MZ:2014:08

The household budget and expenditure data collection module (IOF 2014/2015) within a continuous multipurpose survey system (INCAF)

Report from a fifth short term mission to the National Statistical Institute of Mozambique, Maputo Mozambique

22 November-14 December 2014

within the frame work of the

AGREEMENT ON CONSULTING ON INSTITUTIONAL CAPACITY BUILDING, ECONOMIC STATISTICS AND RELATED AREAS

between

INE and Scanstat

David J. Megill

INSTITUTO NACIONAL DE ESTATÍSTICA

Ref: Contract DARH/2008 /004 October, 2012

Address: *David J. Megill*
1504 Kenwood Ave.
Alexandria, VA 22302, US
E-Mail: *davidmegill@yahoo.com*
Telephone: *1-703-824-0292*

Table of Contents

1. INTRODUCTION AND TERMS OF REFERENCE

The *Instituto Nacional de Estatística* (INE) is conducting the *Inquérito sobre o Orçamento Familiar* (IOF) 2014/15, or Household Budget Survey (HBS), in a nationally-representative sample of 11,592 households in 1,236 sample census enumeration areas (EAs) over the 12-month period from August 2014 to July 2015. This survey was designed as a module of the *Inquérito Contínuo de Agregados Familiares* (INCAF), or Continuous Household Survey, which is a multipurpose household survey with a quarterly employment component, and the HBS, designed to obtain income and expenditure data for all four quarters to represent seasonality. One of the objectives of the IOF is to provide measures of poverty and other socioeconomic indicators, and to provide information on consumption needed for national accounts. The sample of households for IOF is treated as a panel, and each sample household is interviewed each quarter in a different period of the month.

The first quarter of data collection for IOF was conducted between 8 August and 7 November 2014. A small team of Scanstat short-term consultants began working with the INE staff on 24 November to review the first quarter of IOF data and to develop procedures for producing preliminary results from these data. Based on this review, the team also made recommendations for improving the data quality for the remaining three quarters of data collection.

The Terms of Reference for this first mission of the Sampling Consultant were stated as follows:

- **Objective:** During the first mission the Sampling Consultant will focus on assessing the current status of the INCAF/IOF sampling and estimation procedures following the first quarter of data collection, in order to identify any issues that need to be addressed. One objective of this mission will be to finalize the weighting procedures for the INCAF/IOF first quarter results, including the adjustment of weights to take into account non-response.
- **Activities:** The assessment will include a review of summary information from the listing of households in sample enumeration areas (EAs), and the distribution of the completed household interviews for the first quarter of INCAF/IOF. Sampling errors and design effects for key INCAF/IOF estimates such as the quarterly unemployment rate and average household expenditures will be tabulated and reviewed to assess the level of precision and the efficiency of the sample design. The methodology for maintaining the panel of households and the longitudinal analysis will be reviewed. The replacement of non-interview households will also be examined.
- **Expected outputs:** Based on the findings, the Sampling Consultant will make recommendations for improving different aspects of the methodology. Throughout this visit the Sampling Consultant will work closely with the INE Statisticians to provide on-the-job training. A half day seminar on the INCAF/IOF sampling and estimation methodology will be presented for the INE staff at the end of this visit.
- **Reporting:** The Sampling Consultant will submit a report on findings and recommendations on the INCAF/IOF sampling and estimation methodology.

One purpose of this mission report is to document the methodology for calculating the weights for IOF. Since the weighting procedures depend on the sampling methodology, this report also summarizes the IOF sample design.

The calculation of sampling errors for selected IOF indicators can only be accomplished once the IOF data edits are complete and the weighted survey data file for the first quarter is considered final. Therefore this activity will be followed up during the second mission of the Sampling Consultant around March 2015. The main activity of this mission, the calculation of the IOF weights, was only completed by the middle of the third week, given that it took time to compile the sampling frame information for all the sample clusters needed for the calculation of the weights.

The sampling consultant worked closely with Arão Balate, Director, *Direcção de Censos e Inquéritos*, Basílio Cubula, INE Sampling Statistician, and other INE staff in implementing the weighting procedures for the IOF 2014/15. He also collaborated with his Scanstat consultant colleagues, Lars Lundgren and Anne Abelseth. He appreciates their collaboration, and he would also like to thank Dr. João Loureiro, INE President, Manuel Gaspar, INE Vice-President, Antônio Adriano, Director Adjunto, Direcção de Censos e Inquéritos, Cristóvão Muahio, Chief, *Departamento de Metodologia e Amostragem* (DMA), and Eugénio Matavel, Programmer, for their support.

2. ACTIVITIES DURING THE MISSION

The data collection for the first quarter of IOF 2014/15 was completed around 7 November. Since computer-assisted personal interviewing (CAPI) was conducted using tablets that included initial edits in the field, the IOF data files should be available soon after the end of the first quarter. The occupation and industry information for employment was coded later in the office, but it would still be possible to produce some of the key unemployment tables soon after the data files are received from the field. In the case of the expenditure data, paper questionnaires were used, and the data were then supposed to be entered in the field. However, INE decided to enter all the expenditure data again in the office, which resulted in some delay in the availability of these data.

Once the full data set from the household interviews for the first quarter of IOF became available during the first week of this mission, Megill produced aggregated data at the sample cluster level to verify the concatenation of the data and determine whether data from any sample clusters were missing. Initially it was found that nine sample clusters (enumeration areas, or EAs) were missing, so the data processing staff checked on the status of these sample EAs. There were three EAs in Sofala Province that could not be enumerated because of security reasons. In the case of the other missing EAs, the INE staff contacted the field staff to obtain back-up copies of the data files. By the end of the second week of this mission the data files for all the EAs except for the three in Sofala were obtained and merged with the national data file for the first quarter of IOF. The weights were calculated based on the final IOF data set that will be used for the first quarter analysis.

The weighting procedures for the IOF 2014/15 depend on the sample design, so first it was necessary to review the sampling methodology used for the survey. This sample design is described below.

2.1. Summary of Sample Design for IOF 2014/15

A stratified multi-stage sample design was used for selecting the sample for the IOF 2014/15. The sampling frame was based on the Master Sample (*Amostra Mãe*) developed by INE from the 2007 Mozambique Census of Population and Housing (*Recenseamento Geral da População e Habitação*, RGPH 2007). The sampling methodology for the Master Sample is described in the report on"*Recomendações Metodológicas para o Desenho da Amostra Mãe em Base ao RGPH 2007 de Moçambique*" (David J. Megill and Carlos Creva Singano, November 2010). The sampling methodology involves three stages of selection. The primary sampling units (PSUs) selected at the first stage are based on the supervisory areas (*áreas de controlo*) defined for the RGPH 2007. Each supervisory area has about 3 to 5 enumeration areas (EAs), which are operational segments defined on maps for the census enumeration. At the first sampling stage the PSUs were selected systematically with probability proportional to size (PPS) within each stratum. The measure of size for each PSU was based on the number of households in the RGPH 2007 frame. At the second stage one EA was selected with PPS within each PSU. A listing of households was conducted within each sample EA, which is the frame for selecting a sample of households at the third sampling stage.

The PSUs in the Master Sample are stratified by province, urban and rural areas. The urban stratum of each province is further divided into substrata consisting of cities and other urban areas. A few very large cities are also divided into socioeconomic substrata, which were defined based on the RGPH 2007 socioeconomic data. In the case of the rural stratum of each province, the PSUs were classified by agro-ecological zone, which was used as a sorting variable to provide implicit stratification. The sampling frame was also sorted geographically to provide additional implicit stratification.

The main component of the first stage sample for IOF 2014/15 consisted of the 752 sample EAs selected for the INCAF 2012. The sample EAs for that survey had been previously selected from the Master Sample, stratified by province, urban and rural stratum. In order to improve the level of precision for the provincial-level estimates of key indicators, the total number of sample EAs for IOF was increased to 1,236, so 484 additional sample EAs were selected from the Master Sample, using the same systematic PPS selection procedures within each stratum. One reason that the INCAF sample EAs were used for the IOF is that the listing of households for that survey conducted in 2012 could be used again for the IOF in order to reduce the cost of the fieldwork. In this case it was only necessary to conduct a new listing in the additional sample of 484 sample EAs. The overlap in the sample EAs between the INCAF and IOF would also provide a greater correlation between the two samples, which should improve the level of precision for the estimates of trends (differences) over time for the unemployment rate labor force indicators. At the last sampling stage 11 sample households were selected from the listing for each urban EA, and 8 households were selected for each rural EA. A reserve of sample households was also selected in each EA for replacing any sample household that could not be interviewed for any reason. Table 1 shows the distribution of sample EAs and households selected for the IOF 2014/15 by province, urban and rural stratum.

Table 1. Distribution of Sample EAs and Sample Households for IOF 2014/15, by Province and Urban/Rural Stratum

Province	Urban		Rural		Total	
	No. of EAs	No. of Households	No. of EAs	No. of Households	No. of EAs	No. of Households
Niassa	32	352	64	512	96	864
Cabo Delgado	44	484	60	480	104	964
Nampula	60	660	104	832	164	1,492
Zambézia	52	572	124	992	176	1,564
Tete	40	440	68	544	108	984
Manica	40	440	56	448	96	888
Sofala	60	660	44	352	104	1,012
Inhambane	40	440	52	416	92	856
Gaza	40	440	48	384	88	824
Maputo Província	60	660	48	384	108	1,044
Maputo Cidade	100	1,100	-	-	100	1,100
Total	568	6,248	668	5,344	1,236	11,592

One additional step used in the sampling implementation for IOF was that small EAs (for example, with less than 50 households) were combined with adjacent EAs in the census frame to form a larger cluster that was listed. Some large EAs (for example, with more than 200 households) were subdivided into smaller segments, and one segment was randomly selected for the listing. Although the original listing form was designed to include this information for combined and sub-divided EAs, unfortunately the tablet system used for capturing the data in the field did not keep this information. This affects the calculation of the weights, as described later in the section on weighting procedures.

Although the final sample of EAs for the IOF 2014/15 was selected in different phases for the INCAF and the additional IOF sample, the same sampling procedures were used for each phase. That is, at the first stage the PSUs were selected systematically with PPS within each stratum, and one EA was selected within each PSU with PPS. Therefore the estimation procedures for calculating the weights and the sampling errors will be based on the assumption that all the IOF sample EAs within each stratum were selected at the same time using these procedures.

In most EAs all the original sample households that could not be interviewed were replaced, in which case there were exactly 11 completed household interviews for sample urban EAs and 8 completed household interviews for sample rural EAs. However, in some EAs there were more non-interview households than replacements, so that less sample households had completed interviews. In a few cases the interviewers completed 12 household interviews in an urban sample EA. This is not a problem, since the number of completed interviews is taken into account in the weighting procedures, as described later in this section.

2.2. Weighting Procedures for INCAF/IOF 2014/15

In order for the sample estimates from the IOF 2014/15 to be representative of the population, it is necessary to multiply the data by a sampling weight, or expansion factor. The basic weight for each sample household would be equal to the inverse of its probability of selection (calculated by multiplying the probabilities at each sampling stage). The sampling probabilities at each stage of selection are maintained in an Excel spreadsheet with information from the sampling frame for each sample EA so that the corresponding overall probabilities and corresponding weights can be calculated. This section first describes the weights based on the actual probabilities of selection, followed by weight adjustment procedures that were needed to compensate for deficiencies in the sampling information.

Based on the sampling procedures for the Master Sample and the IOF 2014/15, the overall probability of selection for the IOF sample households can be expressed as follows:

$$p_{hij} = \frac{n_h \times M_{hi}}{M_h} \times \frac{M_{hij}}{M_{hi}} \times \frac{n'_h}{n_h} \times p_{Shij} \times \frac{m_{hij}}{M'_{hij}} = \frac{n'_h \times M_{hij}}{M_h} \times p_{Shi} \times \frac{m_{hij}}{M'_{hij}},$$

where:

p_{hij} = probability of selection for the sample households in the j-th sample EA of the i-th sample PSU in stratum h

n_h = number of sample EAs selected in stratum h for the Master Sample

M_h = total number of households in the RGPH 2007 frame for stratum h

M_{hi} = total number of households in the RGPH 2007 frame for the i-th sample PSU in stratum h

M_{hij} = total number of households in the RGPH 2007 frame for the j-th sample EA of the i-th sample PSU in stratum h

n'_h = number of EAs selected in stratum h for the IOF 2014/15

p_{Shij} = probability of selection for the selected segment in large sample EA that is sub-divided; this probability is equal to 1 for all EAs that are not segmented

m_{hij} = number of sample households selected in the j-th sample EA of the i-th sample PSU in stratum h

M'_{hij} = total number of households listed in the j-th sample EA of the i-th sample PSU in stratum h

The different components of this probability of selection correspond to the individual sampling stages. The probability of selecting a segment in a large EA (p_{Shi}) depends on the type of selection procedure that is used. If one segment is selected randomly with equal probability, this probability would be calculated as follows:

$$p_{Shij} = \frac{1}{S_{hij}},$$

where:

S_{hij} = total number of segments in the j-th large sample EA of the i-th sample PSU in stratum h

In the case of a small EA that was combined with another EA in the same PSU for the listing, the measure of size M_{hij} was based on the sum of the number of households in the Census frame for the combined EAs.

The basic sampling weight, or expansion factor, is calculated as the inverse of this probability of selection. Based on the previous expression for the probability, the weight can be simplified as follows:

$$W_{hij} = \frac{M_h \times M'_{hij}}{n'_h \times p_{Shij} \times M_{hij} \times m_{hij}},$$

where:

W_{hij} = basic weight for the sample households in the j-th sample EA of the i-th sample PSU in stratum h

During the first quarter of data collection for the IOF 2014/15, three of the 1,236 sample EAs could not be enumerated because of security problems. In this case it is necessary to adjust the weights for the corresponding strata. The weights are also adjusted to take into account any non-interviews that could not be replaced. The weight adjusted for missing sample EAs and sample households that could not be replaced can be expressed as follows:

$$W'_{hij} = \frac{M_h \times M'_{hij}}{n'_h \times p_{Shij} \times M_{hij} \times m_{hij}} \times \frac{n'_h}{n''_h} \times \frac{m_{hij}}{m'_{hij}} = \frac{M_h \times M'_{hij}}{n''_h \times p_{Shij} \times M_{hij} \times m'_{hij}},$$

where:

W'_{hij} = adjusted basic weight for the sample households in the j-th sample EA of the i-th sample PSU in stratum h

n''_h = number of EAs enumerated in stratum h for the IOF 2014/15

m'_{hij} = number of sample households with completed interviews in the j-th sample EA of the i-th sample PSU in stratum h, including replacement households

Although an attempt was made to obtain all the information needed to calculate this adjusted basic weight based on the probabilities of selection, it was not possible to obtain the information on the EAs that were combined or sub-divided. The spreadsheet

with the information from the frame for each sample EA did not include the number of households for each sample EA in the RGPH 2007 (M_{hij}), so it was necessary to merge this information from a database with all the Census EAs by matching the geographic codes. All the EAs were matched to the Census database except for about 13 EAs. However, these measures of size did not take into account the small EAs that were combined with other EAs in the same PSU. Another problem is that it was not possible to obtain information on the subdivided EAs in order to calculate the probability p_{Shij}. For this reason it was necessary to calculate approximate weights based on the available information. The approximate probabilities were based on using the number of households from the listing for each EA as an approximate measure of size. In this case the approximate adjusted weights for the IOF 2014/15 sample households were calculated as follows:

$$W''_{hij} = \frac{M_h \times M'_{hij}}{n''_h \times M'_{hij} \times m'_{hij}} = \frac{M_h}{n''_h \times m'_{hij}},$$

where:

W''_{hij} = approximate adjusted basic weight for the sample households in the j-th sample EA of the i-th sample PSU in stratum h

It can be seen in this formula that the final adjusted weight is similar for all sample households within each stratum, varying only by the number of completed household interviews in each EA. Given the procedure for replacing non-interview sample households, the number of completed household interviews is exactly 11 for most sample urban EAs, and 8 for most sample rural EAs.

The effect of this approximate weighting procedure is to adjust the weights to the distribution of the frame based on the RGPH 2007. Therefore this weighting procedure does not take into account any differential growth rate of the urban and rural strata by province following the RGPH 2007. However, in the next step these approximate weights are adjusted based on population projections by province, urban and rural stratum, as described later in this weighting section of the report. As long as these population projections are reasonably accurate, the weighted estimates from the IOF 2014/15 will reflect the actual distribution of the population by province, urban and rural stratum. Therefore the final adjusted weights will reduce some of the bias in the distribution of the weighted population by stratum. Another reason to have confidence in the final adjusted weights is that the probability of selection of the sample PSUs at the first sampling stage is known, and the last stage probability of selection of the households from the listing is known. Although we do not know the exact probability of selection of the EA within the sample PSU for the cases where the EA was combined or sub-divided, we know that the final cluster was randomly selected with PPS (based on the EA) within the sample PSU. In this case we use the number of households listed in the cluster as the approximate measure of size.

It should also be pointed out that apparently the weights for the first quarter of INCAF 2012 suffered from a similar problem with the lack of information for sample EAs that were combined or sub-divided. Since the basic weights were not adjusted to take this problem into account, the INCAF 2012 weights were more variable. Using the basic design weights (prior to the adjustment based on population projections), the weighted total population from the INCAF 2012 data was only about 17.1 million, considerably

lower than the corresponding population from the RGPH 2007. This illustrates the problem with a potential under-count in the listing, as well as the lack of information for EAs that were sub-divided for the listing. However, the INCAF 2012 weights were also adjusted based on the population projections, so this will improve the comparability of the weighted estimates from the IOF 2014/15 with those from the INCAF 2012.

As mentioned above, the adjusted basic weights for the IOF sample households will provide a weighted distribution by province, urban and rural stratum that is consistent with the RGPH 2007. In order to reflect the growth in the population by stratum between 2007 and the time of the IOF 2014/15 data collection, the preliminary weights were adjusted based on population projections. The INE demographers had used a demographic analysis model with the data from the RGPH 1997 and 2007 and estimates of different parameters from the 2013 Demographic and Health Survey (DHS) and other sources to produce tables on population projections for each province, urban and rural stratum, by individual year up to 2040.

The weight adjustment factor based on the projected total population by province, urban and rural stratum can be expressed as follows:

$$A_h = \frac{P_h}{\sum_{i \in h} \sum_j \sum_k W''_{hij} \times p_{hijk}},$$

where:

A_h = adjustment factor for the weights of the IOF sample households in stratum h (province, urban/rural)

P_h = projected total population for stratum h for the mid-point of the data collection period for the first quarter of IOF, based on demographic analysis

W''_{hij} = adjusted basic design weight for the sample households in the j-th sample EA of the i-th sample PSU in stratum h

p_{hijk} = number of persons in the k-th sample household in the j-th sample EA of the i-th sample PSU in stratum h

The denominator of the adjustment factor A_h is the estimated weighted total population in stratum h from the IOF data using the preliminary adjusted basic design weights. The preliminary weights for all the sample households within a stratum were multiplied by the corresponding adjustment factor for the stratum to obtain the final adjusted weights, as follows:

$$W_{Ahij} = W''_{hij} \times A_h,$$

where:

W_{Ahij} = final adjusted weight for the sample households in the j-th sample EA of the i-th sample PSU in stratum h

After the adjustment factors were applied to the weights of each stratum, the final weighted survey estimates of total population by stratum were consistent with the

corresponding population projections. Of course the accuracy of the estimates of total population based on the adjusted weights depends on the quality of the population projections by stratum.

The population projections which INE generated for each year reflect the mid-point of the year, or 1 July. For the adjustment of the weights, it is ideal to use the population projections for the mid-point of the data collection period for the survey. In the case of the first quarter of IOF, the data collection was conducted between 8 August and 7 November, so the mid-point was estimated as 23 September 2014. Using the population projections by province, urban and rural stratum for 1 July 2014 and 1 July 2015, an interpolation based on exponential growth was used to estimate the population for 23 September 2014, using the following formula:

$$P_h = P_{14h} \times e^{\ln\left[\left(\frac{P_{15h}}{P_{14h}}\right)\times\left(\frac{t_{IOF}-t_{14}}{t_{15}-t_{14}}\right)\right]}$$

where:

P_h = projected total population for stratum h on 23 September 2014

P_{14h} = population projection for stratum h on 1 July 2014

P_{15h} = population projection for stratum h on 1 July 2015

$t_{IOF} - t_{14}$ = number of days between 1 July 2014 and 23 September 2014 (that is, 84 days)

$t_{15} - t_{14}$ = number of days between 1 July 2014 and 1 July 2015 (that is, 365 days)

Table 2 presents the INE population projections by province, urban and rural stratum, for 1 July 2014 and 1 July 2015, and the corresponding interpolated population estimates for 23 September 2014.

Table 2. Mozambique Population Projections by Province, Urban and Rural Stratum for 2014 and 2015, and Interpolated Population for Mid-Point of IOF Data Collection Period for First Quarter

Province and Stratum	2014 1 July	2015 1 July	IOF - 2014 23 Sept.
Niassa Urban	372,176	388,202	375,805
Niassa Rural	1,221,307	1,268,704	1,232,055
Cabo Delgado Urban	444,864	463,038	448,982
Cabo Delgado Rural	1,417,221	1,430,118	1,420,179
Nampula Urban	1,549,414	1,615,298	1,564,334
Nampula Rural	3,338,425	3,393,495	3,351,019
Zambézia Urban	958,355	1,008,281	969,621
Zambézia Rural	3,724,080	3,794,084	3,740,075
Tete Urban	327,752	341,385	330,840
Tete Rural	2,090,829	2,176,059	2,110,143
Manica Urban	447,430	460,597	450,426

Manica Rural	1,418,871	1,472,925	1,431,132
Sofala Urban	725,458	737,503	728,212
Sofala Rural	1,273,851	1,311,173	1,282,345
Inhambane Urban	349,499	359,253	351,720
Inhambane Rural	1,125,819	1,140,226	1,129,118
Gaza Urban	358,546	365,350	360,101
Gaza Rural	1,033,526	1,051,460	1,037,626
Maputo Province Urban	1,145,642	1,200,866	1,158,122
Maputo Province Rural	492,989	508,192	496,447
Maputo City	1,225,868	1,241,702	1,229,494
Mozambique	25,041,922	25,727,911	25,198,155

Table 3 shows the population projections for the mid-point of the IOF data collection period for the first quarter, the IOF weighted estimates of total population by stratum based on the adjusted design weights, and the corresponding weight adjustment factor for the sample household weights in each stratum. It can be seen in Table 3 that the weight adjustment factors vary from 0.8885 for Cabo Delgado Rural to 1.4808 for Maputo Province Urban.

Table 3. Mozambique Population Projections and IOF Weighted Estimates of Total Population by Province, Urban and Rural Stratum, and Corresponding Weight Adjustment Factors

Province and Stratum	Projected Population 23-09-14	Weighted Population IOF, First Quarter	Weight Adjustment Factor
Niassa Urban	375,805	274,659	1.3683
Niassa Rural	1,232,055	1,157,637	1.0643
Cabo Delgado Urban	448,982	386,203	1.1626
Cabo Delgado Rural	1,420,179	1,598,312	0.8885
Nampula Urban	1,564,334	1,206,509	1.2966
Nampula Rural	3,351,019	3,298,091	1.0160
Zambézia Urban	969,621	697,511	1.3901
Zambézia Rural	3,740,075	3,422,909	1.0927
Tete Urban	330,840	223,590	1.4797
Tete Rural	2,110,143	1,647,588	1.2807
Manica Urban	450,426	373,314	1.2066
Manica Rural	1,431,132	1,179,326	1.2135
Sofala Urban	728,212	771,575	0.9438
Sofala Rural	1,282,345	1,242,540	1.0320
Inhambane Urban	351,720	289,589	1.2145
Inhambane Rural	1,129,118	1,009,156	1.1189
Gaza Urban	360,101	297,715	1.2095
Gaza Rural	1,037,626	942,200	1.1013
Maputo Province Urban	1,158,122	782,079	1.4808

Maputo Province Rural	496,447	396,739	1.2513
Maputo City	1,229,494	1,138,478	1.0799

Megill worked closely with Basílio Cubula on the calculation of weights for the first quarter of data for IOF 2014/15 using these procedures. First Cubula compiled a spreadsheet with the information from the frame for the 1,236 sample EAs selected for IOF 2014/15, including the number of households listed in each sample EA. Megill used the IOF household data file for the first quarter to tabulate the number of sample households with completed interviews in each EA. He also identified sample EAs that were missing, as described previously in this report. He consulted with various INE staff to try to obtain information on the EAs that were combined or sub-divided. Since this information was not available, Megill developed the approximate weighting procedures described above. The final weights were produced in the middle of the third week of this mission.

2.3. Procedures for Calculating Sampling Errors

In the publication of the results for the IOF 2014/15 it is important to include a statement on the accuracy of the survey data. In addition to presenting tables with calculated sampling errors and confidence intervals for the most important survey estimates, the different sources of nonsampling error should be described.

The most common estimates to be calculated from the data for IOF will be in the form of totals and ratios. The survey estimate of a total can be expressed as follows:

$$\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} W_{Ahij}\, y_{hik} \ ,$$

where:

L = number of strata (province, urban/rural) in the domain

y_{hik} = value of variable y for the k-th sample household in the i-th sample PSU in stratum h

The survey estimate of a ratio is defined as follows:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \ ,$$

where $\hat{Y}$ and $\hat{X}$ are estimates of totals for variables y and x, respectively, calculated as specified previously.

In the case of a stratified multi-stage sample design, means and proportions are special types of ratios. In the case of the mean, the variable x, in the denominator of the ratio, is defined to equal 1 for each element so that the denominator is the sum of the weights. For a proportion, the variable x in the denominator is also defined to equal 1 for all elements; the variable y in the numerator is binomial and is defined to equal either 0 or

1, depending on the absence or presence, respectively, of a specified characteristic for the element.

The standard error, or square root of the variance, is used to measure the sampling error, although it may also include a small variable part of the nonsampling error. The variance estimator should take into account the different aspects of the sample design, such as the stratification and clustering. Programs available for calculating the variances for survey data from stratified multi-stage sample designs, such as IOF, include Stata and the Complex Samples module of SPSS. Both of these software packages use a linearized Taylor series variance estimator.

The Complex Samples module of SPSS can be used for calculating the sampling errors for survey estimates of totals, means, proportions and other types of ratios. For each estimate, the SPSS tables show the standard error, coefficient of variation (CV), 95 percent confidence interval, the design effect (DEFF) and the number of observations. The design effect is defined as the ratio of the variance of an estimate from a complex (stratified, multi-stage) sample to the variance of a simple random sample of the same size. It is a relative measure of the sampling efficiency. Most of the design effects are greater than 1 given the clustering effects in the sample design.

The variance estimator for a total used by SPSS Complex Samples and Stata can be expressed as follows:

Variance Estimator of a Total

$$V(\hat{Y}) = \sum_{h=1}^{L} \left[\frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \right],$$

where:

$$\hat{Y}_{hi} = \sum_{k=1}^{m_{hi}} W_{Ahi}\, y_{hik}$$

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}$$

The variance estimator of a ratio used by these statistical software packages can be expressed as follows:

Variance Estimator of a Ratio

$$V(\hat{R}) = \frac{1}{\hat{X}^2} \left[V(\hat{Y}) + \hat{R}^2\, V(\hat{X}) - 2\, \hat{R}\, COV(\hat{X}, \hat{Y}) \right],$$

where:

$$COV(\hat{X}, \hat{Y}) = \sum_{h=1}^{L} \left[\frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{X}_{hi} - \frac{\hat{X}_h}{n_h} \right) \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right) \right]$$

$V(\hat{Y})$ and $V(\hat{X})$ are calculated according to the formula for the variance of a total.

Since the final weighted data set for the first quarter of IOF 2014/15 were not available during this mission, it was not possible to tabulate the sampling errors for selected indicators. However, this will be followed up during the second mission, with the tabulation of sampling errors for key estimates from the first and second quarters of IOF data.

2.4. Capacity Building

As mentioned previously in this report, Megill worked closely with Basílio Cubula and other INE counterparts to provide on-the-job training in developing the weighting procedures for IOF 2014/15. On the last day of this mission Megill gave a half-day training session to the INE staff on the methodology used for calculating the weights for the IOF first quarter. There should be more time available for such training in the next mission, when IOF panel data will be available for the second quarter. Although the final weighted data file for the IOF first quarter was not available during the first mission for the calculation of sampling errors, this will also be followed up in the next mission.

3. FINDINGS AND RECOMMENDATIONS

The main findings during this mission are discussed in the previous section, and the highlights are summarized here. Although the data collection for the first quarter of IOF 2014/15 was successful and the survey data appear to have reasonable quality, there were some important lessons learned. Sampling information related to combining small sample EAs and sub-dividing large sample EAs prior to the listing operation appears to have been lost. This information would be needed to calculate the exact probabilities and corresponding weights for the IOF sample households. Since this information was not available, it was necessary to calculate approximate weights which were then adjusted based on the population projections by province, urban and rural stratum, as described in this report. Since the IOF is based on a panel of households that are enumerated each quarter, it will be necessary to use the same approximate weighting procedures for all quarters. However, it is recommended that for future surveys the information from each sampling stage should be carefully recorded and maintained for the calculation of the probabilities of selection and corresponding design weights.

Conceptually, a complete listing of households in the sample EAs reflects the overall average growth in the number of households across all the sample EAs, so the weighted estimates of the total population would also show a corresponding increase. Therefore the design weights depend on the updating of the sampling frame based on the listing, and if the listing for some sample EAs is not complete, this will lead to a downward bias in the weighted population estimates from the survey data. It is important to note that it is ideal to rely on a high-quality updated listing of households in each sample EA and weights based on the sampling probabilities to reflect the differential population growth by province, urban and rural stratum. Although it is too late to correct this for IOF 2014/15, this is an important lesson learned for

improving future surveys. The population projections are based on the growth rates between the last two censuses and general demographic assumptions, so they do not always accurately reflect the actual differential growth rates by urban and rural stratum within each province. For this reason it is not good to always rely on the population projections for adjusting the probability-based weights.

The second mission is tentatively scheduled for March 2015, following the data collection for the second quarter of IOF 2014/15. It is recommended that the IOF data files for the first and second quarters of IOF be ready prior to that mission, so that the second quarter weights can be calculated during the first week, and more time will be available for the calculation of sampling errors and other aspects of the analysis.

APPENDIX 1. Persons Contacted

Instituto Nacional de Estatística (INE)

Dr. João Loureiro, INE President
Manuel Gaspar, INE Vice-President
Arão Balate, Director, *Direcção de Censos e Inquéritos*
Antônio Adriano, Deputy Director, *Direcção de Censos e Inquéritos*
Cristóvão Muahio, Chief, *Departamento de Metodologia e Amostragem*
Basílio Cubula, Sampling Statistician
Eugênio Matavel, Programmer, INE
Carlos Creva, former INE Sampling Statistician
Tomás Bernardo

Scanstat
Lars Lundgren, Household Surveys Consultant
Anne Abelseth, IT Consultant
Lars Carlsson, Resident Advisor