BANCA D'ITALIA
EUROSISTEMA

Statistics

Methods and Sources:
Methodological Notes

3 July 2017

For further information: statistiche@bancaditalia.it
www.bancaditalia.it/statistiche/index.html

Survey on industrial and service firms

General remarks

The Bank of Italy has conducted sample surveys of firms since 1972[1]. Prior to 1998 the survey only covered industrial processing firms with 50 or more workers. In recent years the field of observation has been extended to include, since 1999, all manufacturing firms as well as energy and extractive industries, and since 2001 firms with 20 to 49 workers. In 2002 a similar survey of non-financial private service firms[2] with 20 or more workers was begun.

In 2006 the survey was enlarged to include construction companies with 20 or more workers. Starting from 2013, construction firms with 10-19 workers have also been included in the sample, this way improving the survey's ability to understand the performance of smaller firms, which are prevalent in a sector characterized by small firm dimensions (in terms of workers), considerably more than industry and non-financial private services.

The sampling design was completely overhauled in 1987, with the adoption of a stratified design based on non-proportional selection probabilities. This has remained largely unchanged, although the overall number of firms sampled has gradually increased over the years. Horvitz-Thompson estimators are employed, with appropriate re-weighting for the treatment of outliers. Over the years, the robustness of the estimates, particularly of variables with a large percentage of extreme data, has been improved by resorting to winsorization techniques.

Until 2003 firms were classified into size groups according to their workforce at the end of the year. In 2004, it was decided to use the average annual workforce for reasons of uniformity with external sources.

Below is a description of various methodological aspects of the present survey[3].

1 This is the reference year of the survey, which is actually conducted in the early months of the following year.

2 The survey does not include financial intermediation (on which the Bank of Italy has a vast dataset collected for purposes of banking supervision and monetary policy) and insurance, general government, the school and health sectors, and other social and personal public services.

3 For further details the reader is referred to the first periodical publication of the survey, Banca d'Italia, 'Survey of Industrial and Service Firms – Year 2003', Supplements to the Statistical Bulletin – Sample Surveys, 55, 20 October 2005.

Composition of the population and the sample

The reference population (Table I1) is composed of firms whose registered head office is in Italy, having 20 or more workers and belonging to various branches of activity in industry (excluding construction) and non-financial private services, or having 10 or more workers and belonging to the construction sector. Starting from the 2010 survey, the NACE 2007 classification of economic activities has been adopted[4].

The sample of the most recent surveys has been composed of around 5,000 firms (about 3,000 industrial firms, 1000 service firms and 600 construction firms). Tables I1 and I2 in the Statistical Annex show the population sizes and the sampling fractions. The sampling fractions are higher for firms with 50 or more workers and firms in the South and Islands.

Table 1

Branches of economic activity

	NACE 2007 section	NACE 2007 divisions	Sectors of economic activity	Sector aggregations used in the tables
Industry excluding construction	**C**	10–12	Food products, beverages and tobacco	Other manufactures
		13–15	Textiles, clothing, leather and footwear	Textiles, clothing, leather and footwear
		19–22	Chemical, rubber and plastic products	Chemical, rubber and plastic products
		23	Non-metallic minerals	Other manufactures
		24–30; 33	Basic metals and engineering	Basic metals and engineering
		16–18; 31–32	Other manufactures (wood, pulp and other)	Other manufactures
	B	05–09	Mining and Quarrying	Energy and extractive industries
	D	35	Electricity supply	Energy and extractive industries
	E	36–39	Water supply	Energy and extractive industries
Construction	**F**	41–43	Construction	Construction
Private non-financial services	**G**	45–47	Wholesale and retail trade, repair services	Distribution, lodging and catering
	I	55–56	Lodging and catering	
	H	49–53	Transport and storage	Transport, storage and communication
	J	58–63	Information and communication services	
	L, M, N (a)	68–75; 77–82	Other services provided to enterprises and households	Other services provided to enterprises and households

(a) Includes: L = real-estate activities; M = Professional, scientific and technical activities; N = renting, travel agencies, support services to enterprises.

4 See Istat (2009), Metodi e Norme, no. 40.

Sample design

The survey adopts a one-stage stratified sample design. The strata are combinations of branch of activity (according to the classification in Table 1), size class (in terms of number of employees)[5] and region in which the firm's head office is located[6].

The sample size is determined in two stages. First, the number of units in the size classes is identified using the method known as optimum allocation to strata[7], which minimizes the variance of the sample means and variations of the main variables observed (employment, turnover and investments). Second, the number of units in each size class is divided among regions and branches of activity in proportion to the number of firms in the target population belonging to that stratum[8].

The firms are selected from the databases of INPS, Infocamere and other lists obtained independently by the Bank's branches in order to minimize the risk of under-coverage. The firms observed in the previous edition of the survey are always contacted again if they are still part of the target population[9], while those no longer wishing to take part are replaced with others in the same branch of activity and size class.

Data collection, the questionnaire and response behaviour

The data for a survey referring to one year are collected in the course of interviews conducted by the Bank of Italy's branches between February and April of the following year.

Since the 2010 edition all the survey data have been uploaded using a web application.

The questionnaires are composed of (i) a fixed part that contains general information on the firm and its structure, investments, employment, turnover, operating result, capacity utilization and financing[10]; (ii) a variable part that covers different themes each year, which are the subject of detailed cyclical or structural analysis. In order to hold the costs incurred by respondents down, some of the questions in the monographic section are asked to a random subset of firms.

The content and structure of the questionnaires are decided shortly before these are administered. New questions are tested by the Bank's branches by means of pilot questionnaires designed to assess whether they are easy to understand and whether the information is effectively available from the firms.

---

5 The size classes, which are based on the end-of-year workforce until 2003 and on the average annual workforce thereafter, are: 20-49(from 2013 the class of 10-19 employees is been added to the sample of construction), 50-99, 100-199, 200-499, 500-999, 1,000-4,999, and 5,000 and over employees. Firms in the last class form a self-representative (census) stratum.

6 Piedmont and Valle d'Aosta are regarded as a single region.

7 See for example W. G. Cochran, Sampling Techniques, New York, John Wiley & Sons, 1977.

8 The number of firms assigned for each region is then divided among the Bank's branches, which collect the data. In some areas of the country over-sampling may be performed where necessary for use in studies of the local economy.

9 Typical reasons for leaving the target population are change of activity and staff cutbacks to below the entry threshold. See Section A4.

10 Investment and turnover are not recorded for construction firms; they are replaced by the value of production (total and relating to public works).

In the case of employment, investments and turnover, information is requested for three periods: the year just ended (preliminary results), the previous year (final results) and the following year (forecasts). Table I3 provides the survey participation rates.

Data quality checks and imputation of missing data

The collected data are subjected to a system of quality checks. In particular, we check for the compatibility of values with the range allowed by the question, the time consistency of panel data, the balancing of certain figures and the presence of outliers. The data are verified at different levels, with each check being run on data that have passed the previous level.

The questionnaires are first checked by the Bank of Italy officers responsible for the interviews, who apply their technical skills and knowledge of the local economy to assess the quality of the data collected[11].

The data-entry procedure then rejects everything outside the defined range of the variable or incompatible with the internal consistency of the questionnaires. Data accepted by the procedure may still be outside certain thresholds based on past statistics or external information. In this case the data are highlighted for the attention of the interviewers, who check them and, if necessary, contact the firm for clarification. Confirmation is recorded in a special field. Only at this point is the checked data entered in the database.

The second set of checks uses editing techniques founded on statistical assumptions that can detect extreme observations with respect to the frequencies based on (natural or logarithmic scale) cross-section distributions.

A further quality check, called selective editing, produces a list of priorities for checking outliers according to their importance for the final estimate. The values of each variable are compared with the predicted value of a simple regression model. On the basis of this statistic a Taylor approximation is used to build a score for each firm according to the impact each value would have on the final estimate in the model. The higher the score (i.e. the greater the impact of the observed value on the final estimate according to the model), the more urgent it is to check the value of that variable for that firm[12]. A ranking drawn up with respect to a function that summarizes the scores for the individual variables integrates the process. The method improves the quality of the estimates while reducing the respondent burden in the final stage of data processing, because it is necessary to re-examine and possibly re-contact only the firms with a significant impact on the final estimates.

A treatment apart is reserved for data on firms affected by extraordinary events, such as mergers or splits. These firms are only included in the estimate if the data for final results, preliminary results and forecasts refer to a set of factories and workers that is homogeneous with the data collected in the same survey. The respondent ensures their homogeneity either by considering the extraordinary event to have taken place at the beginning of the year of the final results, or by pretending the event never occurred and reconstructing the data accordingly; if this is not possible, the firm is excluded from further

11 At this stage, the interviewers use a form containing the main variables provided by the firms taking part in the previous survey to run a preliminary check on data consistency.

12 See P. Battipaglia, 'Selective Editing to Increase Efficiency in Survey Data Processing. An Application to the Bank of Italy's Business Survey on Industrial Firms', *Irving Fisher Committee Bulletin,* 13, December 2002, 149-154.

processing. Although this practice may cause distortions in the estimates of levels (such as the total of investments), it does produce more stable estimates of changes and average values per employee, which are the main objective of the survey.

The firms taking part in the survey may have difficulty answering some of the questions. If the missing answer concerns the main variables, such as expected investment spending or turnover, the missing data are imputed.

In general, ratio estimators are used to impute data, setting the number of the firm's employees as denominator (since this information is always available, otherwise the firm is excluded from the survey) in order to capture the scale effect[13]. In some cases the firm's time series data are used for the reconstruction, in the form of individual effects. This method gives an estimate of a level per employee that is obtained by combining a general cross-section mean and an average based on the firm's time series. The levels at time t and t+1 are reconstructed in sequence, by calculating average changes in appropriate cells of homogeneous firms[14].

The percentage of imputed data is usually small for both preliminary and final results for the fundamental variables. A higher rate of non-response tends to occur with questions involving forecasts, particularly of investments. Even difficult questions belonging to monographic sections may be affected by a high partial nonresponse rate.

The weighting process

The weighting procedure is performed in two stages. In step one, the combinations of branch of activity and size class are used as strata. Each firm is assigned an initial weight, given by the ratio of number of firms in the stratum cell to number of firms in the sample. Let h be the general stratum cell and, within it, N_h the number of firms in the target population and n_h the sample size.[15] The first stage weight of each firm in stratum h is therefore:

$$(1) \qquad w_h^{(0)} = \frac{N_h}{n_h}$$

By construction, the sum of the weights of each cell therefore gives the size of the target population it contains.

In step two, post-stratification is performed using raking[16] to take into account also the geographical area k where the firm's head office is located. The initial weight is modified by an adjustment factor f_k, so that the final weights can be obtained:

13 For an analysis of the situations in which a ratio estimator is preferable to the mean see F. Cicchitelli et al., Il campionamento statistico, Bologna, Il Mulino Editore, 1994.

14 Other devices are also used when imputing data. For example, if a cell in which an average is to be calculated contains a very small number of firms, it is added to neighbouring cells according to size class or geographical area. Moreover, robust averages are calculated in the cells, limiting the influence of outliers in the reconstruction. In some cases specific solutions are adopted that exploit data collected in the questionnaire, as when there are arithmetic constraints between the variables or weaker links that nonetheless allow a reliable reconstruction of the missing datum. For instance, if a firm does not report the number of hirings during the year, this is calculated by adding the number of terminations to the difference between the workforce at the beginning and end of the year; the same method is used for the number of terminations.

15 The symbol *nh* indicates the actual sample size. This allows the weights to be implicitly corrected to take account of the total of missing responses.

16 Iterative proportional fitting (or raking) simultaneously aligns the sample weights to the distribution of certain characteristics known from outside sources. See for instance V. Verma, *Advanced Sampling Method: Manual for Statistical Trainers*, Statistical Institute for Asia and the Pacific, Tokyo, 2000, 6.13-6.21.

(2) $$w_{hk}^{(1)} = w_h^{(0)} f_k$$

the sum of which coincides, in the post-stratum cells, with the number of firms in the target population they contain.

Therefore the system of final weights does not take into account the complete combinations of sector, branch of activity and geographical area, partly because some of them may contain few sample units, and partly because such a system of weights might lead to overly variable estimates for some domains. Consequently, post-stratification is limited to reconstructing the combinations of:

a) North-West, North-East, Centre, South and Islands;

b) firms with 20 to 49 and 50 and over employees (for construction, firms with 10-19 workers are also separately included);

c) manufacturing, energy and extraction, and service firms (the latter divided by section; see Section 2).

At every survey the weights are recalculated according to the distribution of the population on the latest available date[17].

The sample estimates

For a generic variable x, the aggregate $(\mathcal{T})$ is estimated with a Horvitz-Thompson unbiased estimator of the total[18], given by:

(3) $$\hat{\mathcal{T}} = \sum_{h=1}^{L} \hat{\mathcal{T}}_h = \sum_{h=1}^{L} w_h \left(\sum_{i=1}^{n_h} x_{i,h} \right) = \sum_{h=1}^{L} N_h \bar{x}_h$$

where L is the total number of strata and $\bar{x}_h$ is the estimator of the sample mean for the *h-th* stratum.

The estimate of rates of change, for instance of turnover or investments, is obtained using as estimator the ratio of the sums of the levels for each firm, weighted with the inverse sampling fraction according to a formula of the following type:

(4) $$\hat{r}_{t,t-1} = \frac{\hat{\mathcal{T}}^t}{\hat{\mathcal{T}}^{t-1}}$$

where $\hat{\mathcal{T}}^t$ e $\hat{\mathcal{T}}^{t-1}$ are the estimators of the aggregate at time *t* and *t-1* respectively.

The levels of the aggregate that are set as numerator and denominator of the formula are collected from the firm in the reference year, even when that firm has already taken part in the previous edition of the survey[19].

17 The population of firms is that obtained from Istat, Archivio statistico delle imprese attive, 2012. Updates are published periodically for the population of about two years earlier. The estimates are revised periodically to take account of updates in the survey reference population.

18 See F. Cicchitelli et al., *Il campionamento statistico*, Bologna, Il Mulino, 1994.

19 The estimate of the trend of phenomena based on the data of a single survey has proved much more stable than the estimate obtained from a comparison of the values recorded in adjacent surveys, which sometimes reflect structural changes in the firms that are difficult to take into account, as well as problems of classification and measurement. These aspects are monitored more closely within a same questionnaire, leading to more accurate estimates of changes. However, this method does not take full account of the entry and exit of firms in the target population.

In the specific case of percentage changes in investment, robust estimation techniques have been used since the 1999 survey[20]. The empirical distribution of this variable is ridden with outliers, partly owing to the nature of the phenomenon of interest: measurement of the levels, the ratio of which gives the rate of change, is complicated by the simultaneous presence of factors such as the typical long-term nature of spending, uncertainty, classification errors, and other sources of measurement error.

Using the method known as 'type II Winsorization', the rates of change above and below the cut-offs fixed on the basis of the empirical distribution are squashed against the cut-off, in proportion to the sampling fraction, according to the following formula:

$$(5) \qquad y_i^{wins} = \begin{cases} f * y_i + (1-f) * J & if \quad y_i < J \\ f * y_i + (1-f) * K & if \quad y_i > K \\ y_i & otherwise \end{cases}$$

in which y_i^{wins} is the Winsorized rate, y_i the observed rate, *f* the sampling fraction, *J* and *K* respectively the lower and upper cut-offs[21]. This method has been used for turnover since 2005 and, only in the case of construction firms, for total production and output of public works since 2006.

In a few cases, the survey collects directly the rates of change in economic phenomena: this happens, for instance, with the rate of change of capacity utilization or of percentages (such as the percentage of hours worked overtime). In this case, the estimate for the whole population is calculated as an average of the individual rates of change, weighted with the inverse sampling fraction times the amount of the phenomenon (or, if unavailable, a proxy).

The deflators for the levels of investment and turnover are calculated as sector means of the individual deflators collected directly from firms, weighted with the product of the coefficient of the ratio to the population and the amount of turnover[22].

Estimating standard errors

Finding the analytical expressions of the variance of the estimators obtained from a non-proportional stratified sampling design, with weights adjusted to take account of post-stratified variables, can be a complex task[23] that suggests resorting to simulation methods able to take account of the original sample design[24].

---

20 On robust estimation techniques in general see for example D.F. Andrews et al., Robust estimates of location, Princeton, Princeton University Press, 1972 or D.C. Hoaglin et al. (eds.), Understanding robust and exploratory data, New York, John Wiley & Sons, 1983. A classic reference for the theory is P. J. Huber, Robust statistics, New York, John Wiley & Sons, 1981.

21 In the literature, changing the values based on (5) is called 'type II Winsorization'; when the sampling fraction is not taken into account it is called 'type I Winsorization': in the latter case, the values beyond the cut-off are completely squashed against it. For a detailed description of the method used to estimate the changes in investments see P. Battipaglia, 'Robust Estimates of Investments from the Bank of Italy's Business Survey', Statistics Research Report, London, London School of Economics, 2000. For computation of per capita investment and turnover, the winsorization is carried out directly on this variable.

22 For construction firms, total production at constant prices is calculated on the basis of Istat's latest deflator for construction, while the production of public works is calculated using Istat's updated deflator for 'other construction', which does not include housing.

23 See Chapter 7 in C. Särndal et al., Model Assisted Survey Sampling, New York, Springer-Verlag, 1992.

24 A classic reference is K. M. Wolter, Introduction to Variance Estimation, New York, Springer Verlag, 1985.

The values are estimated using the jack-knife method, which is particularly well-suited to take account of the imposed structure of the data due to the nature of the sampling design, while ensuring that appropriate asymptotic properties are maintained[25].

If T_n is the value of the estimator for a sample containing *n* units and $T_{n-1,i}$ is the value of the same estimator calculated for the sample in which the ith unit has been left out (*leave-one-out method*), we first calculate the 'pseudo-values' $\tilde{T}_{n,i}$ defined as:

$$(6) \qquad \tilde{T}_{n,i} = nT_n - (n-1)T_{n-1,i} \quad 1 \le i \le n$$

the jack-knife estimator of the variance of T_n is (Tukey, 1958)[26]:

$$(7) \qquad \hat{V}_{JACK}(T_n) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{T}_{n,i} - \frac{1}{n} \sum_{j=1}^{n} \tilde{T}_{n,j} \right)^2$$

In non-construction industry and services, the highest standard errors are found for the estimate of the variance of investments owing to their intrinsic variability (Table I4). The lowest standard errors occur in the estimate of changes in turnover and employment. In the domain analyses, for example by firm size class or geographical area, the estimates are less accurate than the estimates for the total sample, a circumstance that should be taken into account when analyzing the results.

The standard errors are also estimated for the forecasts of the main variables surveyed. These calculations take into account the fact that a fair portion of the data is affected by partial non-responses, which are imputed using the method described in Section 5. Multiple imputation is used, which entails replicating independently a given number of datasets containing the complete observations. Twenty-five bootstrap samples are extracted from the original sample, only for observations based on complete original data, maintaining the sampling design. The imputation of the forecasting variables is replicated after the design weights have been realigned with the post-stratified variables. If $\hat{t}$ is used to denote the estimator and m the number of replicated samples, the variance of $\hat{t}$ can be estimated by the following expression[27]:

$$(8) \qquad \hat{v}(\hat{t}) = \frac{1}{m} \sum_{j=1}^{m} \hat{v}_{p(s)}(\hat{t}_j^*) + \left(1 + \frac{1}{m}\right) \sum_{j=1}^{m} \frac{(\hat{t}_j^* - \hat{t}_{mi})^2}{m-1}, \text{ where } \hat{t}_{mi} = \frac{1}{m} \sum_{j=1}^{m} \hat{t}_j^*$$

The term $\hat{v}_{p(s)}(\hat{t}_j^*)$ indicates the variance estimated on the j-th sample replicated for the estimator $\hat{t}$, using the information from the sample plan *p(s)*. The first sum is the average within imputation variance, while the second sum, known as the between imputation variance, is interpreted as the variability produced by the imputation. Table I5 gives the results. Clearly, the forecasting data show a greater degree of variability than the final results.

25 This method is well suited to the purpose, as comparisons of its theoretical and empirical properties with those of other replication methods show. See, for example, J. Shao and D. Tu, The Jackknife and Bootstrap, New York, Springer, 1995, which contains a thorough comparative analysis of the two methods of estimating variance.

26 See Tukey, J. W. (1958). Bias and confidence in not quite large samples. Annals of Mathematical Statistics, 29, 614.

27 See, for example, Chapter 4 in H. Lehtonen and E. Pahkinen, Practical Methods for Design and Analysis of Complex Surveys, New York, Wiley, 2004.

Finally, Table I6 contains the standard errors for the forecasts and final results of the main variables recorded for construction firms. Here again, the forecasting data show greater variability.