EMPIRE COTTON GROWING CORPORATION

PRINCIPLES AND PRACTICE OF FIELD EXPERIMENTATION

By

J. WISHART, M.A., D.Sc., and H. G. SANDERS, M.A., Ph.D.

School of Agriculture, Cambridge

Published by

THE EMPIRE COTTON GROWING CORPORATION

KING'S BUILDINGS, DEAN STANLEY STREET, LONDON, S.W.1

1935

PRICE 3/-, POST FREE

PREFACE

In 1926 an article entitled "The Principles and Practice of Yield Trials," by F. L. Engledow and G. Udny Yule, was published in *The Empire Cotton Growing Review* (Vol. III., Nos. 2 and 3), and subsequently issued separately by the Empire Cotton Growing Corporation. To a revised edition published in 1930 an appendix was added, stating that an entirely new technique of plot arrangement and field experimentation had been built up by R. A. Fisher, and referring in this connection to various publications for details. A further advance noted was in connection with the technique of sampling a crop. The present publication is on similar lines to the last, but has been entirely recast to incorporate all the improvements in method and practice that have been brought about during the last ten years.

The outstanding developments over this period in statistical science and field experimental methods are due to R. A. Fisher, whose book "Statistical Methods for Research Workers" (Oliver and Boyd, 1925, 5th edition, 1934) has been freely drawn upon. The authors gladly take this opportunity of acknowledging the inspiration they have at all times received from this source and from the author in person. They are specifically indebted to Professor Fisher and his publishers, Messrs. Oliver and Boyd, for permission to reproduce in Appendices I. and II. the two main tables, those of t and z, which are necessary for the statistical analysis of the data of field experimentation.

More than ever is it found necessary for the full interpretation of field experimental data to have recourse to developmental observations taken during the growth of the crop. For this a reliable and accurate sampling procedure is needed. Much progress has been made in the development of adequate methods, which are described in detail, but it must be acknowledged that this is still one of the greatest difficulties facing the experimenter. Experimental work on the problem, in relation to cereal and root crops, is being continued.

CONTENTS

PART I.—PRINCIPLES

PART II.—PRACTICAL CONSIDERATIONS

PRINCIPLES AND PRACTICE OF FIELD EXPERIMENTATION

PART I.—PRINCIPLES

Alice soon came to the conclusion that it was a very difficult game indeed.
ALICE IN WONDERLAND.

1. WHAT IS THE DIFFICULTY?

Much has been written since the first edition of this book was published* to endorse the views therein expressed, that field experimentation was by no means a simple business. Why is this? If we take the simplest case of all, the comparison of two treatments or varieties, it is not enough to lay down side by side two plots, given over respectively to the treatments to be tested, and argue as to results on the basis of the yields or other observable characters of the plots. *Common sense dictates that the plots shall be identical in size and shape, and treated alike in all other respects except the factor to be tested.* Such perfect experimental control is an ideal desideratum, never capable of being fully carried out. The plot dimensions are subject to errors of measurement; determinations of yield have their inevitable errors; the incidence of disease, weather vagaries, insect attack and the like, are not under man's control and cannot be identical for both plots. Furthermore, the inherent soil fertility, known to be an important factor affecting performance, is not constant, for even if the plots are brought as close together as possible by having them long and narrow, they are nevertheless on different sites. These factors, which are susceptible to some degree of control by the experimenter in the care with which he works, but cannot be entirely eliminated owing to chance fluctuations, make up what is known as *experimental error*. If we may begin at once to introduce statistical ideas, we can postulate that a given area has a certain "true" yield under the conditions of the experiment, while the actual yield obtained by measurement is an estimate of the true yield, being subject to an error. The true yield is an abstraction. Technically it is the mean of the "population" of yields, generated by an infinite repetition of the experiment on the area under standard conditions. Because repetition could only in practice be carried out

* Engledow and Yule, "The Principles and Practice of Yield Trials," 1926 (Empire Cotton Growing Corporation).

in successive seasons, with a consequent alteration in essential conditions, we see that we have no means of determining our population exactly. We might try laying down a number of equal plots, all having the same treatment, in a given year. This would give us an idea of the extent of the variation that is possible within the area examined, but it would be a very imperfect representation of our population, not only because the plots we could deal with would be finite in number, but also because a new factor, that of soil fertility variation between plots, would enter in. This factor of *soil heterogeneity* is of very great importance, and it is usual to consider it separately owing to its magnitude and systematic nature, whereas many of the other factors affecting performance are small in size and random in incidence. In fact, as we shall see later on, improvement in experimental method has been brought about by devising special ways of eliminating much of the soil fertility variation from the actual comparisons being made, thus bringing this factor into special prominence.

An experimenter can lessen or average out the experimental errors by taking larger and larger plots, but since he is never concerned with single plots, but rather with the comparison of two or more, as when he is testing different treatments, it is easy to see that to increase the plot size is to increase the distance between the centres of adjoining plots. The factor of soil heterogeneity then assumes greater importance, working in fact in the opposite direction to the beneficial effect of increasing the plot size. It does not follow, therefore, that the experimenter's troubles are at an end when he decides to work with large plots. Much work has been directed towards finding the minimum size for a given crop, consistent with a certain predetermined standard of accuracy, and it will illustrate our point if we quote from Mercer and Hall,* who harvested at Rothamsted in 1910 a " very uniform area " of one acre of wheat in 500 small plots. The results, for the yield of grain, are summarized in the form of a frequency distribution in Table I., by grouping to the nearest one-fifth of a pound.

TABLE I.—YIELD OF GRAIN IN POUNDS FROM PLOTS OF $\frac{1}{500}$ ACRE.

Yield.	*Frequency.*	*Yield.*	*Frequency.*
2·7–	4	4·1–	69
2·9–	15	4·3–	59
3·1–	20	4·5–	35
3·3–	47	4·7–	10
3·5–	63	4·9–	8
3·7–	78	5·1–	4
3·9–	88		500

* *J. Agric. Sci.*, 1911, iv., 107.

The interpretation of these figures, and of the word " frequency," is that 4 plots had yields within the group 2·7 up to but not including 2·9, 15 were within the group 2·9 up to but not including 3·1, and so on. The range of variation, from 2·7 to 5·2, is very great, but admittedly the plots are very small. It happens that the data of the table are exceedingly well fitted by the normal curve of frequency*, with estimated mean 3·95 lb. and standard deviation 0·46 lb., or 11·7 per cent. of the mean.

TABLE II.—YIELDS OF GRAIN IN POUNDS FROM PLOTS OF $\frac{1}{50}$ ACRE (TOTALS OF TEN SMALL PLOTS).

						Total.
	41·1	42·5	40·3	38·5	36·6	199·0
	41·8	40·5	38·3	40·2	38·0	198·8
	40·4	41·9	37·8	40·0	39·5	199·6
	37·8	42·4	37·8	40·3	35·4	193·7
	40·4	42·0	36·7	41·8	38·8	199·7
	39·4	42·7	38·2	39·7	38·5	198·5
	42·8	42·2	38·1	38·0	40·2	201·3
	41·6	40·2	35·5	33·3	35·6	186·2
	41·8	41·4	40·1	34·0	38·1	195·4
	43·4	43·1	42·1	34·5	38·5	201·6
Total	410·5	418·9	384·9	380 3	379·2	1973·8

Now let us take 10 of these small plots together to form a plot of $\frac{1}{5}$ acre, more in keeping with what the authors say as to the best size of plot to adopt for experimental purposes. This can be done from the original data in the paper cited, by taking 5 adjacent plots along the rows and 2 across. The aggregate yields of grain are given in Table II. A much better idea is obtained from this table of the systematic nature of the fertility differences than from Mercer and Hall's original table, owing to the partial smoothing out of the yields of the smaller plots. The marginal totals of Table II. serve as a convenient summary of the fertility variations in two directions at right angles. Fifty is a convenient number from which to calculate estimates of mean and standard deviation without grouping. This is done as follows:

Estimate of true mean $(m) = \bar{x} = S(x)/p$, where
$S(x)$ = sum of all yields,
and p = number of plots.

Thus $\bar{x} = 1973{\cdot}8/50 = 39{\cdot}476$ lb.

* By this we mean that the data follow the Gaussian law whereby the logarithm of the ordinate at any point distant x from the mean is less than that at the mean by a quantity proportional to x^2. It will be impossible to avoid assuming some acquaintance with statistical methods—in the sequel, however, little will be given that is not fully explained.

Estimate of true variance $(\sigma^2) = s^2 = S(x - \bar{x})^2/(p - 1)$,
and $S(x - \bar{x})^2 = S(x^2) - p\bar{x}^2$, where
$S(x^2)$ = sum of squares of all plot yields, and
$S(x - \bar{x})^2$ = sum of squares of deviations of all plot yields from estimated mean.

Thus
$$s^2 = (78224{\cdot}62 - 77917{\cdot}73)/49$$
$$= 6{\cdot}263$$
$s = 2{\cdot}50$ lb. or 6·3 per cent. of the mean.

The comparable figures for standard deviation calculated from Tables I. and II. are the percentages 11·7 and 6·3, and we see that the error has been much reduced by taking together 10 unit plots of $\frac{1}{500}$ acre to form a larger plot for experimental purposes. The reason why the reduction is not even greater will appear shortly. Mercer and Hall obtained the following standard deviations for different sizes of aggregates:

No. of Plots in Block.	*Area (Acres).*	*Standard Deviation as Percentage of Mean.*
1	$\frac{1}{500}$	11·6
2	$\frac{1}{250}$	10·0
4	$\frac{1}{125}$	8·9
10	$\frac{1}{50}$	6·3
10	$\frac{1}{50}$	7·8
20	$\frac{1}{25}$	5·7
50	$\frac{1}{10}$	5·1

Ten unit plots were aggregated in two different ways, one being that shown in our illustration above. The authors concluded that little was to be gained in accuracy by increasing the plot size beyond $\frac{1}{10}$ acre, and they accordingly recommended this as a convenient size. The way in which further reduction of error can be brought about by suitable methods of plot arrangement, so as to eliminate part of the remaining soil heterogeneity, will be described in the following sections.

II.—The Statistical Methods Used.

We have already indicated that we postulate the existence of an infinite population of yields, which may be taken to be normal in form unless we have good reason to think the contrary, and may therefore be summarized by a statement of the mean (m) and variance (σ^2) or standard deviation (σ). From a sample of data, however, we can only make estimates of these quantities, which will be subject to an error of sampling. To distinguish these estimates from the population parameters different symbols may be used, and

it is customary to use $\bar{x}$ (*x bar*) for the arithmetic mean of the sample, which is the appropriate estimate of the population mean, and s^2 for the estimate of variance. With homogeneous material the variation is measured by means of $S(x-\bar{x})^2$. This is the sum of the squares of p quantities which are not independent, being connected by the relation that their sum is zero, since $\bar{x}$ is the sample mean. The quantity $n=p-1$ is known as the number of *degrees of freedom*, being in fact the number of independent squares to which $S(x-\bar{x})^2$ is equivalent, and if we divide by n we have the appropriate estimate of variance, one whose mean value over the infinite number of samples of size p that might be obtained from the population we are considering is equal to σ^2. The estimate of standard deviation is obtained by taking the square root of s^2. Note the method used above of obtaining $S(x-\bar{x})^2$ by squaring and adding the yields as they stand, and subtracting p times the square of the mean. Alternative ways of calculating this last term are to multiply together the total $S(x)$ and the mean $\bar{x}$, or to divide the square of the total by p. Still another way in which to get $S(x-\bar{x})^2$, yielding smaller numbers, is to subtract some round number near the mean, say 40, from all the yields, square and add the deviations, and subtract p times the square of the difference between 40 and the mean 39·476. If the assumed mean is denoted by a, the formula is evidently

$$S(x-\bar{x})^2 = S(x-a)^2 - p(\bar{x}-a)^2.$$

In the last section a was taken as zero.

We now come to the important question of *tests of significance*. The first proposition that is used is the following: If a variable x is normally distributed round a mean m with standard deviation σ, then the mean $\bar{x}$ of a random sample of p items is normally distributed round m as mean with standard deviation $\sigma/\sqrt{p}$. Thus different random samples of size p from the same population will all yield estimates of m, but these will differ from one another, although not to the same extent that the original observations differ, or if by chance a high value of x is included in the sample other values not so high will also be included, and the sample mean will be nearer to the true mean than this extreme value. In practice we do not obtain a large number of random samples, but we can imagine the sampling procedure to be repeated indefinitely, thereby generating a population of sample means which will have the same mean as the original population, and a standard deviation equal to σ divided by the square root of the number of items in the sample. If the original population departs somewhat from normality, we can

still with a fair measure of confidence make use of the proposition, for the population of sample means will be more nearly normally distributed than that of the original values, the more so as the sample size is increased. It follows that the chance of exceeding $\bar{x}$, were the true mean m, is obtained by calculating

$$t_\infty = (\bar{x} - m)\sqrt{p}/\sigma$$

and finding the area under the curve of distribution of $\bar{x}$ from this point to infinity, a thing which is most conveniently done by consulting tables of the *normal probability integral*, of which Table III., adapted from the earlier edition by subtracting the values there given from unity, is an example.

TABLE III.—TABLE SHOWING THE AREA UNDER THE NORMAL CURVE FROM AN ORDINATE BEYOND THAT AT THE MEAN TO INFINITY.

x/σ.	*Area to Infinity.*	x/σ.	*Area to Infinity.*
0	0·5000	1·9	0·0287
0·1	0·4602	2·0	0·0228
0·2	0·4207	2·1	0·0179
0·3	0·3821	2·2	0·0139
0·4	0·3446	2·3	0·0107
0·5	0·3085	2·4	0·0082
0·6	0·2743	2·5	0·0062
0·7	0·2420	2·6	0·0047
0·8	0·2119	2·7	0·0035
0·9	0·1841	2·8	0·0026
1·0	0·1587	2·9	0·0019
1·1	0·1357	3·0	0·0013
1·2	0·1151	3·1	0·0010
1·3	0·0968	3·2	0·0007
1·4	0·0808	3·3	0·0005
1·5	0·0668	3·4	0·0003
1·6	0·0548	3·5	0·0002
1·7	0·0446	3·6	0·0002
1·8	0·0359	3·7	0·0001

The first and third columns of this table give the deviation from the mean expressed in terms of the standard deviation as unit, while the second and fourth give the corresponding areas. Thus, going back to our formula and denoting the required chance by P, we see that when $t_\infty=1$, P=0·16; when $t_\infty=1·64$, P=0·05, and when $t_\infty=2·33$, P=0·01. Such a calculation enables us to work out the chance that a sample with mean $\bar{x}$ or greater has come from a population of specified mean m, and when we consider that the data of the sample are all that we can obtain by experiment, we see the importance of such a calculation as enabling us to assign limits within

which it is likely that the true mean lies. Suppose for the time being that σ is known; then by solving the equations

$$\bar{x} - m_1 = \sigma t_\infty / \sqrt{p} = m_2 - \bar{x}$$

for the unknowns m_1 and m_2, taking $t_\infty = 1{\cdot}64$ corresponding to the probability level 0·05, we obtain a lower limit m_1 such that the chance of exceeding $\bar{x}$, were m_1 the true mean, is 0·05. m_2 lies a corresponding distance on the other side of $\bar{x}$, and these limits have been termed by Fisher the lower and upper *fiducial* 95 per cent. values of m corresponding to the given $\bar{x}$. The probability level chosen is, of course, arbitrary, but is a convenient one.

Usually, however, σ is as much unknown as m, and we have the much more useful proposition now to be stated. The quantity

$$t = (\bar{x} - m)\sqrt{p}/s,$$

where s is the estimated standard deviation, obtained as already described, is distributed, under the conditions previously stated, in a particular frequency curve first discovered by "Student," and called after him the "Student" distribution, or simply the t-distribution. It depends on a parameter n, the number of degrees of freedom (here, though not generally, equal to $p-1$), and only tends to the normal distribution as n approaches infinity, although in practice there is little to choose between the distributions when n is over 30. Tables of the probability integral of this distribution have been constructed by "Student" (*Metron*, v., 1925) and R. A. Fisher ("Statistical Methods for Research Workers"). Table IV. below is a short adaptation of "Student's" table in *Metron*, by kind permission, and shows how, for a given level of probability, the value of t gets less as n becomes greater until in the limit, with $n = \infty$, it becomes equal to t_∞. This is an indication of the greater uncertainty attending our inferences from small samples owing to the inaccuracy of our estimate of σ, while the infinity value shows us that an infinite sample—*i.e.*, the whole population—yields us the exact value of σ. The use of the table is exactly as outlined above for the normal table, except that we enter the column corresponding to the number of degrees of freedom available for estimating σ. Thus for a sample of 11 ($n=10$), t is equal to 1·81 for P=0·05, and this value would be used instead of 1·64 in obtaining fiducial values of m from a sample of this size, if we were content with this level of probability.

still with a fair measure of confidence make use of the proposition, for the population of sample means will be more nearly normally distributed than that of the original values, the more so as the sample size is increased. It follows that the chance of exceeding $\bar{x}$, were the true mean m, is obtained by calculating

$$t_\infty = (\bar{x} - m)\sqrt{p}/\sigma$$

and finding the area under the curve of distribution of $\bar{x}$ from this point to infinity, a thing which is most conveniently done by consulting tables of the *normal probability integral*, of which Table III., adapted from the earlier edition by subtracting the values there given from unity, is an example.

TABLE III.—TABLE SHOWING THE AREA UNDER THE NORMAL CURVE FROM AN ORDINATE BEYOND THAT AT THE MEAN TO INFINITY.

x/σ.	*Area to Infinity.*	x/σ.	*Area to Infinity.*
0	0·5000	1·9	0·0287
0·1	0·4602	2·0	0·0228
0·2	0·4207	2·1	0·0179
0·3	0·3821	2·2	0·0139
0·4	0·3446	2·3	0·0107
0·5	0·3085	2·4	0·0082
0·6	0·2743	2·5	0·0062
0·7	0·2420	2·6	0·0047
0·8	0·2119	2·7	0·0035
0·9	0·1841	2·8	0·0026
1·0	0·1587	2·9	0·0019
1·1	0·1357	3·0	0·0013
1·2	0·1151	3·1	0·0010
1·3	0·0968	3·2	0·0007
1·4	0·0808	3·3	0·0005
1·5	0·0668	3·4	0·0003
1·6	0·0548	3·5	0·0002
1·7	0·0446	3·6	0·0002
1·8	0·0359	3·7	0·0001

The first and third columns of this table give the deviation from the mean expressed in terms of the standard deviation as unit, while the second and fourth give the corresponding areas. Thus, going back to our formula and denoting the required chance by P, we see that when t_∞=1, P=0·16; when t_∞=1·64, P=0·05, and when t_∞=2·33, P=0·01. Such a calculation enables us to work out the chance that a sample with mean $\bar{x}$ or greater has come from a population of specified mean m, and when we consider that the data of the sample are all that we can obtain by experiment, we see the importance of such a calculation as enabling us to assign limits within

which it is likely that the true mean lies. Suppose for the time being that σ is known; then by solving the equations

$$\bar{x} - m_1 = \sigma t_\infty / \sqrt{p} = m_2 - \bar{x}$$

for the unknowns m_1 and m_2, taking $t_\infty = 1{\cdot}64$ corresponding to the probability level 0·05, we obtain a lower limit m_1 such that the chance of exceeding $\bar{x}$, were m_1 the true mean, is 0·05. m_2 lies a corresponding distance on the other side of $\bar{x}$, and these limits have been termed by Fisher the lower and upper *fiducial* 95 per cent. values of m corresponding to the given $\bar{x}$. The probability level chosen is, of course, arbitrary, but is a convenient one.

Usually, however, σ is as much unknown as m, and we have the much more useful proposition now to be stated. The quantity

$$t = (\bar{x} - m)\sqrt{p}/s,$$

where s is the estimated standard deviation, obtained as already described, is distributed, under the conditions previously stated, in a particular frequency curve first discovered by "Student," and called after him the "Student" distribution, or simply the t-distribution. It depends on a parameter n, the number of degrees of freedom (here, though not generally, equal to $p-1$), and only tends to the normal distribution as n approaches infinity, although in practice there is little to choose between the distributions when n is over 30. Tables of the probability integral of this distribution have been constructed by "Student" (*Metron*, v., 1925) and R. A. Fisher ("Statistical Methods for Research Workers"). Table IV. below is a short adaptation of "Student's" table in *Metron*, by kind permission, and shows how, for a given level of probability, the value of t gets less as n becomes greater until in the limit, with $n=\infty$, it becomes equal to t_∞. This is an indication of the greater uncertainty attending our inferences from small samples owing to the inaccuracy of our estimate of σ, while the infinity value shows us that an infinite sample—*i.e.*, the whole population—yields us the exact value of σ. The use of the table is exactly as outlined above for the normal table, except that we enter the column corresponding to the number of degrees of freedom available for estimating σ. Thus for a sample of 11 ($n=10$), t is equal to 1·81 for P=0·05, and this value would be used instead of 1·64 in obtaining fiducial values of m from a sample of this size, if we were content with this level of probability.

TABLE IV.—TABLE SHOWING THE AREA UNDER THE "STUDENT" CURVE FROM AN ORDINATE BEYOND THAT AT THE MEAN TO INFINITY.

				n			
t	1	2	5	10	15	20	∞
0·0	0·5000	0·5000	0·5000	0·5000	0·5000	0·5000	0·5000
0·2	0·4372	0·4300	0·4247	0·4227	0·4221	0·4218	0·4207
0·4	0·3789	0·3639	0·3528	0·3488	0·3474	0·3467	0·3446
0·6	0·3280	0·3047	0·2873	0·2809	0·2787	0·2776	0·2743
0·8	0·2852	0·2538	0·2300	0·2212	0·2181	0·2166	0·2119
1·0	0·2500	0·2113	0·1816	0·1704	0·1666	0·1646	0·1587
1·2	0·2211	0·1765	0·1419	0·1289	0·1244	0·1221	0·1151
1·4	0·1974	0·1482	0·1102	0·0959	0·0909	0·0884	0·0808
1·6	0·1778	0·1254	0·0852	0·0703	0·0652	0·0626	0·0548
1·8	0·1614	0·1068	0·0659	0·0510	0·0460	0·0435	0·0359
2·0	0·1476	0·0918	0·0510	0·0367	0·0320	0·0296	0·0228
2·2	0·1358	0·0794	0·0395	0·0262	0·0219	0·0199	0·0139
2·4	0·1257	0·0692	0·0308	0·0187	0·0149	0·0131	0·0082
2·6	0·1169	0·0608	0·0241	0·0132	0·0100	0·0086	0·0047
2·8	0·1092	0·0537	0·0190	0·0094	0·0067	0·0055	0·0026
3·0	0·1024	0·0477	0·0150	0·0067	0·0045	0·0035	0·0013
3·2	0·0964	0·0427	0·0120	0·0047	0·0030	0·0022	0·0007
3·4	0·0911	0·0383	0·0096	0·0034	0·0020	0·0014	0·0003
3·6	0·0862	0·0346	0·0078	0·0024	0·0013	0·0009	0·0002
3·8	0·0819	0·0314	0·0063	0·0017	0·0009	0·0006	0·0001
4·0	0·0780	0·0286	0·0052	0·0013	0·0006	0·0004	0·0000

Fisher's more detailed t-table is reproduced as Appendix I., p. 57.

So far we have only considered that we have the data of a single sample. We begin to make contact with the practical problems of field experimentation when we see that methods are required for comparing two or more samples. This will be made clear in the next section, but meantime we shall state the necessary additional propositions. Let $\bar{x}_1$ be the mean of a sample of p_1 observations, supposed taken from a normal population of mean m_1 and standard deviation σ_1. Similarly, let $\bar{x}_2$ be the mean of a second sample of p_2 observations, the population mean and standard deviation being m_2 and σ_2. If the samples are independent, the variance of the difference $\bar{x}_1 - \bar{x}_2$ is equal to $\sigma_1^2/p_1 + \sigma_2^2/p_2$, and the difference is normally distributed round a mean $m_1 - m_2$ with standard deviation

$$\sigma_d = \sqrt{(\sigma^2_1/p_1 + \sigma_2^2/p_2)}.$$

A test of the significance of the difference is a test of how far $m_1 - m_2$ may be supposed zero—*i.e.*, of whether the samples may be supposed to have come from populations of identical means. Such a test is carried out by calculating

$$t_\infty = (\bar{x}_1 - \bar{x}_2)/\sigma_d$$

and finding from the table of the normal probability integral, or from the infinity line of the t-table, the chance that t_∞ should exceed this numerical value, irrespective of sign. Inspection of Table III. shows

that when $t_\infty=1{\cdot}96$ this chance is 0·05, the value 0·025 obtained from the table being doubled, since we have to consider the sum of the areas in the two "tails" of the distribution. Since this chance denotes odds of 19 to 1 against the population means being the same, it is customary to say that when a difference is greater than twice its standard error it is *significant*. For P=0·01, or odds of 99 to 1 against, the value of t_∞ is 2·58.

Differences in standard deviation are unlikely to be of much practical importance in our work. Let us suppose, then, that $\sigma_1=\sigma_2$. σ_d now becomes $\sigma\sqrt{(1/p_1+1/p_2)}$, or $\sigma\sqrt{(2/p)}$ if the samples are equal in size. If we compare this result with that for a single mean we see that we have merely multiplied by $\sqrt{2}$, or 1·414, and a difference between the means of two equal samples may therefore be taken as significant if it exceeds $2\sqrt{2}$, or approximately 3, times the standard deviation of either mean. This test is very commonly made.

Now let us consider what happens when we have to estimate the common standard deviation from the data of the two samples. Let

$$s^2=\frac{\mathrm{S}(x_1-\bar{x}_1)^2+\mathrm{S}(x_2-\bar{x}_2)^2}{n_1+n_2}$$

be taken as our estimate, where $n_1=p_1-1$ and $n_2=p_2-1$ are the numbers of degrees of freedom for the separate samples. If we now calculate

$$t=(\bar{x}_1-\bar{x}_2)/\{s\sqrt{(1/p_1+1/p_2)}\}$$

we find that t is distributed in "Student's" distribution, the parameter n now being equal to n_1+n_2. A similar test of significance to that already described may be made by reading off the required chance from the appropriate line of the t-table. Thus for two samples of 11, with $n=20$, the significance level P=0·05 is reached when $t=2{\cdot}09$, and the level P=0·01 when $t=2{\cdot}84$. (See Appendix I., p. 57.)

Although this test has been described as one for the significance of a difference in means, it is possible for a difference in true variances to contribute to the effect observed. As a supplementary test on this point the two variances may be separately estimated. Thus, let

$$s_1^2=\mathrm{S}(x_1-\bar{x}_1)^2/n_1,\quad s_2^2=\mathrm{S}(x_2-\bar{x}_2)^2/n_2.$$

We may then test the significance of the difference between these two independent estimates by calculating

$$\begin{aligned} z&=\tfrac{1}{2}\log_e(s_1^2/s_2^2)\\ &=\tfrac{1}{2}(\log_e s_1^2-\log_e s_2^2)\end{aligned}$$

If $\sigma_1^2=\sigma_2^2$, the true value of z is zero, and we know the nature of the curve of distribution of z, a curve which contains in its equation the two parameters n_1 and n_2. The test is due to Fisher, who has provided tables of the distribution (reproduced as Appendix II., pp. 58-9) for the probability points P=0·05 and P=0·01 for positive z. To use these we must calculate z by taking s_1^2 as the larger of the two estimates and enter the table with the appropriate n_1 and n_2. The value given therein for P=0·05 is one which would only be exceeded once in twenty times on the average, if the samples were from populations having the same variance. If our value of z is greater than this we may take it that the estimates of variance are significantly different, so that the population variances are probably not identical. The level P=0·01 provides a more stringent test.

We shall find this test exceedingly useful in the sequel, and the reader should make himself familiar with the nature of the calculation. Log_e denotes the natural logarithm, which may be obtained with ample accuracy from four-figure tables giving this function, or from any table of common logarithms by multiplying these by $\log_e 10$, or 2·3026.

This is not a treatise on statistics, and we must therefore be content with this brief statement of the main statistical methods which we shall require to use, reserving for separate consideration their application in the special technique of the *analysis of variance*. Having got so far, we can now return to the more practical consideration of the requirements of a good experimental technique.

III.—The Object of an Experiment.

Put quite simply, our object is to compare different treatments of the land on which our plots are laid out, or of the crops grown thereon. Thus we might wish to test the yield performance of a number of new varieties in comparison with a standard, or the response of a crop to graded applications of one or more fertilizer treatments, or we might be interested in comparing different cultivation processes. The word " treatment " will be used quite generally to designate the thing tested, and the word " plot " for the object of the test. The plot, for example, might sometimes be an experimental animal. If we ignore for a little the factor of soil heterogeneity, we can say that the data collected from a number of equal-sized plots will represent a sample of data from some homogeneous population if the plots are all treated alike, or if the treatments given to the plots have exactly equivalent effects. If, on the other hand, the treatments

are dissimilar in their effects, the total sample will be heterogeneous. The general statistical method is to begin by assuming homogeneity, then test this hypothesis from the data of the samples having different treatments by finding the probability that such divergencies as exist among, say, the sample means are due purely to chance causes. If the probability so found is very low, so that the odds are very much against the hypothesis tested, then we take the hypothesis as having been disproved, and conclude that there are significant differences between the treatments.

We may now lay down certain conditions to be observed. To isolate the factor which is being tested, and to make sure that any significance established is for this factor and no other, it is essential that in all other factors the plots should be as much alike as it is possible to make them. Thus the plots should all be the same size, and the area on which they are placed should be as uniform as possible. If some basal fertilizer is to be given, it should be applied in exactly equal dressings to all plots. Cultivation methods should be identical. Further, a suitable size should be chosen for the unit plot, as already indicated, and a reasonable number of plots should be given over to *each* treatment to be tested. Going back for a moment to Tables I. and II., we saw how the standard error per plot was reduced from 11·7 per cent. to 6·3 per cent. by taking 10 adjacent small plots together to form one larger plot. Had we chosen 10 *at random* from the field in each case we should have expected the standard error of the totals, or means, to be reduced to $11{\cdot}7/\sqrt{10}$, or 3·7 per cent. The reason why the figure is not as small as this is because of the systematic nature of the soil fertility variations. There is a sensible degree of correlation, in fact, existing between neighbouring plot yields, and this applies to the large plots as well as the small ones, as anyone can see by examining Table II. Now the object of having a number of plots laid down to the *same* treatment is (*a*) to average out the experimental errors, and so give us in the mean a better indication of the performance of this treatment than any single plot could provide, and (*b*) to give us the data from which to calculate an estimate of the experimental error. This process of repeating plots of the same treatment is called *replication*. But if we are to be able to make use of the formulæ of the last section respecting the reduction of the standard error, it is essential that the plot yields averaged shall be independent of one another, and independent of those of any other treatment. Each set of plots should be a truly random sample from the area covered by the experiment if it is to yield an unbiased estimate of the productivity of the area under

experiment, apart from deliberately imposed differences of treatment. This necessitates a *random arrangement* of the plots, a point which will be made clearer in the next section.

IV.—A Suggested Method.

To fix our ideas let us suppose that we have 5 treatments to test, and 50 plots of $\frac{1}{10}$ acre on which to work. This means that we can have tenfold replication of each treatment. Now choose 10 plots out of the total *at random* and allot these to treatment No. 1, then 10 more at random for treatment No. 2, and so on, until all the plots are used up. The experimenter may choose his own method of making the random selection, so long as chance is allowed free play and no opportunity for bias enters in. Thus he could take 50 numbered counters, and after mixing them thoroughly separate them into 5 equal heaps. If the plots are consecutively numbered from 1 to 50, the numbers in the first heap will indicate the plots which are to be allotted to treatment No. 1, while those in the second heap give the plots for treatment No. 2, and so on.

When the experiment is completed and yield or other data collected for each of the 50 plots, the next thing is to analyse the figures to see what can be learnt from them. Considered from the standpoint of homogeneity, we must distinguish two sources of variation: (*a*) real errors, in the sense that a difference in treatment may cause one plot to yield differently from another, and (*b*) experimental or random errors. Quite obviously, the lessons to be learnt from the trial will involve an examination of the 5 treatment means, each of 10 plots. The variation among these means will certainly contain the second of these two sources of variation, and may possibly contain the first as well. That is what we have to find out. We therefore work out the amount of the variation present between the 5 treatment means, and compare it with the variation among the 10 replicate plots of each treatment. This latter can only contain the second of the two sources of variation, namely experimental error pure and simple, if our requirements have been followed. It follows that if the amount of variation in the first case is markedly greater than that in the second it points to the existence of real errors. In other words, treatment differences exist of a greater magnitude than would be expected to occur by chance. Essentially, then, the statistical process consists in analyzing the total variation of our experimental material into the two parts mentioned, and comparing these parts to see whether the difference is greater than could reasonably be due to chance fluctuations. If it is concluded that this is

so, then the final stage is to examine the treatment means in the light of a standard error derived for them by consideration of the amount of experimental error present, and so decide what positive conclusions have emerged from the experiment.

Let us illustrate from the data of Table II. This table presents the results of a uniformity trial on 50 plots, but we may assign dummy treatments to the plots and analyze the data as a useful exercise in the methods and computations involved, although we shall not, of course, expect our test to give a positive result. 10 plots were chosen at random and marked A, 10 further random selections were made and marked B, and so on, until all the plots were marked. Grouping under the dummy treatments A to E, we have the results shown in Table V.

TABLE V.—DATA OF TABLE II. ARRANGED IN FIVE RANDOM GROUPS OF TEN PLOTS EACH.

	A.	B.	C.	D.	E.	
	41·8	36·7	37·8	35·5	43·1	
	39·7	41·9	38·5	38·5	34·5	
	40·1	40·3	41·8	40·2	40·2	
	41·6	37·8	37·8	42·5	40·4	
	42·2	33·3	41·4	42·0	39·4	
	41·8	40·3	42·7	35·6	43·4	
	38·0	34·0	39·5	40·4	42·4	
	41·1	42·8	38·3	38·5	38·0	
	35·4	40·0	38·2	38·1	38·8	
	38·1	42·1	40·5	36·6	40·2	
Total	399·8	389·2	396·5	387·9	400·4	*Grand Total.* 1973·8

It will be noticed that the process of random selection has resulted in both low and high yielding plots being present in each column. The column or " treatment " means range from 38·79 to 40·04, the mean of the whole being 39·476.

V.—Analysis of Variance.

We know already from the data of Table II. that the sum of squares of deviations of all the plot yields in Table V. from the mean is

$$S(x - \bar{x})^2 = 306{\cdot}89,$$

from which an estimate, s^2, of the variance of a single plot yield was made by dividing by 49, the number of degrees of freedom. As now arranged, however, the table enables us to calculate two independent estimates of this same variance.

1. *From " Treatment " Totals or Means.*

The 5 column totals of 10 plots each furnish us with an estimate of a variance which will be equal to $10\sigma^2$ (σ^2 being the variance of

a single plot). Alternatively we could work on the column means, which will yield an estimate of $\sigma^2/10$, the difference obviously being only one of a numerical factor. To obtain an estimate of σ^2, therefore, we shall find the sum of squares of deviations of the column totals from their mean 394·76, divide by 10, and then take account of the number of degrees of freedom, which will be one less than the number of totals—*i.e.*, 4. The calculation given below illustrates how a working mean (in this case 400) may be used, although if a calculating machine is available the usual method is to square the numbers as they stand, thus taking the working mean as zero.

	Deviation.	*Square.*
	− 0·2	0·04
	− 10·8	116·64
	− 3·5	12·25
	− 12·1	146·41
	0·4	0·16
Total	− 26·2	275·50
Square	686·44	
Divide by 5		137·288
Subtract		138·212
Divide by 10		13·82
Degrees of freedom		4

2. *From Plots of the Same "Treatment."*

Each set of 10 plots in a column furnishes us with an estimate of the variance σ^2, obtained by summing the squared deviations from the column mean, and taking account of the number of degrees of freedom, 9 in each case. A single estimate of some considerable precision is then obtained by adding together the 5 "sums of squares," this total having 9×5 or 45 degrees of freedom. We shall illustrate on the first column only, using 40 as a working mean.

	Deviation.	*Square.*
	1·8	3·24
	− 0·3	0·09
	0·1	0·01
	1·6	2·56
	2·2	4·84
	1·8	3·24
	− 2·0	4·00
	1·1	1·21
	− 4·6	21·16
	− 1·9	3·61
Total	− 0·2	43·96
Square	0·04	
Divide by 10		0·004
Subtract		43·956
Degrees of freedom		9

Carrying out this calculation for all columns, and adding up, we have the following:

Column.	*Degrees of Freedom.*	*Sum of Squares.*
A	9	43·956
B	9	100·996
C	9	29·825
D	9	55·089
E	9	63·204
Total ..	45	293·07

It will be noticed that the total of the two sums of squares, " between treatments " and " within treatments," is 306·89, exactly equal to the total sum of squares previously determined. This is as it should be, and furnishes incidentally a check on the arithmetic. There is a similar identity in the numbers of degrees of freedom, for 4 + 45=49. We may now collect our calculations into a table of analysis of variance, as follows:

TABLE VI.—ANALYSIS OF VARIANCE OF DATA OF TABLE V.

Variation.	*Degrees of Freedom.*	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Between treatments ..	4	13·82	3·455	0·6199
Within treatments ..	45	293·07	6·513	0·9369
Total	49	306·89	6·263	

It is recommended that this standard form of summary table be adopted in all cases. The column " Mean Square " is obtained by dividing the sum of squares by the number of degrees of freedom for each line of the table. This column furnishes three separate estimates of the same variance, σ^2, the first two of which are independent of one another. We note that the treatment means differ among themselves rather less than plots having the same treatment. That this is purely a chance result is seen by calculating one-half the natural logarithms of the mean squares. The results of this calculation are shown in the last column of the table. Since the mean square " within treatments " is the greater, we have

$$z = 0{\cdot}9369 - 0{\cdot}6199 = 0{\cdot}3170$$
$$n_1 = 45, \qquad n_2 = 4.$$

Turning to the z-table of Fisher (Appendix II., pp. 58-9), we find that there is no value for n_1=45, but as the 5 per cent. values for n_1=24 and ∞ are, for the line n_2=4, 0·8767 and 0·8639 respectively, the required value for n_1=45 is seen to be approximately 0·87. The value we have reached is therefore far from being significant.

The table shows how the random error is apportioned between the

two sources of variation, the total sum of squares being divided roughly in the proportion of the respective numbers of degrees of freedom. We are to imagine that with an infinite number of repetitions of the experiment under the same essential conditions, the average values of the two mean squares would both turn out to be σ^2, the true variance of a single plot yield. When a real experiment is undertaken, the first mean square will be got from the treatment means, the second from replicates of the same treatment. If there are real treatment effects, the first mean square will be larger than it is expected to be by chance, and this is detected by seeing whether the z-value obtained is significant or not. If not, then no conclusive results have emerged, either because the effect is small or non-existent, or because the errors are too large in relation to the observed treatment differences for the latter to be detected. In any case we can go no further, unless indeed, as will be shown later, a further analysis of the treatment effect isolates a part, concentrated in a few degrees of freedom, which does show significance by the z-test. Only when z is significant, either in the original or subsequent analysis, may we proceed further. The calculations, and the reasoning involved, may, however, be described in the present case for the sake of illustration. We begin with the error mean square, 6·513. This is our estimate of the variance of a single plot yield. Its square root, 2·55, is the standard error per plot. It follows that the 5 treatment totals of Table V. may each be assigned a standard error of $2{\cdot}55\sqrt{10}$, or, if we like, $\sqrt{(6{\cdot}513\times10)}$. This is 8·07 lb. The results are summarized as follows:

TABLE VII.—SUMMARY OF RESULTS.

Mean Yield.	A.	B.	C.	D.	E.	*Mean.*	*S.E.*
Lb. per $\frac{1}{10}$ acre ..	399·8	389·2	396·5	387·9	400·4	394·76	8·07
Cwt. per acre ..	17·8	17·4	17·7	17·3	17·9	17·62	0·36
Per cent. ..	101·3	98·6	100·4	98·3	101·4	100·0	2·04

In the table *S.E.* stands for the " standard error " of the figures given under the headings A to E. The first line should be clear from our description. Alternatively we could have tabulated the treatment mean yields per $\frac{1}{10}$ acre, for which the standard error would be $\sqrt{(6{\cdot}513\div10)}$, or 0·807. The second line is usually required in order that the results may be available in the common agricultural units, and is obtained by multiplying the figures in the previous line throughout by 5/112. The last line is obtained by expressing the figures of the first line as percentages of the general mean, 394·76 —*i.e.*, we divide throughout by 3·9476. This is a useful method of summarizing the results, since it enables us to compare the responses,

and the standard error, from one experiment to another. Now the difference between any two of the treatment totals in the first line will have for its standard error $8{\cdot}07\sqrt{2}$. Supposing z to have been significant, we should then look for a difference which at least exceeded $8{\cdot}07\sqrt{2}\times 2$, the figure 2 being an approximation to the value of t for $n=45$, the actual value not being tabled, although it is known to be between 1·96 and 2·04 (Appendix I., p. 57). This gives us 22·8, and we see in the present example that no difference is as large as this, the maximum difference being 12·5. Roughly, we may take three times the standard error 8·07 as an appropriate criterion, and, of course, this calculation may be carried out on any of the three lines of Table VII.

Some care is necessary in making the test, for when we consider that we are comparing the treatment totals with a random sample of chance totals of 10 plots, derived from a single population having a certain standard error, we see that certain of the totals, for example the extremes, are quite likely to differ by two or three times the standard error. For example, in samples of 10, the mean range (*i.e.*, difference between highest and lowest values, averaged for all possible samples) is 3·078 times the standard deviation.* This should be a sufficient warning against choosing figures for comparison at the extreme ends of the range, and applying to them a test which is designed for any pair of values taken at random. The experimenter is guarded against making wrong deductions by the requirement that z shall be significant before significant differences are sought. If not, then the data are compatible with the original hypothesis of homogeneity, and it is no use looking for differences. See later on, however, for methods of subdividing the treatment variation.

Before concluding this section, it will be instructive to compare what we have done with the methods in use before the technique of the analysis of variance was elaborated. Supposing the design of the experiment to have been as described, which is very unlikely, because it was the application of statistical methods based on small samples which led to such designs being put forward, then the experimenter would have had the data of 5 samples of 10 values each, as in Table V. From each he would have obtained estimates of the mean and variance. Ten separate comparisons of the samples in pairs are now possible, and for each the standard error of the difference between the two means would be calculated. When we consider, however, the limited range of yields customarily met with in field

* Tippett, L. H. C., "Methods of Statistics," 1931, p. 26 (Williams and Norgate).

experiments, we see that the 5 separate estimates of variance are unlikely to differ except through sampling variation, while the precision of each estimate will not be great, being based upon only 9 degrees of freedom. Those who have had difficulty in interpreting the results of their trials on such a basis will appreciate the single comprehensive test of significance furnished by the analysis of variance, and the further advantage that five separate estimates of variance are replaced by one, based on a very much larger number of degrees of freedom. Cases do occur, not very frequently, where the variance does not seem to remain constant. If, for example, the yields of one treatment were ten times as high as those of another, the variances would be unlikely to be identical. Some light on this point is furnished by a comparison of the 5 components of the " within treatments " variation, by means of the z-test. It may be necessary in certain cases to estimate the variances separately, when the need for having larger samples will become apparent. The difficulty is sometimes met by working on some transformation, such as the logarithms, of the original figures. Irrespective, however, of whether the group variances are the same or not, the test we have described is the perfectly valid one of comparing the estimated variance of the group means with the mean of the separate estimates of variance, and a positive result to the z-test serves in any case to disprove homogeneity, although it may not be clear how we are to compare the group means with their standard errors; the average standard error as customarily calculated will not always be applicable.

A last point to note is that in the special case of two groups, the analysis of variance method, and the use of z, become identical with the t-test for the comparison of two means. If the reader tries both methods on the first two columns of Table V., for example, he will have no difficulty in proving arithmetically the relationship $z=\log_e t$, which holds when $n_1=1$.

VI.—Reduction of Error by Local Control—Method of Randomized Blocks.

The methods of the last section aim at securing a valid estimate of the total experimental error over the area considered. We agreed at the time to ignore the soil errors, which we have reason to suppose will contribute materially to the total. Not only so, but they will increase as the area under experiment increases, a factor which will limit the usefulness of replication and the number of treatments to be tested in any one experiment. Let us see now whether we cannot

do something to eliminate part of the variation due to soil heterogeneity from our comparisons. Begin with the data of Table II., and divide it into 10 "blocks" of 5 plots each, each column of the table being divided half-way down. If we divide the total variation into two parts, between and within blocks, we reach the following table of analysis of variance by the methods of the last section.

TABLE VIII.—ANALYSIS OF VARIANCE OF DATA OF TABLE II., TAKEN IN TEN BLOCKS OF FIVE PLOTS EACH.

Variation.	*Degrees of Freedom.*	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Between blocks ..	9	190·53	21·170	1·5263
Within blocks ..	40	116·36	2·909	0·5339
Total	49	306·89		

For the variation between blocks we sum the squares of the 10 block totals (see Table X), subtract $\frac{1}{10}$ of the square of the grand total, and divide by 5. The remainder, *i.e.*, "within blocks," is obtained most easily by subtraction from the total which we already know. We have now a very different state of affairs from that shown in Table VI. The value of z is 0·9924, with $n_1=9$, $n_2=40$. The z-table does not give us without interpolation the required values at the 1 and 5 per cent. levels of significance, but the former of these may be roughly guessed as being about 0·53, certainly lying between 0·5773 and 0·4574. Our value of 0·9924 is therefore strongly significant, and undoubtedly disproves the hypothesis of homogeneity for this data. Clearly marked soil fertility differences are therefore shown to exist between groups of 5 plots, or areas of $\frac{1}{10}$ acre, and we note that if we were concerned to compare plot yields *within* the blocks we have constructed, these would have an estimated variance of 2·909—*i.e.*, a standard error of 1·71, or 4·3 per cent. of the mean, a very considerable reduction from the previous figure of 6·3 per cent. The variance is less than half its former figure, or the accuracy more than doubled. This is what is meant by a reduction of error being possible by *local control*, and it suggests at once an improvement in our experimental technique. Suppose that we are still concerned to test 5 treatments, with tenfold replication of each, and that we illustrate on the same data. Divide up the area into 10 blocks as described, and within each block assign one plot *at random* to each of the treatments A to E. Details are given in the second part of various methods whereby this operation may be expeditiously carried out, but the allocation should be random, not systematic, as sometimes practised. The process of

random selection of plots within a block ensures the independence of the 5 samples of 10 plots each that are given over to the treatments to be tested, and justifies the application of the extended technique of the analysis of variance now to be described.

The plan resulting from a random draw on the lines indicated is given in Table IX., the method of arrangement being known as *randomized blocks.*

TABLE IX.—RANDOMIZED BLOCK ARRANGEMENT IMPOSED ON THE DATA OF TABLE II., TESTING FIVE "TREATMENTS" IN TEN BLOCKS.

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

Having eliminated the variation between blocks, as in Table VIII., we now turn our attention to the variation within blocks, which, in a real experiment, would be made up of experimental error on the one hand and a possible real error due to differences between treatments on the other. Since the variation within blocks is calculated by summing the squares of deviations from block means, it is these *deviations* that we evidently require to test for homogeneity by the method of Section V, by isolating that part of the variation that is between treatment mean deviations, having 4 degrees of freedom, from the remainder, which has 36 degrees of freedom. The two mean squares obtained on dividing by the respective numbers of degrees of freedom are then tested by calculating z, and the analysis proceeds by comparing the treatment means, exactly as before, using, however, for the standard error an estimate obtained from the variation within treatment deviations, and based on 36 degrees of freedom. This number is always the product of the degrees of freedom due to blocks and treatments, and represents the number of plot yields that could be assigned arbitrarily, the remainder being determined from the fact that block and treatment totals are fixed.

In practice the complete analysis is most easily obtained by setting out the data in columns corresponding to the treatments and in rows corresponding to the blocks, as in Table X.

TABLE X.—DATA OF TABLE II., ARRANGED BY BLOCKS AND "TREATMENTS."

Block.	*Treatments (see Table IX.).* A.	B.	C.	D.	E.	*Total.*
1	40·4	37·8	41·1	40·4	41·8	201·5
2	41·8	43·4	41·6	39·4	42·8	209·0
3	42·5	40·5	42·4	42·0	41·9	209·3
4	42·2	42·7	43·1	40·2	41·4	209·6
5	40·3	37·8	37·8	36·7	38·3	190·9
6	35·5	42·1	40·1	38·2	38·1	194·0
7	38·5	41·8	40·2	40·3	40·0	200·8
8	39·7	33·3	34·0	34·5	38·0	179·5
9	35·4	38·8	36·6	39·5	38·0	188·3
10	40·2	35·6	38·5	38·1	38·5	190·9
Total ..	396·5	393·8	395·4	389·3	398·8	1,973·8

The total variation is already known from Section II, likewise the block variation from the first part of this section, this being the sum of squares of the 10 deviations of the block totals from their mean 197·38, divided by 5, since each is a total of 5 plot yields. In similar fashion the variation due to "treatments" is obtained by calculating the sum of squares of the 5 deviations of the treatment totals from their mean 394·76, and dividing by 10, since each is a total of 10 plot yields. The details of similar calculations have already been given. The remaining variation, which we consider as being due to experimental error, is then obtained by subtraction from the total, and we have the completed analysis as in Table XI.

TABLE XI.—ANALYSIS OF VARIANCE OF THE DATA OF TABLE X.

Variation.	*Degrees of Freedom.*	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Blocks	9	190·53	21·170	
Treatments	4	5·05	1·262	0·1163
Error	36	111·31	3·092	0·5644
Total	49	306·89		

It is clear that there is no significant difference between the mean squares for treatments and error, the treatment variation being in fact subnormal. We do not expect to find significance in the present case, for the trial is a uniformity one, and the two mean squares are both estimates of the experimental error, the latter, of course, having the greater precision. In a real experiment a positive result would be shown by z being significant, owing to the addition in the treatment line of a component of real error. The calculations would then be rounded off by preparing a summary table like our Table VII., beginning with the treatment totals in the last line of Table X., which have for their standard error $\sqrt{(3{\cdot}092 \times 10)}$, or 5·561 lb., 1·41 per cent. of the mean yield of an area $\frac{1}{5}$ acre in size.

Considerable precision has evidently been attained, for in such an experiment as the present one differences between the treatment means of the order of 4 per cent. would be detected as significant. The method of randomized blocks can be applied with any number of treatments and with any desired degree of replication, for we have only to form a compact block containing a number of plots equal to the number of treatments to be tested, and then repeat this pattern the desired number of times, taking care, of course, that the arrangement is random in each block. The reduction of error through elimination of block differences is often quite considerable, although cases arise where the gain appears to be slight. Any soil differences which occur within the block cannot be eliminated, and will contribute to the experimental error. In large blocks this factor may lead to plots within the block differing considerably for this cause alone, and if the replication is limited some of the treatments may be biased, while the error may occasionally be unduly high. The random arrangement safeguards us as much as possible, but there must evidently be a limit to the efficiency of this, and indeed of any, arrangement of plots. Usually the number of plots within a block is not large unless two or more sets of treatments are represented, as in multiple factor experiments. To these we shall return. The experimenter will be well advised in any case to allow for as much replication as is practicable.

Few experimenters take the trouble to calculate the error sum of squares directly, but are content to obtain it, as we did above, by difference. It is instructive to show how the direct calculation is made. If we suppose in Table X. that the yield of any plot is made up of two additive components, one due to the block in which it is situated, and the other to the treatment of which it is a replicate, we may estimate these " expected " or theoretical values in the following way. Our assumption is that the deviation of the expected yield from the true mean is the sum of the deviations of the true block and treatment means from the general mean. The series of expected yields may therefore be estimated by adding the block and treatment means obtained from the data, and subtracting the general mean. Thus for plot A1 the expected yield is estimated as 40·3+39·65 −39·476, or 40·474. The actual yield is 40·4. The " error " is therefore −0·074. There are 50 such errors in all, and the sum of their squares is 111·31, as shown in Table XI. This calculation throws some light on the nature of the error, for it is seen to be a measure of the aggregate deviation of the plot yields from the expected yields on an assumed additive relation of the components for blocks and treat-

ments. As we have grown accustomed to saying, it is the *interaction* between block and treatment effects. This idea of interaction will be found later to be of importance where different series of treatments are included in the same experiment.

This section will be concluded by an illustrative example of an actual experiment carried out on the randomized block plan at the Cambridge University farm. The data represent the weight of green produce in lbs. from a single cut made on old pasture on June 8, 1931. The plots harvested consisted of strips each 4 yards wide and 45 yards long. Each plot was subdivided for 4 manurial treatments, but this complication will be ignored for the present, only the total for the plot being considered. There were three main treatments, including a control (O) consisting of the untreated land. In the other cases the effect of a grass-land rejuvenator (R) was compared with the use of the harrow (H). The block was therefore composed of 3 plots, and the experiment consisted of 6 randomized blocks, placed side by side. The plan and yields are given in Table XII.

TABLE XII.—PLAN AND YIELDS OF GRASS-LAND EXPERIMENT, CAMBRIDGE, 1931.

O.	H.	R.	R.	H.	O.	O.	R.	H.	O.	R.	H.	H.	O.	R.	O.	H.	R.
813	647	713	814	759	795	705	652	598	774	617	559	580	687	539	581	480	437

For calculation purposes the yields are best set out by blocks and treatments, as under:

TABLE XIII.—YIELDS ARRANGED BY BLOCKS AND TREATMENTS.

Treatment.	*Blocks.*						*Total.*
	1.	2.	3.	4.	5.	6.	
O	813	795	705	774	687	581	4,355
H	647	759	598	559	580	480	3,623
R	713	814	652	617	539	437	3,772
Total ..	2,173	2,368	1,955	1,950	1,806	1,498	11,750

There is evidence from the block totals of a pronounced fertility trend. The sum of squares of the 18 yields is 7,888,448, from which must be subtracted $\frac{1}{18}$ of the square of the grand total, 11,750, or 7,670,138·9. The difference is 218,309·1, with 17 degrees of freedom. The sum of squares of the block totals, less $\frac{1}{6}$ of the square of the grand total, is 449,101·3, which is divided by 3 to yield 149,700·4, the part due to blocks, having 5 degrees of freedom. The sum of squares of the treatment totals, less $\frac{1}{3}$ of the square of the grand total, is 299,304·7. On dividing by 6 we have 49,884·1 for the part due to treatments, with 2 degrees of freedom. The error sum of squares is obtained by difference, and the analysis of variance table is shown in Table XIV.

TABLE XIV.—ANALYSIS OF VARIANCE—WEIGHT OF GREEN PRODUCE.

Variation.	*Degrees of Freedom.*	*Sum of Squares.*	*Mean Square.*	*½ Log, (Mean Square).*
Blocks	5	149,700·4	29,940·1	
Treatments	2	49,884·1	24,942·1	1·6083
Error	10	18,724·6	1,872·5	0·3136
Total	17	218,309·1		

$z = 1·2947$. For $P = 0·01$, $n_1 = 2$, $n_2 = 10$, $z = 1·0214$.

Standard error per plot $= \sqrt{1,872·5} = 43·3$, or 6·63 per cent. of the mean yield, 652·8 lb.

Note that the mean squares have each been divided by 1,000 before calculating the natural logarithms, a procedure which is convenient, and makes no difference to the final result. Summarizing the table, it is seen that a large amount of variation has been removed as due to block differences. The treatment effect is strongly significant, as judged by the z-test. The accuracy of the experiment, shown by the standard error per plot, is very satisfactory. The treatment totals in Table XIII. are total yields in lb. of $6 \times 180/4840$ acre, and may therefore be expressed in tons per acre by multiplying by the factor $4840/(6 \times 180 \times 2240)$, which is almost exactly $\frac{1}{500}$ (0·00200066 to be precise). The standard error of these totals is $\sqrt{(6 \times 1872·5)} = 106·0$, while their mean is 3,916·7. With the aid of the conversion factor we have the following table:

TABLE XV.—SUMMARY OF RESULTS—WEIGHT OF GREEN PRODUCE.

Mean Yield.	*No Treatment.*	*Harrowed.*	*Grass-Land Rejuvenator.*	*Mean.*	*S.E.*
Tons per acre ..	8·71	7·25	7·55	7·84	0·212
Per cent. ..	111·2	92·5	96·3	100·0	2·71

It is obvious, without further specific tests, that both treatments have depressed the yield significantly, but the difference in yield between the plots harrowed and those treated by the rejuvenator is not significant.

The practical man often finds it convenient to express the yields of his treated plots as a percentage of the control rather than of the general mean. This is useful for comparative purposes, but conveys no additional information, and it is recommended that the method followed in the above summary tables be used generally.

VII.—Method of the Latin Square.

A useful special method is available when the number of treatments is not too great, and it is arranged to have the same number of plots of each treatment as there are treatments to be tested. Thus

suppose we have 5 treatments, each to occur on 5 plots. The experimental area is laid out in 5 rows and 5 columns of plots in such a way that each treatment occurs once, and once only, in each row and column. Subject to this restriction the treatments are arranged at random. This is the method of the *Latin square*, a term used by Fisher to describe a random selection from all the possible squares of the given size satisfying the conditions imposed, although it comes down to us from the original formulation of the problem of how to set up such an arrangement. Special methods have to be used to ensure complete randomization, and the work can be carried out for moderate-sized squares by using the typical " transformation sets " tabulated by Yates (*Emp. J. Exp. Agric.*, 1933, i., p. 235). The reader is referred to this paper for a description of the technique, and for an indication of the best method of procedure for squares of larger size than 6×6. See also Part II., p. 84.

When the yield data are tabulated and summed by rows and by columns the variation of the resulting totals will give an indication of the amount of soil heterogeneity, running in two directions at right angles to one another. This variation may be removed from the total by calculating the sums of squares of deviations for these two components, in place of the one component of blocks eliminated by the method described in the last section. No element of treatment differences enters into row or column comparisons, for all treatments are equally represented in all rows and in all columns. The variation due to treatments is calculated as before by adding up by treatments, summing the squares of deviations of the totals from their mean, and dividing by the number of plots contributing to the total. In fact, three calculations of identical form are carried out on row, column, and treatment totals. These three sums of squares are independent of one another, and of the remainder left on subtracting from the total sum of squares. This remainder is used to give our estimate of error, one which is often found to be lower than the corresponding one from a randomized blocks arrangement, owing to the more complete elimination of soil heterogeneity through rows and columns. The treatment effect is tested by forming the treatment and error mean squares, and calculating z, while the subsequent calculations are as before. Note that we now have four different parts into which the total variation is divided, irrespective of a possible further division of the treatment sum of squares which is sometimes possible. The degrees of freedom for rows, columns, and treatments are one less than the number of totals to be compared, while those for error are, for a 5×5 square, 4×3 or 12 [in general

$(p-1)(p-2)$ for a $p \times p$ square]. This represents the number o plot yields that could be assigned arbitrarily, the remainder being determined from the fact that row, column, and treatment totals are considered as fixed, being estimated from the totals given in the experiment.

The method will be illustrated by an experiment on the nitrogenous manuring of wheat, carried out at Rothamsted in 1932.* The arrangement was a 5×5 Latin square, each plot being $\frac{1}{10}$ acre in size. Nitrogenous fertilizer was applied to the plots at the rate of 0·3 cwt. N per acre, according to the following schedule of treatments which includes a control:

1. O=no nitrogenous fertilizer.
2. S=sulphate of ammonia, applied in March.
3. SS=sulphate of ammonia in 6 monthly dressings, November to April.
4. C=cyanamide, applied in October.
5. D=half cyanamide, half dicyanodiamide, applied in October.

The plan and yields of grain in lb. are given in Table XVI.

TABLE XVI.—PLAN AND YIELDS OF WHEAT EXPERIMENT, ROTHAMSTED, 1932.

D	SS	O	C	S
72·2	55·4	36·6	67·9	73·0
O	C	SS	S	D
36·4	46·9	46·8	54·9	68·5
SS	S	D	O	C
71·5	55·6	71·6	67·5	78·4
S	O	C	D	SS
68·9	53·2	69·8	79·6	77·2
C	D	S	SS	O
82·0	81·0	76·0	87·9	70·9

The row, column, and treatment totals obtained from the above table are set out in Table XVII.

TABLE XVII.—ROW, COLUMN AND TREATMENT TOTALS FROM TABLE XVI.

	1.	2.	3.	4.	5.	*Mean.*
Row..	305·1	253·5	344·6	348·7	397·8	329·94
Column	331·0	292·1	300·8	357·8	368·0	329·94
Treatment ..	264·6	328·4	338·8	345·0	372·9	329·94

Grand total 1,649·7. General plot mean 65·988.

* Rothamsted Experimental Station, Report for 1932, p. 147.

Proceeding now to analyze these yields, we first find the total sum of squares by adding up the squares of the 25 individual plot yields and subtracting $\frac{1}{25}$ of the square of the grand total. The sums of squares for rows, columns, and treatments are obtained from Table XVII.; in each case the totals are squared and added; from the result is subtracted $\frac{1}{5}$ of the square of the grand total; the remainder is divided by 5, since each figure in the table is a total of 5 plot yields. The sum of squares due to error is obtained by difference, and the complete analysis of variance is shown in Table XVIII.

TABLE XVIII.—ANALYSIS OF VARIANCE—WEIGHT OF GRAIN.

Variation.	*Degrees of Freedom.*	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Rows	4	2,326·39	581·60	
Columns	4	901·37	225·34	
Treatments	4	1,284·51	321·13	1·7346
Error	12	202·06	16·84	0·2606
Total	24	4,714·33		

z=1·4740. For P=0·01, n_1=4, n_2=12, z=0·8443.

Standard error per plot= $\sqrt{16·84}$=4·104, or 6·22 per cent. of the mean yield, 65·988 lb.

The mean squares have been divided by 10 before calculating the natural logarithms.

Summarizing the information to be obtained from this table, we see that a very large amount of variation has been removed in the row and column components. The reader may test for himself that in both cases z is strongly significant. The result is that the error mean square is quite small. The heterogeneity of this area is such that had the 5 treatments been scattered at random over the area without restriction in the Latin square design, the error mean square would have been expected to be of the order of 3,429·82/20, or 171·49, more than 10 times the value obtained in the Latin square. In such a case the treatment effect would have been insignificant, being masked by the high error. As we see, however, the z is well above the 0·01 probability level, and treatments are strongly significant.

The treatment totals in Table XVII. are total yields in lb. per $\frac{5}{40}$ acre, and may therefore be expressed in cwt. per acre by dividing by 14. The standard error of these totals is $\sqrt{(5\times16·84)}$=9·18. The final table is as follows:

TABLE XIX.—SUMMARY OF RESULTS—WEIGHT OF GRAIN.

Mean Yield.	*No Nitrog. Fert.*	*S. Amm. One Application.*	*S. Amm. Divided.*	*Cyan.*	*Cyan.+ Dicyan.*	*Mean.*	*S.E.*
Cwt. per acre ..	18·9	23·5	24·2	24·6	26·6	23·6	0·66
Per cent. ..	80·2	99·5	102·7	104·6	113·0	100·0	2·78

It is obvious that there has been a significant response to nitrogenous fertilizer, however applied. It seems likely, also, that the additional response on those plots which had part of their dressing in the form of dicyanodiamide is a genuine effect. As an exercise we may test the figure 113·0 against the mean of the other three treated plots.

Cyan+Dicyan. ..	113·0	Variance ..	7·73 (*i.e.*, $2·78^2$)
Mean of others ..	102·3	Variance ..	2·58 (*i.e.*, $7·73 \div 3$)
Difference	10·7	Sum	10·31
Standard error ..	3·21 (*i.e.*, $\sqrt{10·31}$)		

With 12 degrees of freedom available for estimating the experimental error the value of t at the 1 per cent. significance level is 3·055. The ratio of the above difference to its estimated standard error is 3·3. The difference is therefore significant. Another way in which the point may be brought out is to test the figure 113·0 against the general mean of all treatments—*i.e.*, 100. The standard error of the difference is $s\sqrt{(\frac{4}{5})}$, where $s=2·78$. This gives 2·49, and the difference 13·0 is seen to be 5·2 times its standard error. This is the only one of the treatment means that is significantly above the general mean. For p treatments the formula we have used for the standard error is $s\sqrt{\{(p-1)/p\}}$, in which account is taken of the correlation between the single mean and the general mean which includes it. Note that this test is mathematically identical with testing 113·0 against the mean of all others, including control.

The negative information that sulphate of ammonia gave the same results as cyanamide, although one was applied in spring and the other in the autumn, and that a divided dressing was neither better nor worse than a single dressing, may be noted. Such results are often quite as important as positive results.

An experimental design, of whatever form, must provide a sufficient number of replications to lead to an error mean square based on an adequate number of degrees of freedom. If not, the estimate of the error variance may occasionally differ greatly from the

true value, even though it is always a valid estimate of it. An endeavour should be made to secure that the number of degrees of freedom is at least greater than 10. This rules out single 3×3 and 4×4 Latin squares, but with fewer than 5 treatments two or more Latin squares might be laid out side by side. There is an upper limit to the number of treatments which can be accommodated within the structure. With more than 7 or 8 the rows and columns tend to be too long, and the efficiency of the design is impaired. It is not advisable in such cases to divide the treatments into two or more sets, each laid out in a Latin square, unless provision is made for at least one control or standard treatment to be included in all sets. Even so, comparisons between the sets do not have the same accuracy as those within sets. Within the limitations stated, however, the Latin square design is usually very efficient, although the soil heterogeneity may occasionally be of such a character that a randomized block design would have eliminated more of the soil variation than the Latin square.

VIII.—Extension of Analysis of Variance—Multiple Factor Experiments.

So far our experimental designs have been of a simple straightforward character, and the procedure has been to determine by means of a z-test whether treatments as a whole have shown significant differences or not. Only when this is shown is it permissible to examine individual differences between treatment means by their standard errors in order to isolate the significant effects. It may sometimes happen, however, that treatments as a whole fail to be significant owing to an undoubtedly significant effect, isolated in a single degree of freedom, or more than one, being watered down by inclusion with the rest. Even where this does not happen we may require to examine independently the different comparisons that are possible. This brings us to our next point, the further analysis of the treatment sum of squares that is sometimes possible. Endless variations are possible on this theme, for the possibility of further analysis depends on the nature of the treatments included in an experiment, while the fact that it can be done often determines the experimenter in his choice of treatments. We cannot do better than give a number of examples, which will be far from exhaustive, but may serve to illustrate the method.

Our Latin square example concerned the comparison of a number of treatments having equivalent nitrogen, but included, as is usual, a control. We may desire to know how much of the significance is due

to a response of all treated plots to nitrogenous fertilizer, and how far the equivalent nitrogenous treatments differed among themselves. We had little doubt at the time as to what was to be learnt from the data, but the issue may not always be so clear cut. The following calculations explain themselves, with the aid of Table XVII.

	0 Plots.	*Others.*	*All Plots.*
Total	264·6	1,385·1	1,649·7
Mean total of 5	264·6	346·275	329·94
Product	70,013·16	479,625·50	544,302·02

Sum of products in columns 1 and 2, minus that *in column* 3 = 5,336·64
Divide by 5 = 1,067·33 (1 d.f.)

Sum of squares of 4 nitrogen totals = 480,711·41
$\frac{1}{4}(1{,}385{\cdot}1^2)$ = 479,625·50
Difference = 1,085·91
Divide by 5 = 217·18 (3 d.f.)

1,067·33+217·18=1,284·51, the total sum of squares due to treatments.

ANALYSIS OF VARIANCE (ROWS AND COLUMNS LEFT OUT).

	Degrees of Freedom.	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
O *v.* N	1	1,067·33	1,067·33	2·3352
Within N ..	3	217·18	72·39	0·9897
Error	12	202·06	16·84	0·2606

Here we have divided the total sum of squares due to treatments into two parts: (1) that due to the difference between the control plots and the mean of the others, and (2) that due to the variation of the 4 treated plot mean yields round their mean. In the first part the totals have to be weighted in an obvious way because of the unequal numbers on which they are based. We now have a z for the first effect of 2·0746, with $n_1=1$, $n_2=12$, the 1 per cent. value being 1·1166. There is thus no question of the response of the crop to the nitrogenous treatment. For the second effect z is 0·7291, with $n_1=3$, $n_2=12$. The 5 per cent. value is 0·6250 and the 1 per cent. value 0·8919. There are therefore significant differences among the treated plots, although the effect is not so marked as in the comparison between treated and untreated plots. An examination of the means makes it clear that the effect is due to the superiority of the plots having dicyanodiamide.

As a second example take the following figures representing the response of a crop of potatoes to graded applications of superphosphate.* The arrangement was a 4×4 Latin square, the treatments being no superphosphate, and 2, 4, and 8 cwt. superphosphate per acre. The yield was high and the response to superphosphate, perhaps

* Rothamsted Experimental Station, Report for 1927-8, p. 171.

on this account, small, although it represented a paying increase. The extract from the analysis of variance table given below shows that z is quite insignificant on the total 3 degrees of freedom. Each plot was $\frac{1}{512}$ acre in size.

	Treatment: 0.	2.	4.	8.	*Total.*
Total of 4 plots (less 4,800) in lb. ..	.. -46	51	113	171	289

	Degrees of Freedom.	*Sum of Squares*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Treatment	3	6,461·68	2,153·89	0·3836
Error	6	8,416·88	1,402·81	0·1692

z=0·2144. For P=0·05, n_1=3, n_2=6, z=0·7798.

SUMMARY OF RESULTS.

Mean Yield.	*No Super.*	2 *Cwt. Super.*	4 *Cwt. Super.*	8 *Cwt. Super.*	*Mean.*	*S.E.*
Tons per acre ..	16·98	17·32	17·55	17·75	17·40	0·27
Per cent. ..	97·6	99·6	100·8	102·0	100·0	1·54

The responses are of a fairly regular character, although there is evidence of a diminishing return from the higher dressings. We may therefore calculate the linear regression of yield on amount of superphosphate applied, and find that part of the treatment sum of squares that is due to this linear component, having a single degree of freedom. If y represents yield and x the values 0, 1, 2, and 4 (one unit of x being 2 cwt. superphosphate), the regression coefficient is

$$\mathrm{S}\{y(x-\bar{x})\}/\mathrm{S}(x-\bar{x})^2,$$

$\bar{x}$ being the mean of the values of x. The numerator is the same thing as $\mathrm{S}(y-\bar{y})\,(x-\bar{x})$, a form which may be more familiar to some readers acquainted with correlation work, for $\bar{y}\ \mathrm{S}\,(x-\bar{x})=0$. The linear component of which we spoke (see also p. 47) is

$$\{\mathrm{S}y(x-\bar{x})\}^2/\mathrm{S}(x-\bar{x})^2.$$

The term $\mathrm{S}\{y(x-\bar{x})\}$ we calculate from the treatment totals (less 4,800), by multiplying by 0, 1, 2, and 4, and subtracting the product of the total of the y's, namely 289, and the mean of the x's, which is 7/4. We then square this result and divide by $\mathrm{S}(x-\bar{x})^2$, or 35/4, obtaining 23,686·01. The required result is then obtained by dividing by 4, just as the sum of squares for treatments, obtained from the totals of 4 plot yields, had to be divided by 4. The remainder of the treatment sum of squares represents deviations from the linear regression function, and the analysis of variance table may now be set out as follows:

	Degrees of Freedom.	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Linear component ..	1	5,921·50	5,921·50	0·8893
Deviations	2	540·18	270·09	
Error	6	8,416·88	1,402·81	0·1692

z for linear component=0·7201. For P=0·05, n_1=1, n_2=6, z=0·8948.

Now although the z is still insignificant (in fact, had the whole 6,462 been isolated in this single degree of freedom the z would not have been significant, owing to the small number of degrees of freedom available for estimating the error), it is much larger than before, and we see how a significant result may emerge from such a subdivision of the treatment sum of squares. If this turns out to be the case, the conclusion would be that there had been a response to the fertilizer proportional to the amount applied.

The necessary calculations become especially simple when the observations go by equal steps, as in our example had a plot having 6 cwt. superphosphate been added. With n treatments $S(x-\bar{x})^2$ is equal to $n(n^2-1)/12$, while $S\{y(x-\bar{x})\}$ is readily determined by starting from the middle, and multiplying the yields y by $\pm\frac{1}{2}$, $\pm\frac{3}{2}$, etc., for an even set of treatments, or ± 1, ± 2, etc., for an odd set, as we proceed outwards to the extremes, finally adding the results. Not only so, but the quadratic, cubic, etc., terms can be determined in addition to the linear, and these effects isolated in the treatment sum of squares.* A description of the somewhat elaborate methods required would be out of place here.

Multiple Factor Experiments.—In a wide class of cases two or more sets of treatments are introduced in all combinations into a single experiment. Thus if we were concerned to test the effects of nitrogenous and phosphatic fertilizers on a crop, we might choose 0, 1, 2 and 3 units of nitrogenous, and perhaps 0, 1 and 2 units of phosphatic, fertilizer. This gives us 12 treatments in all, which might be arranged in a randomized block experiment with a suitable number of replications. The sum of squares due to treatments may now be divided into three parts. If the treatment totals are arranged in rows corresponding to one fertilizer, and in columns corresponding to the other, it is evident that the methods we have described under the heading " randomized blocks " enable us to isolate a part of the treatment sum of squares as being due to differences between rows, and therefore to this particular manurial comparison. A further part will be due to column differences, and will therefore measure the other manurial effect, while the remainder is the interaction, as already described, between rows and columns. With p rows and q columns the

* R. A. Fisher, " Statistical Methods for Research Workers," § 27.

degrees of freedom for the respective parts are $p-1$, $q-1$, and $(p-1)(q-1)$. The three mean squares thus derived can be tested separately against the error mean square, which latter is, of course, a measure of the interaction between blocks and treatments. Not only, then, can we examine the significance of the manurial effects separately, and so see how far one fertilizer has an effect on plots which have received all the dressings of the other fertilizer in equal amounts, but also information is provided on the combined effect. If the effects are additive and independent, the interaction mean square may be expected to be of the same order as the error mean square, or at least not to differ from it significantly. If this is found to be the case, then we may take it that, within the limits of experimental error, the effect of the one fertilizer is constant over the chosen range of the other. If, however, the interaction is significant when compared with error, it evidently tells us that the combined effect is something more, or less, than the sum of the separate effects. For example, if the effect of adding nitrogenous, or phosphatic, fertilizer to plots not otherwise receiving these substances was to raise the yield from 100 to 110, while if both were applied in combination the yield was 140, that would be evidence of interaction, in the sense that the plots tended to respond better to one fertilizer in the presence than in the absence of the other.

In this latter case there is evidently gained from the experiment involving both factors, information that could not possibly have been deduced from two experiments laid down to test them separately. In the former case—*i.e.*, where the interaction is insignificant—our conclusions evidently have a greater generality than if an experiment had been laid out on one factor with a fixed basal dressing of the other, for we can assert that the result holds over the whole range of such dressing incorporated in the experiment. Not only so, but greater precision is attained on these main comparisons owing to the averaging for one while testing the other. Suppose the 12 treatments which we chose for the purpose of illustration had been laid out in 6 blocks. If the error mean square is denoted by s^2, then the standard error of the means of the nitrogenous treatments will be $s/\sqrt{18}$, and that for the means of the phosphatic treatments $s/\sqrt{24}$, instead of the $s/\sqrt{6}$ which is the standard error of the individual treatment means. This circumstance often more than makes up for the disadvantage that only a limited amount of soil heterogeneity can be eliminated owing to the blocks being of large size. Naturally, the full advantage in this respect will only be enjoyed when interaction is absent.

The reader would be well advised to make himself acquainted with these processes by working out a number of examples, chosen, for example, from the wealth of material contained in the Rothamsted Station Reports since 1925. Any number of factors may be introduced so long as all combinations of treatments are present, and this will multiply the number of interactions that can be worked out. With three factors, there are not only three direct effects to examine, and three first-order interactions of the treatments taken in pairs, but a last component remaining which is called the second-order interaction, and represents the effect of a changing first-order interaction between two of the factors as the third factor is varied.

Experiments with Split Plots.—In discussing the randomized blocks example of section VI it was stated that the main plots, 4 yards in width, were subdivided for 4 manurial comparisons. This is another way of introducing an extra factor, and is particularly convenient where it is necessary to have large plots for the one series of comparisons, as, for example, with cultivation treatments, whereas other comparisons, such as manurial ones, can be investigated by allotting at random the sub-plots within the main plot to the treatments of the second series. With 3 main treatments and 4 sub-treatments there are 12 combinations in all, replicated in this case 6 times. The comparisons are not, however, made all with the same degree of precision. Manurial comparisons are made between closely adjacent small plots which may be expected to differ less in soil fertility than the larger main plots whose centres are further apart. The simplest way to see how the statistical analysis is to be carried out is to regard the experimental data as consisting of 18 classes, each class being a main plot containing 4 members. A first analysis of variance will be into a part between class means, having 17 degrees of freedom, and a part within classes, having 18×3 or 54 degrees of freedom. The first part is identical with the total sum of squares calculated in section VI, and can be divided into parts due to blocks, main treatments and the error appropriate to these main plot comparisons, as there shown, except that our previous figures, being calculated on *totals* of 4 sub-plots, require now to be divided by 4 to get them on a sub-plot basis, and so make them comparable with the rest of the analysis. Let the error so calculated, and based on 10 degrees of freedom, be called error (*a*). This is the error with which to compare the main plot treatments by the z-test, and from which to obtain a standard error for the O, H, and R mean yields. The second part, having 54 degrees of freedom, is now further analyzed into a component due to manurial treatment differences, with 3

degrees of freedom, and calculated in the ordinary way, one for the interaction of sub-plot and main plot treatments, obtained as described above from the two-way table of treatment means, or totals, and having 6 degrees of freedom, and finally an error (*b*), with 45 degrees of freedom, with which these two effects are to be compared by means of the *z*-test, and which will furnish a standard error for the comparison of treatments within the same main plot. Essentially, then, there are two standard errors, one for use along the rows of the table of treatment means, and another for use down the columns.

The full data for this experiment, and the complete analysis, are given below:

TABLE XX.—PLAN, AND YIELDS IN LB., OF SUB-PLOTS. GRASS-LAND EXPERIMENT, CAMBRIDGE, 1931.

O	H	R	R	H	O	O	R	H	O	R	H	H	O	R	O	H	R
F 198	180	200	228	203	247	190	174	168	225	162	149	175	184	144	164	145	116
A 266	213	208	266	222	210	220	184	184	216	207	178	175	202	184	169	142	151
S 184	127	150	157	167	188	140	141	128	174	113	107	112	154	113	116	89	101
C 165	127	155	163	167	150	155	153	118	159	135	125	118	147	98	132	104	69

F= farmyard manure. A=equivalent artificials.
S=equivalent dry matter in straw. C=control, *i.e.*, no manure.

The arrangement of the sub-plots was random within each main plot, but the figures have been rearranged for tabulation purposes.

TABLE XXI.—TABLE OF TREATMENT TOTALS.

	O.	H.	R.	*Total.*
F	1,208	1,020	1,024	3,252
A	1,283	1,114	1,200	3,597
S	956	730	775	2,461
C	908	759	773	2,440
Total ..	4,355	3,623	3,772	11,750

TABLE XXII.—ANALYSIS OF VARIANCE.

Variation.	*Degrees of Freedom.*	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Blocks	5	37,425·11	7,485·02	
Mechanical treatments	2	12,471·03	6,235·52	2·0664
Error (*a*)	10	4,681·14	468·11	0·7718
Manurial treatments ..	3	56,022·72	18,674·24	2·6149
Interaction	6	781·53	130·26	0·1321
Error (*b*)	45	9,091·75	202·04	0·3516
Total	71	120,473·28		

Mechanical treatments have been considered in Section VI.

For manurial treatments $z=2·2633$. For P=0·01, $n_1=3$, $n_2=45$, $z=0·725$ approximately.

The interaction is insignificant.

Standard error of the O, H and R totals of Table XXI. $= \sqrt{(24 \times 468{\cdot}11)} = 106{\cdot}0$, or 2·71 per cent.

Standard error of the F, A, S and C totals $= \sqrt{(18 \times 202{\cdot}04)} = 60{\cdot}31$, or 2·05 per cent.

Since the interaction is insignificant, a fact which is quite evident from Table XXI., though now shown statistically, the summary of results will consist of two tables, showing separately the average effects of the two sets of treatments. The first of these is given in Table XV., and need not be repeated. The other is given below:

TABLE XXIII.—EFFECT OF MANURIAL TREATMENTS.

Mean Yield.	*F.Y.M.*	*Equivalent Artificials.*	*Equivalent D.M. as Straw.*	*Control.*	*Mean.*	*S.E.*
Tons per acre	8·67	9·60	6·56	6·51	7·84	0·161
Per cent. ..	110·7	122·5	83·8	83·1	100·0	2·05

It should be noted that error (*b*) is significantly lower than error (*a*) at the 5 per cent. level. The experiment, in fact, shows great precision on the sub-plot comparisons, and indicates that we may often be able to test in this way treatments whose effect may be small. The differences between manurial treatments in the above experiment are very marked. At the lower end of the scale we have the control and the straw plots, which are practically identical in their yield. The yield of the farmyard manure plots is significantly higher, but there is a further significant response where equivalent artificials were used instead.

The method should not be used when all comparisons are wanted with equal precision, but is useful in cases like the one considered, while a later comparison can often be superimposed on an already existing experiment by this means. The statistical analysis is seen at its simplest when there are only two sub-treatments, for the working out then involves two parallel series of calculations, one on the sums of sub-plot yields and the other on the differences. This is due to the fact that in the case of a sample of 2 the sum of squares of deviations from the mean is equal to one-half of the square of the difference between the sample values. The same method, in principle, is used in the method of sampling, to which we shall devote the next section.

IX.—Sampling Methods.

The practical aspects of sampling will be dealt with in some detail in the second part, but this is the place in which to discuss the principles and the relevant formulæ. Let us suppose that it is desired

to obtain a numerical determination of some characteristic of our experimental material, such as tiller number, yield, or percentage of dry matter, nitrogen or sugar in the crop, by sampling only a proportion of the whole. The object is to obtain as close an estimate as we can of the measure, which would be obtained accurately, within the limits of experimental error, had the produce of the whole plot been counted, weighed or analyzed. It is obvious that the sample must be representative of the whole, and methods have to be devised for securing this end by determining what is a reasonable fraction of the whole to take, and how this fraction shall be selected. In addition we must take account of a further source of error, due to the process of sampling. Suppose a total of $\frac{1}{10}$ of the plot is taken. The measure obtained from the sample, *e.g.* yield, will not when multiplied by 10 agree exactly with the measure from the whole, and so to the ordinary plot error will be added a component of sampling error. This will mean that field experimentation carried out by means of a sampling procedure will never be so accurate as corresponding work based on complete determinations, but it is often advantageous to sacrifice a little to save labour, while it is clear that the sampling method is the only feasible one to use for developmental counts on, for example, a cereal crop, and the only possible one when analytical determinations have to be made at given stages, subsequent work requiring fresh samples.

What we obtain from the sample is an estimate of the plot measure, and we want to ensure that our estimate shall be an unbiased one —that is, neither too high nor too low on the average. Not only so, but a method of procedure which enables us to estimate the amount of sampling error will furnish a valuable check on the adequacy of our technique. The size of sample has an important bearing on this question of sampling error. It is clear, for example, that if we took the whole plot as our sample the error on this account would be reduced to zero.

From the theoretical point of view it is necessary for the *sample* to be made up of a number, say p, of *sampling units*, these being of equal size and selected *at random* from the bulk of the material at our disposal. Let $x_1, x_2, \ldots x_p$ be the measures obtained from the sampling units. These form a statistical sample, from which can be calculated the estimate $\bar{x}$ of the mean, and the estimate s of the standard deviation, of the population of sampling units of which we observe part. $\bar{x}$ is then taken as our best determination of the required measure. Its standard error $s/\sqrt{p}$, which supplies us with a measure of the accuracy of our estimate of the true mean, may be taken as the

sampling error, although there are reasons for preferring the form $s\sqrt{(1-f)/p}$, where f is the fraction of the plot sampled. Thus, for a given size of sampling unit, the larger the number taken the smaller is the error. It is only possible by experimental investigation to determine the sampling error likely to be met with in any given case. If the standard error per plot due to causes other than sampling is s_a, while the sampling error is s_b, we may expect the standard error per plot of the measures determined by sampling to be $\sqrt{(s_a^2+s_b^2)}$, so that, for example, if the plot error is 10 per cent. and we work to a sampling error of 5 per cent., the aggregate plot error will be $\sqrt{125}$, or 11·2 per cent. It is this latter quantity that we can determine from the experimental data, and its size, when compared with the corresponding quantity derived from previous experience of complete experiments, will be a measure of the success of the sampling technique. Having taken the sample, therefore, it is only necessary to concentrate on the mean $\bar{x}$ [or on the total $S(x)$], and pursue the ordinary analysis with these means or totals as the plot data. When it can be done, however, it is advisable to obtain a measure for each sampling unit separately, and so to have the data for calculating the sampling error directly.

The sampling unit may be made up in any way we please by a systematic selection of small areas, or small lengths of row, these latter being termed *units*, and being distributed as far as possible over the entire area of the plot, their total, however, being the only thing it is necessary to record. This is obviously more satisfactory than having single larger areas, which may differ materially in such things as fertility, disease infestation, amount of lodging, etc. It is always open to the investigator to divide his plot into a number of sections, from each of which he may select a minimum of two sampling units at random. Differences between parts of the same plot may thus be eliminated from the estimate of sampling error, just as in field trials part of the soil heterogeneity is eliminated by the block arrangement. The number of sampling units to take from each section will depend on the number of sections, for the sampling error will be determined from " within sections," and we ought to have a reasonably large number of degrees of freedom for its estimation.

All that we have said relates to sampling a single plot, but this last device furnishes us with a reliable and easy method of sampling several plots, as when a field experiment is undertaken. The sampling error is unlikely to be different for different plots of the same experiment, and so by taking a minimum of two random sampling units from each plot we have the requisite data for the calculation of

sampling error with the minimum of trouble. With p plot measures, each determined as the total of q sampling units, the calculation is made from the variation within plots, based on $p(q-1)$ degrees of freedom. With p as large as it usually is, q may be as low as 2. When q is 2, the calculation is especially simple, for, as pointed out in the last section, if x and y are the measures from the sampling units, the plot error is determined from an analysis of variance of $x+y$, while the sampling error is based on the variation within *plots*, and is therefore calculable from the differences $x-y$. The sum of squares of deviations is, for each plot, $\frac{1}{2}(x-y)^2$, and the procedure is therefore to square each difference, add up and divide by $2p$, the number of degrees of freedom being p (1 for each plot). The variance of the mean of the two sampling units is $S(x-y)^2/4p$, and the square root of this is the sampling error. The method is evidently akin to that described in the last section for split plots, except that the whole of the variation within classes of 2 is attributable to sampling error, and is not further divided. Just as in the former case there were two errors, one for comparisons between main plot treatments and the other for the sub-plot comparisons, so in the sampling case the first error is the ordinary plot or experimental error, while the second is the sampling error.

X.—Analysis of Covariance.

We have discussed analysis of variance from the particular point of view of the field experiment, and have seen how the method essentially resolves itself into a test of homogeneity on a sample of data in one variable, namely the yield or other numerical data calculated from the plot. One of the valuable features of the modern design is the way in which the precision of the experiment may be increased by a process which amounts to an equalizing of certain major sources of error among the different treatments. With more than one variable to consider a new field is entered. Now, why is it sometimes advantageous to introduce additional variables? There may be factors which it is impossible to equalize satisfactorily between the different treatments, and yet we may have reason to suppose that greater accuracy would arise from their equalization, were that possible. For example, it is not possible to eliminate fertility differences between the individual plots of a block given over to different treatments, yet a rough assessment of their fertility may be made if the experiment is run for a preliminary year as a uniformity trial, and the plot yields resulting are used as a measure of fertility in the experimental year, assuming fertility to be constant from year to year.

Some other correlated variable may be used instead. Thus it may happen that no preliminary records are available, but a count of the plant population at the beginning of the experiment, or even at the end, if it has been unaffected by treatment, may provide a good index of plot variability. Experimenters are often inclined to distrust the yield figures obtained from plots with very unequal plant numbers, and to insist that a " correction for stand " must be made—*i.e.*, an adjustment of the figures to what they would be if all plots had the same numbers of plants, assuming proportionality. If yield is related to plant number, then evidently the experimental error will be decreased by taking this factor into account and making a correction, but the logical procedure is to see first if such a relation exists. Plant number is here a second variable, and with this brought in we *enter the field of regression and correlation.*

The analysis of covariance is the name given to the technique of testing for homogeneity in problems dealing with two or more correlated variables, and the development of the method in field experimentation has been directed towards further reducing the errors and refining the technique. It has also been used successfully in investigating the interrelations that may exist between different stages in the development of a plant. If a fertilizer treatment has produced more ears, and a greater yield, than the control, we may want to know whether the increased yield was merely a consequence of the larger number of ears, or whether in addition the ears themselves were bigger, or smaller, than those on the control plot. In other words, has treatment had a significant effect on plots adjusted to have the same number of ears ? Particularly where enough reasonably homogeneous material for experimental purposes is hard to come by is the method likely to prove useful. For instance, in animal experimentation the subjects may differ somewhat in initial age and weight, factors which will influence the result materially if a feeding trial is contemplated. These factors may, however, be taken into account as correlated variables, and growth rate, or whatever else is to be measured after treatment, can be corrected to give as nearly as possible the results of a comparable experiment in which the initial factors were standardized.

Let us use x for the independent variable, and y for the dependent variable. It may at times be useful to think of x as, say, plant number and y as yield, to lend concreteness to what some may otherwise consider as a heavy piece of mathematics. The linear regression of y on x is estimated by finding from a sample of p pairs of values the line

$$Y = a + b(x - \bar{x})$$

in which $a=\bar{y}$, $b=S\{y(x-\bar{x})\}/S(x-\bar{x})^2$, are the sample estimates of the unknown parameters α, β in the true regression equation. The numerator of b is the *sum of products* of deviations of x and y from their means, and is most conveniently calculated by choosing *any* arbitrary origins for these variables, and using the formula

$$S\{y(x-\bar{x}\} = S(xy) - \bar{x}\ S(y)$$

where the x and y here stand for the deviations from the assumed means. We therefore sum the products of the respective members of the two series, and subtract the product of the total of one series and the mean of the other. [Note that $\bar{x}\ S(y)=\bar{y}\ S(x)$.] The result will be positive if the variates tend to go up or down together—*i.e.*, if there is positive correlation—and negative if they tend to go in opposite directions.

Now it can readily be shown that

$$S(y-\bar{y})^2 = b^2S(x-\bar{x})^2 + S(y-Y)^2$$

so that the sum of squares of the y's has been analyzed *into* a single square depending on b, and therefore ascribable to the regression, and a sum of squares representing deviations from the regression function. The degrees of freedom appropriate to these two parts are 1 and $p-2$, and as the two are distributed independently of one another, a test of the significance of the regression—*i.e.*, a calculation of the probability that b should exceed this observed value, had β been really zero—is obtained by calculating

$$z=\tfrac{1}{2}\log_e\{b^2S(x-\bar{x})^2/s^2\}, \text{ where } s^2=S(y-Y)^2/(p-2)$$

and consulting the z-table with $n_1=1$, $n_2=p-2$.

Alternatively we have t equal to the square root of the expression under the $\log_e$ sign, with $n=p-2$. If z, or t, is significant, a relationship between the variables has been established, and as this is due to a large part of the variation being isolated in the single degree of freedom due to regression, it is evident that the error of deviations from the regression, estimated by s, will be reduced below the value obtained from $S(y-\bar{y})^2$. Some readers may be more familiar with the calculation of the sample estimate of the coefficient of correlation from the formula

$$r = \frac{S(x-\bar{x})\ (y-\bar{y})}{\sqrt{\{S(x-\bar{x})^2S(y-\bar{y})^2\}}}.$$

The test of significance of r from Fisher's table ("Statistical Methods," Table V.*a*) is exactly equivalent to the above test.

To see how this applies in our tests for homogeneity, consider the simplest kind of experiment as dealt with in section V, where p treatments are tested in q-fold replication, the whole pq plots being randomized over the area. The x-variation, and equally the y-variation, may be analyzed into a part between the means of treatments and a part within treatments. The xy covariation, measured by $S(x-\bar{x})(y-\bar{y})$ and called for short the sum of products, may similarly be analyzed into two parts. The first part is q times the sum of products of the deviations of the x and y treatment means from the general means, while the second is the sum of products of deviations of individual x and y values from the means for the particular treatment of which they are replicate observations, summed for all treatments. The degrees of freedom are $p-1$ for the first part and $p(q-1)$ for the second. The calculation is exactly as described in section V except that we work on products instead of squares, and the formula given above can be used to shorten the work by taking deviations first about some assumed mean. It has already been stated that $Sy(x-\bar{x})$ is the same thing as $S(x-\bar{x})(y-\bar{y})$. The two parts in an analysis of variance are always positive, and add up to the total; here, however, they may be either positive or negative, but they must still add up to the total sum of products, with $pq-1$ degrees of freedom.

Now let us assemble our calculations in an analysis of variance and covariance table shown below (Table XXIV.), denoting for short the sums of squares for x and y by A and B respectively, and the sum of products by C. Undashed letters will denote the variation between groups, dashed letters that within groups, while the corresponding letter with a double dash will be used for the total. It is obvious that $A''=A+A'$, and so on for the others. This relationship will be used when convenient.

TABLE XXIV.—ANALYSIS OF VARIANCE AND COVARIANCE.

Variation.	*Degrees of Freedom.*	(x^2).	(xy).	(y^2).	*Regression Coefficient.*
Between groups ..	$p-1$	A	C	B	$b = C/A$
Within groups ..	$p(q-1)$	A′	C′	B′	$b' = C'/A'$
Total	$pq-1$	A″	C″	B″	$b'' = C''/A''$

Heterogeneity in the simultaneous variation in x and y may show itself in the regression line fitted to the group means having a different slope from the average of the regression lines within groups, or these latter may differ among themselves, or the deviations of the group means from the regression line fitted to them may be more than chance

fluctuations, as, for example, if the regression were non-linear. The regression coefficients appropriate to each line have been calculated in the table, while the sum of squares of deviations $S(y - Y)^2$ is of the form $B - C^2/A$, an expression of this type being derivable from each line. Suppose we calculate this quantity for the " total " and " within groups " lines and subtract. Allowing for the loss of a degree of freedom in each case, we get the following:

TABLE XXV.—ANALYSIS OF RESIDUAL VARIANCE.

	Degrees of Freedom.	*Residual Sum of Squares.*
Total	$pq - 2$	$B'' - C''^2/A''$
Within groups	$pq - p - 1$	$B' - C'^2/A'$
Difference	$p - 1$	$B + C'^2/A' - C''^2/A''$

First note that the significance of the regression relationship between x and y should be tested from the " within groups " line, by comparing C'^2/A', having 1 degree of freedom, with the mean square, s^2, derived from $B' - C'^2/A'$ on dividing by $pq - p - 1$. Only if z is significant will any material advantage be derived from adjusting y for the correlated variable x. The next thing is to test the mean squares obtained from the " difference " and " within groups " lines of Table XXV. z is here also the appropriate criterion, and a significant value indicates that there are differences between the group means after these have been adjusted by means of the " within groups " or error regression. To analyze this effect further, we note that the residual sum of squares between groups, after correcting by the regression b, obtained from the group means, is $B - C^2/A$, with $p - 2$ degrees of freedom. Thus the above " difference " is composed of two parts, as follows:

Degrees of Freedom.	*Sum of Squares.*
1	$C^2/A + C'^2/A' - C''^2/A''$
$p - 2$	$B - C^2/A$

The first part is readily seen, by means of a little algebraic manipulation, to be a single square dependent on the difference between the regression coefficients between and within groups, of the form

$$\frac{A\,A'}{A + A'}(b - b')^2$$

The two parts are distributed independently of one another, and of the residual sum of squares within groups, in such a way that the

mean squares got by dividing by the appropriate numbers of degrees of freedom may be compared with s^2 by the z-test. If the first part is significant, it evidently means that the two regression coefficients are significantly different; in the second case a positive result to the z-test shows that the residuals of the group means from the line of regression fitted to these means are of a greater magnitude than would be expected by chance.

To illustrate with yield and plant number, it may happen that b' is positive and significant, indicating a relation between the variables in sets of plots having the same treatment. Suppose further that b and b' are significantly different. This will mean that there are significant

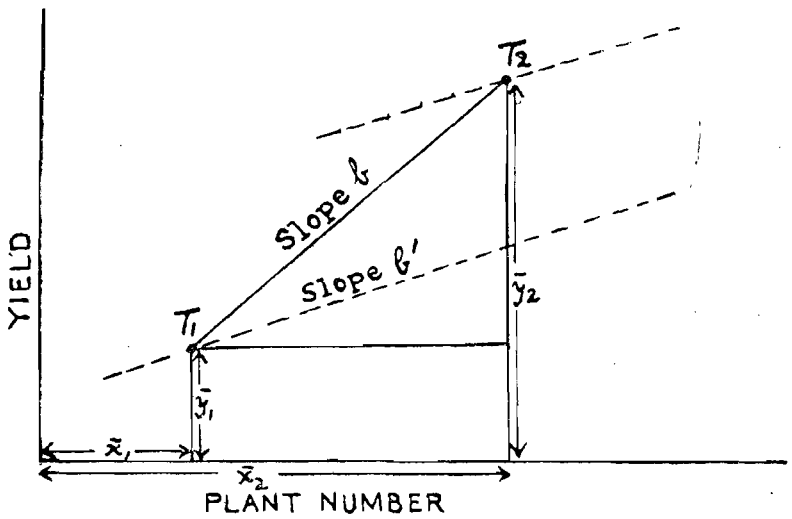


FIG. 1.

differences between treatment means after allowing for the higher value that a mean may have in a group with high plant number—*i.e.*, certain treatments have given a higher yield *per plant*. The diagram illustrates the state of affairs in the case of an experiment with two treatments, in which case the regression line with slope b is the line joining the treatment means.

A possible case in which the residuals of the treatment means from their own line of regression are significantly different is illustrated in the second diagram. Here there has been a significant increase in yield per plant where the plant number was low, an increase which is not shown for those treatments which are associated with high plant number.

We have made the assumption that the separate regression lines within each group can be replaced for the purpose of testing by an "average" line determined from all. It is, of course, possible to test whether the separate lines are significantly different in slope or not, but this is a state of affairs that is not likely to arise in agricultural experimentation, and we shall pass it over, especially as we are generally concerned, not with the simple method so far dealt with in this section, but with experiments of the randomized blocks or Latin square variety, where only one composite error term can be calculated. The way in which to apply the covariance analysis in such cases is to work out the sums of squares and products for blocks (or rows and

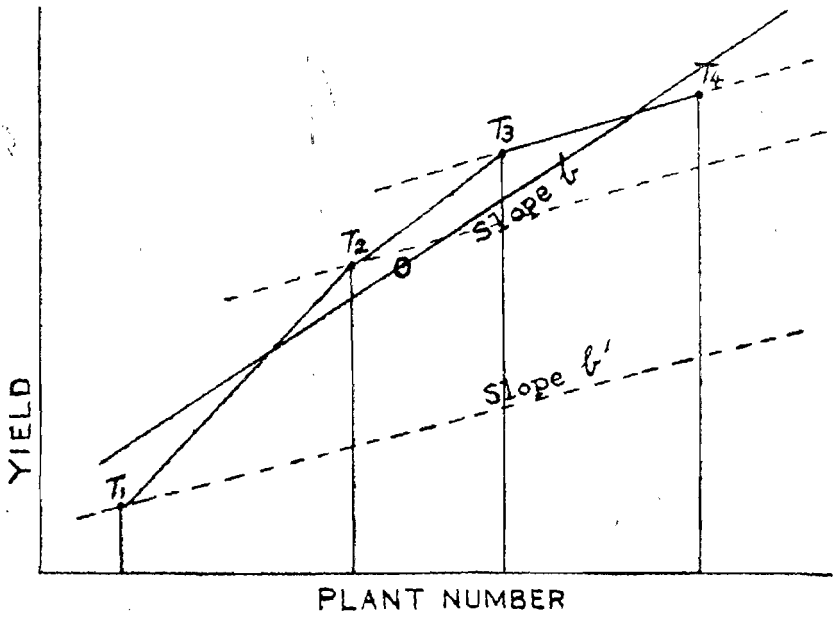


FIG. 2.

columns), treatments and error, ignore the first, and concentrate on the treatment sums with, say, n_1 degrees of freedom, and the error sums with n_2 degrees of freedom. After testing for the significance of the regression calculated from the error term, a similar table to our Table XXV. is drawn up by correcting "treatments" and "treatments+error" by means of their regression coefficients. The tests are then as described, and as a final stage we may draw up a table of the treatment means of the y's, corrected by means of the error regression on x, namely b'. An examination of this table in the light of standard errors which have to be specially calculated for each difference generally tells the whole story.

An example will make the procedure clear. We shall consider the relation between yield (y) and number of stems at harvest (x) in a Cambridge experiment set out to test four sorts of beans, Hunter's (H) and Common (C), and new (N) and old (O) seed, in all combinations.* The experiment had 10 randomized blocks of 4 plots each. Experimental details will be found in the paper cited, and the following calculations are in terms of the units of the data as collected.

TABLE XXVI.—ANALYSIS OF VARIANCE AND COVARIANCE.

Variation.	*Degrees of Freedom.*	(x^2).	(xy).	(y^2).	b.	C^2/A.
Blocks	9					
Treatments ..	3	6,639·3	35,674	239,004	5·373	191,682
Error	27	4,139·2	39,849	798,979	9·627	383,635
Treatments + error	30	10,778·5	75,523	1,037,983		529,176

The blocks contribution may be ignored. From each of the lines "treatments" and "error" the quantities b and C^2/A are calculated from the data of that line. Then the sums of squares and products for treatments and error are added, and a similar calculation made from the data of the "total" line.

We now first test the significance of the error regression, as follows:

	Degrees of Freedom.	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Regression	1	383,635	383,635	1·8236
Deviations	26	415,344	15,975	0·2342
Total	27	798,979		

$z=1·5894$. For $P=0·01$, $n_1=1$, $n_2=26$, $z=1·0220$.

There is thus a strongly significant relation between yield and number of stems at harvest.

The next stage is to prepare a table of analysis of residual variance, which is worked out from the bottom upwards, using the figures of Table XXVI.

TABLE XXVII.—ANALYSIS OF RESIDUAL VARIANCE.

	Degrees of Freedom.	*Sum of Squares.*	*Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).
Difference of regressions	1	46,141	46,141	0·7646
Deviations	2	47,322	23,661	0·4306
Treatments	3	93,463	31,154	0·5682
Error deviations ..	26	415,344	15,975	0·2342
Treatments + error ..	29	508,807		

* Garner, Grantham and Sanders, *J. Agric. Sci.*, 1934, xxiv., 250.

The sums in the second, fourth and fifth lines are obtained by subtracting the C^3/A from the (y^2) component in the lines of Table XXVI. The third sum is obtained by subtraction of the sums in the fourth and fifth lines, while the first is most easily obtained by noting that the excess of the sum of the C^2/A components for treatments and error in Table XXVI. over the corresponding component for treatments+error is 46,141, and measures the difference between the two regression coefficients. As usual, the mean squares are obtained by dividing the sums of squares by the corresponding numbers of degrees of freedom.

We now calculate the various z values, and note that in no case is a significant value reached. This is a point which we shall consider again presently.

Note that the effect of correcting yield for variable number of stems has been to reduce the error variance from 798,979/27, or 29,592, to 15,975—*i.e.*, to about half its former value. We are now ready to consider the treatment means. These are tabulated below.

TABLE XXVIII.—TABLE OF TREATMENT MEANS (OF TEN PLOTS EACH).

	Stems (x).			*Yield* (y).		
	N.	O.	*Mean.*	N.	O.	*Mean.*
H ..	122·8	99·1	110·95	1,066·6	976·1	1,021·35
C ..	110·8	88·3	99·55	947·3	850·0	898·65
Mean	116·8	93·7		1,006·95	913·05	

The ordinary analysis of variance of yield gave no significant results when total treatments, with 3 degrees of freedom, were considered. But on further analysis into H *v.* C, N *v.* O and interaction, with 1 degree of freedom each, it emerged that Hunter's gave a significantly higher yield than Common. New seed gave a higher yield than Old, but the difference was not significant. It should be noticed that Hunter's gave a larger number of stems than Common, and New than Old. The question now is what differences to expect in yield per stem. Is, for example, the higher yield with Hunter's merely a consequence of the higher stem number, consequent on a more favourable germination or seeding rate on plots of this variety? Ignoring for the moment the insignificant z-values arising from the analysis of residual variance on total treatments, let us show how to correct the yield means for variable stem number. Our best estimate of the regression coefficient, freed from possible treatment complications, is $b'=9·627$ (see Table XXVI.), and the calculations are as follows:

TABLE XXIX.—CALCULATION OF MEAN YIELDS, CORRECTED FOR STEM NUMBER.

Treatment.	*Mean No. of Stems* (x).	$x-\bar{x}$.	$b'(x-\bar{x})$.	*Mean Yield* (y).	*Corrected Yield* $y-b'(x-\bar{x})$.
HN ..	122·8	17·55	169·0	1,066·6	897·6
HO ..	99·1	- 6·15	- 59·2	976·1	1,035·3
CN ..	110·8	5·55	53·4	947·3	893·9
CO ..	88·3	- 16·95	- 163·2	850·0	1,013·2
Mean ..	$\bar{x}$=105·25			$\bar{y}$=960·0	960·0

The last column of the table gives the comparative figures for yield of the four combinations of seed when adjusted to equal stem number on the basis of the regression relation. It is the variation of these adjusted yields round their mean that is judged non-significant in Table XXVII., but when arranged as follows, with marginal means added, we see that the principal difference lies in the superiority of Old seed over New, and an examination of this difference by means of its standard error will now be undertaken, a process which will illustrate the method of calculation.

TABLE XXX.—TABLE OF MEAN YIELDS, CORRECTED FOR STEM NUMBER.

	N.	O.	*Mean.*
H	897·6	1,035·3	966·45
C	893·9	1,013·2	953·55
Mean	895·75	1,024·25	960·00

The difference between the N and O means (of 20 plots each) is 128·5. The above calculations show that this is the difference between the N *v.* O yield difference of Table XXVIII., namely 93·9, and the product of the corresponding stem difference, 23·1, and the error regression coefficient, 9·627. The estimated variance of the first part of this difference is evidently $2s^2/20$, where s^2=15,975, while that of b' is s^2/A', A′ being 4,139·2 (see Table XXVI.). The estimated variance of the required difference is therefore

$$s^2\,(2/20+23{\cdot}1^2/4139{\cdot}2).$$

This yields 3,656·9, and its square root, 60·5, is the required standard error. On dividing the difference, 128·5, by 60·5 we obtain t=2·125, with n=26. As the 5 per cent. value is 2·056, we see that Old seed has given a significantly greater yield per stem than New. This is an additional factor of importance to be learnt from the experiment, for we saw that the difference was the other way on total yield, the true facts being obscured by the significantly smaller number of stems produced on the plots having Old seed. It is obvious from Table XXX. without further examination that the difference between

Hunter's and Common is not significant, the earlier significance on uncorrected yields evidently resting entirely on the differences in stem number.

The justification for claiming the difference between the N and O means as significant, although the z of Table XXVII. was not, lies in the fact that the 3 degrees of freedom for treatment can be split up into separate degrees of freedom for testing the H *v.* C and N *v.* O comparisons, and the interaction. Since in each case there is only 1 degree of freedom, and 2 means to compare, the line of regression fitted to these means will degenerate into the line joining the points whose co-ordinates are $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_2, \bar{y}_2)$ (see Fig. 1). The slope is therefore the ratio of $\bar{y}_1 - \bar{y}_2$ to $\bar{x}_1 - \bar{x}_2$, and the treatment effect in the analysis of residual variance, if carried out for the three components separately, will measure the significance of the difference of the slope of this line from that determined from the error regression, while there is nothing left for deviations. Since in each case $n_1=1$, the z-test will evidently be equivalent to the t-test just given, and in fact it was by this means that the standard error of the difference between the corrected means was first worked out. It may assist the reader to understand this point if we give finally the calculations for the N *v.* O comparison.

TABLE XXXI.—ANALYSIS OF VARIANCE AND COVARIANCE (N *v.* O).

	Degrees of Freedom.	(x^2).	(xy).	(y^2).	b.	C^2/A.
N *v.* O	1	5,336·1	21,691	88,172	4·065	88,172
Error	27	4,139·2	39,849	798,979	9·627	383,635
N *v.* O + error ..	28	9,475·3	61,540	887,151		399,689

TABLE XXXII.—ANALYSIS OF RESIDUAL VARIANCE.

	Degrees of Freedom.	*Sum of Squares.*	*Mean Square.*	½ *Log* (*Mean Square*).
Difference of regressions	1	72,118	72,118	0·9879
Error deviations ..	26	415,344	15,975	0·2342
N *v.* O + error ..	27	487,462		

$z=0·7537$. For P=0·05, $n_1=1$, $n_2=26$, $z=0·7205$.

This z is just the natural logarithm of 2·125, the value of t reached earlier.

Similar analyses would be needed for the H *v.* C and interaction comparisons, but inspection shows that there is no need for further calculation.

It is evident that the analysis of covariance technique can be

extended to deal with non-linear regression, and with more than one independent variable, but to go into details would take us somewhat beyond the scope of this book, and we shall conclude this part by expressing the hope that even if the game as here played is found by some readers to be a difficult one, they have at any rate learnt many of the rules.

TABLE OF t.

n.	P=·9	·8	·7	·6	·5	·4	·3	·2	·1	·05	·02	·01
1	·158	·325	·510	·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841
4	·134	·271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4·604
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169
11	·129	·260	·396	·540	·697	·876	1·088	1·363	1·796	2·201	2·718	3·106
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055
13	·128	·259	·394	·538	·694	·870	1·079	1·350	1·771	2·160	2·650	3·012
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977
15	·128	·258	·393	·536	·691	·866	1·074	1·341	1·753	2·131	2·602	2·947
16	·128	·258	·392	·535	·690	·865	1·071	1·337	1·746	2·120	2·583	2·921
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807
24	·127	·256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787
26	·127	·256	·390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750
∞	·12566	·25335	·38532	·52440	·67449	·84162	1·03643	1·28155	1·64485	1·95996	2·32634	2·57582

This table and that in Appendix II. are reproduced, with the kind permission of the author, from "Statistical Methods for Research Workers," by R. A. Fisher. Publishers, Messrs. Oliver and Boyd.

APPENDIX II.

5 PER CENT. POINTS OF THE DISTRIBUTION OF z.

Values of n_2	Values of n_1: 1	2	3	4	5	6	8	12	24	∞
1	2·5421	2·6479	2·6870	2·7071	2·7194	2·7276	2·7380	2·7484	2·7588	2·7693
2	1·4592	1·4722	1·4765	1·4787	1·4800	1·4808	1·4819	1·4830	1·4840	1·4851
3	1·1577	1·1284	1·1137	1·1051	1·0994	1·0953	1·0899	1·0842	1·0781	1·0716
4	1·0212	·9690	·9429	·9272	·9168	·9093	·8993	·8885	·8767	·8639
5	·9441	·8777	·8441	·8236	·8097	·7997	·7862	·7714	·7550	·7368
6	·8948	·8188	·7798	·7558	·7394	·7274	·7112	·6931	·6729	·6499
7	·8606	·7777	·7347	·7080	·6896	·6761	·6576	·6369	·6134	·5862
8	·8355	·7475	·7014	·6725	·6525	·6378	·6175	·5945	·5682	·5371
9	·8163	·7242	·6757	·6450	·6238	·6080	·5862	·5613	·5324	·4979
10	·8012	·7058	·6553	·6232	·6009	·5843	·5611	·5346	·5035	·4657
11	·7889	·6909	·6387	·6055	·5822	·5648	·5406	·5126	·4795	·4387
12	·7788	·6786	·6250	·5907	·5666	·5487	·5234	·4941	·4592	·4156
13	·7703	·6682	·6134	·5783	·5535	·5350	·5089	·4785	·4419	·3957
14	·7630	·6594	·6036	·5677	·5423	·5233	·4964	·4649	·4269	·3782
15	·7568	·6518	·5950	·5585	·5326	·5131	·4855	·4532	·4138	·3628
16	·7514	·6451	·5876	·5505	·5241	·5042	·4760	·4428	·4022	·3490
17	·7466	·6393	·5811	·5434	·5166	·4964	·4676	·4337	·3919	·3366
18	·7424	·6341	·5753	·5371	·5099	·4894	·4602	·4255	·3827	·3253
19	·7386	·6295	·5701	·5315	·5040	·4832	·4535	·4182	·3743	·3151
20	·7352	·6254	·5654	·5265	·4986	·4776	·4474	·4116	·3668	·3057
21	·7322	·6216	·5612	·5219	·4938	·4725	·4420	·4055	·3599	·2971
22	·7294	·6182	·5574	·5178	·4894	·4679	·4370	·4001	·3536	·2892
23	·7269	·6151	·5540	·5140	·4854	·4636	·4325	·3950	·3478	·2818
24	·7246	·6123	·5508	·5106	·4817	·4598	·4283	·3904	·3425	·2749
25	·7225	·6097	·5478	·5074	·4783	·4562	·4244	·3862	·3376	·2685
26	·7205	·6073	·5451	·5045	·4752	·4529	·4209	·3823	·3330	·2625
27	·7187	·6051	·5427	·5017	·4723	·4499	·4176	·3786	·3287	·2569
28	·7171	·6030	·5403	·4992	·4696	·4471	·4146	·3752	·3248	·2516
29	·7155	·6011	·5382	·4969	·4671	·4444	·4117	·3720	·3211	·2466
30	·7141	·5994	·5362	·4947	·4648	·4420	·4090	·3691	·3176	·2419
60	·6933	·5738	·5073	·4632	·4311	·4064	·3702	·3255	·2654	·1644
∞	·6729	·5486	·4787	·4319	·3974	·3706	·3309	·2804	·2085	0

APPENDIX II.

1 PER CENT. POINTS OF THE DISTRIBUTION OF z.

Values of n_2	Values of n_1: 1	2	3	4	5	6	8	12	24	∞
1	4·1535	4·2585	4·2974	4·3175	4·3297	4·3379	4·3482	4·3585	4·3689	4·3794
2	2·2950	2·2976	2·2984	2·2988	2·2991	2·2992	2·2994	2·2997	2·2999	2·3001
3	1·7649	1·7140	1·6915	1·6786	1·6703	1·6645	1·6569	1·6489	1·6404	1·6314
4	1·5270	1·4452	1·4075	1·3856	1·3711	1·3609	1·3473	1·3327	1·3170	1·3000
5	1·3943	1·2929	1·2449	1·2164	1·1974	1·1838	1·1644	1·1457	1·1239	1·0997
6	1·3103	1·1955	1·1401	1·1068	1·0843	1·0680	1·0460	1·0218	·9948	·9643
7	1·2526	1·1281	1·0672	1·0300	1·0048	·9864	·9614	·9335	·9020	·8658
8	1·2106	1·0787	1·0135	·9734	·9459	·9259	·8983	·8673	·8319	·7904
9	1·1786	1·0411	·9724	·9299	·9006	·8791	·8494	·8157	·7769	·7305
10	1·1535	1·0114	·9399	·8954	·8646	·8419	·8104	·7744	·7324	·6816
11	1·1333	·9874	·9136	·8674	·8354	·8116	·7785	·7405	·6958	·6408
12	1·1166	·9677	·8919	·8443	·8111	·7864	·7520	·7122	·6649	·6061
13	1·1027	·9511	·8737	·8248	·7907	·7652	·7295	·6882	·6386	·5761
14	1·0909	·9370	·8581	·8082	·7732	·7471	·7103	·6675	·6159	·5500
15	1·0807	·9249	·8448	·7939	·7582	·7314	·6937	·6496	·5961	·5269
16	1·0719	·9144	·8331	·7814	·7450	·7177	·6791	·6339	·5786	·5064
17	1·0641	·9051	·8229	·7705	·7335	·7057	·6663	·6199	·5630	·4879
18	1·0572	·8970	·8138	·7607	·7232	·6950	·6549	·6075	·5516	·4712
19	1·0511	·8897	·8057	·7521	·7140	·6854	·6447	·5964	·5366	·4560
20	1·0457	·8831	·7985	·7443	·7058	·6768	·6355	·5864	·5253	·4421
21	1·0408	·8772	·7920	·7372	·6984	·6690	·6272	·5773	·5150	·4294
22	1·0363	·8719	·7860	·7309	·6916	·6620	·6196	·5691	·5056	·4176
23	1·0322	·8670	·7806	·7251	·6855	·6555	·6127	·5615	·4969	·4068
24	1·0285	·8626	·7757	·7197	·6799	·6496	·6064	·5545	·4890	·3967
25	1·0251	·8585	·7712	·7148	·6747	·6442	·6006	·5481	·4816	·3872
26	1·0220	·8548	·7670	·7103	·6699	·6392	·5952	·5422	·4748	·3784
27	1·0191	·8513	·7631	·7062	·6655	·6346	·5902	·5367	·4685	·3701
28	1·0164	·8481	·7595	·7023	·6614	·6303	·5856	·5316	·4626	·3624
29	1·0139	·8451	·7562	·6987	·6576	·6263	·5813	·5269	·4570	·3550
30	1·0116	·8423	·7531	·6954	·6540	·6226	·5773	·5224	·4519	·3481
60	·9784	·8025	·7086	·6472	·6028	·5687	·5189	·4574	·3746	·2352
∞	·9462	·7636	·6651	·5999	·5522	·5152	·4604	·3908	·2913	0

PART II.—PRACTICAL CONSIDERATIONS

I.—Considerations of Policy and General Procedure.

Part I. deals with the statistical principles underlying modern field experimentation; in Part II. we shall be concerned with the considerations which arise when those principles are put into practice. Illustrations will be drawn from English experience, but tropical readers can be assured that the problems arising with their crops will be essentially similar, and so no apology is offered for basing the discussion on English crops.

In these days it is difficult, but very important, to keep a sense of proportion over this question of experimentation. The statistical side has been given so much prominence in recent years that there is a real danger of statistics being regarded as the main interest in experimentation. The science of statistics is, however, only a weapon in the experimenter's equipment, and it must be allowed no greater place in his thoughts than the chemical or botanical techniques he employs. There is much room (and this must be specially true in a new country) for enquiries of a general nature, in which statistics can play little or no part. It may be that an agricultural officer can do most good by working out and demonstrating a new farming system; such work will necessitate much knowledge of the country he is serving, and will involve wide economic and social considerations, but will call for no knowledge of statistics. In an endeavour of that nature he will have no need nor time for formal trials, and his freedom from them will enable him to make his activities more widespread.

In no case should precise experiments be laid down before a thorough survey has been made of existing methods. In this respect a famous military maxim should be remembered: "Time spent in reconnaissance is seldom wasted." It may be that methods proved successful in other countries can be adapted: that a finger can be placed on a weak spot such as seed supply, and that the greatest scope for usefulness lies in strengthening this weak spot by organizational, rather than experimental, methods: there may even be room for doubt as to which are the best crops to grow. In all such matters statistics has neither lot nor part, yet they are matters of great moment. The statistical technique is designed to measure something, to reduce a problem to figures on which a known degree of reliance can be placed;

a knowledge of this technique should not blind us to the fact that there are many enquiries to which it is quite unsuited.

Even when broad questions of policy have been settled, there often remains a necessity for exploratory trials before definite experiments can logically be undertaken. A formal experiment is a very rigid thing, designed and carried through with meticulous care to provide precise answers to a few definite questions; the value of the experiment must depend on how wisely the questions have been formulated. It will rarely prove wise to conduct a definite experiment on a crop with which the experimenter is unfamiliar. Common sense dictates that with a new crop the first step should be to grow it normally, keeping it under observation, for a year or two; in this, though different methods of husbandry may be tried, there should be no thought of arriving at comparative figures. From such a general observation of the crop specific questions will arise; if these can be stated in clear-cut terms then they may reasonably form the subject of an experiment. It is very salutary to sit down before commencing an experiment and to write out its objects; vague hopes will not do, and if these are all that materialize the time is not ripe for a precise trial. Both the strength and the weakness of a formal experiment lie in its rigidity; to use it to the best advantage its objects must be definite, and this destroys its value for enquiries of a preliminary nature.

How many questions should a single experiment be expected to answer? Should it be concerned solely with one comparison, or can a number be included? This is a matter where modern ideas differ sharply from earlier ones. Until recently it was regarded as essential that an experiment should be simple, but the methods described in Part I. are clearly designed for experiments comparing a number of treatments. A great argument in favour of complicated experiments lies in the fact that they give results of wider applicability than do simple ones. To take an illustration, the comparative merits of a strong and a weak strawed variety of wheat may well depend on the fertility of the soil on which they are grown; on rich land the weak strawed variety may lodge long before harvest and give a low yield, whereas on poor land it may equal or surpass the strong strawed variety. The result of a simple comparison of the varieties will obviously depend very largely on the fertility of the experimental area, but if high and low manuring be introduced into the trial a much fuller knowledge of the relative merits of the two will be obtained. It must be admitted that this argument, carried to its logical conclusion, almost appears to favour complication *per se*, but in general the number of different

treatments introduced into one experiment is limited rather severely by considerations of space and design. The point will have to be discussed again later; here it is only desired to make it clear that if exploratory trials lead to the formulation of several definite questions they can often be incorporated with advantage in one and the same experiment.

It is very rarely that a worker well versed in his subject has any difficulty in thinking of suitably definite problems for experimentation. His trouble is usually the reverse, and the common difficulty is to keep an experimental programme within reasonable limits. It is very unwise to undertake too much; one experiment properly conducted is infinitely preferable to two that are scamped. An experiment is very exacting. It needs protection at all stages; in England this often only entails bird scaring, but cases of elephants ruining plots have been reported from Africa. Difficulties of storage often arise, and may tax the resources of a small field station very severely. The chief difficulty in most cases, however, is the labour one; all work on a formal trial must be done at the right time, and any one operation should not occupy more than a day or two. It is beside the point to argue that a normal crop would not receive such precedence; a normal crop is not required to give reliable comparative figures, and an experiment will not be made more normal by haphazard attention which may affect treatments differentially. In addition to the periods when it entails bursts of work, an experiment should be kept very closely under observation throughout its existence, for otherwise results may be wrongly interpreted. The importance of restricting experimental work within reasonable limits is very obvious, but, having had some experience, we feel justified in stressing the point; the whole tendency is for the work to expand, and this must be sternly controlled.

Much care must be taken in choosing the site for an experiment. Some people have the impression that uniformity of soil over an experimental area is no longer important, because they imagine that soil inequalities will all come out in the statistical wash. This is not true. The great strength of modern methods is that they protect the experimenter from drawing wrong conclusions when differences are merely due to soil irregularities. But we want more than this negative assistance; we want to detect true differences when they do exist, and this is often prevented by soil patchiness. Two things make for significant results—large differences between the yields from different treatments, and a low experimental error. The placing and design of an experiment do not affect the former, but they should

reduce the error as much as possible; unfortunately no design can give a low error where great soil heterogeneity exists. In a simple randomized blocks experiment the design makes it possible to take away from error the variation between the mean yields from the different blocks, and so to allow for fertility differences from block to block; but fertility differences within a block give rise to irregularities for which allowance cannot be made, and which therefore go to swell the error. Thus the more even the soil the greater will be the efficiency of the experiment, and many cases occur where large differences between treatments are shown, but where significance is missed because of irregularities of soil and the consequent large experimental errors. It is sometimes argued that experiments should be conducted on uneven soil because thus they will be more representative. Representative of what? Presumably of soil with a certain degree of unevenness, but it is difficult to see how this helps. If we take heaviness and lightness as the unevenness concerned, would the average result over the field be applicable to loam? Similar fallacies will always appear in trying to apply results from a soil that is very mixed, and mixed in a random and undefined manner.

Uniformity must be interpreted very widely in selecting an experimental site. It is not only the surface soil that must be considered; the subsoil must also be investigated, and a number of holes should be dug before placing an experiment on an unknown piece of land. Drainage must also be taken into account, together with such obvious features as slopes. The previous management of the field should be carefully studied; if possible an area that has been farmed "in one piece" for several years should be selected, so that cropping, cultivation and manuring will have been the same all over it. In some cases it may be necessary to overlap parts of a field which have been differently cropped recently; with care in arranging that the line of division becomes a line between blocks in the experiment, any fertility difference may be eliminated. In general any inequalities should be avoided, and it is very helpful to keep an accurate plan of all fields, on which the sites of experiments are carefully marked, so that future experiments may be kept clear of recent plots.

Picking a uniform piece of land for an experiment is often a difficult matter, but it is easier when the experimenter has been familiar with the field under normal cultivation for a few years. Experience of "straight" crops will have indicated where the bad soil patches lie. In some cases it may be worth while to carry out a uniformity trial in a previous year on the actual plots that are to be used experimentally, with a view to using the covariance method of

statistical analysis described in Part I. It will be remembered that the principle of the procedure is that the figure for a plot is corrected for the figure given by it in another set of observations. In the present case the yield of a plot in the experimental year is corrected for its yield in the preliminary uniformity trial; if a high correlation exists between the yields of particular plots in the two years (that is, if a high yielding plot in the first year proves to be high yielding in the second, and *vice versa*) this may lead to a large reduction of error, but in some cases the correlation is low and then the preliminary labour will have been thrown away. It is easy to see how a low correlation may arise. The soil may vary in heaviness, and the first year may be a dry one while the second is wet; the heavier patches of the field will yield relatively well in the first year and relatively poorly in the second, and thus a low correlation, or even a negative one, may emerge. As regards the efficacy of the correction the sign of the correlation is immaterial, but in practice a negative one numerically high enough to be serviceable could scarcely arise. The hope lies, then, in the possibility that a patch of land proving itself fertile or infertile in one year may show approximately the same relative fertility or infertility in the next year. The possibility has not been very widely explored, but it appears that with annual crops the correlation is generally low, and hence that the method has little to recommend it. In order to use the method the actual experimental plots must be severally harvested in the first year, when they have all been treated alike, with the same care as in the second, or experimental, year; thus the total labour is very materially increased and only a considerable gain in precision could justify the extra work. With perennial crops, however, greater possibilities appear to exist, for with them not only the soil but also the plants remain the same on each plot, and higher correlations are usually found. In a manurial experiment on a perennial crop the plots might be marked out a year before applying the manures whose effects are to be compared, and the collection of the yields from the separate plots in that year might not involve much dislocation of normal practice. Thus with perennial crops the method is worth a trial, though sufficient experience has not yet accumulated to warrant the expression of a definite opinion. It may be said that it will scarcely justify itself unless high correlation coefficients emerge.

Care must also be bestowed on the selection of the seed for an experiment. In this the same considerations arise as in the selection of the site, the great aim being to avoid introducing chance irregularities. The object in planning an experiment must always

be to achieve the utmost evenness over the plots in all factors other than the treatments, in order that nothing may blur the experimental comparisons. With all factors, save one, which are likely to cause inequality, common sense will usually indicate the correct procedure; the exception is the soil, which despite careful selection will inevitably be irregular, and inequalities in that are minimized, but unfortunately cannot be obliterated, by correct experimental design. In some enquiries the treatments are not imposed until the crop is well up; in such circumstances the error is usually considerably reduced, because the early hazards are past when the plots are marked out, and an area can be selected which avoids inequalities in plant establishment and early growth.

II.—The Agricultural Significance of Experimental Results.

The pure scientist is in a happy position in that he has only to seek for knowledge, but the agricultural worker, to justify himself, must make discoveries of direct value to agriculture, and on him there usually devolves the task of disseminating the information he gains, so that it may influence farming as quickly and as widely as possible. Many cases could be cited where an experiment has been successful, in that it has given definite answers to certain specific questions, but where it has led to no improvement of common practice. It is clearly necessary to make some remarks on the limitations of experimental results when it comes to their practical application.

The best place for an experiment is on a normal farm or estate where the standard of husbandry is high. In general this will prove the cheapest plan, and it will ensure that the crop experimented upon will be in its usual place in a common rotation or, where there is no rotation, treated in a normal fashion. If an experimental farm is to be run normally the experiments on it must only occupy a small proportion of its total area. This will have two distinct advantages, for it will make it easier to avoid the sites of recent experiments, and it will generally mean that the plots will be surrounded by considerable areas of the same crop; this latter is important, because small isolated patches of a crop invariably suffer from a variety of depredations. That the farm should be farmed well is of great importance if it is intended to use it for demonstration as well as for experiment; nothing impresses farmers more than a high standard of husbandry, and the better run the farm the greater will be the conviction that the demonstrations and experiments upon it will

carry. A high standard of farming will guard against the temptation to put down manurial experiments on especially poor land, to ensure striking results. In some manurial enquiries the selection of poor land may be defensible, as, for instance, when it is required to compare the responses from applications at different dates, but in general manurial trials should be conducted on land well up to the average; the object is to obtain knowledge applicable to such land, not merely to make the trials " come off."

In order to link an experiment directly to practice it is a good plan to include an established variety or practice as one of the treatments, to act as a standard. The soil type of an experimental farm should, of course, be representative of a wide expanse of country. If the area it is intended to serve is predominantly sticky clay the experimenter is unlucky; he is not justified in choosing an isolated piece of free-working land, on the plea that only on such soil can he carry through his experiments as planned. On sticky land it is sometimes impossible to sow an experiment exactly when required, and the taking of detailed observations on the plots may occasionally be difficult or even impossible, but the major limitations will be those to which the area is subject, and personal inconvenience should not weigh against direct applicability of results. After considerable experience of running experiments on some of the stickiest land in the world, and in a tricky climate, it can be said that only rarely is any precision necessarily lost, and that a programme of work can be adhered to pretty closely; what is required is extra care to avoid damaging soil texture, and an adequate labour force so that the maximum advantage can be taken of favourable weather. In the case of the plant breeder, dealing in great detail with very small plots, specially selected soil may be necessary, but in field trials difficulties of soil must be overcome, not evaded.

The outstanding agricultural limitation of an experiment lies in the fact that its result is only strictly applicable to the particular field in that particular year. When the variation that exists between different fields—in soil type, in fertility, in cleanliness, in drainage—is considered, together with the vagaries of climate and the diverse methods of management used for the crop concerned, the greatest hesitancy must be felt in predicting similar results in other situations. Nevertheless, experimental results must be applied widely. It follows that a single experiment can be of little agricultural value, and that practical recommendations can be safely based only on an extensive series of experiments. It should be a rigid rule to continue one enquiry for at least three years before drawing definite conclusions,

in order that a fair sample of the weather may be encountered. Three years is the absolute minimum, and even though the soil type and general condition of the fields used may be approximately the same in each year, results will often be inconsistent. Other soil types and situations must also be covered, and the only satisfactory procedure is for the experiment to be repeated at a number of widely scattered centres for several years; this may appear a counsel of perfection, but it is possible where two or three experimenters can agree to work on the same problem, and it is probable that in future greater use will be made of this " community " method of enquiry.

There is another reason why an experiment must be repeated—namely, that the result, though significant, may be untrustworthy. This is a disturbing thought, but it must be faced. In field experiments it is usual to take a low standard for significance, and the one most commonly adopted is a 19 : 1 chance; it must be realized that of twenty differences just reaching this standard one, on the average, will have arisen purely by chance, and the experimenter does not know which is the one. It is comfortable to regard a significant result as something mathematically established, and therefore as certain as that two sides of an isosceles triangle really are equal; but statistical proof is a matter of chances, and the possibility that a result judged significant on the low standard of 19 : 1 may be an odd chance cannot be ignored. In actual fact some unknown factor may act differentially on the treatments, so that a greater proportion of results may be unreliable. In any case a bare 19 : 1 chance must not be allowed too much credence, and the best safeguard against the chance results which do occur is to repeat the experiment.

Where an experiment is repeated at a number of centres and over several years it is possible to combine all the results in one table of the analysis of variance. If this is to be done the size and shape of plot must remain constant, as also must the form of the lay-out, though the actual randomization of the treatments must be done separately for each experiment. Statistically there is much to be said for combining experiments in this way, because in the final table there will be more degrees of freedom appropriate to error, which therefore will be estimated with greater precision. If it is regarded as unimportant to obtain significance at any one centre, the lay-out may be very simple, with very limited replication; thus at each centre the trial will really be only a demonstration, but the whole series considered together may provide adequate replication, and precise results may emerge. An instance where this method has been used successfully is mentioned on page 73. The average result

a series of experiments may not, however, be very valuable. To lustrate this point the wheat varieties Little Joss and Rivet will erve. It is well known that Little Joss succeeds on light land and s not suited to sticky clay, whereas Rivet is particularly valuable for its ability to yield well on the stickiest clay, but is quite unsuited to light land. If the two varieties were compared at a number of different centres the average result might indicate that their yielding capabilities were very much alike, and it is conceivable that the really important agricultural difference—that in their adaptability to heavy and light land—might be missed. It is true that a full statistical analysis of the results might bring the point out in the form of a significant interaction, but this would necessitate the grouping of the centres according to soil type. It follows that to ensure obtaining all the information that the series had to offer the centres would have to be classified according to a number of different factors, and probably it would be easier, assuming that each experiment contained adequate replication within itself, to extract the full information by keeping the centres separate in the statistical work. In such cases the best procedure is to work the experiments out individually first; if they point fairly consistently in one direction greater certainty may be obtained by combining them, but if they give varying results it is futile to compute an average applicable to no one centre.

The practical man often scoffs at modern methods of fiel experimentation on the ground that they are designed to measur differences so small that they do not really matter. This is n serious argument against the methods, for if they will detect sma differences they will show up larger ones with great certainty; but tl question as to whether a difference, though it may be significant, large enough to have any practical value is one that merits care consideration. It is a question most difficult to decide in the case a new variety. On many grounds it is very undesirable that r varieties should be introduced unless they show some real adva on those already established in practice. Whether any y difference of less than 10 per cent. would justify the introduction new variety is very much open to doubt; in any case the decision w have to be made not on the yield difference alone, but also on c characteristics, for yield, though important, is only part of the s

That a variety or treatment cannot be assessed on yield alc one of the great agricultural limitations of experimental results is true to say that no variety nor treatment can be good unl ...ges a high yield, but there are many other factors whic

contribute to the one great aim—remunerativeness. Quality may be as important as yield. In some cases due allowance for quality may be made; thus, in work on sugar beet, analysis of samples will enable the yield to be expressed in terms of sugar, and thus quality will be incorporated in the final figure. In many cases, however, quality cannot be satisfactorily appraised. With barley quality is fairly well understood and can even be expressed with reasonable accuracy in one figure (the nitrogen percentage of the grain), but this cannot be combined with yield. The difficulty is that maltsters in England only require a proportion of the barley grown in the country, and will only buy the best. It is no use demonstrating that some barley is but slightly below the usual malting standard, and calculating an extract figure to allow for this; the point is that if the quality is not quite good enough the maltsters will not buy it, and then it must be sold for feeding at a much lower price.

Pasture is probably the most difficult crop on which to assess quality. The yield of green herbage from a field is obviously quite inadequate as a measure of its productivity. In experimental work it is common to sample the herbage from each plot and to determine the dry matter percentage, so that the yield may be expressed in terms of dry matter. But this is by no means satisfactory, because the nutritive value of the dry matter depends, among other things, on the botanical composition of the sward; no reasonable method of incorporating this has been devised, and even if it were the very important characteristics of palatability and digestibility would remain uncovered. The great difficulties presented by this method of approach have led some experimenters to try to measure the productivity of a grass plot directly, by feeding it off with stock and expressing the yield in terms of live weight gain. There are, however, many drawbacks to that method. In the first place, only a few animals can be folded on each plot of a replicated experiment, so that in the final result the experimental error will be augmented very greatly by a component due to the variation between animals; the error term, in fact, is usually so large that no significant results emerge. Another grave difficulty arises in deciding how many animals should be put on each plot. If one treatment produces much herbage, should more animals be used for its plots than for those of other treatments, and, if so, what allowance can be made for the extra food required for the maintenance of the greater number of animals? There appear to be no satisfactory answers to these questions, and there are many other difficulties, such as those of fencing and water, or that occurring when one of the animals on the

plots is ill. It is very clear that the true result of an experiment on grass is extremely difficult to obtain, and hence the agricultural significance of such trials is very doubtful. This is probably the experimenter's most difficult crop, and the above brief sketch of the problems it offers will serve to illustrate the sort of embarrassments which arise in assessing that elusive thing quality.

There are many other features which must be considered before a final opinion of a variety or treatment is formed. With cereal varieties ability to stand is very important; a variety may show an appreciable advance in yield and yet be unworthy of a place in farming if it is very liable to lodge. In a similar way disease resistance may be a characteristic of overriding importance. The decisive part that earliness may play is well illustrated by the great strides made in Canadian wheat farming during the present century. Considerations of this nature are liable to arise when comparing cultural treatments as well as with varieties. It can only be reiterated that yield is not everything, and in the whole story due space must be allotted to those ancillary features which may, in the end, determine remunerativeness. That they cannot be measured, and are thus unsuitable for precise experiment, is no reason for leaving them out of account.

Ancillary features must, therefore, form the subject of much careful observation on experimental plots, and must be duly considered in applying results to practice. Small plots are apt to be unsuitable in this respect, and for this and other reasons it is very desirable that precise experiments should be accompanied by observation plots. These form the subject of the next section.

III.—Observation Plots.

An observation plot differs from an experimental plot in that it is usually much larger, and that it is treated in all respects in a routine practical manner, so that it need involve no appreciable expense. An observation plot may be of any size, but to derive the full benefit from it an area of about an acre is desirable, to ensure normal cultivations, and because on smaller areas mass effects will not appear. The object is not to obtain a figure. It is generally best not to weigh the produce of observation plots, for the results can have no value for comparative purposes, since no standard errors are attached to them; if the produce is not weighed there is no temptation to give to figures for yield credence to which they are not entitled.

When a new line of enquiry is initiated the observation plot

should be the first step. It may be that a treatment will prove unpractical under ordinary farming conditions, that unforeseen difficulties will preclude its adoption, and it is much better to learn the fact with little expenditure of time and money from an observation plot, than from a costly and laborious experiment. A new method may raise yield so much that the difference has only to be seen to be believed; in such a case it is wasteful to spend time proving the significance of the increase, and the better procedure would be to withhold precise trials for comparing various modifications of the method. It is in questions of this nature that a sense of proportion must be preserved; the aim of the experimenter is to ascertain new facts, and if he can do so without having to resort to statistics so much the better. One of the great advances made in agricultural science in England during the present century was the discovery of the value of basic slag for pasture on heavy land. This was successfully demonstrated on single large plots at Cockle Park, the benefits being so great that no replication was necessary; in this case the main advance in knowledge was made with the minimum of effort, but it led to queries that could only be solved by more precise methods of experimentation.

Observation plots are freely used by plant breeders as a means of sorting out their new varieties, and deciding which are worth carrying forward to precise trials. Some elimination must be made in the early stages, and with the large number of progeny which arises from a single cross it is only possible to grow single plots of a few which appear hopeful. It is worth noting, as an indication of the limitations of single plots, that the decision as to which to continue for trial is always difficult, and that undoubtedly many mistakes are made; the successful plant breeder is still the man who has an instinct for "spotting a winner."

It has been said that observation plots have an important place in the early stages of an enquiry. They often have considerable value throughout an investigation, and it is always a wise practice, when putting down an experiment, to apply the treatments severally to large areas or observation plots. Any abnormality in cultural method which may be necessary on the small experimental plots will be avoided on the observation plots, and in some cases the experimenter may be safeguarded from faulty conclusions. Grassland is very much influenced by the animals which graze upon it, and the treatment of experimental plots as regards grazing can rarely be kept even approximately normal; this is a case where it is very desirable that observation plots should accompany a precise trial.

With cereal crops standing ability is very important, and is most easily appreciated by watching the movement of large areas of the crop under tempestuous conditions.

Much attention has been directed during the last decade to spacing, which has been shown to affect yield very considerably; on small plots spacing usually has some peculiarities as compared with field scale husbandry. With sugar beet, practical experience has shown that a full " plant " is necessary if a full yield is to be obtained, but experiments comparing populations of 20,000 and 40,000 per acre have not shown appreciable differences in yield. The discrepancy between experiment and practice probably arises because where an experimental plot has only 20,000 plants per acre, they are regularly spaced, whilst such a low population on a normal field means that large gaps occur; in the plot the widely spaced beet compensate for their low number by greater individual growth, but with the irregularities in the field some areas are thickly populated, whilst on others occur bare spaces so large that the neighbouring roots cannot compensate for them. Here is a case where the very care taken over the plots differentiates them from normality, and produces results inapplicable to practice. The solution of the problem requires fully replicated experiments conducted under field conditions, and observation plots accompanying the initial precise trials would have served to show the need for this further step before making practical recommendations.

The cereal plant breeder requires to observe each plant of a new cross very closely, and for this it is necessary to dibble small plots with the plants spaced regularly along the row. It will be readily appreciated that results in such trials must be checked under more normal conditions. The tillering of regularly spaced plants may not parallel that of plants in a drilled field, in which there may be anything from 1 or 2 up to 30 or 40 per foot of drill row, and tillering may affect both yield and quality to a high degree. To illustrate the possibilities as regards quality an experience with barley will be quoted. A variety was grown in a chessboard trial with others, the plots being small and regularly dibbled, and during the same year it was grown in an observation plot, which was drilled, on similar soil in a neighbouring field. The quality of the barley from the observation plot was high—a result agreeing with the experience of the variety in general farming—but the barley from the small plots was of poor quality, containing a large proportion of extremely small and very imperfectly filled grain. The disagreement was probably due to differential response to weather conditions, which were very

favourable to tillering in spring. On the regularly spaced plots a great profusion of late tillers produced ears, and when a drought set in about midsummer many of the late formed ears dried out prematurely, to give small grain. This profuse tillering did not occur under field conditions, nor did it occur on the chessboard with the other varieties. The observation plot in this case was extremely valuable, for without it a really high quality variety would have been judged a poor quality one, through the abnormality introduced by the regular spacing of the dibbled plots.

If observation plots are to accompany a precise experiment there is much to be said for placing them in the same field. When an experiment is being demonstrated to farmers it is very impressive to ask them to turn round and observe the results of applying the treatments over large areas; in most cases it is found that more interest is taken in the large than in the small plots. The major portion of the field is made to serve for demonstration, and to act as a practical check on the precise results obtained from the small experimental area. The idea of replication is so ingrained in agricultural workers nowadays that observation plots are often duplicated in the same field. There is little to be said for this, except that in comparing treatments by eye it is convenient to have them immediately adjacent, and duplication allows more chance for direct comparison.

An alternative to arranging observation plots alongside experimental areas is to scatter them over as wide an expanse of country as possible. If new varieties are being compared it may be possible to arrange for them to be grown on, say, half-acre plots at a number of different centres, and this may entail little organization or expense. The wide scatter will ensure that many farmers will see the varieties, and will provide a number of opportunities for observing such features as winter hardiness. With some crops these plots may be used for "growing on" the seed. In England and Wales the National Institute of Agricultural Botany has used single plots at a number of centres with great success, and has arranged them so that, in addition to observations as to adaptability to various situations, precise yield comparisons could be obtained. The statistical principle is, of course, the simple one of treating each centre as a block in an ordinary experiment, the replication being given by the number of centres. At each centre all that is required is a single plot, of a definite size and shape, of each variety; to keep the experimental error as low as possible these single plots should be contiguous, and the seed of each variety used at all centres should be from a common source. The dislocation of farm practice involved

is very minor, and the Institute has been able to arrange for nearly a hundred centres, on ordinary farms, in one trial. The varieties included have already been tested carefully for several years at a few experimental stations, so that their merits are fairly well known; what the large number of single plot centres accomplishes is to advertise them, and to check the earlier results under very diverse conditions, so that subsequent recommendations can be made with great confidence. Readers are reminded of what was said on page 68 as to the limitations of the average result from such a series. With a large number of centres, however, much more than a bare average can be obtained. The centres can be classified by district, by altitude, by soil type, by rainfall, by fertility (on the basis of the average yield of all the varieties), by date of sowing, or by any other factor to which the varieties might conceivably show differential response. By making the classifications broad enough there may be sufficient centres in each group to give precise results, so that it is possible to pick out the best variety, say, in Wales or for heavy land, and so on; even with a hundred centres there would hardly be enough to permit of more than a single classification at once, so that it would not be possible to get an accurate comparison of the varieties, say, on heavy land in Wales. In such work, of course, the investigator must be wary of any inter-relations which may exist between the factors concerned. Any group of centres can be studied separately, but obviously care must be taken in grouping to avoid introducing a bias in favour of any one variety; it would be unreasonable, for instance, to pick out centres in which a particular variety had done well, and to group them together for that reason alone, and any result obtained from such a group would be quite meaningless; but avoiding such pitfalls is a mere matter of common sense.

The above is a special case of observation plots organized so that together they form a definite experiment; it is a method well worthy of extended trial, though it is doubtful if the plots can rightly be termed observation plots. In general, observation plots are no more than their name suggests; they are for observation alone, and are not expected to give comparative figures. It is, indeed, very easy to be misled by observation plots if attempts are made to draw conclusions as to relative yield. Experimenters hear so much, nowadays, of soil heterogeneity, that few will base any conclusions on the yields of single plots, so that, quite rightly, the produce of observation plots is rarely weighed. It is difficult, however, to avoid reviewing them comparatively and, more or less unconsciously, drawing conclusions on eye judgment. Very commonly no obvious

difference is seen within a series of observation plots, and it is concluded that no appreciable yield differences exist. Readers are particularly warned, however, of the fallibility of eye judgment; yield differences of 20 per cent. and more are often missed by experienced practical men. Observation plots must not be regarded as forming even a rough experiment; for comparison of yield they are useless in all cases except where phenomenal differences are involved.

IV.—Size, Shape and Arrangement of Plots.

The statistical technique described in Part I. is equally applicable to large and small plots, so that size of plot is essentially a question of convenience. An important, but by no means the only, consideration must be soil heterogeneity, for precision will largely depend on equalizing fertility between the plots of a block. Uniformity trials in various countries have shown that soil heterogeneity is very complex, but generally two main types can be discerned. There are general trends in fertility over fairly big areas, such as from one end of a ten-acre field to the other, and there are small, irregular patches of high and low fertility; these patches may be very diminutive, such as may be caused by the droppings of animals on a grass field, or they may be depressions or other irregularities of varying extent. A small plot may fall almost entirely on a patch of high or low fertility, so that a series of small plots is liable to show very great variation; as the size of the plot is increased any one patch will have less and less effect on the plot yield, and hence the variation between plots will be reduced. The classical uniformity trials with wheat and mangolds of Mercer and Hall indicated that the variation was considerably reduced until the size reached about $\frac{1}{40}$ acre, but that with larger plots than $\frac{1}{40}$ acre little further decrease occurred. It was therefore concluded—and the conclusion has been supported by other uniformity trials—that the optimum size of plot was somewhere around $\frac{1}{40}$ acre—that is, for a square plot, 11 yards by 11 yards. In general, then, it is a good plan to arrange for plots of this size, since smaller ones will usually give larger experimental errors, but it is often difficult to handle the produce of $\frac{1}{40}$ acre plots with the requisite accuracy; errors introduced by lack of precision in working may easily outweigh the advantages of increased plot size. It might appear that the difficulty of handling large plots could be overcome by a sampling procedure at harvest, but, again, sampling introduces a large error, which easily nullifies any gain from having large plots. Whilst $\frac{1}{40}$ acre is a useful ideal to bear in mind, it is better to have smaller plots with adequate replication rather than cling to the ideal

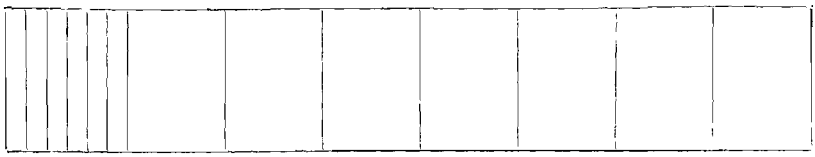
and sacrifice replication. With tree crops, however, there is a fairly high lower limit, for a plot should be large enough to include a number of separate plants, because otherwise genetic variation between individuals will act in the same way as soil patchiness in raising experimental error; in such cases the plots may have to be considerably larger than $\frac{1}{40}$ acre. There is, then, no answer to the question—what is the best size of plot? Cereal plant breeders have used plots of 1 square yard with good results; grass plots of 4 square yards have given reasonable precision; wheat plots of 12 square yards, in the field, but treated with great care, have shown a low experimental error; with sugar cane, however, evidence has been produced that plots should be as large as $\frac{1}{5}$ acre. It may be quite true that, other things being equal, $\frac{1}{40}$ acre is the best size; but the other things rarely are equal, and convenience often demands a smaller, and sometimes a larger, plot.

As regards shape of plot, the choice lies between square and oblong. The square plot has one great advantage, in that for any given size it has the least perimeter; in some enquiries edge effects between different treatments may be serious (in cereals, for instance, a plot may be severely affected by the lodging of its neighbours), and then a square plot is desirable. On the other hand, oblong plots are often much more convenient, and in many cases narrow discards are sufficient to avoid edge effects. Long narrow plots have the advantage that by walking on the discards any part of the plot becomes readily accessible, and this may be a determining consideration where it is hoped to take detailed observations during growth. In cultivation experiments it is almost imperative to have long narrow plots, and a handy arrangement is to have the width a definite multiple of the width of the broadest implement to be used. The same applies to variety trials, and to all experiments where the plots have to be separately drilled; in fact, experience with cereals has shown that there is always much to be said for a plot one drill width across, the required size determining the length. Such plots are not only much easier to manage at all stages of growth than square plots, but are also more convenient at harvest, whether they are cut by hand or by machine. Long narrow plots usually give lower experimental errors than square plots. In taking random sampling units from pasture for botanical analysis it has been shown that units 24 by 6 inches vary less between themselves than units 12 by 12 inches; the explanation is that, in the latter, one plant of a tufty species like cocksfoot may occupy the whole unit, whereas the oblong unit will cut across the plant and will be less dominated by it. The same thing applies to plots; long thin plots will be less liable to

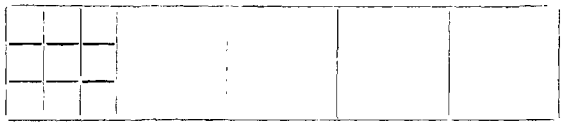
domination by soil patches, and consequently will vary less, than square plots. Sometimes a marked fertility trend may be expected over an experimental site, as when it is on a slope; in such cases the experimental error will be reduced by having long narrow plots with their long sides running up the slope, so that each plot will include some of the soil at the top, and some at the bottom, of the slope. This arrangement is more efficient than running the plots across the slope and leaving the blocks to take out the variability down the slope, because there will then be appreciable differences between the fertility of the upper and the lower plots in each block.

With simple lay-outs the choice of design lies between the Latin Square and Randomized Blocks. It must be realized that a Latin Square does not necessitate square plots; the plots themselves may be oblong and yet be arranged in rows and columns. The Latin Square is generally regarded as the better design, because it allows for fertility trends in two directions, and consequently should give the lower experimental error. It is also a convenient form for a manurial experiment, where the manures are to be applied by hand, and for small dibbled plots in plant breeding work. These are the only advantages it possesses, and in point of fact the first is not always true; few figures comparing the precision of the two designs have been published, but what results are available indicate that the advantage is not invariably with the Latin Square. Its disadvantages are considerable. It is only available for a restricted number of treatments. With 3 treatments it only provides 2 degrees of freedom for error, and so is incapable of detecting any but very large differences, whilst no great precision can be obtained with 4 treatments; with more than 7 or 8 treatments it covers too much ground to be very efficient, and provides for more replication than is usually required. Another grave disadvantage lies in the inaccessibility of the plots in the middle. In some experiments this may be overcome by narrow paths between either rows or columns, but in cultivation experiments large discards must be allowed in one or both directions for turning the implements; thus the square is spaced out, and if rows are separated by wide discards columns will often cease to be effective, and hence the precision of the Latin Square will be diminished.

As regards convenience of working, Randomized Blocks are usually preferable. For cultivation experiments the following arrangement, drawn for 8 blocks and 6 treatments (the actual plots being shown for one block), is admirable:



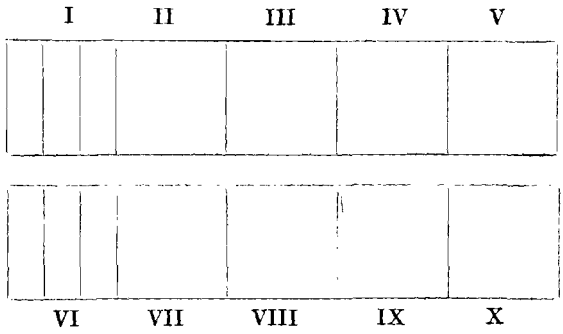
This arrangement has the advantage that each plot stretches right across the experimental area, and the rest of the field is available for turning the implements; in fact, it is often best to avoid all short work and to carry the cultivations right through the field (only a transverse strip being taken as the actual experiment, since experimental handling of the whole field is impossible) so that intercultivation operations may be normally executed. What is said of cultivations applies equally to drilling, so that this design is very commonly adopted for field scale variety trials. This same design has the further advantage that each plot is readily accessible, both of its ends being on the outside of the experimental area; thus the plots are good for demonstration and handy for detailed observations and for harvesting. There are, of course, many other possible arrangements of plots in randomized blocks. The plots given by the above design may be regarded as too narrow, and it may be better to arrange for the plots within each block to run in the other direction, though this will mean loss of accessibility. It may be considered that where there are, say, 9 treatments they will stretch too far if placed side by side, giving the blocks poor control, in which case they may be arranged in 3 tiers of 3 plots each—giving the form:



and it is clear that there are many other possible variants.

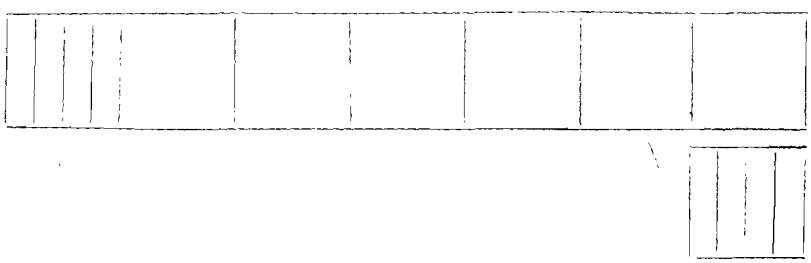
This is an example of the outstanding characteristic of the Randomized Blocks design—its flexibility. There is no absolute limit to the number of treatments that can be accommodated; from 2 to 10 or more can be included in a block, though with a large number some modification of design (see below) may be desirable. The number of blocks required depends to some extent on the number of treatments included (where the treatments are few more blocks are required to provide enough degrees of freedom for the estimate of error), but here again there is great flexibility; with 2 treatments at least 10 blocks are advisable, whereas with 4 treatments 7 or 8

should suffice. Finally, the flexibility of the Randomized Block design is of great service in fitting an experiment to an awkward shaped field. The long line of blocks illustrated above is very common, but, if necessary, they may be placed in two lines, like this:



(N.B.—There must be separate randomization for each block, so that the same treatments *must not* run straight through Blocks I. and VI., II. and VII., etc., except where it occurs through the luck of the draw.)

Another possibility is to mass the blocks together to make a sort of square, or, as long as trees and headlands do not encroach, one or more blocks may be placed in an odd corner of the field, giving the experiment a form of this nature:



In any case discards (see page 96) between blocks or between plots in a block may be desirable, and these must be considered in deciding on what design to adopt.

Where it is required to include a large number of treatments in one experiment, it is usually best to use some more complicated design than those already discussed. A common device is to split the plots of a Latin Square or Randomized Blocks. A large number of treatments can usually be divided into two series; for instance, there may be 4 cultural and 3 manurial treatments to be compared, it being

required to test all combinations, so that there will be 12 different treatments in all. The considerations outlined above may lead to the conclusion that the 4 cultural treatments can best be compared in 6 randomized blocks, giving 24 plots. Each of these plots will then be divided into 3 sub-plots, and over each trio the 3 manurial treatments will be separately randomized; thus the experiment will consist of 72 sub-plots in all. Such an experiment does not prove troublesome, nor is the statistical work difficult or laborious. It should be noted that in this design the cultural treatments are applied to large plots; this may prove a great advantage, for the implements may not be suitable for small areas. The design would be expected to give a more precise comparison of the manures than of the cultivations, because it provides 24 replications of the former, but only 6 distinct replications of the latter.

Many workers fear and distrust experiments of complicated design, but there is much to be said for them. They provide a wider basis for drawing conclusions than do simpler experiments, the manures in the above example, for instance, being compared on the 4 different cultural treatments; thus not only are the average effects of the manures compared, but differential responses on the various cultural treatments will be brought to light in the interaction term of the analysis of variance. Interactions may be supremely important, as when manurial treatments are superimposed on a variety trial. Furthermore, it is solace to a hard-working experimenter to think that every plot he harvests contributes to two comparisons.

In recent years much has been heard of yet more complicated designs, appropriately designated " confounded " experiments. In these each block (or row of a Latin Square) does not include all the treatments, so that the statistical reduction of the results becomes more involved. The lay-out described above is, in fact, a simple case of confounding, for it may be regarded as consisting of 24 blocks of 3 plots each for the manurial comparison, the cultural treatments being confounded with these 24 blocks. A general discussion of confounded experiments is, however, beyond the scope of the present book.

The extreme flexibility of modern methods of field experimentation makes the mastering of them somewhat difficult, because it is not merely a matter of learning a few formulæ; but this flexibility is necessary in view of the very diverse crops and conditions to which they have to be applied. The three great principles underlying [illegible] been mentioned in Part I.—firstly replication, so that the

of the different treatments may be intermingled and in order to obtain an estimate of experimental error, secondly randomization, in order that the estimate of experimental error may be valid, and thirdly local control (as by blocks or by rows and columns in a Latin Square), in order to reduce experimental error. It is sometimes said that the third of these principles is the negation of the second, because when some form of local control is introduced the arrangement ceases to be wholly a random one. It is true that the Latin Square (or Randomized Blocks) design precludes a large number of arrangements which might arise in a draw, but there are still a great number of possibilities remaining, and the random element is introduced when one of these is arrived at by chance. It is interesting to note that the principles were being applied in the earliest centuries of the Christian era. When parcelling out the land in the old open field system, great pains were taken to ensure that each farmer was treated fairly as regards the fertility of the soil allotted to him. The method was that each man had, not one, but several strips, that lots were drawn for the strips, and that each man had to have some strips on the poor and some on the rich soil areas. Thus primitive man applied the principles in order to secure fairness between farmer and farmer; in a precise experiment it is not unreasonable to take the same care to ensure that treatments are compared with equal fairness.

Having designed an experiment, it is very desirable to write down the skeleton form of the table of the analysis of variance, and to make sure that the appropriate statistical operations are thoroughly understood. In particular, note must be taken of the number of degrees of freedom available for error. If they are few, significance will be attained only if the treatments produce wide differences, because in entering the z-table n_2 will be small, and hence a high value of z will be required. In general there should be at least 10 degrees of freedom for error, and if the skeleton table shows less than this number the question of further replication must be seriously considered. The problem of the number of replications required may also be approached on the basis of the standard error of the mean. Past experience may indicate that a standard error of one plot (*i.e.*, the root mean square in the error line of the analysis of variance) of about 10 per cent. of the mean may be expected. If this level of precision is maintained the standard error of the difference between two means of n replicates will be $10\sqrt{(2/n)}$ per cent. of the mean; a significant difference will be this figure multiplied by t, which will vary with the number of degrees of freedom for error, but which will usually be a little over 2 (for $P=\cdot05$). Suppose that the experiment

has been planned to give 8 replications, and that the skeleton table shows 14 degrees of freedom for error. The expected standard error of the difference between two means will be $10\sqrt{(2/8)}$, or 5 per cent. of the mean, and, since t with 14 degrees of freedom for error is 2·145, the lowest difference that can be expected to be significant will be $5 \times 2{\cdot}145$, or 10·725 per cent. of the mean. If it is hoped to detect smaller differences than this more replications will be necessary, unless greater accuracy in working can be achieved.

It must be stated most explicitly that it is imperative to have separate randomization for every experiment. Some workers have used the same design with the same randomization for several experiments, but the procedure is indefensible, and invalidates the statistical tests used in working out the results because of the possibility of a slight bias arising in any particular randomization. A more common mistake is to divide long plots transversely (*i.e.*, with the treatments the same in each half), and to imagine that the replication is thereby doubled. In some experiments it is convenient to carry the treatments, randomized in blocks, right through a field, and to take a strip across the middle as the experimental area. This is a perfectly proper procedure, but there is sometimes a temptation to take two strips across and regard the replication as being doubled, and this must be sternly resisted, because the placings of the treatments will be the same in each strip. It has been pointed out in Part I. that the principle of the Randomized Block design is that the treatments are distributed entirely at random in each block, irrespective of their placing in any other block, and this principle would clearly not be followed in the case of the two strips. The point will be easily appreciated if the idea is carried to its logical conclusion, which is, assuming two treatments, to apply each treatment to half the field, and harvest the lot in, say, 10 strips running across the halves. It is possible to go through the motions of statistical analysis treating the strips as replications, but the result would be utterly valueless, for, since the 10 strips together constitute the field, it is obvious that they can give no greater precision than the two halves of the field. The point is that in each strip the treatments are in the same relative position, so that differences between treatments may be partially due to fertility trends in the soil. Similarly single plots of several treatments cannot be converted into a replicated experiment by taking a number of random samples from each plot. With some crops a plurality of cuts or pickings are taken in one season, and here again some workers have fallen into error by regarding the separate cuts or pickings as replicates; for the comparison of the treatments

the yield at each harvesting may be worked out separately, or the total yields of the several plots over the season may be used, but if there are 5 plots of each treatment and every plot is harvested 4 times, the treatments are not replicated 20 times. The purist might frown on the working out of the figures for each cut separately, since the randomization of the treatments remains the same at all harvests; the answer to this is that it is not proposed to attribute greater certainty to a result because it emerges at all harvests than would be attributed to it if it only emerged once—it is essentially one experiment throughout. Once the point that every experiment must have its own unique randomization is appreciated the statements in this paragraph become self-evident, but reasonable men have fallen into all the errors here mentioned.

It will be realized that designing an experiment is an interesting task, involving much thought. Even the decision as to exactly what treatments to include may be difficult, and often some important ones have to be rejected. Where two series of treatments are included it is best, however, to have them in all combinations. To take a simple example, in an experiment on nitrogenous and phosphatic manuring the 4 treatments—(*a*) no manure, (*b*) nitrogen, (*c*) phosphate, and (*d*) nitrogen and phosphate—should all be included, as there will then be 1 degree of freedom for the nitrogen effect (*b* and *d* against *a* and *c*), one for the phosphate effect (*c* and *d* against *a* and *b*), and one for the interaction (*a* and *d* against *b* and *c*). Thus all the plots of the experiment will contribute to the answers to the following questions:

(1) Does nitrogen have any effect ?
(2) Does phosphate have any effect ?
(3) Does nitrogen have the same effect in the presence as in the absence of phosphate ?

It may be that the application of nitrogen alone would rarely occur in practice, but that treatment is worth including for the sake of completeness; if it is left out the symmetry of the comparisons will be lost, and the results will not be expressible in the same clear-cut terms, neither will it be possible to test the interaction.

There remains only the draw, in which the rules of the game must be scrupulously observed. Tippett's tables* of random numbers are very useful at this stage. Suppose that there are 7 treatments in a Randomized Block experiment, the first step is to number them 1 to in any way, systematic or otherwise. The tables are then opened

* "Tracts for Computers, XV.—Random Sampling Numbers," by L. H. C. Tippett, 1927 (Cambridge University Press).

and two adjacent columns are selected at random; the method of drawing the numbers 1 to 7 is to work down the page dividing the figures by 7 and writing down the remainders. The figures shown might be as follows:

65	giving	a	remainder	of 2
73	,,	,,	,,	3
12	,,	,,	,,	5
35	,,	,,	,,	7
52	,,	,,	,,	3
23	,,	,,	,,	2
92	,,	,,	,,	1
46	,,	,,	,,	4

Thus the arrangement in the first block becomes 2357146. It will be noted that the remainders from 52 and 23 are neglected, because 3 and 2 have already occurred, and that treatment 6 must fall into the last place, after the others have been drawn, to complete the block. The other blocks are randomized by continuing on down the columns in the tables. Care must be taken to avoid any bias. Thus with 7 treatments the figures 99 and 00 must be skipped if they are encountered, because 98 is the highest multiple of 7 obtainable, and to allow the next two to occur would be to give treatments 1 and 2 a slightly higher chance than the others of appearing early; with 6 treatments 97, 98, 99 and 00 would have to be skipped, and so on. If Tippett's tables are not available some other method of drawing must be adopted, and coins, cards, dice, and even roulette wheels have all been used; whatever method is employed great care must be taken to avoid bias, cheap dice, for instance, being very unsatisfactory.

In the case of the Latin Square it is best to draw for one of the types given by Yates, and proceed in the manner described in his paper.* If the paper is not available it can still be done, though less satisfactorily, by drawing random numbers from Tippett's tables or otherwise. Let us take the case of the 4×4 Latin Square, and designate the treatments A, B, C and D. The first row may be drawn as CABD. If the second row emerges as BADC it will have to be rejected and another drawn, because treatment A would occur twice in the second column. The third row will often have to be rejected several times before a permissible arrangement is obtained, but the last row fills itself in automatically. With larger squares the penultimate row (and even those earlier than that) is often very

* " The Formation of Latin Squares for Use in Field Experiments," F. Yates, *Emp. J. Exp. Agric.*, 1933, i., 235.

difficult to obtain, and, though at each failure a fresh series for the row should be drawn, it is excusable to leave in those that will fit and redraw the others; herein lies the weakness of this method, for it is difficult to avoid any "adjustment." To return to the 4×4 square, the arrangement obtained might be:

CABD
BCDA
ADCB
DBAC

The next step is to randomize the rows and columns, for which more numbers must be drawn. Supposing the numbers 3421 are drawn for the rows, then the third row is moved to the first place, the fourth to the second, and so on, and the square becomes:

ADCB
DBAC
BCDA
CABD

Then supposing that 2314 are drawn for the columns, the second column will be written first and so on, and the final form is:

DCAB
BADC
CDBA
ABCD

It is possible to carry the randomization one stage further, and, having obtained this square, to draw the 4 treatments against the letters A, B, C and D; but this is a work of supererogation, and the usual procedure is to allot letters to the treatments before starting the randomization.

A final word might be said in connection with the design of experiments. Modern methods are very flexible, but cases do occur in places with limited facilities where practical considerations conflict seriously with statistical requirements. The two interests can usually be reconciled with a little ingenuity, but where they cannot it is well to remember that statistics is the servant, not the master, of experimentation.

V.—Observations on Plots—Sampling.

In earlier pages it has been insisted that a precise experiment is designed to answer specific questions, and that a numerical result is sought. The reader must not imagine, however, that it is con-

cerned solely with the final yield, and that after an experiment has been initiated it may be left alone and scarcely visited until ready for harvesting. Such negligence may be necessary where the worker is a busy man with several experiments at various places, but it is far from ideal, because features ancillary to yield will not be observed, and no insight into the action of the treatments will be obtained. Weekly visits throughout the period of growth are by no means excessive, and the plots should be carefully inspected every time, so that any differences in appearance between the plots of the various treatments may be noted as soon as they develop.

It is sometimes helpful to make numerical estimates of such features as stage of growth, and, if each plot is separately estimated, there is something to be said for working out the results statistically. Thus it may be suspected that the plots of one treatment carry particularly vigorous plants; the vigour of growth on each plot may be assessed on a scale of points from, say, 1 (very weak) to 10 (very strong) and a table of the analysis of variance drawn up on the figures. Where it is possible it will generally be found easier to make direct estimates, such as percentage lodging, or percentage of plants attacked by disease, than to work to an arbitrary scale. Estimates of the percentage of clover in pasture plots have been employed with marked success, and have proved surprisingly trustworthy when checked by botanical analyses of the herbage cut. It is hardly necessary to point out that this sort of work must be carried out without reference to a plan of the arrangement of the treatments, so that, when estimating a particular plot, the treatment to which it was subjected is not known. Simple honesty is not enough, and great care must be taken to avoid any unconscious bias; it is a great advantage if it can be arranged for several observers, viewing the plots from different angles, to make independent estimates, and after the whole has been covered the first few plots should be estimated afresh, to ensure that the standard of judgment has not changed during the process. The writers are not prepared to maintain the strict validity of applying statistical tests to figures obtained in this way, but the method does serve to give a certain degree of precision, and to provide a permanent record of visible differences which cannot be measured.

With most crops there are some developmental observations to which definite figures can be assigned, and counts or measurements made during growth may be extremely valuable in explaining subsequent differences in yield; if it is possible to obtain, by means of early observations, an idea of how yield is synthesized, and of where

divergences arise between differently treated plots, the whole problem will be approached more philosophically, and elaborations of the treatments will often suggest themselves. The type of observation to make will obviously depend on the crop. With wheat, the number of plants present and the number of tillers produced are two developmental observations which are commonly made. One count of each is often inadequate, as a treatment may hasten germination or lead to a better plant survival during the winter, or, again, it may lead to increase in early or in late tiller formation; in order to provide the full story two or three counts of each may be required. Counts, essentially developmental in nature, may be taken at harvest time; with beans, the number of pods, and even the number of beans, may serve to give valuable information as to the action of the treatments. Developmental studies have played an important part in recent enquiries as to the effect of applying nitrogenous manure to wheat. It has been shown that early applications increase the number of tillers, and that tillers appearing before the end of March have a good chance of carrying ears at harvest; nitrogen applied late in growth has no effect on the number of ears, but increases the weight of grain per ear. Thus there has emerged a fuller understanding of the subject, which has proved much more valuable in practice than any bald average result on the increase to be expected from applications at various dates. If the practices of husbandry, whether in cultivations or manuring or pasture management, are to be raised from an empirical to a scientific basis, a " philosophy " of each, based on developmental studies to interpret final effects, must be built up.

Where a series of counts is made during the season it is helpful not only to watch the appearance of differences between treatments, but also to test for the significance of changes which occur from count to count. This can only be done by the covariance method of statistical analysis. To revert to wheat, it has been shown repeatedly that tiller formation is highly dependent on spacing, the plants on a sparsely populated area tillering much more freely than those on a crowded one. If one treatment produces a better plant establishment than another, it will be expected that the plants on the latter plots will tiller more profusely, because of their relative gappiness. If a plant and a tiller count are taken, the natural thing to do to compare the rate of tillering is to divide the number of tillers by the number of plants, to find the average number of tillers per plant; in the present case the second treatment will give a higher quotient and might be judged to have affected tillering beneficially, but that conclusion would be quite fallacious, because the real difference lies

in the depression of plant number. It is very easy to fall into errors of this nature, and, indeed, until very recently there was no method of getting a fair comparison of the rate of tiller formation in this case. The beauty of the covariance method lies in the fact that it allows for inequalities at the earlier count, and gives an estimate (complete with test of significance of difference) of what the number of tillers would have been had the number of plants been the same for each treatment. This, of course, is just what is wanted, and provides a real means of judging the effects of the treatments stage by stage throughout development. In an experiment with beans,* significant differences were found in early germination, in late germination, in the number of early produced stems, in the number of stems at harvest, in number of pods, in number of beans and in weight of beans; in all cases the significance could be judged after due allowance for earlier inequalities. Such a full analysis brings to light the complexity of yield, and shows how little of the whole story the end result, by itself, reveals. Experimenters will find it well worth while to master the covariance method of statistical analysis. It has been stated earlier that there are probably cases with perennial crops where it may be profitably employed in conjunction with preliminary uniformity trials, but it is in analyzing effects stage by stage that the method has most to offer. Nearly every experiment yields more than one series of observations (with cereals weight of grain and weight of straw provide two, apart from any developmental counts), and the worker cannot be sure that he has extracted the fullest possible information from his experiment until he has tried the method. At Cambridge it is now almost a matter of routine after any analysis of variance (except, of course, the first) has been worked out to carry through the covariance analysis on the immediately preceding series of observations from the experiment, at least to the point of calculating r and b; if these are insignificant it may be clearly futile to proceed further.

It is utterly impossible to count all the plants or tillers on a normal field experiment. It follows that some system of sampling must be adopted. In order to represent the plot there must be a number of sampling units scattered over it, and scattered in a random

* " The Value of Covariance in Analysing Field Experimental Data," F. H. Garner, J. Grantham and H. G. Sanders, *J. Agric. Sci.*, 1934, xxiv., 250. Attention is directed to the second footnote on page 254 of that paper. Since the paper was published the correct test for significance of differences in $(y - bx)$ has been evolved (see pp. 45-56).

fashion. A great deal of work has been carried out to determine the best sampling unit to take, but much still remains to be done. The unit must not be too small, or a lot of zeros will occur and the distribution of the figures will be far from normal; furthermore, with very small units slight inexactitudes of measurement may introduce considerable error. On the other hand, a number of small units is preferable to a few large ones, as the former will give a better representation of the plot. Clearly the optimum size of unit must be determined for each type of crop separately. Wheat is the crop that has been most studied in this connection, and units from a foot to a divided metre (*i.e.*, two half-metres with a gap of half a metre between them) of drill row have been widely used. With beans 1 yard of drill row has proved satisfactory. With pasture plots an area of 1 square foot (preferably 2 feet by 6 inches, see page 76) is very commonly used for botanical analyses, but there is generally very great variation between units, and it is rarely possible to cope with enough to get a good representation of the plot. With tree crops the problem is not easy, as the unit will usually be part of a tree, and to select part of a tree in a random manner is difficult, though not impossible.*

Scattering sampling units over a plot may be done in a variety of ways. Suppose that each plot of a cereal experiment consists of a 20-yard length of 10 drill rows. If 30 foot-lengths are to be taken, they may be placed entirely at random, in which case, for each sample, a draw will be made to decide the row and another draw for the distance down the row; if the same position is arrived at twice it will generally be rejected the second time and another drawn, though there is no very sound reason against its being counted and the figure entered twice. It is only reasonable to lay down certain restrictions to ensure that the units shall be adequately dispersed over the plot, and one method is to divide the plot into sections, and make the rule that a definite number of units must fall in each section. In the present case a sensible division would be into drill rows, and this would have the advantage that an equal number of units would come from each coulter of the drill; thus 3 draws would be made for each row, to determine the distance down the row of 3 random sampling units. Alternatively the plot might be divided in any way into, say, 6 sections and 5 units taken at random positions in each section. For sampling sugar beet plots at lifting time a suitable unit is one beet, and a pattern method of sampling has been used.

* See " Precision Records in Horticulture," J. O. Irwin, *J. Pomology and Hort. Sci.*, 1931, ix., 149.

Suppose that there are 200 beet in a plot, and that it is proposed to take 2 sampling units of 10 beet each. Two numbers between 1 and 20 (inclusive) are drawn—say 4 and 16. The plot is then covered by walking along one row, back on the next, and so on, and the 4th, 24th, 44th . . . beet are pulled for one sampling unit, and the 16th, 36th, 56th . . . beet for the other sampling unit, totals only being recorded in each case. It is claimed that this pattern method of sampling facilitates the work, and it undoubtedly saves much drawing of numbers. Randomization for position, however, is a very minor task if Tippett's tables are used, the numbers being entered into field notebooks as they are drawn.

It is possible to divide the plot into many sections, and to take only one sampling unit from each section, but it is better to arrange for at least two sampling units per section, in order that an estimate may be obtained of the sampling error—that is, of the ultimate error within a section. It must be confessed that the calculation of the sampling error does not advance matters much from the practical point of view, but the information may prove valuable in future work, as showing the number of units required to represent a plot with a given degree of precision; in any case it involves no extra work to arrange that more than one sampling unit occurs in each section.

An example of the calculation of a sampling error may prove helpful, and for this we shall take a 7 by 7 Latin Square concerned with the nitrogenous manuring of wheat. The plots were small, consisting of 8 drill rows for a length of 15 feet. The method of sampling was to take two random foot lengths from each drill row, and the yields from the 16 units for each plot were added together to give the plot yields, which were analyzed in the usual way to test the significance of the treatment differences. The total "sum of squares" (*i.e.*, $S(x-\bar{x})^2$) for the 48 degrees of freedom between plots was 233,067. But each plot figure was the total of 16 foot-lengths, and for the variability within the plots we shall have to work on the basis of 1 foot-length. The first step, therefore, will be to divide 233,067 by 16, to give 14,567 as the "sum of squares" between plots. There being 8 rows per plot and 49 plots, there will be 8 by 49, or 392 row totals of 2 foot-lengths each. Squaring each of these 392 and adding the squares together, the figure reached is 1,201,261. The grand total is 20,652·6, so that the "sum of squares" between rows becomes $1{,}201{,}261/2-(20{,}652{\cdot}6)^2/784$, which works out to 56,588. This, of course, is associated with 391 degrees of freedom. By squaring the figures from the 784 separate foot-lengths and adding,

the sum of 653,090 is reached, whence the " sum of squares " between feet is calculated as $653,090 - (20,652{\cdot}6)^2/784$, or 109,047. The following little table can now be constructed:

	Degrees of Freedom.	*Sum of Squares.*	*Mean Square.*	*Root Mean Square.*	$\frac{1}{2}$ *Log*$_e$ (*Mean Square*).	z.
1. Between plots ..	48	14,567	—	—	—	
2. Within plots and between rows	343	42,021	122·51	11·068	2·4038	0·0445
3. Between rows ..	391	56,588	—	—	—	
4. Within rows and between feet	392	52,459	133·82	11·568	2·4483	
5. Between feet ..	783	109,047				

The degrees of freedom and " sums of squares " in the second and fourth lines are obtainable by subtraction. The figure 343 for the degrees of freedom in the second line and its corresponding sum of squares can be calculated as a check, because each plot contains 8 rows, giving 7 degrees of freedom, and $49 \times 7 = 343$; similarly for the fourth line there are 392 rows, each with 2 foot-lengths—that is, each with 1 degree of freedom. The second and fourth lines are the interesting ones, for the former gives the variation between rows, when allowance is made for differences between plots, and the latter the variation between random foot-lengths in the same row—that is, the sampling error. In this case we see that the sampling error of 1 foot-length is 11·568, which amounts to 43·91 per cent. of the general mean yield per foot-length (26·343 gm.). The sampling error per plot is therefore $43{\cdot}91/\sqrt{16}$ or 10·98 per cent. of the mean. Some interest lies in the comparison of the variation between and within rows, for it provides a test of whether the coulters of the drill were all sowing alike; appreciable differences between coulters would be shown by greater variation between than within rows. The table shows practically the same variation in each case, so that, as far as can be told from yields, all the coulters were delivering seed at the same rate and depth. The value of z for the comparison is 0·0445; this cannot be looked up in the z-table, for values of n_1 and n_2 of 392 and 343 are far beyond the limits of the table. In such cases a standard error for z may be calculated; it is given by $\sqrt{\frac{1}{2}(1/n_1 + 1/n_2)}$, which, with the present figures, reduces to 0·0523. Hence z is less than its standard error and so quite insignificant, indicating that the variation between rows did not differ significantly from the variation within them.

Calculations of variation within plots are interesting, but very laborious, because so many figures have to be squared; fortunately the interest is more academic than practical, so that the worker who

wants experimental results rather than information on experimental design need not trouble himself with them. All he will need will be the plot totals from the samples, which totals can be used for statistical analysis in the same way as plot yields. The important practical point is that sufficient units should be taken to give a reasonably accurate representation of the plots; to achieve this the samples should amount to 5 per cent. of the plot at the very least. In the above example 16 units were taken from small plots, each of which only comprised 120 foot-lengths of drill row; thus the samples constituted over 13 per cent. of the plots, but nevertheless they were not very closely representative. At the same time that the samples were cut the remainders of the plots were harvested, and the correlation coefficient between the two sets of 49 figures was only +0·557. It is true that the relation was close enough for the samples to place the treatments in much the same order (though with a larger experimental error) as obtained from the full plots, but irregularities between plots (after allowance for treatment, row and column) were not paralleled by the samples. It is feared that this will always be so—feared because it means that a low correlation coefficient will be given by the error line of the analysis, and hence that it will be impossible to get an accurate correction of plot yields for inequalities during growth which can only be determined on samples.

The low correlation between samples and the plots from which they are drawn is a serious matter in connection with harvesting by sample. With cereal experiments on ordinary farms it is usually impossible to thresh the produce from each plot separately, and attempts have been made to overcome the difficulty by cutting only samples which can be transported to a research station for threshing; in this it is obvious that only a small proportion of the plots can be taken in the samples, and this often precludes the detection of any but large differences. Harvest samples have their uses, however, where it is intended to cut and weigh the whole plots, because the sample can be examined in greater detail than the bulk; thus with wheat number of ears can be counted on the samples, and the results may prove valuable in explaining yield differences which are determined more accurately from the whole plot yields.

Where several numerical observations are obtained by sampling during the growth of a crop the question arises as to whether the samples should be the same, or whether a fresh randomization should be made, at each count. The usual procedure is to keep to the same positions, and experience has shown that if it is proposed to use covariance in the statistical analysis the units must be precisely

the same, for a shift of a few inches along the row greatly decreases the correlations between the figures at different counts, thereby diminishing the accuracy of correction. Units may be marked out with small lath pegs at the first count without material damage to the plants. Even then the correlations are not usually very high, except between counts near together in time, so that, for instance, number of tillers in March can, but yield at harvest cannot, be accurately corrected for number of plants in January.

It is very difficult to avoid damaging plots when making counts during growth. Treading may vitiate the results of an experiment, especially if it is concerned with cultivations. Efforts to avoid treading have been made by restricting the sampling to a yard or two at each end of each plot, but experience has shown the futility of this, the ends utterly failing to represent a plot. With cereal experiments the problem has been solved by having plots one drill width broad and by using wooden bridges (about 15 inches wide) to span them; counts are made by assuming a recumbent position upon a bridge. There is another form of damage which is more insidious because less patent to the eye, and that is the injury the plants suffer in being pulled about in the process of counting. There is not much that can be done about this, except to treat them as gently as possible, not to make any more counts than are necessary, and to avoid counts late in the period of growth. In some cases it may be necessary to uproot plants in making developmental studies; with sampling adequate in extent to be representative this will mean that the plots are ruined, and such studies should be made on an experiment which is specially laid down for them and which is abandoned after the upheaval.

So far we have dealt only with the sampling of plots during growth or at harvest; there is another form of sampling which plays an important part in experimentation—namely, the sampling of produce after it has been harvested. An attractive, but wholly unsound, procedure is to weigh the produce of each plot and then to bulk all from each treatment for sampling, and to apply the results from the samples to the plots; for instance, if the crop is grass, and the dry matter percentage of a bulk sample of one treatment is found to be 20 per cent., the green weights of all plots of that treatment will be multiplied by 20/100 to estimate the yields of dry matter, which yields will then be analyzed statistically. Apart from the more theoretical objections to this method, it will be realized that an inaccurate determination of dry matter percentage for one treatment may give rise to a significant difference, and such cases have been known to occur. The produce of each plot must be separately sampled to give a distinct

estimate of the dry matter yield from that plot, and then statistical tests may be applied to the figures. It is preferable, though rarely possible, to take two samples from the produce of each plot, and where this is done the error of sampling may be determined, there being 1 degree of freedom for its estimate from each plot. It is generally found that the error of sampling is disconcertingly high, so that the greatest care must always be taken in mixing the produce and in drawing the sample. A good method is to mix the heap of material thoroughly, to divide it into four quarters, and to reject, say, the N.E. and S.W. quarters, mixing the other two together again; the process is repeated until the bulk is reduced to the size required for a sample.

An interesting case of produce sampling has arisen with cereals. It has been stated above that experiments at outlying centres present difficulties with threshing, and that these are not satisfactorily overcome by sampling the growing plots because that leads to a high experimental error. An alternative is to weigh all the sheaves from each plot (a simple matter with a spring balance), and then to choose at random a few sheaves from each plot for threshing separately on a small scale thresher at a research station; the percentage grain obtained from the sample sheaves is then applied to the total sheaf weight from the plot. In trying a new method it is important to determine what sampling error it involves, and until that is done no definite opinion on the method can be reached.

The application of percentages obtained from samples to the fresh produce of plots may not increase the error of the yield figures. Thus with grass it is commonly found that the experimental error emerging when dry weights are analyzed is rather less than that with green weights; the explanation is that a plot with an abnormally high weight of green produce carries a luscious growth with a low dry matter percentage, so that application of the percentages tends to even the plots out.

There are many diverse conditions in which it is required to sample produce, but the principles are the same in all cases. The important points are to sample each plot separately and to mix the material thoroughly before drawing a sample. The former is necessary if statistical tests are to be applied to the results, but there are some cases where that is not the aim, and then bulk samples (preferably in duplicate or triplicate) from each treatment may suffice. In the case of sugar beet, for example, in some experiments it may be confidently predicted that the treatments will not affect sugar percentage nor dirt tare; bulk samples may serve as a check on this, and to support

the validity of drawing conclusions from the weights of dirty beet harvested, but if the samples do indicate differences between the treatments an impasse will be reached, for there is then nothing that can be done about it.

VI.—Matters of Detail.

From beginning to end an experiment will produce a large number of observations and figures, so that some system must be introduced into the keeping of records. A field notebook containing more or less legible pencilled scribblings is not sufficient, and any notes made on the plots should be copied, as soon as possible, into some more permanent record; a loose-leaved file is very convenient, and it is best to keep a separate file for each experiment—there is usually enough material to fill it. All descriptions in the file should be extremely full. Immediately after making a series of observations details of the methods employed are clearly imprinted on the mind and it may be felt that it is unnecessary to write out a complete account of the procedure; it must be remembered, however, that reference may be made to the file five years later, and it is surprising what can be forgotten in a much shorter time than that. Matters obviously important, such as the condition of the soil when cultivations are carried out, must be dealt with at length, but quite minor points are worthy of a place in the record; thus if different sized or coloured pegs are used in marking out the plots, the system adopted should be described in the file, for it will probably be forgotten before the plots are harvested.

A number of sections will be required in an experimental file. At the front should be one dealing with the objects of the experiment, and giving precise details of the treatments employed, together with a scale plan (oriented in respect to some permanent topographical feature of the field) of the actual lay-out. Copies of the plan on postcards are very handy for field work, and there should always be several available; if these matters of detail are not attended to conscientiously it is possible to commit the supreme absurdity of laying down an experiment and losing the only plan of it. A brief section of the file should be devoted to the previous history of the field, and should give an account of the cropping and manuring, and anything else likely to affect the condition of the soil, during the preceding four or five years. One section should be the experimental diary, and it is important to head all entries in this with the date; a well kept diary, describing clearly all operations, methods and observations, proves of incalculable value when the results are reviewed at the

end of the experiment. At some stations a system involving the use of blank forms for weekly routine observations has been adopted; there is something to be said for this, but its rigidity must not be allowed to stultify the worker's initiative. The other sections of the file should be devoted to the various phases of work on the experiment, such as drill testing to determine seed rate, developmental counts, studies on soil condition, harvest results, and so on.

Edge effects between plots may be serious, and they often present a difficult problem. Cases vary very much, and only experience of the crop concerned can decide whether discards between plots are necessary, and, if so, how large the discards must be. Sometimes the plants on the plots of one treatment overshadow their neighbours, sometimes cultivations cannot be carried accurately to a line, sometimes manures are liable to get worked slightly out of their right plots, or plant roots from adjacent plots may penetrate and benefit from them. All things of this nature must be considered in planning an experiment, for they may necessitate modification of design. It is true that if unsuspected edge effects are seen during growth it is not too late to avoid them, by harvesting only the unaffected parts of the plots—that is, to decide on discards at harvest time—but it must be remembered that only large differences are visible, and that edge effects may be considerable and yet unrecognized. Discards must be big enough to take care of any interference, but it is important not to have them larger than is absolutely necessary, because they space out the plots and consequently tend to increase the experimental error. In the case of variety trials with root crops a single long row of plants is a convenient form of plot, but unfortunately varieties differ greatly in size of top so that discards to avoid shading are indispensable. A discard row each side of each plot row (*i.e.*, a plot to consist of three rows, the middle one to count) will space the plots to a high degree, since only one row in three of the experimental area will be used; the position can be improved by making the plots shorter and broader, but this sacrifices some convenience in working. Where a cereal plot consists of one drill width of, say, 12 crop rows, either 1 or 2 discard rows each side may be allowed, leaving 10 or 8 crop rows as the true plot; if the bridges referred to on p. 93 are to be used, discards of 2 rows each side are the better, as the feet of the bridge, when placed between plots, will affect the outside rows and there should be one normal guard row.

As an example of discards necessitated by cultivations and drill widths a case encountered with gyrotilling and combine drilling will be described. The gyrotiller is a large implement with a rather ill-

defined edge to its operation; the combine drill sows manure down the coulters with the seed, and it was required to compare the yield of oats drilled in this way with the yield from oats with an equal amount of manure broadcast on the surface of the ground, and from oats unmanured. Four breadths of the gyrotiller (which in practice amounted to a total width varying from 36 to 44 feet) formed a main plot, and 6 randomized blocks were allowed for the comparison of gyrotilling and normal cultivations. The drill was only 7½ feet wide, so that the 3 drill breadths necessary for the manurial comparison did not fully occupy the main plots. The arrangement was to allow one drill width discard in the middle of each block (*i.e.*, between the gyrotilled and non-gyrotilled plots) to overcome the ragged edge left by the gyrotiller, and to take up the remainder of the spare ground in large discards between blocks. Since bridges were to be used for counts during development, further discards of 2 rows on each side of each drill plot were necessary; these served also as a safeguard against any manurial "creep." The important point to notice is that the arrangement kept the plots in a block as close together as possible, the large discards occurring between blocks; this spacing of the blocks would not increase the experimental error, whereas large discards within a block would tend in that direction. The difficulty with this experiment was that the gyrotilling had been carried out eighteen months previously, and the combine drill trial had to be superimposed on main plots already existing.

When an experiment is harvested it may be necessary to discard a whole plot. In such a case it is still possible to obtain a result, the method being to calculate a figure for the yield of the missing plot;* the yields of a number of missing plots can be calculated,† though in field work the experimenter should very rarely be forced to such wholesale rejection. For the statistical techniques used in calculating yields for missing plots readers are referred to the papers cited, but a few words may be said on the morality of rejecting plots. After a plot has been harvested and the produce weighed, it is too late to reject it. The figure obtained may be under suspicion, but once a start is made in rejecting actual figures there is no knowing where to stop; specious arguments may sometimes be found for putting aside any plots whose yields are irregular, and a little skill

* "A Method of Estimating the Yield of a Missing Plot in Field Experimental Work," F. E. Allan and J. Wishart, *J. Agric. Sci.*, 1930, xx., 399.

† "The Analysis of Replicated Experiments when the Field Results are Incomplete," F. Yates, *Emp. J. Exp. Agric.*, 1933, i., 129.

at this game will lead to very significant, but quite untrustworthy, results. There is no wish to impugn the reader's honesty, but no man is so virtuous that he can afford to treat temptation with disdain. A plot may be discarded, before harvesting, if there is any definite extraneous reason for doing so, but the mere fact that it looks like giving a poor yield is certainly not a sufficient reason; above all it must not be rejected because it is a poor plot of the experimenter's pet treatment. If a plot is ruined by a marauding animal, or if a control plot has manure applied to it by mistake, then it is right and proper to reject it, but such cases occur very infrequently. In animal experimentation the problem of rejection is a continual source of worry; if an animal dies it must clearly be rejected, but how near death must an animal be to qualify ?

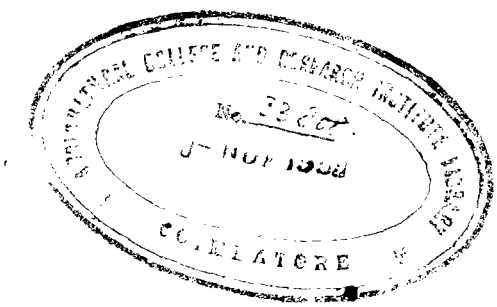
A high standard of accuracy should be set in all experimental work. It is much better to aim for a standard verging on the ridiculous than to be content with one which, it is hoped, will just avoid inaccuracies large enough to have appreciable effect on experimental error; it must be remembered that at one time or another we all fall short of the standard we set ourselves, and it is best to allow a good margin for human frailty. Great care should be taken in marking out plots and in applying artificial manures; as regards the latter, when applying by hand it is desirable to divide the allowance for each plot into several parts, and the plot likewise, to ensure that each part of the plot receives its aliquot portion. Lack of care in harvesting may raise the experimental error materially. With cereal crops experience has shown that cutting by sickle is preferable to cutting by reaper and binder, and, in the end, involves hardly any more work; to avoid loss of corn in carting from the stook to the thresher each sheaf is put, ears foremost, into a sack, and the adoption of this precaution has been attended by a pleasing drop in experimental errors. In weighing it is no extra trouble to read the scale to the last gram, and this will at least ensure accuracy to ten grams. In computations the number of figures retained will depend on whether or not a calculating machine is available; where one is being used it is little extra trouble to work with large numbers, and it is always better to drop figures in the final table, rather than at the beginning of the calculation.

It will be appreciated that inaccuracy at any stage of the work acts in the same way as soil heterogeneity in increasing experimental error. Occasionally ludicrously high errors occur, as when part of the produce of one plot is mixed with that of the next, but it is a fortunate fact that they very rarely lead to false conclusions; their

is done. Some workers in the tropics have to conduct experiments with very unskilled labour, and they will find that trivialities of this nature will merit their most careful consideration. It will be realized that a full programme of experimental work will call for a variety of impedimenta; most of these will be cheap and readily obtainable, but some foresight should be shown in collecting them.

The arrangement of work on an experiment calls for a little thought. Where an operation, such as sowing or harvesting, cannot be completed in one day, work should not be stopped in the middle of a block; if whole blocks are dealt with together all treatments will be influenced alike, and differences caused by delay will be eliminated as part of the differences between blocks. If a number of workers is being employed, the same individuals should work right through a block, so that personal idiosyncrasies, which are by no means negligible even in the simplest operations, may affect all treatments to the same extent.

It is hoped that one thing that will have impressed itself on the reader of these pages is the conviction that an understanding of the principles underlying modern methods is necessary at all stages of an experiment. Statistical knowledge is not only required for working out the results; it is essential in designing, and without it an experiment cannot be properly conducted. The presence of one trained computer at each station does not meet the situation, and it is clear that the experimenter himself must gain a working knowledge of the technique he employs.



Printed in Great Britain by Billing and Sons Ltd., Guildford and Esher