Model of the Sampling Procedure and Sampling Estimators for the Integrated Household Survey CSD of The Gambia

Christophe Muller

(Preliminary Version: March 2004)

Capacity Building for Economic Management

Project (CBEMP)

Banjul, The Gambia

1 Introduction

This document has several aims: (1) to derive a model of the sampling procedure of the IHS (Integrated Household Survey); (2) to derive the estimators of the total and the means for the characteristics of interest from the IHS; (3) to derive the estimators of the sampling standard errors for the latter estimators.

The model of the sampling procedure is necessary to correct for: (1) some non-randomness at the stage of the systematic sampling method that has been used to draw enumeration areas and households, (2) the drawing of only one primary unit in several strata, which prevents the calculation of intra-strata variance with classical asymptotic formulae of standard errors and (3) other minor inaccuracies in the sampling procedure. It is also necessary to enable analysts to work only with the subsample of households who have been surveyed with daily diaries, which will speed up the arrival of results.

The estimators of the total and the means are important for the tables of consumption, CPI weights and poverty indicators. I introduce an additional post-stratification to increase the accuracy of estimators and take advantage of the coincidence of collection operations from the IHS and the 2003 Census. The estimators of the standard errors will provide a measure of the accuracy of the statistics of interest based on the IHS sample of household.

2 The Sampling Scheme

2.1 The method

The sampling procedure was designed by the CSD. 4800 households have been drawn in the whole country. As a matter of fact, the whole country has been covered at every quarter of the survey. Since the drawing procedures at different quarters were quasi-identical and almost independently implemented, one may also consider that there are four identical sampling schemes, one for each quarter.

The sample scheme has two levels: enumeration areas (EAs) and households. The enumeration areas were drawn from the 2003 updated Census EA list using a systematic drawing method. The listing of households in each drawn EA was then updated by an enumerator for subsequent household selection. The drawing of EAs was stratified by rural-urban areas (12 strata for six geographical divisions that are divided in rural and urban domains + Banjul and Kaninfing that are wholly urban). 240 EAs have been obtained, consisting of 4 sub-samples of 60 EAs surveyed at each quarter. Households in 'protected areas' were not surveyed (mostly military populations).

At the second level, in each enumeration area two subsamples of 10 households each in each EA have been drawn independently according to a systematic

drawing method. The first subsample of 10 households was submitted to daily diaries, but not the second one.

Six teams made of 6 supervisors and 30 enumerators were assigned to the different geographical locations. Each enumerator covered 40 households in two EAs by quarter. Among these 40 households, 20 households were selected (10 per EA) and were subject to interviews based on the daily diary. In total the enumerators stayed 6 weeks in each EA.

2.2 Some issues

During the systematic sampling of EAs and households, the lists from which the units were drawn were characterised by non-random geographical organisation. That is, the list of EAs to be drawn from was ranked according to the geographical location of these EAs, neighbouring EAs occurring together in the list. The same issue occurred for the lists of households in each EA. In that case, the sampling of households was separated into two independent drawings to constitute two subsamples of 10 households each in each EA. As mentioned above, only one of these subsamples of 10 households was surveyed using daily diaries.

Thus, because of the systematic sampling approach that has been used for EAs and households, the probability of drawing units that are well spread geographically is augmented as compared with random systematic drawing. This situation can be dealt with by modelling the actual sampling process. I shall present a model of it in the next section.

For future surveys, it is advisable to randomly rerank the list from which units are drawn, before the beginning of the systematic drawing process. This can be done very easily with any spread sheet software (Excell for example) or any statistical software.

I found another shortcoming in the way the number of surveyed EAs was decided in each stratum at each quarter. Indeed, in some cases only one EA was drawn in the stratum. In that case, it is not possible to calculate the intra-stratum variances. Therefore, the total variance of the estimators of the mean (or the total) of the characteristics of interest cannot be derived from the initial sampling scheme using the usual asymptotic formulae for standard errors. Note however that when considering all quarters together there is no isolated EA in any given stratum since at least four EAs can be found in any stratum over the whole year.

A third issue is that the EAs were drawn without replacements, not only at each quarters, but for all subsequent quarters. That is, the EAs that had been selected for previous quarters were removed from the list to draw from at a given quarter. This creates a slight temporal dependence between the four quarterly sampling schemes.

Finally, a minor problem is that the number of drawn units at any stage was almost proportional to the size of the population in which it was drawn.

This procedure is suboptimal and better accuracy would have been reached by drawing more than proportionally in subpopulations with suspected higher variability.

These types of issues are frequent in African sampling schemes[1]. They may lead to disaster if samples with individuals of too close characteristics are drawn as a result of the systematic ranking in the list and of the other imperfections. However, this does not seem to be the case here since far apart units have a larger probability of being drawn than geographically close units. In fact, the representativity of the final sample is likely to be better than what had been anticipated. The opportunity to improve efficiency is good news, although this also implies that the sample weights and the estimators have to be revised accordingly.

All this necessitates to explicitly develop a model of the sampling scheme that could be used to derive estimators and estimators of their variances. A posteriori modelling of sampling schemes is not unknown in Africa (e.g., Roy, 1984).

Another issue that may bear consequences for the sampling estimators is that delays have occurred for the collection in various EAs. Indeed, the collection may end about three months later than expected in some EAs. Then, the collection quarters may have to be stretched to groups of five months in some cases. Alternatively, some households of the initial sample may have to be dropped. It is difficult to assess at the moment what is the extent of this problem which will have to be investigated at the end of the collection.

Despite all these shortcomings the work done by the agents of the CSD must be recognized. The sampling scheme, although not optimal remains basically sound and should not lead to notable bias. If anything, the variance of estimators of means or totals for the survey should be smaller than expected.

3 The Model of the Sampling Procedure

The model of the sampling procedure can be presented in six steps with only two sampling stages:

(1) Exhaustive selection of strata that are the six geographical divisions of The Gambia each divided into urban and rural sectors plus Banjul and Kanifing that are totally urban. One obtains 14 strata.

(2) Definition of Pseudo-Strata of EAs (*EA-pseudo-strata* or *EAPS*) that equiprobably divide the above strata into groups of EAs according to the ranking used for the systematic sampling of the EAs. The 2003 updated Census EA list is used as a sampling basis.

(3) Equiprobable drawing of 2 or 3 EAs in each pseudo-strata of EAs.

(4) Definition of Pseudo-Strata of households (*HH-pseudo-strata* or *HHPS*) in each drawn EA.

[1] For example, I found it in countries as different as Rwanda and Tunisia.

(5) Equiprobable drawing of 2 households in each pseudo-strata of households.

Some additional comments about steps (2) and (4) are useful. The definition of pseudo-strata is a way to approximate the mixture of randomness and spatial stratification in the systematic sampling. Then, it is natural to sequentially define the pseudo-strata by following the sampling list of EAs in each stratum so that each pseudo-stratum has approximately the same size.

The number of pseudo-strata of EAs in each stratum is limited by the facts that only 60 EAs in total have been drawn for each quarter, and at least 2 EAs are necessary to be able to estimate the intra-pseudo-stratum variance for the characteristics of interest. This leaves us with 2 or 3 EAs to draw in each pseudo-stratum. It is not necessary to decompose into pseudo-stata the strata including less than 5 EAs for the whole year.

Similar issues arise for the definition of the pseudo-strata of households in stage (4). Here, it is possible to draw groups of 2 households in each pseudo-stratum of households. This is possible because no clustering has been used at the stage of household drawing. The two-households groups must be taken each in one of the two subsamples of 10 households to preserve the possibility of working only with the households surveyed with daily diaries.

Since only 1 household is actually *randomly* selected in each 10-households subsamples during the systematic sampling procedure, two sampling strategies could be distinguished for (1) analyses based only on the daily diaries (using the subsample of 10 households in each EA) and for (2) analyses based on the whole questionnaire for which records associated with all the 20 households must be used. However, I do not explicitly develop this possibility in this document. At this stage I only deal with the selection of EAs and I can neglect the division of the household samples in two.

To simplify, I present the EAPSs and the HHPSs for the daily diary subsample only and the whole year. Similar pseudo-strata can be defined for the second subsample of households or for quarterly subsamples. Consider the following graph describing how EAPS can be defined from the list of EAs in a given stratum.

Graph of the pseudo-strata:

Numbers of the EAs in strata h:
1 2 3 4 5 6 .....25.... 50
<—pseudo-strata 1————->< ——pseudo-strata 2——>

51.....100.....249 250
<— —pseudo-strata 10—>

The above graph corresponds for example to the strata Erewan-Rural. This stratum has 250 EAs in total and 20 EAs have been drawn[2]. Since one assumes

[2] In total 20 EAs for the year, but only 5 EAs per quarter.

that the sequence of the numbers of the EAs (from 1 to 250) corresponds to a geographical link, it is possible to divide this sequence of the numbers into 10 consecutive domains so as to obtain two drawn EAs in each of these domains. Then, the first pseudo-strata corresponds to EAs with numbers from 1 to 25, the second strata to EAs with numbers from 26 to 50, etc.

The method with 5 EAs per quarter is still valid. In that case the list of EAs in the stratum Erewan-rural would go from number 1 to a number about 62. Thus, for each quarter only 5 EAs have been drawn in this strata and only two EAPSs can be defined if one wants work with quarters. This also implies to calculate different sampling weights depending on if one wants yearly and quarterly indicators.

The same process can be implemented to define the HHPSs. In that case, the number of drawn households in each HHPS is 2 and the number of drawn households in each drawn EA is 20 among which there are the 10 households surveyed with the daily diary records. Let us first focus on this subsample of households surveyed with the diaries. Then, exactly 5 pseudo-strata have to be defined in each EA. As above, this can be done by sequentially dividing the list of the household numbers into five equiprobable parts as follows.

In the graphic example, I assume that there is exactly 100 households in the EA where one wants to select the households to survey.

Numbers of the households in the EA:

1 2 3 4 5 6 7 8 9 10 .....20 21 .....40 41.....

<—pseudo-strata 1————————-> <–pseudo-strata 2 —> ....

..... 81.....100

<—pseudo-strata 5 ———>

For the cases such that the other subsample of 10 households (the ones only surveyed with retrospective questionnaire) is also used in the analysis, the HHPSs are the same. It is only necessary to allocate the new drawn households to these pseudo-strata, which should deliver 4 selected households in total by HHPS. With this model in hand, I can define the estimators of the indicators of interest.

4 The Estimators of the Totals and Means for the Characteristics of Interest

4.1 The basic estimators

At the beginning of the mission, estimators for the characteristics to estimate and for the sampling standard errors were missing. I now fill this gap. I start with a few definitions.

Let y_i be the value of a numerical variable of interest for household i. Let N be the sample size. $Y = \sum_{i=1}^{N} y_i$ is the total of the variable over the population. $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$ is the mean of the variable over the population.

$S^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{Y})^2$ is the variance of the variable over the population.

In both stages the units have been drawn without replacement. Working with the usual approximation of drawing with replacement implies that the intra-pseudo-strata variances at the stage of drawing EAs and the intra-pseudo-strata variances at the stage of drawing households may be overestimated. However, given the problems to deal with in this scheme and the need to obtain with relatively simple formulae that will be easy to implement at the CSD, I choose to neglect these variance overestimation factors, but uniquely in the case of replacements of EAs from one quarter subsample to another. Indeed, in practice the fraction population correction can be ignored when the sampling fraction does not exceed 5 percent, and often even 10 percent. This situation should be close to what I shall obtain after all non-response problems are accounted for, although it is difficult to ascertain it before the collection is completed. Then, I shall know the exact extent of non-response problems caused by delays in the collection. Hopefully, the approximation approach should not substantially affect the inference results of this survey.

Alternatively, deriving formulae to account for the absence of replacements across quarter subsamples would be possible, although cumbersome. An adjustment of formulae can be implemented once the data is entered and what is available is exactly known. Indeed, there exists a possibility that some EAs and some households would be eliminated if the timing of the corresponding interviews has occurred too far from what had been planned. Also, the refinement of using formulae corresponding to non-replacement may lose much of its interest as compared with the arising non-response problem.

Because of the stratification, summarized in the definition of the pseudo-strata, the values of the drawing probabilities depend on the considered pseudo-strata $h = 1, ..., H$ for the EA level and $h' = 1, ..., H_j^h$ for the household level in EA j of the EAPS h. To indicate this dependence, I add *upper indices* h and h' to the relevant symbols. For example, N^h is the sample size in EAPS h, $\bar{Y}^h$ is the mean of variable y in EAPS h, Y^h is the total of y over EAPS h, P_j^h is the probability of drawing EA j in EAPS h, $S^{2,h}$ is the variance of variable y in EAPS h, etc. Respective sample analogs are denoted $n^h, \bar{y}^h, s^{2,h}$. Naturally, $Y = \sum_{h=1}^{H} Y^h$, then the knowledge of all EAPS totals gives the total over the whole population. Therefore, one can start by solving the problem for a given EAPS h.

The estimators of the totals are based on the popular Horwitz-Thompson formulae. In these conditions and for this pseudo-stratum, I can define the estimator of the total of y as a function of the totals of y for each drawn EA

and of their inclusion probabilities:

$$\hat{Y}^h = \sum_{j=1}^{J^h} \frac{Y_j^h}{P_j^h} I \left[\begin{array}{c} \text{EA } j \text{ has been drawn} \\ \text{in EAPS } h \end{array} \right], \tag{1}$$

where $I[.]$ is the Kronecker index, J^h is the number of EAs in the EA-pseudo-strata h and P_j^h is the inclusion probability of EA j in EAPS h. Later on, I shall just need to complete this formula by specifying how the Y_j^h are estimated.

If $\frac{n_h}{n} = \frac{N_h}{N}$ it is a *representative* stratified sample. This situation is close to what has been done in the actually implemented sampling procedure. Then, simple means over the observed sample could give a first idea about the mean over the whole population. This may be convenient although this approach delivers inefficient estimators. In future surveys, it would be better for the precision of the results to select relatively more units in strata with high expected dispersion.

The Horwitz-Thompson estimator of the total of a variable y for EAPS h ($h = 1, ..., H$) and for HHPS h' in each EA j of the considered EAPS ($h' = 1, ..., H_j^h$) is:

$$y^{hh'} = \sum_{j=1}^{M^h} \frac{M^h}{m^h} \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}}{n_j^{hh'}} y_{jk}^{hh'} I \left[\begin{array}{c} (j,k) \text{ has been drawn in EAPS } h \\ \text{and in HHPS } h' \end{array} \right],$$

where $j(= 1, ..., M^h)$ is the index of the EAs in the EAPS h,

$k(= 1, ..., N_j^{hh'})$ is the index of the households in the HHPS h' of EA j in the EAPS h,

M^h = number of EAs in EAPS h,

m^h = number of drawn EAs in EAPS h (in general 2),

$N_j^{hh'}$ = number of households in HHPS $h'$of EA j in EAPS h,

$n_j^{hh'}$ = number of drawn households (normally 2) in HHPS h' of EA j in EAPS h.

$y_{jk}^{hh'}$ is the level of variable y for household k of HHPS h' in EA j of EAPS h. $y_{jk}^{hh'}$ is its total over HHPS h' of EAPS h.

Finally, the Horwitz-Thompson estimator of the population total is:

$$\begin{aligned} \hat{Y} &= \sum_{h=1}^{H} \sum_{j=1}^{M^h} \frac{M^h}{m^h} \sum_{h'=1}^{H_j^h} \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}}{n_j^{hh'}} y_{jk}^{hh'} . I \left[\begin{array}{c} (j,k) \text{ has been drawn in EAPS } h \\ \text{and in HHPS } h' \end{array} \right] \\ &= \sum_{i=1}^{n} w_i y_i, \end{aligned}$$

where $w_i = \frac{M^h}{m^h} \frac{N_j^{hh'}}{n_j^{hh'}}$ is the *sampling weight* of household i. n is the sample size. h is the index of the EAPS, j is the index of the EA, h' is the index of the HHPS, k is the index of the household. H is the number of EAPS, H_j^h is the number of HHPSs in EA j of EAPS h.

Then, the quantities $m^h, M^h, N_j^{hh'}, n_j^{hh'}$ must be calculated for all values of j, h, h' to be able to calculate the sampling weights w_i. The estimator of the mean, $\hat{y}$, can be obtained by dividing $\hat{Y}$ by the population size or by a consistent estimator of the population size.

These estimators that are unbiased may be sufficient to yield satisfactory estimation results. However, it is possible to improve on them by using census information about household size.

4.2 The estimators based on post-stratification

Although, there is a lot of a priori stratification for the sampling stage of the EAs, this is not so much the case for the sampling stage of households. Since the 2003 Census data has been collected in the middle of the collection period of the IHS, there is an opportunity to add a degree of post-stratification to the drawing of households in each EA. Alternatively, we could use the exhaustive list of households with a few household characteristics that has been established in each surveyed EA of the IHS, before the beginning of the collection. In any case, the household size information can be used to anchor the post-stratification since household size is expected to be very correlated with household consumption, which is the central variable for the survey objectives. In practice the post-strata should merely be defined so as to separate large and small households.

Post-stratifying may be important because the short-track strategy of producing early economic analyses from the IHS is based on daily diary records that have been collected only for 10 households in each EA, instead of 20 households for the other questionnaires. Then, the used sample (and the subset of questionnaires) in that case is smaller than initially planned and something should be done to compensate the subsequent loss in prevision. It seems worthwhile to improve sampling estimators because the intensity of the collection process with these diaries suggests that data contamination is less than usual. Finally, as mentioned before, the sampling procedure remains close to representative sampling, i.e. the number of drawn units is roughly proportional to the size of the population in which they are drawn. In this situation, opportunities of improving the efficiency of the estimation by over-representing strata with high dispersion have been lost and the post-stratification would help to compensate for it.

In the case of post-stratification, the formulae of the Horwitz-Thompson estimators are the same than for a priori stratification (changing HHPS into post-strata), but here the subsample size in the post-strata is random. Consequently, the formulae of the variances and of their estimators need to be revised.

The definition of the pseudo-strata also need to be revised. Indeed, a too thin spatial pseudo-stratification is not compatible with an additional degree of post-stratification with respect to household size. Then, in that case the pseudo-strata defined to correct for the imperfections of the systematic sampling will simply be replaced by two sequential and (quasi-)equiprobable parts of the list of households in the selected EAs. Each of these new pseudo-strata will be further divided in two post-strata that separate small and large households.

A preliminary estimate of the median household size will provide the criterion for defining these post-strata. Moreover, the weights for the drawing of households in each EA (i.e. calculated from new values of the $N_j^{hh'}$ and $n_j^{hh'}$) need to be recalculated by using the Census data. With this approach the probability of not observing any household in any post-strata, which would prevent the calculus of inclusion probabilities, is negligible.

In theory, it is possible to compare formulae of the variances of the estimators with and without post-stratification[3]. The condition for preferring post-stratification estimators is likely to be satisfied when a rough nomenclature is used for the post-stratification. Then, using household size only to separate large and small households seems appropriate to define the post-stratification.

The Census (April 2003) provides estimates of household size at a time in the middle of the IHS collection operations (starting in February 2003). Then, the proposed post-stratification should work well without noticeable bias. To be able to apply it the number of household of large and small sizes must be calculated in each 'new HHPS' in the drawn EAs from the 2003 Census data. As mentioned above, there are only two 'new HHPSs' per drawn EA.. As a matter of fact, it is even possible and probably better to use the lists of households established by the IHS enumerators at the beginning of the survey in the surveyed EAs. Indeed, these lists include information about the household size.

Finally, the post-stratification will be the opportunity of preparing the Census data for the poverty mapping analysis.

4.3 The clusters

Using clusters saves a lot of resources for the collection operations, although it may imply some loss of precision. Often, the group of drawn households in each EA are neighbours and can be considered as a sampling cluster. However, in this survey the selected households have not been drawn as clusters. Then, lower variances of estimators than usual should be obtained and no correction for clusters is necessary. I discuss the sampling standard errors in the next section.

[3] as in Grosbras (1987).

5 The Estimators of the Sampling Standard Errors

5.1 Classical asymptotic formulae

The derivation of the classical estimators of sampling standard errors is as follows. The formula of an unbiased estimator of the variance of the Horwitz-Thompson estimator for a two-degrees scheme is

$$\hat{V}(\hat{T}_{HT}) = \sum_h \frac{\hat{\Lambda}^h}{P^h} I\left[\text{Primary unit } h \text{ is drawn}\right] + \hat{\delta},$$

$$\text{where } \hat{\Lambda}^h = \frac{1}{2} \sum_{\substack{i \neq j, \\ i,j \in <h>}} \sum \frac{P_i^h P_j^h - P_{i,j}^h}{P_{i,j}^h} \left(\frac{Y_i^h}{P_{i,}^h} - \frac{Y_j^h}{P_j^h} \right)^2$$

$$\text{and } \hat{\delta} = \sum_h \left(\hat{T}_h \right)^2 \frac{\left(1 - P^h\right)}{\left(P^h\right)^2} I\left[\text{Primary unit } h \text{ is drawn}\right]$$

$$+ \sum_{h \neq l,} \sum \frac{\left(P^{hl} - P^h P^l\right)}{P^h P^l P^{hl}} \hat{T}^h \hat{T}^l I\left[\text{Primary units } h \text{ and } l \text{ are drawn}\right].$$

where $\hat{T}_{HT}$ is the corresponding Horwitz-Thompson estimator of the total, $< h >$ is the set of indices of households in the primary unit h, P^h, P^l, P^{hl} are respectively the probability of inclusion of strata h, l and (h, l), $P_i^h P_j^h$ and $P_{i,j}^h$ are the inclusion probability of second stage units (respectively for $i, j, (i, j)$ conditionally on drawing the first stage unit h.

When the drawings at the second degree are equiprobable without replacement, we obtain $\hat{\Lambda}^h = \frac{N_h(N_h - n_h)}{n_h} s^{h,2}$ and $\hat{T}^h = N_h \bar{y}_h$, where N_h is the total number of primary unit, n_h is the drawn number of primary unit, and $\bar{y}_h$ is the mean of y on the primary unit h. If I decide to neglect the replacements, it simplifies to $\hat{\Lambda}_h = \frac{(N_h)^2}{n_h} s_h^2$.

It remains (1) to introduce the two additional levels of pseudo-stratification, and (2) to write the formulae of P^h, P^l, P^{hl} corresponding the the first stage of the sampling. In the case of a simple drawing stage, the latter corresponds to P^h =(number of drawn primary units)/(total number of primary units) and P^{hl} =(number of drawn pairs of primary units)/(total number of pairs of primary units) . Introducing the pseudo-stratifications is straightforward since in that case all pseudo-strata are drawn and inter-pseudo-strata variances cancel out[4]. However, the number of total primary units and the number of drawn primary units will vary with the EAPS.

The calculation goes as follows

$$\hat{Y} = \sum_{h=1}^{H} \sum_{j=1}^{M^h} \frac{M^h}{m^h} \sum_{h'=1}^{H_j^h} \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}}{n_j^{hh'}} y_{jk}^{hh'} . I\left[.\right], \text{ where in this calculation } I\left[.\right] \text{ plays}$$

[4] This can be seen for example in the analog formula with $P_h = P_l = P_{hl} = 1$, or simply by noticing that drawing in different strata are independent, which permits the decomposition of the variance as the sum of the variances in the strata.

the role of the dummy indicating that the considered units have been drawn. $\hat{Y} = \sum_{h=1}^{H} \hat{Y}^h$, which defines $\hat{Y}^h$. Since all EAPS h are drawn, the formula of the variance decomposition simplifies and I obtain a decomposition of the *estimator* of the variance of $\hat{Y}$ that is $\hat{V}(\hat{Y}) = \sum_{h}^{H} \hat{V}(\hat{Y}^h)$. The calculation of the estimator

of each $\hat{V}(\hat{Y}^h)$ is more complicated and involves variances inter-EAs.:

$$\hat{V}(\hat{Y}^h) = \sum_{j=1}^{M^h} \frac{M^h}{m^h} \hat{V}(\hat{Y}_j^h).I\,[.] + \sum_{j=1}^{M^h} \frac{1-P_j^h}{P_j^h} \left(\hat{Y}_j^h\right)^2 .I\,[.] + \sum_{j=1}^{M^h} \sum_{\substack{j'=1 \\ j\neq j'}}^{M^h} \frac{P_{jj'}^h - P_j^h P_{j'}^h}{P_j^h P_{j'}^h P_{jj'}^h} \hat{Y}_j^h \hat{Y}_{j'}^h .I\,[.].\text{with}$$

obvious notations for the probabilities that are conditional on the drawing of EAPS h (not indicated). In that case, $P_j^h = P_{j'}^h = m^h/M^h$ and $P_{jj'}^h = \frac{m^h(m^h-1)}{M^h(M^h-1)}$ because the EAs are drawn without replacements. This yields

$$\hat{V}(\hat{Y}^h) = \sum_{j=1}^{M^h} \frac{M^h}{m^h} \hat{V}(\hat{Y}_j^h).I\,[.] + \sum_{j=1}^{M^h} \left(\frac{M^h}{m^h} - 1\right) \left(\hat{Y}_j^h\right)^2 .I\,[.] + \sum_{j=1}^{M^h} \sum_{\substack{j'=1 \\ j\neq j'}}^{M^h} \frac{(m^h - M^h).M^h}{(m^h)^2(m^h-1)} \hat{Y}_j^h \hat{Y}_{j'}^h .I\,[.].$$

This is finally equivalent to

$$\hat{V}(\hat{Y}^h) = \sum_{j=1}^{M^h} \frac{M^h}{m^h} \hat{V}(\hat{Y}_j^h).I\,[.] + \sum_{j=1}^{M^h} \frac{(M^h - m^h).M^h}{m^h(m^h-1)} \left(\hat{Y}_j^h - \bar{Y}^h\right)^2 .I\,[.], \text{ where } \bar{Y}^h$$

is the mean of the $\hat{Y}_j^h$.

I now give the formula for the estimator $\hat{V}(\hat{Y}_j^h)$ where $\hat{Y}_j^h = \sum_{h'=1}^{H_j^h} \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}}{n_j^{hh'}} y_{jk}^{hh'} .I\,[.]$:

$\hat{V}(\hat{Y}_j^h) = \sum_{h'=1}^{H_j^h} \hat{V}(\hat{Y}_j^{hh'})$ for all j and h, because all HHPS are drawn with probability one in each drawn EA, and the calculus simplifies.

Then, $\hat{Y}_j^{hh'} = \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}}{n_j^{hh'}} y_{jk}^{hh'} .I\,[.]$ and and estimator of its variance is

$\hat{V}(\hat{Y}_j^{hh'}) = \frac{1}{2} \sum_{k=1}^{N_j^{hh'}} \sum_{k'=1}^{N_j^{hh'}} \frac{\left(P_k^{hh'j} P_{k'}^{hh'j} - P_{kk'}^{hh'j}\right)}{P_{kk'}^{hh'j}} \left(\frac{y_{jk}^{hh'}}{P_k^{hh'j}} - \frac{y_{jk'}^{hh'}}{P_{k'}^{hh'j}}\right)^2 .I\,[.]$ again with obvious notations for the conditional probabilities such that $P_{k'}^{hh'j} = P_k^{hh'j} = \frac{n_j^{hh'}}{N_j^{hh'}}$.

In the case of drawing without replacements it simplifies to the following unbiased estimator

$\hat{V}(\hat{Y}_j^{hh'}) = \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}\left(N_j^{hh'} - n_j^{hh'}\right)}{n_j^{hh'}\left(n_j^{hh'}-1\right)} \left(y_{jk}^{hh'} - \bar{y}_{j'}^{hh'}\right)^2 .I\,[.]$, where $\bar{y}_{j'}^{hh'}$ is the mean of the $y_{jk}^{hh'}$.

Gathering the various above expressions provides the final estimator

$$\hat{V}(\hat{Y}) = \sum_{h}^{H} \left\{ \begin{array}{c} \sum_{j=1}^{M^h} \frac{M^h}{m^h} \sum_{h'=1}^{H_j^h} \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}\left(N_j^{hh'}-n_j^{hh'}\right)}{n_j^{hh'}\left(n_j^{hh'}-1\right)} \left(y_{jk}^{hh'} - \bar{y}_{j'}^{hh'}\right)^2 .I\left[.\right] \\ + \sum_{j=1}^{M^h} \frac{(M^h-m^h).M^h}{m^h(m^h-1)} \left(\hat{Y}_j^h - \bar{Y}^h\right)^2 .I\left[.\right]. \end{array} \right\}$$

The variance of the Horwitz-Thompson estimator of the mean is the variance of the Horwitz-Thompson estimator of the total divided by N^2.

5.2 Balanced repeated replications based on EAPS

The aggregation of the pseudo-strata in which only one EA has been selected with another strata complicates matters and, strictly speaking, makes usual asymptotic formulae imperfect.

Another issue is that of the drawing without replacement which have been used at several places in the sampling procedure. All these complications and the approximations that are used to deal with them may make the classical formulae of standard errors less robust than one would like.

To adapt to this situation, I now propose an estimator for sampling standard errors that is a combination of 'linearization' estimators obtained using balanced repeated replications[5] and that is simpler and quicker than stratified bootstrap procedures. Howes and Lanjouw (1998) have shown that the sampling design can substantially modify the estimated standard errors for poverty measures. Consequently, my estimators for the sampling standard errors account for the sample design, here mostly coming from the stratification.

A mean indicator of a given variable y for a sub-population is estimated by a ratio of the type $\bar{y}_x^* = \frac{z^*}{x^*}$, where * denotes the Horwitz-Thompson estimator for a total (sum of values for the variable of interest weighed by the inverse of the inclusion probability). For example, z is the sum of the poverty in the sub-population and x is the size of the sub-population.

An approximation of the variance associated with the sampling error is then $V\left(\bar{y}_x^*\right) = \left[V\left(z^*\right) - 2\bar{y}_x^* Cov\left(z^*, x^*\right) + \left(\bar{y}_x^*\right)^2 V\left(x^*\right)\right] / \left(x^*\right)^2$,

obtained from a Taylor expansion at the first order from function $Y = f(Z/X)$ around (Ey^*, Ex^*) and because $Ez^* \neq 0$ and x^* does not cancel. Here, the appropriate expectations are estimated by x^* and $\bar{y}_x^*$.

Using balanced repeated replications implies to define balanced sub-samples that are consistent with the stratification. As mentioned before, I have divided the sample of EAs (first actual stage of the sampling since all the strata are drawn) in $NP1$ EA-pseudo-strata ($a = 1$ to $NP1$)[6]. This enables us to group together the EAs sharing similar characteristics, and to a priori reduce the variance intra-strata. The fact that several EAs are assumed to have been drawn

[5] Krewski and Rao (1981), Shao and Rao (1993) .

[6] $NP1$ needs to be calculated.

in each pseudo-strata allows the estimation of the variance intra-pseudo-strata. However, I neglect here the HHPS which should deliver highly overestimated standard errors.

As seen above, the Horwitz-Thompson formula leads to:

$$z^* = \sum_{h=1}^{H} \sum_{j=1}^{M^h} \frac{M^h}{m^h} \sum_{h'=1}^{H'_h} \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}}{n_j^{hh'}} z_{jk}^{hh'} = \sum_{i=1}^{n} w_i z_i$$

and

$$x^* = \sum_{h=1}^{H} \sum_{j=1}^{M^h} \frac{M^h}{m^h} \sum_{h'=1}^{H'_h} \sum_{k=1}^{N_j^{hh'}} \frac{N_j^{hh'}}{n_j^{hh'}} x_{jk}^{hh'} = \sum_{i=1}^{n} w_i x_i.$$

$Cov(z^*, x^*)$ is estimated by

$$\begin{aligned} \hat{C}ov(z^*, x^*) &= \frac{1}{NP1(NP1-1)} \sum_{h=1}^{H} \sum_{j=1}^{M^h} (z_j^{h*} - z^*)(x_j^{h*} - x^*) \\ &= \frac{1}{NP1(NP1-1)} \sum_{a=1}^{NP_1} (z_a^* - z^*)(x_a^* - x^*). \end{aligned}$$

where z_a^* and x_a^* are the Horwitz-Thompson estimators in EAPS a.

$V(x^*)$ is estimated by

$$\begin{aligned} \hat{V}(x^*) &= \frac{1}{NP1(NP1-1)} \sum_{h=1}^{H} \sum_{j=1}^{M^h} (x_j^{h*} - x^*)^2 \\ &= \frac{1}{NP1(NP1-1)} \sum_{a=1}^{NP1} (x_a^* - x^*)^2. \end{aligned}$$

5.3 Balanced repeated replications based on household-pseudo-strata

The approach is similar to that of the previous sub-section, but now the strata definition for applying the balanced repeated replications are the HHPS. This leads to

$$\begin{aligned}\hat{Cov}(z^*, x^*) &= \frac{1}{NP2(NP2-1)} \sum_{h=1}^{H} \sum_{j=1}^{M^h} \sum_{h'=1}^{H'_h} (z_j^{hh'*} - z^*)(x_j^{hh'*} - x^*) \\ &= \frac{1}{NP2(NP2-1)} \sum_{a=1}^{NP2} (z_a^* - z^*)(x_a^* - x^*)\end{aligned}$$

and

$$\begin{aligned}\hat{V}(x^*) &= \frac{1}{NP2(NP2-1)} \sum_{h=1}^{H} \sum_{j=1}^{M^h} \sum_{h'=1}^{H'_h} (x_j^{hh'*} - x^*)^2 \\ &= \frac{1}{NP2(NP2-1)} \sum_{a=1}^{NP2} (x_a^{h*} - x^*)^2,\end{aligned}$$

where $NP2$ is the number of HHPS (to be calculated).

Experiments could be done with the estimators of standard errors from the three methods: the two formulae of balanced repeated replications and approximate classical asymptotic formulae. This would constitute a control of the validity of these estimators.

5.4 Adjustment of standard errors in the case of the post-stratification

When post-stratification is introduced, the adjusted formulae are as follows for the case of a single stage sampling.

$\hat{M}_{post} = \sum_{h=1}^{K} \frac{N_h}{N} \bar{y}_h$ as before and $E(\hat{M}_{post}) = \bar{Y}$,

but now $V\left(\hat{M}_{post}\right) = \sum_{h=1}^{K} \frac{(N_h)^2}{N} S_h^2 \left[E\left(\frac{1}{n_h}\right) - \frac{1}{N_h}\right]$

$\simeq \frac{N-n}{nN} \sum_{h=1}^{K} \frac{N_h}{N} S_h^2 + \frac{N-n}{n^2 N} \sum_{h=1}^{K} S_h^2$ if n is large, where $S_h^2 = \frac{1}{N_h} \sum_{h=1}^{N^h} (y_i - \bar{y}^h)^2$.

This type of formula can be used to modify the previous estimators of the variances for total and mean estimators.

Appendices

Appendix 1: References:

Cochran, W.G., "*Sampling Techniques*," John Wiley and Sons, 1977.

Gouriéroux, "*Théorie des sondages*," Economica, Paris, 1981.

Grosbras, J.-M., "*Méthodes Statistiques des Sondages*," Economics Ed., Paris, 1987.

Howes, S. and Lanjouw, J.O, "Does Sample Design Matter for Poverty Rate Comparisons," *Review of Income and Wealth*, Ser. 44, No. 1, March 1998.

Kish, L., "*Survey Sampling*," John Wiley and Sons, 1967.

Krewski, D., and J.N.K., Rao, "Inference from Stratified Sample," *Annals of Statistics*, vol. 9, pp 1010-1019, 1981.

Roy, G., "Enquête Nationale Budget Consommation Rwanda: Plan de sondage," INSEE, Département de la Coopération, Paris, 1984.

Shao, J. and J.N.K. Rao, "Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples," *Sankhya: The Indian Journal of Statistics*, vol. 55, Series B, Part 3, 393-414, 1993.

Appendix 2: File of population data:

to insert in pdf

Appendix 3: Information to search for the calculus of post-stratified estimators:

- Establishment of the list of non-respondent households or households that will be excluded because the corresponding collection has taken place with too much delay.

- Definition of the EA-pseudo-strata and the Household-pseudo-strata for the year samples from the survey listings.

- Definition of the EA-pseudo-strata and the Household-pseudo-strata for the quarter sub-samples from the survey listings.

- Calculation of the number of large and small households in the new dichotomic household-pseudo-strata.

As before, assume that there is exactly 100 households in the EA where one wants to select the households to survey.

Numbers of the households in EA a:

1 2 3 4 5 6 7 8 9 10 .....50 51 100

<—pseudo-strata 1————-> <–pseudo-strata 2 ————————>

- Calculation of final weights from the population data (see Appendix 2), accounting for the corrections for non-responses and the post-stratification information.

The Gambia

INTEGRATED HOUSEHOLD SURVEY - 2002/03

Sampling Distribution of Households for 2003 Integrated Household Survey

Strata	Area	No. of EAs	No. of HHs	Population	% of HHs	Target No. of HHs.	Determined No. of EAs	Determined No. of Sampled HHs	No. of EAs Adjusted for Quarterly Spreads	No. of HHs Adjusted for Quarterly Spreads	No . of Quarterly EA Spreads				No. of Quarterly-lh Spreads			
											1	2	3	4	1	2	3	4
	Banjul-Urban	**92**	**6756**		**0,04405**	**211**	**11**	**220**	**12**	**240**	**3**	**3**	**3**	**3**	**60**	**60**	**60**	**60**
	Banjul South	19	1314		0,008567	**41**	**2**	40										
	Banjul Central	19	1395		0,009096	**44**	**2**	40										
Banjul	**Banjul North**	52	4047		0,026387	**127**	**6**	120										
Kanifing	**Kanifing-Urban**	**634**	**44647**		**0,291105**	**1397**	**70**	**1400**	**68**	**1360**	**17**	**17**	**17**	**17**	**340**	**340**	**340**	**340**
	TOTAL	**724**	**43214**		**0,281761**	**1352**	**68**	**1360**	**68**	**1360**	**17**	**17**	**17**	**17**	**340**	**340**	**340**	**340**
	Urban	**102**	**6809**		**0,044396**	**213**	**11**	**220**	**12**	**240**	**3**	**3**	**3**	**3**	**60**	**60**	**60**	**60**
Brikama	**Rural**	**622**	**36405**		**0,237366**	**1139**	**57**	**1140**	**56**	**1120**	**14**	**14**	**14**	**14**	**280**	**280**	**280**	**280**
	Kombo North	317	20406		0,13305	**639**	**32**	640										
	Kombo South	112	6130		0,039968	**192**	**10**	200										
	Kombo Central	158	9277		0,060487	**290**	**15**	300										
	K-Central Urban	102	6809		0,044396	**213**	**11**	220										
	K-Central Rural	56	2468		0,016092	**77**	**4**	80										
	Kombo East	48	2691		0,017546	**84**	**4**	80										
	Foni Brefet	22	1297		0,008457	**41**	**2**	40										
	Foni Karanai	25	1238		0,008072	**39**	**2**	40										
	Foni Kansala	20	1087		0,007087	**34**	**2**	40										
	Foni Bondali	11	539		0,003514	**17**	**1**	20										
	Foni Jarrol	11	549		0,00358	**17**	**1**	20										
	TOTAL	**151**	**8746**		**0,057025**	**274**	**14**	**280**	**16**	**320**	**4**	**4**	**4**	**4**	**80**	**80**	**80**	**80**
	Urban	**28**	**2139**		**0,013947**	**67**	**3**	**60**	**4**	**80**	**1**	**1**	**1**	**1**	**20**	**20**	**20**	**20**
Mansakonko	**Rural**	**123**	**6607**		**0,043079**	**207**	**10**	**200**	**12**	**240**	**3**	**3**	**3**	**3**	**60**	**60**	**60**	**60**
	Kiang West	33	1680		0,010954	**53**	**3**	60										
	Kiang Central	15	767		0,005001	**24**	**1**	20										
	Kiang East	12	596		0,003886	**19**	**1**	20										
	Jarra West	52	3592		0,02342	**112**	**6**	120										
	J-West Urban	28	2139		0,013947	**67**	**3**	60										
	J-West Rural	24	1453		0,009474	**45**	**2**	40										
	Jarra Central	13	608		0,003964	**19**	**1**	20										
	Jarra East	26	1503		0,0098	**47**	**2**	40										
	TOTAL	**316**	**19693**		**0,128401**	**616**	**31**	**620**	**28**	**560**	**7**	**7**	**7**	**7**	**140**	**140**	**140**	**140**
	Urban	**66**	**4931**		**0,032151**	**154**	**8**	**160**	**8**	**160**	**2**	**2**	**2**	**2**	**40**	**40**	**40**	**40**
Kerewan	**Rural**	**250**	**14762**		**0,09625**	**462**	**23**	**460**	**20**	**400**	**5**	**5**	**5**	**5**	**100**	**100**	**100**	**100**
	Lower Niumi	80	4756		0,03101	**149**	**7**	140										
	LN-Urban	24	1637		0,010673	**51**	**3**	60										
	LN-Rural	56	3119		0,020336	**98**	**5**	100										
	Upper Niumi	50	2602		0,016965	**81**	**4**	80										
	Jokadu	31	1527		0,009956	**48**	**2**	40										
	Lower Baddibu	31	3310		0,021582	**104**	**5**	100										
	LB-Urban	8	505		0,003293	**16**	**1**	20										
	LB-Rural	23	2805		0,018289	**88**	**4**	80										
	Central Baddibu	30	1620		0,010563	**51**	**3**	60										
	Upper Baddibu	94	5878		0,038325	**184**	**9**	180										
	UB-Urban	34	2789		0,018185	**87**	**4**	80										
	UB-Rural	60	3089		0,020141	**97**	**5**	100										
	TOTAL	**122**	**6620**		**0,043163**	**207**	**10**	**200**	**12**	**240**	**3**	**3**	**3**	**3**	**60**	**60**	**60**	**60**
	Urban	**11**	**615**		**0,00401**	**19**	**1**	**20**	**4**	**80**	**1**	**1**	**1**	**1**	**20**	**20**	**20**	**20**
Kuntaur	**Rural**	**111**	**6005**		**0,039153**	**188**	**9**	**180**	**8**	**160**	**2**	**2**	**2**	**2**	**40**	**40**	**40**	**40**
	Lower Saloum	24	1321		0,008613	**41**	**2**	40										
	LS-Urban	11	615		0,00401	**19**	**1**	20										
	LS-Rural	13	706		0,004603	**22**	**1**	20										
	Upper Saloum	21	1204		0,00785	**38**	**2**	40										
	Nianija	12	641		0,004179	**20**	**1**	20										
	Niani	34	1913		0,012473	**60**	**3**	60										
	Sami	31	1541		0,010048	**48**	**2**	40										

	TOTAL	**179**	**10582**		**0,068996**	**331**	**17**	**340**	**16**	**320**	**4**	**4**	**4**	**4**	**80**	**80**	**80**	**80**
	Urban	**26**	**1906**		**0,012427**	**60**	**3**	**60**	**4**	**80**	**1**	**1**	**1**	**1**	**20**	**20**	**20**	**20**
Janjanbureh	**Rural**	**153**	**8676**		**0,056569**	**272**	**14**	**280**	**12**	**240**	**3**	**3**	**3**	**3**	**60**	**60**	**60**	**60**
	Niamina Dankunku	11	630		0,004108	**20**	**1**	20										
	Niamina West	11	669		0,004362	**21**	**1**	20										
	Niamina East	36	2030		0,013236	**64**	**3**	60										
	Fulladu West	113	6705		0,043718	**210**	**10**	200										
	FW-Urban	18	1358		0,008854	**43**	**2**	40										
	FW-Rural	95	5347		0,034863	**167**	**8**	160										
	Janjangbureh	8	548		0,003573	**17**	**1**	20										
	Janjang-Urban	8	548		0,003573	**17**	**1**	20										
	Janjang-Rural	0	0		0	**0**	**0**	0										
	TOTAL	**246**	**13113**		**0,085499**	**410**	**21**	**420**	**20**	**400**	**5**	**5**	**5**	**5**	**100**	**100**	**100**	**100**
	Urban	**53**	**4050**		**0,026407**	**127**	**6**	**120**	**8**	**160**	**2**	**2**	**2**	**2**	**40**	**40**	**40**	**40**
	Rural	**193**	**9063**		**0,111178**	**284**	**14**	**280**	**12**	**240**	**3**	**3**	**3**	**3**	**60**	**60**	**60**	**60**
	Fulladu East	146	8188		0,053387	**256**	**13**	260										
	FE-Urban	53	4050		0,026407	**127**	**6**	120										
	FE-Rural	93	4138		0,02698	**130**	**6**	120										
	Kantora	32	1559		0,010165	**49**	**2**	40										
	Wuli	43	2256		0,014709	**71**	**4**	80										
Basse	**Sandu**	25	1110		0,007237	**35**	**2**	40										
	Both	**2464**	**153371**		**1**	**4800**	**240**	**4800**										
	Urban	**1012**	**71853**		**0,468491**	**2249**	**112**	**2240**	**120**	**2400**	**30**	**30**	**30**	**30**	**600**	**600**	**600**	**600**
All LGAs	**Rural**	**1452**	**81518**		**0,531509**	**2551**	**128**	**2560**	**120**	**2400**	**30**	**30**	**30**	**30**	**600**	**600**	**600**	**600**
Total	**TOTAL**	**2464**	**153371**		**1**	**4800**	**240**	**4800**	**240**	**4800**	**60**	**60**	**60**	**60**	**1200**	**1200**	**1200**	**1200**

Note 1: A district that constitute urban and rural has the urban and rural distribution, otherwise the district is considered rural.

Note 2: Natural groups to create quarterly EAPS are (1) Janjanbureh Urban + Kerewan Urban + Kuntour Urban; (2) Mansakonko Urban + Basse Urban