ANNEX J

FORMULAS FOR ESTIMATING MEANS AND VARIANCES

Formulas for estimating means and variances

Exact formulas for variances using our sampling strategy are not available, since we are using a systematic random sampling procedure (with a random start) in the first stage.[1] As noted in the text, the estimated variance based on a simple random sample provides a conservative estimate for the variance with systematic sampling, so we use the formulas associated with (stratified) simple random sampling in the first stage. The formulas below were derived for our case of a two stage simple random sample without replacement, with stratification in both stages (i.e., stratification of regions in the first stage and stratification of respondent groups in the second stage), and are consistent with the approach and formulas found in Särndal, et al. (2003), chapter 4.[2]

For a population of K potential respondents in our sample universe, a consistent and asymptotically unbiased estimate of the population mean of response variable y is given by:

$$1)\ \hat{\bar{y}} = \frac{\hat{t}}{\hat{K}},$$

where $\hat{t}$ is the estimate of the population total of y and $\hat{K}$ is the estimate of K. These estimates are determined by the following formulas:

$$2)\ \hat{t} = \sum_{h=1}^{H} \frac{N_h}{n_h} ¿¿$$

$$3)\ \hat{K} = \sum_{h=1}^{H} \frac{N_h}{n_h} ¿¿,$$

where h is the first stage stratum number, H is the total number of first stage strata (= 6 for the full population), N_h is the total number of towns in stratum h, n_h is the number of sample towns in

[1] See Särndal, et al. (2003), section 3.4.4 for an explanation of this point.
[2] Our exact case is not shown in Särndal, et al. (2003), so the formulas were derived using the same approach.

stratum h, M_{ig} is the total number of potential respondents of type g in town i, and $\hat{\bar{y}}_{ig}$ is the sample mean of y for respondent group g in town i ($\hat{\bar{y}}_{ig} = \frac{1}{m_{ig}} \sum_{k=1}^{m_{ig}} y_{igk}$, where m_{ig} is the number of sample respondents of type g in town i and y_{igk} is the value of y for respondent k of type g in town i).

A consistent and asymptotically unbiased estimator of the variance of $\hat{\bar{y}}$ is

$$4)\ \hat{V}(\hat{\bar{y}}) = \frac{\hat{V}(\hat{t})}{\hat{K}^2},$$

where $\hat{V}(\hat{t})$ is given by

$$5)\ \hat{V}(\hat{t}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{t,h}^2}{n_h} + ¿ \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} [\sum_{g=1}^{G} M_{ig}^2 \left(1 - \frac{m_{ig}}{M_{ig}}\right) \frac{S_{y,ig}^2}{m_{ig}}] ¿$$

and where

$$S_{t,h}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (¿ \hat{t}_i - \bar{\hat{t}}_h)^2 ¿,$$

$$S_{\hat{y},ig}^2 = \frac{1}{m_{ig} - 1} \sum_{i=1}^{n_h} (¿ y_{igk} - \hat{\bar{y}}_{ig})^2 ¿,$$

$$\bar{\hat{t}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{t}_i, \text{ and}$$

$$\hat{t}_i = \sum_{g=1}^{G} M_{ig} \hat{\bar{y}}_{ig}.$$

Equations 1) to 5) can also be used to estimate the mean and variance of y for subpopulations of the respondent universe. For example, to estimate the mean and variance for a subset of the first stage strata (h), the sums in these equations will be over the selected strata, rather than all six strata. Similarly, to estimate the mean and variance for a subset of the

respondent groups (g), the sums will be over the selected groups rather than all G groups. We use this fact to estimate the means and variances for subpopulations discussed in the text, including communities with vs. without a hospital and communities in different regions.