A TEST OF THE SUPPOSED PRECISION OF SYSTEMATIC ARRANGEMENTS

By Dr S. BARBACKI and R. A. FISHER, Sc.D., F.R.S.

I. Introduction

In human as in biological experimentation frequent use is made of the method of pairing, often, though erroneously, ascribed to "Student", who has, however, expressly disclaimed this invention.

Possibly because of its early introduction, pairing has been so frequently applied without the precaution of randomization that, by force of example, it seems often to have been thought that this method would provide a valid estimate of error even when systematic pairs were used. Recently, indeed, in connexion with Beavan's split drill method of testing cereal varieties "Student" has claimed explicitly that higher precision is attainable with systematic than with randomized arrangements.

The only method of testing such an assertion is by the direct application of the two alternative methods to yields harvested in half-drill strips from a trial using only a single variety. For, though the precision attainable with the aid of randomization is well known from the many trials carried out by workers who have taken this precaution of obtaining unequivocally valid estimates of error, the precision of comparisons using a systematic arrangement of split drills is not known, since the estimates of error derived from this or any other systematic arrangement cannot be relied upon not to be biased in one direction or the other. "Student" indeed expresses the opinion that with the split-drill method the error is slightly over-estimated, and this, though contrary to what the following data will show, may, in other cases, be true. It is, however, a somewhat back-handed compliment to the method, for it implies, as "Student" does not appear to realize, that, using the systematic split-drill method, results which, with randomization, would have been recognized as significant will be passed over as without statistical significance. This has been demonstrated with certain systematic square arrangements (O. Tedin, 1931).

When this occurs, the effort and outlay expended in carrying through an experimental programme may have been frustrated merely through neglecting to take the precautions needed to obtain a valid estimation of error.

II. Experimental data

Wiebe gives yields in grams of grain for 1500 15 ft. rows of wheat. In order to parallel the situation in which the split-drill method is used these rows have been totalled in groups of six, omitting one row between each consecutive pair of groups. Each group thus gives the yields of a half-drill strip, of which there are sixteen running side by side, and twelve end to end, as shown in Table I. As in the systematic split-drill method these are lettered *A B B A*, *A B B A*, ..., across the field. The differences of the yields between the half-drill strips, taking *A–B* in each case, are shown in Table II.

Table I

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	(xi)	(xii)
A	4410	4035	3865	3640	3650	3985	3490	3330	3358	3712	3487	3781
B	3950	3865	3295	2960	2925	3685	3400	3040	2889	3195	3496	3576
B	4185	4075	3325	2860	2965	3770	3240	2735	2764	3460	3273	3442
A	3785	3515	3255	2815	2630	3295	2875	2630	2775	3040	2940	3152
A	3870	3780	3660	2980	2650	3250	2925	2915	2933	3277	3042	3363
B	3910	3690	3705	3050	2910	3630	2985	3130	2986	3040	2778	3123
B	3890	3695	3720	2990	2970	3315	2910	2985	2851	2635	2906	3081
A	4190	3970	4335	3350	3325	3870	3120	3015	3097	2909	2936	3628
A	4170	4070	4455	3610	3365	3460	2970	2855	2877	2834	3020	3632
B	4015	4480	4730	3805	3375	3545	3080	2810	2794	2974	2770	3805
B	4150	4755	5065	4125	3550	3740	3425	2690	2789	2810	2895	3695
A	4190	4740	5265	4415	3675	3965	3685	3030	2782	2904	3080	2798
A	4095	5075	5495	4270	3760	4010	3695	3255	2759	3118	3287	3547
B	3805	4360	4415	3870	3585	3785	4025	3300	3199	3407	3473	3572
B	4005	4225	3840	3800	3780	3780	4025	3710	3564	3616	3539	3853
A	3700	4325	3550	3455	3540	3660	3980	3705	3577	3759	3558	3673

Table II. *Differences (A–B) between the yields of pairs of half-drill strips*

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	(xi)	(xii)
	460	170	570	680	725	300	90	290	469	517	– 9	205
	–400	–560	– 70	– 45	–335	–475	–365	–105	11	–420	–333	–290
	– 40	90	– 45	– 70	–260	–380	– 60	–215	– 53	237	264	240
	300	275	615	360	355	555	210	30	246	274	30	547
	155	–410	–275	–195	– 10	– 85	–110	45	83	–140	250	–173
	40	– 15	200	290	125	225	260	340	– 7	94	185	103
	290	715	1080	400	175	225	–330	– 45	–440	–289	–186	– 25
	–305	100	–290	–345	–240	–120	– 45	– 5	13	143	19	–180
Total	500	365	1785	1075	535	245	–350	335	322	416	220	427

Since in using randomized half-drill strips it would usually be thought preferable to maintain the sandwich arrangement *A B B A* or *B A A B*, and to choose between these alternatives at random for each sandwich, Table III shows in a similar arrangement the differences in yield for the forty-eight sandwiches so obtained.

Table III. *Differences* $(A-B-B+A)$ *given by the yields of sandwiches*

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	(xi)	(xii)
	60	–390	500	635	390	–175	–275	185	480	97	–342	– 85
	260	365	570	290	95	175	150	–185	193	511	294	787
	195	–425	– 75	95	115	140	150	385	76	– 46	435	– 70
	– 15	815	790	55	– 65	105	–375	– 50	–427	–146	–167	–205
Total	500	365	1785	1075	535	245	–350	335	322	416	220	427

The statistical effects of using a systematic or a randomized arrangement are now easily compared. The sum of the yields from the *A* strips is 339,535 g. That from the *B* strips is 333,660, the difference in favour of *A* being 5875. The mean yield is 336,598, so that the actual error, using the systematic arrangement, is 1·745 per cent.

The sum of the squares of the forty-eight differences is 5,538,279, and this will be equal to the sampling variance of the difference between the totals, if the sandwiches are randomized. In such a case, therefore, the standard error of random sampling is 2353·35 g., or 0·699 per cent. of the average yield. The actual error of the systematic arrangement is thus nearly 2·5 times as great as the standard error obtainable by a randomized experiment of the same scope.

The analysis of variance of the results of the systematic trial is as follows:

	Degrees of freedom	Sums of squares	Mean square
Varieties	1	719,076	719,076
Estimated error	47	4,819,203	102,536
Total	48	5,538,279	

The standard error of the difference between the total yields, as estimated from the experiment, is 2218·50, or 0·659 per cent. It will be noticed that, not only has the systematic experiment the higher *real error*, but that it yields a lower *estimate of error* than the randomized experiment. The test of significance is vitiated for both reasons. Moreover, if, as "Student" appears to expect, the real errors of the systematic arrangement had been lower than those of randomized arrangement, it is evident that the estimate of error would have been correspondingly raised, so that such experiments would be less sensitive than random experiments in detecting doubtfully significant differences. An inaccurate estimate of error is a disadvantage in whichever direction it is biased.

III. Randomized pairs

It is interesting for comparison to examine the results of randomizing not sandwiches of four half-drill strips, but pairs of half-drill strips. This would, of course, be usually expected to be a less favourable form of randomization. We have, however, seldom so good an opportunity of examining the exact advantage of the sandwich arrangement.

The sum of the squares of the 96 values in Table I is 9,387,099. The standard error of the differences in total yields is, therefore, 3063·84, or 0·912 per cent. The randomized pairs of plots have only 77 per cent. of the efficiency of the randomized sandwiches. This emphasizes the value of the current opinion.

For the systematic experiment the analysis of variance, using pairs, is as follows:

	Degrees of freedom	Sums of squares	Mean square
Varieties	1	359,538	359,538
Estimated error	95	9,027,561	95,027
Total	96	9,387,099	

The standard error for the systematic experiment is again under-estimated, being 3020·35 g., or 0·897 per cent. It should be noticed that when a systematic experiment has been carried out, there is no more reason for estimating the error from pairs than from sandwiches. Different workers might, with equal justification, arrive at the estimate 0·897 per cent., or at the estimate 0·659 per cent. Neither has, in fact, any objective justification. Numerous alternative estimates can be equally suggested, as, for example, by "Student", who from 96 pairs of half-drill strips proposes to use 94 degrees of freedom for error, i.e. those used above less one representing "fertility slope". The true error, however, remains unaltered, in this case at 1·745 per cent. When randomization has been practised there is no such ambiguity. When pairs have been randomized they must be used in the estimate of error; when sandwiches have been randomized they supply the only justified basis. When the whole arrangement is systematic no valid estimate is possible, and any estimate arrived at is due to the arbitrary choice of the estimator.

IV. "Student's" test of significance

Since the uniformity trial under discussion is more extensive than most practical tests, and achieves a higher real precision than is usually attained, it may conveniently be subdivided into six minor experiments. To do this, and to minimize the effect of neighbourhood end to end, we may take together the first and seventh series, the second and

Table IV. *Values of "Student's" t observed compared with their theoretical distributions*

Values of t	Sandwiches, $n=7$		Values of t	Pairs, $n=15$	
	Number expected	Number observed		Number expected	Number observed
$-\infty$ to $-1·415$	0·6	—	$-\infty$ to $-1·341$	0·6	—
$-1·415$ to $-0·896$	0·6	—	$-1·341$ to $-0·866$	0·6	—
$-0·896$ to $-0·549$	0·6	—	$-0·866$ to $-0·536$	0·6	—
$-0·549$ to $-0·263$	0·6	—	$-0·536$ to $-0·258$	0·6	—
$-0·263$ to 0	0·6	—	$-0·258$ to 0	0·6	—
0 to $+0·263$	0·6	1	0 to $+0·258$	0·6	1
$+0·263$ to $+0·549$	0·6	—	$+0·258$ to $+0·536$	0·6	—
$+0·549$ to $+0·896$	0·6	2	$+0·536$ to $+0·866$	0·6	3
$+0·896$ to $+1·415$	0·6	1	$+0·866$ to $+1·341$	0·6	1
$+1·415$ to $+\infty$	0·6	2	$+1·341$ to $+\infty$	0·6	1
Total	6	6		6	6

eighth, etc., so as to make six experiments each with 32 half-drill strips. For each of these six experiments "Student's" t was calculated as a test of significance, and the values observed are compared with the theoretical distribution given by "Student". In both trials the six values of t are all positive, whereas in theory they should with equal frequency be positive and negative. The positive bias of the values is evidently sufficient to ruin the exactitude of the test of significance for which we are indebted to "Student".

Summary

1. This enquiry was carried out to test the truth of the opinion expressed by "Student" that randomization achieves its object "usually at the expense of increasing the variability when compared with balanced arrangements", and that one of the means available to experimenters of reducing the error is by adopting "a regular balanced arrangement".

2. Using an extensive uniformity test it is found that the arrangements randomizing either pairs or sandwiches of half-drill strips give smaller errors than the systematic arrangement advocated as more precise.

3. As a consequence experimenters using the systematic arrangement systematically underestimate their errors.

4. The error estimated from a systematic arrangement is ambiguous, and the experimenter has an arbitrary choice between several widely different estimates.

5. Owing to the failure to furnish a valid estimate of error, "Student's" test of significance is not approximately correct for systematic arrangements.

REFERENCES

"Student" (1936). "Co-operation in large-scale experiments." *J. roy. statist. Soc.* Supplement.

Tedin, O. (1931). "The influence of systematic plot arrangement upon the estimate of error in field experiments." *J. agric. Sci.* **21**, 191–208.

Wiebe, G. A. (1935). "Variation and correlation in grain yield among 1500 wheat nursery plots." *J. agric. Res.* **50**, 331–57.