ANALYSIS OF THE 1939 MODEL SAMPLE SURVEY RESULTS FROM THE VIEWPOINT OF INTEGRAL GEOMETRY

By MOTOSABURO MASUYAMA
Tokyo University

[*Editorial Note :* When exploratory work on the sample survey of the area under jute in Bengal started in 1937, it became necessary to make rough calculations about the cost of operations which in many respects depended on the number of fields or plots requiring to be enumerated for each sample unit. Working with sample units of square shape but of different sizes, it was clear that the average number of fields or plots which would lie either partly or wholly within a sample unit of a given size would increase with the size. On the basis of some experimental observations I found that a very simple graduation formula of the type: $p = 2A + 4\sqrt{A}$ where 'A' is size of the sample unit in acres. As the average size of a plot was something less than half acre in Bengal, I thought two such plots on an average would probably lie within the sampling unit while the number of plots which would cut the boundary of the sample unit would be proportional to the length of the perimeter. This was the basis of my formula which gave reasonable values for a wide range of size of sample units. I am glad M. Masuyama has now investigated the theoretical basis and J. M. Sen Gupta has collaborated with him on the experimental side. — P. C. Mahalanobis]

1. When I told Professor P. C. Mahalanobis about the outline of my previous paper published in this Journal[3], he said that he had obtained a similar formula empirically. He had used a square, instead of the circle in our case, as a moving oval in his survey of crop-area in Bengal in 1939. The present paper is a detailed analysis of his unpublished data from the viewpoint of integral geometry with a new model experiment. These unpublished data were kindly supplied by his colleague Mr. J. M. Sen Gupta.

Let the expected number of plots which are completely or partially included in a grid, i.e. a movable square, of size $A=a^2$ be p; and let T be the total area covered by the survey in question, ϕ the total area of plots, λ the total length of their perimeters, and ν the total number of plots in T. Then, as was shown in my previous paper, we have

$$p = \frac{\phi}{T} + \frac{2\lambda}{\pi T} a + \frac{\nu}{T} a^2 \qquad \ldots (1)$$

when the union of each grid and each fixed plot is contained in T. The convexity of each plot is assumed herein. Otherwise, we should consider an extended area T' of T. The method of constructing T' is given in § 4 of the previous paper. However, if the total area surveyed is sufficiently large and not too narrow in any direction we need not consider, for practical purposes, such an extended area T'.

Mahalanobis's intention was to estimate p for given size of grid $A = a^2$, the coefficients being obtained empirically. He thought that this would be quite valuable for estimating the survey cost, so that the explicit general form of each coefficient was not given at that time.

2. The following figures were obtained in the 1939 experiment:

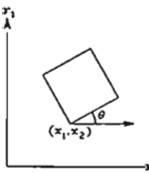
TABLE 1. NUMBER OF PLOTS PER GRID FOR DIFFERENT GRID-SIZES

size of grids in acres	no. of grids	no. of plots included per grid		
		fully	partly	total
1	7687	0.58	6.71	7.29
4	2260	4.52	13.55	18.07
9	1521	13.91	20.87	34.78
16	529	26.81	27.90	54.71
36	314	75.00	42.18	117.18

The grids for different sizes were independently sampled and not nested.

The population in this experiment was all the plots on map in 31 police stations scattered over West Bengal, India, but not at random. This non-randomness has nothing to do with the following conclusions.

However, in this 1939 experiment, only the coordinates of one vertex of the moving square, i.e. grid, were selected at random but not the orientation, i.e. in the kinematic density of this moving oval $dO_2 = dx_1 dx_2 d0$, 0 remained unchanged.



This may affect the validity of the conclusion.[4] The next practical restriction was that one digit in a table of random sampling numbers (00—00 or something like this) corresponded to a distance of 396 m. on the ground, because we could not use the random numbers for sampling from a continuum.

At any rate, in this experiment it was found from actual measurements that

$$\phi = 1{,}627{,}488 \text{ acres} \doteq 1.6275 \times 10^6 \text{ acres} \quad \ldots (2)$$

and

$$\nu = 4{,}207{,}410 \text{ plots} \doteq 4.2074 \times 10^6 \text{ plots} \quad \ldots (3)$$

3. If we now neglect the footpaths and the so-called "boundary" effect, we may put

$$\phi = T \quad \ldots (4)$$

in this case, so that we have from (1), (2), (3) and (4),

$$p = 1 + 3.9117 \times 10^{-7} \lambda a + 2.5852 a^2 \quad \ldots (5)$$

in the acre-scale.

As a_i and the corresponding p_i are given in the Table 1, we can estimate λ for each i. The result is given in Table 2.

TABLE 2. THE PERIMETERS OF PLOTS FOR DIFFERENT GRID-SIZES

a	$p - \frac{\phi}{T} - \frac{\nu}{T} a^2$	$\frac{2\lambda}{\pi T}$	λ
1	3.68	3.68	9.408×10^6
2	6.70	3.40	8.692×10^6
3	10.65	3.55	9.075×10^6
4	12.31	3.08	7.874×10^6
6	23.00	3.83	9.781×10^6

The agreement between the independent estimates of λ seems to be not quite satisfactory, but not bad either, for such large samples.

In any way, taking the mean of the first three λ's, i.e. only the mean of estimates of the larger samples we have

$$\lambda = 9.058 \times 10^6 \sqrt{\text{acre}} = 5.762 \times 10^3 \text{ m.} \quad \ldots (6)$$

The formula (5) will be put in the following form

$$p = 1 + 3.543\, a + 2.585\, a^2 \quad \ldots (7)$$

In this empirical equation, the first figure 1 may be negligibly small for large a, say $a \geqslant 2$. Then we have an approximation

$$p = 3.5\, a + 2.6\, a^2 \quad \ldots (8)$$

which is nothing but the Mahalanobis's empirical formula quoted in the previous paper with slight differences in its numerical coefficients, but these differences do not matter for a rough estimation of p.

4. The mean perimeter will be

$$\frac{\lambda}{\nu} = 2.153 \sqrt{\text{acre}} = 137.0 \text{ m.} \quad \ldots (9)$$

Writing the relation between the area of the α-th plot F_α and its perimeter L_α as

$$F_\alpha = CL_\alpha^2 + \epsilon_\alpha, \qquad \sum_{\alpha=1}^{s} \epsilon_\alpha = 0 \quad \ldots (10)$$

we introduce the shape factor C.

Then by the isoperimetric inequality

$$C \leqslant 1/4\pi. \quad \ldots (11)$$

To estimate the actual value of C, a new experiment was performed by Mr. J. M. Sen Gupta in January, 1953. Some results are shown below:

TABLE 3

character	1st sample		2nd sample	
	$\bar{x}$	s.d.	$\bar{x}$	s.d.
area per plot in acres	0.4234	0.6253	0.3850	0.4673
perimeter per plot in $\sqrt{\text{acre}}$	2.5059	1.7908	2.3100	1.4709
C	1/22.5		1/19.5	

Two independent random samples were drawn with equal probability from a compact area of about 9 square miles in the police station, Memari District, Burdwan, West Bengal. The size of each sample was 100.

From these observations we may assume

$$C = 1/21 \quad \ldots (12)$$

which is naturally slightly smaller than the shape factor of the square, i.e. $C = 1/16$.

Then the variance of perimeters will be

$$\sigma_L^2 = \frac{\phi}{C\nu} - \left(\frac{\lambda}{\nu}\right)^2 = 3.488 \text{ acres} \qquad \ldots (13)$$

or

$$\sigma_L = 1.868\sqrt{\text{acre}} = 118.8 \text{ m.} \qquad \ldots (14)$$

which is nearly equal to the actual observations in Table 3.

The variance of plot area is approximately equal to

$$\sigma_F^2 = 4C\left(\frac{\phi}{\nu}\right)^3 \sigma_L^2 = 0.2570 \text{ acre}^2 \qquad \ldots (15)$$

or

$$\sigma_F = 0.5070 \text{ acre} \qquad \ldots (16)$$

which is also nearly equal to the actual observations in Table 3.

The shape factor C plays an important role in a rapid method of estimating mean area, because we can estimate this mean area by merely measuring perimeters, which is easier than measuring corresponding areas.

The formula (15) shows at a glance that the information about the parameter λ in our integral geometrical method seems to be a nuisance but actually it is not so.

5. We note also that the figures in the second column of Table 2 are nearly equal to half of the figures in the fourth column of Table 1. That it should be so can be shown from the view point of integral geometry.

According to Poincaré's formula[1, 4] which is a special case of the Blaschke—Santaló's kinematic principal formula, we have $\int n dO_2 = 4L_1L_2$ ... (17)

where n is the number of intersection points of two curves, one of them say, of length L_1, being fixed and the another of length L_2 movable. In our case the boundary of each plot and that of grid are closed curves so that if we put $n/2 = \omega$, then the expectation of the number of plots partly included in a grid, i.e. on the boundary of plots is seen to be equal to

$$\frac{\int \omega dO_2}{2\pi T} = \frac{2L_1L_2}{2\pi T^2}\omega = \frac{4\lambda a}{\pi T} \qquad \ldots (18)$$

In this case if the number of points of intersection of a plot and a grid be n, it should be counted $n/2 = \omega$ times but not once.

TABLE 4. AGREEMENT BETWEEN THEORETICAL AND OBSERVED NUMBER OF PLOTS PARTLY INCLUDED IN GRIDS OF DIFFERENT SIZES

a=side of grid	observed number	$\frac{4\lambda a}{\pi T}$
1	6.71	7.36
2	13.55	13.68
3	20.87	21.30
4	27.90	24.62
6	42.18	46.00

From Table 4 it can be seen that the agreement between theory and observations is fairly good for large samples, except for the case for $a = 1$.

6. Let us suppose now that ϕ, λ and ν are all unknown but a_i and p_i are known. Then we can estimate these three unknowns as solutions of linear simultaneous equations

$$A_i = p_i T = \phi + \frac{2\lambda}{\pi} a_i + \nu a_i^3, \qquad i = 1, 2, 3. \qquad \ldots \quad (19)$$

The solution is

$$\phi = A_2 - 3(A_3 - A_1) \qquad \ldots \quad (20)$$

$$\lambda = \frac{\pi}{4a}(8A_2 - 5A_1 - 3A_3), \qquad \ldots \quad (21)$$

and

$$\nu = \frac{1}{2a^3}(A_1 + A_3 - 2A), \qquad \ldots \quad (22)$$

where we put $a_1 = a$, $a_2 = 2a$ and $a_3 = 3a$.

The actual solutions and true values are given in Table 5 below. The agreement is not quite satisfactory.

TABLE 5

	ϕ	λ	ν
true value	1.627×10^4	9.058×10^4	4.207×10^6
estimate	2.181×10^4	5.024×10^4	4.826×10^6

Comparing this table and the formulae (20), (21) and (22), we could not help suspecting the creeping in of some biases in the method of counting plots. A_1 or A_3 seems to be relatively over-estimated and A_2 relatively under-estimated, i.e. p_1 or p_3 seems to be over-estimated and p_2 under-estimated.

7. There may be various sources of bias. To know these sources, we shall analyse at first the number of plots which are fully included in a grid.

We know that this number is equal to

$$\frac{\phi}{T} + \frac{\nu}{T} a^2 - \frac{2}{\pi T} \lambda a = p - \frac{4\lambda a}{\pi T} \qquad \ldots \quad (23)$$

Putting $\frac{2\lambda}{\pi T} = f$, we have the following independent estimates of f for $a = 1$, 2, 3, and 4.

$$\left.\begin{aligned} 3.585 - f &= 0.58, & f &= 3.01 \\ 11.341 - 2f &= 4.62, & f &= 3.42 \\ 24.267 - 3f &= 13.01, & f &= 3.44 \\ 42.363 - 4f &= 26.81, & f &= 3.88 \end{aligned}\right| \qquad \ldots \quad (24)$$

There is a tendency that f increases approximately linearly with increasing values of a (c.f. the 3rd column, Table 2.)

It seems that the number of plots fully included in a grid is relatively over-estimated for smaller a, that is to say, some of the plots which should be estimated as plots partly included may have been counted as fully included. Of course, this is only one of the many possible interpretations. It is easily seen that this sort of bias

may occur if we count as one plot completely included such a plot which is in a grid that it has only one point or only one line, may be sometimes two points or two lines or even more, in common with the boundary, because the actual points has non-zero area and the actual line has non-zero width on map. In the 1939 experiment, this sort of bias was not taken into consideration, because the purpose was not to check the integral geometrical formulae. Then because of the above reason, the number of plots partly included in a grid must have been under-estimated.

To remove this sort of bias we should mark two sides of a grid by colour say red and the plots which have one point or one side in common with these red sides should be counted always as plots partly included. If a plot has two or more points or lines in common with a grid a similar technique is advisable to reduce bias.*

REFERENCES

[1] BLASCHKE, W. (1935): *Vorlesungen über Integralgeometrie*, Heft 1, Teubner, Leipzig.

[2] CROFTON, M. W. (1868): On the theory of local probability applied to straight lines drawn at random in a plane, etc. *Phil. Trans. Roy. Soc.*, 158, 181.

[3] MASUYAMA, M. (1953): A rapid method of estimating basal area in timber survey, an application of integral geometry to the areal sample problems, *Sankhyā*, 12, 291-302.

[4] MASUYAMA, M. (1953): Rapid method of estimating the sum of specified areas in a field of given size, *Rep. Stat. Appl. Res.*, 2, 113-119.

[5] POINCARÉ, H. (1912): *Calcul des Probabilités*, Gauthier-Villars, Paris.

Paper received : January, 1953.

*This is a preliminary report to test the possibility of using integral geometry to sample surveys of areas, and it is intended to make new experiments for this purpose. The author offers his sincere thanks to Professor Mahalanobis and Mr. J. M. Sengupta for the facilities they gave him when he worked on this problem in the Indian Statistical Institute in 1952.