CHAPTER 5

Randomization

5.1 INTRODUCTION

In Chapter 3 designs called randomized blocks and Latin squares were introduced. Their object is to increase precision. In both cases some constraint was introduced into the allocation of treatments, for example by requiring each treatment to occur once in each block of the randomized block design. Subject to the constraint we said that the arrangement of treatments is to be randomized and we now must consider this process of randomization in detail.

The discussion falls into three main parts dealing with the practical details of carrying out the randomization, with the justification of the procedure and with various detailed points that occasionally arise in applications.

5.2 THE MECHANICS OF RANDOMIZATION

The basic operation is that of arranging in "random order" a series of numbered objects. In the more complicated designs this process has to be applied several times, but we shall begin with the simple cases. One of the essential features of randomization is that it should be an objective impersonal procedure; to arrange things in random order does *not* mean just to manipulate them into some order that looks haphazard.

One method of randomizing is to shuffle numbered cards or to draw numbered balls out of a well-shaken bag. Such methods are sometimes useful, but we shall not discuss them further. The main method, and the one we shall deal with, is the use of numerical random tables. Such tables for experimental design take two forms: tables of random permutations and tables of random digits. Short examples of both are given for illustration in the Appendix, Tables *A*.1 and *A*.2 of random permutations being taken from Cochran and Cox's book (1957, § 15.5) and Table *A*.3 of random digits from Kendall and Babington Smith's (1939) tables. These sources give much more extensive tables.

The use of tables will be illustrated by examples.

Example 5.1. Consider the randomization in Example 3.2; all that needs to be recalled is that this was a randomized block experiment with three blocks of five plots each and five treatments $T_1, T_2, \ldots, T_5$.

(*a*) Number the plots in each block 1, . . . , 5 in any convenient way.
(*b*) Use Table *A*.1, Random Permutations of 9, random permutations of 5, which are all that we need, not being available. Choose a starting point in a haphazard way without looking at the tables. For example write down a number (1 or 2) for the page, a number (1 to 5) for the row and a number (1 to 7) for the column block. Thus 2, 3, 6 gives the group beginning 9, 3, 4, 6, 2, 7, 5, 8, 1.
(*c*) Read off the first permutation, omitting the numbers 6 to 9 since there are only five treatments. This gives 3, 4, 2, 5, 1, and determines the allocation of treatments in the first block. Thus T_3 goes on plot 1, T_4 on plot 2, and so on.
(*d*) For the next block use the next permutation in the Table, which is 7, 4, 6, . . . and leads to the order $T_4\ T_2\ T_5\ T_3\ T_1$. Similarly for the third block, using, of course, different tabular permutations for each block.

A useful alternative device for selecting a starting point is to begin, on the first application, at the beginning of the table and to mark the last permutation used with a light pencil mark. At the next application carry on from where the last application finished, and so on. This assumes that recollection from a previous reading of the table is unimportant.

Example 5.2. Suppose that we have a randomized block experiment with 10 units per block and 7 treatments, T_1 occurring four times in each block and $T_2, \ldots, T_7$ each once. In this case let the digits 1, 8, 9, 10 represent T_1 and the digits 2, . . . , 7 represent $T_2, \ldots, T_7$ in order. The whole process of randomization is now analogous to that in Example 5.1, except that Table *A*.2, Random Permutations of 16, is used.
Thus if the first permutation is 7, 12, 1, 5, 16, 4, 11, 8, 2, 9, 10, 13, 15, 3, 6, 14, the numbers above 10 are rejected and the remainder replaced by the appropriate treatments to give

$$T_7\ T_1\ T_5\ T_4\ T_1\ T_2\ T_1\ T_1\ T_3\ T_6.$$

The randomization of Latin square designs is done by similar methods but since one or two special points are involved the discussion is postponed to Chapter 10.

A design that was only mentioned implicitly in Chapter 3 is the completely randomized arrangement, in which no grouping of the units is made and the treatments assigned at random subject only to the condition that each occurs the required number of times in all. The method, which is very simple and flexible, may be used in very small experiments, in order to get the maximum number of degrees of freedom for estimating error (Chapter 8), in experiments in which no reasonable grouping into blocks suggests itself, or when all attempts to increase precision are to be made by adjustment for concomitant variables.

Example 5.3. Consider the randomization of such an experiment over 21 units with three treatments each occurring seven times. There are several ways of proceeding; we cannot use the tables of random permutations in their simple form since more than sixteen objects are involved. (If sixteen or fewer were involved, we would write down a random permutation of the units and assign the first so many to T_1, etc.)

The following method is as quick as any.

(*a*) Number the units in any convenient way 00, 01, 02 through 20.

(*b*) Select haphazardly a starting point in the Table of Random Digits, Table *A*.3, and write out pairs of digits as they occur, subtracting 30 from two digit numbers from 30 to 59 and 60 from those that are between 60 and 89. Numbers 90 through 99 are rejected. Thus, 53 would be recorded as 53 − 30 = 23. If the starting point chosen is the block in row 24 and column 12 on the first page of Table *A*.3, the first numbers are 5, 10, 6, 10 and 5 again, omitted, 21, and so on. When this device is used it is essential to check that each of the final set of numbers, in this case 00 through 29, has equal chance of selection.

(*c*) The first seven numbers determine the units to receive T_1, the next seven are to receive T_2 and the remainder T_3.

The process can be modified in various ways; for example, when 5 have been selected the remaining units could be ordered by a random permutation of 16. Another method (Cochran and Cox, 1957, § 15.3) is to produce a random permutation of $1, \ldots, 21$ by writing under each of the 21 figures a 3-figure random digit. The arrangement of the numbers is then changed until the random digits are in order of increasing magnitude.

Other tricks will occur to the reader when he has some experience of the tables. The randomization of some of the more complicated designs will be described when we come to them.

The tables in the Appendix are given primarily for illustrative purposes. They may be used for the randomization of small experiments but on no account should they be used for experiments so large that the same permutation or digit would have to be employed twice. A more extensive table is necessary in such cases.

5.3 NATURE OF RANDOM NUMBERS AND RANDOMNESS

A table of random digits is a series of digits $0, \ldots, 9$ in which each occurs approximately equally frequently and in which there is no recognizable pattern. A recognizable pattern means, for example, a tendency for some digits to follow let us say a 5 more frequently than others. Some readers may feel that no amplification of this statement is called for, but there are in fact a few general points worth making, although the reader satisfied with the statement may omit this section. It is simpler to discuss tables of random digits, although the same remarks would, with minor changes, apply to tables of random permutations.

A completely random sequence of digits is a mathematical idealization in which we think of a mechanism capable of producing an infinite sequence of digits. The sequence is to conform completely to the mathematical laws of probability, as applied to a set of mutually independent events, each equally likely to be $0, \ldots, 9$. That is to say, if we make any calculation of probability connected with the sequence, e.g., the probability of five adjacent pairs 01 occurring in a block of fifty, then the resulting probability is to equal the proportional frequency of times with which this event occurs in the infinite sequence. The main properties of such a completely random series would be that

(*a*) each digit would occur equally frequently in the whole sequence;

(*b*) adjacent digits, or adjacent sets of digits, would be completely independent of one another, so that, for example, if we knew one digit, we would have no basis for predicting the next one;

(*c*) moderately long sections of the whole would show substantial regularity, e.g., the number of 1's in a set of 1000 digits would not deviate much from 100, and so on.

A table of random digits is a finite collection of digits which

(*a*) is produced by a process which it is reasonable to expect will give results closely approximating to the above mathematical idealization;

(*b*) has been tested to check that in several important respects, e.g., in the relative frequencies of 0's, 1's, . . . , and in the simple independence properties, it does behave as a finite section from a completely random series should.

The first conclusion from this is that randomness is a property of the table as a whole; thus to be accurate we should talk about permutations produced by a random method, rather than about random permutations, as if the individual permutations were random. Thus, any permutation of $1, \ldots, 12$ is a possible random permutation and any two such, for example

1	2	3	4	5	6	7	8	9	10	11	12		(*a*)
8	3	9	7	6	11	10	1	12	4	2	5		(*b*)

are equally likely to occur. Whether or not they are legitimate random permutations is to be decided by the properties of the methods by which they were produced and not by inspecting them as individuals.

This, of course, conflicts with the every-day usage of the word random, and there are two related reasons for this. First, if we come across (*a*) in an application, we can usually think of good physical hypotheses that will explain the precise ordering; a hypothesis that will explain (*b*) is

likely to be difficult to find. The second point is that the great majority of the permutations of 12 are disordered like (*b*) and not highly ordered like (*a*). These remarks have some bearing on the problem of the rejection of "unsatisfactory" randomizations (§ 5.7).

The second point to note from the general discussion is the independence of different numbers in the table. This is important where several randomizations have to be done in one experiment. Thus, we might have three Latin squares forming one experiment; we randomize these separately and then the random errors in the treatment estimates from the three squares are independent of one another. It would, of course, be wrong to randomize a Latin square and then to use it three times in the same form on each occasion. The reason is that if a very similar pattern of uncontrolled variation occurs in the three squares, the chance that it will produce a serious distortion in the treatment comparisons is much less if the squares are randomized independently than if a single common randomization is used.

5.4 JUSTIFICATION OF RANDOMIZATION

(i) Introduction

Having considered how to randomize, and briefly what a random series is, we must now consider why we randomize. Consider any of the designs we have discussed so far, for example, the randomized block, the Latin square, or one of its generalizations. When we have fixed the general type of design we are going to use, say a randomized block design with a certain number of blocks and treatments, we could determine the precise arrangement of treatments

(*a*) by adopting a particular systematic arrangement that seems unlikely to fit in with a pattern in the uncontrolled variation;
(*b*) by subjectively assigning the treatments in a way that seems haphazard;
(*c*) by randomization.

The dangers of (*a*) and (*b*) will be illustrated by examples.

(ii) Systematic Arrangements

Example 5.4. Greenberg (1951) has discussed an experiment in parasitology which illustrates the drawbacks of systematic arrangement. The experimental units were mice arranged in pairs of the same sex, one member of each pair receiving a series of stimulating injections, *T*, and the other member acting as a control (untreated, *U*). The observation consisted in challenging each mouse with 0.05 cc of solution, supposed to contain a standard number of larvae, and noting any response.

The point of the discussion depends on this. We have a series of pairs of mice *T*, *U*; *T*, *U*; *T*, *U*; ... In what order are the mice to be taken for inoculation? Greenberg reports that it was common to use the systematic order *TU*; *TU*; ... in the hope of cancelling variations in dosage. He produces data and experimental reasons, however, to show that during the course of the experiment, the number of larvae per injection increases steadily, and that therefore the ordering gives a systematically greater injection to the untreated mice. The consequences of this are:

(*a*) there is a systematic error in the estimated treatment effect, which would persist in a long experiment and even over several different experiments, if the above order were always used;
(*b*) the estimate of the error, based on the false hypothesis of the randomness of uncontrolled variations, is misleading.

When a systematic uncontrolled variation, such as this, is discovered in the experimental technique, it is, of course, important to take steps to eliminate the variation. However the aspect that concerns us now is the effect of such variations that we do not know about when the experiment is planned.

It is easy to be wise after the event and to say that a steady trend is a priori quite likely and that therefore some other pattern such as *TU*; *UT*; *TU*; *UT*; ... should have been used. However, whatever such pattern is chosen, there is the possibility that it coincides with some pattern in the uncontrolled variation, maybe one of obscure origin, producing a systematic error persisting even in a long experiment. To put the point another way, if a systematic arrangement of treatments is chosen, the presumption that it does not coincide with a pattern in the uncontrolled variation is a statement of the experimenter's opinion, which may well be justified, but which cannot be assessed quantitatively and which it is difficult for others to check on. If a surprising result is obtained, the experimenter may begin to doubt the validity of the systematic arrangement; if the results appear surprising to a later worker in the field, he will probably have no way of checking on the reasonableness of the pattern used.

Randomization, on the other hand, is an objective procedure, equally convincing to all and dealing equally with any pattern of uncontrolled variation that may present itself. The disadvantages of the systematic arrangements do not apply.

Example 5.5. When Latin squares were first introduced into experimental design, there was some discussion on whether a randomized square should be used or a square chosen deliberately for its balanced properties. An example of such a systematic square is the so-called knight's move 5 × 5 square,

A	*B*	*C*	*D*	*E*
D	*E*	*A*	*B*	*C*
B	*C*	*D*	*E*	*A*
E	*A*	*B*	*C*	*D*
C	*D*	*E*	*A*	*B*

which has the treatments evenly spread out with respect to the diagonals of the square.

The disadvantages of this design are less striking than those of the systematic arrangement of Example 5.4. However there would certainly be objections to repeating the design unchanged in a series of trials and also, as Yates (1951) has pointed out, difficulties could arise from the fact that in four cases out of five E, for example, is immediately to the right of D. Thus if the experiment is an agricultural field trial and D is a "tall" variety, the effect would be, with certain orientations of the square, to depress the yield of E.

However the main objection to the systematic design in this case is not the appreciable possibility of serious systematic error in the treatment estimates but the difficulty of estimating the amount of random error (Fisher, 1951, p. 74). In the discussion of the analysis of the randomized block and Latin square designs in Chapter 3, the principle was mentioned that whenever a source of uncontrolled variation is eliminated from the error in the design of the experiment, it must also be eliminated in the analysis, if a correct estimate of random error is to be obtained. Now in the systematic square some variation parallel to the diagonals of the square is eliminated in virtue of the balanced property of the design, but the ordinary analysis of the Latin square takes no account of this. Therefore a biased estimate of error is to be expected and Tedin (1931) confirmed this by examining uniformity data from field experiments. The bias he found was small, although there is, of course, no guarantee that this would always be so.

The bias in the estimate of error could probably be removed by a modification of the method of analysis, but several different, although plausible, ways of doing this are available and it is not clear which should be used, so that there is some loss of objectivity.

To sum up, the objection to lack of randomization in this case is mainly connected with the estimation of error. If the design is not randomized, and particularly if it is chosen for its "balanced" nature, it is quite likely that the estimate of error from the conventional method of statistical analysis will not be appropriate, even if it is very unlikely that there is appreciable systematic error in the treatment estimates themselves. This consideration of the correctness of the estimate of error applies also to Example 5.4, but there it is rather overshadowed by the occurrence of substantial bias in the treatment estimates themselves.

The conclusion from these two examples is that systematic arrangements suffer from the disadvantages that

(*a*) the arrangement of treatments may combine with a pattern in the uncontrolled variation to produce a systematic error in the estimated treatment effects, persisting over a long experiment or even over a series of experiments. We may begin by thinking this possibility sufficiently unlikely to be disregarded, but this is a matter of personal judgement which cannot be put on an objective basis;

(*b*) there is likely, even in the most favorable cases, to be difficulty connected with the estimation of error from such designs.

Randomization removes these disadvantages and hence is, other things being equal, to be preferred to systematization. That is, we aim, by

controlled grouping of the units (as in randomized blocks or the Latin square) to eliminate the effect of as much of the uncontrolled variation as possible and then to randomize the remainder.

It should not be thought, however, that these remarks mean that systematic designs are never to be tolerated. If we have good knowledge of the form of the uncontrolled variation and if a systematic arrangement is much easier to work with, as when the different treatments represent ordered changes of a machine, it may be right not to randomize. For example, suppose that there are at some stage of an experiment 24 test tubes of solution, to which are to be added x cc of reagent for treatment T_1, $2x$ cc for treatment T_2, and $3x$ cc for treatment T_3. Imagine further that this operation needs to be completed as quickly as possible. Clearly, the procedure that is quickest and least likely to lead to gross errors is to set the work out in systematic order, dealing first with all units receiving T_1, etc. If it is known that negligible error is involved in pipetting and if the whole set of 24 units can be completed in such a short time that the first and last tubes can be considered as dealt with simultaneously, it would be quite wrong to attempt randomization. If a number of repetitions are involved, a sensible precaution, however, would be to change the order of treatments for each repetition and, if practicable, to include some check on the assumptions.

Another reason for not randomizing is that in certain very special short experiments there is an appreciable gain in precision in using a special systematic arrangement (§ 14.2). What is important, however, is to randomize except when there is a very good reason not to, to understand that the conclusions from a nonrandomized experiment depend on the correctness of what is assumed about the uncontrolled variation, and to state this explicitly in reporting the experiment. The reader interested in a further discussion of systematic arrangements should read the papers by "Student" (1938), Yates (1939), and, for an account of some more recent work, Cox (1951). If a systematic design is adopted for an experiment to be repeated several times with the same design, the names of the treatments should be randomized independently in each repetition.

(iii) Subjective Assignment

We now consider the second alternative, the assignment of treatments not by strict randomization but in a subjective way that seems haphazard. The following is an example of an experiment spoiled by a procedure of this sort.

Example 5.6. In 1930 a very extensive experiment was carried out in the schools of Lanarkshire, in which 5000 school children received 3/4 pint of raw milk per day, 5000 received 3/4 pint pasteurized milk, and 10,000 children were selected as controls to receive no milk. The children were weighed and

measured for height at the beginning and end of the experiment, which lasted four months. The discussion below is based on "Student's" critique of the experiment ("Student," 1931).

The following method was used in determining, for each school, which children should receive milk and which not, only one type of milk being used in any one school. A division into two groups was made either by ballot or using an alphabetical system. If this appeared to give a group with an undue proportion of well-fed or ill-nourished children, others were substituted in order to obtain a more level selection. In other words a random, or nearly random, assignment of treatments was made and then "improved" by subjective assessment.

This resulted in the final observations on the control group exceeding those on the treated group by an amount equivalent to three-months' growth in weight and four-months' growth in height. The explanation of this is presumably that the teachers were unconsciously influenced by the greater need of the poorer children, and that this led to the substitution of too many ill-nourished among the feeders and too few among the controls.

This had a particularly serious effect on the comparisons of weight, because the children were weighed in their indoor clothes in February at the beginning of the experiment and in June at the end. Thus the difference in weight between their winter and summer clothing is subtracted from their actual increase in weight. Had the control and treated groups been random this difference in weight due to the clothing would have decreased the precision of the results but would not have introduced bias; however there was the suggestion that the treated group contained more poor children, who probably lost less weight from this cause, so that the experiment was biased.

Although it was possible to draw certain conclusions, these were of a very approximate and tentative nature, even though the number of children taking part was large. The failure of the experiment to yield clear-cut conclusions was due to the failure to adopt an impersonal procedure in allocating treatments to experimental units.

"Student" pointed out that a much more economical and precise way of comparing two treatments, say the two types of milk, would have been to work with pairs of identical twins, in a randomized paired comparison experiment (§ 3.2). Very probably a comparatively small number of such pairs would give high precision and it would then have been practical to make detailed and carefully controlled measurements on each child.

The conclusion from this is that an experiment is in danger of being very seriously affected if the personal judgement of people taking part is allowed to determine the allocation of treatments to units. There is abundant evidence that observer biases occur even in apparently unlikely circumstances, and moreover, even if the arrangement chosen is in fact satisfactory, there is always the suspicion that it may not be, and this will detract considerably from the cogency of the experiment if surprising conclusions are found. The time taken to carry through a process of objective randomization by the methods of § 5.2 is trivial under all ordinary circumstances, so that there is no argument for subjective assignment on the grounds of simplicity.

(iv) Randomization as a Device for Concealment

In most of the previous examples, randomization has been used to deal with variations in space, in time, between different animals or subjects, and so on to ensure that any patterns of variation that may exist in the experimental material cause no systematic error in treatment comparisons. A very important further use of randomization, however, is in situations where a substantial amount of the uncontrolled variation arises from subjective effects due to personal biases of the people taking part in the experiment, including the experimenter himself. In such applications randomization achieves its aim by concealing from the persons involved which treatment is applied to each unit.

Consider first an application where bias may enter the selection of units to take part in the experiment.

Example 5.7. Suppose that a clinical trial is set up to compare two or more methods (drugs, surgical treatments, etc.) of treating a disease. Experimental units, i.e., patients, are included in the experiment as suitable individuals appear at the various centers taking part. There will often be doubt as to whether a particular person should, in fact, be included.

If the doctor responsible for the decision about inclusion knows that the patient, if included, will receive say treatment A, this may easily influence, consciously or unconsciously, the decision reached in doubtful cases, and if this happens, the groups of experimental units receiving different treatments will not be genuinely comparable.

If the allotment of treatments is determined by a systematic pattern, this may soon become apparent; equally, if the order of treatments is determined by an initial randomization and if the full key is available to the doctor concerned, the necessary concealment will not be achieved. The satisfactory method is either to do the randomization after the patient has been chosen for inclusion, or to arrange that the treatment a particular patient is to receive is named in a sealed envelope that is not opened until after the patient has definitely been selected. The order of treatments in successive envelopes is randomized by the controller of the experiment and not revealed.

This device is now widely used in the design of clinical trials. Similar remarks apply to any experiment in which a subjective element enters into the selection of experimental units for inclusion in the experiment.

Randomization to achieve concealment may also be necessary in applying the treatments, particularly where the units are people likely to be influenced in an irrelevant way if they knew the treatment which they have actually received.

Example 5.8. Consider an experiment on school children to assess the effect of a new tooth paste, say one with an added fluoride. We need a group of children with whom to compare the children who receive the experimental tooth paste, F. A quite unsatisfactory way of obtaining such a control group would be to give F to half the children chosen by a randomized method and to give the other half no special treatment. In order to obtain worth-while results,

steps to encourage correct and frequent use of F would be necessary and any relative improvement in the experimental groups' teeth might well be due to the additional attention given to the cleaning of teeth rather than to the particular merits of F.

A better, but still unsatisfactory, procedure would be to issue the control group with a standard brand of tooth paste. The objection here is that the experimental group, knowing that they are receiving special treatment, may tend to be more diligent than the control group. The only satisfactory way of ensuring that no such effects occur, is to have identical tubes of control and experimental tooth pastes, so far as is possible differing only in the absence or presence of the special ingredient, and if possible indistinguishable in flavor, etc. The treatments are randomly assigned to the children, the key to the randomization being available only to the controller of the experiment. The children, their parents, and the staff responsible for instruction in the use of the tooth paste and for assessing the children's teeth at the end of the trial, should know neither which treatment any particular child has received, nor which groups of children have received the same treatment. The final observation on each child at the end of the experimental period would be some such index as the number of defective and missing teeth. This would also be determined for each child before the start of the experimental period and this initial value would enable adjusted treatment means to be calculated, eliminating from the error much of the variation connected with the initial state of the teeth (see § 4.3). It would also be possible to examine whether any difference between the treatments was more or less marked with children with good teeth.

These considerations are important in any experiment in which the application of one or more treatments may tend to be influenced by personal attitudes towards the treatments, these attitudes being considered irrelevant to the purpose of the experiment.* Thus in comparing a new and an old experimental technique or a modified and an unmodified industrial process, bias may arise due say to devoting greater attention to the running of the modified process. If, owing to the nature of the processes, this possible bias cannot be eliminated by concealment, whatever steps that may be practicable should be taken to remove such a bias. For example, the biases may tend to disappear if the experiment is spread over an appreciable time or is preceded by a practice period, or if the people taking part are deliberately misinformed as to the object of the investigation.

In some applications it may be required to conceal the nature of the treatment from the person who has to apply the treatments to a large number or all of the units. For example, one treatment might require an impure chemical, another an analytically pure source of the same substance. It would not in such cases be satisfactory to have two supplies one labelled A, and the other B, and to conceal which was which, since the necessary

* Note that for some purposes we might consider such attitudes as part of the treatments and would then not wish to eliminate their effect.

requirement of independence from unit to unit would not be satisfied. Thus a guess, possibly incorrect, might be made as to which treatment A was and a systematic error introduced. A better arrangement would be to have say at least six sources of material labelled A through F, three of which are impure, three not. For each experimental unit the source to use is given in the experimental instructions, having been determined by appropriate randomization.

The final stage in which concealment may be advisable is in the making of the observation itself. There are many fields where substantial personal biases may arise and in all these randomization of the order of presentation of the units for measurement is desirable. Thus in taste-testing experiments, where a judge is asked to state his preference among a number of products, it is most inadvisable that the judge should know the treatment to which each of the objects has been subjected. Again, in experiments in which the reproducibility of, say, an analytical technique is of interest, it is best for the analyst not to know which of the items submitted for analysis have received identical treatment. Quite generally, in any experiment in which personal judgement enters to a considerable extent into the determination of the final observation, concealment is desirable. Sometimes this is impracticable, but quite often randomization does achieve concealment in a simple and satisfactory way.

(v) Summing Up

To sum up, it seems fair to say that subjective allocation of treatments to units should never be used, because the method has serious disadvantages and no compensating advantages when compared with objective randomization. Of course, subjective allocation may work out perfectly well in some applications, but this is no argument for using it, since randomization is just as simple and has definite advantages.

Therefore, our general conclusion is that, with the minor exceptions noted in the discussion of systematic arrangements, randomization is to be preferred to alternative methods. This conclusion has been reached in a rather negative way by showing the disadvantages of other methods. In § 5.6 there is a brief discussion of the positive advantages of randomization from a more statistical point of view. That section may be omitted if desired. First, however, we deal with an important general matter concerned with the scheme of randomization to be used.

5.5 ERRORS ARISING IN SEVERAL STAGES

In many, if not most, experiments important uncontrolled variation may arise from several sources, notably in the experimental material, in

the various stages of applying the treatments, and in the taking of the observations. It is important that the randomization should cover all important sources of variation connected with the experimental units and that, so far as is practicable, the different experimental units receiving the same treatment should be dealt with separately and independently at all stages at which important errors may arise, one such stage being in the application of the treatments.

This point has already been discussed to some extent in Example 3.2. It will be dealt with now in more detail using a somewhat fictitious example connected with the shrinkage of socks. The example should be studied carefully because it can be paralleled in many fields, particularly in that type of laboratory work in which a whole sequence of operations has to be carried out on each batch of experimental material.

Example 5.9. In an experiment to compare four treatments applied to knitted socks to reduce shrinkage, something like the following might be done. Forty-eight socks are divided into 4 sets of 12, each set to receive one of the 4 treatments, say a control and 3 different chlorination processes. The treatments are applied and the socks measured. Normal wear and washing is then simulated in a controlled way in a machine that can take from 1 to 12 socks at a time. At the end the socks are remeasured and the percentage shrinkage calculated.

The treatment comparisons can be affected by (*a*) the variation of intrinsic properties from sock to sock, (*b*) measurement errors, (*c*) variations arising during the application of the chlorination processes, (*d*) lack of complete uniformity in the simulation of wear. We may decide, after investigation, that measurement errors can be treated as completely random and are in any case small compared with other sources of variation. If this is done, the measurements may be obtained in any convenient order. We shall consider several randomization procedures in the light of the remaining three sources of variation.

Method I. The socks are divided randomly into 4 sets of 12. Each set is processed as one batch and after measurement, dealt with in one run of the simulation machine. There are, thus, 4 runs of the simulation machine, each run dealing with socks that have all received the same treatment.

Method II. The socks are divided randomly into 4 sets as before, but the chlorination processes are applied independently to single socks, for example by including each sock in a separate batch for processing. After measurement the simulation of wear and washing is carried through as in Method I.

Method III. This is the same as Method II up to the simulation stage. Here the socks are grouped into blocks of 4, one from each treatment, each block being used for one run of the machine.

Method IV. The socks are divided into 12 sets of 4, 3 sets per process. The chlorination processes are applied independently to each set, so that 3 separate batches need to be run for each process. The runs of the simulation machine are arranged by Method III.

Method I is adequate only if negligible variation arises from sources (*c*) and (*d*), the chlorination and simulation stages. If, for example, there is some

variation in the performance of the simulation machine from run to run, this will appear as systematic error, since all the socks having one treatment are dealt with in a single run. If there happens to be appreciable variation between the conditions appertaining in different runs of the same chlorination process, the conclusions from Method I will apply solely to the particular runs of each process used and this will be a serious restriction. For example, it will be impossible from the experimental results alone to distinguish between real differences in the processes and variation from one application to another of one process.

Randomization is no help for treatment errors and the right procedure is to have independent applications of the treatment for each unit. This is Method II; any variation connected with the simulation process is still inadequately dealt with. Method III gives one way of dealing with this, by the randomized block principle, each run of the machine forming a block.

Method IV is a compromise version of Method III which meets the practical objection that would often be made that an independent run of a chlorination process for each sock is uneconomic. The method here is essentially to take experimental units consisting of 4 socks each and to set up a randomized block design of 3 blocks each with 4 treatments.

There are numerous further possibilities. Moreover, if the measurement process, instead of being fairly straightforward, involved a substantial subjective element, it would be necessary to measure the socks in randomized order, taking the sort of precaution discussed in the preceding section to conceal the treatment applied to the sock being measured. Of course, if this stage of randomization can be omitted, the work of measurement may be much simplified.

This discussion can be summarized as follows. Variation may arise from several sources, and randomization should cover all those at which the variation cannot be assumed negligible or completely random. It is frequently not good enough to randomize just at one stage of the experimental procedure and to leave the treatments systematically arrayed at other stages.

5.6 STATISTICAL DISCUSSION OF RANDOMIZATION

In this section, which may be omitted at a first reading, the statistical consequences of randomization are discussed. We start from the assumption,* § 2.2, equation (1), which says that the observation obtained when a particular treatment is applied to a particular experimental unit is

$$\begin{pmatrix}\text{a quantity}\\ \text{depending only}\\ \text{on the unit}\end{pmatrix} + \begin{pmatrix}\text{a quantity}\\ \text{depending on the}\\ \text{treatment}\end{pmatrix}. \qquad (1)$$

Consider for definiteness a randomized block experiment, although the

* An analysis based on a more general assumption, allowing variations in treatment effect from unit to unit, has been made by Wilk and Kempthorne (1956).

following remarks apply with minor changes to nearly all the designs described in this book.

If we randomize the treatments within blocks, the unit quantities associated with a particular treatment T_1 consist of a random sample of one from the set of unit quantities for the first block, a random sample of one from the unit quantities for the second block, and so on. Similarly for the other treatments, the only complication being that the samples for T_2, T_3, . . . are drawn "without replacement," since no unit receives more than one treatment. Hence we may apply the mathematical theory of random sampling to the behavior of our observations, and moreover the theory is rigorously applicable, provided that the assumption (1) holds and that the table of random numbers, or random permutations, used in randomizing the treatments, is adequate. The latter point need cause us no trouble.

In this way we reach the following conclusions, without further assumptions about the nature of the uncontrolled variation.

(*a*) The estimated treatment effects are *unbiased*, in the sense that the average of the estimates over a large number of independent repetitions of the experiment would be equal to the true treatment effects defined from (1).

(*b*) In a single experiment with a fixed amount of uncontrolled variation, as measured by the standard deviation, the error in the estimated treatment effects would almost certainly be very small if the number of units were sufficiently large. That is, there is a negligible chance of appreciable error persisting in a very long experiment; this state of affairs should be contrasted with the situation for a nonrandomized experiment, such as Example 5.6, where there was appreciable systematic error even though the number of units was very large.

(*c*) The square of the standard error of the estimated treatment effects, calculated by the method described in § 3.3, is unbiased, in the sense that, averaged over a number of independent repetitions of the experiment, it would equal the average of the square of the actual error, i.e., estimated effect minus true effect, squared.

(*d*) In principle it is possible (Fisher, 1951, p. 43) to make exact significance tests concerning the treatment effects and to calculate limits within which the true effects lie at any assigned level of probability. Thus we can build up a distribution, inferred from the data, for the magnitude of the true treatment effect. In practice these calculations are almost always done, not by the "exact" method, but by introducing certain assumptions about the shape of the distribution of the unit quantities in (1). This enables the significance calculations to be made very simply by

the t test and related methods. It is known that, except in small experiments, results obtained in this way agree satisfactorily with those based on the "exact" argument. In any case the assumptions about the uncontrolled variation concern the shape of the overall distribution of the unit quantities and not the nonexistence of patterns.

To return to a less statistical description, the positive advantages of randomization are assurances

(*a*) that in a large experiment it is very unlikely that the estimated treatment effects will be appreciably in error. In other words a randomized experiment may be more accurate than a corresponding nonrandomized one in which an unskilful assignment of treatments to units has led to systematic bias. Randomization achieves this mechanically;

(*b*) that the random error of the estimated treatment effects can be measured and their level of statistical significance examined, taking into account all possible forms of uncontrolled variation subject to (1).

Thus, to take a simple case to illustrate (*b*), we might conclude from a randomized experiment that there is a difference between two treatments that is statistically significant at a very high level. The corresponding conclusion for an experiment laid out in a systematic arrangement might be that the difference is very unlikely to be due to random uncontrolled variation (this is shown by the significance test) and that it is considered very improbable that the systematic arrangement is responsible for the apparent effect. This last statement has no measurable uncertainty, nor is there any guarantee that the standard error and significance test measure anything very relevant about the system. It is not that the systematic arrangement is necessarily less precise than the randomized one, but that the assessment of the results is on a less objective basis.

One or both points (*a*) and (*b*) may apply in any particular case.

This concludes the general discussion of the arguments for randomization in the allotment of treatments to experimental units. There is a second very important use for randomization in experimental work, namely in sampling, i.e., in selecting from a given bulk a portion for detailed study and measurement, the portion to be representative of the whole. The arguments for randomization in sampling are parallel to those developed above, but will not be discussed here.

5.7 SOME FURTHER POINTS

There are some difficulties that arise in the application of randomization, particularly to small experiments, and these will now be discussed.

The first point concerns the rejection of an arrangement produced by

the randomization when it seems particularly unsuitable. As an example, consider the paired comparison experiment, Example 3.1, with eight pairs of units. Suppose that, as in our first account of this experiment, the units are arranged in a definite order within each pair, but that it is decided that this ordering is not of sufficient importance to warrant balancing it in the design of the experiment by the method of Example 3.10. Now it will happen, actually about once in 128 times in the long run, that the ordering of treatments is the same for every pair, either $T_1 T_2$ every time or $T_2 T_1$ every time. Further, once in about 14 times the arrangement is either of this type or has just one pair showing a different ordering from the remaining 7.

It is clearly undesirable to use these arrangements. Even though we think that there is probably not an important order effect, there are likely to be various things, connected say with the experimental technique, that could produce such an effect. In other words a pattern of uncontrolled variation with a substantial systematic difference between the first and second unit in the pair, is a priori considerably more probable than other *particular* patterns we can think of.

Similar considerations apply in other experiments where the randomization produces an arrangement that fits in with some physically meaningful pattern in the experimental material, even though this pattern is thought probably unimportant. Other examples are if a Latin square on randomization has a line of treatment T_1, say, down a diagonal, or if a randomized block experiment gives the same order of treatments within each block. The chances of these particular arrangements occurring are extremely small, except in experiments with a small total number of units.

There are three ways of dealing with the difficulty, all depending on curtailing the randomization. The first method is to incorporate a condition about order into the formal design of the experiment, as was done in Example 3.10, where T_1 and T_2 each occurred four times in the first position. This is probably the best solution in the present case, but it is certainly not a general answer to the problem, since there are various reasons why it may be impracticable or undesirable to introduce further constraints into the design. For example we lose degrees of freedom for residual in eliminating a source of variation that is probably not important, we make the experiment more complicated and there may already be several different systems of grouping in the design, making the introduction of further conditions difficult or impossible.

The second method is to reject extreme arrangements whenever they occur, i.e., to rerandomize. For example in the paired comparison experiment, we may decide to reject all arrangements with seven or more pairs in the same order. A highly desirable condition in using this method, if observer biases like those of Example 5.6 are to be avoided, is that if any arrangement is to be rejected, so must all other arrangements obtained by permuting the names of the treatments. Thus if the arrangement with eight $T_1 T_2$'s is rejected, so must the arrangement with eight $T_2 T_1$'s. There would be little likelihood of disagreement over such an extreme case, but since the decision as to what arrangements to regard as unsatisfactory is arbitrary, there could be disagreement with less extreme cases. The best plan is, if possible, to decide which arrangements are to be rejected before randomization. It is difficult to give general advice about which arrangements to reject, but the best rule is probably to have no hesitation in rejecting any arrangement that seems on general common-sense grounds to be unsatisfactory. Fortunately this matter is not nearly so important in practice as might be thought, since, as remarked above, extreme arrangements occur with appreciable chance only in very small experiments.

The third method is to use a special device, known technically as restricted randomization (Grundy and Healy, 1951; Youden, 1958). This is a very ingenious idea, in which a design is selected at random from a very special set of arrangements. The set is chosen to exclude both the extreme arrangements and the very balanced arrangements, in such a way that the full mathematical consequences of ordinary randomization follow. The method is probably of most value for a special design called the quasi-Latin square (Chapter 12), for which the method was first introduced, and otherwise in a series of small experiments, each of some interest in itself, but which also need to be considered collectively. The method is however too specialized to discuss here and its full implications have not yet been worked out; the nonstatistical reader requiring more information about it should consult a statistician.

The reader may object that the second method, the rejection of extreme arrangements, will falsify the mathematical consequences of randomization described in § 5.6. This is true of the estimation of error, although not of the absence of bias in the treatment estimates themselves. The estimate of error will only be unbiased if there is in fact no systematic order effect. However in single small experiments the estimate of error is very inaccurate anyway. More importantly we have here a mathematical interpretation of randomization: that it leads to desirable properties in the long run, or on the average, and on the other hand a practical problem—namely the designing and drawing of useful conclusions from a particular single experiment that we are now in the process of considering. Usually the concept that our procedures will work out well in the long run is a very helpful one, both qualitatively and in giving a vivid physical picture of the meaning of probabilities calculated in connection with a

particular experiment. However to adopt arrangements that we suspect are bad, simply because things will be all right in the long run, is to force our behavior into the Procrustean bed of a mathematical theory. Our object is the design of individual experiments that will work well: good long-run properties are concepts that help us in doing this, but the exact fulfillment of long-run mathematical conditions is not the ultimate aim.

The second general matter is closely related to the first. Suppose that we design and carry out a randomized experiment, and that when we come to analyze and interpret the results we realize either that the arrangement we have used is probably an unfortunate one and should have been rejected, or, by inspection of the results, that there is some particular form of uncontrolled variation. For example, we might have the above paired comparison experiment with, say, six pairs receiving the order $T_1 T_2$ and two receiving the order $T_2 T_1$. Inspection of the results may suggest a substantial order effect comparable to the treatment effect. Another example would be if an agricultural field trial arranged in randomized blocks shows a systematic trend from one end to the other of the experimental area. What do we do in such situations?

In some cases, possibly in the first, we may decide that the data should be regarded with suspicion. Suppose, however, that we do wish to draw what conclusions we can. The previous discussion shows that it is not good enough to say that the long-run properties are valid whatever the form of the uncontrolled variation and on those grounds to analyze the experimental results by the usual methods. On the other hand, to introduce modifications into the analysis based on inspection of the results and on personal judgement about the design must lead to some loss of objectivity. The following procedure is suggested.

(*a*) Work through the conventional analysis of the observations ignoring the suspected complication.

(*b*) Make a special statistical analysis of the observations taking account of the complication in whatever seems the most reasonable way. The reader who is not familiar with fairly advanced statistical methods will probably need statistical advice in this. The method will usually involve the analysis of what is known technically as a nonorthogonal least-squares situation.

(*c*) If the conclusions of the two analyses are for practical purposes equivalent there is no difficulty. If the conclusions do differ, care is needed. The assumptions underlying the second analysis should be carefully thought over, and if they seem reasonable, the second analysis should be regarded as correct.

(*d*) In reporting on the experiment, conclusions from both analyses

should be given, at any rate briefly. If the first analysis is rejected, reasons should be outlined. The general idea should be to make it clear to the reader what has been done and to give him the opportunity of forming his own conclusions as far as practicable.

Fortunately these difficulties tend to occur infrequently in practice.

Another difficulty that occasionally arises is that there is some practical reason why certain treatment arrangements are not allowable. One example arises in raspberry variety trials (Taylor, 1950). The point here is that additional canes spring up near many of the canes originally planted and it is necessary to remove these new canes from each plot. For this to be possible varieties that resemble each other closely must not occur close together, thus restricting the randomization. Another example occurs in carpet wearing trials, in which dyed and undyed carpets are under comparison. An experimental carpet is formed by sewing together squares of carpet of different types and the whole carpet placed say in a busy corridor. It would often be desirable that the carpet should look presentable and this would preclude full randomization of the dyed and undyed sections. The procedure in such cases is either to do as much randomization as possible or to use a systematic arrangement taking whatever steps are practicable to avoid bias.

SUMMARY

When a particular type of design, say a Latin square, has been chosen as likely to give precise treatment comparisons, the arrangement of treatments should be determined by impersonal randomization. This is done by shuffling cards, etc. or, much more usually, by tables of random permutations or of random digits.

Systematic arrangement is very occasionally to be preferred to randomization, for example on the grounds of simplicity; subjective assignment of treatments in a haphazard way should never be done. The justification for randomization is that it makes the chance negligible that systematic differences between units receiving different treatments will persist in a long experiment and that it enables the error to be estimated whatever the form of the uncontrolled variation. In effect the randomization rearranges the experimental units into random order and converts uncontrolled variation of whatever pattern into completely random variation. It is very important that randomization should cover all stages at which major errors may arise.

Care is needed, particularly in very small experiments, whenever unsatisfactory arrangements are produced in the randomization.

REFERENCES

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

Cox, D. R. (1951). Some recent work on systematic experimental designs. *J. R. Statist. Soc.*, B, **14,** 211.

Fisher, R. A. (1951). *The design of experiments.* 6th ed. Edinburgh: Oliver & Boyd.

Greenberg, B. G. (1951). Why randomize? *Biometrics,* **7,** 309.

Grundy, P. M., and M. J. R. Healy. (1951). Restricted randomization and quasi-Latin squares. *J. R. Statist. Soc.*, B, **12,** 286.

Kendall, M. G., and B. Babington Smith. (1939). *Tables of random sampling numbers.* Tracts for computers, No. XXIV. Cambridge: Cambridge University Press.

"Student". (1931). The Lanarkshire milk experiment. *Biometrika,* **23,** 398. Reprinted in "*Student's*" *Collected Papers.* Cambridge, 1942.

——— (1938). Comparison between balanced and random arrangements of field plots. *Biometrika,* **29,** 363. Reprinted in "*Student's*" *Collected Papers.* Cambridge, 1942.

Taylor, J. (1950). A valid restriction of randomization for certain field experiments. *J. Agric. Sci.*, **39,** 303.

Tedin, O. (1931). The influence of systematic plot arrangement upon the estimate of error in field experiments. *J. Agric. Sci.*, **21,** 191.

Wilk, M. B., and O. Kempthorne. (1956). Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. Math. Statist.*, **26,** 950.

Yates, F. (1939). The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments. *Biometrika,* **30,** 440.

——— (1951). Bases logiques de la planification des experiences. *Ann. de l'Institut H. Poincaré,* **12,** 97.

Youden, W. J. (1958). Randomization and experimentation. *Ann. Math. Statist.* To appear.