ON SELECTING RANDOM NUMBERS FOR LARGE-SCALE SAMPLING

By ABRAHAM MATTHAI
Indian Statistical Institute, Calcutta

A proper choice of random numbers from any of the standard random number tables can be done in one of many ways without any cost considerations, when the number of figures to be selected is small. In large-scale work, however, a certain amount of planning and proper methods for choosing the numbers are of importance.

Extensive random sampling number tables are now available to meet the requirements of almost any practical large-scale investigation; such as, those by Tippett, Kendall and Babington-Smith, Fisher and Yates, and the Calcutta Statistical Tables. It may not be an uncommon experience that in using any of these tables, most selectors differ in their methods of selecting numbers, without necessarily giving regard to the labour involved in the method adopted, though both of them do not commit any error in giving equal chance to all units concerned.

When large-scale sampling methods are becoming of more and more use every day, time and money-saving methods of selecting have their place. In this paper certain methods and details in the matter of selecting random numbers, on a large-scale, are given.

Suppose now that a random sample of 800 is to be selected from a total of about say 70000 individual units. This would involve the selection of 800 numbers from among the numbers 1—70000 or the numbers 0—69999 (00000 being reckoned as 70000). One method of selection would be to pick out five digited numbers from the random number tables omitting all numbers 70000 and above. This would entail rejection of about three in ten of the numbers met with.

An alternate method would be to take four digited numbers from the tables and prefix to each a fifth digit that is random in the set 0, 1, ..., 6. This might suggest the preparation of random number tables of one digited figures for the different sets 0-1, 0-2, ..., 0-9. But such an auxiliary table would seem unnecessary in view of the extensiveness of random number tables available now.

What can be done then is to assign, say, ten digits in the table, to give a safe margin, for every five-digited random number to be selected, and to take the last four digits prefixed with the *first* figure less than 7 met with in the remaining set of six digits proceeding towards the left. Thus, if in the tables, the following are three consecutive rows of ten-digited figures:

67	6345	0912
72	3419	1256
25	2987	4391

the three random numbers less than 70000, provided by the above procedure would be 50912, 11256 and 24391. Obviously the selection is done this way with almost no trouble of rejection.

In order just to verify the appropriateness of such a selection from the existing random number tables, a test was made on 898 five-digited random numbers so selected from some of the random number tables mentioned earlier. Table 1 below gives the joint distribution of the fifth and the fourth digits of these numbers.

TABLE 1

		fourth digit										
		0	1	2	3	4	5	6	7	8	9	total
fifth digit	0	22	15	8	13	6	9	11	13	19	9	125
	1	11	13	13	6	6	13	14	14	10	9	109
	2	16	8	14	14	14	16	14	10	17	16	139
	3	16	5	11	17	13	11	11	14	13	12	123
	4	11	13	14	21	14	8	14	12	18	17	142
	5	11	19	11	14	17	9	13	14	15	21	133
	6	10	13	12	12	16	15	16	10	15	8	127
	total	97	77	83	97	86	81	91	87	107	92	898

The χ^2 with 69 degrees of freedom calculated on the basis of the expected value, $(898 \div (10 \times 7)$, of the internal cell frequencies works out to be $\frac{(22)^2+(15)^2+\ldots+(8)^2}{898 \div 70} - 898 = 68.125$. This χ^2 comprises deviations of the fourth digit figures, of the fifth-digit figures and of the two-digit combinations, from their respective expectations, The χ^2 relating to the fifth digit figures is $\frac{(125)^2+(109)^2+\ldots(127)^2}{898 \div 7} - 898 = 5.748$ and that of the fourth digit figures similarly is 7.069. We then have the following analysis of χ^2 table.

TABLE 2

due to	d.f.	χ^2	probability (approx.)
fifth digit	6	5.748	50%
fourth digit	9	7.069	60%
patchiness (combination of the digits).	54	54.408	93%
total	69	68.125	

The probabilities of the χ^2's are low enough to regard the numbers selected to be satisfactory.

If the problem is to select a set of random numbers less than any number, say 2853, not necessarily a round number like 70000, a procedure that is often adopted is to select four-digited figures from a table omitting all figures 2853 (0000 would stand for 2853) and above that are come across, or alternatively to divide the numbers by 3000 and note the remainders, rejecting those from 2853 to 2999 and also leaving the numbers above 9000 out of consideration.

A good amount of rejection labour can be saved if the following method is adopted, namely, to choose three digited figures prefixed as in the former case by a fourth digit obtained as the first figure met with on the left which belongs to the set 0, 1, and 2, taking care however to reject cases when the four digited figure formed happens to be 2853 or above.

Thus starting with the last column of 12 digits of random numbers (I) of Fisher and Yates Tables, the first few rows of which are reproduced below we select as the required random numbers the figures shown in bold type. The last case where the figures shown in italics will be come across, is to be rejected

...	...	...	...	45		60	11	14	10	**95**
...	...	...	...	51		24	51	79	89	**73**
...	...	...	...	59		88	07	54	14	**10**
...	...	...	...	21		88	26	49	81	**76**
...	...	...	...	53		23	83	01	30	**30**
...	...	...	...	43		84	26	34	03	**64**
...	...	...	...	25		83	91	12	06	**76**
...	...	...	...	07		44	39	*52*	*38*	*79*
...	...	...	...	...		...	...	...	...	...
...	...	...	...	...		...	...	...	...	...

It will be seen that this procedure brings about considerable reduction in the amount of rejection, but involves some rejection when the figure is not a round number. The expected rejection in the above example will be $\frac{1}{(2+1)}\left(\frac{1000-853}{1000}\right)=4.9\%$ as compared to 71.5% in the procedure of rejecting all numbers 2853 and above and 14.4% in the procedure of division by 3000 and taking remainders. It will also be seen that the rejection will be more, $\frac{1}{8}\left(\frac{844}{1000}\right)=10.6\%$ if the number were say, 7156 and still more, $\frac{1}{2}\left(\frac{868}{1000}\right)=43.4\%$ if the number was 1132. But it will be observed that even in the extreme case it will be 50% as compared to a maximum rejection of 90% for the alternative procedure. Also it will be noted that a chief merit of this method is that it dispenses with the labour of dividing numbers.

As regards the above methods which require reserve figures to be kept, it will be observed that they do not necessarily involve, as compared to the alternatives, a reduction in the total number of random numbers suitable for selection, available in any table. But the need for a proper choice of the number of digits to be reserved is obvious. The smaller the range of the prefixing set the more the number of digits to be reserved.

We may now touch upon certain other aspects of large-scale selection of random sampling numbers. Usually a team of workers will have to be engaged to select the numbers and any possible overlapping or duplication is to be avoided by allotting

separate sets of pages of the different random number tables to different workers. So also, in spite of the fact that the tables have passed tests of randomness before their publication, it would be worth while as a safety measure, to subject random numbers selected for any investigation, to at least a rough check for randomness, to ensure that a bad set has not been selected either through chance possibilities or due to unconscious errors on the part of the selector.

When the random selection in question is to be made from a set represented on punched-cards, the selection is best made on a Collator with the help of a set of random number cards on which their serial numbers together with a suitable range of random numbers from any of the standard tables have been punched and kept. An account of the use of the punched card machines for random selection, useful in large-scale work, is given by Vickery (1939).

REFERENCES

FISHER, R. A. and YATES, F. (1938): *Statistical Tables for Biological, Agricultural and Medical Research*, Edinburgh.

KENDALL, M. G. and BABINGTON-SMITH, B. (1939): *Random Sampling Numbers, Tracts for Computers XXIV*, Cambridge University Press.

Statistical Laboratory, Calcutta (1941): Random Sampling Numbers, Calcutta Statistical Tables (1).

TIPPETT, L. H. C. (1927): *Random Sampling Numbers, Tracts for Computers XV*, Cambridge University Press.

VICKERY, C. W. (1939): On drawing a random sample from a set of punched cards. *J. Roy. Stat. Soc., Suppl.*, 6.

Paper received : August, 1952.