SAMPLING DESIGN AND ESTIMATION

A household is a group of persons (or a single person) who usually live together and have a common arrangement for food, such as using a common kitchen or a common food budget. The persons may be related to each other or may be non-relatives, including servants or other employees, staying with the employer.

Sampling Design

The preparation of separate estimates for the different geographic sub-divisions of the country was an important consideration in determining the sampling design. The gains in efficiency through stratification based on economic and geographic criteria would result in lowering sampling errors when the number of strata were increased. In order to produce separate estimates for urban and rural sectors, they were treated as two domains. The capital city of Ulaanbaatar, which carried almost one third of the country's population, was accepted as a separate stratum. The Central, East, West and Khangai Regions of the country were treated as separate domains. The sub-division of these regions in terms of their urban - rural stratifications resulted in 8 strata.

A two-stage probability sample design with enumeration areas as primary sampling units (PSU)'s and households as the secondary sampling units (SSU's) was adopted. Apart from these major innovations, circular systematic sampling with probability proportional to size (CSSPPS) techniques were adopted in the selection of enumeration areas which were the primary sampling units (PSU's) and households which formed the secondary sampling units (SSU's). The sampling strategy adopted is described in the paragraphs that follow.

Sampling Frame

The sampling frame from the Census of Population 2000 was used as the sampling frame. There had been no major changes in the boundaries of geographic sub-divisions of Mongolia accept for a re-grouping of aimags (provinces) to create the Hangai region suppressing the Southern region. There were enumeration areas without any households, with mainly male population figures recorded against them, which were demarcated with institutional living quarters. These were checked and were deleted from the frame. In the Census 2000 sampling frame, the basic unit of enumeration was the enumeration area, with identifiers and measure of size in households and population totals and breakdowns by males and females. The census enumeration areas were chosen as primary sampling units (PSU) in both the urban and rural areas.

Sample Size Determination

The inclusion of such topics as unemployment, child labor demonstrated that the sample size should be adequate to produce statistically reliable estimates for the main stratifications. Although several household surveys have been conducted during the past decade, it does not appear that the survey dimensions have been determined on the basis of desired precision of the estimates. The unemployment rates disclosed in the census 2000 and from other surveys conducted recently were used in deriving the coefficient of variation in the unemployment estimate. These levels of precision with assumed values of DEFF was used in ascertaining the sample size that will be required to produce the desired level of precision.

Survey Taking Capacity

The discussions with the NSO officials reflected that the main consideration in deciding on the sample size is funding rather than other considerations including staff capacity. Thee previous experience of undertaking sample surveys during the 5-7 years indicated that the NSO had been able to successfully carry out the implementation of several surveys. Further the Census of Population 2000 had been undertaken while some household surveys and regular data collection and compilation through the administrative network had been on-going. The availability of statistical staff of the *aimags* for field operations and the decentralized arrangements used in data processing have made it possible to extend the survey taking capacity much beyond what would have been feasible if the NSO was dependent on its own cadre for the entirety of survey operations.

Sample Selection

First Stage Selection

In the first stage the EA's or primary sampling units (PSU's) were drawn from each stratum. The frame was arranged so that all the enumeration areas within a stratum were listed in the order of aimag, soum, and komiss, with their identification codes and the number of households in the EA. The number of households in the enumeration area M_{hi} was used as the measure of size (MOS) in the probability proportional to size method of selection of first stage units. The method of circular systematic sampling with probability of inclusion of an enumeration area proportional to its size (CSSPPS) method was used to select the sample of wards from each stratum. The procedure adopted is described below.

The selection probability for enumeration area i in stratum **h** is given by the formula,

$$p_h^{(i)} = a_h M_{hi} / M_h \qquad \text{(Eq. 1)}$$

where

a_h = number of EA's or PSU's to be drawn from the stratum

M_{hi} = number of households in the i^{th} EA as reported in the frame

M_h = $\sum M_{hi}$ = total number of households in the stratum as recorded in the frame

The selection of PSU's was performed by arranging the EA's in the **h** th stratum according to aimag, soum, and komiss and the estimated number of households was used as the measure of size M_{hi}. The values of M_{hi} were then cumulated and Cu M_{hi} was recorded against each PSU. The sampling interval I_{h1} was computed which is given by

$$I_{h1} = M_h / a_h \quad \text{rounded of to the nearest integer.}$$

A random number R_h that falls between 0 and M_h was then selected using the random number generator in the Excel programme. The sequence of a_h selector numbers were generated by the addition of I_{h1} to the previous number selected. If the total exceeds $M_{h,}$ then M_h was subtracted from the total to derive the number.

Let $R_{h1} = R_h$ and for $j = 2, 3\ldots\ldots a_{h,}$ R_{hj} ,= $R_{hj-1} + I_{h1}$, if this does not exceed $M_{h;}$ R_{hj} ,= $R_{hj-1} + I_{h1} - M_h$ otherwise.

Accordingly, the selector numbers will be of the form

$R_{h,}$ $(R_h + I_{h1})$, $(R_h + 2\ I_{h1})$, $(R_h + 3\ I_{h1})$, $(\ R_h + 4\ I_{h1})$, ... $(R_h + (j\text{-}1)\ I_{h1})$, $(R_h + (a_h\text{-}1)\ I_{h1})$. when R_{hj} ,= $R_{hj-1} + I_{h1}$, does not exceed $M_{h.}$ The expressions should be replaced with the terms R_{hj} ,= $R_{hj-1} + I_{h1} - M_h$ when R_{hj} exceeds $M_{h.}$

Selection of EA's in the first stage by CSSPPS was done using Excel rogramme. The details of the samples selected for Ulaanbaatar and the urban and rural sub-divisions of the 4 regions are copied to PC' of the SSD staff members.

Circular Systematic Sample with Equal Probabilities of selection, **CSSEQP** was used for the selection of households from a selected EA, is a simpler version of **CSSPPS.**

The list of households in the sample EA prepared at the house listing stage was used as the frame, and a sample reference number was assigned sequentially to each household. The last number assigned should be equal to the total number households in the enumeration area M_{hi}^***.** Then the probability of selecting a household in the **i** th PSU in the **h** th domain is

$$p_h^{(j/i)} \quad = \quad n_h / M_{hi}^* \qquad \text{(Eq. 2)}$$

where n_h is equal to 10 in this instance. The sampling interval $I = M_{hi}^* / 10$ was computed and rounded off to the nearest integer. I should be computed after the listing operation when the actual number of households is determined. A random number R_{hj} in the interval 1 to M_{hi}^* was taken as the first selector number. The remaining 9 selector numbers were calculated one after the other by adding I to the previous number. If the sum exceeded M_{hi}^* the remainder after subtracting M_{hi}^* from the sum was taken as the selector number. These selector numbers were the serial numbers of the selected households. Selection of households from the sampled EA's can be done in the EA itself by enumerators under the supervision of supervisors.

The design provides for estimators to be computed for the 9 strata, namely Ulaanbaatar, urban and rural areas sub-divisions of the 4 regions into which Mongolia is divided. These estimates are in respect of the all four quarterly rounds of the survey. However, some estimates will have to be prepared based on the quarterly rounds of the

survey. The method to be applied is the same. Most of the estimators that will be computed from the survey will be ratio estimates but frequently estimates of stratum totals are required for use by policy makers and administrators. The estimation procedure for these estimators are set out in the paragraphs
that follow.

Design Weights

The design weights are used to compensate for differences in the selection probabilities. The weight for the PSU is inversely proportional to its selection probability.

The probability of selection of **j** th household in normal size PSU's and blocks in the **h** th domain is

$$p_h^{(i)} \times p_h^{(j/i)} = p_h^{(ij)} \qquad \text{(Eq. 3)}$$

$$\text{where} \quad p_h^{(i)} = a_h M_{hi} / M_h$$

$$\text{and} \quad p_h^{(j/i)} = n_h / M_{hi}^{*}$$

Thus the design weights w_{hij} for households are

$$w_{hij} = 1 / p_h^{(ij)}$$

$$= \frac{M_h \times M_{hi}^{*}}{a_h \times M_{hi} \times n_h} \qquad \text{(Eq. 4)}$$

The design for LFS is not self-weighting and therefore it is necessary to compute weight for each PSU selected in the sample and these weights have to be used in the estimation procedure.

Estimation Procedure for Household Information

The estimate of the stratum total of a characteristic y is given by the following formula.

$$\hat{Y}_h = \sum_i \sum_j w_{hij}\, y_{hij} \quad \text{for} \quad \begin{array}{l} i = 1, 2, 3, \ldots\ldots\ldots\ldots a_h \\ j = 1, 2, 3, \ldots\ldots\ldots n_{hi} \end{array} \qquad \text{(Eq. 5)}$$

where

$\hat{Y}_{\mathbf{h}}$ = estimate of characteristic y for stratum h

y_{hij} = any characteristic of person k in household j in sample enumeration area i in stratum h

n_{hi} = number of sample households in enumeration area i

a_h = number of sample enumeration areas in stratum h

$w_{hij} = 1 / f_h$

$f_h = 1 / w_{hij}$

The estimate for the total for all 9 strata $\hat{Y}$ was computed as the sum of the estimates for each domain viz.

$$\hat{Y} = \sum \hat{Y}_h \qquad h = 1, 2, 3, \ldots .9. \qquad \text{(Eq. 6)}$$

Most of the estimators to be computed from the LFS are in the form of averages and proportions. In general these estimators are combined ratio estimators which take the form set out below. The estimated stratum mean is a ratio and it is given by

$$r_h = \frac{\hat{Y}_h}{\hat{X}_h} = \frac{\sum_i \sum_j w_{hij} \; y_{hij}}{\sum_i \sum_j w_{hij} \; x_{hij}} \qquad \text{(Eq. 7)}$$

where

$y_{hij},$ $a_{h},$ $n_{hi},$ w_{hij} are as defined earlier.

$x_{hij} = 1$ for $j = 1, 2, 3, \ldots\ldots\ldots.n_{hi}$
$i = 1, 2, 3, \ldots\ldots\ldots.a_h$

The population mean is also a ratio, say r, which was estimated using the following formula.

$$r = \frac{\sum_h \sum_i \sum_j w_{hij} \quad y_{hij}}{\sum_h \sum_i \sum_j w_{hij} \, x_{hij}} \qquad \text{(Eq. 8)}$$

where

y_{hij} , a_h , n_{hi} , w_{hij} are as defined in Eq. 7

X_{hij} is as defned in Eq. 7

Estimation of Variances and Standard Errors

The computation procedure will be incomplete without establishing the procedure for assessing the precision or reliability of the survey estimates. The variances of the ratio estimates will be of the form

$$\mathrm{var}\,(r) \;=\; \frac{1\,/}{\mathbf{X}^2}\;\sum\,(\,1-f_h)\;\;(a_h\,/\,a_h\;-\;1)\;\sum\,(\,z_{hi}^2\;-\;z_h^{\;2}\,/\,a_h) \qquad (\text{ Eq. } 9\text{ })$$

where

$r \;=\; \mathbf{y}\;/\;\mathbf{x}$

$\mathbf{y}_{hi} \;=\; \sum_j w_{hij}\;\; y_{hij}$

$x_{hi} \;=\; \sum_j w_{hij}\;\; x_{hij} = \sum_j w_{hij}\;\; x_{\,hij}$

$r \;=\; \sum\sum\sum w_{hij}\;\; y_{hij}\;\;/\;\;\sum\sum\sum\;\; w_{hij}\;\; x_{hij}$

$\hat{x}^2 \;=\; X^2 \;=\; (\;\sum_h\sum_i\sum_j w_{hij}\;\; x_{hij}\;)^2$

$z_{hi} \;=\; y_{hi}\; - \; r\,x_{hi}$

a_h = number of sample enumeration areas from stratum h

w_{hij} = weight for each individual in the sample household

Variance of Ratio of rh in Stratum h

The variance of ratio estimate r_h in stratum h is of the form:

$$\mathrm{var}\,(r_h\,) \;=\; (1\,/\;\;x_h^{\;2}\,)(1-f_h)\;\;(a_h\,/\,a_h\;-1)\;\sum\,(\,z_{hi}^2 -\;z_h^{\;2}\,/\,a_h)\quad (\text{ Eq.10 })$$

where

$\hat{X}_h \;=\; x_h \;=\; \sum_i\sum_j\;\; w_{hij}\;\; x_{hij}$

and f_h, a_h, and z_{hi} are as defined earlier.

Standard Error and Coefficient of Variation

The standard error of a survey estimate provides a measure of how far the survey estimate is likely to vary from the true population value (i.e. parameter) as a result of having collected the data on a sample basis rather through a complete census. The standard error se(r) of a survey estimate is by definition

$$\mathrm{se}(\,r\,) \;=\; \mathrm{var}(\,r\,)^{1/2}$$

The relative standard error or coefficient of variation (cv), on the other hand provides a measure of the relative variance of a survey estimate; that is the magnitude of the estimated sampling error relative to the magnitude of the estimate itself. The cv that is expressed as a proportional error enables the data user to compare the relative reliability or precision with which different types of survey characteristics have been measured eg. Means versus proportions, where direct comparisons of standard errors are uninformative since the magnitude of the standard error is dependent upon the magnitude of the estimate.

Computationally, the coefficient of variation is calculated as

$$cv(r) = se(r) / r.$$

Since only a sample of enumeration areas were included in the LFS the estimates prepared from the survey are subject to sampling errors. The sampling error indicates the extent to which an estimate from the LFS would vary by chance because only a sample of EA's is included rather than all the EA's into which the country is divided. The sample size and survey design determine the magnitude of sampling errors and in respect of some items the sampling errors are expected to be high and the users are cautioned to note this fact in using the data.