ON SAMPLING SCHEMES TO ESTIMATE DIFFERENCE IN PRICES ON TWO OCCASIONS

By G. DIVAKARA RAO

Indian Statistical Institute

SUMMARY. In the case where the price collection survey is integrated with an existing nation-wide multi-subject survey, it requires examination as to how the sampling procedures of (i) independent samples on each occasion and (ii) sampling on successive occasions with partial replacement of units compare with the fixed sampling procedure for price collection. This paper examines this question after taking into consideration precision and cost involved.

1. Introduction

In price collection surveys aimed at estimating price differences the practice is to collect price data from a fixed set of sample points at constant intervals of time. The reason for preferring a fixed sample to an independent sample is apparently the gain in precision that is obtained by taking a fixed sample due to existence of positive correlation between prices in the sample points over period of time. However, a fixed sample procedure, if applied over a long period of time vitiates the statistical inference. In countries having continuing nation-wide multi-subject surveys a substantial reduction in cost can be obtained by integrating the price collection survey with other surveys. Two types of sampling procedures, viz., (1) independent samples or moving samples on each occasion, and (ii) sampling on successive occasions with partial replacement of units are considered here as possible alternatives to the fixed sampling procedure for price collection.

In Section 2 fixed and moving sampling procedures are compared whereas in Section 3 fixed sampling procedure is compared with the procedure of sampling on successive occasions with partial replacement of units. Throughout this paper, the sampling adopted is supposed to be simple random sampling without replacement. Also the sample size is sufficiently small compared to population size so that finite population correction could be omitted.

2. Fixed and moving sampling procedures

We shall first proceed to find the relation between the sizes of fixed and moving samples that would ensure the same precision of estimates of price change. Though it is usual that price changes are measured by means of the ratio of prices at two different periods, for simplicity the estimator of price change is here chosen as the difference in the mean prices of two different periods for studying the relation between the sizes of samples.

If $S^2_{(1)}$ and $S^2_{(2)}$ are the variances on the two occasions, n_m and n_f are the sample sizes of the moving sample and fixed sample respectively and ρ the correlation coefficient between prices on the same two occasions, then the ratio of variance of the moving sample (V_m) to the variance of fixed sample (V_f) is

$$\frac{V_m}{V_f} = \frac{\frac{S^2_{(1)}}{n_m}+\frac{S^2_{(2)}}{n_m}}{\frac{S^2_{(1)}}{n_f}+\frac{S^2_{(2)}}{n_f}-2\rho\frac{S_{(1)}S_{(2)}}{n_f}}$$

$$= \frac{n_f}{n_m(1-\rho)}, \quad \text{if } S^2_{(1)} = S^2_{(2)} = S^2 \qquad \dots \quad (1)$$

giving the relation between equivalent sample sizes to be

$$n_f = (1-\rho)n_m. \qquad \dots \quad (2)$$

This functional relationship enables us to determine the ratio of sample sizes n_f and n_m from the value of ρ.

It is now proposed to compare the two types of samples, viz., fixed and moving, by setting n_m and n_f to satisfy (2). The value of ρ and consequently the equivalent sample size (n_m) for a fixed n_f differs from item to item. Since our interest is mainly in the general price level we may choose a unique sample size for all items which can be either a maximum value of n_m obtained over different items or a sample size such that the variance of a given linear estimate of change in the prices of all items is the same for moving sample and fixed sample. We choose the latter for illustration.

Let $\bar{d}_i$ be the estimated price change in the i-th item. Consider the weighted average of these price changes as $\frac{\Sigma w_i \bar{d}_i}{\Sigma w_i}$ where w_i is the per-capita expenditure on the item-group to which the i-th item belongs. We assume that $\bar{d}_i$ and $\bar{d}_j$ are independent. The expression for the ratio of variances V_m and V_f for the above weighted average is given by

$$\frac{V_m}{V_f} = \frac{n_f}{n_m} \quad \frac{\Sigma\, w_i^2 S_i^2}{\Sigma\, w_i^2(1-\rho_i)S_i^2} \qquad \dots \quad (3)$$

where S_i^2 and ρ_i are the variance and correlation coefficient between prices on two occasions of the i-th item ($i = 1, 2, \dots, m$). From (3) we have $V_m = V_f$ if

$$\frac{n_m}{n_f} = \frac{\Sigma\, w_i^2 S_i^2}{\Sigma\, w_i^2(1-\rho_i)S_i^2}. \qquad \dots \quad (4)$$

To estimate (4), the population parameters S_i^2 and ρ_i can be replaced by their sample estimates s_i^2 and r_i respectively.

Since the ratio n_m/n_f also depends on units of price quotation, in practice only standard units of quotation should be used in (4). In fact we use only standard units in our examples given at the end.

Another method of determining n_m/n_f would be to consider the ratio,

$$\frac{V_m}{V_f} = \frac{n_f}{n_m} \Sigma w_i^2 \frac{V_{mi}}{V_{fi}}$$

$$= \frac{n_f}{n_m} \Sigma \frac{w_i^2}{1-\rho_i}$$

where V_{mi} and V_{fi} are the moving sample and fixed sample variances respectively of the price difference for the i-th item. Although these $\frac{n_m}{n_f}$ do not change with the units of price quotation, they may get unusually high when any ρ_i is close to 1. To reduce the effect of such cases $\frac{n_m}{n_f}$ can be obtained by setting $V_m = V_f$ in

$$\frac{V_f}{V_m} = \frac{n_m}{n_f} \Sigma w_i^2(1-\rho_i).$$

The second method eliminates the effect of changes in the units of price quotation on the values of n_m/n_f.

Cost functions. In this section cost functions are developed for a fixed sample of villages exclusively meant for price collection and for a moving sample of villages where the price enquiry is integrated with other enquiries of a survey.

Consider now an investigator zone where all the socio-economic enquiries are to be canvassed. Consider the following two methods of collecting monthly prices. The first method (A) consists in earmarking a particular village (P) (different from the m villages chosen in the zone) for price enquiry. This requires the investigator to visit the price village, (P), at a fixed time of every month. The second method (B) would be to collect prices on a fixed day of every month from that sample village (among the m villages) where the investigator would be conducting socio-economic enquiries.

Considering a total of 12 monthly price collections the cost function for price collection under method (A) is expected to be the following.

$$C_f = (2t_1+t_2)12 \qquad \dots \quad (5)$$

while under method (B), we expect the following cost function

$$C_m = \frac{mt_1'}{k} + 12t_2' \qquad \dots \quad (6)$$

where t_1 is time required for journey to the price village from camp, t_1' is the average time required for travel from one sample village to another, t_2 and t_2' are the times required for price collection. Evidently $t_2 = t_2'$. k is the number of socio-economic enquiries including the price enquiry. Since in method (B) only m journeys are required (including the journey from headquarters to the first sample village) for all the k enquiries, the journey time attributable to price collection is mt_1'/k.

Further, since the total distance travelled between m random points in a given area is approximately $\alpha\sqrt{m}$ (Mahalanobis, 1940; and Jessen, 1942) we expect

$$\theta = \frac{t_1}{t'_1} = \frac{\sqrt{2/3}}{\sqrt{m/m+1}}\ .$$

Similarly it can be seen that for comparisons over intervals of six months or one year the ratio $\frac{C_m}{C_f}$ will be the same as that given by (5) and (6).

It is seen that $\frac{C_m}{C_f}$ gives the ratio of costs under the two sampling procedures. Therefore the relative cost (R) of these two sampling procedures which ensure the same precision for estimates under the two methods, is given by

$$R = \left(\frac{n_m}{n_f}\right)\left(\frac{C_m}{C_f}\right) \qquad \ldots \quad (7)$$

where n_m and n_f are so chosen as to make $V_m = V_f$. Thus if C_1 is the amount spent under method (A), then for equal precision one must spend an amount of RC_1 under method (B).

In the preceding discussion the comparison made was between the costs attributable to price collection under each sampling scheme. In certain cases where it is sought to find the additional cost incurred in including a price collection survey with an already existing multi-purpose survey and where one has to choose between the two methods of price collection then comparisons will be more meaningful if additional costs of price collection under the two schemes are compared. In this case additional cost of 12 monthly price collections under the fixed sampling scheme will be same as in (6). But the additional cost of price collection under the moving sampling scheme where the price enquiry is integrated with the general survey will be only the cost of enquiry viz. $C_m = 12t'_2$ for 12 monthly price collections. Then in this case relative cost,

$$R_1 = \left(\frac{n_m}{n_f}\right)\left(\frac{t'_2}{2\theta t'_1+t'_2}\right)$$

3. Comparison between fixed sampling procedure and procedure of sampling on successive occasions with partial replacement of units

In this section we consider the second alternative viz., sampling on successive occasions with partial replacement of units, hereinafter called replacement sampling scheme. In a multi-subject survey whenever the work-load in each sample village requires the investigator to stay in the sample village for a period of about one month it is possible to retain a certain proportion of the villages selected for price collection on any occasion for the next occasion also by properly adjusting the programme for socio-economic enquiries in these sample villages.

Consider a sample of size n_1 on the first occasion and of n_2 on the second occasion, the units common to the two occasions being n^1.

Denote the sample values on the first occasion by x's and those on the second occasion by y's.

Let $\bar{y}_{1i}$ be the estimate of the population mean of the i-th item on the second occasion obtained by the regression method from the same units which are retained and $\bar{y}_{2i}$ is the estimate of the population mean of the i-th item derived from the sample units included on the second occasion only, then

$$\bar{y}_{wi} = \frac{\lambda}{1-\mu^2 r_i^2}\bar{y}_{1i} + \frac{\mu(1-\mu r_i^2)}{(1-\mu^2 r_i^2)}\bar{y}_{2i} \quad \ldots \quad (8)$$

where λ and μ denote the fraction of the sample units being retained and replaced respectively when a sample of the same size is taken on each occasion. $\bar{x}_{wi}$ is the revised estimate of the population mean of the i-th item on the first occasion having known the results on the second occasion.

An estimate of the change between the population means on the two occasions is $\hat{d}_i = \bar{y}_{wi} - \bar{x}_{wi}$. The sampling variance of $\hat{d}_i$ is given by Yates (1949) and Tikkiwal (1951).

$$V(\hat{d}_i) = \frac{1}{n_1 n_2}\left(1-\rho_i^2 \frac{n_1^{"} n_2^{"}}{n_1 n_2}\right)^{-1}\left[\left(1-\frac{n_1^{"}}{n_1}\rho_i^2\right)n_1\sigma_{yi}^2 + \left(1-\frac{n_2^{"}}{n_2}\rho_i^2\right)n_2\sigma_{xi}^2 - 2n'\rho_i\sigma_{xi}\sigma_{yi}\right] \quad \ldots \quad (9)$$

where $n_1^{"} = n_1 - n'$, $n_2^{"} = n_2 - n'$; ρ_i is the coefficient of correlation between x's and y's of i-th item σ_{xi} and σ_{yi} the variances of the population for the i-th item on the two occasions.

When $\quad n_1 = n_2 = n_s$ and $\frac{n'}{n_s} = \lambda$, $\frac{n_1^{"}}{n_s} = \frac{n_2^{"}}{n_s} = \mu \quad (\lambda+\mu = 1)$

$$V(\hat{d}_i) = \frac{(1-\mu\rho_i^2)(\sigma_{xi}^2+\sigma_{yi}^2) - 2\lambda\rho_i\sigma_{xi}\sigma_{yi}}{n_s(1-\mu^2\rho_i^2)}. \quad \ldots \quad (10)$$

Considering the weighted average of the estimate of price change of all items (the weights used being the same as in the preceding section) and assuming $\hat{d}_i$ and $\hat{d}_j$ to be independent we find that the variance of this weighted mean is given by

$$V_s = \frac{1}{(\Sigma w_i)^2}\,\Sigma\, w_i^2\, \frac{(1-\mu\rho_i^2)(\sigma_{xi}^2+\sigma_{yi}^2) - 2\lambda\rho_i\sigma_{xi}\sigma_{yi}}{n_s(1-\mu^2\rho_i^2)}. \quad \ldots \quad (11)$$

As seen already in the case of a fixed sampling procedure the variance of the weighted average of the differences on the mean prices in the two occasions is given by

$$V_f = \frac{1}{(\Sigma\, w_i)^2} \cdot \Sigma\, w_i^2 \frac{1}{n_f}\,\{\sigma_{xi}^2+\sigma_{yi}^2 - 2\rho_i\sigma_{xi}\sigma_{yi}\}. \quad \ldots \quad (12)$$

From (11) and (12) we have $V_s = V_f$ if

$$\frac{n_s}{n_f} = \Sigma\, w_i^2\, \frac{(1-\mu\rho_i^2)(\sigma_{xi}^2+\sigma_{yi}^2)-2\lambda\rho_i\sigma_{xi}\sigma_{yi}}{1-\mu^2\rho_i^2} \Big/ \Sigma\, w_i^2(\sigma_{xi}^2+\sigma_{yi}^2-2\rho_i\sigma_{xi}\sigma_{yi}). \quad \ldots \quad (13)$$

The r.h.s. of (13) can be estimated by replacing population parameters σ_{xi}^2, σ_{yi}^2, ρ_i by their estimates s_{xi}^2, s_{yi}^2 and r_i respectively.

Cost function. For comparison of cost attributable to either sampling scheme let us consider the same components of cost as considered in Section 2. For the fixed sampling procedure the scheme of price collection and the consequent cost function adopted in the previous section are adopted here also. But for the replacement sampling procedure the scheme consists in considering n investigator zones of m sample villages each where all the socio-economic enquiries will be conducted. On any specified occasion each of the n investigators will conduct the price enquiry from one sample village in his zone thus making available for price collection n villages in all. But which sample village will be covered for price enquiry on a given occasion will be pre-determined so as to ensure that a fraction λ of the sample villages covered on any occasion be retained for the next occasion.

Consider for example 10 investigator zones numbered 1, 2, ..., 10, so that there will be 10 villages for price collection on each occasion. Let the m sample villages in each investigator zone be serially numbered 1, 2, 3, ..., m. When $\lambda = \frac{1}{2}$ the following programme of price collection in the sample villages can be followed. On the first occasion the price collection will be done in the sample village-1 in all the investigator zones. On the second occasion in the investigator zones with odd numbers the sample village-1 will be repeated for price collection and in the even numbered zones price collection will be done in sample village–2. On the third occasion the sample village will be changed to '2' in odd numbered zones whereas in even numbered zones the sample village-2 will be retained. Thus on each occasion only in half the zones the sample village is changed or retained. The same procedure continues till all the sample villages are covered. Under this programme by following a strict time schedule prices can be collected at intervals of one month while attending to other socio-economic surveys. It is, however, necessary that the total work-load in each village should require the investigator to stay for not more than one month. When $\lambda = \frac{1}{4}$ similar programme can be drawn but in that case the period of stay of the investigator in each sample village will not be uniformly equal to the interval of price collection. If 0, t_1', t_2' and k are the same as in the previous section and if the time gap is one month, since fresh journeys to sample villages on each occasion are required in respect of $(1-\lambda)n$ villages and since such journey time is shared by all the k socio-economic enquiries the cost attributable to price collection on any occasion from the second occasion onwards in n sample villages is

$$C_s = \frac{(1-\lambda)nt_1'}{k} + nt_2'. \quad \ldots \quad (14)$$

In general C_s can be taken as the cost of price collection at intervals which are almost the same as the duration of stay of the investigator in a sample village.

The cost attributable to price collection in n sample villages under the fixed sampling procedure follows from (5) as

$$C_f = (2t_1+t_2)n. \qquad \dots \quad (15)$$

The above cost functions hold as long as the gap between the two occasions is one month only. If gaps as long as six months or one year are considered the type of integration of price collection with other socio-economic enquiries has to be different because on the second occasion, which in the present case comes after a lapse of six months/one year, the investigator would have moved to a sample village different from the village surveyed on the first occasion. So in respect of those sample villages repeated for price enquiry on the second occasion fresh to and fro journeys to the village from the camp village are required. The rest of the sample villages which are $(1-\lambda)n$ in number to be selected newly can be chosen from among the villages being surveyed by the investigators for socio-economic enquiries on the second occasion. Thus the journey time to those $(1-\lambda)n$ sample villages can be taken as shared by all the socio-economic enquiries including price enquiry. Now the cost attributable to price collection on any occasion from the second occasion onwards is given by

$$C_s' = 2\theta t_1'(\lambda n)+\frac{t_1'(1-\lambda)n}{k}+nt_2'$$

$$C_s' = \left\{2\theta\lambda+\frac{1-\lambda}{k}\right\} nt_1'+nt_2'. \qquad \dots \quad (16)$$

Hence $\left(\frac{n_s}{n_f}\right)\left(\frac{C_s}{C_f}\right)$ when the interval is one month and $\left(\frac{n_s}{n_f}\right)\left(\frac{C_s'}{C_f}\right)$ when the interval is six months/one year, give the relative cost, R^1, of these two sampling procedures which ensure the same precision of estimates.

4. Numerical example

In this section the different schemes of price collection discussed in the earlier sections are compared empirically. In the first place we consider the fixed sampling procedure and the moving sampling procedure.

In the table below the pooled estimates of S^2's and the estimates of the coefficients of correlation (s_1^2 and r_t respectively) observed for the different time intervals of one month, six months and one year were obtained from the village-wise monthly retail prices collected by the National Sample Survey.

TABLE 1. ESTIMATES OF COEFFICIENT OF CORRELATION (r) AND VARIANCE (s^2)

item	weight	time interval					
		one month[1]		six months[2]		one year[3]	
		s^2	r	s^2	r	s^2	r
1. rice (coarse)	7.42	0.0081	0.97	0.0088	0.84	0.0088	0.64
2. arhar dal	0.68	0.0098	0.93	0.0135	0.82	0.0074	0.82
3. potato	0.49	0.0241	0.87	0.0205	0.36	0.0156	0.62
4. turmeric	0.41	0.0744	0.82	0.0968	0.68	0.0620	0.36
5. groundnut oil	0.48	0.0097	0.87	0.0352	0.73	0.0758	0.40
6. milk (buffalow)	0.95	0.0301	1.00	0.0276	0.78	0.0152	0.76
7. gur (cane)	0.52	0.0172	0.92	0.0106	0.65	0.0188	0.39
8. meat (goat)	0.47	0.1714	0.94	0.1390	0.83	0.1348	0.73
9. tea	0.44	0.1567	0.82	0.1619	0.65	0.1023	0.63
10. tobacco (leaf)	0.38	1.3370	0.96	1.1460	0.95	1.8331	0.86
11. firewood	1.11	0.3758	0.97	0.5030	0.63	0.4074	0.71
12. dhoti (mill)	1.58	4.0282	0.93	3.6498	0.88	3.7225	0.78

Note: The unit of price quotation is seer (= 0.933 kg.) for item nos. (1)–(8) and (10); lb. for item no. (9); maund (= 0.735 cwt) for item no. (11) and pair for item no. (12).

[1] Based on the prices of February 1961 and March 1961.

[2] Based on the prices of September 1960 and March 1961

[3] Based on the prices of September 1960 and Septemer 1961.

From (5) and (6) we have

$$\frac{C_m}{C_f} = \frac{mt_1'/k+12t_2'}{(20t_1'+t_2')12}\,.$$

Using the values of $t_1' = 10.5$ hours, $t_2' = 5.5$ hours, $m = 6$ or 9 and $k = 10$ observed from recent National Sample Survey rounds, the relative costs are given below.

time interval	$\frac{n_m}{n_f}$[1]	$m = 6$			$m = 9$		
		$\frac{C_m}{C_f}$	R	R'	$\frac{C_m}{C_f}$	R	R'
1. one month	14.88	0.178	2.649	2.418	0.163	2.423	2.128
2. six months	7.34	0.178	1.307	1.193	0.163	1.196	1.050
3. one year	4.38	0.178	0.776	0.708	0.163	0.711	0.623

[1] These values were obtained using formula (4). The values obtained from the second method were 29.88, 6.11 and 2.84 respectively.

We observe from the above table that the ratio of sample sizes as well as the relative cost decrease with increase in the gap between the two occasions. It may be generally observed that for measuring price changes over gaps of one month fixed sampling porcedure is preferable to moving sampling procedure if precision and cost are jointly considered unless t_1'/t_2' and k take sufficiently high values. This is true whether R or R_1 is considered. But as the interval increases the relative cost decreases because coefficients of correlation decrease. Thus when the interval is six months $R \leqslant 1$ for all $t_1' \geqslant 3t_2'$, $k \geqslant 7$ and $m \geqslant 6$ and when the interval is one year $R \leqslant 1$ for all $t_1' \geqslant 2t_2'$, $k \geqslant 3$ and $m \geqslant 6$.

We now consider the fixed sampling procedure and the procedure of sampling on successive occasions with partial replacement of units.

Taking the same values of t_1', t_2', m and k as in the preceding illustration and $\lambda = \frac{1}{2}$ and $\frac{1}{4}$ the relative costs, R', are given below.

	$\frac{n_s}{n_f}$		$m = 6$		$m = 0$	
	$\lambda = \frac{1}{2}$	$\lambda = \frac{1}{4}$	$\lambda = \frac{1}{2}$	$\lambda = \frac{1}{4}$	$\lambda = \frac{1}{2}$	$\lambda = \frac{1}{4}$
1. C_s/C_f			0.178	0.186	0.157	0.163
2. C_s'/C_f			0.665	0.420	0.585	0.378
3. R' for intervals of						
(a) one month	2.06	3.22	0.367	0.599	0.323	0.523
(b) six months	1.69	2.71	1.124	1.163	0.989	1.024
(c) one year	1.58	2.24	1.051	0.961	0.924	0.847

Note : The values of r_i used are those given in Table 1, while those of s^2_{ei} and s^2_{yi} used are the same values which have been pooled to obtain s^2 in Table 1.

It can be seen that in general for intervals of one month the relative cost $R' \leqslant 1$ for all $k \geqslant 2$ when $t_1' = 2t_2'$ or for all $t_1' \geqslant t_2'$ when $k \geqslant 5$ and $m \geqslant 6$ provided $\lambda = \frac{1}{2}$ or $\frac{1}{4}$. For intervals of six months or one year the relative cost, $R' \leqslant 1$ for all $k \geqslant 7$ when $t_1' = 3t_2'$, $m \geqslant 6$ and $\lambda = \frac{1}{2}$ and $\frac{1}{4}$. In conclusion it can be said that (i) for measuring price changes over intervals of one month the procedure of sampling on successive occasions with partial replacement of units is better than the moving sampling procedure for all $m \geqslant 2$, k, t_1' and t_2' when $\lambda = \frac{1}{2}$ or $\frac{1}{4}$ and better than the fixed sampling procedure only under the conditions stated above, (ii) for measuring price changes over intervals of six months or one year nothing can be said in general about the relative merits of moving sampling procedure and the procedure of sampling on successive occasions. It depends on the particular set of values of m, k, t_1', t_2' and λ.

In practice ρ decreases very rapidly as lag increases and consequently the relative cost will decrease considerably. Assuming that the correlation coefficient for lags of one, two, three years etc. decreases as ρ, ρ^2, ρ^3 etc. (Patterson, 1950) it may be said that for longer intervals the relative cost of moving sampling procedure to fixed sampling procedure and procedure of sampling on successive occasions with partial replacement of units to fixed sampling procedure would decrease considerably.

I am thankful to Dr. Saibal Banerjee for his useful suggestions.

REFERENCES

JESSEN, R. J. (1942): Statistical investigation of a sample survey for obtaining farm facts. *IOWA Agr. Stat. Res. Bull.*, 304.

MAHALANOBIS, P. C. (1940): A sample survey of acreage under jute in Bengal. *Sankhyā*, 4, 511–530.

——— (1944): On large scale sample surveys. *Phil. Trans. Roy. Soc.*, Series B, 231, 584.

PATTERSON, H. D. (1950): Sampling on successive occasions with partial replacement of units. *J. Roy. Stat. Soc.*, 12, No. 2.

TIKKIWAL, B. D. (1951): Diploma thesis on theory of successive sampling, (mimeographed), 2-4, University of Karnatak.

YATES, F. (1953): Sampling methods for censuses and surveys, Charles Griffin & Co., London.

Paper received : June, 1965.
Paper revised : August, 1965.