CHAPTER 8

The Choice of the Number of Observations

8.1 INTRODUCTION

We now turn to a matter more specifically statistical, namely the relation between the number of experimental units and the precision of the estimates of the treatment effects. There are two aspects to this. First the scale of effort that can be devoted to the experiment may be fixed by circumstances outside the experimenter's control. In this case it is nearly always helpful to have, before doing the experiment, some rough estimate of the precision that is likely to result. This rough estimate may, for example, show that the final estimates will probably be subject to such large errors that no effective conclusions are likely to result, thus suggesting that the experiment is not worth doing until more resources can be assembled. Or it may appear that adequate precision can be obtained with less than the full number of experimental units.

The second aspect is more positive. If the number of units is to an important extent under the experimenter's control, we may work out the precision corresponding to a range of values of the number of units. Hence reasonable compromise may be reached between, on the one hand, having too few units and low precision, and on the other, wasting time and experimental material in attaining unnecessary precision.

In the general discussion of § 1.2(ii) we noted that the final precision of the estimated treatment effects depends on

(*a*) the intrinsic variability of the experimental material and the accuracy of the experimental work;

(*b*) the number of experimental units (and the number of repeat observations per experimental unit);

(*c*) the design of the experiment (and on the method of analysis if this is not fully efficient).

In the present chapter we are concerned solely with (*b*) and so we shall

assume that all practicable steps have been taken to increase precision by methods (*a*) and (*c*). We also assume that all important sources of systematic error have been removed, for example by randomization, and that the treatment comparisons are therefore subject only to random errors.

One most important point concerns the definition of an experimental unit for the purposes of the following calculations. Two observations on the same treatment are considered to come from different units only if the design of the experiment is such that the experimental material corresponding to the two observations might have received different treatments, and moreover if the corresponding material has been dealt with independently at all stages at which important variation may enter. For example, imagine that a consignment of material is divided into eight parts, four to receive each of two treatments. Let these eight parts be dealt with separately in the experimental procedure, and at the end let triplicate observations be made on each part, for example by sampling each part three times. There are then twelve observations for each treatment, but only four units. It would be legitimate to regard the twelve observations as twelve units only in the unlikely event that we may assume that negligible variation enters the experiment prior to the last stage, the taking of the observations. There is a further discussion in § 8.3(iv).

The details that follow are inevitably somewhat statistical, and the reader who wishes primarily to learn the general nature and scope of the subject may omit the following sections. He should be aware, however, of the importance of making, before an experiment is started, some estimate of the precision that is likely to be obtained.

8.2 THE MEASUREMENT OF PRECISION

We begin by discussing the way in which precision can be measured. Suppose that we have the situation of § 2.2, so that we are interested in comparisons, or *contrasts*, of the treatment constants $a_1, \ldots, a_t$. Contrasts can take various forms, for example:

(*a*) the difference, $a_1 - a_2$, between the effects of two particular treatments, say the first and the second, T_1 and T_2. This is usually estimated by the mean observation on units receiving treatment T_1 minus the corresponding mean for T_2;

(*b*) the mean difference between one group of treatments and another treatment or group of treatments. For example, in a nutritional experiment, T_1 might represent a basic diet and the remaining treatments various forms of supplemented diet. One contrast of interest might then be the average of the a's for all the supplemented diets minus a_1. The

contrast would usually be estimated by the corresponding difference of the observed treatment means;

(*c*) if the treatments correspond to different levels of one or more quantitative carrier variables, we may be interested in the particular combinations of the *a*'s that measure, for example, the slope and curvature of the response curve. Methods of estimating these contrasts have been considered in Chapter 6.

From the observations we construct an estimate of the particular contrast* that interests us. In general the estimate from the observations will not equal the true value of the contrast calculated from the treatment constants $a_1, \ldots, a_t$, and it is with the magnitude of the difference between the true and estimated values that we are concerned. We call this difference the *error* in the estimated contrast. Of course in any particular instance the true treatment constants $a_1, \ldots, a_t$ are unknown and so is the error in the estimated contrast. We have to work with a probability distribution of errors derived from the fact that the treatment arrangement for use has been selected from a set of possible arrangements in a random way. It can be shown that the average error is zero; this is another way of expressing the elimination of systematic error achieved by randomization. The general size of the errors for a particular contrast is usually best measured by the *standard error*, which is defined formally to be equal to the square root of the average of the squared errors. The interpretation of the standard error given in § 1.2 depends somewhat on the form of the frequency distribution of the uncontrolled variation, but has a more direct practical meaning. This interpretation is as follows. In about one occasion out of three, the randomization will lead to a design in which the error is more than plus 0.97 times standard error or less than minus 0.97 times standard error, i.e., is in absolute magnitude more than about one standard error. In about one occasion out of twenty, the randomization will lead to a design in which the error is in absolute magnitude more than 1.96 times the standard error (for practical purposes 1.96 may be replaced by 2); in about one occasion out of a hundred, the randomization will lead to a design in which the error is in absolute magnitude more than 2.58 times the standard error.

In view of these facts, the standard error is a measure of precision with a direct practical interpretation. The multiple of the standard error corresponding to other frequencies of errors can be obtained from tables of the normal distribution given in books on statistical methods; see also Table 8.1.

* The common mathematical feature of these contrasts is that they are linear combinations of the *a*'s with the sum of the coefficients zero.

It would be expected on general grounds that the standard error of a particular contrast would depend in part on the form of the contrast, in part on the numbers of observations involved, and in part on the amount of uncontrolled variation. This is confirmed by mathematical calculation which shows that, for example, the standard error of any estimate formed by taking the mean of one set of observations minus the mean of a different set, is

$$\sqrt{\left\{\left(\frac{1}{\text{no. of observations in first set}}\right) + \left(\frac{1}{\text{no. of observations in second set}}\right)\right\}} \times \begin{pmatrix}\text{residual}\\ \text{standard}\\ \text{deviation}\end{pmatrix}, \quad (1)$$

where the residual standard deviation is a measure of the amount of that part of the uncontrolled variation which affects the error of the treatment contrasts. More precisely imagine that we could obtain observations from the experimental units in the absence of true treatment effects. We then remove that part of the variation in these observations that can be accounted for by the blocks in a randomized block design or by the rows and columns in a Latin square design, i.e., we take residuals eliminating blocks in the randomized block design and eliminating rows and columns in the Latin square. The residual standard deviation measures the amount of variation in the residuals, in a way analogous to that in which the standard error measures the error in an estimated contrast, i.e., we can think of the standard error and the standard deviation as similar quantities, the first referring to estimated effects, the second to individual observations. As we saw in the discussion of split plot designs in Chapter 7, different contrasts may have different residual standard deviations.

Equation (1) gives the standard error of the difference between two means and the formulas for slopes and curvatures were given in Chapter 6. A fairly nonmathematical discussion for a general contrast* is given by Cochran and Cox (1957, § 3.5). For the particular case when the two sets contain equal numbers of observations, (1) becomes

$$\sqrt{\left\{\left(\frac{2}{\text{no. of observations per set}}\right)\right\}} \times \begin{pmatrix}\text{residual}\\ \text{standard deviation}\end{pmatrix}. \quad (2)$$

It follows that, if we can determine or estimate the standard deviation,

* The general formula, from which all those given here and in Chapter 6 follow as special cases, is that the standard error of $l_1\bar{x}_1 + \ldots + l_k\bar{x}_k$ is $\sigma\sqrt{\{l_1^2/n_1 + \ldots + l_k^2/n_k\}}$, where $\bar{x}_1$ is the mean of n_1 observations, etc., no two $\bar{x}$'s having observations in common, and where σ is the residual standard deviation.

we are able to find the standard error for any particular contrast and hence can

(*a*) determine from the observations limits within which the true value of the contrast lies, at any given level of probability. Thus, the true value lies within the estimated value plus or minus two standard errors, with a probability of 95%;

(*b*) determine, before the experiment is done, the width of the interval of uncertainty at any given level of probability.

In this chapter we are mainly concerned with (*b*), but (*a*) will be briefly illustrated by an example.

Example 8.1. Suppose that in comparing the growth rates of animals receiving two diets, T_1 and T_2, it is known from previous experience of similar experiments that the residual standard deviation is likely to be about 2.5 units. Let ten animals be devoted to each diet. Then the standard error of the estimated difference between the growth rates for the two diets is, by formula (2), $\sqrt{(2/10)} \times 2.5 = 1.12$ units.

If now the observed mean observation on T_1 is 6.10 units more than that on T_2, we can calculate limits for the true difference between the diets as follows: with a chance of 2/3, the true difference lies between $6.10 - 0.97 \times 1.12$ and $6.10 + 0.97 \times 1.12$, i.e., between 5.01 and 7.19; with a chance of 19/20, the true difference lies between $6.10 - 1.96 \times 1.12$ and $6.10 + 1.96 \times 1.12$, i.e., between 3.90 and 8.30. With a chance of 99/100, the true difference lies between $6.10 - 2.58 \times 1.12$ and $6.10 + 2.58 \times 1.12$, i.e., between 3.21 and 8.99.

These statements enable us to form an objective picture of what can be inferred about the true difference from the results of the experiment.*

Quite often we are interested not just in estimating a particular contrast but also in examining its *statistical significance*. This idea needs careful explanation.

Suppose for definiteness that we are interested in the relative effect of two particular treatments T_1 and T_2, i.e., in the true contrast $a_1 - a_2$. Now imagine that using the results of a particular experiment we find that the estimated difference is roughly equal in magnitude to the standard error. Then the frequency interpretation of the standard error tells us that even if there were no real difference between the treatments, a difference as large or larger than the one observed would occur by chance about once in every three times. That is, the difference is just such as would be expected to occur if the true treatment effect were zero. The consequence of this is not that the true difference is asserted to be zero, but that *on the basis of the results under analysis* we would not be justified in claiming that there is a real difference between the treatments, or, in other words, that the data are consistent with a zero true treatment difference.

* The precise interpretation of these probability statements is explained in textbooks on statistics and needs careful qualification.

To put this slightly differently, the data do not, at an interesting level of significance, establish the sign of the true treatment effect. For the positive estimated difference is reasonably consistent with a zero or negative true difference. Significance tests, from this point of view, measure the adequacy of the data to support the qualitative conclusion that there is a true effect in the direction of the apparent difference.

Imagine next that the estimated contrast is just over twice its standard error. An apparent treatment difference as great or greater than this would occur by chance less than one time in twenty and we say that the difference is statistically significant at the 5 per cent (1 in 20) level. Similarly, if the estimated contrast is more than about 2.6 times its standard error, an apparent difference as great or greater than the observed one would occur by chance less than one time in one hundred, and we say that the difference is statistically significant at the 1 per cent level. The level of significance corresponding to other values of the estimated contrast is shown in Table 8.1. Any estimate statistically significant at the 1 per cent level is automatically statistically significant at the 2 per cent, 5 per cent, etc. levels.

The level of statistical significance attained measures the uncertainty involved in taking say an apparent difference between two treatments to be real. For example, if high statistical significance is attained, e.g., the 0.1 per cent level, the statistical uncertainty involved in treating the apparent difference as real is very slight. The following is a general guide to the practical meaning of the various levels:

not statistically significant at 10% level	data are consistent with a zero true contrast.
statistically significant at or near the 5% level but not near the 1% level	data give good evidence that the true contrast is not zero.
statistically significant at or near the 1% level	data give strong evidence that the true contrast is not zero.

Example 8.2 illustrates some of these ideas.

Example 8.2. Consider again the experiment described in Example 8.1. The estimated difference is 6.10 units and the standard error is 1.12 units. The ratio of these is $6.10/1.12 = 5.45$, and this is considerably larger than the largest value in Table 8.1. Therefore the difference is very highly significant statistically, and negligible uncertainty is involved in taking there to be a real difference between the two sets of observations in the direction indicated.

If the estimated difference had been 1.50 units the ratio to the standard error would have been $1.50/1.12 = 1.34$, and from Table 8.1 there is a probability of rather less than 20 per cent of obtaining as great or greater a difference just by chance, the true difference being zero. Since this probability is quite appreciable, we may consider the data as consistent with the absence of a true treatment

difference. The 95 per cent limits of error for the true effect are 1.50 plus and minus 1.96 × 1.12, i.e., (−0.70, 3.70). Note first that this includes negative values, so that we cannot, at this level of probability, infer that the true difference is positive. Note also that it depends entirely on the circumstances of the application whether the data are consistent with the existence of practically important true treatment differences.

As a final example, note that if the estimated difference had been 2.38 units, it would have been statistically significant at the 5 per cent but not at the 2 per cent level. Usually this would be taken as moderately good evidence that the true treatment difference is positive.

TABLE 8.1

LIMITS OF STATISTICAL SIGNIFICANCE FOR AN ESTIMATED CONTRAST WITH THE STANDARD ERROR KNOWN

Ratio of Estimated Contrast to its Standard Error	Level of Statistical Significance
1.28	20%
1.64	10%
1.96	5%
2.33	2%
2.58	1%
2.81	0.5%
3.09	0.2%
3.29	0.1%

If the ratio exceeds the value in the left-hand column, it is statistically significant at the level given in the right-hand column.

Derived, by permission of the Biometrika trust and the authors, from Table 1 of *Biometrika Tables for Statisticians* by E. S. Pearson and H. O. Hartley, Cambridge University Press, 1954.

The following further points should be noted.

(*a*) The significance test is concerned with what the data under analysis tell us. If further data become available, or if we have relevant information about the contrast from general experience or theoretical considerations, our overall conclusions about the contrast may be changed.

(*b*) If the contrast is statistically significant at say the 1 per cent level, we are in little doubt that there is a nonzero true contrast. This is only part of the matter. It will be necessary to consider also the magnitude of the true contrast, not just whether or not it is zero, and we do this by the method indicated at the beginning of this section, that is, by working out the estimated contrast plus and minus appropriate multiples of the standard error, in order to give limits between which the true contrast lies at assigned levels of probability.

(*c*) It can happen that although a contrast is statistically significant,

the limits, at a reasonable probability level, for the true value correspond to differences of no practical importance. That is, we may sometimes conclude that although two treatments differ, the difference between them is of no importance. Statistical significance is not the same as technical importance.

(*d*) On the other hand, if the estimated contrast is consistent with a zero true contrast, it is nevertheless still possible that an important true contrast may exist. For instance in the case discussed above in which the estimated difference is equal to the standard error, and equal, say, to one unit, the limits for the true difference at the 5 per cent probability level are

$$\text{estimated contrast} \pm 1.96 \times \text{standard error},$$

i.e., are very nearly −1 and 3. That is, there is a 95 per cent chance that the true contrast lies between the limits worked out in this way. Now depending entirely on what magnitude of difference we regard as of practical importance, this range from −1 to 3 may or may not include differences of practical concern to us. All that the absence of statistical significance tells us is that we cannot reasonably claim that these data show that T_1 gives a higher observation than T_2, because the range −1 to 3 includes negative as well as positive differences. Further data may, or may not, show that a practically important difference exists, unless we can say from practical knowledge that differences in the range −1 to 3 are of no interest. It follows that we should ordinarily consider the limits of error for a true contrast, even when the estimated difference is not statistically significant. Significance tests fulfil an important, but limited, role in the analysis of data.

This last point (*d*) suggests the need to consider the sensitivity or *power* of a significance test. The whole idea of statistical significance hinges around the desire to protect ourselves against claiming that our data show a treatment contrast in a particular direction, when, in fact, the true contrast is zero (or in the opposite direction). But it is also important to arrange if possible that if the true contrast is sufficiently different from zero to be of practical importance, then the estimated contrast should stand a good chance of being judged statistically significant. This suggests that we should consider the probability that for a given value of the true contrast, the estimated contrast should be statistically significant at some particular level, for example the 5 or 1 per cent levels. This gives what is called the power of the significance test, i.e., it measures the chance of detecting a certain true contrast at a specified level of significance. Power is important in choosing between alternative methods of analysing data and in deciding on an appropriate size of experiment. It is quite irrelevant in the actual analysis of data.

Table 8.2 shows the results of such calculations. We illustrate the meaning of the Table on the same situation that has been used for Examples 8.1 and 8.2.

TABLE 8.2

POWER OF THE SIGNIFICANCE TEST FOR A CONTRAST WHERE STANDARD ERROR IS KNOWN

Magnitude of True Contrast/Standard Error	Probability that Estimated Contrast will be Positive and Statistically Significant at the Following Level		
	10 per cent	5 per cent	1 per cent
0	5 per cent	2½ per cent	½ per cent
0.5	13	7	2
1.0	26	17	6
1.5	44	32	14
2.0	64	52	28
3.0	91	85	66
4.0	99	98	92

Example 8.3. The standard error in the example under discussion is 1.12 units. Therefore the third line of Table 8.2 tells us that if a true difference of this magnitude existed, with say T_1 giving a greater observation than T_2, there is a 26 per cent chance that the estimated difference will show the mean for T_1 to be greater than the mean for T_2, the difference being statistically significant at the 10 per cent level. Similarly, from the next to the last line, if the true difference is $3 \times 1.12 = 3.36$ units, there is a 91 per cent chance that the estimated difference will be statistically significant at the 10 per cent level, etc.

We can express these quantitative statements roughly by saying that if the true difference is equal to one standard error, there is not a high chance that the sample difference will be statistically significant at a useful level, but that if the true difference is equal to three times the standard error, there is a reasonably high chance that statistical significance will be attained.

Suppose now that in the example the number of experimental units for each diet is increased from 10 to 40, the residual standard deviation being unchanged. This halves the standard error of the difference between diets and so the true difference of $3 \times 0.56 = 1.68$ units has the same probability of leading to statistically significant differences as a true difference of 3.36 units had before, and so on.

We can sum up the discussion so far as follows. The standard error of an estimated contrast is a measure of the difference that is likely to arise between the estimate of the contrast and its true value. If we know, before doing the experiment, what the standard error will be, we can predict the resulting width of the interval of uncertainty for the true contrast and can also work out the power of the test of the statistical significance of the estimated contrast. The standard error depends on the form of the contrast, the number of experimental units involved, and

the residual standard deviation, which measures the relevant portion of the uncontrolled variation of the observations.

It has been assumed throughout that the amount of uncontrolled variation is constant. If, for example, some treatments lead to more uncontrolled variation than others, the formulas are changed. The possibility that the standard deviation is different in different sections of the experiment has quite often to be allowed for in complicated statistical analyses, but only affects the planning of the experiment, when we know beforehand roughly what variations in standard deviation will occur. The general effect is that we should take relatively more observations where the variability is expected to be high.

The immediate object of most experiments of the type we are considering is the estimation of the magnitude of certain contrasts among the treatments, and often also the examination of the statistical significance of the resulting estimates. We usually require, therefore, that the standard errors of our estimated contrasts should not be too large. Very occasionally, however, the idea that we require to estimate the magnitude of a particular contrast, say the difference between two treatments, is misleading. We may be interested solely in deciding which of the two treatments gives the higher observation, or in picking out from a number of treatments a small set with particular properties. In such cases we want to end with a simple recommendation that, for example, T_1 gives a higher observation than T_2. Although we want assurance that this is in some sense the proper decision to reach, we do not necessarily ask for a measure of the uncertainty of the final decision or for an estimate of the magnitude of the difference between the treatments. The effect of this on the design of the experiment can be seen from the following example.

Consider an experiment to determine whether a proposed new medical treatment effects a higher proportion of cures than a standard treatment. Suppose that on the basis of the results of the experiment we propose to reach one or other of two possible decisions, namely to use in future either always the standard treatment or always the new treatment. Suppose also that observations become available sequentially in time, as suitable patients present themselves.

Now if one treatment is markedly superior to the other, this may become apparent very soon in the experiment. Then, provided that the evidence is statistically convincing, there are compelling reasons for discontinuing the experiment and using the better treatment on all future patients. On the other hand, if the difference between the treatments is slight, many observations will be usually required to reach a decision. In the first case the estimate of the magnitude of the difference between the treatments may, owing to the small number of observations, be very

imprecise; this is the price that has to be paid for carrying out the experiment with just the choice between the two particular decisions in mind. In the medical application just described, it would often be very reasonable to regard the problem as purely one of reaching a decision between two (or more generally a small number) of alternative courses of action. However experiments that can be profitably regarded in this way are not so common as might be thought. Although many experiments, particularly in technology, are done primarily to determine some course of action, for example to decide which of a number of industrial processes or experimental methods to use, it does seem to be the case that we nearly always need a reasonably precise estimate of the differences involved. There are various reasons for this, such as the following:

(*a*) Decisions are rarely as simple as in the case outlined above, in that they may depend on several types of observation and also on the relative expensiveness of the alternative treatments or processes. The final decision has often to be made by an act of judgement, weighting these different factors in a rather intuitive way. Estimates of the magnitude of the treatment effect for each type of observation are needed for this process to be at all satisfactory.

(*b*) Even in experiments with an immediate practical aim, it is usually advisable to try to reach some understanding of the system under investigation in addition to the decision of immediate concern. For this, quantitative estimation of the magnitudes of treatment effects is usually desirable.

(*c*) It often happens that the results of an experiment are useful in a somewhat unexpected way, for example in helping to settle a question different from that for which the investigation was first set up. If the results are obtained in a form bearing solely on the immediate point at issue, much of this potential usefulness may be lost.

To sum up this discussion, an estimate of the relevant treatment contrasts is nearly always required in experiments designed to add to fundamental knowledge. In experiments intended to decide between alternative courses of action, it is important to consider in designing the experiment exactly what the possible decisions are and how they are related to the observations to be made. The experiment should, within reason, be designed to give just information relevant to the decision, and considerable economy is sometimes achieved by determining the total number of units in the light of the initial results of the experiment.* Usually, however, it will again be necessary to estimate the magnitudes of the relevant contrasts.

* The statistical technique for doing this is called *sequential sampling* and is described briefly in § 8.5.

8.3 THE ESTIMATION OF PRECISION

In the preceding section we saw that the precision with which a contrast is estimated is measured by the standard error. This depends in a known way on the number of observations and the form of the contrast and on the residual standard deviation, which measures the amount of the relevant part of the uncontrolled variation. Therefore the numerical determination of the standard error in any particular case depends largely on finding or estimating the residual standard deviation, and this we now consider.

We have to consider the estimation of the residual standard deviation both in the analysis of the final results and in the preliminary calculations to determine the appropriate number of experimental units. There are essentially five methods of determining the residual standard deviation:

(i) by the observed dispersion, in the experiment itself, of the observations on different units receiving the same treatment. This can be used only in the final analysis and not in preliminary calculations;
(ii) from the magnitude of high-order interactions in factorial experiments;
(iii) by theoretical considerations;
(iv) from within-unit sampling variation. This will be described later;
(v) from past experience of similar experiments.

These methods will be considered in turn.

(i) Use of Observed Variation between Experimental Units

This is the most frequently used method and has already been described in Chapter 3 in connection with randomized block and Latin square designs. An experimental unit was defined to correspond to the smallest subdivision of the experimental material, such that it is possible for different units to receive different treatments. By considering the variation between different units receiving the same treatment, we have a direct measure of the reproducibility of the observations obtained in independent repetitions of the experiment.

The requirements for this method to give a correct estimate of the residual standard deviation are

(*a*) that the different units should respond independently of one another (see § 2.4); and
(*b*) that any source of uncontrolled variation balanced out in the design of the experiment should also be removed before calculating the standard deviation (see § 3.3).

The reader should re-read §§ 3.3, 3.4 for an account of the method by which the estimate of the residual standard deviation can be calculated.

The estimate of standard deviation can now be used to obtain an *estimated standard error* for any particular contrast. Thus for the difference between two treatment means, both based on the same number of observations, the estimated standard error is

$$\sqrt{\left\{\frac{2}{\text{no. of observations per treatment}}\right\}} \times \left(\begin{matrix}\text{estimate of residual} \\ \text{standard deviation}\end{matrix}\right). \qquad (3)$$

This formula is obtained from formula (2) for the true standard error by replacing the standard deviation by our estimate of it.

In the discussion in § 8.2 of the meaning and use of the standard error, it has been assumed that the true standard deviation is known. Thus in Example 8.1, the interpretation of the limits given is correct only if the standard deviation of 2.5 units is not itself subject to random errors. If the standard deviation is only an estimate, it is intuitively clear that the limits for the true contrast, at a given level of probability, must be pushed further apart to allow for the additional uncertainty in the system.

The additional allowance in the uncertainty that has to be made depends on the *degrees of freedom* for the residual which measure, roughly speaking, the number of independent pieces of information available to estimate the standard deviation. There is some further discussion in Example 3.2. The residual degrees of freedom depend to a considerable extent on the total number of units in the experiment and to a lesser extent on the particular design adopted. The most important cases will be put down here for reference:

For a completely randomized experiment in which t treatments are tested on N experimental units, not necessarily with the same number of units for each treatment, the residual degrees of freedom are $(N - t)$.

In a randomized block experiment in which t treatments are tested on N experimental units arranged in k randomized blocks with N/k units in each, the residual degrees of freedom are $N - t - k + 1$. In particular if each block contains each treatment just once, $N = tk$ and the residual degrees of freedom are $(k - 1)(t - 1)$.

In a single $t \times t$ Latin square experiment in which t treatments are compared on t^2 experimental units, the residual degrees of freedom are $(t - 1)(t - 2)$.

In a composite Latin square design with r separate $t \times t$ squares, the residual degrees of freedom are $(t - 1)(rt - r - 1)$.

In a composite Latin square design with r squares each of size $t \times t$ and in which the rows, say, are intermixed, the residual degrees of freedom are $(rt - 2)(t - 1)$.

The derivation of these formulas need not concern us, although the reader who has followed the discussion in Example 3.2 should be able to work them out.

The effect of errors of estimation in the standard deviation is illustrated in Table 8.3.

TABLE 8.3

MULTIPLIERS TO OBTAIN LIMITS OF ERROR AT ASSIGNED LEVELS OF PROBABILITY WHEN THE STANDARD DEVIATION HAS TO BE ESTIMATED

Degrees of Freedom for Residual	Level of Probability		
	90 per cent	95 per cent	99 per cent
5	2.02	2.57	4.03
10	1.81	2.23	3.17
15	1.75	2.13	2.95
20	1.72	2.09	2.85
25	1.71	2.06	2.80
30	1.70	2.04	2.75
Standard deviation known	1.64	1.96	2.58

The precise interpretation of these limits depends on an assumption about the form of the frequency distribution of the uncontrolled variation.

Extracted, by permission of the Biometrika trust and the authors, from Table 12, *Biometrika Tables for Statisticians* by E. S. Pearson and H. O. Hartley, Cambridge University Press, 1954.

Thus, in Example 8.1, with a known standard error of 1.12, and an estimated difference of 6.10 there is a chance of 19/20 that the true difference lies between $6.10 - 1.96 \times 1.12$ and $6.10 + 1.96 \times 1.12$, i.e., between 3.90 and 8.30. If the standard error had been obtained from an estimated standard deviation with 20 degrees of freedom and had happened again to have the value 1.12, the limits would have been $6.10 - 2.09 \times 1.12$ and $6.10 + 2.09 \times 1.12$, that is, 3.76 and 8.44. Similarly with 10 degrees of freedom the limits are 3.60 and 8.60. Although Table 8.3 could be extended down to a single residual degree of freedom, general experience suggests that standard deviations based on less than about five degrees of freedom should not be used for the estimation of standard errors.

In the analysis of experiments, Table 8.3 is used for calculating statistical

significance and finding limits of error. If it is required, during the design stage of the experiment, to estimate the precision to be expected, the following rough rule is useful. The effect on the estimated precision of contrasts of having to estimate the residual standard deviation is approximately to multiply the standard deviation by

$$1 + \frac{1}{(\text{residual degrees of freedom})}. \qquad (4)$$

This rule tends to underestimate the effect when the residual degrees of freedom are small and, as noted above, the degrees of freedom should not, if possible, fall below five.

The increase in error arises from errors in the estimation of the residual standard deviation. If we were solely concerned with obtaining estimated contrasts as close as possible to the true values, and not with estimating the precision of our conclusions, the residual degrees of freedom would be irrelevant. That is, the factor in formula (4) applies to the estimated precision and not to the true precision.

As an example of the use of the rule, suppose that we have 5 treatments and 20 experimental units and wish to choose between a completely randomized experiment and a design in 4 randomized blocks. In the first design the standard deviation is effectively multiplied by $1 + 1/15 = 1.067$, and in the second by $1 + 1/12 = 1.083$; the degrees of freedom 15 and 12 have been obtained from the general formulas given above. The ratio of these factors is 1.015, so that the completely randomized design is the more accurate unless at least a $1\frac{1}{2}$ per cent reduction in the residual standard deviation is attained by blocking. Usually skilful use of the randomized block design would produce an appreciably greater reduction in standard deviation than this. In general it is clear that little information is lost by having to estimate the residual standard deviation provided that the residual degrees of freedom exceed 15 or 20.

In a more complex design, such as a split plot experiment, there are two or more residual standard deviations and these have to be estimated separately, each having its appropriate degrees of freedom.

To sum up, we can estimate the residual standard deviation directly from the results of the experiment, by considering the dispersion of the observations on different units receiving the same treatment. Any portion of the uncontrolled variation whose effect has been eliminated in the design of the experiment must likewise be eliminated before calculating the standard deviation. If we are designing an experiment in which the residual standard deviation is to be estimated in this way, the ultimate precision to be expected is lower than if the standard deviation had been known exactly; an allowance for this can be made. The advantage of

this method of determining precision is that it makes the interpretation of the experiment selfcontained, in that the standard deviation is determined under the actual conditions of the experiment, and is not dependent on any assumption that, for example, the standard deviation is the same as in previous similar experiments.

(ii) Use of High-Order Interactions in a Factorial Experiment

This has been discussed in § 6.11. In a complicated factorial experiment we may attain sufficient precision from one replicate, and in this case all experimental units receive different treatments, so that method (i) is inapplicable. We can however estimate the residual standard deviation, if it can be assumed that the true values of certain high-order interactions are negligible. Once this assumption has been made, an estimate can be obtained with a certain number of degrees of freedom, and the discussion in (i) applies. If the assumption about the high-order interactions is false, the true residual standard deviation will be overestimated.

(iii) From Theoretical Considerations

It is sometimes possible to calculate theoretically what the residual standard deviation should be under idealized conditions. Such a calculation is useful

(*a*) to estimate the residual standard deviation in the analysis of small experiments, in which very few degrees of freedom are available for residual;

(*b*) to provide, in the planning of the experiment, an estimate of the precision that is likely to be attained;

(*c*) to use in the interpretation of an estimate of standard deviation obtained by methods (i) or (ii). It is often instructive to compare the observed standard deviation with a theoretical value. If the theoretical value is too small there are important sources of variation present not accounted for in the theoretical analysis.

The calculation of theoretical standard deviations is, of course, a matter of statistical technique and will not be gone into in detail here. The following are the most important cases.

First the observation on each experimental unit may be that out of, say, N individuals, r have a certain property and the remaining $N - r$ do not. For example on each plot of an agricultural field trial, we may examine 100 randomly selected plants and count the number diseased. In this case N is 100 and r is the number with the disease actually counted on the plot. Suppose that the only source of uncontrolled variation

present arises from sampling the plot rather than counting all plants, i.e., in general, arises from examining N randomly selected individuals rather than an indefinitely large number. Then it can be shown mathematically that the residual standard deviation of the observed proportion with the property is equal to

$$\sqrt{\left\{\frac{\text{true proportion with property} \times \text{true proportion without property}}{N}\right\}}. \quad (5)$$

For instance, if the true proportion of diseased plants was 0.3 in the above example, the standard deviation would be $\sqrt{(0.3 \times 0.7/100)} = 0.046$.

The second case is when the observation made on each experimental unit is the rate of occurrence of a randomly occurring event. Examples are counts of radioactive particles, or accidents, or breakdowns of a machine, or mutations of genetic material. In all these cases we have events occurring in a haphazard way in time, and the observation on each experimental unit is that a certain number, n, of events occur in a period of observation T. The rate of occurrence is thus n/T. Suppose that the only source of uncontrolled variation arises from having observed each unit only for a time T, rather than for a much longer time,* and that on each unit events occur completely randomly, the occurrence of one event being entirely independent of the occurrence of all other events. Then it can be shown mathematically that the residual standard deviation of the rate of occurrence is equal to

$$\sqrt{\left\{\frac{\text{true rate of occurrence}}{T}\right\}} = \frac{\sqrt{\left(\begin{matrix}\text{expected number of}\\ \text{events observed}\end{matrix}\right)}}{\text{period of observation, } T}. \quad (6)$$

For example, suppose that each experimental unit is a batch of wool and that the observation made on each unit is the end breakage rate in spinning. If the true end breakage rate is expected to be about 10 per 1000 spindle hours, and the period of observation is 3000 spindle hours, the standard deviation will, under the above assumptions, be $\sqrt{(10/3)} = 1.83$ if the unit of time is taken to be 1000 spindle hours. Similarly with a period of observation of 1000 spindle hours per experimental unit, the residual standard deviation would be $\sqrt{10} = 3.16$. Other applications of this formula are to counting problems in bacteriology and serology.

A third situation is when the observation for each experimental unit is a measure of dispersion. For example the treatments may be different experimental methods and one object of the experiment may be to compare

* That is, we assume that all units receiving the same treatment would give effectively the same rate of occurrence, if observed for a sufficiently long time.

the reproducibilities of the different methods. For one batch of material several observations are made by a particular method, and the dispersion of these observations measured, for example by the standard deviation. We now have for each experimental unit a standard deviation and we treat these as the "observations" for analysis. These "observations" have a residual standard deviation, i.e., we have to consider the standard deviation of a standard deviation. It can be shown that if the frequency distribution of the readings on any one unit approximates to a special mathematical form, called the normal or Gaussian distribution, the residual standard deviation is approximately equal to

$$\frac{\text{true value of the standard deviation}}{\sqrt{(2 \times \text{number of readings per determination of st. dev.})}}. \quad (7)$$

These are not the only cases where a theoretical calculation of the residual standard deviation can be made; whenever the observation under analysis can be considered as originating from a probability model, a theoretical calculation of standard deviation may be possible. The disadvantage of using the theoretical standard deviation is that the theoretical model assumed, for example the sampling of a completely random series of events with no other sources of variability, may be quite inaccurate as a representation of the uncontrolled variation actually occurring. If it can be obtained, an estimate of the residual standard deviation calculated from the observed variation between units is to be preferred for the direct assessment of the precision of treatment effects. In analyzing data on proportions, counts, and variabilities it is frequently desirable to work with mathematically transformed values. The theoretical standard deviation is different after transformation but can always be found.

(iv) From Within-Unit Sampling Variation

It frequently happens that the observation of main interest for any one unit is the mean of independent readings obtained from randomly selected portions of the unit. Some examples should make this clear.

In an agricultural field trial, it may be required to analyze the total yield of product per plot. If each plot is large, the yield may be estimated by selecting a number of small areas within each plot and weighing the product only from these. From the total yield of the sampled areas, we can estimate the yield of the whole plot.

In many types of industrial experiment, we are interested, among other things, in comparing the mean strengths of articles produced by alternative processes. Each experimental unit consists of a batch of articles, processed at one time by one method, and the mean strength

will usually be estimated by testing a relatively small number of articles, randomly selected from the batch. For example in a textile experiment each batch of yarn, i.e., each experimental unit, would probably consist of many miles of yarn, the mean strength being estimated from tests on say 100 1-ft lengths randomly selected from the batch.

More generally, it is a common characteristic of methods of chemical and biological analysis that duplicate or triplicate independent determinations are made on samples of material from the same experimental unit, the final observation for the unit being the average of the separate determinations. In complex methods of chemical analysis there may be several stages of sampling, corresponding to the different stages of the analysis.

In this type of situation we can distinguish several components of uncontrolled variation, as shown by the following example.

Example 8.4. Consider as a typical example a simplified form of an experiment for comparing two methods S_1 and S_2 of spinning wool yarn. There will be several experimental units, each being a batch of raw material from which yarn is to be spun. Suppose that the batches are processed in random order, half by process S_1 and half by process S_2. Finally from each batch a number of lengths are randomly selected and tested for strength. This gives us a collection of observations of the following general type:

Unit 1	Unit 2	Unit 3	...
S_1	S_2	S_2	...
—	—	—	
—	—	—	
—	—	—	
.	.	.	
.	.	.	
.	.	.	

Quite generally think of a situation in which the primary observation on each unit is the average of several readings.

Now it would in principle be possible to make a large number of observations on each experimental unit. From the variation of the observations on one unit, we could then obtain a measure of the within-unit variability. If we measure variation by the standard deviation we thus obtain the *within-unit standard deviation*. In the present example this is a measure of the variation of strength within a batch of yarn spun in one lot; it takes no account of variation in mean strength from batch to batch. Next if we had the true mean strength for each experimental unit, we could define the *between-unit standard deviation* to measure the uncontrolled variation between units receiving the same treatment in the true mean strength for each unit. This would measure the effect of variations between batches of raw material and of nonconstancy in the conditions of processing. Notice that the between-unit standard deviation is unaffected by variations of strength within a batch, since it refers to the mean of a very large number of observations per batch.

In a practical situation we usually have only a small or moderate number of observations on each unit, and it can be seen that the comparison of the mean

strength for the two processes is then subject to errors arising from both sources. The effective residual standard deviation for the comparison of mean strengths can be shown to be

$$\sqrt{\left\{(\text{between-unit st. dev.})^2 + \frac{(\text{within-unit st. dev.})^2}{\text{no. of obs. per unit}}\right\}}.$$

There is a discussion of this formula with numerical applications in Example 8.9; for the present, note that if we can estimate both component standard deviations, we can predict the residual standard deviation corresponding to any number of observations per experimental unit.

The two components of standard deviation can be estimated from the results of an experiment in which there are at least two observations per experimental unit; the analysis of variance for doing this is described in textbooks on statistical methods (Goulden, 1952, p. 67). The estimation of the separate components is, however, only necessary in order either to examine the nature of the uncontrolled variation or to predict what the standard deviation would have been with a different number of observations per unit. If all that is required is to estimate the precision of the process comparisons in the experiment as performed, it is enough to analyze the mean strengths per batch as if they were single observations.

Now consider the type of experiment, which is sometimes done, in which there is only one experimental unit for each treatment. That is, one batch of material is processed by S_1 and one by S_2, and several measurements of strength are made for each batch. Clearly no fully satisfactory estimate of precision can be obtained from such an experiment, because there is no way of estimating the between-unit standard deviation. The most that can be shown is that the two units have different mean strengths; whether or not this difference is due to the processes or is just random between-unit variation cannot be determined from the observations themselves. An essential condition for a selfcontained analysis of the observations and for the correct estimation of precision is that for each treatment there should be several experimental units, run independently. Nevertheless, in cases where it is impracticable to have more than one, or a small number, of experimental units for each treatment, and in which prior knowledge suggests that the between-unit component of variation is relatively unimportant, the estimation of precision from the within-unit standard deviation is permissible, i.e., we in effect assume that the between-unit standard deviation is zero. This is not a good procedure, however, and should be avoided wherever possible, by running enough independent experimental units for each treatment to provide a satisfactory estimate of the residual standard deviation by method (i).

In more complicated cases, with several stages of sampling, there will be several components of standard deviation but the general principles involved remain the same.

The use of within-unit sampling variation to measure the precision of treatment contrasts is therefore in general undesirable. However, in experiments with a very small number of units, so that no effective estimate of the correct residual standard deviation can be made, the within-unit may, if used with caution, be useful in giving the minimum error to which the treatment contrasts are subject. More generally the

magnitudes of the within-unit standard deviation and the between-unit standard deviation give information about the importance of the different sources of uncontrolled variation, and also determine, for future experiments, what is a suitable value for the number of readings per unit.

Notice that the use of within-unit variation is analogous to that of certain theoretical values for the standard deviation, in that both are obtained assuming that some sources of variability are negligible.

(v) From the Results of Previous Similar Experiments

The last method of estimating the residual standard deviation is from the statistical analysis of the results of previous similar experiments. Particularly in routine laboratory work, large bodies of previous data may be available for such an analysis from which an estimate based on a large number of degrees of freedom may be obtained.

Such an estimate is particularly useful in the determination of an appropriate size for an experiment being designed. It is also useful in the analysis of the results of an experiment

(*a*) to estimate the residual standard deviation when few degrees of freedom for residual are available in the experiment;

(*b*) to compare with a residual standard deviation obtained from the experiment itself. It is frequently a good check on the experimental work to see how the standard deviation compares with that in previous similar experiments.

If a reasonably accurate estimate of the residual standard deviation can be obtained from the experiment itself, we would normally use this for the calculation of standard errors rather than the estimate from prior work, even though the latter is nominally more accurate. We thereby avoid the assumption that the amount of uncontrolled variation is the same as in previous work, make the interpretation of the experiment more selfcontained, and, other things being equal, the conclusions more cogent. A possible exception to the use of the observed residual standard deviation is when it is appreciably less than the value from prior work, and yet it is fairly certain from knowledge of the system that no real increase in precision can have occurred.

(vi) Summing Up

We have seen that methods (i), (ii), and (iv) of estimating the standard deviation are applicable only in the analysis of the observations, not in the design of the experiment. In order to obtain an estimate of precision prior to the performance of the experiment we must use methods (iii) or (v), theoretical calculation or the analysis of the results of previous similar

experiments. Occasionally, as for example when the experiment is the first of its type, neither of these methods can be used; in such a case the size of the experiment must either be settled by general judgement or, alternatively, if a prior calculation of precision is very desirable, the experiment must be done in two or more stages, the observations from the first stage being used to determine the appropriate size of the second stage. This technique is discussed briefly in § 8.5.

8.4 SOME STANDARD FORMULAS

We can now give some formulas for deciding on an appropriate number of observations to take. First determine the residual standard deviation as well as possible, by one or another of the methods of the previous section.

If we have a set of treatments, and the comparisons of all pairs are of equal importance, we devote an equal number of units to each treatment.* Then the standard error of the estimate of the difference between any two treatments is

$$\sqrt{\left\{\frac{2}{\text{no. of units per treatment}}\right\}} \times \text{residual standard deviation}$$

and therefore the number of units per treatment leading to a preassigned standard error is

$$2 \times \left\{\frac{\text{residual standard deviation}}{\text{required standard error}}\right\}^2. \qquad (8)$$

If we can now decide what standard error we require, either by considering the width of the limits of error for the true contrast, or by considering the power of the associated significance test, the appropriate number of units per treatment is determined. Similar calculations may be made for contrasts other than simple differences between pairs of treatments.

Example 8.5. In a certain type of agricultural field trial it may be known that the residual standard deviation is about 10 per cent of the mean yield. Suppose that we require to make the limits of uncertainty for a true difference at the 95 per cent level of probability extend 5 per cent on each side of the estimated difference. This implies a standard error of $2\frac{1}{2}$ per cent, and hence from (8) the appropriate number of plots per treatment is $2 \times (10/2\frac{1}{2})^2 = 32$.

With even a moderate number of treatments this represents a large experiment, and it might well be decided that our requirements on precision have to be

* An exception to this would be if it were expected that observations on different treatments would have different amounts of uncontrolled variation. This possibility is noted briefly in § 8.2. It would then be reasonable to take more observations on those treatments for which the variability is expected to be high.

weakened. The next step would then be to work out the extent of the interval of uncertainty at the 95 per cent level for various sizes of experiment. We get

Number of Plots per Treatment	Approximate 95 Per Cent Limits, plus and minus
32	5 per cent
25	5.7 per cent
16	7.1 per cent
9	9.4 per cent

It is now usually a matter of intuitive judgement to decide what to do. The additional expense of an increase in the size of the experiment has to be balanced against the resulting gain in precision in the conclusions. We shall give an example below in which these considerations can be weighed quantitatively, but this is rather unusual.

Alternatively it may seem better to think in terms of the power of the significance test of the difference between two treatments. For example suppose that a true difference between two treatments of 10 per cent is considered of appreciable practical importance. Then it will be desirable that if a true difference of this magnitude exists there should be a good chance that a reasonable degree of statistical significance should be attained by the observed values. Now Table 8.2 shows that if the true difference is three times the standard error, there is 91 per cent chance of attaining statistical significance at the 10 per cent level, an 85 per cent chance of attaining statistical significance at the 5 per cent level, and so on. If the ratio to the standard error is much less than three the chance of attaining significance is appreciably reduced. Therefore it would be reasonable to arrange that the true difference, 10 per cent, is three times the standard error, i.e., to arrange that the standard error is 10/3 per cent. If we substitute this value into formula (8), we get for the number of plots per treatment $2 \times (10 \times 3/10)^2 = 18$. Again, if this calculation makes the total number of plots in the experiment intolerably large, the effect of weakening the requirements can be investigated.

If it seems that an experiment with a small number of experimental units will be adequate, the residual degrees of freedom in the design will be small, and this, as we have seen, increases the effective standard deviation. Usually, however, the allowance for this is relatively small compared with the general uncertainty involved in the whole calculation of the appropriate number of units. An exception is when the size of the experiment as determined from the first calculation would leave five or fewer degrees of freedom for residual; it would then be impracticable to estimate the standard deviation from the observed dispersion of the observations, and it may be desirable to increase the number of units solely in order to get enough degrees of freedom for residual. This is especially the case when methods of estimating the standard deviation other than from the observed dispersion of the observations are unreliable.

It must be stressed, however, that the condition that there should be enough degrees of freedom for residual is not to be used as a general criterion for determining the size of experiments. The main consideration

is the standard error of the contrasts, with the degrees of freedom for residual a subsidiary matter.

Similar methods apply when the residual standard deviation is calculated theoretically, or when the contrast of interest is not a simple difference.

Example 8.6. Suppose that in an investigation in nuclear physics it is desired to examine the frequency with which certain conditions which can be set up experimentally lead to specified types of transition. Imagine that the mechanism of the transition is unknown and that to find out something about it, the effect on the frequency of transitions of various modifications to the experimental conditions is to be investigated.

Suppose that initially the transition occurs in about 20 per cent of occasions and that it is thought desirable that if a certain treatment increases this true proportion to 30 per cent, there should be a good chance of attaining statistical significance. Let each experimental unit consist of n trials, the proportion of these leading to transitions being observed. If we calculate on the basis of a 25 per cent transition rate, the residual standard deviation is, from (5), $\sqrt{(0.25 \times 0.75/n)} = \sqrt{(0.1875/n)}$. If r units are tested with each treatment, the standard error of the estimated difference between the treatments will be $\sqrt{(2/r)} \times$ standard deviation, which equals $\sqrt{[0.375/(rn)]}$. Arguments similar to those used for the previous example suggest arranging that the true difference of interest 30–20 per cent, i.e., 0.1, is three times the standard error. This gives the equation

$$3 \times \sqrt{\left\{\frac{0.375}{rn}\right\}} = 0.1, \qquad \text{that is } rn \simeq 340.$$

Thus we need about 340 trials for each treatment. Tables and a nomogram for the appropriate number of units, calculated by a more refined method, are available (Eisenhart et al., 1947, p. 247).

Now the calculation has given just the total number of trials that should be carried out for each treatment; from the point of view of the calculation it makes no difference whether we have for each treatment one unit with 340 trials, two units with 170 each, and so on. This is because formula (5) for the standard deviation is based on the assumption that there are no sources of uncontrolled variation to make two trials on different units receiving the same treatment any less alike than two trials on the same unit. In practice this would be at best a good approximation and it would be preferable to have as many different units as practicable, in order to attain the best possible sampling of other sources of uncontrolled variation that may be present. In the present example a good arrangement would possibly be to have, for each treatment, seven experimental units, each unit consisting of, say, 50 trials made as far as possible under identical conditions.

Example 8.7. Suppose that we are particularly interested in the slope of the response curve for a certain quantitative factor and that the factor is investigated at three equally spaced levels, with equal numbers of units at each level. Table 7.1 shows that the standard error of the slope is

$$1.225 \times \frac{\text{residual standard deviation}}{\sqrt{(\text{total number of experimental units at the three levels})}}.$$

If we can estimate the residual standard deviation and decide on the standard error that we require, we can determine the total number of experimental units as before.

In determining the number of experimental units, we are compromising between, on the one hand, having high precision and an expensive experiment and, on the other, having an economical experiment giving low precision. Usually this compromise has to be reached in a somewhat intuitive way, but if it is possible to measure in the same units, for example of money, the cost of the experiment and the loss caused when the conclusions are inaccurate, it will be possible to calculate explicitly the appropriate number of units. Yates (1952) has provided an interesting discussion of this.

Example 8.8. Yates's example is of the determination of the optimum dressing of nitrogen for sugar beet. The optimum dressing will be such that the cost of a small additional dressing just equals the value of the additional yield produced; the determination of this optimum will be subject to random errors of experimentation and it is possible to express the average "loss" arising from the use of an incorrect dressing in terms of the square of the standard error of the estimated optimum, and of the total area of crop to which the conclusions are to be applied, etc. The standard error will depend in part on the size, and hence on the cost, of the experiment, and if the cost of an experiment of given size is known, the average "loss" from an incorrect recommendation plus the cost of experimenting can be minimized, and so the most economical size of experiment calculated.

To carry through the calculation it is necessary to know approximately the cost of experimenting, the loss per unit of experimental material due to a given departure from optimum conditions, the residual standard deviation, and the quantity of material to which the conclusions are to be applied.

There are several things that may complicate such a calculation. It may be advisable to determine optimum treatments separately for various portions of the experimental material. Again, it often happens that the conditions under which the experimental work is done are not fully representative of conditions under which the results are to be applied, i.e., there may be a bias. Effort spent in removing this bias rather than in increasing the size of the experiment is often worth-while.

The main discussion at the beginning of this section has been of the case where the comparisons of all pairs of treatments are of equal importance, so that we arrange that each treatment occurs the same number of times. It may happen, however, that some comparisons are of more interest than others. There are two main possibilities.

We may have one control treatment and a number, m, of alternative treatments. Sometimes the most interesting thing is to compare the alternative treatments individually with the control, comparisons of the alternative treatments among themselves being of secondary importance.

It can be shown that if all observations have the same precision the best procedure is to arrange that for each unit receiving a particular alternative treatment there are approximately $\sqrt{m}$ units receiving the control treatment. A second case is when the main interest is in the difference between the control and the average of the other treatments. Then we should have m observations on the control for each observation on a particular alternative treatment.

For example, in a nutritional experiment we may compare a control diet deficient in a certain constituent with, say, three other diets, all containing substantial amounts of the constituent, but differing in the form in which it is presented. Since the nearest whole number to $\sqrt{3}$ is 2, the recommended arrangement is to have two observations on the control to one on each of the other three treatments, whenever the main interest is in comparing an individual supplemented diet with the control. If the main interest is in comparing the average of the three supplemented diets with the control, the recommendation would be three observations on the control to one on each of the other three treatments. The experiment could be set out in randomized blocks with five units per block in the first case and six in the second case. The total number of blocks would probably be determined so as to reduce the standard error of the principal comparison to an acceptable level.

A different situation arises when the treatments can be divided into two groups, one relatively more important than the other. Here trial and error combined with the use of formula (1) will usually indicate a suitable arrangement. For example if the total number of available units is severely limited, we may prefer to attain a specified precision for comparisons within the important group, accepting whatever precision can be obtained from the remaining units for the remaining comparisons. Or we may decide to have the standard errors for comparisons within the more important group and within the less important group to be in the ratio of say 1 : 2. This would be achieved by having the corresponding ratio for the number of observations per treatment be 4 : 1. In fact, by formula (1), if there are $4n$ observations on each of one group of treatments and n observations on each of the second group, the following standard errors for estimated differences are obtained:

for two treatments in first group,	$\sqrt{[2/(4n)]} \times$ standard deviation $= \sqrt{[1/(2n)]} \times$ standard deviation;
for two treatments in second group,	$\sqrt{(2/n)} \times$ standard deviation;
for a treatment in one group compared with a treatment in another group,	$\sqrt{[1/n + 1/(4n)]} \times$ standard deviation $= \sqrt{[5/(4n)]} \times$ standard deviation.

Again the number of units per second group treatment, n, can be determined if the required value of the standard error is known.

The final group of problems for consideration is connected with the within-unit sampling variation of the type illustrated in Example 8.4. Here we have to decide not only on the total number of experimental units but also on the number of repeat observations per experimental unit.

The general principle involved is the obvious one that if the main expense and time are in taking the observations, so that repeat observations on the same unit cost as much as the testing of the same number of new units, the best procedure is to have as many units as are necessary, making either one observation on each unit or two if the variation within units is of intrinsic interest. On the other hand, if, as is commonly the case, the main expense is in the provision and testing of the experimental units, it will be best to use a small number of experimental units, making a relatively large number of observations on each unit. For instance, in Example 8.4, an increase in the number of experimental units would involve the processing of fresh batches of wool and would be expensive, whereas an increase in the number of observations per unit merely involves selecting further lengths for test and carrying out the strength-testing, and the expense of this is slight.

A statistician should be consulted for details on how to proceed in such cases. Two components of standard deviation are involved (see Example 8.4). The within-unit standard deviation measures the variation that would be obtained if a large number of observations were made all on one unit. The between-unit standard deviation measures the variation, in the absence of treatment effects, when the average of a large number of observations on each unit is analyzed. The effective residual standard deviation, for the comparison of the treatment means, is

$$\sqrt{\left\{(\text{between-unit st. dev.})^2 + \frac{(\text{within-unit st. dev.})^2}{\text{no. of repeat obs. per unit}}\right\}}, \qquad (9)$$

and once approximate values for the two components of standard deviation can be found, we can determine as before the standard error that will result from any given number of units and number of observations per unit.

Example 8.9. Suppose that in an experiment such as that of Example 8.4, it is known from previous work that the between-unit and within-unit standard deviations are, respectively, about 1 and 2 units. The standard error of the estimated difference between two treatments is therefore

$$\sqrt{\left\{\frac{2}{\text{no. of units per treatment}}\left(1 + \frac{4}{\text{no. of repeat obs. per unit}}\right)\right\}}.$$

The numerical values of this are shown in Table 8.4.

TABLE 8.4

STANDARD ERROR OF DIFFERENCE BETWEEN TWO TREATMENTS IN A HYPOTHETICAL EXPERIMENT

No. of Repeat Obs. per Unit	1	2	4	8	16	32	Infinity
No. of Units per Treatment							
2	2.24	1.73	1.41	1.22	1.12	1.06	1.00
4	1.58	1.22	1.00	0.87	0.79	0.75	0.71
6	1.29	1.00	0.82	0.71	0.65	0.61	0.58
8	1.12	0.87	0.71	0.61	0.56	0.53	0.50
10	1.00	0.77	0.63	0.55	0.50	0.47	0.45
12	0.91	0.71	0.58	0.50	0.46	0.43	0.41

The following conclusions can be drawn from Table 8.4 and can be paralleled in a more general case.

(*a*) Not much decrease in the standard error is produced by increasing the number of repeat observations per unit beyond about 16 (the corresponding number in the general case is about four times the square of the ratio of the within-unit standard deviation to the between-unit standard deviation).

(*b*) A particular standard error can be produced in various ways. For example a standard error of 0.71 can be obtained with 20 units per treatment and 1 observation per unit (not shown), with 12 units per treatment and 2 observations per unit, with 8 units per treatment and 4 observations per unit, with 6 units per treatment and 8 observations per unit, and also with 4 units per treatment and a very large number of observations per unit. This is the smallest number of units per treatment that will give the required standard error; no allowance has been made for the effective loss of precision consequent on a reduction in the degrees of freedom for residual.

(*c*) If it is possible to assess the relative costs of a unit and of an observation, the most economical combination to produce a given standard error can be found, either mathematically or by direct examination of a table such as Table 8.4.

In more complicated cases with several stages of sampling, there will be several components of standard deviation. The general remarks above are applicable, but details are too complicated to go into here.

8.5 SOME SEQUENTIAL TECHNIQUES

In the methods described in § 8.4, a single calculation is made of the number of observations to be made to attain given precision. It is sometimes useful to fix the number of observations not in one step at the beginning of the experiment, but in several steps, that is, to make the number of observations depend on the actual outcome of the experiment. An experiment set up in this way is called *sequential*.

The general idea of working in stages, deciding what to do at one stage only after examination of all the results obtained up to that point, is of

course widely used; here we are concerned with a specialized aspect of it in which the results already obtained determine, not the objective of further work, but simply how many more observations shall be taken. There are four situations in which these methods may be useful:

(*a*) when initially no reliable estimate of the residual standard deviation is available;

(*b*) when the residual standard deviation depends in a known way on the quantity to be estimated;

(*c*) when a clear-cut decision is required between a small number of courses of action;

(*d*) when an estimate is required with a precision depending on the value of the contrast under estimation.

Some experiments lend themselves naturally to a sequential approach. For example if experimental units have to be dealt with singly, so that observations become available at intervals in time, a sequential determination of the number of observations is often perfectly feasible. In other situations, notably in agricultural field trials, the experimental work has to be planned and started at one time and the results become available together much later. A sequential method is not then practicable for an individual experiment, although it may well be applicable to a series of similar experiments, for example in the repetition for several years of an important variety trial. The general point that the following discussion applies mainly to the first type of experiment should be borne in mind throughout.

In any experiment in which the total number of observations is influenced by the values of the observations care is needed in applying the conventional statistical methods of analysis (Anscombe, 1954), since the practical interpretation of statistical significance limits, etc., requires that the total number of observations is chosen without regard to the outcome of the experiment. If, for example, units are tested in small groups and statistical significance is calculated after each step, the experiment being stopped as soon as, say, the 1 per cent level of statistical significance is reached, the true statistical significance of the conclusions is usually considerably exaggerated. This consideration means that ideally the appropriate method of statistical analysis has to be worked out theoretically for each method of determining the number of observations; in practice the point is usually of importance for the decision problem, (*c*), and sometimes for the fourth type of problem, but not in the other cases.

Consider first the type of problem that would be dealt with by a simple calculation like that of Example 8.5 were a sufficiently reliable estimate of the residual standard deviation available. If there is no such estimate,

a common-sense procedure is to do a preliminary experiment using as many units as possible, subject to the proviso that the final required precision is unlikely to be attained. From the observations, the residual standard deviation and the standard error of the interesting contrasts are estimated in the usual way. If the standard error is already as small or smaller than the required value, the experiment is complete; if the standard error is too large, the residual standard deviation is used, as in Example 8.5, to calculate the total number of observations required, and then the appropriate number of additional units is taken. Such a two-stage procedure is called *double sampling;* it is the simplest form of sequential technique.

Example 8.10. In an experiment to compare three methods of measuring the percentage of red cells in blood, it was required to estimate the true mean difference between any two methods with a standard error of about $\frac{1}{2}$ per cent. For each subject the percentage is measured, in random order, by all methods, i.e., we use a randomized block design. Suppose that from the results of a preliminary test of fifteen subjects, the residual standard deviation is estimated to be 1.85 per cent.* The standard error of the difference between two treatments is $1.85\sqrt{(2/15)} = 0.68$. This is somewhat greater than the standard error originally required, and to calculate how many more subjects should be tested to attain the required precision we argue as follows. If the residual standard deviation is approximately the same for future subjects, the standard error for any particular total number of subjects is about

$$1.85\sqrt{\left(\frac{2}{\text{no. of subjects}}\right)}$$

and if this is to equal $\frac{1}{2}$, we have the equation

$$1.85\sqrt{\left(\frac{2}{\text{no. of subjects}}\right)} = \frac{1}{2},$$

or

$$1.85^2\left(\frac{2}{\text{no. of subjects}}\right) = \frac{1}{4},$$

i.e., number of subjects $= 8 \times 1.85^2$, which is approximately 27. This is the total number of subjects; we have 15 already and so need to test about another 12.

The observations on the 27 subjects are analyzed as a whole by the ordinary methods of analysis for a randomized block design and, provided that the second set of observations are not markedly different from the first, approximately the required precision will result.†

* The methods did not depend on the direct counting of cells, so that the theoretical formula (5) of § 8.3(iii) was not applicable.

† There is an approximation involved in applying ordinary methods of analysis to the results of a sequential experiment, but this is unlikely to be important. The common-sense procedure here is a modification of one due to Stein (1945), who, at the cost of some loss of information, arranged that the precision should exactly equal the required value.

Double sampling is a simple procedure in that only one intermediate stage of calculation is involved, and the experiment falls into just two parts. In some situations, particularly when observations become available singly at say weekly intervals, it may be worth elaborating the scheme somewhat. This can be done by calculating, at frequent intervals, the estimated standard error of the important contrasts. So long as this is greater than the required standard error the experiment is continued, but as soon as it becomes less than or equal to the required value, the experiment is stopped and a full analysis of the results is made by conventional statistical methods. There is an approximation involved in using conventional methods of analysis in a sequential problem, but the effect here is small. The disadvantage of the fully sequential procedure, as compared with double sampling, is that it involves more intermediate calculation; the main advantage is that the rather arbitrary choice of an initial sample size is avoided.

The second type of problem referred to above arises mostly in connection with the special theoretical formulas (5) and (6) of § 8.3(iii).

Example 8.11. In some types of work the observation to be made on each unit is a proportion, for example the proportion of a particular subject's blood cells showing a certain abnormality. The question arises of determining not the number of subjects (units), but the number of observations per subject. A reasonable requirement is often to obtain a certain *fractional* precision in the estimate for each subject. If this were say 20 per cent, we require that if the true proportion abnormal is 5 per cent, the standard error of our estimate should be 1 per cent, whereas if the true proportion abnormal is 1 per cent, the standard error should be 0.2 per cent, and so on.

Now from formula (5) it may be shown that to attain this standard error the number of cells to be counted depends markedly on the true value of the proportion to be estimated. Thus if the true proportion is 5 per cent, the standard error is, from (5), $\sqrt{(0.05 \times 0.95/\text{number of cells per subject})}$. This is equal to the required value, 0.01, when the number of cells counted per subject is 475. Similarly, if the true proportion is 1 per cent, the number to be counted should be slightly less than 2500.

Now if we have to begin with a fairly good idea of the value of the proportion to be estimated, this calculation will determine the number of cells that should be counted. But if all we know is that the proportion lies somewhere between, say, 5 and 1 per cent, the calculation is not helpful, because the appropriate number of cells is so critically dependent on the unknown proportion to be estimated.

Therefore some sequential method seems called for. Double sampling is one possibility. Another method was proposed by Haldane (1945) and is called inverse sampling. The idea here is that instead of fixing the total number of cells to be examined and recording the number of abnormals, counting should continue until a certain predetermined number of abnormals have been obtained. This number is the reciprocal of the square of the fractional error required, and in the above case is $1/(0.2)^2 = 25$. That is, counting should continue until 25 abnormals have been obtained. The resulting proportion of abnormals is an estimate with approximately the required precision.

This case has been described as a simple illustration of a technique for adjusting the number of observations to produce an estimate with the desired precision. The same type of method can be used in more complicated cases; a statistician should be consulted for details.

There is a review of the statistical literature on methods of this type by Anscombe (1953) and a discussion of double sampling methods for these problems by Cox (1952).

The third type of problem, in which a decision is required between two, or sometimes more, courses of action, raises fresh points. There are two approaches; first we may attempt a direct balancing of the monetary loss due to reaching the wrong decision and the cost of experimenting. This is analogous to the approach of Example 8.8. A double-sampling scheme for choosing between two alternative decisions, based on these considerations, has been given by Grundy et al. (1954). The general idea is that a choice has to be made between say two alternative processes; if an initial experiment indicates a substantial superiority for one process, the appropriate decision is reached. If not, further units are tested, the number of new units being chosen to minimize the sum of the average loss arising from reaching the wrong decision and the cost of testing the further units. For these results to be applied a reasonably accurate economic analysis of the problem has to be possible and this is frequently not so. In that case different and more intuitive methods have to be used; the following is an example.

Example 8.12. Kilpatrick and Oldham (1954) have described an experiment to decide between two methods of relieving bronchial spasm in patients with a chronic pulmonary disease. The treatments were the inhalation of either adrenaline, the customary treatment, or calcium chloride, which had been suggested as a possible substitute with certain advantages. The assessment of the drugs was made in terms of an objective measure, the expiratory flow rate (e.f.r.).

The experimental procedure was of the paired comparison type, as follows. In the morning the e.f.r. of a subject was determined and then a 15-minute inhalation of one or other substance was given. Neither patient nor observer knew which, the decision having been reached by randomization. The e.f.r. was again determined on completion of inhalation, and on the evening of the same day the procedure was repeated using the other substance. The difference between the gain of e.f.r. resulting from calcium chloride inhalation and that resulting from adrenaline inhalation was worked out as each subject's results were obtained.

Suitable subjects were expected to become available for the experiment only infrequently, so that it was of some importance to reach the appropriate decision in the most economical way. The problem is of the same general type as that discussed immediately above, but any economic analysis of the consequences of

a wrong decision is out of the question. Instead the following considerations were formulated, after careful thought:

(*a*) If calcium chloride caused subjects to gain, in the long-run average, 10 liters per min of e.f.r. more than they gained on adrenaline, it was desired to reach the decision to prefer calcium chloride:

(*b*) If the gain of e.f.r. with calcium chloride was no greater than with adrenaline, in the long-run average, it was desired to reach the decision to prefer adrenaline, in view of its well-established virtues.

(*c*) The chance that, in the situations described in (*a*) and (*b*), we reach the wrong decision as a result of chance fluctuations in the observations is to be only 1 per cent.

For any scheme that is set up, there will be an operating characteristic of the

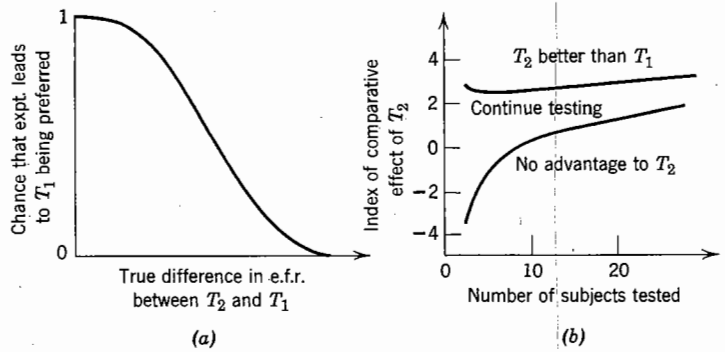

Fig. 8.1. A sequential scheme for comparing two treatments. (a) Typical operating characteristic curve; (b) Boundaries defining the test.

general shape shown in Fig. 8.1(*a*); what we are doing in (*a*)–(*c*) is to find two points on this curve, one towards each end. It is implicit that, for example, if the true difference in gains exceeds 10 per cent in favor of calcium chloride, the probability of deciding to prefer adrenaline is even less than 1 per cent.

When the requirements (*a*)–(*c*) have been formulated, it is a statistical problem to determine a rule for conducting the experiment, so that the requirements (*a*)–(*c*) are satisfied. The rule will say, after each subject's results become available, either

(i) that the experiment should be stopped and adrenaline preferred; or
(ii) that the experiment should be stopped and calcium chloride preferred; or
(iii) that a further subject should be tested.

That is, the experiment is allowed to continue until sufficient results are available to justify a decision.

The most convenient form for representing the sampling rule is on a diagram, such as Fig. 8.1(*b*). As explained above, for each subject the difference is calculated between the gains in e.f.r. on calcium chloride and on adrenaline. After each subject is tested, the cumulative total of the differences is worked out

and divided by the square root of the cumulative total of the squares of the differences. Thus if the first three differences are 2.6, 7.3, and −1.4, the quantity calculated is $(2.6 + 7.3 - 1.4)/\sqrt{(2.6^2 + 7.3^2 + 1.4^2)} = 1.080$. This is an index of the comparative effect of calcium chloride. Its use is mathematically equivalent to that of the mean difference divided by its standard error. The index is large and positive if calcium chloride is much the better treatment, large and negative if adrenaline is much the better treatment, and zero if the average of the observed differences is zero. This index is plotted step by step on Fig. 8.1(*b*). On this diagram are two boundaries; as soon as one or other of them is crossed, the experiment stops and the corresponding decision is reached. So long as the plotted point remains between the boundaries, the experiment is continued. If calcium chloride is much the superior treatment, it is very probable that the index will rapidly cross the upper boundary, leading to a speedy decision; similarly if adrenaline is much superior, the lower boundary is likely to be crossed after the testing of only a small number of subjects. If, however, the situation is less clear-cut, it is likely that the index will remain between the boundaries until an appreciable number of subjects has been tested. That is, the boundaries arrange that the experiment is speedily ended in the clear-cut cases, and is continued in the doubtful cases. The formulas for determining the boundaries are given by Rushton (1950).

The use of a sequential scheme of this type, which suits the number of observations to the requirements of choosing between two alternatives with preassigned chances of error, leads on the average to substantial economies in the number of subjects needed to reach a decision. This is particularly so when there is a big difference between the treatments.

In the experiment described by Kilpatrick and Oldham, the observations were all substantially negative, suggesting adrenaline to be the superior substance, and after only four subjects had been tested, the lower boundary was crossed, see Fig. 8.1(*b*), and the experiment was ended with the decision to prefer adrenaline.

Now in a strict statistical sense, this is the only conclusion that can be given formal justification, i.e., that in accordance with the criteria (*a*), (*b*), and (*c*) the appropriate decision is to prefer adrenaline. Yet the fact that all results were negative suggests not only that adrenaline is to be preferred, but also that it is actually superior; in accordance with (*b*), adrenaline would have been preferred even if there had been no difference between the substances. It is natural to try to estimate, with limits of error, the amount of the difference between the substances, but this cannot be done, at any rate until there has been further research into the statistical problems involved. Nor can we measure the statistical significance of the observed difference, beyond saying that there is a significant departure from (*a*) at the 1 per cent level. Thus, the economy of the sequential scheme has been achieved at the cost of restricting the conclusions to the choice between two decisions. Kilpatrick and Oldham point out that the experiment might better have been designed to choose between three decisions: adrenaline superior, no difference, and calcium chloride superior. However, even then the statistical conclusions are severely limited, and no estimation of limits for the amount of the true difference is at present possible.*

* This raises interesting issues of general statistical theory. It is not clear whether an essential defect in conventional theory is involved, or an essential property of the design, or merely a mathematical difficulty in working out details of appropriate techniques.

This example has been discussed in some detail because it illustrates the advantages and disadvantages of these sequential decision procedures. The formulas for determining the boundaries in the sampling diagram have been worked out for many standard statistical situations (Wald, 1947); this was done with the application to industrial inspection problems in mind, where a clear-cut decision between accepting and rejecting a batch has to be made. There have not been many applications of the methods in research work. An excellent account of the methods with a view to their application in medicine has been given by Armitage (1954), who has also given (Armitage, 1957) some very interesting new types of

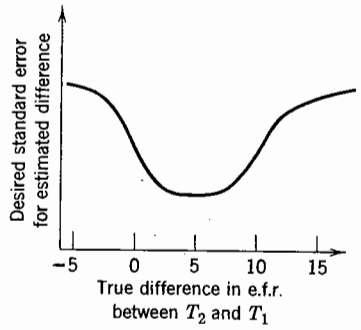

Fig. 8.2. Desired standard error in a situation where the precision required depends on the true value being estimated.

sequential scheme that may be more suited to some sorts of experimental work than Wald's methods.

To sum up, sequential decision procedures merit serious consideration whenever a choice between two or three (or any small number) of courses of action is required, when experimental units are tested in order and it is practicable to do a certain amount, usually small, of calculation after the testing of each unit or group of units, and when requirements analogous to (*a*), (*b*), and (*c*) of Example 8.12 can be formulated. The disadvantage is that, at present, the statements at the end of the experiment that can be given statistical justification are rather limited.

This last point means that the sequential decision procedures just described are inappropriate whenever it is part of the object of the experiment to estimate, with limits of error, the magnitude of contrasts. A natural method to use in cases where such an estimate is required, even though the main objective is the decision, is to try to set up a scheme for estimating, say, the difference between treatments with a standard error

depending on the true value of the difference. This is the fourth type of sequential scheme. For example, in the situation of Example 8.12 it might have been reasonable to require an estimate of the true difference in e.f.r. between substances with a standard error of the general form in Fig. 8.2. If the true difference lies in the "doubtful" range (0, 10), we require a low standard error, so that a suitably precise significance test can be made for reaching the appropriate decision; if the difference lies outside this range, we still require an estimate of the difference, but are content with a much higher standard error. Double sampling schemes for achieving these objectives have been discussed theoretically by Cox (1952), but no examples of their use in practice are known.

SUMMARY

It is very often desirable either to make a preliminary calculation of the precision to be expected from an experiment of the size contemplated, or, preferably, to determine the size so as to just attain a desired precision. Precision can, for most purposes, be measured by the *standard error* of the contrasts of interest, for example by the standard error of the estimated difference between two treatments.

The value of the standard error depends on the contrast to be estimated, on the number of units, and on the amount of uncontrolled variation, which is itself measured by the residual *standard deviation*. This can be estimated

(*a*) from the observed dispersion of the observations between units receiving the same treatment, eliminating any part of the variation that is balanced out by the design of the experiment;

(*b*) from the magnitudes of high-order interactions in a factorial system;

(*c*) from theoretical considerations, as for example when the observation is the count of a number of occurrences of a randomly occurring event;

(*d*) from the magnitude of the within-unit sampling variation, when the main observation on each unit is the mean of several readings;

(*e*) from the results of previous similar experiments.

Method (*a*) is usually the best for the analysis of data, although the comparison of measures of residual variation from several sources is frequently instructive. Methods (*c*) and (*e*) are the ones applicable for the preliminary estimation of precision.

Once an approximate value for the standard deviation has been obtained, the standard error of any particular contrast can be worked out corresponding to any given number of observations, or, alternatively, the

number of observations to achieve a given standard error can be predicted. If some comparisons are more important than others, the number of units should not be the same for each treatment. In more complicated cases both the number of experimental units and the number of observations per unit have to be determined.

Sequential methods, of which the simplest is double sampling, are sometimes useful, particularly when the experiment can conveniently be done in stages with some intermediate calculation between stages. The idea here is that the number of units should be determined in the light of the observations actually obtained and not settled definitely in advance. This technique is especially worth considering when (*a*) the standard deviation is initially completely unknown, or is appreciably dependent on the quantities being estimated or when, (*b*) a clear-cut decision is required between two or more courses of action or when, (*c*) an estimate of a contrast is required with a precision depending markedly on the value of the contrast.

REFERENCES

Anscombe, F. J. (1953). Sequential estimation. *J. R. Statist. Soc.* B, **15,** 1.

——— (1954). Fixed-sample-size analysis of sequential observations. *Biometrics,* **10,** 89.

Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *Q. J. of Medicine,* **23,** 255.

——— (1957). Restricted sequential procedures. *Biometrika,* **44,** 9.

Cochran, W. G., and G. M. Cox. (1957). *Experimental designs.* 2nd ed. New York: Wiley.

Cox, D. R. (1952). Estimation by double sampling. *Biometrika,* **39,** 217.

Eisenhart, C., M. W. Hastay, and W. A. Wallis. (1947). *Selected techniques of statistical analysis.* New York: McGraw-Hill.

Goulden, C. H. (1952). *Methods of statistical analysis.* 2nd ed. New York: Wiley.

Grundy, P. M., D. H. Rees, and M. J. R. Healy. (1954). Decision between two alternatives—how many experiments? *Biometrics,* **10,** 317.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika,* **33,** 222.

Kilpatrick, G. S., and P. D. Oldham. (1954). Calcium chloride and adrenaline as bronchial dilators compared by sequential analysis. *Brit. Med. J.,* part ii, 1388.

Rushton, S. (1950). On a sequential *t*-test. *Biometrika,* **37,** 326.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.,* **16,** 243.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Yates, F. (1952). Principles governing the amount of experimentation in developmental work. *Nature,* **170,** 138.