STATISTICAL NOTES FOR AGRICULTURAL WORKERS.

NO. 14.—THE USE OF RANDOM SAMPLING NUMBERS IN AGRICULTURAL EXPERIMENTS.

BY

P. C. MAHALANOBIS,

Statistical Laboratory, Presidency College, Calcutta.

(Received for publication on 28th June 1933)

(With three text-figures)

We often receive enquiries regarding the most convenient method of collecting random samples or randomizing the lay-out of plots in field trials. The old method of drawing tickets from a bag or urn can, of course, be always used. But the process is extremely laborious, and in practice it is almost never possible to shuffle the tickets adequately between successive draws. In order to get over the difficulty of random sampling, Prof. Karl Pearson suggested some time ago to Mr. L. H. C. Tippett that the system of tickets might be replaced by a random system of numbers. Over 10,000 sets of 4 random numbers arranged by Mr. Tippett in 26 pages was published in 1927 (Tracts for Computers No. XV, Cambridge University Press). It contains a valuable introduction by Prof. Pearson and is practically indispensable in a statistical laboratory. But as all agricultural field workers do not have access to this tract, I am giving here a short list of 2,000 random numbers arranged in 500 sets of 4, which were obtained by taking random samples from Tippett's numbers and then re-arranging (*i.e.*, shuffling) them again in a random manner. The following examples will indicate some of the ways in which these random numbers may be used in agricultural experiments.

Example 1.—Let us suppose there are 100 plants arranged in 10 rows and 10 columns in a particular field experiment as shown in Fig. 1. It is desired to select a random sample of five plants.

—	1	2	3	4	5	6	7	8	9	0
1	*	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*	*	*	*
8	*	*	*	*	*	*	*	*	*	*
9	*	*	*	*	*	*	*	*	*	*
0	*	*	*	*	*	*	*	*	*	*

Fig. 1.

We can identify any particular plant by the number of the row and the number of the column in which it occurs. Thus the numbers (1,1) will represent the plant in the 1st row and the 1st column; (1,9) the plant in the 1st row and 9th column; (5,4) the plant in the 5th row and the 4th column; and so on. We can conveniently settle that (0,0) will represent the plant in the 10th row and 10th column.

It is clear that any set of two figures will represent one particular plant. Instead of using two separate figures, we can also use a single number of two figures provided we adopt a convention that the first figure will represent the row and the second figure will represent the column (or *vice versa*). Any random number of two figures will then represent a plant chosen in a random manner.

In Block 1 of Table I we have, for example, the following random numbers:—

2082; 1494; 7012; 0095; 6866.

We can choose 5 numbers of 2 figures each in any way we like:—vertically downwards, from the first two columns: 21, 70, 60, 40, 08; or from the last two columns: 89, 19, 62, 42, 56; or from the 2nd and 3rd columns: 04, 80, 88, 91, 96; horizontally from first three rows: 20, 82, 14, 94, 70; from the 2nd, 3rd and the 4th rows: 14, 94, 70, 12, 00; from the 1st, 3rd and the 5th rows: 20, 82, 70, 12, 68, etc. Now each number of two figures will represent a single plant, and hence any set of 5 random numbers (of two figures each) will give us a set of 5 plants selected at random, in other words a random sample of 5 plants.

It is obvious that we can easily extend the same method to give us random samples of 6, 8, 10, 20 or any number of plants.

Example 2.—Suppose instead of 100 plants we have 10,000 plants arranged in 100 rows and 100 columns. We can now use a number of two figures to represent the number of the row, and another number of two figures to represent the number of the column. Thus the plant in the 42nd row and 37th column will be labelled by the two numbers 42 and 37; or the plant in the 3rd row and 15th column by the two numbers 3 and 15 (which may be more conveniently written as 03 and 15); the plant in the 4th row and 6th column by the numbers 04, 06; and so on, with the convention that (00,00) will represent the plant in the 100th row and 100th column. It is clear that any number of 4 figures will now suffice for identifying a particular plant, and hence any random number of 4 figures will give a plant selected at random.

Example 3.—If we have, say, 1,000 plants arranged in 10 rows and 100 columns, we can obviously use a number of 3 figures to represent a particular plant with the convention that the first figure will give the number of the row, and the last two figures the number of the column, while (000) will represent the plant in the 10th row and 100th column.

Example 4.—Suppose 1,000 plants are arranged in 50 rows and 20 columns. A number of 4 figures will again identify a plant. For example 4316 will represent the plant in the 43rd row and the 16th column. But it is clear that the number representing the row, that is the number given by the first two figures will never exceed 50; and similarly the number representing the column given by the last two figures will never exceed 20. Thus although each plant will have a particular number of 4 figures, all numbers of 4 figures will not represent particular plant. Any number greater than 5020 or any number like 4334 will not represent anything.

A slight modification in our procedure is now necessary. We can, of course, ignore all numbers which do not represent plants. But this will involve the rejection of many numbers. A better method will be to assign more than one number to each plant. For example, in this particular case we can decide that 01 and 51 will both represent the 1st row, 02 and 52 the 2nd row, 11 and 61 the 11th row, 24 and 74 the 24th row, 49 and 99 the 49th row, and 50 and 00 the 50th row. Similarly we can settle that 01, 21, 41, 61 and 81 will all represent the 1st column; 02, 22, 42, 62 and 82 the 2nd column; 11, 31, 51, 71 and 91 the 11th column; and 20, 40, 60, 80 and 00 the 20th column. There will not be any empty numbers left, so that each number will identify one particular plant (although each plant will have more than one number assigned to it). We can now use any set of random numbers of 4 figures each to represent a random sample of plants.

Example 5.—Further modifications will be necessary when the number of rows or columns is not a multiple of 10. Suppose we have 17 rows. We can now assign 5 numbers to each row. For example 1, 18, 35, 52 and 69 will all represent the 1st row; 17, 34, 51, 68, and 85, the 17th row, the rule being that only the remainder after division by 17 is to be taken into consideration (a zero remainder standing for the 17th row). But this will leave the 15 numbers 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 00 as blanks. We can, however, still continue to divide by 17 and use the remainder as before. Provided this is done, each number of 2 figures will again represent unambiguously one particular row, and we shall be in a position to use random numbers of 2 figures to specify sample rows selected in a random manner.*

* A slight inequality will however be introduced in this case. Rows 1 to 15 will each be represented by six numbers, while rows 16 and 17 will be represented by 5 numbers. The chance of occurrence of rows 16 and 17 will, therefore, be slightly less in the long run. We can better equalise the chance if we use 3 figures (*i.e.* numbers 1—999 and 000) instead of only two figures. Dividing 1,000 by 17 we get 58 as dividend and 14 as remainder. It is clear that each of the rows 1—14 will be represented by 59 figures while rows 15, 16 and 17 will have only 58 figures each. The inequality is now reduced to 1 in 58, while with two figures the inequality was 1 in 5. If we use 4 figures the chances will be still better equalised. For agricultural experiments (in which the size of samples is usually small) such refinements will not, however, be usually required. The difficulty may be avoided by rejecting all numbers above 85, so that all the rows will be represented by five numbers.

Example 6.—It is required to distribute 7 varieties (A, B, C, D, E, F, and G) to 7 plots within a "block" in a field experiment. Let us assign the numbers 1, 2, 3, 4, 5, 6 and 0 to the 7 varieties, *i.e.*, settle that the number 1 will represent A, *i.e.*, 1=A, 2=B, 3=C, 4=D, 5=E, 6=F, and 0=G. Let us use random numbers of 3 figures, and adopt the convention that only the remainder is to be taken into consideration after division by 7. Consider any particular number of 3 figures, say 725; dividing by 7 we obtain a remainder of 4; this will then represent variety D.

We can now draw 2 random numbers of 3 figures each from our plates, say 569 and 411. Dividing by 7 we obtain remainders 2 and 4. We may, therefore, proceed to allot 2=B to plot No. 4.

Plots	1	2	3	4	5	6	7
Random No.	(334)	(073)	(636)	(569)		(451)	(932)
Varieties	5=E	3=C	6=F	2=B	G	4=D	1=A

Fig. 2.

We next draw two other random numbers, say 342 and 451 which give remainders 6 and 4. We can use 4=D for plot No. 6 (because the order in which the two remainders are used is obviously immaterial). The next two random numbers are 334 and 316 with remainders 5 and 1; we can, therefore, assign 5=E to plot No. 1.

The next 2 numbers are 068, 690 with remainders 5 and 4. As we have already used both the varieties 4=D and 5=E, we ignore this set. We draw a fresh set of two numbers 073, 275 with remainders 3 and 2. We can allot variety 3=C to plot No. 2. The next 2 numbers are 014, 932 with remainders 0, 1; which enable us to assign variety No. 1=A to plot No. 7 (which will be represented by the number 0).

We have next 636, 948 with remainders 6, 3; we assign variety No. 6=F to plot No. 3. This leaves us variety G for plot No. 5.

The plots are now effectively randomized. It will be noticed that the whole procedure is extremely simple and quick; it only requires the use of a set of random numbers.

Example 7.—It is desired to select a random sample of plants from an experimental plot. The plants were not arranged in rows and columns, or in any regular manner, but were sown in a haphazard way.

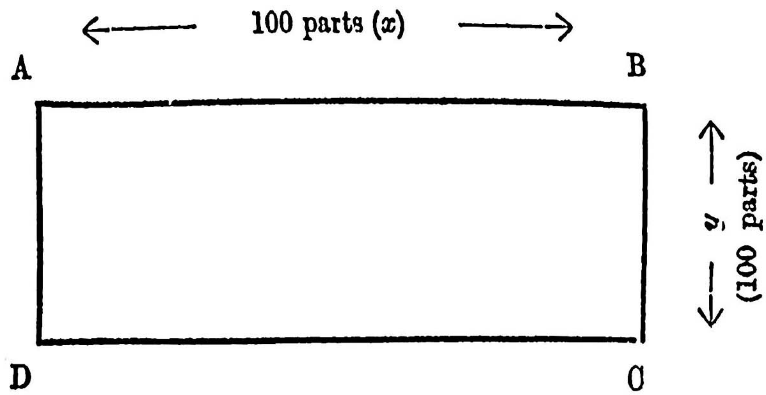

Fig. 3.

Let A, B C, D, be the four corner points of the plot (which may be square or rectangular in shape). We can divide all the sides AB, BC, CD and DA into 100 segments, and draw (on a plan) lines through each segment parallel to the sides. The whole plot will then be divided into 100 × 100 = 10,000 small cells (each of which will be of the same shape as the plot itself). We can consider the segments along AD (or BC) to give the successive rows, and the segments along AB (or CD) the successive columns. Each small cell can then be labelled by the number of the row (or segment along AD or BC) and the number of the column (or segment along AB or CD) in which it occurs.

Thus, as in Example 1, each small cell will be represented by a number of 4 figures; and since there are 100 × 100 = 10,000 cells, each number of 4 figures will represent a particular cell. We can, therefore, proceed to use random numbers of 4 figures each to give random selections of cells. The corresponding plants lying within these random cells will obviously furnish a random selection of plants.

If the length of one side is considerably greater than the other side, we can divide the longer side into 1,000 parts and the shorter side into 100 parts. We shall have then to use 3 figures to represent segments for the longer side, and 2 figures for the shorter side, that is numbers of 5 figures each to represent particular cells.

It will be noticed that the coarseness or fineness of the division is entirely at our choice. With 10 × 10 = 100 cells, we require only 2 figures; but the division is very coarse. But if necessary we can use $10 \times 100 = 1,000$ cells; or $100 \times 100 = 10,000$ cells; or $1,000 \times 1,000 = 1,000,000$ cells; and so on to any desired degree of fineness. It will be noticed that in the limit this process will reduce to using coordinates (x, y) to specify a point in the plot, where x will denote the distance from one end of the field AD, measured along say the side AB, and y the distance measured along the side AD from the end AB, the length of each side being taken equal to 10, or 100, or 1,000, or some other convenient multiple of 10.

These are only some of the uses to which random numbers may be put. Other examples will easily occur to field workers. Prof. Pearson's foreword to Tippett's tract should be consulted by every one having access to the tract. The random numbers may, of course, be taken backwards or diagonally or in any other way. The same set of numbers taken in different ways can thus furnish many more sets. One word of caution, however, is necessary. It is true that from the same set of numbers we can obtain a large number or combinations by taking the individual numbers in different ways, but we should not use the same set over and over again beyond a certain limit. In Example 1, we have altogether $5 \times 4 = 20$ random figures in Block 1, Table I. Now each plant requires 2 figures to specify its position. The set of 20 figures in Block 1, Table I, will therefore yield at the most 10 (since $20/2 = 10$) independent random plants. If we try to draw more than 10 samples from the same set of numbers (say from Block 1, Table I), it is clear that some of the numbers will no longer remain independent as some of the figures will occur over and over again and some bias will be thus introduced. The general rule is that the total number of random samples must not exceed the total number of sets of figures available. Provided this condition is not violated it is, of course, immaterial how the particular numbers are selected.

TABLE I.

2082	5006	2688	4629	3551
1494	5749	2330	8307	9238
7012	3114	1447	9732	0432
0095	3314	3419	5627	8113
6866	2145	0382	9555	8768
7286	1643	4390	5507	1407
2218	5373	0150	3655	8588
8774	6808	6639	2857	2889
0240	6061	3841	8302	9957
8793	6401	9600	6442	0172

TABLE I—*contd.*

9996	7519	8940	4275	0187
5901	6804	8329	6822	3275
1490	4995	6587	5403	1121
6887	3226	7515	0334	9889
4876	7103	2298	4117	7621
3392	3987	9739	2667	2732
0476	1250	0513	7847	6960
3256	0224	6925	3011	2118
6747	1073	0905	1945	4238
2330	4853	0647	3362	6904
7565	0328	3044	1912	1501
3561	7684	7741	4569	2655
3004	9895	6380	8898	4694
2103	0797	4079	8557	3068
8161	1106	4885	4649	4500

TABLE II.

0845	1501	7041	8726	8884
1382	2736	3981	9230	1195
6956	5478	5575	9360	6447
6037	5636	0310	0359	6595
0589	4927	1730	4497	8995
2102	2390	9498	3343	4428
3561	4996	8820	0496	5865
5397	2341	0297	5356	3074
7691	1365	3090	2427	7839
7872	7937	5748	0962	1434
8495	5160	9628	9465	3209
2096	4152	6965	6152	4258
1465	2245	0818	3836	3275
1498	0809	3911	1760	1693
2837	2642	2454	3508	0772
3830	1927	8163	0331	7553
9157	5933	7759	6911	8729
1079	0032	9151	1106	6038
1497	0782	2798	2649	8135
1175	7452	3581	4532	2850
0658	6910	4432	0176	6436
0668	7398	8369	4097	3844
2052	4564	0577	7022	4994
7052	9129	7152	3788	3993
3690	7895	5722	2031	5215

TABLE III.

9607	8606	2182	9039	9403
2130	2445	9289	3350	2317
4764	2839	1687	9803	3551
2557	7875	5816	6867	6090
7276	9687	1698	4807	7534
8986	4736	1074	8623	5911
1306	4733	5438	5208	3817
0674	1429	1161	9063	6104
5315	8416	3880	2542	5267
9917	9917	8756	2896	9215
2902	7844	5663	5943	6897
3641	0367	9614	4905	4874
5334	9693	4072	6084	4461
7149	0056	9681	3795	1840
1338	2004	7031	2548	4930
4130	1299	9253	1904	2169
5291	8839	4594	4246	8382
2580	8098	6867	1666	5813
2159	5827	5748	7605	8800
0386	3421	5866	3828	1400
0608	7866	1459	7561	9246
1498	3359	2707	2446	8793
7930	4024	9284	1058	7762
4105	4261	9978	1407	2538
8819	0557	1834	2450	1929

TABLE IV.

3939	7628	2056	3852	4277
1658	4695	5218	8680	5970
7201	9589	9408	2626	3254
7261	4891	5967	1457	5635
2270	7538	6340	8247	2924
9862	1627	8774	3008	7778
7262	7040	2726	8520	6217
5421	5619	5601	9904	1571
0572	3791	1652	2355	6299
5021	4593	2307	1350	1592
8078	4293	0315	1375	1275
3256	0262	6259	7778	4374
9004	5223	9682	8131	0909
9334	1992	8646	7421	1581
7639	7842	4939	3601	0771
8419	8457	5751	3389	5751
4920	5006	1541	5122	1541
5699	1717	3666	5655	3666
9310	9079	9821	7705	9821
6561	6467	5851	8065	5855
0523	1698	5996	4326	7217
8253	5667	9590	0778	2422
6095	3442	1173	9072	8365
1807	9825	7566	9676	9153
1920	0481	9517	1276	4466