3 Statistical inference

Constant and random error

In a simple experiment, we are trying to find the effect, if any, that the independent variable has on the dependent variable. As discussed in the last chapter, the dependent variable may be affected by variables other than the independent variable. We concentrated there on variables producing an order effect, but there is a whole host of other variables which might affect the dependent variable. Such variables can be thought of as producing two kinds of errors when we are trying to work out the effect of the independent variable – **random errors** and **constant errors**.

For example, in experiments with animals, food is often used as a reward for food-deprived animals. The method commonly used at one time to ensure that they were appropriately food-deprived was to allow access to food for only a short time, say thirty minutes, per day. However, different animals would take in different amounts of food during this time. If performance in some task were related to food deprivation, then there would be variations in performance which would contribute unpredictable errors, likely to be random in their effect, to any experiment using this method of food deprivation.

A constant error would occur if, for some reason, all the animals in one experimental condition were able to eat for longer than those in another condition. Hence the direction of the error would always be the same and constant in its effect.

Notice that random and constant errors have different effects: **a random error obscures** the experimental effect we are interested in, **a constant error biases** or distorts the results.

Constant errors must go!

Our business in designing an experiment is to hunt down all possible sources of constant error. In some cases it is possible to eliminate them completely by means of *direct control.* To take a fanciful but hopefully clear example: a constant error would be introduced in comparing heights of eskimo and pygmy children if a ruler were used which expanded as the temperature increased (assuming pygmies live in warmer climates than eskimos!). This could be controlled and completely eliminated by using a non-expanding ruler. Less fancifully, if we are carrying out an experiment where there are likely to be gender differences in performance which are not in themselves the focus of our interest then we could use direct control by working solely with females (or solely with males). In other cases one may not be able to control directly, but the biasing effect can be removed either by counterbalancing (as in the case of simple order effects) or, more generally, by randomization. However, neither counterbalancing nor randomization eliminates the error. They merely have the effect of transforming it from constant error to random error. This means that they remove the error as a source of bias, but it still remains to obscure the experimental effect in which we are interested. Whilst it may appear desirable to eliminate all possible constant errors by direct control, there are arguments against this.

Consider what might be called the 'left-handed, fifty-three-year-old introverted Isle of Wight rat-catcher' approach to experimentation. In setting up a particular hypothetical experiment it might appear likely that the handedness of the subjects would be related to their performance. Using direct control we would decide to work either entirely with left-handers, or entirely with right-handers – say the former. Similarly, age could be seen as a possible variable, and using direct control we would opt for a particular age or age range for our subjects. In like manner, personality variables, geographical location and profession might also be seen as having a potential effect in our experiment. Clearly the end result is ludicrous. It is highly unlikely that we would be able to find even a single individual to fit the bill, let alone a viable group to carry out the experiment. And even if by some miracle this were possible, the generality of any results we obtained would be highly questionable.

Put in more general terms it may well be that the effects of an IV on a DV can be demonstrated when everything within sight is held constant. However, it may possibly happen that this effect is dependent on the particular values of one or more of the variables held constant. If we had held them constant at a different value, then the experimental effect might have disappeared.

The main alternative to direct control is randomization, i.e. we allow things to vary but seek to ensure their random allocation to the different conditons of the experiment. If the experimental effect still stands when these variables have been randomized, it indicates that the effect is reasonably robust.

Random errors will not go!

Some random errors can be eliminated as, for example, those caused by the animals eating different amounts during their thirty-minute feeding period. A better method, now widely used, is to feed the animals a carefully measured amount of food which maintains them at a given percentage of their free-feeding body weight. Thus if we can assume that the effect of food deprivation is directly dependent on this percentage, the random error attributable to differences in food deprivation can be completely eliminated.

However, there are many random errors which cannot be eliminated in this way. Consider the many things which might affect a human participant's performance on a memory or a learning task. In order to control these effects one would need a set of participants with identical heredity and environment. Their learning and other experiences before the experiment would have to be equated. They would need to be of the same intelligence and have the same personality and attitudes, to be in the same state of health, etc., etc. The list is endless and it would be impossible to contemplate even starting any experiments if this kind of control were a necessary prerequisite. One is forced to conclude that random error is here to stay and that our methods will have to take this into account.

To this end, the basic strategy is to make sure that the allocation of participants to the different experimental conditions is random so that any potential constant errors end up as random errors.

Statistical inference and probability

Granted, then, that random error will be present, both in its own right and as a result of our having randomized constant errors, how can it be disentangled from the experimental effect that we are after? The answer is that we make use of **statistical inference**.

What we do is:

1 *Estimate how probable it is that the random error by itself could produce the changes in the dependent variable observed in the experiment.*

If

2 *It seems unlikely that random error by itself could produce these changes*

then

3 *We decide that it is the independent variable which is having an effect on the dependent variable.*

You should work through this argument several times. The idea is very important and the process is the reverse of what many people expect. Instead of coming to a decision about the independent variable's effect directly, we approach it indirectly by discounting the likelihood that the effect was produced by random error.

Statistics is used to make inferences about these effects – hence the term 'statistical inference'. Before we can do this you need to have some understanding of the concept of 'probability'.

Probability

The concept of probability is controversial among both statisticians and philosophers. It is used in at least three different ways. In everyday life the reference is usually to how likely or unlikely it is that a future event will occur. Thus we have statements such as 'Huddersfield Town will probably win on Saturday' or 'You'll probably be sick if you eat that third cream cake'. Sometimes this feeling of doubt or uncertainty is expressed in numerical terms as 'I think it's 10 to 1 against him stopping smoking in the New Year' but even so these are subjective estimates and as such this use of the term is commonly referred to as **subjective probability**.

A second use of the term derives from analyses of card games and other games of chance. In cutting a well-shuffled pack of playing cards what is the probability that the card turned over is an ace? This approach defines **probability as the ratio of the number of favourable cases** (here the four aces) **to the total number of equally likely cases** (here the fifty-two cards, assuming a normal pack with no jokers). The probability is then 1 in 13, otherwise expressed as 1/13. This idea, particularly the notion of 'equally likely cases', enables one to work out theoretically the probability of various events occurring in quite complex situations and is of considerable value to casino owners and the like. You should note, however, that this is a formal, theoretical approach to probability sometimes referred to as **mathematical probability**, and the extent to which it corresponds to real life in any situation depends on whether the theoretical assumptions (and particularly the idea of equally likely cases) apply in that actual situation. When playing with dice it seems reasonable to assume that each of the six alternatives resulting from rolling a single die is equally likely. However, such things as loaded dice are not unknown and if over a period we find in practice that, say, a 1 comes up in over half the rolls, then the applicability of the theory in this case is cast into considerable doubt.

This last example illustrates a third approach to probability, the so-called 'relative frequency' approach, otherwise known as **empirical probability**. Here the probability of an event is estimated by the ratio of the number of times the event occurs to the total number of trials which have taken place. It is an estimate because the actual number of trials which have taken place are regarded as a sample from the almost infinitely large population of trials which could theoretically take place. The probability is the state of affairs in this population and will tend to be more and more accurately estimated as we increase the size of the sample. Anyone with a few days to spare might like to test this by tossing a coin repeatedly and noting the proportion of heads obtained after, say, ten trials, a hundred trials, a thousand trials, etc.

These three approaches to probability are not necessarily incompatible. In particular the mathematical and empirical approach often helpfully complement each other. The mathematical approach to coin tossing obviously gives a probability of a half for heads and various lengthy series of coin tosses which have been carried out give estimates, as one might expect subjectively, that are essentially the same (although some types of coin give a very small but consistent 'heads' bias, presumably because of a slight asymmetry in the coin itself).

Probability (given the symbol *p*) is commonly expressed numerically with a minimum value of 0 and a maximum value of 1: 0 refers to something which never occurs and 1 to something which always occurs. In practice, of course, most of the things in which we are interested are somewhere between these extremes and hence have values larger than 0 and smaller than 1.

Consider an experiment. Suppose that in a design where we have pairs of scores (either the matched pairs or repeated measures design) we found that in seven out of eight pairs, scores are larger in condition A than in condition B. And that in only one out of eight pairs the score was higher in condition B than in condition A. (We are going to assume, to make things simpler, that it is not possible to get a tie.) What is the probability that we would have obtained that result on a chance basis, i.e. if only random effects are involved and there is no effect due to the experimental conditions (i.e. no effect of the IV on the DV)?

To simplify the explanation let us refer to the pair where condition A scores are higher than condition B as a '+'; and the pair where they score lower as a '−'. Using mathematical probability we can take the assumption that only random effects are involved as equivalent to the assumption of equally likely outcomes. That is, the probability of getting a + with any pair is $\frac{1}{2}$ (a half), and the probability of getting a − is also $\frac{1}{2}$. If you find it easier, think in terms of the probability of getting 'heads' when you toss a coin.

Suppose we consider two pairs together. There are then three possibilities: two +s (i.e. both pairs A scores larger than B scores); one + (one pair A larger than B, the other B larger than A); and zero +s (both pairs B larger than A). What are the probabilities associated with these? A 'family tree' helps to demonstrate this (Figure 2).

There is a total of *four* possible outcomes. It is reasonable to assume that each of the four are equally likely. They are − first, pair 1 getting + and pair 2 getting +; second, pair 1 getting +

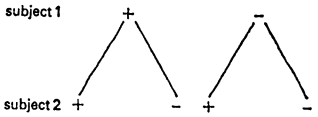

Figure 2 'Family tree' for two participants

and pair 2 getting –; third, pair 1 getting – and pair 2 getting +; fourth, pair 1 getting – and pair 2 getting –. Considering these four possible outcomes, two +s occurs once in four (i.e. $p = \frac{1}{4} =$ 0·25), one + occurs twice in four (i.e. $p = \frac{2}{4} =$ 0·5) and zero +s occurs once in four (i.e. $p = \frac{1}{4} =$ 0·25).

In this way it is possible to work out the probability, on a chance basis, of getting any given number of +s with any total number of pairs. For example, with four pairs, the 'family tree' looks like Figure 3.

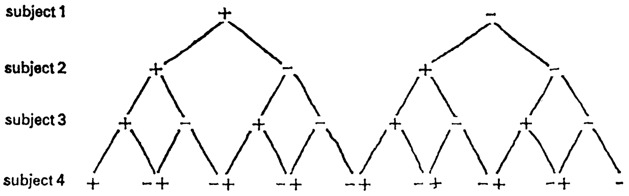

Figure 3 'Family tree' for four participants

There is a total of sixteen possible outcomes and a table of probabilities can be obtained as shown in Table 5.

Table 5 Probabilities for different numbers of +s with four pairs

Number of +s	Probability (= fraction of outcomes)
4	$\frac{1}{16}$ = 0·0625
3	$\frac{4}{16}$ = 0·2500
2	$\frac{6}{16}$ = 0·3750
1	$\frac{4}{16}$ = 0·2500
0	$\frac{1}{16}$ = 0·0625

Returning to our original example of eight pairs, the drawing of the family tree is left to the reader, the total number of possible

Table 6 Probabilities for different numbers of +s with eight pairs

Number of +s	Probability (= fraction of outcomes)
8	$\frac{1}{256}$ – 0·004
7	$\frac{8}{256}$ = 0·031
6	$\frac{28}{256}$ = 0·110
5	$\frac{56}{256}$ = 0·220
4	$\frac{70}{256}$ = 0·270
3	$\frac{56}{256}$ = 0·220
2	$\frac{28}{256}$ = 0·110
1	$\frac{8}{256}$ = 0·031
0	$\frac{1}{256}$ = 0·004

outcomes now being 256. The table of probabilities is given in Table 6. We can see from the table that the probability of obtaining seven +s out of eight is 0·031.

But how does this relate to our analysis of the experiment? This can be shown in graphical form in what is called a histogram or bar chart (discussed in more detail in the next chapter, p. 40). A bar is drawn for each number of +s, the height of the bar representing the probability of that number of +s (Figure 4).

Significance level

Recall that the decision is made that the independent variable has affected the dependent variable when the probability of getting the result obtained, if random errors only are involved, is sufficiently low.

The histogram shown in Figure 4 gives the distribution of the number of +s out of eight when it is pure chance whether or not any particular result ends up as + or −. Look at the extremes of this distribution. The probability of getting no +s at all is 0·004, i.e very low. Similarly the probability of getting all +s is 0·004. We might feel that these probabilities are so low that we are justified in deciding that results as extreme as this (i.e. all of the eight going in the same direction) are not due to random errors alone. If they are not due to the random errors, and we have got rid of the constant

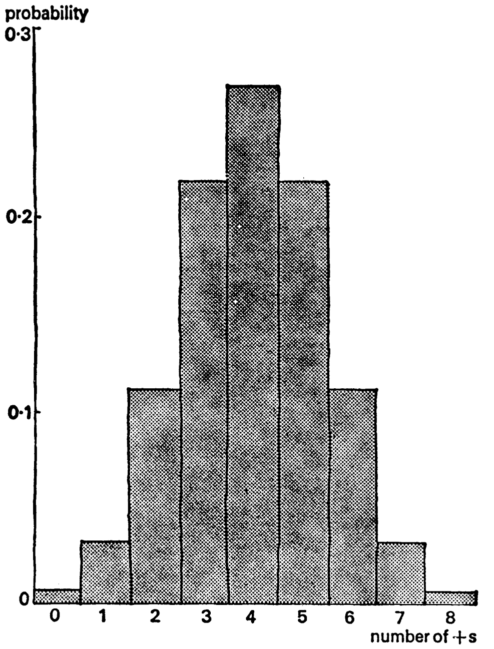

Figure 4 Histogram showing the probabilities for different number of +s with eight participants

errors, then this result must be due to the effect of the independent variable.

How low does the probability have to be for us to make this kind of decision? There is no definitive answer to the question. Whatever level we choose, it is possible to make an error. In fact, there are two possible kinds of error. What is usually called a **type 1 error** occurs when we decide that the independent variable had an effect on the dependent variable when it did not have an effect (i.e. when, in fact, the change in the dependent variable was due to the random effects alone). A **type 2 error** occurs when we conclude that the independent variable had no effect on the dependent variable when, in fact, there was a genuine relationship.

What is normally done is to choose a **significance level**. The significance level is simply the probability of making a type 1 error.

The meaning of significance level is quite often misunderstood, and it will perhaps be useful to talk around this a little more.

Refer back to Figure 4. Suppose that we set a significance level of $p < 0{\cdot}01$ (i.e. probability of making a type 1 error less than 1 in a hundred) then, with a result of 8 +s or of 0 +s, we would come to the decision that the IV had an effect on the DV, i.e. that the result was sufficiently improbable for us to decide that it was not due to random errors alone. Note that this is because the probability of 8 +s and that of 0 +s adds up to 0·008 which is smaller than 0·01.

Suppose that we are willing to take a somewhat higher risk of making a type 1 error, say $p < 0{\cdot}1$. You can see that with results of 8, 7, 1 or 0 +s we would decide that the IV had an effect on the DV. This is because their combined probability adds up to 0·070, which is smaller than 0·1. Notice here that, if you are going to say that with 7 +s the decision is that the IV has an effect, you obviously must also say this for 8 +s, i.e. for any more extreme result. Also that we are deciding this irrespective of the direction of the difference, i.e. that both a large proportion of +s and a small proportion of +s (hence a large proportion of −s) are evidence of an effect.

The lower the probability set for the significance level – and hence the less chance of making a type 1 error – the greater the chance of making a type 2 error. In other words, by limiting your decision that the IV affects the DV to the very extreme cases, you are making it more likely that in some cases you will decide incorrectly that there was no effect. So some kind of balance has to be struck between these two errors.

There is a convention whereby a significance level of probability $p < 0{\cdot}05$ (the 5 per cent significance level) is referred to as **statistically significant** (sometimes simply referred to as **significant**). It must be stressed that this is simply a convention, an agreement between consenting experimentalists. There is nothing magic about the 5 per cent figure. It may be (for example, in exploratory research into an area) that one is worried about type 2 errors, i.e. about regarding as non-significant a relationship which perhaps ought to be followed up, and hence that a significance level of $p < 0{\cdot}1$ might be preferable. On the other hand, there are situations

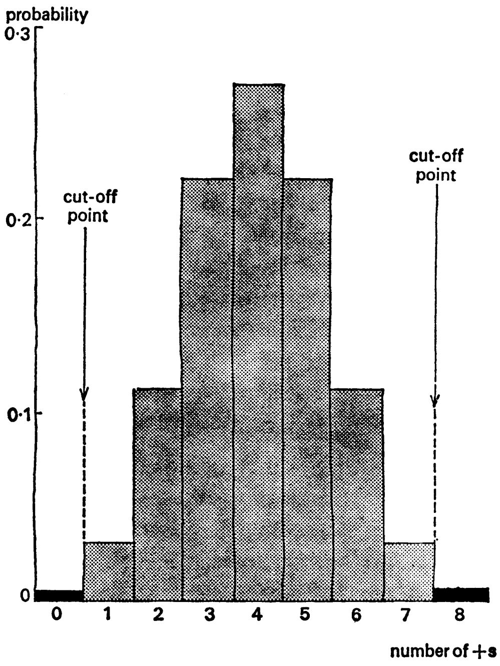

Figure 5 Histogram from Figure 4 with cut-off points added

where the consequences of making a type 1 error might be particularly worrying (owing, perhaps, to one's findings and conclusions being at variance with other published work), and a significance level of $p < 0{\cdot}01$ or smaller might be indicated. Incidentally the $p < 0{\cdot}01$ or 1 per cent significance level is sometimes referred to as **highly significant**.

If we decide, however, to be conventional and to use, say, a 5 per cent ($p < 0{\cdot}05$) significance level, what decision do we come to in our experiment? Referring back to the histogram showing the distribution of +s (Figure 4), the significance level is used to divide the dependent variable into two regions – a region where we will decide that random effects alone are involved and one where we will decide that the independent variable did have an effect on the variable. A glance at Table 6 (p. 32) reveals that, if the cut-off points are marked as shown in Figure 5, the total probability of making a type 1 error is $p = 0{\cdot}008$.

If the cut-off point had been moved in to include the 1 + and 7 + cases, the total probability of making a type 1 error increases to $p = 0{\cdot}070$. This latter value would exceed the significance level, which we had set at $p < 0{\cdot}05$ – i.e. too much of the distribution is being cut off and the cut-off points actually shown on the diagram should be used. As our observed number of +s in the experiment was 7, we see that this lies in the region 'decide IV had no effect on DV' and hence we cannot regard this as evidence for a relationship between independent variable and dependent variable.

Hypotheses and hypothesis testing

You will often find that issues about whether or not the IV affects the DV are referred to in terms of the **null hypothesis** (H_0) and the **alternative hypothesis** (H_1). In this language, the null hypothesis is that the IV does not affect the DV. Various alternative hypotheses are possible, but the most general one would be that the IV does affect the DV (stated in terms of the particular IV and DV in your experiment).

When the experiment and analysis are completed we then

either reject H_0 and accept H_1, if the result is less probable than the chosen significance level;

or accept H_0 and reject H_1 if the result is equal to or more probable than the chosen significance level.

Some statisticians consider it inappropriate to talk about 'accepting' H_0 and prefer 'fail to reject' H_0 instead. This is because in 'accepting' H_0 we don't mean that it is likely that H_0 is true, only that we don't have evidence to reject it.

A warning about the (lack of) significance of statistical significance

There is some tendency for experimenters to worship statistical significance. This is in part because it is much easier to secure publication for a 'significant' finding than for one which is 'non-significant'. In some ways this is very understandable. The discovery of causal relationships between variables is central to much science. Also, poorly performed and controlled experiments are likely to produce non-significant findings.

The difficulty with the concept is that there is a tendency to jump from 'statistical significance' to 'significance' in the sense of 'importance'. All that statistical significance tells you is that what you have found is unlikely to be explicable in terms of random errors. Given a well-designed experiment you can make the leap to saying that it is likely that the IV has had a causal effect on the DV. It says nothing about the size or importance of the effect. In fact, an almost sure-fire way of achieving statistical significance is simply to increase the size of the sample taking part in your experiment. Larger samples provide a more sensitive test of differences between the experimental conditions and there will almost inevitably be some kind of effect which is detectable with a sufficiently large sample. So while it is tempting, and true, to say that your non-significant result in an experiment may point to the need for a larger-scale study, it is actually the significant results from well-designed relatively small-scale studies that are going to pick up the more important 'robust' experimental findings.

The sign test

It is perfectly possible to work out the probabilities of different outcomes for any number of pluses and minuses, through the 'family tree' method. This does, however, become somewhat laborious – particularly for large samples. The **sign test** provides a simple way of reaching the same conclusions. It involves looking up the number of pluses (or the number of minuses; whichever is the smaller) against the total number of pluses and minuses in a table. The table then tells you whether the result you have obtained is statistically significant at the 5 per cent level.

A step-by-step procedure and worked example for doing this are given overleaf on pp. 38 and 39.

Step-by-step procedure

Sign test

Use this test when you have pairs of scores (i.e. matched pairs or repeated measures designs)

Step 1 Give each pair of scores a plus (+) if the score in the left-hand condition exceeds that in the right-hand condition, a minus (−) if the score in the left-hand condition is less than that in the right-hand condition, a zero (0) if there is no difference.

Step 2 Note the number of times (L) the less frequent sign occurs and the total number (T) of pluses and minuses. *Ignore all zeros*, i.e. do not include them in T.

Step 3 Look up in Table B (p. 161) the highest value of L which is significant at the 5% level for this value of T.

Step 4 If your value of L is equal to or lower than the value obtained from the table, the decision is made that the IV had an effect on the DV – the results are referred to as 'significant at the 5% level'. If your value of L is greater than the table value, then the decision is made that the independent variable had no effect on the dependent variable – the results are 'not significant'.

Step 5 Translate the result of the statistical test back in terms of the experiment.

Worked example

Sign test

The following scores were obtained in a matched pairs design with nine pairs of participants.

A	*B*	**Step 1**
12	7	+
10	8	+
15	11	+
8	8	0
7	8	−
10	9	+
8	4	+
7	5	+
13	9	+

Step 2 $L = 1, T = 8$

Step 3 From Table B, highest value of $L = 0$ for significance when $T = 8$.

Step 4 As our value of L is greater than this we decide that the IV has no effect on the DV – i.e. the results are not significant.

Step 5 The difference between condition A and condition B is not significant at the 5% level.

Note (a) The results obtained are identical to those obtained by a direct calculation of probabilities (p. 36).

(b) Although the results are not significant, there is a suggestion that there are higher scores under condition A; this should perhaps be explored in a more extensive experiment (or with a more sensitive test).