error**. The margin of error depends on the sample size and tells you how much error you can expect from random sampling variation. For example, consider the quality control example:

> A random sample of 50 grommets was selected from the production line of this factory and ten of them were found to be faulty. So it is likely that about 20% of the grommets made at this factory are faulty.

Assuming the sampling procedure was really random, the margin of error for this sample size can be calculated and is approximately 5% at a 95% confidence level. What does that mean? It means that if you took repeated random samples of size 50 from the population, then 95% of the time, the sample proportion would be within 5% of the true proportion. So in the above example, we can be fairly confident (95% confident if you like) that the true proportion of faulty grommets at the factor is somewhere between 15% and 25%. And that might be accurate enough for practical purposes. If you wanted to hone in on the true proportion more accurately, say with a margin of error of just 2% you would need a sample size of about 1,540 grommets.

The details of how this calculation is done need not concern us here, though it is not complicated. It is important to understand what the term 'margin of error' *means* though, since you will often come across it when encountering arguments based on samples. Organisations that conduct opinion polls for example, like to report the margin of error of the sample size. The crucial thing to remember is that the 'margin of error' only accounts for *one* source of error in a sample. It is a way of quantifying the variation you would expect to get if you took repeated random samples of that size. However, in opinion polls and other arguments based on non-random samples, there are many *other* sources of error that are likely to be far more significant than sampling variability. The 'margin of error' does not take those sources of error into account.

The size of the population does not matter

Before leaving the topic of random samples, it is worth pointing out a very common misconception about sampling. People often think that the reliability of a sample depends on the size of the population. For example, when asked how large a sample of people you would need to accurately determine something about a population, a common answer is 'you need a sample of about 10% of the population'. Not so. In the example above, the margin of error is the same, no matter how many grommets the factory produces every day. With random samples you do not need a larger sample for a larger population. A random sample of 50 individuals from a population of 2 thousand is just as accurate as a random sample of 50 individuals from a population of 2 million, or 2 billion. This is counter intuitive, but true nonetheless as any statistics textbook will tell you.

To make this plausible, imagine a huge vat full of equal numbers of red and blue marbles. The vat is thoroughly mixed. You have a small cup you can dip in to take a random sample (random because the marbles are so thoroughly mixed together). Obviously if you have a bigger cup, you can get a larger sample, which would be more reliable: the number of red and blue marbles you get in your cup will be equal more often than with the smaller cup. But how often you will get equal numbers of red and blue marbles depends only on the size of the cup, not the size of the vat. The cup does not 'care' how many marbles there are in the vat: it doesn't matter if in fact we are taking a sample from an enormous swimming pool with millions of marbles in it. Provided the marbles are always thoroughly mixed together (which is just to assume that the sample is always random) it makes no difference how many there are in total.

The general principle is that the reliability of a sample depends on the sample size, but not on the size of the population. The exception to this is when the population size and the sample size are close to each other. For example, if I have 60 students in my class and I take a random sample of 50 of them, I would have sampled 83% of the entire population. This would be a very reliable sample indeed but slightly less reliable than a sample of the same size from a class of 100 students. However, once the size of the population becomes significantly larger than the sample size (more than 100 times larger is an often used rule of thumb) it begins to make no difference at all what the total size of the population is.

Biased samples

If the sample is not random (and usually it will not be) you must use your background knowledge to decide whether the sample is large enough. In general, the more variation in the target property there is likely to be in the population, the larger the sample needs to be.

With non-random samples though, sample size is unlikely to be the main issue. Much more important will be issues to do with the way in which the sample was selected. Even very large samples can be highly inaccurate if they are selected in a way that is likely to introduce a bias into the sample. Here is a very famous example.

In 1936, the U.S. magazine *Literary Digest* conducted a poll to predict the winner of the forthcoming presidential election. Questionnaires were sent out to people selected at random from the Digest's subscription list and from the telephone directory. Two million people responded. The poll predicted that Republican Alf Landon would win by a significant majority. In fact, Franklin D. Roosevelt won by a landslide.

A sample of 2 million people is enormous. A *random* sample of that size would have a very high degree of accuracy. So why did this poll, despite the huge sample, get things so wrong? The problem was to do with the way the sample was selected: it certainly was not random or even close and was almost certain to be biased. 1936 was the height of the great depression in the United States; a time of great financial hardship and unemployment for a significant proportion of the population. You can probably see for yourself how the process the magazine used to select its sample was likely to exclude the less well-off members of the society. So despite its great size, this sample was not representative of the population as a whole.

This is just one example of bias introduced into a sample because of the specific way the sample was selected. We turn now to two very common sources of selection bias in samples; self-selection and availability samples.

Self-selected samples

Sometimes people select themselves to be part of a sample. For example, newspapers and television news programs like to conduct call-in polls of public opinion. The newspaper or program publishes a question and asks viewers to call one number (or click a button on a website) to respond 'yes' and another for 'no'. Polls of this kind are often quite large. It is not unusual for example to get responses from several thousand people. Nevertheless, such polls are very unreliable because the sample is almost certain to be heavily biased.

Why? Because people who make the effort to respond to the question are probably not representative of the general population. People with strong opinions about the issue or people who are very angry about it are far more likely to respond to the question than others. So the sample will be biased.

Samples of this kind are called *self-selected* or *voluntary response* samples.

> A **self-selected sample** consists of people who choose themselves to be part of the sample by responding to a general appeal.
>
> In opinion polls, self-selected samples are likely to biased because people with strong opinions are more likely than others to respond.

Properly conducted opinion poll or surveys are not self-selected in this way. The organisation conducting the poll contacts people and asks them to complete the survey. Reputable organisations that conduct surveys try to obtain a representative sample by ensuring that their sample contains representative proportions of people of different ages, levels of income, gender and all other variables they think might be relevant. Surveys like this, even with a modest sample size (about 1,500 is typical) can be quite accurate. Even so, people can refuse to complete the survey and this can create **non-response** bias. If the people who do not agree to take part are dissimilar in some relevant way to those who do, the sample can become biased.

Although they can introduce a bias, self-selected samples are quite common in scientific research. Many studies in fields such as psychology, education, economics and even medicine use samples of people who have volunteered to take part in the study. This might or might not be a problem, depending on the target property. Consider for example a psychologist investigating the relationship between personality types (as measured by a questionnaire) and professions. She might publicise her study and ask for volunteers to complete the questionnaire and indicate their profession. Even if she obtains a large sample, it would probably be biased because people of certain personality types could be more willing than others to volunteer their time to take part in such a study.

Contrast this with a medical study to test the effectiveness of a new treatment for a certain disease. By necessity, such a study will consist of volunteers. But if there's no reason to think that willingness volunteer is in any way related to the process of the disease, this would not be a problem.

Availability samples

A large proportion of experiments in psychology are carried out on psychology students. Why? Because most psychologists work in universities and so their students are readily available source of experimental subjects. But of course, psychology students might in various ways be unrepresentative of the general population of all human beings. A medical researcher might look for volunteers for her studies from the patients she sees in her clinic. Again, the patients attending her clinic might not be typical in various ways. A chemist studies almost entirely samples of chemical elements found on the planet earth; extraterrestrial samples are hard to come by. Physical theory assures us that the elements are the same everywhere in the universe, but is at least conceivable that this is not always the case.

These are all examples of *availability samples*. Another name for them is *convenience samples*.

> An **availability sample** is one in which individuals are selected because they are easily accessible, rather than through a random process.

The majority of samples used in both scientific research and everyday reasoning are availability samples, rather than truly random samples. Availability samples can be biased because the *available* individuals may not be representative of the population as a whole.

A particular kind of availability sample is one that consists of cases that come easily to mind; psychologically available cases you might say. Reasoning based on such samples is probably the most common source of error in everyday inductive reasoning. Consider these examples:

> Your friend Gabriel is the only one of your close friends who rides a bike and he is quite overweight. This case therefore readily springs to mind whenever you hear the claim that cycling has health benefits.
>
> You have very fond memories of your grandfather, who smoked heavily for most of

> his life. Since he never got sick, you conclude that cigarettes can't be all that dangerous.
>
> You have vivid memories of a really unpleasant person you once met who was a brain surgeon. As a result, you try to convince your best friend not to marry one.

In all these cases, you make a judgement based on a case that easily comes to mind. There is a well-document psychological tendency for people to confuse how easily certain cases come to mind with how likely or frequent they are. This psychological tendency might well explain why arguments based on a sample size of one are so common and so compelling to people, especially if we don't stop to consider exactly what evidence we are basing our judgement on. I call this the *availability fallacy*.

> The **availability fallacy** occurs when we base a conclusion on a sample of cases which are memorable, striking or in some other way psychologically salient.
>
> Samples obtained in this way are likely to be biased because what is psychological salient is not necessarily representative.

In fact, memorable or psychological salient cases might be memorable or psychologically salient exactly because they are *not* typical. The typical, average or normal case is often uninteresting and forgettable and importantly in our culture, not *newsworthy*. You're a far more likely to read the headline

> **MAN HIT BY TRAIN AT LEVEL CROSSING**

than the headline

> **ZERO FATAL ACCIDENTS ON LEVEL CROSSINGS AGAIN TODAY**

A great deal of everyday reasoning, especially about risk, is prone to the availability fallacy because our judgements of frequency and likelihood are strongly influenced by the ease with which we can imagine or recall examples. But vivid and unusual events are much more easily remembered and many of the examples we know have come to us through the media, which tends report, often in vivid detail, a carefully selected sample of events.

Hence people are likely to think that car accidents kill more people every year than strokes, whereas in reality, stroke is the more common cause of death. But strokes are not reported on television and the newspapers very often, whereas car accidents are. For that reason, car accidents are easier to bring to mind than cases of stroke, making them seem more frequent. On the other hand, if someone in your family has suffered a stroke, that case will be highly salient for you and so you might well judge that strokes are more frequent killers than car accidents. You would be right but not for the right kind of reason: the evidence for your judgement is based on a biased availability sample. The only way to really know whether strokes case more or less deaths than car accidents is to look at reliably obtained statistics.

Whenever you find yourself reasoning like this: "I can remember lots of cases where A happened, but hardly any where B happened. Therefore, A is probably more common than B" you have just committed the availability fallacy. The cases you can remember are likely to be a biased sample.

Problems with measuring the target property

Suppose a criminologist is interested in studying petty theft amongst office workers. She randomly selects a large sample of offers workers and asks them questions like 'Have you ever stolen stationery from the office?' To her surprise a very significant majority of people say 'no'. The problem with this study of course is that the method used to measure the target property is not very reliable. People are not likely to admit to criminal or unethical behaviour, even if they are told that their answers will be anonymous.

In surveys, people can lie, exaggerate or misrepresent themselves for al sorts of reasons. In studies which look at the effects of diet on health, people are often asked to estimate how often they ate certain foods during the previous month. This way of measuring a target report is known as *self-report* and is quite common in all kinds of research. But people might not remember very well exactly what they ate last month and as we have just seen, memory is not a reliable indicator of frequency. And people might tend to understate the quantity of unhealthy snacks they have consumed and overstate their consumption of virtuous fruit and vegetables. Asking people to keep a 'food diary' helps to alleviate some, but not all, of these problems.

It is well known that the wording of a question, or the order in which questions are asked in a survey, can influence the answer people give. One survey carried out in Scotland asked people the question 'Do you support independence for Scotland?' and 51% of the sample answered 'yes'. Another survey asked what was intended to be the same question, but with different wording 'Do you support an independent Scotland, separate from the United Kingdom?' and this time only 34% of the sample answered 'yes'. To critically evaluate the reliability of an opinion poll or survey, it is always important to ask, 'What question was asked exactly? Could the wording of the question have introduced a bias?'

Summary

An **inference from a sample** is an argument of the following kind:

1. X% of cases in a sample have the target property.
2. The sample is representative of the population.
Therefore:
C. It is likely that X% of the population have the target property.

If the target property has been measured in a reliable and accurate way and the sample really is representative of the population, then both premises are true and the conclusion will be inductively supported. To determine this, some important questions to ask are:

1. What was the size of the sample?
    > Is it large enough to support the conclusion?
2. How was the sample selected?
    > Could the selection process introduce bias?
    > Was the sample self-selected or based on memorable or available cases?
    > What was the response rate?
3. How was the target property measured?
    > What questions were asked?
    > Are people likely to respond untruthfully or inaccurately?

Exercise 7.1. Inference from a sample

Read the following arguments and identify the *sample*, *target property* and *population*. Then comment on the strength of the argument. Assuming that the premises are true, is the conclusion sufficiently supported? Why or why not?

1. Farmers in Queensland will not be too seriously affected by the floods. A recent phone-in poll conducted by ABC News Radio asked the question: 'Will you still buy fruit and veggies, if the prices go up as a result of the flood?' 1,254 people called, sent an SMS or went online to respond. 74% of respondents voted 'yes' and only 26% voted 'no'. So a majority of Australians are willing to pay more for fruit and veggies because of the floods. weak

2. Newspaper report: Quebec environment minister Lise Bacon pledged the PCBs would be moved out and broken down somehow within 18 months. She also said that PCBs couldn't be all that dangerous because her father had washed his hands in PCBs but lived to an old age.

3. The most common cause of violence in Melbourne is excessive drinking. I myself, and many of my friends have often been attacked by drunks coming out of pubs and clubs late at night in the city.

4. Dr. Smith: We should teach this course at the Caulfield Campus, rather than at Clayton. It's nearer to the city, so we might attract more students.
Dr. Jones: I don't think so. I've asked most of the students in the course about that and they all said they're perfectly happy to come out to Clayton.

5. Letter to a newspaper: The implication that Canada's airport security is more lax than that of the United States is wrong. I managed to clear security in a US airport last November with a pair of scissors (1.5 centimeter blades, flat sides, rounded tips) although it did create a stir and a 15 minute delay. The same potential weapon couldn't get through a Canadian airport I travelled through last month. The scissors were confiscated and my fellow passengers could rest easy.