APPENDIX SEVEN

SAMPLING DETAILS

This appendix contains more technical information and is intended mainly for sample specialists. It contains the following sections, all of which are referred to in Chapter 4:

- Sample size calculation.
- Procedures for sampling with PPS – Option 2.
- Procedures for sampling with PPS – Option 3.
- Computation of the post-stratification weights.

SAMPLE SIZE CALCULATION

The current section describes how the sample size can be calculated when the survey situation fits neither that used for Table 4.9 nor for Table 4.10 in Chapter 4. The sample size calculation applies only to persons, since the most important indicators for end-decade assessment are person-based. Household sample size calculations would not only require a different formula, but also very different design effect, or *deff* values, of 10 or more.

The calculating formula, taking into account the parameters and assumptions discussed in Chapter 4, is given by

$$n = [4\,(r)\,(1 - r)\,(f)\,(1.1)] / [(e^2)\,(p)\,(n_h)], \text{ where} \qquad (1)$$

(taking the components in order)

n is the required sample size for the KEY (rarest) indicator,

4 is a factor to achieve the 95 percent level of confidence,

r is the predicted or anticipated prevalence (coverage rate) for the key indicator, which is based upon the *smallest target group* (in terms of its proportion of the total population),

1.1 is the factor necessary to raise the sample size by 10 percent for nonresponse,

f is the *deff*,

e is the margin of error to be tolerated,

p is the proportion of the total population that the smallest group comprises, and

n_h is the average household size.

A numerical example is provided to illustrate the calculation.

. EXAMPLE (MODERATE-TO-HIGH COVERAGE RATE):

Suppose the target group in your country that comprises the smallest percentage of the total population is one-year-old children (recall that we are purposely excluding the four-month age groups that form the base for the breastfeeding indicators) and this group comprises 3 percent of the population. Further suppose that their DPT coverage is anticipated to be the lowest of all the indicator coverages – 50 percent, for which we want our margin of error to be 5 percentage points. If your average household size is 6 persons and we assume the sample *deff* is moderate, or 1.75, then the values of your parameters will be as follows:

$r = 0.5$

$p = .03$

$f = 1.75$

$e = .05$

$n_h = 6$

Substituting, you have

$$n = \{4 \times 0.5 \times (1-0.5) \times 1.75 \times 1.1\} / \{(.05)^2 \times .03 \times 6\}$$

$$= 4{,}278.$$

This is the number of households you would need to survey in order to estimate DPT coverage of about 50 percent with a margin of error of 5 percentage points. Those households would contain about 25,667 persons, of which about 770 would be one-year-old children.

Formula (1) can be rewritten in shortcut version for easy calculation whenever the values of p, f, e, and n_h are fixed at .03, 1.75, .05, and 6, respectively, and when the 95 percent level of confidence and nonresponse adjustment (factors of 4 and 1.1, respectively) are not changed. In that case the shortcut version is given by

$$n = (17{,}111)\, r\, (1- r). \qquad (2)$$

It is recommended to use the formulas (long or shortcut) instead of Table 4.9 in Chapter 4 if your moderate-to-high prevalence rate is not close to 50 percent, which is the value that the table is based upon. You would have to use the long version (formula 1) if you want to change one or more of the p, e, f, or n_h values.

You might also want to consider using the long version if your nonresponse is not expected to be as high as 10 percent, in which case you would substitute for the factor of 1.1 accordingly.

It is recommended that you use the formula instead of Table 4.9 if your least coverage indicator is quite high (for example, 75 percent), because the sample size will be considerably less. For an r value of 0.75, for example, n would be 3,208 (short formula).

Another example is provided for the case where your key indicator has low coverage.

. EXAMPLE (LOW COVERAGE RATE):

Suppose your polio coverage is expected to be about 25 percent. In this case you would want your margin of error to be 3 percentage points instead of 5 (so that the confidence interval for the coverage estimate is 22 to 28 percent, as opposed to 20 to 30 percent). The other parameter values are the same as in the first example. Substituting, you would have

$$n = \{4 \times 0.25 \times (1-0.25) \times 1.75 \times 1.1\} / \{(.03)^2 \times .03 \times 6\}$$
$$= 8{,}912$$

You can readily see that with stricter tolerance for the margin of error, necessary for the low coverage indicator, the sample size is much larger. This is why it is important in selecting the key indicator upon which to base your sample size that both the smallest target group be identified, and, within that group, the indicator that has the lowest coverage.

The shortcut version for calculating sample sizes for different low coverage indicators is given by:

$$n = 47{,}531\, r\, (1- r), \text{ whenever} \quad (3)$$

p, e, f, and n_h are fixed at .03, .03, 1.75, and 6, respectively.

The formulas should be used instead of Table 4.10 in Chapter 4 if your low coverage indicator has a value that departs significantly from 25 percent, since the latter is the value that Table 4.10 is based upon.

PROCEDURES FOR SAMPLING WITH PPS – OPTION 2

In this section we give an illustration of how to select the first-stage units using *pps*. The illustration also shows you how to combine systematic *pps* sampling with geographic arrangement of the sampling frame to achieve *implicit* stratification.

For the illustration we take Option 2 from Chapter 4, the standard segment design, and we select a national sample. Suppose (1) the standard segment size under Option 2 is to be 500 persons, or about 100 households; (2) census enumeration areas (EAs) are to be the sample frame; and (3) the number of PSUs to be selected is 300. The steps of the first-stage selection, which follow, *should be done as a computer operation*, although it is possible to do them manually.

Step 1: Sort the file of EAs by urban and rural.

Step 2: Within the urban category, further sort the file in geographic serpentine order according to the administrative subdivisions of your country (for example, province or state, district, commune, etc.).

Step 3: Repeat Step 2 for the rural category.

Step 4: In one column show the census population count of the EA.

Step 5: In the next column compute the number of standard segments, which is equal to the population count divided by 500, and rounded to the nearest integer. This is the measure of size for the EA.

Step 6: Cumulate the measures of size in the next column.

Step 7: Compute the sampling interval, I, by dividing the total cumulant by 300, to one decimal place. In this illustration suppose the total cumulant is 5,281. Then the sampling interval, I, would be equal to 5,281/300, or 17.6.

Step 8: Select a random start between 0 and 17.6. The way to do this, in practice, is to use a table of random numbers and select a three-digit number between 001 and 176 and insert the decimal afterward. Suppose you select 042; then your random start is 4.2. Then your first sample PSU would be the one for which the cumulant measure of size is the smallest value equal to or greater than 4.2.[1]

Step 9: Add 4.2 to I, or 4.2 + 17.6 = 21.8; then your next sample PSU would be the one whose cumulant corresponds to the smallest value equal to or greater than 21.8.

Step 10: Add 21.8 to I, or 21.8 + 17.6 = 39.4; the next sample PSU is the one with cumulant corresponding to the smallest value equal to or greater than 39.4.

Step 11: Continue as above, through the urban EAs followed by the rural ones, until all 300 PSUs have been selected.

The procedure is further demonstrated in Table A7.1.

The two sample PSUs that are depicted in the illustration are those in EAs 003 of commune 01 and EA 002 of commune 03, both in district 01 and province 01. In the case of the first EA, its measure of size is 3, which would mean that three segments would have to be created, each of roughly 540 persons (1,630 divided by 3), and then one of the segments would be selected at random for listing and subsampling of households. In the second sample EA, two segments would be created, each containing about 590 persons, before selecting one of them at random.

The illustration demonstrates the many advantages of implicit stratification. First, it is very easy to achieve, merely requiring that the frame of enumeration areas be sorted geographically before then selecting the sample systematically with *pps*. Second, it automatically provides a sample of PSUs that is proportionately distributed by urban and rural and by province (or other geographic subdivisions). For example, if 10 percent of your population is located in province 12, then 10 percent of your sample will also be selected in that province. Third, it can be easily implemented on the computer.

[1] Kish recommends rounding down when the sampling interval is fractional. See Kish, L. (1965) *Survey Sampling*,

Wiley, New York, page 116.

Table A7.1

Illustration of Systematic *pps* Sampling and Implicit Stratification – Sample Option 2

Urban	Population	Measure of size (segments of 500 population)	Cumulative
Province 01			
District 01			
Commune 01			
EA 001	1,470	3	3
EA 002	562	1	4
EA 003	1,630	3	7 selected
EA 004	1,006	2	9
Commune 02			
EA 001	412	1	10
EA 002	1,537	3	13
EA 003	1,312	3	16
EA 004	397	1	17
Commune 03			
EA 001	1,540	3	20
EA 002	1,181	2	22 selected
EA 003	1,025	2	24
District 02			
Commune 01			
EA 001	567	1	25
EA 002	1,111	2	27
EA 003	409	1	28
*			
*			
etc.			
Rural			
Province 12			
District 05			
Commune 05			
EA 001	512	1	5,280
EA 002	493	1	5,281

Once the PSUs have been selected, under Option 2, segmentation will have to be carried out in those PSUs where the measure of size (number of segments) is two or more, followed by one segment being selected at random in each PSU. Then, a new household listing will have to be made in the selected segments plus the one-segment PSUs. The final step in the selection procedure for

Option 2 is to select the sample households within the selected segments. *This procedure is described in Table A7.2 with an illustration.*

Table A7.2

Selecting the Households – Option 2

Suppose your standard segment size is 500 persons. Let your desired cluster size for the survey be designated as ñ households.

1. Calculate the average households per segment by dividing 500 by the average household size in your country. Let this be s_h.
2. Divide s_h by ñ. This is your sampling interval, I, for selecting households within each sample segment.

   (Note, if your standard segment size is other than 500, that value must be used, of course.)

Illustration:

Suppose your average household size is 5.5. Then s_h is 500/5.5, or 90.9. Suppose you want your cluster size, ñ, to be 25. Divide 90.9 by 25 (1 decimal place) = 90.9/25, or 3.6. Then, select households in each segment at the rate of 1 in 3.6, starting with a random number between 01 and 36 (inserting the decimal after selecting the number).

PROCEDURES FOR SAMPLING WITH PPS – OPTION 3

If Option 3, the modified segment design described in Chapter 4, is used instead of Option 2, implicit stratification is done in the same way, *although the measure of size is different*. Under option 3, if we suppose, as an example, that our segment size is going to be 20 households (on average), then the measure of size would be calculated by dividing the census count of households by 20, rounded to the nearest whole number. Note that under Option 3 the second column in Table A7.3 must be number of households rather than population. You would calculate the sampling interval, I, by dividing the total cumulant – suppose it is 26,425 – by the desired number of PSUs, again let it be 300. So, you would have 26,425/300 = 88.1. If the random start is chosen to be 19.4, the first two PSUs selected, as illustrated in Table A7.3, would be those corresponding to the smallest cumulants exceeding the values, 19.4 and 107.5 (88.0 +

19.4), respectively. They are EA 002 in commune 01 and EA 002 in commune 03 of province 01, district 01.

Recall that under option 3 the measure of size is equivalent to the number of segments of predesignated size that must be created (in our illustration, that is 20). So, for the sample PSUs chosen, 6 segments of approximate size 20 households each must be formed in the first PSU and 12 in the second. Again, one of the segments would then be selected at random within each sample PSU, *and all of the households within that segment would be interviewed for the survey*, even if the actual number of households in the segment departs significantly from its expected size.

Chapter 6 details the procedures for creating segments both for Option 2 and Option 3.

Table A7.3

Illustration of Systematic *pps* Sampling and Implicit Stratification – Sample Option 3

Urban	Population	Measure of size (segments of 500 population)	Cumulative
Province 01			
District 01			
Commune 01			
EA 001	290	14	14
EA 002	120	6	20 selected
EA 003	325	16	36
EA 004	200	10	46
Commune 02			
EA 001	81	4	50
EA 002	307	15	65
EA 003	261	13	78
EA 004	80	4	82
Commune 03			
EA 001	308	15	97
EA 002	236	12	109 selected
EA 003	205	10	119
*			
*			
*			
etc.			
Rural			
Province 12			
District 05			

Commune 05			
EA 001	102	5	26,400
EA 002	99	5	26,405

COMPUTATION OF POST-STRATIFICATION WEIGHTS

The procedure for applying the post-stratification weights depends on whether (i) weighting is done at the level of each sample person (or household) before the computation of any aggregate measures such as separate indicators, or (ii) if separate indicators (such as by urban-rural or region) are computed first and then put together with appropriate weights.

Alternative (i): Weighting at the Sample Case Level

This involves the assignment of a weight to each sample person (or household). Within each weighting domain (e.g., by region or urban-rural), all records are given the same uniform weight equal to the *ratio* of the domain's population as a proportion of the national population according to the census or some other reliable source, *to* the domain's sample as a proportion of the total sample. For non-self-weighting samples, the denominator is computed after the application of the design weights to the sample units.

Let:

P_i = population of domain i as a proportion of the total national population according to the census or other reliable source (such as current population projections)

p_i = sample population of domain i as a proportion of the total population enumerated in the survey (weighted by design weights plus nonresponse adjustment, if applicable)

The weights applied at the level of the individual sample case are $W_i = P_i / p_i$, which implies post-stratification by domain. The sample population distribution by the weighting domains is adjusted to match the corresponding census distribution. No post-stratification adjustment is implied by other subgroups of the population, since it is expected that the sample would provide a reasonable representation of the proportion of these subgroups in the population. All indicators, whether at the

national or domain level, and whether applicable to the whole population or to specific subgroups (such as children of a particular age or gender), automatically incorporate the post-stratification adjustment since these weights have been applied at the level of the individual cases.

Alternative (ii): Weighting at the Aggregation of Indicators Level

Indicators may be computed separately by domain and then aggregated across the domains. This may be done to simplify data processing – to reduce the size of the data file to be dealt with at a given time, to avoid weighting the data at the micro-level, to permit simple hand computations after the data have been processed, etc. One reason for seeking such simplifications is that post-stratification weights often become available only after some initial tabulation of the data.

Let:

I_i	=	an indicator computed for domain i
P_i	=	population of domain i as a proportion of the total national population according to the census or other reliable source
I	=	weighted index at the national level, where $I = 3\ P_i\ I_i$ [the summation is over all domains]

Note that P_i refers only to the population that is used as the base in computing the domain indicators I_i. For indicators based on children of a particular age or gender (such as proportion immunized), P_i is the number of children in the domain as a proportion of the total number of children in that category in the whole country. For indicators based on households (e.g., the proportion of households with access to safe water), P_i refers to the number of households in the domain as a proportion of the total households. Hence, the weights to be used for this purpose depend on the base population involved in the computation of the indicators concerned.

In all cases, P_i is computed on the basis of external data, and not on the basis of numbers in the sample. Unlike weighting at the level of individual units, alternative (i), the sample numbers are not involved as such.

The computation of indicator I requires reliable external information in more detail than alternative (i). Thus, it requires external information on the domain distribution of households and population subgroups used in the computation of various indicators, such as children by age and gender. It is more precise than alternative (i) in the sense that it implies post-stratification not only in terms of the distribution of the total population by domain, but also in terms of various population subgroups.

Note also that alternatives (i) and (ii) are not mutually exclusive. Indicators at the domain level may first be computed using individual-level weights if applicable. When those weights do not vary within domains, but vary only across domains, then weighting has no effect on the computation of domain-level indicators. Then the domain-level indicators may be aggregated using alternative (ii).