BANCO DE PORTUGAL
EUROSYSTEM

Practical aspects of using the ISFF database

1. Database files

The database of the Portuguese Household Finance and Consumption Survey (ISFF) made available by Statistics Portugal is organised by type of variable and includes nine files with the following content:

1) **S variables**: variables that identify the initial sample (country, wave, household number, etc.), that record the interviewer contacts and the final outcome and that include some characteristics of the dwelling; these data refer to all dwellings included in the initial sample (gross sample).
2) **H variables:** variables of the interview at the household level and the technical variable with the final weight; these data refer to respondent households (net sample).
3) **R variables:** variables of the interview at individual level (demographics); these data refer to all members of respondent households.
4) **P variables:** variables of the interview at individual level (other than demographics); these data refer to all household members aged 16 and over

5, 6, 7 e 8) **FS, FH, FR and FP variables:** shadow variables (flags) of the S, H, R and P variables, respectively.

9) **W variables:** 1000 replicates of the final weight; these data refer to respondent households.

A list of all variables comprising the S, H, P and R files and the flag codes can be found on the section 'ISFF database' on the ISFF page of Banco de Portugal's website.

In the ISFF database, there are no missing values for most variables, as missing values were imputed. A stochastic multiple imputation method was applied. The method produces five alternative estimates (five implicates) for each missing value.

Consequently, for each H, R, P, FH, FR and FR variable the number of observations is five times the number of respondents, either households or individuals. For the same household, the five observations of H, R and P variables are the same when the response was provided during the interview and different when it was imputed.

2. Identification variables and links between files

The ISFF database contains the following identification variables:

- SA0010: household identification number
- RA0010: personal identification number, inside each household.
- IM0100: implicate identification number

Variable SA0010 is present in all nine files. Variable RA0010 is present in all files with data at individual level (R, FR, P and FP). Variable IM0100 is present in all files that include variables collected from households and in all respective flag files (H, FH, R, FR, P and FP). For example, variables SA0010 and IM0100 must be used to merge H and P files and variables SA0010, RA0010 and IM0100 must be used to merge the R and P files.

3. Naming conventions of the H, P and R variables

The **first character** of the name of the variables of the interview identifies the variable reference unit: H-Household; R-All household members; P-Household members aged 16 and over.

In the ISFF, the H, P and R variables may be core, non-core or national variables. Core and non-core variables have harmonized definitions across the countries participating in the HFCS. Core variables are available for all countries and non-core variables are only available for the countries that have decided to adopt them. National variables are specific to the ISFF and are not included in the HFCS database. For non-core and national variables, the **second character** of the name is N and O respectively.

The **second (third) character** of the core variables (non-core and national variables) identifies the respective section of the questionnaire, in accordance with the following list:

- A – Socio-demographic aspects
- B – Real assets and loans collateralised by real estate
- C – Other liabilities and access to credit
- D – Private businesses and financial assets
- E – Labour status
- F – Pension rights
- G – Income
- H – Inheritances and gifts
- I – Consumption and saving

M**ultiple choice questions** originate multiple variables whose names only differ in their **last character** (a letter).

In the ISFF, some questions are repeated in **loops with three iterations** (for example, questions on loans using the main residence as collateral are asked for each of the three loans with the highest outstanding amount). Responses to the same question in each of the three iterations are recorded in variables whose names differ only in the **last character** (1, 2 or 3, which identifies the iteration). For instance, HB1701, HB1702 and HB1703 correspond to the outstanding amount of each of the three major loans using the main residence as collateral. In multiple choice questions that are part of a loop, the iteration number corresponds to the **second to last character** of the variable’s name. For instance, the HB1201A,…, HB1201I variables correspond to the reasons for taking out the loan with the highest outstanding amount.

4. Use of multiple imputation, weights and replicates

When using ISFF data, the following technical characteristics of the survey should be taken into account:

i) The ISFF sample is a probabilistic sample, which means that, in order to obtain extrapolated results for the households living in Portugal, each observation should be multiplied by the weight of the household in question;

ii) The five implicates should be used together in order to take into account the uncertainty of the imputation process both in obtaining point estimates and estimating variances.

iii) Replicate weights should be used in variance estimation in order to take into account the uncertainty associated with the sample selection.

In order to calculate an estimate of a given parameter θ (for example, the mean, median, or the parameter of a regression), first the values of the estimates of that parameter are obtained for each of the five implicates ($\hat{\theta}_m$) and then the arithmetic average of the five estimates is calculated, i.e.:

$$\bar{\theta} = \tfrac{1}{5}\textstyle\sum_{m=1}^{5} \hat{\theta}_m \qquad (1)$$

In turn, in order to take into account the uncertainty of the imputation process, the variance of this estimator (T) corresponds to a combination of the variance within each implicate (within variance - W) and of the variance between implicates (between variance - B).

$$T = W + \left(1 + \tfrac{1}{5}\right) B \qquad (2)$$

$$W = \tfrac{1}{5}\textstyle\sum_{m=1}^{5} \hat{V}_m(\,\hat{\theta}_m) \qquad (3)$$

$$B = \tfrac{1}{4}\textstyle\sum_{m=1}^{5} (\hat{\theta}_m - \bar{\theta})^2 \qquad (4)$$

When estimating parameters for the population, the sample final weight should be used to calculate the point estimates and the 1000 replicates of the final weight should be used to calculate the variance of the parameter estimate. For example, the estimate of the mean income of the households obtained with implicate *m* is given by:

$$\hat{\theta}_m = \frac{1}{N}\sum_{i=1}^{n} w_i y_{im} \qquad (5)$$

where $\mathrm{w_i}$ is the sample final weight for household *i*, $\mathrm{y_{im}}$ is the income of household *i* according to implicate *m*, *n* is the number of households in the net sample, *N* is the number of households living in Portugal.

The variance of the estimate of the mean income is given by:

$$V_m\left(\hat{\theta}_m\right) = \frac{1}{1000-1}\sum_{b=1}^{1000}(\theta^*_{mb} - \bar{\theta}^*_m)^2 \qquad (6)$$

θ^*_{mb} corresponds to the mean income estimated with implicate *m* and using as weights the replicate *b:*

$$\theta^*_{mb} = \frac{1}{N}\sum_{i=1}^{n} w_{ib} y_{im} \qquad (7)$$

$\bar{\theta}^*_m$ is the average of the estimates for the mean income obtained with each of the 1000 replicate weights and using the income data of implicate *m*:

$$\bar{\theta}^*_m = \frac{1}{1000}\sum_{b=1}^{1000} \theta^*_{mb} \qquad (8)$$

To facilitate the use of this type of data, a number of statistical software include specific commands for the use of survey data and multiple imputed data. These commands take into account the fact that the five implicates should not be treated as independent, as that would result in an overestimation of the significance of estimated coefficients.[1]

[1] For some useful examples of Stata codes, see Albacete, N., S. Thandi Dippenaar, P. Lindner, K. Wagner, *Methodological notes for Austria: Eurosystem Household Finance and Consumption Survey 2017*, January 2019.