INVARIANTS IN SAMPLE SURVEYS

By
PAUL JACOB AND K. SANKARANARAYANAN
National Sample Survey Organisation, Calcutta

(Received : May, 1973)

INTRODUCTION

Sample survey literature abounds in various uses of information on auxiliary variates to improve the estimates of characteristics of main interest. A wide variety of techniques—ranging from varying probability selection to ratio, product and regression methods of estimation—are available for the practioner who wants to use such information.

All these techniques aim at reducing the sampling variance of the usual unbiased estimators. In this paper we describe a special kind of auxiliary variate which can be used as a check on the estimates obtained in a sample survey—as an indicator of their reliability. They are variates which remain invariant over a period of time and can be estimated from the same sample along with the main characteristics. The National Sample Survey of India (NSS) has been collecting data on a number of such items in their experimental crop surveys carried out during the past two or three decades in West Bengal. In Section 2 we describe the nature and use of the estimates of such "invariants" in assessing the overall reliability of survey data. We also present and discuss some results based on the above surveys in Section 3.

Further, it is seen that estimates of invariant characteristics have a role to play also in reducing the sampling error of the customary estimators of change. A new estimator is presented in Section 4 which makes use of them to improve estimates of change of study variables. While this is always better than the usual estimator of change (the difference between the unbiased estimates of the means on two occasions), its performance will almost equal that of another estimator (the difference between the regression estimates of the two means) under common circumstances.

2. Invariants and Their Use

In repetitive surveys reliability of results is often judged on the basis of the consistency of the estimates over successive rounds. But when the estimates relate to characteristics that are subject to change, this check fails. The differences observed among the estimates of different rounds would include, in such cases real changes as well as survey errors. In such circumstances reliability of the survey results may be appraised by introducing a few items which are known to remain invariant over long or short periods. Data on these invariant items will be collected from the same sample along with data on the survey variables. Estimates of such items obtained from the samples of different rounds should remain near to one another, thus providing a check on the overall reliability of the surveys. Whether or not any difference is observed between the estimates of a survey item obtained from two rounds, the estimates of an 'invariant', also obtained from the same two rounds, can indtcate evantualities such as the sample being unrepresentative (a freak), or the presence of large non-sampling errors and so on. While agreement between the coresponding estimates of the invariant item will reassure one of the validity of the main survey results, a divergence on the other hand should be taken as a signal to show that there is something wrong somewhere. Thereafter further probes may be made to locate the actual errors and corrective measures may be taken.

In repetitive crop surveys, for example, to estimate agricultural production, certain items can be covered which are known to stay constant over a long or short period. For instance, the area under 'river' would stay constant over a long period. So if the area under river can be estimated every round and if it is found to be more or less constant, one can reasonalby be sure that one's sample was not a freak one or that no gross inconsistencies have affected the estimate. Alternatively, items which are expected to remain constant for short periods of 3-4 years may be used to serve as short term invariants. Some such items are : (*i*) road and path ; (*ii*) house and house sites ; (*iii*) water areas etc. Even area under some agricultural crops like bamboo also may serve the purpose.

3. Studies in West Bengal on Short Term Invariants

Parallel to the NSS surveys, but independent of them, experimental land utilisation surveys were being carried out in West Bengal on a regular basis since 1957-58 during each of the autumn, winter and spring seasons. Prior to that also, beginning from the late forties, these were carried out for a short period. The primary objectives of these surveys were to try out different sampling methodologies, field techniques for data collection etc. to give valid estimates of agricultural production. In these surveys some of the above-mentioned *short term invariants* were also introduced to study their behavioural pattern and to see how they help in determining the consistency of the estimates of crop area over years.

Results of studies during 1947-50 : During these three years crop surveys were planned to give estimates of the major crops of the three different seasons in West Bengal. Area estimates were obtained for jute, autumn paddy, winter paddy and spring crops. The area under some short term invariants like 'road', 'tank', 'homestead', 'temple' and 'bamboo' were also worked out for the three seasons. It may be noted that the area under these invariants is expected to be constant over seasons and over the years. The results are given in Table 1.

TABLE 1

Estimates of area under major crops and some invariants in ('000) acres for different years

Year/ season	*area in ('000) acres under*							
	jute	*paddy*	*spring crops*	*road*	*tank*	*home stead*	*temple*	*bamboo*
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*
1947-48								
autumn	273	1279		288	615	420	4.6	165
winter		8026		230	643	446	4.0	174
spring			1746	235	621	423	5.0	199
1948-49								
autumn	383	1241		195	496	307	4.1	131
winter		7854		222	591	383	4.0	159
spring			1842	184	529	389	4.0	177
1949-50								
autumn	458	1217		231	600	412	3.7	165
winter		8180		232	603	416	4.8	173
spring			1839	228	680	458	4.9	159

It is seen from columns (5) to (9) that the estimated magnitudes of the invariants are generally of similar order indicating that the trend in the pattern of cultivation revealed by the survey is a genuine one. An increasing trend in jute area is noticed over years wheras the area under autumn paddy, winter paddy and rabi crops have remained almost constant over these three years.

Results of the studies during 1962-65 : The surveys carried out during the years 1962-65 aimed also to bring out the pattern of short term invariants like 'water areas' and 'house and house sites'. The areas under these two items are estimated for the two major seasons, autumn and winter for the three years. The results are given in Table 2.

TABLE 2

Estimates of area under 'water areas' and 'house and house sites' for different years (1962-65)

year	*estimated area in ('000) acres during*			
	autumn		*winter*	
	water area	*house and house sites*	*water area*	*house and house sites*
(1)	*(2)*	*(3)*	*(4)*	*(5)*
1962-63	1576	667	1396	613
1963-64	1528	597	1358	622
1964-65	1526	591	1372	628

It is clearly seen that over the three years the area under both the 'short terms', 'invariants' under study were more or less constant.

Changes in short term invariants over decades : Short term invariants are expected to stay constant over a given short period only. Over decades, therefore, one may expect that the magnitude of these 'invariants' will undergo changes. Table (3) gives the estimated area under 'water areas' and 'house and house sites' for three consecutive years in the 1940's and 1960's.

TABLE 3

Area under 'water area' and 'house and house sites' for three consecutive years of 1940's and 1960's

year	*area in ('000) acres under**	
	water areas	*house and houses ites*
(1)	*(2)*	*(3)*
1947-48	1066	434
1948-49	942	363
1949-50	1154	433
1962-63	1486	640
1963-64	1443	601
1964-65	1451	610

* The annual estimates are the mean of the three seasonal estimates for the years 1947-50 and of the two seasons autumn and winter for the years 1962-65.

It is seen that over a decade the water areas have increased by about 37% and the house and house sites by about 46%. Thus, these items can serve as checks only for short periods.

We have given illustrations of invariants whose estimates can serve as checks on estimates obtained from crop surveys. It would be interesting to study whether such suitable invariants can be identified in the case of socio-economic surveys also.

4. An Estimator of Change

Besides serving as an indicator of the reliablility of survey estimates, especially observed changes in estimates obtained from surveys conducted at two or more different points of time, estimates of invariant characters can be used to improve upon estimates of changes of the parameters of study variables. In this section we present an estimator of the difference between the means of the study variable on two occasions.

We are using the customary notations only. Any new symbol or notation will be explained as it is introduced.

The study variable is y ; and the auxiliary variate, x is assumed to be invariant. $\bar{Y}_1$ and $\bar{Y}_2$ are the actual y means at times t_1 and t_2. The aim is to estimate $\bar{Y}_2-\bar{Y}_1$. We assume that $\bar{X}_1=\bar{X}_2$. Two samples have been selected independently to obtain data on the two occasions. $(\bar{y}_1, \bar{x}_1)$ and $(\bar{y}_2, \bar{x}_2)$ are the unbiased estimates of the population means of y and x obtained from two samples.

To estimate $\bar{Y}_2-\bar{Y}_1$ the following estimator is suggested :—

$$T=\bar{Y}_{CI}=(\bar{y}_2-\bar{y}_1)+\lambda(\bar{x}_2-x_1) \qquad \ldots(1)$$

where λ is a constant chosen such that $V(T)$ is minimum, as in the case of the ordinary regression estimator. It can of course be easily seen that T is an unbiased estimator of $\bar{Y}_2-\bar{Y}_1$ for any as $\bar{X}_1=\bar{X}_2$. The idea here is that while the difference $\bar{y}_2-\bar{y}_1$ confounds $\bar{Y}_2-\bar{Y}_1$ and the sampling error, $\bar{x}_2-\bar{x}_1$ is purely due to sampling error alone. Hence one can try to isolate, so to speak, the real difference $\bar{Y}_2-\bar{Y}_1$ by using $\bar{x}_2-\bar{x}_1$; and the degree of success in this attempt may be expected to depend on the closeness of the relationship between y and x on the two occasions. We have, then, the following :—

Theorem : $V(T)$ is minimum when

$$\lambda=-\frac{Cov\ (\bar{y}_2, \bar{x}_2)+Cov\ (\bar{y}_1, \bar{x}_1)}{V(\bar{x}_2)+V(\bar{x}_1)} \qquad \ldots(2)$$

Proof :

$$T=(\bar{y}_2-\bar{y}_1)+\lambda(\bar{x}_2-\bar{x}_1)$$

$$V(T)=V(\bar{y}_2-\bar{y}_1)+\lambda^2V(\bar{x}_2-\bar{x}_1)+2\lambda Cov.\ (\bar{y}_2-\bar{y}_1,\ \bar{x}_2-\bar{x}_1)$$

$$\frac{\partial V(T)}{\partial X}=2\lambda.\ V(\bar{x}_2-\bar{x}_1)+2\ Cov\ (\bar{y}_2-\bar{y}_1,\ \bar{x}_2-\bar{x}_1)$$

Equating this to zero, we have

$$\lambda=-\frac{Cov\ (\bar{y}_2-\bar{y}_1,\ \bar{x}_2-\bar{x}_1)}{V(\bar{x}_2-\bar{x}_1)} \qquad ...(3)$$

Now,

$$Cov\ (\bar{y}_2-\bar{y}_1,\ \bar{x}_2-\bar{x}_1)=E(\bar{y}_2-\bar{y}_1)(\bar{x}_2-\bar{x}_1)-E(\bar{y}_2-\bar{y}_1)E(\bar{x}_2-\bar{x}_1)$$
$$=E(\bar{y}_2\bar{x}_2+\bar{y}_1\bar{x}_1-\bar{y}_1\bar{x}_2-\bar{y}_2\bar{x}_1),$$

the second term vanishing since $E(\bar{x}_2-\bar{x}_1)=0$

$$=E(\bar{y}_2\bar{x}_2)+E(\bar{y}_1\bar{x}_1)-E(\bar{y}_1)E(\bar{x}_2)-E(\bar{y}_2)E(\bar{x}_1),$$

since the two samples are independent.

$$=E(\bar{y}_2\bar{x}_2)+E(\bar{y}_1\bar{x}_1)-\bar{Y}_1\bar{X}_2-\bar{Y}_2X_1$$
$$=Cov\ (\bar{y}_2,\ \bar{x}_2)+Cov\ (\bar{y}_1,\ \bar{x}_1) \qquad ...(4)$$

And since the two samples are independent,

$$V(\bar{x}_2-\bar{x}_1)=V(\bar{x}_2)+V(\bar{x}_1) \qquad ...(5)$$

Substituting (4) and (5) in (3) we get that $V(T)$ is minimum when

$$\lambda=-\frac{Cov\ (\bar{y}_2,\ \bar{x}_2)+Cov\ (\bar{y}_1,\ \bar{x}_1)}{V(\bar{x}_2)+V(\bar{x}_1)} \qquad ...(6)$$

If the sample designs are the same on both the occasions, this will reduce to:

$$\lambda=-\frac{Cov\ (\bar{y}_2,\ \bar{x}_2)+Cov\ (\bar{y}_1,\ \bar{x}_1)}{2V(\bar{x})} \qquad ...(7)$$

Thus the suggested estimator is

$$\hat{Y}_{CI}=(\bar{y}_2-\bar{y}_1)-\frac{Cov\ (\bar{y}_2,\ \bar{x}_2)+Cov\ (\bar{y}_1,\ \bar{x}_1)}{V(\bar{x}_2)+V(\bar{x}_1)}(\bar{x}_2-\bar{x}_1) \qquad ...(8)$$

In practice, however, $Cov\ (\bar{y}_2,\ \bar{x}_2)$, $Cov\ (\bar{y}_1,\ \bar{x}_1)$, $V(\bar{x}_2)$ and $V(\bar{x}_1)$ will have to substituted by their sample estimates.

It is interesting to note that $-\lambda$ is the weighted average of $\beta(\bar{y}_1,\ \bar{x}_1)$ and $\beta(\bar{y}_2,\ \bar{x}_2)$ the weights being $V(\bar{x}_1)$ and $V(\bar{x}_2)$ respectively.

Now, clearly :

$$V(\hat{Y}_{CI})=V(\bar{y}_2-\bar{y}_1)(1-\rho^2_{\bar{y}_2-\bar{y}_1,\ \bar{x}_2-\bar{x}_1}) \qquad ...(9)$$

which shows that $V(\hat{Y}_{CI})$ is always less than $V(\bar{y}_2-\bar{y}_1)$ unless $\rho_{\bar{y}_1-\bar{y}_2,\, \bar{x}_2-\bar{x}_1}=0$, which is unlikely if y and x are correlated.

It can be easily verified that

$$V(\hat{Y}_{CI})=V(\bar{y}_2)+V(\bar{y}_1)-\frac{\{Cov\,(\bar{y}_2,\bar{x}_2)+Cov\,(\bar{y}_1\,\bar{x}_1)\}^2}{V(\bar{x}_2)+V(\bar{x}_1)} \qquad ...(10)$$

When information on an auxiliary variate is available for all units in the population, the best possible estimator from the point of view of lowest variance is the regression estimator. One can construct a regression estimator for each of the two occasions, $\hat{Y}_{1r}$ and $\hat{Y}_{2r}$ and use

$$\hat{Y}_{cr}=\hat{Y}_{2r}-\hat{Y}_{1r} \qquad ...(11)$$

as an estimator of change.

That is

$$\hat{Y}_{cr}=\{\bar{y}_2-\beta(\bar{y}_2,\bar{x}_2)(\bar{x}_2-\bar{x})\}-\{\bar{y}_1-\beta(\bar{y}_1,\bar{x}_1)(\bar{x}_1-\bar{x})\} \qquad ...(12)$$

Now,

$$\hat{V}(Y_{cr})=V(\bar{y}_2)(1-\rho^2_{\bar{y}_2,\bar{x}_2})+V(\bar{y}_1)(1-\rho^2_{\bar{y}_1,\bar{x}_1})$$

$$=V(\bar{y}_2)+V(\bar{y}_1)-\left(\frac{Cov^2(\bar{y}_2,\bar{x}_2)}{V(\bar{x}_2)}+\frac{Cov^2(\bar{y}_1,\bar{x}_1)}{V(\bar{x}_1)}\right) \qquad ...(13)$$

(10)—(13) gives :

$$V(\hat{Y}_{CI})-V(\hat{Y}_{CR})=\frac{\{V(\bar{x}_1)\,Cov\,(\bar{y}_2,\bar{x}_2)-V(\bar{x}_2)\,Cov\,(\bar{y}_1,\bar{x}_1)\}^2}{V(\bar{x}_1)\,V(\bar{x}_2)\{V(\bar{x}_1)+V(\bar{x}_2)\}} \qquad ...(14)$$

Thus $\hat{Y}_{cr}$ is better than $\hat{Y}_{CI}$.

If the sample designs are the same in the two rounds, $V(\bar{x}_1)$ will equal $V(\bar{x}_2)$ and hence (14) will reduce to

$$V(\hat{Y}_{CI})-V(\hat{Y}_{cr})=\{Cov\,(\bar{y}_2,\bar{x}_2)-Cov\,(\bar{y}_1,\bar{x}_1)\}^2/2V(\bar{x}) \qquad ...(15)$$

Thus if the difference between the two covariances are not large relative to $V(\bar{x})$, or, in other words, the regression lines of $\bar{y}_2$ and $\bar{y}_1$ on $\bar{x}$ are nearly parallel to each other, the advantage of $\hat{Y}_{cr}$ over $\hat{Y}_{CI}$ is negligible. However, $\hat{Y}_{CI}$ has one important advantage over the other. To calculate $\hat{Y}_{CI}$ no knowledge of $\bar{X}$ is necessary. That is to say, data on x need not be available in the frame; it is sufficient that it is known for the units included in the sample. And as this can very easily be collected at the time of actual survey, this appears to be no small advantage.

Summary

The use of collecting data on a kind of auxiliary variates, which remain invariant over a period of time, for checking the validity of estimates of main survey items in the case of repetitive sample surveys is discussed in this paper with results from NSS Surveys conducted in West Bengal. Further, an estimator of change over two occasions which makes use of the estimates of such invariant auxiliary variates is also given and its performance its briefly compared with that of two other estimators.