48

THE ARRANGEMENT OF FIELD EXPERIMENTS

Author's Note (CMS 17.502a) *

From about 1923 onwards the Statistical Department at Rothamsted had been much concerned with the precision of field experiments in agriculture, and with modifications in their design, having the dual aim of increasing the precision and of providing a valid estimate of error.

These two desiderata had been somewhat confused in the minds of experimenters, and the present paper was the author's first attempt at setting out the rational principles on which he might proceed. The paper is a precursor to the book on the Design of Experiments published nine years later.

Journal of the Ministry of Agriculture of Great Britain, 33: 503-513, (1926).

THE ARRANGEMENT OF FIELD EXPERIMENTS

R. A. FISHER, SC.D.,
Rothamsted Experimental Station.

The Present Position.—The present position of the art of field experimentation is one of rather special interest. For more than fifteen years the attention of agriculturalists has been turned to the *errors* of field experiments. During this period, experiments of the uniformity trial type have demonstrated the magnitude and ubiquity of that class of error which cannot be ascribed to carelessness in measuring the land or weighing the produce, and which is consequently described as due to " soil heterogeneity " ; much ingenuity has been expended in devising plans for the proper arrangement of the plots ; and not without result, for there can be little doubt that the standard of accuracy has been materially, though very irregularly, raised. What makes the present position interesting is that it is now possible to demonstrate (*a*) that the actual position of the problem is very much more intricate than was till recently imagined, but that realising this (*b*) the problem itself becomes much more definite and (*c*) its solution correspondingly more rigorous.

The conception which has made it possible to develop a new and critical technique of plot arrangement is that an estimate of field errors derived from any particular experiment may or may not be a valid estimate, and in actual field practice is usually not a valid estimate, of the actual errors affecting the averages or differences of averages of which it is required to estimate the error.

When is a Result Significant ?—What is meant by a valid estimate of error ? The answer must be sought in the use to which an estimate of error is to be put. Let us imagine in the broadest outline the process by which a field trial, such as the testing of a material of real or supposed manurial value, is conducted. To an acre of ground the manure is applied ; a second acre, sown with similar seed and treated in all other ways like the first, receives none of the manure. When the produce is weighed it is found that the acre which received the manure has yielded a crop larger indeed by, say, 10 per cent. The manure has scored a success, but the confidence with which such a result should be received by the purchasing

public depends wholly upon the manner in which the experiment was carried out.

The first criticism to be answered is—" What reason is there to think that, even if no manure had been applied, the acre which actually received it would not still have given the higher yield ? " The early experimenter would have had to reply merely that he had chosen the land fairly, that he had no reason to expect one acre to be better than the other, and (possibly) that he had weighed the produce from these two acres in previous years and had never known them to differ by 10 per cent. The last argument alone carries any weight. It will illustrate the meaning of tests of significance if we consider for how many years the produce should have been recorded in order to make the evidence convincing.

First, if the experimenter could say that in twenty years experience with uniform treatment the difference in favour of the acre treated with manure had never before touched 10 per cent., the evidence would have reached a point which may be called the verge of significance ; for it is convenient to draw the line at about the level at which we can say : " Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." This level, which we may call the 5 per cent. point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials. To locate the 5 per cent. point with any accuracy we should need about 500 years' experience, for we could then, supposing no progressive changes in fertility were in progress, count out the twenty-five largest deviations and draw the line between the twenty-fifth and the twenty-sixth largest deviation. If the difference between the two acres in our experimental year exceeded this value, we should have reasonable grounds for calling the result significant.

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard of significance at the 5 per cent. point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance. The very high odds sometimes claimed for experimental results should usually be discounted, for inaccurate methods of estimating error

have far more influence than has the particular standard of significance chosen.

Since the early experimenter certainly could not have produced a record of 500 years' yields, the direct test of significance fails; nevertheless if he had only ten previous years' records he might still make out a case, if he could claim that under uniform treatment, the difference had never come *near* to 10 per cent. His argument is now much less direct; he wishes to convince us that such an error as 10 per cent. would occur by chance in less than 5 per cent. of fair trials, and he can only appeal to ten trials. On the other hand, for those ten years he knows the actual value of the error. From these he can calculate a standard error, or rather an estimate of the standard error, to which the experiment is subject; and, if the observed difference is many times greater than this standard error, he claims that it is significant. At how many times greater should he draw the line? This factor depends on the amount of experience upon which the standard error is based. If on ten values, we look in the appropriate published table for "the 5 per cent. value of t, when $n=10$" and find (1 p. 137) the value 2·228. If, then, the standard error is only 3 per cent., the 5 per cent. point is at 6·684 per cent., and we can admit significance for a difference of 10 per cent.

If we thus put our trust in the theory of errors, all the calculation necessary is to find the standard error. In the simple case chosen above (in which, for simplicity, it is assumed that each of the two acres beats the other equally often) all that is necessary is to multiply each of the ten errors by itself, thus forming its square, to find the average of the ten squares and to find the square root of the average. The average of the ten squares is called the variance, and its square root is called the standard error. The procedure outlined above, relying upon the theory of errors, involves some assumptions about the nature of field errors; but these assumptions are not in fact disputed, and have been extensively verified in the examination of the results of uniformity trials.

Measurement of Accuracy by Replication.—It would be exceedingly inconvenient if every field trial had to be preceded by a succession of even ten uniformity trials; consequently, since the only purpose of these trials is to provide an estimate of the standard error, means have been devised for obtaining such an estimate from the actual yields of the trial year.

The method adopted is that of replication. If we had challenged, as before, the result of an experiment performed, say, ten years ago, we should not probably have been referred to the experience of previous years, but should have learnt that each trial acre was divided into, say, four separate quarters ; and that the two acres were systematically intermingled in eight strips arranged ABBAABBA, where A is the manured portion, and B the unmanured.*

Besides affording an estimate of error such intermingling of experimental plots is of value in diminishing the actual error representing the difference in actual fertility between the two acres. For it is obvious that such differences in fertility will generally be greater in whole blocks of land widely separated, than in narrow adjacent strips. This important advantage of reducing the standard error of the experiment has often been confused with the main purpose of replication in providing an estimate of error ; and, in this confusion, types of systematic arrangement have been introduced and widely employed which provide altogether false estimates of error, because the conditions, upon which a replicated experiment provides a valid estimate of error, have not been adhered to.

Errors Wrongly Estimated.—The error of which an estimate is required is that in the difference in yield between the area marked A and the area marked B, *i.e.*, it is an error in the difference between plots treated differently in respect of the manure tested. The *estimate* of error afforded by the replicated trial depends upon differences between plots treated alike. An estimate of error so derived will only be valid for its purpose if we make sure that, in the plot arrangement, pairs of plots treated alike are not nearer together, or further apart than, or in any other relevant way, distinguishable from pairs of plots treated differently. Now in nearly all systematic arrangements of replicated plots care is taken to put the unlike plots as close together as possible, and the like plots consequently as far apart as possible, thus introducing a flagrant violation of the conditions upon which a valid estimate is possible.

One way of making sure that a valid estimate of error will be obtained is to arrange the plots deliberately at random,

* This principle was employed in an experiment on the influence of weather on the effectiveness of phosphates and nitrogen alluded to by Sir John Russell (3). The author must disclaim all responsibility for the design of this experiment, which is, however, a good example of its class.

so that no distinction can creep in between pairs of plots treated alike and pairs treated differently; in such a case an estimate of error, derived in the usual way from the variations of sets of plots treated alike, may be applied to test the significance of the observed difference between the averages of plots treated differently.

The estimate of error is valid, because, if we imagine a large number of different results obtained by different random arrangements, the ratio of the real to the estimated error, calculated afresh for each of these arrangements, will be actually distributed in the theoretical distribution by which the significance of the result is tested. Whereas if a group of arrangements is chosen such that the real errors in this group are on the whole less than those appropriate to random arrangements, it has now been demonstrated that the errors, as estimated, will, in such a group, be higher than is usual in random arrangements, and that, in consequence, within such a group, the test of significance is vitiated. It is particularly to be noted that those methods of arrangement, at which experimenters have consciously aimed, and which reduce the real errors, will appear from their (falsely) estimated standard errors to be not more but less accurate than if a random arrangement had been applied; whereas, if the experimenter is sufficiently unlucky, as must often be the case, to *increase* by his systematic arrangement the real errors, then the (falsely) estimated standard error will now be smaller, and will indicate that the experiment is not less, but more accurate. Opinions will differ as to which event is, in the long run, the more unfortunate; it is evident that in both cases quite misleading conclusions will be drawn from the experiment.

A Necessary Distinction.—The important question will be asked at this point as to whether it is necessary, in order to obtain a valid estimate of error, to give up all the advantage in accuracy to be obtained from growing plots, which it is desired to compare, as closely adjacent as possible. The answer is that it is not necessary to give up any such advantage. Two things are necessary, however: (a) that a sharp distinction should be drawn between those components of error which are to be eliminated in the field, and those which are not to be eliminated; and that while the elimination of the one class shall be complete, no attempt shall be made to eliminate the other; (b) that the statistical process of the estimation of error shall be modified so as to take account of

the field arrangement, and so that the components of error actually eliminated in the field shall equally be eliminated in the statistical laboratory.

In reconciling thus the two *desiderata* of the *reduction of error* and of the *valid estimation* of error, it should be emphasised that no principle is in the smallest degree compromised. An experiment either admits of a valid estimate of error, or it does not; whether it does so, or not, depends not on the actual arrangement of plots, but only on the way in which that arrangement was arrived at. If the arrangement ABBAABBA was arrived at by writing down a succession of "sandwiches" ABBA, it does not admit of any estimate of certain validity, although "Student" (2) has shown reasons to think that by treating each "sandwich" as a unit, the uncertainties of the situation are much reduced. If, however, the same arrangement happened to occur subject to the conditions that each pair of strips shall contain an A and a B, but that which came first shall be decided by the toss of a coin, then a valid estimate may be obtained from the four differences in yield in the four pairs of strips. It is not now the "sandwiches" but the pairs of strips which provide independent units of information, and these units are double the number of the "sandwiches."

Moreover, if the experiment is repeated, either by replication on the same field, or at different farms scattered over the country, the arrangement must be obtained afresh by chance for each replication, so that in only a small and calculable proportion of cases will the sandwich arrangement be reproduced.

Thus validity of estimation can be guaranteed by appropriate methods of arrangement, and on the other hand there is reason to think that well-designed experiments, yielding a valid estimate of error, and therefore capable of genuine significance tests, will give actual errors as small as even the most ingenious of systematic arrangements. It is difficult to prove this assertion save by experimenting on the data provided by uniformity trials, because, in the absence of any satisfactory estimate of error, it is impossible to tell for certain how accurate, or inaccurate, such systematic arrangements really are; while the aggregate of the uniformity trial data, hitherto available, is scarcely adequate for any such test. What can be said for certain is, that experiments capable of genuine tests of significance can easily be designed to be

very much more accurate than any experiments ordinarily conducted.

A Useful Method.—The distinction between errors eliminated in the field, and the errors which are to be carefully randomized in order to provide a valid estimate of the errors which cannot be eliminated, may be made most clear by one of the most useful and flexible types of arrangement, namely, the arrangement in "randomized blocks." Let us suppose that five different varieties are to be tested, and that it is decided to give each variety seven plots, making thirty-five in all. It would be a perfectly valid experiment to divide the land into thirty-five equal portions, *in any way one pleased*, and then to assign seven portions chosen wholly at random to each treatment. In such a case, as has been stated above, no modification is introduced in the process of estimating the standard error from the results, for no portion of the field heterogeneity has been eliminated. On most land, however, we shall obtain a smaller standard error, and consequently a more valuable experiment, if we proceed otherwise. The land is divided first into seven blocks, which, for the present purpose, should be as compact as possible; each of these blocks is divided into five plots, and these are assigned in each case to the five varieties, independently, and wholly at random. If this is done, those components of soil heterogeneity which produce differences in fertility *between plots of the same block* will be completely randomized, while those components which produce differences in fertility between different blocks will be completely eliminated. In calculating an estimate of error from such an experiment, care must of course be taken to eliminate the variance due to differences between blocks, and for this purpose exact methods have been developed (1. pp. 176-232).

Most experimenters on carrying out a random assignment of plots will be shocked to find how far from equally the plots distribute themselves; three or four plots of the same variety, for instance, may fall together at the corner where four blocks meet. This feeling affords some measure of the extent to which estimates of error are vitiated by systematic regular arrangements, for, as we have seen, if the experimenter rejects the arrangement arrived at by chance as altogether "too bad," or in other ways "cooks" the arrangement to suit his preconceived ideas, he will either (and most probably) increase the standard error as estimated from the yields;

or, if his luck or his judgment is bad, he will increase the real errors while diminishing his estimate of them.

The Latin Square.—For the purpose of variety trials, and of those simple types of manurial trial in which every possible comparison is of equal importance, the problem of designing economical and effective field experiments, reduces to two main principles (*i*) the division of the experimental area into the plots as small as possible subject to the type of farm machinery used, and to adequate precautions against edge effect; (*ii*) the use of arrangements which eliminate a maximum fraction of the soil heterogeneity, and yet provide a valid estimate of the residual errors. Of these arrangements, by far the most efficient, as judged by experiments upon uniformity trial data, is that which the writer has named the Latin Square.

Systematic arrangements in a square, in which the number of rows and of columns is equal to the number of varieties, such as

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

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

have been used previously for variety trials in, for example, Ireland and Denmark; but the term " Latin Square " should not be applied to any such systematic arrangements. The problem of the Latin Square, from which the name was borrowed, as formulated by Euler, consists in the enumeration of *every possible* arrangement, subject to the conditions that each row and each column shall contain one plot of each variety. Consequently, the term Latin Square should only be applied to a process of randomization by which one is selected at random out of the total number of Latin Squares possible; or, at least, to specify the agricultural requirement more strictly, out of a number of Latin Squares in the aggregate, of which every pair of plots, not in the same row or column, belongs equally frequently to the same treatment.

The actual laboratory technique for obtaining a Latin Square of this random type, will not be of very general interest, since it differs for 5×5 and 6×6 squares, these being by far the most useful sizes. They may be obtained quite rapidly, and the Statistical Laboratory at Rothamsted is prepared to supply these, or other types of randomized arrangements, to intending experimenters; this procedure is considered the

more desirable since it is only too probable that new principles will, at their inception, be, in some detail or other, misunderstood and misapplied; a consequence for which their originator, who has made himself responsible for explaining them, cannot be held entirely free from blame.

Complex Experimentation.—Only a minority of field experiments are of the simple type, typified by variety trials, in which all possible comparisons are of equal importance. In most experiments involving manuring or cultural treatment, the comparisons involving single factors, *e.g.*, with or without phosphate, are of far higher interest and practical importance than the much more numerous possible comparisons involving several factors. This circumstance, through a process of reasoning, which can best be illustrated by a practical example, leads to the remarkable consequence that large and complex experiments have a much higher efficiency than simple ones. No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed.

A good example of a complex experiment with winter oats is being carried out by Mr. Eden at Rothamsted this year, and is shown in the diagram.

Nitrogenous manure in the form of Sulphate (S), or Muriate (M) of ammonia, is applied as a top dressing *early*, or *late* in the season, in quantities represented by 0, 1, 2. When no manure is applied, we cannot, of course, distinguish between sulphate and chloride, or between early and late applications; nevertheless, since the general comparison 0 *versus* 1 dose is one of the important comparisons to be made, the number of plots receiving no nitrogenous manure (corresponding roughly to the so-called "control" plots of the older experiments) are made to be equal in number to those plots receiving one or two doses. This makes twelve treatments, and these are replicated in the above sketch in eight randomized blocks. Note what a "bad" distribution chance often supplies; the chloride plots are all bunched together in the middle of the first block, while they form a solid band across the top block on the right; in the bottom block on the right, too, all the early plots are on one side, and all the late plots on the other.

	2 M EARLY	2 S LATE		2 S LATE			1 S EARLY
1 S EARLY	1M EARLY	1M LATE	1 S LATE	2M EARLY	2M LATE	1M EARLY	1 M LATE
	2M LATE		2 S EARLY		1 S LATE		2 S EARLY
2 S EARLY	2M EARLY		1M LATE		2 S EARLY	2 S LATE	2M LATE
	1 S LATE	1 S EARLY	1 M EARLY	1M LATE			1 S LATE
2 M LATE		2 S LATE		2M EARLY		1 M EARLY	1 S EARLY
2 S EARLY	2M LATE	1 S EARLY	2M EARLY	2 S LATE	2 S EARLY	2 M EARLY	
		1M LATE		1 M EARLY	2M LATE		1M LATE
2 S LATE	1M EARLY		1 S LATE			1 S EARLY	1 S LATE
2M EARLY	1M EARLY	2M LATE	2 S LATE	1 S EARLY			1 S LATE
1 S LATE			1M LATE	1M EARLY	2 S EARLY	2M LATE	
1 S EARLY		2 S EARLY			2M EARLY	2 S LATE	1M LATE

FIG. 1.—A COMPLEX EXPERIMENT WITH WINTER OATS.

The value of such large and complex experiments is that all the necessary comparisons can be made with known and with, probably, high accuracy ; any general difference between sulphate and chloride, between early and late application, or ascribable to quantity of nitrogenous manure, can be based on thirty-two comparisons, each of which is affected only by such soil heterogeneity as exists between plots in the same block. To make these three sets of comparisons only, with the same accuracy, by single question methods, would require 224 plots, against our 96 ; but in addition many other comparisons can also be made with equal accuracy, for all combinations of the factors concerned have been explored. Most important of all, the conclusions drawn from the single-factor comparisons will be given, by the variation of non-essential conditions, a very much wider inductive basis than could be obtained, by single question methods, without extensive repetitions of the experiment.

In the above instance no possible interaction of the factors is

disregarded ; in other cases it will sometimes be advantageous deliberately to sacrifice all possibility of obtaining information on some points, these being believed confidently to be unimportant, and thus to increase the accuracy attainable on questions of greater moment. The comparisons to be sacrificed will be deliberately confounded with certain elements of the soil heterogeneity, and with them eliminated. Some additional care should, however, be taken in reporting and explaining the results of such experiments.

References.—(1) R. A. Fisher : *Statistical Methods for Research Workers.* (Oliver & Boyd, Edinburgh, 1925) ; (2) "Student" : *On Testing Varieties of Cereals.* (*Biometrika*, XV, pp. 271-293, 1923) ; (3) Sir John Russell : Field Experiments : How They are Made and What They are. (*Jour. Min. Agric.*, XXXII, 1926, pp. 989-1001.)