CHAPTER 7

READING THE RESULTS OF EXPERIMENTAL RESEARCH

Introduction — 1 Statistical significance — 2 χ^2 as an example of a statistical significance test — 3 Tests, questions and hypotheses — 4 Confidence intervals — 5 Confidence intervals and meta-analyses — 6 Non-parametric and parametric statistics — 7 Sample size — 8 Statistical power — 9 Measures of central tendency, dispersion and diversity: modes, medians, means, variance and standard deviation — 10 Expressions of effect size — 11 Counting costs — 12 Sensitivity analysis — 13 Further reading on understanding the results of research — References and further reading

Introduction

This chapter provides information to help in deciphering the numerical presentation of research results, which people often find daunting.

1 Statistical significance

The results of experiments are usually tested for statistical significance. Results that are statistically significant are results that are most unlikely to have arisen by chance. Your friend deals you ten red playing cards from what she says is a standard shuffled pack. You know that this is very unlikely but not impossible. Having shuffled the pack again she deals you four reds. How suspicious would you be now? In fact, of all four-card deals 5.5 per cent of them will be all reds. One such deal should come up roughly every 18 deals. Ten reds running will only occur once in 3,333 deals. Knowing these odds, you know that ten reds is much more unlikely to be a chance deal than four reds is likely to be.

Testing for statistical significance is a matter of checking what actually happened against an estimate of how often it might have happened by chance. The principle is easy to understand, but lots of people get bogged down in the mechanics.

2 χ^2 as an example of a statistical significance test

The most transparent of all tests is one used in the research featured in the exemplar study for Chapter 1. It is called χ^2 and pronounced 'Ki-square'. How it is calculated is demonstrated below.

The Observed (or 'O') figures are what actually happened. The O figures in Table 7.1 show that of 35 ulcers 16 healed. Eight of these were treated with one kind of bandage and eight with another. It should be fairly obvious from 'eye-balling' the data that the differences in healing rates (8/17 and 8/18) are just the kind which might have occurred by chance with two treatments of equal effectiveness. Though this is obvious, the text below will show how the same conclusion can be reached statistically.

The Expected (or 'E') figures are what would be most expected to happen by chance. In this case they are calculated simply on a fair-shares basis, sharing out the healings, non-healings and withdrawals proportionately between the two bandaging systems. Thus 7.77 = seventeen thirty-fifths of 16, which is the Charing Cross 'fair-share' of all healings. Fair-shares is what would be most expected if the results were due to chance. The further calculation is simple:

For each pair of cells calculate $\frac{(O - E)^2}{E}$

Then add up all the results:

$$\frac{(8 - 7.77)^2}{7.77} + \frac{(8 - 8.23)^2}{8.23} + \frac{(5 - 6.32)^2}{6.32} + \frac{(8 - 6.69)^2}{6.69} + \frac{(4 - 2.91)^2}{2.91} + \frac{(2 - 3.08)^2}{3.08} = 1.3 = \chi^2$$

In this calculation the subtractions are the comparisons between the actual figure (O) and what is most to be expected to happen by chance (E). Logically then, the bigger the resulting figure the more

Table 7.1 A χ^2 calculation with the data from the leg ulcer bandaging trial (Chapter 1)

	Charing Cross Bandaging System		Trial Bandaging System		
	Observed	Expected	Observed	Expected	Total
Healed	8	7.77	8	8.23	16
Not healed	5	6.32	8	6.69	13
Withdrawn	4	2.91	2	3.08	6
Totals	17	17.00	18	18.00	35

$\chi^2 = 1.3$; df = 2; $p = 0.51$.

different the observed figures will be from the expected figures and the less likely the results were due to chance.

The figure for χ^2 is then looked up in a ready-reckoner table, two rows of which are given below as Table 7.2

In Table 7.2 df stands for 'degrees of freedom'. There are two degrees of freedom in Table 7.1, since once the row totals and the column totals are filled in, filling in *two* of the remaining O cells determines all the content of all the others: think of crossword puzzles. The content of two cells are free to vary; hence two degrees of freedom. Usually the formula for calculating degrees of freedom is (columns − 1) × (rows − 1) = df: in this case (2 − 1)(3 − 1) = 2.

The *p* stands for probability and the *p* values provide an estimate of the likelihood of a particular value of χ^2 occurring by chance. In terms of the card deals referred to in section 1, four reds running would have a probability of just a bit more than 20 per cent ($p = 0.20$) and 10 reds running have a probability of just over 1/3000. Usually this latter would be expressed as $p < 0.001$ (less than one in one thousand). Probability values are ubiquitous in tables of experimental results and the top line of Table 7.2 will serve as a useful resource for you if you find it difficult to remember what *p* values mean.

Table 7.2 shows that, at two degrees of freedom, a value for χ^2 of 13.82 or more will occur by chance less than once per 1000, i.e. $p = 0.001$. If we had a result like that we could be very confident that it was not due to chance. It would be a highly statistically significant result.

A value of 5.99 or more will occur less than 5 times in 100, i.e. $p = 0.05$. By convention statisticians will not accept as significant any value of *p* greater than 0.05 (or 'the 5% level'). The question to be asked about the value of χ^2 obtained in the calculation above is 'is it equal to or bigger than the 0.05 value?'

The 0.05 (5%) value is 5.99. The value given by the calculation was 1.3. This is much, much smaller. This tells us what we already knew, that the result is not statistically significant. It also tells us how

Table 7.2 Critical values of χ^2 level of significance for a two-tailed test*

p =	0.90 90% likely by chance	0.70 70% likely by chance	0.50 50% likely by chance	0.20 20% likely by chance	0.10 10% likely by chance	0.05 5% likely by chance	0.02 2% likely by chance	0.01 1% likely by chance	0.001 1/1000 likely by chance
df = 1	0.02	0.15	0.45	1.64	2.71	3.84	5.41	0.64	10.83
df = 2	0.20	0.71	1.39	3.22	4.61	5.99	7.38	9.21	13.82

* 'Two-tailed' means that we are interested in a difference in any direction. In terms of Table 7.1 that means any difference whether this is in favour of the Charing Cross bandages or in favour of the trial bandages. Most significance testing is two-tailed (Wright, 1997: 39)

statistically *in*significant it is. It lies between the value where $p = 0.70$ and where it equals 0.50. So if χ^2 equals 1.3 then differences of this size would crop up as chance variations somewhere between 50 and 70 per cent of times if the experiment were repeated again and again.

3 Tests, questions and hypotheses

Researchers often frame their questions in terms of hypotheses. In the leg ulcer bandaging trial (Chapter 1) the so-called *null hypothesis* (or H_0) would have been that: 'there is no difference in effectiveness between the two bandaging systems greater than might have been expected by chance'. The so-called *experimental hypothesis* (H_E) would be that there is a statistically significant difference in effectiveness between the two bandaging systems. Saying that results show no statistical significance is the same as saying that the null hypothesis should not be rejected, and that the experimental hypothesis falls.

There is actually an opportunity for two null hypotheses here:

H_{0i} – there is no difference in drop-out rates as between the two bandaging systems greater than might be expected by chance.

H_{0ii} – there is no difference in healing rates as between the two bandaging systems, greater than might be expected by chance, when the effect of drop-out has been accounted for.

The way the calculation above (Table 7.1) was done tested both null hypotheses in a single test. But there might have been a statistically significant difference in drop out rates between the two bandaging systems, but no significant difference in healing rates for those ulcers remaining in the trial, *or* no significant difference in drop out rates, but a significant difference in healing rates for those ulcers remaining in the trial. In fact this is not so, but it is important to note that a single statistic, χ^2 in this case, may be measuring several differences at once. To compare the two bandaging systems for drop out rates only it would be necessary to use a two row table amalgamating the healed and the not-healed ulcers. Then the only difference visible to the statistical test would have been that between withdrawals and non-withdrawals. The lesson here is that the way the data should be set up for testing will depend on the hypothesis being tested.

4 Confidence intervals

The TV pathologist says 'Time of death twelve midnight, give or take two hours either side'. The 'give or take two hours' defines a confidence interval. Similarly, experimental results are always regarded as estimates of the true state of affairs and are usually cited with confidence limits. The 95% confidence intervals are those most usually used, but

the 99% intervals are not uncommon. The assumption being made is that if the experiment were repeated again and again and again it would produce a series of different results, but 95% of these would fall within the 95% confidence limits. For example, for the leg ulcer bandaging trial it can be calculated that 95 per cent of such results would fall between 5 healings for the Charing Cross Bandages and 11 for the Trial Bandages at one extreme, and 11 healings for the Charing Cross Bandages and 5 for the Trial Bandages at the other extreme. This is another way of saying that in this trial one bandaging system would have to show at least 7 more healings than the other to be regarded as superior, which is much the same as saying that anything less than a difference of 7 out of 16 here would not be statistically significant at the 0.05 (5%) level.

5 Confidence intervals and meta-analyses

Confidence intervals are particularly useful for making 'at a glance' judgements about the meaning of experimental results. This is shown particularly when they are used in meta-analyses, but you can regard this section as an illustration of how to interpret confidence intervals however they are used.

Meta-analyses involve bringing together the findings of a number of experiments or trials and comparing the results. They are dealt with in further detail in Chapter 4. Very often they include a diagrammatic synopsis of results, like Figure 7.1, which puts confidence intervals to good use.

Figure 7.1 The effect of home visiting in preventing childhood injury: eight randomised controlled trials reviewed by Roberts et al., 1996: odds ratios and 95% confidence intervals (see Chapter 4)

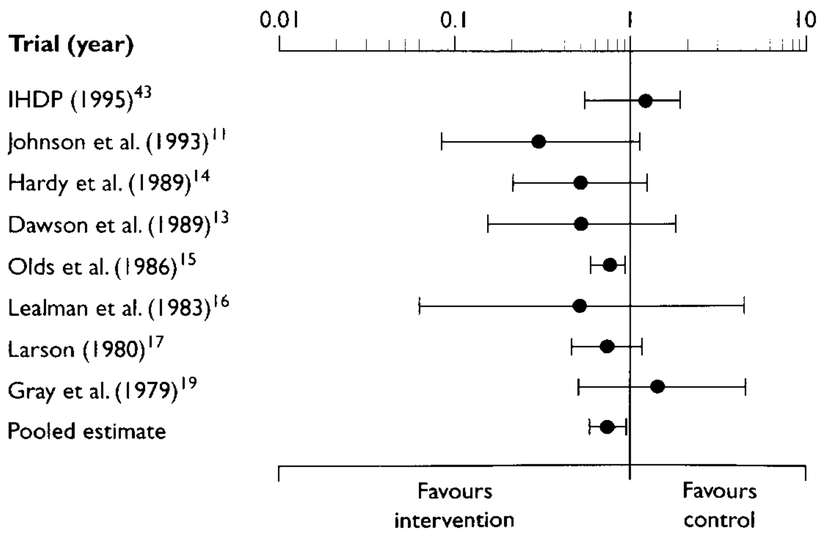
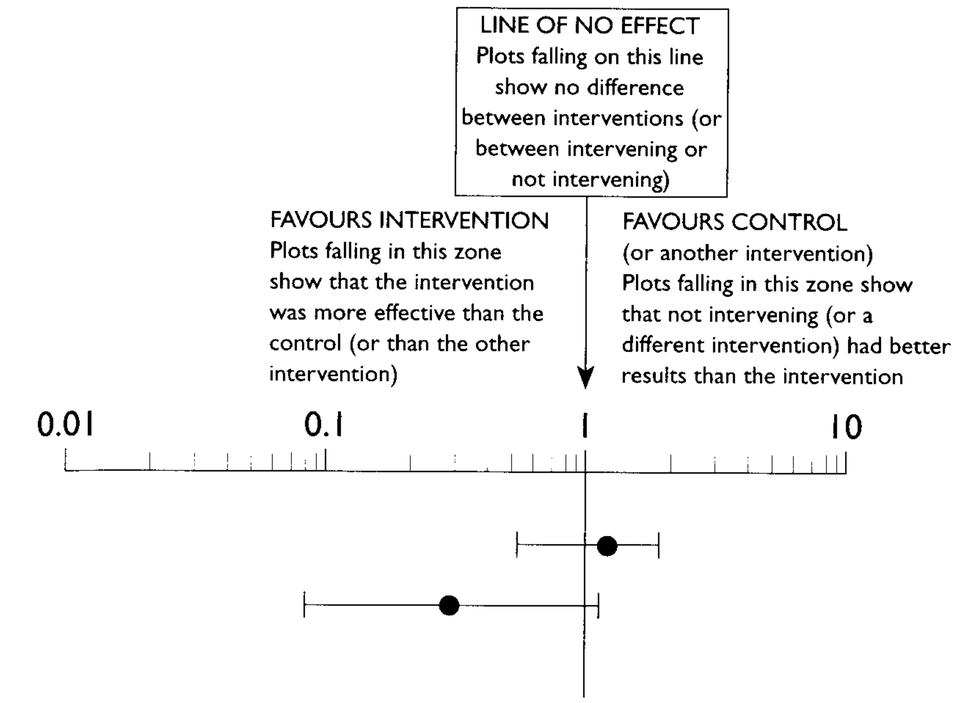


Figure 7.2 First two lines from Figure 7.1 annotated



Each of the trials involved comparing rates of injury for children who were and children who were not visited as part of a visiting scheme. The diagram shows the differences between the children visited and the controls as odds ratios for each of the trials reviewed. Odds ratios are explained in section 10.4, but understanding odds ratios is not necessary for understanding diagrams like this.

The first line on Figures 7.1 and 7.2 is for a trial conducted as part of the USA Infant Health and Development Program (IHDP) in 1995. The dot, or 'plot' on the line gives the actual result. It falls to the right of the vertical line on the side marked 'Favours Control'. In this trial there was less childhood injury among those who were not visited, than among those who were. The vertical line is called 'the line of no effect'. If the actual results of a trial fell on this line this would indicate that there was roughly the same amount of childhood injury among those who were visited as among those who were not: no greater difference than might have been expected to occur by chance. The next trial on the diagram has its results well to the left of the line of no effect, on the side which 'Favours Intervention'.

The plots show the actual results of the trials. But the results of a trial are always regarded as estimates, on the assumption that had the trial been conducted over and over again in the same way its results would tend to cluster round a mid-point, but would not be identical. The horizontal lines (or whiskers) show how wide the

estimates are. In this case the estimates are the 95% confidence limits (see section 4). What the limits suggest is that we can be 95% certain that the true result lies somewhere on the horizontal line. Sometimes 99% confidence intervals are used instead.

The plot for the IHDP study was to the right of the midline, favouring not visiting (Figure 7.2). But one of the whiskers strays over the line into the area favouring intervention. Maybe if the IHDP study had been done again in exactly the same way except that the intervention and the control groups were made up of different families, its results would be to the left of the line rather than to the right, favouring intervention. Confidence limits that stray across the midline indicate that no confident judgement can be made as to whether two interventions are different in their effects.

For relatively rare events such as childhood injury the sample sizes for each trial here are rather small. The IHDP is the largest, with 345 visited and 551 controls, and the smallest is Gray et al., with 26 visited and 25 controls. Small samples tend to give wide confidence intervals – long whiskers, and hence imprecise estimates. Figure 7.1 shows seven of the eight trials with confidence intervals crossing the line of no effect. The last entry on Figure 7.1 pools the results of all trials. It treats all the trials together as if they were just one randomised controlled trial involving several thousand subjects. The plot for the pooled results falls to the left, favouring intervention. But, more importantly, its confidence intervals fall entirely to the left, so that there is a 95 per cent chance that the results of all the trials together indicate in favour of visiting rather than not visiting, though the benefits of visiting schemes still seem to be rather small.

The same logic can be used in interpreting confidence intervals where effects are shown in different ways. But it is important to be clear as to how the different ways of reporting results differ in expressing equivalence. With odds ratios and risk ratios (section 10.4) equal effectiveness is expressed as 1, and the line of no effect (Figure 7.2) runs through the 1s. But where effects are shown by subtracting averages (section 10.1), or subtracting proportions (section 10.2), or using a standard deviation (section 10.3), equivalent effectiveness is shown as zero. Hence with these other ways of interpreting effects it is 0 (as it were) which marks the 'line of no effect', and confidence intervals which stray over that line are likely to indicate non-significant results.

6 Non-parametric and parametric statistics

Statistical tests are divided into two kinds:

- Non-parametric tests, which are the only kinds which should be used with nominal data (see Box 6.1 in Chapter 6). Using non-parametric tests with higher level data is possible, but non-parametric tests cannot 'see' all the information included in interval or ratio level data.
- Parametric tests, which should not be used with nominal level data, which are, but perhaps should not be, used with ordinal level data, and which make the best use of all the information contained in interval and ratio level data.

Parametric tests should only be used when a number of conditions are satisfied. These include that:

- A sample is drawn from a population in which the characteristic of interest is *normally distributed*. If it were graphed it would show a bell-shaped curve, with the mean or average somewhere very close to the middle of the range. Some of the complicated statistical manoeuvres to be found in the literature derive from converting non-normally distributed data into normally distributed data: for example, by using the logarithms or square roots of scores rather than the scores themselves. In Chapter 4, for example, Roberts et al. quote their results in terms of 'an inverse variance weighted average of the study specific odds ratios'. The 'inverse variance weighted average' is a way of converting scores that are not normally distributed into scores that are, so that they are amenable to processing using parametric statistics.
- When comparing two samples of dissimilar size, the variance of the two samples should be similar. Roughly speaking, the statistical concept of variance refers to the degree of variability *within* samples (see section 9).

In Chapter 2 Marshall, Lockwood and Gath write: 'The data were first evaluated to ensure normality of sampling distributions, linearity and homogeneity of variance.' This means that they checked to see whether it was permissible to use the parametric statistics they proposed to use.

7 Sample size

The extent to which confident conclusions can be drawn from experiments depends on sample size. Unfortunately, it is very difficult to lay down general rules as to what is an adequate sample size, but Table 7.3 gives some rules of thumb for judging whether a sample size was adequate.

Most of the discussion of sample size revolves around the second item in Table 7.3: the size of the difference which can be detected in

Table 7.3 Some considerations in judging adequate sample size

A sample size of 40 would be adequate if:	A larger, sometimes much larger, sample would be needed if:
The experiment only has two arms and there are no less than 20 in each arm	The experiment has more than two arms (20 for each arm would be a usual minimum)
There is no interest in differences smaller than 10 per cent	There is an interest in differences smaller than 10 per cent
And/or the events of interest are common	And/or the events of interest are rare (an experimental evaluation of a suicide prevention scheme would require a very large sample, since suicide is rare even among those of highest risk)
The subjects very similar to each other in the ways relevant to the experiment	The subjects are very diverse in ways relevant to the experiment – then a large sample is needed to ensure the same range of characteristics in each arm of the experiment (see Chapter 5, section 3 on forming comparison groups)
The interventions are highly standardised within each arm	The interventions within each arm are diverse. A 'two arm' trial with diverse interventions within each arm is not really a 'two arm trial' but a many armed trial and there should be at least 20 subjects for each of the diverse interventions *within* an arm
There is no interest in what happens to sub-groups within the different arms of the trial	There is an interest in what happens to sub-groups. The logic here is to think of the trial as having as many arms as there are sub-groups, e.g. male controls, female controls, males treated, females treated. In that case there should be at least 20 in each group, disregarding other considerations in this table
The range of possible outcomes is limited and/or the measuring scale to be used has a limited number of measurement categories	There is a large array of possible outcomes/the measuring scale has a large number of categories. Obviously if 100 grades of outcome are possible a sample of 40 is too small

an experiment with a sample of a given size, which is the idea of 'statistical power'.

8 Statistical power

Table 7.4 shows that two different kinds of errors can occur in interpreting the results of statistical tests. For example, there may really be no significant difference in the healing power of two bandaging systems, but it is assumed erroneously that there is – a type I error. Or there may really be a difference and it is assumed erroneously that there is not: a type II error. In the leg ulcer bandaging trial (Chapter 1) this is the more likely error because of the small sample size.

The more one kind of error is avoided, the more likely the other is to be committed. However, researchers would prefer not to make any

Table 7.4 Type I (alpha or α) and type II (beta or β) errors

	We assume that:	
	There is no difference (we do not reject the null hypothesis)	**There is a difference (we reject the null hypothesis)**
In reality there is no difference	Our judgement is correct	We make a type I error (an alpha error): we reject the null hypothesis when we should uphold it
In reality there is a difference	We make a type II error (a beta error): we do not reject the null hypothesis when we should reject it	Our judgement is correct

errors at all. There is something of an art about choosing both a sample size and a statistical test which will produce the optimum balance in avoiding both kinds of error, while at the same time not requiring a sample size which is prohibitively expensive.

Improving the power of a test means lowering the risk of failing to detect real differences – that is lowering the risks of making type II errors. This might be done in various ways:

- Increasing the sample size.
- Using a higher level of data (see Box 6.1 in Chapter 6).
- Using a more powerful statistical test.
- Improving the design of the experiment.

But usually the power is set by choosing an appropriate sample size. The calculation of an appropriate sample size involves the researcher:

- Declaring the size of the risk he or she is willing to take of making a type I error: the error of assuming a difference when really there isn't one. The choice is often 5 per cent (the 95 per cent limits again). In this context this is often called the alpha level.
- Declaring the size of the difference of interest. In health and social care often only largish differences will be of practical significance. No one is going to invest large sums of money, or undertake large scale service reorganisations, to boost their success rate by 1 per cent for non-life-threatening matters.
- Declaring the size of the risk he or she is willing to take of committing a type II error: the error of mistaking a real difference for a non-significant one – setting the beta level.

Sometimes you will read something like this:

> Sample size for patients having a hysterectomy had a power of 80% with an α of 0.05, to detect a change of 10 points on a physical functioning domain of the SF-36, based on a standard deviation of 18.7. (Shepperd et al., 1998: 1787; see also Chapter 3 in this volume)

This means that the researchers have decided that they are not interested in differences of less than 10 points on part of a standard instrument called SF-36 which measures health and well-being (Chapter 6, section 1). Ten points was chosen as being a *clinically* important difference: the kind of outcome difference which, if it were shown between two different treatments, would be a persuasive case for choosing the better one. The SF-36 is a very widely used instrument and usually produces results with a standard deviation of around 18.7 (see section 9). The authors are willing to take a 5 per cent risk of mistaking a chance difference for a real one: α is 0.05 But they want to be 80 per cent certain that they will not mistake a real difference of 10 points for a chance difference. The minimum sample size is set by entering all this information into a calculation (Cohen, 1988; Altman, 1991). For readers this information is useful insofar as it warns them not to take too much notice of any comments by the researchers about differences smaller than those nominated – here smaller than 10 points on this scale, because their research design makes it unlikely that they can say anything sensible about differences smaller than this. It also tells readers that if there are real differences of 10 points or more then the researcher has an appropriate sample size for detecting them.

Decisions about power are not just about statistical significance. They are also about *substantive*, or 'clinical' significance. The appropriate power depends on the practical implications of making a decision on the basis of results that turn out to be wrong. For example, making the assumption that there is no difference in survival rates between two clinical operations, when there really is a difference – a type II error – may have very serious consequences if a clinician then chooses to use the technique with the higher death rate. In an experiment providing data for making a decision like this it would be wise to set a high beta level to give the best chance of detecting any difference between the treatments. By contrast, in situations where the consequences of making the wrong decision are not particularly serious, or where it would take a big difference to persuade someone to change their practice, experiments with relatively little power may be quite adequate.

9 Measures of central tendency, dispersion and diversity: modes, medians, means, variance and standard deviation

Measures of central tendency express the extent to which results cluster, or the extent to which they are spread out. The arithmetical average or 'mean' is the most familiar of these – add up all the scores

and divide by the number of scores. However, quoting the average/mean can sometimes be misleading, for the following reasons.

- A handful of very small or very large scores can skew the mean so that most results fall above or below it. With a skewed distribution it is often better to cite the *median* – which is the score below which 50 per cent of scores fall, and above which 50 per cent of scores fall. Sometimes when the median is used, statistical calculations are done only with the *inter-quartile* range, that is with the middle 50 per cent of scores: the 25 per cent of all scores which fall immediately above and the 25 per cent of all scores which fall immediately below the middle score: 25 per cents are 'quartiles'. The reason for doing calculations with the score in the inter-quartile range is that the further scores are from the median the more erratic they are likely to be. For an example see the paper by Shepperd et al. in Chapter 3 of this volume (Table 3 on p. 30).
- A measure of central tendency has little meaning without reference to the *range*, which is the distance between the highest and lowest scores.

The other main measure of central tendency is the *mode*, which is the most commonly occurring datum. This is the only measure of central tendency possible with nominal level data (see Box 6.1, Chapter 6). In Table 1 in Chapter 1, for example, the mode for all ulcers is 'healed', as it is for the Charing Cross bandages alone. The outcome for the Trial bandages is bimodal, with 8 healed and 8 not healed. The mode is rarely a very useful statistic.

The *variance* is a measure of how variable the data are in terms of whatever is being measured. The *standard deviation* combines in one statistic both a measure of how much scores are clustered around the mean, and how much they are dispersed: the smaller the standard deviation, the more clustered the scores are around the mean. The calculation of the standard deviation is embedded in most statistical tests and cited in many tables of results (SD, sd, s, S or δ) and is involved in calculations of statistical power. Any standard statistics textbook will explain how these are calculated. You don't have to know how to calculate the variance or standard deviation in order to read research as a practitioner. However, the two statistics do give some 'at a glance' information which is useful. For example:

- If all the subjects had the same score then the variance and the standard deviation would be zero. In comparing two groups of subjects, the one with highest variance or standard deviation is the one showing most diversity in scores. Thus, if two groups of subjects in an experiment show much the same mean, but very different standard deviations or variances, then they differ from each other

in that one group is more homogeneous than the other. Knowing this can be important in judging the possibility of regression to the mean effects (see Chapter 5, section 6) and judging whether it is permissible to use parametric statistics (section 6 above).

- With variables that show a normal distribution the figure for the standard deviation can be used to calculate the confidence intervals. The 95% confidence limits will lie at the mean plus two standard deviations and at the mean minus two standard deviations. Thus if the mean were 62 and the standard deviation were 22, the 95% confidence limits would be 18 and 106 respectively, and it would be a reasonable expectation that 95 per cent of all scores would fall in that range. The 99% confidence limits are the mean plus and minus three standard deviations. Though they are rarely used, the 68% confidence limits are the mean plus and minus one standard deviation.

10 Expressions of effect size

Experiments are usually designed to demonstrate the size of the effect of doing one thing rather than another. For example, it might be the difference between treating and not treating, or the difference in outcomes between two treatments. Some of the more usual ways of expressing effect size are given below (sections 10.1–10.5). Note, however, that a large effect size is not necessarily a significant effect size. A small effect shown by comparing two large groups will be more significant (and probably more real) than a large effect shown by comparing two small groups. To be important effect size needs to be both statistically significant and to be important in a practice-relevant way.

10.1 Showing effect size by subtracting averages

The simplest way of expressing an effect is to subtract the average results of a group who have been treated in one way from the average results of a group who have been treated in another. This is the approach adopted by Marshall et al. in the exemplar study in Chapter 2. Two lines of their table of results are reproduced below as Table 7.5. Column 6 is the result of subtracting group averages.

In Table 7.5 REHAB GB (column 1) is a rating scale for observations of social competence, the observations being conducted by observers trained to do so. Several observers were involved so this raises issues about inter-rater reliability (see Chapter 6, sections 5 and 6). There are two arms to the experiment, hence two groups: a control group receiving ordinary mental health care and a group experiencing case management (column 2). Baseline scores (column 3) refer to the

Table 7.5 Outcomes of a trial comparing case management and normal practice in mental health care (see Chapter 2)

1	2	3	4	5	6	7	8
Measure	Group	Baseline score (*n*)	Change at 7 mth (*n*)	Change at 14 mth (*n*)	Mean difference at 14 mth (95% CI)	*F*	Clinically relevant difference
REHAB GB	Control	44.7 (40)	4.3 (35)	4.9 (30)	4.3 (−4.9 to +13.4)	0.87	15
	'Case-man.'	42.2 (40)	5.3 (34)	7.5 (31)			

average score on REHAB GB for each group, with the number of subjects in each group in brackets – 40 each at this stage. Columns 4 and 5 show changes to these average scores, which are improvements for both groups, though greater for the case managed group. But these columns also show drop out, so that by 14 months there were only 30 controls and 31 case managed clients left in the experiment. Improvements might be due to those least likely to recover disappearing (see Chapter 5, section 8). The improvements quoted here will not be improvements for all 40 in each group, but improvements on the average baseline scores of just those left in the trial in each group. Column 6 subtracts the mean scores of the two groups at 14 months. Those left in the case managed group showed a greater improvement than those left in the control group to the extent of 4.3 REHAB GB points. As always, the results are regarded as an estimate. The 95% confidence limits quoted suggest that there is only a 5 per cent chance that the true figure lies beyond 13.4 points greater improvement for the case managed group on the one hand, or beyond 4.9 points greater improvement for the controls on the other. Since the confidence interval spans scores which would show more improvement for the controls, and scores which would show more improvement for the case managed group, the conclusion to draw is that there is no significant difference in improvement between the two groups. The statistic for *F* (column 7) is akin to χ^2: the result of a test for statistical significance, which can be looked up in a ready-reckoner of critical values for *F* (similar to Table 7.2). At 0.87, *F* is not statistically significant, this being shown (as is often done) by the absence of any note to say it is, although sometimes researchers will give *p* values for non-significant results (for *p* values see section 2). Hence the difference between the two groups on this line of the table is within the range that might be expected to have occurred by chance. Column 8 gives a figure for the size of the difference between groups which practitioners would regard as clinically important – a statement of substantive, rather than statistical significance. Roughly speaking, this means that practitioners would think something important had happened if someone improved their REHAB score by 15 points. The measured difference between the two groups was only 4.3 points, and 15 is not even within

the confidence limits range. From a practitioner's point of view the two groups were just as much like each other at the end as they were at the beginning in terms of social competencies.

10.2 Showing effect size by subtracting proportions and rates

Another way of expressing effects is to convert results to proportions, such as percentages or rates out of 1000, and subtract. Thus one of the trials contained in the meta-analysis diagram (Figure 7.1 section 5) showed that there were eight head injuries among the 131 visited, and 15 among the 132 controls. This converts to injury rates of 6.1 and 11.36 per 100 children respectively. Subtracting yields 2.7. This might be interpreted by saying that home visiting may have reduced the injury rate by 5.26 per 100 children; or by 5.26 per cent.

10.3 Showing effect size using standard deviation

This is sometimes simply called 'effect size'. It is calculated with the formula

$$\frac{\text{(Mean of change shown by treatment group) − (Mean of change shown by alternative treatment/control group)}}{\text{Standard deviation of mean change shown by alternative treatment/control group}}$$

There is no example of this calculation in the studies in this volume, but if you do encounter it, a result of zero means no difference. Assuming that improvement is shown in terms of higher scores, a result of 0.2 would be a small difference in favour of the treatment group, a result of 0.5 a medium sized difference, and one of 0.8 or more, a large difference. Minus figures indicate that the control or alternative treatment group has fared better than the treatment group. The smaller scores might or might not be statistically significant depending on the result of a test of statistical significance.

10.4 Showing effect sizes using risk ratio and odds ratio

Using the figures for home visiting trials again, as above (section 10.2):

Risk ratio
(number to whom event happened in one group divided by total in group) divided by (number to whom event happened in the other group divided by total in group)
(8/131) / (15/132) = 0.061/0.1136 = 0.537 = 0.54 = Risk ratio

Odds ratio
(number to whom event happened in one group divided by number to whom event did not happen in that group) divided by (number to whom event happened in the other group divided by number to whom event did not happen in that group)
(8/123) / (15/117) = 0.065/0.128 = 0.508 = 0.51 = Odds ratio

Neither of these statistics is particularly easy to describe in non-numerical terms, but odds ratios are a particularly common statistic for expressing the results of experimental work. In either case a ratio of 1 would mean that there was no difference between the two groups. A ratio of 1 would put the results on the line of no effect in a diagram summarising a meta-analysis (see Figures 7.1 and 7.2). Where an intervention group is being compared with a control group it is usual, though not universal, to divide the results for the intervention group by the results for the control (as above). Then ratios of less than 1 suggest that something was less likely to happen to the intervention group, and ratios above 1 mean that something was more likely to happen to the intervention group. In the example above there were fewer injuries in the group visited, hence the ratio was less than 1, and this was evidence in favour of home visiting. However, this trial was about preventing something happening. Thus if these figures had come from a trial to see whether a drug cured an illness, a ratio of less than 1 would be evidence against the drug's effectiveness. Where two treatments are being compared with each other, it is important look carefully to see what has been divided into what.

In the leg ulcer bandaging trial (Chapter 1) the authors cite an odds ratio of 1.11, which is short for 1.11 healings for the Charing Cross bandages for every 1 of the trial bandages, or '1 to 1.11'. If you look at Table 7.1 you will not be surprised to see that this is the result of comparing 8 out of 17 with 8 out of 18. The authors also cite confidence intervals of 0.24 to 5.19. Since the limits fall on either side of 1, the 'true' result might be in favour of either bandaging system (see Section 5 for the way this interpretation is made).

10.5 Numbers needed to treat (NNT)

For practitioners, one of the most useful expressions of effect size is the calculation of the numbers needed to treat (NNT) (Chattelier et al., 1996). This is a measure of how many people would, on average need to be treated to produce *one additional* successful outcome by comparison with the alternative treatment. In the trials on home visiting and childhood injury (Chapter 4) this is the question of how many families would need to be visited in order to prevent one injury, since the alternative is simply not visiting at all. One of the trials showed that there were 8 head injuries among the 131 visited, and 15

among the 132 controls. This converts to injury rates of 6.1 and 11.36 per 100 children respectively. But in this case the interest is in *non-injury* rates, which are 93.9 and 88.64 respectively.

The calculation for NNT is:

> 100 divided by (percentage or rate of desired outcome in intervention group) minus (percentage or rate of desired outcome in non-intervention group)
>
> 100/ (93.9 − 88.64) = 19

Thus this particular trial suggests that 19 families (of the kinds featured in the trial) need to be put on the visiting list to prevent one head injury; that visiting 100 families might prevent between 5 and 6 injuries. Or, put another way, visiting one of these families would reduce the chances of a child there experiencing a head injury by 1 in 19, or by 5.26 per cent.

NNT figures should be interpreted with care. If a practitioner had a case mix identical to that featured in the research and was able to do precisely what was done in the research then the NNT figure would provide a close estimate of the number of these clients who would have to be treated in this way to produce an additional benign outcome. However, it is highly likely that any particular practitioner will have a case mix that is different from that which featured in the research, and may not be able to do exactly what was done in the research.

11 Counting costs

NNT calculations convert the results of an experiment into a form where the cost per desirable outcome can be calculated. If the NNT is 95 then the additional cost of getting one additional benign outcome by adopting this intervention will be the cost of intervening 95 times. The exemplar study in Chapter 3 illustrates the way in which the costs of services are calculated, and the further reading for this chapter gives a list of useful sources.

12 Sensitivity analysis

Experiments are always to be regarded as producing estimates as to the true state of affairs. Confidence intervals (sections 4 and 5) provide some purchase on the extent to which experimental results are likely to be misleading because of the play of chance factors. However, experimental research may also provide a misleading basis for practice decision-making because the circumstances under which the research was conducted are unlike those in some practice setting

(Chapter 5, section 12). A *sensitivity analysis* is a 'what if' analysis. The researcher says, 'things were thus and thus in the research, what if they had been different?' Sometimes sensitivity analyses are conducted as a way of managing the problem of subjects lost to an experiment. The researcher will say, 'if there were data for these lost subjects it might be like this, and then the results would be thus, or alternatively it might be like that, and then the results might be like this.' However, sensitivity analyses are most common as accompaniments of economic analyses. Chapter 3 in this volume is an economic analysis, which compares the cost-effectiveness of a hospital at home scheme with inpatient care. The first phase of the study, which is not reprinted in this volume, was a randomised controlled trial showing that both modes of postoperative care were equally effective, leaving the way clear to choose the one with the lowest cost. In fact, the study found that inpatient care was cheaper than hospital at home care. However, costs depend on a great many factors, such that costs in one place may be different from costs in another and change quickly over a period of time (Briggs and Gray, 1999). In their study, Shepperd et al. (1998) carry out a number of sensitivity analyses to investigate this. Table 7.6 gives part of one of these by way of illustration.

The research found that care was delivered for a much longer period to hospital at home patients than to inpatients. This made it more expensive. The researchers were suspicious that this was an 'experiment effect' (see Chapter 5, sections 5 and 12) arising from the staff delivering care at home behaving differently because they were involved in an experiment. If this were so, then the results of the research might be misleading insofar as staff delivering hospital at home care routinely, and not as an experiment, might discharge patients quicker and provide a cheaper service. Table 7.6 gives a number of estimates for the relative costs of hospital at home and inpatient care according to different lengths of time spent in hospital at home care.

Table 7.6 Sensitivity analysis of relative costs of inpatient care and hospital at home (HaH) care varying average length of hospital at home treatment: hysterectomy patients only (see Chapter 3)

Cost per case of hospital at home care, above or below cost of inpatient care	
+£92.40 (HaH is more expensive than inpatient care)	Average time as actually recorded
−£21.75 (HaH would be cheaper than inpatient care)	Estimate if HaH was on average one day shorter
−£80.84 (HaH would be much cheaper than inpatient care)	Estimate if HaH was on average two days shorter

13 Further reading on understanding the results of research

On the way experimental data are analysed and presented

There are many excellent textbooks on the statistics relevant to research in health and social care. Coolican (1994) is recommended for its user-friendliness and as a good primer on research methods as well, and Wright (1997) provides some accessible commentary on the theory and philosophy of statistics in addition to basic information. Most recent textbooks assume readers have access to computer software for doing statistical calculations. The *Statistical Package for the Social Sciences* (SPSS) is most widely used by sociologists and psychologists and very widely used in medical and nursing research as well (Norusis, 1993). Computer packages often turn out to be an easy way of producing results which are incomprehensible to the user. Wright (1994) and Pett (1997) (and many other writers) explain how to interpret the computer print-outs. For a book that is specifically a tutor on how to use SPSS, readers might try the book and disk kit by Babbie and Halley (1994). Pett (1997) is a particularly useful text for dealing with the problems of small samples and unusual distributions and takes its examples from health care.

Two widely used texts for the statistics of medical research are Altman (1991) and Bland (1995).

On costings

Netton and Beecham (1993), Clark and Lapsley (1996) and Yates (1996) all provide details about costing methodologies. Jefferson et al. (1996) is a good primer on economic analysis in health and social care generally.

The Further Reading in Chapter 5 includes references to conducting cost-effectiveness research.

References and further reading

Altman, D. (1991) *Practical Statistics for Medical Research*. London: Chapman & Hall.

Babbie, E. and Halley, H. (1994) *Adventures in Social Research: Data Analysis using SPSS®*. London: Pine Forge Press.

Bland, M. (1995) *An Introduction to Medical Statistics*. Oxford: Oxford University Press.

Briggs, A. and Gray, A. (1999) 'Handling uncertainty in economic evaluations of healthcare interventions', *British Medical Journal*, 319: 635–8.

Chattelier, G., Zapletal, E. and Lemaitre, D. (1996) 'The number needed to treat: a clinically useful nonogram in its proper context', *British Medical Journal*, 312: 426–9.

Clark, C. and Lapsley, I. (eds) (1996) *Planning and Costing Community Care*. London: Jessica Kingsley Publishers.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum.

Coolican, H. (1994) *Research Methods and Statistics in Psychology*, 2nd edn. London: Hodder and Stoughton.

Jefferson, T., Demichelli, V. and Mugford, M. (1996) *Elementary Economic Evaluation in Health Care*. London: BMJ Publishing Group.

Netton, A. and Beecham, J. (1993) *Costing Community Care*. Canterbury: PSSRU University of Kent.

Norusis, M. (1993) *SPSS for Windows Base Systems: Users' Guide*. Chicago: SPSS Inc.

Pett, M. (1997) *Non-Parametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions*. London: Sage.

Shepperd, S., Harwood, D., Jenkinson, C., Gray, A., Vessey, M., and Morgan, P. (1998) 'Randomised controlled trial comparing hospital at home care with inpatient hospital care: I. three month follow up of health outcomes', *British Medical Journal*, 316: 1786–91.

Wright, D. (1997) *Understanding Statistics: an Introduction for the Social Sciences*. London: Sage.

Wright, L. (1994) 'The long and the short of it: the development of the SF-36 General Health Survey', in C. Jenkinson, (ed.), *Measuring Health and Medical Outcomes*. London: UCL Press. pp. 89–109.

Yates, B. (1996) *Analyzing Costs, Procedures, Processes, and Outcomes in Human Services*. Thousand Oaks, CA: Sage.