1

Samples and populations

Introduction

In this book we present the correct ways to analyse data in appropriate situations, but it is also important that the conclusions made from the data and the analyses are valid and relevant.

When we make a determination on a substance in the laboratory we seek reassurance about a batch from a small quantity from that batch; we do not make every possible determination we could from the batch. In a consumer study we obtain evaluations from a panel of people; we do not seek opinions from everyone in the consumer population. In these and other situations where we obtain data, we base our conclusions on a **sample**.

There are many reasons for sampling. It may be to reduce the effort that would be required if we selected every item produced. It may be because the sampling is destructive (for example, tensile testing) or that the items would then be unfit for use (food). Either way, sampling is an essential feature of process-based industries.

However, when we take a sample it must be for a purpose. It should be followed by measurement, statistical analysis and a decision.

The decision must not be about the sample but the group of items, material or people from which the item was drawn. We refer to this as the **population**.

Thus the starting point for sampling must be the population about which some inference needs to be made.

Let us now look at four examples.

Statistical Methods in Practice: for Scientists and Technologists R. Boddy, G. Smith,

What a lottery!

The key to assessing the validity of sampling is **random sampling**, although practically it can rarely be applied.

A lottery is typical of random sampling, although its purpose is certainly not sampling.

In the lottery draw every number has an equal chance of being chosen. Furthermore, the selection of a number (say, 36) will not alter or lead to any bias in the chance of any other number being chosen. When the next number is drawn, 1 or 37, for example, all numbers will have equal probabilities of being chosen.

No can do

Fizico Ltd produce many thousands of cans of drink per year. They receive the cans in pallets of 1000 cans; a typical delivery will be 20 pallets.

From each delivery they use a sampling inspection scheme to test whether the cans will leak when subjected to pressure. The test uses a higher pressure than is used in the filling process.

To select the sample they use random numbers to choose a pallet and then random numbers again to choose the position within the top layer of the pallet.

They test a sample of 100 cans and find one that leaks.

What can we say about the **population** and the **sampling**?

The sample results can be applied to the population: 'In the delivery it is estimated that 1% of cans will be defective when tested at the high pressure.' Clearly another sample may give a different result, so we do well to give a margin of error on the 1%. This is the subject of a later chapter.

The population is well defined. It is the **delivery**. Furthermore, a decision will be made about the delivery **on the basis of the sample**.

The sampling method contains two elements – choosing the pallet and choosing the cans from the top layer – both of which use random sampling. They always choose the top layer of the pallet. This should not be a problem since there should be no pattern within the pallet. Thus that layer should be representative of the pallet, so is equivalent to randomly sampling the whole pallet.

The only problem would be with an unscrupulous supplier who knows Fizico's sampling method and never puts doubtful cans on the top layer.

In this situation it is safe to assume that we have a valid sampling method.

Nobody is listening to me

Midlands Radio have a regular programme about motoring, and one edition featured safe driving using the speed limit. There is a government proposal to instal in new cars sensors which can receive input from GPS on speed limits. The sensor then sends a message to the engine management system, limiting the speed of the car to within the limit. The host of the programme has asked listeners to phone in and say whether or not they are in favour of the proposal.

Altogether 612 listeners phone to take part in the survey, of whom 511 are against the proposal.

Thus the presenter concludes that 'in the Midlands 83% are against having speed limiters in cars'.

Should we have any doubts about this conclusion?

Let us first consider the **population**. This seems undefined. Is it car drivers? Is it adults? Does it refer to the geographical area classed as the Midlands or the coverage area of the radio station? Does it just refer to radio listeners?

If we do not define the population it is not possible to ascertain whether the sample is representative.

However, even if the population was well defined, the sample is clearly not representative since it is self-selecting and as such is almost certainly biased.

Thus the sample fails on two accounts – a poorly defined population and a biased sample.

It may make good entertainment but is certainly not good science!

How clean is my river?

The biological oxygen demand (BOD) and chemical oxygen demand (COD) are measures that indicate that a river is able to support aquatic life. An agency is assessing the quality of water in an urban river and has decided to sample the water.

One of their inspectors takes three measurements. All three samples give similar values for BOD and similar values for COD. On the basis that the values showed little variability and were sufficiently low the agency concluded that the river can support aquatic life.

Let us now consider the situation. To maintain aquatic life it will be necessary that BOD and COD values are low at all times. The sampling procedure was a

snapshot of the quality at one time. Thus it was representative of the quality of river water only at the sampling site at a stated time. It is well known that the quality will depend on discharges from industrial and effluent plants and these will occur spasmodically and at different times. It is therefore important that the river is sampled over many different hours, days and weeks, and at different sampling sites, to ensure that the sampling is truly representative of the river quality.

Discussion

We have emphasised the importance of considering the **population**. In later chapters we shall be using **significance tests**. All significance tests make **inferences** about the population, and it is important that the population is well defined as well as considering whether the sample is valid before any inferences are made from the data.