ESTIMATION OF PARAMETERS FROM INCOMPLETE DATA WITH APPLICATION TO DESIGN OF SAMPLE SURVEYS

By ABRAHAM MATTHAI
Statistical Laboratory, Calcutta

1. Introduction

Suppose that a sample from a population characterised by two measurements x, y, is such that, for some of the individuals, only one or the other of the two measurements is available. Thus if $N = n+n_1+n_2$ be the total number of individuals observed, then n_1 of them may provide the first measurement alone, n_2 the second alone and n both. If the two measurements are correlated then it is possible that the observations on either one of the measurements may throw some information on the characteristics of the other.

Assuming normal population, and starting with the simultaneous distribution of the sample of $n+n_1+n_2$ as a whole, Wilks (1932) has shown that the maximum likelihood estimates so obtained of the parameters μ_1, μ_2, σ_1, σ_2 and ρ are more accurate than the alternate estimates: $\bar{x}$ based on $n+n_1$ values for μ_1, $\bar{y}$ based on $n+n_2$ values for μ_2 and so on. Extension of this method of estimation to the case of more than two variables is given in section 2 of this paper.

Efficiency of estimates and certain details of estimation are discussed with example in section 3, and finally in section 4 certain applications to planning of sampling investigations on correlated variables are indicated with instances from crop acreage surveys.

2. Extension to more than two Variables

Without loss of generality, we shall consider here the case of three variables x_1, x_2, x_3. The sample in the trivariate case may be considered to consist of $N = n_{123}+n_{120}+n_{103}+n_{023}+n_{100}+n_{020}+n_{003}$, individuals.

Let the covariance matrix of the variables x_1, x_2, x_3 be denoted by

$$\Gamma_{123} = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{12} & V_{22} & V_{23} \\ V_{13} & V_{23} & V_{33} \end{pmatrix}$$

Similarly let Γ_{12}, Γ_{13} etc., denote the covariance matrices of (x_1, x_2), (x_1, x_3) etc., where the elements of these matrices are from Γ_{123} itself.

Let another set of matrices A_{ijk} be introduced so that a matrix like A_{103} will stand for the third order matrix obtained by adding a third middle zero column and row each to the second order matrix Γ_{13}^{-1} (the inverse of Γ_{13}) and multiplied by n_{103}.

Now, assuming normal distributions for the variables, the logarithm of the likelihood of the sample can be written as,

$$\begin{aligned}L=\text{const}-\tfrac{1}{2}\{(X_{123}-\mu)A_{123}(X'_{123}-\mu')&+(X_{120}-\mu)A_{120}(X'_{120}-\mu')+(X_{103}-\mu)A_{103}(X'_{103}-\mu')\\ &+(X_{023}-\mu)A_{023}(X'_{023}-\mu')+(X_{100}-\mu)A_{100}(X'_{100}-\mu')\\ &+(X_{020}-\mu)A_{020}(X'_{020}-\mu')+(X_{003}-\mu)A_{003}(X'_{003}-\mu')\end{aligned}$$

where the matrix $\mu=(\mu_1, \mu_2, \mu_3)$ and a matrix like X_{023} stands for $(0, \bar{x}_2, \bar{x}_3)$, $\bar{x}_2$ and $\bar{x}_3$ here being based on the n_{023} observations.

For the estimation of the μ's, differentiating L with respect to μ_1, μ_2, μ_3 and equating to zero we have

$$\left\|\begin{matrix}\dfrac{\partial L}{\partial\mu_1}\\[2mm] \dfrac{\partial L}{\partial\mu_2}\\[2mm] \dfrac{\partial L}{\partial\mu_3}\end{matrix}\right\| = \left\|\begin{matrix}0\\[2mm] 0\\[2mm] 0\end{matrix}\right\| = \begin{aligned}&(X_{123}-\mu)A_{123}+(X_{120}-\mu)A_{120}+(X_{103}-\mu)A_{103}\\ &+(X_{023}-\mu)A_{023}+(X_{100}-\mu)A_{100}+(X_{020}-\mu)A_{020}\\ &+(X_{003}-\mu)A_{003}.\end{aligned}$$

These maximising equations can then be rewritten as

$$\mu A = \Sigma(X_{ijk}A_{ijk})$$

where A represents the sum of all the A_{ijk} matrices.

We thus have

$$\mu = \Sigma(X_{ijk}A_{ijk})A^{-1} \qquad \dots \quad (2.1)$$

giving the estimates $M=(m_1, m_2, m_3)$ when Γ_{ij}'s are known.

The information matrix relating to the maximum likelihood estimates of the μ's is given by

$$I = -E\left\|\frac{\partial^2 L}{\partial\mu_i\partial\mu_j}\right\| = A \qquad \dots \quad (2.2)$$

so that the dispersion matrix is A^{-1}. ... (2.3)

For the estimation of A, the maximum likelihood estimate will be complicated. But for practical purposes estimates of variances and covariances in the modified form suggested at the end of section 3 can however be satisfactorily used.

3. EFFICIENCY OF ESTIMATES

Before considering practical applications, it will be useful to make a study of the degree of efficiency of the estimates.

In the case of two variables, if m_1 represents the estimate of μ_1 then it has variance (Wilks, 1932) :

$$V(m_1) = \frac{n+n_2(1-\rho^2)}{(n+n_1)(n+n_2)-\rho^2 n_1 n_2}\ \sigma_1^2$$

An estimate m'_1 of μ_1 obtained as the arithmetic mean of the $n+n_1$ observations on the first measurement has the variance :

$$V(m'_1) = \frac{\sigma_1^2}{(n+n_1)}$$

so that the efficiency of the estimate m'_1 is

$$\frac{V(m_1)}{V(m'_1)} = \left(1-\frac{n_2}{n+n_2}\rho^2 \right) \div \left(1-\frac{n_1 n_2}{(n+n_1)(n+n_2)}\rho^2 \right)$$

The efficiencies of the estimate m'_1 for certain values of ρ and certain proportions $n : n_1 : n_2$ are given in Table 1.

TABLE 1

PERCENTAGE EFFICIENCIES OF m_1'

values of				efficiency of m_1'	values of				efficiency of m_1'
n :	n_1 :	n_2	ρ		n :	n_1 :	n_2	ρ	
all	values		0	100.00	10 :	3 :	1	0.7	98.53
all	values :	0	all values	100.00					
0 : all	values		all values	100.00	10 :	3 :	3	0.7	89.61
10 :	1 :	3	0.3	98.11	10 :	0 :	3	0.7	88.09
			0.5	84.73					
			0.7	90.02	1 :	3 :	2	0.7	89.18
			0.9	82.71					
			1.0	78.57	1 :	3 :	3	0.7	87.32
10 :	1 :	15	0.7	72.54	1 :	2 :	3	0.7	83.77

To take an actual example, Münter (1936) gives data in which out of 187 skeletons of Anglosaxon males, 106 provide maximum lengths of both right and left femora, 47 of the right femur alone and 34 of the left alone.

In this example,

$$n = 106, \qquad \Sigma x = 49494.7, \qquad \Sigma x^2 = 23160979.61,$$
$$\Sigma y = 49443.0, \qquad \Sigma y^2 = 23210233.32,$$
$$\Sigma xy = 23187716.93,$$
$$n_1 = 47, \qquad \Sigma_1 x = 21385.5, \qquad \Sigma_1 x^2 = 9752550.23,$$
$$n_2 = 34, \qquad \Sigma_2 y = 15643.3, \qquad \Sigma_2 y^2 = 7224807.89.$$

The estimates $\hat{\theta}_i$ of the five parameters θ_i: μ_1, μ_2, σ_1, σ_2 and ρ, obtained by solving the five maximising equations, $\frac{\partial L}{\partial \theta_i}=0$, with their standard errors are given in Table 2. The set of estimates θ_i' given by the means, standard deviations and the product moment coefficient of correlation are also shown in the table. The figures given in bracket with each estimate is the number of observations on which the estimate is based. In column (4) of the table is given n' which is the number of observations required for the estimate θ_i' to attain the precision of $\hat{\theta}_i$.

TABLE 2

Sets of estimates and their relative accuracy*

parameter	estimates with standard errors						n'
	$\hat{\theta}_i$	s.e.		θ'_i	s.e.		
(1)		(2)			(3)		(4)
μ_1	462.7	1.73	(187)	463.3	1.81	(153)	185
μ_2	462.0	1.75	(187)	465.7	2.06	(140)	184
σ_1	23.6	1.23	(187)	22.4	1.28	(153)	183
σ_2	23.8	1.25	(187)	24.4	1.46	(140)	181
ρ	0.0786	0.0036	(187)	0.0835	0.0032	(106)	131

In the application of this kind of estimation a word of caution, however, is necessary with regard to the estimate of ρ. In practice, the estimate of ρ obtained from all the five maximising equations together, may sometimes prove to be unreliable, depending on the type of data used. To illustrate the nature of unreliability of this estimate, let us consider the quantity,

$$\frac{\left\{\Sigma(xy)-\frac{\Sigma x\,\Sigma y}{n}\right\}\Big/(n-1)}{\sqrt{\left\{\Sigma_1x^2+\Sigma x^2-\frac{(\Sigma_1x+\Sigma x)^2}{n+n_1}\right\}\Big/(n+n_1-1)}\sqrt{\left\{\Sigma_2y^2+\Sigma y^2-\frac{(\Sigma_2y+\Sigma y)^2}{n+n_2}\right\}\Big/(n+n_2-1)}} \quad \cdots \quad (3.1)$$

as a possible estimate of ρ. Now the data available may be such that the additional

*In solving the maximising equations a convenient method is to start with a set of approximate values of the estimates, and multiply the dispersion matrix by the vector $\frac{\partial L}{\partial \theta_i}$ based on these values. The resulting vector gives the corrections to the approximate values. Repeating the process with the corrected vector and dispersion matrix each time, estimates correct to the required number of decimal places can be obtained. This method has been followed to one stage of approximation to get the estimates in this example.

n_1 and n_2 values entering the denominator of the above formula can give rise to an estimate that is even greater than unity. (Such a value of the estimate might also result in the breakdown of the algebra involved). This feature of the estimate of ρ can be ascribed to lack of homogeneity or in other words absence of perfect randomness in data that occur in practice. In situations, therefore, where due to lack of ideal sampling conditions a vitiated estimate is likely, it will be advisable to estimate ρ by the product moment coefficient of correlation based on the n pairs of values and estimate the other four parameters from the four respective maximising equations.

If however the problem is to estimate the μ's only, it would suffice, especially when the sample is large, to obtain the estimates from equations $\frac{\partial L}{\partial \mu_1} = 0$ and $\frac{\partial L}{\partial \mu_2} = 0$ in which σ_1, σ_2 and ρ are equated to statistics s_1, s_2 and r respectively, given by

$$s_1^2 = \left\{\Sigma x^2 + \Sigma_1 x^2 - \frac{(\Sigma x + \Sigma_1 x)^2}{n+n_1}\right\} \div (n+n_1-1) \qquad \ldots \quad (3.2)$$

$$s_2^2 = \left\{\Sigma y^2 + \Sigma_2 y^2 - \frac{(\Sigma y + \Sigma_2 y)^2}{n+n_2}\right\} \div (n+n_2'-1) \qquad \ldots \quad (3.3)$$

$$r = \left\{\Sigma xy - \frac{\Sigma x \Sigma y}{n}\right\} \div \sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} \qquad \ldots \quad (3.4)$$

Thus in the previous example of Anglosaxon femora we have $s_1 = 22.4$, $s_2 = 24.4$, $r = .9835$. Substituting these values in the two maximising equations and solving for μ_1 and μ_2 we get an estimate of 462.4 for the right femur and 463.1 for the left femur. The variances of these estimates being given by

$$\frac{n+n_2(1-\rho^2)}{(n+n_1)(n+n_2)-n_1n_2\rho^2}\,\sigma_1^2 \text{ and } \frac{n+n_1(1-\rho^2)}{(n+n_1)(n+n_2)-n_1n_2\rho^2}\,\sigma_2^2, \quad \text{respectively,}$$

we have for their standard errors 1.64 and 1.79, the estimates being as efficient as the arithmetic means of about 184 values.

Similar considerations would apply to estimation in the case of more than two variables also. Following the notation in section 2, for the purpose of estimating A we can have,

$$\hat{A} = \Sigma n_{ijk}\,(n_{ijk} - p_{ijk} - 2)\; S_{ijk}^{-1} \qquad \ldots \quad (3.5)$$

where p_{ijk} denotes the number of non-zero ijk's and S_{ijk} is the sample dispersion matrix. In (3.5) we observe

$$E(S^{-1}_{ijk}) = \Sigma^{-1}_{ijk}/(n_{ijk}-p_{ijk}-2),$$

as shown by Seal (1951).

4. Application to sampling investigations on correlated variables

The results discussed in the preceding sections find certain useful applications in the design of sample surveys and several other investigations.

In the first place, when it is inconvenient or not possible to obtain sufficient number of observations of a variable, estimates with desired accuracy can be obtained by taking additional observations on other correlated variables. For instance, in the measurement of maximum length of femora, lengths including spines are not obtained in all cases as the spine usually gets broken. Then, if maximum lengths including and excluding spines are available for some femora and for others only that excluding spines, the best estimates of both the measurements can be obtained by following the procedure in section 3. (Here the measurements including spine alone not arising, n_2 is zero.)

Consider, again, for example an yearly sample survey for estimating the acreage under a crop. Without going into the possibilities of improving past years' estimates, we may here consider problems in designing such surveys.

Let N_1 be the total number of sample units surveyed in the first year. Let $V(p)$ denote the variance of the estimate of the crop-proportion in the second year. Then,

$$\frac{V(p)}{\sigma^2} = \frac{n\rho^2+N_1(1-\rho^2)}{N_1(n+n_2)-\rho^2(N_1-n)n_2} = M \qquad \dots \quad (4.1)$$

where n is the number of sample units common for the two years, n_2 the number of units taken fresh in the second year, σ the standard deviation of the second year's crop proportion and ρ the correlation that exists between the first and second years' proportions. M is a quantity proportional to the margin of error of the estimate.

Let g_1 denote the cost for surveying a retained sample unit and g_2 that for a fresh unit. Then if C be the total cost

$$g_1n+g_2n_2=C \qquad \dots \quad (4.2)$$

The optimum allocation of n and n_2 for minimum margin of error or for minimum cost determined by

$$\frac{\partial}{\partial n}(M+\lambda C) = 0$$

and $$\frac{\partial}{\partial n_2}\,(M+\lambda C) = 0,$$

leads to $$g_2[N_1^2(1-\rho^2)] = g_1[n\rho^2+N_1(1-\rho^2)]^2$$

The admissible part of this condition is

$$n = N_1\,\frac{\sqrt{g_1g_2(1-\rho^2)} - g_1(1-\rho^2)}{g_1\rho^2} \qquad \ldots \quad (4.3)$$

It will be of some interest to observe that the equation (4.1) in n and n_2 defines a system of hyperbolas for different values of M. The equation $g_1n+g_2n_2=C$, then representing the 'budget line', the optimum solutions for n and n_2 would correspond to coordinates of points of contact of budget lines with curves in the system. For any particular survey, if the locus of the point of contact, namely the familiar "offer curve", is drawn and graduated in terms of the margin of error, it can be readily used to determine the optimum number of fresh sample units to be taken for a given cost or for a given margin of error. It can also be used to determine, under optimum allocation of n and n_2, the accuracy possible of the estimate for a given cost or the cost required for obtaining a given precision.

Example: In a crop acreage survey 1000 sample units were surveyed in the first year. For purposes of designing the second year survey the following values are available: $\mu=0.25$, $\sigma=0.04$, $\rho=0.5$. The cost of surveying 19 fresh units is equal to that of 20 retained units. What is the number of sample units to be retained and what is the number to be taken fresh, in order to give an estimate within a 1% margin of error?

Equation (4.3) in this case gives

$$n = 1000\,\frac{\sqrt{19\times 20\,(0.75)} - 19(0.75)}{19(0.25)} = 554.$$

Measuring the margin of error in terms of twice the coefficient of variation, (4.1) becomes

$$\left(\frac{0.25\times 1}{2\times 0.04\times 100}\right)^2[1000(554+n_2)-0.25\,(1000-554)n_2]$$

$$-(554(0.25)-1000(0.75) = 0$$

$$n_2 = 401$$

If the second year survey is designed independently without taking into account the first year survey the total number required for the same margin of error would be $\left(\frac{2\times 0.04\times 100}{1\times .26}\right)^2 = 1024$ instead of 955.

In the above example if the total cost was fixed at that of 900 fresh sample units then (4.2) becomes

$$10 \times 554 + 20 n_2 = 900 \times 20$$
$$n_2 = 874$$

and the resulting margin of error of the estimate would be 1.03% as against 1.14% which a design independent of the first year would have given.

I am thankful to Dr. C. R. Rao under whose guidance the foregoing work was done.

References

Matthai, A. (1949): Estimation of parameters from incomplete data with application to design of sample surveys. *Proceedings of the 36th Indian Science Congress.*

Münter, A. H. (1936): A study of the lengths of the long bones of the arms and legs in man with special reference to Anglosaxon skeletons. *Biometrika* 28, 258.

Seal, K. C. (1951): On errors of estimates in various types of double sampling procedure (Appendix A). *Sankhyā* 11, 2.

Wilks, S. S. (1932): Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Math. Stat.* 3, 163.

Paper received: March, 1951.