6. Can your results be believed? Tests of significance and the analysis of variance

'. . . before anything was known of Lydgate's skill, the judgements on it had naturally been divided, depending on a sense of likelihood, situated perhaps in the pit of the stomach, or in the pineal gland, and differing in its verdicts, but not less valuable as a guide in the total deficit of evidence.'

GEORGE ELIOT
(*Middlemarch*, Chap. 45)

6.1. The interpretation of tests of significance

THIS has already been discussed in Chapter 1. It was pointed out that the function of significance tests is to prevent you from making a fool of yourself, and not to make unpublishable results publishable. Some rather more technical points can now be discussed.

(1) *Aids to judgement*

Tests of significance are only aids to judgement. The responsibility for interpreting the results and making decisions always lies with the experimenter whatever statistical calculations have been done.

The result of a test of significance is always a probability and should always be given as such, along with enough information for the reader to understand what method was used to obtain the result. Terms such as 'significant' and 'very significant' should never be used. If the reader is unlikely to understand the result of a significance test then either explain it fully or omit reference to it altogether.

(2) *Assumptions*

Assumptions about, for example, the distribution of errors, must always be made before a significance test can be done. Sometimes some of the assumptions are tested but usually none of them are (see §§ 4.2 and 11.2). This means that the uncertainty indicated by the test can be taken as only a minimum value (see §§ 1.1 and 7.2). The assumptions of tests involving the Gaussian (normal) distribution are discussed in §§ 11.2 and 12.2. Other assumptions are discussed when the methods are described.

Some tests (*nonparametric* tests), which make fewer assumptions than those based on a specified, for example normal, distribution (*parametric* tests such as the *t* test and analysis of variance), are described in the following sections. Their relative merits are discussed in § 6.2. Note, however, that whatever test is used, it remains true that if the test indicates that there is *no* evidence that, for example, an experimental group differs from a control group then the experimenter cannot reasonably suppose, on the basis of the experiment, that a real difference exists.

(3) *The basis and the results of tests*

No statements of *inverse probability* (see § 1.3) are, or at any rate need be, made as a result of significance tests. The result, P, is always the probability that certain observations would be made given a particular hypothesis, i.e. if that hypothesis were true. It is *not* the probability that a particular hypothesis is true given the observations.

It is often convenient to start from the hypothesis that the effect for which one is looking does not exist.† This is called a *null hypothesis*. For example, if one wanted to compare two means (e.g. the mean response of a group of patients to drug A with the mean response of another group, randomly selected from the same population, to drug B) the variable of interest would be the difference between the two means. The null hypothesis would be that the *true* value of the difference was zero. The amount of scatter that would be expected in the difference between means if the experiment were repeated many times can be predicted from the experimental observations (see § 2.7 for a full discussion of this process), and a distribution constructed with this amount of scatter and with the hypothetical mean value of zero, as illustrated in Fig. 6.1.1. From this it can be predicted what *would* happen *if* the null hypothesis that the true difference is zero were true. In practice it will be necessary to allow for the inexactness of the experimental estimate of error by considering, for example, the distribution of Student's *t*, see §§ 4.4 and 9.4, rather than the distribution of the difference between means itself. If the differences are supposed to have a continuous distribution, as in Fig. 6.1.1, it is clearly not possible to calculate the probability of seeing *exactly* the observed difference (see § 4.1); but it is possible to calculate the probability of seeing a difference equal to or *larger than* the observed value. In the example illustrated this is $P = 0{\cdot}04$ (the vertically shaded area) and

† See p. 93 for a more critical discussion.

this figure is described as the result of a *one-tail significance test*. Its interpretation is discussed in (4) below. It is the figure that would be used to test the null hypothesis against the alternative hypothesis that the *true* difference is positive. When the alternative hypothesis is that true difference is *positive*, the result of a one-tail test for the difference between two means always has the following form.

> *If* there were no difference between the true (population) means *then* the probability of observing, because of random sampling error, a difference between sample means equal to or greater than that observed in the experiment would be P (assuming the assumptions made in carrying out the test to be true).

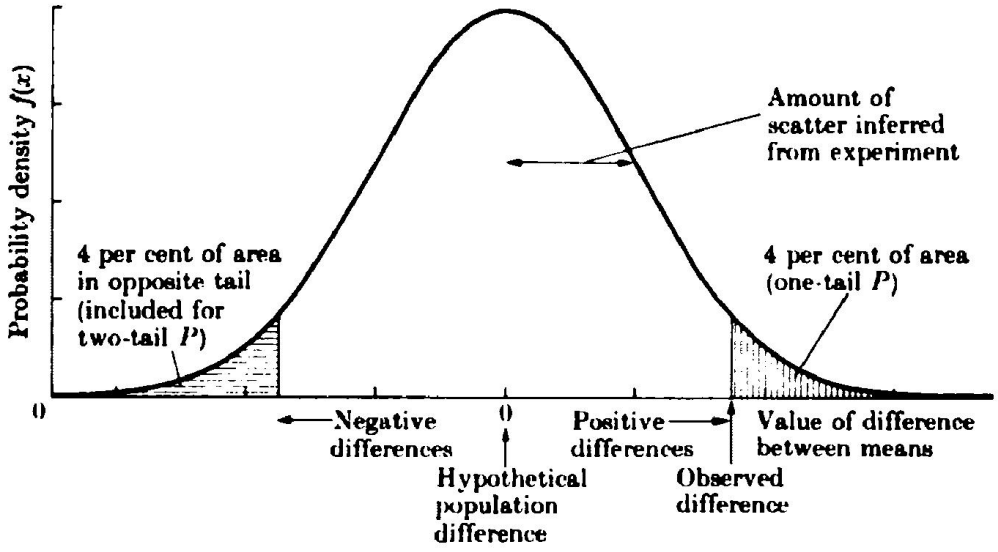

FIG. 6.1.1. Basis of significance tests. See text for explanation.

If the only possible alternative to the null hypothesis is that the true difference is negative, then the interpretation is the same, except that it is the probability (on the null hypothesis) of a difference being equal to or *less than* the observed one that is of interest.

In practice, in research problems at least, the alternative to the null hypothesis is usually *not* that the true difference is positive (or that it is negative) but simply that it differs from zero† (in either direction), because it is usually not reasonable to say in advance that only positive (or negative) differences are possible (or that only positive differences are of interest so the test is not required to detect negative differences).

† See also p. 93.

If the alternative to the null hypothesis is the hypothesis that the true difference between means is, say, positive, this implies that however large a negative difference was *observed* it would be attributed to chance rather than a *true* (population) negative difference (or at least that it would be considered of no interest if real).

Suppose now that it cannot be specified beforehand whether the *true* difference between means is positive, zero, or negative. In the example above there would be probability of 0·04 of seeing a difference at least as large as the positive difference observed in the experiment *if* the null hypothesis were true. But there would also be a probability of 0·04 (the horizontally shaded area) of seeing a deviation from the null hypothesis at least as extreme as that actually observed but in the opposite direction. The total probability of observing a deviation from the null hypothesis (in either direction) at least as extreme as that actually observed would be $P = 0{\cdot}04 + 0{\cdot}04 = 0{\cdot}08$ *if* the null hypothesis were true. This is the appropriate probability because, if it were resolved to reject the null hypothesis as false every time an experiment gave a difference between means as large as, or larger than that observed in this experiment, then, *if* the null hypothesis were actually true it would be rejected (wrongly) not in 4 per cent of repeated experiments, but in 8 per cent. This is because negative observed differences in the lower tail of Fig. 6.1.1, which would also lead to wrong rejection of the null hypothesis, would be just as common, in the long run, as positive differences. The probability is chosen so as to control the frequency of this sort of error. This is discussed in more detail in subsection (6) below.

The value $P = 0{\cdot}08$ is described as the result of a *two-tail test of significance*. Its interpretation is discussed in subsection (4) below. The value of P is usually† twice that for a one-tail test. The result of a two-tail test always has the following form.

> *If* the null hypothesis were actually true *then* the probability of a sample showing a deviation from it, in either direction, as extreme,† or more extreme, than that observed in the experiment would be P (assuming the assumptions made in carrying out the test to be true).

† In the case of the normal distribution (§ 4.2), or any other distribution that is symmetrical, whether continuous or discontinuous, for example the binomial distribution with $\mathscr{P} = 0{\cdot}5$ (§§ 3·2 and 3·4) or Student's distribution, (§ 4.4), one could say here '. . . a deviation from it, in either direction, *as large as, or larger than*, that observed in the

Notice that P is not the probability that the null hypothesis *is* true but the probability that certain observations would be made *if it were*.

Perhaps the best popular interpretation of P is that it is the 'probability of the results occurring by chance'. Although this is inaccurate and vague, and should therefore be avoided, it is not too misleading.

(4) *Interpretation of the results*

If P is very small the conclusion drawn is that either

(a) an unlikely event has taken place, the null hypothesis being true. As Fisher (1951) said: ' . . . no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*,' or

(b) the assumptions on which the test was based were faulty, for example the samples were not drawn randomly, or

(c) the null hypothesis is not true, for example the true (population) means in the above example are different, so that the drugs do in fact differ in their effects on patients (see also subsection (7), below).

Whether (b) can be ruled out, and what level of improbability is enough to make one favour explanation (c) rather than (a), are

---

experiment . . . ' In general, this simpler statement is not possible, however. Two other cases must be considered. (1) The sampling distribution (e.g. Fig. 6.1.1) is continuous but unsymmetrical (see § 4.5). In this case different sized positive and negative deviations will be needed to cut off equal areas in the upper and lower tails (respectively) of the distribution. It is the *extremeness* (i.e. *rarity*) of the deviation measured by the area it cuts off in the tail of the distribution (rather than its *size*) that matters. The two-tail probability is still twice the one-tail probability, however. (2) The sampling distribution is both unsymmetrical and discontinuous (as often happens in the very important sort of tests known as randomization tests, see §§ 8.2, 9.2, 9.3, and 10.2–10.4). A greater difficulty arises in this case because the most extreme observations in the opposite tail of the distribution (that not containing the observation) will not generally cut off an area exactly the same as that cut off by the observation in its own tail so P for the two-tail test cannot be exactly twice that for the one-tail test. There is no definite rule about what to do in this case. Most commonly a deviation is chosen in the opposite direction to that observed that cuts cuts off an area in the opposite tail *not greater than* the value found in the one-tail test, so the two-tail P is not greater than twice the one-tail P. However, it may be decided to choose a deviation that cuts off an area in the opposite tail that is *as near as possible* to that of the one-tail test. This is exemplified at the end of § 8.2 where the deviations of a from the null hypothetical value are stated, to show exactly what has been done. With small unequal samples the most extreme possible observation in the opposite tail may cut off an area far greater than that in the one tail test. This problem is discussed in § 8.2.

entirely matters for personal judgement. The calculations throw no light whatsoever on these problems. It is often found in the biomedical literature that $P = 0{\cdot}05$ is taken as evidence for a 'significant difference'. However 1 in 20 is not a level of odds at which most people would want to stake their reputations as an experimenters and, if there is no other evidence, it would be wiser to demand a much smaller value before choosing explanation (c).

A twofold change in the value of P given by a test should make little difference to the inference made in practice. For example, $P = 0{\cdot}03$ and $P = 0{\cdot}06$ mean much the same sort of thing, although one is below and the other above the conventional 'significance level' of 0·05. They both suggest that the null hypothesis may not be true without being small enough for this conclusion to be reached with any great confidence.

In any case, as mentioned above, no single test is ever enough. To quote Fisher (1951) again: 'In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result'.

(5) *Generalization of the result*

Whatever the interpretation of the statistical calculations it is tempting to generalize the conclusion from the experimental sample to other samples (e.g. to other patients); in fact this is usually the purpose of the experiment. To do this it is necessary to assume that the new samples are drawn randomly from the same population as that from which the experimental samples were drawn. However, because of differences of, for example, time or place this must usually remain an untested assumption which will introduce an unknown amount of bias into the generalization (see §§ 1.1 and 2.3).

(6) *Types of error and the power of tests*

If the null hypothesis is not rejected on the basis of the experimental results (see subsection (7), below) this does *not* mean that it can be accepted. It is only possible to say that the difference between two means is *not demonstrable*, or that a biological assay is *not demonstrably invalid*. The converse, that the means are identical or that the assay is valid, can never be shown. If it could it would always be possible to find that there was, for example, 'no difference between two means' but doing such a bad experiment that even a large real difference was not

apparent. Although this may seem gloomy, it is only common sense. To show that two population means are identical *exactly*, the whole population, usually infinite, is obviously needed.

An example. The supposition that a large P value constitutes evidence in favour of the null hypothesis is, perhaps, one of the most frequent abuses of 'significance' tests. A nice example appears in a paper just received. The essence of it is as follows. Differences between membrane potentials before and after applying three drugs were measured. The mean differences ($\bar{d}$) are shown in Table 6.1.1.

TABLE 6.1.1

d stands for the difference between the membrane potentials (millivolts) in the presence and absence of the specified drug. The mean of n such differences is $\bar{d}$, and the observed standard deviation of d is $s(d)$. The standard deviation of the mean difference is $s(\bar{d}) = s(d)/\sqrt{n}$ and values of Student's t are calculated as in § 10.6.

	$\bar{d}$	$s(d)$	n	$s(\bar{d})$	t	P (approx)
Noradrenaline	2·7	10·1	40	1·60	1·7	0·1
Adrenaline	3·4	12·2	80	1·36	2·5	<0·02
Isoprenaline	3·9	10·8	60	1·39	2·8	<0·01

The potentials were about 90 mV so the percentage change is small, but by doing many (n = 40–80) pairs of measurements, evidence was found against the null hypothesis that adrenaline has no effect, using the paired t test (see § 10.6). Similarly it was inferred that isoprenaline increases membrane potential. These inferences are reasonable, though the order in which treatments were applied was not randomized. In contrast, the P value for noradrenaline was 0·1 and the authors therefore inferred that 'noradrenaline had no effect on membrane potential', i.e. that the null hypothesis was true. This is completely unjustified. The apparent effect of noradrenaline, 2·7 mV, was not much smaller than that for other drugs, and, although the significance test shows that we cannot be sure that repeating the measurements would give a similar result, it certainly does not show that we *would not* get similar results. Suppose, perfectly plausibly, that 80 experiments had been done with noradrenaline (as with adrenaline) instead of 40. And suppose the mean difference was 2·7 mV and the standard deviation of the differences was 10·1. In this case $t = 2{\cdot}7/(10{\cdot}1/\sqrt{80}) = 2{\cdot}4$ giving $P < 0{\cdot}02$ a 'significant' result. The size of the difference $\bar{d} = 2{\cdot}7$ mV, and the scatter of the observations $s(d) = 10{\cdot}1$, is just the same as in

Table 6.1.1, but despite this the authors would presumably have come to the opposite conclusion. This is clearly absurd. But if the original experiment with $n = 40$ differences had been interpreted as 'no evidence for a real effect of noradrenaline' or 'effect, if any, masked by experimental error' there would have been no trouble. It is reasonable that the larger experiment should be capable of detecting differences that escape detection in the smaller experiments.

These ideas can be formalized by considering the *power* of a significance test which is defined as the probability that the test will reject the null hypothesis (e.g. that two population means are equal), this probability being considered as a function of the *true* difference between the means. For example, if the null hypothesis was always rejected whenever a test gave $P \leqslant 0{\cdot}05$ then, if the null hypothesis really were true it would be rejected (*wrongly*) in 5 per cent of trials, as explained in subsection (3) above (see subsection (7), below). The wrong rejection of a correct hypothesis is *called an error of the first kind*, and, in this case, the probability (α) of an error of the first kind would be $\alpha = 0{\cdot}05$. If in fact there was a difference between true population means, and this real difference was, for example, equal in size to the true standard deviation of the difference between means (see §§ 2.7 and 9·4) (i.e. the difference, although real, is similar in size to the experimental errors), then it can be shown that a two-tail normal deviate test† would reject the null hypothesis (this time correctly) in 17 per cent of experiments. However, if the null hypothesis was accepted as true every time it was not rejected then it would be *wrongly* accepted in 83 per cent of experiments. The wrong acceptance of a false hypothesis is called *an error of the second kind*, and, in this case, the probability (β) of this sort of error is $\beta = 0{\cdot}83$.

The power curve for a two-tailed normal deviate test for the difference between two means is shown in Fig. 6.1.2 and compared with the power curve for the (non-existent) ideal test that would always accept true hypotheses and reject false ones. The power of even the best tests to detect real differences that are similar in size to the experimental error is quite small.

(7) *Some more subtle points about significance tests*

The critical reader will, no doubt, have some objections to the arguments presented in this section. It is difficult to give a consensus of informed opinion

† A *t* test (see § 9.4) in which the standard deviation is accurately known (e.g. because the samples are large) so the standard normal deviate, *u* (see § 4.3), can be used in place of *t* (see §4.4).

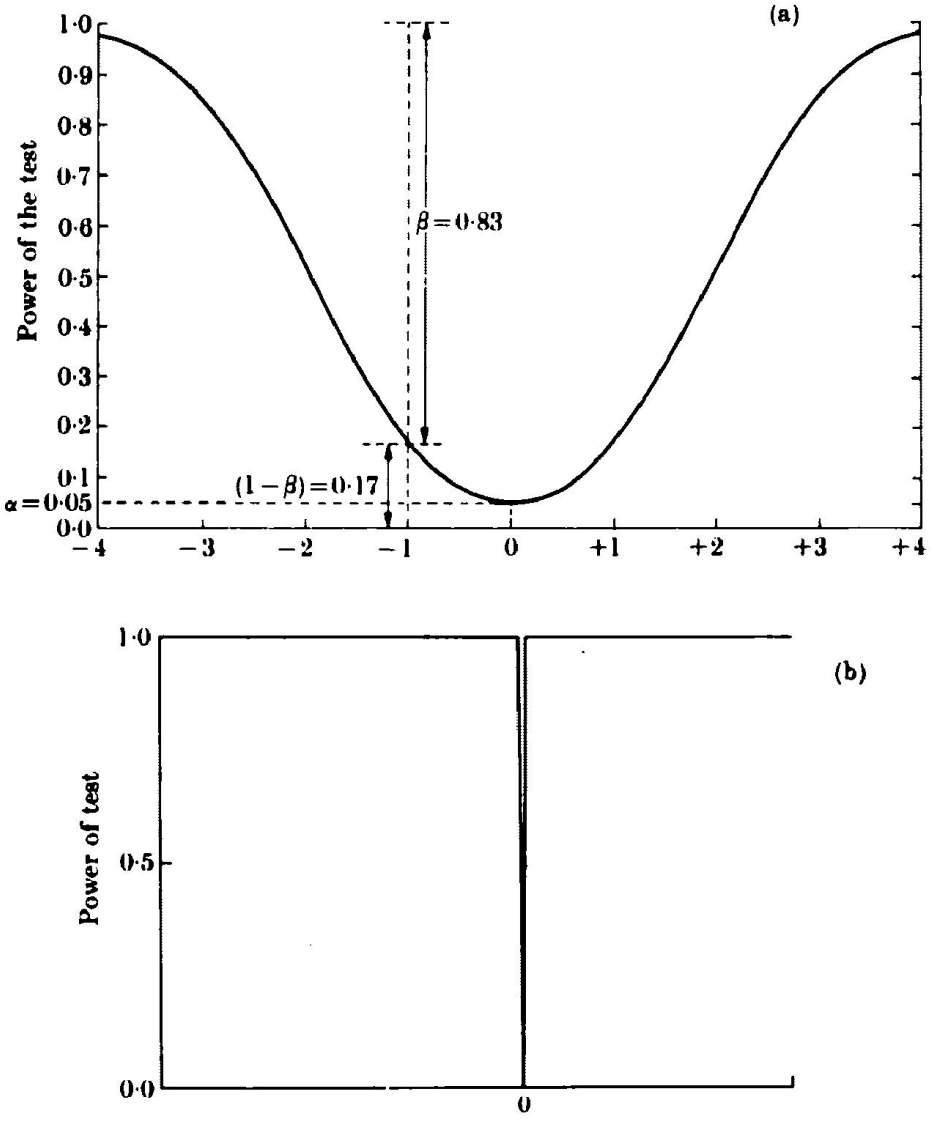

FIG. 6.1.2. In both figures the abscissa gives the difference between the *population* means (expressed as a multiple of the standard deviation of the difference between means: see § 9.4). (a) The power curve for a two-tail normal deviate test for difference between two means (see text) when $\alpha = 0{\cdot}05$, i.e. the null hypothesis is rejected whenever $P < 0{\cdot}05$, so if it were actually true it would be wrongly rejected in 5 per cent of repeated experiments. If the null hypothesis were false, i.e. there *is* a difference between the *population* means (in this example, a difference equal in size to one standard deviation of the difference between means: see § 9.4) the null hypothesis would be rejected (correctly) in 17 per cent of experiments and not rejected (wrongly) in $\beta = 83$ per cent of experiments. (b) Power curve for the (non-existent) ideal test that always rejects a hypothesis (population means equal) when it is false, and never rejects it when it is true.

because, although there is much informed opinion, there is rather little consensus. A personal view follows.

The first point concerns the role of the null hypothesis and the role of prior knowledge, i.e. knowledge available before the experiment was done. It is widely advocated nowadays (particularly by Bayesians, see §§ 1.3 and 2.4) that prior information should be used in making statistical decisions. There is no doubt that this is desirable. All relevant information should be taken into account in the search for truth, and in some fields there are reasonable ways of doing this. But in this book the view is taken that attention must be restricted to the information that can be provided by the experiment itself. This is forced on us because, in the sort of small-scale laboratory or clinical experiment with which we are mostly concerned, no one has yet devised a way that is acceptable to the scientist, as opposed to the mathematician, of putting prior information in a quantitative form.

Now it has been mentioned already that in most real experiments it is unrealistic to suppose that the null hypothesis† could ever be true, that two treatments could be *exactly* equi-effective. So is it reasonable to construct an experiment to test a null hypothesis? The answer is that it is a perfectly reasonable way of approaching our aim of preventing the experimenter from making a fool of himself if, as recommended above, we say only that 'the *experiment* provides evidence against the null hypothesis' (if P is small enough), or that 'the *experiment* does not provide evidence against the null hypothesis' (if P is large enough). The fact that there may be prior evidence, *not* from the experiment, against the null hypothesis does not make it unreasonable to say that the experiment itself provides no evidence against it, in those cases where the observations in the experiment (or more extreme ones) would not have been unusual in the (admittedly improbable) event that the null hypothesis was exactly true.

And, because it has been stressed that if there is no evidence against the null hypothesis it does not imply that the null hypothesis is true, the inference from a large P value does not contradict the prior ideas about the null hypothesis. We may still be convinced on prior grounds that there is a real difference of some sort, but as it is apparently not large enough, relative to the experimental error and method of analysis, to be detected in the experiment, we have no idea of its size *or direction*. So the prior knowledge is of no practical importance.

Another point concerns the discussion of power. It has been recommended that the result of significance test should be given as a value of P. It would be silly to reject the null hypothesis automatically whenever P fell below arbitrary level (0·05 say). Each case must be judged on its merits. So what is the justification for discussing in subsections (3) and (6) above, what would happen 'if the null hypothesis were always rejected when $P \leqslant 0{\cdot}05$'? As usual, the aim is to prevent the experimenter making a fool of himself. Suppose, in a particular case, that a significance test gave $P = 0{\cdot}007$, and the experimenter decided that, all things considered, this should be interpreted as meaning that the experiment provided evidence against the null hypothesis, then it is certainly of interest to the experimenter to known what would be the consequences of acting consistently in this way, in a series of imaginary repetitions of the experiment in question. This does not in any way imply that given a different experiment, under different circumstances, the experimenter should behave in the same way, i.e. use $P = 0{\cdot}007$ as a critical level.

† This remark applies to point hypotheses, i.e. those stating that means, populations, etc., are *identical*. All the null hypotheses used in this book are of this sort.

6.2. Which sort of test should be used, parametric or nonparametric?

Parametric tests, such as the t test and the analysis of variance are those based on an assumed form of distribution, usually the normal distribution, for the population from which the experimental samples are drawn. Nonparametric tests are those that, although they involve some assumptions, do not assume a particular distribution. A discussion of the relative 'advantages' of the tests is ludicrous. If the distribution is known (not assumed, but *known;* see § 4.6 for tests of normality), then use the appropriate parametric test. Otherwise do not. Nevertheless the following observations are relevant.

Characteristics of nonparametric methods

(1) Fewer untested assumptions are needed for nonparametric methods. This is the main advantage, because, as emphasized in § 4.2, there is rarely any substantial evidence that observations follow a normal, or any other, distribution. The assumptions involved in parametric methods are discussed in § 11.2. Nonparametric methods do involve some assumptions (e.g. that two distributions are of the same, but unspecified, form), and these are mentioned in connection with individual methods.

(2) Nonparametric methods can be used for classification (Chapter 8) or rank (Chapters 9–11) measurements. Parametric methods cannot.

(3) Nonparametric methods are usually easier to understand and use.

Characteristics of parametric methods

(1) Parametric methods are available for analysing for more sorts of experimental results. For example there are, at the moment, no widely available nonparametric methods for the more complex sort of analysis of variance or curve fitting problems. This is not relevant when choosing which method to use, because there is only a choice if a nonparametric method *is* available.

(2) Many problems involving the estimation of population parameters from a sample of observations have so far only been dealt with by parametric methods.

(3) It is sometimes listed as an advantage of parametric methods that *if* the assumptions they involve (see § 11.2) are true, they are more powerful (see § 6.1, para. (6)), i.e. more sensitive detectors of real differences, than nonparametric. However, if the assumptions are not true, which is normally not known, the nonparametric methods may

well be more powerful, so this cannot really be considered an advantage. In any case, even when the assumptions of parametric methods *are* fulfilled the nonparametric methods are often only slightly less powerful. In fact the randomization tests described in §§ 9.2 and 10.3 are as powerful as parametric tests even when the assumptions of the latter are true, at least for large samples.

There is a considerable volume of knowledge about the *asymptotic relative efficiencies* of various tests. These results refer to infinite sample sizes and are therefore of no interest to the experimenter. There is less knowledge about the relative efficiencies of tests in small samples. In any case, it is always necessary to specify, among other things, the distribution of the observations before the relative efficiencies of tests can be deduced; and because it is part of the problem that nothing is known about this distribution, even the results for small samples are not of much practical help. Of the alternative tests to be described, each can, for certain sorts of distribution, be more efficient than the others.

There is, however, one rather distressing consequence of lack of knowledge of the distribution of error, which is, of course, not abolished by assuming the distribution known when it is not.

As an example of the problem, consider the comparison of the effects of two treatments, A and B. The experimenter will be very pleased if a large and consistent difference between the effects of A and B is observed, and will feel, reasonably, that not many observations are necessary. But it turns out that with very small samples it is impossible to find evidence against the hypothesis that A and B are equi-effective, however large, and however consistent, the difference observed between their effects, unless something is known about the distributions of the observations. Suppose, for the sake of argument, that the experimenter is prepared to accept $P = 1/20$ (two tail) as small enough to constitute evidence against the hypothesis of equi-effectiveness (see § 6.1). If the experiment is conducted on two independent samples, each sample must contain at least 4 observations (for all the nonparametric tests described in Chapter 9, q.v., the minimum possible two-tail P value with samples of 3 and 4 would be $2.3!4!/7! = 1/17\frac{1}{2}$, however large and consistent the difference between the samples). Similarly, if the observations are paired, at least 6 pairs of observations are needed; with 5 pairs of observations the observations on the nonparametric methods described in Chapter 10, q.v., can *never* give a two-tail P less than $2.(\frac{1}{2})^5 = 1/16$. (See also the discussion in §§ 10.5 and 11.9.)

In contrast, the parametric methods can give a very low P with the smallest samples if the difference between A and B is sufficiently large and consistent. ***Nevertheless, these facts mean that it is a disadvantage not to know the distribution of the observations.*** They do not constitute a disadvantage of nonparametric tests. The problem is less acute with samples larger than the minimum sizes mentioned.

In view of these remarks it may be wondered why parametric tests are used at all when there are nonparametric alternatives. In fact they are still widely used even now. This is partly because of familiarity. The t test and analysis of variance were in use for many years before most nonparametric methods were developed. It probably also results from the sacrifice of relevance to the real world for the sake of mathematical elegance. Methods based on the assumption of a normal distribution have been developed to cover a wide range of problems within a single, admittedly elegant, mathematical framework.

It is not uncommon for those who are dubious about the assumptions necessary for parametric tests to be told something along the lines 'experience has shown that the t test (for example) will not mislead us'. Unfortunately, as Mainland (1963) has pointed out, this is just wishful thinking. There is no knowledge at all of the number of times people have been misled by using the t test when they would not have been misled by a nonparametric test (see §§ 4.2 and 4.6).

A plausible reason for using tests based on the normal distribution is that some of them have been shown to be fairly insensitive to some sorts of deviations from the assumptions on which they are based if the samples are reasonably big. The tests are said to be fairly *robust*. But this knowledge can usually be used only by intuition. One is never sure how large is large enough for the purposes in hand. When the nature and extent of deviations from the assumptions is unknown, the amount of error resulting from assuming them true is also unknown. It is much simpler to avoid as many as possible of the assumptions.

> If a nonparametric test is available it should be used in preference to the parametric test, unless there is *experimental* evidence about the distribution of errors.

In spite of what has just been said parametric methods are discussed in the following chapters, even when nonparametric methods exist. This is necessary as an approach to the more complex experimental designs, curve-fitting problems, and biological assay for which there are

still hardly any nonparametric methods available, so parametric tests or nothing must be used. Whichever test is used, it should be interpreted as suggested in §§ 1.1, 1.2, 6.1, and 7.2, the uncertainty indicated by the test being taken as the minimum uncertainty that it is reasonable to feel.

6.3. Randomization tests

The principle of *randomization tests*, also known as *permutation* tests, is of great importance because these tests are among the most powerful of nonparametric tests (see § 6.1 and 6.2). Moreover, they are easier to understand, at the present level, than almost all other sorts of test and they make very clear the fundamental importance of randomization. Examples are encountered in §§ 8.2, 8.3, 9.2, 9.3, 10.2, 10.3, 10.4, 11.5, 11.7, and 11.9.

6.4. Types of sample and types of measurement

When comparing two groups the groups may be related or independent. For example, to compare drugs A and B two groups could be selected *randomly* (see § 2.3) from the population of patients, and one group given A, the other B. The two samples are independent. Independent samples are discussed in Chapters 8 and 9, and in §§ 11.4, 11.5, and 11.9. On the other hand, the two drugs might both be given, in *random* order, to the same patient, or to a patient randomly selected from a pair of patients who had been matched in some way (e.g. by age, sex, or prognosis). The samples of observations on drug A and drug B are said to be related in this case. This is usually a preferable arrangement if it is possible; but it may not be possible because, for example, the effects of treatments are too long-lasting, or because of ignorance of what characteristics to match. Related samples are discussed in Chapter 10 and in §§ 8.6, 11.6, 11.7, and 11.9.

The method of analysis will also depend on what sort of measurements are made. The three basic types of measurement are (1) classification (the nominal scale), (2) ranking (the ordinal scale), and (3) numerical measurements (the interval and ratio scales). For further details see, for example, Siegel (1956*a*, pp. 21–30). If the best that can be done is *classification* as, for example, improved or not improved, worse or no change or better, passed or failed, above or below median, then the methods of analysis in Chapter 8 are appropriate. If the measurements cannot be interpreted in a quantitative numerical way but can

be *arranged (ranked) in order of magnitude* (as, for example, with arbitrary scores such as those used for subjective measurements of the intensity of pain) then the rank methods described in §§ 9.3, 10.4, 10.5, 11.5, 11.7, and 11.9 should be used. For *quantitative numerical measurements* the methods described in the remaining sections of Chapters 9–11 are appropriate.

Methods for dealing with a single sample are discussed in Chapter 7 and those for more than two samples in Chapter 11.