## Annex 6. Statistical vocabulary

Terms included in this Annex are part of the statistical vocabulary developed by the International initiative carried on by the Bank for International Settlements (BIS), the European Central Bank (ECB), the Statistical Office of the European Communities (EUROSTAT), the International Monetary Fund (IMF), the Organization for Economic Co-operation and Development (OECD), the United Nations Statistics Division (UNSD), and the World Bank (WB) known as SDMX (Statistical Data and Metadata eXchange). Interested readers can access the complete set of 397 terms in http://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf

Definitions (highlighted in cursive) and the corresponding illustrative information are reproduced as in such document. Exceptionally, some terms include instead as illustrative information a text prepared by UNWTO (those terms are identified with an asterisk)

### Administrative data use (*)

*Administrative records are data collected for the purpose of carrying out various programs, for example, income tax collection. As such, the records are collected with a specific decision-taking purpose in mind, and so the identity of the unit corresponding to a given record is crucial. In contrast, in the case of statistical records, on the basis of which no action concerning an individual is intended or even allowed, the identity of individuals is of no interest once the database has been created.*

Administrative records present a number of advantages to a statistical agency or to analysts. Since these records already exist, costs of direct data collection and a further burden on respondents are avoided. They are usually available for the complete universe and, hence, for the most part unconstrained by sampling error considerations. Most importantly, they can be used in numerous ways in the production of statistical outputs. Examples of their uses include:

- the creation and maintenance of frames;
- the complete or partial (via record linkage) replacement of statistical collection;
- the editing, imputation and weighting of data from statistical collection; and
- the evaluation of statistical outputs.

Administrative datasets are not designed nor are the data collected with any specific statistical purposes in mind. The use of such data sources may require some compromises to be made with respect to population definition and coverage.

UNWTO is firmly convinced of the need to promote the use of administrative sources, among other reasons because it is impossible to base the development of the System of Tourism Statistics and the TSA on strictly statistical operations. And there are three areas on which attention should be focused: the information generated by traffic regulation authorities, fiscal sources and the “electronic fingerprints” left by tourists (toll motorways, bank cards, mobile telephones, use of the Internet to consult tourism websites, etc.).

### Bias

*An effect which deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error which may distort on any one occasion but balances out on the average.*

The bias of an estimator is the difference between its mathematical expectation and the true value it estimates. In the case it is zero, the estimator is said to be unbiased.

## Census

*A survey conducted on the full set of observation objects belonging to a given population or universe.*

A census is the complete enumeration of a population or groups at a point in time with respect to well defined characteristics: for example, Population, Production, Traffic on particular roads. In some connection the term is associated with the data collected rather than the extent of the collection so that the term sample census has a distinct meaning. The partial enumeration resulting from a failure to cover the whole population, as distinct from a designed sample enquiry, may be referred to as an "incomplete census". (The International Statistical Institute, "The Oxford Dictionary of Statistical Terms", edited by Yadolah Dodge, Oxford University Press 2003).

## Coherence

*Adequacy of statistics to be combined in different ways and for various uses.*

When originating from different sources, and in particular from statistical surveys using different methodology, statistics are often not completely identical, but show differences in results due to different approaches, classifications and methodological standards. There are several areas where the assessment of coherence is regularly conducted: between provisional and final statistics, between annual and short-term statistics, between statistics from the same socio-economic domain, and between survey statistics and national accounts.

The concept of coherence is closely related to the concept of comparability between statistical domains. Both coherence and comparability refer to a data set with respect to another. The difference between the two is that comparability refers to comparisons between statistics based on usually unrelated statistical populations and coherence refers to comparisons between statistics for the same or largely similar populations.

## Consistency

*Logical and numerical coherence*.

An estimator is called consistent if it converges in probability to its estimand as sample increases. Consistency over time, within datasets, and across datasets (often referred to as inter-sectoral consistency) is major aspects of consistency. In each, consistency in a looser sense carries the notion of "at least reconcilable". For example, if two series purporting to cover the same phenomena differ, the differences in time of recording, valuation, and coverage should be identified so that the series can be reconciled. Inconsistency over time refers to changes that lead to breaks in series stemming from, for example, changes in concepts, definitions, and methodology. Inconsistency within datasets may exist, for example, when two sides of an implied balancing statement-assets and liabilities or inflows and outflows-do not balance. Inconsistency across datasets may exist when, for example, exports and imports in the national accounts do not reconcile with exports and imports within the balance or payments.

## Data checking

*Activity whereby the correctness conditions of the data are verified. It also includes the specification of the type of error or of the condition not met, and the qualification of the data and their division into “error-free data” and “erroneous data”.*

It also includes the specification of the type of the error or condition not met, and the qualification of the data and its division into the "error free" and "erroneous data". Data checking may be aimed at detecting error-free data or at detecting erroneous data.

## Data collection (*)

*Data collection is the process of gathering data. Data may be observed, measured or collected by means of questionnaires, as in a survey or census response.*

In the case of border surveys the information should be collected when visitors are returning to their countries of origin. In airports, the ideal place for the survey is the departure lounge for international flights and, if possible, the Frequent Flier Lounges as well.

Broadly speaking, the following points should be taken into account so as to avoid a high incidence of non-response:

- the ease with which contact can be established with a potential respondent (it may, for instance, be difficult to attract the attention of a person collecting his luggage);
- the place chosen for the interview (a good example being the flight departure lounge);
- the ease with which the target respondents can be located (sometimes it is difficult to determine specific countries of residence);
- the language in which the survey is carried out (for border surveys aimed at non-residents not only is the translation of the questionnaire particularly important but also the language in which the interview is conducted);
- the use of incentives (some countries take the extreme step of paying respondents who return completed questionnaires);
- the extent of cooperation from officers responsible for border traffic (their involvement in the survey may be decisive for achieving specific response levels).

## Data confrontation

*The process of comparing data that has generally been derived from different surveys or other sources, especially those of different frequencies, in order to assess and possibly improve their coherency, and identify the reasons for any differences.*

Such data may not be coherent for a number of reasons including the use of different data item definitions, classifications, scope, reference period, etc.

## Data processing

*Data processing is the operation performed on data by the organization, institute, agency, etc, responsible for undertaking the collection, tabulation, manipulation and preparation of data and metadata output.*

As with surveys of other types, when publishing the findings of a border survey it would be desirable for them to be accompanied by:

- a set of sample counters, as would be the case of the number of questionnaires used and how many of them provide full information for estimating tourist expenditure (this would serve to determine whether the response rate is reasonable or if it could be increased by a different kind of initiative in respect of the interview);

- the values corresponding to visitors and expenditure according to the associated characteristics (organization of the trip, country of residence, length of stay, purpose of the visit, type of accommodation used, etc.);

- a basic set of indicators (average length of stay, extent to which private accommodation is used, etc.).

Special care should be taken to select the most appropriate reference period for the various types of characteristic researched. It does not always make sense to publish all the available data for each and every reference period of the survey, owing to either the small number of observations available, the inherent seasonality of tourism, or other considerations.

## Data reconciliation

*The process of adjusting data derived from two different sources to remove, or at least reduce, the impact of differences identified.*

Editing and reconciliation may involve fixing errors or adopting alternative sources and methods that are aimed at improving the process of reviewing or understanding data.

This may entail the reconciliation of stocks and transactions data; reconciliation of reported data with money and banking statistics, custodian data; differences with partner data or preshipment inspection data; the treatment of differences between GDP compiled for the production approach and GDP compiled from the expenditure approach. It is a special kind of editing done after initial compilation.

## Documentation (*)

*Processes and procedures for imputation, weighting, confidentiality and suppression rules, outlier treatment and data capture should be fully documented by the survey provider. Such documentation should be made available to at least the body financing the survey.*

Because tourism statistics include a wide range of data produced by different types of institution (at both national and international levels), there is a need for standards to be observed in the presentation of *metadata* (the term used for documentation of the coverage, temporal reference, distribution and a whole series of other technical characteristics of the data collection methodology applied).

The goal of documentation is to provide a complete, unambiguous and multi-purpose record of the survey, including the data produced from the survey. (The term *survey* is used generically to cover any activity that collects or acquires statistical data.) Providing documentation that is up to date, well organized, easily retrievable, concise and precise should be one of the objectives of any type of initiative in this regard.

## Estimation

*Estimation is concerned with inference about the numerical value of unknown population values from incomplete data such as a sample. If a single figure is calculated for each unknown parameter the process is called "point estimation". If an interval is calculated within which the parameter is likely, in some sense, to lie, the process is called "interval estimation".*

## Frame

*A list, map or other specification of the units which define a population to be completely enumerated or sampled.*

The frame consists of previously available descriptions of the objects or material related to the physical field in the form of maps, lists, directories, etc., from which sampling units may be constructed and a set of sampling units selected (Eurostat, "Assessment of Quality in Statistics: Glossary", Working Group, Luxembourg, October 2003). The frame may or may not contain information about the size or other supplementary information about the units, but should have enough details so that a unit, if included in the sample, may be located and taken up for inquiry. The nature of the frame exerts a considerable influence over the structure of a sample survey. It is rarely perfect, and may be inaccurate, incomplete, inadequately described, out of date or subject to some degree of duplication. Reasonable reliability in the frame is a desirable condition for the reliability of a sample survey based on it.

## Grossing up

*Activity aimed at transforming, based on statistical methodology, micro-data from samples into aggregate-level information representative for the target population.*

## Imputation (*)

*Imputation is a procedure for entering a value for a specific data item where the response is missing or unusable*

## Measurement error

*Error in reading, calculating or recording numerical value.*

Measurement errors occur when the response provided differs from the real value. Such errors may be attributable to the respondent, the interviewer, the questionnaire, the collection method or the respondent's record-keeping system. Errors may be random or they may result in a systematic bias if they are not random.

Measurement error in a survey response may result from respondents' confusion, ignorance, carelessness or dishonesty; error attributable to the interviewer, may be a consequence of poor or inadequate training, prior expectations regarding respondents' responses, or deliberate errors; and error attributable to the wording of the questions in the questionnaire, the order or context in which the questions are presented, and the method used to obtain the responses. Statistics Canada, "Statistics Canada Quality Guidelines", 4th edition, October 2003, page 5).

## Pilot survey

*The aim of a pilot survey is to test the questionnaire (pertinence of the questions, understanding of questions by those being interviewed, duration of the interview) and to check various potential sources for sampling and non- sampling errors: for instance, the place in which the surveys are carried out and the method used, the identification of any omitted answers and the reason for the omission, problems of communicating in various languages, translation, the mechanics of data collection, the organization of field work, etc.*

It is highly recommended that a pilot survey be carried out prior to launching a border survey geared to gauging visitor expenditure owing to the complexity and specific nature of certain aspects of research, including attracting the attention of potential respondents (particularly tricky in the case of road travel) and selecting them according to the sample design, choosing the best survey areas, organizing fieldwork at points that are geographically remote, and coping with problems arising on account of the type of entry point and means of transport used, securing the cooperation and permission of the authorities responsible for border traffic at the various points of entry into a given country, and coordinating their cooperation with the work carried out by those conducting the survey, determining the data collection mechanisms to be used (questionnaire completed by the respondents themselves as opposed to one-to-one interviews) and, finally, coping with the difficulties inherent in the collection of data whose quantitative nature presents an additional problem (e.g. main items of expenditure).

## Probability sample

A *sample selected by a method based on the theory of probability (random process), that is, by a method involving knowledge of the likelihood of any nit being selected.*

## Reference period

*The period of time or point in time to which the measured observation is intended to refer.*

In many cases, the reference period and time period will be identical, but there are also cases where they are different. This can happen if data are not available for the target reference period, but are available for a time period which is judged to be sufficiently close. For example, the reference period may be a calendar year, whereas data may only be available for a fiscal year. In such cases, “reference period” should refer to the target reference period rather than the actual time period of the data. The difference between target and actual reference period can be highlighted in a free text note.

## Response and non-response

*Response and non-response to various elements of a survey entail potential errors:*

*Response errors may be defined as those arising from the interviewing process. Such errors may be due to a number of circumstances, such as the following:*

- Inadequate concepts or questions;
- Inadequate training;
- Interviewer failures;
- Respondent failures.

Response error may result from the failure of the respondent to report the correct value (respondent error), the failure of the interviewer to record the value reported correctly (interviewer error), or the failure of the instrument to measure the value correctly (instrument error). (United States Federal Committee on Statistical Methodology, "Statistical Policy Working Paper 15: Quality in Establishment Surveys", Washington D.C., July 1988, page 57).

Non-response errors occur when the survey elicits no response to one, or possibly all, of the questions.

## Sample

*A subset of a frame where elements are selected based on a process with a known probability of selection.*

## Sample survey

*A survey which is carried out using a sampling method.*
In sample survey only a portion, and not the whole population is surveyed.

## Sampling error

> *That part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a subset of the population is enumerated.*
> Sampling errors are distinct from errors due to imperfect selection, bias in response or estimation, errors of observation and recording, etc. For probability sampling, the random variation due to sampling can be calculated. For non-probability sampling, random errors cannot be calculated without reference to some kind of model. The totality of sampling errors in all possible samples of the same size generates the sampling distribution of the statistic which is being used to estimate the parent value.

## Seasonal adjustment (*)

*Seasonal adjustment is a statistical technique to remove the effects of seasonal calendar influences on a series. Seasonal effects usually reflect the influence of the seasons themselves, either directly or through production series related to them, or social conventions. Other types of calendar variation occur as a result of influences such as number of days in the calendar period, the accounting or recording practices adopted or the incidence of moving holidays (such as Easter).*

The seasonal adjustment or deseasonalization of a series is the process whereby the seasonal variations of a series are estimated and eliminated. The variations reflect short-term seasonal fluctuations occurring with certain regularity (seasonality) that have a direct effect on most travel movements at certain periods every year, owing to such fundamental factors as holiday periods and weather patterns.

The seasonal component of a tourism series is extremely important, because time plays a more prominent role here than does any other component.

The advantage of a deseasonalized series is that it reveals the underlying trend-cycle movements, thereby facilitating data analysis.

The possibility of removing the seasonal component of a series may prove fundamental for analysing the trend in, for example, tourist arrivals throughout the year, irrespective of moveable feasts such as Easter or the enormous importance traditionally attached to the summer months compared to those of the low season.

Because tourism series covering periods of less than a year usually display a clear seasonal trend, their analysis yields information that is seriously affected by seasonality, and it is consequently advisable to smooth the series by making a seasonal adjustment.

In any case, whenever seasonally adjusted visitor arrival series are published together with trend-cycle series it is important to ensure that the estimations of both are consistent with each other.

One of the statistical tools most commonly used for calculating the seasonal components of a series are the ARIMA models. Some programs automatically incorporate specific treatments for certain effects, for instance the one known as the “Easter factor”, owing to the impact, mentioned earlier, of the moveable feast of Easter on a country’s visitor arrival series from year to year.

One factor to be borne in mind when seeking to deseasonalize a series is the effect on that series of incorporating additional data, because each new piece of information added at the end of the series changes the previous estimation. It is therefore advisable to try to confine any revisions to those that will considerably improve on the previous estimations so as to keep them to a minimum.

## Standard classification

*Classifications that follow prescribed rules and are generally recommended and accepted.*

Standard classifications aim to ensure that information is classified consistently regardless of the collection, source, point of time etc.

In the international context, standard classifications include ISIC, ISCO, CPC, NACE, etc. Many national statistical systems also have their own versions of standard classifications, which in the main are consistent with international standard classifications, though modified to meet national circumstances.

## Statistical error

*The unknown difference between the retained value and the true value*.

It is immediately associated with accuracy since accuracy is used to mean "the inverse of the total error, including bias and variance" (Kish L., "Survey Sampling", John Wiley, New York 1965). The larger the error, the lower the accuracy.

## Statistical indicator

*A data element that represents statistical data for a specified time, place, and other characteristics, and is corrected for at least one dimension (usually size) to allow for meaningful comparisons*.

A simple aggregation such as the number of accidents, total income or women Members of Parliament, is not in itself an indicator, as it is not comparable between populations. However, if these values are standardized, e.g. number of accidents per thousand of population, average income, or women Members of Parliament as a percentage of the total, the result meets the criteria for an indicator.

## Statistical metadata

*Data about statistical data.*

Statistical metadata comprise data and other documentation that describe objects in a formalised way (Economic Commission for Europe of the United Nations (UNECE), "Terminology on Statistical Metadata", Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf). They provide information on data and about processes of producing and using data. Statistical metadata describe statistical data and - to some extent - processes and tools involved in the production and usage of statistical data (UNECE, "Guidelines for the Modelling of Statistical Data and Metadata", 1995).

Statistical metadata can be classified in various ways, but there is a clear high-level distinction between the metadata needed to search for and display data (Structural metadata) and the metadata that give more information on definitions, methodologies, processes and quality (Reference metadata).

## Survey

*An investigation about the characteristics of a given population by means of collecting data from a sample of that population and estimating their characteristics through the systematic use of statistical methodology.*

Included are:

- censuses, which attempt to collect data from all members of a population;
- sample surveys, in which data are collected from a (usually random) sample of population members.

Surveys can be unique in time or repeated with regular or irregular periodicity. A single wave of a repeated survey is called survey instance.

A wider definition under which the term survey covers any activity that collects or acquires statistical data (including censuses, sample surveys, the collection of data from administrative records and derived statistical activities) has also been proposed. (see Statistics Canada, "Statistics Canada Quality Guidelines", 4th edition, October 2003, page 7, available at http://www.statcan.ca:8096/bsolc/english/bsolc?catno=12-539-X&CHROPG=1).