Glossary

12

Abby Israëls and Sander Scholtus

Statistical Methods (201212)

Statistics Netherlands

The Hague/Heerlen, 2012

Explanation of symbols

.	data not available
*	provisional figure
**	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/'12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/'10– 2011/'12	crop year, financial year, etc. 2009/'10 to 2011/'12 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1876-0333

60165201212 X-37

Concept	Description
Accuracy	The extent to which the distribution of an estimator is concentrated around the true value of the estimated population parameter. The accuracy of an estimator is quantified by the mean square error. See also precision.
Additive weighting	See linear weighting.
Administrative data	Data that originate from the administrative process of an external party. Example: VAT data from the Tax Administration.
Allocation of sample size	Distributing a preselected total sample size across different strata in a stratified sample.
Allocation of sample units to interviewers	Scheduling interviews by space (addresses for CAPI, telephone numbers for CATI) and time (shifts).
ARGUS	Software for statistical disclosure control, developed at Statistics Netherlands. There are two versions: μ-ARGUS for disclosure control of microdata sets, and □-ARGUS for disclosure control of tables.
ARIMA model	Autoregressive Integrated Moving Average model, a method for seasonal adjustment.
Automatic coding	Coding by means of a computer program. All coding decisions are made by the computer.
Automatic editing	An umbrella term for editing methods in which the data are checked and corrected by a computer.
Automatic integration	A process step in macro integration, during which values are made consistent with respect to a set of restrictions.
Auxiliary variable	A variable that correlates with the target variable and is known for the entire population. Auxiliary variables can be used to improve the quality of estimates.
Bascula	Software developed at Statistics Netherlands for weighting respondent data to correct for sampling errors and non-response, by calibration to known population or sample totals.
Base register	A register that complies or works towards full compliance with the requirements for authentic status of the records. Base registers are at the foundation of the Dutch system of government registers.
Bias (of an estimator)	The expectation of the difference between an estimator and the true value of the population parameter, also called 'systematic error'. If the bias equals zero, then the estimator is called unbiased. Otherwise, it is called biased.
Calendar effect	The occurrence of calendar-related events, often with an irregular pattern, that influence a time series.
Calibration	A form of weighting for which the weights are determined in such a way that a set of known population or sample totals is reproduced exactly from the observed data.
CAPI	Computer Assisted Personal Interviewing: a type of survey for which an interviewer visits the respondent and conducts the survey by means of a computer-controlled questionnaire.
CASAQ	Computer Assisted Self-Administered Questionnaire: a type of survey for which the respondent completes a computer-controlled questionnaire without an interviewer, e.g. online. See also web survey.
CATI	Computer Assisted Telephone Interviewing: a type of survey for which an interviewer contacts the respondent by telephone and conducts the survey by means of a computer-controlled questionnaire.
CAWI	Computer Assisted Web Interviewing: a term that is often used instead of CASAQ, although this type of survey does not involve an interviewer. See also web survey.
Checking	See editing.
Cluster effect	The design effect of a cluster sample or two-stage sample in comparison with a simple random sample of the same size. The precision of estimates based on a cluster sample decreases if units within clusters are relatively homogeneous.
Cluster sample	A sampling design for which the population is assumed to consist of clusters of units. A sample is drawn from the clusters, and all units within the sampled clusters are observed. See also two-stage sample.
Coding	An activity in the statistical process that attaches a code from a classification to a description, if this is possible.
Completing	Correcting undercoverage and overcoverage of the target population in a survey by means of a set of correction rules.
Confidence interval	An interval calculated from the sample data, which is known to contain the true value of an estimated population parameter with a predetermined

	probability (often 95%).
Correction	See editing.
Correction for measurement errors	Using a system of decision rules, resolving inconsistencies in data to improve the quality.
Coverage error	A deviation in an estimator caused by differences between the sampling frame and the target population, either in the form of undercoverage, or overcoverage.
Data supplier	A person or entity that provides data to a National Statistical Institute (NSI); see also respondent.
Deductive correction	An umbrella term for editing methods that use logical reasoning to derive corrections from the uncorrected data.
Deductive imputation	An umbrella term for imputation methods that use logical reasoning to derive imputed values in a deterministic manner.
Design effect (DEFF)	The ratio of the variance of the estimator under a chosen sampling design and the variance of the corresponding estimator under simple random sampling.
Deterministic matching	A matching method that does not use a stochastic model; counterpart of probabilistic matching.
Direct standardisation	A standardisation method in which mortality rates of one or several populations are weighted by a characteristic of one particular population (the 'standard population'), in order to make a fair comparison of these populations possible.
Disclosure	The act of retrieving information on identifiable individual persons, households, companies, or institutions from statistical data.
Donor imputation	An imputation method for which the imputed value is copied from a donor record that closely matches the recipient record on many characteristics.
Edit / Edit rule	A restriction on the values in a dataset; data that violate an edit rule either certainly contain an error (hard edit rule), or very probably contain an error (soft edit rule).
Editing	Detecting and correcting missing values and erroneous values in a dataset.
Efficiency of an estimator	The ratio of the variances of the direct estimator (Horvitz-Thompson estimator) and a chosen estimator under the sampling design. An estimator with an efficiency larger than 1 is more precise than the direct estimator.
Face-to-face interviewing	See CAPI and PAPI.
Harmonising	The act of making data on a particular concept from different sources comparable by means of a set of rules.
Imputation	(1) The process of determining and filling in new values for occurrences of missing or discarded values in a dataset. (2) A value filled in during the process described under (1).
Imputation class	A subpopulation for which imputation is carried out, without using any information from the rest of the population. Different imputation methods can be used for different imputation classes.
Imputed value	See imputation (2).
Imputing	See imputation (1).
Inclusion probability	For a sampling design without replacement, the probability that a particular unit from the population is drawn; this probability may vary between units, depending on the sampling design.
Index number	A figure that measures the ratio of a current value to a reference value, often expressed as a percentage.
Index reference period	The period for which an index number equals 100 by definition.
Indirect standardisation	A standardisation method for which an observed mortality rate is compared to the corresponding rate that is obtained by adopting the age-specific mortality rates of an external population.
Interactive coding	Coding performed manually, aided by an interactive computer program, which displays necessary background information to the coder. The program also processes the decisions made by the coder.
Interactive editing	An editing method for which a computer program checks the data and the data are corrected manually.
Interviewer administered	A form of data collection for which interviews are conducted by an interviewer; counterpart of self-administered.
Interviewer area	A region associated with an interviewer, where he or she conducts interviews.
Item non-response	The occurrence of wrongfully missing values in the data of a respondent.
Life table	A table that presents how many people from a group of, for example, 100,000 newborns (the radix) are still alive at certain ages.
Linear weighting	A form of weighting for which the weights come from a linear regression model that tries to explain the target variable of the survey using auxiliary

	variables.
Logical imputation	See deductive imputation.
Longitudinal imputation	An umbrella term for imputation methods that make use of observed values for the same variable at other times, either for the same object or for different objects. This can also be a form of multivariate imputation.
Macro editing	An umbrella term for editing methods that (initially) check the data on an aggregate level.
Macro integration	Integrating data from different sources on an aggregate level, to enable a coherent analysis of the data, and to increase the accuracy of estimates.
Mail survey	A survey for which a self-administered paper questionnaire is sent to and from respondents by mail.
Manual coding	Coding performed manually, without the assistance of a computer program.
Manual editing	See interactive editing.
Matching	The process of bringing together records of data on units from two different datasets, based on (nearly) identical features (matching keys).
Matching key	A set of key variables that occur in two or more datasets that have to be matched, which is used to relate records from different sets to each other.
Mean square error of an estimator	The expected value of the squared difference between the estimator and the true value of the population parameter. It can be shown that the mean square error is equal to the sum of the variance and the squared bias of the estimator.
Measurement error	(1) A deviation in an observed response from the true value. A measurement error may be caused by the questionnaire design, the interviewer, the mode of data collection, the respondent, or a combination of these. (2) A deviation in an estimate from the true population parameter, caused by (1).
Micro editing	An umbrella term for editing methods that check and correct the data on the level of individual units.
Micro integration	A method that matches data on individual statistical units from different sources, to obtain a combined dataset with better information. The quality of the data relates to validity, reliability and consistency. Micro-integration techniques include completing, harmonising and correction for measurement errors.
Mismatch	An error that occurs during matching when two records are incorrectly matched.
Missed match	An error that occurs during matching when two unmatched records should have been matched.
Mixed mode	A survey design for which largely identical questionnaires are used in different modes of data collection.
Mode	See mode of data collection.
Mode effect	The influence of the mode of data collection on the response given by the respondent.
Mode of data collection	The way in which information is obtained from units of observation in a survey. This refers both to the way the questions are presented to a respondent, as well as the way the answers are recorded. Examples of modes of data collection are CAPI, CATI, CASAQ, and PAPI.
Mortality rate	The average mortality probability for an age cohort between the current age and the next age.
Multiplicative weighting	A form of weighting for which the weights are obtained by multiplying relevant weight factors, determined in an iterative process. Also called raking or iterative proportional fitting.
Multivariate imputation	Imputing several missing values in a record.
Non-representative outlier	An outlier in the sample which is either incorrectly observed or unique in the population; counterpart of representative outlier.
Non-response	A unit of observation for which no data are obtained (this is unit non-response). See also item non-response.
Non-sampling error	The part of the total estimation error that would also have occurred in the case of complete observation; counterpart of sampling error.
Outlier	An observation that deviates strongly from the average pattern, and hence requires special attention, in particular during the editing and raising stages. See also representative outlier and non-representative outlier.
Overcoverage	The fact that a register or sampling frame contains units that do not belong to the target population, or multiple entries for the same unit.
Panel imputation	See longitudinal imputation.
Panel survey	A survey for which the same units of observation are approached several times in subsequent survey rounds, to measure a development over time.
PAPI	Paper-and-pencil Interview: an interview conducted by means of a paper questionnaire. The term is also commonly (mis)used to refer to a mail survey.
Parallel or simultaneous mixed mode survey	A mixed mode survey for which the sampled units are contacted through different modes of data collection from the outset; counterpart of sequential mixed mode survey.
Population parameter	A characteristic of the population to be estimated, e.g. the average score on a particular variable.
Post-stratification	An estimation method for which separate estimates for subpopulations (post-strata) are formed and subsequently combined into an estimate for the whole

	population. This method is only applicable if the number of units per stratum is known. Post-stratification can also be seen as a special case of weighting.
Precision	The extent to which the distribution of an estimator is concentrated around its expected value. The precision of an estimator is quantified by the variance or the standard error. See also accuracy.
Primary data	Data collected on behalf of a National Statistical Institute (NSI), that are used for the production of statistics, and for which the NSI has defined the conceptual metadata and process metadata; counterpart of secondary data.
Primary data collection	The act of obtaining primary data.
Primary sensitive cell	A table cell that is not safe according to the rules for statistical disclosure control.
Probabilistic matching	A matching method that uses a stochastic model. Two records may be matched even when their scores on the matching key are not identical. Records with different scores may belong to the same unit, because 1. measurement or processing errors occur in the data; 2. the two datasets have been observed at different times; 3. the definition of the key variables differs between the two datasets. Counterpart of deterministic matching.
Proxy interview	An interview in which the survey unit does not respond him or herself; another person (e.g. a household member) answers the questions on his/her behalf.
Questionnaire design	Producing the texts, structure and layout of a questionnaire.
Radix	The size of the population of newborns in a life table. At Statistics Netherlands, this size is usually set at 100,000.
Raising	A form of weighting for which the sum of the individual weights equals the size of the population.
Ratio estimator	An estimator that uses one quantitative auxiliary variable. This estimator has a high accuracy if there is little variation in the ratio of the target variable and the auxiliary variable.
Register	A collection of data recorded and maintained in a structured way.
Reminder strategy	A strategy for handling non-response during the period of data collection. It concerns instructions to interviewers, and also – particularly for business statistics – a strategy for sending reminder letters to non-respondents.
Repeated weighting	A method that repeatedly assigns raising weights to respondents, by adjusting the old weights to (possibly estimated) population totals. This adjustment – often cosmetic – leads to estimates that are consistent with previously published figures or register totals.
Representative outlier	An outlier in the sample, for which it is assumed that it has been observed correctly and that similar units exist in the non-sampled part of the population; counterpart of non-representative outlier.
Respondent	A unit of observation that answers the questions on the questionnaire. In business surveys, the business unit is considered as the respondent, but the actual reporting may be done by accountants or administrative offices.
Sample	A subpopulation used in a survey to estimate parameters of the whole population.
Sampling design	A prescription for drawing a sample from the sampling frame, which allows the calculation of a (positive) drawing probability for every possible sample. In the case of sampling without replacement, the design is usually defined by attributing an inclusion probability to every unit in the sampling frame.
Sampling error	The part of the total estimation error that can be attributed to the fact that data are only available for a sample from the population; counterpart of non-sampling error.
Sampling frame	The frame from which a sample is drawn, i.e. an administrative description - not necessarily perfect - of the target population.
Seasonal adjustment	The act of correcting a time series for seasonal patterns.
Seasonal pattern	A set of periodically (e.g. annually) recurring upward and downward movements in a time series.
Secondary data	Data collected by an authority outside a National Statistical Institute (NSI), that are used for the production of statistics, and for which the NSI has not defined the conceptual metadata and process metadata; counterpart of primary data.
Secondary data collection	The acquisition of secondary data.
Secondary sensitive cell	A table cell that is safe according to the rules for statistical disclosure control, but has to be suppressed in order to protect primary sensitive cells.
Selective editing	An umbrella term for methods that select records that are likely to contain influential errors for interactive editing.
Self-administered	A form of data collection for which respondents fill in questionnaires without the aid of an interviewer; counterpart of interviewer administered.
Self-completion; Self-	Mode of data collection for which the questionnaire is filled in by the respondent him or herself, e.g. in the case of sensitive questions.

report	
Sequential mixed mode survey	A mixed mode survey for which all sampled units are initially contacted through one mode of data collection, and another mode is used to contact non-respondents from the first wave; counterpart of parallel mixed mode survey.
Simple random sample	A sampling design for which all units in the sampling frame have an equal probability to be drawn. In the case of sampling without replacement, every subset of *n* units from the sampling frame must have the same probability to be drawn (where *n* denotes the sample size). In the case of sampling with replacement, the sampled units must be drawn independently.
SLICE	A software package for automatic editing, developed at Statistics Netherlands.
Small area	A publication cell for which the quantity of observed data is so small that an estimate based only on these data would be too inaccurate.
Small area estimation	Methods that can produce accurate estimates for small areas. They are based on a model that relates different areas to each other, so that data from different areas contribute to the estimate for a particular small area.
Standardisation	(1) Using a linear transformation to obtain a variable with mean 0 and standard deviation 1. (2) The act of correcting aggregate figures for the influence of background features; see direct standardisation and indirect standardisation. (3) The act of making concepts uniform.
Statistical disclosure control	Protecting statistical information in such a way that the risk of disclosure remains within preset bounds.
Statistical unit	A type of unit for which survey results are published.
Stratified sample	A sampling design in which the sampling frame is divided into subpopulations (strata), and a sample is drawn from each stratum independently.
Survey	In a survey, selected units of observation are studied in a standardised way.
Survey population	The population of sampling units for which data can potentially be obtained. The survey population may differ from the intended target population, e.g. due to coverage errors.
Survey unit	The object about which data are collected, non-respondents included.
Survival model	A model for analysing the length of periods between two events (e.g. birth and death).
Synthetic estimations	Parameter estimates based on a (linear) regression model, for which scores on the target variable are predicted from the model for all non-observed units.
Systematic sample	A sampling design for which units are selected by going through the sampling frame in a systematic way. A random starting point is drawn, and units are selected for which the distance from the starting point is a multiple of a fixed step size.
Target population	The intended population, i.e. the collection of statistical units for which the survey is to give results.
Target variable	An observed or derived variable that measures an aspect of a phenomenon of interest in a survey; one purpose of the survey will be to estimate population parameters for such a variable.
Time series	A collection of measurements of a particular variable at different time points.
TRAMO-SEATS	A method and software package for time series analysis and seasonal adjustment, developed by the National Bank of Spain.
Trend discontinuity	The phenomenon that a change in the design of a survey (e.g. in the questionnaire or the mode of data collection) disturbs a time series. A time series model can be used to try to correct the trend discontinuity.
Trimmed mean	The α-trimmed mean of a set of values is the mean of the subset found by leaving out the α/2 percent largest and the α/2 percent smallest values.
Two-stage sample	A sampling design for which the sample is drawn in two steps. In the first step, a sample of clusters of units is drawn. In the second step, a sample of units is drawn from each selected cluster. See also cluster sample.
Undercoverage	The fact that a register or sampling frame does not contain all units of the target population.
Unit non-response	See non-response.
Unit of observation	A unit for which data are collected in a survey. These units may be transformed to statistical units later in the statistical process.
Vivaldi	A shell around X-12-ARIMA, developed by Statistics Netherlands.
Web survey	A survey for which a self-administered questionnaire is accessible via a website, either to be filled in online or to be downloaded and filled in off-line by respondents.
Weighting	The act of assigning weights to survey respondents, which are then used to obtain estimates of population parameters by calculating weighted sums of observed values.

Winsorised mean	The α-winsorised mean of a set of values is the mean of the associated set of values found by replacing the α/2 percent largest and the α/2 percent smallest values by the largest and the smallest non-replaced value respectively.
X-12-ARIMA	A method and software package for seasonal adjustment, developed by the US Census Bureau.
XBRL	eXtensible Business Reporting Language: an open standard for exchanging financial data, based on XML.
XML	eXtensible Markup Language: a standard for defining formal markup languages to represent structured data in the form of flat text.

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Glossary				
1.0	15-09-2010	First Dutch version	Abby Israëls Sander Scholtus	Editorial board
English version: Glossary				
1.0E	28-08-2012	First English version	Abby Israëls Sander Scholtus	