

Annex A

SAMPLE DESIGN AND IMPLEMENTATION¹

Introduction

Nepal's first labour force survey, Nepal Labour Force Survey 1998/99 (NLFS-I) was carried out by the Central Bureau of Statistics (CBS) in 1998/99. The survey provided comprehensive information on employment and labour market conditions of the country. For conducting the survey, technical inputs were supplied by the International Labour Organization (ILO). The survey covered the broad areas of current and usual economic activities, informal sector activities, economic activities of children, non-economic activities, unemployment and underemployment.

During the last nine years, Nepal encountered great challenges towards the development of physical infrastructure, poverty reduction, unemployment and other social phenomenon due to a decade long insurgency. So, it is deemed necessary for assessing a follow up survey of the current labour situation for comparing with benchmark NLFS 1998/99. Further the demand of this survey is realized everywhere by the government, various development organizations and stakeholders involved in planning and policy formulation.

CBS has conducted the Nepal Labour Force Survey 2008 (NLFS-II) mainly with its own resources; however, the external assistance is being used partially. The survey results are very useful to monitor the current employment policies, vocational training and income generation schemes that are directly related to the poverty alleviation programs. Besides this, it will also assist in explaining demographic changes, economic upheavals and impacts of the conflicting situation in the country.

This note summarizes the sampling procedures employed to carry out NLFS-II comparable to the previous NLFS-I. It comprises the process of determining sample size and describing the characteristics of the sample, its size and distribution, and the possible effects on analysis. Moreover, it describes the steps involved in selecting the Primary Sampling Units (PSUs), the allocation of PSUs by season to the field teams, and the method of calculating raising factor to be used in analysis of the survey data.

Sample design

The sampling design adopted in NLSS-II is slightly modified from NLFS-I. The design is based on two stages stratified sampling technique with equal PSUs or households distributed between urban and rural areas as done in NLFS-I considering the heterogeneous labour force activities to provide a detailed picture of employment situation in the urban areas. So the prescribed 800 PSUs are divided equally in two parts, i.e., 400 PSUs each for urban and rural. Urban areas are stratified into three strata as Urban Kathmandu Valley, Other Urban Hills and Urban Tarai, and rural areas are stratified into Rural Mountains, Rural Hills and Rural Tarai. The sample size of 400 PSUs in each urban and

¹ This note is prepared by the Household Survey Section of CBS during the preparatory work of Nepal Labour Force Survey 2008.

rural area will be proportionately distributed within their respective strata. In the first stage, the prescribed PSUs (wards or sub-wards or combination of wards) are selected by Probability Proportional to Size (PPS) from each stratum, i.e. the number of households in the ward as the measure of the size. Using PPS at the first stage is followed by selecting a fixed number of households at the second stage that provides a convenient workload for each field team. This design has the benefit that the sample, in principle, is self-weighted. In the second stage, 20 households are selected from each PSU (ward) of urban and rural strata with systematic random sampling procedure giving a total of 16000 households from 800 PSUs.

The minimum sample size required is estimated as 244 households for urban and 361 households for rural domain based on the currently economically active population taken from Nepal Living Standards Survey 2003-04. Based on employment status of the population of this fresh survey information, the minimum households to be selected turns out to be 265 for urban and 277 for rural. However, 400 households each in urban and rural area have been allocated in this survey to provide reliable estimates for lower level of disaggregation up to 5 years age group (10 groups) by sex (2 groups). There will be altogether 8000 households in each urban and rural area.

Sample frame

The sample frame is updated for this survey, which was developed during the sampling design of the recent Living Standards Survey 2003/04, which was based on data from the 2001 National Population Census.

The size of each ward (as measured by number of households) was taken as a unit of sample frame. Some larger wards were divided into smaller units (sub-wards) of clearly defined territorial areas supported by reliable cartography while some of the smaller wards with fewer than 30 households were appended to adjoining wards in the same VDC to make size equal to or greater than 30 households in the frame. The resulting sampling frame consisted of 35,069 enumeration areas (wards or sub-wards or merged wards) spread over 3 ecological zones, 5 development regions, 75 districts, 58 Municipalities and 3,914 Village Development Committees (VDCs) of the country. The sample frame was sorted by district, VDC, ward and sub-ward and districts were numbered from East to West.

Table 1: Primary sampling units of the NLFS 2008 by development region and zone

Ecological Zone	Development Region					Total
	East	Central	West	Mid West	Far West	
Nepal	176	329	152	81	62	800
Mountains	11	15	1	6	8	41
Hills	43	201	102	34	22	402
Tarai	122	113	49	41	32	357
Urban	78	203	67	27	25	400
Mountains	3	4	0	0	0	7
Hills	11	152	47	6	7	223
Tarai	64	47	20	21	18	170
Rural	98	126	85	54	37	400
Mountains	8	11	1	6	8	34
Hills	32	49	55	28	15	179
Tarai	58	66	29	20	14	187

A major consideration in the design is to ensure adequate samples in the different geographical subgroups that are likely to be used for the analysis of the survey data. Certainly data will need to be provided separately for each season of the year, and separately for urban and rural areas. For the survey purpose, the whole year is divided into three seasons namely: rainy (mid-May to mid-September), winter (mid-September to mid-January) and dry (mid-January to mid-May). Estimates will also be required for each of the three ecological zones and five development regions. However, if estimates were required for each subgroup (i.e. season by urban/rural by ecological zone by development region) this would involve a **3 X 2 X 3 X 5** table with 90 cells, though 9 of the cells would obviously be empty since the Mountains of West, Mid West and Far West development regions do not contain any urban households.

Selecting the wards from the sampling frame

The existing sample frame provides an excellent starting point for drawing the sample. All the wards are listed in ascending order of district, and sorted by village development committees (VDCs)/municipalities and ward codes in the sample frame. The following steps are taken in series:

1. The sample frame has been split out into six strata consisting of three urban and three rural strata, each in ascending order of district code and sample selection procedure is applied separately for each stratum with prescribed number of PSUs.
2. The enumeration areas are ordered by districts separately for each stratum, so as to separate groupings required for analysis. We are introducing an implicit stratification, which should help to ensure that the distribution of the final sample is extremely close to the expected distribution. In this new grouping, all the districts located within a particular intersection of development regions and ecological belts (e.g. the Mountains of Western development region) are placed next to each other.
3. The measure of size (number of households) for each ward in each stratum needs to be examined at the time of listing households. Since the 'take' in both urban and rural wards will be 20 households, there might be practical problem in the fieldwork. If the selected ward is found to have less than 20 households, it needs to be linked up with its smallest neighbouring ward, but always within the same VDC or municipality.

Table 2: Primary sampling units and sample households of the NLFS 2008

Stratum	PSUs	Sample Households
Rural Mountains	34	680
Urban Kathmandu Valley	131	2620
Other Urban Hills	99	1980
Rural Hills	179	3580
Urban Tarai	170	3400
Rural Tarai	187	3740
Nepal	800	16000

Estimation methodology

The estimation procedure will be developed on the basis of each stratum with attention given to the following:

- For the estimation of the population as a whole, we need to consider the selection probabilities at each stage of selection. In doing this, each stratum needs to be considered separately as different selection probabilities were used.
- Adjustment for non-response should be handled carefully since no substitute households are allowed.
- Estimates are required separately for each season of the year (rainy, winter and dry), and one-third of the total sample is allocated for each season so that each sub-sample (seasonal) provides a representative sample for different certain level of disaggregation of the country.

The weighting of sample data for the estimation procedure is given as the following stages.

$$1. \text{ First stage: Probability of selecting a particular ward } i = S_i * \frac{N_i}{\sum N_i}$$

where

N_i is the number of households in i^{th} ward in the sample frame,

$\sum N_i$ is the total number of households in the i^{th} stratum of sampling frame, and

S_i is the number of the PSUs in the i^{th} stratum.

When cartographic work is done in the selected ward i , a number of sub-wards are created with size $K_{i1}, K_{i2}, K_{i3},$ etc, where the total size of the ward $\sum K_{ij}$ is unlikely to be equal to the original size of the ward N_i . One of these sub-wards is then selected with PPS for inclusion in the survey.

$$2. \text{ Intermediate stage: Probability of selecting the } j^{\text{th}} \text{ sub-ward} = \frac{K_{ij}}{\sum_j K_{ij}}$$

where

K_{ij} is the number of households counted in the j^{th} sub-ward of ward i in the cartographic survey, and

$\sum_j K_{ij}$ is the total number of households counted in the i^{th} ward in the cartographic survey.

The interviewer then visits the j^{th} sub-ward during the survey, lists all the households, and attempts to interview 20 selected households (no substitutes are allowed).

$$3. \text{ Final stage: Probability of selection of a household} = \frac{n_{ij}}{K_{ij}^*}$$

where

n_{ij} is the number of households successfully interviewed, and

K_{ij}^* is the number of households counted at the listing stage.

The overall probability of selection for an individual household is the multiple of the selection probabilities at the three stages.

$$\text{Overall selection probability} = S_i * \frac{N_i}{\sum N_i} * \frac{K_{ij}}{\sum_j K_{ij}} * \frac{n_{ij}}{K_{ij}^*}$$

The raising factor is the inverse of this probability, and the formula can be rewritten as follows:

$$\text{Raising up factor} = \frac{\sum N_i}{h_i} * \frac{\sum K_{ij}}{N_i} * \frac{K_{ij}^*}{K_{ij}} * \frac{20}{n_{ij}}$$

(1) (2) (3) (4)

where $h_i = S_i * 20$, total number of households selected in the i th PSU.

Each of the numbered terms in this formula has a clear meaning should be explained as:

- (1) is the simple grossing up factor required to go from the original sample size of households in each stratum to the total number of households in the original sampling frame;
- (2) reflects the change in the number of households found in the i th ward during the cartographic work, as compared with the corresponding figure in the original frame;
- (3) reflects the change in the number of households found in the j th sub-ward of the i th ward during the listing work, as compared with the corresponding figure during the cartographic work;
- (4) is the adjustment factor required to make allowance for any non-response which occurs during the fieldwork in a particular ward or sub-ward.

Sampling errors

As with all surveys, the results from the NLFS II are subject to two major sources of error. These can be broadly termed as sampling error and non-sampling error. In a survey of this size, the robustness of the sample design means that the sampling errors for statistics at the national level are likely to be fairly small. Non-sampling errors are likely to be the major source of concern, and every effort has been made at all stages of the survey to try to minimise these non-sampling errors.

Table B.2 gives an indication of the likely sampling errors for some of the key aggregates measured in this survey. These sampling errors have been calculated in STATA application that was used for processing this survey. In order to derive these estimates of sampling error, account was taken of the structural design of the survey, with PSUs being assigned to six strata consisting three urban and three rural stratum, and with different sampling fractions being used in each stratum. However the resulting sampling errors probably substantially overstate the width of the true confidence intervals, since they take no account of the very strong implicit stratification by region and ecological zone incorporated into the design. The true confidence intervals for sampling errors will therefore be much narrower than those shown in Table B.2 .

All the estimates shown in Table B.2 can be found in the appropriate chapter of the survey report. In the case of estimate for the total currently active population aged 15 years and above (shown in Table 6.1 as being 12.032 million), the 95 percent lower and upper bounds for this estimate are 11.649 and

12.414 million respectively. This means that we can be 95 percent confident that the number of people currently economically active lies within this range. Put another way, we can say that we are 95 percent confident that the true value lies within the range 11.649 million plus or minus 382 thousand.

Table B.2 Approximate sampling errors for some key aggregates in the NLFS

	Estimate 95% confidence interval			Estimate 95% confidence interval		
	Currently active (thousand)			Currently inactive (thousand)		
	Lower	Upper		Lower	Upper	
Nepal	12032	11649	12414	2303	2149	2457
Male	5646	5455	5838	783	721	845
Female	6385	6167	6603	1520	1413	1626
Urban	1660	1537	1783	801	715	887
Rural	10400	10000	10700	1502	1373	1630
	Estimate 95% confidence interval					
	Labour force participation rate (percent)					
	Lower	Upper				
Nepal	83.4	82.4	84.4			
Male	87.5	86.7	88.4			
Female	80.1	78.8	81.4			
Urban	67.3	65.7	68.8			
Rural	86.8	85.7	87.8			
	Estimate 95% confidence interval			Estimate 95% confidence interval		
	Currently employed (thousand)			Currently unemployed (thousand)		
	Lower	Upper		Lower	Upper	
Nepal	11779	11400	12158	253	220	285
Male	5519	5330	5709	127	112	141
Female	6259	6042	6477	126	.	.
Urban	1535	1427	1643	125	105	145
Rural	10244	9880	10607	128	101	154
	Estimate 95% confidence interval			Estimate 95% confidence interval		
	Unemployment rate (percent)			Labour underutilization rate (percent)		
	Lower	Upper		Lower	Upper	
Nepal	2.1	1.8	2.4	30.0	28.8	31.3
Male	2.2	1.9	2.6	38.2	36.8	39.6
Female	2.0	1.6	2.4	22.8	21.5	24.2
Urban	7.5	6.7	8.4	49.9	48.2	51.6
Rural	1.2	0.9	1.5	26.9	25.5	28.2