Section Six

Inferential statistics

SECTION CONTENTS

The statistical analysis of the data is an essential stage in the process of quantitative research. The principles of inferential statistics, as introduced in this section, are applied for deciding if the obtained sample data show the differences and patterns we set out to demonstrate in the population. This is the case for all types of research where we are using sample data to make inferences concerning populations.

Quantitative research in the health sciences mostly involves working with samples drawn from populations. In order to generalize our findings, we must draw generalizations and inferences from sample statistics (e.g. $\bar{X}$, s) to the population parameters (e.g. μ, σ). Inferences are always probabilistic, because even with random samples there is always the chance of sampling error. The finite probability of sampling error implies that the differences or patterns identified in our sample data could represent random variations or chance patterns rather than 'real' ones which are true for the population as a whole.

As an illustration, imagine that we have collected data in a study aimed at identifying age-related differences in the use of sedatives and tranquillizers in a given population. Our participants (n = 200) kept diaries over a period of 1 year, recording each time they had taken a sedative or tranquillizer. The research question is: is there a difference between sedative or tranquillizer use in older and younger people? Assume that the following hypothetical data were obtained:

Sedative and tranquillizer consumption

Age group	n	$\bar{X}$	s
20–39	100	20	5
40–59	100	30	5

Three important and interrelated questions are examined in Section 6 concerning the evidence

provided by sample data such as those shown in the above table.

1. Even if we used an adequate sampling procedure (see Ch. 3), how confident are we to infer that the true population parameters (μ) are the same as, or are at least close to, $\bar{X}$? For example, is it true that the mean tranquillizer intake for the 20–39 age group is $\mu = 20$?
2. It appears that there is a large difference between the two sample means; but is this difference also true for the populations? In other words is this difference 'real' or significant, or is it simply due to sampling error?
3. What we are inferring is μ older $> \mu$ younger. In other words, are we justified in concluding that the mean sedative/tranquillizer intake for the older age group is greater than for the younger group in the population?

The key issue here is that we are using sample data for decision making. Sampling error refers to the difference between sample statistics and the actual state of the population.

We cannot eliminate sampling error even with large and well-chosen samples. Rather, as outlined in Section 6, we can apply the principles of inferential statistics to calculate the probability of error. We then use this information to minimize the probability of making errors when we generalize from sample statistics to population parameters.

In Chapter 17 we examine how sampling distributions are derived and used for calculating the probability of obtaining a given sample statistic. This information can be applied to the calculation of confidence intervals, which represent a range of scores which contain the true population parameter at a given level or probability.

In Chapter 18 we outline the logic of hypothesis testing, using single sample z and t tests as exemplars. Hypothesis testing is a procedure used to decide if a difference or pattern identified in our sample data is statistically significant. If the outcome of our analysis is significant, then we are in a position to decide that the patterns or differences found in our data may be generalized to the populations from which the samples were drawn.

There are numerous statistical tests available for analysing the significance of our data. In Chapter 19 we discuss basic criteria for selecting an appropriate statistical test, including (i) scale of measurement used to collect the data, (ii) the number of groups being compared and (iii) the dependence or independence of measurements. We will use the χ^2 (chi-squared) test to demonstrate how statistical tests are selected and used to analyse the data.

Ultimately, statistical decision making is probabilistic, implying that the possibility of making decision errors cannot be eliminated. Decisions may be correct, or involve what are called Type I and Type II errors. In Chapter 20, we examine how these errors may influence our interpretation of the obtained data in relation to the aims or hypotheses guiding our research, and we will examine strategies which can be employed to reduce the probability of making such errors. In this chapter we outline the relationship between effect size and the clinical or practical significance of research findings.

Chapter Seventeen 17

Probability and sampling distributions

CHAPTER CONTENTS

Introduction

Sample statistics (such as $\bar{X}$, s) are estimates of the actual population parameters μ, σ. Even where adequate sampling procedures are adopted, there is no guarantee that the sample statistics are the exact representations of the true parameters of the population from which the samples were drawn. Therefore, inferences from sample statistics to population parameters necessarily involve the possibility of sampling error. As stated in Chapter 3, sampling errors represent the discrepancy between sample statistics and true population parameters. Given that investigators usually have no knowledge of the true population parameters, inferential statistics are employed to estimate the probable sampling errors. While sampling error cannot be completely eliminated, its probable magnitude can be calculated using *inferential statistics*. In this way investigators are in a position to calculate the probability of being accurate in their estimations of the actual population parameters.

The aims of this chapter are to examine how probability theory is applied to generating sampling distributions and how sampling distributions are used for estimating population parameters. Sampling distributions can be used for specifying confidence intervals, as discussed in this chapter, as well as for testing hypotheses, as further discussed in Chapter 18.

The specific aims of this chapter are to:

1. Define probability.
2. Demonstrate how sampling distributions are generated.
3. Show how sampling distributions of the mean are used to calculate the probability of a sample mean.
4. Explain how confidence intervals are calculated for continuous data.
5. Distinguish between *z* and *t* distributions.

Probability

The concept of probability is central to the understanding of inferential statistics. Probability is expressed as a proportion between 0 and 1, where 0 means an event is certain not to occur, and 1 means an event is certain to occur. Therefore if the probability (p) is 0.01 for an event then it is unlikely to occur (chance is 1 in a hundred). If $p = 0.99$ then the event is highly likely to occur (chance is 99 in a hundred). The probability of any event (say event A) occurring is given by the formula:

$$p(\text{A}) = \frac{\text{number of occurrences of A}}{\text{total number of possible occurrences}}$$

Sometimes the probability of an event can be calculated a priori (before the event) by reasoning alone. For example, we can predict that the probability of throwing a head (H) with a fair coin is:

$$p(\text{H}) = \frac{\text{number of occurrences of H}}{\text{total number of possible occurrences}} = \frac{\text{H}}{\text{H} + \text{T(tails)}} = \frac{1}{2} = 0.5$$

Or, if we buy a lottery ticket in a draw where there are 100 000 tickets, the probability of winning first prize is:

$$p(\text{1st prize}) = \frac{1}{100\,000} = 0.00001$$

This is true only if the lottery is fair, if all tickets have an equal chance of being drawn by random selection.

In some situations, there is no model which we can apply to calculate the occurrence of an event a priori. For instance, how can we calculate the probability of an individual dying of a specific condition? In such instances, we use previously obtained empirical evidence to calculate probabilities a posteriori (after the event).

For example, if it is known that the percentages (or proportions) for causes of death are distributed in a particular way, then the probability of a particular cause of death for a given individual can be predicted. Table 17.1 represents a set of hypothetical statistics for a community.

Given the data in Table 17.1, we are in a position to calculate the probability of a selected individual over 65 dying of any of the specified causes. For example, the probability of a given individual dying of coronary heart disease is:

$$p(\text{dying of heart disease}) = \frac{\%\text{ of cases dying of heart disease}}{100\%} = \frac{50\%}{100\%} = 0.5$$

This approach ignores individual risk factors and assumes that the environmental conditions under which the data were obtained are still pertinent. However, the example illustrates the principle that once we have organized the data into a frequency distribution we can calculate the probability of selecting any of the tabulated values. This is true whether the variable was measured on a nominal, ordinal, interval or ratio scale. Here, we will

Table 17.1 Causes of death for persons over 65

Cause of death	Percentage of deaths
Coronary heart disease	50
Cancer	25
Stroke	10
Accidents	5
Infections	5
Other causes	5

examine how to calculate the probability of values for normally distributed, continuous variables.

We can use the normal curve model, as outlined in Chapter 15, to determine the proportion or percentage of cases up to, or between, any specified scores. In this instance, probability is defined as the proportion of the total area cut off by the specified scores under the normal curve. The greater the proportion, the higher the probability of selecting the specified values.

For example, say that on the basis of previous evidence we can specify the frequency distribution of neonates' weight. Let us assume that the distribution is approximately normal, with the mean ($\bar{X}$) of 5.0kg and a standard deviation (s) of 1.5. Now, say that we are interested in the probability of a randomly selected neonate having a birth weight of 2.0kg or under. Figure 17.1 illustrates the above situation.

The area A1 under the curve in Figure 17.1 corresponds to the probability of obtaining a score of 2 or under. Using the principles outlined in Chapter 15 to calculate proportions or areas under the normal curve, we first translate the raw score of 2 into a z score:

$$z = \frac{x - \bar{X}}{s} = \frac{2 - 5}{1.5} = -2$$

Now we look up the area under the normal curve corresponding to $z = -2$ (Appendix A). Here we find that A1 is 0.0228. This area corresponds to a probability, and we can say that 'The probability of a neonate having a birth weight of 2kg or less is 0.0228'. Another way of stating this outcome is that the chances are approximately 2 in 100, or 2%, for a child having such a birth weight.

We can also use the normal curve model to calculate the probability of selecting scores between any given values of a normally distributed continuous variable. For example, if we are interested in the probability of birth weights being between 6 and 8kg, then this can be represented on the normal curve (area A2 on Fig. 17.2). To determine this area, we proceed as outlined in Chapter 15. Let $s = 1.5$.

$$z_1 = \frac{6 - 5}{1.5} = 0.67$$

$$z_2 = \frac{8 - 5}{1.5} = 2.0$$

Therefore: the area between z_1 and $\bar{X}$ is 0.2486 (from Appendix A) and the area between z_2 and $\bar{X} = 0.4772$ (from Appendix A). Therefore, the required area A2 is:

$$A2 = 0.4772 - 0.2486 = 0.2286$$

It can be concluded that the probability of a randomly selected child having a birth weight between 6 and 8kg is $p = 0.2286$. Another way of saying this is that there is a chance of 23 in 100 or a 23% chance that the birth weight will be between 6 and 8kg.

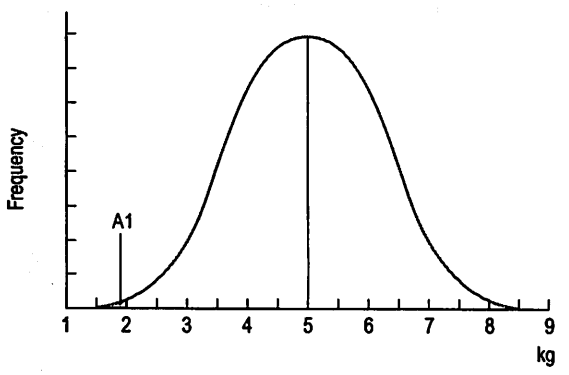

Figure 17.1 • Frequency distribution of neonate birth weights. Area A1 corresponds to $z \leq -2$.

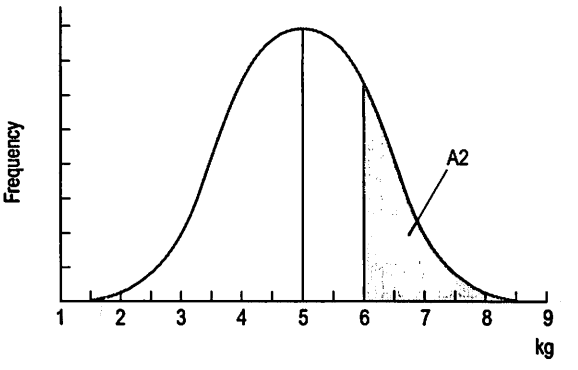

Figure 17.2 • Frequency distribution of neonate birth weights. Area A2 corresponds to probability of weight being 6–8kg.

The above examples demonstrate that when the mean and standard deviation are known for a normally distributed continuous variable, this information can be applied to calculating the probability of events related to this distribution. Of course, probabilities can be calculated for other than normal data but this requires integral calculus which is beyond the scope of this text. In general, regardless of the shape or scaling of a distribution, scores which are common or 'average' are more likely to be selected than those which are atypical, being unusually high or low.

Sampling distributions

Probability theory can also be applied to calculate the probability of obtaining specific samples from populations.

Consider a container with a very large number of identically sized marbles. Imagine that there are two kinds of marbles present, black (B) and white (W), and that these colours are present in equal proportions, so that $p(B) = p(W) = 0.5$.

Given the above population, say that samples of marbles are drawn randomly and with replacement. (By 'replacement' we mean the samples are put back into the population, in order to maintain as a constant the proportion of B = W = 0.5.) If we draw samples of four (that is, $n = 4$) then the possible proportions of black and white marbles in the samples can be deduced a priori as shown in Figure 17.3.

Possible outcomes	Number black	Number white	Proportions black
● ● ● ●	4	0	1.00
● ● ● ○	3	1	0.75
● ● ○ ○	2	2	0.50
● ○ ○ ○	1	3	0.25
○ ○ ○ ○	0	4	0.00

Figure 17.3 • Characteristics of possible samples of $n = 4$, drawn from a population of black and white marbles.

Ignoring the order in which marbles are chosen, Figure 17.3 demonstrates all the possible outcomes for the composition of samples of $n = 4$. It is logically possible to draw any of the samples shown. However, only one of the samples (2B, 2W), is representative of the true population parameter. The other samples would generate incorrect inferences concerning the state of the population. In general, if we know or assume (hypothesize) the true population parameters, we can generate distributions of the probability of obtaining samples of a given characteristic.

In this instance, when attempting to predict the probability of specific samples drawn from a population with two discrete elements, the binomial theorem can be applied. The expansion of the binomial expression, $(P + Q)^n$, generates the probability of all the possible samples which can be drawn from a given population. The general equation for expanding the binomial expression is:

$$(P + Q)^n = P^n + \frac{n}{1}P^{n-1}Q + \frac{n(n-1)}{2}P^{n-2}Q^2 + \ldots Q^n$$

P is the probability of the first outcome, Q is the probability of the second outcome and n is the number of trials (or the sample size).

In this instance, P = proportion black (B) = 0.5; Q = proportion white (W) = 0.5; $n = 4$ (sample size). Therefore, substituting into the binomial expression:

$$(B+W)^4 = B^4 + 4B^3W + 6B^2W^2 + 4BW^3 + W^4$$

Note that each part of the expansion stands for a probability of obtaining a specific sample. For the present case:

Sample 1 p(4B0W)	$= B^4$	$= (0.5)^4$	$= 0.0625$
Sample 2 p(3B1W)	$= 4B^3W$	$= 4 \times (0.5)^3(0.5)$	$= 0.2500$
Sample 3 p(2B2W)	$= 6B^2W^2$	$= 6 \times (0.5)^2(0.5)^2$	$= 0.3750$
Sample 4 p(1B3W)	$= 4BW^3$	$= 4 \times (0.5)^3(0.5)$	$= 0.2500$
Sample 5 p(0B4W)	$= W^4$	$= (0.5)^4$	$= 0.0625$

The calculated probabilities add up to 1, indicating that all the possible sample outcomes have been accounted for. However, the important issue here is not so much the mathematical details but the general principle being illustrated by the example. For a given sample size (n) we can employ a mathematical formula to calculate the probability of obtaining all the possible samples from a population with known parameters. The relationship between the possible samples and their probabilities can be graphed, as shown in Figure 17.4.

Taking the statistic 'number of black marbles in the sample', the graph in Figure 17.4 shows the probability of obtaining any of the outcomes. The distribution shown is called a 'sampling distribution'. In general, a sampling distribution for a statistic indicates the probability of obtaining any of the possible values of a statistic.

Therefore, having obtained our sampling distribution we can see that some sample outcomes have low probability while others are more likely. Although there is a finite chance of obtaining a sample such as 'all blacks', the probability of this happening is rather small ($p = 0.0625$). Conversely, a sample of 2B2W, which is equal to the true population proportions, is far more probable ($p = 0.375$). Generating sampling distributions for calculating the probability of given sample statistics is a basic practice in inferential statistics. The sampling distributions enable researchers to infer (with a determined level of confidence) the true population parameters from the sample statistics.

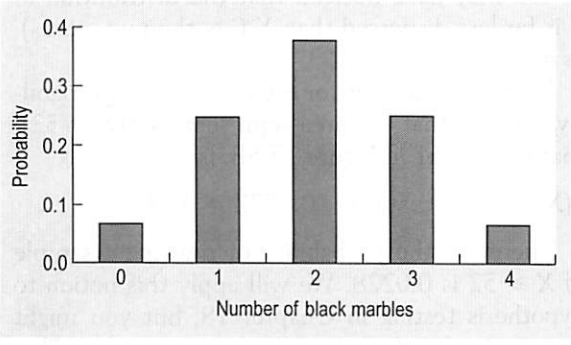

Figure 17.4 • Sampling distribution of black marbles; $n = 4$.

Sampling distribution of the mean

The binomial theorem is appropriate for generating sampling distributions for discontinuous nominal scale data. However, when measurements are continuous, the mean and standard deviations are appropriate as sample statistics and are measured on interval or ratio scales. The sampling distribution of the mean represents the frequency distribution of sample means obtained from random samples drawn from the population. The sampling distribution of the mean enables the calculation of the probability of obtaining any given sample mean ($\bar{X}$). This is essential for testing hypotheses about sample means (Ch. 18).

In order to generate the sampling distribution of the mean, we use a mathematical theorem called the central limit theorem. This theorem provides a set of rules which relate the parameters (μ, σ) of the population from which samples are drawn to the distribution of sample means ($\bar{X}$).

The central limit theorem states that if random samples of a fixed n are drawn from any population, as n becomes large the distribution of sample means approaches a normal distribution, with the mean of the sample means ($\bar{X}_{\bar{x}}$ or $\mu_{\bar{x}}$) being equal to the population mean (μ) and the standard error of estimate ($s_{\bar{x}}$ or $\sigma_{\bar{x}}$) being equal to $\sigma/\sqrt{n}$. The standard error of the estimate is the standard deviation of the distribution of sample means.

Let us follow the above step by step.

1. Imagine we have a population of continuous scores or measurements with a mean of μ and a standard deviation of σ.
2. We select a very large number of random samples, each sample being of a size n.
3. Having obtained our samples, for each sample we calculate the sample mean ($\bar{X}_1$, $\bar{X}_2$, . . . and so on).
4. Each sample mean, $\bar{X}$, is a number. The sampling distribution of the mean is a frequency distribution representing the large number of sample means.
5. The central limit theorem predicts theoretically the shape (normal for large n), mean ($\bar{X}_{\bar{x}}$ or $\mu_{\bar{x}}$) and standard deviation ($s_{\bar{x}}$ or $\sigma_{\bar{x}}$) of a large number of sample means.

It should be noted that:

1. The sampling distribution of the mean is a frequency distribution of a large number of sample means of size n drawn from a given population. When n increases, the sampling distributions approach normal.
2. The mean of the sample means ($\mu_{\bar{x}}$ or $\bar{X}_{\bar{x}}$) is the mean of the distribution of sample means. $\bar{X}_{\bar{x}}$ and $\mu_{\bar{x}}$ are equal to μ, the population mean.
3. The standard error of the mean ($s_{\bar{x}}$ or $\sigma_{\bar{x}}$) is the standard deviation of the frequency distribution of sample means drawn from a population. The magnitude of $s_{\bar{x}}$ or $\sigma_{\bar{x}}$ is equal to $\sigma/\sqrt{n}$, the population standard deviation divided by the square root of the sample size.
4. $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ are used in reference to a sampling distribution based on all the possible samples drawn from a population (a population of samples, would you believe?), while $\bar{X}$ and $s_{\bar{x}}$ are used when the sampling distribution is based on a 'sample' of samples.

Let us have a look at an example. Assume for a hypothetical test of motor function that $\mu = 50$ and $\sigma = 10$. What is the probability of drawing a random sample from this population with $\bar{X} = 52$ or greater (i.e. $\bar{X} \geqslant 52$) given that $n = 100$? The central limit theorem predicts that when we draw samples of $n = 100$ from the above population, the sampling distribution of the means will be as follows:

- The shape of the sampling distribution will be approximately normal.
- The mean of the sampling distribution will be equal to μ:

$$\mu_{\bar{x}} = \mu = 50$$

- The standard error of estimate ($\sigma_{\bar{x}}$) will be:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

(We can show this as in Fig. 17.5.)

Previously, we saw how we can use normal frequency distributions for estimating probabilities. Using the same principles as in Chapter 15, we can calculate the z score corresponding to $\bar{X} = 52$, and look up Appendix A to find out the area representing the probability in question:

$$z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{52 - 50}{1} = 2$$

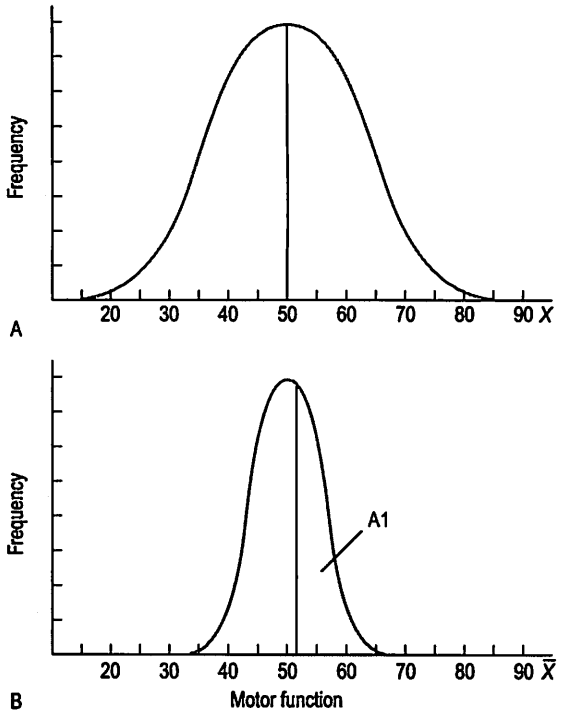

Figure 17.5 • Relationship between original population and sampling distribution of the mean, where $\mu = 50$ and $\sigma = 10$; (A) original population, (B) sampling distribution of the mean.

That is, $\bar{X} = 52$ is two standard error units above $\mu_{\bar{x}}$, the population mean for the sample means.

You may have noticed that the distribution of $\bar{X}$ is far less dispersed than X (i.e. the raw scores), as $\sigma_{\bar{x}} = 1$ and $\sigma = 10$.

Using Appendix A for establishing the probability, we find that the area representing $p(\bar{X} \geqslant 52)$, that is, area A1 in Figure 17.5B, is:

$$p(\bar{X} \geqslant 52) = 0.5000 - 0.4772 = 0.0228$$

Therefore the probability of drawing a sample of $\bar{X} \geqslant 52$ is 0.0228. We will apply this notion to hypothesis testing in Chapter 18, but you might have noticed that it is a rather low probability. That is, it is unlikely ($p = 0.0228$) randomly to draw a sample with $\bar{X} \geqslant 52$ where $n = 100$ and $\mu = 50$.

Application of the central limit theorem to calculating confidence intervals

Let us assume that you are asked to estimate the weight of a newborn baby. If you are experienced in working with neonates, you should be able to make a reasonable guess. You might say 'The baby is 6 kg'. Someone might ask 'How certain are you that the baby is exactly 6 kg?' You might then say 'Well, the baby might not be exactly 6 kg, but I'm very confident that it weighs somewhere between 5.5 and 6.5 kg'. This statement expresses a confidence interval – a range of values which probably include the true value. Of course, the more certain or confident you want to be of including the true value, the bigger the range of values you might give: you are unlikely to be wrong if you guess that the baby weighs between 4 and 8 kg.

Confidence intervals can be calculated for a large range of statistics such as proportions, ratios or correlation coefficients. In this chapter we will look at confidence intervals for single samples as an illustration for using confidence intervals.

We have seen previously that if we know the population parameters we can estimate the probability of selecting from that population a sample mean of a given magnitude. Conversely, if we know the sample mean we can estimate the population parameters from which the sample might have come, at a given level of probability. Let us take an example to illustrate this point.

A researcher is interested in the systolic blood pressure (BP) levels of smokers of more than 10 cigarettes per day. She takes a random sample of 100 10+ smokers in her district and finds that the mean BP = 148 mmHg for the sample, with a standard deviation of $s = 10$.

She wants to generalize to the population of smokers of more than 10 cigarettes per day in their district. The best estimate of μ (the population parameter) is 148, but it is possible that, because of sampling error, 148 is not the exact population parameter. However she can calculate a confidence interval (a range of blood pressures that will include the true population mean at a given level of probability). A confidence interval is a range of scores which includes the true population parameter at a specified level of probability. The precise probability is decided by the researcher and indicates how certain she can be that the population mean is actually within the calculated range. Common confidence intervals used in statistics are 95% confidence intervals, which offer a probability of $p = 0.95$ for including the true population mean, and 99% confidence intervals, which include the true population mean at a probability of $p = 0.99$.

Calculating the confidence interval requires the use of the following formula:

$$\bar{X} - zs_{\bar{x}} \leq \mu \leq \bar{X} + zs_{\bar{x}}$$

where $\bar{X}$ is the sample mean; z is the z score obtained from the normal curve table such that it cuts off the area of the normal curve corresponding to the required probability; $s_{\bar{x}}$ is the sample standard error, which is equal to the sample standard deviation divided by $\sqrt{n}$, that is, $s_{\bar{x}} = s/\sqrt{n}$.

Let us turn to the previous example to illustrate the use of the above equation. Here $\bar{X} = 148$, and $s_{\bar{x}} = 10/\sqrt{100}$. Assume that we want to calculate a 95% confidence interval. We are looking for a pair of z scores which have 95% of the standard normal curve between them. In this case, 1.96 is the value for z which cuts off 95% of a normal distribution. That is, we looked up the value of z corresponding to an area (probability) of 0.4750, since the 0.05 has to be divided among the two tails of the distribution, giving 0.025 at either end. Substituting into the equation above we have:

$$148 - (1.96)(10/\sqrt{100}) \leq \mu \leq 148 + (1.96) \times (10/\sqrt{100}) = 146.04 \leq \mu \leq 149.96$$

That is, the investigator is 95% confident that the true population mean, the true mean BP of smokers, lies between 146.04 (lower limit) and 149.96 (upper limit). There is only a 5% or 0.05 probability that it lies outside this range. If we chose a 99% confidence interval, then using the formula as above, we have:

$$148 - (2.58)(10/\sqrt{100}) \leq \mu \leq 148 + (2.58) \times (10/\sqrt{100}) = 145.42 \leq \mu \leq 150.58$$

Here, 2.58 is the value of z which cuts off 99% of a normal distribution (Fig. 17.6). That is, the investigator is 99% confident that the true population mean lies somewhere between 145.42 and 150.58. Clearly, the 99% interval is wider than the 95% interval; after all, here the probability of including the true mean is greater than for the 95% interval. Conventionally, health sciences publications report 95% confidence intervals.

Confidence intervals where *n* is small: the *t* distribution

It was previously stated that: 'as *n* becomes large, the distribution of sample means approaches a normal distribution' (central limit theorem). The questions left to explain are:

- How large must the sample size, *n*, be before the sampling distribution of the mean can be considered a normal curve?
- What are the implications for the sampling distribution if *n* is small?

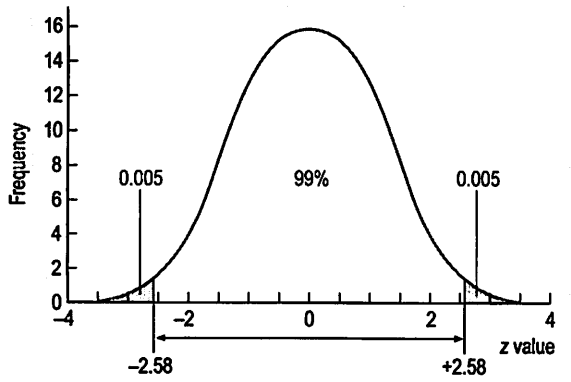

Figure 17.6 • z scores for 99% confidence interval.

It has been shown by mathematicians that the sample size, *n*, for which the sampling distribution of the mean can be considered an approximation of a normal distribution is $n \geqslant 30$. That is, if *n* is 30 or more, we can use the standard normal curve to describe the sampling distribution of the mean. However, when $n < 30$, the sampling distribution of the mean is a rather rough approximation to the normal distribution. Instead of using the normal distribution, we use the *t distribution*, which takes into account the variability of the shape of the sampling distribution due to low *n*.

The *t* distributions (Fig. 17.7) are a family of curves representing the sampling distributions of means drawn from a population when sample size, *n*, is small ($n < 30$). A 'family of curves' means that the shape of the *t* distribution varies with sample size. It has been found that the distribution is determined by the *degrees of freedom* of the statistic.

The degrees of freedom (df) for a statistic represents the number of scores which are free to vary when calculating the statistic. Since the statistic we are calculating in this case is the mean,

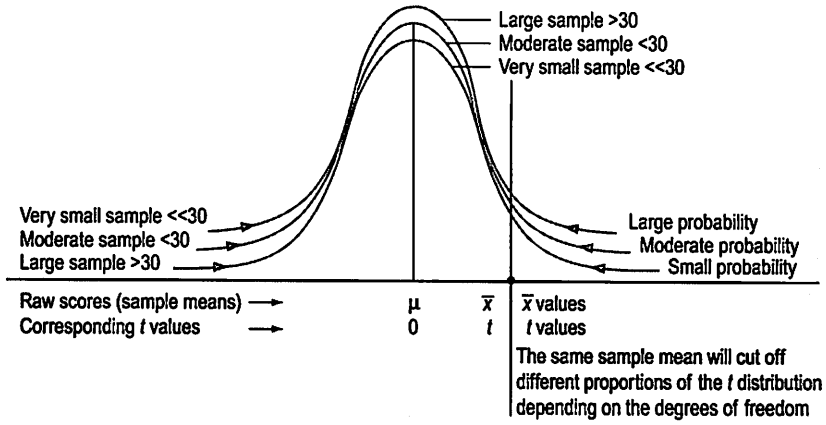

Figure 17.7 • *t* distributions.

all but one of the scores could vary. That is, if you were inventing scores in a sample with a known mean, you would have a free hand until the very last score. There df is equal to $n - 1$ (the sample size minus one). Each row of figures shown in Appendix B represents the critical values of t for a given distribution.

1. The t distribution is symmetrical about the mean.
2. The values of t along the x-axis cut off specific areas under the curve, just as for z. These areas are given at the top of the page in Appendix B, under 'Directional' and 'Non-directional' probabilities.
3. The t distribution approaches a normal distribution as n becomes larger. As stated earlier, when $n \geq 30$, for all practical purposes the t and z distributions coincide.

The t distribution, just as the z distribution, can be used to approximate the probability of drawing sample means of a given magnitude from a population; t can also be used for calculating confidence intervals. Let us re-examine the example relating to the blood pressure of smokers presented earlier. Let us assume that $n = 25$, with the other statistics remaining the same: $\bar{X} = 148$, $s = 10$. The general formula for calculating the confidence intervals for small samples is:

$$\bar{X} - ts_{\bar{x}} \leq \mu \leq \bar{X} + ts_{\bar{x}}$$

You will note the similarity to the equation on page 199; here t replaces z. If we want to show the 95% confidence interval, then we use the same logic as for z distributions (Fig. 17.8).

To look up the t values from the tables (Appendix B) consider (i) direction, (ii) probabilities and (iii) degrees of freedom.

We are looking at a 'non-directional' or 'two-tail' probability in the sense that the t values cut off 95% of the area of the t curve between them, leaving 5% distributed at the two tails of the t distribution: $p = 0.05$; df $= 25 - 1 = 24$. Therefore $t = 2.064$ (from Appendix B). Substituting into the equation for calculating confidence intervals:

$$148 - (2.1)(10/\sqrt{25}) \leq \mu \leq 148 + (2.1)(10/\sqrt{25})$$
$$148 - 4.2 \leq \mu \leq 148 + 4.2$$
$$143.8 \leq \mu \leq 152.2$$

Consider the width of the confidence interval defined as the distance between the upper and lower limits. Note that this is a wider interval than that which was obtained when n was 100. As sample size, n, becomes smaller, our confidence interval becomes wider, reflecting a greater probability of sampling error.

To calculate the 99% confidence interval (Fig. 17.9) we need to look up $p = 0.01$, non-directional, df $= 24$ in Appendix B to obtain the critical value of t, which is 2.797.

$$148 - (2.8)(10/\sqrt{25}) \leq \mu \leq 148 + (2.8)(10/\sqrt{25})$$
$$142.4 \leq \mu \leq 153.6$$

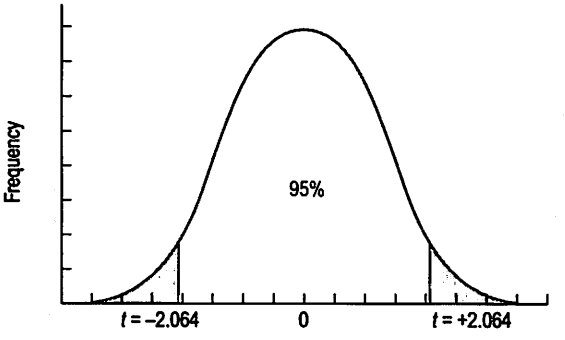

Figure 17.8 • The 95% confidence interval for sample size of 25.

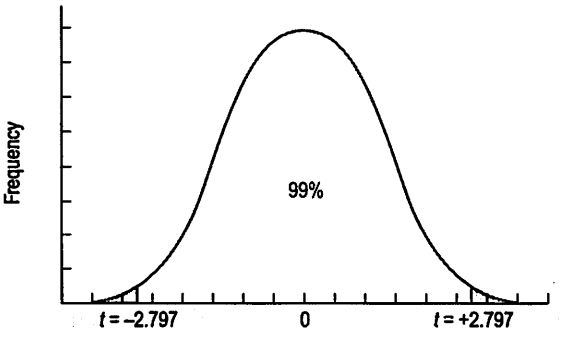

Figure 17.9 • The 99% confidence interval for sample size of 25.

We can see that, when $n = 25$, the 99% confidence interval is wider than for $n = 100$. That is, the bigger our sample size, the narrower (more precise) our estimate of the range of values (which includes the true population parameter) becomes when our sample size is large.

Summary

It was argued in this chapter that, even with randomly selected samples, the possibility of sampling error must be taken into account when making inferences from sample statistics to population parameters. It was shown that probability theory can be applied to generating sampling distributions, which express the probability of obtaining a given sample from a population. With discontinuous, nominal data the binomial theorem provides an adequate mathematical distribution for estimating the probability of obtaining possible samples. However, with continuous data, the central limit theorem is applied to generate the sampling distributions of the mean. The standard distribution of the mean enables the calculation of the probability-specified sample mean(s) by random selection. The sampling error of the mean ($s_{\bar{x}}$ or $\sigma_{\bar{x}}$), which expresses statistically the range of the sampling error, depends inversely on the sample size, such that the larger the n, the smaller the $s_{\bar{x}}$ or $\sigma_{\bar{x}}$.

One of the applications of sampling distributions is for calculating confidence intervals for continuous data. Confidence intervals represent a range of scores which specify, from sample data, the probability of capturing the true population parameters.

Health researchers usually report 95% confidence intervals. When sample sizes are small ($n < 30$), the t distribution is appropriate for representing the sampling distribution of the mean. With large sample sizes, the two distributions merge together. As the next chapter will demonstrate, sampling distributions are essential for testing hypotheses, a procedure which uses inferential statistics to calculate the level of probability at which sample statistics support the predictions of hypotheses.

Self-assessment

Explain the meaning of the following terms:

central limit theorem
confidence interval
degrees of freedom
population parameter
probability
sampling distribution
sampling distribution of the mean
sampling error
standard error of the mean
t distribution

True or false

1. If death and taxes are an absolute certainty, then the probability of their occurrence is infinite.
2. Probability values fall between 0 and 1.
3. When a score occurs at a relatively high level of frequency the probability of randomly selecting it from a distribution is $p = 1$.
4. The higher the corresponding *z* score, the lower the probability of randomly selecting the score from a distribution.
5. For a normal distribution, σ is the score which has the highest probability of random selection.
6. It is possible to have a negative *z* score but not a negative probability.
7. For a continuous normal distribution, the probability of selecting a score up to and including the mode is $p = 0.5$.
8. The probability of randomly selecting a score 2 standard deviations above the mean is $p = 0.2$.
9. We can generate appropriate sampling distributions for statistics derived from nominal or ordinal data.
10. Statistical inference involves the estimation of population parameters from sample statistics.
11. Statistical inference depends on knowing the true population parameters before beginning research.
12. The sampling distribution of the mean is a frequency polygon of mean scores.
13. Using the sample mean as an estimate for the population mean is an example of statistical inference.

14. The sampling distribution of the mean is always normally distributed.
15. As n increases, the variability of the sampling distribution of the mean increases.
16. $\mu_{\bar{x}} = \mu$ regardless of the shape of the sampling distribution of the mean.
17. The characteristics of the sampling distribution of the mean vary with the size of the sample.
18. If a random selection method is used, sampling error will be zero.
19. The bigger the sampling error, the smaller the confidence interval.
20. Generally speaking, the higher the level of confidence we have that an interval contains the population mean, the larger is that interval.

Multiple choice

1. The mean age of the Canadian population is known to be 31 years. A small randomly selected sample of Canadians is found to have a mean age of 32. This discrepancy is an example of:
   a a sampling error
   b a measurement error
   c a problem in ecological validity
   d failure to control for the effects of maturation.

At a large maternity hospital the following hypothetical data are compiled concerning the birth weights and survival of neonates.

Weight of neonates (g)	Numbers (n)	Percentage of neonates surviving
0–499	25	40
500–999	50	60
1000–1499	75	80
1500–1999	150	90
2000–2499	250	95
2500+	450	98

Questions 2–5 refer to this table.

2. What is the probability that a randomly selected neonate will weigh under 500 g?
   a 0.4000
   b 0.1700
   c 0.0250
   d 0.0025
3. What is the probability that a randomly selected neonate will have a birth weight of 2000 g or over?
   a 0.7000
   b 0.9638
   c 0.3000
   d 0.9500
4. What is the probability that a neonate weighing 499 g or under will survive?
   a 0.4000
   b 0.1700
   c 0.040
   d 0.025
5. What is the probability that a neonate weighing 2500 g or over will not survive?
   a 0.400
   b 0.300
   c 0.006
   d 0.020

A test of reaction times has a mean of 10 and a standard deviation of 4 in the normal adult population with a normal distribution. Questions 6–12 refer to this information.

6. A person scores 8. That person's z score is:
   a 2
   b −2
   c −0.5
   d −1
7. What percentage of the population would have scores up to and including 14 on this test?
   a 84.13
   b 15.87
   c 65.87
   d 34.13
8. What is the percentile rank of a score of 8 on this test?
   a 19.15
   b 30.85
   c 80.85
   d 53.28
9. What score (to the nearest whole number) would cut off the highest 10% of scores?
   a 1
   b 14

c 15
d 18

10. What is the probability that a randomly selected individual will score greater than 6 on this test?
    a 0.4987
    b 0.3413
    c 0.8413
    d 0.6587
11. What is the probability that a randomly selected individual will score between 8 and 14 on this test?
    a 0.1498
    b 0.5328
    c 0.6816
    d 0.4671
12. What is the probability that a randomly selected individual will score either more than 14 or less than 10 on this test?
    a 0.6587
    b 0.6816
    c 0.3184
    d 0.8184
13. Sampling error of the mean:
    a occurs because of poor sampling techniques
    b decreases as sample size increases
    c is independent of the standard deviation
    d is always equal to 1.
14. Samples of 100 are drawn from a normally distributed population with a mean of 50 and a standard deviation of 10. The distribution of sample means will:
    a have a mean of 50
    b have a standard deviation of 1
    c be normally distributed
    d all of the above.
15. Increasing the sample size, *n*:
    a decreases sampling error
    b increases sampling error
    c has no effect on standard error of the mean
    d requires increasing correction of sample estimates of population parameters
    e none of these.
16. If the dispersion of the raw score population increases while *n* is held constant, $\sigma_{\bar{x}}$:
    a decreases
    b increases
    c remains the same
    d cannot be ascertained without more information.
17. The sampling distribution of the mean:
    a is always positively skewed for continuous data
    b is normally distributed if the population raw scores are not normally distributed
    c is approximately normally distributed if sample size is large
    d all of the above.
18. As the degrees of freedom decrease, the similarity between the *t* and *z* distributions:
    a increases
    b decreases
    c remains the same
    d approaches infinity.
19. The theoretical sampling distributions of the *t* statistics depend on:
    a *p*
    b *r*
    c $s_{\bar{x}}$
    d df.

A normally distributed population has a mean of 80 and a standard deviation of 12. Questions 20–24 refer to this information.

20. For samples of $n = 36$, what is the standard error of the mean?
    a 12
    b 0.33
    c 2
    d 3
21. A sample of 64 cases is found to have a mean of 83. What is the *z* score of this mean?
    a 2
    b 0.25
    c 4
    d 1.5
22. A sample of $n = 144$ has a mean of 77. What is the probability that a mean this low would occur by chance?
    a 0.0300
    b 0.0013
    c 0.4989
    d 0.9987
23. A sample of 36 cases is selected. What is the probability that its mean falls outside the range of 79–81?
    a 0.6813
    b 0.8085
    c 0.3085
    d 0.6170

24. Which is more likely: that a randomly selected sample of $n = 36$ will have a mean greater than 82 or that it will have a mean less than 77?
    a Greater than 82.
    b Less than 77.
    c Both are equally probable.
    d Impossible to tell.
25. The *t* distribution differs from the *z* distribution in which of these ways?
    a Its mean is not exactly equal to 0.
    b It is not quite symmetrical.
    c It is somewhat wider and flatter.
    d All of the above.
26. The 95% confidence interval arrived at from a particular experiment is 72–79. Therefore:
    a the probability is 0.05 that μ falls between 72–79
    b the probability is 0.95 that the interval 72–79 contains $\bar{X}$
    c the probability is 0.95 that the interval 72–79 contains μ
    d *a* and *c*.
27. Compared with a 99% confidence interval, a 95% confidence interval is:
    a larger
    b smaller
    c more likely to contain the population mean
    d less likely to contain the sample mean.

The following information should be used in answering questions 28–31: a random sample of 25 clients is selected, and their systolic blood pressures measured; the mean BP is 115 mmHg, with a standard deviation of 10.

28. This is an example of:
    a an experiment
    b a natural comparison study
    c a survey
    d field research.
29. What is the standard error of the mean for a sample of this size?
    a 10
    b 20
    c 2
    d 2.5
30. In order to calculate the 99% confidence interval of the mean, what *t* score will be used?
    a 2.492
    b 2.787
    c 2.797
    d 1.711
31. What is the 99% confidence interval of the mean in this example?
    a $110.0 \leqslant \mu \leqslant 120.0$
    b $109.4 \leqslant \mu \leqslant 120.6$
    c $111.6 \leqslant \mu \leqslant 118.4$
    d $113.0 \leqslant \mu \leqslant 117.0$
32. A random sample of 25 university students is found to have a mean IQ of 110, with a standard deviation of 10. Between what two possible scores can we be 99% confident that the true mean IQ for the students at the university lies?
    a 95.5–112.5
    b 85–135
    c 102.4–117.6
    d 104.1–115.9

In order to establish the mean weight of newborn babies at a large maternity hospital, a random sample of 64 babies is weighed. Their mean weight is 2500 g, with a standard deviation of 80 g.

Questions 33–35 refer to this information.

33. What is the standard error of the mean?
    a 80
    b 10
    c 1.25
    d 0.1
34. In calculating the 95% confidence interval of the mean, what *z* value is used?
    a 1.96
    b 2.33
    c 2.58
    d 1.64
35. What is the 95% confidence interval of the mean in this example (to the nearest whole number)?
    a $2480 \leqslant \mu \leqslant 2520$
    b $2477 \leqslant \mu \leqslant 2523$
    c $2474 \leqslant \mu \leqslant 2526$
    d $2484 \leqslant \mu \leqslant 2516$

Chapter Eighteen 18

Hypothesis testing

CHAPTER CONTENTS

Introduction

In the previous chapter we introduced the use of inferential statistics for estimating population parameters from sample statistics. In the case of some non-experimental research projects, such as surveys and descriptive statistics, parameter estimation is adequate for analysing the data. After all, these investigations aim at describing the characteristics of specific populations. However, other research strategies involve data collection for the purpose of testing hypotheses. Here the investigator has to establish if the data support or refute the hypotheses being investigated. The key issue is that hypotheses are generalizations addressing differences in patterns and associations in populations. Inferential statistics enables us to calculate the probability (level of significance) for asserting that what we are seeing in our sample data is generalizable to the population. This probability is related to the statistical significance of the sample data.

The aim of this chapter is to introduce the logical steps involved in hypothesis testing in quantitative research. Given that hypothesis testing is probabilistic, special attention must be paid to the possibility of making erroneous decisions, and to the implications of making such errors.

The specific aims of the chapter are to:

1. Examine the logic of hypothesis testing for retaining or rejecting null hypotheses.

2. Outline how decisions are made with directional and non-directional alternative hypotheses.
3. Define the concept of statistical significance.
4. Introduce the use of the single sample *z* and *t* test for analysing the statistical significance of the data.
5. Outline the probability and implications of making Type I and Type II decision errors.

A simple illustration of hypothesis testing

One of the simplest forms of gambling is betting on the fall of a coin. Let us play a little game. We, the authors, will toss a coin. If it comes out heads (H) you will give us £1; if tails (T) we will give you £1. To make things interesting, let us have 10 tosses. The results are:

Toss	1	2	3	4	5	6	7	8	9	10
Outcome	H	H	H	H	H	H	H	H	H	H

Oh dear, you seem to have lost. Never mind, we were just lucky, so send along your cheque for £10. What is that, you are a little hesitant? Are you saying that we 'fixed' the game? There is a systematic procedure for demonstrating the probable truth of your allegations:

1. We can state two competing hypotheses concerning the outcome of the game:
   (a) The authors fixed the game; that is, the outcome does not reflect the fair throwing of a coin. Let us call this statement the 'alternative hypothesis', H_A. In effect, the H_A claims that the sample of 10 heads came from a population other than P (probability of heads) = Q (probability of tails) = 0.5.
   (b) The authors did not fix the game; that is, the outcome is due to the tossing of a fair coin. Let us call this statement the 'null hypothesis', or H_0. H_0 suggests that the sample of 10 heads was a random sample from a population where $P = Q = 0.5$.
2. It can be shown that the probability of tossing 10 consecutive heads with a fair coin is $p = 0.001$, as discussed previously (see Ch. 17). That is, the probability of obtaining such a sample from a population where $P = Q = 0.5$ is extremely low.
3. Now we can decide between H_0 and H_A. It was shown that the probability of H_0 being true was $p = 0.001$ (1 in a 1000). Therefore, in the balance of probabilities, we can reject it as being true and accept H_A, which is the logical alternative. In other words, it is likely that the game was fixed and no £10 cheque needs to be posted.

Table 18.1 Probability of obtaining all heads in coin tosses

n (number of tosses)	*p* (all heads)
1	0.5000
2	0.2500
3	0.1250
4	0.0625
5	0.0313

The probability of calculating the truth of H_0 depended on the number of tosses (n = the sample size). For instance, the probabilities of obtaining all heads with up to five tosses, according to the binomial theorem (Ch. 17), are shown in Table 18.1. The table shows that, as the sample size (n) becomes larger, the probability at which it is possible to reject H_0 becomes smaller. With only a few tosses we really cannot be sure if the game is fixed or not: without sufficient information it becomes hard to reject H_0 at a reasonable level of probability.

A question emerges: 'What is a reasonable level of probability for rejecting H_0?' As we shall see, there are conventions for specifying these probabilities. One way to proceed, however, is to set the appropriate probability for rejecting H_0 on the basis of the implications of erroneous decisions.

Obviously, any decision made on a probabilistic basis might be erroneous. Two types of elementary decision errors are identified in statistics as *Type I* and *Type II errors*. A Type I error involves mistakenly rejecting H_0, while a Type II error involves mistakenly retaining H_0.

In the above example, a Type I error would involve deciding that the outcome was not due to chance when in fact it was. The practical outcome of this would be to accuse the authors falsely of fixing the game. A Type II error would represent the decision that the outcome was due to chance, when in fact it was due to a 'fix'. The practical

outcome of this would be to send your hard-earned £10 to a couple of crooks. Clearly, in a situation like this, a Type II error would be more odious than a Type I error, and you would set a fairly high probability for rejecting H_0. However, if you were gambling with a villain, who had a loaded revolver handy, you would tend to set a very low probability for rejecting H_0. We will examine these ideas more formally in subsequent parts of this chapter.

The logic of hypothesis testing

Hypothesis testing is the process of deciding statistically whether the findings of an investigation reflect chance or real effects at a given level of probability. If the results do not represent chance effects then we say that the results are statistically significant. That is, when we say that our results are statistically significant we mean that the patterns or differences seen in the sample data are generalizable to the population.

The mathematical procedures for hypothesis testing are based on the application of probability theory and sampling, as discussed previously. Because of the probabilistic nature of the process, decision errors in hypothesis testing cannot be entirely eliminated. However, the procedures outlined in this section enable us to specify the probability level at which we can claim that the data obtained in an investigation support experimental hypotheses. This procedure is fundamental for determining the statistical significance of the data as well as being relevant to the logic of clinical decision making.

Steps in hypothesis testing

The following steps are conventionally followed in hypothesis testing:

1. State the alternative hypothesis (H_A), which is the prediction intended for evaluation. The H_A claims that the results are 'real' or 'significant', i.e. that the independent variable influenced the dependent variable, or that there is a real difference among groups. The important point here is that H_A is a statement concerning the population. A real or significant effect means that the results in the sample data can be generalized to the population.
2. State the null hypothesis (H_0), which is the logical opposite of the H_A. The H_0 claims that any differences in the data were just due to chance: that the independent variable had no effect on the dependent variable, or that any difference among groups is due to random effects. In other words, if the H_0 is retained, differences or patterns seen in the sample data should not be generalized to the population.
3. Set the decision level, α (alpha). There are two mutually exclusive hypotheses (H_A and H_0) competing to explain the results of an investigation. Hypothesis testing, or statistical decision making, involves establishing the probability of H_0 being true. If this probability is very small, we are in a position to reject the H_0. You might ask 'how small should be the probability (α) for rejecting H_0?' By convention, we use the probability of $\alpha = 0.05$. If the H_0 being true is less than 0.05 we can reject H_0. We can choose an α of < 0.05, but not more, That is, by convention among researchers, results are not characterized as significant if $p > 0.05$.
4. Calculate the probability of H_0 being true. That is, we assume that H_0 is true and calculate the probability of the outcome of the investigation being due to chance alone, i.e. due to random effects. We must use an appropriate sampling distribution for this calculation.
5. Make a decision concerning H_0. The following decision rule is used. If the probability of H_0 being true is less than α, then we reject H_0 at the level of significance set by α. However, if the probability of H_0 is greater than α, then we must retain H_0. In other words, if:

   p (H_0 is true) $\leq \alpha$; reject H_0

   p (H_0 is true) $> \alpha$; retain H_0

   It follows that if we reject H_0 we are in a position to accept H_A, its logical alternative. If we reject H_0, we decide that H_A is probably true.

Let us look at an example. A rehabilitation therapist devises an exercise programme which is expected to reduce the time taken for people to leave hospital following orthopaedic surgery. Previous records show that the recovery time for patients has been $\mu = 30$ days, with $\sigma = 8$ days. A sample of 64 patients are treated with the exercise programme, and their mean recovery time is found to be $\bar{X} = 24$ days. Do these results show

that patients who had the treatment recovered significantly faster than previous patients? We can apply the steps for hypothesis testing to make our decision.

1. State H_A: 'The exercise programme reduces the time taken for patients to recover from orthopaedic surgery'. That is, the researcher claims that the independent variable (the treatment) has a 'real' or 'generalizable' effect on the dependent variable (time to recover).
2. State H_0: 'The exercise programme does not reduce the time taken for patients to recover from orthopaedic surgery'. That is, the statement claims that the independent variable has no effect on the dependent variable. The statement implies that the treated sample with $\bar{X} = 24$, and $n = 64$ is in fact a random sample from the population $\mu = 30$, $\sigma = 8$. Any difference between $\bar{X}$ and μ can be attributed to sampling error.
3. The decision level, α, is set before the results are analysed. The probability of α depends on how certain the investigator wants to be that the results show real differences. If he set $\alpha = 0.01$, then the probability of falsely rejecting a true H_0 is less than or equal to 0.01 (1/100). If he set $\alpha = 0.05$, then the probability of falsely rejecting a true H_0 is less than or equal to 0.05 or (1/20). That is, the smaller the α, the more confident the researcher is that the results support the alternative hypothesis. We also call α the level of significance. The smaller the α, the more significant the findings for a study, if we can reject H_0. In this case, say that the researcher sets $\alpha = 0.01$. (Note: by convention, α should not be greater than 0.05.)
4. Calculate the probability of H_0 being true. As stated above, H_0 implies that the sample with $\bar{X} = 24$ is a random sample from the population with $\mu = 30$, $\sigma = 8$. How probable is it that this statement is true? To calculate this probability, we must generate an appropriate sampling distribution. As we have seen in Chapter 17, the sampling distribution of the mean will enable us to calculate the probability of obtaining a sample mean of $\bar{X} = 24$ or more extreme from a population with known parameters. As shown in Figure 18.1, we can calculate the probability of drawing a sample mean of $\bar{X} = 24$ or less. Using the table of normal curves (Appendix A), as outlined previously, we find that the probability of randomly selecting a sample mean of $\bar{X} = 24$ (or less) is extremely small. In terms of our table, which only shows the exact probability of up to $z = 4.00$, we can see that the present probability is less than 0.00003. Therefore, the probability that H_0 is true is less than 0.00003.
5. Make a decision. We have set $\alpha = 0.01$. The calculated probability was less than 0.0001. Clearly, the calculated probability is far less than α. Therefore, the investigator can reject the statement that H_0 is true and accept H_A, that patients treated with the exercise programme recover earlier than the population of untreated patients.

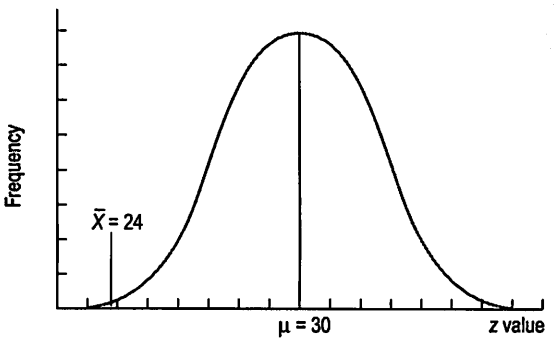

Figure 18.1 • Sampling distribution of means. Sample size = 64; population mean = 30; standard deviation = 8.

Directional and non-directional hypotheses and corresponding critical values of statistics

In the previous example, H_A was directional in that we asserted that the difference between the mean of the treated sample and the population mean was expected to be in a particular direction. If we state that there was some effect due to the dependent variable, but do not specify which way, then H_A is called non-directional. In the previous example, if the investigator stated H_A as 'The exercise programme *changes* the time taken to recover following surgery' then H_A would have been non-directional.

In general, an alternative hypothesis is directional if it predicts a specific outcome concerning the direction of the findings by stating that one group mean will be higher or lower than the other(s). An alternative hypothesis is non-directional if it predicts a difference, without specifying which

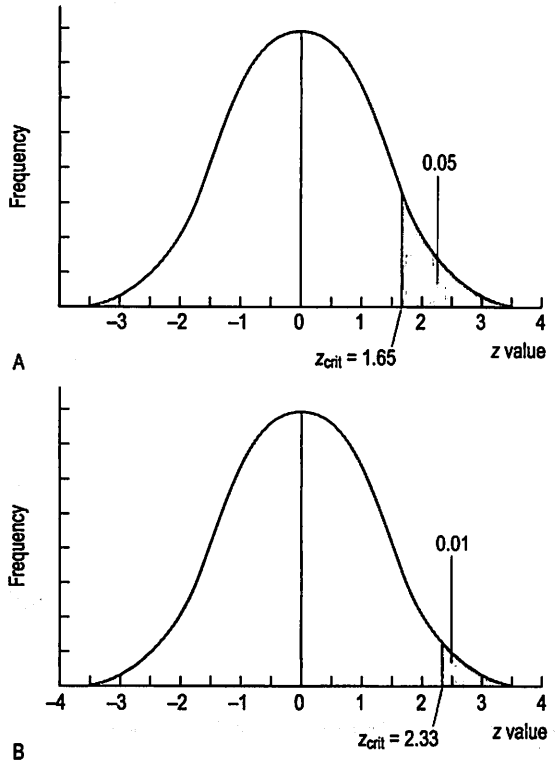

Figure 18.2 • Two examples of statistical decision making with directional (one-tail) hypothesis, H_A.

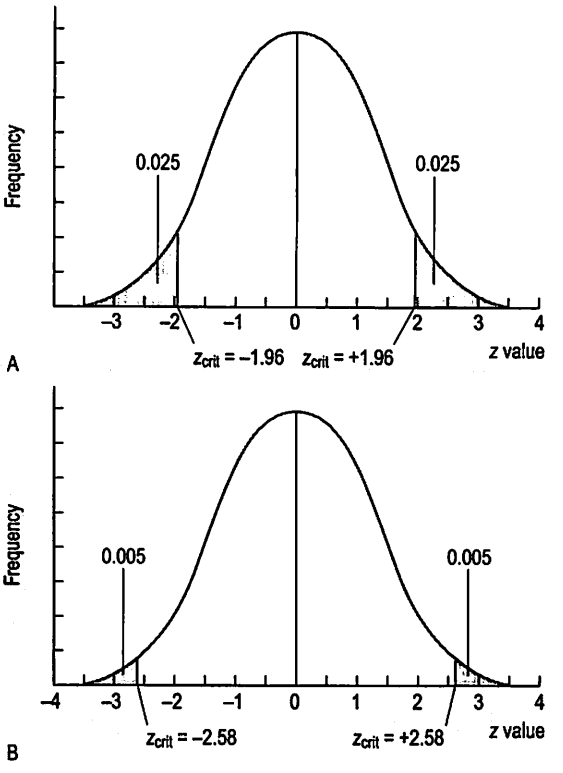

Figure 18.3 • Two examples of statistical decision making with non-directional (two-tail) hypothesis, H_A.

group mean is expected to be higher or lower than the others.

If we propose a directional H_A, it is understood that we have reasonable information on the basis of pilot studies or previously published research for predicting the direction of the outcome. The advantage of a directional H_A is that it increases the probability of rejecting H_0. However, the decision of the directionality of H_A must be decided before the data are collected and analysed.

Let us now examine the concept of the 'critical' value of a statistic. The critical value of a statistic is the value of the statistic which bounds the proportion of the sampling distribution specified by α. The critical value of the statistic is influenced by whether H_A is directional or non-directional.

Figures 18.2 and 18.3 represent the sampling distributions of the mean where n is large; that is, the sampling distribution for the statistic $\bar{X}$.

As we have seen in Chapter 17, these are the sampling distributions for $\bar{X}$ we would expect by the random selection of samples, as specified by H_0. Therefore, we can estimate from the distributions the probability of selecting any sample mean, $\bar{X}$, by chance alone. The α value (the level of significance) specifies the criterion for rejecting H_0. We can see that the critical value for the statistic (in this case z_{crit}) cuts off an area of the distribution corresponding to α ($p = 0.05$ or $p = 0.01$).

In Figure 18.2, we can see that $z_{crit} = 1.65$ (for $\alpha = 0.05$) and $z_{crit} = 2.33$ (for $\alpha = 0.01$). (These values are obtained from Appendix A.) Therefore, for any sample mean, $\bar{X}$, where the transformed (z) value is greater than or equal to z_{crit}, we will reject H_0 (that the sample mean was a random sample). However, if the absolute value of the transformed statistic is less than z_{crit}, then we

must retain H_0. Note that when $\alpha = 0.01$, the z_{crit} is greater than when $\alpha = 0.05$. Clearly, the higher the level of significance set for rejecting H_0, the greater the absolute critical value of the statistic. Figure 18.2 shows statistical decision making with a directional H_A, where the probabilities associated with only one of the tails of the distribution are used.

Figure 18.3 shows the critical values for z with a non-directional H_A. Here, the probabilities associated with α (0.05 or 0.01) are divided between the two tails of the distribution. That is, where $\alpha = 0.05$, half (0.025) goes into each tail, and where $\alpha = 0.01$, half (0.005) also goes into each tail. This changes the values of z_{crit}, which becomes ± 1.96 or ± 2.58, respectively, as shown in Figure 18.3. Here, we reject H_0 if the calculated transformed z value of $\bar{X}$ falls beyond the values of z_{crit}. When we compare the values of z_{crit} for the one-tail and two-tail decisions, we find that the critical values are greater for the two-tail decisions. This implies that it is more difficult to reject H_0 if we are making two-tail decisions on the basis of a non-directional HA.

Decision rules

In general, Figures 18.2 and 18.3 illustrate the decision rules for statistical decision making for hypotheses concerning sample means. These rules are:

$|z_{obt}| \geq |z_{crit}|$; reject H_0
$|z_{obt}| < |z_{crit}|$; retain H_0

The same decision rules hold for the *t* distributions associated with the sampling distribution of the mean when *n* (the sample size) is small (see Ch. 17).

$|t_{obt}| \geq |t_{crit}|$; reject H_0
$|t_{obt}| < |t_{crit}|$; retain H_0

z_{obt} and t_{obt} refer to the calculated value of the statistic, based on the data:

$$z_{obt} = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

$$t_{obt} = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

z_{crit} and t_{crit} are the critical values of the statistic obtained from the tables in Appendices A and B. As we have seen, the values of these depend on α and the directionality of H_A. | | is the symbol for modulus, implying that we should look at the absolute value of a statistic. Of course, the sign is important when considering if $\bar{X}$ is greater or smaller than μ. However we can ignore the sign (+ or −) when making statistical decisions. In effect, the greater z_{obt} or t_{obt}, the more deviant or improbable the particular sample mean, $\bar{X}$, is under the sampling distribution specified by H_0.

Statistical decisions with single sample means

The following examples illustrate the use of statistical decision making concerning a single sample mean, $\bar{X}$. Such decisions are relevant when our data consist of a single sample and we are to decide if the $\bar{X}$ of the sample is significantly different to a given population, with a mean of μ.

A statistical test is a procedure appropriate for making decisions concerning the significance of the data. The *z* test and the *t* test are procedures appropriate for making decisions concerning the probability that sample means reflect population differences. (As shown in Ch. 19, there is a variety of statistical tests available for hypothesis testing.)

Example 1

A researcher hypothesizes that males now weigh more than in previous years. To investigate this hypothesis he randomly selects 100 adult males and records their weights. The measurements for the sample have a mean of $\bar{X} = 70$ kg. In a census taken several years ago, the mean weight of males was $\mu = 68$ kg, with a standard deviation of 8 kg.

1. Directional H_A: males are heavier. That is, $\bar{X} = 70$ is not a random sample from population $\mu = 68$.
2. H_0: males are not heavier. That is, $\bar{X} = 70$ is a random sample from population $\mu = 68$.
3. Decision level: $\alpha = 0.01$.
4. Calculate probability of H_0 being true. Here, $\alpha = 0.01$, one-tail. We can find from the tables

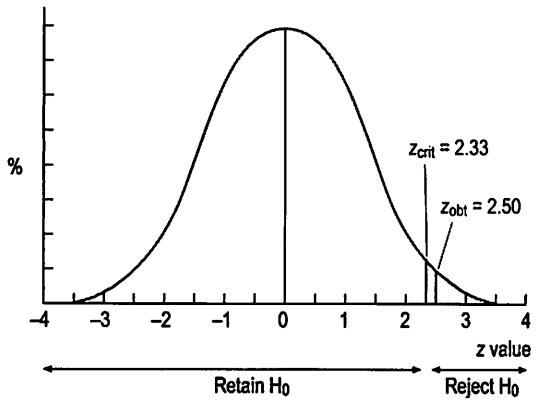

Figure 18.4 • Hypothesis testing: directional.

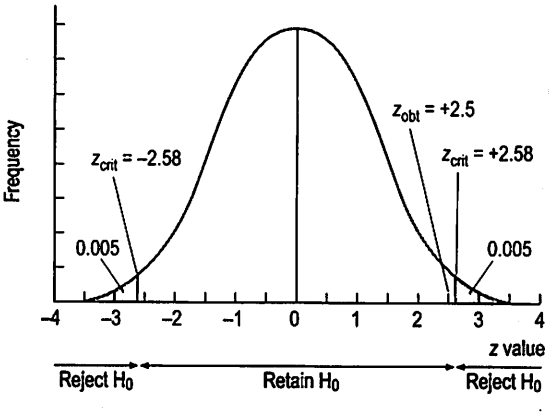

Figure 18.5 • Hypothesis testing: non-directional.

(Appendix A) z_{crit}, the z score which cuts off an area of 0.01 of the total curve. $z_{crit} = +2.33$ (α = 0.01; one tail).

Calculating the z score (z_{obt}) representing the probability of the sample being drawn from the population under H_0 ($\mu = 68$), we use the formula:

$$z_{obt} = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

where $s_{\bar{x}} = s/\sqrt{n}$

$$z_{obt} = \frac{70 - 68}{8/\sqrt{100}}$$
$$= 2.5$$

Here, $z_{crit} = 2.33$

5. The decision rule is that if:

$|z_{obt}| \geq |z_{crit}|$; reject H_0
$|z_{obt}| < |z_{crit}|$; retain H_0

2.5 > 2.33, so the z_{obt} falls into the area of rejection, as shown in Figure 18.4. Therefore, the researcher can reject H_0, and accept H_A at a 0.01 level of significance. That is, the results of the investigation indicate that the mean weight of males has increased (consistent with the predictions of the research hypothesis). We conclude that the results are statistically significant at $p \leq 0.01$.

Example 2

A researcher hypothesized that men today have different weights (either more or less) than in previous years (assume the same information as for Example 1).

1. Non-directional H_A: males are of different weight, that is $\bar{X} = 70$ is not a random sample from population $\mu = 68$.
2. H_0: males are not different, that is $\bar{X}$ is a random sample from population $\mu = 68$.
3. Decision level: $\alpha = 0.01$.
4. Calculate the probability of H_0 being true. Here, $\alpha = 0.01$ (two-tail); the value of $z_{crit} = 2.58$ (from Appendix A); the value of $z_{obt} = 2.5$ (as calculated in Example 1).
5. Decision: applying the decision rule as outlined in Example 1:

$|z_{obt}| < |z_{crit}|$; as 2.5 < 2.58

z_{obt} falls into the area of acceptance, as shown in Figure 18.5. Therefore, the researcher must retain the H_0, and conclude that the study did not support H_A at a 0.01 level of significance. The investigation has not provided evidence that the mean weight of males has increased. The results are reported as not being statistically significant.

Example 3

The previous two examples involved sample sizes of $n > 30$. However, as we saw in Chapter 17, if

$n < 30$, the distribution of sample means is not a normal, but a t distribution. This point must be taken into account when we calculate the probability of H_0 being true. That is, for small samples, we use the t test to evaluate the significance of our data.

Assume exactly the same information as in Example 1, except that sample size is $n = 16$.

1. Directional H_A: as in Example 1.
2. H_0: as in Example 1.
3. $\alpha = 0.01$, one tail.
4. We can find from the t table (Appendix B) the value for t_{crit}. To look up t_{crit} we must have the following information:
   (a) α, the level of significance (0.05 or 0.01)
   (b) direction of H_A (directional or non-directional)
   (c) the degrees of freedom (df).
   In this instance:
   (a) $\alpha = 0.01$
   (b) H_A is directional, therefore we must look up a one-tail probability
   (c) $df = n - 1 = 16 - 1 = 15$
   Looking up the appropriate value for t; $t_{crit} = 2.602$. Calculating the t score (t_{obt}) representing the probability of the sample being drawn from the population under H_0, we use the formula:

$$t_{obt} = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$
$$= \frac{70 - 68}{8/\sqrt{16}}$$
$$= 1.0$$

5. As we stated earlier, the decision rule is identical to that of the z test:

$|z_{obt}| \geq |z_{crit}|$; reject H_0
$|z_{obt}| < |z_{crit}|$; retain H_0

Here, $1.0 < 2.602$, such that t_{obt} falls into the area of retention (Fig. 18.6). Therefore, we must retain H_0 at a 0.01 level of significance. Clearly, when $n = 16$, the investigation did not show a significant weight increase for the males.

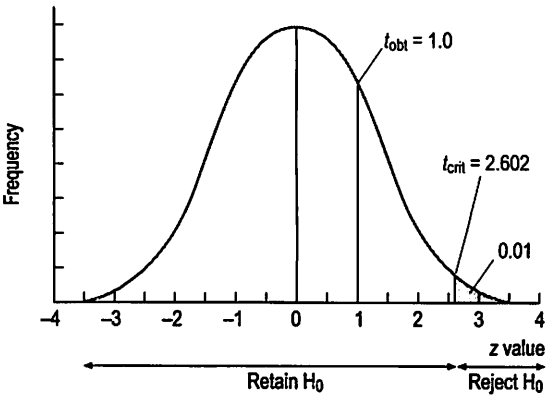

Figure 18.6 • Hypothesis testing: directional.

Conclusion

The above examples demonstrate the following points about statistical decision making:

- We are more likely to reject H_0 if we use a one-tail test (directional H_A) than a two-tail test (non-directional H_A). In effect, we are using the prediction of which way the differences will go to increase the probability of rejecting H_0 and therefore accepting H_A. Examples 1 and 2 demonstrate this point; in Example 1 we rejected H_0 with a directional H_A, while we retained H_0 in Example 2, with exactly the same data.
- The larger the sample size, n, the more likely we are to reject H_0 for a given set of data. Comparing Examples 1 and 3 demonstrates this; although μ, σ, and $\bar{X}$ were the same, where n was small we had to retain H_0. Also, when n is small, ($n < 30$), we must use the t test to analyse the significance of our sample mean being different.
- The more demanding the decision level (that is, if α is small), the less likely we are to reject H_0. To illustrate this point, repeat Example 2, but set $\alpha = 0.05$. Here, $z_{crit} = 1.96$ so that z_{obt} is greater than z_{crit}. Therefore, we can reject H_0, and accept H_A at a 0.05 level of significance. That is, with exactly the same data, we have rejected or accepted H_0, depending on the level of significance, α.

Errors in inference

When we say our results are statistically significant, we are making the inference that the results for

Table 18.2 Decision outcomes

Reality	Decision: reject H_0	Decision: retain H_0
H_0 correct (no difference or effect)	'False alarm' Type I error	Correct decision
H_0 incorrect (real difference or effect)	Correct decision	'Miss' Type II error

our sample are true for the population. It should be evident from the previous discussion that statistical decision making can result in incorrect decisions. There are two main types of inferential error: Type I and Type II.

A Type I error occurs when we mistakenly reject H_0; that is, when we claim that our experimental hypothesis was supported when it is, in fact, false. The probability of a Type I error occurring is less than or equal to α. For instance, in the previous Example 1 we set $\alpha = 0.01$. The probability of making a Type I error is less than or equal to 0.01; the chances are equal to or less than 1/100 that our decision in rejecting H_0 was mistaken. Therefore, the smaller α, the less the chance of making a Type I error. We can set α as low as possible, but by convention it must be less than or equal to 0.05.

A Type II error occurs when we mistakenly retain H_0; that is, when we falsely conclude that the experimental hypothesis was not supported by our data. The probability of a Type II error occurring is denoted by β (beta). In Example 3 we retained H_0, perhaps falsely. If n was larger, we might well have rejected H_0, as in Example 1. Type I errors represent a 'false alarm' and Type II errors represent a 'miss'. Table 18.2 illustrates this.

Table 18.2 illustrates that, if we reject H_0, we are making either a correct decision or a Type I error. If we retain H_0, we are making either a correct decision or a Type II error. While we cannot, in principle, eliminate these from scientific decision making, we can take steps to minimize their occurrence.

We minimize the occurrence of Type I error by setting an acceptable level for α. In scientific research, editors of most scientific journals require

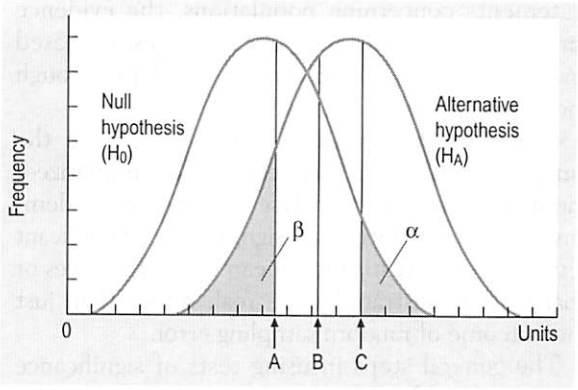

Figure 18.7 • Change of decision level to increase β and decrease α. As the decision criterion is moved from A to B to C, the relative frequency of Type I and Type II errors alters.

that α should be set at 0.05 or less. This convention helps to reduce false alarms to a rate of less than 1/20. Replication of the findings by other independent investigators provides important evidence that the original decision to reject H_0 was correct.

How do we minimize the probability of Type II error?

1. Increase the sample size, n.
2. Reduce the variability of measurements ($s_{\bar{x}}$, either by increasing accuracy (Ch. 12) or by using samples that are not highly variable for the measurement producing the data).
3. Use a directional H_A, on the basis of previous evidence about the nature of the effect.
4. Set a less demanding α, Type I error rate. There is a relationship between α and β, such that the smaller α, the greater β. This relationship is illustrated in Figure 18.7. Figure 18.7 shows that, as α decreases, β increases. Inevitably, as we decrease the Type I error rate, we increase the probability of Type II error. This is the reason why we do not normally set α lower than $p = 0.01$. Although a significance level such as $\alpha = 0.001$ would reduce 'false alarms' it would also increase the probability of a 'miss'.

Summary

The problem addressed in the previous two chapters was that, although our hypotheses are general

statements concerning populations, the evidence for verifying or supporting our hypotheses is based on sample data. We solve this problem through the use of inferential statistics.

It was argued in this chapter that, once the sample data have been collected and summarized, the investigator must analyse the findings to demonstrate their statistical significance. Significant results for an investigation mean that differences or changes demonstrated were real, rather than just the outcome of random sampling error.

The general steps in using tests of significance were explained, and several illustrative examples using the z and t tests for single sample designs were presented. A critical value is set for the statistic (in this case z_{crit}, t_{crit}) as specified by α. If the magnitude of the obtained value of the statistic (z_{obt}, t_{obt}) exceeds the critical value, H_0 is rejected. In this case, the investigator concludes that the data supported the differences predicted by the alternative hypothesis (at the level of significance specified by α). However, if the obtained value of the statistic is calculated to be less than the critical value, then the investigator must conclude that the data did not support the hypothesis. It was noted that following these steps does not guarantee the absolute truth of decisions made about the rejection or acceptance of the alternative hypotheses, but rather specifies the probability of the decisions being correct.

Two types of erroneous decisions were specified, Type I and Type II errors. A Type I error involves falsely concluding that differences or changes found in a study were real, that is, concluding that the data supported a hypothesis which is, in fact, false. A Type II error involves falsely concluding that no differences or changes exist, that is, concluding that the data did not support a hypothesis which is, in fact, true. It was demonstrated that the probability of these errors depends on factors such as the size of n, the directionality of H_A and the variability of the data.

The procedures of hypothesis testing and error were related to the logic of clinical decision making. The probabilities (α and β) of making Type I and Type II errors are interrelated. In this way, both researchers and clinicians must take into account the implications of possible error when setting levels of significance for interpreting the data.

Self-assessment

Explain the meaning of the following terms:

alternative hypothesis
critical value of a statistic
decision rule
directional or non-directional alternative hypothesis
null hypothesis
one-tail or two-tail test of significance
region of acceptance
region of rejection
significance (level of)
Type I and Type II error
z test or *t* test

True or false

1. The alternative hypothesis states that there is an effect or difference in the results.
2. If the probability of H_0 being true is greater than β, we can reject H_0.
3. Sampling distributions are used to enable the calculation of H_0 being true.
4. The critical value of a statistic is the value which cuts off the region for the rejection of H_0.
5. If the critical value of a statistic is less than the obtained or calculated value, we can reject H_0.
6. α is a probability, usually set at 0.01 or 0.05.
7. The *t* test requires that the sampling distribution of *t* should be normally distributed.
8. Hypothesis testing involves choosing between two mutually exclusive hypotheses, H_0 and H_A.
9. If α is set at 0.01 instead of 0.05, then the probability of making a Type I error decreases.
10. If we retain H_0, then we must conclude that the investigation did not produce significant results.
11. If *n* is greater than 30, the *t* test is more appropriate than the *z* test.
12. If results are statistically significant, the independent variable must have had a very large effect.
13. A directional H_A should be used when there is theoretical justification for the existence of a directional effect in the data.
14. When the results are statistically significant, they are unlikely to reflect sampling error.

15. It is impossible to prove the truth of H_A when using sample data as opposed to population data.
16. If we reject H_0 then we are in a position to accept H_A.
17. If α decreases (is made more stringent), then β increases.
18. If H_0 is true and we reject it, we have made a Type I error.
19. If H_0 is false and we reject it, we have made a Type II error.
20. If H_0 is false, and we fail to reject it, we have made a Type II error.

Multiple choice

1. Hypothesis testing involves:
   a deciding between two mutually exclusive hypotheses, H_0 and H_A
   b deciding if the investigation was internally and externally valid
   c deciding if the differences between groups was large or small
   d none of the above.
2. An α level of 0.01 indicates that:
   a the probability of falsely rejecting H_0 is limited to 0.05
   b the probability of Type II error is 0.01
   c the probability of a correct decision is 0.01
   d none of the above.
3. If α is changed from 0.01 to 0.001:
   a the probability of making a Type II error decreases
   b the probability of a Type I error increases
   c the error probabilities stay the same
   d the probability of a Type I error decreases.
4. If we reject the null hypothesis, we might be making:
   a a Type II error
   b a Type I error
   c a correct decision
   d *a* or *c*
   e *b* or *c*.
5. Statistical tests are used:
   a only when the investigation involves a true experimental design
   b to increase the internal validity of experiments
   c to establish the probability of the outcome of an investigation being due to chance alone
   d *a* and *b*.
6. The outcome of a statistical analysis is found to be $p = 0.02$. This means that:
   a the alternative hypothesis was directional
   b we can reject H_0 at $\alpha = 0.05$
   c we must conclude that H_A must be true
   d *a* and *c*.
7. When the results of an experiment are non-significant, the proper conclusion is:
   a the experiment fails to show a real effect for the independent variable
   b chance alone is at work
   c to accept H_0
   d to accept H_A.
8. It is important to know the possible errors (Type I or Type II) we might make when rejecting or failing to reject H_0:
   a to minimize these errors when designing the experiment
   b to be aware of the fallacy of accepting H_0
   c to maximize the probability of making a correct decision by proper design
   d all of the above.
9. An α level of 0.05 indicates that:
   a if H_0 is true, the probability of falsely rejecting it is limited to 0.05
   b 95% of the time chance is operating
   c the probability of a Type II error is 0.05
   d the probability of a correct decision is 0.05.
10. A directional alternative hypothesis asserts that:
    a the independent variable has no effect on the dependent variable
    b a random effect is responsible for the differences between conditions
    c the independent variable does not have an effect
    d there are differences in the data in a given direction.
11. If α is changed from 0.05 to 0.01:
    a the probability of a Type II error decreases
    b the probability of a Type I error increases
    c the error probabilities stay the same
    d the probability of Type II error increases.
12. If the null hypothesis is retained, you may be making:
    a a correct decision about the data
    b a Type I error
    c a Type II error
    d *a* or *c*
    e *a* or *b*.

13. When the results are statistically significant, this means:
    a the obtained probability is equal to or less than α
    b the independent variable has had a large effect
    c we can reject H_0
    d all of the above
    e *a* and *c*.
14. β refers to:
    a the probability of making a Type I error
    b the probability of $(1 - \alpha)$
    c the inverse of the probability of sampling error
    d the probability of making a Type II error.
15. Setting $\alpha = 0.0001$ would reduce the probability of Type I error. However, it would:
    a increase Type II error probability
    b increase the standard error of variance
    c reduce external validity
    d all of the above.
16. We retain H_0 if:
    a $|t_{obt}| \leq |t_{crit}|$
    b $|t_{obt}| > |t_{crit}|$
    c $|t_{obt}| < |t_{crit}|$
    d none of the above.
17. If α is changed from 0.01 to 0.001:
    a the probability of a Type II error decreases
    b the probability of a Type I error increases
    c the error probabilities stay the same
    d none of the above.

A researcher believes that the average age of unemployed people has changed. To test this hypothesis, the ages of 150 randomly selected unemployed people are determined (A). The mean age is 23.5 years. A complete census taken a few years before showed a mean age of 22.4 years, with a standard deviation of 7.6 (B).

Questions 18–22 refer to these data.

18. The alternative hypothesis should be:
    a $\bar{X}_A = \bar{X}_B$
    b $\mu_A = \mu_B$
    c $\mu_A \neq \mu_B$
    d $\bar{X}_A \neq \bar{X}_B$
19. The z_{crit} where $\alpha = 0.01$ is:
    a +2.58
    b +1.64
    c +2.33
    d −1.64
20. The obtained value of the appropriate statistic for testing H_0 is:
    a 2.88
    b 2.35
    c 1.84
    d 1.77
21. What do you decide, using $\alpha = 0.01$?
    a retain H_0
    b reject H_0
    c it is not possible to decide
    d *a* and *b*.
22. Therefore, the researcher should conclude that:
    a unemployed persons are getting older on average
    b there is no evidence supporting the hypothesis that the average age of unemployed people has changed
    c too many young people are unemployed
    d *b* and *c*.
23. When the results are not statistically significant, this means that:
    a the experimental hypothesis was not supported by the data at a given level of probability
    b the null hypothesis was retained at a given level of probability
    c the alternative hypothesis must have been directional
    d the investigation was internally valid
    e *a* and *b*.
24. If $\alpha = 0.05$ and the probability of the statistic calculated from the data is $p = 0.02$, then:
    a we should retain H_0
    b we should reject H_A
    c we should reject H_0 at $\alpha = 0.05$
    d we should restate H_0 so that the findings will become significant at the 0.05 level.

Chapter Nineteen

19

Selection and use of statistical tests

CHAPTER CONTENTS

Introduction

In the previous chapter, we examined the logic of hypothesis testing and the use of *z* and *t* tests for testing hypotheses about single sample means. There are numerous statistical tests available which are used in a conceptually similar fashion to analyse the statistical significance of the data. That is, all statistical tests involve setting up the relevant hypotheses, H_0 and H_A, and then, on the basis of the appropriate inferential statistics, computing the probability of the sample statistics obtained occurring by chance alone. We are not going to attempt to examine all statistical tests in this introductory book. These are described in various statistics text books or in data analysis manuals. Rather, in this chapter we will examine the criteria used for selecting tests appropriate for the analysis of the data obtained in specific investigations. To illustrate the use of statistical tests we will examine the use of the chi-square test (χ^2). This is a statistical test commonly employed to analyse nominally scaled data. Finally, we will briefly examine the uses of the Statistical Package for Social Sciences (SPSS) for data analysis in general.

The aims of this chapter are to:

1. Discuss the criteria by which a statistical test is selected for analysing the data for a specific study.

2. Demonstrate the use of the χ^2 test for analysing nominal scale data.
3. Explain how statistical packages are used for quantitative data analysis.

The relationship between descriptive and inferential statistics

As we have seen in the previous chapters, statistics may be classified as descriptive or inferential. Descriptive statistics are concerned with issues such as 'What is the average length of hospitalization of a group of patients?' Inferential statistics are used to address issues such as whether the differences in average lengths of hospitalization of patients in two groups are statistically significantly different. Thus, descriptive statistics describe aspects of the data such as the frequencies of scores, the average or the range of values for samples, whereas when using inferential statistics, one attempts to infer whether differences between groups or relationships between variables represent persistent and reproducible trends in the populations.

In Section 5 we saw that the selection of appropriate descriptive statistics depends on the characteristics of the data being described. For example, in a variable such as incomes of patients, the best statistics to represent the typical income would be the mean and/or the median. If you had a millionaire in the group of patients, the mean would give a distorted impression of the central tendency. In this situation the median would be most appropriate. The mode is most commonly used when the data being described are categorical. For example, if in a questionnaire respondents were asked to indicate their sex and 65 said they were male and 35 female, then 'male' is the modal response. It is quite unusual to use the mode only with data that are not nominal. As a rule, the scale of measurement used to obtain the data and its distribution determine which descriptive statistics are selected.

In the same way, the appropriate inferential statistics are determined by the characteristics of the data being analysed. For example, where the mean is the appropriate descriptive statistic, the inferential statistics will determine if the differences between the means are statistically significant. In the case of ordinal data, the appropriate inferential statistics will make it possible to decide if either the medians or the rank orders are significantly different. With nominal data, the appropriate inferential statistic will decide if proportions of cases falling into specific categories are significantly different.

Thus, when the data have been adequately described, the appropriate inferential statistic will follow logically. However, when selecting an appropriate statistical test, the design of the investigation must also be taken into account.

Selection of the appropriate inferential test

Before addressing the issue of the selection of the appropriate inferential statistical test, it is useful to reiterate the reason why a statistical test should be employed.

In many studies, inferential statistical tests are not required. For example, if a health care needs-assessment survey is conducted in a particular community, using a full population, the investigator might not be overly concerned with generalizing the results to other communities, or with demonstrating that certain relationships between variables are reliable. It may be enough to be able to say, for example, that '35% of the respondents indicated that they were dissatisfied with the existing level of medical services'. In this instance, descriptive statistics are all that the investigator requires since the complete population was studied. If, however, the investigator wishes to argue that certain differences between groups or that certain correlations between variables for a sample are generalizable to the population, then inferential statistical tests are necessary.

The inferential statistic provides the investigator with a means of determining how reproducible the obtained results are, by enabling access to a probability. The probability associated with the

value of an inferential statistic informs the investigator of the likelihood that the results obtained were due to chance factors, or if they are significant at a given level of probability.

Please note that we are not going to examine all of the numerous statistical tests available for decision making. Rather, the aim of this chapter is to examine the criteria used for selecting tests appropriate for the analysis of data obtained in investigations. To illustrate the use of statistical tests we will look at the χ^2 test, commonly employed for analysing nominal data. We examine the interpretation of findings which do not reach statistical significance and the relationship between statistical and clinical significance. In Chapter 20 we will consider some of the personal and social values implicit in making decisions concerning the actual adoption and use of treatments and diagnostic tests in clinical settings.

There is a variety of statistical tests, some of which are named in Table 19.1. The selection of the appropriate statistical test is determined by the following considerations:

1. The scale of measurement used to obtain the data (nominal, ordinal, interval, or ratio).
2. The number of groups used in an investigation (one or more).
3. Whether the measurements were obtained from independent subjects or from related samples, such as those involving repeated measurements of the same subjects.
4. The assumptions involved in using a statistical test, such as the distribution of the scores or the minimum required sample size.

Table 19.1 offers a sample of statistical tests in order to illustrate how statistical tests are selected for analysing data. Several points are worth noting.

1. It can be seen that appropriate tests are selected on the basis of the four criteria outlined above. When we have determined these four criteria for a given investigation, the cell containing the appropriate test can be readily selected. We might need additional criteria for deciding between two tests within a cell. For instance, we saw in the previous section that, if $n < 30$, we use the *t* test rather than the *z* test.
2. The tests appropriate for analysing ordinal and nominal data are called non-parametric or distribution-free. The tests for analysing interval or ratio data are called parametric tests. The parametric tests (for example, *z*, *t* or *F*) require that certain assumptions (such as normality and equal variance) be valid for the populations from which the samples were drawn. The non-parametric tests (e.g. χ^2, Mann–Whitney *U*) require few, if any, assumptions about the underlying population distributions.
3. Even before the data are collected, an investigator should have a good idea of which statistical test is appropriate for analysing the data. Sometimes, however, the distribution of the data is such that the test that was initially selected is found to be inappropriate.

Let us look at some examples to illustrate how statistical tests are selected.

An investigator wishes to evaluate the effectiveness of a new treatment in contrast to a conventionally used treatment. Assume that the outcome (dependent variable) is measured on a five-point

Table 19.1 Selection of tests of significance

Scale	Two groups		Three or more groups	
	Independent	Dependent	Independent	Dependent
Nominal	χ^2 test	McNemar's test	χ^2 test	Cochran's *Q* test
Ordinal	Mann–Whitney *U* test	Sign test	Kruskal–Wallis *H* test	Friedman two-way analysis of variance
Interval or ratio	*t* test (independent groups)	*t* test (dependent groups)	ANOVA (*F*) (independent groups)	ANOVA (*F*) (dependent groups)

ordinal scale. Each subject is assigned to one of the two treatment groups. Which test would the investigator use to analyse the significance of sample data when:

1. the measurement was ordinal?
2. there were two groups (new treatment, conventional treatment)?
3. subjects were independently assigned to a specific group?

By inspection of Table 19.1 the investigator would select the Mann–Whitney *U* test to analyse the significance of the data.

If we change the above example by stating that the dependent variable was measured on an interval scale, the appropriate test would now be a *t* test (for independent groups). Let us say that three groups were used (by the inclusion of a placebo group) by the investigator. Now, if the outcome measurement remained ordinal, the appropriate test for analysing the data is the Kruskal–Wallis *H* test. If, however, the outcome measures were interval, it follows from Table 19.1 that the appropriate test for analysing the results would be ANOVA (analysis of variance).

Finally, say that in the original example each of the subjects was treated with both the new and old treatments. Now, the data would have been obtained from the repeated measurement of the same subjects, and the appropriate statistical test would be the Sign test (ordinal, two groups, dependent).

Table 19.1 does not include all the available statistical tests and their uses. In fact, mathematical statisticians can generate inferential tests appropriate for a whole variety of designs. The basic idea is to use probability theory to generate appropriate sampling distributions in terms of which the probability of H_0 being true can be calculated, and the statistical significance of the findings evaluated.

Rather than examining all the tests and their underlying assumptions, we will look at the use of the χ^2 test in some detail. As well as being a very useful test for analysing nominal data, it (along with the *z* and *t*) illustrates how statistical tests are carried out to test hypotheses.

The χ^2 test

As shown in Table 19.1, χ^2 (chi-square) is appropriate for statistical analysis when:

1. variables were measured on a nominal scale
2. measurements were of independent subjects.

The χ^2 test is appropriate for deciding if proportions of cases falling into categories are different at a given level of significance.

The statistic, χ^2, is given by the formula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o = observed frequency for a given category and f_e = expected frequency for a given category, assuming H_0 was true.

The sampling distribution for χ^2 is a family of curves, which, like *t*, vary with degrees of freedom. The use of this inferential statistic is best illustrated by an example.

Suppose that an investigator is interested in finding out whether there is a difference in the relative frequency of different kinds of treatments currently offered to extremely depressed patients. A random sample of 150 patients is selected from a population of patients in Australia, and the type of treatment offered to them is determined from their medical records, as shown in Table 19.2.

The entries in each cell represent the frequency with which patients were given the various treatments. Thus, 45 patients were offered psychotherapy, 40 drugs and 65 electroconvulsive therapy. The χ^2 is the appropriate test for analysing these data. Let us follow the steps involved in hypothesis testing, as outlined in Chapter 18.

1. H_A: there is a difference in the population proportions for the three treatments. H_A is non-directional when we use χ^2.

Table 19.2 Treatments offered

Psychotherapy	Drugs	Electroconvulsive therapy
$n = 45$	$n = 40$	$n = 65$

2. H_0: there is no difference in the population proportions of the three treatments. The frequencies shown in each cell in the table occurred through random sampling from a population where there is an equal frequency of the three treatments.
3. Decision level, α: say the investigator sets a significance level of 0.05 for rejecting H_A ($\alpha = 0.05$).
4. Calculation of the statistic: χ^2_{obt} is the value of χ^2 calculated from the data obtained. To calculate χ^2_{obt}, we must determine f_e for each cell (f_o is, of course, determined by the data). If the null hypothesis is true, then our expectation is that the frequencies in each cell should be the same. In this case, $n = 150$, so that f_e should be 150/3 = 50, given that there are three cells. Let us show this in tabular form (Table 19.3).

We can now calculate χ^2_{obt}, by calculating $(f_o - f_e)^2/f_e$ for each cell, and then summing the values.

$$\chi^2_{obt} = \sum \frac{(f_o - f_e)^2}{f_e}$$
$$= \frac{(45 - 50)^2}{50} + \frac{(40 - 50)^2}{50} + \frac{(65 - 50)^2}{50}$$
$$= 0.5 + 2.0 + 4.5$$
$$= 7.0$$

The greater the discrepancy between f_e and f_o, the greater the calculated value of the chi-square statistic (χ^2_{obt}). The direction of the difference is of no account as the difference between f_e and f_o is squared.

5. Making decisions concerning H_0: the decision rule for χ^2 is similar to that of the *z* and *t* tests, as shown in the previous chapter:

$\chi^2_{obt} \geq \chi^2_{crit}$; reject H_0

$\chi^2_{obt} < \chi^2_{crit}$; retain H_0

Here, χ^2_{crit} is the critical value of the statistic χ^2, which cuts off a proportion of the sampling distribution equal to α. The value of χ^2_{crit} is obtained from the tables in Appendix C. To look up this statistic, we need to know:

(a) α, which was set at 0.05 for this example
(b) the degrees of freedom, df.

Note that with χ^2 the degrees of freedom with one variable is $k - 1$, where k stands for the number of categories or groups. In this instance, we have $k = 3$ (three treatments) so that df = 3 − 1 = 2. Now we can look up the tables in Appendix C. In this case, $\alpha = 0.05$ and df = 2, therefore $\chi^2_{crit} = 5.99$.

Here, since $\chi^2_{obt} > \chi^2_{crit}$ we can reject H_0 at a 0.05 level of significance. The investigator is in a position to accept H_A (that the three treatments are offered at different frequencies to depressed patients). Clearly, electroconvulsive therapy is given most frequently for the condition (in this hypothetical example).

Table 19.3 Treatments; f_e is shown in parentheses

Psychotherapy	Drugs	Electroconvulsive therapy
45	40	65
(50)	(50)	(50)

χ^2 and contingency tables

In the previous example of χ^2 we had clear expectations of the expected frequencies (f_e) and were dealing with only one variable. The χ^2 test is also relevant for analysing nominal data where f_e is not known, and where we are interested in the effects of more than one variable. Thus, χ^2 is a statistical test appropriate for deciding whether two variables are significantly related.

For example, an investigator compares the effectiveness of drug therapy with coronary artery surgery in males 55–60 years old, suffering from coronary heart disease. A sample of 40 patients consenting to the investigation is selected from this population, and randomly divided into the two treatment groups (drugs only or coronary artery surgery). The treatment outcome is measured in terms of survival over 5 years. The outcome of this hypothetical study is shown in Table 19.4.

Table 19.4 is called a contingency table. A contingency table is a two-way table showing the relationship between two or more variables. Note that the levels of the variables have been classified into mutually exclusive categories ('drugs or surgery' for the independent variable, and 'dead or alive' for the dependent variable, in this instance). The cells

Table 19.4 Contingency table showing obtained frequencies for a hypothetical study comparing survival after treatments

	Drugs	Surgery	*Row marginal*
Dead	11	8	***19***
Alive	9	12	***21***
Column marginal	***20***	***20***	***n = 40***

in the contingency table show the frequency of cases falling into each joint category (for example, 11 people who had 'drugs only' died during the 5 years). The row and column marginal scores are the sums of the frequencies. The row and column marginals necessarily add up to n, the sample size ($n = 40$ for this example).

Table 19.4 is called a two-by-two (2 × 2) contingency table. Depending on the number of categories (or levels) in each of the two variables, we might have 3 × 2 tables, 3 × 3 tables, etc. Let us now turn to analysing the data.

1. H_A: there is a difference in the proportion of patients surviving for 5 years following the two types of treatment.
2. H_0: there is no difference in the frequency of survival rates; any difference between observed and expected frequencies in the sample is due to chance.
3. Decision level: $\alpha = 0.05$.
4. Calculation of the statistic χ^2_{obt}

$$\chi^2_{obt} = \sum \frac{(f_o - f_e)^2}{f_e}$$

To make our explanation of the calculation easier, let us label the cells and the marginal values, as shown in Table 19.5. We calculate χ^2_{obt} by calculating f_e for each of the cells and then substituting this value into the equation for χ^2_{obt}. In order to calculate the expected frequencies, f_e, for each of the cells, we use the formula:

$$\frac{\text{Row total} \times \text{column total}}{n}$$

Table 19.5 General format for 2 × 2 contingency table

A	B	j
C	D	k
l	m	n

Table 19.6 Sample calculation of chi-squared

	f_o	f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
A	11	9.5	2.25	0.236
B	8	9.5	2.25	0.236
C	9	10.5	2.25	0.214
D	12	10.5	2.25	0.214
n	40			$\chi^2 = 0.90$

Substituting into the above formula for each of the cells:

$$A: f_e = \frac{j \times l}{n}, \quad B: f_e = \frac{j \times m}{n},$$

$$C: f_e = \frac{k \times l}{n}, \quad D: \frac{k \times m}{n}$$

Now f_o are the observed frequencies as in the data, summarized in contingency Table 19.5. Substituting the values for f_e and f_o for each cell is shown in Table 19.6.

5. Making decisions concerning H_0: the degrees of freedom for a contingency table are calculated by the following formula:

$$df = (r - 1)(c - 1)$$

where r = the number of rows and c = the number of columns. In this instance, given a 2 × 2 contingency table:

$$df = (2 - 1)(2 - 1) = 1$$

Now we can look up χ^2_{crit}, for $\alpha = 0.05$ and df = 1. From Appendix C, $\chi^2_{crit} = 3.84$. Therefore, since $\chi^2_{obt} < \chi^2_{crit}$ we must retain H_0: there is no statistically significant difference in the frequencies of survival over 5 years following the two kinds of treatments. Our data show that either there is no

difference in the outcomes of the two treatments or we made a Type II error.

The χ^2 test can be used to analyse the statistical significance of nominal data arising from experimental or non-experimental investigations. This non-parametric test can be used provided that two simple assumptions are met:

1. Each subject has provided only one entry into the χ^2 table; that is, each of the entries is independent.
2. The expected frequency (f_e) in each cell is at least five. Therefore, if the sample size is too small, χ^2 may not be used.

If either of these assumptions is violated, the use of χ^2 is inappropriate for statistical decision making. Assumption 2 is particularly important when the degrees of freedom is one (df = 1) for a contingency table.

Statistical packages

We have been looking at a few simple examples of establishing the statistical significance of the results. These calculations were presented only for teaching purposes. Some older applied statistics text books are crammed with complicated formulae for calculating dozens of different inferential statistics. Researchers now use statistical packages which have made statistical analysis simpler and more accessible to all researchers. The following steps are followed when using statistical packages:

Select a statistical package

There are many packages on the market, including Statistical Package for Social Sciences ('SPSS'), 'STATISTICA', 'Statsview' or various spreadsheets with useful statistical functions. Each program has its strengths and weaknesses and some researchers have formed strong attachment to specific programs. If you are a beginner, you should be guided by your thesis supervisor or your workplace mentor about the availability of packages.

Training

It is useful to learn how to use a package before you begin data analysis. Depending on your aptitude and experience, training sessions take 1 or 2 days. With more complex scientific packages such as 'STATISTICA' it might require long-term usage before one feels like an expert user.

Encode the raw data

Using all packages we begin by encoding the data. You have to be clear about issues such as which are your independent or dependent variables or the scaling of the data (continuous/discontinuous). That is, designs and measurement procedures used in your research project influence the encoding process (e.g. Schwartz & Polgar 2003, Ch. 1).

Identifying the statistical analysis required

Some degree of statistical knowledge is required beyond that covered in this book to enable you confidently to select and interpret the appropriate statistical analyses. Keep in mind that our book is introductory; it is expected that you will complete a more advanced statistics subject. For more complex analyses it might be useful to seek expert help.

Printout

You will select the appropriate statistical analysis from the 'menu' and print out the results of the analysis.

Interpretation

Finally you will need to interpret the 'printout' in relation to your research questions and/or hypotheses.

Let us look at a simple example. Say we are interested in the benefits of exercise for improving mobility in nursing home residents. A sample of 24 (n = 24) residents with reduced mobility are selected and give informed consent to participate. The participants are randomly assigned to either of two groups: 'E', which involves them undertaking an exercise programme suitable for improving mobility in elderly patients, and 'C', which involves undertaking alternative activities of equal duration. The outcome or dependent variable is measured as the distance in metres safely walked by the residents unassisted at the completion of the exercise and control programmes. The research hypothesis here is: exercise improves mobility in nursing home residents. We will look at a set of

created data to illustrate hypothesis testing using a statistical package (SPSS).

Let us follow the steps for analysing the data:

1. Assume that the SPSS package is readily available in your workplace and you are informed by your computing expert or supervisor.
2. You have had some training. Even if no formal training is available, self-help books such as *SPSS: Analysis Without Anguish. Version 11.0 for Windows* (Coates & Steed 2003) are very useful.
3. After accessing the spreadsheet for the program, we encode the data in the two columns representing 'E' and 'C'. We inform the program that the data are 'continuous' (ratio-scale data).
4. We refer to Table 19.1 to identify the required statistical analysis. Here we have:
   - two groups
   - independent groups
   - ratio-scale data.

   Therefore, we select the independent *t* test for analysing the data from the 'menu' of inferential tests available.
5. Printouts: Tables 19.7 and 19.8 for SPSS are based on printouts presented in Coates & Steed (2003, p.73). We changed the original example which contains additional information: the interested reader might wish to examine it further in Coakes & Steed (2003).

Table 19.7 Printout for descriptive statistics

Treatment	*n*	Mean	Standard deviation	Standard error mean
Control (C)	11	8.00	2.864	0.863
Exercise (E)	11	9.00	3.821	1.152

Table 19.7 shows key descriptive statistics:

- *n* refers to the sample size for each group
- mean ($\bar{X}$) for each group
- standard deviation (*s*) for each group
- standard error mean ($s_{\bar{x}}$) for each group. You will recall from Chapter 17 that $s_{\bar{x}} = s/\sqrt{n}$. Here, for the control group 'C' the standard error mean = $2.864/\sqrt{11} = 0.863$, as shown in the printout.

The group mean for the exercise group does in fact indicate a better overall performance. However, we are justified in concluding that exercise in general produces improved mobility in nursing home residents. To answer this question we must look at the results of the *t* test, as shown in Table 19.8.

The following information was presented:

- The *t* obtained was −0.695, the minus sign simply reflecting that the mean for the exercise group was less than the mean for the control group. A *t* value of 0.695 is quite 'small' in relation to the critical values for *t* shown in Appendix B.
- 'df' refers to degrees of freedom. For the control group's *t* test, $df = n_1 + n_2 - 2$ (Schwartz & Polgar 2003). Here, $df = 11 + 11 - 2 = 20$, as shown in Table 19.8.
- 'Sig. (2-tailed)' refers to the probability of a difference between the means being obtained by chance, i.e. *p* (H_0 is true) (see Ch.18). The decision rule is that we reject H_0 if $p \leq 0.05$. Here the calculated probability is 0.495 and therefore we retain H_0. We conclude that the results were not statistically significant.
- The 'mean difference' refers to the difference between the two sample means; that is, 8 − 9 = −1, as shown in Table 19.8.
- 'Standard error difference' refers to the standard error of the distribution of sample means. We will

Table 19.8 *t* test for equality of means

Treatment	*t*	df	Sig. (2-tailed)	Mean difference	Standard error difference	95% confidence interval of the difference	
						Lower	Upper
Control (C)	−0.695	20	0.495	−1.00	1.440	−4.003	2.003
Exercise (E)	−0.695	18.539	0.496	−1.00	1.440	−4.018	2.013

not discuss this statistic in detail. However, as shown in Chapter 17, the standard error enables us to calculate a 95% confidence interval; in this case for $\bar{X}_C - \bar{X}_E$.

- The lower and upper limits show that the true (i.e. population) difference ($\mu_C - \mu_E$) probably ($p = 0.95$) falls between −4.003 and 2.003. This is a wide range for a confidence interval and therefore indicates that residents who exercise might be able to walk either four extra metres or, for that matter, two metres less than those who don't exercise. Of course, any other difference between the two limits is possible. Clearly, such results are far too variable to attribute any clinical benefits to the exercise programme.

Summary

There are a variety of statistical tests available for analysing the significance of the obtained data. The statistical test appropriate for analysing a given set of data is selected on the basis of:

1. the scaling of the data
2. the dependence/independence of the measurements
3. the number of groups being studied
4. specific requirements for using a statistical test.

Generally, parametric and non-parametric statistical tests were distinguished on the grounds of the scaling of the data and the assumptions underlying the sampling distributions.

None of the individual statistical tests was discussed in detail, except the χ^2 test, which was presented as an example. Together with the discussion on the z and t tests in Chapter 18, the χ^2 test illustrates the principle that theoretical sampling distributions can be generated, and the probability of obtaining specific outcomes can be calculated. If the obtained value of the inferential statistic is greater than or equal to the critical value, the null hypothesis can be rejected at the level of significance specified by the Type I error rate (α). This is the case regardless of which particular statistical test is being used.

The retention of H_0 might reflect a correct decision, or a Type II error. Sample size is a factor which contributes to Type II error rate, as shown in both Chapters 18 and 20.

Self-assessment

Explain the meaning of the following terms:

chi-square
contingency table
expected frequency
non-parametric test
observed frequency
parametric test

True or false

1. Inferential statistics are used to decide if differences obtained in sample data are persistent, 'real' trends.
2. The selection of descriptive and inferential statistics is independent of the scaling of the data.
3. Inferential statistics must be used regardless of the nature and aims of an investigation.
4. Parametric tests are used to analyse the significance of interval or ratio data.
5. The use of non-parametric tests depends on the normal distribution of the underlying population.
6. Each statistical test entails the use of sampling distributions for calculating the probability of the obtained sample outcomes.
7. It is impossible to select an appropriate statistical test before the data are collected.
8. The number of groups being compared in an investigation influences the selection of the appropriate statistical test.
9. A basic assumption for using t is that the samples were drawn from a normally distributed population. A basic assumption of χ^2 is that the scores in each cell are independent.
10. When using χ^2, the closer the observed frequency for each cell is to the expected frequency, the higher the probability of rejecting H_0.
11. In order to reject the null hypothesis, $\chi^2_{obt} > \chi^2_{crit}$.
12. The χ^2 sampling distribution is a family of curves, the distribution of which varies with the degrees of freedom.
13. The χ^2 test is appropriate for testing hypotheses about proportions.
14. Each entry in a χ^2 table is a frequency.
15. The value of f_e is looked up in the appropriate χ^2 table.
16. If the f_e and f_o values are the same for each cell, χ^2_{obt} will not be statistically significant.

17. The decision level, α, is generally set at 0.05 or 0.01 with χ^2.
18. If we use sample data to calculate the values of f_e, then we use contingency tables for calculating χ^2_{obt}.
19. A 2 × 2 contingency table shows the relationship between two variables.
20. '*rc*' stands for the degrees of freedom for a 2 × 2 contingency table.

Multiple choice

1. In a study, three independent samples are compared and the dependent variable is measured on a ratio scale. A statistical test appropriate for analysing these findings is:
   a χ^2
   b Mann–Whitney *U*
   c *t*
   d ANOVA (analysis of variance).
2. In a study, two independent samples are compared and the dependent variable is measured on an ordinal scale. A statistical test appropriate for analysing these findings is:
   a χ^2
   b Mann–Whitney *U*
   c *t*
   d Wilcoxon.
3. Which of the following is a 'non-parametric' test?
   a ANOVA (analysis of variance)
   b *t*
   c *z*
   d Kruskal–Wallis *H*.
4. Which of the following is a 'parametric' test?
   a Median test
   b McNemar's test
   c *z*
   d Cochran's *Q*.
5. Which of the following tests is appropriate for analysing data where three or more groups were used?
   a *z*
   b *t*
   c χ^2
   d Sign test.
6. The larger the discrepancy between f_o and f_e for each cell in a contingency table:
   a the more likely it is that the results will not be significant
   b the more likely it is that H_0 will be rejected
   c the more likely it is that the population proportions are the same
   d the more likely it is that the population proportions are different
   e *a* and *c*
   f *b* and *d*.
7. For any given level of significance, χ^2_{crit}:
   a increases with increases in sample size
   b decreases with increases in degrees of freedom
   c increases with increases in degrees of freedom
   d decreases with increases in sample size.
8. A contingency table:
   a always involves two degrees of freedom
   b always involves two dependent frequencies
   c always involves two variables
   d all of the above
   e *a* and *b*.
9. Entries into the cells of a contingency table should be:
   a frequencies
   b means
   c percentages
   d degrees of freedom.
10. The degrees of freedom for a contingency table:
    a equal $n - 1$
    b equal $rc - 1$
    c cannot be determined if $r = c$
    d equal $(r - 1)(c - 1)$.
11. χ^2 should not be used with a 2 × 2 contingency table if:
    a df > 1
    b f_e is below 5 in any cell
    c f_o is below 5 in any cell
    d $f_e = f_o$
    e *b* and *c*.

An investigator is interested in determining whether there is a relationship between gender and susceptibility to a substance known to trigger an allergic response. 'Susceptibility' is measured as yes or no.

Questions 12–15 refer to this example. The raw data are presented in the contingency table below.

	Not susceptible	Susceptible	*Total*
Female	90	110	***200***
Male	60	140	***200***
Total	***150***	***250***	***400***

12. The value of χ^2_{obt} is:
    a 2.50
    b 8.09
    c 9.60
    d 11.05
13. The value of df is:
    a 2
    b 1
    c 3
    d need more information.
14. Using $\alpha = 0.05$, χ^2_{crit} is:
    a 3.841
    b 5.412
    c 2.706
    d −3.841
15. Using $\alpha = 0.05$, what is your conclusion?
    a Accept H_0: there is no relationship between gender and susceptibility.
    b Reject H_0: there is a significant relationship between gender and susceptibility.
    c Fail to reject H_0: the study does not show a significant relationship between gender and susceptibility.
    d Fail to reject H_0: this study shows a significant relationship between gender and susceptibility.
16. In selecting an appropriate statistical test:
    a *z* should be used as it is most powerful
    b *t* should be used as it takes the sample size into account
    c the choice depends on the design of the study
    d χ^2 should be avoided.
17. The χ^2 test requires that:
    a data be measured on a nominal scale
    b data conform to a normal distribution
    c expected frequencies are equal in all cells
    d all of the above occur.

The following information should be used in answering questions 18–25. Aerobics classes are conducted by the student union of a tertiary institution; although there are equal numbers of male and female students enrolled at the institution, it is observed that far more female than male students attend. A test is performed to see whether the proportion of the two sexes at the class is representative of the proportion of the two sexes enrolled at the institution as a whole. Of the 50 students who attend the classes, 10 are male. A χ^2 is conducted on these data.

18. What type of χ^2 will be conducted?
    a one-way
    b two-way
    c contingency analysis
    d parametric.
19. How many cells will there be in the χ^2 table?
    a 1
    b 2
    c 3
    d 4
20. What is the expected frequency of male students at the aerobic classes?
    a 10
    b 40
    c 25
    d 8
21. What is the obtained value of χ^2?
    a 4.5
    b 2.5
    c 18.0
    d 9.0
22. What are the degrees of freedom?
    a 1
    b 2
    c 49
    d 48
23. If α is set at 0.01, what is the critical value of χ^2?
    a 0.0201
    b 4.605
    c 9.210
    d 6.635
24. What statistical decision should be made on the basis of these data?
    a Reject null hypothesis.
    b Retain null hypothesis.
    c Increase α.
    d Increase size of sample.

25. What conclusion can be drawn on the basis of these data?
    a Overall, the tendency for more females than males to attend the classes is not statistically significant.
    b There is a statistically significant tendency for more females than males to attend the classes ($\alpha = 0.01$).
    c The aerobics classes should have their format changed to attract more male students.
    d There is a statistically significant trend ($\alpha = 0.01$) for differential attendance by the two sexes, but it is impossible to state the direction of this trend.

The following information should be used in answering questions 26–31. In a test of the effectiveness of phenothiazine in treating schizophrenia, 60 patients are randomly assigned to receive either the drug or a placebo; after 2 weeks of daily treatment each patient is assessed by the chief psychiatrist as 'improved' or 'not improved'. A 2×2 table is constructed to indicate improved number of patients falling into each category:

Assessment	Treatment	
	Phenothiazine	Placebo
Improved	20	10
Not improved	10	20

26. The degrees of freedom in this table are:
    a 1
    b 2
    c 3
    d 4

27. The obtained χ^2 is:
    a 1.33
    b 13.5
    c 8.71
    d 6.67

28. With α set at 0.05, the critical value of χ^2 is:
    a 3.841
    b 5.991
    c 6.635
    d 0.013

29. The correct statistical decision in this case is to:
    a reject H_0
    b retain H_0
    c decrease α
    d increase α.

30. The appropriate conclusion to be drawn from these data is:
    a Patients receiving the active drug are not significantly more likely to get an 'improved' rating than those receiving the placebo.
    b Those receiving the active drug are significantly more likely to be rated as 'improved' ($\alpha = 0.05$).
    c The drug cures schizophrenia.
    d The improvements cannot be due to the drug, as some people received the drug and did not improve.

31. Following the publication of this study, it is revealed that the psychiatrist who did the ratings of improvement was also the person who had assigned the patients to phenothiazine or placebo groups. What type of problem could have invalidated these findings?
    a Rosenthal effects.
    b Placebo effects.
    c Instrumentation effects.
    d All of the above.

Chapter Twenty 20

The interpretation of research evidence

CHAPTER CONTENTS

Introduction

When researchers have demonstrated statistical significance within a study, what have they actually established? Statistical significance suggests that it is likely that there is a similar effect or phenomenon in the population from which the study sample has been drawn. However, it is not correct to assume that statistical significance necessarily implies clinical importance or usefulness. We must also establish the *clinical* or *practical significance* of our results following the step of demonstrating statistical significance.

The aims of this chapter are to discuss the following:

1. Effect size, that is, how large the relationships or differences observed in the data are.
2. The relationship between effect size and statistical and clinical significance.
3. The determinants of statistical power.
4. Basic principles of clinical decision making.

Effect size

In a clinical intervention study, the effect size is the size of the effects that can be attributed to the intervention. The term *effect size* is also used more broadly in statistics to refer to the size of the phenomenon under study. For example, if

we were studying gender effects on longevity, a measure of effect size could be the difference in mean longevity between males and females. In a correlational study, the effect size could be represented by the size of the correlation between the selected variables under study. There are many measures or indicators of effect size selected on the basis of the scaling of the outcome or dependent variable (Sackett et al 2000).

The concept of effect size can be illustrated by results from two student research projects supervised by the authors.

Study 1: Test–retest reliability of a force measurement machine

In the first study, the student was concerned with demonstrating the test–retest reliability of a device designed to measure maximum voluntary forces being produced by patients' leg muscles under two conditions (flexion and extension). Twenty one patients took part and the reliability of the measurement process was tested by calculating the Pearson correlation between the readings obtained from the machine in question during two trials separated by an hour for each patient. The results are shown in Table 20.1. Both results reach the 0.01 level of significance.

The student was ecstatic when the computer data analysis program informed that the correlations were statistically significant at the 0.01 level (indicating that there was less than a 1 in 100 chance that the correlations were illusory or actually zero). We were somewhat less ecstatic because, in fact, the results indicated that approximately 69% ($1 - 0.56^2$) and 71% ($1 - 0.54^2$) of the variation was not shared between the measurements of the first and second trial. In other words, the measures were 'all over the place', despite statistical significance being reached. Thus, far from being an endorsement of the measurement process, these results were somewhat of a condemnation. This is a classic example of the need for careful interpretation of effect size in conjunction with statistical significance.

Table 20.1 Pearson correlations between trials 1 and 2

Flexion	Extension
0.56	0.54

Study 2: A comparative study of improvement in two treatment groups

The second project was a comparative study of two groups: one group suffering from suspected repetition strain injuries (RSI) induced by computer keyboard input and a group of 'normals'. An Activities of Daily Living (ADL) assessment scale was used and yielded a 'disability' index of between 0 and 50. There were 60 people in each group. The results are shown in Table 20.2.

The appropriate statistic for analysing these data happens to be the independent groups t test, although this is not important for understanding this example. The t value for these data was significant at the 0.05 level. Does this finding indicate that the difference is clinically meaningful or significant? There are two steps in interpreting the clinical significance of the results.

First, we calculate the effect size. For interval or ratio-scaled data the effect size 'd' is defined as:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

where $\mu_1 - \mu_2$ refers to the difference between the population means and σ, the population standard deviation.

Since we rarely have access to population data we use sample statistics for estimating population differences. The formula becomes:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_1}$$

Table 20.2 Mean ADL disability scores

	RSI group	Normals
Mean	33.2	30.4
Standard deviation	1.6	1.2

where $\bar{X}_1 - \bar{X}_2$ indicates the difference between the sample means and s_1 refers to the standard deviation of the 'normal' or 'control' group. Therefore, for the above example, substituting into the equation yields:

$$d = \frac{30.4 - 33.2}{1.2} = -2.33$$

In other words, the average ADL score of the people with suspected RSI was 2.33 standard deviations under the mean of the distribution of 'normal' scores. The meaning of d can be interpreted by using z scores. The greater the value of d, the larger the effect size.

Second, we need to consider the clinical implications of the evidence. It might be that the difference of 2.8 units of ADL scores is important and clinically meaningful. However, if one inspects the means, the differences are slight, notwithstanding the statistical significance of the results. This example further illustrates the problems of interpretation that may arise from focusing on the level of statistical significance and not on the effect sizes shown by the data.

When we say that the findings are clinically significant we mean that the effect is sufficiently large to influence clinical practices. It is the health workers rather than statisticians who need to set the standards for each health and illness determinant or treatment outcome. After all, even relatively small changes can be of enormous value in the prevention and treatment of illnesses. There are many statistics currently in use for determining effect size. The selection, calculation and interpretation of various measures of effect are beyond our introductory book, but interested readers can refer to Sackett et al (2000).

How to interpret null (non-significant) results

As we discussed in previous chapters, the researcher will sometimes analyse data that show no relationships or differences according to the chosen statistical test and criteria. In other words, the researcher cannot reject the null hypothesis. There are several reasons why the researcher may obtain a null result.

1. The trend or difference that the researcher originally hypothesized is incorrect.
2. The sample included in the analysis is unrepresentative, so the effect does not show up, although it exists in the wider population.
3. There are insufficient cases in the sample to detect the trend; this is especially a problem if the trends are subtle (i.e. the effect size is small).
4. The measurements chosen have very high or inherent random variability.

Therefore, if the researcher obtains a null result, one or more of the above explanations may be appropriate. There are, however, steps that can be taken to minimize the chance of missing real effects. In order to understand these measures, it is necessary again to invoke the table illustrating the possible outcomes of a statistical decision (as shown in Table 20.3).

There are four possible outcomes. On the basis of the statistical evidence you may: (i) correctly conclude there is an effect when there is indeed an effect; (ii) decide that there is an effect when there is not (this is a false alarm or Type I error); (iii) decide that there is not an effect when there really is (this is a miss or Type II error); or (iv) correctly decide there is not an effect when indeed there is not. The probability that researchers derive from their statistical tables is in fact the probability of making a false alarm or Type I error. This value does not tell us, however, how many times we will fail to identify a real effect. The probability of missing a real effect (or making a Type II error) is affected by the size of the actual effect and the number of cases included in our study. If you have large effects and large samples, the number of misses will be small. To detect a small effect size larger samples are needed.

Table 20.3 Statistical decision outcomes

Reality	Decision: effect	Decision: no effect
Effect	Correct	'Miss' Type II error
No effect	'False alarm' Type I error	Correct

Table 20.4 The relationship between effect size, sample size and decision making

Effect size	Sample size	
	Large	*Small*
High	Both statistical and clinical significance are likely to be demonstrated	Statistical significance might not be demonstrated, but clinical significance would be indicated
Low	Statistical significance would be likely, but the results might not indicate clinically applicable outcomes	Neither statistical nor clinical significance is likely. Statistically significant results might result in Type I error

Statistical power analysis

In statistical analysis we need to know how likely a miss is to occur. We can do this by calculating the statistical power of a design. The statistical power for a given effect size is defined as 1 – probability of a miss (Type II error or β). Thus, if the power of a particular analysis is 0.95, for a given effect size we will correctly detect the existence of the effect 95 times out of 100. Power is an important concept in the interpretation of null results. For example, if a researcher compared the improvements of two groups of only five patients under different treatment circumstances, the power of the analysis would almost certainly be low, say 0.1. Thus 9 times out of 10, even with an effect really present, the researcher would be unable to detect it.

It is essential to be careful in the interpretation of null results where they are used to demonstrate a lack of superiority of one treatment method over another, especially when there is a low number of cases. This may be purely a function of low statistical power rather than the equivalence of the two treatments. Unfortunately the calculation of statistical power is complicated and beyond the scope of this text. However, there are technical texts, such as Cohen (1988), that are available to look up the power of various analyses. There are also statistical programs that perform the same function. The best defence against low statistical power is a good-sized sample. Before quantitative research projects are approved by funding bodies or ethics committees, there is the requirement that sufficient data will be collected to identify real effects.

Effect size is a key determinant of both statistical and clinical significance. We are more likely to detect a significant pattern or trend in our sample data when a factor has a strong influence on health or illness outcomes. A very powerful treatment such as the use of antibiotics for bacterial infections could be demonstrated even in a small sample. Table 20.4 shows the association between effect size and sample size for determining statistical and clinical significance for the results of research and evaluation projects. The most useful results for clear decision making occur when both effect size and sample size are large. Where the effect size is large but the results are not statistically significant, it might be useful to replicate the study with a larger sample size. Unfortunately, in real research it might be difficult to obtain a large sample and the effect size, we discover, might be disappointingly small. It is for this reason that researchers make the best use of previous research and, if possible, complete pilot studies. The evidence from previous research and the results of pilot studies enable us to conduct power analyses for estimating the minimum sample size for detecting an effect if it is really there.

Table 20.5 shows how evidence for statistical and clinical significance can be combined to interpret the findings of a study. A clear positive outcome is when there is strong evidence for both clinical and statistical significance. In this case we are confident that the information obtained is clinically useful and generalizable to the population.

Table 20.5 How to interpret findings

Clinical significance	Statistical significance	
	Yes	*No*
Yes	Clear: strong evidence for treatment effect	Inconclusive: need for further research (e.g. with larger samples)
No	Inconclusive: suggests findings might not be meaningful – need further research	Clear: strong evidence for lack of a treatment effect

Acknowledgement: We are grateful to Dr. Paul O'Halloran (La Trobe University) for this table.

Another clear finding, provided that the sample size was adequate, is the lack of clear treatment effect. Such negative findings can be very useful in eliminating false hypotheses or ineffectual treatments.

Clinical decision making

The decisions confronting a clinician making a diagnosis on the basis of uncertain information are similar to the scientist's hypothesis testing procedure. As a hypothetical example, assume that a clinician wishes to decide whether a patient has heart disease on the basis of the cholesterol concentration in a sample of the patient's blood. Previous research of patients with heart disease and 'normals' has shown that, indeed, heart patients tend to have a higher level of cholesterol than normals.

When the frequency distributions of cholesterol concentrations of a large group of heart patients and a group of normals are graphed, they appear as shown in Figure 20.1. You will note that, if patients present with a cholesterol concentration between 1.0 and 2.4 mg/cc, it is not possible to determine with complete certainty whether they are normal or have heart disease, due to the overlap of the normal and heart disease groups in the cholesterol distribution.

Therefore, the clinician, like the scientist, has to make a decision under uncertainty: to diagnose pathology (that is, reject the null hypothesis) or normality (that is, retain the null hypothesis). The clinician risks the same errors as the scientist, as shown in Table 20.6.

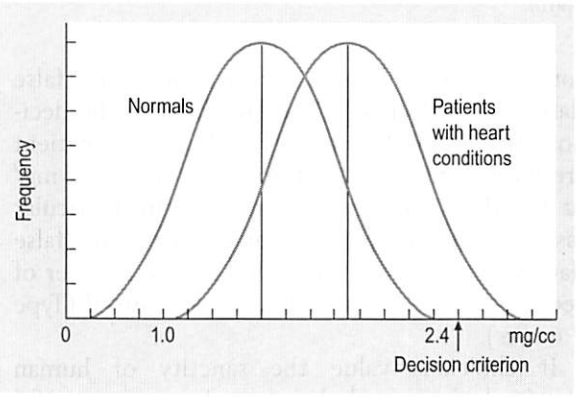

Figure 20.1 • Decision criterion: risk of Type II error (miss).

Table 20.6 Clinical decision outcomes

Reality	Decision: pathology	Decision: no pathology
No pathology	'False alarm' Type I error	Correct decision
Pathology	Correct decision	'Miss' Type II error

The relative frequency of the type of errors made by the clinician can be altered by moving the point above which the clinician will decide that pathology is indicated (that is, the decision criterion). For example, if clinicians did not

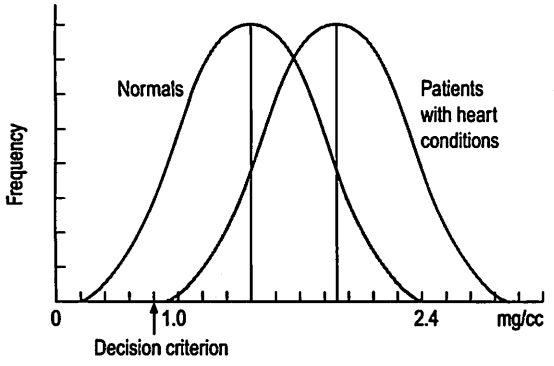

Figure 20.2 • Decision criterion: risk of Type I error (false alarm).

bother their colleagues or patients with false alarms (Type I errors), they might shift the decision criterion to 2.5 mg/cc (Fig. 20.1). Any patient presenting with a cholesterol level below 2.5 mg/cc would be considered normal. In this particular case, with a decision point of 2.5 mg/cc, no 'false alarms' would occur. However, a huge number of people with real pathology would be missed (Type II errors).

If clinicians value the sanctity of human life (and their bank balance after a successful malpractice suit) they will probably adjust the decision criterion to the point shown in Figure 20.2. In this case, there would be no misses but lots of false alarms.

Thus, most clinicians are rewarded for adopting a conservative decision rule, where misses are minimized, by receiving lots of false alarms. Unfortunately, this generates a lot of useless, expensive and sometimes even dangerous clinical interventions.

Summary

In the interpretation of a statistical test, the researcher calculates the statistical value and then compares this value against the appropriate table to determine the probability level. If the probability is below a certain value (0.05 is a commonly chosen value), the researcher has established the statistical significance of the analysis in question. The researcher must then interpret the implications of the results by determining the actual size of the effects observed. If these are small, the results may be statistically significant but clinically unimportant. Statistical significance does not imply clinical importance.

A null result (indicating no effects) must be carefully interpreted. It is possible that the researcher has missed an effect because of its small size and/or insufficient cases in the analysis. The statistical analysis measures the chance of correctly detecting a real effect of a given size. Thus, a null result may be a function of low statistical power, rather than there being no real effect. It is necessary to carry out a statistical power analysis *before* undertaking a research project.

There are several criteria, beyond statistical significance, which need to be considered before making decisions concerning the clinical relevance of investigations. The most important criterion is a large and consistent effect size. In addition to effect size, the determination of clinical significance is influenced by values and economic limitations concerning the administration of health care in a given community.

Self-assessment

Explain the meaning of the following terms:

clinical significance
effect size
null result
power analysis
social significance
statistical significance

True or false

1. If the effect size is small, clinical significance will be large.
2. In order to establish statistical significance, clinical significance must first be established.
3. In order to establish clinical significance, statistical significance must first be established.
4. The effect size in an analysis is directly measured by the size of the *p* value associated with the statistic.

5. High inherent variability in measures will promote the detection of effects within data.
6. If a statistical analysis has high power, this means that β will be low.
7. A 'miss' is a correct rejection of the null hypothesis.
8. If a statistical analysis has low power, the null hypothesis will be accepted more frequently.
9. Power = $1 - \beta$ (Type II error).
10. It is more difficult to detect small effects in data where the statistical power is high.

Multiple choice

1. If $\beta = 0.80$ and $\alpha = 0.1$, the power of an analysis equals:
   a 0.9
   b 0.7
   c 0.2
   d 0.65
2. In a study, there was a 1% difference in improvement of systolic blood pressure for two groups of patients receiving different treatments. This was statistically significant at $p = 0.05$. The results probably demonstrate:
   a clinical and statistical significance for the difference
   b clinical significance only
   c statistical significance only
   d neither clinical nor statistical significance.
3. A study of the relationship between family income and probability of occurrence of nutritionally related disorders demonstrated a correlation of 0.8 with $p < 0.001$. The results probably demonstrate:
   a clinical and statistical significance of the relationship
   b clinical significance only
   c statistical significance only
   d neither clinical nor statistical significance.
4. If the effect size in a study is large, the results are likely to have:
   a clinical and statistical significance for the difference
   b clinical significance only
   c statistical significance only
   d neither clinical nor statistical significance.
5. If the power of the statistical analysis of a study is high there will be:
   a fewer misses
   b fewer correct rejections
   c fewer correct acceptances
   d more misses.
6. The effect of large sample sizes in a study upon statistical power is generally to:
   a increase it
   b decrease it
   c not affect it.
7. In a study, the effect of larger sample sizes upon clinical significance is generally to:
   a increase it
   b decrease it
   c not affect it.
8. In a study, the effect of a larger sample size upon obtained statistical significance as measured by p is generally to:
   a increase it
   b decrease it
   c not affect it.
9. If a null result is obtained in an experimental clinical study, the clinical significance of any observed differences between treatment groups:
   a cannot be supported
   b can be supported if it is big
   c can be supported if it is small
   d should be determined by power analysis.
10. If a power analysis is not performed, is it sensible to accept a null result from a study at face value?
    a Yes
    b No.

Section Six

Discussion, questions and answers

Inferential statistical tests arise from the desire of clinical researchers to generalize from the data they have collected in a sample to the population from which the sample has been drawn. 'Is what I have found in my sample a true representation of the population (and hence other samples)?' is the basic question to be answered through the use of inferential statistical tests.

Inferential statistical tests all have the same basic format. The data are processed using the appropriate calculation procedure (often with the support of a computer program) and the value of the statistic is calculated. This value obtained is then compared with a table of known values in order to interpret the outcome of the statistical test. This is very much like the application of clinical tests where, in order to interpret the value of the test result, it is compared with a known standard. As with the clinician, the clinical researcher needs to know which test to choose in which circumstance. It would not be appropriate to try to measure the weight of a patient by giving her an X-ray. Similarly, it is not appropriate to use a χ^2 test when the *t* test is required. It is beyond the scope of an introductory text to have an extended discussion of the various types of statistical tests and when they might be used (although it should be noted that there are many fewer statistical than clinical tests). However, it is essential that the student understands the basic use of inferential tests.

Consider the following analysis using χ^2. This statistic is designed to test the relationship between variables with nominal or categorical scales (i.e. the values are categories).

The clinical researcher is using the χ^2 test to examine the relationship between length of stay in hospital and the rate of unplanned readmissions. These data are described more fully in Section 5. The goal is to determine whether there is a statistically significant association between the two variables. The raw data appear in Table D20.1.

As demonstrated in Section 5, we could use the Pearson correlation to analyse these data. However, to illustrate the use of χ^2 we will recode the data to categorical data and use this technique. The data will be recoded using the averages for each variable to convert the data from ratio data to categorical data. For example, all those cases (hospitals) with a mean length of stay of 13.6 days or greater will be considered as having an 'above-average' length of stay. Those cases (hospitals) with a stay below 13.6 days will be considered as having a 'below-average' length of stay. The same procedure will be followed for readmission rates of 4.47 or greater. These are the respective means for the two variables shown in Table D20.1. The recoded data appear as Table D20.2.

From these data we can construct a contingency table which shows the relationship between

Table D20.1 Average lengths of stay and readmission rates per 100 patients for patients with fractured neck of femur at 30 hospitals

Hospital	Average length of stay (days)	Unplanned readmission rates per 100 patients	Hospital	Average length of stay (days)	Unplanned readmission rates per 100 patients
1	11.100	7.800	16	13.200	4.500
2	11.200	6.500	17	13.200	5.500
3	11.200	4.300	18	13.300	4.100
4	11.200	5.500	19	13.700	3.200
5	11.700	5.100	20	13.900	3.400
6	12.100	5.200	21	14.100	3.500
7	12.100	5.000	22	14.200	3.400
8	12.100	4.900	23	14.200	6.000
9	12.300	4.800	24	14.900	4.400
10	12.400	3.400	25	15.300	3.300
11	12.400	5.000	26	15.400	4.200
12	12.500	4.300	27	15.400	4.300
13	13.100	3.900	28	15.500	4.100
14	13.100	3.300	29	16.300	3.200
15	13.200	4.700	30	22.400	3.300

the two newly coded variables. We do this by counting the number of times the 30 cases fall into the appropriate categories.

As can be seen from Table D20.3, only one hospital with an above-average length of stay had an above-average readmission rate, while 11 hospitals with above-average lengths of stay had below-average readmission rates.

These data can be subjected to χ^2 analysis. If these calculations are performed, we obtain a χ^2 value of 9.98, df = 1, $p < 0.01$. In other words there is a statistically significant association between length of stay and readmission rates for the 30 hospitals. This confirms the analysis conducted in Section 5.

Questions

1. From Table D20.1, how many hospitals have a below-average length of stay, if the average length of stay is 13.6 days? How many have an above-average length of stay? Why is it not 15 above and below?
2. From Table D20.3, why is there one degree of freedom in this analysis?
3. On the basis of this analysis, what would you conclude about the relationship between average length of stay in hospital and unplanned readmission rates for patients with fractured neck of femur at the 30 hospitals?
4. To what other groups of patients could these findings be generalized?

Table D20.2 Recoded average lengths of stay and readmission rates per 100 patients for patients with fractured neck of femur at 30 hospitals

Hospital	Average length of stay (days)	Unplanned readmission rates per 100 patients	Hospital	Average length of stay (days)	Unplanned readmission rates per 100 patients
1	Below average	Above average	16	Below average	Above average
2	Below average	Above average	17	Below average	Above average
3	Below average	Below average	18	Below average	Below average
4	Below average	Above average	19	Above average	Below average
5	Below average	Above average	20	Above average	Below average
6	Below average	Above average	21	Above average	Below average
7	Below average	Above average	22	Above average	Below average
8	Below average	Above average	23	Above average	Above average
9	Below average	Above average	24	Above average	Below average
10	Below average	Below average	25	Above average	Below average
11	Below average	Above average	26	Above average	Below average
12	Below average	Below average	27	Above average	Below average
13	Below average	Below average	28	Above average	Below average
14	Below average	Below average	29	Above average	Below average
15	Below average	Above average	30	Above average	Below average

Table D20.3 Contingency table of relationship between average length of stay and readmission rates at 30 hospitals for patients with fractured neck of femur

	Unplanned readmission rate	
Length of stay	Above average	Below average
Above average	1	11
Below average	12	6

Answers

1. In this sample, 18 hospitals have a below-average length of stay. Twelve hospitals have an above-average length of stay. Although that is how the median is defined, i.e. the score above which and below which half of the cases fall, the mean does not always fall at the exact half-way point of the sample.
2. The number of degrees of freedom in a contingency table is calculated by the formula:

$$\begin{aligned} df &= (\text{number of rows} - 1) \times \\ &\quad (\text{number of columns} - 1) \\ &= (2 - 1) \times (2 - 1) \\ &= 1 \end{aligned}$$

3. There is a moderately sized statistical association between average length of stay in hospital and unplanned readmission rates. That is, those hospitals with shorter lengths of stay for patients with a fractured neck of femur tend to have higher unplanned readmission rates.

4. It is difficult to say. The current data include patients with one condition only, i.e. fractured neck of femur. These patients may be atypical of other acute/surgical patients; they are likely to be older and perhaps more debilitated. These analyses would need to be extended to other types of patients before the results could be generalized. The country in which the study has been performed also needs to be considered, as procedures and incentives may vary considerably from one country to the next.

Section Seven

Dissemination and critical evaluation of research

SECTION CONTENTS

Having completed the analysis and interpretation of our data, we are now ready to communicate our results to the community of health scientists and professionals. Depending on the context in which our research was carried out, this entails the writing up of a report, a thesis or a 'paper' for a health sciences journal. The most common way of communicating research findings by established researchers is first to report the results at a professional conference and then to write a more formal paper for a relevant journal.

Each journal has its particular set of rules and requirements for how research projects should be written up for publication. In general, at least for quantitative research, the format for presenting our research follows the sequential stages of the research process outlined in the present book. This general format is outlined in Chapter 22, which includes a detailed discussion of the specific sections of a research paper and outlines some 'stylistic' considerations required by journal editors.

It is an ethical requirement that we report our results in an accurate and honest fashion. Before a paper is published in a reputable journal, it is critically evaluated by experts in the area (called referees) for errors or problems. However, sometimes problems remain unidentified. Ultimately, it is our task as health professionals to read important publications in a critical fashion. We owe it to our patients and clients to be cautious and critical concerning recent developments in theories and practices. However, being critical does not imply the adoption of a cynical or derogatory approach towards the work of other health researchers. We are aware that ethical and economic constraints, and the complex nature of the subject matter, as discussed in Section 2, make it difficult to ensure the external and internal validity of research projects.

The critical evaluation of a paper is not like judging a dog show; we do not simply award or subtract points for the strengths and weaknesses of a research project. Rather, if the information

is relevant to advancing the effectiveness of our practices, we have a stake in the project (even as readers). In this way, we take an active role in trying to 'repair' the problems which might cloud or invalidate the evidence.

In Chapter 23, we outline some of the criteria which we generally apply to evaluate specific sections of a research paper. We also discuss the implications of finding serious problems with the design, data collection and analysis and interpretations of a research project.

In effect, a single research project is rarely sufficient either to verify or falsify a theory, or to demonstrate the effectiveness of a treatment programme convincingly. Rather, we need to evaluate and summarize the literature as a whole, that is, conduct a literature review. Conflicting findings or gaps in the knowledge for a given area of health care identified in our literature review provide the impetus for further research, as outlined in Section 2. In this way research is a circular process.

Chapter Twenty One

21

Qualitative data analysis

CHAPTER CONTENTS

Introduction

Qualitative data are collected through techniques such as in-depth interviews, focus groups, participant observation and narratives spoken and written by the participants and researchers involved in the study. Qualitative data analysis refers to the processes by which researchers organize the information collected and analyse the meanings of what was said and done by the participants. In qualitative data analysis we bring our values, experiences and social understanding into analysing and constructing the meaning of what our respondents were telling us about their lives. At the same time, qualitative data analysis is principled; there are various explicit and shared strategies for summarizing and making sense of the data and checking the accuracy of our interpretation.

The aims of this chapter are to:

1. Describe the process of interpreting qualitative research data.
2. Describe the basic procedures involved in conducting content analysis, thematic analysis and semiotic analysis.
3. Discuss the comparative advantages and disadvantages of using different types of qualitative analyses.
4. Explain basic strategies for ensuring the accuracy of interpretations.

Understanding meaning in everyday life

Understanding people involves discovering the contents of people's minds – their beliefs, desires, intentions. There is nothing remarkable or supernatural being implied by this, simply that we infer mental contents by listening to and observing what people say and do, taking into account the social settings in which these actions occur.

For instance, one person says to another: 'Would you like to come in for a coffee?' What are the intentions of the speaker? Does he or she simply want to prepare the dark beverage and consume it in silence? Or should we look for 'hidden' or 'latent' meanings in order to understand the speaker's true intentions? Consider these two everyday scenarios (Polgar & Swerissen 2000):

1. You have been given a lift by a work mate who had to go out of his way to drive you home. Although you are tired, it seems the right thing to offer the driver refreshments. However, your intention is to be polite and acknowledge the colleague's effort; in fact you are hoping that the invitation will be refused. The 'hidden message' here is: 'Thank you and goodbye!' The worst-case outcome is that the colleague is too insensitive to read your intentions and stays around gossiping until midnight. Bad luck!
2. The British film 'Brassed Off' (1996) has a scene where a young woman is escorted home after a date by a young man. A dialogue was (approximately) as follows:

   She: Come up for a cup of coffee.
   He: I don't drink coffee.
   She: That's alright; I don't have any.

The above dialogue shows the nuances in the everyday use of language. Just as we sometimes misunderstand meanings and intentions in everyday life, we can also misinterpret the data produced by qualitative data collection. To avoid error we need to cross-check the accuracy of our interpretation.

Coding qualitative data

Qualitative data analysis frequently involves analysis of verbatim transcripts of dialogues and narratives. A common point of departure for analysing the transcript data is to develop a coding system. A coding system is to organize the data into specific classes or categories. There are two fundamental approaches to coding: predetermined and emerging with thematic analysis.

Predetermined coding

Predetermined coding uses predetermined categories to organize and analyse the transcripts.

For example, you might be conducting a survey to determine how clients experienced a rehabilitation programme at your workplace. Say that you conducted 20 in-depth interviews and produced a 100-page transcript representing what people said in these interviews. Considering your research aims, you might code the statements into three categories:

1. Satisfaction with the rehabilitation programme.
2. Dissatisfaction with the rehabilitation programme.
3. Neutral statements.

At the simplest level, analysing the first two categories would enable us to understand the reasons why the clients found the rehabilitation programmes to be satisfactory or unsatisfactory. This information could be useful for improving the programme. In most studies the coding system would be a good deal more elaborate.

Coding and thematic analysis

An alternative approach to using predetermined codes is to develop a coding system that identifies common themes as they emerge from the text. Different qualitative researchers advocate different approaches to coding but it typically involves the following steps. The researchers first study their materials, in this case transcripts, and develop a close familiarity with the material. During this process, all the concepts, themes and ideas are noted to form major categories. Often, the researcher will then attach a label and/or number to each category and record their positions in the transcript. Coding is an iterative process (we retrace our steps), with the researcher coding and recoding as the scheme develops. The researchers,

having developed the codes and coded the transcripts, then attempt to interpret their meanings in the context in which they appeared. The reporting of this process typically involves 'thick' or detailed description of the categories and their context, with liberal use of examples from the original transcripts.

Content analysis

Content analysis allows the quantification of units of meaning occurring in a text or a number of texts. Content analysis can be seen as a blending of quantitative and qualitative methods. The recognition and coding of meaning are qualitative, while the counting of the meaningful 'chunks' is quantitative. The 'meaningful chunks' can be words, sentences or paragraphs, that is, the units of language that were coded by the researchers from the narratives and dialogues.

For example, in an unpublished study one of the authors was interested in how leading newspapers were representing the use of stem cells in medical research. The following research questions were asked:

1. How extensive was the newspaper coverage of the medical use of stem cells?
2. What was the attitude of the newspapers (positive, negative or neutral) to the use of stem cells?

In relation to question 1, the data were collected by identifying relevant newspaper articles published on the topic and measuring the length of the columns. They were quantified by counting the number of articles published per month across the selected time interval. The data relevant to question 2 were obtained by identifying statements supportive or critical of using stem cells or simply 'neutral' descriptions of the nature and possible uses of stem cells. The column lengths for each of the three categories were measured and the percentages devoted to each were graphed across the months. Therefore, the content analysis provided evidence for the level of interest and changing attitudes of the media towards the use of stem cells. This evidence was relevant to understanding the cultural context in which government policy for using stem cells was being formulated.

The discussion of content analysis provides a good opportunity for raising the issue of computer-assisted data analysis. As the texts are often transcribed using personal computers the text is available in electronic form. This means that the text can be fed into a software package to assist with its analysis. For example, say that the data representing the contents of a hundred newspaper articles were transcribed into a software package. We could now introduce our codes and identify the segments of the whole text which use the relevant words/phrases/sentences. The segments can then be retrieved, examined or modified (cut and paste) on screen. Also, various frequency counts can be readily performed using software tools.

A detailed discussion for selecting and using computer packages is beyond the scope of the present book. Interested readers might find Liamputtong Rice & Ezzy (1999, pp 202–210) a useful introduction to selecting and using currently available software packages for expediting and improving qualitative data analysis in general (not only for content analysis). Liamputtong Rice & Ezzy have discussed the ambivalent attitude among qualitative researchers to computer-assisted data analysis. A key objection has been the distancing of the researcher from the creativity and surprising insights afforded by the more hands-on approaches. Another objection is that meanings of words and sentences sometimes do not follow dictionary definitions but rather have to be understood in the general context. The true meaning of certain subtle and ambiguous communications can be missed in crude and electronically conducted data analyses.

Content analysis is a technique that combines elements of both qualitative and quantitative approaches. We interpret the meaning of the text for developing our coding strategy for organizing or 'chunking' the text and then we use statistics to describe the quantities of text devoted to a specific point of view. Content analysis can be used to test hypotheses, for example hypotheses addressing media perspectives on embryonic stem cell research.

Thematic analysis, verstehen and grounded theory

Counting and hypothesis testing is not the essence of the qualitative approach. What we are trying to do is to see things from the perspectives of our informants and to explain their actions from their points of view. The German word *verstehen* is often used in phenomenological research to express the notion of 'putting ourselves in someone else's shoes' or attaining a strong empathy with their situation. Empathy with other people might seem quite simple, just something we do as human beings. It is worthwhile remembering, however, that sometimes we misunderstand how people feel or think, even when they are our close friends or family. In the same way, we might misunderstand the points of view of persons who are very different to us in age, gender, education, language and culture. Yet, it is essential to understand the points of views of the people to whom we offer health services. So how does 'verstehen' arise through qualitative health research?

First, as we described earlier, our data collection must use a technique (in-depth interviews, written materials, focus groups, etc.) which enables our respondents to express their point of view. Second, we can adopt a theoretical framework for explaining our understanding of the respondents' experiences. The key point, in the context of *grounded theory*, is that our explanations or theories must emerge inductively from the information provided by our informants. The theory is constructed gradually as more evidence is provided by additional informants. Third, the data are often analysed by coding and thematic analysis as we outlined earlier in this chapter.

A theme is a grouping of ideas or meanings which emerge consistently in the text. The themes emerging from the data illuminate the experiences of the informants and enable us to understand their points of view (verstehen). Let us consider an example of thematic analysis.

In a study titled 'The plight of rural parents caring for adult children with HIV', Fred McGinn (1996) studied the experiences of parents caring for their adult children with acquired immuno-deficiency syndrome/human immuno-deficiency virus (AIDS/HIV). In-depth interviews were conducted with eight mothers and two fathers from rural families involved in this task. The interview transcripts were analysed using a thematic analysis/grounded theory approach (Miles & Huberman 1984).

McGinn extracted three major themes:

1. *Physical and mental problems related to HIV/AIDS.* Here the parents discussed their experiences of their children's problems and the emotional consequences of physical decline and death, e.g:

   'He would fall over, so I would sit him in the wheelchair. And then from within a week in November he went from not being able to sit in the wheelchair to not getting out of bed. And he went from eating little bits of food along with taking a liquid nutrition to just liquid nutrition . . . and then he got to where he wouldn't swallow the liquid nutrition and he subsisted on just water and juices and Pepsi . . . and then in the end he even refused them: he wouldn't take anything . . . He just wasted away.'

2. *Stigma associated with having AIDS.* Because of the mode of transmission of AIDS and superstitious fears of contacting the condition, many of the parents found themselves socially isolated at such a very difficult time of their lives, e.g:

   'That Sunday, I never will forget. I asked him, 'Do you want anybody to know?' And I don't remember if he said no, but his head . . . he almost shook it off. No way did he want anybody to know what the real problem was. But I want you to know that that was a terrible, stressful time. People who came, who normally would be support for me . . . weren't. It was a real traumatic experience.'

3. *Health care.* This theme summarizes the difficulties of accessing necessary health services in rural settings. Even though there were serious deficiencies in health services, one mother reported:

   'As for the hospital, I couldn't have asked for a better hospital. There may have been

nurses who refused to work with him, I don't know, but the nurses that did come in were great . . . They even hugged and kissed him goodbye whenever he got well and left. They didn't act like they didn't want to be around him and I appreciated that. I think that's important.'

These three themes enable us to understand and empathize with the parents of these very sick young people. Also, they were the bases for recommending improvements in rural health care which directly address the needs of AIDS sufferers and their families in non-metropolitan environments.

Also, we must note that McGinn's paper reported the experiences of people in the mid-1990s, living in rural Canada. With improvements in the treatment and prevention of AIDS and a decrease in the stigma attached to the condition, the experiences of families caring for sufferers have improved. Because of differences and changes in practices and the cultural context, it is always important to note the time at and place in which interpretive research was carried out.

Interpretation and social context

As we have seen, qualitative data analysis is a systematic way of interpreting texts. There are many areas of study (e.g. history, politics, theology) where the interpretation of texts is an essential part of the research process. What these diverse disciplines have in common with qualitative health research is the recognition that the meaning of language and texts must be interpreted in a cultural context.

An example is *hermeneutics*, which is a method that was originally used to analyse the meaning of religious texts. Consider the meaning of the term 'god'. When the Romans spoke of Augustus Caesar as a 'god', they were referring to him as a hero who was immortal in the history of Rome. The use of the term 'god' by a polytheistic is quite different to meanings in the context of contemporary Judaeo–Christian or Muslim traditions. The meaning of the term must be interpreted in the context of the religious tradition (polytheistic, monotheistic) and the position of the speaker (believer, non-believer).

An important issue in reading texts is that they might have implicit (in addition to explicit) meanings. Semiotics is a method of textual interpretation which seeks to uncover the hidden, omitted meanings implicit in a text. In order to do this, we must adopt a theoretical framework in terms of which we can 'deconstruct' a text. The theoretical framework reflects our understanding of the culture within which the text was produced. You have probably read the book *Animal Farm* by George Orwell. There are several levels at which one can read this story; for example:

- A fairy tale about the imaginary lives of farm animals where animals have human traits and concerns.
- A morality tale in Aesop's style about how power corrupts and leads to betrayal.
- A critique of Stalinism and a retelling of the bloody history of the Bolshevik revolution and its social consequences in the Soviet Union.

In order to identify Orwell's book as a political critique, one needs to understand the historical/cultural context in which the author worked and lived.

To illustrate these points, we will examine a letter to the editor in a Melbourne newspaper by a woman writer who was apparently concerned about the physical and mental health of young men:

They're just asking for it.

Since the weather improved, it seems that young men all over the place are discarding their shirts and going about half-naked. I worry for them. Do they have any idea of what a provocative and inviting image they put across?

To my mind, they would be doing themselves a far greater service if they would just compromise a little and get dressed properly. It might not seem fair, and it might be less comfortable, but at least then there wouldn't any

longer be the danger of urge-driven women raping young men because of the confusing visual signals they so often put across.

(In Polgar & Swerissen 2000).

Let us analyse the text consistent with a procedure outlined in Daly et al (1997). First, let us analyse the explicit content of the letter.

1. *Tone*. Serious and condescending as that used by authority figures such as teachers and magistrates; '. . . now, see here young man, this is for your own benefit . . .' type of communication.
2. *Language*. Moralistic (e.g. 'dressed properly', 'going about half-naked') and calling for responsibility ('compromising a little'). Also the language is alarmist, predicting that men non-compliant to a dress code will be assaulted.
3. *The aim*. The explicit aim of the letter is to warn young men of the dire consequences of dressing immodestly and thereby inviting attention by 'urge-driven women'.
4. *Repetition of ideas*. The main idea seems to be that a scantily dressed man is sexually provocative to women. Another notion is that women are struggling to control powerful sexual urges. It is implied that men should accept responsibility for suppressing these urges in women. If men dress immodestly then they have to accept the consequences.
5. *Themes*. The first basic explicit theme is the importance of men taking responsibility in projecting a safe, chaste image. The second is the power and danger of women's sexual urges which can explode into assault when provoked by scantily dressed males. An underlying theme which you might have detected is one of 'blaming the victim'; if men are assaulted it is their fault, they should have been more careful.
6. *Oppositional elements*. If men ignore the letter writer's message and move towards the choice of scanty dress then they are putting themselves at risk. That is, modest dress means safety while immodest dress means assault. Another dichotomy is gender: women are sexually powerful and dangerous; men are presented as naive victims lacking any defined sexuality. In different ways the overt themes emerging from the text are demeaning of both males and females.

You may have different views about how best to interpret the text. As Daly and her colleagues (1997, p. 183) noted: 'Let us now make some basic semiotic moves across the data'. Let us interrogate the 'data' further using the six points suggested by Daly et al (1997).

- Is the content of the letter preposterous? Are there scantily dressed male construction workers being dragged into alleys by out-of-control schoolgirls? Are there gangs of libidinous females cruising our streets with evil purposes on their minds? Preposterous! The incidence of assault by women is, to all intents and purposes, very low, regardless of how men choose to dress. Therefore, the letter is unsound or it may be a parody.
- In order to understand the meaning of the letter, we play a language game as follows: read 'male' for 'female' in the text. The story now reads quite differently; in fact it resembles a more usual story told to women concerning their responsibility for ensuring that men don't assault them.
- One might propose that the latent agenda for the letter was to ridicule the notion that victims are in some way responsible for the violence of the perpetrator.
- The apparent hero of the explicit story was the author, the caring woman dispensing advice to young men to keep themselves safe by dressing in a chaste fashion. In the implicit story the villains are people who blame women for contributing to violence simply by the clothes they wear.
- What is missing from the original story? Or what was introduced? The writer introduced the notion of female sexuality as an urge that could transform at the slightest provocation into violence. If this notion is ludicrous for females, the question is, how can it be tenable for males? You might ask that if the true intention of the author was to denounce myths of male sexuality then why didn't she say so directly? This is like asking why George Orwell wrote a fairy tale with talking farm animals rather than a direct denouncement of totalitarianism and Stalinist terror. It is a question of how we use language; we use metaphors, parables, hyperboles and so on for expressing ourselves in an interesting, colourful fashion. Semiotics is one of the ways for interpreting the meanings that might be hidden or camouflaged in the original narrative.
- A basic principle for semiotic analysis is selecting a theoretical framework in terms of which we can deconstruct the original narrative and identify its hidden, repressed or mystifying elements. The key to the previous analysis was that the basic idea underpinning the argument (that immodestly dressed males are in danger of being assaulted by out-of-control women) was false and absurd.

Our interpretation of the meaning is that the text is a parody of the victim-blaming discourses in patriarchal societies. By adopting a feminist theoretical framework we are in a position to identify the hidden meaning of the text and infer the intentions of the author.

- There is always the possibility that we have misinterpreted the text and misrepresented the intentions of the author. What if she was genuinely concerned about the welfare of young men? The fact of the matter is that there are no absolute guarantees. It might be that, regardless of his well-known interest in political affairs, George Orwell was simply intending to create a children's story when he wrote *Animal Farm*.

In the next part of this chapter we will outline some strategies for ensuring validity and reliability for qualitative research.

The accuracy of qualitative data analysis

How can we be sure that the themes we identified in a text accurately reflect the actual views of the participants? Also, how do we know that similar themes would emerge from the reports of other people who had similar experiences to our sample?

There are a number of qualitative researchers who ensure that the collection and interpretation of their evidence are carried out in a methodologically rigorous fashion. The following represent some of the key methodological criteria for conducting qualitative research (Lincoln & Guba 1985):

1. *Data saturation*. This refers to ensuring that we have collected sufficient data from our respondents. Saturation occurs when the themes and ideas emerging from the text become repetitive and we are confident that the inclusion of new participants or further engagement with current participants will not lead to novel themes or interpretations.
2. *Credibility*. Checking if the interpretation of the evidence is judged as accurate by both the research participants and also independent clinicians or scholars. In other words, does your interpretation make sense and if not, why not?
3. *Auditability*. This refers to each of the steps of the research process being clearly described, so that an independent scholar can critique the research process from its beginning to the analysis and interpretation of the data. The auditor confirms or rejects the researcher's methodology.
4. *Triangulation*. This strategy involves the use of multiple independent methods for collecting data and checking if the themes and interpretations emerging from these different methods are consistent and matching. For example, in order to evaluate client satisfaction with a health service you might use three different data collection strategies (Fig. 21.1). It is useful (but not essential) to use three different data collection methods. The researcher might, if appropriate, use both qualitative (e.g. in-depth interviews) and quantitative (e.g. structured questionnaires) methods for data collection. These are called 'mixed' designs (see also Chs 1 and 9).

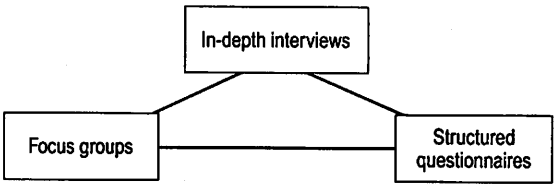

Figure 21.1 • Triangulation using different approaches to data collection.

As you can see, the methodological concerns in qualitative research are parallel to those of quantitative research (i.e. reliability and validity). However, because of the differences in the way the two types of research are conducted, the terminology for describing the methodological principles is somewhat different.

Summary

There are different approaches to analysing qualitative data depending on the theoretical framework and data collection strategies adopted by the researchers. However, as we saw in this chapter, there are several common aspects to qualitative data analysis:

- 'Immersion' in the data; reading and re-reading the texts to develop a sense of what it is that respondents are trying to say.

- Developing a coding system and identifying the themes emerging from the text.
- Using these themes as a basis for insight, empathy (verstehen) with the experiences, emotions and thinking of the respondents.
- Interpreting and theorizing the respondents' experiences in the context of cultural or historical settings.
- Ensuring the accuracy of the interpretations by cross-checking themes and explanations with other sources of data, other researchers' interpretations of the data, and the respondents.

Self-assessment

Explain the meaning of the following terms:

auditability
coding
coding and thematic analysis
content analysis
credibility
data saturation
grounded theory
hermeneutics
predetermined coding categories
semiotics
theme
theoretical framework
triangulation
verstehen

True or false

1. Hermeneutics refers to understanding meanings in their cultural contexts.
2. Quantitative data analysis is a far more subjective process than qualitative analysis.
3. Qualitative data analysis is a systematic way of 'reading' another person's mental states.
4. In the context of qualitative data analysis, the term 'text' refers to the transcripts of in-depth interviews, focus groups, etc.
5. Predetermined coding is based on categories of meaning emerging as the texts are analysed.
6. When coding and thematic analysis are carried out together, it is essential to maintain the first coding system chosen to avoid thematic confusion.
7. Content analysis often employs statistical analyses of the data.
8. A basic objection to computer-assisted coding and analysis is that this technique might obscure more subtle meanings communicated by the respondents.
9. 'Verstehen' is a German word referring to the precise dictionary definitions of the words used in a text.
10. A fundamental objective of qualitative data analysis is to enable the researcher to see health-related events from the perspectives of the respondents.
11. A theme is a specific idea which reflects the unique experiences of only one respondent.
12. Hermeneutics is a form of religious practice requiring withdrawal from everyday life.
13. Semiotics requires the adoption of a theoretical framework for identifying hidden meanings in a text.
14. The results of a qualitative study represent experiences which are useless unless they are true for all times and places.
15. 'Auditability' refers to the clarity of the methodology employed in conducting a qualitative research project.
16. 'Data saturation' addresses an issue analogous to external validity in quantitative research.
17. The results of quantitative and qualitative research cannot be compared with each other.
18. As there are no absolute guarantees for the truth of our interpretation of other people's experiences, all qualitative research lacks credibility.
19. Qualitative analyses produce evidence relevant for constructing social theories, but have no relevance to evidence-based health care.
20. In conducting evidence-based health care, we should combine the results of qualitative and quantitative research to identify the best practice for our clients.

Multiple choice

1. Which of the following statements is *not* an aspect of 'intepretivist' approaches?
   a Society exists as the result of meaningful social interactions.
   b It is the actors themselves who are best able to define social situations.
   c A society is defined by the sum of the total behaviours of individuals constituting a population.

d Data collection requires a degree of empathy with the way in which people experience their social situation.

2. Themes are:
   a the planning process used by the researcher
   b organized around interview data
   c the result of the research question
   d ideas represented in a pattern.
3. A qualitative researcher could develop codes for interview data:
   a after conducting all of the interviews
   b before collecting data
   c both before and during data collection
   d all of the above.
4. Latent content of themes in qualitative data refers to:
   a the hidden or underlying themes
   b the second level of data collection
   c what was not said directly in the data
   d both *a* and *c*.
5. Which of the following statements is true in coding data?
   a Words, concepts and themes are selected.
   b The researcher must wait until all data are collected before establishing codes.
   c Concepts are weighted as far more important than themes.
   d The researcher must be trained in using a qualitative computer program.
6. In a qualitative study of public perceptions of people with mental illness, you intend to identify the social significance of being perceived as mentally ill. You are well aware that most of your respondents will attempt to *appear* caring and tolerant, even when they hold strong prejudices against people with mental illness. A useful approach for analysing the data in relation to the above question is called:
   a grounded theory
   b semiotic analysis
   c content analysis
   d typological categorization.
7. The term 'grounded theory' refers to:
   a any sociological theory that is based on empirical evidence
   b any theory, sociological or otherwise, that is based on empirical evidence
   c a systematic way of formulating theories which are the sources for empirically testable hypotheses
   d generating and testing interpretive theories during the course of data collection.
8. Generally, an interpretive theory aims to explain:
   a the social causes of human action
   b the experiences of people in the context of their cultural settings
   c why people do what they do
   d the factors which interact in a society for generating human personality.
9. The process of interpreting texts (such as the Bible) in the cultural contexts in which they were written is called:
   a hermeneutics
   b content analysis
   c semiotics
   d discourse analysis.
10. The term 'data saturation' as used in qualitative research is most closely related to what in quantitative research?
    a Sample size.
    b Validity of the evidence.
    c Reliability of the evidence.
    d 'Scaling' of the data (nominal, ordinal, etc.)
11. The term 'triangulation' as used in qualitative research is most closely related to what in quantitative research?
    a Reliability.
    b Validity.
    c Hypothesis testing.
    d Descriptive data analysis.
12. Which of the following is *not* a traditional approach to qualitative research?
    a Ethnomethodology.
    b Grounded theory.
    c Phrenology.
    d Phenomenology.

Chapter Twenty Two

22

Presentation of health science research

CHAPTER CONTENTS

Introduction

Knowledge in the health sciences is the sum of the individual efforts of investigators working all over the world. Professional journals in science and health care provide the dominant medium for disseminating information about the outcome of specific investigations. Investigators must report their procedures and results in an accurate and complete fashion. In this chapter, we outline the format and style generally followed for presenting the results of empirical investigations.

The specific aims of this chapter are to:

1. Describe the conventional way in which quantitative research is presented for publication.
2. Discuss the style or language used to describe research.
3. Outline briefly the way in which research papers are selected for publication.

The structure of research publications

The format of a professional publication reporting empirical research reflects the stages of the

Table 22.1 Format of research publications and the research process

Publication format	Research process
Title	
Abstract	
Introduction	Research planning
Method: Subjects Apparatus Procedure	Design Measurement
Results	Descriptive statistics Inferential statistics
Discussion	Interpretation of the data
References	
Appendices	

research process discussed in this book. Table 21.1 represents the relationship between the stages of research and the commonly used publication format. This format is generally used to report quantitative empirical research, although you will find that some variations on this theme are adopted by some professional journals. This format is not necessarily followed for certain types of scholarly communications, such as for qualitative research, theoretical papers or literature reviews. In the subsections following, we examine in detail each of the components of a research report shown in Table 21.1.

Title and abstract

The *title* is a descriptive sentence stating the exact topic of the report. Many titles of research reports take one of the following two forms:

- *y* as a function of *x*
- the effect of *x* upon *y*.

In causal research, such as experiments, *y* refers to the dependent variable being measured and *x* refers to the independent variable being manipulated. For example:

- The incidence of alcoholism in health professionals as a function of work-related stress.
- The effect of major tranquillizers on the cognitive functioning of persons with schizophrenia.

For descriptive or qualitative research the title should inform the reader about the groups being studied and the characteristics being reported, for example: 'The attitudes of physicians to the professional functions of podiatrists'. In general, titles should be concise and informative, enabling a prospective reader to identify the nature of the investigation. Immediately below the title should appear the name(s) of the investigator(s) and affiliation.

The *abstract* is a short (not more than 250 words) description of the entire report. The purpose of this section is to provide the reader with a general overview of the communication. It should provide enough details to enable the reader to decide whether or not the article is of interest. This section can be difficult to write because of its precise nature. When writing an abstract you should include:

1. A brief statement about previous findings which led you to conduct your own research.
2. The hypothesis and/or aim of your research.
3. Methods, including subjects, apparatus and procedure.
4. A short description of what you found and how you interpreted your results.
5. What you concluded.

In some journals, this section may appear at the end of the manuscript in the form of a summary. For our purposes, however, we will treat this section as an abstract.

The title and the abstract together are important and should contain key words that enable the efficient retrieval of the information.

Introduction

The introduction is equivalent to the planning stages of research, discussed in Section 2. A good introduction will set the stage for the hypotheses being tested. It should do this by discussing the

theoretical background of the problem under consideration and evaluating the relevant research done previously. The introduction thus serves as a link between the past and the present.

Generally, all aspects of the literature cannot be covered in a relatively brief research paper, therefore the review of past research is done with a bias towards only those aspects of the problem which are of direct relevance to your report. In this way the hypotheses being tested can be derived in a logical manner. For this reason, a good introduction starts out by making a few general statements about the field of research, leading logically to a narrow and specific set of statements which represent the aims or hypotheses. The final paragraph of the introduction should state the precise aims or the hypotheses being investigated.

Method

The purpose of the method section is to inform the reader of how the investigation was carried out. It is important to remember that the method section should contain enough detail to enable another researcher to replicate your investigation. (Of course, replications may not be feasible for a unique event, such as a case study of a specific individual.) Conventionally, three subsections are used: subjects, apparatus and procedure.

- *Subjects*. Three questions must be answered concerning the subjects: who were they, how many were there and how were they selected? Specific information must be given concerning the subjects, as results may vary from one sample to another.
- *Apparatus*. A description of all equipment, including questionnaires, etc., used in the research must be provided. If it is commercially available, provide the reader with the manufacturer's name and the commercial identification of the equipment. Alternatively, if the equipment was privately made, provide the reader with enough information to allow replication. Measurements and perhaps a diagram will be necessary.
- *Procedure*. Once again, this section should provide enough information for other researchers to replicate the investigation. Details of how the research was carried out should include how subjects were assigned to groups, how many subjects per group, the experimental procedure and a description of how the data were collected.

In a sense, the method section should read like a cookbook. The 'subjects' subsection describes the ingredients. The 'apparatus' subsection describes the equipment necessary for baking (note we did not say 'cooking' the experiment) and, finally, the 'procedure' subsection describes how the ingredients were mixed to produce the final outcome: the results section.

Results

The results section presents the findings of the investigation and draws attention to points of interest. Raw data and statistical calculations are not presented in this section. Rather, we use the principles of descriptive and inferential statistics to present the summarized and analysed data: graphs, tables and the outcomes of statistical tests are presented in this section. It is essential that all the findings are presented and that the graphs and tables are correctly identified.

Discussion

The discussion section restates the aim(s) of the investigation and discusses your results with reference to the aims or experimental hypothesis stated in the introduction. Did you find what you expected? How do the present results relate to previous research?

It is important to remember that one experiment in isolation cannot make or break a theory or establish the effectiveness of a practice. Thus, the discussion should connect the findings with similar studies and especially with the theory underlying such studies. If unexpected results were obtained, possible reasons for the outcome (such as faulty design and controls) should be discussed. By this, the discussion will point the way to further problems which remain to be solved. Unconstructive, negative or unimportant criticism should be avoided, so that the report does not end with long discussions of possible reasons for the outcome. Brief, concise discussion is more appropriate.

In the conclusion, which is usually the last paragraph of the discussion section, the main findings

are summarized and suggestions made for further research. For example, you may have demonstrated certain phenomena which may have implications for explaining broader concepts which can be empirically tested. You are therefore taking your findings and generalizing them to phenomena not directly tested in the present research.

References and appendices

It is expected that all the literature discussed in the paper is listed in the references section. This enables your reader to evaluate your sources. You should refer to appropriate style manuals for information on how references should be listed. Sufficient information must be provided for an interested reader to be able to identify and retrieve the sources. In addition, a report may include labelled appendices. These might include a full description of questionnaires or other measuring instruments, raw data or statistical calculations if required.

The style of research publications

It is essential that you read research publications in your professional area to gain a 'feel' for the appropriate style of writing. In general, the following points should be kept in mind when writing reports:

1. Avoid long phrases or complicated sentences. Short, simple sentences are far more easily understood by your reader. In other words, try not to posture but to communicate.
2. Use quotations sparingly; put ideas in your own words. Quotations are only used when it is necessary to convey precisely the ideas of another researcher, for instance while conducting a critique of a paper.
3. Use past tense when writing your research report.
4. Use an objective style, avoiding personal pronouns wherever possible.
5. Make sure you are writing to your audience; if the material is specialized or difficult, explain it clearly.
6. Make sure that you are concise and clear; do not introduce issues and concepts which are not strictly relevant to reporting your investigation. Raising interesting but superfluous issues might distract and confuse your reader.

In general, you should aim to improve your report writing and your ability to communicate your findings and ideas by seeking constructive criticism from your colleagues and supervisors.

The publication process

The formal knowledge representing the empirical and professional basis for your professional practice is in large part stored in journals, books and conference reports. Journals are published by appropriate professional associations, government departments or private companies. Having completed a research project, how does one publish it in a professional journal? After all, the value of research is negligible if it is not made public.

In general, the prospective author will:

1. Select a professional or scientific journal appropriate for the material.
2. Present the research report in a format required by the journal.
3. Send the completed manuscript to the journal's editor.

The editor is generally a person of high standing in a given scientific or professional area. If the article is judged as being appropriate for the journal, the editor will send the article to two or more referees and, on the basis of the referees' reports, publish or reject the manuscript. Sometimes the referees recommend certain additions or changes which have to be made by the author before the manuscript is judged to be publishable.

Therefore, when you read research publications in refereed journals, you can be confident that the articles have been scrutinized by experts. However, as shown in the next subsection, this does not necessarily guarantee the truth of either the evidence or the conclusions.

Ethics of presenting research

The health science researcher has an obligation to publish honest and accurate results that would

not harm those people who participated in the research.

Most ethics committees in health care institutions and universities have the twin objectives of not only advancing knowledge for the common good but also preventing harm to those participating in the research. This is particularly so in the situation where the participants may have a diminished capacity to consent freely to their involvement (e.g. children or people who are unconscious or seriously ill). It is crucial to maintain the dignity and confidentiality of participants in health research.

Therefore, in the process of ethical evaluation of health science research, the researcher can expect to be closely questioned on these issues. If the researcher cannot convince the ethics committee that the research will deliver knowledge for the common good and that it will not harm the participants, then the research will not usually proceed.

In research performed for a higher degree, many universities will not accept a thesis without an accompanying ethical clearance from the relevant ethics committee. Most hospitals and universities have strict ethical procedures that must be followed before any research work is commenced by their staff. Most, if not all, health research grant bodies require an ethical clearance before they will release the funds to successful applicants. Many journals also require certification from the researcher that the work complies with ethical principles. It is likely that this trend towards tightening of procedures will continue.

The ultimate unethical act is to manufacture data. Broad & Wade (1982), in their book *Betrayers of the Truth*, describe this problem. It would seem to be a growing problem that may be associated with the 'publish or perish' requirements placed upon health science researchers by granting bodies and employers.

In the health sciences, it is not only the participants in research who may be harmed or assisted by the research. If an erroneous research finding is widely applied, it may harm many thousands of people. Ethics are therefore not simply concerned with whether the researcher has good intentions and treats the research participants well; there is also the issue of competence. Poorly designed research is unethical in that it may bring great harm to others. Thus, the ethical researcher must also be a competent researcher.

Summary

In this chapter we outlined the general format followed by researchers for publishing their results. The format is related to the logical steps of planning, conducting and interpreting research. The style involves clarity, accuracy and sufficient completeness for colleagues to understand or replicate the research project. Research is published in journals, which are generally edited by persons of high standing in the field. Every effort is made by editors to ensure the validity of the research published in their journals. The individual researcher is also ethically bound to report findings in an unbiased and truthful fashion.

Although the format and style outlined in this chapter might seem rather arduous, poor presentation may destroy the intrinsic value of a research project.

Self-assessment

Explain the meaning of the following terms:

abstract
apparatus
discussion
method
plagiarism
procedure
refereed journal
subjects

True or false

1. As a rule, the title of a research investigation should not contain more than seven words.
2. Generally, the research hypothesis should be presented in the introduction.
3. A research report should contain sufficient information so that the investigation can be replicated.

4. The results section should contain all computational details for each statistic.
5. The abstract should normally contain the key tables of the results.
6. The design of an investigation influences the content of the method section.
7. All the names and addresses of your subjects must be published to enable replication of your investigation.
8. Quotations should be used sparingly in a research report.
9. A research report should be written in the past tense.
10. The outcomes of statistical analyses are reported in the results section.
11. Scientists do not normally report the results of their investigations, in case their work is stolen or misrepresented.
12. Calculations are best presented in appendices.
13. The 'referees' are hired by the investigator in order to convince the editor that an investigation should be published.
14. The role of an editor for a scientific journal is to censor research publications for pornographic, blasphemous or politically undesirable material.
15. Good research is unique and cannot be replicated.
16. A researcher should report data even if it is inconsistent with the researcher's original preconceptions.
17. Scientific and professional journals are important for disseminating and storing knowledge.
18. Fortunately, there have been no major scandals concerning scientists publishing fabricated data.
19. Provided that the results are statistically significant, there is no need to present descriptive statistics.

Multiple choice

1. Scientific journals:
   a only publish empirical evidence
   b depend on the services of referees to comment on the validity of the research project
   c publish only true knowledge
   d *b* and *c*.
2. The literature review is normally found in which section of a research report?
   a Abstract.
   b Introduction.
   c Discussion.
   d References.
3. A literature review for a research report should:
   a contain a detailed review of all previously published reports
   b contain a selective review of evidence pertinent to the current research project
   c be at least 5000 words
   d *a* and *c*
   e *b* and *c*.
4. Which of the following is most inadequate as a title for a research report?
   a The effects of the twentieth century culture on being human: an empirical evaluation of personal functioning in declining cultures.
   b Electrical stimulation of the limbic system: effects on emotion and memory.
   c A survey of the incidence of mental illness in the London metropolitan area.
   d Popularity, friendship selection and specific peer interaction among children.
5. The methods section of a research report:
   a informs the reader of the purpose of an investigation
   b informs the reader about the state of methodological advances in the subject area
   c informs the reader as to how the investigation was carried out
   d informs the reader as to how the hypothesis or aim of the investigation was formulated.
6. When writing a scientific report one should:
   a make sure the introduction contains 250 words or less
   b use personal pronouns as much as possible
   c try to impress the readers by one's level of general knowledge
   d use the past tense.
7. In which part of a research report are the descriptive and inferential statistics normally reported?
   a Abstract.
   b Results.
   c Discussion.
   d Appendices.
8. In writing a discussion, one should:
   a relate the results to findings reported in previous publications
   b establish if the results of the investigation supported the hypothesis

c neither *a* nor *b*
d both *a* and *b*.

9. Which of the following statements is true?
   a The discussion section should relate present findings to previous research.
   b The literature review should be conducted in a special appendix labelled 'references'.
   c The results section should contain only tables and graphs, but not any verbal descriptions of the data.
   d All the above statements are true.
   e None of the above statements are true.

10. Which of the following statements is false?
    a The abstract should be a brief summary of the research.
    b It is unethical to fabricate data.
    c A refereed journal is one in which experts independently evaluate a research report before it is published.
    d A well-designed research project need not have a procedure section.

Chapter Twenty Three

23

Critical evaluation of published research

CHAPTER CONTENTS

Introduction

By the time a research report is published in a refereed journal, it has been critically scrutinized by several experts and, usually, changes have been made to the initial text by the author(s) to respond to the referees' comments. Nevertheless, even this thorough evaluation procedure doesn't necessarily guarantee the validity of the design or the conclusions presented in a published paper. Ultimately, you as a health professional must be responsible for judging the validity and relevance of published material to your own clinical activities. The evidence-based practice movement focuses on the ways in which practitioners can incorporate better procedures into their practice based upon well-founded research and evaluation evidence. The systematic review processes employed by bodies such as the Cochrane and Campbell Collaborations are intended to assist clinicians in the selection of interventions that are well proven (Ch. 24).

The proper attitude to take with published material, including systematic reviews, is hard-nosed scepticism, notwithstanding the authority of the source. This attitude is based on our understanding of the uncertain and provisional nature of scientific and professional knowledge. In addition,

health researchers deal with the investigation of complex phenomena, where it is often impossible for ethical reasons to exercise the desired levels of control or to collect crucial information required to arrive at definitive conclusions. The aim of critical evaluation is to identify the strengths and weaknesses of a research publication, so as to ensure that patients receive assessment and treatment based on the best available evidence.

The aim of this chapter is to demonstrate how select concepts in research design, analysis and measurement can be applied to the critical evaluation of published research. The chapter is organized around the evaluation of specific sections of research publications.

The specific aims of this chapter are to:

1. Examine the criteria used for the critical evaluation of a research paper.
2. Discuss the implications of identifying problems in design and analysis in a given publication.
3. Outline briefly strategies for summarizing and analysing evidence from a set of papers.
4. Discuss the implications of critical evaluation of research for health care practices.

Critical evaluation of the introduction

The introduction of a paper essentially reflects the planning of the research. Inadequacies in this section might signal that the research project was erroneously conceived or poorly planned. The following issues are essential for evaluating this section.

Adequacy of the literature review

The literature review must be sufficiently complete so as to reflect the current state of knowledge in the area. Key papers should not be omitted, particularly when their results could have direct relevance to the research hypotheses or aims. Researchers must be unbiased in presenting evidence that is unfavourable to their personal points of view. This is why we now have systematic review procedures, such as those utilized by the Cochrane Collaboration, so as to avoid inappropriate and biased exclusion or inclusion of work that supports or challenges a point of view favoured by the researcher or other researchers who hold contrary opinions. Poor review of the literature could lead to the unfortunate situation of repeating research or making mistakes that could have been avoided if the previous work's findings had been incorporated into formulation of the research design.

Clearly defined aims or hypotheses

As stated in Chapter 2, the aims or hypotheses of the research should be clearly stated. If this clarity in expression of the aims is lacking, then the rest of the paper will be compromised. In a quantitative research project, it is usual to see a statement of the hypotheses as well as the research aims. All research, whether qualitative or quantitative, should have a clear and recognizable statement of aim(s).

Selection of an appropriate research strategy

In formulating the aims of the investigation, the researcher must have taken into account the appropriate research strategy. For instance, if the demonstration of causal effects is required, a survey may be inappropriate for satisfying the aims of the research. If the purpose of the study is to explore the personal interpretations and meanings of participants then a qualitative strategy will be best. Some researchers now advocate mixed designs where multiple studies are performed to examine different perspectives of the same issues. Thus in a study of views concerning health practices, a focus group discussion may also be accompanied by a structured questionnaire even within the same study sample, so that the findings from each may be used to inform the total understanding of the research issue(s) under study.

Selection of appropriate variables/ information to be collected

In a quantitative study, if the selection of the variables is inappropriate to the aims or questions being investigated, then the investigation will not

produce useful results. Similarly, in a qualitative study, the information to be collected must be appropriate to the research aims and questions.

Critical evaluation of the methods section

A well-documented methods section is a necessary condition for understanding, evaluating and perhaps replicating a research project. In general, the critical evaluation of this section will allow a judgment of the validity of the investigation to be made.

Research subjects/participants

This section shows if the study participants were representative of the intended target group or population and the adequacy of the sampling model used.

Sampling model used

In Chapter 3, we outlined a number of sampling models that can be employed to optimize the representativeness of a study sample. If the sampling model is inappropriate, then the sample might be unrepresentative, raising questions concerning the external validity of the research findings. In qualitative research, although the participant sampling method may be less formal than in a quantitative study, the issue of participant representativeness is still pertinent in terms of being able to apply the results more broadly.

Sample size/number of participants

Use of a small sample is not necessarily a fatal flaw of an investigation, if the sample is representative. However, given a highly variable, heterogeneous population, a small sample will not be adequate to ensure representativeness (Ch. 3). Also, a small sample size could decrease the power of the statistical analysis in a quantitative study (Ch. 20). As discussed in the qualitative sampling section of this text, unlike in quantitative sampling procedures, there is not widespread agreement among qualitative researchers as to the issue of how many participants are needed in such studies.

Description of the study participants

A clear description of key participant characteristics (for example age, sex, type and severity of their condition) should be provided. When necessary and possible, demographic information concerning the population from which the participants have been drawn should be provided. If not, the reader cannot adequately judge the representativeness of the sample.

Instruments/apparatus/tools

The validity and reliability of observations and/or measurements are fundamental characteristics of good research. In this section, the investigator must demonstrate the adequacy of the tools used for the data collection.

Validity and reliability

The investigator should use standardized tools, or establish the validity and reliability of new tools used. A lack of proven validity and reliability will raise questions about the adequacy of the research findings.

Description of tools

A full description of the structure and use of novel tools should be presented so that they can be replicated by independent parties.

Procedure

A full description of how the investigation was carried out is required for both replication and for the evaluation of its internal and external validity. This requirement applies to both qualitative and quantitative studies.

Adequacy of the design

It was stated previously that a good design should minimize alternative conflicting interpretations of the data collected. For quantitative research aimed at studying causal relationships, poor design will result in uncontrolled influences by extraneous variables, muddying the identification of causal effects. In Section 3, we looked at a variety of threats to internal validity which must be considered when critically evaluating an investigation.

In a qualitative study the theoretical approach taken in the study design or approach should be clearly stated.

Control groups

In quantitative research a common way of controlling for extraneous effects is the use of control groups (such as placebo, no treatment, conventional treatment). If control groups are not employed, then the internal validity of the investigation might be questioned. Also, if placebo or untreated groups are not present, the size of the effect due to the treatments might be difficult to estimate.

Subject assignment

When using an experimental design, care must be taken in the assignment of subjects so as to avoid significant initial differences between treatment groups. Even when quasi-experimental or natural comparison strategies are used, care must be taken to establish the equivalence of the groups.

Treatment parameters

It is important to describe all the treatments given to the different groups. If the treatments differ in intensity, or the administering personnel take different approaches, the internal validity of the project is threatened. The adherence of the study in the delivery of the intervention to the intended intervention is sometimes called *treatment fidelity*.

Rosenthal and Hawthorne effects

Whenever possible, intervention studies should use double- or single-blind procedures. If the participants, researchers or observers are aware of the aims and predicted outcomes of the investigation, then the validity of the investigation will be threatened through bias and expectancy effects. In qualitative research, it is very important that the research findings are not unduly influenced by the personal positions of the researchers in a way that obscures the meanings and interpretations of the research participants. Of course, the position of the researcher in any study, whether qualitative or quantitative, will to some extent influence the findings but this needs to be kept to a minimum.

Settings

The setting in which a study is carried out has implications for external (ecological) validity. An adequate description of the setting is necessary for evaluating the generalizability of the findings. The context of the investigation may have important effects on the study outcomes. Research conducted in the investigator's lab or office may yield different results to the same work conducted in the field.

Times of treatments and observations

In intervention studies the sequence of any treatments and observations must be clearly indicated, so that issues such as series and confounding effects can be detected. Identification of variability in treatment and observation times can influence the internal validity of experimental, quasi-experimental or $n = 1$ designs, resulting in, for instance, internal validity problems.

Critical evaluation of the results

The results should represent a sound and, where appropriate, statistically correct summary and analysis of the data. Inadequacies in this section could indicate that inferences drawn by the investigator were erroneous.

Tables and graphs

Data should be correctly tabulated or drawn and adequately labelled for interpretation. Complete summaries of all the relevant findings should be presented.

Selection of statistics

Where appropriate both descriptive and inferential statistics must be selected according to specific rules. The selection of inappropriate statistics could distort the findings and lead to inappropriate inferences.

Calculation of statistics

Clearly, both descriptive and inferential statistics must be correctly calculated. The use of computers generally ensures this, although some attention must be paid to gross errors when evaluating the data presented.

Methods of qualitative analysis

The methods chosen must complement the theoretical approach taken in the study and be performed according to the specified protocols.

Critical evaluation of the discussion

In the discussion, investigators draw inferences from the information or data they have collected in relation to the initial aims, questions, and/or hypotheses of the investigation. Unless the inferences are correctly made, the conclusions drawn might lead to useless and dangerous treatments being offered to clients.

Drawing correct inferences from the collected information/data

The inferences from the collected information or data must take into account the limitations of the study and the analytical methods used to analyse them. In the quantitative domain we have seen, for instance in Chapter 16, that correlations do not necessarily imply causation, or that a lack of significance in the statistical analysis could imply a Type II error or incorrect missing of a real trend or finding (see Ch. 20). In the qualitative domain, the findings must follow reasonably from the information collected in the investigation according to the paradigm used.

Logically correct interpretations of the findings

Interpretations of the findings must follow from the information collected, without extraneous evidence being introduced. For instance, if the investigation used a single-participant design, the conclusions should not claim that a procedure is generally useful for the entire population.

Research protocol deviations

In interpreting the data or information collected in a study, the investigator must indicate, and take into account, unexpected deviations from the intended research protocols. For instance, in a quantitative study a placebo/active treatment code might be broken, or 'contamination' between control and experimental groups might be discovered. In a qualitative study, it could be that participants have conversed with each other about the research prior to one of the participants completing participation. If such deviations are discovered by investigators they are obliged to report these, so that the implications for the results might be taken into account.

Generalization from the findings

Strictly speaking, the data obtained from a given sample are generalizable only to the population from which the participants were drawn. This point is sometimes ignored by investigators and the findings are generalized to subjects or situations which were not considered in the original sampling plan. Qualitative researchers may vary in their willingness to claim generalizability of their findings outside the actual research participants but this must also be systematically considered.

Statistical and clinical significance

As was explained in Chapter 22, in quantitative studies, obtaining statistical significance does not necessarily imply that the results of an investigation are clinically applicable or useful. In deciding on clinical significance, factors such as the size of the effect, side effects and cost-effectiveness, as well as value judgments concerning outcome, must be considered.

Theoretical significance

It is necessary to relate the results of an investigation to previous relevant findings that have been identified in the literature review. Unless the results are logically related to the literature, the theoretical significance of the investigation remains unclear. The processes involved in comparing the findings of a set of related papers are introduced in the next subsection.

Table 23.1 summarizes some of the potential problems, and their implications, which might emerge in the context-critical evaluation of an investigation. A point which must be kept in mind is that, even where an investigation is flawed, useful knowledge might be drawn from it. The aim of critical analysis is not to discredit or tear down published work, but to ensure that the reader

Table 23.1 Checklist for evaluating published research

Problems which might be identified	Possible implications in a research article
1. Inadequate literature review	Misrepresentation of the conceptual basis for the research
2. Vague aims or hypotheses	Research might lack direction; interpretation of evidence might be ambiguous
3. Inappropriate research strategy	Findings might not be relevant to the problem being investigated
4. Inappropriate variables selected	Measurements might not be related to concepts being investigated
5. Inadequate sampling method	Sample might be biased; investigation could lack external validity
6. Inadequate sample size	Sample might be biased; statistical analysis might lack power
7. Inadequate description of sample	Application of findings to specific groups or individuals might be difficult
8. Instruments lack validity or reliability	Findings might represent measurement errors
9. Inadequate design	Investigation might lack internal validity; i.e. outcomes might be due to uncontrolled extraneous variables
10. Lack of adequate control groups	Investigation might lack internal validity; size of the effect difficult to estimate
11. Biased subject assignment	Investigation might lack internal validity
12. Variations or lack of control of treatment parameters	Investigation might lack internal validity
13. Observer bias not controlled (Rosenthal effects)	Investigation might lack internal and external validity
14. Subject expectations not controlled (Hawthorne effects)	Investigation might lack internal and external validity
15. Research carried out in inappropriate setting	Investigation might lack ecological validity
16. Confounding of times at which observations and treatments are carried out	Possible series effects; investigation might lack internal validity
17. Inadequate presentation of descriptive statistics	The nature of the empirical findings might not be comprehensible
18. Inappropriate statistics used to describe and/or analyse data	Distortion of the decision process; false inferences might be drawn
19. Erroneous calculation of statistics	False inferences might be drawn
20. Drawing incorrect inferences from the data analysis (e.g. Type II error)	False conclusions might be made concerning the outcome of an investigation
21. Protocol deviations	Investigation might lack external or internal validity
22. Over-generalization of findings	External validity might be threatened
23. Confusing statistical and clinical significance	Treatments lacking clinical usefulness might be encouraged
24. Findings not logically related to previous research findings	Theoretical significance of the investigation remains doubtful

understands its implications and limitations with respect to theory and practice.

Summary

The critical evaluation of published material at a level of detail suggested by this chapter can be a time-consuming, even pedantic, task. One undertakes such detailed analysis only when professional communications are of key importance, for example, when writing a formal literature review or when evaluating current evidence for adopting a new intervention or approach. Nevertheless, it is a necessary process for an in-depth understanding of the empirical and theoretical basis of your clinical practice.

Even when some problems are identified with a given research report, it is nevertheless likely that the report will provide some useful additional knowledge. Given the problems of generalization, an individual research project is usually insufficient for firmly deciding upon the truth of a hypothesis or the usefulness of a clinical intervention. Rather, as we will see in Chapter 24, the reader needs to scrutinize the range of relevant research and summarize the evidence using qualitative and quantitative review methods. In this way, individual research results can be evaluated in the context of the research area. Disagreements or controversies are ultimately useful for generating hypotheses for guiding new research and for advancing theory and practice.

Self-assessment

Explain the meaning of the following terms:

critical evaluation
protocol deviation

True or false

1. Critical analysis of a publication aims to identify the internal and external validity of the investigation.
2. If an investigation is published in a reputable journal by established investigators then the validity of the investigation can be taken for granted.
3. Random assignment of subjects to treatment groups ensures that the investigation uncovers causal effects.
4. The outcome of an investigation can be useful even with a small sample size.
5. If an investigation produces statistically significant results, its design must have been adequate.
6. Obtaining statistical significance in an investigation is a condition for the demonstration of the clinical significance of a quantitative study.
7. The replication of an investigation demonstrates the internal validity of the original investigation.
8. Without adequate controls the size of an effect might be difficult to estimate.
9. If a study is internally valid, the investigator is justified in generalizing the results to any other population.
10. Provided that the outcomes are statistically significant, it doesn't matter which statistical tests were chosen to analyse the data.
11. If the design of an investigation is inadequate, none of the empirical findings are of scientific or clinical use.
12. Controversies in an area of science usually reflect the presence of fraudulently published evidence.
13. One of the problems with using human subjects for research is the expectations of the subjects concerning the purpose of the investigation.
14. Even poorly planned research can provide some useful results.
15. The application of the scientific method ensures the validity of a researcher's conclusions.
16. Disagreements among researchers in an area are useful for generating new hypotheses.

Multiple choice

1. The aim of the critical analysis of a publication is to:
   a identify the relevance of the results for clinical practice
   b identify the internal and external validity of the investigation
   c identify and attack incompetent researchers in one's area of interest
   d *a* and *b*.

2. If the internal validity of a study is adequate, then:
   a the results will be statistically significant
   b the results will be clinically useful
   c the investigation may demonstrate causal effects
   d *a* and *b*.
3. Say that an investigation has generated some interesting findings. However, you find that the investigators selected an inappropriate statistical test to analyse their findings. You should:
   a regretfully discard the study as useless
   b re-analyse the data from the descriptive statistics provided
   c write to the investigators for their raw data, and re-analyse yourself
   d *b* or *c*.
4. The reason one should evaluate the 'literature' as a whole is to:
   a identify general patterns of findings in the area
   b condense results from related papers into a single statistic
   c identify and attempt to explain controversies in the area
   d all of the above.
5. In judging the clinical significance of a well-designed investigation one should consider:
   a the cost-effectiveness of the interventions
   b the size of the therapeutic effects
   c the possible undesirable side effects of the treatment
   d all of the above.

An investigation was carried out in order to show that 'prepared childbirth' was an effective method for reducing pain during delivery. Ninety women attending a large hospital constituted the sample. Sixty of the women chose to participate in childbirth preparation, based on the Lamaze method, provided by trained instructors working at the hospital. This method encourages 'natural' (drug-free) childbirth through teaching physical and mental strategies for coping with pain or discomfort occurring during childbirth. The other 30 women chose not to attend the childbirth preparation programme. The level of pain experienced was assessed on the McGill Pain Questionnaire, which has been shown to be a valid and reliable interval scale for pain. It was administered following the childbirth. In addition the number of women seeking analgesia during childbirth was recorded as a measure of levels of discomfort experienced. The results for the investigation are as shown in the table below.

Questions 6–14 refer to the above investigation.

6. The strategy for the investigation is best described as:
   a an experiment
   b a quasi-experiment
   c a correlational study
   d an $n = 90$ design.
7. One of the problems with the above investigation was that:
   a the subjects could not be randomly assigned to treatment groups
   b the dependent variable was irrelevant to the aims
   c basic ethical issues were not considered
   d the instructors teaching the Lamaze method were incompetent.
8. From the information given above, it is clear that the investigators controlled for:
   a Hawthorne effects
   b Rosenthal effects
   c subject assignment
   d none of the above.

Groups	Mean pain scores	Number given medication
Women with no training ($n = 30$)	38	24
Women with childbirth preparation ($n = 60$)	32	49
	(The difference was statistically significant at $\alpha = 0.05$)	(The difference was not statistically significant at $\alpha = 0.05$)

9. If you wanted to calculate the proportion of women with no training who had greater McGill pain scores than women with childbirth preparation, then the required statistics are:
   a the distribution of t for $n = 98$
   b the normal distribution
   c the indices for reliability and validity
   d the standard deviations for the two groups.
10. Which of the following statistical tests is most appropriate for analysing the significance of the data for the McGill pain scores?
   a Mann–Whitney *U*
   b Sign test
   c *z* test for two means
   d χ^2 test.
11. Which of the following statistical tests is most appropriate for analysing the significance of the data for women requiring medication?
   a Mann–Whitney *U*
   b Sign test
   c *t* test for two means
   d χ^2 test.
12. The lack of statistical significance for the data on medication implies that:
   a the power for the test may have been too low
   b equal sample sizes should have been used
   c training has no effect
   d both *a* and *c*.
13. The outcome of this investigation can be generalized to:
   a women having children and undergoing Lamaze training
   b women having children without Lamaze training
   c women who choose the type of childbirth they undergo
   d none of the above groups.
14. Considering the evidence provided, one concludes that:
   a prepared childbirth is a waste of time
   b there is evidence that Lamaze preparation at this hospital results in statistically significant reductions in pain during delivery
   c women undergoing childbirth find Lamaze preparation useless at this hospital
   d *a* and *c*.

Chapter Twenty Four

24

Synthesis of research evidence

CHAPTER CONTENTS

Introduction

The research papers published in scientific and professional peer-reviewed journals contain the basic information currently available for understanding the causes and consequences of health problems. Each research publication contributes to the overall knowledge regarding a health problem. However, given the problems of generalization, an individual research project is usually insufficient for firmly deciding the truth of a hypothesis or the efficacy of a clinical intervention. In order to understand the progress of a research programme (Ch. 1) we need to compare, contrast and synthesize the results of related research papers. Consistent results across diverse research reports are the most appropriate bases for justifying the delivery of health services.

The term literature refers to the set of publications containing the network of theories, hypotheses, practices and evidence representing the current state of knowledge in a specific area of the health sciences. A literature review contains both the critical evaluation of the individual publications and the identification of emergent trends and patterns of evidence. The literature review is a synthesis of the available knowledge in an area and therefore constitutes the strongest foundations for initiating further advances in theory and practice.

The overall aim of this chapter is to outline the basic strategies used for synthesizing evidence and producing a literature review.

The specific aims of the chapter are to:

1. Identify the basic methodological principles relevant to writing a health sciences review.
2. Describe the process for conducting a systematic review.
3. Explain how to interpret the findings of a published meta-analysis.
4. Explain the uses and limitations of systematic reviews and meta-analyses.

Basic principles

The first thing to consider is that writing a literature review is a demanding intellectual challenge. The facts do not 'speak' for themselves. Rather, the evidence has to be extracted, critically evaluated, organized and synthesized into a logical, coherent representation of the current state of knowledge. For example, consider the review by Olanow (2004) tilted 'The scientific basis for the current treatment of Parkinson's disease'. This relatively brief 15-page review is based on only 75 references, although there are thousands of research papers, articles and reports available on the anatomy, physiology and treatment of Parkinson's disease. In writing the review, the author had to make a series of expert judgments regarding which were the nine key papers, containing the most salient, up-to-date information for understanding the medical treatment of Parkinson's disease.

Second, the outcome of the review process is influenced by the theoretical orientation and professional background of the reviewer. Olanow, a leading neurologist, provided an authoritative review written from a biomedical perspective. In contrast, a physiotherapist working in neurological rehabilitation might take a different conceptual approach to the causes and treatment of Parkinson's disease. He or she might place more emphasis on psychological and social factors as integral components of the etiology and treatment of Parkinson's disease.

Third, the selection, analysis, critique and synthesis of the materials is an active, interpretive process drawing on the personal experiences, interests and values of the reviewer. Even if their professional backgrounds were identical, there are no guarantees that two reviewers interpreting the evidence from the same set of publications will arrive at exactly the same conclusions. Depending on how these reviewers approached the subject matter, they might emphasize different aspects of the evidence or select different patterns in the data as being important, resulting in different syntheses. In Chapter 1, we discussed the post-positivist position that theories and preconceptions can influence our perceptions of what is happening in the world, and therefore shape the way in which we construct knowledge.

Last, the notion that we all have our experiences, opinions and prejudices does not imply that 'anything goes' when writing health sciences reviews. On the contrary, we need to apply the principles of scientific methodology to ensure that we provide an accurate overview of the literature. In other words, there are principles which we must follow in preparing a literature review.

As stated before, in preparing literature reviews and evaluating research findings, a multiplicity of papers must be considered, at least according to the following general steps:

1. Identify relevant literature; select key papers.
2. Critically evaluate key papers, as discussed in this section. You might decide to discard some papers if irreparable problems are discovered.
3. Identify general patterns of findings in the literature. Tabulate findings, where appropriate.
4. Identify crucial disagreements and controversies.
5. Propose valid explanations for the disagreements. Such explanations provide a theoretical framework for resolving controversies and proposing future research.
6. Provide a clear summary concerning the state of the literature, identifying progress, obstacles and further research.

Systematic reviews

There are several approaches to conducting health sciences reviews. For example, the previously mentioned review by Olanow (2004) can be classified as a 'narrative' review. This approach

entails producing a 'story-like' overview of the state of current evidence and theories. Although reviewers adhere to the principles of science and logic in conducting a narrative review, there have been concerns about bias and lack of clarity. More recently, systematic reviews and meta-analyses have been introduced to enhance the rigour for combining and interpreting the state of the literature. Systematic reviews rely on the explicit use of the methodological principles discussed in previous chapters. In effect, systematic reviews follow the problem-solving approach as used for conducting empirical research (Schwartz & Polgar 2003, Ch. 14). Let us examine a published example (Polgar et al 2003) to illustrate the logic and principles underlying the conduct of systematic reviews.

Research questions

When conducting a systematic review, we are expected to formulate a clear research question which will be answered by the outcomes of the review. You might have read about the current research using embryonic or stem cells for reconstructing the brains of people suffering from neurological conditions such as stroke and Parkinson's disease. We were interested in reviewing the evidence for answering the research question 'How effective is reconstructive neurosurgery (i.e. the grafting of immature cells) for improving the signs and symptoms of Parkinson's disease?' (Polgar et al 2003).

Theoretical framework

Although this principle is not adhered to by all reviewers, it is very useful to specify the theoretical framework(s) which guides a specific review. The theoretical framework used for conducting the review was identified as the 'Repair Model'. This is a purely biomedical, quantitative view of neural reconstruction based on the notion that recovery is due to the replacement of dopaminergic cells damaged in Parkinson's disease. An explicit theoretical framework is essential for understanding a given area of health as a coherent research programme (Ch. 1).

Search strategy

The next step is to identify the relevant publications. For example, in Polgar et al (2003) we conducted the following search:

> *A search of MEDLINE (1994–2000 and 2000/01–2000/10) using the exploded terms fetal tissue transplantation, Parkinson's disease, human or fetal tissue transplants, Parkinson's disease, human, also using the author Kopyov was conducted. In conjunction with this search, abstracts from the American Society for Neural Transplantation and Repair conferences (1999; 2000) were hand searched for authors that may have been overlooked. The reference lists of key papers were searched to identify papers that might not have been identified through on-line search mechanisms. An additional electronic search was conducted using the following databases: Medline (1966–March Week 2, 2001), Embase (1994–April Week 1, 2001), CINAHL (1982–March Week 2, 2001).*

The general point here is that reviewers must be diligent in identifying all the publications which constitute the literature in the area targeted for review. The search should include both 'electronic, on-line searches' and 'hand' searches of key journals for cross-checking if the relevant papers were identified by the search engines. It is essential to have a working knowledge of who the key researchers are in a particular field and what critical issues exist in a research programme before we can undertake a formal published review.

Selection of key papers

Depending on the area of health sciences being reviewed, literature searches might yield any number of publications, from one or two to many hundreds. Relying on the outcome of the search, the reviewers might restate the research questions and redefine the scope of the proposed systematic review. This will expand or restrict the number of studies for further searchers. In addition, explicit inclusion/exclusion criteria are used to identify the

most relevant sources. For example, Polgar et al (2003) used the following criteria for including a study in a review. In order to be included in the review, a study had to:

- be published in a peer-reviewed journal
- have transplantation surgery performed after 1993, following consensus for optimal donor characteristics
- have grafted human or embryonic cells
- have followed 'best practice' stereotactic surgery procedures
- have followed standard post-surgical assessment protocols.

It would be tedious in the present context to explain the specific reasons for each of the above selection criteria; interested readers will find that they were justified in the original review. The point is that we must have explicit, objective criteria for including or excluding studies.

The key papers function as a sample of the best available information accessible in the literature. Parallel to empirical research, we use the evidence from key papers to draw inferences about the state of knowledge in the area under review. The use of diagnostic searches and explicit inclusion/exclusion criteria ensure that the 'sample' of papers produces a representative, rather than biased, sample of the overall knowledge.

Coding the research studies

When the process of selecting and collecting the publications is complete, we have the available evidence relevant to answering our research question. The information that we seek is embedded in the text of research publications. We need to extract the key information from the text of each of the papers selected for the review. We can approach the analysis of the meaning of this information in a way that is similar to qualitative data analysis (Ch. 21). Similarly to the predetermined codes used in content analysis, reviewers use the constant features of quantitative research for identifying the categories appropriate for analysing the key papers (Schwartz & Polgar 2003). These features include:

1. the design of the studies
2. the sampling strategies used and the sample characteristics
3. the ways in which the treatment or intervention was administered
4. the data collection strategy used (i.e. the measurement strategy)
5. the statistical and clinical significance of the results.

Other dimensions or features might also be coded or the above features can be modified according to the judgments of the authors. These aspects of the research process are often presented in a table form, as shown in Table 24.1. We used a hypothetical example rather than the previous published review.

To illustrate coding, consider a set of four hypothetical studies reporting on levels of diabetics' compliance to insulin administration. The results and key features of the hypothetical studies are tabulated in Table 24.1.

Table 24.1 represents how findings from several publications might be tabulated. Key information

Table 24.1 Compliance with insulin use by diabetics

Publication	Sample size	Average age of patients (years)	Method of measuring compliance	Percentage of patients compliant
Smith (2000)	50	55	Self-report	85
Jones (2001)	60	58	Self-report	82
Brown (2001)	50	59	Blood sugar level	40
Miller (2002)	55	56	Blood sugar level	35

about each study, as well as the outcomes, is presented in the table, enabling the emergence and demonstration of an overall pattern.

Interpretation of the evidence

The coding of the hypothetical studies enables the reviewer to answer the research question: 'How compliant are patients with diabetes with using insulin?' The hypothetical summarized evidence illustrates that sometimes no clear overall trends emerge from the tabulated findings.

In the simple hypothetical example shown in Table 24.1, the percentage of compliance reported by Smith (2000) and Jones (2001) is over twice that reported by Brown (2001) and Miller (2002). Clearly, there is an inconsistency in the literature. A possible explanation for this discrepancy might emerge by the inspection of Table 24.1. Neither differences in sample size nor the average ages of the patients provide an explanation for the difference. However, the method by which compliance was measured emerges as a plausible explanation. The investigators Smith and Jones relied on the patients' self-reports and might have overestimated compliance levels, in contrast to Brown and Miller, who used a more objective method and found poor levels of compliance. Of course, this explanation is not necessarily true, but is simply a hypothesis to guide future investigations of the problem. There are other possibilities which might account for the pattern of findings. It appears that more research is needed.

Meta-analysis

Dooley (1984) discussed the availability of two general types of strategies for summarizing research findings from multiple papers:

1. *Qualitative*. A qualitative review involves the selection of key features of related publications such as designs, subject characteristics or measures used in the studies. These features are presented in a table form, such that differences in the features of the research can be related to outcomes. The qualitative reviews identified by Dooley (1984) are related to the systematic reviews, as we discussed above.
2. *Quantitative*. A quantitative review calls for the condensation of the results from several papers into a single statistic. This statistic represents an overall or pooled effect size. These procedures are meta-analyses which are systematic procedures for summarizing the results published in a set of research papers.

Although many statistical procedures can be used for synthesizing data, meta-analysis also refers to an active area of statistics examining strategies most suitable for synthesizing published evidence. Statisticians have developed software packages such as 'Comprehensive Meta-Analysis' which expedite the computational difficulties entailed in synthesizing evidence from diverse studies (http://www.metaanalysis.com, info@metaanalysis.com).

Combining data from diverse studies

How are results synthesized? Let us look at a simple example for combining data. Say that you are interested in the average age of the participants in three related studies: A, B and C (Table 24.2).

Say we wish to calculate the average age for all the 230 participants in the three studies. Could we calculate the overall mean, $\bar{X}$, simply by adding up the three means and dividing by three? The answer is no, because there are different numbers of participants across the groups. We must give a weight to each of the statistics depending on 'n', the sample size for each study. The equation which we use is:

$$\bar{X} = \frac{(\bar{X}_A \times n_A) + (\bar{X}_B \times n_B) + (\bar{X}_C \times n_C)}{n_A + n_B + n_C}$$

$$= \frac{(40 \times 80) + (45 \times 50) + (60 \times 100)}{80 + 50 + 100}$$

$$= 49.8$$

Table 24.2 Participants' ages in three hypothetical studies

Study	Number of participants (n)	Average age ($\bar{X}$)
A	80	40
B	50	45
C	100	60

The point here is that, in order to calculate the correct overall statistics, we must give a weight to each study. In general, the weight assigned to a study represents the proportion of information the study contributes to the overall analysis. For the above calculation, weight was determined by the sample size used in each study. In general the weight assigned to a study represents the proportion of the information contributing to the calculation of the overall statistic.

Even the calculation of a simple statistic like 'average overall age' can be useful for understanding the state of a research programme. For instance, in Polgar et al (2003) we found that the mean age of Parkinson's disease sufferers was 56 years, with the mean overall duration of the illness being 13 years. These results indicated that the average age of onset of the disease was only about 43 years, indicating that experimental reconstructive neurosurgery has been offered to an unrepresentative sample of Parkinson's disease sufferers. Typically people with Parkinson's disease are in their late sixties or early seventies.

Of course we need to synthesize other clinically and theoretically relevant statistics appearing across the papers constituting a research programme, including the overall pooled standard deviation, the overall statistical significance, the overall effect size and the confidence intervals for the overall effect sizes. These analyses are best carried out using statistical software packages. You can check the Internet for further information regarding the logic of meta-analysis and currently available software packages.

Interpreting the results of a meta-analysis

There are different ways for conducting and reporting the results of meta-analyses which have become quite frequent in the health sciences literature. A typical way of presenting the results is shown in Figures 24.1 and 24.2, which show the hypothetical outcome of a computer-assisted meta-analysis (modified from www.metaanalysis.com). The graphics shown in these figures are referred to as 'forest plots'. A forest plot is a visual representation of the results of a meta-analysis.

Say that the printout showed the results of five randomized double-blind trials (see Ch. 5) aiming to demonstrate the effectiveness of a vaccine for influenza. In each of the studies volunteers were given either the vaccine (treated) or a placebo (control) injection. The following features of Figures 24.1 and 24.2 are important for interpreting the outcomes of a meta-analysis.

Studies

Following searching and critical analysis of the relevant literature, as discussed previously, the reviewer selected the five papers shown. If you look carefully at the sample sizes in each of the groups under Treated (odds) and Control (odds), you will find that the sample sizes are not equal. If they represent randomized trials, you might ask why the groups were unequal in the studies. Sometimes unequal groups represent people dropping out because of harmful side effects to the treatment.

Effect

The effect size was represented as an odds ratio (OR), which is a commonly used statistic for outcomes measured on a nominal scale. For this measure the outcome is: diagnosed with influenza following vaccination or placebo over, say, 6 months: 'yes' or 'no'. Looking at the first (English) study, the odds of 'yes' to 'no' are 30/530 for the treated group and 40/540 in the control group. The statistical software package computed an OR of 0.750, indicating a slight reduction in the odds of contracting influenza. Note that an OR = 1 means equal odds or no difference at all, while decreasing OR under 1.0 favours treatment. For example, an OR of 0.5 would indicate that the vaccination halved the odds for contracting influenza in the sample.

Confidence intervals

As we discussed in Chapter 17, a 95% confidence interval contains the true population parameter at $p = 0.95$. The lines produced from the 'squares' containing the sample OR represent the 95% confidence intervals. You can see in Figure 24.1 that all the confidence intervals for the five studies overlap with 1.0, indicating that we cannot infer that the studies favour the treatment.

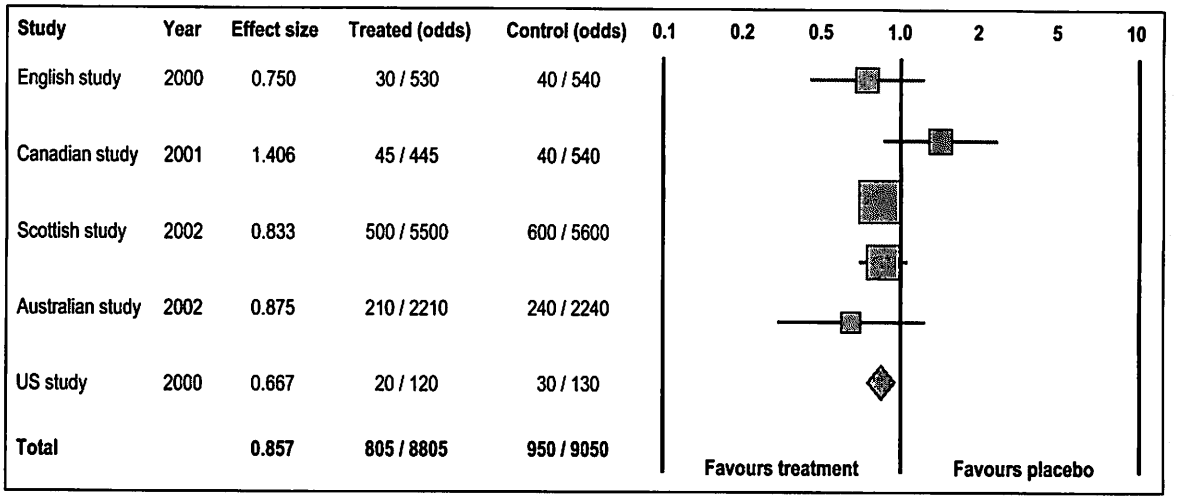

Figure 24.1 • Results of a hypothetical meta-analysis: negative findings.

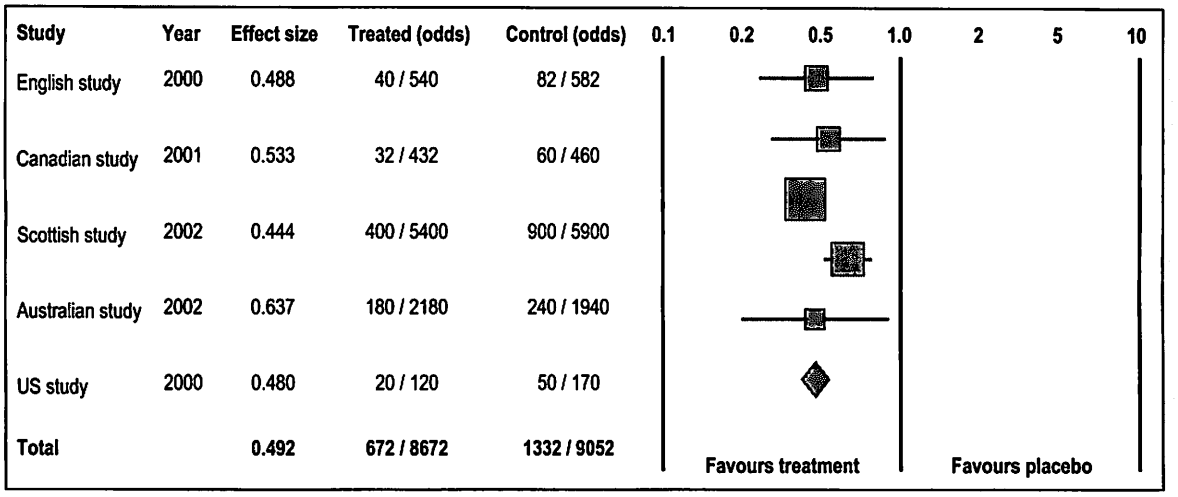

Figure 24.2 • Results of a hypothetical meta-analysis: positive findings.

Weights and totals

When you look at the 'odds' columns in Figures 24.1 and 24.2, you can see that there is a variation in the sample sizes used in the hypothetical studies. For example, the 'Scottish' study included 12 200 participants, while the 'US' study included only 300 participants. These differences contribute to the relative 'weight' of the study represented graphically by the area of the 'squares' in the forest plots. Clearly, the larger the square, the greater the sample size. You will also note that the confidence intervals are wider with the smaller squares in comparison to the larger squares. As discussed in Chapter 20, the larger the sample size, the more 'power' we have for making accurate inferences.

The 'Totals' in Figures 24.1 and 24.2 refer to the overall statistics synthesized from the results of the five hypothetical studies.. These statistics are represented by the diamond shapes on the

forest plots. In Figure 24.1 the total OR was 0.857, indicating a very weak effect for the vaccination. The OR is close to one or equal 'odds' for having influenza. For the results shown in Figure 24.2, the total OR was 0.492. This represents a strong effect, indicating that the odds for contracting influenza would have been more than halved by the vaccination. Such results are suggestive of the clinical or practical significance of introducing the treatment.

Validity of systematic reviews

The interpretation of the results of published meta-analyses is far more difficult than indicated in the above example. Let us look at some of the sources of difficulties.

Sampling

The results reported in the papers selected for review generally represent a sample of the total information published in a field. This leads us to the first problem: studies that do not report statistically significant findings are often not submitted or accepted for publication. By not having access to these 'negative' findings, the selection of papers becomes biased towards those with reported 'positive' outcomes. Also, in published research papers where outcomes for multiple dependent variables are reported, only the statistically significant outcomes are reported, undermining attempts to synthesize the evidence accurately (Polgar et al 2003).

In addition, the exclusion/inclusion criteria used for selecting the studies can result in a sampling bias. Some practitioners of evidence-based medicine (e.g. Sackett et al 2000) are reluctant to include studies which have not adopted a randomized experimental design. While this approach has strong methodological justifications, valuable information can be lost by using highly selective inclusion criteria. Of course, the more information that is lost, the weaker the external validity of the review or meta-analysis in relation to the 'population' of research results constituting a research programme. The reduced external validity is a trade-off for including only methodologically stronger studies in the review.

Extracting the data

Another source of error arises when extracting the data from an individual meta-analysis. Some authors report very clear, descriptive statistics but others report their results in an obscure, uninterpretable fashion. Also, some journals and authors only discuss the statistical significance of the results. Obscure and incomplete reporting of the evidence leads to errors in synthesizing overall statistics.

The state of the research programme

The validity of a systematic review or meta-analysis is limited by the methodological rigour and statistical accuracy of the studies selected for review. To put it bluntly, many health-related problems cannot be resolved and questions cannot be answered on the basis of the currently available evidence. An inconclusive systematic review or meta-analysis is not necessarily a waste of time. Although inconclusive attempts to synthesize data cannot be used to make valid clinical decisions, they provide strong evidence for gaps in knowledge and provide objective grounds for identifying further research required to advance the research programme (e.g. Polgar et al 2003). Until better evidence becomes available we simply provide the best practices suggested by tradition and experience.

The Cochrane Collaboration

Cochrane (1972), a Scottish medical practitioner, was one of the first influential practitioners in the modern era to advocate the systematic use of evidence to inform clinical practice.

In recognition of Cochrane's pioneering work, the Cochrane Collaboration, the Cochrane Library and the Cochrane Database of Systematic Reviews were established. The Cochrane Collaboration is now a large international venture with a series of special-interest groups commissioning and maintaining reviews on a wide range of topics. There are also detailed protocols that have been established

for the conduct and presentation of Cochrane systematic reviews. Although the scope of the Cochrane database is very broad, much of it is quite focused on intervention research, i.e. what intervention approaches work best for specific health problems and populations. The database is now expanding into other areas but there is a strong intervention focus.

The Cochrane approach adheres to a hierarchy of evidence. There are five levels of evidence, with systematic reviews of multiple randomized controlled trials at the top, followed by single randomized controlled trials, evidence from trials without randomization but with pre–post, cohort or time series measurement, evidence from non-experimental studies and, at the bottom level, opinions of respected authorities. Many reviews, however, only focus on the top two levels of the evidence categories. Obviously, because of the intervention-oriented nature of many Cochrane reviews, qualitative research, case study and policy research do not yet figure prominently in this system.

The Cochrane Database of Systematic Reviews is available on a wide range of websites in different countries. Access arrangements for the reviews vary widely from country to country. In some countries some fees are payable, but in others, such as Australia, the government has taken out a national subscription so that access may be freely available. In order to access the database, we suggest that you use a search engine to search for 'Cochrane Collaboration' and follow the links or consult your librarian.

Evidence-based practice

Although the name of evidence-based medicine or evidence-based practice is relatively new, the idea of using research evidence to inform the design of clinical interventions is very old. Muir Gray (1997) provides a very good review of this approach. The quality movement in health care, at which commentators such as Donabedian (1990) have been at the forefront, has also been influential in promoting the need for the systematic use of evidence to promote the delivery of high-quality health care.

As Muir Gray (1997) notes, evidence-based practice has at its centre three linked ideas. These are how to find and appraise evidence, how to develop the capacity for evidence-based decision making and how to get research evidence implemented into practice. The finding and appraising of evidence draw heavily upon systems such as the Cochrane Collaboration approach where evidence is systematically collected and appraised according to pre-defined principles and protocols. However, the facilitation and implementation elements of evidence-based practice are important additions to the basic establishment of research evidence to support particular approaches in the provision of health services. The evidence-based practice movement is based upon the recognition that the mere existence of evidence for the effectiveness of particular interventions does not mean that it will necessarily be effectively implemented. This is the new element of evidence-based practice, the systematic implementation of programmes based upon sound research evidence.

Summary

We have seen in this chapter that the advancement of knowledge and practice depends not only on the results of individual research projects but also on the information provided by the synthesis of the results across the literature. Where the projects share the same clinical aims and theoretical frameworks, they are said to constitute a research programme. In this chapter we outlined the process of identifying and selecting publications which contain theoretically or practically relevant research findings. We argued that reviewing a research programme is an active, creative process which is influenced by our expectations and attitudes. While absolute objectivity is not a realistic requirement of a reviewer, there are basic rules to ensure that the review of the evidence is carried out with a minimal degree of bias.

A systematic review of the literature proceeds by identifying the basic components of research papers for organizing the evidence. We use these dimensions to identify patterns or trends which enable us to synthesize the information and answer the research questions. When studies are sufficiently similar, their results can be condensed into

single statistics such as overall effect size. In this chapter, we examined how the results of meta-analyses are interpreted.

The relationships between the state of a research programme and practice are very complex. We examined hypothetical situations where there was a strong consistency in the research findings and clear trends emerging in the literature could be identified. In these cases, applying the results is relatively straightforward in that we can either adopt or reject the use of a treatment on the basis of the evidence. Systematic reviews and meta-analyses are essential means for identifying best practices available for our patients. However, even well-conceived reviews and meta-analyses can fail to identify clear trends or clinically meaningful effect sizes. When the evidence is inconclusive, we simply continue with traditional practices and identify further research required for resolving unanswered questions concerning improved efficacy. In this way research in the health sciences is a continuous process, producing better information for advancing theory and practice, as we discussed in Chapter 1 of this book.

Self-assessment

Explain the meaning of the following terms:

Cochrane Collaboration
combined or pooled results
evidence-based practice
exclusion criteria
extraction (of data)
forest plots
inclusion criteria
literature
literature review
meta-analysis
odds ratio
systematic review
weight (for a study)

True or false

1. The term 'literature' refers to the set of papers purposefully selected by a reviewer.
2. 'Trends' in the literature refer to consistent outcomes across a number of publications.
3. The conduct and outcomes of a carefully written review are not influenced by the theoretical positions of the reviewers.
4. The methodological principles relevant to conducting empirical research are not applicable to writing literature reviews.
5. Contemporary literature searches use both 'online' and traditional means for identifying relevant papers.
6. Explicitly stated inclusion and exclusion criteria ensure the objective selection of publications.
7. When conducting a systematic review, the 'population' is defined as the literature for an area of research.
8. Only studies employing randomized double-blind designs should be included in a systematic review.
9. Odds ratios are used to represent effect sizes for interval and ratio-scaled data.
10. An odds ratio of 0.95 indicates a large effect size.
11. An overall weighted mean is a useful statistic for synthesizing the results of studies using continuous outcome measures.
12. Inconsistent findings should not be included in a meta-analysis.
13. The results of a meta-analysis can be graphed as a forest plot.
14. The 'weight' of a study in a meta-analysis is inversely proportional to the sample size.
15. Confidence intervals are used to estimate the range for the probable true values of overall statistics.
16. The outcomes of systematic reviews and meta-analyses enable researchers and practitioners to make decisions based on overall evidence.

Multiple choice

1. Which of the following statements is true about systematic reviews?
   - *a* Systematic reviews are the same as narrative reviews.
   - *b* The 'system' referred to follows the problem-solving strategy of applied empirical research.
   - *c* Systematic reviews are appropriate for synthesizing evidence from qualitative research.
   - *d* Narrative reviews are preferable to systematic reviews for minimizing reviewers' bias in selecting key papers.
   - *e* Systematic reviews are alternative strategies to meta-analyses for synthesizing evidence.

2. Which of the following statements is false regarding the use of meta-analyses?
   a Meta-analyses are statistical strategies for synthesizing evidence from several empirical studies.
   b Meta-analyses can provide significant outcomes even when the individual studies lacked statistical significance.
   c Meta-analyses are alternative strategies to systematic reviews for synthesizing evidence.
   d Using meta-analyses increases the statistical power for identifying therapeutically useful outcomes.
   e Meta-analyses are an important component of evidence-based health care.
3. Consider the forest plots shown in Figure 24.1 and 24.2. These figures indicate that:
   a the larger the sample size in a study, the greater the weight contributed to the overall results
   b the narrower the confidence interval, the larger the sample size in an individual study
   c the key outcome measure is the total or overall effect size for the studies included in the analysis
   d all of the above (*a*, *b* and *c*) are true
   e none of the above (*a*, *b* and *c*) are true.
4. Which of the following is problematic for conducting a meta-analysis?
   a 'Negative' findings are sometimes not reported in health sciences journals.
   b Some outcomes are measured on a nominal scale.
   c Not all published research involves randomized controlled designs.
   d The Cochrane Collaboration favours the use of narrative reviews for synthesizing evidence.
   e Differences such as samples, measurement strategies, etc. in each of the studies makes the synthesis of the evidence 'invalid'.
5. The best evidence for implementing a treatment based on a meta-analysis is:
   a a large overall effect size
   b very wide confidence intervals
   c large overall sample sizes
   d both *a* and *b*
   e none of the above.

Assume that you are working as a rehabilitation provider in your area of professional practice. Your supervisor asks you to review the relevant literature in order to decide if 'Rehabilitation A' or 'Rehabilitation B' should be implemented at your centre. Having identified and critically evaluated the relevant studies in the area, you find only two well-designed comparative experiments in which the outcomes were measured on the same standardized inventory. The outcomes were as follows:

The *t* values were calculated for an independent groups *t* test, with non-directional H_A. Questions 6–10 refer to the above information.

	Type of rehabilitation	$\bar{X}$	df	*t*	*p*
Experiment 1	A	50	10	2.0	>0.05
	B	30			
Experiment 2	A	45	58	2.8	<0.01
	B	30			

6. A standardized inventory was used to measure rehabilitation outcome, where $\mu = 40$ and $\sigma = 10$. What was the effect size for Experiment 1? (Hint: use σ for calculating effect size.)
   a 0.05
   b 0.2
   c 1.67
   d 2.0
   e 10.0
7. On the basis of the evidence provided, a plausible explanation for the results of Experiment 1 being non-significant is that:
   a the sample size was too small
   b the variability was too large
   c the effect size was very small
   d both *a* and *b*
   e all of the above (*a*, *b* and *c*).
8. The results of Experiment 1 indicate the importance for evaluating the _____ of a statistical analysis.
   a application
   b sampling error
   c variability
   d power
   e reproducibility.
9. Assume that the higher the scores on the inventory, the better the rehabilitation outcome. If the clinical significance of the results is set

at *d* (effect size) ⩾1, then the trends shown in Experiment 1 and Experiment 2 indicate that:
   a both studies demonstrate clinical significance
   b both studies demonstrate statistical significance
   c Rehabilitation A should be implemented
   d both *a* and *b*
   e *a*, *b* and *c*.

10. Your supervisor is concerned that the literature did not demonstrate a clear trend, considering the different values for *p* in the two studies. If you are not in a position to carry out further research, you could test the hypothesis 'Rehabilitation A is superior to Rehabilitation B' by:
   a using a different statistical test
   b using your clinical knowledge
   c applying narrative analysis
   d using a computer model of the outcome
   e carrying out a quantitative meta-analysis.

Consider the following two hypothetical scenarios representing research findings for randomized controlled trials (RCTs) evaluating the differences between the outcomes of two treatments, X and Y. Assume that an independent samples *t* test was used to analyse the data. A positive effect size represents outcomes in favour of treatment X.

Questions 11–13 are based on the information below:

Scenario A

Study	Effect size (*d*)	df	*p*
A	1.0	30	<0.01
B	1.0	30	<0.01
C	1.2	14	>0.05
D	1.2	14	>0.05

Scenario B

Study	Effect size (*d*)	df	*p*
A	0.2	80	>0.05
B	0.2	100	>0.05
C	−0.1	200	>0.05
D	−0.2	200	>0.05

11. Which of the following is correct for Scenario A?
   a There was a large effect size.
   b The results were consistently in favour of treatment X.
   c The outcomes for the four studies were statistically significant.
   d Both *a* and *b*.
   e All of the above (*a*, *b* and *c*) are correct.

12. Which of the following is correct for Scenario B?
   a There was a large effect size.
   b The results were consistently in favour of treatment X.
   c The outcomes for the four studies were not statistically significant.
   d Both *a* and *b*.
   e All of the above (*a*, *b* and *c*) are correct.

13. Which scenario demonstrates that treatment X is more effective than treatment Y?
   a Scenario A, because all the studies provided statistically significant outcomes.
   b Scenario B, because all the studies provided *p* values greater than 0.05.
   c Scenario A, because the studies provided evidence for large and consistent effect sizes.
   d Scenario B, because the studies provided evidence for large and consistent effect sizes.
   e Scenario B, because the sample sizes in the studies were significantly larger than those in Scenario A.

14. A well-known project for constructing and disseminating the results of systematic reviews and meta-analyses online is called:
   a the Cochrane Collaboration
   b the Meta-Analyses Database (MAD)
   c Medline
   d CINAHL
   e Embase.

Section Seven

Discussion, questions and answers

This question is based on a survey which was published in an Australian newspaper. Of course, such surveys do not represent research published in scientific journals, but they are important sources for public knowledge and/or attitudes towards health sciences issues. The survey questioned a sample of adults concerning their smoking habits. Only one of the questions asked is discussed here and the results are hypothetical.

Table D24.1 Survey characteristics

Sample	1000 voters
Coverage	Australia-wide
Method	Telephone
Question	Do you smoke? (Yes or No)

Table D24.2 Results

Percentage of replies to the question in two major cities		
	Melbourne	Sydney
Yes	24	18
No	76	82

Questions

The following questions involve the critical analysis of the above survey.

1. If we assume that cigarette smoking is now a 'stigmatized' behaviour, do you think the telephone survey produced valid answers?
2. A total of 180 people were interviewed in Melbourne and 220 in Sydney. If the population of Australia is 17 million and the populations of Melbourne and Sydney are 2.5 and 3.2 million respectively, do the samples appear to be quota samples?
3. Which categories of smokers may not have been reached by this survey? What implications might this have for the external validity of the survey?
4. A journalist commented on the results, saying: 'This difference is ironic, given that anti-smoking lobbyists have applauded Melbourne as a pacesetter for smoking law reform, such as tobacco tax-funded health promotion'.
   (a) Explain why this comment is inappropriate given the design of the survey.
   (b) What research design would be appropriate to show a causal effect on smoking due to health promotion on smoking? (Hint: see Ch. 6.)
5. Explain why the comment quoted in question 4 is inappropriate, given that the statistical significance of the results was not calculated.
6. Which statistical test should be used to analyse the significance of the results concerning

differences in smoking between the two cities? Justify your selection.

7. Setting $\alpha = 0.05$, calculate the statistic and decide if the results were significant (note that we gave the results in percentages).
8. Do you think the sample size ($n = 1000$) was adequate? Explain.

Answers

1. Although telephone interviews and mailed-out questionnaires are a relatively cost-efficient strategy for collecting data, we have problems validating the responses. This is particularly true for conditions and behaviours which are socially stigmatized: why should respondents disclose such information about themselves? In face-to-face interviews, we can explore issues, for example if the respondents have nicotine-stained fingers or smell of cigarettes, we may pursue the issue further to establish the accuracy of the replies.
2. Given that $n = 1000$, 18% of the respondents were from Melbourne and 22% from Sydney. For a quota sample, the expected samples would be:

$$\text{Melbourne: } \frac{2.5}{17} \times 100 = 14.7\%$$

and

$$\text{Sydney: } \frac{3.2}{17} \times 100 = 18.8\%$$

Assuming that the information used to calculate the above figures is correct, it seems that the sample included more respondents from Melbourne. This may reflect the different proportion of 'voters' in the two cities, or a rather poor quota sample.

3. People who are not on the electoral roll, such as persons under 18 years of age, and people who do not have or do not answer their telephones, would not have been contacted. In this way, the sample may not be representative of all the smokers in the city (e.g. young people, poor or itinerant people, people with unlisted telephone numbers). Therefore, the survey may not be externally valid if we generalize to all persons in Australia who smoke.
4. (a) The present survey did not tell us how rates of smoking have changed over a period of time.
   (b) We may use a quasi-experimental design and introduce the programme in one city, A, but not in the other equivalent city, B. If the reduction is greater over time in A than in B, we may argue that this difference could reflect the causal effect of health promotion.
5. Although results for the samples show a difference between the two cities, this may simply reflect sampling error. We must establish the significance of the results before we can draw inferences ('ironic' or otherwise) about populations.
6. χ^2; nominal data and independent measurements or samples.
7. Convert the data into frequencies (see Ch. 19) before entering obtained values into a 2×2 contingency table (values rounded to closest whole number).

Table D24.3

	Melbourne	Sydney	Total
Smokers	43 (cell 1)	40 (cell 2)	83
Non-smokers	137 (cell 3)	180 (cell 4)	317
Total	180	220	400

Expected values (for calculation procedure, see Ch. 19):

$$f_e \text{ (cell 1)} = \frac{83 \times 180}{400} = 37.4$$

$$f_e \text{ (cell 2)} = \frac{83 \times 220}{400} = 45.7$$

$$f_e \text{ (cell 3)} = \frac{317 \times 180}{400} = 142.6$$

$$f_e \text{ (cell 4)} = \frac{317 \times 220}{400} = 174.3$$

$$\chi^2_{obt} = \Sigma \frac{(f_o - f_e)^2}{f_e} = 1.96$$

Table D24.4 Calculation of χ^2

Cell	f_o	f_e	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_o}$
1	43	37.4	31.36	0.84
2	40	45.7	32.49	0.71
3	137	142.6	31.36	0.22
4	180	174.3	32.49	0.19

Critical value of χ^2; $\alpha = 0.05$ where degrees of freedom (df) = 1 (Appendix C)

$\chi^2_{crit} = 3.84$

In this case we would retain H_0: there is no association between the variables 'city' (Melbourne or Sydney) and smoking (Yes or No). (For details of the decision-making process, refer to Ch. 19.) It is apparent that the results are not significant, therefore we are not justified in drawing any inferences concerning the different proportions of smokers in Melbourne and Sydney.

8. Although a sample size of $n = 1000$ appears quite large, this was an Australia-wide sample which was divided up to represent regions. It is possible that the null results obtained in question 7 are because there are no differences in smoking rates between the two cities, but there are other possibilities (see Ch. 20). Perhaps the sample size was inadequate and we made a Type II error in our decision. Replicating the study with larger sample sizes might enable us to show significant differences in smoking rates.