

Survey Sample for the Tunisia Household and Youth Survey in Urban Areas

Mario Navarrete and Abla Safir
September 2012

BACKGROUND

This Technical support and ESW is to collect timely data to produce a diagnostic of socio-economic inequality and vulnerability of Tunisians, especially youth, in the urban areas of the country.

This ESW involves a survey of 4,224 households, representative of 7 regions, the 6 administrative regions of Tunisia and Tunis as a separate region. To ensure accurate responses, individual interviews of household members are held, focusing on:

- Spatial patterns of unemployment, economic activity, migration, access to basic services and shocks.
- Job search methods and the effectiveness of active labor market programs.
- Attitudes of young people to government and the economy.
- Time use and activities of young people.

In the wake of the historical changes that have marked Tunisia since the revolution, this survey was born out of the initiative of a group called Projet Citoyen, comprising 70 students and teachers from various universities in Tunisia. This project builds the capacity of the entire project team so that the approach can continue to be applied in future beyond the scope of the current exercise.

For this study, Sistemas Integrales was in charge of drawing the sample, in collaboration with the Tunisian Institut National de la Statistique (INS), which provided the sampling frame.

The final sample was drawn during a mission of Mario Navarrete in February 2012. Subsequently, because maps were missing for two enumeration areas, they had to be replaced. This note documents the sampling strategy, including for the replacement of EAs with missing map. It also provides the probability of selection for each EA in the sample.

The sample must be representative of urban areas, both nationally and regionally for the seven regions in which the country has been divided.

Implemented Activities

The main activities implemented by Sistemas Integrales in the course of two missions (one in November 2011 by Beatriz Godoy; one in February 2012 by Mario Navarrete) regarding the determination of the sample were as follows:

- Review of the sampling provided by the INS;
- Discussion and clarification of the regional allocation;
- Review of the samples obtained within each region;
- Calculation of sample weights

The sampling for the youth survey is shown below.

SAMPLING

The HYSU-2012 survey has a national coverage of urban areas, and is regionally representative. The total sample size of 4224 households was established to allow good inference at the regional level, within the budget constraints of the project.

Taking into account the foregoing, 7 explicit survey strata were selected:

- District of Tunis
- North East
- North West
- Central East
- Central West
- South East
- South West

The target population consists of urban households from the 7 regions of the country and their members, in particular those in the 15 to 29 age group. ¹

The sample is selected through an EAaw at two levels:

- In the first stage, 352 Enumeration areas (EAs – in French Districts de recensement, EAs) have been drawn with Probability Proportional to Size, using the General Census of Population and Housing in 2004 as sampling frame, and the number of households as the measure of size. ²
- In the second stage, in each EA selected in the first stage, 12 households will be drawn with equal probabilities. In each EA, the survey team assigned to this EA starts with listing, a comprehensive enumeration of the households in the EA.

¹ Naturally, as it is a household survey, it excludes collective dwellings that do not fall within the definition of a household.

² Some urban areas, in particular those on the outskirts of urban centers, have been updated in 2009. That is to say that new maps and a new household listing have been conducted in these areas in 2009.

The number of EAs to consider within each stratum was calculated using as main criterion the need to obtain estimates of acceptable quality in the key areas of estimation for the investigation. Two scenarios have been discussed. The first scenario follows a distribution proportional to the size of each region. The second scenario simply splits the total sample in an equal way between the 7 regions. For the smallest regions, the first scenario results in a sample that is too small to allow correct inference. On the other hand, it seems of interest to allow a larger sample in larger regions. Therefore, the sampling adopted follows the rule of Markwardt, which allows for a regionally representative sample. We compare the three scenarios in the following table.

Table 1 – Sampling frame and comparison of sampling rules

REGIONS	TOTAL NUMBER OF EAs	TOTAL NUMBER OF URBAN HOUSEHOLDS	DISTRIBUTION OF HOUSEHOLDS (%)	PROPORTIONAL DISTRIBUTION OF 352 EAs	EQUITABLE DISTRIBUTION OF 352 EAs	MARKWARDT DISTRIBUTION OF 352 EAs
District of Tunis	6884	594,618	34.1	120	50.3	85
North East	2666	228,885	13.1	46	50.3	48
North West	1426	118,986	6.8	24	50.3	37
Central East	5248	447,653	25.7	90	50.3	70
Central West	1318	105,464	6.0	21	50.3	36
South East	1730	162,167	9.3	33	50.3	42
South West	1074	87,002	5.0	18	50.2	34
Total	20346	1,744,775	100.0	352	352	352

If we consider that 12 households will be selected in the EA, the proportional distribution does not allow us to have a sufficient number of households to make an inference of good quality in the North West, Central West and South West regions (total sample of 288, 252 and 216 households respectively). In the case of the equitable distribution, each region has 14% of the EAs of the total sample, and therefore does not take into account that, for instance, the District of Tunis has 34.1 % of the urban households in the country. Therefore the average between the proportional distribution and the equitable distribution seems to be appropriate, and takes into account the need to have a sample size to make the inference of good quality for each region, with a certain proportionality that allows for a lesser total error than if using equitable distribution. The distribution of households to be surveyed is shown in table 2.

Table 2 - Distribution of households in the sample

REGION	DISTRIBUTION OF HOUSEHOLDS TO BE SURVEYED
DISTRICT OF TUNIS	1020
North East	576
North West	444
Central East	840
Central West	432
South East	504
South West	408
Total	4224

Once we established the allocation of the enumeration areas (EAs) by region and once we built the sampling frame for this survey with all enumeration areas of urban areas, we selected the 352 EAs of our sample. The selection of EAs within each region was made with a probability proportional to size (PPS), using the number of households as a measure of size. A detailed list of the selected EAs is given in Appendix 1.

The following table shows the distribution of the sample of the Enumeration areas (EAs) by governorate in the interior of every region. The actual distribution of the enumeration areas (EAs) chosen is equal to the theoretical distribution which was made by distributing the Enumeration areas (EAs) in relation to the percentage of the number of households in each governorate to the number of households in the region. Finding the same distribution is a confirmation, since the choice was made with PPS.

Table 3 - Distribution of enumeration areas by governorate

Governorate / Region	SAMPLE			
	HOUSEHOLD	% household in region	Theoretical districts distribution	Distribution
DISTRICT DE TUNIS	582.646	100,0	85	85
Tunis	266.305	45,7	39	39
Ariana	120.321	20,7	18	17
Ben Arous	131.068	22,5	19	19
Manouba	64.952	11,1	9	10
Nord Est	224.873	100,0	48	48
Nabeul	116.275	51,7	25	25
Zaghuan	15.162	6,7	3	3
Bizerte	93.436	41,6	20	20
Nord ouest	116.256	100,0	37	37
Beja	31.422	27,0	10	10
Jendouba	28.880	24,8	9	9
Le Kef	35.535	30,6	11	11
Siliana	20.419	17,6	6	7
Centre est	430.964	100,0	70	70
Sousse	122.372	28,4	20	20
Monastir	116.985	27,1	19	19
Mahdia	41.227	9,6	7	6
Sfax	150.380	34,9	24	25
centre ouest	102.736	100,0	36	36
Kairouan	43.760	42,6	15	15
Kasserine	36.031	35,1	13	13
Sidi Bouzide	22.945	22,3	8	8
sud est	153.901	100,0	42	42
Gabes	56.997	37,0	16	15
Mednine	78.632	51,1	21	22
Tataouine	18.272	11,9	5	5
sud ouest	88.729	100,0	34	34
Gafsa	52.341	59,0	20	20
Tozeur	16.132	18,2	6	6
Kebili	20.256	22,8	8	8
Total	1.700.105		352	352

REPLACEMENT OF EAs WITH MISSING MAPS

After the sample was drawn, new problems arose with obtaining the maps of enumeration areas. As maps were missing at the INS, the INS “replaced” enumeration areas with neighboring enumeration areas. This is problematic because it means that the sample is no longer a random sample. Solving this problem has involved another phase of intense work with the INS to recover the location of most EAs with missing maps except 2. For 2 EAs, 2 replacement EAs have been drawn with a random method.

The general idea was to take an area larger than the EA and to which the EA belonged, divide this area into EAs, and then select randomly an EA within these new EAs. This is because it was possible to recover the map for the larger area while it was not possible to recover the map of the specific EAs in our sample.

The EAs with missing maps are EA 1304502 and EA 1310502. Each of those EAs is in a different CR. Each of those EAs belongs to sections that have been updated in 2009. Lastly, CR 1310 includes an EA that is in our sample, in a section not updated.

Each of CRs 1304 and 1310 includes both a part updated in 2009 and a part that was not updated. This made our life easier since we only had to proceed to a new division of the CR into EAs in the areas that had been updated in 2009. To have the complete list of EAs in the CR before proceeding to selecting an EA randomly, we simply needed to append the list of EAs not updated in 2009.

With the North-East regional direction of the INS, which includes Tunis, the World Bank team (Zied Ouelhazi and Abba Safir) took the *cahiers de sections* that have been updated in CRs 1304 and 1310 and divided the CRs into new EAs, new in the sense that these EAs do not correspond to the EAs that the INS has. While in the end the same households will be included overall, their grouping into the EAs as done by the World Bank team and the INS in this case may differ from the grouping done by INS in its sampling frame.

A *section* is delimited by streets or boulevards and each *section* includes *ilots*. An EA can include *ilots* that belong to different sections. However, an *ilot* can never belong to more than one EA. We also checked that there is no overlapping between the areas updated in 2009 and the areas that have not been updated. This is a concern for the whole sample, to make sure that there is no double counting of households.

Hence, the World Bank with the INS, grouped *ilots* in order to build EAs. This was done keeping in mind that each EA includes between 50 and 100 households, the rule usually followed by the INS.

The team had the *cahiers de section* in front of it and Zied completed an Excel spreadsheet in which he entered all the *ilots* of the areas updated in 2009, assigning each *ilot* to an EA and numbering the EAs hence created as well counting the number of households in them.

This excel sheet was supplemented with a list of EAs not updated. The complete sheets for CR1310 and CR1304 were then sent to Mario Navarrete, who selected randomly, following a PPS

rule, an EA in each CR. In CR1310, he excluded from the sampling frame the EA that was already part of the sample.

Selection probabilities and sampling weights

With the sampling strategy chosen, the probability p_{hij} to draw the household hij in the Census District hi of the stratum h is

$$p_{hij} = \frac{k_h n_{hi}}{\sum_{\alpha} n_{h\alpha}} \frac{12}{n'_{hi}}$$

Where

k_h is the number of Enumeration areas (EAs) Drawn in stratum h ,

n_{hi} is the number of households recorded by the census of 2004 in the Enumeration areas (EAs), and

n'_{hi} is the number of households in the Census District (EA) during the listing operation.

The weight of w_{hij} survey of household hij is simply the opposite of the probability of selection p_{hij} . If n_{hi} and n'_{hi} were identical in all the Enumeration areas (EAs), the formula above would simplify to a constant for each stratum. In practice, the two numbers will rarely be identical, but often close enough. The poll strategy chosen will thus provide samples that are self-ponderated in each stratum.

For EAs 1304502 and 1310502, the probability selection has been calculated as the product of:

- the probability of selection of the CR (which contains the EA) in the stratum;
- and the probability of selection of the EA in the CR.

We assume that the selection of these units implied the selection of its higher level of aggregation - the CR -, which were divided into EAs again, to select one EA in each of these CR with probability proportional to size.

