**METHODOLOGICAL NOTE**

**HOUSEHOLD BUDGET SURVEY 2021**

## 1. Sample Design

### 1.1 Type of sample design and sampling units

The two-stage area sampling was applied for the Household Budget Survey 2021. The sample of private households was selected in two stages. The primary units are the areas (one or more unified building blocks) and the ultimate sampling units selected in each sampling area are the households.

### 1.2 Stratification and sub-stratification criteria

There are two levels of area stratification in the sampling design. The first level is the geographical stratification based on the partition of the total country area into thirteen (13) Regions corresponding to the European NUTS 2 level. The two former major city agglomerations of Greater Athens and Greater Thessaloniki constitute separate major geographical strata.

The second level of stratification entails grouping municipal and local communities within each NUTS 2 Region by degree of urbanization, i.e., according to their population size. The scaling of urbanization was finally designed in three groups:

| | |
|---|---|
| Urban | Municipal or Local Communities with 10.000 inhabitants or more |
| Semi-urban | Municipal or Local Communities with 2.000 to 9.999 inhabitants |
| Rural | Local Communities up to 1.999 inhabitants |

The number of the final strata in the thirteen (13) Regions of the Country (except Greater Athens Area & Greater Thessaloniki Area) is 39. These were derived from the crossing of Region by the degree of urbanization. Additionally, the Greater Athens Area was divided into 31 strata taking into consideration socio-economic criteria. Similarly, the Greater Thessaloniki Area was divided into 9 strata. Thus, the total number of strata of the survey is 79. The two Major City Agglomerations account for about 37% of total population.

### 1.3 Sample size and allocation criteria

The initial sample size was 6,450 households (sampling fraction $\frac{1}{\lambda} \cong 0.14\%$). This fraction was the same in each Region.

### 1.4 Sample selection schemes

***1st stage of sampling***

In this stage, from any final stratum, say stratum $h$, $n_h$ primary units (areas) were drawn. The number $n_h$ of draws was approximately proportional to the population size $X_h$ of the stratum. The population size $X_h$ of the stratum is defined by the number of households according to the population census of the year 2011.

In each final stratum attention was paid so as the primary units drawn, to be a multiple of four. Thus, the sample of primary units can be divided in 4 sub-samples of equal size. The reference period for the household data of each one of the 4 sub-samples corresponds to each one of the 4 quarters of the year, in order to allow for full representativeness of the household consumption expenditures.

Each area unit (primary sampling unit) of the stratum had a selection probability proportional to its size. So, if $X_{hi}$ was the number of households, according to the 2011 census population, of the area of order $i$ in the sample, then the selection probability of the unit was:

$$P_{hi} = \frac{X_{hi}}{X_h} \quad (1)$$

The total number of the primary sampling units is 1,068. Due to non-response, the actual total number of primary sampling units is 1,064.

Additionally, as in each year the 25% of the sample households is replaced, the new households belong to different primary sampling units

***2nd stage of sampling***

In this stage from each primary sampling unit (selected area) the sample of ultimate units (households) is selected. Actually, in the second stage we draw a sample of dwellings. However, in most cases, there is one to one relation between household and dwelling. If the selected dwelling consists of one or more households then all of them are interviewed.

Let $M_{hi}$ be the number of households during the survey period in the selected area *i* of stratum $h$. Out of them a systematic sample of $m_{hi}$ households is selected with equal probabilities. Each of the $m_{hi}$ households has the same chance to be included in the survey, equal to: $\frac{m_{hi}}{M_{hi}}$

In every selected primary unit, remains the determination of the sample size $m_{hi}$. The total number of households to be interviewed of the $n_h$ selected primary sampling units will be

$$m_h = \sum_{i=1}^{n_h} m_{hi} \quad (2)$$

i.e. finally by applying the two stage sampling procedure, the sampling rate of households in stratum $h$ is $\frac{m_h}{M_h}$, where $M_h = \sum_{i=1}^{n_h} M_{hi}$.

In repeated sampling, the numerator of this fraction will vary from sample to sample; to be more specific the fraction $\frac{m_h}{M_h}$ is a random variable. Within each primary sampling unit the calculation of the sampling interval $\delta_{hi} = \frac{M_{hi}}{m_{hi}}$ is carried out, so that the following two desired conditions are satisfied.

a) The expected result $\frac{m_h}{M_h}$ is the predetermined over sampling fraction $\frac{1}{\lambda}$ in each Region (NUTS 2): $E\left(\frac{m_h}{M_h}\right)=\frac{1}{\lambda}$

b) The estimator of the stratum total $Y_h$ (for any characteristic) should be self-weighting. In other words, the calculated estimator is the result derived from the sum of the values of the characteristic over the $m_h$ sample households by the overall raising factor $\lambda$, which is the same in each Region.

The conditions (a) and (b) are satisfied when:

$$\frac{1}{n_h}\cdot\frac{1}{P_{hi}}\cdot\frac{M_{hi}}{m_{hi}}=\lambda \quad \Rightarrow \tag{3}$$

$$\frac{1}{n_h}\cdot\frac{1}{P_{hi}}\cdot\delta_{hi}=\lambda \Rightarrow$$

$$\delta_{hi}=\frac{M_{hi}}{m_{hi}}=\lambda\cdot n_h\cdot P_{hi} \tag{4}$$

### 1.5 Renewal of the sample: rotational groups

The survey is a *simple rotational design* survey. The sample for any year consists of 4 replications, which have been in the survey for 1-4 years. With the exception of the first three years of the survey, any particular replication remains in the survey for 4 years. Each year, one of the 4 replications from the previous year is dropped and a new one is added. Between year T and T+1 the sample overlap is 75%; the overlap between year T and year T+2 is 50%; and it is reduced to 25% from year T to year T+3, and to zero for longer intervals.

## 2. Weightings

Let $w_{hij}$ (>0) stand for the survey weight attached to the sample ultimate unit (household) of order $j$ ( $j=1,...,m_{hi}$ ), belonging to the selected area of order $i$, of stratum $h$. The $w_{hij}$ is the product of three factors: a) the inversion of the inclusion probabilities of the ultimate sampling units, b) the inversion of the weighted response rate $r_h$ in stratum $h$ and c) a factor $t_{hij}$, which makes weighted sample estimates to conform to external total values (values from known totals from censuses, administrative sources, population projections etc). The weight $w_{hij}$ is defined as follows:

$$w_{hij}=p_{hij}^{-1}\cdot r_h^{-1}\cdot t_{hij}$$

where:

$p_{hij}$: Inclusion probability of the $hij$ ultimate unit

$r_h$: Weighted response rate of the ultimate units in stratum $h$

$t_{hij}$ : Factor that adjusts the total of households and individuals to external data

### 2.1 Inclusion probabilities of households

A two-stage sampling scheme was applied, according to which in the final strata the areas were selected with probabilities proportional to their sizes and within the selected areas the households were selected with equal probabilities. Then the inclusion probabilities of households are defined, as follows:

$$p_{hij} = n_h \cdot P_{hi} \cdot \frac{m_{hi}}{M_{hi}} \Rightarrow p_{hij}^{-1} = \frac{1}{n_h} \cdot \frac{1}{P_{hi}} \cdot \frac{M_{hi}}{m_{hi}} \quad (5)$$

where:

$P_{hi} = \frac{X_{hi}}{X_h}$ : Selection probability of the $hi$ area

$X_{hi}$ : The number of households that belong to the $hi$ area, according to the population census of 2011

$X_h$ : The number of households that belong to stratum $h$, according to the population census of 2011

$M_{hi}$ : The number of households in the $hi$ area that are recorded in the updated sampling frame

$m_{hi}$ : The initial sample size of households in the $hi$ area that were selected from the $M_{hi}$ units

### 2.2 Non-response adjustments

Within each final stratum non-response adjustment of the responding households was carried out by the inverse of the weighted response rate, so as to adjust for non-responding cases in that stratum.

### 2.3 Adjustment to external data

The adjustment to external data was conducted. This involves the calibration of the household weights in conjunction with external sources. It enables the distribution of auxiliary variables at both household and individual level to coincide with the corresponding population distribution of the external data. The auxiliary variables used at household level are the household size and at individual level the gender and age (ten years age groups).
By applying calibration: a) the estimated households by size conform to the number of households of the reference period resulting from the projection of the trend observed between the population 2017 and 2018 and b) the estimated population by gender and age conforms to the population projections for the reference period. These projections are based on vital statistics (population census, births, deaths, migration) and the Population Census 2011.

### 2.4 Triming

The final weights were trimmed iteratively so that to avoid the existence of extreme (large) weights which lead to increment of estimations' variance

## 3. Sampling Errors

### 3.1 Estimation of survey characteristics

Let $y_{hij}$ be the value of the characteristic **y** of the sampling household of order $j$, in the $hi$ primary sampling unit (area). Moreover, $Y_h$ stands for the stratum total, which results when adding the characteristic $y$ for all households or household members included in stratum $h$.
The form of the estimator on the basis of the two-stage design is:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \cdot y_{hij} \tag{6}$$

Where $w_{hij}$ is the final (adjusted) weight of the household

For estimating the characteristic **y** at country level, all stratum estimates $\hat{Y}_h$ should be added, as follows:

$$\hat{Y} = \sum_{h} \hat{Y}_h \tag{7}$$

### 3.2 Estimation of a Ratio

The estimation of the number of households $X_h$ in stratum h is calculated using the formula:

$$\hat{X}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \tag{8}$$

while the estimation of the relevant characteristic in country level is calculated by adding all strata estimations, that is:

$$\hat{X} = \sum_{h} \hat{X}_h \tag{9}$$

The form of the estimator $\hat{R}$ (mean household consumption expenditure) on the basis of the two-stage design is:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}}$$

### 3.3 Variance Estimation

In order to estimate the variances of the required characteristics (mean household consumption expenditure for the various categories of expenditures), the following steps should be applied.

**a.** For every selected PSU $i$ of the stratum $h$, we calculate the quantities $T_{hi}$ and $F_{hi}$ using the following formulas:

$$T_{hi} = n_h \cdot \sum_{j=1}^{m_{hi}} w_{hij} \cdot y_{hij} \tag{10}$$

$$F_{hi} = n_h \cdot \sum_{j=1}^{m_{hi}} w_{hij} \tag{11}$$

**b.** After having calculated $T_{hi}$ and $F_{hi}$ for every PSU $i$ ($i = 1,2,...,n_h$) of stratum $h$, then :

$V\left(\hat{Y}_h\right)$ is calculated as:

$$V\left(\hat{Y}_h\right) = \frac{1}{n_h \cdot (n_h - 1)} \cdot \left[ \sum_{i=1}^{n_h} T_{hi}^2 - \frac{1}{n_h} \cdot \left( \sum_{i=1}^{nh} T_{hi} \right)^2 \right] \tag{12}$$

and

$V\left(\hat{Y}\right)$ (country level) is calculated by adding $V\left(\hat{Y}_h\right)$ for all strata $h$, that is:

$$V\left(\hat{Y}\right) = \sum_h V\left(\hat{Y}_h\right) \tag{13}$$

Correspondingly, $V\left(\hat{X}_h\right)$ is given by:

$$V\left(\hat{X}_h\right) = \frac{1}{n_h \cdot (n_h - 1)} \cdot \left[ \sum_{i=1}^{n_h} F_{hi}^2 - \frac{1}{n_h} \cdot \left( \sum_{i=1}^{nh} F_{hi} \right)^2 \right] \tag{14}$$

and

$V\left(\hat{X}\right)$ (country level) is calculated by adding $V\left(\hat{X}_h\right)$ for all strata $h$, that is:

$$V\left(\hat{X}\right) = \sum_h V\left(\hat{X}_h\right) \tag{15}$$

The variance of $\bar{R}$ can be calculated using the formula below

$$V(\hat{R}) = \frac{V(\hat{Y}) + \hat{R}^2 V(\hat{X}) - 2\hat{R}Cov(\hat{X},\hat{Y})}{\hat{X}^2}$$

(166)

where

$$Cov(\hat{X}_h, \hat{Y}_h) = \frac{1}{n_h(n_h-1)}\left[\sum_{i=1}^{n_h} T_{hi}F_{hi} - \frac{1}{n_h}\left(\sum_{i=1}^{n_h} T_{hi}\right)\left(\sum_{i=1}^{n_h} F_{hi}\right)\right] \tag{17}$$

and

(17)

$$Cov(\hat{X}, \hat{Y}) = \sum_h Cov(\hat{X}_h, \hat{Y}_h)$$

In order to estimate the variances for mean household consumption expenditure for certain population subsets, the same procedure described above is followed. For that case, we also defined domain indicator variables in order to represent the specific population subsets (domains) required, (e.g. age of the household' s reference person: less than 30, 30-44, 45-59 and 60+ years)

Let,

- the specific population subset (the domain) be denoted $U_d$, where $U_d \subset U$ (whole population)
- the size of $U_d$ be denoted $N_d$

then the value for the j unit (household or household reference person) in the selected area i of the final stratum h of the domain indicator variable is denoted as:

$$y_{hij} = \begin{cases} y_{hij} & if \quad i \in U_d \\ 0 & otherwise \end{cases}$$

$$w_{hij} = \begin{cases} w_{hij} & if \quad i \in U_d \\ 0 & otherwise \end{cases}$$

With the use of the domain indicators above and the procedure and formulas already described we estimated the characteristics and the sampling errors of the mean household final consumption expenditure of the specific subpopulations.

### 3.4 Standard Errors and Coefficients of Variation

Standard errors and coefficients of variation were calculated for mean household consumption expenditure for certain expenditure categories and population subsets. They are presented in the following tables.

For an estimate $\hat{R}$, the coefficient of variation is defined as:

$$CV\left(\hat{R}\right) = \frac{\sqrt{V\left(\hat{R}\right)}}{\hat{R}} * 100 \qquad (18)$$

## 4. Design Effect

The design effect for survey estimates is used as a tool to measure sample efficiency and to assess the effect of sample design beyond the variability in Simple Random Sampling. The design effect is defined as the ratio of the variance of an estimate under the complex sample design to the variance of the same estimate that would have been obtained from a simple random sample of the same size. The Household Budget Survey employs complex sample design that involves stratification, unequal weighting and clustering.

The design effect was calculated by the following formula:

$$\boldsymbol{deft^2}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{swc}}\right) = \frac{\boldsymbol{V}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{swc}}\right)}{\boldsymbol{V}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{srs}}\right)}$$

where:

$\theta$ : parameter such as $R$ (Ratio)

s : represents stratification

w : represents weighting

c : represents clustering

SRS: Simple Random Sampling

In our study, $\hat{\theta} \equiv \hat{R}$, therefore for the calculation of $V\left(\hat{R}\right)$ in the nominator, we use formulae (16) above.

For the calculation of the denominator we apply the formula,

$$V\left(\hat{R}_{SRS}\right) = \frac{V\left(\hat{Y}_{SRS}\right) + \hat{R}_{SRS}^2 V\left(\hat{X}_{SRS}\right) - 2\hat{R}_{SRS} Cov\left(\hat{Y}_{SRS}, \hat{X}_{SRS}\right)}{\hat{X}_{SRS}^2}$$

where,

$\hat{Y}_{SRS}$ : Estimation of a characteristic **y** after applying SRS

$\hat{X}_{SRS}$ : Estimation of the number of households after applying

**Coefficients of variation of estimation of mean annual total expenditure of 12 main categories of goods and services: 2021 HBS**

| *Goods and services* | *Coefficient of variation* *%* |
|---|---|
| *Total* | **1.7** |
| Food | 1.0 |
| Alcoholic beverages and tobacco | 2.3 |
| Clothing and footwear | 4.4 |
| Housing | 1.3 |
| Durables | 4.3 |
| Health | 2.6 |
| Transport | 3.4 |
| Communications | 1.2 |
| Recreation and culture | 7.1 |
| Education | 4.1 |
| Hotels, cafes, and restaurants | 2.8 |
| Miscellaneous goods and services | 2.4 |