UNITED NATIONS SECRETARIAT **ESA/STAT/AC.93/5**
Statistics Division **03 November 2003**

English only

Expert Group Meeting to
Review the Draft Handbook on
Designing of Household Sample Surveys
3-5 December 2003

D R A F T

Construction and use of sample weights*

by
Ibrahim S. Yansaneh**

* This document is being issued without formal editing.

** The views expressed in this paper are those of the author and do not imply the expression of any opinion on the part of the United Nations Secretariat.

Table of contents

Chapter Five: Construction and use of sample weights 3
5.1. The need for sampling weights 3
5.2. The development of sampling weights 3
5.2.1. Adjustments of sample weights for unknown eligibility 4
5.2.2. Adjustments of sample weights for duplicates 4
5.3. Weighting for unequal probabilities of selection 4
5.3.1. A case study in construction of weights: the Viet Nam National Health Survey 6
5.3.2. Self weighting samples 6
5.4. The adjustment of sample weights for non-response 6
5.4.1. Reducing non-response bias in household surveys 7
5.4.2. Compensating for non-response bias 7
5.4.3. Non-response adjustment of sample weights 8
5.5. The adjustment of sample weights for non-coverage 10
5.5.1. Sources of non-coverage in household surveys 10
5.5.2. Compensating for non-coverage in household surveys 11
5.6. Increase in variance due to weighting 12
5.7. Concluding remarks 12
References and further reading 14

Abstract

This chapter provides a brief overview of the various stages in the construction and adjustment of sample weights to be used in the analysis of survey data. In particular, the adjustment of sample weights to compensate for non-coverage and non-response is described. In addition, the chapter discusses how sample weights are used in the development of estimates of characteristics of interest. The important ideas presented are illustrated using real examples of current surveys conducted in developing countries, or ones that mimic real survey situations.

Key Words. Base weight; non-response adjustment; post-stratification; domain estimation

Chapter Five: Construction and use of sample weights

5.1. The need for sampling weights

1. Sampling weights are needed to correct for imperfections in the sample that might lead to bias and other departures between the sample and the reference population. Such imperfections include the selection of units with unequal probabilities, non-coverage of the population, and non-response. In other words, the purposes of weighting are:

a. To compensate for unequal probabilities of selection.
b. To compensate for (unit) non-response.
c. To adjust the weighted sample distribution for key variables of interest (for example, age, race, and sex) to make it conform to a known population distribution.

2. We shall discuss in detail the procedures underlying each of these scenarios in the sections that follow. Once the imperfections in the sample are compensated for, weights can then be used in the estimation of population characteristics of interest and also in the estimation of the sampling errors of the survey estimates generated.

(Give an example to illustrate what happens when weights are not used)

5.2. The development of sampling weights

3. The development of sampling weights usually starts with the construction of the *base weight* for each sampled unit, to correct for their unequal probabilities of selection. In general, the base weight of a sampled unit is the reciprocal of its probability of selection into the sample. In mathematical notation, if a unit is included in the sample with probability P_i, then its base weight, denoted by w_i, is given by

$$w_i = 1/p_i.$$

4. For example, a sampled unit selected with probability 1/50 represents 50 units in the population from which the sample was drawn. Thus sample weights act as inflation factors to represent the number of units in the survey population that are accounted for by the sample unit to which the weight is assigned. The sum of the sample weights provides an unbiased estimate of the total number of individuals in the target population.

5. For multi-stage designs, the base weights must reflect the probabilities of selection at each stage. For instance, in the case of a two-stage design in which the *i-th* PSU is selected with probability p_i at the first stage, and the *j-th* household is selected within a selected PSU with probability $p_{j(i)}$ at the second stage, then the overall probability of selection of the every household in the sample is given by

$$p_{ij} = p_i * p_{j(i)}$$

and the overall base weight the household is obtained as before, by taking the reciprocal of its overall probability of selection. Correspondingly, if the base weight for the j-th household is $w_{ij,b}$, and the weight attributable to compensation for non-response is $w_{ij,nr}$, and the weight attributable to the compensation for non-coverage is $w_{ij,nc}$, then the overall weight of the household is given by:

$$w_{ij} = w_{ij,b} * w_{ij,nr} * w_{ij,nc}$$

5.2.1. Adjustments of sample weights for unknown eligibility

(Discuss weighting for unknown eligibility)

5.2.2. Adjustments of sample weights for duplicates

6. If it is known a priori that some units have duplicates on the frame, then increased probability of selection of such units can be compensated for by assigning to them weighting factors that are reciprocals of the number of duplicate listings on the frame if such units end up in the sample. Often however, duplicates are discovered only after the sample is selected, and the probabilities of selection of such sampled units need to be adjusted to account for the duplication. This adjustment is implemented as follows:
Suppose the *i-th* sampled unit has a probability of selection, denoted by p_{i1} and suppose there are *k-1* additional records on the sampling frame that are identified by this sampled unit as duplicates, each with selection probabilities given by $p_{i2}, \ldots, p_{ik}$. Then, the adjusted probability of selection of the sampled unit in question is given by

$$p_i = 1 - (1 - p_{i1})(1 - p_{i2}) \ldots (1 - p_{ik})$$

The sampled unit is then weighted accordingly, that is, by $1/p_i$.

7. We now illustrate the procedures for constructing sample weights under scenarios outlined above, with specific examples.

5.3. Weighting for unequal probabilities of selection

8. An *epsem* sample of 5 households is selected from 250. One adult is selected at random in each sampled household. The monthly income (y_{ij}) and the level of education (z_{ij}= 1, if secondary or higher; 0 otherwise) of the *j-th* sampled adult in the *i-th* household are recorded. Let M_i denote the number of adults in household *i*. Then, the overall probability of selection of a sampled adult is given by:

$$p_{ij} = p_i \times p_{j(i)} = \frac{5}{250} \times \frac{1}{M_i} = \frac{1}{50} \times \frac{1}{M_i}$$

Therefore, the weight of a sampled adult is given by:

$$w_i = \frac{1}{p_{ij}} = 50 \times M_i$$

- ***Example***

9. To illustrate the estimation procedure, let us assume a first-stage sample of 5 households with data obtained from the single sampled adult for each household as given in the table below:

Sampled Household	M_i	w_i	y_{ij}	z_{ij}	$w_i y_{ij}$	$w_i z_{ij}$	$w_i z_{ij} y_{ij}$
1	3	150	70	1	10,500	150	10,500
2	1	50	30	0	1,500	0	0
3	3	150	90	1	13,500	150	13,500
4	5	250	50	1	12,500	250	12,500
5	4	200	60	0	12,000	0	0
TOTAL	**16**	**800**	**300**	**3**	**50,000**	**550**	**36,500**

10. Estimates of various characteristics can then be obtained from the above table as follows:

a. The estimate of monthly income is

$$\bar{y}_w = \frac{\sum w_i y_{ij}}{\sum w_i} = \frac{50{,}000}{800} = 62.5$$

If weights were not used, this estimate would be 60 (=300/5)

b. The estimate of the proportion of people with secondary or higher education is

$$\bar{y}_w = \frac{\sum w_i z_{ij}}{\sum w_i} = \frac{550}{800} = 0.6875 \text{ or } 68.75\%$$

If weights are not used, this estimate would be 3/5 or 0.60 or 60%.

c. The estimate of the total number of people with secondary or higher education is

$$\hat{t} = \sum w_i z_{ij} = 550$$

d. The estimate of the mean monthly income of adults with secondary or higher education is

$$\bar{y}_w = \frac{\sum w_i z_{ij} y_{ij}}{\sum w_i z_{ij}} = \frac{36{,}500}{550} = 66.36$$

11. Note that for estimating totals, sampled elements need to be weighted by the reciprocal of their selection probabilities. For estimating means and proportions, the weights need only be

proportional to the reciprocals of the selection probabilities. Thus, in the preceding example, the weights w_i's are proportional to Mi (w_i=50* M_i). If we use M_i as the weights, then the estimate of the proportion with secondary or higher education is

$$\hat{p} = \frac{\sum M_i z_{ij}}{\sum M_i} \frac{3\times1+1\times0+3\times1+5\times1+4\times0}{3+1+3+5+4} = \frac{11}{16} = 0.6875 \text{ or } 68.75\%,$$

as before. However, the estimate of the total number of adults with secondary or higher education is

$$\hat{p}_s = 50\sum M_i z_{ij} = 50\times11 = 550$$

5.3.1. A case study in construction of weights: the Viet Nam National Health Survey

12. We now proceed to illustrate the construction of the sampling weights for an actual survey, the National Health Survey conducted in Viet Nam in 2001.

(Insert a case study of weight construction for the VNHS)

5.3.2. Self weighting samples

13. When the weights of all sampled units are the same, the sample is referred to as *self-weighting*. Samples are rarely self-weighting at the national level for several reasons. First, sampling units are selected with unequal probabilities of selection. Indeed, even though the PSUs are often selected with probability proportional to size, and households selected at an appropriate rate within PSUs to yield a self-weighting design, this may be nullified by the selection of one person for interview in each sampled household. Second, the selected sample often has deficiencies including non-response and under-coverage (see sections 5.4 and 5.5). Third, the need for precise estimates for domains and special subpopulations often requires over-sampling these domains (see section 5.5).

5.4. The adjustment of sample weights for non-response

14. It is rarely the case that all desired information is obtained from all sampled units in surveys. For instance, some households may provide no data at all while other households may provide only partial data, that is, data on some but not all questions in the survey. The former type of non-response is called *unit* or *total non-response*, while the latter is called *item non-response*. If there are any systematic differences between the respondents and non-respondents, then naïve estimates based solely on the respondents will be biased. It is important to keep survey non-response as low as possible, in order to reduce the possibility that the survey estimates could be biased in some way by failing to include (or including a disproportionately small percentage of) a particular portion of the population. For example, persons who live in

urban areas and have relatively high incomes might be less likely to participate in a multi-purpose survey that includes income modules. Failure to include a large segment of this portion of the population could affect national estimates of average household income, educational attainment, literacy, etc.

5.4.1. Reducing non-response bias in household surveys

15. The size of the non-response bias for a sample mean, for instance, is a function of two factors:

- The proportion of the population that does not respond.
- The size of the difference in population means between respondent and non-respondent groups.

16. Reducing the bias due to non-response therefore requires that either the non-response rate be small, or that there are small differences between responding and non-responding households and persons. With proper record keeping of every sampled unit that is selected for the survey, it is possible to estimate directly from the survey data, the non-response rate for the entire sample and for sub-domains of interest. Furthermore, special carefully designed studies can be carried out to evaluate the differences between respondents and non-respondents (Groves and Couper, 1998).

17. For panel surveys (in which data are collected from the same panel of sampled units repeatedly over time) the survey designer has access to more data for studying and adjusting for the effects of potential non-response bias than in one time or cross-sectional surveys. Here, non-response may arise from units being lost over the course of the survey, or refusing to participate in the survey after a while due to respondent fatigue or other reasons, and so on. Data collected on previous panel waves can then be used to learn more about differences between respondents and non-respondents, and to serve as the basis for the kind of adjustments described below. More details on various techniques used for compensating for non-response in survey research are provided in Brick and Kalton (1996), Lepkowski (1988), and references cited therein.

5.4.2. Compensating for non-response bias

18. A number of techniques can be employed to reduce the potential for non-response bias in household surveys. The standard method of compensating for partial or item non-response is *imputation*, which is not discussed in this volume. A good introduction to imputation methods for large complex datasets is provided by Yansaneh et. al. (1998). For unit or total non-response, there are three basic procedures for compensation:

a. Non-response adjustment of the weights.
b. Drawing a larger sample than needed and creating a reserve sample from which replacements are selected in case of non-response.
c. Substitution, the process of replacing a non-responding household with another household that was not sampled which is in close proximity to the non-responding household with respect to the characteristic of interest.

19. It is advisable that unit non-response in household surveys be always handled by adjusting the sample weights to account for non-responding households. In many surveys in developing countries, substitution is frequently used. However, this procedure increases the probabilities of selection for the potential substitutes, because non-sampled households close to non-responding sampled households have a higher probability of selection than those close to responding sampled households. Furthermore, attempts to substitute for non-responding households are time-consuming, prone to errors and bias, and very difficult to check or monitor. For example, a substitution may be made using a convenient household rather than the household specifically designated to serve as the substitute or replacement for a non-responding household, thereby introducing bias.

5.4.3. Non-response adjustment of sample weights

20. The procedure of adjusting sample weights for non-response is the preferred practice in major household surveys throughout the world. Essentially, the adjustment transfers the base weights of all eligible non-responding sampled units to the responding units, and is implemented in the following steps:

Step 1: Apply the initial weights (for unequal selection probabilities and other adjustments discussed in Section 5.2, if applicable);

Step 2: Partition the sample into subgroups and compute weighted response rates for each subgroup;

Step 3: Use the reciprocal of the subgroup response rates for non-response adjustments; and

Step 4: Calculate the non-response adjusted weight for the *i*-th unit as:

$$w_i = w_{1i} * w_{2i},$$

where w_{1i} is the initial weight and w_{2i} is the non-response adjustment weight. Note that the weighted non-response rate can be defined as the ration of the weighted number of interviews completed with eligible sampled cases to the weighted number of eligible sampled cases.

21. We now illustrate the ideas presented in this section with an example.

- ***Example***

22. A stratified multi-stage sample of 1000 households is selected from two regions (North and South) of a country. Households in the North are sampled at a rate of 1/100 and those in the south at a rate of 1/200. Response rates in urban areas are lower that those in rural areas. Let n_h denote the number of households sampled in stratum h, let r_h denote the number of eligible households that responded to the survey, and let t_h denote the number of responding households

with access to primary health care. Then, the non-response adjusted weight for the households in stratum h is given by:

$$w_h = w_{1h} * w_{2h,}$$

where $w_{2h} = n_h / r_h$. Assume that the stratum-level data are as given in the following table:

Stratum	n_h	r_h	t_h	w_{1h}	w_{2h}	w_h	$w_h r_h$	$w_h t_h$
North-Urban	100	80	70	100	1.25	125	10,000	8,750
North-Rural	300	120	100	100	2.50	250	30,000	25,000
South-Urban	200	170	150	200	1.18	236	40,120	35,400
South-Rural	400	360	180	200	1.11	222	79,920	39,960
Total	**1,000**	**730**	**500**				**160,040**	**109,110**

23. Therefore, the estimated proportion of households with access to primary health care is:

$$\hat{p} = \frac{\sum w_h t_h}{\sum w_h r_h} = \frac{109{,}110}{160{,}040} = 0.682 \quad \text{or } 68.2\%$$

The estimated number of households with access to primary health care is

$$\hat{t} = \sum w_h t_h = 109{,}110 \text{ = } 68.2\% \text{ of } 160{,}040$$

Note that the unweighted estimated proportion of households with access to primary health care, using only the respondent data is

$$\hat{p}_{uw} = \frac{\sum t_h}{\sum r_h} = \frac{500}{730} = 0.685 \text{ or } 68.5\%,$$

and the estimated proportion using the initial weights without non-response adjustment is

$$\hat{p}_1 = \frac{\sum w_{1h} t_h}{\sum w_{1h} r_h} = \frac{83{,}000}{126{,}000} = 0.659 \text{ or } 65.9\%.$$

24. Note also this example is provided for the purpose of illustrating how initial weights are adjusted to compensate for non-response. The results show considerable disparity between the estimated proportion using only the initial weights compared to that using non-response adjusted weights, but the difference between the unweighted proportion and the non-response-adjusted proportion appears to be negligible.

25. After non-response adjustments of the weights, further adjustments can be made to the weights as appropriate. In the next section, we consider adjustment of the weights to account for non-coverage.

5.5. The adjustment of sample weights for non-coverage

26. Non-coverage refers to the failure of the sampling frame to cover all of the target population and thus some sampling units have no probability of selection into the sample selected for the household survey. This is just one of many possible deficiencies of sampling frames used to select samples for surveys in developing countries. See Yansaneh (2003), and references cited therein, for a detailed discussion of sampling frame problems and some possible solutions.

5.5.1. Sources of non-coverage in household surveys

27. Most household surveys in developing countries are based on stratified multi-stage area probability designs. The first-stage units, or primary sampling units, are usually geographic area units. At the second stage, a list of households or dwelling units is created, from which the sample of households is selected. At the last stage, a list of house members or residents is created, from which the sample of persons is selected. Thus non-coverage may occur at three levels: the PSU level, the household level, and the person level.

28. Since PSUs are generally based on enumeration areas identified and used in a preceding population and housing census, they are expected to cover the entire geographic extent of the target population. Thus, the size of PSU non-coverage is generally small. For household surveys in developing countries, PSU non-coverage is not as serious as non-coverage at subsequent stages of the design. However, non-coverage of PSUs does occur in most surveys. For instance, a survey may be designed to provide estimates for the entire population in a country, or a region of a country, but some PSUs may be excluded on purpose at the design stage, because some regions of a country are inaccessible due to civil war or unrest, a natural disaster, or other reasons. Also, remote areas with very few households or persons are sometimes removed from the sampling frames for household surveys because they are too costly to cover, and they represent a small proportion of the population and so have very little effect on the population figures. In reporting results for such a survey, the exclusion of these areas must be explicitly stated. The impression should not be created that survey results apply to the entire country or region, when in fact a portion of the population is not covered. The non-coverage properties of the survey must be fully reported in the survey report.

29. Non-coverage becomes a more serious problem at the household level. Most surveys consider households to be the collection of persons who are usually related in some way, and who usually reside in a dwelling or housing unit. There are important definitional issues to resolve, such as who is a usual resident; and what is a dwelling unit? How are multi-unit structures (such as apartment buildings) and dwelling units with multiple households handled? It may be easy to identify the dwelling unit, but complex social structures may make it difficult to identify the households within the dwelling unit. There is thus a lot of potential for misinterpretation or inconsistent interpretation of these concepts by different interviewers, or in

different countries or cultures. In any event, strict operational instructions are needed to guide interviewers on whom to consider a household member or what to consider a dwelling unit.

30. Other factors that contribute to non-coverage include the inadvertent omission of dwelling units from listings prepared during field operations, or sub-populations of interest (for example, young children or the elderly), and omissions due to errors in measurement, non-inclusion of absent household members, and omissions due to misunderstanding of survey concepts. There is also a temporal dimension to the problem, that is, dwelling units may be unoccupied or under construction at the time of listing, but become occupied at the time of data collection. For household surveys in developing countries, the non-coverage problem is exacerbated by the fact that most censuses in developing countries, the unique basis for constructing sampling frames, do not provide detailed addresses of sampling units at the household and person levels. Frequently, out of date or inaccurate administrative household listings are used, and individuals within a household are deliberately or accidentally omitted from a household listing of residents. More details on sources of non-coverage are provided in Lepkowski (2003) and references cited therein.

5.5.2. Compensating for non-coverage in household surveys

31. Non-coverage is a major concern for household surveys conducted in developing countries. Evidence of the impact of non-coverage can be seen from the fact that sample estimates of population counts based on most developing-country surveys fall well short of population estimates from other sources. There are several procedures for handling the problem of non-coverage in household surveys (Lepkowski, 2003). These include:

a. Improved field procedures such as the use of multiple frames and improved listing procedures.
b. Compensating for the non-coverage through a statistical adjustment of the weights.

32. In this chapter, we shall concentrate on the second procedure. If reliable control totals are available for the entire population and for specified subgroups of the population, one could attempt to adjust the weights of the sample units in such a way as to make the sum of weights match the control totals within the specified subgroups. The subgroups are called *post-strata,* and the statistical adjustment procedure is called *post-stratification*. This procedure simultaneously compensates for non-response and non-coverage. It adjusts the weighted sampling distribution for certain variables so as to conform to a known population distribution. See Lehtonen and Pahkinen (1995) for some practical examples of how to analyze survey data with poststratification. A simple example is provided below, just to aid understanding of the procedure.

- ***Example***

33. In the preceding example, suppose that the number of households is known to be 45,025 in the North and 115,800 in the South. Suppose further that the weighted sample totals are respectively 40,000 and 120,040.

Step 1: Compute the post-stratification factors.

For the North region, we have: $w_{3h} = \frac{45{,}025}{40{,}000} = 1.126$; and

For the South region, we have: $w_{3h} = \frac{115{,}800}{120{,}040} = 0.965$.

Step 2: Compute final, adjusted weight: $w_f = w_h \times w_{3h}$

34. The numerical results are summarized in the following table:

Stratum	r_h	t_h	w_h	w_f	$w_f * r_h$	$w_f * t_h$
North-Urban	80	70	125	140.75	11,256	9,849
North-Rural	120	100	250	281.40	33,768	28,140
South-Urban	170	150	236	227.77	38,709	34,155
South-Rural	360	180	222	214.20	77,112	38,556
Total	**730**	**500**			**160,845**	**110,700**

Therefore, the estimated proportion of households with access to primary health care is:

$$\hat{p}_f = \frac{\sum w_f t_h}{\sum w_f r_h} = \frac{110{,}700}{160{,}845} = 0.688 \quad \text{or } 68.8\%$$

35. Note that with the weights adjusted by post-stratification, the weighted sample counts for the North and South regions are respectively 45,024 (11,256+33,768) and 115,821 (38,709+77,112), which closely match the control totals given above.

5.6. Increase in variance due to weighting

(…)

5.7. Concluding remarks

36. Sample weights have now come to be regarded as an integral part of the analysis of household survey data in developing countries, as in the rest of the world. Most survey programmes now advocate the use of weights even in the rare situations involving self-weighting samples (in which case the weights would be 1). In the past, tremendous efforts were expended by survey designers for the virtually unattainable goal of achieving self-weighting samples and making weights unnecessary in the analysis of survey data. The conventional wisdom was that the use of weights made the analyses too complicated, and that there was very little, if any, computing infrastructure for weighted analysis of survey data. However, advances in computer technology in past decade have invalidated this argument. Computer hardware and software are

now affordable and available in many developing countries. In addition, many specialized computer software packages are now available specifically for the analysis of survey data. These are reviewed and compared in chapter 6. More details can be obtained from the references cited in chapter 6.

37. As discussed, the use of weights reduces biases due to imperfections in the sample related to non-coverage and non-response. Non-response and non-coverage are different types of error due to the failure of a designed survey to obtain information from some units in the target population. For household surveys in developing countries, non-coverage is a more serious problem than non-response. The chapter provides examples of procedures for developing and statistically adjusting the basic weights to compensate for these unavoidable problems of household surveys, and for using the adjusted weights in the estimation of parameters of interest. The advent of fast-speed computers and affordable of free statistical software should make the use of weights a routine aspect of the analysis of household survey data even in developing countries.

References and further reading

Brick, J.M. and Kalton, G. (1996). "Handling missing data in survey research," *Statistical Methods in Medical Research*, Vol. 5, 215-238.

Cochran, W.G. 1977. *Sampling Techniques*, 3rd edition. New York: John Wiley & Sons.

Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. 2002. *Survey Non-response*. New York: John Wiley & Sons

Groves, R.M., and Couper, M.P. 1998. *Non-response in Household Interview Surveys*. New York: John Wiley & Sons.

Kalton, G, and Kasprzyk, D. 1986. "The treatment of missing survey data," *Survey Methodology*, Vol. 12, pp. 1-16

Kalton, G. (1983). *Introduction to Survey Sampling*. Sage Publications Series, No. 35.

Kish, Leslie. 1965. *Survey Sampling*. New York: Wiley.

Kish, L., and Hess, I. 1950. "On noncoverage of sample dwellings," *Journal of the American Statistical Association*, Vol. 53, pp. 509-524.

Lehtonen, R., and E. J. Pahkinen (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: Wiley.

Lepkowski, James M. *Non-observation Error in Household Surveys in Developing Countries*, Technical Report on Surveys in Developing and Transition Countries, United Nations, 2003.

Lessler, J., and Kalsbeek, W. 1992. *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

Levy, Paul S. and Stanley Lemeshow (1999). Sampling of Populations: Methods and Applications. Third edition. John Wiley & Sons, New York.

Lohr, Sharon. 1999. *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Yansaneh, I. S. *An Overview of Sample Design Issues for Household Surveys in Developing Countries*, Technical Report on Surveys in Developing and Transition Countries, United Nations, 2003.