Regional Workshop on the Use of Sampling in Agricultural Surveys

MONTEVIDEO, URUGUAY, 20 – 24 June 2011

REFERENCE MATERIAL ON SAMPLING METHODS[*]

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

[*] Prepared by A.K. Srivastava, FAO Consultant

Index

I. SAMPLING SCHEME

Here we briefly describe some of the sampling procedures commonly followed in actual sample surveys. A sample must be representative of the population. The representative character of a sample is normally ensured through random selection procedures.

1.1 Random Sampling

In random sampling procedures, selection of sample is done through a probability mechanism, so that all samples are provided a definite probability of selection. This is why, Random sampling is also known as probability sampling. The simplest of all random sampling procedures is the one in which all possible samples are provided equal probability of selection. It is not always feasible, nor desirable, to generate all possible samples and then select one sample with equal probability. The procedure of selection, normally followed, is *one after another draw* procedure in which units are selected one after the other draw with equal probability of selection at each draw. This procedure is known as *simple random sampling.* In this procedure all population units have equal chance of selection. Thus, simple random sampling may be defined as follows:

Simple random sampling is the method of selecting the units from the population where all possible samples are equally likely to get selected.

It follows that in *simple random sampling* every population unit has the same chance of being selected in the sample. Such sampling procedures are known as Equal Probability Selection Methods (EPSEM). It may be noted that simple random sampling is an EPSEM procedure, but all EPSEMs are not necessarily simple random sampling methods.

1.2 Systematic Sampling

Systematic sampling is yet another method of selecting a sample. In simple random sampling the units were selected randomly at each draw. In systematic sampling the whole sample is selected with just one random number. The procedure is defined as follows:

Systematic sampling is a method of sample selection in which only the first unit is selected at random and the rest of the units are automatically selected according to a predetermined pattern.

The most common predetermined pattern is the one in which after the random start, units are selected at equal intervals. This method is also known as *linear systematic sampling*. Suppose we want to select a systematic sample of size n from a population consisting of N units. The method of linear systematic sampling is employed when N is a multiple of n i.e. N=n.k where k is an integer. Let us assume that the nk serial numbers of the population units in the frame are arranged as follows:

1	2	3	..	r	..	k
k+1	k+2	k+3		k+r		2k
2k+1	2k+2	2k+3		2k+r		3k
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
(n-1)k+1	(n-1)k+2	(n-1)k+3		(n-1)k+r		nk

Select a random number r such that $1 \le r \le k$. The number r is called a random start and k is called sampling interval. The selected sample is the population units with serial numbers r, r+k, …, r+(n-1)k.

Systematic sampling is an EPSEM procedure. However, it is not equivalent to simple random sampling. The method has got advantage of simplicity in selection. However, the efficiency of estimation depends on the ordering of the units. With a suitable choice of arrangement, keeping in view the trends in the population, the method has got the potential of performing very well. It has got a limitation that an unbiased estimation of variance is not feasible with this method of sampling.

There are several other variants of systematic sampling, depending upon the systematic pattern used for selection. One such procedure is ***Circular Systematic Sampling***, in which random start is taken between 1 and N and then subsequent units are selected with equal interval after arranging the population units in a circular way. This method takes care of the situation when N is not multiple of n i. e. $N \neq nk$.

1.3 Unequal Probability sampling

Simple random sampling and systematic sampling are EPSEM procedures. However, when units vary considerably in sizes, providing equal chance of selection for every unit may not be a desirable proposition. Under such situations, selection of units with unequal probabilities may provide more efficient estimators. In this scheme, the units are selected with probability proportional to a given measure of size. The size measure is an auxiliary variable closely associated with the study variable. This method is known as varying probability sampling or probability proportional to size (PPS) sampling. For selecting a population unit with PPS, following methods are used:

Cumulative Total Method:

Let the size of the ith unit be denoted by X_i. Let the total size for N population units be, $X = \sum X_i$, i=1,…., N. Then the selection procedure consists of the following steps:

Step 1. Define cumulative totals

$$T_{i-1} = X_1 + X_2 + \ldots\ldots\ldots + X_{i-1}$$
$$T_i = T_{i-1} + X_i ; \qquad\qquad i=2,\ldots, N$$

Step 2. Chose a random number r such that $1 \leq r \leq k$

Step 3. Select i^{th} population unit if $T_{i-1} < r \leq T_i$

The probability of selecting the ith population unit, using this procedure is given by $P_i=X_i/X$. This procedure is described for selecting one unit with probability of selection P_i.

It is observed that in this method, it is required to cumulate the sizes and write down these cumulative totals. The procedure becomes a bit tedious when population size N is large. A procedure which does not involve cumulating the sizes was given by D. B. Lahiri (1951) and is described below:

Lahiri's Method

Step 1. Select a random number (say) I from 1 to N

Step 2. Select another random number (say) j, fro 1 to M, where M is either equal to the maximum of the sizes X_i; i=1, 2, ….., N or is more than the maximum size in the population.

Step 3. If $j \leq X_i$, the i^{th} unit is selected, otherwise, the pair (i,j) of random numbers is rejected and another pair is chosen by repeating the steps 1 and 2.

The procedure is repeated till a unit is selected. This methods ensures that the probability of selection for i^{th} population unit is $P_i=X_i/X$.

Probability Proportional to Size with and without replacement

If n units are to be selected with replacement, the procedure is to be applied independently n times. Thus, conceptually, every selected unit is replaced to the population before next unit is selected again with the same probability measures. The estimation procedure for estimating population total and for estimating the sampling variances is simpler in this case. We are not describing the estimation procedure here.

For PPS sampling without replacement, the selected units are to be excluded from the population for subsequent draws and the selection and estimation becomes much more complex. Since estimation become somewhat involved, attempts have been made to make the estimation procedures somewhat simpler. We shall not get into vast area of sampling literature in varying probability without replacement. However, it is worthwhile to mention about a special category of varying probability without replacement procedure, which is commonly used in Agricultural censuses as well as annual surveys. For sampling procedures with varying probabilities without replacement, if inclusion probabilities of sampling units are proportional to size measures, then estimation becomes very simple. Such procedures are called Inclusion Probability Proportional to Size (IPPS) procedures or πps procedure, since inclusion probabilities are sometimes denoted by $\pi_i's$. One such procedure is PPS systematic sampling which is described below.

Probability Proportional to Size (PPS) Systematic Sampling

The procedure is described here for selection of EAs within a specific stratum. Define
N = number of EAs in the stratum
n = number of EAs to be selected in the stratum
z_i = the measure of size (MoS - number of agricultural households in this case) for the ith EA in the stratum

$$Z = \sum_{i=1}^{N} z_i$$

$$p_i = z_i / Z \qquad \text{i=1, …………, N}$$

$$\pi_i = np_i \qquad \text{i=1, ………….., N}$$

The π_i values are the selection probabilities for the ith EA. The pps systematic sampling selection procedure is described in following steps:

Step 1: In this step, the procedure of implicit stratification is described. We consider a sub-administrative (say, SA) level as implicit strata. Sort the list of EAs in the stratum by SA level. Within a PA, arrange the EAs in ascending order of MoS; then in the next SA arrange the EAs in descending order of MoS. Continue this sorting by alternating between ascending and descending sorting from one SA to the next. This type of sorting helps in improving the efficiency of pps systematic sampling.

Step 2: Check that $np_i < 1$, i.e. z_i is less than Z/n for all i in the stratum.

Step 3: Compute cumulative totals

$C_1 = \pi_1$
$C_2 = C_1 + \pi_2$
…..
$C_{N-1} = C_{N-2} + \pi_{N-1}$
$C_N = C_{N-1} + \pi_N$ (Note that C_N= n)

Step 4: Generate a random number "r" between 0 and 1. Compute the numbers r_i = r+i-1 with i = 1, 2, 3, …..,n+2.

Step 5: Select the n EAs with the labels $i_1, i_2, i_3, …….,i_n$ such that

$C_{i_1-1} < r_1 \leq C_{i_1}$
$C_{i_2-1} < r_2 \leq C_{i_2}$
$C_{i_3-1} < r_3 \leq C_{i_3}$
……………..
$C_{i_n-1} < r_n \leq C_{i_n}$

The procedure yields a sample of size n with pps systematic sampling and the selection probabilities are given by $\pi_i = np_i$; i=1, 2, 3, ………,N.

In this procedure, the estimation of population total becomes very simple and the estimator is given by

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

1.4 Stratification

The selection procedures described so far i.e. simple random sampling; systematic sampling and unequal probability sampling are all methods of selection. In order to improve the efficiency of estimators, some times the population is divided into certain number of groups, such that the variability within the group is minimum, whereas it is maximum between groups and then smaller numbers of units are selected within each group following any of the random selection methods as described earlier.

The procedure of partitioning the population into groups, called strata and then drawing a random sample independently from each stratum, is known as **stratified random sampling.** In fact, stratification is a control measure applied in the process of selecting a sample for improving the precision of estimates.

Certain considerations need to be addressed while deciding about the stratification plans:

1. How many strata to be formed?
2. How to allocate the total sample size to different strata?
3. How to form the strata?

The basic considerations for all these questions are that the sampling variance should be minimized for a given cost or vice versa i.e. minimize the cost for a specified precision level. Following considerations should be kept in mind for stratified sampling.

1. The strata should be non-overlapping and should together comprise the whole population.
2. The strata should be homogeneous with respect to the study variable.
3. Several rules are available in literature for demarcating the strata boundaries. However, when it is difficult to stratify the population with respect to the study variable or a highly correlated auxiliary variable, the administrative convenience may be considered as a basis for stratification.

Number of strata

Regarding the question of number of strata, it may be observed that the normally the efficiency of stratified sampling estimators increases with increase in number of strata. But rate of reduction in the variance decreases as the number of strata increases. Also, the cost of the survey is affected by increase in the number of strata. Based on certain theoretical and empirical considerations, Cochran (1977) observed that if an increase in number of strata (say, L) beyond 6 necessitates any substantial decrease in sample size n in order to keep the cost constant, the increase will seldom be profitable. This indicates that if one is interested in overall estimates for the population mean/total, number of strata around six or seven should be reasonable enough. However, if estimates are wanted also for geographical subdivisions of the population, the argument for larger number of strata is stronger.

Allocation of sample to different strata

Although the total sample size is generally obtained based on cost and variance considerations, the decision about allocating the total sample size to different strata has to be made before selecting the sample. Following methods are commonly available:

1. Equal allocation
2. Proportional allocation
3. Compromise allocation
4. Optimum allocation

We explain the allocations with the help of an example in which districts are not only strata, but the interest also lies in getting reliable district level estimates and also national level estimates. Enumeration Areas (EAs) are the sampling units. A given sample size of EAs is to be allocated to different strata (districts).

Equal Allocation: In this approach, the sample is allocated equally to each stratum. The strata sizes vary considerably and equal allocation will provide not so efficient estimates at higher levels as the districts will not get due representation in the sample. This allocation is therefore, not a suitable alternative.

Proportional Allocation: This allocation will provide sample sizes in proportion to strata sizes. This is a good alternative for provincial and national level estimates. Estimates for larger districts should be good enough, but smaller districts will have poor estimates.

Compromise Allocation: In this approach we try to get a balance between producing reliable district level estimates and reliable national level estimates. Sometimes a "square root" allocation is used, in which the sample is allocated in proportion to $x^{1/2}$, where x is the measure of size. A more general allocation plan is the "power allocation" in which the sample is allocated in proportion to x^{λ}, where λ can take values between zero and 1. A suitable value of λ may be obtained by obtaining the design effects for national level estimates and also keeping in mind the requirement of getting reasonable district level estimates. Normally, λ =0.4 or 0.5 is considered good enough in many situations.

Optimum Allocation: In this method variance of the stratified estimator is minimized with respect to a given cost. Let us consider a simple cost function

$$C = c_0 + \sum_{h=1}^{L} c_h n_h$$

where c_0 is the overhead cost, c_h is the cost of observing study variable y for each unit selected in the sample from hth stratum, h=1,…., L. After optimization, a fixed cost – minimum variance allocation is given by

$$n_h = \frac{(C - c_0) W_h S_h \big/ \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}}$$

If the cost per unit is same for all the strata then the variance is minimized with respect to the restriction

$n = \sum_{h=1}^{L} n_h$ and n_h is given by

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$$

$$= n \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

This allocation is also known as **Neyman's optimum allocation.**

1.5 Cluster sampling and Multi-stage sampling

Consider the situation of agricultural censuses, in which agricultural holdings are the sampling units. In case, a list of all the holdings in a stratum (say district) is not available, a sample of holdings can not be selected. Even if the list is available, a sample of holdings straightaway selected from the entire stratum will be scattered all over the stratum. This will involve lot of travel expenditure.

A list of EAs is usually available. Each EA is a group or cluster of households. If EAs are selected and all the agricultural households in the selected households are enumerated, then a considerably reduced number of EAs will account for the same number of agricultural households to be selected inn the sample. The spread of selected households will be limited to the selected EAs only, thereby reducing the travel expenditure.

The cluster sampling consists of forming suitable clusters of contiguous population units and completely enumerating all the units in a sample of clusters, selected according to a suitable sampling scheme.

In terms of efficiency, cluster sampling is advantageous if clusters are heterogeneous with respect to study variable. In this respect, cluster sampling is converse of stratified sampling in the sense that both constitute of groups of units but strata should be homogeneous whereas clusters should be heterogeneous.

Multi stage sampling is a natural extension of cluster sampling. If the clusters are not completely enumerated, but units are further selected within selected clusters then it is called two stage sampling. Thus, in agricultural census example, if EAs are selected and then within each selected EA, agricultural holdings are selected, the sampling is done in two stages. The selection may be extended to more than two stages and the procedure is termed as multistage sampling. The sampling units at the first stage are called first stage units (FSUs) or also primary sampling units (PSUs). The units at the second stage are termed as second stage units or secondary sampling units (SSUs).

An important feature of multi-stage sampling is that at different stages, samples are selected independently and different methods of selection may be used at different stages. For example, in two-stage sampling, SRSWOR method may be followed at both the stages. Particular cases of special interest are when PPS with replacement or PPS systematic sampling is followed at first stage and SRSWOR are followed at the second stage. The later case, i.e. when PPS systematic sampling is followed at the first stage and SRSWOR at the second stage, is quite common in agricultural censuses and surveys. In this case, EAs are selected at the first stage with measures of sizes as the number of agricultural households in EAs and a given number of agricultural households are selected at the second stage in each of the selected EAs. This approach yields an EPSEM method of selection for each house holds. However, EPSEM nature of the selected sample is sometimes vitiated slightly due to differences in the size measure used in the selection process and actual number of agricultural households at the time of field work, when second stage selection of agricultural households actually takes place.

1.6 Multivariate Probability Proportional to Size (MPPS) Sampling

In PPS sampling, samples are selected with probability proportional to a size measure. The size measure is normally some auxiliary variable, which is highly correlated with the study characteristics. If there are several characteristics of interest, there may be a number of variables which may be correlated to the study variables. However, for sample selection with PPS, only one variable may be used. This variable could be a combination of auxiliary variables in order to generate a probability measure for selection. Multivariate approaches for generating a common index which could be used for selection purposes are sometimes used. The situation of several study characteristics of interest is very common in agricultural censuses and surveys. The characteristics related to different themes of supplementary modules are simple examples of multiple characteristics of interest.

An approach of MPPS as used in Censuses and surveys used in China (Steiner (2000)) is described below.

Define N= Number of units in the population

N_k= Number of units in the population having kth characteristic (k=1, …, K)

n_k= Number of units to be selected for the kth characteristic).

X_{ik}= value of the kth auxiliary variable for ith population unit.

$$X_k = \sum_{i=1}^{N_k} X_{ik}$$

$$p_{ik} = {}^{X_{ik}}\!\!\diagup_{X_k}$$

$$\pi_{ik} = n_k \, p_{ik}$$

$$\pi_i = Max(\pi_{i1}, \pi_{i2}, \ldots\ldots., \pi_{iK})$$

In MPPS procedure, for selecting ith unit, select a random number r_i (say) between 0 to1. If $r_i \le \pi_i$, then ith unit is selected, otherwise rejected. Continue this procedure independently for all the N units in the population. Essentially, the procedure is a Bernoulli's trial experiment with π_i as the probability of selection for ith unit.

The procedure ensures that the individual selection probabilities for different characteristics are taken into account and maximum one is taken as the selection probability for ith unit. The probability π_i serves as an index value based on all the auxiliary characteristics.

For estimation purposes, $w_i = \dfrac{1}{\pi_i}$ serves as the basic weight.

II. DETERMINATION OF SAMPLE SIZE

Determination of sample size is one of the initial questions which a survey statistician has to face while planning any sample survey. Cost and variance are the prime considerations while working out the sample size requirement. In random sampling, sampling variances are generally expressed as a function of sample size it reduces with increase in sample sizes. Cost of the survey is an increasing function of the sample size. Thus, increasing the sample size reduces the variance but it increases the cost. For a desirable sampling size, a balance is needed between cost and variance.

The principal steps involved in the choice of a sample size are as follows:

1. There must be some statement concerning the desired limits of error. In other words some statement is needed as to what is the tolerable margin of error in the estimates. This statement has to come from the persons, who wish to use the results.
2. Some statement that connects the sample size n with the desired precision of the sample must be found. One of the advantages of probability sampling is that sampling variances which measure the precision can be expressed in terms of n.
3. Sampling variances are population parameters and it contains some parametric values which need to be estimated in order to give specific results. For example, in simple random sampling, the sampling variance is a function of n but it has also got mean squares i.e. S^2.
4. Finally, the chosen value of n must be appraised to see whether it is consistent with the resources available to take the sample.

We consider the case of simple random sampling for quantitative character y, to demonstrate the steps needed for determining the sample size. Let r be the margin of relative error to be tolerated in estimating the population mean $\overline{Y}$. An unbiased estimator of population mean $\overline{Y}$ is sample mean $\overline{y}$. We want

$$\Pr\left(\left|\frac{\overline{y}-\overline{Y}}{\overline{Y}}\right| \geq r\right) = \Pr\left(\left|\overline{y}-\overline{Y}\right| \geq r\overline{Y}\right) = \alpha \,,$$

where α is a small probability. We assume that $\overline{y}$ is normally distributed. Also, the standard error of $\overline{y}$ is

$$\sigma_{\overline{y}} = \sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}$$

Hence

$$r\overline{Y} = t\sigma_{\overline{y}} = t\sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}$$

Solving for n gives

$$n = \left(\frac{tS}{r\overline{Y}}\right)^2 \Bigg/ \left[1+\frac{1}{N}\left(\frac{tS}{r\overline{Y}}\right)^2\right]$$

Here, $\frac{S}{\overline{Y}}$ is the coefficient of variation which is a fairly stable quantity. In order to calculate n, we need an approximate idea about this coefficient of variation.

As a first approximation, we take

$$n_0 = \left(\frac{tS}{r\bar{Y}}\right)^2 = \frac{1}{C}\left(\frac{S}{\bar{Y}}\right)^2$$

If n_0/N is appreciable, we compute n as

$$n = \frac{n_0}{1+(n_0/N)}$$

In case of **qualitative characteristics if a proportion P** is to be estimated and p is the sample proportion, the sample size is given by

$$n = \frac{(t^2PQ/d^2)}{1+\left(\frac{t^2PQ}{d^2}-1\right)\Big/N}$$

If N is large, a first approximation is

$$n_0 = \frac{t^2PQ}{d^2}$$

and $$n = \frac{n_0}{1+(n_0/N)}$$

Design Effect (Deff) and its role in sample size determination for complex designs

What has been described above is a procedure for determining the sample sizes in simple random sampling without replacement. In actual practice designs are much more complex. Kish (1965) described Deff as ratio of the variance of the estimate obtained from the (more complex) sample to the variance of the estimate obtained from a simple random sample of the same number of units. The sample size as obtained for simple random sampling is multiplied by Deff in order to get the required sample size for the complex design. The concept was initially given in the context of cluster sampling. In cluster sampling with equal clusters, the design effect is given by $\{1+(M-1)\rho\}$, where M is the cluster size and ρ is the intra-class correlation. In actual practice, the design effect is worked out from previous surveys and is used to determine the required sample size for the current survey. If the complex design is more efficient than the simple random sampling, value of Deff will be less than one and the required sample size will be smaller than the one obtained for simple random sampling. On the other hand if Deff is more than one, the required sample size will be more than the one obtained on the basis of simple random sampling.

III. ESTIMATION PROCEDURES

One of the main objectives of conducting sample surveys is to estimates population parameters of interest. Quite often, the interest lies in estimating parameters like population mean/total, sampling variances etc. Keeping in view the parameter of interest, estimators are chosen satisfying desirable properties like unbiased-ness, efficiency etc. For every sampling design, the estimation procedure invariably includes estimator of the parameter and estimators for sampling variance, which is a measure of the precision of the estimator.

Let us consider the estimators of population mean $\overline{Y}$ or population total $Y = N\overline{Y}$ and estimators of sampling variances in case of some of the prevalent sampling designs.

3.1 Simple random sampling (SRS):

For both with replacement (WR) as well as without replacement (WOR) cases, sample mean $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is an unbiased estimator of population mean $\overline{Y}$.

Estimator of sampling variance in case of SRS WR is given by

$\hat{V}(\overline{y}) = \frac{s^2}{n}$; where s^2 is the sample mean square given by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

In case of SRSWOR, estimator of variance is given by

$$\hat{V}(\overline{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)s^2$$

3.2 Systematic sampling:

For linear systematic sampling, with N being a multiple of n, systematic sampling is an EOSEM procedure and sample mean $\overline{y}$ is an unbiased estimator of population mean $\overline{Y}$.

Unbiased estimation of variance is not possible in this case. However some approximations are available. One such approximation is as follows:

$$\hat{V}(\overline{y}) = \frac{1}{2}\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n-1}\left(y_{i+1} - y_i\right)^2$$

However, if it is assumed that population is randomly distributed, then the same expression as used in case of simple random sampling may be used , i.e.

$$\hat{V}(\overline{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)s^2$$

3.3 Probability proportional to size (with replacement) sampling:

For PPSWR, estimator of population total is given by

$$\hat{Y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} ,$$

where p_i's are the initial probabilities of selection. Unbiased estimator of the sampling variance is given by,

$$\hat{V}(\hat{Y}_{pps}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{pps} \right)^2$$

3.4 Varying probability sampling (without replacement):

Most common estimator of population total Y in case of PPS WOR schemes is due to Horvitz and Thompson and is given as follows

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

where π_i is the selection probability or the probability of inclusion of ith population unit in the sample. Calculation of inclusion probabilities in general PPSWOR schemes is quite complicated and efforts have been made to either suggest estimators which do not require calculation of inclusion probabilities or to suggest varying probability without replacement schemes in which selection probabilities are proportional to size measures used for selection. These schemes are known as IPPS or πps schemes. One of the IPPS schemes is PPS-systematic sampling, which has already been described in the section 5.3. In IPPS schemes, $\pi_i = np_i$. Thus,

$$\hat{Y}_{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

which is of the same form as the estimator in PPSWR case. As in the case of systematic sampling, in this case also unbiased estimation of sampling variance is not possible. However, variances are estimated under some approximations and assumptions. Quite often, software packages for estimation of variances for complex sample surveys are used. Some of the methods used are sample re-use procedures, which are quite computer intensive methods. Simple expressions for estimators of variances are not available.

3.5 Stratified sampling:

In stratified sampling, samples are selected independently within each stratum. The estimation procedure depends on the method of sampling used within each stratum. Here, we consider the estimation procedure when SRSWOR method has been used within strata. An unbiased estimator of population mean is given as

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$$

An unbiased estimator of sampling variance is given by

$$\hat{V}(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{L} W_h{}^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) s_h{}^2$$

The estimators will depend on the method of sampling used within strata. If it is PPS selection, the formulae will change accordingly.

3.6 Cluster sampling:

We consider the case of equal clusters of size M each. Let n clusters be selected from N clusters using SRSWOR. Define

Y_{ij} = value of the character under study for jth unit in the ith cluster

$Y_{i.}$ = total for the ith cluster

$Y_{..}$ = total of the y-values for all the units in the population

$\overline{Y}_i$ = per unit ith cluster mean

$y_{i.}$ = ith sample cluster total

$\overline{y}_i = \frac{1}{M}\sum_{i=1}^{M} y_{ij}$ = per unit ith sample cluster mean

$\overline{y}_c = \frac{1}{n}\sum_{i=1}^{n} y_{i.}$ mean per cluster in the sample

$\overline{Y} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$

$\overline{Y}_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = M\overline{Y}$ = Population mean per cluster

An unbiased estimator of population mean is

$$\hat{\overline{Y}}_{cl} = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i$$

Variance of this estimator is given by

$$V(\hat{\overline{Y}}_{cl}) = (\frac{1}{n}-\frac{1}{N})\frac{1}{N-1}\sum_{i=1}^{N}(\overline{Y}_i-\overline{Y})^2$$

An estimator of variance is given by

$$\hat{V}(\hat{\overline{Y}}_{cl}) = (\frac{1}{n}-\frac{1}{N})\frac{1}{n-1}\sum_{i=1}^{n}(\overline{y}_i-\hat{\overline{Y}}_{cl})^2$$

For unequal clusters also estimation procedure is available, but several alternative estimators are considered depending upon whether population total is known or not.

An alternative form for the $V(\hat{\overline{Y}}_{cl})$ is approximately given as

$$V(\hat{\overline{Y}}_{cl}) = \frac{S^2}{n}\{1+(M-1)\rho\}$$

where ρ is the intra class correlation. In fact this very form of the variance leads to the well known form of Design Effect as $\{1+(M-1)\rho\}$.

3.7 Multi stage sampling

Consider the case of two stage sampling of unequal PSUs where selection at both the stages is done with SRSWOR. Define the following

N = number of PSUs in the population
n = number of PSUs selected in the sample
M_i = number of SSUs in the ith PSU
m_i = number of SSUs selected in the ith PSU
Y_{ij} = value of the study variable y for the (ij)th SSU
y_{ij} = value for the study variable in the jth selected SSU in the ith selected PSU
Y_i = total of y values in the ith PSU
Y = total of y values in the entire population
$\overline{Y}_i$ = Mean per SSU in the ith PSU
$\overline{y}_i$ = mean per SSU as obtained in the sample

An unbiased estimator of the population total Y is given by

$$\hat{Y}_{ts} = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{ij} = \frac{N}{n}\sum_{i=1}^{n} M_i \overline{y}_i$$

Variance of this estimator is given by

$$V(\hat{Y}_{ts}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_{bt}^2 + \frac{N}{n}\sum_{i=1}^{N} M_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2$$

where $S_{bt}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \frac{1}{N}\sum_{i=1}^{N} Y_i\right)^2$ and

$$S_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(Y_{ij} - \frac{1}{M_i}\sum_{j=1}^{M_i} Y_{ij}\right)^2$$

Estimator of variance is given by

$$\hat{V}(\hat{Y}_{ts}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_b{}^2 + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)s_i^2$$

where $s_{bt}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{Y}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i\right)^2$; $\hat{Y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$

and $s_i^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}\left(y_{ij} - \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}\right)^2$

In case of equal clusters where M_i=M and m_i=m, if mean is to be estimated the above formulae reduce to a simpler form as follows:

$$\hat{\bar{Y}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \sum_{i=1}^{n}\bar{y}_i$$

$$V\left(\hat{\bar{Y}}\right) = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2$$

where $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{Y}\right)^2$

and $\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2$; with $S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(Y_{ij} - \bar{Y}_i\right)^2$

In the variance formula as given above, the two components denote the contributions towards the total sampling variance due to PSUs and SSUs respectively. This splitting the variance into parts representing different stages of selection is very helpful in optimizing the sample sizes for two stages.

Estimator of variance is given by

$$\hat{V}\left(\hat{\bar{Y}}\right) = \left(\frac{1}{n}-\frac{1}{N}\right)s_b^2 + \frac{1}{N}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2$$

where

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{Y}_i - \bar{Y}\right)^2$$

$\bar{s}_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2$; with $s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_i\right)^2$

If N is large enough to ignore terms of $O\left(\frac{1}{N}\right)$, we obtain a simple expression for estimator of variance as

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{s_b^2}{n}$$

3.8 Role of sample weights in estimating population totals/means

It may be observed from the above discussions that sampling weights have got important roles to play in estimation of various parameters. Quite often we are interested in parameters like totals and means which are linear in nature. The estimates for such parameters are also linear in nature with sample observations suitably weighted with appropriate sampling weights. In agricultural censuses, since sampling is done for small and medium agricultural households only, the weighting procedure is considered only for such households. Large farms and institutional holdings are anyway completely enumerated. Corresponding weights for such holdings will be one only. The weighting procedure is essentially based on following three types of weights:

1) Base weights
2) Non response adjustments
3) Post-stratification adjustments

Base weights

It may be observed that in varying probability sampling without replacement, Horvitz-Thompson estimator is given as:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} w_i y_i$$

where the weights are inverse of selection probabilities of the ultimate units. In agricultural censuses, operational holdings are the units of observation. Since the selection probabilities are associated with the units of selection, which are agricultural households in this case, the agricultural households associated with the holding provide the base weights for the holdings. If there is one to one correspondence between agricultural households and holdings, the selection probabilities of the holdings are straightaway the selection probabilities of the agricultural households. There is no problem for multiple-holding households, as the selection probabilities for such holdings are straight forward. In the cases where one operational holding corresponds to several agricultural households, pro-rata adjustments may be done by considering pseudo-holdings corresponding to each household. However, such cases are likely to be very rare.

In two stage sampling, the selection probability of a SSU is the product of selection probability of corresponding PSU and the conditional selection probability of SSU for the given PSU. In the present case, EAs are PSUs which are selected with pps systematic sampling and agricultural households are SSUs which are selected with equal probability sampling.

Let π_i be the probability of selection for ith PSU (i.e. EA) and $\pi_{j|i}$ be the conditional probability for selecting jth SSU (household) in ith PSU, then the probability of selection for jth SSU in ith PSU is given by $\pi_{ij} = \pi_i \pi_{j|i}$. In this case,

$\pi_i = n\frac{X_i}{X}$; X_i is the measure of size (number of Agricultural households in ith EA as per 2007/08 PHC) and X is the sum of X_i in the specific stratum to which ith EA belongs. Also, $\pi_{j|i} = \frac{m}{M_i}$ where M$_i$ is the number of agricultural households in ith EA as observed at the time of field work for preparing the frame and m is the number of households selected in each EA.

Thus, $\pi_{ij} = \frac{nmX_i}{XM_i}$

In case when $X_i = M_i$, $\pi_{ij} = \frac{nm}{X}$ and the sample design is EPSEM. However, when $X_i \neq M_i$, the design is no more EPSEM and the base should be calculated carefully.

In general, the Base weights for each household in the ith EA is $\frac{XM_i}{nmX_i}$

Non-response adjustment

Invariably, there is some amount of non-response in every survey, which disturbs the weights. Therefore there is a need for adjusting for non-response. Normally, the non-response adjustments are done within each EA. The adjustment factor is (m/r), where m is the number of sampled holdings while r is the number of responding households.

Post-stratification adjustment

Sometimes it is felt desirable that the estimated totals for certain characteristics (auxiliary variables) in some population groups (which may as well be post-strata) are in conformity with the known totals for these groups. Some characteristics from PHC 2007 for which information is also collected in CAP-II may serve as a suitable variable for this adjustment. For example, number of households in a district may be known from PHC. An estimate for this characteristic may also be developed from the survey. The weights may be adjusted in such a way that the estimated value is equal to the known value fro PHC. This type of adjustment provides a check on the face value of the estimates with respect to known characteristics. Since the auxiliary characteristic is also correlated to the main study variable, the adjustment is also expected to provide more reliability to the estimates.

The final weights are the product of Base weight, non-response adjustment and the post-stratification adjustment.

IV. ESTIMATION OF SAMPLING ERRORS

4.1 . Variance Estimation

As described in the previous section, estimation of variance should be an integral part of estimation procedure. Estimation of variance provides a measure of precision for the estimates. It provides a level of confidence to the estimates developed.

Coefficient of Variation: Another convenient method for measuring precision of an estimator for a parameter (say, $\overline{Y}$) is Coefficient of Variation (CV), which is defined as follows

$$C.V. = \left(\frac{\sqrt{sampling \operatorname{var} iance}}{Parameter} \right)$$

and is estimated as

$$Est(CV) = \left(\frac{Est(SE)}{Estimate} \right)$$; where SE is the Standard Error.

Since it is a unit free measure, it is often used to compare the precision levels of estimators in different populations.

Confidence Intervals: Another useful concept associated with precision levels is the concept of Confidence Intervals. The concept of confidence limits and confidence interval is closely linked to interval estimation. A point estimate is a single value given as the estimate of a population parameter that is of interest, for example the mean of some quantity. An interval estimate specifies instead a range within which the parameter is estimated to lie. A confidence interval (CI) can be used to describe how reliable survey results are. For a given point estimate, 90% or 95% confidence intervals can be generated depending upon whether the level of confidence is 10% or 5%. When mean is to be estimated, in case of simple random sampling, sample mean $\overline{y}$ is a point estimate for population mean $\overline{Y}$ and the confidence limits are $\overline{y} \pm t_{\alpha} SE(\overline{y})$, where t_{α} is the t-value with $(1-\alpha)$ per cent level of confidence. At a given level of confidence, and all other things being equal, a result with a smaller CI is more reliable than a result with a larger CI. A major factor determining the width of a confidence interval is the size of the sample used in the estimation procedure.

Confidence intervals are closely related to statistical significance testing. In many situations, if the point estimate of a parameter is *X*, with confidence interval [*a*,*b*] at confidence level *P*, then any value outside the interval [*a*,*b*] will be significantly different from *X* at significance level α = 1 − *P*, under the same distributional assumptions that were made to generate the confidence interval.

Variance Estimation in Complex Surveys

In the estimation procedures corresponding to different sampling designs as described above, formulae for estimates of variances for estimated means/totals are provided. In the case of linear estimates it is simple. However, in more complex survey situations, it is not always possible to express the estimated variances in terms of simple formulae. Even in more familiar situations of estimating variances for ratio and regression estimators, which are non linear in nature, expressing variance estimators, similar to linear estimators is not feasible.

Several alternate methods for estimating variances in complex survey situations are available. Some of these methods are

1) Linearization (Taylor's series)
2) Random Group Methods
3) Balanced Repeated Replication (BRR)
4) Re-sampling techniques
   - Jackknife, Bootstrap

Taylor's Series Linearization Method

In this method, non-linear statistics are approximated to linear form using Taylor's series expansion. This involves expressing the estimate in terms of a Taylor's series expansion, and then approximating the variance of the estimate by the variance of the first-order or linear part of the Taylor series expansion. This method requires the assumption that all higher-order terms are of negligible size. If this assumption is correct, then the variance approximation works well. In this linearization approach to variance estimation, a separate formula for the linearized estimate must be developed for each type of estimator. We are already familiar with this approach in a simple form in case of ratio estimator.

Random Group Methods

This concept is based on the concept of replicating the survey design. The earliest form of this method is available in the concept of Interpenetrating samples. However it is usually not possible to replicate the survey. In such cases, survey can be divided into R groups so that each group forms a miniature version of the survey. Based on each of the R groups estimates can be developed for the parameter of interest θ,(say). Let $\hat{\theta}_r$ be the estimate based on r^th sample. Considering the groups as independent, an unbiased estimate of variance of $\hat{\theta}=\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$ is given as

$$\hat{V}\left(\hat{\theta}\right)=\frac{1}{R(R-1)}\sum_{r=1}^{R}\left(\hat{\theta}_r-\hat{\theta}\right)^2$$

Advantages of this method are that it is easy to calculate and it is a general method in which complex functions can be tackled easily. However, the assumption of independent samples may be somewhat restrictive, if samples are not selected independently.

Balanced Repeated Replications (BRR) methods

Consider that there are H strata with two units selected per stratum. There are 2^H ways to pick 1 from each stratum. Each combination could be treated as a sample. Pick R samples. Which samples should we include? Following steps may be followed:

- Assign each value either 1 or –1 within the stratum
- Select samples that are orthogonal to one another to create balance
- One can use the design matrix for a fraction factorial

- Specify a vector α_r of 1,-1 values for each stratum

An estimator of variance based on BRR method is given by

$$\hat{V}_{BRR}\left(\hat{\theta}\right)=\frac{1}{R(R-1)}\sum_{r=1}^{R}\left(\hat{\theta}(\alpha_r)-\hat{\theta}\right)^2$$

where $\hat{\theta}=\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}(\alpha_r)$

Some of the advantages in the method is that they are relatively few computations ant it is asymptotically equivalent to linearization methods for smooth functions of population totals.

Jack-knife Method

The method was initially developed (Quenouille (1949)) in the context of reducing the bias of ratio estimator. The procedure was to randomly divide the sample (SRS) in two parts. Each part could provide an estimate of the estimator (ratio estimator in this case). The third estimate could be developed from the whole sample. These three estimators could be combined linearly such that first order term in the bias expression vanishes. If units are dropped individually, the corresponding statistics was found (conjectured first by Tukey (1958)) to be uncorrelated. This property has been exploited for variance estimation.

Let $\hat{\theta}^i$ be the estimator of θ after omitting the ith observation. Define

$$\tilde{\theta}^i = n\hat{\theta}-(n-1)\hat{\theta}^i$$

The Jackknife estimate is given by $\hat{\theta}_J = \frac{1}{n}\sum\tilde{\theta}^i$

Jackknife estimator of the variance is given by

$$\hat{V}_J(\hat{\theta}_J)=\frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\theta}^i-\hat{\theta}_J)^2$$

Bootstrap Method

This is also a re-sample technique, in which large numbers of samples are selected by equal probability sampling with replacement from the main sample. Similar to the estimate for the main sample, independent estimates are prepared for each sample. Estimate of variance is obtained from the repeated samples. An advantage of this method is that estimates of variance for complex statistics like quantiles and median can be obtained.

All these variance estimation techniques are highly computer intensive. Most of the survey data analysis packages utilize one of these methods. Some of these packages are as follows:

- OSIRIS – BRR, Jackknife
- SAS – Linearization
- Stata – Linearization
- SUDAAN – Linearization, Bootstrap, Jackknife
- WesVar – BRR, JackKnife, Bootstrap

V. EVALUATION AND TREATMENT OF NON SAMPLING ERRORS

5.1 . Non Sampling Errors (NSEs)

In the previous discussions on survey design and estimation methodology, the focus was on sampling errors only. There are, however, other sources of variation in surveys caused by non-sampling errors. All survey data are subject to error from various sources. The difference in the true value of the parameter and survey results is an error due to one reason or the other. The sampling variance and mean square errors are measures of error due to sampling. All other types of errors from various sources are termed as non-sampling errors. Non-sampling errors arise mainly due to misleading definitions and concepts, inadequate frames, unsatisfactory questionnaires, defective methods of data collection, tabulation, coding, incomplete coverage of sample units etc. Sampling errors arise solely as a result of drawing a probability sample rather than conducting a complete enumeration. Non-sampling errors, on the other hand, are mainly associated to data collection and processing procedures.

5.1.1 Types of Non-sampling Errors

Non-sampling errors arise due to various causes right from initial stage when the survey is being planned and designed to the final stage when data are processed and analyzed. Some of the factors contributing towards Non-sampling error are as follows:

1) Data specification being inadequate and/or inconsistent with respect to objectives of the survey.
2) Duplication or omission of units due to imprecise definition of the boundaries of area units, incomplete or wrong identification particulars of units or faulty methods of enumeration.
3) Inappropriate methods of interview, observation or measurement using ambiguous questionnaires, definitions or instructions.
4) Lack of trained and experienced field enumerators including lack of good quality field supervision
5) Inadequate scrutiny of the basic data.
6) Errors in data processing operations such as coding, keying, verification, tabulation etc.
7) Errors during presentation and publication of tabulated results.

Five prominent components of NSEs are known as:

1. Specification errors,
2. Coverage errors,
3. Measurement or response errors,
4. Non-response errors and
5. Processing error.

These types or error are briefly discussed below:

Specification errors

This occurs when the concept implied by the question is different from the underlying construct that should be measured. A simple question such as whether a household is an agricultural household can be subject to different interpretations. A person may be doing agriculture as an own account holder, he may be involved in agricultural activities as a part time activity. The meaning of the questions must be conveyed in an unambiguous way and must be properly understood by the respondent. Unless the right screening and filter questions are included in the questionnaire, the answers may not fully bring out the message behind the question.

Coverage errors

In most area surveys primary sampling units comprise clusters of geographic units generally called enumeration areas (EAs). It is not uncommon that the demarcation of EAs is not properly carried out during census mapping. Thus households may be omitted or duplicated in the second stage frame. Updating of EA boundaries before the conduct of agricultural census becomes very important. Cartography of EAs is normally available from the population censuses, but updating of the selected EAs is an essential part of the cartography for agricultural censuses. Otherwise, exclusion of sample units in some EAs and duplication of units in other EAs are highly probable. Frame imperfections can bias the estimates in several ways: If units are not represented in the frame but should have been part of the frame, this results in zero probability of: selection for those units omitted from the frame. This leads to under-coverage. On the other hand if some units are duplicated; this results in over-coverage with such units having larger probabilities of selection.

It is important to note that sometimes there is a deliberate and explicit exclusion of sections of a larger population from survey population. Survey objectives and practical difficulties determine such deliberate exclusions. For example, when we define the agricultural households by putting certain cut-offs, some households are deliberately excluded. When computing non-coverage rates, members of the group deliberately and explicitly excluded should not be counted either in the survey population or under non-coverage. In this regard defining the survey population should be part of the clearly stated essential survey conditions. Non-coverage is often associated with problems of incomplete and faulty frames. If the flames are not updated or old frames are used as a device to save time or money, it may lead to serious bias.

The most effective way to reduce coverage error is to improve the frame by excluding erroneous units and duplicates and updating the frame through filed work to identify units missing from the frame. It is also important to undertake a good mapping exercise during the preparatory stages of a population and housing census. However, the frame prepared during the census should be updated periodically. It is also imperative to put in place procedures that will ensure the coverage of all selected sample units.

Measurement errors

These errors arise from the fact that what is observed or measured departs from the actual values of sample units. These errors centre on the sustentative content of the survey such as definition of survey objectives, their transformation into able questions, and the obtaining, recording, coding and processing of responses. These errors concern the accuracy of measurement at the level of individual units. When we get responses from the selected units through a questionnaire and there the responses are different than the true values, these errors are called response errors. Inadequate instructions to field staff and inadequate training normally lead to response errors.

Mathematical treatment of measurement errors is available in the form of linear response error models (Refer Cochran, W. G. (1977)). Such models have also been used in the treatment of interpenetrating net-work of sub-sampling which is used for estimating the enumerators’ effect. The mathematical details are not given here.

Non-response errors

Non-response refers to the failure to measure some of the sample units. Thus failure to obtain observations on some units selected for the sample. It is instructive to think of the sample population as split into two strata, one consisting of all sample units for which measurements can be obtained and the second for which no measurements could be obtained.

In most cases non-response is not evenly spread across the sample units but is heavily concentrated among subgroups. As a result of differential non-response, the distribution of the achieved sample across the subgroups will deviate from that of the selected sample. This deviation is likely to give rise to non-response bias if the survey variables are also related to the subgroups. While non-response can not be completely eliminated in practice, it could be overcome to a great extent by persuasion or by some other methods. One way of dealing this problem was due to Hansen and Hurwitz (1946). In this method the population was conceived as divided in two strata – respondents and non respondents. From the non respondents, a sub-sample is selected and special efforts are made to get response from these units. An estimation procedure is developed on the basis of suitably pooling the results of respondent and non-respondent groups. Yet another technique was developed by Politz and Simon (1949) for reducing the bias without call backs by asking to the respondent as to how many times he was at home during previous week.

There are two types of non-responses: unit non-response and item non-response. Unit non-response implies that no information is obtained from certain sample units. This may be because respondents refuse to participate in the survey when contacted or they cannot be contacted. Item non-response refers to a situation where for some units the information collected is incomplete. Item non-response is therefore, evidenced by gaps in the data records for responding sample units. Reasons may be due to refusals, omissions by enumerators and incapacity.

Causes of non-response

Respondents to provide information can cause non-response error if they are being not- at home or by sample units not being accessible. This introduces errors in the survey results because sample units excluded may have different characteristics from the sample units for which information was collected. Refusal by a prospective respondent to take part in a survey may be influenced by many factors, among them, lack of motivation, shortage of time, sensitivities of the study to certain questions, etc.

Errors arise from the exclusion of some of the units in the sample. This may not be a serious problem if the characteristics of the non-responding units are similar to those of the responding units. But such similarity is not common in practice.

With specific reference to item non-response, questions in the survey may be perceived by the respondent as being embarrassing, sensitive or/and irrelevant to the stated objective.

The enumerator may skip a question or ignore recording an answer. In addition, a response may be rejected during editing. For sensitive questions a technique of randomized response is available.

In personal interview surveys, the enumerator can play an important role in maximizing response from respondents. The way interviewers introduce themselves, what they say about the survey, the identity they carry, and the courtesy they show to respondents matter. In most surveys the enumerator is the only link between the survey organization and respondent. It is for this reason that enumerators and their supervisors should be carefully selected, well trained and motivated. Close supervision of enumerator's work and feedback on achieved response rate is of paramount importance.

Processing errors
Processing errors comprise:
— Editing errors.
— Coding errors.
— Data entry errors.
— Programming errors etc.
The above errors arise during the data processing stage. For example in coding open ended answers related to economic characteristics, coders may deviate from the laid out procedures in coding manuals, and therefore assign wrong codes to occupations. In addition, the weighting procedures may be wrongly applied during the processing stage, etc.

5.1.2 Interpenetrating sub-sampling

It is worthwhile to mention about this technique which was initially developed (Mahalanobis (1946)) in the context of study of correlated errors. In this technique a random sample of n units is divided at random into k sub-samples, each sub sample containing m=n/k units. The field work and processing of the sample are planned so that there is no correlation between the errors of measurement of any two units of in two different sub-samples. For instance, suppose that the correlation with which we have to deal arises solely from biases of the enumerators. If each of k enumerators 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 this technique. With a suitable model it is possible to estimate the relative amount which the correlated component (in this case due to interviewer's effect) of the response variance contributes to the total variance.
The technique has also been very helpful in estimation of variances for complex statistics.

5.1.3 Evaluation of non-sampling errors

Consistency checks
In designing the survey instruments (questionnaires), care should be taken to include certain items of information that will serve as a check on the quality of the data to be collected. If the additional items of information are easy to obtain, they may be canvassed for all units covered in the survey, otherwise, they may be canvassed only for a sub-sample of units.
It is also desirable to follow some external consistency checks on salient results thorough comparable data sources. It is important for validity as well as acceptability of the estimates.

Sample check/verification

One way of assessing and controlling non-sampling errors in surveys is to independently duplicate the work at the different stages of operation with a view to facilitating the detection and rectification of errors. For practical reasons the duplicate checking can only be carried out on a sample of the work by using a smaller group of well-trained and experienced staff If the s is properly designed and if the checking operation is efficiently carried out, it would be possible, not only to detect the presence of non-sampling errors, but also to get an idea of their magnitude. If it were possible to completely check the survey work, the quality of the final results could be considerably improved. With the sample check, rectification work can only be carried out on the sample checked. This difficulty can be overcome by dividing the output at different stages of the survey, e.g. filled in schedules, coded schedules, computation sheets, etc., into lots and checking samples from each lot. In this case, when the error rate in a particular lot is more than the specified level, the whole lot may check and corrected fir the errors, thereby improving the quality of the final results.

Post-survey checks

An important sample check, which may be used to assess non-sampling errors, consists of selecting a sub-sample, or a sample in the case of a census, and re-enumerating it by using better trained and more experienced staff than those employed for the main investigation.

Usually the check-survey is designed to facilitate the assessment of both coverage and content errors. For this purpose, it is first desirable to re-enumerate all the units in the sample at the high stages, e.g. EAs and villages, with the view of detecting coverage errors and then to re survey only a sample of ultimate units ensuring proper representation for different parts of the population which have special significance from the point of view of non-sampling errors.

VI. BULDING AND USING AREA SAMPLING FRAMES FOR AGRICULTURAL CENSUSES AND SURVEYS

General presentation on Building, Using and Maintaining Sampling Frames: Area Frame, Multiple Frame, Master Sampling Frames, ppt.