Sample Design

Parameters of Sample Design

The Yemen MICS sample design was a two-stage stratified cluster sample. The following parameters were accounted for in designing the sample:

1- The sample is to provide estimates with reasonable precision at national and urban/rural levels.
2- The residents of the Yemeni islands and the nomadic population are excluded from survey coverage.
3- The size of ultimate cluster is 20 households
4- It is approximately self-weighted design.

Determination of Sample Size

The sample size has been figured out on the basis of the recommendations given in the MICS Manual of "Designing and Selection the Sample". The size of the Sample has been estimated using the following formula:

$$n = \frac{4r\,(1-r)\,f\,(1.1)}{(0.12r)^{2}\,p\,n_h}$$

where:

n: is the required sample size
4: is a factor to achieve the 95% confidence level
r: is the predicted or anticipated prevalence (coverage rate) for the indicator being estimated
1.1: is a factor necessary to raise the sample size by 10% for non-response
f: is a shortened symbol for design effect (*deff)*
0.12r: is the margin of error to be tolerated at the 95% level of confidence, defined as 12 percent of r (12 percent thus represents the relative sampling error of r)
p: is the proportion of the total population upon which the indicator, r, is based, and
n_h : is average household size.

As the percentage of immunized children aged between 1-2 years is one of the most important indicators that the survey aims to estimate, it will be relied upon to determine the sample size. The Family Health Survey (PAPFAM) conducted in Yemen in 2003 revealed that the percentage of fully immunized children in the age group 12-23 months is 37.2%. Based on the same survey, the proportion of children in this age group (p) is 0.031 approximately. The 2004 Population Census indicates that the average household size (n_h) is 7.1 persons. Assuming that the design effect (f) is about 1.5, the sample size has been estimated as 3516 households. It was deemed useful to increase the sample size to 4000 households so as to provide

estimates for urban and rural strata with no much less precision. Increasing the sample will also be useful in case of measuring phenomena with less prevalence at the national level. In other words if a prevalence of another phenomenon measured in the survey is higher than 37.2%, the sample will provide more precise estimates (less sampling errors) for the prevalence rates of such phenomena. Conversely, if a phenomenon is less prevalent than the 37.2% level, the sample will provide an estimate of the prevalence rate with lower precision than that of the immunization rate.

Sample allocation

The sample is allocated proportionally between urban and rural strata; the percentage of households that should be allocated to urban and rural areas was obtained from the 2004 Census. As the ultimate cluster is determined to be 20 households, the number of sample clusters is therefore 200. The proportional allocation of the sample is such that 142 for rural stratum and 58 for urban stratum.

Sample Selection

The sample is to be selected in two stages. The Primary Sampling Unit (PSU) is a village (or a group of villages) in rural areas and a lane (hara) in urban. The micro data of the 2004 Census at these administrative levels has been relied upon to create frames for the first stage sample. The following provides a description of the sample selection in both stages:

First Stage Sample

The 2004 Census data (numbers of households and population) for all urban and rural agglomerations have been utilized to create appropriate frames for the first stage sample of urban and rural strata. It was taken into account that the PSU size would be in the range 150-300 households approximately. The creation of a rural frame has entailed grouping neighboring small villages so as to create PSUs in the range of 150-300 households each. Hence, a rural PSU is in most cases a group of small villages. The whole village is considered a PSU as long as its size is in the range 150-300 households.

The situation in urban areas is quite different from rural areas since most lanes (Haras) are much larger than the indicated range of the desired PSU size. For this reason, a second (dummy) sampling stage is necessary to reduce the burden of field listing whenever the lane size is above 300 households. The first urban stage sample included 41 PSU's that required division into equally sized parts. Whereas only 4 PSU's in the rural sample needed to be divided into equal parts.

An implicit stratification has been introduced in both rural and urban frames of the PSUs. Governorates were ordered geographically in a serpentine fashion starting from the northwest corner moving to the northeast corner and back to the west, then to the east and so on till the last governorate. Moreover, as governorate are further

divided into a number of directorates (modyriate), another process of implicit stratification within each governorate was implemented by geographically ordering directorates following the same way as for governorates. Undoubtedly, implicit stratification will contribute to more precise sample estimates at both national and urban/rural levels.

The selection of rural and urban first stage samples was made following the Probability Proportionate to Size (PPS) selection method. The employed measure of size (MOS) is the number of Households in each PSU as measured in the 2004 Census.

Accordingly, the probability of selection of the first stage sample can be represented as follows:

$$p(\alpha) = \frac{\lambda M_\alpha}{\sum_\alpha M_\alpha}$$

Where:

$p(\alpha)$ is the probability of selecting the α^{th} PSU in the sample

λ is size of first stage sample : λ = 58 in Urban, and λ = 142 in Rural

M_α is the number of households of the α^{th} PSU

As indicated above, a second (dummy) sampling stage is implemented in larger urban PSU's. The large urban PSU selected in the sample is divided into equal parts of about 150-300 households each. The cartographic facilities of the MOPHP as will as its Geographic Information System (GIS) have been utilized in dividing such PSU's into parts of equal population size. One part was then selected with equal probability method. If the α^{th} large urban PSU was selected in the sample, it was then divided into "k" parts of equal size, the selection probability of a certain part in the sample is defined as: (1/k)* $P(\alpha)$.

The distribution of the first stage urban and rural samples according to governorates is shown in table 1, while the lists of first stage urban and rural samples are given in the appendix.

Table1
Distribution of Urban/Rural first stage sample by governorates

Governorate	Urban Sample	Rural Sample	Total
Sa'da	1	5	6
Algouf	1	3	4

Hadarmout	4	5	9
Ma'reb	-	2	2
Sana'a	2	7	9
El Amana	17	1	18
Omran	1	6	7
Heggah	2	12	14
Al Hodayedah	8	19	27
Al Mahweet	-	5	5
Remah	-	3	3
Damar	2	12	14
Al BAydaa'	1	4	5
Shabowa	-	3	3
Abyan	1	3	4
Al Dhalee'	1	4	5
Ibb	4	21	25
Taez	6	20	26
Aden	7	-	7
Laheg	-	7	7
Total	**58**	**142**	**200**

Evidently, the above table shows that the distribution of first stage sample among different governorates is well balanced. Few governorates were not represented in the urban sample because of the extremely lower weight of their urban populations relative to the total urban population of the country.

Second stage sample

The selected PSU from the first sample stage, whether it was the whole PSU or a part of one, was updated in the field. A field operation was carried out in each PSU (or a part of it), which has been selected in the first stage sample so as to create an updated list of households for each sample PSU. These lists were used as sample frames for selecting the second stage sample.

The proposed selection method was determined in such a way so as to create compact ultimate clusters of 20 households in the rural sample, and non-compact ultimate cluster of the same size in the urban sample. The reason for selecting compact clusters for rural sample is that most of the rural sample PSU's are composed of several small villages which are, in most cases, located at the tops of adjacent mountains. The spread of the household sample over several small villages, within the same PSU, that would result from the systematic selection, would impose much difficulty in the main survey fieldwork. Hence it has been deemed operationally efficient to deal with the household list for each rural sample PSU as forming a circle. The selection of a single random number in the

range of 1 - the total number of households in the list, will determine the entire household sample to be selected from the sample PSU. The household indicated by the selected random number and the subsequent 19 households in the list constitute the household sample to be selected from rural sample PSU's (keeping in mind the circular nature of the list).

As an example, assume that the list of a certain rural sample PSU includes 220 households. The selected random number (in the range of 1-220) is 206. Therefore, the household sample constitutes the households with the serial numbers:

206-207-208-209-210-211-212-213-214-215-216-217-218-219-220-1-2-3-4-5.

In the case of the urban sample, however, an ordinary random systematic selection is suggested, so as to produce a non-compact cluster of 20 households. The households forming urban PSU (or a part of it) are not dispersed over a large area; hence the compact cluster is not justifiable.

The conditional selection probability of a certain household given the selection of the PSU in the first stage sample is given as follows:

$$P(\beta \mid \alpha) = \frac{20}{M_{\alpha}^{*}}$$

Where $P(\beta \mid \alpha)$ is the selection probability of the β^{th} household given that the α^{th} PSU was selected in the first stage sample, M_{α}^{*} is the updated number of households of the α^{th} PSU(or a part of it).

Sampling Rate

The overall sampling rate is the non-conditional probability of selecting a given household in the sample. It is given by the following formula:

$P(\alpha\beta) = P(\alpha)P(\beta \mid \alpha)$, where :

$p(\alpha)$ is the probability of the first stage sample, i.e, $p(\alpha) = \frac{\lambda M_\alpha}{K\sum_\alpha M_\alpha}$, and k = number of parts of equal size into which the PSU is divided (urban sample), $K = 1$ if the PSU is not divided. $P(\beta \mid \alpha) = \frac{20}{M_\alpha^*}$. Thus :

$$P(\alpha\beta) = \frac{\lambda M_\alpha}{K\sum_\alpha M_\alpha}\frac{20}{M_\alpha^*}$$

Evidently the sample is strictly self-weighted if $M_\alpha = M_\alpha^*$ for all sample PSU's. Since the updating process will most probably result in a different PSU size, the sample is approximately self-weighted as long as the updated PSU size does not deviate very much from the non-updated (census) size.

Sample Weights

Weights were used in deriving survey estimates to account for the expected differences between the updated household lists of the sample PSU and the Measure of Size (the 2004 number of households) as well as non-response which is inevitable in surveys of this nature. If non-response varies substantially over the sample PSU's weights are needed for data tuning. The final weight (W) is the product of design weight (W_1) and non-response weight (W_2), where the design weight is the inverse of the overall selection probability and the non-response weight is the inverse of response rate.
Thus:

$$W = \frac{1}{W_1}\frac{1}{W_2}$$

$$= \frac{1}{P(\alpha\beta)}\frac{1}{\text{response rate}}$$

Where:

$P(\alpha\beta)$ is as defined above and the respnse rate $= \dfrac{\text{number of surveyed households}}{\text{number of sample households}}$