CHAPTER 9

Choice of Units, Treatments, and Observations

9.1 INTRODUCTION

The central techniques of the subject have been described in the preceding chapters. Ideally, randomization is used to achieve the absence of systematic error; the devices of blocking and of adjusting for concomitant observations lead to increased precision; and the idea of factorial experiments allows both the effective precision to be increased and also the range of validity of the conclusions to be extended. Finally, the methods outlined in the preceding chapter enable a suitable size of experiment to be set up. In this way the criterion for a satisfactorily designed experiment, set out in § 1.2, are in principle satisfied.

All this, however, assumes that three matters have been settled: that we have decided on the treatments to be compared, the types of observation to be made, and the nature of the experimental units to be used. These issues are clearly of central importance and it might fairly be claimed that they are the essence of experimental design. They are, however, generally regarded as being technical questions specific to the subject matter of the experiment and so are not considered in statistical studies of experimental design. In the following, a few general points will nevertheless be made.

9.2 CHOICE OF EXPERIMENTAL UNITS

An experimental unit was defined in § 1.1 to correspond to the smallest subdivision of the experimental material such that different units may receive different treatments. We consider to be included in the definition of the unit all those aspects of the experimental set-up not involved in the treatments, i.e., those that are independent of the particular assignment of treatments adopted. Thus in an industrial investigation an experimental unit might be defined as a particular batch of raw material processed at a certain time, tested on a particular apparatus by

a certain observer, etc., a treatment being a particular method of processing.

In setting up experimental units, the following are among the questions to be settled:

(i) an appropriate size for the experimental units;

(ii) whether it is important that the conditions investigated should be representative of "practical" conditions, and whether a wide range of validity for the conclusions is desirable. In particular, it may be desirable to introduce more variation than is present among the experimental units initially available, or to introduce special treatments to represent the effect of variations in external conditions;

(iii) whether one physical object can profitably be used as a unit several times and whether this, or any other aspect of the experiment, introduces important lack of independence in the responses of different units.

(i) Size of Units

In some experiments a suitable choice for the amount of material to be included in each unit is important. An example is the selection of plot size (and shape) in an agricultural field trial.

Technical considerations enter into such choices to a considerable extent. Sometimes, however, the following situation arises. A certain total amount of material is available and, within certain limits, can be divided into any number of experimental units. The number of repeat observations to be made on each unit is also at our disposal. What is the optimum procedure? Alternatively, how much experimental material is needed to attain a specified precision in the most economical way? This problem is closely related to that of § 8.3, where the idea of components of dispersion between and within units was introduced.

Suppose, to take a specific case, material is available for 100 hours' production, with four treatments for comparison, the minimum run on one process being say 5 hours. One possibility is to have just four units, i.e., to run on one process for the first 25 hours, and so on. Numerous observations of each type are taken within each unit; for example if the production flow is continuous, single observations might be taken at hourly intervals, to give finally 25 observations on each process. This is a bad design, since a proper estimate of error, based on the comparison of different whole units receiving the same treatment, is not available. To evaluate the precision of comparisons between treatments, quite specific assumptions have to be introduced about the form of the haphazard variation. Not only are such assumptions best avoided, but also the resulting precision is likely to be rather low. Hence we should use

this arrangement only if there are strong practical reasons for keeping the number of treatment changes to a minimum.

At the other extreme we could have 20 experimental units, each corresponding to 5 hours' production, and these might, for instance, be arranged in 5 randomized blocks, grouping into blocks on the basis of order in time and still taking hourly observations. Similarly, there are various intermediate possibilities. According to the formula of § 8.4, the standard error of the estimated difference between two treatments is

$$\sqrt{\left[\frac{2}{\left(\begin{matrix}\text{total no. of}\\ \text{obs. per treatment}\end{matrix}\right)}\left\{\left(\begin{matrix}\text{st. dev.}\\ \text{within units}\end{matrix}\right)^2 + \left(\begin{matrix}\text{no. of obs.}\\ \text{per unit}\end{matrix}\right) \times \left(\begin{matrix}\text{st. dev.}\\ \text{between units}\end{matrix}\right)^2\right\}\right]},$$

the total number of observations per treatment being the product of the number of experimental units per treatment and the number of repeat observations per unit.

Note first that if the standard deviation between units is zero, the precision depends only on the total number of observations per treatment and not on the number of units per treatment. Second, the precision corresponding to any desired distribution of effort can always be worked out, provided that we can specify how the two standard deviations in the formula depend on unit size. As a first approximation, the standard deviation within units may be treated as constant and the standard deviation between units as varying according to the law

$$\text{standard deviation} = A \times (\text{size of unit})^{-B},$$

where B is a constant usually between 0 and $\frac{1}{2}$; for the application of this to agricultural experiments, see Fairfield Smith (1938). Both A and B and the standard deviation between units are characteristics of the experimental material to be estimated by the analysis of suitable data from previous experiments. Once they have been determined, the precision of any set-up can be determined. For a given total number of observations and a given total quantity of material, maximum precision will be attained with minimum block size, but this is not usually the relevant comparison. A more realistic analysis is to assume that the total cost of the experiment is approximately

$$\begin{aligned} &C_1 \times \text{total amount of material used}\\ &+ C_2 \times \text{total number of experimental units used}\\ &+ C_3 \times \text{total number of observations made,}\end{aligned}$$

where C_1, C_2, and C_3 are constants expressing in the same units the costs of a unit of material, of an experimental unit (e.g., of changing from one process to another), and of an observation.

If all the quantities mentioned above can be determined approximately, the information is thus available from which to decide on the most efficient arrangement of resources. It is clear that to do this with any precision calls for quite detailed knowledge of the system.

(ii) Representative Nature of Units

In many technological investigations, particularly in the final stages immediately preceding practical application, it is important that the conditions investigated should be as representative as possible of the conditions under which the results are to be applied. This can have an important influence on the design of the experiment, especially if the practical conditions are very variable or if the treatment effects are likely to be rather sensitive to changes in the external conditions. In any case proper control treatments should of course be included to check that an apparent treatment effect is not solely a consequence of, for example, increased attention paid to the units during the experiment. The point at issue now is that a treatment effect which may be perfectly genuine under experimental conditions may be quite changed under working conditions.

There are various steps which may be taken. Thus, apart from direct efforts to reproduce practical conditions as closely as possible, a factorial experiment may be set up in which one or more factor levels correspond to artificially severe forms of complications likely to arise in practice, that is, we may insert a separate treatment "conditions" and examine the interaction of this with the treatment effects of direct concern.

If the main variation connected with the experimental units lies in the experimental material itself rather than in the external conditions, it will be important to consider where the units used come from. For instance, if in an animal feeding experiment it is required to apply the conclusions to pigs of a certain breed, then ideally the pigs in the experiment should be a sample of pigs of this breed chosen by a sound statistical sampling procedure. Again, in an industrial experiment, if it is suspected that the treatment effects depend somewhat on the particular consignment of raw material used, then the material used in the experiment should be chosen appropriately from the whole set or population of consignments to which the conclusions are to apply. This is a counsel of perfection which is probably practicable only very rarely. It is, however, desirable in such cases to check so far as possible that the units used do not differ in an obvious respect from the population to which it is required to extend the conclusions. Further, it is almost always advantageous to arrange the

experiment so that any variation in the treatment effects from unit to unit can be detected and its nature analyzed, for if the treatment effects may be shown to be effectively constant over the experiment, confidence in extrapolation of the conclusions is much greater.

In scientific work, on the other hand, the representative nature of the experimental units is often not of great interest. The choice of experimental material so that the treatment effects can be observed in a simple and illuminating form will be of considerable importance; however, the emphasis is, to begin with, usually on the deliberate choice of the most suitable material. Even here it may be desirable to include a range of experimental units, both in order to get conclusions with a broader basis and possibly also to provide a link with earlier work.

In both sorts of experiment it may, therefore, be desirable to include deliberately additional variations between experimental units with the related objects of examining nonconstancy of the treatment effects and of extending the range of validity of the conclusions. These additional variations, such as those between sexes or between units of two very different types, are best introduced as a classification factor in a factorial experiment (Chapter 6). With four main treatments, T_1, T_2, T_3, and T_4 and two levels of the classification factor, M and F, the two main types of design are recalled in Table 9.1. In the first type

TABLE 9.1

TWO DESIGNS FOR THE INCLUSION OF A SECOND FACTOR

(*a*) *Simple Factorial Arrangement in Randomized Blocks*

Block 1:	MT_4;	MT_1;	FT_2;	MT_3;	FT_1;	MT_2;	FT_4;	FT_3
Block 2:	FT_1;	FT_3;	MT_4;	FT_2;	MT_1;	MT_3;	FT_4;	MT_2

(*b*) *Split Unit Arrangement*

Whole Unit 1:	F;	T_2	T_1	T_3	T_4
Whole Unit 2:	M;	T_4	T_3	T_1	T_2
Whole Unit 3:	M;	T_1	T_3	T_4	T_2
Whole Unit 4:	F;	T_2	T_4	T_3	T_1

a simple factorial arrangement is used, with units of the two types mixed together in each block. In the second, we have in effect a randomized block design for the treatments $T_1, \ldots, T_4$, with each block consisting of units of one type, blocks of different types being randomly intermixed. (In the language of § 7.4, this is a split unit experiment with the treatment of direct interest considered as the subunit treatment.)

The second arrangement is to be preferred if the units of the two types

are most conveniently dealt with separately, or if the smaller block size of the second design is likely to lead to an increase of precision in the estimation of the important effects.

(iii) Independence of Different Units

It is very desirable that the different experimental units should respond independently of one another, in the senses that there should be no way in which the treatment applied to one unit can affect the observation obtained on another unit, and that the occurrence of, say, an unusually high or low observation on one unit should have no effect on what is likely to occur on another unit. The first requirement is necessary to allow the effects of the different treatments to be sorted out from one another; the second ensures that a proper estimate of error is obtained from the comparison of observations on units receiving the same treatment.

The precautions to be taken depend on the nature of the experiment, but they usually consist in physical isolation of the different units and, in particular, of the units receiving the same treatment. This needs to be done at all stages of the experiment at which important variations may be introduced. Thus if appreciable variation is likely to occur in obtaining the observations, e.g., in testing an industrial product, it will be desirable to deal with different units in an order involving some randomization. This will ensure that systematic errors arising in the testing procedure do not bias the comparisons and will also tend to minimize any subjective tendency to make observations on the same treatment more alike.

The main situation in which observations on one unit may be affected by the treatment applied to a different unit is when the same physical object (subject, animal, etc.) is used as a unit several times. A considerable increase in precision may often be obtained by doing this, because of the elimination of the effect of differences between subjects, but even if special precautions are taken, it may not be possible to avoid a carry-over of treatment effects from one unit to another. Designs that allow for carry-over effects are discussed in Chapter 13. A further difficulty with this sort of design is that it may involve comparing the treatments under conditions rather different from those that apply in practice.

9.3 CHOICE OF TREATMENTS

So far in this book, we have been discussing how to plan experiments so that reliable and precise comparisons of treatments can be made when uncontrolled variation is present. The selection of treatments to be compared falls almost entirely outside the discussion as being partly a

technical question specific to the field under study and partly a question of general scientific procedure.

Thus the investigation of a rather complicated phenomenon might take the form first of a general survey of the effect on the system of a variety of changes; a factorial experiment may be very appropriate here. Then one or more ideas about how the system, or part of it, "really works" are tested as rigorously as possible by suitable modification of the system (treatments). Nearly always a series of experiments is necessary, the initial ideas being modified at each stage, wherever necessary. In so far as we have to disentangle the treatment effects from irrelevant variations, the methods we have been discussing are, of course, applicable; treatments are chosen so as to give as direct an indication as possible of the underlying mechanism. In suitable cases many different types of modified systems may be used, possibly involving quite distinct experimental techniques, and possibly using systems quite different from the initial one.

Even in experiments with a direct practical aim, it may be possible to include special treatments intended to give fundamental knowledge about the process. Thus in an industrial experiment to compare new and standard processes, various modified forms of the new process might be used, at any rate in a small-scale trial, even though there might be no intention of putting these additional processes to direct practical use.

In many experiments the inclusion of a control treatment is essential; the correct specification of the control can be very important too. It should consist in applying to the experimental units a procedure identical to that received by the "treated" units in all respects except that which it is desired to test. To take a simple example, in assessing a new drug we would not usually wish to consider as part of the treatment effect, the improvement that normally results from receiving treatment with pharmacologically inactive substances. Therefore, the control treatment should be a placebo, indistinguishable to the subject from the new treatment* (see Chapter 5); it may be profitable to include an untreated control group as well. Again, if it is required to assess the effect of the excision of a portion of an experimental animal, the control group of animals should be subjected to as much as possible of the procedure applied to the experimental group; here again a group of untreated animals can be included too, but they should not be regarded as the main control.

Example 2.5, concerned with the comparison of drugs for the relief of

* Rutstein (1957) has described consequences, in some experiments of this type, of failing to include proper control groups.

headaches, illustrates the importance of controls in a rather different way. There the control treatment was in effect used to divide the subjects into two types and hence to show a nonuniformity in response. Had this experiment been done without a control group, it is extremely likely that misleading conclusions would have been obtained. Example 1.8 is another illustration of the importance of controls.

Where the treatments differ qualitatively, it is usually desirable that the treatments differ in single specifically identified ways. Otherwise the implications of any treatment differences found will not be clear. Thus, suppose that we compare a standard industrial process A with a process B, modified in several respects. The resulting comparison may be of immediate practical interest, but is unlikely to add much to understanding of the process, since the cause of any change that is observed cannot be identified. In scientific experiments it is very desirable that any treatment differences found should have as far as possible unique interpretations. It often happens that an experiment establishes clear-cut treatment effects, but that on consideration it is seen that these have two or more quite different interpretations. Further treatments ought to have been included to discriminate between the different explanations. The initial choice of treatments so that ambiguities of this sort are avoided is one of the most important and difficult steps in experimental design. It will often be an advantage to set up a factorial structure for the treatments under investigation.

The discussion and examples of § 7.2 concern the choice of factors for inclusion in factorial experiments, and this section should be re-read at this point. The selection of factor levels when the factors are quantitative has been discussed in § 7.3; see also § 14.3.

A final general remark is that one of the best checks of the general reliability of an experiment is to show agreement with previously established results in that field. Hence it is quite often worth including a treatment solely with this object.

9.4 CHOICE OF OBSERVATIONS

The preceding chapter dealt with the number of repeat observations of a particular type that should be made on each unit, and with the number of units. In most experiments, however, observations of several different types are made on each unit and we now discuss briefly the selection of quantities for observation. They can be classified first according to the purpose for which they are made and second according to their mathematical nature, for example whether they are quantitative or whether they amount solely to an ordering of different objects.

Observations may be classed roughly into the following types.

(i) Primary Observations

These either measure properties for which the treatment comparisons are of direct interest, or are needed in the calculation of such quantities. Yield of product is an obvious example; another is the set of observations that are needed on each plot of a sugar-beet experiment in order to calculate the yield of sugar in cwt/acre.

(ii) Substitute Primary Observations

If a particular primary observation is difficult to obtain, it may be profitable to use a substitute observation that is easier to get. For example, the primary properties in a textile experiment may be the wearing properties, handle, and appearance of the woven fabric. All these in principle can be measured by consumer trials, but it is much easier to measure physical properties of the fabric, such as life in a laboratory wearing test, flexural rigidity, and yarn irregularity, which are thought to be closely related to the more nebulous qualities judged by the consumer. Another textile example is that in a spinning experiment, a primary quantity might well be the behavior of the yarn in the next process, weaving, as measured, say, by the number of yarn breaks per loom running hour. A substitute observation for this is the yarn strength as measured in a laboratory winding or strength test. In rather small experiments of this sort we would probably rely entirely in such substitute observations.

We normally regard an observation as a substitute one only if there is no theoretical or empirical relation converting the observation taken into the primary one, and more particularly if there is far from complete correlation between the two. The principle of measuring one thing by observing another closely dependent on it is of course the basis of many methods of measurement.

(iii) Explanatory Observations

These are taken to attempt an explanation of any treatment effects found on the primary observations. Thus suppose that in the experiment discussed briefly in (ii), the primary observations that we are directly interested in are taken, and are subjective in nature. Then we shall want to explain treatment differences found for these observations in terms of the physical or chemical behavior during processing, and to do this additional observations need to be made; some of these may, in fact, be those considered as substitute observations under (ii).

(iv) Supplementary Observations for Increasing Precision

In § 4.3, a method was explained for reducing the effect of variation between experimental units by a process of *adjustment*. In this, the mean value, for each treatment, of the main observation is adjusted to what it would have been had the units had the same value of a secondary or concomitant observation. In an alternative simpler, but in general less efficient, procedure an index of response is used. Thus in a psychological experiment the main observation might be the score obtained by a subject after exposure to a treatment, the concomitant observation being the score obtained in a similar test administered before applying the treatment. A natural index of response is the difference between the two scores for the subject.

The condition for the validity of this is that the observation used as a basis for adjustment should be unaffected by the treatments. This is so if the observation is made before the treatments are applied. For the procedure to be useful there should, of course, be high correlation between the two types of observation. If it is suspected that the main source of uncontrolled variation lies in the individual nature of the experimental units, rather than in the variation between natural groups of units, the skilful choice of concomitant observations can lead to an appreciable increase in precision.

(v) Supplementary Observations for Detecting Interactions

A further important use of the type of observation discussed in (iv) is to examine whether treatment effects vary systematically between units, for example whether subjects with a high initial score tend to respond differently to the treatments from those with low initial scores. The general importance of examining whether treatment effects are constant has been discussed in Chapter 2.

(vi) Observations for Checking the Application of the Treatments

For example if the treatments correspond to different temperatures, it will be natural to check independently that the appropriate temperature has in fact been achieved. If there is a serious discrepancy, and corrective action is impossible, it may be worth adjusting for the error in the statistical analysis of the results.

(vii) Observations to Check on External Conditions

Routine observations are often necessary to check that no unexpected gross change in external conditions occurs, and that no mishap, irrelevant to the treatment comparisons, occurs to any of the units.

A systematic consideration of these types might often be a good thing. Of course it remains true that the best experiments are often simple in conception and that many of the types of observation may not be required.

In some situations there may be no doubt how a particular property should be measured, a generally accepted objective and quantitative scale of measurement being available. This is the case with such things as yield of product, etc. Once what is to be considered as effective product is clearly defined, there will usually be little difficulty, in principle, in measuring what is wanted. Similarly, classical research has established reliable methods of measuring the standard physical and chemical quantities. In new fields, however, the situation may be quite different and some discussion will now be given of some of the general issues that may need consideration.

First, problems may arise in reducing a complex response to a manageable form. This, however, usually is a question more of analysis and interpretation than of experimental design. Thus in studies of the smoothness of metallic surfaces or the irregularity of textile yarns the initial observation will be an irregular trace showing the variation of thickness. In learning experiments the observation on one animal may consist of a series of "successes" and "failures," corresponding to successive attempts at a task. During learning the proportion of successes increases to near unity, during extinction the proportion decreases again. For further study it is usually very desirable to reduce such data and two general procedures can be used for this. One is to define one or two quantities which have a direct interpretation in terms of the more complex response. For instance, the rate of learning is often measured by the number of trials necessary before, say, five consecutive successes are obtained; the irregularity traces might be summarized by the coefficient of variation of thickness. The second general procedure is to estimate the parameters in a mathematical model that is thought to represent the system. In the examples just discussed these models would be probabilistic, for example one of the various stochastic models that have been advanced to represent learning. If such models give real insight into the system, their use is of course very desirable; if not (if the model is purely empirical and if appreciable extra labor is involved in fitting the model) it will be worth considering whether some simpler method can be used. The danger of the first method is that if the response can change in various ways, the empirical indices may be misleading. For example if a learning curve should really be described by the limiting proportion of correct responses when learning is complete and by the rate at which this limit is reached, it is clear that any single measure of

learning may be misleading. This sort of consideration is particularly important when it is planned to record directly the final indices of interest.

Sometimes, instead of recording observations on a scale, we may take dichotomous observations, for example, on whether the diameter of a circular cylinder is or is not greater than a critical value, whether a foodstuff is judged satisfactory or unsatisfactory, whether one foodstuff is judged preferable to another, and so on. Such observations may be considered either for convenience and simplicity, or because only qualitative judgements are possible.

An ingenious example of the first use is a technique due to Anderson (1954) for comparing the strengths of two textile yarns A and B. Two 10 in. lengths, one of A and one of B, are joined and the resulting 20 in. length pulled until it breaks. The position of the break shows which is the weaker of the two lengths. If this is repeated, the weaker specimen will be from yarn A in about one-half the trials if the strengths are really the same, and if a significant departure from equal proportions is observed a difference between the strengths of the yarns has been established. This gives a simple comparative test without special apparatus and without quantitative measurement. Such a method has a statistical efficiency of about 64 per cent for establishing the existence of small differences, as compared with the quantitative method in which the strength of each section is measured. That is, a certain number of dichotomous observations and 64 per cent of that number of quantitative observations give about the same precision. This may be considered quite a high efficiency if appreciable simplification is attained; the main disadvantage of the qualitative method is that if a moderate or large difference exists, the method will indicate its presence but its magnitude will be estimated only with low precision. Similar remarks apply to comparison with a standard (as when the diameter of a cylinder is compared with a critical value) although this sort of procedure is perhaps not very likely to be used in experimental work.

A natural generalization of the pairing method just discussed is the *ranking* of more than two objects in order either of scale value or of subjective preference. Again quite sensitive tests can be made of the existence of small differences, but only poor estimates are obtained of the size of large differences. One problem of design that sometimes arises is that of choosing between either ranking, say, five objects or placing in order all possible pairs of objects. In general, ranking is likely to be the more economical method provided that it is practicable to judge several objects simultaneously without loss of precision.

An alternative to the qualitative judging of pairs and to ranking is the

scoring of differences on a simple scale, for instance the five-point scale,

+2: A is much preferred to B,
+1: A is preferred to B,
0: no difference,
−1: B is preferred to A,
−2: B is much preferred to A.

General instructions are given to the judges to try to standardize the use of the scale as much as possible. The advantage of this over the previous method is partly that an analysis, even of complex designs, may usually be made by standard statistical methods (Scheffé, 1952) and partly the added sensitivity that should, in theory at any rate, result from the distinction that may be drawn between strong preference and preference. No comparative studies seem to have been published of how the differences between the two methods work out in practice. A tentative recommendation is to use a five or more point scale whenever the comparison of results from different judges is not of particular importance.

Statistical techniques are available for determining from the data a system of scoring the qualitatively different responses that will give most sensitive discrimination between treatments (Fisher, 1954, p. 289).

SUMMARY

The choice of experimental units, of treatments for comparison, and of types of observation is considered briefly.

The size and representative character of the units may be important and the insertion of additional variations among the units profitable. Treatments are chosen in some cases because of their direct interest, in others because of the information they may give about the mechanism underlying the system and in others to attain wider range of validity for the conclusions about the main treatments.

The main purposes for taking observations of a particular quantity are: to get information of direct technological or scientific importance, to substitute for such primary observations, to explain treatment effects occurring with the primary observations, and to use as a supplementary variable to increase precision or to detect variations in the treatment effects.

A brief discussion is given of the forms in which observations may be obtained.

REFERENCES

Anderson, S. L. (1954). A simple method of comparing the breaking load of two yarns. *J. Text. Inst.*, **45**, T472.

Fairfield Smith, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *J. Agric. Sci.*, **26**, 1.

Fisher, R. A. (1954). *Statistical methods for research workers.* 12th ed. Edinburgh: Oliver and Boyd.

Rutstein, D. D. (1957). The cold-cure merry-go-round. *Atlantic Monthly*, **199**, No. 4, 63.

Scheffé, H. (1952). An analysis of variance for paired comparisons. *J. Am. Statist. Assoc.*, **47**, 381.