INTERPENETRATING SAMPLES

Dr. A. C. Kulshreshtha
UN Statistical Institute for Asia and Pacific

Regional Training Course on
Sampling Methods for Producing Core Data Items for
Agricultural and Rural Statistics
BPS-Statistics Indonesia, Jakarta, Indonesia
29 September - 10 October 2014

Interpenetrating Sampling – Who developed ?

- Technique of interpenetrating sub-sampling was originally developed by P.C. Mahalanobis in 1936 as Inter-Penetrating Network of Sub-samples (IPNS)
- It was in the context of studying correlated errors within same samples arising due to enumerators' effect in large scale surveys conducted in India
- A similar method called 'Tukey Plan' was used by Deming in 1966 -
- Also known as method of **Replicated Sampling**

What is Interpenetrating Sampling ?

- Interpenetrating sampling technique consists of drawing the samples from the same universe,
  - in the form of two or more samples, selected in an identical manner and
  - each capable of providing a valid estimate of the population parameter
- Sub-samples may or may not overlap one another at different phases of sampling as could be in multi-stage sampling scheme
- In some studies Mahalanobis applied different treatments to different sets of sub-samples to compare effects of treatments

Interpenetrating Sampling- Advantage

- It simplifies the computation of variance in complicated survey designs such as multi-stage designs
- This method is incorporated at the design stage itself, unlike other general methods of estimating variance such as jackknife and bootstrap methods

Interpenetrating Sampling - Uses

- Examine factors for different sources of variation, e.g, enumerators, filed schedules, different methods of data collection and processing
- Provide control in data collection and processing stage
- In computing the sampling error from the FSUs if these are selected independently with replacement
- Provide correction for bias in ratio type estimator

Interpenetrating Sampling – Uses (Contd.)

- Provide a non-parametric method of computing
- Probability of sub-sample range covering median of estimator

- Supply advance estimates on the basis of one or more sub-samples and provide estimates based on one or more sub-samples when the total sample cannot be covered due to some emergency

- Provides basis of analytical studies

Main Purpose of Interpenetrating Sampling

- This technique helps in providing “a means of control of quality of information”
- Interpenetrating sub-samples can be used to secure information on non-sampling errors such as difference arising from differential interviewer bias, different method to elicit information, etc. (United Nations, 1964)

Rationale

- After the sub-samples have been surveyed by different groups of investigators and processed by different teams of workers at the tabulation stage
- A comparison of the final estimates based on the sub – samples provides a broad check on the quality of the survey results

Rationale (Contd.)

- For instance, in comparing the estimates based on four sub – samples surveyed and processed by different groups of survey personnel
  - if three estimates agree among themselves and the other estimate differs widely from them in spite of the sample sizes being large enough
  - then normally one would suspect the quality of work in the discrepant sub – sample

Rationale (Contd.)

- When measurement errors are correlated on different units (by same interviewer), we want to obtain
  - a sample estimate of variance of mean that is unbiased
  - a way of finding out the extent to which the correlations decrease the precision

A Caution

- It may be noted that comparison of several sub – samples only provides an idea about the differential non -sampling errors and not an idea of the magnitude of the non-sampling error itself
- That is, if the magnitude and the direction of the biases of two investigators were of the same order, a comparison of the sub– sample figures would generally show an agreement even when the magnitude of the bias of each investigator is considerable
- However, this point can be met by getting mean of the sub– samples surveyed by specially trained and experienced investigators

A Common Practice

- In actual practice, sometimes, instead of obtaining independent sub–samples, a random sample of size n is divided at random into sub-samples, each sub–sample containing m = n / k units
- This is justified to some extent because the correlation between sample means for randomly obtained sub – samples are negligibly small

A Common Practice (Contd.)

- The field work and processing of the samples are planned so that there is no correlation between the errors of measurement of any two units in different sub–samples
- suppose the correlation with which we have to deal arises solely from biases of the interviewers, then
- If each of k interviewers is assigned to a different sub–sample and if there is no correlation between errors of measurement for different interviewers, we have an example of the technique

A Model

$$Y_{ij\alpha} = \mu'_{ij} + d_{ij\alpha}$$

i : sub – sample (interviewer)

j : member within the sub – sample

$Y_{ij\alpha}$: α^{th} measurement on $(i, j)^{th}$ unit

$$\mu'_{ij} = \mu_{ij} + \beta_{ij}$$

μ'_{ij} : conceptual average of repeated measurement

μ_{ij} : true value of $(i, j)^{th}$ unit

β_{ij} : bias in measurement

$$V(\bar{Y}_{i\alpha}) = \frac{1}{m}\left\{S^2_{\mu'} + \sigma_d^2\left[1+(m-1)\rho_w\right]\right\}$$

ρ_w : correlation between $d_{ij\alpha}$ for the same interviewer

$\sigma_d{}^2$: average variance of errors

$$V\left(\bar{Y}_\alpha\right) = \frac{1}{k}V\left(\bar{y}_{i\alpha}\right) = \frac{1}{n}\left\{S^2{}_{\mu'} + \sigma_d{}^2\left[1+(m-1)\rho_w\right]\right\}$$

Analysis of Variance

EXPECTATIONS OF THE MEAN SQUARES
(ON A SINGLE- UNIT BASIS)

	d.f.	**M.S.**	**E(M.S.)**
Between interviewers (sub – samples)	k-1	$s_b^2 = \frac{m\sum\left(\bar{y}_{i\alpha} - \bar{y}_{\alpha}\right)^2}{k-1}$	$S_{\mu}^2, + \sigma_d^2\left[1+(m-1)\rho_w\right]$
Within interviewers	k(m-1)	$s_w^2 = \frac{\sum\sum\left(y_{ij\alpha} - \bar{y}_{i\alpha}\right)^2}{k(m-1)}$	$S_{\mu}^2, + \sigma_d^2(1-\rho_w)$

Two Important Results

- $\frac{s_b^2}{n}$ is an unbiased estimate of $V(\bar{y}_\alpha)$

- ANOVA also enables us to estimate the correlated component, since

$$\frac{E(s_b^2 - s_w^2)}{m} = \rho_w \sigma_d^2$$

Contribution of Response Variance to Total Variance

- comparison of $(m-1)(s_b^2 - s_w^2)/m$ with s_b^2 estimates the relative amount which the correlated component of the response variance contributes to the total variance of y_α.

- With measurements in which the correlated component is much larger than the simple response variance, the ratio $(m-1)(s_b^2 - s_w^2)/m s_b^2$ has been used alternatively as a measure of the relative contribution of the total response variance to the total variance of $\bar{y}_\alpha$.

- If the primary interest is in an unbiased estimate of $V\left(\bar{y}_{\alpha}\right)$ that takes proper account of the effects of errors of measurement, all that is necessary is that the sample consist of a number of sub – samples of the same structure in which we are sure that errors of measurement are independent in different sub – samples.

- Strictly, this requires that different interviewing teams, supervisors and data processors be used in different sub – samples.

- If $\bar{y}_{i\alpha}$ is the mean of the i^{th} sub-sample, the quantity $\sum(\bar{y}_{i\alpha} - \bar{y}_{\alpha})^2 / k(k-1)$ is an unbiased estimate of $V(\bar{y}_{\alpha})$ with k(k-1) d.f.

- This result holds because the sub-sample can be regarded as a single complex sampling unit, the sample being in effect a simple random sample of these complex units, with uncorrelated errors of measurement between different complex units

- Interpenetrating sub-sampling is also known as replicated sampling. It has got a distinct advantage in variance estimation.
- Suppose each sub-sample is drawn independently through an identical sample selection process.
- From each of the sub-samples an estimate of the parameter is obtained.
- When these estimates are compared with the overall estimate (derived from all the sub-samples combined) then an indication of the variability of the independent estimates of sub-samples can be ascertained

Result : If $t_1, t_2, \ldots t_k$ are unbiased and mutually uncorrelated estimators of θ, with variance σ^2, then $\bar{t} = \frac{1}{k}\sum_{i}^{k} t_i$ is an unbiased estimator of θ, with variance as $V(\bar{t}) = \sigma^2 / k$ and unbiased estimator of $V(\bar{t})$ as

$$v(\bar{t}) = \frac{1}{k(k-1)} \sum^{k} (t_i - \bar{t})^2$$

In replicated sampling t_i's are uncorrelated due to independent selection. For many complex sample surveys variance estimators are complex. However, estimation of this variance becomes extremely simple in replicated sampling.

The estimator $v(\bar{t})$ can be applied to any situation where we have independent samples (replicates) with any sample design Thus for k=2, we have the following useful simplification,

$$v(\bar{t}) = \frac{\sum_{i=1}^{2}(t_i - \bar{t})^2}{2} = \frac{(t_1 - t_2)^2}{4} \quad or, \quad s.e.(\bar{t}) = \frac{1}{2}|t_1 - t_2|$$

Provides basis of Analytical Studies

- The variance estimate with k replicates is based on (k-1) degrees of freedom. In analytical studies a large number of replicates, say between 20 and 30, may be required to draw reasonably reliable statistical inference [Kalton (1983)]
- In the early days, Mahalanobis used replicated sampling in a method called graphical fractile analysis

Provides basis of Analytical Studies (Contd.)

- Interpenetrating sampling method has been used in standard large-scale sample surveys in number of countries like, India Peru, Philippines and USA
- It has been a regular feature in almost all survey designs used by the Indian National Sample Survey Organization (NSSO)
- However, there are certain problems associated with this method

Provides basis of Analytical Studies (Contd.)

- Selection and data collection for a series of independent replicated samples may turn out to be more costly and cumbersome then drawing and observing a single large sample
- One must be careful not to create an undesired dependence between the sub-sample estimates. Such dependence could be introduced through the interviewers or at the data handling stage by processing staff

Provides basis of Analytical Studies (Contd.)

- To have a stable variance estimator $v(\bar{t})$ the number of independent samples, k should be large
- In practice, however, k cannot be so large usually, which makes a variance estimator unstable
- Mahalanobis proposed to use as few as four groups, whereas Deming suggests ten

Probability of Sub-Sample Range Covering Median of Estimator

- The sub-samples can be used to derive a simple measure of uncertainty of the estimator in a non-parametric way
- Suppose each of the t independent sub-samples provides an independent and identically derived valid estimate of the parameter.
- Further assume that the estimator is symmetrical about its median

Probability of Sub-Sample Range Covering Median of Estimator (Contd.)

- Then the probability that the median of the distribution of the sub-sample estimators lies within the sub-sample range is

$$1 - \left(\frac{1}{2}\right)^{t-1}$$

- Thus with two sub-sample estimates, the range between the two estimates provides 50% probability of coverage for the median of the estimator under the given conditions

Correction of Bias in a Ratio Type Estimator

- Suppose t independent sub-samples are drawn from the population
- and each of the sub-samples is of the same size, m (thus, n = m t)
- From each of the sub-samples two unbiased estimators y_j^* and x_j^* (j=1,2,…, m) of the universe total Y and X are obtained respectively

Correction of Bias in a Ratio Type Estimator

- Two estimators of the universe ratio $R = \frac{Y}{X}$ (combined and separate type) are

$$r^c = \frac{\sum_{j=1}^{m} y_j^* / m}{\sum_{j=}^{m} x_j^* / m} \quad \text{and} \quad r^s = \frac{1}{m} \sum_{j=1}^{m} \frac{y_j^*}{x_j^*}$$

Correction of Bias in a Ratio Type Estimator

- In large samples, the bias of the estimator r^s is m times that of r^c, and so an approximately unbiased estimator of bias of r^c is

$$Bias(r^c) = \frac{(r^s - r^c)}{(m-1)}$$

Correction of Bias in a Ratio Type Estimator

- An almost unbiased estimator of the universe ratio (unbiased up to the second order of approximation) is

$$r^{au} = \frac{(m r^{c} - r^{s})}{m-1}$$

- So an almost unbiased estimator (Quenouille – Durbin – Murthy – Nanjamma estimator) of Y is given by

$$\hat{Y}^{au} = X \cdot r^{au}$$

THANK

YOU