CHAPTER 12

INTRODUCTION TO FIELD EXPERIMENTS

12.1 Soil heterogeneity

12.1.1 The earliest field experiments were observational (§ 2.7). If the difficulty due to differences in fertility between adjacent plots was realized, it was countered by repetitions of the experiment in a number of localities over a period of years. It was not until about 1910 that the first **uniformity trial** was conducted to investigate the nature of soil heterogeneity and its effects on field experiments.

12.1.2 A uniformity trial, sometimes called a blank trial or dummy experiment, consists of an experiment in which all the plots receive identical treatment. There are two types of uniformity trial:

Type 1 The area planted is one which is to be used the following season for a proper experiment, and the plots harvested are the same as will be used in the subsequent experiment. Such yields provide pre-knowledge of the relative fertility of the plots and this information may prove useful in the analysis of the experiment proper. (See "Analysis of covariance", Chapter 18.)

Type 2 A rectangular area of land is planted uniformly to a given crop and receives uniform treatment throughout. At harvest, the area (probably with border areas excluded) is divided up into a large number of very small rectangular areas equal in size and similarly oriented, the yields of which are separately recorded. The yields of Table 4.1 are from a trial of this sort.

12.1.3 The original data from which Table 4.1 was compiled would consist of the yields entered in a rectangular table representing a plan of the trial and showing the geographical locality of each plot yield relative to the others (cf. Exercise 4.1). Such data provide information concerning the pattern of soil heterogeneity of the area concerned and such information can be put to various uses. What concerns us here, however, is the nature of this pattern of heterogeneity.

12.1.4 Numerous uniformity trials in various localities and with different crops have confirmed that even on areas specially selected for their uniformity the yields of identically treated plots vary considerably. Some of this variation is due to chance error arising from sources other than soil fertility, but the fact that *the yields do not vary randomly over the area* leads us to believe that this non-randomness is due to non-random changes in soil fertility. Also,

although it is impossible to make any ultimate separation between variation due to soil fertility and variation due to other sources, it is suspected that the former plays a dominant role. This non-random or systematic character of the yields (and consequently of soil fertility) manifests itself in two main ways:

(1) Identically treated adjacent plots tend to yield more alike as a rule than plots further apart.

(2) Although the fertility is frequently very patchy and fertility contours are very erratic in shape, there is often a tendency for the general level of fertility to change in a particular direction. This tendency is referred to as a **fertility gradient** or **fertility slope**, though it must not be thought that there is anything very regular about such gradients. It is a convenient way of referring to the main direction in which fertility changes (up, down, or fluctuating) occur.

12.2 Replication

12.2.1 Once the idea of the inevitable variability between identically treated plots was recognized, elementary statistical reasoning suggested the inclusion in an experiment of several plots of each treatment. This has, as we have seen, two important results:

(1) It makes possible the estimation of the variance of plots treated alike, i.e. of the error variance, thus enabling the precision of the treatment means to be fixed and the significance of treatment differences to be tested.

(2) The more plots of each treatment we have, i.e. the greater the number of **replications**, the greater the precision of the estimates, both of the treatment means and of the error variance itself. The former arises from the fact that, if σ^2 is the error variance and r the number of replications, the variance of a treatment mean is σ^2/r (Theorem 6.8, corollary) which is obviously reduced by an increase in the number of replications (cf. §§ 6.6.3 and 9.10.1; see, however, § 12.11.7). The latter is due to the increased number of D.F. available for the Error M.S. (cf. § 10.6.7). These quantities are two factors in determining the **efficiency** of an experiment, the third being the magnitude of σ^2 itself, over which there is no direct control.

12.2.2 We speak of an experiment of this type as a **replicated experiment.** By replication we mean the repetition of plots of a treatment within a single experiment. If an experiment is self-sufficient within itself, but is repeated in a number of localities or seasons, it is usual to regard these as *repetitions* of the experiment rather than as additional replications within the same experiment. Occasionally, however, a definite decision is made to place replications of the treatments in different localities but within a single experiment. By "a replication" we usually imply a set of plots in which each of the treatments in an experiment occurs once; yet an experiment with one such set is regarded as "unreplicated", even though there is really one replication, and an experiment with two such sets is regarded as being replicated twice, not once.

12.3 The principle of randomization

12.3.1 Replication is, however, not enough. Some early replicated experiments were laid out in a *regular* order, e.g. $ABCD : ABCD : ABCD : \ldots$, but it is obvious that, if there were any consistent upward or downward trend in soil fertility along the field, comparisons between the treatment means would be biased, because, for example, A always occurs on the left of B, and so on. Such a trend is, in fact, almost inevitable, because any fertility gradient, unless at right angles to the line of the plots, will have some component along this line. While it could possibly be claimed that in a series of identical experiments carried out in many localities and seasons these biases would even out, it does not alter the fact that the results of individual experiments would be almost certainly subject to biases of unknown magnitude, and the results of individual experiments are important in themselves. Also, one cannot always have a large number of repetitions of every experiment even though it is admittedly desirable that the treatments should be tried out in various localities and over several seasons in view of the fact that results can vary widely in different seasons and localities.

12.3.2 Given a set of plots, therefore, on which an experiment is to be conducted, we must reject the regular type of lay-out in favour of one which will provide a fairer allocation. This allocation must not be left to the haphazard whim of the experimentalist. Even if we do him the justice that he will not deliberately pick out the best-looking plots for his pet treatment, subconscious influences might still have unpredictable consequences. As evidence of this sort of thing, consider the following table:

Table 12.1: Nos. of students selecting an integer from 1 to 4 "at random"

Digit selected	Frequency
1	3
2	8
3	32
4	5
Total	48

Each student was asked to select an integer from 1 to 4 "at random", i.e. on the spur of the moment. The curious popularity of the digit 3 shows that such a selection is not, in fact, random in the statistical sense, or else the choices would be much more evenly spread. (See also Example 15.1.)

12.3.3 A solution which at first sight seems to have merit is to try and "balance" the plots allotted to various treatments so that each treatment starts off on plots of equal, or nearly equal, average fertility. One way of doing this would be to use the data from a uniformity trial carried out on the same plots the previous season. It is then a matter of juggling the plots around until each treatment group contains plots whose average fertility, as judged by the uniformity data, is approximately the same. Although uniformity data may not often be available in field experiments, the method appears feasible

in animal nutrition experiments, for example, where initial weights of the animals are readily obtainable.

12.3.4 An alternative method of balancing which does not require uniformity data is the use of *systematic* designs constructed so that each treatment will share as equally as possible in the fertility of the experimental area. A simple example is a design such as

$$ABCD : DCBA : ABCD : DCBA,$$

where the plots are set out in one long line.

Without doubt balancing procedures will tend to minimize the errors to which estimates of the treatment effects will be subject.

12.3.5 *Fisher's solution (1925) to the problem of securing an unbiased allocation of plots to the treatments was to insist that the allocation be random.* As already explained, this is not to be regarded as equivalent to a "haphazard" allocation, and, in fact, *an actual physical process of randomization* is insisted on, so that the allocation will depend solely on the laws of chance and so that each treatment has an exactly equal *a priori* chance of being allotted to any particular plot. Tables of random numbers are usually employed for this purpose. *Each experiment must be separately randomized whether belonging to the same series or not.*

12.3.6 The proponents of regular (and systematic) designs objected to randomization on the grounds that the laying out of an experiment would be more difficult and mistakes in the field work more common. Experience has shown that with proper organization quite complex randomized designs can be managed without difficulty, and the realization that unbiased treatment comparisons were essential led to the early disappearance of the regular designs.

12.4 Simple random design

12.4.1 For reasons of convenience and in order to keep the plot variability within reasonable limits, it is conventional to lay out the plots of an experiment in a compact rectangle known as the **experimental area.**

12.4.2 A design in which each of the set of plots in the experimental area is allotted at random in turn to a particular treatment (i.e. where each treatment has an exactly equal chance of being allotted to any plot, not only *a priori*, but also at any stage during the draw) would be justifiably called a "completely random" design. In practice, however, it is usual to specify in advance the number of plots to be allotted to each treatment, usually equal numbers. In that case, once the required quota of plots for a given treatment is exhausted, the probability that this treatment will be allotted to one of the remaining plots is zero. It has become customary to refer to a design of this sort as a completely random design, but as indicated above, such a design is not really *completely* random, and it is felt that the name **simple random design** is better in view of the close analogy with the simple random sample (cf. § 12.4.3).

12.4.3 Although it is possible to perform the randomization by considering each plot in turn and allotting a treatment to it at random until the quota for each treatment is exhausted, it is sometimes found that this method will, if the usual procedure with tables of random numbers is followed (viz. dividing pairs of numbers by the number of treatments and taking remainders, as in § 3.13.2), show up the defects of individual pages of random number tables by revealing a scarcity of pairs of numbers giving a particular remainder. Consequently one may find that, after the quotas of plots for all but one of the treatments have been filled, the remaining plots in one corner of the plan are left for the last treatment. In an experiment with 8 replications one might find as many as 4 or 5 plots of the last treatment occurring in a bunch. Since this is undesirable and because it is really the consequence of a defect in the randomization process rather than a genuine, but unusual, randomization, it is recommended that the random allotment of treatments be made by following the method set down for drawing a simple random sample (§ 3.12). Suppose we have 48 plots for an experiment with 8 replications of 6 treatments. We would first select a simple random sample of 8 plots from the 48 and allot these plots to one of the treatments—either at random or to the first treatment in order, it is immaterial, so long as one doesn't inspect the selected plots and decide that these will favour a particular pet treatment! Next, a further simple random sample of 8 plots is drawn from the remaining 32, and allotted to a second treatment. It is here probably simplest to keep regarding the sampling as being from the original 48 plots, but rejecting any plots already drawn. The process is continued until 5 samples of 8 plots have been drawn and allotted to one treatment each. The last 8 are then allotted to the sixth treatment. As a precaution against any "funny business", the randomization should be done in an office, not on the site of the experiment.

12.4.4 The data from a simple random experiment will have the same appearance as the data from Example 10.4, and we shall naturally wish to employ the same convenient method of analysis of variance as used there. Our immediate concern, therefore, is with the validity of the analysis of variance in connection with a randomized field experiment.

12.5 Randomization and unbiased estimates of error

12.5.1 By contrast with the early disappearance of "regular" designs, the use of systematic designs (as defined in § 12.3.4) continued to flourish, and the controversy between the advocates of systematic designs (among whom was "Student") and the Rothamsted school came to a head about 1936. Although the issue was quite clearly settled (for the second time) in favour of randomization, systematic designs continued to be used. The author has himself seen fairly recent designs in which the balancing of the various treatments had been worked out with great, but nevertheless misguided, ingenuity. Systematic designs are still used on the Continent to some extent even today. The reasons for this persistent survival of systematic designs lie no doubt in their allegedly greater convenience in the field, in the argument that systematic

designs have lower errors than randomized designs, and in the apparent fairness of the procedure of "parcelling out" the fertility among the various treatments as equally as possible as compared with leaving this equality to chance, which might happen to bring about a very uneven allocation. Yet the possibility that the fertility pattern of the experimental area might coincide to some extent with the pattern of a prescribed systematic design, thus upsetting the fairness of allocation or the desired balance, must not be overlooked.

12.5.2 The advantages of randomization go much deeper than the question of ensuring unbiased treatment comparisons. When Fisher introduced his analysis of variance technique in conjunction with the principle of randomization, this method of analysis was taken over for systematic designs as well. To illustrate what happens in such a case, let us consider the yields from a set of plots to which no treatments have been applied, and superimpose a dummy experiment on these yields by allocating a set of "treatments" to the plots. The analysis of variance of this experiment will then be of the type discussed in Chapter 10, and the null hypothesis is known to be true. If the allocation of treatments is according to a systematic design and if this design is successful in achieving "balance" among the treatments, differences between treatment means will be small. The Total S.S. of the plot yields, however, remains the same irrespective of the allocation. Consequently, if we have succeeded in reducing the errors of the treatment means, the Treatments M.S. will be reduced, and it inevitably follows that the Error M.S. will be increased. On the null hypothesis these two M.S.'s are "expected" to be equal. Hence although the treatment effects will be very precisely estimated, they will *appear* to be much less precisely estimated because of the increased Error M.S. This could even account for a significantly small value of F (cf. § 10.6.5).

12.5.3 *Valid experimental designs are ones in which the randomization of treatments to plots is such that on the null hypothesis the mean of the Treatments M.S.'s under all possible randomizations is equal to the mean of the Error M.S.'s.* This ensures that the Treatments M.S. is able to be tested against a comparable Error M.S., or, in other words, *that the estimate of error is unbiased.*

Notice that there are two elements in ensuring an unbiased estimate of error—the allocation of treatments to plots must be random and the design must belong to a class of valid experimental designs. Randomization is not alone sufficient, but it is an essential element. We shall discuss only valid designs.

12.5.4 It follows that, if the analysis of variance is applied to data from a systematic design, the estimate of error will be biased. This alone does not condemn the use of systematic designs, but it does condemn the use of analysis of variance in analysing the data from such designs, which Kempthorne has described as "pure opportunism". Actually, the defect of all designs involving balancing is that there is no method of analysis which is theoretically acceptable and which will provide an unbiased estimate of the errors of treatment means. Hence the great argument against randomization as opposed to systematic designs, that randomization increases the error of treatment

means compared with balancing techniques, is seen to have no substance in view of the impossibility of being able to assess the lower errors of the systematic designs. Moreover, as we shall see later, experimental designs are available in which the error can be kept low, while still not forfeiting the essential element of randomization.

12.5.5 A great advantage of randomization (in conjunction with replication) is that it makes possible the estimation of an unbiased estimate of experimental error in each individual experiment. It is therefore unnecessary to assess treatment differences in relation to past experience, for example, of experimental error with the crop concerned. This is just as well, since experimental error can vary widely according to season and locality.

12.6 Randomization and the independence of the errors of plot yields

12.6.1 It has already been pointed out that adjacent plots under identical treatment tend to yield alike, i.e. the errors of adjacent plots tend to be alike. Quite clearly this is contrary to the requirement that the ϵ's in Model I should be statistically independent; instead, individual plot deviations result from a component due to chance causes upon which is superimposed a strongly systematic and probably dominant component due to soil heterogeneity. Furthermore, the plots are fixed on the ground, so that the systematic nature of the variability cannot be broken up by shifting the plots round as one might shift a group of animals, for example.

12.6.2 If, however, the allocation of treatments to plots is made by a process of randomization, it becomes a matter of pure chance which of the set of systematic components due to soil heterogeneity becomes associated with any particular treatment. It is just as if the plots have been shuffled into a random order on the ground; the systematic nature of the variability due to soil heterogeneity is effectively broken up, and it becomes possible to regard the components of error from this source as independent chance errors, which are intermingled with independent chance errors from other sources in the experimental error. This accords with our assumption that the ϵ's in Model I are independent. Should there be any other source of error which tends to vary systematically according to plot position on the ground (cf. § 2.8.7), randomization will have a similar effect.

12.6.3 In a field experiment, therefore, each error component, ϵ_{ij}, in the model may be regarded as the sum of two independent components, the one associated with a particular plot due to position on the ground, differential soil fertility, etc., and allotted to a particular treatment by the randomization process, the other the resultant of a multiplicity of purely chance errors, which would occur even if it were physically possible to observe two yields on the same plot under identical conditions. If the variance of the former is denoted by σ_P^2 (P for plot) and the variance of the latter by σ_D^2 (D for duplicate) then, because the two components are independent we may regard the error variance σ^2 as equal to $\sigma_P^2 + \sigma_D^2$. As indicated in § 12.1.4, it is impossible to estimate σ_P^2 and σ_D^2 separately.

12.6.4 Independence of the ϵ's in Model I ensures unbiased estimates of treatment effects and of the error variance, advantages which we have previously noted as being associated with randomization. The effect of randomization on the errors of the plots of a valid experimental design is therefore sufficient guarantee of the unbiasedness of the estimates of the treatment effects and of experimental error. On the other hand, where randomization is not possible or is not done, there is no such guarantee. In particular, successive readings on the same plots, e.g. yields in successive years, will not be independent. Another example is given in § 13.8.3.

12.7 Other assumptions in the analysis of variance in relation to field experiments

12.7.1 *Normal distribution.* In a simple random field experiment the yields for each treatment constitute a random sample from a finite population of yields. If we suppose that this finite population is itself a sample from some infinite population of yields and if this infinite population is N.D., then the yields for each treatment are random samples from a normal distribution, i.e. the ϵ_{ij} are N.I.D. as postulated in Model I.

That this is seldom far from the truth in respect of crop yields is evident from the results of many uniformity trials. The frequency table of the data in Table 4.1, for example, has an appearance consistent with being a random sample from a normal population, and we shall demonstrate statistically in Chapter 15 (Example 15.6) that this is a reasonable hypothesis. Uniformity trial yields will not necessarily be N.D.; for example, if the trial is conducted on a field one half of which is much more fertile than the other, then the frequency distribution of yields would tend to be dimodal. Nevertheless on a reasonably uniform area the yields, when considered in a random order, do seem to bear a close resemblance to a random normal sample.

When the variate is not yield, other distributions are possible (cf. § 7.1.11, for example). It is worth remembering, however, that in an analysis of variance it is only in respect of tests of significance that the assumption of normality is required, and even in this respect these tests are sufficiently robust to withstand some degree of non-normality (e.g. skewness) without any great loss of accuracy. (See, however, § 12.7.2).

12.7.2 *Constant variance.* In Model I it is assumed that the variance of the ϵ_{ij} is the same irrespective of the values of the τ_i. Although in most experimental data this assumption will not be far from the truth, there are occasions where it will almost certainly not be true. Where an experiment is concerned with a single crop and the treatment means do not differ too markedly, there will seldom be cause to examine this assumption, although in a simple random design it is an easy matter to compute sample estimates for comparison if required. If, however, there are treatment differences of the order of 100%, then we may suspect unequal variances. This is because there is in general a tendency for variance to increase with mean, and, although this will be immaterial when treatment differences are small, the opposite is true when treat-

ment differences are large. Where a certain treatment is a complete failure and all yields are zero, the variance of this treatment is zero; on the other hand, a poor treatment may result in partial failure with widely fluctuating yields, causing the variance to be increased as compared with higher yielding treatments.

The remedy in these instances is to reject from the analysis the low-yielding treatment or group of treatments. This may seem drastic, but the technique of analysis of variance is essentially a fine instrument for detecting and determining relatively small differences. We do not need statistics to tell us that a treatment which has failed completely is inferior to the others.

Inequality of variances is also associated with non-normal distributions where the variance is dependent on the mean, e.g. the binomial and Poisson distributions. This situation may be dealt with by a transformation of the variate (see Chapter 24). Since equality of variances is an essential condition for the validity of the analysis of variance, it works out in practice that the assumption of normality is implicated to a greater extent than might be thought from the last paragraph of § 12.7.1.

Experiments involving a number of different crops (e.g. animal greenfeed experiments) should be suspect from the point of view of equality of variances since it is an unlikely assumption that different crops will have the same variance.

12.7.3 *Additivity.* It is implicit in Model I that the treatment effects τ_i are constant for a given treatment irrespective of the fertility of the plots. Thus for any plot under a given treatment the true mean is the result of adding a constant quantity to the over-all true mean μ; the treatment effects are therefore described as additive.

In practice this assumption is unlikely to be strictly true. A fertilizer treatment, for example, is likely to have a greater effect on a plot of low fertility than on one of high fertility.

12.8 Difficulties associated with randomization

12.8.1 From § 12.7 it will have been gathered that the assumptions of normality, equality of variances, and additivity required for an analysis of variance based on Model I are unlikely to be exactly fulfilled by data from a field experiment. We may now add that in reality the same applies also to the independence of plot errors under randomization. Whereas Model I assumes that the ϵ_{ij} are random samples from some hypothetical infinite population, so far as the systematic error components due to soil fertility are concerned, there is actually only a finite population (n) to draw from. Once $n - 1$ of these have been allocated, the last is automatically allocated.

12.8.2 *On the other hand, for a valid experimental design (§ 12.5.3) the act of randomization itself provides, with no assumption other than that of additivity mentioned in § 12.7.3, a completely valid basis not only for estimating treatment effects but also for testing their significance.* The fundamental idea behind these tests is that the yields in the experiment are regarded as a finite

population, and all possible randomizations of these yields among the treatments in accordance with the prescribed design are considered. For an overall test of the significance of treatment differences we would note the number of randomizations in which the Treatments S.S. exceeded or equalled the Treatments S.S. observed with the randomization actually used in the experiment. If this Treatments S.S. is equalled or exceeded in 5% or fewer of all possible randomizations, then the differences between treatments are regarded as significant at the 5% level. Such tests are called randomization tests, based on randomization theory, under which the model for a design is called the finite model. Model I, on the other hand, is called the infinite model for the simple random design, and tests such as the *F*- and *t*-tests are said to be based on infinite or normal-law theory. Randomization tests are an example of non-parametric tests, a name given to tests which involve no assumption in respect of the probability distribution of the variates, such as that they are normally distributed, and hence do not involve the estimation of the parameters of any assumed probability distribution. Earlier examples are provided by the tests of §§ 9.6.8 and 9.8.8, alternatives to the *t*-test which require no assumption of normality.

12.8.3 Proper randomization tests would be impossibly laborious to perform but they are the only ones which stem directly from the data of randomized experiments. Although recent studies have shown certain discrepancies between the two theories, for a valid experimental design (§ 12.5.3) the estimates of treatment effects are the same and it is accepted that the tests of significance of the analysis of variance provide a close approximation to the corresponding tests of randomization theory provided the assumptions of Model I are at least reasonably approximately fulfilled. *This is the real rock on which the principle of randomization rests.*

12.8.4 There is one additional proviso, and that is that the experiment must be "large" enough to make the use of randomization tests possible. To illustrate this latter point, consider an experiment with two replications of two treatments, *A* and *B*, in a simple random design. There are only 6 different ways in which the treatments can be allotted to the 4 plots, and, if (in accordance with a two-tail test) interchanges between *A* and *B* are ignored, only 3 different ways, viz. *ABAB*, *AABB*, and *ABBA*. A randomization test for such a design would therefore have a maximum significance level of 1 in 3 or $33\frac{1}{3}\%$, and would be quite useless for ordinary purposes. With the analysis of variance, however, there are 2 D.F. for error and any level of significance is possible under the assumptions of Model I. There is a close connection between the number of D.F. for error and the "size" of an experiment, and it seems to be generally accepted that *12 D.F. for error is a reasonable minimum* for ensuring an adequate concurrence between randomization tests and the tests of analysis of variance. This rules out certain designs for field experiments, e.g. a simple random design with 2 treatments and fewer than 7 replications. A minimum of 12 D.F. for error is also desirable purely from considerations relating to Model I, viz. to provide a sufficiently precise estimate of the

error variance and to keep the minimum F- or t-ratios for significance at a reasonably low level.

12.8.5 It is an interesting point that in randomized experiments it is necessary to call upon randomization theory to the extent set out in § 12.5.3 even though it is proposed to treat the data by analysis of variance. This is an addition to Model I which is necessary to ensure an unbiased estimate of error, even though Model I is in itself, if exactly complied with, sufficient to ensure this.

12.8.6 Although randomization tests make no assumptions beyond those implicit in the body of the data, it is clear that the conclusions are based solely on the finite population of plots used in the experiment under the conditions of weather, etc. actually encountered; the population, or reference set, from which the data comprise a single sample is the large finite number of all possible randomizations. Under the Model I assumptions there is an appeal to a certain hypothetical infinite population of which the ϵ_{ij} are a random sample. In Example 10.4 it is easy to visualize clearly a hypothetical infinite population of determinations made by each analyst working indefinitely, but how can we specify with any semblance of reality the infinite population of plots postulated in § 12.7.1? It is sometimes stated that, although we make more assumptions using analysis of variance, we can draw wider conclusions, if the assumptions are valid, by regarding the results as true for an infinite population of similar plots under similar conditions. But this argument seems to ignore the fact that this population of plots is not only hypothetical but is also lacking in any semblance of reality, for what we have actually done is to select a rectangle of land and divide it up into the required number of plots. Not far away the same procedure may provide a set of plots for which the parameters τ_i and σ^2 of Model I cannot be regarded as pertaining to the same population. It would therefore appear that the only infinite population of repetitions of the experiment to which we can make justifiable appeal is the set of repetitions of the experiment on the same plots under identical conditions, i.e. all possible randomizations with repetitions allowed. If so, then the supposedly broader conclusions of analysis of variance as opposed to randomization theory vanish, and the generalization of the experimental results can rest on nothing stronger than the considerations set out in § 2.2.2.

12.8.7 We may sum up by stating that, provided the assumptions of § 12.7 are reasonably well fulfilled and there are sufficient D.F. for error, we will never obtain results far from the truth by using the analysis of variance on data from valid randomized designs. Considerations of convenience will accordingly far outweigh the possibility of minor inaccuracies in tests of significance. The analysis of a simple random design will therefore be conducted according to the methods set out in Chapters 10 and 11, and there is in consequence no need to make any fresh demonstration of the method.

12.8.8 It will be appreciated that occasionally the randomization process will generate a design with a strongly regular or systematic allocation

of treatments. Even lay-outs such as *ABCDABCDABCD* . . . . or *AAAAA-BBBBBCCCCC* . . . . have a certain probability of turning up in a simple random design. This is a difficult situation. Theoretically, each arrangement must be allowed an equal probability of being used, but nevertheless one is virtually faced with the certainty that such lay-outs will be affected by fertility trends so that any significant results could easily be due only to the occurrence of the $\frac{1}{20}$ or $\frac{1}{100}$ chance by which odd randomizations are allowed for in the tests of significance. Refuge seems to be generally taken in the low probability with which such extreme randomizations occur in a way which amounts almost to an emulation of the supposed activities of the ostrich. In one type of design at least it has been proposed to admit some restriction on the set of possible randomizations so that certain awkward arrangements will not be permitted. Possibly this idea may be extended. In the meantime the author's advice, despite theory, would be to discard an extreme randomization and draw again. Be sure, however, before doing so, that it is really extreme, and do not start rearranging treatments for better balance, or else the result will be equivalent to a systematic design.

12.8.9 The point may be raised that in a series of identical designs fresh randomization on each occasion as stipulated in § 12.3.5 is really unnecessary because owing to the pattern of soil heterogeneity being different on each occasion the same arrangement is virtually equivalent to a fresh one every time. Apart from the essential connection with randomization theory explained above, a defect in such a proposal is that in one particular randomization the plots of one treatment may be grouped more closely on the ground than those of another, so that, if this arrangement is used throughout, the treatment means will in reality have different errors and tests of significance will be biased if the same S.E. is used for all comparisons. In single experiments allowance is made in the tests for such chance effects of randomization, and in a series of experiments separately randomized an effect of this sort would even out, but, if the same randomization is used throughout, the repetition of the same tendency amounts to a bias.

12.8.10 Sometimes a mistake is made in the field and the correct treatment shown in the randomized plan is accidentally replaced by another. In a simple random design, provided there is no premeditation involved, it would be conceded that an interchange of treatments between two plots merely replaces one randomization by another without prejudice to the original randomization. In more complex designs such an interchange may have more awkward consequences, but, if only the randomization is affected, may be disregarded.

12.8.11 There are circumstances when randomization may be for practical reasons impossible or undesirable. If a series of spray treatments against a crop disease is being tested, the inclusion of a control is essential (see § 11.8.2). Yet, if the control treatments were randomized to the plots along with the other treatments, neighbouring plots might be unfairly subjected to the influence of a strong source of infection close at hand, unless the incidence of

the disease on the controls happens to be slight (which might mean that the experiment had failed). The only reasonable possibility here would seem to be to site the controls a fair distance from the remainder of the experiment—far enough not to be an unfair source of infection, but near enough to show the incidence of the disease on untreated plots in the neighbourhood of the experiment. The control would then be outside the experiment proper and the comparison with the sprayed plots would be on an observational basis, but if the incidence of the disease is as heavy as expected this will not matter very much.

12.9 Principle of realism in field experiments

12.9.1 Any field experiment carried out for the purpose of comparing treatments or varieties *which are thought suitable for adoption into agricultural practices* should comply in all relevant aspects with the ordinary farming methods of the region concerned, except where the latter may themselves be the object of study. For example, if different types of nitrogenous fertilizer are being compared on a small-grain crop, the basis of comparison could be distorted if the fertilizers were broadcast when the general farming practice is to drill the fertilizer with the seed; if, however, the question of the best method of placement were to be raised as a matter to be studied in the experiment, then it would be in order to include such methods as broadcasting in addition to the normal farming method.

12.9.2 The reasons for adopting the above principle are more or less obvious. The difficulty of convincing farmers of results from small-plot experiments has already been mentioned; one type of objection to such trials will be overcome if the plots are treated in every possible respect in the way a farmer would treat them. In that case, also, the experimentalist can be surer that a successful treatment in a small-plot trial will be successful when tried on a large scale. Where trials are conducted on farmers' private land (so-called *co-operative trials*), it is essential for ensuring the continued goodwill of the farmer that the experimental activities interfere as little as possible with the farmer's own operations. Where feasible, therefore, machinery should be used for sowing, cultivating, harvesting, etc. if it is used by the farmer.

12.9.3 Although this principle is primarily one of technique, it does have repercussions on lay-out (size and shape of plots, etc.) and hence indirectly on design. Plot size, for example, has to be adapted in relation to the type of machinery used.

12.10 Experiments other than field experiments

12.10.1 *Generic terms.* It was mentioned in Chapter 2 that we would deal mainly with field experiments. This is convenient because not only is this the most familiar type of application, but also it was in this type of application that the principles of experimental design were originally developed. For this reason it is common to use the words "plot" and "yield", especially the

latter, as generic terms. Nevertheless, the need for alternative generic terms has brought into common use "experimental material" for "soil" (plus planting material), and "experimental unit" for "plot"; for "yield" we may use "observation" or "determination", and, if these seem less precise, it must be remembered that yield is not always the variate even in field experiments (cf. § 6.12.2). Experimental units are defined as those units of the experimental material to which treatments are allocated in consequence of each single act of randomization, e.g. the drawing of a single random number. If chickens in 8 houses are allotted two rations, A and B, in such a way that all chickens in 4 of the houses chosen at random receive A and all chickens in the other 4 houses receive B in a simple random design, then houses are the experimental unit, not chickens, and there will be only 6 D.F. for error. If there were only 2 houses, one receiving A and the other B, a test of significance in the manner of Example 9.4 (regarding individual chickens as the unit) would not be a valid test of the difference between rations, since this difference is indistinguishable from the difference between the houses. The experiment is then only an observational one, even if the chickens were randomly allocated to houses. A comparison with the experiment of Exercise 9.7 is of interest here.

12.10.2 The particular difficulties of a statistical nature associated with field experimentation may be summarized as under:

(1) It is impossible to carry out different treatments on the same unit (plot) under similar or identical conditions. The crop takes a season to grow and the following season will present a different set of conditions altogether. In an industrial experiment, however, with machines as experimental units and various settings of the machine as treatments, the treatments can be applied to the same unit in rapid succession under conditions which can possibly be made identical or which do not vary much.

(2) Repetitions of the experiment in time must coincide with separate seasons, whereas in industrial experimentation, for example, repetitions can be carried out as quickly as the degree of organization required permits.

(3) The experimental units are subject to a strongly systematic source of variation.

(4) There is a limitation on the number of experimental units available.

12.10.3 Once we get away from field experiments, agricultural experimentation is not necessarily tied down by the first two difficulties. Thus, a possible design for testing the differential effect of two rations on the milk yields of cows as in Examples 9.4 and 9.6 might have been to give each cow short equal periods on both rations in turn, the yields being observed for each period separately with an interval of one or two weeks between the periods for the cows to settle down on the new ration. Provided the periods were not too long, the experiment could possibly be repeated on the same cows during the same lactation, or else a second group of cows in milk could be started on a repetition of the experiment at any convenient time.

12.10.4 *Replication.* For reasons set out in § 12.2, replication is essential in all types of experiment. Even if the material is perfectly homogeneous as in Example 10.4, where it is chemically impossible for the actual iron concentration of the 50 samples to vary, there are still measurement errors in the determinations. In biological experimentation the material must be expected to be heterogeneous. On the other hand, the fourth difficulty associated with field experiments may have no relevance, and this may make possible a relatively large number of replications, if the cost per determination is not a serious limiting factor.

12.10.5 *Independence of errors.* The caution relating to non-independence of successive readings on the same unit (§ 12.6.4) is still applicable, e.g. weights of the same animals taken on several occasions. In other words the animal is the unit, not the weights.

12.10.6 *Randomization.* With perfectly homogeneous material there is theoretically no need to allocate the treatments to the units at random. Thus in Example 10.4, which is formally a simple random design, the iron concentration of the samples is identical, and so it does not matter how the allocation to analysts was made. Similarly, when chemical substances, for example, can be so well mixed that samples from the mixture are completely homogeneous, it is virtually immaterial how the samples are allocated to different treatments. It is for this sort of reason that the chemist and physicist do not have to bother much about proper randomization; with their material σ_P^2 (§ 12.6.3) is zero or negligible.

In most cases, however, the experimental material is not homogeneous and the experimentalist may even be acutely aware of initial measurable variation in the units. Randomization brings all such causes of variation "under statistical control", i.e. they contribute to chance errors only, with the result that the applied treatments are a "unique special condition" (cf. § 8.1.4) to which a significant deviation from H_0 may be attributed. On the other hand, where sex is a "treatment", as in many animal experiments, no randomization is, of course, possible. Consequently, "sex" may not be a unique special condition; the two sexes may have been previously subjected to different environments which may be the cause of observed treatment effects. Similarly with survey data (cf. Example 10.5 and Exercise 9.3) the units occur in the different categories of a classification because of their particular properties and not because of a randomization process; consequently, there can be no guarantee that observed differences between categories arise from the ostensible source (cf. § 18.10.1).

In particular, it would in general be highly dangerous to assume the absence of any systematic component of variation affecting the experimental material, and the allocation of treatments to units should be random.

For example, if a group of cows is to be allocated to a set of treatments (r cows per treatment) it might be thought that it would be in order to allocate the first r cows passing through a gate to one treatment, the next r to a second treatment, and so on, on the grounds that the order of presentation at the

gate is itself a randomization. Such an assumption ignores the fact that cows are known to observe certain social distinctions, and unknown to the experimentalist the order of presentation at the gate may correspond to some grouping or other which might entirely vitiate the experiment.

Similarly with a trial of different settings of a machine in succession, there may be some systematic change in conditions, such as a rise in temperature during the duration of the experiment. In the example of § 12.10.3 it is known that after a certain interval the milk yields of a lactating cow follow a decreasing trend, so that, if each cow received a particular ration first, the treatment difference would be confounded with the natural fall in yield over the duration of the experiment. *A remedy in all these and other cases, where systematic trends are known or suspected to exist in the experimental material, is to randomize. Even where no such trend is suspected, the act of randomization serves as an insurance in respect of the validity of the experiment. When in doubt randomize. Even where there is no cause for doubt, randomize! Where any allocation of experimental units is made, ask yourself whether the allocation is random: if it is not, then randomize!*

12.10.7 *An example of the effects of failure to randomize.* The oxygen uptake of a sample of poultry semen may be measured by the fall of pressure in a Warburg flask into which an aliquot from a prepared suspension has been pipetted. Owing to the rapid ageing of the semen, errors with this method have to be determined by using a number of flasks simultaneously, heated to constant temperature on the same water-bath. Since the flasks cannot be made absolutely identical, a flask constant (different for each flask and based primarily on volume) is used to convert the recorded fall in pressure into the required oxygen uptake. In an experiment in which a number of different samples were being compared, using the same flasks in succession, the analysis of variance showed highly significant differences between flasks, which led to a suspicion that the flask constants had been wrongly determined. It was revealed, however, that in pipetting from the suspension the first aliquot was always allocated to one particular flask, the second to another particular flask, and so on. It was also found that as the pipetting proceeded the density of the suspension became lower as the particles settled. This systematic variation, coupled with a failure to randomize the allocation of aliquots to flasks, was the reason for the significant differences between flasks. In practice the trouble was easily overcome by stirring the suspension prior to withdrawing each aliquot, but such a convenient alternative to randomization is not always possible.

12.11 Size and shape of plots in field experiments

12.11.1 The experimental area should be selected so as to be as uniform as possible. Selection of an uneven site in no way adds to the generality of the results, but will increase the experimental error and militate against the obtaining of significant results. It may not be an easy matter to judge the uniformity of a given area, but in some kinds of experiments with fruit trees,

for example, the trees are often well established before treatments are applied, and uniformity yields or other measurements may well be available.

12.11.2 In these circumstances, provided no border effects are anticipated, it would be feasible to form plots made up of groups of trees chosen so that differences between the averages of the groups are as small as possible; thus one plot might consist of trees from different parts of the experimental area. Such a procedure is perfectly legitimate and is not to be confused with the balanced designs of §§ 12.3.3 and 12.3.4; the treatments would still be allotted randomly to the plots thus demarcated. On the other hand, unless the trees were reasonably contiguous (though the plots might still have odd shapes), this would not be a very convenient arrangement. Normally plots consist of rectangles or squares. Naturally they must be of equal area within a single experiment. They should also be of consistent shape and similarly oriented, i.e. rectangular plots must always have their longer side running in the same direction. The reason for this lies in the assumption of constant variance set out in § 12.7.2; the fertility pattern of a field is often such that the variability of rectangular plots laid out in one direction is different from that of plots of the same size and shape laid out in a direction at right angles. Although there is no assurance that the fertility pattern will be consistent, it is a more reasonable assumption to make about a relatively small area than that the direction of rectangular plots makes no difference to their variability.

12.11.3 Uniformity trials of the second type (§ 12.1.2) may be used to compare different sizes and shapes of plot. The small rectangular units harvested separately are used as a basis for "building up" plots equal in size to some multiple of the original units and of various rectangular shapes, the yields of the original units being added to give the yields of the plots so constructed. On these different types of plots are superimposed dummy designs of representative types, the data for which are analysed to give the C.V.'s, which are traditionally used in this connection to compare the efficiency of the different types of plot. Masses of uniformity trials of this sort have been carried out with a great variety of crops in attempts to specify standard or optimal sizes and shapes of plots for different crops. It must not be forgotten, however, that the results of such a trial are applicable only to the type of fertility pattern encountered in the field concerned. Also factors other than soil heterogeneity have to be taken into account. Nevertheless, certain trends are apparent which can be confirmed by theoretical arguments as well.

12.11.4 It is usually found that the greatest C.V. is obtained with the smallest plot sizes and that the C.V. diminishes as plot size is increased—rapidly at first and then more slowly. A typical idealized graph is shown in Figure 12.1. The optimal plot size can be fixed as that size beyond which any further increase causes very little diminution of the C.V. It has been objected that, if this determination is made from a graph, the apparent optimal value can depend on the scales used for the axes. However, if it is kept in view that a C.V. of 10% is a desirable figure (cf. § 6.12.2), one is not likely to go far wrong. Thus, if doubling the plot size gives a decrease in C.V. from 10% to

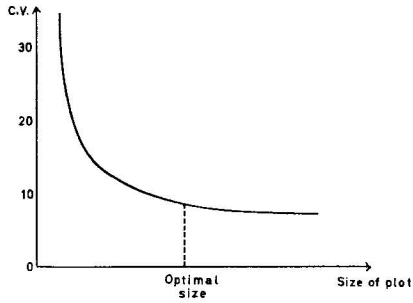

Figure 12.1: Relationship between C.V. and plot size.

9%, then the larger plot is probably not worth while because the same land could be used for doubling the replication, and thus decreasing the S.E. of a treatment mean in the ratio $\sqrt{2} : 1$.

12.11.5 Plot sizes commonly range between about $\frac{1}{20}$ acre and $\frac{1}{100}$ acre, but larger and smaller sizes are also used. Large plots may be required for experiments involving the grazing animal owing to the necessity for having at least a few animals on each plot for the sake of company since single animals may behave abnormally (cf. § 2.7.3). The author has seen experiments on carrots with plots of 1 square yard or less which have given a C.V. as low as 7%; this is probably exceptional, however, since one would expect a C.V. of the order of 20% with such small plots.

12.11.6 Size of plot has to be considered in conjunction with limitations such as the amount of land available, shortage of seed, labour costs, etc. *For a given area of land* it will usually be found that the smallest possible plot is the most efficient in the sense that it will give the lowest S.E. for a treatment mean (cf. § 12.2.1). This is because, although the C.V. is high, the number of replications is so large. With uniformity trial data it is often found that even the very large drop in C.V. which accompanies the doubling of the smallest plot size when plots made up of two original units are considered, does not offset the halving of the replications. Such small plots must, of course, be rejected on other grounds. For example, the labour of harvesting and recording so many separate plots would be very great; if border areas are necessary (cf. § 2.8.3), these will take up too large a proportion of the area planted, and, in fact, the validity of the statement above based on a fixed area of land would be upset if border areas had to be taken into account; also the principle of realism (§ 12.9) would require larger plots, especially if machine operations come into the picture.

12.11.7 In practice the area of land available is not more than occasionally likely to be a critical factor, since, unless very large plots are to be used, an experiment is seldom so large that the question of the land taken up becomes really pressing. We need to have adequate D.F. for error (20 D.F. or more are desirable), but the number of replications required for this diminishes as the number of treatments increases, so that the total number of

plots required remains fairly constant. An average experiment entails 40–60 plots; one with 100 plots may be regarded as fairly large. For 2 treatments then, 20 replications might be considered desirable, but for 5 treatments, only 10. Usually we would have in mind a plot size within the range suggested in § 12.11.5, and, should the question of total land area need to be considered, it would probably be feasible to adjust the size of plot while still keeping within this range. Of course, if we use 48 plots for an experiment with 8 treatments as well as for an experiment with 6 treatments, the S.E. of a treatment mean in the former case will be larger owing to the fewer replications, but there has to be a compromise somewhere since experiments cannot be so large that they become unwieldy to manage. Also, although, for example, we may halve the S.E. of a treatment mean by taking 12 replications in place of 3, the addition of a further 9 replications does not bring commensurate gains since actually a further 36 replications would be needed to halve the S.E. again. Once adequate replication has been decided upon and assuming plenty of land is available, there would be little point in increasing the plot size beyond the optimal for reasons explained in § 12.11.4; in addition, as the total area increases, so the experimental area may become less uniform—one cannot go on indefinitely. However, large plots may be convenient for some other reason; in New Zealand, for example, the standard length of plot in co-operative wheat trials was increased from 2 chains to 3 chains primarily because the greater length is more suitable for harvesting by a combine or header-harvester. In such cases convenient arrangements for bagging and weighing the produce of the enlarged plot also need consideration.

12.11.8 The objective in siting plots on the ground is to make them as alike in fertility as possible so that the experimental error will be as low as possible. This will be aided by the availability of uniformity data (even if we do not go beyond a simple division of the experimental area into rectangles), but usually no such data are available, and we may have no better guide than vague ideas about possible fertility gradients, though on experimental farms the previous history of the area may prove helpful. In these circumstances theoretical considerations may be able to assist the choice of a suitable plot shape, since the shape of rectangular plots, i.e. whether long and narrow, moderately oblong, or square, can influence their variability. It is intuitive to consider long narrow plots lying alongside as likely to be less variable than square plots of the same area on the grounds that the centres of long narrow plots are closer together than the centres of square plots. A long narrow plot is likely to slice through patches of varying fertility which it will tend to share with its neighbours lying alongside. A moderate number of long narrow plots lying alongside one another are therefore likely to be of similar fertility; with square plots, however, one plot may occupy largely a patch of high fertility and another largely a patch of low fertility. Nevertheless, square plots are not always more variable than oblong plots. When a fertility gradient is present, it is possible for long narrow plots to offer the extremes of variability for a given plot size. Thus in diagram (*a*) of Figure 12.2 the plots share equally

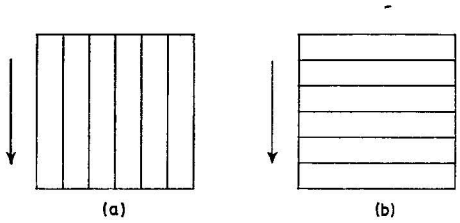

Figure 12.2: Long narrow plots with a fertility gradient.

in the fertility changes (the direction of which is indicated by the arrow) and will be of similar fertility. On the other hand in diagram (*b*) the plots will tend to differ in fertility. Square or squarish plots *on the same area* would tend to occupy an intermediate position as regards variability.

12.11.9 Where the direction of a fertility gradient is known, therefore, it would be advisable from the point of view of low experimental error to have long narrow plots in the direction of the gradient as in diagram (*a*), though actually the superiority of (*a*) over (*b*) is not usually as marked as might be expected from the theoretical argument, amounting to a drop of, say, from 14% to 12% in the C.V. However, it is not always easy to detect beforehand (in default of any previous history of the area) the direction of even quite marked fertility gradients which may later be readily apparent, and, of course, the direction of fertility changes is not just as straightforward as seems to be indicated by the theoretical representation by arrows. In the expectation that a fertility gradient will be present, and in the absence of any certainty about its direction, an experimenter may prefer to adopt a middle course and use squarish plots.

12.11.10 In practice other considerations can affect the choice of shape of plot and have to be taken into account. Of these the most important is that square plots are unsuitable for machinery since several short runs have to be made over a plot, whereas if the width of an oblong plot is suitably chosen only a single run will be necessary. Thus the standard plot for co-operative wheat trials in New Zealand is 49 in. × 3 chains, the width corresponding to 7 coulters 7 inches apart and being such that the seed and fertilizer can be drilled by traversing the plot only once. A special 7-coulter drill is used (the commercial drill being much larger), and only the central 5 coulters are ordinarily counted for the plot yields. Oblong plots are therefore used for convenience despite the possibility of an unfortunate fertility gradient as in diagram (*b*) of Figure 12.2, unless the plots are going to be worked entirely by hand, in which case square plots may be preferred for the reason explained in § 12.11.9. Even with plots worked by hand, however, there is some advantage with row crops in having rows parallel to the longer side of an oblong plot.

Another point in this connection is that long narrow plots lend themselves to a field lay-out such as that in Figure 12.3, which is exceptionally convenient where machinery is to be used since there can be unlimited turn-

ing space at the end of the plots without any necessity for leaving pathways between plots.

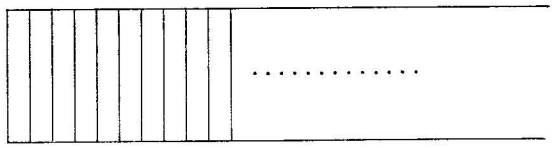

Figure 12.3: Possible lay-out with long narrow plots.

12.11.11 A common exception to the rule that oblong plots should have their longer sides parallel to the fertility gradient occurs when the experiment is placed on sloping ground. The fertility gradient usually runs down the slope, but, if long narrow plots were to be placed parallel to the slope, soil washaways would be encouraged. Also where plots are placed on contoured land, it may be impossible to keep the plots similarly oriented owing to the curve of the contour strips, but this need not cause concern since it is a reasonable assumption that the fertility gradient keeps at right angles to the contour.

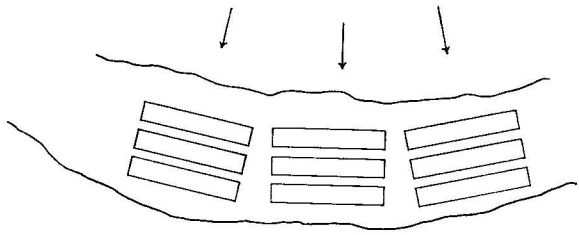

Figure 12.4: Arrangement of plots on a contour strip.

12.11.12 Lastly there is the question of border areas. A square plot possesses an advantage here since for a given area of plot, a square plot has the smallest perimeter. Hence square plots need the least amount of border areas, if these are necessary. With row crops a single-row plot may have to be bordered by rows on either side, which means that only about $\frac{1}{3}$ of the area would contribute to the experimental yields. From the point of view of the principle of realism, too, excessively narrow plots may not be desirable.

12.11.13 It is interesting to note some minor difficulties which arise in experiments where the treatments include different spacings between rows or between plants within rows. Ordinarily with row crops it is convenient to regard the gross plot as consisting of so many rows from which certain edge rows are rejected to give the net plot. (End plants will have to be rejected as well, but let us confine attention to the rows.) If the treatments are spacings between rows, e.g. 15 in. and 18 in., and the gross plot sizes are equal (which can be achieved with plots 90 in. wide), then the net plot sizes will not be equal. If this inequality is not very pronounced it may be possible to break the rule about equal plot sizes and adjust the yields by simple proportion so

that they can be treated on an equal basis; this is preferable to varying the length of the plot to equalize the area and so interfering with the shape of the plots.

It is better, however, to start with equal net plots, taking care that the correct competition according to the spacing treatment is given to each plot by the border rows between the plots. Thus in the example above, if there are no other treatments such as fertilizers, etc. involved, it would be possible to have only a single border row 15 in. from the one (net) plot and 18 in. from the other. Occasionally there may be difficulty in obtaining a suitable plot width to ensure equal areas. For example, with 12 in., 15 in., and 18 in. spacings the minimum plot width for equal net plots is 180 in., accommodating 15, 12, and 10 rows, respectively. A possible remedy if these plots are considered too wide would be to make the net plots (say) 72 in. wide with 6 rows at the 12 in. spacing, 4 at the 18 in., and 5 rows at 14·4 in. instead of 15 in. The use of this device means that plots will have to be very accurately laid out. The objections that a 14·4 in. spacing would not be used in practice and that it is not exactly the required 15 in. are relatively trivial.