DERIVATION AND ESTIMATION OF PARAMETERS OF THE MAN-POWER SURVEY FOR EGYPT (II)

BY

M. A. EL-BADRY and M. D. MOUSTAFA

Institute of Statistics, Cairo University

During a first reading of the paper on this subject (1) to the Egyptian Statistical Association and in the light of results reached there, the authors presented a number of suggestions for possible changes in sample design which would increase the efficiency of the estimates. It was also pointed out that the mathematical difficulties encountered in the statistical analysis in the first paper resulted solely from the fact that the sample size for each stratum was distributed over the sample of p.s.u's in proportion to their sizes. This resulted in the difficulty that the total number of families in the selected p.s.u's, which appeared frequently as a denomenator, is a random variable. Expected values and variances of terms including $N_{i.}$ in the denomenator were thus difficult to handle with absolute exactness This difficulty vanishes if we take a fixed proportion of families in each of the selected p.s.u's in the stratum.

This resulted in a re-consideration of the sample design. It was decided to replace the sampling procedure for group I, (namely regions No. 1, 2, 3, 4 & 7) which was described in paper (I) by the following new technique for each stratum :

1.—All primary sample units, rather than a sample thereof, are taken.

2.—A sample of proportion λ (the sampling fraction in the region) is taken from the families of each p.s.u. in the stratum.

These alterations will give rise to the following changes in the previous mathematical treatment. Using the same notations as in paper (I) and referring to equations there we have.

(*a*) Since $\frac{n_{ij}}{N_{ij}} = \frac{n_i}{N_i} = \lambda$, equation (3)′ becomes exact and the unbiased estimate T (x) of the total number of individuals having the criterion x will be given by the exact formula

$$T(x) = \frac{1}{\lambda} \sum_i x_i , \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where x_i is the number of individuals having x in the sample of the i^{th} stratum.

(*b*) The variance of x_i will be given by (10) in paper (I) after putting $\frac{M_{i.}}{N_{i.}} = \frac{M_i}{N_i}$ and $\frac{n_i}{N_{i.}} = \frac{n_i}{N_i} = \lambda$.

Hence formula (10) becomes :

$$V(x_i) = \lambda M_i p_i(x) q_i(x) + (1 - \lambda) p_i^2(x) \sum_i^{u_i} n_{ij} \sigma_{ij}^2, \quad (2)$$

which is also exact. Consequently the variance of the total T (x) will be given by the exact formula :

$$V[T(x)] = \frac{1}{\lambda^2} \sum_i \Big[\lambda M_i p_i(x) q_i(x) + (1-\lambda) p_i^2(x) \sum^{u_i} n_{ij} \sigma_{ij}^2 \Big] \quad (3)$$

The variance is estimated by replacing $p_i(x)$ by

$$\hat{p}_i(x) = \frac{x_i}{\lambda M_i} .$$

By comparing (2) and (14) of paper (I), it is easy to see that the modified design has increased the efficiency to a considerable extent.

The variance given by (I4) has been reduced by the amount

$$\sum_i \left[\frac{n_i^2}{\lambda_i} \times p_i^2 (x) \underset{j}{V} \left(\frac{M_{i.}}{N_{i.}} \right) \right] \text{ which is equal to}$$

$$\sum_i \left[N_i^2 \; p_i^2 (x) \underset{j}{V} \left(\frac{M_{i.}}{N_{i.}} \right) \right].$$

1.—*An Alternative Design :*

In regions like group II (regions 5, 8) where sampling each primary unit gives rise to difficulties in the execution of the field work, and where it may be necessary to sample the p. s. u's, it was decided to adopt the mentioned recommendation of the authors, namely to take a fixed proportion of all p. s. u's selected in the stratum. If this is carried out then $\frac{n_{ij}}{N_{ij}} = \frac{n_{i.}}{N_{i.}} = \frac{\lambda}{\alpha_i}$.

Thus the expected value of $x_{i.}$ as given by (3)′ in (I) becomes :

$$E (x_{i.}) = p_i (x) \; n_i \underset{j}{E} \left(\frac{M_{i.}}{N_{i.}} \right) = \frac{\lambda}{\alpha_i} \; p_i (x) \underset{j}{E} \left(M_{i.} \right)$$

$$= \frac{\lambda}{\alpha_i} \; p_i (x) \times \alpha_i M_i = \lambda \; M_i \; p_i (x)$$

Formula (3)′ thus becomes exact, and consequently equation (6) giving an unbiased estimate of T (x) is in this case free from any approximation.

Turning now to the variance of $x_{i.}$, we find that equation (10) becomes in this case :

$$V (x_{i.} \mid \underline{j}) = \frac{\lambda}{\alpha_i} \; p_i (x) \; q_i (x) \; M_{i.} +$$

$$\left(1 - \frac{\lambda}{\alpha_i} \right) p_i^2 (x) \sum_i^{\alpha_i \; u_i} n_{ij} \; \sigma_{ij}^2 \qquad (4)$$

The expected value of the first term on the right hand side is equal to $M_i\ p_i\ (x)\ q_i\ (x)$ and that of the second term is equal to

$$\left(1-\frac{\lambda}{\alpha_i}\right)\ n_i\ p_i^2\ (x)\ \sigma_i^2,$$

assuming that $E\ (\sigma_{ij}^2) = \sigma_i^2$. Hence

$$\underset{j}{E}\ (x_{i.} \mid \underline{j}) = \lambda\ M_i\ p_i(x)\ q_i\ (x) +$$

$$\left(1-\frac{\lambda}{\alpha_i}\right)\ n_i\ p_i^2\ (x)\ \sigma_i^2 \quad . \quad . \quad (5)$$

Also, since $E\ (x_{i.} \mid \underline{j}) = \frac{\lambda}{\alpha_{i.}}\ p_i\ (x)\ M_{i.}$ (from (3), I)

then $$\underset{j}{V}\ E\ (x_{i.} \mid \underline{j}) = \left(\frac{\lambda}{\alpha_i}\ p_i\ (x)\right)^2\ \underset{j}{V}\ (M_{i.}) = \left[\frac{\lambda}{\alpha_i} p_i(x)\right]^2$$

$$x\ \alpha_i\ u_i\ x\ \frac{u_i - \alpha_i\ u_i}{u_i - 1}\ V(M_{ij}) \quad . \quad (6)$$

Adding (5) and (6), we get the following formula for the unconditional variance of $x_{i.}$:

$$V\ (x_{i.}) = \lambda\ M_i\ p_i(x)\ q_i(x) + \left(1-\frac{\lambda}{\alpha_i}\right)\ n_i\ p_i^2\ (x)\ \sigma_i^2$$

$$+\lambda^2\ p_i^2(x)\ \frac{u_i}{\alpha_i} \times \frac{u_i - \alpha_i\ u_i}{u_i - 1} V\ (M_{ij}) \quad . \quad . \quad (7)$$

which is the form that (13) would take if $N_{i.}$ were constant. Consequently

$$V\ [T\ (x)] = \frac{1}{\lambda^2} \Sigma_i \left[\lambda\ M_i\ p_i\ (x)\ q_i(x) + \left(1-\frac{\lambda}{\alpha_i}\right) n_i\ p_i^2(x)\ \sigma_i^2\right.$$

$$\left. + \lambda^2\ p_i^2\ (x)\ \frac{u_i}{\alpha_i}\ \frac{u_i - \alpha_i\ u_i}{u_i - 1}\ V\ (M_{ij})\right] \quad . \quad (7)$$

This variance is estimated by substituting the estimate:

$$\hat{p}_i\ (x) = \frac{x_{i.}}{\lambda\ M_i} \quad \text{for } p_i\ (x).$$

2.— *The Required sample size :*

We are going to calculate here the samlpe size required for the following two purposes:—

1.— to estimate the probability $p_i\ (x)$ that an individual in the region will have the criterion x_i, with a 95% confidence level that the error will not exceed. 0.005

2.— to render significant any change in the value of the estimated p from one round to another, wherever the observed change exceeds. 0.005

In group I, namely regions No. 1, 2, 3, 4 and 7 where each Kism in the region is sampled with a constant proportion, the variance of the total $T\ (x)$, is given by formula (3).

$$V\ [\ T\ (x)] = \frac{1}{\lambda^2} \sum_i \left[\lambda\ M_i\ p_i\ (x)\ q_i\ (x) + (1-\lambda)\ p_i^2 \sum_1^{n_i} n_{ij}\ \sigma_{ij}^2 \right]$$

if we make the substitution $\frac{n_{ij}}{N_{ij}} \lambda$, then the above formula becomes :

$$V\ [T\ (x)] = \frac{1}{\lambda^2} \sum_i \Big[\lambda\ M_i\ p_i\ (x)\ q_i\ (x) + (1-\lambda)\ p_i^2\ (x)$$

$$\times \sum_1^{ni} N_{ij}\ \sigma_{ij}^2 = \sum_i \Big[\frac{1}{\lambda} \Big\{ M_i\ p_i\ (x)\ q_i\ (x) + S_i\ p_i^2\ (x) \Big\}$$

$$- S_i\ p_i^2\ (x) \Big] \qquad . \quad . \quad . \quad . \quad (8)$$

where

$$S_i = \sum_i^{ni} N_{ij}\ \sigma_{ij}^2 \qquad . \quad . \quad . \quad . \quad (9)$$

But $Mp = T(x)$ and consequently the variance of the proportion P of individuals having x in random sample of size λ will be given by,

$$M^2 V(p) = V[T(x)] = \frac{1}{\lambda}\left[\sum_i M_i p_i(x) q_i(x) + \sum_i S_i p_i^2(x)\right] - \sum_i S_i p_i^2(x) \quad (10)$$

$$\lambda = \frac{\sum_i [M_i p_i(x) q_i(x)] + \sum_i S_i p_i^2(x)}{M^2 V(p) + \sum_i S_i p_i^2(x)} \quad (11)$$

1—Now if we want the difference $D = |P - p| \leqslant d$ with a certain level of confidence and if a is the value of the standard normal deviate corresponding to this level of confidence, then

$$\frac{d}{[V(p)]^{\frac{1}{2}}} = a$$

Thus λ as given by (11) becomes

$$\lambda = \frac{\sum_i [M_i p_i(x) q_i(x)] + \sum_i S_i p_i^2(x)}{M^2 \frac{d^2}{a^2} + \sum_i S_i p_i^2(x)} \quad (12)$$

The numbers M, M_i and S_i are population statistics. Their most recent sourse is the data collected in preparation for the 1957 census. The quantities $p_i(x)$ and $q_i(x)$ can only be estimated for the criterion under consideration from the information already obtained in the previous rounds.

Once the right hand side of (12) is computed, the sampling fraction (or the sample size) will follow immediately since $n_i = \lambda N_i$ for each Kims.

2.—We move now to consider the sample size that would detect as significant a change in P (x) from one round to the next, whenever this change is $\geqslant$ d.

When the two consecutive samples are independent. The variance of the difference is equal to the sum of the two variances. Each of these two varainces will be given by (10).

Now the variance of a forthcoming round can never be computed with complete exactness because p_i (x) and q_i (x) are not known, in addition to the difficulties involved in the estimation of the changes in M, M_i and S_i. For these reasons we are going to assume that the variance of P in the forthcoming round is equal to that of the preceding one. The varaince of the difference between the two P (x) 's will thus be twice V (P) as given by (10), where p_i (x) and q_i (x) are estimated from the results of the last round.

Consequently, if we replace V (P) in (11) by 2 V (P) and use the substitution given in (12), λ becomes :

$$\lambda = \frac{\sum_i [M_i\, p_i\,(x)\, q_i\,(x)] + \sum_i S_i\, p_i^2\,(x)}{\frac{M^2}{2} \times \frac{d^2}{a^2} + \sum_i S_i\, p_i^2\,(x)} \quad . \quad . \quad (13)$$

3.—*The Sample size when the p. s. u's are sampled :*

From (7)', by putting $n_i = \lambda_i N_i$ and $P(x) = \frac{T\,(x)}{M}$ we get :

$$M^2 \, V\,(p) = \frac{1}{\lambda} \sum_i^{\alpha_i u_i} \left\{ M_i\, p_i\,(x)\, q_i\,(x) + N_i\, p_i^2\,(x)\, \sigma_i^2 \right\}$$

$$+ \sum_i^{\alpha_i u_i} \left\{ p_i^2\,(x), \frac{u_i}{\alpha_i} \times \frac{u_i - \alpha_i u_i}{u_i - 1}\, V\,(M_{ij}) - \frac{N_i}{\alpha_i}\, p_i^2\,(x)\, \sigma_i^2 \right\}$$

$$\therefore \lambda = \frac{\sum_i \left\{ M_i\, p_i(x)\, q_i(x) + N_i\, p_i^2(x)\, \sigma_i^2 \right\}}{M^2\, V(p) + \sum_i \frac{M_i}{\alpha_i} p_i^2(x)\sigma_i^2 - \sum_i p_i^2(x) . \frac{u_{i.}}{\alpha_i} \frac{u_i - {}_i u_i}{u_i - 1} V(M_{ij})}$$

Putting: $$A(x) = \sum_i^{\alpha_i u_i} \left\{ M_i\, p_i(x)\, q_i(x) + N_i\, p_i^2(x)\, \sigma_i^2 \right\}$$

and $$B(x) = \sum_i^{\alpha_i u_i} \left\{ \frac{N_i}{\alpha_i} p_i^2(x)\sigma_i^2 - p_i^2(x) \frac{u_{i.}}{\alpha_i} \times \frac{u_i - \alpha_i u_i}{u_i - 1} V(M_{ij}) \right\}$$

$$\therefore \lambda = \frac{A(x)}{M^2 \frac{d^2}{a^2} + B(x)}$$

This is the sampling fraction that would keep the random sampling error below d, with the confidence level corresponding to a.

If we want to be able to detect as significant a change in $P(x)$ equal to d or more from one round to the next, the necessary sample size is given by

$$\lambda = \frac{A(x)}{\frac{M^2}{2} . \frac{d^2}{a^2} + B(x)}$$

References

(1) Derivation and estimation of the Parameters of the Man—Power Survey for Egypt (I), M. D. Moustafa and M. A. El-Bardy.

(2) Sample Survey, Methods and Theory, M. H. Hansen, W. N. Hurwitz and W. G. Madow, Parts I and II. 1953.