THE PROBABLE ERROR IN FIELD EXPERIMENTS IN AGRICULTURE.

BY

P. C. MAHALANOBIS, B.Sc., M.A. (CANTAB.), I.E.S.,

Professor of Physics, Presidency College, and Meteorologist, Calcutta.

IN a recent number (Vol. XVIII, Pt. 5) of the "Agricultural Journal of India" Mr. B. N. Sarkar has discussed the question of estimation and value of "probable errors in variety trials." As the problem is essentially one of a statistical nature, a discussion from the statistical standpoint may prove useful.

From the statistical point of view the factors which affect the yield of paddy (or other crops) may be analysed into several distinct groups.

Consider any particular variety. Even if external conditions are kept absolutely uniform, slight variations in yield will still occur from plant to plant. Such deviations constitute the organic variability of the plant. They will however cancel out if we take the average of a large number of plants and the "mean yield" may be considered to be a stable constant for the variety. The "mean yield" of different varieties will however be different, and the difference in mean yield of two varieties may be conveniently called the "mean organic difference in yield."

Now consider the external factors. Innumerable small fluctuations will occur (from plot to plot or from experiment to experiment) in puddling or levelling the land, manuring, watering, drainage, exposure to rain and sun, etc. These external fluctuations also will however cancel out on averaging for a large number of experiments and may be grouped under "random fluctuations."

"Systematic variations" in the external factors, such as differences in the composition of the soil, in the nature of manure used, in methods of cultivation or in the fertility of land, will, in general, also occur and cannot always be removed. In addition, large accidental errors such as destruction of crops by crabs, rats, birds or other pests, or mistakes in manuring, harvesting or threshing may affect different experiments or different plots in different degrees. Mr. Sarkar says that such accidental errors can be considerably reduced by careful supervision. I shall assume that they are negligible.[1]

There will also remain the random errors of measurements and certain other purely statistical errors. If the number of experiments, *i.e.*, the size of the sample,

[1] These errors are of such a particular nature that a general discussion is theoretically impossible. My assumption, although never strictly true, is therefore unavoidable.

be very large, we can obtain precise information about the value of the "mean yield" and its probability. But in agricultural work the number of experiments is usually small, very often less than 10 or 12, and, as "Student" pointed out some time ago,[1] the probable error found from such small samples is subject to great uncertainty and "judgments reached in this way may become altogether misleading."

Grouping together the different random fluctuations and neglecting large accidental errors we thus have :—

(A) Organic differences in yield.
(B) Systematic variations of the external factors.
(C) Random fluctuations and statistical errors.

Field experiments in agriculture are of two different kinds :—

(1) "Variety trials" in which different varieties are used and, keeping external factors as uniform as possible, the organic differences in yield (A) are determined. The chief point here is to reduce (B) to zero. But this is often impossible in practice; for example, variations in fertility of the land cannot possibly be removed. It then becomes necessary to estimate the effects of (B) and allow for them as accurately as possible.

(2) In another class of experiments, (A) is reduced to zero, *i.e.*, only one single variety is used and the effect of (B) is sought to be determined with accuracy. Experiments with different kinds of soil, or different kinds of manure or different methods of cultivation are typical illustrations.[2]

The problem of estimating (C) remains the same in either class of experiments. It is entirely a statistical question.

In the case of "variety trials" we have the additional problem of estimating (B). This is partly empirical and partly statistical.

I shall first consider (C).

Random Fluctuations and Statistical Errors.

The "mean difference in yield" of any two varieties may be calculated in two slightly different ways. We may first calculate the mean yield of each variety and then find the difference in the mean or we may first calculate the individual difference in yield of adjoining plots and directly find the mean of these differences. So far

---

[1] "The Probable Error of a Mean." *Biometrika*, VI (i), 1908, 1-25.

[2] The generally accepted experimental procedure is apparently as follows. Long narrow strips of experimental plots are prepared and sown with the different varieties in regular recurring series. In the example given by Mr. Sarkar there were 60 strips, each 80′ long and 4′ wide, sown with 6 different varieties. Each variety thus occurred 10 times altogether. Calling the different varieties a, b, c, d, e and f, we may refer to the yields from the different plots in terms of the varieties. Thus the yield from the 1st, 7th, 13th, 19th, 25th, 31st, 37th, 43rd, 49th and 55th plot may be called a (1), a (2), a (3),.... a (9), a (10), from the 2nd, 8th, 14th ..50th and 56th will be b (1), b (2), b (3), .. b (9), b (10) and so on for the other plots.

as the value of the " mean difference" is concerned, the result will obviously be the same, since

$$\text{Mean of (a)—Mean of (b)=Mean of (a—b).}$$

I quote the figures given by Mr. Sarkar :—

	Kalamdan (a)	Indrasal (b)	Difference (a—b)
	703	670	+33
	705	630	+75
	653	560	+93
	640	615	+25
	700	542	+158
	715	667	+48
	647	702	—55
	848	750	+98
	918	758	+160
	870	830	+40
MEAN	739·9	672·4	+67·5

The " mean difference " of course is the same whether calculated from 739·9 —672·4 or directly from col. 3.

Apparently however there is some confusion of ideas about the probable error of the difference. The probable error of the difference calculated directly is $\pm 14 \cdot 1$, while the probable error calculated with the help of the formula

$$e\ (a—b)=\sqrt{e^2\ (a)+e^2\ (b)}\ \ldots\ldots\ldots\ldots(1),$$

[where e^2 (a) and e^2 (b) are the squares of the probable errors of mean (a) and mean (b) respectively] is $\pm 28 \cdot 5$. Mr. Sarkar evidently prefers the value obtained by the direct difference method, apparently because it gives a *lower value* in this particular example. This however may not always be the case. The direct difference method can easily give *higher values* of the probable error under other conditions.

The real point is that the direct difference method gives the correct value, while the formula (1) is only valid when the two experiments are entirely independent. The complete expression for the probable error of a mean difference (a—b) is

$$e\ (a—b) = \sqrt{e^2\ (a) + e^2\ (b) - 2\ r\ (a,\ b).\ e\ (a).\ e\ (b)}\ \ldots\ldots\ldots\ldots(2)$$

where r (a, b) is the coefficient of correlation between a and b.

Now in variety trials systematic variation in, say, the fertility of the land is bound to introduce correlation between the different variates and hence, *unless (B) is zero, i.e., unless the external conditions are absolutely uniform and there is no correlation between the variates, the abbreviated formula for calculating probable errors of a difference from the probable error of the " means " will give totally misleading results.*

We conclude therefore that *under usual experimental conditions the probable error of a difference should wherever possible be calculated by the direct difference method.*

I now pass on to the question of the restricted size of samples. The subject has been fully discussed by "Student" in a paper already cited and I need merely quote "Student's" results. He has determined the distribution of a quantity Z, which is obtained by dividing the difference between the mean of small sample and the true mean by the standard deviation of the sample and has also constructed a table for estimating the probability of occurrence of Z.[1] Let us consider the example given above. The mean difference is 67·5 and the standard deviation of the 10 differences is 61·26. Dividing 67·5 by 61·26 we get Z=1·10.

From "Student's" Table for n=10, p=·99539 and 1—p=·00461.

The odds are therefore 99539 to 461, or 216 to 1, that Kalamdan gives a greater yield than Indrasal.

Mr. Sarkar finds the value of the probable error of 67·5 to be 14·1. The standard deviation of the mean difference is therefore 20·9 and the mean difference in terms of its standard deviation is 3·238 nearly. From Tables of the Probability Integral, I find that $\frac{1}{2}$ (1+a) is ·999397 and $\frac{1}{2}$ (1—a) is ·000603. The odds are therefore nearly 1700 to 1 in favour of Kalamdan and are much greater than the odds obtained by "Student's" formula.

Of course in the present example it is practically certain that Kalamdan gives a greater yield than Indrasal, and it matters little whether the odds are 1700 to 1 or merely 216 to 1. But the need for caution is obvious and, since "Student" has shown that the probability integral gives too large a value for p when the probability is large, it is extremely important that the correct formula should be used otherwise misleading results may easily be obtained. I conclude that *in estimating the probable error of "mean difference in yield" the table given by "Student" should be used wherever possible.*

I wish to point out that the calculations involved are practically the same as the standard deviation of the differences must be found in either method. "Student's" Table is also easily available. There is no reason therefore why "Student's" method should not be used more extensively.

The "direct difference" method cannot however be always used. In such cases it then becomes necessary either (i) to determine the correlation between the two variates and use formula (2) or (ii) get rid of the correlation by eliminating the variations in the external factors and then use formula (1).

In either case a further statistical correction will be necessary if the size of the sample is small. As I have already pointed out, this question was first investigated by "Student" in the paper cited above. Two further papers, one by Karl Pearson[2] and another by A. W. Young,[3] have completed "Student's" work in this subject.

---

[1] This table has been reproduced as Table XXV, p. 36 of *Tables for Statisticians and Biometricians* (Cambridge University Press).

[2] "On the Distribution of the Standard Deviations of Small Samples." *Biometrika*, Vol. X, p. 522, 1915.

[3] "Standard Deviations of Samples of Two and Three." *Biometrika*, Vol. XI, p. 277, 1916.

Pearson says : " We think it must be concluded that for samples of 50 the usual theory of the probable error of the standard deviation holds satisfactorily, and that to apply it for the case of n=25 would not lead to any error which would be of importance in the majority of statistical problems. On the other hand, if a small sample, n=20 say, of a population be taken, the value of the standard deviation found from it will be usually *less* than the standard deviation of the true population " (p. 528). Tables were constructed by Pearson and Young for making necessary corrections. The correcting factors are given in the following table for easy reference. They are taken from Pearson's and Young's Tables ; but I have put them in a slightly more convenient form for actual use.

TABLE A.

Correcting factors for standard deviations of small samples.

Size of sample	Correcting factor	Size of sample	Correcting factor	Size of sample	Correcting factor
2	Indeterminate	11	1·1056	20	1·0541
3	1·7319	12	1·0955	25	1·0425
4	1·4142	13	1·0871	30	1·0351
5	1·2910	14	1·0801	35	1·0297
6	1·2247	15	1·0742	40	1·0260
7	1·1832	16	1·0691	45	1·0230
8	1·1547	17	1·0646	50	1·0206
9	1·1339	18	1·0607	75	1·0136
10	1·1181	19	1·0572	100	1·0120

To obtain the " corrected " standard deviation we multiply the observed standard deviation of the sample by the appropriate correcting factor taken from the above table. The probable error of the mean will then be obtained from the corrected standard deviation by ordinary methods. The use of the above table will be sufficiently illustrated in later sections.

Construction of the " Normal Fertility Curve."

I shall now consider the problem of estimating systematic variations in the external factors. In certain experiments such systematic variations of external factors are known to exist and cannot be removed. For example, in the illustration given by Mr. Sarkar there is apparently a variation of 40 per cent. in the fertility of the land from one end of the field to the other. It is obviously necessary to make allowances for such variation. Mr. Sarkar has sought to eliminate the effects of such variations by two slightly different methods. In the first he uses one variety as a standard and with its help constructs a "normal fertility curve" for the whole field, while in the second he uses all the different varieties for the same purpose. He then uses the "normal fertility curve" as a standard and considers the difference in yield, *i.e.*, "the departures" from this normal.

The problem will be recognized by statisticians as one of "smoothing." What we want is the "smoothed normal" yield curve of the field as a whole. Each single plot gives a reading, a reading which is made up of the "normal yield" together with a certain deviation imposed upon it by the factors producing variation.

In general there is no reason why any particular variety should give more reliable readings than others. It therefore seems pretty clear that, *unless there is any special reason to the contrary, all the different varieties should be used to determine the "normal fertility curve" of the field.*[1]

The problem of smoothing has received a good deal of attention during the last few years[2] and a large number of formulæ are available for this purpose. It is not however an easy task to choose the most suitable method for any particular case. Mr. Sarkar has used the method of "moving averages"; it is certainly simple, but has no other special merit; it also suffers from the disadvantage that the normal yield for the end-plots cannot be determined by this method.

From the nature of the problem it seems clear that *a very smooth* rather than a very close fit is desirable. My own feeling is that Whittaker's probability method of smoothing would probably give very good results with the present type of material. It possesses several advantages; the total of the variates and their first and second moments (which are often required for statistical purposes) are the same in the smoothed table as in the actual statistics on which the smoothing is based; it has satisfactory logical basis in the mathematical theory of probability; it makes use of the whole material available to graduate each individual value; there is no difficulty near the beginning or end of the plot; and finally the computations are fairly easy and straightforward. The numerical processes are described in detail in Whittaker's book, but I am quoting the necessary formulæ in Appendix I for easy reference.

Taking 6 varieties and all 60 plots I get the following expression for the smoothed or graduated values

$$y=9\ 17{\cdot}54-1{\cdot}34\ 65\ x-0{\cdot}05\ 87\ 88\ x^2$$

[1] This would also serve to eliminate factors of differential fertility. Let us take an extreme case. Consider a field in which the fertility increases for one variety, say (a), and decreases for another variety, say (b), from one end to the other. The fertility curve for (a) and the fertility curve for (b) will then be two single straight lines inclined to one another. Using either the a-curve or the b-curve alone we may get fallacious results, but using both we can get rid of the differential factors.

[2] For an excellent account see "Smoothing" by E. C. Rhodes, *Tracts for Computers*, No. VI (Cambridge University Press, 1921) and also Chapter XI of Whittaker and Robinson's *Calculus of Observations* (Blackie & Co., Ld., 1924).

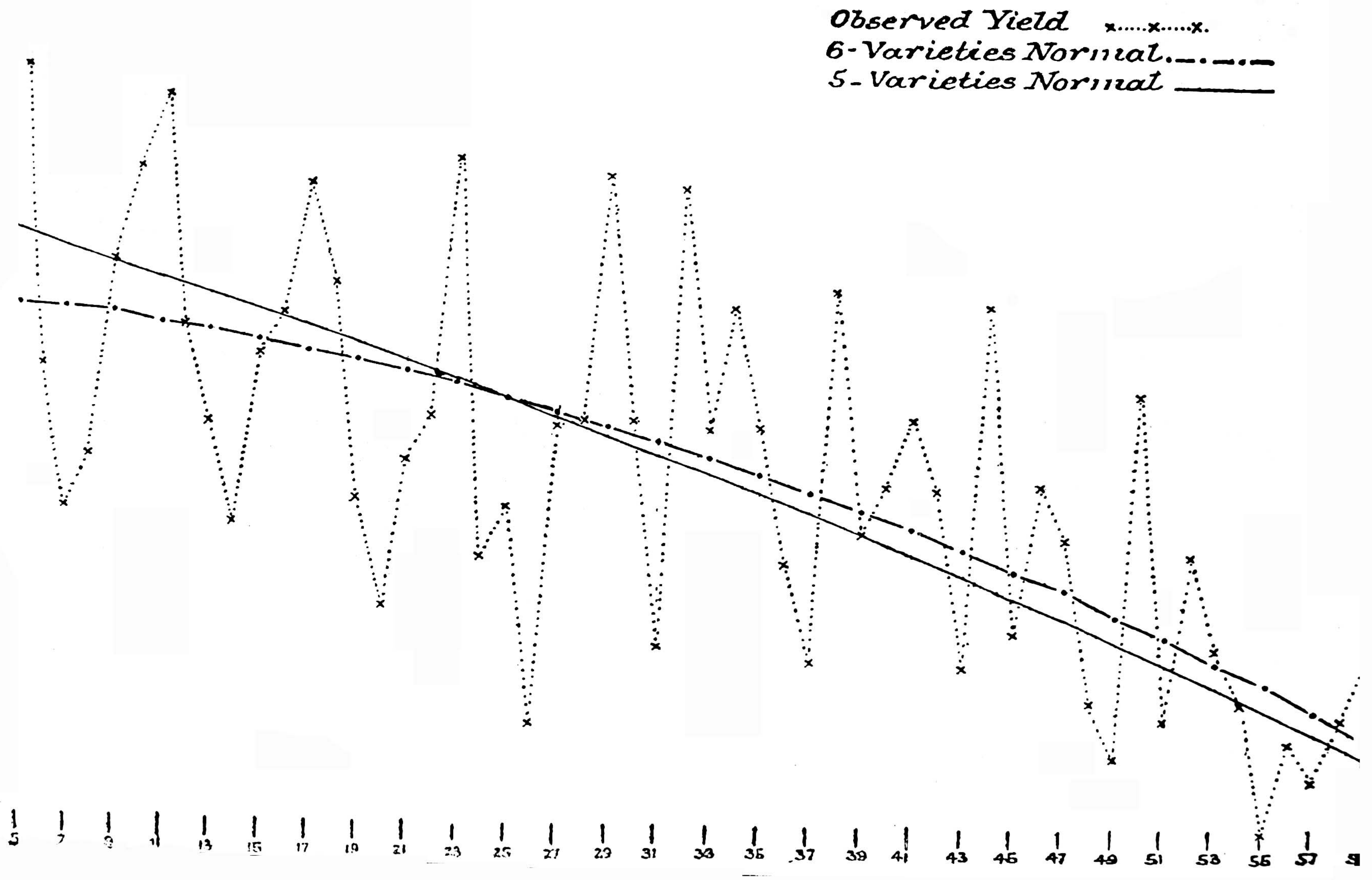
The graduated values[1] are given in Table I (col. 2) and are plotted as a graph in Diagram 1. Here y gives the yield in *tolas* and x represents the serial number of plots.

TABLE I.

"Normal yield" (by Whittaker's method).

No. of plot	Normal based on 6 varieties	Normal based on 5 varieties	No. of plot	Normal based on 6 varieties	Normal based on 5 varieties	No. of plot	Normal based on 6 varieties	Normal based on 5 varieties
1	916	976	21	863	869	41	761	744
2	915	971	22	859	864	42	755	738
3	913	966	23	855	858	43	749	731
4	911	961	24	851	852	44	742	724
5	909	956	25	846	846	45	735	717
6	907	951	26	842	840	46	728	710
7	905	946	27	837	834	47	722	703
8	903	941	28	833	828	48	715	696
9	901	935	29	828	822	49	707	689
10	898	930	30	823	816	50	700	682
11	895	925	31	818	809	51	693	675
12	893	920	32	813	803	52	685	668
13	890	914	33	808	797	53	677	661
14	887	909	34	802	790	54	670	653
15	884	903	35	797	784	55	662	646
16	881	898	36	791	777	56	654	639
17	877	892	37	785	771	57	646	631
18	874	887	38	780	764	58	637	624
19	870	881	39	774	758	59	629	616
20	867	875	40	768	751	60	621	608

[1] It will be noticed that the "normal yield curve" is nearly linear, showing a very steady decrease of fertility from one end of the field to the other. I have also calculated the correlation between the actual yield and the position of the plot as indicated by its serial number. The coefficient of correlation comes out to be r= —0·67 58. This gives a convenient measure of the variation of fertility. If we assume that all the different varieties are equally affected by this variation, then the correlation between any two varieties will also be 0·67 58 but *positive, i.e.*, the correlation between any two varieties is r (ab)= +0·68 approximately. Mr. Sarkar gives (p. 481) the p. e. of Kalamdan as 21·5 and of Early Indrasal as 19·33. Let us take e (a)= e (b)= 20 approximately. Using the complete expression for the probable error of mean (a—b), we get $\sqrt{20^2+20^2-2\times0{\cdot}68\times20\times20}=\pm16{\cdot}0$ approximately, which compares very favourably with the value 14·1 found directly by Mr. Sarkar. Evidently the correlation between Kalamdan and Indrasal is higher than +0·68 but is of the same order.



Observed Yield x.....x.....x.
6-Varieties Normal
5-Varieties Normal
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57

Subtracting the normal value (*i.e.*, the graduated value) from the observed value of the yield we get the "departure from normal"[1] and expressing it as a percentage of the normal, we obtain the percentage departure. Tables II and III (which correspond to Mr. Sarkar's Table VII, p. 485) give these percentage departures and the means and other constants.

TABLE II.

Percentage departure from the first normal.

No.	Indrasal	No.	Lochai	No.	Dudhsar	No.	No. 26	No.	No. 51	No.	Kalamdan
1	—13·2	2	— 6·3	3	+10·4	4	+18·6	5	+16·3	6	— 4·4
7	—14·6	8	—10·5	9	+ 8·8	10	+11·1	11	+16·8	12	..
13	— 6·7	14	—14·0	15	—1·0	16	+ 2·2	17	+12·3	18	+ 5·3
19	—10·3	20	—18·1	21	—6·7	22	— 3·0	23	+17·0	24	—12·8
25	— 8·4	26	—24·6	27	—1·1	28	— 0·4	29	+19·3	30	+ 0·9
31	—16·3	32	+20·5	33	+2·1	34	+12·5	35	+ 3·8	36	— 6·7
37	—14·0	38	+17·2	39	—0·8	40	+ 2·6	41	+ 9·2	42	+ 4·1
43	—10·3	44	+22·0	45	—5·4	46	+ 8·5	47	+ 4·6	48	— 9·1
49	—13·0	50	+21·1	51	—7·6	52	+ 8·8	53	+ 0·9	54	— 3·0
55	—14·4	56	— 4·4	57	—7·1	58	+ 0·5	59	+ 8·1	60	+ 0·6

TABLE III.

Mean percentage departures, etc.

No.	(1) Variety	(2) Mean percentage departure and corrected prob. error	(3) Standard deviation	(4) Z=M/s	(5) Odds based on Z
1	No. 51	+10·83 ±1·46	6·12	1·77	>5×10[3]
2	No. 26	+ 6·14 ±1·54	6·47	0·95	101·0
3	C. P. Lochai	+ 0·29 ±4·09	17·16	0·02	1·10
4	Dudhsar	— 1·34 ±1·29	5·42	0·25	3·19
5	Kalamdan	— 2·51 ±1·31	5·50	0·46	8·76
6	Indrasal	— 12·2 ±0·69	2·90	4·18	>10[4]

Table III, Col. (2) gives the mean percentage departure with "corrected" probable error (explained below). Col. (3) gives the standard deviation and Col. (4) is "Student's" function $Z=M/s$ (described on p. 99), while Col. (5) gives the odds based on Col. (4) and "Student's" Table.

[1] Mr. Sarkar subtracts the actual value from the normal; his departures are therefore opposite in sign to mine.

The odds are 10,000 to 1 or overwhelmingly against Indrasal and 5,000 to 1 or overwhelmingly in favour of No. 51. The odds are 100 to 1 in favour of No. 26; it is therefore fairly certain that No. 26 gives a better yield than the normal. Dudhsar and C. P. Lochai are more or less average, while Kalamdan is probably slightly inferior.

It will be noticed that C. P. Lochai gives a very high standard deviation 17·16, showing abnormal variations. Mr. Sarkar has noted this; he thinks that it is due to some "accident." I shall come back to this point a little later.

The observed standard deviations given in Col. (3), Table III, are corrected by multiplying them by the factor 1·11 87 (which is the appropriate value for $n=10$ in Table A above). The corrected probable errors are then obtained by multiplying the "corrected" standard deviations by $\cdot 6745/\sqrt{10}$; they are given in Col. (2), Table III.

We can now proceed to compare any two varieties with the help of the corrected probable errors given in Table III, Col. (2) above. We have presumably got rid of the systematic variations of the external factors, *i.e.*, of the fertility of the land, so that we shall be now justified in using the abbreviated formula $e\,(a-b)=\sqrt{e^2\,(a)+e^2\,(b)}$ for finding the probable error of differences.

Mr. Sarkar has also used a modified form of this formula[1] for constructing his Table VII, but he has not applied the correction for smallness of the size of samples. In the following Table IV, I show the odds calculated by using both the "corrected" and the "uncorrected" standard deviations.

We can also use the percentage departures given in Table II for a direct comparison of any two varieties by the "difference method," using "Student's" Table. I have calculated the odds by this method also and have shown them in Col. (5) of Table IV. The odds are reduced to unity in each case.

TABLE IV.

Varieties compared	Mean percentage difference with prob. error	Odds based on		
		"Uncorrected" prob. error	"Corrected" prob. error	Difference method
(1)	(2)	(3)	(4)	(5)
No. 51 and—				
No. 26 . . .	$+4\cdot69\pm2\cdot12$	19·62	13·68	9·73
C. P. Lochai . . .	$+10\cdot54\pm4\cdot34$	28·76	18·80	8·76
Dudhsar . . .	$+12\cdot17\pm1\cdot95$	$>10^6$	$7\cdot7\times10^4$	$8\cdot5\times10^2$
Kalamdan . . .	$+13\cdot34\pm1\cdot96$	$>10^7$	$5\cdot0\times10^5$	$3\cdot3\times10^3$
Indrasal . . .	$+22\cdot95\pm1\cdot61$	$>10^{25}$	$>10^{20}$	$3\cdot0\times10^4$

[1] Mr. Sarkar uses the approximation $\sqrt{2}\left\{\frac{e\,(a)+e\,(b)}{2}\right\}$ This introduces a small error in his results. The abbreviated formula (1) is actually simpler in use, as it avoids the multiplication by $\sqrt{2}$ The p. e. of a difference can be written down in a few seconds with the help of Barlow's Table (or any other Table) of squares.

TABLE IV—*concld.*

Varieties compared	Mean percentage difference with prob. error	Odds based on		
		" Uncorrected " prob. error	" Corrected " prob. error	Difference method
(1)	(2)	(3)	(4)	(5)
No. 26 and—				
C. P. Lochai . . .	+ 5·85 ± 4·37	5·40	4·43	5·67
Dudhsar . . .	+ 7·48 ± 2·01	390	164·7	$1{\cdot}8 \times 10^3$
Kalamdan . . .	+ 8·65 ± 2·03	$1{\cdot}6 \times 10^3$	502	100
Indrasal . . .	+ 18·26 ± 1·69	$>10^{15}$	$>10^{11}$	$2{\cdot}0 \times 10^4$
C. P. Lochai and—				
Dudhsar . . .	+ 1·63 ± 4·29	1·59	1·52	1·52
Kalamdan . . .	+ 2·80 ± 4·30	2·20	2·03	5·13
Indrasal . . .	+ 12·41 ± 4·15	83	45·10	23·6
Dudhsar and—				
Kalamdan . . .	+ 1·17 ± 1·84	2·17	2·00	2·19
Indrasal . . .	+ 10·78 ± 1·47	$>10^{10}$	$2{\cdot}8 \times 10^6$	$2{\cdot}04 \times 10^5$
Kalamdan and Indrasal .	+ 9·61 ± 1·48	$>10^8$	$>10^5$	$2{\cdot}18 \times 10^3$

The " uncorrected " probable errors give too high values for the odds and may easily create a false sense of security. The " difference method " appears to give the lowest odds and hence is probably the safest, but the " corrected " probable errors are not likely to lead to serious mistakes.

The construction of the " normal yield curve " for the whole field is admittedly an empirical process. A very careful examination of the raw material is therefore essential. One way of securing this would be to graduate each variety separately and then compare the results to see if there is any general agreement.

Adopting Whittaker's method I get the following expressions for the graduated values.[1] " y " in each case gives the yield in *tolas* and " x " the serial number of the plot.

No. 51	$y = 11\,23{\cdot}46 - 6{\cdot}76\,81\,x - 0{\cdot}01\,64\,84\,x^2$
No. 26	$y = 10\,71{\cdot}95 - 8{\cdot}53\,72\,x + 0{\cdot}03\,17\,61\,x^2$
Dudhsar	$y = 9\,95{\cdot}94 - 6{\cdot}16\,87\,x - 0{\cdot}01\,08\,38\,x^2$
Kalamdan	$y = 9\,10{\cdot}73 - 2{\cdot}54\,70\,x - 0{\cdot}04\,08\,25\,x^2$
Indrasal	$y = 8\,00{\cdot}73 + 0{\cdot}10\,46\,x - 0{\cdot}08\,04\,58\,x^2$
TOTAL .	$y = 49\,02{\cdot}81 - 24{\cdot}0\,64\,x - 0{\cdot}11\,66\,44\,x^2$

[1] I give in an appendix full details of the arithmetical work for one variety, Indrasal.

Dividing by 5, we get

$$Mean\ y = 9\ 80{\cdot}56 - 4{\cdot}80\ 13\ x - 0{\cdot}02\ 33\ 29\ x^2$$

The graduated values as well as the departures of the observed values from the graduated values are given in Table V and are plotted in Diagrams 2 and 3. The standard deviations of the departures are also given at the bottom of the columns.

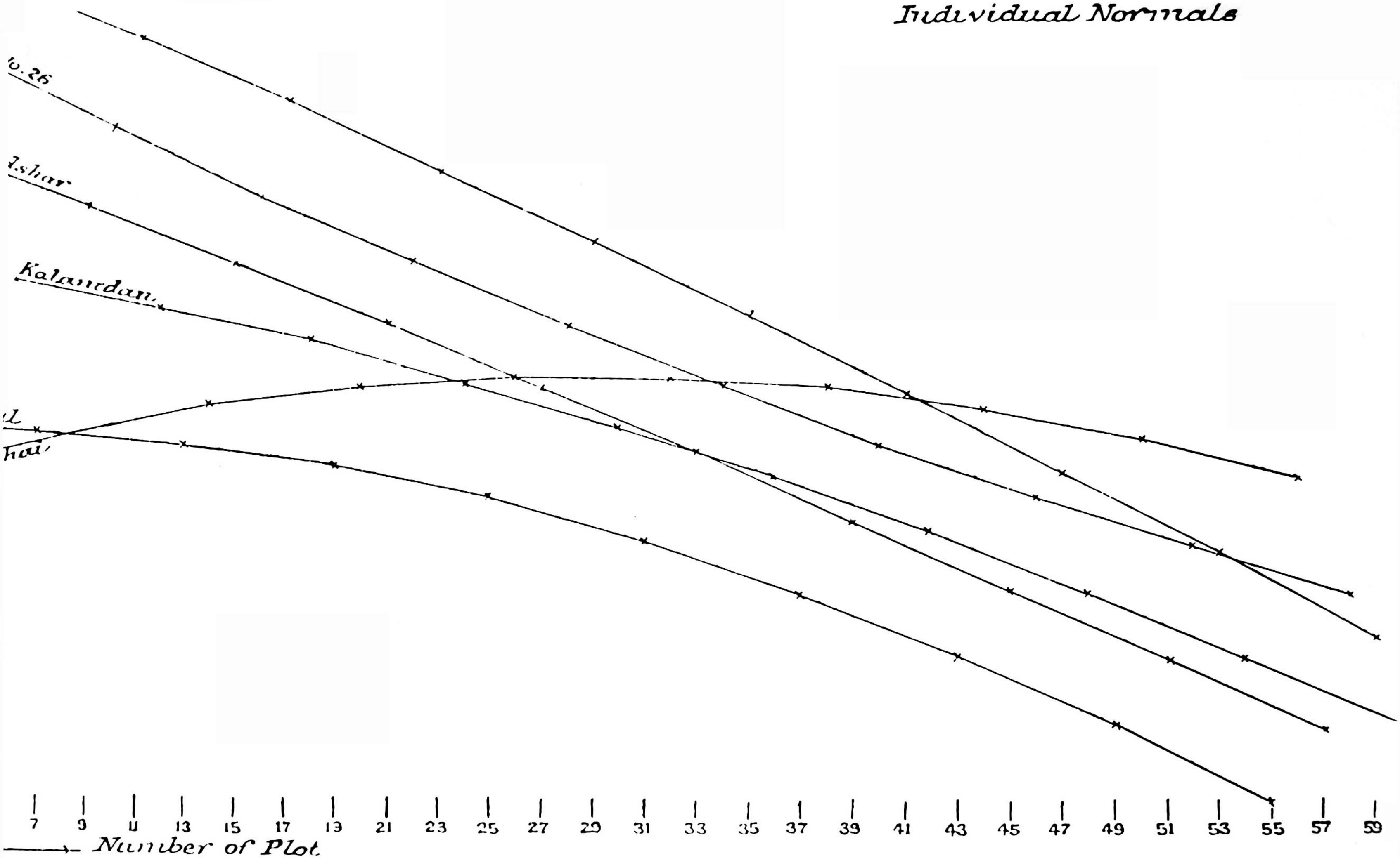
TABLE V.

Departures from normal.

No. 51		No. 26		C. P. Lochai		Dudhsar		Kalamdan		Indrasal	
(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1089	−32	1038	+42	772	+ 85	977	+31	894	−27	801	−6
1047	− 2	990	+ 8	795	+ 13	940	− 5	874	+19	798	−25
1004	−19	944	44	813	− 50	901	−26	852	+68	789	+41
959	+41	901	−68	824	−114	862	−57	826	−84	774	+ 6
913	+75	860	−30	829	−104	821	+ 7	798	+32	753	+22
866	−39	822	+80	828	+152	781	+44	766	−28	727	−42
816	+13	786	+ 2	821	+ 93	739	+29	732	+54	694	−19
769	−14	752	+38	808	+ 97	696	− 1	694	−44	656	+16
718	−46	721	+24	780	+ 59	653	−13	654	− 4	613	+ 2
667	+13	692	−52	764	−139	609	− 9	611	+14	563	+ 4
S. D.=34·6		S. D.=45·1		S. D.=107·0		S. D.=28·3		S. D.=44·3		S. D.=23·1	

It will be seen from Diagram 2 that all the different varieties with the one single exception of C. P. Lochai are in satisfactory agreement. The departure curve for C. P. Lochai, as well as its high standard deviation, 107·0, show that it is most probably affected by "accidental" (or sudden and discontinuous) errors. It is obvious that we cannot use it for constructing the field normal. It is too irregular and must be rejected.

We may now combine the 5 concordant varieties and construct the "normal curve." The simplest way of doing this will be to take the arithmetic mean of the 5 component curves. We can obtain the equation for the "normal yield" by simply taking the mean values of the three constants a, b and c. The arithmetic mean has already been given above. The "normal yield" is given in Col. 3 of Table I and is shown as a graph in Diagram 1.



Individual Normals
o.26
dshar
Kalamdan
d
hai
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59
Number of Plot

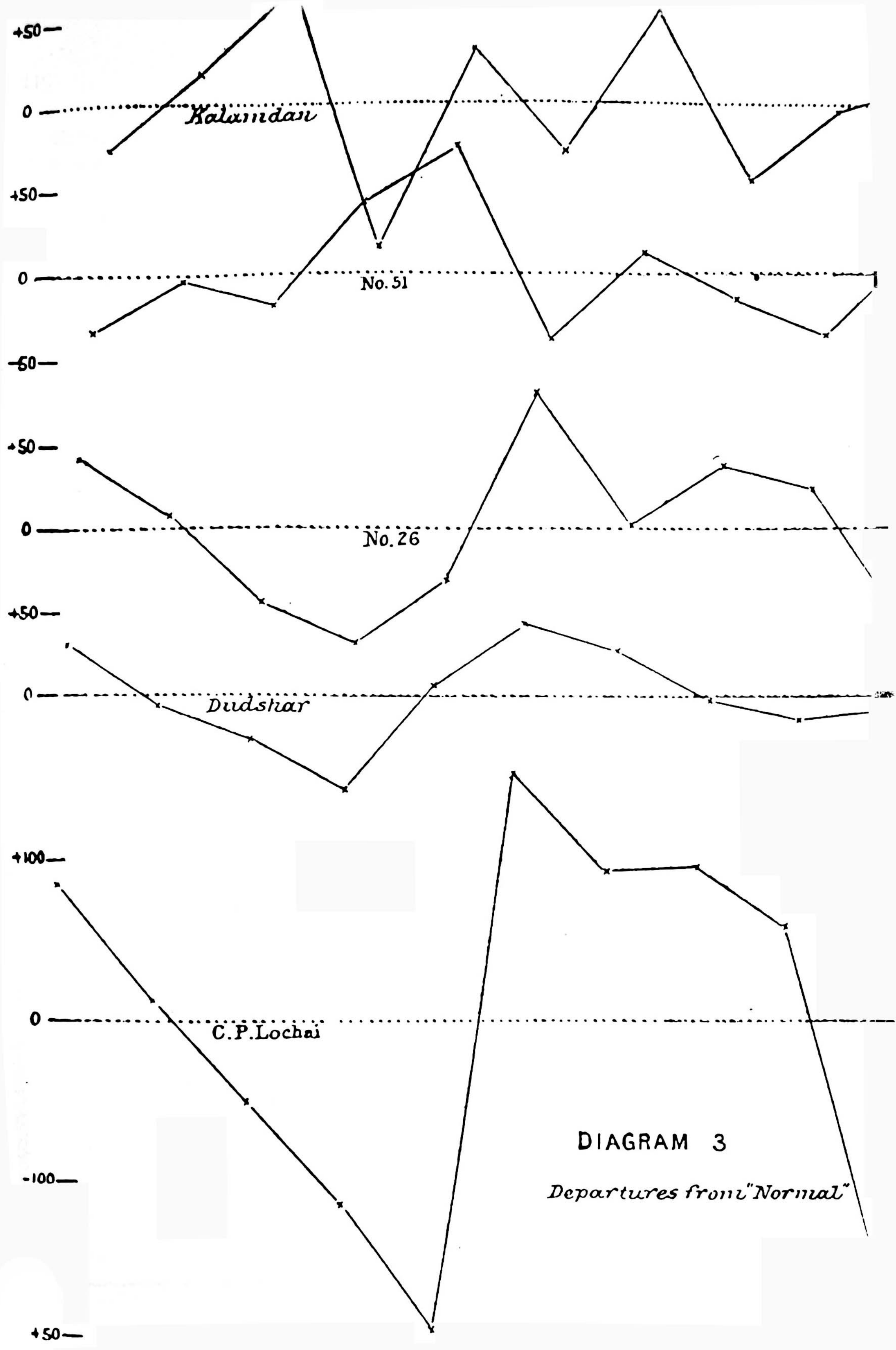


DIAGRAM 3

Departures from "Normal"

Having obtained our "normal" we can now proceed to calculate the percentage departures from normal, mean percentage differences and probability in the same way as described above. Corresponding to Tables II, III and IV we get a new set of Tables VI, VII and VIII which are given below.

Table VI.

Percentage departure from the second normal.

No.	No. 51	No.	No. 26	No.	Dudhsar	No.	Kalamdan	No.	Indrasal
5	+10·56	4	+12·38	3	+4·35	6	— 8·83	1	—18·55
11	+12·97	10	+ 7·31	9	..	12	— 2·93	7	—18·18
17	+10·43	16	+ 0·22	15	—3·10	18	+ 3·72	13	— 8·75
23	+16·55	22	— 3·59	21	—7·36	24	—12·91	19	—11·46
29	+20·20	28	+ 0·24	27	—0·72	30	+ 1·72	25	— 8·39
35	+ 5·48	34	+14·18	33	+3·59	36	— 5·02	31	—15·33
41	+11·69	40	+ 4·93	39	+1·32	42	+ 6·50	37	—12·45
47	+ 7·40	46	+11·27	45	—3·07	48	— 6·61	43	— 8·07
53	+ 3·33	52	+11·53	51	—5·19	54	— 0·46	49	—10·74
59	+10·39	58	+ 2·56	57	—4·91	60	+ 2·80	55	—12·38

Table VII.

Mean percentage difference.

Variety	Mean percentage departure with corrected P. E.	Corrected standard deviation
No. 51	+10·90 ± 1·13	5·23
No. 26	+ 6·10 ± 1·39	6·51
Dudhsar	— 1·52 ± 0·88	4·12
Kalamdan	— 2·20 ± 1·39	6·53
Indrasal	—12·43 ± 0·87	4·05

TABLE VIII.

(1)	(2)	ODDS BASED ON	
Varieties compared	Percentage difference with corrected P. E.	Corrected P. E.	"Difference" method.
No. 51 and—			
No. 26	$+4{\cdot}80 \pm 1{\cdot}79$	27·45	10·29
Dudhsar	$+12{\cdot}42 \pm 1{\cdot}43$	$>4{\cdot}3\times10^{7}$	$>1{\cdot}0\times10^{4}$
Kalamdan . . .	$+13{\cdot}10 \pm 1{\cdot}79$	$>2{\cdot}6\times10^{6}$	$<3{\cdot}4\times10^{3}$
Indrasal	$+23{\cdot}33 \pm 1{\cdot}42$	Very large	Very large
No. 26 and—			
Dudhsar	$+7{\cdot}62 \pm 1{\cdot}64$	$>1{\cdot}1\times10^{3}$	$>1{\cdot}9\times10^{3}$
Kalamdan . . .	$+8{\cdot}30 \pm 1{\cdot}97$	456·5	91·76
Indrasal	$+18{\cdot}53 \pm 1{\cdot}64$	Very large	$>2{\cdot}4\times10^{4}$
Dudhsar and—			
Kalamdan . . .	$+0{\cdot}68 \pm 1{\cdot}65$	1·57	1·57
Indrasal	$+10{\cdot}91 \pm 1{\cdot}23$	$>7{\cdot}7\times10^{7}$	$>2{\cdot}7\times10^{3}$
Kalamdan and Indrasal .	$+10{\cdot}23 \pm 1{\cdot}64$	$>7{\cdot}8\times10^{4}$	$>3{\cdot}8\times10^{3}$

We see again that the "difference method" gives the lowest odds. It should be remembered however that the plots compared are in certain cases widely separated, *e.g.*, Indrasal and Kalamdan, where the first plot of Kalamdan is separated from first plot of Indrasal by 4 intervening plots but is actually contiguous to the *second* plot of Indrasal.

The "Sub-plot" Method.

In order that the present method may give reliable results it is obviously necessary that the "normal yield curve" should be a reasonable description of the actual variation of fertility of the field. In fact the reliability of the present method depends entirely on the probability that the normal yield curve is a true description of the actual situation. In order to appreciate the real probability of the results obtained by the present method it is therefore necessary to determine the probability of the normal yield curve.

Now the normal yield curve is built up from the separate individual yield curves which in their turn are obtained from the values of the yield for each individual plot. The probability of the normal yield curve thus ultimately rests on the probability of the yield for each individual plot. It is therefore essential that we should have some basis on which we can determine the probable error of the yield for single plots. This point is of vital importance, but unfortunately in the procedure usually adopted in agricultural experiments absolutely no way is left for determining the probable error of the yield of single plots.

The only possible basis on which this can be done is to divide each plot into a number of sub-plots and secure the yield from each separate sub-plot. The scheme can be diagrammatically represented as follows :---

a (11)	b (11)	c (11)	..	..	..	a (21)	b (21)	..	..
a (12)	b (12)	c (12)	..	..	..	a (22)	b (22)	..	..
a (13)	b (13)	c (13)	..	..	..	a (23)	b (23)	..	..
a (14)	b (14)	c (14)	..	..	..	a (24)	b (24)	..	..
a (15)	b (15)	c (15)	..	..	..	a (25)	b (25)	..	..
a (1·).	b (1·)	c (1·)	..	..	..	a (2·)	b (2·)	..	..

As at present arranged the yield for the whole plot a (1) is determined integrally. In the method proposed the plot a will be divided into a number of sub-plots a (11), a (12), a (13), a (14) .. .. etc., and the yield for each sub-plot will be determined separately. The total yield for a (1) will then be obtained by adding the yield for all the sub-plots.

The essential point, however, is that this method will enable us to determine the probable error of a (1), a (2), a (3) .. .. . The probability of the graduated yield-curves for a, b, c (*i.e.*, for all the different varieties) can then be found and finally the probability of the normal yield curve itself.

The additional labour involved in the purely agricultural portion of the work is negligible. The only thing necessary will be to mark out the sub-plots a (11), a (12), a (13) .. b (11), b (12), b (13) .. .. etc.

Harvesting, threshing and weighing will however have to be carried out individually for each of the sub-plots a (11), a (12) .. .. .. b (11), b (12) .. .. etc., and will entail a much larger number of measurements. But the additional labour involved will probably be fully justified in view of the additional accuracy and the greater significance of the results which may be secured by this method.

I conclude therefore that *in order to obtain reliable results it is absolutely essential to adopt the above " sub-plot " (or " chess-board pattern ") method of laying out the experimental plots.*[1]

SUMMARY OF CONCLUSIONS.

It may be useful to indicate briefly the chief conclusions of my discussion.

(*a*) It is desirable to adopt, wherever possible, the " direct difference " method for finding the probable error of a mean difference and to use " Student's " Table for finding the probability.

(*b*) Wherever external variations are known to occur, it is desirable to eliminate their effect by considering departures from the " normal."

[1] I have indicated briefly in Appendix III suitable statistical formulæ which may be conveniently used in connection with the " sub-plot " method.

(*c*) In constructing the "normal," as many varieties as possible should be used ; and

(*d*) preferably, the results for each variety should be graduated separately and then combined together for constructing the "normal," after elimination of irregular varieties.

(*e*) In finding the probable error from small samples it is desirable to "correct" the observed standard deviations by multiplying them by suitable correcting factors. (A table of correcting factors has been given above.)

(*f*) Statistically speaking, it is absolutely essential to adopt a "sub-plot" method of laying out the field in order to determine the reliability of the "normal" used.

I note with interest that Mr. Sarkar proposes to test different statistical methods in connection with the results of further trials. May I suggest that in doing so he will keep in view the results offered above. Personally I shall be only too glad to give such statistical help as may lay in my power.

I am grateful to Dr. C. W. B. Normand, M.A., D.Sc., Officiating Director-General of Observatories, for drawing my attention to Mr. Sarkar's paper. I am also much indebted to my Assistant Babu Devendranath Chakravarti for arithmetical aid.

APPENDIX I.

Whittaker's method of smoothing.

Let a (1), a (2), a (3), ... a (n) be the successive yields. We then find the following fundamental constants by straightforward summation.

$$M = a(1) + a(2) + a(3) + \ldots + a(n)$$

$$M = 1.\,a(1) + 2.\,a(2) + \ldots + n.\,a(n)$$

$$M = 1^2.\,a(1) + 2^2.\,a(2) + \ldots + n^2.\,a(n)$$

Arranging the work in tabular form :—

	Yield		
(1)	(2)	(3)	(4)
Number of Indrasal plot = n	= a (n)	n. a (n)	n². a (n)
1	7 95	7 95	7 95
2	7 73	15 46	30 92
3	8 30	24 90	74 70
4	7 80	31 20	1 24 80
5	7 75	38 75	1 93 75
6	6 85	41 10	2 46 60
7	6 75	47 25	3 30 75
8	6 72	53 76	4 30 08
9	6 15	55 35	4 98 15
10	5 67	56 70	5 67 60
Sum	71 67	3 72 42	25 04 70

We get $M_0 = 71\ 67$. $M_1 = 3\ 72\ 42$. $M_2 = 25\ 04\ 70$.

From these we next obtain the following constants (remembering $n=10$).

$$p = \frac{M_0}{n} = \frac{7167}{10} = 716\cdot7$$

$$q = \frac{2\ M_1}{n\ (n+1)} = \frac{2 \times 372\ 42}{10 \times 11} = 677\cdot09$$

$$r = \frac{6\ M_2}{n\ (n+1)} = \frac{6 \times 25\ 04\ 70}{10 \times 11} = 1\ 36\ 60\cdot18\ 18$$

$$s = \frac{6\ (q-p)}{n-1} = \frac{6\ (677\cdot09-716\cdot7)}{9} = -26\cdot40\ 60\ 61$$

$$t = \frac{2[r-(2n+1)\ p]}{n-1} = \frac{2\ [13600\cdot1818-21 \times 716\cdot7]}{9} = -309\cdot00\ 40\ 40$$

From these we get finally—

$$c = \frac{15\ [t-(n+1)\ s]}{(n+2)\ (n-2)} = \frac{15\ [-309\cdot00\ 40\ 40+11 \times 26\cdot40\ 60\ 61]}{12 \times 8} = -2\cdot89\ 64\ 63$$

$$b = s-(n+1)\ c = -26\cdot40\ 60\ 61+11 \times 2\cdot89\ 64\ 63 = +5\cdot45\ 50\ 32$$

and $a = q - \frac{2n+1}{3} \cdot b - \frac{n\ (n+1)}{2} \cdot c = 677\cdot09 - 7 \times 5\cdot45\ 50\ 32 + 55 \times 2\cdot89\ 69\ 63 + 798\cdot21$.

The graduated values will then be given by :—

$$y = a + bZ + cZ^2 = 798\cdot21 + 5\cdot45\ 50.\ Z - 2\cdot89\ 64\ 63.\ Z^2$$

where Z refers to the number of Indrasal plots :—1, 2, 3, ... 10.

But these correspond to the serial number x :—1. 7. 13, ... 55. Evidently $Z = \frac{x+5}{6}$.

Substituting this value in the above equation,

$$y = 7\ 98\cdot21 + 5\cdot46\ 50\ \frac{(x+5)}{6} - 2\cdot89\ 64\ 63\ \frac{(x+5)^2}{6^2}$$

or $\quad y = 8\ 00\cdot73 + 0\cdot10\ 46\ x \quad -0\cdot08\ 01\ 58\ x^2$

the value quoted above.

As I have already pointed out, we want a very smooth fit. I therefore stop here and do not try to improve the fit by the method described by Whittaker.

APPENDIX II.

The mean difference for graduated values can be obtained very easily with the help of simple algebraic expressions

Let $y_1 = a_1 + b_1\ x + c_1\ x^2$ and $y_2 = a_2 + b_2\ x + c_2\ x^2$ be any two graduated yield-curves where "y" gives the yield and "x" the serial number of the plot. Then $Z = y_1 - y_2 = A + B\ x + C\ x^2$ where $A = a_1 - a_2$, $B = b_1 - b_2$ and $C = c_1 - c$.

It can be easily shown that the mean value of the difference

$$Z = A + B.\ \frac{1}{n}\ S(x) + C.\ \frac{1}{n}.\ S(x^2)$$

and the square of the standard deviation of Z,

$$s^2(z) = \left[\frac{1}{n}.\ S(x^2) - \left\{\frac{1}{n}.\ S(x)\right\}^2\right] B^2$$
$$+ \left[\frac{1}{n} S(x^3) - \frac{1}{n^2} S(x)\ S(x^2)\right] BC$$
$$+ \left[\frac{1}{n} S(x^4) - \left\{\frac{1}{n} S(x^2)\right\}^2\right] C^2$$

where S denotes a summation for the different values of n.

Now in the present type of problem x will usually take the form a + (n—1) d where, " a " = general serial number of the first plot. " n " = the number of plots sown with each variety and " d " = number of different varieties. For example, in the illustration given by Mr. Sarkar n = 10 and d = 6. For Indrasal a = 1, and we get 1, 7, 13 .. .. 55 for successive values of x.

Thus putting x = a + (n—1) d, we get

$$\frac{1}{n}\ S(x^2) - \left\{\frac{1}{n}\ S(x)\right\}^2 = \frac{n^2-1}{12}\ d^2$$

$$\frac{1}{n} S(x^3) - \frac{1}{n^2} S(x).\ S(x^2) = \frac{n^2-1}{12}\ d^2 \{2a + (n-1)\ d\}$$

$$\frac{1}{n} S(x^4) - \left\{\frac{1}{n} S(x^2)\right\}^2 = \frac{n^2-1}{3}\ d^2 \left\{a^2 + (n-1)\ a.\ d + \frac{16n^2 - 30n - 1}{60}\right\}$$

from which Z and s (z) can be easily found.

APPENDIX III.

Goodness of fit of the " normal curve."

We may apply the X^2 test for " goodness of fit " devised by Karl Pearson. Let a (1), a (2), a (3) .. .. , in general, a (n) be the yield of the nth a—plot, let a(O) = mean yield of all a. plots taken together and s^2 (a) = the square of the standard deviation of all a. plots irrespective of the positions in the field. Let p be the number of sub-plots into which *each* plot is divided and let n = total number of a. plots altogether.

Then the correlation ratio η (of yield on position of the plot) is defined by the equation

$$\eta^2 = \frac{p}{n}.\ S\left\{\frac{(n) - a\ (o)}{s^2\ (a)}\right\}^2$$

where S denotes a summation for all n values of a.

Let a^1 (n) be the graduated value corresponding to observed value a (n).

Karl Pearson has suggested the following value* for X^2 :—

$$X^2 = p.\ \frac{S\{a\ (n) - a^1\ (n)\}^2}{(1-\eta^2).\ s^2\ (a)}$$

The value of X^2 can be easily calculated with the help of the above two formulæ.

* " On the application of ' Goodness of Fit ' Tables to test Regression Curves and Theoretical Curves used to describe Observational or Experimental Data." *Biometrika*, XI (1915-1917), 239-261.

Tables are available for finding the probability of the fit from observed values of X^2, the quantity defined above.*

It should be remembered that here the total number of independent variables is n in the present case and hence the probability Tables for X^2 should be looked up under $n^1 = n+1$.

Having obtained the probability for each separate curve we can assign suitable statistical weights to the different curves in constructing the normal. The probability of the normal can then be found either from the constants of the component curves or directly from the correlation ratio of the yield on position in the field for all plots taken together. In the absence of actual material I am unable to give a numerical illustration.

---

* *Biometrika*, I (1903), 155-163, reprinted as Table XII, p. 26, *Tables for Statisticians and Biometricians*.

FURNACES FOR THE MANUFACTURE OF JAGGERY OR GUR

BY

W. SMITH-ROLLO, A.M.I.E. (India),

Agricultural Engineer to Government, Burma.

In Burma, the area under sugarcane has more than doubled itself within the last 10 years and, as a result of this, the enormous amount of timber consumed in the furnaces in the province has forced the Forest Department to restrict the supply of fuel.

The cane-grower here ordinarily uses fuel 5–8 inches in diameter and 6–9 feet long and costing from Rs. 1-8 to Rs. 3 per cartload of approximately 1,000 lb.

One of the greatest problems, therefore, in the sugarcane districts is the supply of fuel, and for the last 18 months the author and his staff have devoted much time to the designing of a furnace which would be able to burn the megass or crushed cane and at the same time evaporate the juice without the addition of any wood fuel.

In the experiments hereafter related none of the leaves or trash was burnt but megass only, thus releasing a certain quantity of valuable manure which can be ploughed into the land, although it is usually set on fire before ploughing.

This year in the Yamethin District near Pyinmana which is the centre of the most important cane-growing district in Burma, two experimental furnaces were erected and the object aimed at was to produce a furnace which would not differ too greatly from the Burmese furnaces to be popular with the growers and at the same time which would effect a saving of fuel and time.

1. The Ordinary Burmese Furnace.

This consists of a pit 6 feet long by 3 feet broad by 3 feet deep, and at one end of this a tunnel about 18–24 inches in diameter is started for the fire-box and continued for a distance of about 20 feet, holes being cut in the roof for the placing of pans as shown in Fig. 1. These pans are usually about 45 inches in diameter and 16 inches deep at the centre and hold 32 gallons or 8 kerosine tins of juice.

No chimney is usually built but if so it is never more than 2 or 3 feet high, and logs about 6 inches in diameter and 6 feet long are always burnt.