ON THEORETICAL QUESTIONS UNDERLYING THE TECHNIQUE OF REPLICATED OR INTERPENETRATING SAMPLES

J. C. Koop, North Carolina State College
Raleigh, North Carolina

1. Introduction

The technique of interpenetrating samples was introduced by Mahalanobis in 1937 in jute acreage surveys in Bengal mainly for statistical control. Mention of its use in checking the work of field investigators is made in his 1939 paper where the theory behind his method of crop acreage estimation is discussed. Subsequently it was developed (Mahalanobis, 1944, 1946) and it is now used in practically all surveys conducted by the Indian Statistical Institute. The United Nations Subcommission on Statistical Sampling (1949) has recommended its use and has also suggested the alternative term "replicated sampling."

The technique consists in drawing two or more sets of samples from the same population using the same sampling procedure for each set of samples. There are two variants of the technique; one when the samples are statistically independent and the other when they are not so.

The sets of samples will be independent if each set is replaced after the entire drawing is completed. Consider for example, a multi-stage sampling procedure when k sets each containing m first-stage units, are to be drawn from a given stratum. Starting from the first set, all of the m first-stage units (as a set) must be replaced before the second set is drawn and so on; this will ensure statistical independence between the k sets in the context of any given probability system. Within a set the sampling at the first stage and each subsequent stage may be any type provided only that the units are selected according to some probability system constructed for that stage. This appears to be the most general definition of independent interpenetrating or replicated samples defined by Lahiri (1954) in a terminology slightly different from mine. It will be seen that when sampling is carried out with replacement of each first-stage unit, independent interpenetrating samples of a special type are obtained. (In this situation m = 1.)

The sets of samples will not be statistically independent if the constituent units of one sample are in some way associated or linked with those of another. Regardless of whether or not the sets of interpenetrating samples are independent, appreciable disagreement between the estimates from each set will indicate, in some sense, the discrepancies in observation.

There has been much controversial discussion (particularly in India) on the use of the technique in the assessment or control of observational, enumeration and response errors* (Panse and Sukhatme 1948, Yates 1949; Ghosh 1949, 1957; Mokashi, 1950; Sukhatme, 1952; Sukhatme and Seth, 1952; Cochrane, 1953) and it will not be discussed in this paper. Yates also discusses the question of estimating sampling errors using estimates from the independent samples. In both review articles Ghosh comments on the estimation of sampling errors and the "margin of uncertainty" (Indian National Sample Survey, 1956). His second article contains some references not cited here. Lahiri (1954) in his monograph discusses the practical and theoretical aspects of the problem of estimating sampling errors of various population estimates (totals, means, ratios) and the problem of confidence limits for them, but he gives more attention to the problems of design very relevant to Indian conditions.

Apart from the problem of non-sampling errors which will always be enmeshed in the estimate of the sampling variance, Yates, Ghosh, Cochrane and Lahiri are all of the view that the estimation of sampling errors, using estimates obtained from the independent samples, will not be precise because of the small number of degrees of freedom. This view will be examined later.

Finally, in regard to previous work related to this subject, Deming (1956) and Flores (1957) have reported on and discussed theoretically a type of sample design with very desirable properties, which in the writers view, does not strictly fall into the category of sample designs resulting from the principle of the technique, i.e., drawing two or more samples using the same sampling procedure. Jones (1956) has elaborated on Deming's ideas for this type of design.

In this paper (i) the technique of independent interpenetrating samples in its most general application will be examined from a theoretical standpoint; (ii) the problem of efficiency of linear estimates for multi-stage designs will be discussed; (iii) confidence intervals which do not depend on the distribution of the estimates will be given; and (iv) applications in which there is much theoretical uncertainty about the use of the method will also be discussed.

---

*This question has been discussed by Hansen et al. (1951) from a different point of view.

2. Theoretical basis of the technique

There is a certain probability system which shows the probabilities according to which each sampling unit is to be selected at each step (stage or phase) in sampling down to the ultimate unit which may or may not be the unit of analysis. This probability system defines the way by which each set of samples is to be drawn. It is abstract and we may say that the frame, which describes the units at each step in the sampling procedure, and the method of sampling which ensures that the units are selected with probabilities prescribed by the system, are its real counter parts.

We draw k sets of independent samples each by the same sampling procedure and with probabilities prescribed by the system and we recall that each set is replaced prior to drawing the next set. Because of the identity of the sampling procedure each set of samples will be identical in structure. This means that at a given step in sampling, a given set will have the same number of sampling units as the other sets. Hence each set of samples will have the same number of ultimate sampling units.

An example at this stage may be useful. Suppose we have a frame which shows two strata. In Stratum I there are N elements. In Stratum II there are M clusters each consisting of N_i elements (i=1, 2,...,M). We desire two sets of independent interpenetrating samples, s and s'. Each set is to consist of n elements from Stratum I and n elements each (at the second stage) from two clusters selected at the first stage from Stratum II. It is given that n is smaller than the size of the smallest cluster of Stratum II. The probability system is defined as follows:

Stratum I: Probabilities are equal and the units are not replaced after each draw.

Stratum II: Probabilities are equal at each stage and the units are not replaced after each draw and in each stage.

Consider the drawing of sample s according to the probability system just defined. In Stratum I n elements are selected; in Stratum II clusters i and j are selected at the first stage, and n elements from each of these clusters are selected at the second stage. Then the total probability of drawing the sample s (of an aggregate size 3n) is

$$P_s = \frac{1}{\binom{N}{n}} \cdot \frac{1}{\binom{M}{2}\binom{N_i}{n}\binom{N_j}{n}} . \qquad (1)$$

After this drawing, in Stratum I the n elements are replaced and in Stratum II the selected elements of clusters i and j are replaced. Another drawing is now made by the same sampling procedure to obtain sample s'. Suppose this time clusters i' and j' are obtained. Then

$$P_{s'} = \frac{1}{\binom{N}{n}} \cdot \frac{1}{\binom{M}{2}\binom{N_{i'}}{n}\binom{N_{j'}}{n}} . \qquad (2)$$

Altogether there are

$$S = \binom{N}{n}\sum_{i>j}\binom{N_i}{n}\binom{N_j}{n} \qquad (3)$$

possible samples, and therefore S possible estimates for a given characteristic. Samples s and s' are clearly statistically independent. Further

$$\sum_{s=1}^{S} P_s = 1 . \qquad (4.1)$$

This discussion illustrates the concept of the probability system (which is a construct), and the procedure of selecting the two sets of independent samples. Because of the replacement of the first set, some elements may appear again in the second set. If they appear again, they are not rejected. In the foregoing illustration the selection probabilities are equal but generally they need not be so. The probability system can be constructed to make each estimate more efficient in some sense. So also the frame can be constructed (in respect of the number of strata, the composition of the hierarchy of units incident to the stages (or steps) in sampling) to increase efficiency in some defined sense. The classical theories given in text books have discussed many of these problems.

With these preliminaries we resume where we left off at the second paragraph of this section. Let P_s be the probability of drawing a set of samples from a population or universe $\mathcal{U}$ described by a frame $\mathcal{F}$ according to a most general probability system $\mathcal{P}$. As explained above both $\mathcal{F}$ and $\mathcal{P}$ can be constructed (or chosen) so that their state (or condition) predisposes the resulting estimates to have, in some sense, some desirable property or properties.

There are k mutually independent sets of samples each of which appears with probability P_s not necessarily equal to that of the others. If we need to distinguish between the samples and their respective probabilities we may attach subscripts to s. Thus s_t will denote the $t^{\underline{th}}$ independent set of samples and P_{s_t} its corresponding probability. We will have

$$\sum_{s} P_s = 1 . \qquad (4.2)$$

Let T_s be some estimate of a population value T, computed on the basis of values revealed by sample s. T may be a mean value or a population total, in which case it is a linear function, or a ratio or some function of the values of one or more characteristics of the elements of $\mathcal{U}$. We have

$$E(T_s) = \sum_s P_s T_s, \tag{5}$$

and

$$V(T_s) = \sum_s P_s (T_s - E(T_s))^2. \tag{6}$$

Since the sampling procedure for each set is the same, each of the k estimates T_s (computed by some formula), will have an identical variance given by (6) above.

Let us consider the estimate T_s in itself which is made up of n ultimate sampling units. Suppose we are estimating some population value such as

$$T = \sum_{i\varepsilon\mathcal{U}} a_i f(x_i, y_i, \ldots) \equiv \sum_{i\varepsilon\mathcal{U}} a_i f_i \tag{7}$$

where $f(x_i, y_i, \ldots)$ is a known single-valued function of the measurable characteristics observed on the i^{th} element (ultimate sampling unit) at the last step in sampling and the a_i's are known constants specific to each i. We shall give concrete meaning to (7) shortly. For the purpose of a general discussion it is not necessary to specify the location of the ultimate unit i (e.g., its location in the hierarchy of units) in the frame $\mathcal{F}$. In this situation we can obtain an unbiased estimate of T, linear in f_i, as follows. Let

$$T_s = \sum_{i\varepsilon s} R_i f_i \tag{8}$$

where each R_i is a weight to be attached to each f_i. Now f_i is defined above and can be computed on the basis of values revealed by i. For T_s to be unbiased we must determine the R_i's such that

$$E(T_s) = E\Big(\sum_{i\varepsilon s} R_i f_i\Big) \equiv \sum_{i\varepsilon} a_i f_i \tag{9}$$

i.e.,

$$\sum_{i\varepsilon\mathcal{U}} a_i f_i \equiv \sum_s P_s \sum_{i\varepsilon s} R_i f_i$$

$$\equiv \sum_{i\varepsilon\mathcal{U}} a_i f_i \sum_{s\supset i} P_s \frac{R_i}{a_i} .$$

That is we must have

$$\sum_{s\supset i} P_s \frac{R_i}{a_i} = 1 \text{ for every } i\varepsilon\mathcal{U}. \tag{10}$$

There are as many such equations (10) as there are elements in the universe $\mathcal{U}$. But the total number of R_i's is nS where, it will be recalled, n is the number of ultimate sampling units in each set of samples and S is the number of possible samples. If $\mathcal{N}$ is the total number of elements then

$$nS > \mathcal{N},$$

and therefore the equations (10) can be solved. If S' is the number of samples which include i, then one set of solutions is obtained by choosing

$$R_i = \frac{a_i}{S' P_s} \quad \text{for all } s\supset i. \tag{11}$$

Hence we find

$$T_s = \frac{1}{S' P_s} \sum_{i\varepsilon s} a_i f_i \ . \tag{12}$$

For example if we are estimating the total characteristic

$$T = \sum_{i=1}^{N} x_i$$

of a finite population of size N with a sample of size n drawn with equal probabilities and without replacement, then $a_i = 1$, $f_i = x_i$ for all i and $S' = \binom{N-1}{n-1}$ and $P_s = 1/\binom{N}{n}$, so that we find $T_s = \frac{N}{n} \sum_{i=1}^{n} x_i$ a familiar formula.

By and large the estimating functions we may use may not turn out to be unbiased. Further it can be proved (with essentially the same approach adopted by Koop (1957)) that generally no minimum unbiased estimator to estimate $T = \sum_{i\varepsilon\mathcal{U}} a_i f_i$ exists.* Earlier this was proved by Godambe (1955) with a less general formulation of the problem.

On the basis of the k independent estimates, T_s, we wish to estimate T. The best linear combination of the k estimates will be given by

*Minimum unbiased linear estimators exist only in the special case when the units are selected with equal probabilities.

$$T' = \frac{1}{k}(T_{s_1} + \ldots + T_{s_k}). \qquad (13)$$

Clearly if each T_s is unbiased then T' will be unbiased. Regarding the variance of T' we have

$$V(T') = \frac{1}{k^2}\left[\sum_{t=1}^{k} V(T_{s_t}) + 2\sum_{t>t'} \mathrm{Cov}(T_{s_t}, T_{s_{t'}})\right]$$

We will find

$$V(T') = \frac{1}{k}\ V(T_s) \qquad (14)$$

since the estimates are statistically independent so that $\mathrm{Cov}(T_{s_t}, T_{s_{t'}}) = 0$, and each has uniform variance given by (6) above.

It is not difficult to show that $\hat{V}(T_s)$, the unbiased estimate of $V(T_s)$, is given by

$$\hat{V}(T_s) = \frac{1}{k-1}\sum_{t=1}^{k} (T_{s_t} - T')^2, \qquad (15)$$

so that

$$\hat{V}(T') = \frac{1}{k(k-1)}\sum_{t=1}^{k} (T_{s_t} - T')^2, \qquad (16)$$

which is of the same type as the formula for the sampling variance of the mean of an infinite population which was also noted by Lahiri (1954) but not proved.

Coming back to the estimate T' given by (13), had we adopted an alternative linear estimate

$$\sum_{t=1}^{k} \alpha_t\, T_{s_t}$$

where $\sum_{t=1}^{k} \alpha_t = 1$, we will find that

$$V\left(\sum_{t=1}^{k} \alpha_t\, T_{s_t}\right) = \sum_{t=1}^{k} \alpha_t^2\, V(T_{s_t}) = V.$$

By the use of the Cauchy inequality it can be shown that V attains a minimum when $\alpha_t = \frac{1}{k}$ for all t and this leads to (13) and (14).

Thus we find that the method of independent interpenetrating samples leads to an extremely simple formula (16) for the estimation of sampling variances whatever the design of the sample. As Lahiri has remarked, the calculation of estimates of variance by classical methods will involve, for a normal sample survey, very heavy computational work because of squaring numbers many hundreds of times when the classical formulas underlying the sample design are used. When two independent interpenetrating samples are used, we will find

$$\hat{V}(T') = \frac{1}{4}(T_{s_1} - T_{s_2})^2 \qquad (17)$$

which is very simple for computational purposes.

Also the standard error of T' is given by

$$(1/2)\left|T_{s_1} - T_{s_2}\right| . \qquad (18)$$

This formula is very suggestive. If we are carrying out, say, a demographic survey by the use of two independent interpenetrating samples and we desire to know the standard error of each estimate of the total characteristics, all we have to do is to tabulate the estimates of totals from each interpenetrating sample separately and compare the corresponding cells of the tables and apply (18). In the light of this discussion many of the estimates of the Indian National Sample Survey display a high degree of accuracy despite the possibility of the presence of non-sampling errors, which obviously will be enmeshed in the estimate of the variance. Elsewhere the possibility of evaluating these errors by experimental design has been mentioned (United Nations, 1949).

Finally in discussing the theoretical basis of the technique the barest assumptions about the nature of the frame $\mathcal{F}$ and the probability system $\mathcal{P}$ have been made (i.e., only their existence). If it is admitted that the universe $\mathcal{U}$ is divided up into h strata, then

$$P_s = \prod_{i=1}^{h} P_s^{(i)} \qquad (19)$$

where $P_s^{(i)}$ is the probability of obtaining one set of k independent samples for stratum i, (i=1, 2...h) obtained by the same sampling procedure as the remaining k-1, and we will have

$$\sum_{s} P_s^{(i)} = 1 \qquad (20)$$

for all possible s in stratum i. Also we will have

$$T_s = \sum_{i=1}^{h} T_s^{(i)} , \qquad (21)$$

where the $T_s^{(i)}$ are the separate independent stratum estimates so that (6) reduces to

$$V(T_s) = \sum_{i=1}^{h} V(T_s^{(i)}) \qquad (22)$$

in view of (19). Of course (22) can also be proved starting with (21). The unbiased estimate of $V(T_s)$ will then be

$$\hat{V}(T_s) = \sum_{i=1}^{h} \hat{V}(T_s^{(i)}) \qquad (23)$$

and each $\hat{V}(T_s^{(i)})$ can be estimated by formula of the same algebraic form as (16).

3. Relative efficiencies of linear estimates for multistage designs

We shall first consider a somewhat general sample design where sampling at the first stage is with unequal probabilities and with replacement and leave unspecified the probability system for the subsequent steps (stages and/or phases) in sampling.

Let the population or universe $\mathcal{U}$ consist of M first stage units which can be identified by $\mathcal{F}$. To estimate a certain total characteristic T we select k independent sets* of interpenetrating samples each containing m first-stage units each of which is replaced after drawing. In each set at each subsequent step in sampling a predetermined number of units is drawn. Let $P_i > 0$ be the probability of drawing the i^{th} first-stage unit where $\sum_{i=1}^{M} P_i = 1$. Here the probability system $\mathcal{P}$ consists of the set of probabilities $\{P_i\}$ and the remaining unspecified sets corresponding to each step in sampling.

Let X_i be the true total in unit i and X_i' the unbiased estimate of X_i calculated on the basis of estimates from the subsequent steps in sampling.

*One may wish to know why k sets of samples are taken. One practical reason is that if one or more sets are ruined(say be non-response or other unforeseen difficulties) the others will still permit valid estimation. Therefore the survey is not entirely a loss. We are playing safe.

Let s be one of the set of k interpenetrating samples. Then an unbiased estimate of T will be given by

$$T_s' = \frac{1}{m} \sum_{i \varepsilon s} X_i' \qquad (24)$$

To determine the variance of T_s' we shall apply Madow's theorem (1949). We find

$$V(T_s') = \frac{1}{m^2} \left[V\left(E\left(\sum_{i=1}^{m} X_i' \,\middle|\, i \varepsilon s\right)\right) + E\left(V\left(\sum_{i=1}^{m} X_i' \,\middle|\, i \varepsilon s\right)\right) \right] \qquad (25)$$

We find $E\left(\sum_{i=1}^{m} X_i' \,\middle|\, i \varepsilon s\right) = \sum_{i=1}^{m} X_i$ and

$$V\left(\sum_{i=1}^{m} X_i' \,\middle|\, i \varepsilon s\right) = \sum_{i=1}^{m} V(X_i') = m\, V(X_i') .$$

If we define the probability system and the procedure of sampling after the first stage, $V(X_i')$ can be determined by repeated application of the theorem. We now have

$$V(T_s') = \frac{1}{m^2} \left[m \sum_{i=1}^{M} P_i (X_i - T)^2 + mE(V(X')) \right] = \frac{1}{m} \left[\sum_{i=1}^{M} P_i (X_i - T)^2 + E(V(X_i')) \right] \qquad (26)$$

We have k sets of independent samples. The best estimate of T will be

$$T' = \frac{1}{k} (T_{s_1}' + \ldots + T_{s_k}) \qquad (27)$$

and

$$V(T') = \frac{1}{k} V(T_s') \qquad (28)$$

Had we taken a direct sample of mk first stage units and then sampled the subsequent units by the same procedures (which will mean that the number of samples taken at each step will

be the same) then the unbiased estimate of T, T'', will be given by

$$T'' = \frac{1}{mk} \sum_{i=1}^{mk} X_i' \tag{29}$$

and

$$V(T'') = \frac{1}{mk}\left[\sum_{i=1}^{M} P_i (X_i - T)^2 + E(V(X_i')) \right] \tag{30}$$

Clearly $V(T'') = V(T')$ so that when sampling is carried out with replacement at the first stage the estimate obtained by taking the arithmetic average of the separate independent estimates is as efficient as the estimate from a single sample of equivalent size.

Next consider the case when we sample the first-stage units without replacement and with equal probabilities from the same universe as above. The units at the subsequent stages are selected by some scheme which we shall leave unspecified. Suppose we take k sets of independent interpenetrating samples with m units at the first stage; then using sample, s_t, the estimate of the population total T will be given by

$$T_{s_t} = \frac{M}{m} \sum_{i \in s} \hat{X}_i \tag{31}$$

where $\hat{X}_i$ is an unbiased estimate of X_i. We will find

$$V(T_{s_t}) = M^2 \left[\frac{\sigma^2}{m} \left(1 - \frac{m}{M}\right) + \frac{1}{m} E(V(\hat{X}_i)) \right] \tag{32}$$

where σ^2 is the variance of the first-stage units defined with divisor M-1. If we take an arithmetic average of the k estimates

$$T' = \frac{1}{k} \sum_{t=1}^{k} T_{s_t}, \tag{33}$$

We will find that

$$V(T') = \frac{1}{k} V(T_{s_t}). \tag{34}$$

Had we taken a single sample of first-stage units each without replacement and employed the same sampling scheme at the subsequent steps as in the case of the separate independent samples then we will have

$$T'' = \frac{M}{mk} \sum_{i=1}^{mk} \hat{X}_i \tag{35}$$

and we will find

$$V(T'') = M^2 \left[\frac{\sigma^2}{mk} \left(1 - \frac{mk}{M}\right) + \frac{1}{mk} E(V(\hat{X}_i)) \right]. \tag{36}$$

Comparing the variances we have

$$V(T') - V(T'') = M^2 \frac{\sigma^2}{mk} \left(\frac{mk}{M} - \frac{m}{M}\right)$$

$$= M \frac{\sigma^2}{k} (k-1). \tag{37}$$

Thus the efficiency of the set of independent interpenetrating samples is less than a single sample of equivalent size. But clearly the difference in variances is the least when k=2. Therefore in this situation it is best to take only two independent interpenetrating samples because it will have the best possible efficiency but still below that of the single equivalent sample. It will be noted that this conclusion is also true for single-stage sampling. The estimate of the variance of V(T') will be given by

$$(\tfrac{1}{4})(T_{s_1} - T_{s_2})^2 \tag{38}$$

i.e., as in (17). The kind of argument usually advanced that (38) has "one degree of freedom" and therefore is a "poor estimate" does not seem quite relevant in the context of finite population theory.

If there is stratification in the universe, in both cases considered above it will be seen that the same conclusions hold.

The question of cost so far has not been mentioned in the discussion. Since the k sets of interpenetrating samples and the single sample of equivalent size have the same number of units at each step in sampling, the costs for the former cannot be greater than that for the latter in whatever way the allocation of units takes place. Indeed the k sets of interpenetrating samples are likely to have common units (since the sets are replaced) and therefore certain cost components may even be less.

4. Confidence limits for finite population estimates

Much has been said on this problem by Lahiri (1954) in his monograph and therefore the remarks which follow will be by way of a commentary and little can be added to it. Also a general theoretical discussion of this kind without consideration of this problem

would not be complete.

When the estimates from interpenetrating (or replicated) samples are linearly combined as in (13) the possibility that the resulting estimates for some or all of the several characteristics will have a normal form cannot be ignored particularly if the number of strata is large. Therefore for characteristics which are abundant, confidence limits of population values may be set on the basis of Student's distribution using the k estimates. However, we have just shown that for a given overall sample size for a multistage design without replacement at the first stage and with equal probabilities, it is best to use only two interpenetrating samples. This conclusion may or may not be true for other types of sample designs when units are drawn without replacement at the first stage and with unequal probabilities. Therefore for such designs it may be prudent not to have too many sets of interpenetrating samples. Some balance has to be struck between the opposing attractions of shorter confidence intervals (because of smaller t-values for larger number of degrees of freedom) on the one hand and on the other hand of the possibility of increased efficiency in the estimate of the standard error because of fewer replications.

The above difficulties can be remedied by certain arguments advanced by Savur (1937). He advocated the setting of confidence limits for the population median rather than the population mean. Most characteristics of finite populations are not distributed symmetrically. Now whena population of estimates (for a characteristic in question) each made up of a large number of observations is formed, most of these estimates will cluster round the true value and the median. Here it is necessary to point out that the true value and the expected value of the estimate are not necessarily identical. Suppose the difference of the cumulative probabilities up to the true value T and up to the median is δ, a very small quantity. Let the k independent estimates (which may be of any kind including ratios) be ordered as follows:

$$T_{s_1}, \cdots \quad \cdots, T_{s_k}$$

where s_1 is the least and s_k the greatest. Suppose the true value T is below the median. Then the cumulative probability up to T is $\frac{1}{2} - \delta$. Therefore the probability that all k estimates lie above T is $(\frac{1}{2} + \delta)^k$ and all lie below it is $(\frac{1}{2} - \delta)^k$. Hence the probability that some will lie below T and some above is

$$\alpha = 1 - \left\{(\tfrac{1}{2} - \delta)^k + (\tfrac{1}{2} + \delta)^k\right\} \simeq 1 - (\tfrac{1}{2})^{k-1} . \qquad (39)$$

if δ is very small as compared to its constituent binomial terms.

$$\therefore \quad P(T_{s_1} < T < T_{s_k}) \simeq 1 - (\tfrac{1}{2})^{k-1} = \alpha' . \qquad (40)$$

It will be noted that when k = 5, $\alpha' = .9375$. How small δ should be, can easily be worked out from the relevant terms in the expansion within braces in (39). However, if we cannot ignore δ, we can always make statements of inference conditional on it assuming a certain value as follows:

$$P(T_{s_1} < T < T_{s_k} \mid \delta) = 1 - \left\{(\tfrac{1}{2} - \delta)^k + (\tfrac{1}{2} + \delta)^k\right\} . \qquad (41)$$

Similar arguments have been advanced by Lahiri (1954) on the assumption that the true value and the median are for all practical purposes identical. Thompson (1936) also investigated the problem but from slightly a different point of view.

5. Interpenetrating samples which are not independent

At this stage comment on the use of such samples is appropriate. The samples may have units or elements which are linked by some rule or they may have overlapping units. Thus because the samples are not mutually independent difficulties are created in the investigation of the properties of their estimates. Even combined linear estimates may not be unbiased.

Further the derivation of estimates of variance leading to simple formulas as in (14), (15), and (16) does become possible. Also it is hard to find a way of determining confidence intervals on the same lines as in section 4 simply because the estimates are not independent.

6. Some applications

(i) Price Index Numbers. There are a large number of problems in this area which are familiar to economic statisticians. The question of sampling errors in the "weights" (in the sense of price index theory) used in the computation of index numbers does not seem to have been given attention. Kelly (1921) and Knibbs (1924) considered this problem. Recently Banerjee (1960) and Koop (1952) also considered the problem.

Because a price index number is a complicated ratio, the determination of its variance (if it has elements in its formula which can be treated as random variables) is very complicated. For example, consider an index

number of the type

$$I = \frac{\sum p_o q_o \left(\frac{p_1}{p_o}\right)}{\sum p_o q_o} . \qquad (42)$$

If the base year prices and quantities are determined by sampling methods then they will be subject to at least sampling errors the formula for which will depend on the underlying method. Now

$$V(I) = \sum \left(\frac{p_i}{p_o}\right)^2 V(w_o) \qquad (43)$$

where

$$w_o = \frac{p_o q_o}{\sum p_o q_o} \qquad (44)$$

and assuming that $\frac{p_1}{p_o}$ is determined by some reliable pricing system.

Now the expression for the variance of w_o, a ratio, will be very complicated. Indeed even if it is agreed that the methods used to find an expression for it are valid it is almost intractable.

Had we used two independent interpenetrating samples, then there would have been two independent sets of weights for the set of goods and services consumed by the population under study. With the same price information generated by some system (government) we can separately compute two index numbers which are statistically independent. Suppose they are I_1 and I_2. We will find the average, which has been shown to be the best, to be

$$I' = (\tfrac{1}{2})(I_1 + I_2) \qquad (45)$$

and

$$\hat{V}(I') = (\tfrac{1}{4})(I_1 - I_2)^2, \qquad (46)$$

so that the coefficient of variation of I' will be

$$\frac{|I_1 - I_2|}{I_1 + I_2} \qquad (47)$$

a very simple formula. In the light of the discussion in section 2 the index can be of any kind and not necessarily of a naive type used above for illustration. The sampling error will reflect the variations displayed by the population when it was surveyed. If these variations can be assumed to be conservative for at least some time then the use of $\hat{V}(I')$ will be of some help in practical situations, e.g., in wage adjustments.

Further if more than two independent interpenetrating samples, e.g., five, are used, then a confidence interval, with a high confidence coefficient,

$\alpha' = 1 - (\tfrac{1}{2})^4 = .9375$, for the true index number will be obtained.

(ii) Number of different words in a book. There has been a discussion on a class of problems of this type by Mosteller (1949) who commented on the solutions proposed by various statisticians.

Suppose we sample, by some probability system, n lines out of a total number N in a book* and count the number of different words and find it to be r, then a consistent estimate of the number of different words W will be given by

$$W' = N \frac{r}{n} . \qquad (48)$$

Now

$$V(W') = \frac{N^2}{n^2} V(r) \qquad (49)$$

and we will find that the expression for V(r) is very difficult to determine. Had we taken two independent interpenetrating samples and found the number of different words to be r_1 and r_2, then we will have

$$W' = (\tfrac{1}{2})(W_1' + W_2') \qquad (50)$$

and

$$V(W') = (\tfrac{1}{2}) \frac{N^2}{n^2} V(r). \qquad (51)$$

Now although we do not know V(r) we can estimate it as

$$\hat{V}(r) = (\tfrac{1}{2})(r_1 - r_2)^2 \qquad (52)$$

which is an unbiased estimate of V(r) and therefore we find

$$\hat{V}(W') = \tfrac{1}{4} \frac{N^2}{n^2} (r_1 - r_2)^2 \qquad (53)$$

so that we have a standard error to assess the precision of our estimate. This illustration shows another attractive property of the technique of independent interpenetrating samples. Formula (48) is essentially of the same type as Haenszel's formula. His ultimate units of analysis were elements whereas here the elements are clustered (as lines).

*We assume that the book is not a mathematical text where usually one is hard put to it to define lines.

References

Banerjee, K. S., (1960), "Calculation of sampling errors for index numbers," Sankhya, 22, 119-130.

Cochrane, W. G., (1953), Sampling techniques, New York: John Wiley and Sons, Inc., 212-214.

Deming, W. E., (1956), "On simplifications of sampling design through replication with equal probabilities and without stages," J. Amer. Statis. Ass., 51, 24-53.

Flores, A. M., (1957), "The theory of duplicated samples and its use in Mexico," Int. Statist. Inst. Bull., 36, 120-126.

Ghosh, B., (1949), "Interpenetrating (networks of) samples," Calcutta Statist. Ass. Bull., 2, 108-119.

_____, (1957), "Enumerational errors in surveys," Calcutta Statist. Ass. Bull., 7, 50-59.

Godame, V. P., (1955), "A unified theory of sampling from finite populations," J. R. Statist. Soc., B, 17, 269-278.

Hansen, M. H., Hurwitz, W. N., Marks, E. S., and Mauldin, W. P., "Response errors in surveys," J. Amer. Statist. Ass., 46, 147-190.

Jones, H. L., (1956), "Investigating the properties of a sample mean by employing random subsample means," J. Amer. Statist. Ass., 51, 54-83.

Kelly, T. L., (1921), "Certain properties of index numbers," Quarterly Publication of Amer. Statist. Ass., 17, 826-841.

Knibbs, G. H., (1924), "The nature of an unequivocal price-index and quantity-index," J. Amer. Statist. Ass., 19, 42-60.

Koop, J. C., (1952), "On the use of statistical methods in assessing the precision of index numbers," (Paper read before the first Burma Research Society Conference, August, 1952.)

_____, (1957), "Contributions to the general theory of sampling finite populations without replacement and with unequal probabilities," Ph.D. thesis, N. Carolina State College Library, and Dissertation Abstracts, 18, (7), 144.

Lahiri, D. B., (1954), "National Sample Survey Number 5. Technical paper on some aspects of the sample design," Sankhya, 14, 264-313.

Madow, W. G., (1949), "On the theory of systematic sampling II," Ann. Math Statist., 20, 335.

Mahalanobis, P. C., (1939), "A sample survey of the acreage under jute in Benegal," Sankhya, 4, 511-531.

_____, (1944), "On large-scale sample surveys," Phil. Trans. Roy. Soc., B, 231, 329-451.

_____, (1946), "Recent experiments in statistical sampling in the Indian Statistical Institute," J. R. Statist. Soc., 109, 325-378.

Mokashi, V. K., (1950), "A note on interpenetrating samples," J. Indian Soc. Ag. Statist., 2, 189-195.

Mosteller, F., (1949), "Number of different kinds in a population," American Statistician, 3, (3), 12-13.

National Sample Survey No. 7, (1956), Sankhya, 16, 230-434.

Panse, V. G., and Sukhatme, P. V., (1948), "Crop surveys in India-1," J. Ind. Soc. Agric. Statist., 1, 34-58.

Sukhatme, P. V. and Seth, G. R., (1952), "Non-sampling errors in surveys," J. Ind. Soc. Agric. Statist., 4, 5-41.

Sukhatme, P. V., (1952), "Measurement of observational errors in surveys," Int.Statist. Inst. Review, 20, 121-134.

Savur, S. R., (1937), "The use of the median in tests of significance," Proc. Ind. Academy Sc., A, 5, 564-576.

Thompson, W. R., (1936),"Confidence ranges for the median and other expectation distributions for populations of unknown form," Ann. Math. Statist, 7, 122-128.

United Nations, (1949), "Preparation of Sample Survey Reports," C (1), New York: Statistical Office of the United Nations.

Yates, F., (1949), Sampling methods for censuses and surveys. New York: Hafner Publishing Co., 44, 107, 143, 241-243.

References not cited in text

Dalenius, T., (1950), "The problem of optimum stratification," Skand. Aktuar. Tidskr., 33, 203-213.

Das, A. C., (1951), "On two-phase sampling and sampling with varying probabilities," Bull. Inter. Statist. Inst., 33, 105-112.

Durbin, J., (1953), "Some results in sampling theory when units are selected with unequal probabilities," J. R. Statist. Soc., B, 15, 262-269.

Horvitz, D. G. and Thompson, D. J., (1951), "A generalization of sampling without replacement from a finite universe," Ann. Math. Statist., 22, (2), 315.

Kitagawa, T. (1956), "Some contributions to the design of sample surveys." Sankhya, 17, (1), 1-36.

Koop, J. C. (1956), Sample survey of labour force in Rangoon: A study in methods. Rangoon: Union Government Printing and Stationery.

Midzuno, H., (1950). "An outline of the theory of sampling systems. Ann Inst. Statist. Math. (Japan), 1 (2), 149-156.

Roy, A. D. (1952), " An Exercise in errors," J. R. Statist. Soc., A, 115, 507-520.