Sampling strategy
and degree of precision of the results

1. Knowledge aims

The population included in the multi-purpose survey "Aspects of daily life", i.e. the set of statistical units around which it is designed to inquire, consists of households that are resident in Italy and the members that comprise them; therefore permanent members of communities are excluded. The terms "household" is used to mean *de facto* family, i.e. a group of people who live together and are bound by marriage, kinship, affinity, adoption, guardianship or emotional ties.

The reference period is primarily constituted by the twelve months prior to the interview, although for some questions the reference is to the time of interview.

The domains of study, i.e. the areas to which the parameters of the population being studied refer, are:

- the country overall;
- the five geographical macro-areas (North-west Italy, North-east Italy, Central Italy, South Italy and the Islands);
- the geographical regions (with the exception of Trentino-Alto Adige, for which estimations are produced separately for the provinces of Bolzano and Trento);
- the type of municipality, obtained be dividing Italian municipalities into six classes formed according to socio-economic and demographic characteristics:

A) municipalities belonging to the metropolitan area, divided into:

A_1, municipalities that are metropolitan area centres: Turin, Milan, Venice, Genoa, Bologna, Florence, Rome, Naples, Bari, Palermo, Catania and Cagliari;

A_2, municipalities which gravitate around municipalities that are metropolitan area centres;

B) municipalities not belonging to the metropolitan area, divided into:

B_1 municipalities with up to 2,000 inhabitants;
B_2 municipalities with 2,001-10,000 inhabitants;
B_3 municipalities with10,001-50,000 inhabitants;
B_4 municipalities with over 50,000 inhabitants.

2. Sampling strategy

2.1 General description of the sampling design

The sampling design is of the complex type and makes use of two different sampling schemes. Within each of the domains defined by the crossing of the geographical region with the six areas A_1, A_2, B_1, B_2, B_3 and B_4, the municipalities are divided into two subsets according to resident population:

- the set of self-representative municipalities (which we will denote hereafter as Ar municipalities) constituted by the municipalities of greater demographic dimension;
- the set of non-self-representative municipalities (or Nar) constituted by the remaining municipalities.

Within the set of Ar municipalities, each municipality is considered as a standalone stratum and a design known as cluster sampling is adopted. The primary sampling units are represented by the households extracted systematically from the municipality's own population register. For each household included in the sample the characteristics being studied are gathered for all of the *de facto* members belonging to the household itself.

Within the Nar municipalities a two-stage design is adopted with stratification of the primary units. The primary units (Up) are the municipalities, while the secondary units are the households taken from the municipal population registers. For each household included in the sample the characteristics being studied are gathered for all of the *de facto* members belonging to the household itself.

The municipalities are selected with probabilities that are proportional to their demographic size and without replacement, while the households are extracted with equal probability and without replacement.

2.2 Definition of the sample size

For a survey with multiple objectives, such as the one under consideration, it is unrealistic to expect to be able to design a sampling strategy that ensures predetermined levels of precision for all of the estimations produced. The question is complicated by the fact that the purpose of the survey is to produce estimations for different area levels, which entails the adoption of various, opposing optimal solutions. If, for example, the sole scope of publication of the estimations were the national level, an approximately optimal solution would be to determine the national sample size and to distribute it between regions in proportion to their demographic size; conversely, if the aim were to produce estimates with equal reliability at the regional level, an approximately optimal solution would be to select an equal sample in all regions. The latter solution, however, is inefficient for estimations at the national level. In order to deal with this problem, in keeping with what has been carried out in other countries, use has been made of a strategy that enables the sample size to be established by means of successive approximations.

On the basis of the foregoing considerations it was decided to adopt a mixed approach, based on criteria of both cost and organisation, as well as on an evaluation of sampling errors in the main estimates at the national level and with reference to each of the area domains concerned.

The criteria followed can be summed up in the following points:

- the theoretical sample size in terms of households, which was predetermined at the national level essentially on the basis of cost and operational criteria, was around 24,000 households;
- the number of sample municipalities involved could not be greater than 900 in order to allow adequate checking and supervision.

The allocation of the sample of households and of municipalities between the various regions was therefore calculated by adopting a criterion of compromise that would guarantee both the reliability of the estimates at the national level and the reliability of estimates for each of the area domains described in Section 1.

2.3 Stratification and selection of sample units

The aim of stratification is to form groups (or strata) of units that are characterised, in relation to the target variables of the survey, by the greatest possible internal consistency of the strata and the greatest heterogeneity between strata. The achievement of this aim, in statistical terms, translates into gains in precision of the estimates, or in other words a reduction of the sampling error, sample size remaining equal.

In the survey under consideration, municipalities are stratified according to demographic size and in accordance with the following conditions:

- self-weighting of the sample at the regional level;
- selection of a sample municipality within each stratum defined on the municipalities of the Nar set;
- selection of a minimum number of households to interview in each sample municipality. This number was set at 23;
- formation of strata with approximately constant breadth in terms of resident population.

The stratification procedure, implemented within each area domain identified by the areas A_1, A_2, B_1, B_2, B_3 and B_4 of each geographical region, is structured according to the following stages:

- ordering of the municipalities in the domain in descending order according to their demographic size in terms of resident population;
- establishment of a population threshold for defining Ar municipalities, by means of the relation

$$ {}_r\lambda = \frac{{}_r\overline{m}\ {}_r\delta}{{}_r f} $$

in which for the generic geographical region r, ${}_r\overline{m}$ indicates the minimum number of households to interview in each sample municipality, ${}_r\delta$ the average number of members per household, and ${}_r f$ the sampling fraction;

- division of all municipalities into the two subsets Ar and Nar: municipalities greater than or equal to ${}_r\lambda$ in size are defined as Ar municipalities, and the remainder are defined as Nar municipalities;
- division of municipalities in the Nar set into strata with approximately constant size and roughly equal to the threshold ${}_r\lambda$.

After stratification, the Ar municipalities are included with certainty in the sample, while for the Nar municipalities, within each stratum a sample municipality is extracted with probability proportional to demographic size, by means of the systematic selection procedure proposed by Madow.[1]

Households to interview in each sample municipality are selected from the population register of each municipality without replacement and with equal probabilities.

Specifically, the selection technique is of the systematic type and within each municipality is performed according to the following stages:

- the household records from the municipal population register are placed in sequence;
- the sampling interval e_{hi} is calculated as the ratio between the number of households resident in municipality i of stratum h and the corresponding number of sample households, $e_{hi}=M_{hi}/m_{hi}$;
- m_{hi} households which occupy the following positions in the sequence constructed in point 1) are selected:

$$ 1,\ 1+e_{hi},\ 1+2e_{hi},\ .....,\ 1+(m_{hi}-1)e_{hi}. $$

Summary 1 gives the regional distribution of the universe and the sample of municipalities, households and individuals.

[1] Madow, W.G. "On the theory of systematic sampling II", *Annals of Mathematical Statistics,* 20, (1949): 333-354.

Summary 1 – Regional distribution of municipalities, households and individuals in the universe and in the sample – 2009

REGIONS	Municipalities		Households		Individuals	
	Sample	Universe (a)	Sample	Universe (a)	Sample	Universe (a)
Piemonte	62	1.206	1.426	1.929	3.326	4.391
Valle d'Aosta - Vallée d'Aoste	21	74	453	56	1.032	126
Lombardia	84	1.546	1.716	4.051	4.153	9.674
Trentino-Alto Adige	48	339	1.076	414	2.652	1.008
Bolzano - Bozen	*23*	*116*	*550*	200	1.366	494
Trento	*25*	*223*	*526*	214	1.286	514
Veneto	53	581	1.089	1.930	2.739	4.839
Friuli-Venezia Giulia	32	219	705	528	1.642	1.217
Liguria	26	235	843	764	1.810	1.601
Emilia-Romagna	46	341	1.052	1.877	2.466	4.305
Toscana	50	287	1.120	1.558	2.678	3.682
Umbria	22	92	580	356	1.428	888
Marche	36	246	755	608	1.942	1.560
Lazio	33	378	1.005	2.348	2.378	5.582
Abruzzo	37	305	726	520	1.832	1.328
Molise	23	136	560	123	1.420	319
Campania	55	551	1.316	2.021	3.710	5.794
Puglia	50	258	1.135	1.479	3.125	4.064
Basilicata	27	131	602	228	1.555	588
Calabria	41	409	881	752	2.330	1.999
Sicilia	53	390	1.293	1.916	3.373	5.015
Sardegna	38	377	794	653	2.012	1.662
Italia	**837**	**8.101**	**19.127**	**24.112**	**47.603**	**59.643**

(a) Multi-purpose survey "Aspects of daily life" estimate

2.4 Procedure for calculation of estimates

The estimates produced by the survey are essentially estimates of absolute and relative frequencies, referring to households and individuals.

The estimates are obtained by means of a calibration estimator, which is the estimation method used in most Istat surveys of enterprises and households.

The principle on which each sampling estimation method is based is that the units belonging to the sample also represent the units of the population which are not included in the sample.

This principle is satisfied by assigning to each sampling unit a weight indicating the number of units of the populations represented by that unit. If, for example, a sampling unit is assigned a weight of 30, then this unit represents itself and another 29 units of the population which have not been included in the sample.

In order to make the exposition which follows clearer, we shall introduce the following symbols: d, index of geographical reference level for the estimates; i, index of municipality; j, index of household; p, index of household member; h, index of stratum of municipalities; y, generic survey target variable; Y_{hijp}, value of y observed for member p of household j of municipality i of stratum h; P_{hij}, number of members of household j of municipality i of stratum h; $Y_{hij} = \sum_{p=1}^{P_{hij}} Y_{hijp}$, total for variable y observed for household j of municipality i of stratum h; M_{hi}, number of households resident in municipality i of stratum h; m_{hi}, sample of households in municipality i of stratum h; N_h, total of municipalities in stratum h; n_h, number of sample municipalities in stratum h (in the survey under consideration $n_h = 1$); H_d, total number of strata in the generic area domain d.

It is supposed that we wish to estimate, with reference to a generic domain d, the total of the generic survey target variable y, expressed by the following relation:

$$Y_d = \sum_{h=1}^{H_d} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} \,. \qquad (1)$$

The estimate of the total (1) is given by

$$\hat{Y}_d = \sum_{h=1}^{H_d} \hat{Y}_h \text{ , since } \hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij} Y_{hij} \,, \qquad (2)$$

in which W_{hij} is the final weight to attribute to all of the members of household j of municipality i of stratum h.

From the previous relation it follows that in order to obtain the estimate of the total (1) it is necessary to multiply the value of the variable y assumed by each sampling unit by the weight of that unit[2] and perform, at the level of the domain in question, the sum of the products thus obtained.

The weight to be assigned to the sampling units is obtained by means of a complex procedure which:

- corrects the distorting effect of the lack of total response due to the impossibility of interviewing some of the selected households due to unavailability or refusal to be interviewed;
- takes account of knowledge of known totals of important auxiliary variables (available from sources external to the survey), in the sense that the sampling estimates of known totals of auxiliary variables must coincide with the known values for those auxiliary variables.

In the survey under consideration, for each geographical region we define 18 known totals, which refer to the distributions of the regional population by sex and six age classes[3] and of the regional population in the six areas A_1, A_2, B_1, B_2, B_3 and B_4. Thus by indicating with ${}_kX$ (k=1,…,18) the known total of the k-th auxiliary variable for the generic geographical region and with ${}_kX_{hij}$ the value assumed by the k-th auxiliary variable for the respondent household hij, the condition described above is expressed by the following equation:

$${}_kX = {}_k\hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hijk} X_{hij} \quad (k=1,\ldots., 18)$$

in which H indicates the overall number of strata defined in the region. If, for example, ${}_6X$ indicates the number of males aged sixty-five or over, the auxiliary variable ${}_6X_{hij}$ represents the number of males aged sixty-five or over in the household hij.

The procedure that enables the construction of the *final weights* to assign to respondent sampling units is structured according to the following stages:

1) the *direct weights* are calculated as the reciprocal of the probability of inclusion of the units;
2) the corrective factors due to lack of total response are calculated as the inverse of the response rate for the municipality to which each units belongs;
3) the *base weights*, or weights corrected for lack of total response, are calculated by multiplying the direct weights by the corresponding corrective factors due to lack of total response;
4) the corrective factors that make it possible to satisfy the condition of equality between known totals of auxiliary variables and the corresponding sampling estimates are calculated;
5) lastly, the final weights are calculated by means of the product of the base weights and the corrective factors obtained in stage 4.

The corrective factors in stage 4 are obtained by solving a minimum bound problem, in which the function to minimise is a function of distance (suitably selected in advance) between the base weights and the final weights and the constraints are defined by the condition of equality between sampling estimates of known population totals and known values of those totals. The preselected distance function is the truncated

[2] In order to obtain consistent estimated for individuals and households the final weights are defined so that each household hij and all of the members of the household are assigned the same final weight W_{hij} .

[3] The age classes considered are: 0-5, 6-13, 14-24, 25-44, 45-64, and 65 and over.

logarithmic function; the adoption of this function ensures that the final weights are positive and contained within a predetermined range of possible values, thus eliminating extreme positive weights (too large or too small).

All of the estimation methods that stem from the solution of a minimum bound problem of the type described above fall within a general class of estimators known as calibration estimators.[4] An important estimator belonging to this class, which is obtained using the Euclidean distance function, is the *generalised regression estimator*. As will be clarified in Section 3, this estimator assumes a central role, as it is possible to demonstrate that all calibration estimators converge asymptotically as sampling size increases towards the generalised regression estimator.

3. Evaluation of the level of precision of estimates

3.1 Method of calculating sampling errors

The main statistics of interest for evaluating the sampling variability of estimates produced by a survey are the absolute sampling error and the relative sampling error. By denoting with $\hat{V}ar(\hat{Y}_d)$ the estimate of variance of the generic estimate $\hat{Y}_d$, the estimate of the absolute sampling error for $\hat{Y}_d$ can be obtained by means of the following expression:

$$\hat{\sigma}(\hat{Y}_d) = \sqrt{\hat{V}ar(\hat{Y}_d)}\,; \tag{3}$$

while the estimate of the relative sampling error for $\hat{Y}_d$ is defined by the expression:

$$\hat{\varepsilon}(\hat{Y}_d) = \frac{\hat{\sigma}(\hat{Y}_d)}{\hat{Y}_d}\,. \tag{4}$$

As described in 2.4 above, the estimates produced by the survey were obtained by means of a calibration estimator defined on the basis of a truncated logarithmic distance function. Since the estimator adopted is not a linear function of the sampling data, for the estimation of the variance $\hat{V}ar(\hat{Y}_d)$ the method proposed by Woodruff was used; on the basis of this method, which makes use of the Taylor series linearised expression, the variance of each non-linear estimator can be obtained (regular function of totals) by calculating the variance of the linearised expression obtained. Specifically, for the definition of the linearised expression of the estimator, reference was made to the generalised regression estimator, making use of the asymptotic convergence of all calibration estimators towards this estimator, as in the case of calibration estimators that use distance functions which are different from Euclidean distance (which leads to the generalised regression estimator) it is not possible to derive the linearised expression of the estimator.

The linearised expression of the estimator (2) is thus given by

$$\hat{Y}_d \cong \hat{Z}_d = \sum_{h=1}^{H_d} \hat{Z}_h \quad \text{, since } \hat{Z}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hj}} Z_{hij} W_{hij} \tag{5}$$

where Z_{hij} is the linearised variable expressed as $Z_{hij} = Y_{hij} - \mathbf{X}'_{\mathbf{hij}}\beta$, since $\mathbf{X}_{\mathbf{hij}} = \left({}_1X_{hij},\ldots,{}_kX_{hij},\ldots{}_KX_{hij}\right)'$, the vector containing the values of the K (K=18) auxiliary variables, observed for the generic household hij and $\hat{\beta}$, the vector of the regression coefficients of the linear model that binds the variable in question y to the K

[4] In the Anglo-Saxon literature these estimators are known as *calibration estimators.*

auxiliary variables x. From (5) it thus follows that the estimation of the variance of estimate $\hat{Y}_d$ is obtained by means of the following relation:

$$\hat{V}ar\left(\hat{Y}_d\right) \cong \hat{V}ar\left(\hat{Z}_d\right) = \sum_{h=1}^{H_d} \hat{V}ar\left(\hat{Z}_h\right). \tag{6}$$

From (6) it follows that the estimation of the variance of estimate $\hat{Y}_d$ is calculated as the sum of the estimate of variances of the individual strata, Ar and Nar, belonging to domain d. The formula for calculating the variance, $\hat{V}ar\left(\hat{Z}_h\right)$, of estimate $\hat{Z}_h$, differs according to whether the stratum is Ar or Nar, and may thus be broken down as follows:

$$\hat{V}ar\left(\hat{Y}_d\right) \cong \hat{V}ar\left(\hat{Z}_d\right) = \sum_{h=1}^{H_{AR}} \hat{V}ar\left(\hat{Z}_h\right) + \sum_{h=1}^{H_{NAR}} \hat{V}ar\left(\hat{Z}_h\right), \tag{7}$$

in which H_{AR} and H_{NAR} indicate respectively the number of Ar and Nar strata belonging to domain d.

In the Ar strata (in which each municipality is its own stratum and $N_h = n_h = 1$, the municipality index i being superfluous and therefore omitted) the variance is estimated by means of the following expression:

$$\sum_{h=1}^{H_{AR}} \hat{V}ar\left(\hat{Z}_h\right) = \sum_{h=1}^{H_{AR}} M_h^2 \frac{\left(M_h - m_h\right)}{m_h\left(m_h - 1\right)} \sum_{j=1}^{m_h} \left(Z_{hj} - \overline{Z}_h\right)^2, \tag{8}$$

where $M_h = M_{hi}$, $m_h = m_{hi}$, $Z_{hj} = Z_{hij}$ and $\overline{Z}_h = \frac{1}{m_h} \sum_{j=1}^{m_h} Z_{hj}$.

In the Nar strata, in which a single sample municipality is extracted from each stratum, in order to estimate the sampling variance use is made of the *strata collapsing technique*. This technique consists of forming G groups each containing L_g $(L_g \geq 2)$ strata; the variance is estimated by means of the following formula:

$$\sum_{h=1}^{H_{NAR}} \hat{V}ar\left(\hat{Z}_h\right) = \sum_{g=1}^{G} \hat{V}ar\left(\hat{Z}_g\right) = \sum_{g=1}^{G} \frac{L_g}{L_g - 1} \sum_{h=1}^{L_g} \left(\hat{Z}_{hg} - \frac{\hat{Z}_g}{L_g}\right)^2 \tag{9}$$

where the quantities are expressed as:

$$\hat{Z}_{hg} = \sum_{j=1}^{m_{hi}} Z_{hij} W_{hij} \qquad \text{and} \qquad \hat{Z}_g = \sum_{h=1}^{L_g} \sum_{j=1}^{m_{hi}} Z_{hij} W_{hij}.$$

Using expressions (8) and (9) it is possible, finally, to calculate the sampling variance, $\hat{V}ar\left(\hat{Y}_d\right)$, according to (7) and to calculate, thus, according to (3) and (4) respectively the absolute sampling error and the relative sampling error.

The sampling errors expressed by (3) and by (4) make it possible to evaluate the degree of precision of the estimates; in addition, the absolute error makes it possible to construct a confidence interval, which, with confidence level P contains the target parameter of estimation. The interval is expressed as

$$\left\{\hat{Y}_d - k_p \hat{\sigma}(\hat{Y}_d) \leq Y_d \leq \hat{Y}_d + k_p \hat{\sigma}(\hat{Y}_d)\right\} \tag{10}$$

In (10) the value of k_P depends on the value set for probability P; for example, for P=0.95 we obtain k=1.96.

3.2 Statistical foundations of the procedure for calculating sampling errors

In order to calculate sampling errors in surveys of households and enterprises conducted by Istat, a computerised system developed within the institute is currently used. Section 3.1 described the methodology implemented by the procedure for calculating sampling errors in estimates produced by surveys while the current section discusses the statistical foundations and limitations of such a methodology.

In the Ar strata, in which a cluster sampling design is adopted and in which the primary units (households) are selected without replacement and equal probabilities, the procedure makes it possible to obtain estimates of sampling variance that are correct.

In the Nar strata, for which a two-stage sampling design is adopted with selection of the primary units (municipalities) without replacement and variable probabilities, the procedure makes it possible to obtain correct estimates of sampling variances when:

- in each stratum two or more primary units are selected;
- the primary units are chosen by means of independent extractions.

The first condition is not satisfied, as in the survey under consideration, a single sample municipality is selected from each stratum and to estimate the sampling variance the *strata collapsing* technique is used. This technique, which consists in forming superstrata each containing a number of strata greater than one, generally leads to overestimation of the actual sampling variance.

The second condition implies that the primary units are selected with replacement. Also this assumption is not satisfied for the Nar municipalities and this entails overestimation of the variance. It should be observed, nevertheless, that such overestimation depends on the sampling fraction for each Nar stratum: it is of negligible size in the strata in which the sampling fraction is small, while conversely it may be more significant for those strata in which the sampling fraction is larger.

3.3 Brief presentation on sampling errors

For each estimate $\hat{Y}_d$ there is a corresponding relative sampling error $\hat{\varepsilon}(\hat{Y}_d)$; this means that to enable a correct reading of the published tables it would be necessary to indicate for each published estimate the corresponding relative sampling error. This, however, is not possible, both because if time limitations and data processing costs and because the tables in the publication would be overburdened and not easy for the end user to consult. In addition, the errors for unpublished estimates, which the user can obtain independently, would not be available in any case.

For the reasons set out above, use is frequently made of an abridged presentation of relative errors, based on the *regression models method*. This method is based on the definition of a mathematical function which relates each estimate to its own relative error.

In the current survey, the model used for estimating absolute and relative frequencies is of the following type:

$$\log\left(\hat{\varepsilon}^2(\hat{Y}_d)\right) = a + b\log(\hat{Y}_d) \qquad (11)$$

where the parameters a and b are estimated using the least square method.

Summary 2 shows the values of the coefficients a and b and of the determination index R^2 of the model used for interpolation of sampling errors for absolute and relative frequency estimates, by Italy total, geographical macro-region, type of municipality and region.

On the basis of the information contained in the summary, it is possible to calculate the estimate of the relative sampling error of a specific estimate of absolute frequency $\hat{Y}_d$ by means of the following formula:

$$\hat{\varepsilon}(\hat{Y}_d) = \sqrt{\exp\left(a + b\log(\hat{Y}_d)\right)} \qquad (12)$$

which can easily be obtained from (11).

If, for example, the estimate $\hat{Y}_d$ refers to individuals in North-west Italy, the corresponding relative error is obtained by introducing into (12) the values of parameters a and b indicated in the second row of Summary 2 under Individuals (a = 8.886722, b = -1.121521).

Summaries 3 and 4, presented in addition, enable sampling errors to be calculated more easily. They regard, respectively, households and individuals, and have the following structure: a) to the side the estimate values are listed in ascending order (20,000, 30,000, …, 25,000,000); b) the successive columns contain the relative sampling errors for each area domain concerned, calculated by means of Formula (12), corresponding to the estimates of absolute frequencies in the first column.

The information contained in these summaries make it possible to calculate the relative error for a generic estimate of absolute (or relative) frequency by means of two easily applied procedures, although they lead to less precise results than those obtainable with Expression (12). The first method consists in identifying in the first column of the summary the estimate level which most closely approximates the estimate desired, and in considering as the relative error the value found in the same row, in the column which corresponds to the relevant area domain.

With the second method, the sampling error for the estimate $\hat{Y}_d$ is obtained by means of the following expression:

$$\hat{\varepsilon}(\hat{Y}_d) = \hat{\varepsilon}(\hat{Y}_d^{k-1}) - \frac{\hat{\varepsilon}(\hat{Y}_d^{k-1}) - \hat{\varepsilon}(\hat{Y}_d^{k})}{\hat{Y}_d^{k} - \hat{Y}_d^{k-1}}(\hat{Y}_d - \hat{Y}_d^{k-1}) \qquad (13)$$

where $\hat{Y}_d^{k-1}$ and $\hat{Y}_d^{k}$ are the values of the estimates, shown in the first column, within which the estimate desired $\hat{Y}_d$ is contained, and $\hat{\varepsilon}(\hat{Y}_d^{k-1})$ and $\hat{\varepsilon}(\hat{Y}_d^{k})$ are the corresponding relative errors.

Summary 2 – Values of coefficients a and b and of index of determination R2 (%) of functions used in interpolating sampling errors for estimates referring to households and individuals for Italy total, geographical macro-area, type of municipality and region – 2009

GEOGRAPHICAL AREAS	Households			Individuals		
	a	b	R^2 (%)	a	b	R^2 (%)
ITALY	8,542317828	-1,089033063	97,58267602	9,547620201	-1,159970875	92,64258047
GEOGRAPHICAL MACRO-AREAS						
North	8,312101741	-1,071633118	98,05511266	9,630430131	-1,171775517	91,97376327
North-west	*8,046856751*	*-1,049151526*	*97,3382424*	*9,474727766*	*-1,162400439*	*91,89665479*
North-east	*8,296272206*	*-1,084230923*	*97,61600799*	*9,449040783*	*-1,177595514*	*91,57547384*
Centre	8,084961765	-1,061696733	97,1457397	9,133139177	-1,146839811	92,38771757
South	7,840812894	-1,055613916	95,73051043	8,527021939	-1,107294787	92,13434098
South (mainland)	*7,615568346*	*-1,042579195*	*95,41827258*	*8,557335646*	*-1,115892328*	*92,00176126*
South (islands)	*7,743459294*	*-1,048777562*	*94,84227495*	*8,20128782*	*-1,086823837*	*91,13322283*
TYPE OF MUNICIPALITY						
A1	8,390322484	-1,080732303	98,2109756	9,717474681	-1,192544298	93,49984295
A2	7,924180505	-1,043776826	96,7309111	9,624372164	-1,186283911	92,18289781
B1	7,238306483	-1,024191717	89,21599898	7,49206074	-1,049306541	86,46130213
B2	7,796605888	-1,036260146	95,86209324	8,961372188	-1,132728445	90,94886937
B3	7,913854423	-1,051615433	94,85433041	8,635325265	-1,107572665	90,68137634
B4	8,327381916	-1,10613822	98,29713694	9,090933231	-1,167207004	93,75625462
REGIONS						
Piemonte	7,646708797	-1,047541519	96,11905441	8,174624012	-1,099822398	88,10768688
Valle d'Aosta - Vallée d'Aoste	5,244693266	-1,075555238	94,48990756	6,312762838	-1,216213432	87,84143241
Lombardia	8,481889737	-1,069055864	97,28105871	9,652002852	-1,16662067	91,47935093
Trentino-Alto Adige	6,539105863	-1,074294063	96,83740275	7,644578113	-1,192070374	91,54342946
Bolzano - Bozen	*6,576227063*	*-1,087922979*	*96,14653482*	*7,326588656*	*-1,17799668*	*91,26955629*
Trento	*6,650257168*	*-1,090390861*	*94,74996183*	*7,757455103*	*-1,216614875*	*89,50851967*
Veneto	8,380360111	-1,08498219	97,13521114	9,142395737	-1,152284197	90,05881115
Friuli-Venezia Giulia	7,483601355	-1,093975004	93,79945718	8,163219842	-1,164231002	89,54550844
Liguria	7,467289748	-1,079094912	96,7167393	7,991235518	-1,131855819	92,45110025
Emilia-Romagna	7,886049698	-1,046090558	92,52516152	9,418947844	-1,183204127	90,5339536
Toscana	7,891121011	-1,069022972	96,80824125	8,669421357	-1,140667324	92,02445347
Umbria	7,060839136	-1,078362815	96,3068871	8,333420432	-1,203980504	92,09151429
Marche	7,654593499	-1,110449975	95,78768284	8,098330825	-1,146701935	91,13771696
Lazio	8,532449227	-1,080524874	96,83133857	9,220301488	-1,138908795	92,15024614
Abruzzo	6,927324225	-1,038382547	93,99491701	7,493461864	-1,103840767	88,3863238
Molise	6,217197735	-1,103652585	94,76787437	6,748765658	-1,160689451	90,79661513
Campania	7,915713936	-1,048482498	94,38036876	9,086062888	-1,147787475	91,820623
Puglia	7,847783689	-1,068287135	94,78934556	8,150062658	-1,088716404	90,75073392
Basilicata	6,509493266	-1,075352952	95,82988695	7,11892378	-1,139235499	92,04649626
Calabria	7,383549973	-1,069947611	94,33257096	8,074425379	-1,132074685	88,82593787
Sicilia	7,853981898	-1,049991006	94,24851967	8,247163626	-1,081695771	90,55817952

(a) North-west Italy: Piedmont, Aosta Valley, Lombardy, Liguria; North-east Italy: Bolzano, Trento, Veneto, Friuli-Venezia Giulia, Emilia-Romagna; Central Italy: Tuscany, Umbria, Marche, Lazio; South Italy: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria; Italian Islands: Sicily, Sardinia.

(b) Municipalities, type A1: central metropolitan area; type A2: peripheral metropolitan area; type B1: municipalities with up to 2,000 inhabitants; type B2: from 2,001 to 10,000 inhabitants; type B3: from 10,001 to 50,000 inhabitants; type B4: over 50,000 inhabitants.

Summary 3 – Interpolated percentage values of relative sampling errors for estimates referring to households for Italy total, geographical macro-area, type of municipality and region – 2009

ESTIMATES	Italy	North	North-west	North-east	Centre	South	South (mainland)	South (islands)	A1	A2	B1	B2	B3	B4
20.000	32,6	31,7	31,0	29,5	29,7	27,1	25,8	26,7	31,5	29,9	23,4	29,1	28,6	26,9
30.000	26,1	25,5	25,0	23,7	23,9	21,9	20,9	21,6	25,3	24,2	19,0	23,6	23,1	21,5
40.000	22,3	21,8	21,5	20,3	20,5	18,8	18,0	18,5	21,6	20,8	16,4	20,3	19,9	18,3
50.000	19,8	19,4	19,2	18,0	18,2	16,7	16,0	16,5	19,2	18,6	14,6	18,1	17,7	16,2
60.000	17,9	17,6	17,4	16,3	16,6	15,2	14,6	15,0	17,4	16,9	13,3	16,5	16,1	14,6
70.000	16,5	16,2	16,1	15,0	15,3	14,0	13,4	13,8	16,0	15,6	12,3	15,2	14,8	13,4
80.000	15,3	15,1	15,0	13,9	14,2	13,0	12,5	12,9	14,9	14,5	11,5	14,2	13,8	12,5
90.000	14,4	14,1	14,1	13,1	13,4	12,2	11,8	12,1	14,0	13,7	10,8	13,4	13,0	11,7
100.000	13,6	13,4	13,3	12,3	12,6	11,6	11,1	11,5	13,2	12,9	10,3	12,7	12,3	11,0
200.000	9,3	9,2	9,3	8,5	8,7	8,0	7,8	8,0	9,1	9,0	7,2	8,8	8,5	7,5
300.000	7,5	7,4	7,5	6,8	7,0	6,5	6,3	6,4	7,3	7,3	5,8	7,2	6,9	6,0
400.000	6,4	6,4	6,4	5,8	6,1	5,6	5,4	5,5	6,2	6,3	5,0	6,2	5,9	5,1
500.000	5,6	5,6	5,7	5,2	5,4	5,0	4,8	4,9	5,5	5,6	4,5	5,5	5,3	4,5
750.000	4,5	4,5	4,6	4,1	4,3	4,0	3,9	4,0	4,4	4,5	3,7	4,5	4,3	3,6
1.000.000	3,9	3,9	4,0	3,5	3,7	3,4	3,4	3,4	3,8	3,9	3,2	3,8	3,7	3,1
2.000.000	2,7	2,7	2,8	2,4	2,6	2,4	2,3	2,4	2,6	2,7	2,2	2,7	2,5	2,1
3.000.000	2,1	2,2	2,2	2,0	2,1	1,9	1,9	1,9	2,1	2,2	-	2,2	2,1	1,7
4.000.000	1,8	1,9	1,9	1,7	1,8	1,7	1,6	-	1,8	-	-	1,9	1,8	1,4
5.000.000	1,6	1,6	1,7	1,5	1,6	1,5	1,5	-	-	-	-	1,7	1,6	1,3
7.500.000	1,3	1,3	1,4	-	-	1,2	1,2	-	-	-	-	1,4	1,3	-
10.000.000	1,1	1,1	-	-	-	-	-	-	-	-	-	-	-	-
15.000.000	0,9	0,9	-	-	-	-	-	-	-	-	-	-	-	-
20.000.000	0,8	-	-	-	-	-	-	-	-	-	-	-	-	-
25.000.000	0,7	-	-	-	-	-	-	-	-	-	-	-	-	-

ESTIMATES	Piemonte	Valle d'Aosta - Vallée d'Aoste	Lombardia	Trentino-Alto Adige	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia-Romagna	Toscana	Umbria
20.000	25,6	6,7	34,9	25,2	12,3	12,6	30,7	18,7	20,0	29,0	26,0	16,4
30.000	20,7	5,4	28,1	20,4	9,8	10,1	24,6	15,0	16,1	23,5	20,9	13,2
40.000	17,8	4,6	24,1	17,5	8,4	8,6	21,0	12,8	13,8	20,2	17,9	11,3
50.000	15,8	4,1	21,4	15,6	7,4	7,6	18,6	11,3	12,2	18,0	15,9	10,0
60.000	14,4	3,7	19,4	14,2	6,7	6,9	16,9	10,3	11,1	16,3	14,4	9,1
70.000	13,3	-	17,9	13,1	6,2	6,3	15,5	9,4	10,2	15,1	13,3	8,3
80.000	12,4	-	16,6	12,2	5,8	5,9	14,5	8,8	9,5	14,1	12,4	7,8
90.000	11,6	-	15,6	11,5	5,4	5,5	13,6	8,2	8,9	13,2	11,6	7,3
100.000	11,0	-	14,8	10,9	5,1	5,2	12,8	7,8	8,4	12,5	11,0	6,9
200.000	7,7	-	10,2	7,6	3,5	3,6	8,8	5,3	5,8	8,7	7,6	4,7
300.000	6,2	-	8,2	6,1	-	-	7,1	4,3	4,6	7,0	6,1	3,8
400.000	5,3	-	7,0	5,3	-	-	6,0	3,6	4,0	6,1	5,2	3,3
500.000	4,7	-	6,2	4,7	-	-	5,3	3,2	3,5	5,4	4,6	-
750.000	3,8	-	5,0	3,8	-	-	4,3	2,6	2,8	4,4	3,7	-
1.000.000	3,3	-	4,3	3,3	-	-	3,7	-	2,4	3,8	3,2	-
2.000.000	2,3	-	3,0	2,3	-	-	2,5	-	-	2,6	2,2	-

ESTIMATES	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	18,8	33,8	18,7	9,5	29,1	25,5	12,6	20,1	28,0	20,3
30.000	15,0	27,2	15,1	7,6	23,5	20,5	10,1	16,2	22,6	16,4
40.000	12,8	23,3	13,0	6,5	20,2	17,6	8,7	13,8	19,5	14,1
50.000	11,3	20,6	11,6	5,7	18,0	15,6	7,7	12,3	17,3	12,6
60.000	10,2	18,7	10,6	5,2	16,4	14,2	7,0	11,1	15,7	11,4
70.000	9,4	17,2	9,7	4,7	15,1	13,1	6,4	10,3	14,5	10,6
80.000	8,7	16,0	9,1	4,4	14,1	12,2	6,0	9,6	13,5	9,8
90.000	8,2	15,0	8,6	4,1	13,2	11,4	5,6	9,0	12,7	9,3
100.000	7,7	14,2	8,1	3,9	12,5	10,8	5,3	8,5	12,0	8,8
200.000	5,2	9,7	5,6	2,7	8,7	7,5	3,7	5,9	8,4	6,1
300.000	4,2	7,8	4,6	-	7,0	6,0	2,9	4,7	6,8	4,9
400.000	3,6	6,7	3,9	-	6,1	5,2	-	4,0	5,8	4,2
500.000	3,1	5,9	3,5	-	5,4	4,6	-	3,6	5,2	3,8
750.000	2,5	4,8	2,8	-	4,4	3,7	-	2,9	4,2	3,1
1.000.000	-	4,1	-	-	3,7	3,2	-	-	3,6	-
2.000.000	-	2,8	-	-	2,6	-	-	-	2,5	-

Summary 4 – Interpolated percentage values of relative sampling errors for estimates referring to individuals for Italy total, geographical macro-area, type of municipality and region – 2009

ESTIMATES	Italy	North	North-west	Nord-east	Centre	South	South (mainland)	South (island)	A1	A2	B1	B2	B3	B4
20.000	37,9	37,3	36,1	33,1	32,9	29,5	28,7	27,8	35,1	34,6	23,5	32,4	31,1	29,1
30.000	30,0	29,4	28,5	26,0	26,1	23,6	22,9	22,3	27,6	27,2	19,0	25,7	24,9	23,0
40.000	25,4	24,8	24,1	22,0	22,1	20,1	19,5	19,1	23,2	22,9	16,3	21,9	21,2	19,4
50.000	22,3	21,8	21,2	19,3	19,4	17,8	17,2	16,9	20,3	20,1	14,5	19,3	18,7	17,1
60.000	20,0	19,6	19,1	17,3	17,5	16,1	15,6	15,3	18,2	18,0	13,2	17,4	16,9	15,3
70.000	18,3	17,9	17,4	15,8	16,0	14,8	14,3	14,1	16,6	16,4	12,2	15,9	15,6	14,0
80.000	17,0	16,5	16,1	14,6	14,8	13,7	13,3	13,1	15,4	15,2	11,3	14,8	14,5	13,0
90.000	15,8	15,4	15,1	13,6	13,9	12,8	12,4	12,3	14,3	14,2	10,7	13,8	13,5	12,1
100.000	14,9	14,5	14,2	12,8	13,1	12,1	11,7	11,6	13,5	13,3	10,1	13,0	12,8	11,4
200.000	10,0	9,7	9,5	8,5	8,8	8,3	8,0	7,9	8,9	8,8	7,0	8,8	8,7	7,6
300.000	7,9	7,6	7,5	6,7	7,0	6,6	6,3	6,4	7,0	6,9	5,7	7,0	6,9	6,0
400.000	6,7	6,4	6,3	5,7	5,9	5,6	5,4	5,5	5,9	5,8	4,9	5,9	5,9	5,1
500.000	5,9	5,7	5,6	5,0	5,2	5,0	4,8	4,8	5,2	5,1	4,3	5,2	5,2	4,4
750.000	4,6	4,5	4,4	3,9	4,1	4,0	3,8	3,9	4,0	4,0	3,5	4,2	4,2	3,5
1.000.000	3,9	3,8	3,7	3,3	3,5	3,4	3,2	3,3	3,4	3,4	3,0	3,5	3,6	3,0
2.000.000	2,6	2,5	2,5	2,2	2,3	2,3	2,2	2,3	2,3	2,3	2,1	2,4	2,4	2,0
3.000.000	2,1	2,0	2,0	1,7	1,9	1,8	1,8	1,8	1,8	1,8	1,7	1,9	1,9	1,6
4.000.000	1,8	1,7	1,7	1,5	1,6	1,6	1,5	1,6	1,5	1,5	1,5	1,6	1,7	1,3
5.000.000	1,5	1,5	1,5	1,3	1,4	1,4	1,3	1,4	1,3	1,3	-	1,4	1,5	1,2
7.500.000	1,2	1,2	1,2	1,0	1,1	1,1	1,1	1,1	1,0	1,0	-	1,1	1,2	0,9
10.000.000	1,0	1,0	1,0	0,9	0,9	0,9	0,9	-	0,9	-	-	1,0	1,0	0,8
15.000.000	0,8	0,8	0,8	0,7	0,7	0,8	0,7	-	-	-	-	0,8	0,8	0,6
20.000.000	0,7	0,7	0,7	-	-	0,6	-	-	-	-	-	-	-	-
25.000.000	0,6	0,6	-	-	-	0,6	-	-	-	-	-	-	-	-

ESTIMATES	Piemonte	Valle d'Aosta - Vallée d'Aoste	Lombardia	Trentino-Alto Adige	Bolzano	Trento	Veneto	Friuli-Venezia Giulia	Liguria	Emilia-Romagna	Toscana	Umbria
20.000	25,7	5,7	38,6	12,5	11,4	11,7	32,2	18,6	20,0	31,7	26,9	16,6
30.000	20,6	4,4	30,5	9,8	9,0	9,1	25,5	14,7	15,9	24,9	21,3	13,0
40.000	17,6	3,7	25,8	8,3	7,6	7,7	21,6	12,4	13,5	21,0	18,1	10,9
50.000	15,5	3,3	22,6	7,2	6,7	6,7	19,0	10,9	11,9	18,4	15,9	9,6
60.000	14,0	2,9	20,4	6,5	6,0	6,0	17,1	9,8	10,7	16,5	14,4	8,6
70.000	12,9	2,7	18,6	5,9	5,5	5,5	15,6	9,0	9,8	15,1	13,2	7,8
80.000	12,0	2,5	17,2	5,5	5,0	5,0	14,5	8,3	9,1	14,0	12,2	7,2
90.000	11,2	2,3	16,1	5,1	4,7	4,7	13,5	7,7	8,5	13,0	11,4	6,7
100.000	10,6	2,1	15,1	4,8	4,4	4,4	12,7	7,3	8,0	12,2	10,7	6,3
200.000	7,2	1,4	10,1	3,2	2,9	2,9	8,5	4,9	5,4	8,1	7,2	4,2
300.000	5,8	-	8,0	2,5	2,3	2,3	6,8	3,8	4,3	6,4	5,7	3,3
400.000	4,9	-	6,7	2,1	2,0	1,9	5,7	3,2	3,7	5,4	4,9	2,7
500.000	4,4	-	5,9	-	1,7	1,7	5,0	2,9	3,2	4,7	4,3	2,4
750.000	3,5	-	4,7	-	-	1,3	4,0	2,3	2,6	3,7	3,4	1,9
1.000.000	3,0	-	3,9	-	-	-	3,4	1,9	2,2	3,1	2,9	1,6
2.000.000	2,0	-	2,6	-	-	-	2,3	1,3	1,5	2,1	1,9	-
3.000.000	1,6	-	2,1	-	-	-	1,8	-	-	1,6	1,5	-
4.000.000	1,4	-	1,8	-	-	-	1,5	-	-	1,4	1,3	-
5.000.000	1,2	-	1,5	-	-	-	1,3	-	-	1,2	-	-

Summary 4 *(continued)* – Interpolated percentage values of relative sampling errors for estimates referring to individuals for Italy total, geographical macro-area, type of municipality and region – 2009

ESTIMATES	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
20.000	19,6	35,7	17,9	9,3	32,0	26,8	12,5	20,8	29,2	20,3
30.000	15,5	28,4	14,3	7,4	25,3	21,5	9,9	16,6	23,4	16,1
40.000	13,2	24,1	12,2	6,2	21,5	18,4	8,4	14,1	20,0	13,7
50.000	11,6	21,2	10,8	5,5	18,9	16,3	7,4	12,4	17,8	12,1
60.000	10,4	19,1	9,8	4,9	17,0	14,7	6,7	11,2	16,1	10,9
70.000	9,6	17,5	9,0	4,5	15,6	13,6	6,1	10,3	14,8	10,0
80.000	8,9	16,2	8,3	4,2	14,4	12,6	5,7	9,5	13,8	9,3
90.000	8,3	15,2	7,8	3,9	13,5	11,8	5,3	8,9	12,9	8,7
100.000	7,8	14,3	7,4	3,7	12,7	11,2	5,0	8,4	12,2	8,2
200.000	5,2	9,6	5,0	2,4	8,5	7,7	3,4	5,7	8,4	5,5
300.000	4,2	7,6	4,0	1,9	6,8	6,1	2,7	4,5	6,7	4,4
400.000	3,5	6,5	3,4	-	5,7	5,3	2,3	3,8	5,8	3,7
500.000	3,1	5,7	3,0	-	5,0	4,7	2,0	3,4	5,1	3,3
750.000	2,5	4,5	2,4	-	4,0	3,7	1,6	2,7	4,1	2,6
1.000.000	2,1	3,8	2,1	-	3,4	3,2	-	2,3	3,5	2,2
2.000.000	1,4	2,6	1,4	-	2,3	2,2	-	1,5	2,4	1,5
3.000.000	-	2,1	-	-	1,8	1,8	-	-	1,9	-
4.000.000	-	1,7	-	-	1,5	1,5	-	-	1,7	-
5.000.000	-	1,5	-	-	1,3	1,3	-	-	1,5	-

3.4 Examples of calculation of sampling errors

3.4.1 Examples relating to estimates of households

Example 1

In 2009, 389,000 households in Piedmont owned between 26 and 50 books.

We seek the level of estimate closest to 389,000 in the column corresponding to the Piedmont region in Summary 3.

The relative percentage error of the estimate under consideration, for Piedmont, is 5.3%.

The absolute error is thus: σ (389,000) = 0.053 x 389,000 = 20,617

and the extremes of the confidence interval are: 389,000 - (1.96 x 20,617) = 348,591
389,000 + (1.96 x 20,617) = 429,409

Example 2

Considering the previous estimate, more precise values can be obtained for the sampling error by using linear interpolation of the two consecutive levels of estimation between which the value of the estimate falls.

These levels are 300,000 and 400,000, which correspond to the percentage values 6.2% and 5.3%.

The relative error corresponding to 389,000 is

σ(389,000) = 6.2 - { [(6.2 – 5.3) / (400,000 – 300,000)] x (389,000 – 300,000) } = 5.4%

The corresponding absolute error is: 0.054 x 389,000= 21,006

and the extremes of the confidence interval are: 389,000 - (1.96 x 21,006) = 347,828
389,000 + (1.96 x 21,006) = 430,172

Example 3

The error can be calculated directly by means of the interpolating function

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp\left(a + b\log(\hat{Y})\right)}$$

whose parameters, which are provided in Summary 2 at the Piedmont row, are the following:

a = 7.646708797 b = -1.047541519

For $\hat{Y}$ = 389,000 we obtain

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp\left(7.646708797 - 1.047541519 \times \log(389{,}000)\right)} = 0.054.$$

The relative percentage error is thus 5.4% and the absolute error and confidence interval are calculated in exactly the same way as in examples 1 and 2.

3.4.2 Examples regarding estimates of individuals

Example 1

In 2009, 1,586,000 people in Italy in the 20-24 age class had read at least one book in the 12 months prior to the interview.

We seek the level of estimate closest to 1,586,000 in the first column of Summary 4 for the Italy total.

The relative percentage error of the estimate under consideration is 2.6 %.

The absolute error is thus: σ (1,586,000) = 0.026 x 1,586,000 = 41,236

and the extremes of the confidence interval are: 1,586,000 - (1.96 x 41,236) = 1,505,177

1,586,000 + (1.96 x 41,236) = 1,666,823

Example 2

Considering the previous estimate, more precise values can be obtained for the sampling error by using linear interpolation of the two consecutive levels of estimation between which the value of the estimate falls.

These levels are 1,000,000 and 2,000,000, which correspond to the percentage values 3.9% and 2.6%.

The relative error corresponding to 1,586,000 is

σ (1,586,000) = 3.9 - (3.9 – 2.6) / (2,000,000 – 1,000,000) x (1,586,000 – 1,000,000) = 3.1%

The corresponding absolute error is: 0.031 x 1,586,000 = 49,166
and the extremes of the confidence interval are: 1,586,000 - (1.96 x 49,166) = 1,489,635
1,586,000 + (1.96 x 49,166) = 1,682,365

Example 3

The error can be calculated directly by means of the interpolating function

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp\left(a + b\log(\hat{Y})\right)}$$

whose parameters, which are provided in Summary 2 at the Italy row, are the following:

$$a = 9.547620201 \qquad b = -1.159970875.$$

For $\hat{Y}$ = 1,586,000 we obtain

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp\left(9.547620201 - 1.159970875 \times \log(1{,}586{,}000)\right)} = 0.030$$

The relative percentage error is thus 3.0% and the absolute error and confidence interval are calculated in exactly the same way as in examples 1 and 2.