Chapter 3.
Populations and Samples

Populations

In statistics the term "population" has a slightly different meaning from the one given to it in ordinary speech. It need not refer only to people or to animate creatures - the population of Britain, for instance or the dog population of London. Statisticians also speak of a population of objects, or events, or procedures, or observations, including such things as the quantity of lead in urine, visits to the doctor, or surgical operations. A population is thus an aggregate of creatures, things, cases and so on.

Although a statistician should clearly define the population he or she is dealing with, they may not be able to enumerate it exactly. For instance, in ordinary usage the population of England denotes the number of people within England's boundaries, perhaps as enumerated at a census. But a physician might embark on a study to try to answer the question "What is the average systolic blood pressure of Englishmen aged 40-59?" But who are the "Englishmen" referred to here? Not all Englishmen live in England, and the social and genetic background of those that do may vary. A surgeon may study the effects of two alternative operations for gastric ulcer. But how old are the patients? What sex are they? How severe is their disease? Where do they live? And so on. The reader needs precise information on such matters to draw valid inferences from the sample that was studied to the population being considered. Statistics such as averages and standard deviations, when taken from populations are referred to as population parameters. They are often denoted by Greek letters: the population mean is denoted by μ (mu) and the standard deviation denoted by σ (lower case sigma).

Samples

A population commonly contains too many individuals to study conveniently, so an investigation is often restricted to one or more samples drawn from it. A well chosen sample will contain most of the information about a particular population parameter but the relation between the sample and the population must be such as to allow true inferences to be made about a population from that sample.

Consequently, the first important attribute of a sample is that every individual in the population from which it is drawn must have a known non-zero chance of being included in it; a natural suggestion is that these chances should be equal. We would like the choices to be made independently; in other words, the choice of one subject will not affect the chance of other subjects being chosen. To ensure this we make the choice by means of a process in which

chance alone operates, such as spinning a coin or, more usually, the use of a table of random numbers. A limited table is given in the Table F (Appendix), and more extensive ones have been published.[1-4] A sample so chosen is called a *random sample.* The word "random" does not describe the sample as such but the way in which it is selected.

To draw a satisfactory sample sometimes presents greater problems than to analyze statistically the observations made on it. A full discussion of the topic is beyond the scope of this book, but guidance is readily available[1][2]. In this book only an introduction is offered.

Before drawing a sample the investigator should define the population from which it is to come. Sometimes he or she can completely enumerate its members before beginning analysis - for example, all the livers studied at necropsy over the previous year, all the patients aged 20-44 admitted to hospital with perforated peptic ulcer in the previous 20 months. In retrospective studies of this kind numbers can be allotted serially from any point in the table to each patient or specimen. Suppose we have a population of size 150, and we wish to take a sample of size five. contains a set of computer generated random digits arranged in groups of five. Choose any row and column, say the last column of five digits. Read only the first three digits, and go down the column starting with the first row. Thus we have 265, 881, 722, etc. If a number appears between 001 and 150 then we include it in our sample. Thus, in order, in the sample will be subjects numbered 24, 59, 107, 73, and 65. If necessary we can carry on down the next column to the left until the full sample is chosen.

The use of random numbers in this way is generally preferable to taking every alternate patient or every fifth specimen, or acting on some other such regular plan. The regularity of the plan can occasionally coincide by chance with some unforeseen regularity in the presentation of the material for study - for example, by hospital appointments being made from patients from certain practices on certain days of the week, or specimens being prepared in batches in accordance with some schedule.

As susceptibility to disease generally varies in relation to age, sex, occupation, family history, exposure to risk, inoculation state, country lived in or visited, and many other genetic or environmental factors, it is advisable to examine samples when drawn to see whether they are, on average, comparable in these respects. The random process of selection is intended to make them so, but sometimes it can by chance lead to disparities. To guard against this possibility the sampling may be *stratified.* This means that a framework is laid down initially, and the patients or objects of the study in a random sample are then allotted to the compartments of the framework. For instance, the framework might have a primary division into males and females and then a secondary division of each of those categories into five age groups, the result being a framework with ten compartments. It is then important to bear in mind that the distributions of the categories on two samples made up on such a framework may be truly comparable, but they will not reflect the distribution of these categories in the population from which the sample is drawn unless the compartments in the framework have been designed with that in mind. For instance, equal numbers might be admitted to the male and female categories, but males and females are not equally numerous in the general population, and their relative proportions vary with age. This is known as *stratified random sampling.* For taking a sample from a long list a compromise between strict theory and

practicalities is known as a *systematic random sample.* In this case we choose subjects a fixed interval apart on the list, say every tenth subject, but we choose the starting point within the first interval at random.

Unbiasedness and Precision

The terms unbiased and precision have acquired special meanings in statistics. When we say that a measurement is *unbiased* we mean that the average of a large set of unbiased measurements will be close to the true value. When we say it is *precise* we mean that it is repeatable. Repeated measurements will be close to one another, but not necessarily close to the true value. We would like a measurement that is both accurate and precise. Some authors equate unbiasedness with *accuracy,* but this is not universal and others use the term accuracy to mean a measurement that is both unbiased *and* precise. Strike[5] gives a good discussion of the problem.

An estimate of a parameter taken from a random sample is known to be unbiased. As the sample size increases, it gets more precise.

Randomization

Another use of random number tables is to randomize the allocation of treatments to patients in a clinical trial. This ensures that there is no bias in treatment allocation and, in the long run, the subjects in each treatment group are comparable in both known and unknown prognostic factors. A common method is to use *blocked randomization*. This is to ensure that at regular intervals there are equal numbers in the two groups. Usual sizes for blocks are two, four, six, eight, and ten. Suppose we chose a block size of ten. A simple method using Table F (Appendix) is to choose the first five unique digits in any row. If we chose the first row, the first five unique digits are 3, 5, 6, 8, and 4. Thus we would allocate the third, fourth, fifth, sixth, and eighth subjects to one treatment and the first, second, seventh, ninth, and tenth to the other. If the block size was less than ten we would ignore digits bigger than the block size. To allocate further subjects to treatment, we carry on along the same row, choosing the next five unique digits for the first treatment. In randomized controlled trials it is advisable to change the block size from time to time to make it more difficult to guess what the next treatment is going to be.

It is important to realize that patients in a randomized trial are *not* a random sample from the population of people with the disease in question but rather a highly selected set of eligible and willing patients. However, randomization ensures that in the long run any differences in outcome in the two treatment groups are due solely to differences in treatment.

Variation between samples

Even if we ensure that every member of a population has a known, and usually an equal, chance of being included in a sample, it does not follow that a series of samples drawn from

one population and fulfilling this criterion will be identical. They will show chance variations from one to another, and the variation may be slight or considerable. For example, a series of samples of the body temperature of healthy people would show very little variation from one to another, but the variation between samples of the systolic blood pressure would be considerable. Thus the variation between samples depends partly on the amount of variation in the population from which they are drawn.

Furthermore, it is a matter of common observation that a small sample is a much less certain guide to the population from which it was drawn than a large sample. In other words, the more members of a population that are included in a sample the more chance will that sample have of accurately representing the population, provided a random process is used to construct the sample. A consequence of this is that, if two or more samples are drawn from a population, the larger they are the more likely they are to resemble each other - again provided that the random technique is followed. Thus the variation between samples depends partly also on the size of the sample. Usually, however, we are not in a position to take a random sample; our sample is simply those subjects available for study. This is a "convenience" sample. For valid generalizations to be made we would like to assert that our sample is in some way representative of the population as a whole and for this reason the first stage in a report is to describe the sample, say by age, sex, and disease status, so that other readers can decide if it is representative of the type of patients they encounter.

Standard error of the mean

If we draw a series of samples and calculate the mean of the observations in each, we have a series of means. These means generally conform to a Normal distribution, and they often do so even if the observations from which they were obtained do not (see **Exercise 3.3**). This can be proven mathematically and is known as the "Central Limit Theorem". The series of means, like the series of observations in each sample, has a standard deviation. The standard error of the mean of one sample is an estimate of the standard deviation that would be obtained from the means of a large number of samples drawn from that population.

As noted above, if random samples are drawn from a population their means will vary from one to another. The variation depends on the variation of the population and the size of the sample. We do not know the variation in the population so we use the variation in the sample as an estimate of it. This is expressed in the standard deviation. If we now divide the standard deviation by the square root of the number of observations in the sample we have an estimate of the standard error of the mean, $SEM = SD/\sqrt{n}$. It is important to realize that we do not have to take repeated samples in order to estimate the standard error, there is sufficient information within a single sample. However, the conception is that *if* we were to take repeated random samples from the population, this is how we would expect the mean to vary, purely by chance.

A general practitioner in Yorkshire has a practice which includes part of a town with a large printing works and some of the adjacent sheep farming country. With her patients' informed consent she has been investigating whether the diastolic blood pressure of men aged 20-44 differs between the printers and the farm workers. For this purpose she has obtained a random

sample of 72 printers and 48 farm workers and calculated the mean and standard deviations, as shown in Table 3.1.

To calculate the standard errors of the two mean blood pressures the standard deviation of each sample is divided by the square root of the number of the observations in the sample.

$$\text{Printers:SEM} = 4.5 / \sqrt{72} = 0.53 \text{ mmHg}$$
$$\text{Farmers:SEM} = 4.2 / \sqrt{48} = 0.61 \text{ mmHg}$$

These standard errors may be used to study the significance of the difference between the two means, as described in successive chapters

Table 3.1   Mean diastolic blood pressures of printers and farmers

	Number	Mean diastolic blood pressure (mmHg)	Standard deviation (mmHg)
Printers	72	88	4.5
Farmers	48	79	4.2

Standard error of a proportion or a percentage

Just as we can calculate a standard error associated with a mean so we can also calculate a standard error associated with a percentage or a proportion. Here the size of the sample will affect the size of the standard error but the amount of variation is determined by the value of the percentage or proportion in the population itself, and so we do not need an estimate of the standard deviation. For example, a senior surgical registrar in a large hospital is investigating acute appendicitis in people aged 65 and over. As a preliminary study he examines the hospital case notes over the previous 10 years and finds that of 120 patients in this age group with a diagnosis confirmed at operation 73 (60.8%) were women and 47 (39.2%) were men.

If p represents one percentage, 100 p represents the other. Then the standard error of each of these percentages is obtained by (1) multiplying them together, (2) dividing the product by the number in the sample, and (3) taking the square root:

$$SE\ percentage\ =\ \sqrt{\frac{p(100-p)}{n}}$$

which for the appendicitis data given above is as follows:

$$SE\ percentage\ =\ \sqrt{\frac{60.8\times39.2}{120}}\ =\ 4.46$$

Problems with non-random samples

In general we do not have the luxury of a random sample; we have to make do with what is available, a "*convenience sample*". In order to be able to make generalizations we should investigate whether biases could have crept in, which mean that the patients available are not typical. Common biases are:

- hospital patients are not the same as ones seen in the community;
- volunteers are not typical of non-volunteers;
- patients who return questionnaires are different from those who do not.

In order to persuade the reader that the patients included are typical it is important to give as much detail as possible at the beginning of a report of the selection process and some demographic data such as age, sex, social class and response rate.

Common questions

Given measurements on a sample, what is the difference between a standard deviation and a standard error?

A standard deviation is a sample estimate of the population parameter σ; that is, it is an estimate of the variability of the observations. Since the population is unique, it has a unique standard deviation, which may be large or small depending on how variable the observations are. We would not expect the sample standard deviation to get smaller because the sample gets larger. However, a large sample would provide a more precise estimate of the population standard deviation σ than a small sample.

A standard error, on the other hand, is a measure of precision of an estimate of a population parameter. A standard error is always attached to a parameter, and one can have standard errors of any estimate, such as mean, median, fifth centile, even the standard error of the standard deviation. Since one would expect the precision of the estimate to increase with the sample size, the standard error of an estimate will decrease as the sample size increases.

When should I use a standard deviation to describe data and when should I use a standard error?

It is a common mistake to try and use the standard error to describe data. Usually it is done because the standard error is smaller, and so the study appears more precise. If the purpose is to describe the data (for example so that one can see if the patients are typical) and if the data are plausibly Normal, then one should use the standard deviation (mnemonic D for Description and D for Deviation). If the purpose is to describe the outcome of a study, for example to estimate the prevalence of a disease, or the mean height of a group, then one should use a standard error (or, better, a confidence interval; see Chapter 4) (mnemonic E for Estimate and E for Error).

References

1 Altman DG. *Practical Statistics for Medical Research.* London: Chapman & Hall, 1991
2. Armitage P, Berry G. *Statistical Methods in Medical Research.* Oxford: Blackwell Scientific Publications, 1994.
3. Campbell MJ, Machin D. *Medical Statistics: A Commonsense Approach.* 2nd ed. Chichester: John Wiley, 1993.
4. Fisher RA, Yates F. *Statistical Tables for Biological, Agricultural and Medical Research,* 6th ed. London: Longman, 1974.
5. Strike PW. Measurement and control. *Statistical Methods in Laboratory Medicine.* Oxford: Butterworth-Heinemann, 1991:255.

Exercises

Exercise 3.1   The mean urinary lead concentration in 140 children was 2.18 μ mol/24 h, with standard deviation 0.87. What is the standard error of the mean?

Exercise 3.2   In Table F (Appendix), what is the distribution of the digits, and what are the mean and standard deviation?

Exercise 3.3   For the first column of five digits in Table F take the mean value of the five digits and do this for all rows of five digits in the column.
What would you expect a histogram of the means to look like?
What would you expect the mean and standard deviation to be?