CHAPTER 3

Designs for the Reduction of Error

3.1 INTRODUCTION

In this chapter we consider some ways of reducing the effect of uncontrolled variations on the error of the treatment comparisons. The general idea is the common sense one of grouping the units into sets, all the units in a set being as alike as possible, and then assigning the treatments so that each occurs once in each set. All comparisons are then made within sets of similar units. The success of the method in reducing error depends on using general knowledge of the experimental material to make an appropriate grouping of the units into sets. This method, and various generalizations of it, will be introduced mainly by examples.

3.2 PAIRED COMPARISONS

We begin by considering experiments for the comparison of just two treatments.

Example 3.1. Fertig and Heller (1950) have discussed an experiment for comparing the effect on sewage of two treatments, T_1 and T_2. Both treatments involved 100 per cent chlorination; with T_2 there was no special mixing and with T_1 there was an initial 15-sec period of rapid mixing. The observation made on each unit after processing was the logarithm of the coliform density per ml, and it was required to estimate any additional reduction in coliform density due to the rapid mixing in treatment T_1.

The main source of uncontrolled variation, other than random sampling errors in the determination of the coliform density, arose from variations in the sewage before processing. Therefore to obtain pairs of units as alike as possible, it was natural to take batches of sewage on the same day and as close together as possible in time. This was done on several days giving a series of pairs of similar units and it was then arranged that T_1 and T_2 both occur on each pair. This involves a series of choices between the orders $T_1 T_2$ and $T_2 T_1$. In the present case there was no reason for expecting a systematic difference between the first and second units in the pairs and the appropriate procedure is then to *randomize* the order of the treatments, i.e., to use an objective device such as a table of random numbers to choose, independently for each pair, between $T_1 T_2$

and $T_2\,T_1$, giving each equal probability. The full discussion of this process of randomization is deferred to Chapter 5.

A typical arrangement of treatments resulting from such a randomization is shown in Table 3.1 together with some fictitious observations. For each pair of units the difference between the observation on T_2 and the observation on T_1 is calculated. The treatment effect is estimated by $\bar{d}$, the mean of these differences, and the estimated standard error of $\bar{d}$, and a test of the statistical significance of $\bar{d}$ can be obtained by simple standard statistical calculations (Goulden, 1952, p. 51), the amount of the uncontrolled variation being estimated from the observed dispersion of the differences in the last column of Table 3.1.

TABLE 3.1

PAIRED COMPARISON EXPERIMENT

Day	First Unit	Second Unit	Difference, d
1	T_1:2.8	T_2:3.2	0.4
2	T_2:3.1	T_1:3.1	0.0
3	T_2:3.4	T_1:2.9	0.5
4	T_1:3.0	T_2:3.5	0.5
5	T_2:2.7	T_1:2.4	0.3
6	T_2:2.9	T_1:3.0	−0.1
7	T_2:3.5	T_1:3.2	0.3
8	T_1:2.6	T_2:2.8	0.2
		Mean, $\bar{d}$ =	0.262
		Estimated standard error =	0.078

It is clear in this design that the variation from one day to another has no effect on the experiment, i.e., that if both the observations on one day are changed by the same amount, the estimate of the treatment difference, and its error, are unaffected. The elimination in this way of the effect of part of the uncontrolled variation is, of course, the object of the pairing of the units.

Notice that we take differences between observations which are themselves logarithms. This implies the assumption that T_1 achieves a constant *fractional* change in coliform density over what would be obtained on the same material with T_2.

A natural objection to the randomization used in obtaining the design in Table 3.1 is that T_2 has occurred five times in the first position and three times in the second, and that it would have been better to have arranged for each treatment to occur equally often in each column. This point will be discussed fully later, but in the meantime it should be noted that the objection is really only cogent if there is reason to expect a systematic difference between the first and second units and this was not so in this experiment.

This example could be paralleled from many applied fields. The general method is simply to obtain a number of pairs of experimental units, where the two units in each pair are expected to give as nearly as

possible identical observations in the absence of treatment differences. The treatments T_1 and T_2 are then assigned in random order to each pair. The method will give a comparison of treatments free of systematic error whatever pairing of units is used, but the success of the method in reducing error depends on a skilful grouping of units.

The following are a few examples of methods that can be used to obtain a suitable pairing. Often, as in Example 3.1, the general fact that experimental units close together in some natural arrangement in space or time will tend to be alike suggests an appropriate pairing. Thus, plots close together in a field tend to give yields more alike than plots far apart, the products from one machine at two times nearer together tend to be more alike than products at times far apart or than products from different machines, and so on. If more explicit information is available about differences between units, this should, of course, be used. In experiments with rats, the pairs would probably be taken of the same sex, of approximately the same weight, and, so far as possible, from the same litter. In some experiments on animals it is possible to use twins, especially identical twins, for the pairs. In other animal work it is possible to use paired organs (kidneys, eyes, etc.) from the same animal. A somewhat similar idea in plant experimentation has been put forward by James (1948); he split clover plants through the middle of the tap root and used the two halves as paired units. Another device that is sometimes valuable is the use of the same physical object as a unit twice. This frequently happens in experimental psychology and in those clinical experiments in which the treatments are of a comparatively minor nature, so that each subject can be treated more than once. In such experiments, it may happen, even if there are no complications due to the overlap of treatment effects, that there is a systematic difference between the first and second units. In this case some restriction on the randomization is desirable to balance out the systematic difference; this will be discussed later.

The use of inbred lines of animals or plants has often been advocated in biological work as a method of ensuring uniform material. This use has been questioned, for example by Biggers and Claringbold (1954), who give examples where inbred lines are not more homogeneous than randomly bred material. They suggest that F_1 hybrids between inbred lines may be more suitable than the inbred lines themselves.

A final method depends on using a supplementary observation made on each unit before the experiment starts. For example, in an experiment on animals, the supplementary observation could be the initial weight. In this case the two animals with lowest weight are put in one pair, the two with the next lowest weight in the next pair, and so on. Provided

that animals with extreme weights are omitted and that the final observation is highly correlated with initial weight, this provides a satisfactory grouping into pairs. The methods to be used if two or more supplementary measurements are available will be dealt with later.

One general warning is necessary in connection with the use of artificially uniform material. It may, by the use of such material, be possible to obtain a substantial increase in precision, but sometimes only at the cost of getting conclusions that are not representative of a wider class of units (see also § 9.2). What should be done in such cases depends on the purpose of the investigation; for example, if it is desired to obtain conclusions of immediate practical applicability in industry or agriculture, it will be desirable to use representative material.

3.3 RANDOMIZED BLOCKS

(i) Introduction and Example

If we have more than two treatments to be compared, the method just described can be extended in a straightforward way. If there are t alternative treatments, we group the units into sets of t, the units in each set being expected to give as nearly as possible the same observation if the treatments are equivalent in their effect. It is usual to call each set of t units a *block*. The order of treatments is then independently randomized within each block, arranging that each treatment occurs once in each block. Just as in § 3.2 the effect of variations between pairs is eliminated, so in the present case the effect of variations between blocks is eliminated, so far as treatment comparisons are concerned.

An experiment in which block differences are removed from the error in the way just described is said to be arranged in *randomized blocks*.

Example 3.2. In an experiment discussed by Cochran and Cox* (1957, § 4.23), the treatments were five levels of application of potash, 36, 54, 72, 108, and 144 lb K_2O per acre applied to a cotton crop. One observation analyzed was a measure of single-fiber strength in arbitrary units, obtained as an average of a number of tests on the cotton from each plot.

There were three blocks each containing five plots. The observations are given in the above reference but not full particulars of the arrangement of treatments within blocks, etc. In accordance with the general principle for grouping plots into blocks, the five plots in a block should be chosen to minimize the uncontrolled variation from plot to plot within blocks, and this is usually achieved by arranging the plots within a block in a compact approximately square area. This and the randomization of treatments within blocks allows statistical assessment of uncontrolled variation in the results arising from

* This book is referred to frequently. Section rather than page numbers are given because the section numbering is the same in both editions.

variation between plots. But this is certainly not the only way error can arise; three other possible sources of erratic variation are associated with

(*a*) the cultivation and harvesting of the crop;
(*b*) the selection of fibers for test;
(*c*) the strength testing;

and these will be discussed briefly.

Variations connected with the order in which the plots are cultivated or harvested would ordinarily be assumed negligible; however, if for instance the harvesting takes more than one day a useful precaution would be to harvest all the plots in one block on the same day. In this way constant differences between days become identified with block differences and do not contribute to the error of the experiment.

Only a minute proportion of the fibers on a plot are used in the strength testing; the use of a reliable method of sampling in selecting the fibers is a vital part of the method of testing but will not be discussed here.

There may be uncontrolled variations connected with the behavior of the testing machine, with the temperature or humidity of the testing room, and with the testing operative. The best procedure here is usually to test the cotton from one block in random order in as short a time period as possible. If several operatives or several testing machines are used in the whole experiment it is usually desirable that the results for each block should be obtained by one operative on one machine, i.e., possible differences within blocks that could arise from operative or machine differences should be eliminated.

To sum up, at each stage of the experiment, from the initial planting to the final testing, sources of uncontrolled variations are either identified with blocks and in effect eliminated from the treatment comparisons, or randomized, or possibly assumed negligible. The last course is avoided as far as possible, since, as discussed in Chapter 1, it is usually best to avoid assumptions about the nature of the uncontrolled variation.

The observations are given in Table 3.2(*a*); the five treatments have been denoted in order of increasing amount of K_2O, $T_1, \ldots, T_5$. (The detailed arrangement of treatments within blocks as shown has been obtained by randomizing the values given in Cochran and Cox and is presumably not the order actually used in the experiment.)

To analyze the observations* they are first rearranged as in Table 3.2(*b*) and the totals and means for each treatment (and block) calculated. Thus for the first treatment $7.62 + 8.00 + 7.93 = 23.55$, and this divided by 3 gives the treatment mean of 7.85. The differences between treatment means are the best estimates of the true treatment differences, provided that the basic assumption of Chapter 2 holds and that the amount of the uncontrolled variation does not vary appreciably from block to block.

To estimate the precision of these estimates we use formula (1) of § 1.2, i.e.,

$$\begin{pmatrix}\text{standard error of}\\ \text{difference of two}\\ \text{means of 3}\\ \text{observations each}\end{pmatrix} = \sqrt{\frac{2}{3}} \times \text{standard deviation.} \qquad (1)$$

* The following account of the analysis may be omitted at a first reading.

TABLE 3.2

EXAMPLE OF RANDOMIZED BLOCK EXPERIMENT

(*a*) *Original Design and Observations*

Block 1	T_5:7.46	T_4:7.17	T_1:7.62	T_2:8.14	T_3:7.76
Block 2	T_2:8.15	T_1:8.00	T_5:7.68	T_4:7.57	T_3:7.73
Block 3	T_3:7.74	T_2:7.87	T_1:7.93	T_4:7.80	T_5:7.21

(*b*) *Rearranged Observations*

	T_1	T_2	T_3	T_4	T_5	Total	Mean
Block 1	7.62	8.14	7.76	7.17	7.46	38.15	7.63
Block 2	8.00	8.15	7.73	7.57	7.68	39.13	7.83
Block 3	7.93	7.87	7.74	7.80	7.21	38.55	7.71
Total	23.55	24.16	23.23	22.54	22.35	115.83	7.72
Mean	7.85	8.05	7.74	7.51	7.45	7.72	

(*c*) *Residuals*

	T_1	T_2	T_3	T_4	T_5
Block 1	−0.14	0.18	0.11	−0.25	0.10
Block 2	0.04	−0.01	−0.12	−0.05	0.12
Block 3	0.09	−0.17	0.01	0.30	−0.23

Estimate of standard deviation $= \sqrt{(0.3496/8)} = 0.2090$ (8 degrees of freedom)
Standard error of difference between two treatment means $= 0.2090\sqrt{(2/3)} = 0.171$
Estimated increase in strength per 18 lb K_2O increment is −0.090 with standard error 0.0251

We have first to estimate the standard deviation, that is the amount of uncontrolled variation from unit to unit. This is usually done by an elegant technique called *analysis of variance*; its application to the present problem has been described in full by Cochran and Cox and general accounts of the method will be found in any textbook on statistical methods.

It is, however, worth indicating briefly an equivalent method for estimating the standard deviation which, while rather inconvenient numerically, does indicate the physical basis for the estimate. We require to measure that part of the variation that is not due to real treatment effects and that cannot be regarded as systematic variation between blocks. Therefore it is natural first to express each observation as a difference from the overall mean and then to remove the variation accounted for by block differences. This is done by subtracting

$$\begin{pmatrix}\text{mean observation for}\\ \text{the particular block}\end{pmatrix} - \begin{pmatrix}\text{overall}\\ \text{mean}\end{pmatrix}.$$

Next the variation accounted for by treatments is removed by subtracting

$$\begin{pmatrix}\text{mean observation for}\\ \text{the particular treatment}\end{pmatrix} - \begin{pmatrix}\text{overall}\\ \text{mean}\end{pmatrix}.$$

At the end of this process we get, corresponding to each original observation, a *residual*, which may be defined directly as

$$\text{observation} - \begin{pmatrix}\text{mean observation}\\ \text{for the}\\ \text{particular block}\end{pmatrix} - \begin{pmatrix}\text{mean observation}\\ \text{for the}\\ \text{particular treatment}\end{pmatrix} + \begin{pmatrix}\text{overall}\\ \text{mean}\end{pmatrix}. \quad (2)$$

These are given in Table 3.2(*c*). Thus for the first observation we have that $7.62 - 7.63 - 7.85 + 7.72 = -0.14$. Except for rounding-off errors, the residuals add up to zero for each block and for each treatment.

The standard deviation measures the magnitude of the residuals and is calculated by finding the average of the squared residuals and then square-rooting the answer. However, in averaging the squared residuals it turns out to be appropriate to divide not by the number of residuals (15) but by what are called the *residual degrees of freedom*, (number of blocks − 1) × (number of treatments −1), which in this case is 8. The reason for this is essentially that if the 8 residuals in the upper left hand section of Table 3.2(*c*) were assigned arbitrarily, the condition that the row and column sums must be zero would determine the remaining residuals uniquely; i.e., effectively there are 8 *independent* residuals. Thus, the required estimate of standard deviation is

$$\sqrt{\{\tfrac{1}{8}[(-0.14)^2 + (0.18)^2 + \cdots + (-0.23)^2]\}} = 0.2090,$$

and is said to have 8 degrees of freedom. This is exactly the value that is given more quickly by the analysis of variance; the detailed table of residuals is, however, very useful if it is required to check the assumptions underlying the analysis. For example the occurrence of a single very large residual suggests that the corresponding observation may be suspect, whereas the distribution of the residuals gives information about the frequency distribution of error. It may sometimes happen that some blocks are much more variable than others and in extreme cases this too can be detected from the residuals, although considerable caution is needed in doing this. Important work on the examination of residuals has been done by F. J. Anscombe and J. W. Tukey; their work had not been published when this book went to press.

We now use formula (1) to obtain the estimated standard error of the mean as $0.2090 \times \sqrt{(2/3)} = 0.171$. In accordance with the account of the standard error given in § 1.2, the interpretation of this figure is that, for example, there is only a chance of about 1 in 20 that the estimate of a single preselected effect is in error by more than $\pm 2 \times 0.171 = \pm 0.342$. However, as explained in § 1.2, this interpretation needs some modification when the standard deviation is itself only estimated from a small number of observations, and in fact the residual degrees of freedom determine what this modification should be. The 1 in 20 limits for 8 degrees of freedom are increased to 2.31 × standard error, i.e., to plus or minus 0.395. This increase from 2 to 2.31 to allow for the uncertainty in the estimate of error is explained in textbooks on statistics and is an example of the use of what is known as "Student's" t distribution. Further modification of the multipliers is needed if they are applied solely to differences suggested by the data, such as to the difference between the treatments with highest and lowest mean responses.

The essential general points in this calculation are first the estimation of the treatment effects by a straightforward process of averaging, and second the estimation of the variation of the observations when treatment and block differences are removed. The important principle here, which applies also to

more complicated cases, is that when the effect of a source of variation, for example block differences, is eliminated in the planning of the experiment, it must also be eliminated in the analysis, if an appropriate measure of error is to be obtained.

In this particular experiment the five treatments bear a special relation to one another in that they represent different levels of a continuous quantity, the amount of K_2O per acre. It is, therefore, natural to consider not just the differences between different treatments, but also the curve of mean strength against the amount of K_2O and, in particular, to see whether this curve is effectively a straight line. Standard statistical methods of regression analysis (Goulden, 1952, p. 102) can be used to show that the curve does not depart significantly from a straight line representing a decrease in strength of 0.090 per 18 lb K_2O per acre increment. The standard error of the slope is 0.0251.

Finally it is often worth examining the block means, even though they do not bear directly on the estimation of treatment effects. First, it is possible to assess from the magnitude of the differences between blocks whether the grouping of the units into blocks has appreciably reduced the error and this information may be useful in planning similar experiments in the future. Second, particularly in experiments with more blocks than the present one, a detailed examination of the block differences may be very helpful. For example, if two operatives had been used in the strength testing, a comparison of block means to see whether there is evidence of a systematic difference between operatives might be interesting. No very cogent conclusions would normally be drawn because of the difficulty of disentangling operative differences from other sources of block variation. (However, we shall later consider *split plot experiments*, which are essentially randomized block experiments in which a further set of treatments are applied to whole blocks, and in these reliable conclusions can be drawn from block differences.) It must be repeated that the examination of block means tells us nothing about the treatment effects and is of interest only in adding to the general knowledge of the experimental material.

(ii) Missing Values

The relatively simple analysis just described depends in an essential way on the balanced nature of the randomized block design. For example, it is only because each treatment occurs the same number of times within each block that the mean observations on the treatments can be used to compare treatments in a way unaffected by constant block differences. If, for instance, treatment T_1 did not occur in the first block and if the first block happened to give systematically high results, the mean for T_1 would be depressed relative to the means for the treatments that did occur in the first block, so that the treatment means would no longer give a fair basis for comparing treatments, free of block effects.

It can happen, particularly in experiments with many units, that the results on one or more units are lost, do not become available, or have to be discarded. For instance, if the units are animals, some may die from causes unconnected with the treatments. This loss will destroy the property of balance, i.e., the pattern of observations will no longer be that of a randomized block design.

A special case of a general principle, called the method of least squares, can be used for the efficient analysis of observations grouped into blocks and subject to treatments arranged in an unbalanced scheme, but the calculations tend to be complicated. Fortunately a very simple method is available when observations are missing for only one unit. This is to calculate a so-called estimated missing value by the formula

$$(kB + tT - G)/[(k - 1)(t - 1)],$$

where k is the number of blocks, t is the number of treatments, B is the total of all remaining observations in the block containing the missing observation, T is the total of observations on the missing treatment, and G is the grand total.

Then we analyze by the straightforward method, just as if the estimated missing value is a genuine observation. A small modification is that the degrees of freedom for residual are reduced by one. This procedure gives the same estimated treatment effects as the method of least squares and also the same estimated standard error for comparing two treatments for which no observations are missing. The correct standard error for comparisons involving the missing treatment is slightly greater and is closely approximated by using formula (2) of Chapter 1, allotting the treatment with the gap one fewer observations than the other treatments.

The importance for experimental design of the missing-value formula is that it would have been a serious drawback to the use of randomized blocks, and of course also of more complex designs, had the analysis and interpretation been greatly complicated whenever an observation is lost, or more generally whenever it proves impossible to get data in exactly the form intended. The existence of the formula means that the randomized blocks design can safely be adopted even when the occurrence of occasional missing values is expected.

Extensions of the method can be used if there are several missing observations; it is then necessary to solve a set of simultaneous linear equations, the number of equations being equal to the number of missing values. Similar methods are available if by accident the treatments are not applied exactly according to plan and the pattern of treatments departs somewhat from the randomized blocks form.

Analogous formulas are available for the other designs described in the present book (see Cochran and Cox, 1957; Goulden, 1952).

(iii) Further Examples

Example 3.3. Another application of randomized blocks is in laboratory work with animals such as mice or rats. Suppose for definiteness that there are five treatments under comparison, their nature depending on the particular field of application, but being, for example, different diets, different amounts and types

of drug, different diets fed to rats during pregnancy, etc. The final observations might be of the amount of a certain substance in an organ of the animal at the end of the experimental period or, for the last case, some characteristic of the offspring.

To make successful use of the idea of randomized blocks, we begin by grouping the animals into sets of five in such a way that the final observation that would be obtained under uniform treatment is expected to be as nearly as possible constant within each set. Any special knowledge of the animals, such as their performance in previous experiments, can be used. In the absence of special knowledge it is common to rely either on the correlation that often exists between the final observations and a suitable, easily measured, initial property of the experimental animal, such as body weight, or on the general fact that animals from the same litter tend to respond similarly.

To use the last property, five suitable animals are taken from a number of litters of five or more animals, numbering the animals in each litter $1, \ldots, 5$ in any convenient way. The order of treatments is then randomized within each block to give some arrangement such as

Litter 1. Animal 1, T_3: 2, T_1: 3, T_5: 4, T_2: 5, T_4
Litter 2. Animal 1, T_2: 2, T_5: 3, T_3: 4, T_4: 5, T_1,
etc.

To use a quantitative property, such as body weight, to form blocks, the animals are numbered in order of increasing body weight, say, 1 through 20 if four blocks are required. Animals 1 through 5 form the first block, 6 through 10 the second block, and so on, the order of treatments again being randomized independently within each block. The effect of this is that the animals within any one block have approximately the same body weight. It sometimes happens that the first or last blocks contain one or more animals with very extreme body weights, so that there is appreciable variation of body weight within these blocks. If practicable, this is best avoided, for example, by starting with a few more animals than it is intended to use and discarding those with very extreme weights.

The results from such a design can be analyzed by the method of Example 3.2, the value of initial body weight being ignored once the grouping into blocks has been determined. In the next chapter we shall consider an alternative method of using the initial quantitative variable, in which the actual value is used in the analysis.

Example 3.4. In certain textile investigations it is required to test a number of modifications in a process for producing a thin web of parallel fibers. One important property of the web is the number of fiber entanglements, say per mg of web, and this is measured by passing a section of web slowly over an illuminated strip, when individual entanglements can be noted and the total found. However it is difficult to define precisely what constitutes an entanglement so that, whereas one observer can get reasonably reproducible counts over a short period of time, there are liable to be large systematic differences between observers and between the same observer's counts on different days. This example is typical of an important class of experiments in technology in which the properties under investigation are either rather difficult to define precisely and so are subject to personal errors of measurement, or, in extreme cases, are essentially matters of subjective judgement.

The first step in planning this sort of experiment is to take all reasonable precautions to eliminate the sources of systematic variation, for example by displaying in front of the observer photographs or slides showing typical fiber arrangements, some to be counted as entanglements and some not. The remaining systematic variations can then be reduced by the randomized block principle, as follows:

Suppose for definiteness that we have six different batches $W_1, \ldots, W_6$ of web to be compared. Let us assume to begin with that they have been produced by six different processes under highly controlled conditions, so that any

TABLE 3.4

PLAN FOR COMPARING SIX WEBS FOR ENTANGLEMENTS

		Order of Measurement					
		1	2	3	4	5	6
Block 1	Observer 1 First period	W_4	W_1	W_2	W_5	W_6	W_3
Block 2	Observer 2 First period	W_5	W_6	W_2	W_1	W_4	W_3
Block 3	Observer 1 Second period	W_3	W_6	W_2	W_4	W_1	W_5
Block 4	Observer 2 Second period	W_3	W_1	W_4	W_5	W_6	W_2
Block 5	Observer 1 Third period	W_6	W_5	W_4	W_3	W_1	W_2
Block 6	Observer 2 Third period	W_5	W_1	W_4	W_3	W_6	W_2
Block 7	Observer 1 Fourth period	W_2	W_3	W_6	W_4	W_1	W_5
Block 8	Observer 2 Fourth period	W_2	W_6	W_4	W_5	W_3	W_1

difference between $W_1, \ldots, W_6$ can be confidently attributed to the effect of processes. A similar point arose in connection with the preceding example and this is of course just the sort of assumption that it is so often desirable to avoid; this can be done by having several batches from each process, produced and tested independently.

In each block there will be observations on all six webs and we want it to be possible to complete the observations in a block within a fairly short time, so that time differences are eliminated. Therefore, we take as units small sections of web that can each be examined for entanglements in say 10–15 min, the sections being selected by a random-sampling procedure. The number of sections of each web that it would be desirable to measure depends on the final precision required and on the regularity with which the entanglements are distributed, and they would have to be determined from previous work or from the results of a preliminary experiment. Suppose that eight sections are judged adequate and that two observers are available, each measuring four times. Then an arrangement in randomized blocks is shown in Table 3.4.

The order of webs is randomized independently within each block. Subjective biases of measurement are minimized by concealing from the observer the identity of the section under analysis. In analyzing results it would be desirable to check the consistency of the observers in their comparison of the 6 webs. It is not necessary to randomize the allotment of observers to blocks, because the object of the experiment is not the comparison of observers; the only purpose that would be served by randomizing observers over blocks would be that of ensuring the absence of systematic differences in the external conditions during the two observers' measurements. Note that if the primary object had been an examination of the difference between observers, it would have been advisable to have had each section of web measured by both observers. However when the object is the comparison of webs, the more distinct sections that are taken from each web the better, provided that the main cost of the experiment is in the counting of entanglements and not in the selection of sections or in the cost of the material that is in effect destroyed in sampling the web.*

In this example the randomized block principle has been used to eliminate the effect of systematic variations arising in the actual measurement, rather than variations arising from the experimental material itself. There are other ways of achieving this end. For example we may insert into each series of sections of experimental webs, a section of standard web, which has been counted many times and may be considered to have a known number of entanglements. The observation actually recorded on the standard section is then used to adjust the remaining observations. Another possibility, which may well be the best if large observer and time differences seem unavoidable, is to abandon the idea of directly counting entanglements and instead to measure the amount of entanglement either by assigning a score to each section after a subjective comparison of it with standard sections showing varying degrees of entanglement, or alternatively by direct ranking of a series of sections in order of increasing apparent degree of entanglement. The discussion of the relative advantages of these procedures raises difficult general questions; they are dealt with briefly later (§ 9.4).

We have now had several examples of the use of randomized blocks. The grouping of the material into blocks eliminates the effect of constant differences between blocks and the randomization allows us to treat the remaining variation between units as random variation, so far as assessing treatment comparisons is concerned. The success of the method depends on a good grouping of the units into blocks. The general idea of grouping into blocks is of fundamental importance and is not only frequently used in simple experiments but also forms the basis for most of the more complicated designs.

Sometimes a generalized form of the randomized block design is useful. It may be required to make some treatment comparisons more precisely

* For example, if only a very limited quantity of each web is produced and it is required to leave as much as possible for further processing, it might be advisable to measure each section more than once. If the magnitudes of the different components of variation can be estimated and if the relative costs of the various stages of the experiment can be measured, the optimum distribution of effort can be determined.

than others. For example, we may have a control treatment C and a number of other treatments $T_1, T_2, \ldots$ and the main interest may be in comparing $T_1, T_2, \ldots$ individually with C rather than in comparing $T_1, T_2, \ldots$ among themselves. In such a case it is proper to devote more units to C than to each T treatment. The block principle can still be used with a simple analysis, provided that the number of times a particular treatment occurs in a block is the same for all blocks. Thus, in the above example C might occur four times in each block and $T_1, T_2, \ldots$ once in each block. The difference between two treatment means is again unaffected by constant differences between blocks.

3.4 ELIMINATION OF ERROR BY SEVERAL GROUPINGS OF THE UNITS

(i) Latin Squares

In the preceding section we have seen how a single grouping of the units into blocks can be used to reduce the error of an experiment. Sometimes two or more systems of grouping suggest themselves and it may be desired to use them simultaneously. For instance in a paired comparison experiment, like Example 3.1, it might happen that there is reason to expect a systematic difference between the first unit and the second unit in the pair. Then we should have two systems of grouping, into pairs and into order within pairs, and we would wish to balance out both the associated types of systematic variation. A discussion of this example would involve one or two special points, and it is convenient instead to introduce the basic design, the Latin square, with a somewhat different problem.

Example 3.5. Consider an industrial experiment in which four processes are under comparison and in which it is suspected that there will be systematic changes in external conditions from day to day and also between different times of day, e.g., observations on material processed in the early morning may on the whole be lower than on material processed in the afternoon, etc. Suppose that the number of units that can be dealt with on one day is limited and that four is a convenient number, say two in the morning and two in the afternoon. Suppose also to begin with that four observations on each process are considered likely to give sufficient precision. The sixteen experimental units can then be set out in the square array shown in Table 3.5(a). If we preferred to use "days" as blocks in a randomized block design, we should arrange that each process is used once on each day, the arrangement otherwise being random. If we preferred to use "times of day" as blocks in a randomized block design, we should arrange that each process is used once at each time of day. Therefore if we wish to eliminate both sources of variation simultaneously we must arrange the four processes P_1, P_2, P_3, P_4 in the 4×4 square of Table 3.5(a) so that each letter occurs once in each row and once in each column.

One such arrangement is shown in Table 3.5(*b*) and is an example of a 4 × 4 *Latin square*. In general an $n \times n$ Latin square is an arrangement of n letters in an $n \times n$ square, such that each letter occurs once in each row and once in each column.

The particular Latin square given in Table 3.5(*b*) has been obtained by randomization in a way to be described in Chapter 10. The procedures to be used if it is required to have more experimental units, for example to have eight units for each treatment, will be discussed later.

TABLE 3.5

A LATIN SQUARE DESIGN

(*a*) *General Arrangement of Experimental Units*

	Time of Day			
	Time 1	Time 2	Time 3	Time 4
Day 1	—	—	—	—
Day 2	—	—	—	—
Day 3	—	—	—	—
Day 4	—	—	—	—

(*b*) *A Latin Square*

	Time 1	Time 2	Time 3	Time 4
Day 1	P_2	P_4	P_3	P_1
Day 2	P_3	P_1	P_2	P_4
Day 3	P_1	P_3	P_4	P_2
Day 4	P_4	P_2	P_1	P_3

The analysis of the observations from a Latin square is done by a procedure exactly analogous to that for the randomized blocks design. The treatment effects are estimated by comparing the average observations for the different treatments, and the estimate of the error standard deviation is obtained either from the appropriate analysis of variance or by calculating residuals. In accordance with the principle stated in § 3.3, every source of variation balanced out in the design of the experiment must be removed in the analysis before the standard deviation is estimated. The definition of the residual corresponding to a given observation is thus

$$\text{observation} - \left(\begin{array}{c}\text{mean observation on the corresponding treatment}\end{array}\right) - \left(\begin{array}{c}\text{mean observation on the corresponding row}\end{array}\right) - \left(\begin{array}{c}\text{mean observation on the corresponding column}\end{array}\right)$$

$$+ \text{ twice the overall mean,}$$

and the estimate of the standard deviation is

$$\sqrt{\left\{\frac{1}{(n-1)(n-2)} \times \text{sum of squares of the residuals}\right\}},$$

for an $n \times n$ square. The divisor $(n-1)(n-2)$, the residual degrees of freedom for the Latin square, is the number of independent residuals. The estimated standard error of the difference in mean observation is*

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

Full details of the procedure of analysis are given in textbooks on statistical methods.

The example just discussed shows that the Latin square arrangement is a simple and natural extension of the randomized block design. The example is of wide applicability, since there are many types of work in which the rate at which experimental material can be dealt with is limited and in which it is worth balancing out certain time variations. Another common possibility arises when there are a number of observers or sets of apparatus or machinery, which can be used simultaneously. The Latin square arrangement can then be used with the rows standing for different times, and the columns for the different sets of apparatus, etc. In this way systematic time differences and systematic differences between sets of apparatus, etc., are eliminated.

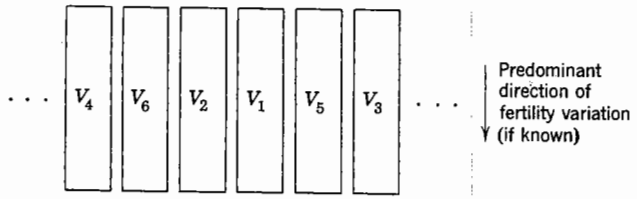
A restriction, which clearly limits the use of the Latin square in its simple form, is that the number of rows, the number of columns, and the number of treatments must all be equal. Arrangements not restricted like this are discussed in Chapter 11. It is convenient now to consider some more examples of the use of the Latin square.

Example 3.6. In an agricultural field trial to compare a fairly small number of varieties or treatments, the best arrangement of plots depends in part on the shape of the plots, which is largely dictated by technical considerations. For example in a variety trial, particularly if it is required to have small plots, the plots would be long and narrow, only a few drills wide. In this case a natural grouping of plots for the use of a randomized block design is that shown for six varieties in Table 3.6(*a*), in which the blocks are approximately square and are, if possible, oriented to minimize the effect of the predominant fertility variations, if the best direction for doing this is known from previous experience of the field. If, on the other hand, the plots are more nearly square, a compact arrangement, such as that illustrated in Table 3.6(*b*), would usually be better. The exception is when it is confidently expected that fertility variations in one direction will be much larger than variations in the perpendicular direction; in this case the arrangement corresponding to Table 3.6(*a*) is likely to be successful. The objection to this design would ordinarily be that with wide plots, the whole block is of considerable extent and is therefore likely to contain excessive variation; all the variation of fertility between plots within the long block would contribute to the error. The design would be particularly bad if the predominant direction of fertility variation was misjudged and happened to lie parallel to the length of the block.

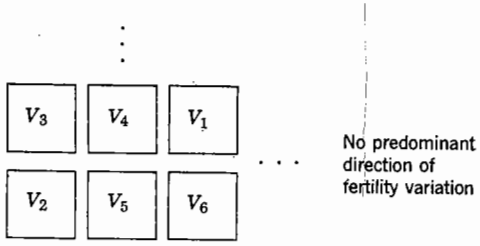
* In fact this formula, or its generalization (Chapter 1, Eq. (2)) applies for all designs in which the treatment effects are estimated by simple treatment means.

TABLE 3.6

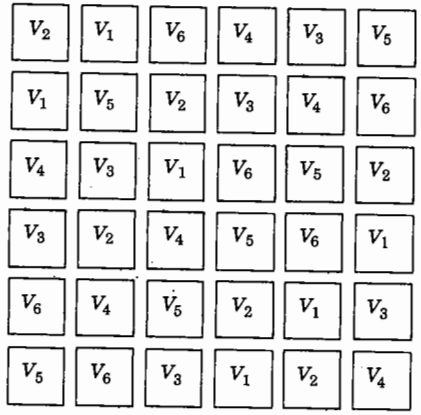
AGRICULTURAL FIELD TRIALS



(a)



(b)



(c)

Consider, however, the Latin square arrangement shown in Table 3.6(*c*). Here we are eliminating fertility variations in two directions and so there is a good chance that one or other, if not both, of the groupings will account for an appreciable portion of uncontrolled variation. Much depends on the particular circumstances and on special information that may be available in particular cases, but it seems that for many types of field experiment with approximately square plots and with not more than about ten to twelve treatments, the Latin square is a good design, and is to be preferred to randomized blocks, provided that it is reasonable to have the number of plots per treatment equal to, or a simple multiple of, the number of treatments. Exceptions to this have, however, been reported in the literature.

Example 3.7. The Latin square principle is employed frequently in experiments in which the same object or person is used several times. Thus for the cattle experiment discussed in Example 2.7, suppose that interference between different two-week periods can be ignored. Then we have two types of systematic error to try to eliminate, arising from variations between animals and from the common time trend between periods. Therefore the Latin square design is indicated. For each group of three similar animals a 3 × 3 Latin square is used, as shown in Table 3.7; the randomization of each 3 × 3 square is done independently.

TABLE 3.7

ANIMAL FEEDING TRIAL

	Two-week Period		
	1	2	3
Cow 1	*C*	*A*	*B*
Cow 2	*B*	*C*	*A*
Cow 3	*A*	*B*	*C*

The diets under comparison are denoted by *A*, *B*, *C*.

The three animals in a square should be chosen so that they are likely to have similar lactation curves and to be at corresponding points on the curve at the start of the experiment. The reason for this is that the balancing of columns in the Latin square removes the effect of a common trend in milk yield, but that if the trends are appreciably different for the three animals, the error of the treatment comparisons is inflated. The whole experiment will consist of several squares of the above type. It does not matter if the trends are different in the different squares; what is important is that as far as possible any trends that do exist should be the same for the three animals in any one square. If this state of affairs is not attained the randomized Latin square arrangement is, of course, still a perfectly valid experiment, giving treatment comparisons of measurable precision and free of systematic error. The point is that precision is lost.

The straightforward use of the Latin square applies when there is no carry-over of the treatment effect from one period to another. We shall see later that a special sort of Latin square is the design to use when there is a carry-over of treatment effect.

Example 3.7 could be paralleled from many applied fields. For

example in some experiments on the fuel consumption of buses, described by Menzler (1954), four vehicles were used to compare four different tire pressures, four different tread thicknesses, or four different methods of operation, the experiment being repeated on four days. There are two types of systematic variation to be balanced out, that between days and that between vehicles, and a 4 × 4 Latin square is appropriate, with the rows representing vehicles and the columns, days.

Example 3.8. The following case of the misuse of the Latin square, quoted by Babington Smith (1951), illustrates the importance of considering the basic assumptions of Chapter 2. Four backward readers, Tom, Dick, Harry, and George undergo in succession four trainings in spelling, denoted by *A*, *B*, *C*, *D*, the treatments being arranged in a Latin square as in Table 3.8. An observation

TABLE 3.8

COMPARISON OF TRAINING METHODS

	Period 1	Period 2	Period 3	Period 4
Tom	*B*	*A*	*D*	*C*
Dick	*C*	*B*	*A*	*D*
Harry	*A*	*D*	*C*	*B*
George	*D*	*C*	*B*	*A*

is made on each subject at the end of each period, using a standard type of spelling test.

The justification for using the Latin square is that it aims at balancing out differences between subjects, and also any systematic effects accounted for by the order in which the methods of training are applied. However, we have an untenable assumption, namely that the observation obtained, say with method *D* for Tom in the third period, is unaffected by the particular choice of treatments for Tom in the preceding periods. The observation obtained on a subject in a particular period is likely to depend, probably in rather a complicated way, on all the training received up to that point. It is not difficult to conceive of situations in which the comparison of methods of training by a simple averaging of observations would give quite misleading results.

A list of standard Latin squares, from which designs can be constructed by randomization, is given in Chapter 10, where there is also a discussion of more complicated designs based on the Latin square principle. Two simple extensions of the Latin square design will be discussed briefly here; the first is that several Latin squares may be used simultaneously and the second is that instead of two sources of systematic variation to be dealt with, there may be three or even more.

(ii) Combined Latin Squares

In any experiment of the type we have been considering, for which an $n \times n$ Latin square is appropriate, it may well happen that more than n

units are required for each treatment to get estimates of adequate precision. This situation can be dealt with simply if one of the sides of the square, say the rows, represents extension in space or time, and if a multiple of n^2 units can be used.

Example 3.9. Consider again Example 3.5, where the rows of the square represent different days, the columns times of the day. If we want to extend the experiment over twelve days instead of four, two procedures are available and are illustrated in Table 3.9(*a*) and (*b*).

In the first design, Table 3.9(*a*), we have three independently randomized Latin squares placed underneath one another. In the second design, Table 3.9(*b*), we have completely randomized the rows of the previous design, so that, for example, the first four rows by themselves no longer necessarily form a Latin square.

The difference in practice between the designs is best seen by considering what types of systematic variation are eliminated from the error by the two designs. In both cases constant differences between days have no effect on the treatment comparisons. In the second design, the effect of constant differences between times of the day persisting throughout the whole experiment is likewise eliminated. In the first design, however, not only is this done, but also time of day effects are eliminated separately from each set of four days. This would be particularly useful if, as might be convenient, there is a considerable gap in time between the sets of four days, or if it were desired to introduce some external change in conditions, either of which things might mean that time of day effects would not be the same in all parts of the experiment.

Therefore in general we prefer the design (*a*) because it achieves all that (*b*) does and more. There are, however, two considerations which prevent (*a*) always being better than (*b*), although neither consideration is likely to be of any importance in the present case. First we would usually need to estimate the error standard deviation from the results of the experiment itself and, as we have seen above, the accuracy with which we do this, measured by the degrees of freedom for residual, affects somewhat the effective precision of the experiment. Now the residual degrees of freedom for the design in Table 3.9(*a*) may be shown to be 24 and the corresponding value for Table 3.9(*b*) is 30. In accordance with the full discussion in Chapter 8, this means that if the true standard deviations corresponding to the two designs were equal, the effective standard deviation for Table 3.9(*b*) would be about 1 per cent less than that for Table 3.9(*a*). This gain is negligible, but in smaller experiments the corresponding gain may be enough to justify the use of the design analogous to Table 3.9(*b*), in a case where there is good reason to expect any column effects to be constant throughout the experiment.

The second consideration, which is of merely academic interest in the present case, is that with very special patterns of uncontrolled variation Table 3.9(*b*) may give the more precise design, even apart from the consideration of the residual degrees of freedom. Suppose, to take the extreme case, the observations in the absence of treatment effects were of the form in Table 3.9(*c*), with only two possible values x, y distributed in the pattern shown. Notice that the systematic variation between days is zero, since for each day the observations add up to $2(x + y)$. Likewise if the columns are taken in sets 1–4, 5–8, 9–12 there is no systematic variation between columns. Of course, if it were known or suspected

TABLE 3.9

EXTENDED LATIN SQUARE DESIGNS

(*a*) *Separate Latin Squares*

	Time of Day				
	Time 1	Time 2	Time 3	Time 4	
Day 1	P_3	P_2	P_1	P_4	
2	P_2	P_3	P_4	P_1	Replicate
3	P_4	P_1	P_3	P_2	1
4	P_1	P_4	P_2	P_3	
5	P_4	P_1	P_2	P_3	
6	P_3	P_4	P_1	P_2	Replicate
7	P_2	P_3	P_4	P_1	2
8	P_1	P_2	P_3	P_4	
9	P_2	P_4	P_3	P_1	
10	P_3	P_1	P_2	P_4	Replicate
11	P_4	P_3	P_1	P_2	3
12	P_1	P_2	P_4	P_3	

(*b*) *Intermixed Latin Squares*

	Time 1	Time 2	Time 3	Time 4
Day 1	P_3	P_1	P_2	P_4
2	P_2	P_3	P_4	P_1
3	P_3	P_4	P_1	P_2
4	P_4	P_1	P_3	P_2
5	P_3	P_2	P_1	P_4
6	P_2	P_3	P_4	P_1
7	P_4	P_1	P_2	P_3
8	P_1	P_2	P_4	P_3
9	P_2	P_4	P_3	P_1
10	P_1	P_4	P_2	P_3
11	P_4	P_3	P_1	P_2
12	P_1	P_2	P_3	P_4

(*c*) *A Very Special Pattern of Uncontrolled Variation*

	Time 1	Time 2	Time 3	Time 4
Day 1	x	x	y	y
2	x	x	y	y
3	y	y	x	x
4	y	y	x	x
5	x	x	y	y
	and so on up to			
12	y	y	x	x

that the uncontrolled variation was of this form, neither of the designs we are considering would be at all appropriate. However, if the pattern of variation in Table 3.9(*c*) did occur, the standard deviation for design (*a*) can be shown to be $\sqrt{(33/27)} = 1.11$ times that for design (*b*). This, the largest factor in favor of (*b*) that can occur, does not represent a large change in precision and in any case only arises in the exceptional circumstances of Table 3.9(*c*).

We can formulate an important general rule for experiments of this type; that it is better to keep separate the sections, e.g., Latin squares, from which the whole design is built, except possibly when the experiment is a small one with few degrees of freedom available for the estimation of error, or when very special patterns of uncontrolled variation may arise.

All the arrangements that we have considered so far give equal precision for the comparison of every pair of treatments. If, however, it is required, say, to have two observations on A for each observation on B, C, etc. this can be done most simply by taking a 5×5 Latin square for treatments A, B, C, D, E and applying the treatment A every time the letters A or E occur. If the conditions of the experiment do not allow the use of five units in one day, the more complicated "unbalanced" arrangements of Chapter 11 must be used, if the additional observations on A are to be obtained.

Example 3.10. As a final example of the simple two-way elimination of error, consider the paired comparison experiment, Example 3.1, in which it is required to compare two treatments T_1, T_2, eliminating from the effective error not only the variation between pairs of units, but also any systematic variation associated with the order in which the units are arranged within pairs.

If we look back at Table 3.1, which shows a particular arrangement of treatments with one-way grouping of the units, we see that T_1 occurs three times in the first position and five times in the second. What we want, if it is expected that the observation on the first unit in a pair will tend to be larger (or smaller) than the observation on the second unit in the pair, is that each treatment should occur four times in each position. This raises essentially the same problems as Example 3.9. The key design in the 2×2 Latin square:

$$\begin{matrix} T_1 & T_2 \\ T_2 & T_1 \end{matrix} \qquad \text{or} \qquad \begin{matrix} T_2 & T_1 \\ T_1 & T_2 \end{matrix}$$

and we want four of these to build up the requisite number of observations. There are three different arrangements that merit consideration. First we might consider taking four Latin squares kept separate, analogous to the arrangement in Table 3.9(*a*). It can be shown that this design has only three degrees of freedom for residual and this would ordinarily be insufficient (see § 8.3), so that this arrangement would be used only in very special circumstances. The second arrangement, analogous to Table 3.9(*b*), is obtained by choosing completely at random four pairs to receive the order $T_1\ T_2$ and assigning the order $T_2\ T_1$ to the remaining four pairs. This may be shown to leave 6 degrees of freedom for residual, and is the arrangement that would normally be used. The third method is an intermediate arrangement, which is worth considering when the pairs fall naturally into two equal sets in which the order effects are quite possibly different. In this design the treatments are randomized separately within each set, so that $T_1\ T_2$ and $T_2\ T_1$ both occur twice in each set. This leaves 5 degrees of freedom for residual. The reader should write out examples of the three methods and consider carefully the types of systematic variation balanced out by each.

In experiments like this in which the degrees of freedom for residual are inevitably small, it will be worth considering whether useful information about the error standard deviation can be derived from the results of previous similar experiments.

(iii) Graeco-Latin Squares

The randomized block design is useful when the experimental units are grouped in one way. The Latin square design is useful when the units are simultaneously grouped in two ways. It is natural to consider what can be done if the units are grouped in three (or even more) ways.

Example 3.11. Consider again Example 3.5 used to illustrate the idea of a Latin square. Suppose that the observations are made by four observers and that each experimental unit is to be measured by one observer. Then, unless the absence of systematic observer differences can confidently be assumed, which would not be often, each observer should measure one unit of each process.

This could be done by a further application of the randomized block principle. One unit could be selected at random from each process for observer 1, a further set for observer 2, and so on. However, it would normally be an advantage to be able, in the analysis, to separate out differences between observers, between days and between times of day. Although this separation is not essential for the immediate purpose of comparing the processes, it may give information about the uncontrolled variation, of value both in attaining a general understanding of the experimental set-up and in designing future experiments.

Therefore we want to superimpose on the Latin square in Table 3.5(*b*) the symbols O_1, O_2, O_3, O_4 for four observers in such a way that

(*a*) each observer occurs once in combination with each process;
(*b*) each observer measures once on each day and once at each time of day.

The second condition is satisfied if the O's, considered by themselves, form a Latin square.

One such arrangement, after randomization, is shown in Table 3.10(*a*). The

TABLE 3.10

EXPERIMENT IN A GRAECO-LATIN SQUARE

(*a*) *Arrangement of Processes and Observers*

	Time 1	Time 2	Time 3	Time 4
Day 1	P_2O_3	P_4O_1	P_3O_2	P_1O_4
Day 2	P_3O_4	P_1O_2	P_2O_1	P_4O_3
Day 3	P_1O_1	P_3O_3	P_4O_4	P_2O_2
Day 4	P_4O_2	P_2O_4	P_1O_3	P_3O_1

(*b*) *The Same with Latin and Greek Letters*

$B\gamma$	$D\alpha$	$C\beta$	$A\delta$
$C\delta$	$A\beta$	$B\alpha$	$D\gamma$
$A\alpha$	$C\gamma$	$D\delta$	$B\beta$
$D\beta$	$B\delta$	$A\gamma$	$C\alpha$

processes are arranged in the same way as in Table 3.5(*b*). Note that the O's form a Latin square and that condition (*a*) is satisfied because, for example, O_3 occurs in combination with P_2 just once. An arrangement like this is called a *Graeco-Latin square*. The reason for this name is that it is a common convention to rewrite the square replacing one set of symbols, say $P_1, \ldots, P_4$, by Latin letters A, B, C, D and the other set of symbols, $O_1, \ldots, O_4$, by Greek letters $\alpha, \beta, \gamma, \delta$. This has been done in Table 3.10(*b*); the general definition is that an $n \times n$ Graeco-Latin square is an arrangement of n Latin letters and n Greek letters in an $n \times n$ square in such a way that each Latin letter (and each Greek letter) occurs once in each row and once in each column, and that each combination of a Latin letter and a Greek letter occurs paired just once.

The statistical analysis of the results of such an experiment is a straightforward extension of that for a Latin square. Process, day, time of day, and observer effects are estimated by averaging and the error standard deviation is estimated either by analysis of variance or, equivalently, by forming and squaring the residuals, defined as

$$\text{observation} - \begin{pmatrix}\text{mean obs. on}\\ \text{corresponding}\\ \text{process}\end{pmatrix} - \begin{pmatrix}\text{mean obs. on}\\ \text{corresponding}\\ \text{day}\end{pmatrix} - \begin{pmatrix}\text{mean obs. on}\\ \text{corresponding}\\ \text{time of day}\end{pmatrix} - \begin{pmatrix}\text{mean obs. on}\\ \text{corresponding}\\ \text{observer}\end{pmatrix} + 3\,(\text{overall mean}).$$

The degrees of freedom for residual in an $n \times n$ square are $(n-1)(n-3)$, so that a 4×4 Graeco-Latin square gives only 3 degrees of freedom for residual. This would not, by itself, lead to an adequate estimate of error and so it would, with one replicate of this design, be necessary to have a supplementary estimate of error.

The Graeco-Latin square, though an important design both in principle and as a basis for constructing further designs, is not itself used very frequently in practice. The same applies to the more complicated squares, in which, for example, a third alphabet is placed in Table 3.10(*b*) in such a way that any alphabet by itself forms a Latin square and any pair of alphabets a Graeco-Latin square. This would be relevant if there were four simultaneous groupings of the experimental units.

Examples of Graeco-Latin and higher-order squares for small and moderate values of n, and instructions for their randomization, are given in Chapter 10. Some ingenious practical applications of these squares have been described by Tippett (1935).

3.5 THE NEED FOR MORE COMPLICATED ARRANGEMENTS

The essential point of randomized block and Latin square designs is that the experimental units are grouped into sets, the grouping being chosen so that the uncontrolled variation within sets is as small as possible.

It quite often happens that if this last condition is to be satisfied the number of units in a block must be small.

Thus, if the experimental units consist of pairs of identical twins, we are restricted to two units per block in order to make effective use of the similarity of the twins. If we wish to make one day's work a block in a randomized block design, this will set an upper limit to the number of units in a block, depending on how many units can be dealt with in a day. In an agricultural field trial there is no such clear upper limit to the number of plots per block, but the more plots in a block, the greater the area of the block and the more likely it is to contain substantial heterogeneity. Hence, there is again reason for limiting the number of plots per block, and twelve to sixteen is usually regarded as a maximum satisfactory number.

Now the randomized block design has at least as many units in a block as there are treatments, and similarly the simple form of Latin square has the number of rows and columns equal to the number of treatments. But what if the number of treatments exceeds the allowable number of units per block or exceeds the permissible number of columns in a Latin square design? For example, suppose that we wish to use the pairs of twins to compare five diets. We need an arrangement similar to randomized blocks, eliminating differences between blocks from the error, but having fewer units per block than the number of treatments. Much of the mathematically advanced work connected with experimental design aims at providing designs that enable this elimination to be achieved efficiently and simply. Some special types are balanced incomplete blocks, lattices, confounded arrangements, and so on. Similarly there are designs, Youden squares, lattice squares, and quasi-Latin squares, which fulfil the purpose of the Latin square but have the number of rows or columns, or both, less than the number of treatments. These will all be described later; the point of the present discussion has been to see just how the need for these more complicated arrangements arises.

SUMMARY

Knowledge available to the experimenter about the probable nature of the uncontrolled variation can be used to increase the precision of the treatment comparisons. The procedures considered in this chapter are:

(*a*) randomized blocks, in which the units are grouped into blocks and the treatments arranged randomly within blocks, each treatment occurring once, or more generally the same number of times, within each block;

(*b*) Latin squares, in which a similar method is used, although with two groupings of the experimental units.

These methods depend for their success on a skilful grouping of the units.

REFERENCES

Babington Smith, B. (1951). On some difficulties encountered in the use of factorial designs and analysis of variance with psychological experiments. *Brit. J. Psychol.*, **42**, 250.

Biggers, J. D., and P. J. Claringbold. (1954). Why use inbred lines? *Nature*, **174**, 596.

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

Fertig, J. W., and A. N. Heller. (1950). The application of statistical techniques to sewage treatment processes. *Biometrics*, **6**, 127.

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

James, E. (1948). Incomplete block experiment with half plants. *Proc. Auburn Conference on Applied Statistics*, 52.

Menzler, F. A. A. (1954). The statistical design of experiments. *Brit. Transport Rev.*, **3**, 49.

Tippett, L. H. C. (1935). Some applications of statistical methods to the study of variation of quality in cotton yarn. *J. R. Statist. Soc., Suppl.*, **2**, 27.