STATISTICAL NOTES FOR AGRICULTURAL WORKERS.*

No. 6.—THE EFFECT OF FERTILIZERS ON THE VARIABILITY OF THE YIELD AND THE RATE OF SHEDDING OF BUDS, FLOWERS AND BOLLS IN THE COTTON PLANT IN SURAT.

BY

P. C. MAHALANOBIS.

Professor of Physics, Statistical Laboratory, Calcutta.

(Received for publication on 3rd October 1932).

1. Mr. K. V. Joshi of the Cotton Research Laboratory, Surat, has sent us for statistical analysis the results of certain experiments † conducted by him in 1930-31 with fertilizers on selected strains of the cotton plant.

The effect of fertilizers may be divided into two distinct groups—(A) changes in the mean value of the yield, or the mean rate of shedding at different stages, and (B) changes in the variability of the yield or the variability of the rate of shedding. Consider the production of " buds ". The application of fertilizers may affect the number of " buds " produced. It may also affect the variability of the bud production from plant to plant under the same treatment. These two effects must be carefully distinguished.

Fisher's method of " analysis of variance " is designed to test whether the mean values are affected or not, that is, to investigate effect (A). In this method it is assumed that the variabilities are identical, *i.e.*, that effect (B) is inappreciable.

It is quite possible, however, that mean values are not affected, *i.e.*, effect (A) is inappreciable, while effect (B) is not negligible, so that variabilities are appreciably altered.

Speaking generally it is desirable to investigate both the effects. In case variabilities are appreciably the same (*i.e.*, effect (B) is neglible), Fisher's *z*-test can be applied to test whether mean values have altered. If effect (B) is not negligible, further studies may become necessary. Neyman and Pearson [1931] have considered this problem very fully in a recent paper and have developed suitable methods for disentangling the different effects. I shall use Mr. Joshi's data to illustrate the use of such methods.

* We are receiving a large number of enquiries of a statistical nature from agricultural workers in different parts of India. Many of these enquiries are of cosiderable general interest, and it is proposed to publish notes on selected topics from time to time. These notes will deal mainly with statiscal methods and procedure, and it is not intended that they should always contain new matter.—Ed.

† The actual data will be found in Appendix I given at the end of the paper and fuller details in another paper, Statistical Notes No. 7. This Journal, Vol. 3, p. 139.

2. Neyman and Pearson [1931] have distinguished three different cases :—

(*i*) The hypothesis H_0 that the samples belong to populations having the same mean value and the same variance, so that so far as the mean value and the variance are concerned, the samples may be considered to have come from the same population [1930].

(*ii*) The hypothesis H_1 that the samples come from populations having the same variance. The mean value may be different for each population (or identical as a special case).

(*iii*) The hypothesis H_2 that the mean values are identical, it being assumed that the variances are also identical.

It will be noticed that (*i*) H_0 will test whether both mean values and variances are identical, (*ii*) H_1 whether variances are equal or not, and (*iii*) H_2 whether mean values are equal or not, it being assumed that variances are identical. Test of H_2 is simply an alternative form of Fisher's z-test and need not be further considered here. It must be emphasized, however, that Fisher's method of analysis of variance in its usual form can only be applied on the assumption that the variances are identical.

3. It will be now necessary to define certain statistical parameters.

Let n_t, m_t and S_t^2 be the size, mean value and variance of the "t"th sample. Then

$$n_t\, m_t = s\,(x) \qquad n_t\, s_t^2 = S\,(x - m_t)^2 \quad . \quad . \quad . \quad . \quad (1)$$

where S represents a summation for all n_t values of x for the "t"th sample. s_t^2 is thus the observed standard deviation for a single sample. Let there be k such samples.

The average variance s_a^2 within all samples is defined by :—

$$N = \Sigma(n_t) \qquad N s_a^2 = \Sigma\,(n_t\, s_t^2) \quad . \quad . \quad . \quad . \quad (2)$$

where Σ represents a summation for all "k" samples, and N is the total number of individual observations available.

Finally the general mean m_0 and the general variance s_0^2 are defined by

$$N m_0 = \Sigma\, S(x), \quad N s_0^2 = \Sigma S\,(x - m_0)^2$$

where ΣS represents the summation for all N values of x. It will be noticed that s_0^2 is the "total" variance, and s_a^2 the mean variance "withins amples" ordinarily used in the analysis of variance.

A case of special importance occurs when the size of the sample is the same for all samples, *i.e.*, $n_1 = n_2 = \ldots\ldots\ldots\ldots n_t = n$, and $N = nk$. In this case there is a considerable simplification in the working formulae. It is therefore extremely desirable to arrange the size of the sample to be the same, whenever this is possible, in field experiments.

* The author has considered this problem in greater detail from a theoretical standpoint in a different paper [1930].

Neyman and Pearson's formulae can then be put in the following form.

$$l_0 = \left(\frac{s_1^2}{s_0^2} \cdot \frac{s_2^2}{s_0^2} \cdot \quad . \quad . \quad . \text{ upto } \frac{s_k^2}{s_0^2}\right)^{1/k} \qquad (4)$$

$$l_1 = \left(\frac{s_1^2}{s_n^2} \cdot \frac{s_2^2}{s_n^2} \cdot \quad . \quad . \quad . \text{ upto } \frac{s_k^2}{s_n^2}\right)^{1/k} \qquad (5)$$

We may introduce the geometric mean of the variances defined by :—

$$s_g^2 = (s_1^2 . s_2^2 \quad . \quad . \text{ upto } s_k^2)^{1/k} \qquad (6)$$

or its logarithmic form :—

$$\log s_g^2 = 1/k \, (\log s_1^2 + \log s_2^2 + \quad . \quad . \quad + \log s_k^2) \qquad (6{\cdot}1)$$

Then $\log \, l_0 = \log (s_g^2) - \log (s_0^2)$ . . . (4·1)

$$\log l_1 = \log (s_g^2) - \log (s_n^2) \qquad (5{\cdot}1)$$

Also $\log \, l_2 = \log (s_n^2) - \log (s_0^2)$ . . . (7)

So that $\log l_0 = \log l_1 + \log l_2$ . . . (8)

The formulae for the distribution of l_0 and l_1 have been given by Neyman and Pearson [1931].

4. The interpretation of l_0 and l_1 is extremely simple. If hypothesis H_0 is true (that is, if all "k" samples are drawn from a population having the same mean value and the same variance) then l_0 will be sensibly equal to unity. In the same way if hypothesis H_1 is true (that is, if the "k" samples are drawn from populations having the same variance but with either the same or different mean values), then l_1 will be sensibly equal to unity. On the other hand as l_0 and l_1 become smaller and smaller the hypothesis H_0 and H_1 respectively become less and less probable. In other words if l_0 is found to be significantly less than unity, the observed samples cannot be considered to have come from the same population. In the same way if l_1 is found to be sensibly lower than unity then the observed samples must be considered to belong to populations having different variabilities.

With the help of the formulae for moment co-efficients, given by Neyman and Pearson, it is possible to calculate the 5 per cent. and one per cent. points for both l_0 and l_1, and hence judge whether l_0 or l_1 is significantly lower than unity, or may be considered sensibly equal to unity.

The statistics η_2 is equal to $(1 - \eta^2)$ where η is the "correlation ratio" of Karl Pearson. When l_2 is small η^2 is large, so that the mean values for the different samples cannot be considered to be identical; l_2 thus furnishes simply an alternative form of Fisher's z-test.

It will be noticed that $l_0 = l_1 . l_2$. so that the value of l_0 may be reduced either due to l_1 or l_2. Thus hypothesis H_0 may become improbable owing to (*i*) the variabilities being different, or (*ii*) the mean values being different, or (*iii*) due to the joint effect of both the factors.

5. I shall now consider Mr. Joshi's data. The "control" plot will be referred to as sample No. 1, the July-manured plot as sample No. 2, and the August-manured plot as sample No. 3.

The calculation of l_0, l_1 and l_3 (or z) is quite simple and straightforward. The variances s_1^2, s_2^2, s^2 for the three samples are determined directly and the weighted geometric mean s_g^2 is calculated with the help of logarithms. The mean variance within samples s_a^2, and the general variance s_0^2 are required for the z-test and are calculated in the usual way.

The calculation for the number of "buds" is shown in Table (I,1). Adding the logarithms of the three individual variances, we get 11·9097214. Dividing by $k = 3$ we get the weighted geometric mean $\log s_g^2 = \bar{3}{\cdot}9699071$. Subtracting the logarithm of s_a^2 (=4·0134755) from $\log s_g^2$, we get $\log l_1 = \bar{1}{\cdot}9564316$. Similarly subtracting $\log s_0^2$ (=4·1392424) we get $\log l_0 = \bar{1}{\cdot}8306647$. The observed value of l_0=0·8741 and of l_1= 0·9465. (For comparison we also find that l_2=0·9236.)

The variances and calculations for the number of "flowers", "bolls" and the proportion of "flowers : buds", "bolls : flowers" and "bolls : buds" are shown in Tables (I, 2)—(I, 6).

The observed values of l_0, l_1, l_2 and z are shown in Table II.

6. We can now use Neyman and Pearson's theory to judge the significance of these observed values of l. Using the formulae for moment co-efficients given by them, we find the following values for the 5 per cent. and one per cent. points of l_0 and l_1 for $n = 20$, $k = 3$. The 5 per cent. and one per cent. values of z are also given for comparison.

Size of sample		Level of significance	l_0	l_1	z
$n = 20$	$N = n \cdot k = 60$	5 per cent.	0·8417	0·8912	0·5761
$k = 3$		1 per cent.	0·7816	0·8334	0·8065

The significance of the expected values is clear If hypothesis H_0 were true, that is, if sets of 3 samples of 20 ($k = 3$, $n = 20$) were repeatedly drawn from the same normal population, then the observed value of l_0 would be less than 0·8417 in 5 per cent. and less than 0·7816 in one per cent. of cases. Similarly if hypothesis H_1 were true, that is, if sets of 3 samples of 20 were drawn from normal populations with an identical variability (but equal or different mean values), then the observed value of l_1 would be less than 0·8912 in 5 per cent. and less than 0·8334 in one per cent. cases.

TABLE (I, 1).

Buds.

Sample	Variance	log
s_1^2	1,69,00·25	4·2278032
s_2^2	70,77·30	3·8498676
s_3^2	67,91·42	3·8319606
Total .		11·9097214
s_g^2	Average .	3·9699071
s_a^2	1,03,15·15	4·0134755
s_0^2	1,37,79·78	4·1392424
l_0	s_g^2/s_0^2	$\bar{1}$·8306647
l_1	s_g^2/s_a^2	$\bar{1}$·9564316
l_2	s_a^2/s_0^2	$\bar{1}$·8742331

TABLE (I, 2).

Flowers.

Sample	Variance	log
s_1^2	14,66·35	3·1662376
s_2^2	1,10,74·65	4·0443065
s_3^2	13,75·89	3·1385838
Total .		10·3491279
s_g^2	..	3·4497093
s_a^2	46,91·07	3·6715496
s_0^2	49,55·73	3·6950501
l_0	s_g^2/s_0^2	$\bar{1}$·7546592
l_1	s_g^2/s_a^2	$\bar{1}$·7781597
l_2	s_a^2/s_0^2	$\bar{1}$·9764995

TABLE (I, 3).

Bolls.

Sample	Variance	log
s_1^2	1,29·01	2·1106234
s_2^2	77·39	1·8886848
s_3^2	1,04·28	2·0182010
Total .		6·0175092
s_g^2	..	2·0058364
s_a^2	1,03·55	2·0151501
s_0^2	1,15·15	2·0612631
l_0	..	$\bar{1}$·9445733
l_1	..	$\bar{1}$·9906863
l_2	..	$\bar{1}$·9538870

TABLE (I, 4).

Flowers : Buds.

Sample	Variance	log
s_1^2	1,04·65	2·0197392
s_2^2	1,09·65	2·0400086
s_3^2	39·26	1·5939503
Total .		5·6536981
s_g^2	..	1·8845660
s_a^2	85·29	1·9308981
s_0^2	101·43	2·0061664
l_0	..	$\bar{1}$·8783996
l_1	..	$\bar{1}$·9536679
l_2	..	$\bar{1}$·9247317

TABLE (I, 5).

Bolls : Flowers.

Sample	Variance	log s^2
s_1^2	13·00	1·1139434
s_2^2	26·05	1·4158077
s_3^2	29·83	1·4746533
Total .		4·0044044
s_g^2	..	1·3348015
s_a^2	..	1·3586961
s_0^2	..	1·3932241
l_0	..	1·9415774
l_1	..	1·9761054
l_2	..	1·9654720

(TABLE I, 6).

Bolls : Buds.

Sample	Variance	log s^2
s_1^2	8·79	0·9439889
s_2^2	11·09	1·0449315
s_3^2	4·53	0·6560982
Total .		2·6450186
s_g^2	..	·8810729
s_a^2	8·20	0·9138139
s_0^2	10·03	1·0013009
l_0	..	1·8803720
l_1	..	1·9678590
l_2	..	1·9125130

TABLE II.

Observed values of l_0, l_1, l_2 *and* z.

Character	l_0	l_1	l_2	z
Buds	0·6771	0·9046	0·7486	1·1207
Flowers	0·5684	0·6000	0·9473	0·2215
Bolls	0·8802	0·9788	0·8993	0·5718
Flowers : Buds	0·7558	0·8988	0·8409	0·8250
Bolls : Flowers	0·8741	0·9465	0·9236	0·4185
Bolls : Buds	0·7579	0·9287	0·8176	0·9179
5 per cent.	0·8417	0·8912	0·8999	0·5764
One per cent.	0·7816	0·8334	0·8503	0·8005

7. Adopting an one per cent. level of significance we notice that l_0 is significantly lower than unity in 4 cases : (i) Buds, (ii) Flowers, (iii) Ratio of flowers to buds, and (iv) Ratio of bolls to buds, showing that the application of fertilizers produces significant effects in the case of "buds", "flowers", and the ratio of shedding of buds.

With the help of l_1 and l_2 (or rather z), we can make a deeper analysis. The z-test shows that the mean values are significantly different in the case of (i) "buds", and the ratio of (ii) "flowers : buds" and (iii) "bolls : buds", while l_1 shows that the variabilities are different in the case of "flowers". In the case of "flowers : buds" the l_1-test is on the verge of significance.

We conclude that the application of fertilizers has had the following effects--

(*i*) The mean number of "buds", the mean proportion of "flowers : buds", and the mean proportion of "bolls : buds" are all altered appreciably. The effect on the production of "buds" is apparently the basic factor. The application of Fisher's z-test is thoroughly justified in this case.

(*ii*) The variability of the production of "flowers" is altered but not the mean number of flowers. It is possible that this has caused a just appreciable effect on the variability of the proportion of "flowers : buds". In such cases the z-test would not reveal any effect.

(*iii*) In the case of "bolls" and the proportion of "bolls : buds" neither the mean values nor the variabilities appear to have altered appreciably.

8. It will be noticed from the above results that in certain cases (*e.g.*, the production of "flowers" in the cotton plant under different manurial treatments), the use of the z-test is not sufficient. It is, therefore, desirable and necessary to use the new tests of Neyman and Pearson whenever there is any suspicion of the variabilities becoming sensibly altered. The separate calculation of the variance for each sample is not difficult (in fact most of the arithmetical work is usually done in the course of the analysis of variance), and the calculation of l_0 and l_1 is also easy and straightforward and should take very little time. The expected values of l_0 and l_1, however, require very laborious calculations with Gamma functions. Tables of 5 per cent. and one per cent. values of l_0 and l_1 for a fairly large range of values of n and k have been prepared in my laboratory, and will be shortly published. With the help of these new tables, the use of the l-tests will be as easy and as simple as the use of the z-tests. It is scarcely necessary to point out that the l_0-and l_1-tests do not supplant but merely supplement the z-test.

REFERENCES.

Mahalanobis, P. C. (1930). *J. Asiat. Soc.*, Bengal, 26, 541.

Mahalanobis, P. C. and Bose, S. S. (1933). *Ind. J. Agric. Sci.* 3, 139.

Neyman, J. and Pearson, E. S. (1931). *Bull. de l'Acad. Pol. Cracow, Seris* A, 460.

APPENDIX I

	Buds			Flowers			Bolls			Bud to flower			Flower to boll			Bud to boll		
	"A"	Control	"B"	"A"	Control	"B"	"A"	Control	"B"	"A"	Control	"B"	"A"	Control	"B"	"A"	Control	"B"
1	378	266	331	136	104	186	42	34	36	36·0	28·4	41·1	30·8	32·7	26·5	11·1	9·3	10·9
2	375	258	156	145	90	74	37	28	19	38·7	34·9	47·4	25·5	31·1	25·7	9·9	10·85	12·2
3	169	184	205	66	79	83	20	31	29	39·1	42·9	40·5	30·3	39·2	34·9	11·8	16·8	14·2
4	622	184	306	234	60	130	67	22	40	37·6	32·6	42·5	28·6	36·7	30·9	10·8	11·95	13·05
5	293	208	402	114	120	176	36	35	44	38·9	39·0	43·8	31·6	29·2	25·0	12·3	11·4	10·9
6	218	159	259	94	56	104	27	19	35	38·5	35·2	40·2	32·1	34·1	33·7	12·4	11·9	13·5
7	410	358	178	141	126	83	41	40	27	34·4	35·2	46·6	29·1	31·1	32·5	10·9	11·2	15·2
8	280	88	300	117	39	147	28	12	38	41·8	44·3	48·0	23·9	30·8	25·9	10·0	13·6	12·4
9	470	356	259	141	133	132	39	41	45	30·0	34·5	45·7	27·7	30·8	34·1	8·3	10·6	15·6
10	431	375	126	133	137	61	40	43	17	30·9	36·5	48·3	30·1	31·4	27·9	9·3	11·5	13·5
11	266	152	..	102	68	..	28	21	..	38·4	44·7	..	27·5	30·8	..	10·5	13·8	..
12	504	234	225	166	94	101	45	82	33	32·9	40·2	44·9	27·1	34·0	32·7	6·9	13·7	14·7
13	639	260	224	186	111	110	58	30	27	29·1	42·7	49·1	31·2	35·1	24·5	9·1	15·0	12·0
14	380	175	221	131	85	107	44	29	33	33·9	48·6	48·4	33·6	34·1	30·8	11·4	16·6	14·9
15	502	295	296	193	117	155	53	41	44	32·6	39·7	52·4	27·5	35·0	28·2	8·95	13·9	14·85
16	219	224	208	99	96	103	24	28	32	45·2	42·9	48·5	24·2	29·2	31·1	10·95	12·5	15·4
17	337	184	276	150	95	139	37	23	39	44·5	51·6	50·0	24·7	21·2	28·1	11·0	12·5	14·1
18	309	315	149	152	151	60	43	42	19	49·2	47·9	53·7	28·5	27·8	23·9	13·9	13·4	12·7
19	397	253	400	169	104	2 7	51	33	60	42·6	41·1	47·2	30·2	31·7	27·6	12·8	13·0	13·0
20	308	328	271	113	120	116	54	40	42	36·7	30·6	42·8	30·1	33·3	36·2	11·0	10·4	15·5