bias Not representative of the population.
A biased sample is not like the population.

non-response bias Bias produced as the result of some subjects not completing a survey.

interviewer bias Bias arising from differences in the way interviewers collect data, such as variations in questions.

design flaw bias Bias arising from faults in the study, such as poor questions or variations in the collection of data.

self-selection bias A kind of non-response bias arising from surveys that seek participation by invitation.

completion bias Bias arising from incomplete data such as unfinished or omitted survey responses.

recall/reporting bias Bias arising from difficulty in recalling information or from the influence that one answer has on another response.

RANDOM SAMPLING

- **understand the concept of a random sample (ACMMM171)**
- **discuss sources of bias in samples and procedures to ensure randomness (ACMMM172)**
- **use graphical displays of simulated data to investigate the variability of random samples from various types of distributions, including uniform, normal and Bernoulli (ACMMM173)**

SAMPLE PROPORTIONS

- **understand the concept of the sample proportion $\hat{p}$ as a random variable whose value varies between samples, and the formulas for the mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ of the sample proportion $\hat{p}$ (ACMMM174)**
- **examine the approximate normality of the distribution of $\hat{p}$ for large samples (ACMMM175)**
- **simulate repeated random sampling, for a variety of values of p and a range of sample sizes, to illustrate the distribution of $\hat{p}$ and the approximate standard normality of $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ where the closeness of the approximation depends on both n and p (ACMMM176)**

9.01 RANDOM SAMPLES AND BIAS

Many variables involve the collection of data from very large groups. It may not be practical to use the whole group. It may be more practical to collect information from only part of the group.

IMPORTANT

For any variable or group of variables, the **population** (or **population of interest**) is the whole group from which data could be collected. It is the universal set for the data.

A **sample** is a part of the population.

In a **census**, data is collected from the whole population.

A **parameter** is a characteristic value of a particular population, such as the mean.

A **statistic** is an estimate of a parameter obtained using a sample.

A **survey** obtains the same information from each member of the sample or population. For a survey of people you would ask each person the same questions.

Example 1

A sample of twenty people waiting in an ATM queue at 7:30 a.m. were asked how much they intended to withdraw. The smallest amount was $20, the average amount was $78, and the greatest was $500. Identify the population, some parameters and statistics.

Getty Images/Bloomberg

Solution

The population is the whole group that could be asked about the amount they withdraw from an ATM.

The population is all the people who use ATMs.

Parameters are clearly defined values from the whole population. You don't need to know the value to define it clearly.

There are 3 parameters: the minimum withdrawal, the average amount withdrawn and the maximum withdrawal.

Statistics are the values you get from the sample. The number of people (20) is not a statistic because it is not an estimate of a parameter.

There are 3 statistics: the minimum withdrawal ($20), the average withdrawal ($78) and the maximum withdrawal ($500).

The sample size is not a parameter because it is not a population property. The size of the population is a parameter.

When you use a sample to find a statistic, you want the statistic to be as close as possible to the population parameter. You need to choose the sample so that it is representative of the population.

Using a very small sample will not usually give you an accurate representation of the population. If you used the extreme case of a sample of size 1, it is obvious that this will not give good results.

You cannot guarantee that a sample will be representative. Suppose that you wanted to find the average income of people in a particular area. Unless you checked everyone, you might miss the one person who was a millionaire. This would obviously have a big effect on the statistic from your sample.

IMPORTANT

A **fair sample** is one that is representative of the population.

A **biased sample** is not representative: it favours some section of the population.

A **random sample** ensures that every member of the population has an equal chance of being chosen.

The statistics from a fair sample are likely to be close to the parameters of the population. Those from a biased sample are unlikely to be close to those of the population.

A random sample is more likely to be fair than one chosen by other means, so statisticians prefer random samples. Unfortunately, random samples of large groups are generally difficult and expensive to obtain.

There are many sources of bias in statistical studies. Investigations involving opinion or feelings are more likely to involve bias than those where you are making measurements.

- **Selection bias** arises from the choice of the sample. This is best avoided by using a random sample.
- **Design flaw bias** arises from faults in the design of the study. Use objective measures wherever possible. If opinion or other subjective measures are required, focus clearly on the question. For example, rather than asking 'Do you support the Liberal party?', ask 'If an election were held this Saturday, which party would you be most likely to vote for?'.

Bias can arise during data collection from differences in the way that data is collected.

- **Interviewer bias** can occur through differences in the way that different interviewers seek information. This can be minimised using standard questions and question options.

In medical trials, special procedures are used to reduce bias. For example, in a double-blind trial some patients are given a drug and others are given a placebo, but the doctors and patients are not told which. This reduces bias caused by the doctor or patient knowledge of treatment.

- **Recall/reporting bias** arises when knowledge of the outcome of one answer affects recall or reporting of the answer. For example, a question about voting intention could affect reporting of past voting practice. Even asking about something that happened the previous week could give false results due to inaccurate recollection.
- **Completion bias** occurs when surveys are incomplete. This can mean that later questions are biased because the questionnaire or interview is abandoned. Surveys should be as short as practical. Longitudinal studies use similar surveys of the same group over an extended period. Completion bias is a big problem, as lost subjects could have a systematic affect on outcomes. A longitudinal study of rural employment in remote areas would be affected by the loss of people who moved away.

Some people will refuse to answer sensitive questions. This is minimised by avoiding questions of potential embarrassment. For example, rather than asking 'how much do you earn in a week?' put it as 'tick the box that shows your income category'.

- **Non-response bias** occurs when some subjects do not take part in the survey. You should be sure that this does not systematically affect results. The most extreme example of non-response bias is a **self-selected sample**. An example of a such a sample is the group of people who respond to a media poll, such as a TV program that asks people to respond 'Yes' or 'No' to a question by ringing up, texting or logging on to a site.

Even if a survey is framed and conducted with minimum bias, bias can still be introduced by inappropriate analysis or reporting of results.

Example 2

In each of the following cases, state whether or not the sampling method is fair, and if it is biased, state the kind(s) of bias.

a An interviewer outside a supermarket on Saturday morning asked people going in: 'Do you prefer *Razzle* dishwasher detergent or an inferior brand?'

b 2000 mobile phone numbers were telephoned at random and people answering were asked: 'What kind of dishwasher detergent do you use?'

c 200 people are chosen at random from the electoral roll of a 'litmus-test' electorate and interviewed at home on a Sunday morning. They are asked 'If an election were held tomorrow, who would you vote for?' Anyone who wasn't home was contacted later at a follow-up that evening. Altogether, 190 people were successfully contacted and only 10 refused to answer. They were put into the 'don't know' category.

Solution

a Those interviewed were available at a particular place and time only. They were asked a leading question.

The survey is biased, with both selection bias and design flaw bias.

b Mobile phones are still not common among elderly people. The question seems fair, but some people would just hang up.

While the design of the question is good, there is some selection bias and there is likely to be some non-response bias.

c Within the electorate, the method and follow-up gives as close to a random selection as practical. However, just because this electorate has gone with the government in the past doesn't mean it will in the future.

The method is fair within the electorate, but if the intention is to predict the result in terms of government, there is selection bias.

Example **2** shows that it is virtually impossible to avoid some bias in a survey, and that this may even be true for a census because of non-response bias. The Australian Bureau of Statistics (ABS) has the legal power to demand answers for its surveys, but this does not guarantee that people will give genuine answers. For important surveys, they use advanced statistical methods to ensure that the final analysis is as free of bias as possible.