Singapore Department of Statistics

Statistical Best Practices

Contents

0. Statistical Best Practices Pre-requisites

The organisation should provide ample governance for the collection and safeguarding of data.

Strongly Recommended

S0.1.1 Establish primary responsibility and authority for the agency to collect, process and disseminate statistics. This arrangement can be in the form of a law, such as a statistical law, or other formal provision.

S0.1.2 Safeguard the confidentiality of individually-identifiable data through a law or other formal provision.

S0.1.3 Ensure that the staff have been equipped with the necessary skills to perform the required tasks.

1. Planning and Preparation

1.1 Initial planning

The first task in planning a statistical activity is to specify the objectives. To meet these objectives, agencies may use existing statistical or administrative data, re-design existing surveys, undertake new surveys, or a combination of these activities. Conducting an initial planning will aid the agency in determining the best course of action.

Strongly Recommended

S1.1.1 State the objectives of the statistical activity.

S1.1.2 Study relevant national standards, good practices, and recommendations by international bodies such as the United Nations (UN) or other National Statistical Organisations (NSOs). Adopting or adapting them helps to ensure that data are collected using sound methodologies and cost effective procedures, and are of good quality and internationally comparable.

S1.1.3 Examine existing data sources to explore if additional tabulations or more advanced statistical techniques such as modelling can be applied to adequately approximate the required data variables instead of undertaking a new survey.

1.2 Planning proposal

When it is determined that the necessary data are to be collected through conducting a survey or through administrative sources, agencies should draw up a detailed plan for the new project and consult relevant parties such as the data source agencies. Planning is crucial to managing the cost effectiveness of a project and assuring the quality of the final output. Increasingly, the use of data from administrative sources is preferred as it minimises compliance costs for businesses, households and individuals.

Strongly Recommended

S1.2.1 Draw up a timeline of key activities, which includes user consultation on data items, questionnaire design, data collection, data processing, data dissemination and publicity campaign.

S1.2.2 Estimate the budget and manpower resources required for the statistical activity.

S1.2.3 Consult data users to review the usefulness of existing statistics and identify data requirements.

S1.2.4 Adopt the methodology that best suits the purpose, stating the reasons for its use. This includes the proposed target population, target response rate, sample frame, sample design, sample size, effective sample size and other methodological issues.

S1.2.5 Establish procedures to promote consistency of methods and results when more than one data producing agency is involved in producing parts of the statistics.

S1.2.6 Follow national and international practices in the use of key statistical concepts and definitions.

S1.2.7 Document the planned data collection/acquisition procedures such as frequency of data collection/acquisition, primary method(s) of collection and, in the case of surveys, methods for achieving acceptable response rates.

S1.2.8 Design information systems for data collection/acquisition to prevent data loss or modification due to system failures or human errors during data capture, transmission and handling.

S1.2.9 Include validation checks in the data collection/acquisition system and process to detect any omissions, unnecessary and erroneous entries. Validation checks that are online should be kept simple so that the performance (i.e. response time) of the system will not be significantly affected.

S1.2.10 Conduct regular reviews of the entire process under which the data are collected/acquired, processed and used.

Recommended

R1.2.1 Work with DOS to review new key statistical concepts, definitions and classifications, where necessary.

R1.2.2 Choose an appropriate data collection method that achieves the data quality objectives such as timeliness and accuracy. This can include telephone and face-to-face interviews.

R1.2.3 Consider adopting new technology that improves the efficiency of data collection and the quality of data collected. Advances in communications and computing technology offer opportunities to reduce costs and respondent burdens. For example, electronic handheld devices in face-to-face interviews help to reduce errors when filling up the questionnaire.

R1.2.4 Outline a plan for quality assurance during each phase of the survey process for monitoring and assessing the performance during implementation. The plan should include contingencies to modify the survey procedures if design parameters are unlikely to be met.

1.3 Resources

Resources refer to the funding, facilities, material, staff and other assets available to the organisation necessary for its effective operation. Ensuring that sufficient resources are deployed for the statistical activity is critical to ensuring the success of the activity.

Strongly Recommended

S1.3.1 Allocate sufficient resources to exploit the potential of computing technology for collecting, compiling and disseminating the statistics.

S1.3.2 Use/Adapt appropriate software to perform tasks for compiling and analysing the statistics.

S1.3.3 Distribute hardware and software adequately to facilitate the efficient collection and processing of data, and management of databases.

S1.3.4 Provide adequate protection for IT systems. For mission critical systems, this includes provision of emergency back-up systems for retrieval and updates of data in the event of natural disasters, accidents and other unusual events.

1.4 **Training**

Training involves a variety of approaches which develop and equip survey officers, interviewers and supervisors with the skills and knowledge, so as to minimise the errors in data capture and data processing, and reduce non-response. This includes formal classroom training as well as on-the-job training.

Training can be conducted in various ways, such as courses and sharing sessions. Offering a wide range of learning opportunities will better meet the training needs of the staff.

Apart from statistical training, staff can also benefit from courses on management, communication and presentation, information technology, customer services, career or personal development.

Strongly Recommended

S1.4.1 Provide basic training to the officers involved in data collection, including:

- Overview of the objectives and principles behind conducting the statistical activity
- Statutory obligations to the respondents (where relevant)
- Security measures to safeguard the data collected
- Definitions of the terms used in the statistical activity
- Interviewing techniques such as establishing trust between respondents and the statistical agency, so as to obtain their cooperation in providing inputs to the present and future surveys
- Techniques in dealing with difficult respondents
- Other relevant topics such as the organisation and management of data collection

S1.4.2 Plan in advance all trainings required. Monitor and evaluate the effectiveness of such trainings.

S1.4.3 Review and update training materials on a regular basis.

S1.4.4 Provide data capture and coding operators with appropriate training and the tools to capture and code the data accurately.

S1.4.5 Equip staff with the skills to use new technologies or methodologies in the editing process.

S1.4.6 Provide appropriate training in the methodology and compilation methods. This may include participation in seminars, courses and workshops arranged by regional and international organisations to improve knowledge of statistical practices and provision of easy access to professional literature.

S1.4.7 Train all staff involved in the imputation process on the principles behind imputation, the different imputation methods and the corresponding procedures.

S1.4.8 Train all staff involved in the seasonal adjustment process on the principles behind seasonal adjustment, the different adjustment methods and the corresponding procedures.

Recommended

R1.4.1 Provide interviewers with a script to follow. This helps to ensure that a uniform answer is given to respondents by all interviewers. Interviewers should practise using the script to ensure that the interviewing process is as natural as possible.

R1.4.2 Provide each interviewer with a document containing all necessary procedures and instructions in handling respondents.

R1.4.3 Instruct interviewers to be punctual and appropriately dressed for all interview appointments and provide advance notice to the respondent if they are unable to keep to the appointment.

R1.4.4 Instruct interviewers to take precautions to prevent contents of the survey from being disclosed during fieldwork. For example, an interviewer should not switch on issued computer devices in public places where there is a chance that the contents can be disclosed accidentally.

R1.4.5 Complete telephone surveys or fieldwork which do not require follow-up in a single appointment wherever possible. If not, remember to fix another appointment to continue the survey.

R1.4.6 Arrange interviewers so that they can converse in a language that the respondent is familiar with.

R1.4.7 Arrange for new interviewers to observe and understudy the more experienced interviewers for the first few visits/calls.

R1.4.8 Equip supervisors with the necessary management and appraisal skills.

R1.4.9 Put up appraisal reports on interviewers to keep track of their performance and identify any need for further training.

R1.4.10 Review survey returns with interviewers and look for possible errors that can be corrected on the ground.

1.5 Coverage and Frame

Once the target population has been determined, a sampling frame should be identified to select the survey population. The chosen sampling frame should conform to the target population and contain minimal under- and over- coverage.

Strongly Recommended

S1.5.1 Assess the most appropriate sampling frame to be used for the survey, taking into account the costs that may be incurred.

S1.5.2 Establish procedures to eliminate duplication, inconsistencies and errors in the sampling frame. For example, build in validation checks to reduce such errors.

S1.5.3 Assess the coverage of the frame and target population to determine any coverage errors.

Recommended

R1.5.1 Update the frames on a regular basis. Put in place a series of checks to preserve integrity of the frames and ensure that the data are of good quality. For example, build in error correction checks during the updating process to ensure data quality and integrity. Records which fail the validation or edit checks should be scrutinised and corrected. Consult source agencies, where applicable.

R1.5.2 Use the same frame for surveys with the same target population and reference period, wherever possible. This is to avoid inconsistencies and enable cross-survey estimates to be made.

R1.5.3 Consider using supplementary data from other sources and using sampling weights to minimise coverage error of the frame.

1.6 Sampling techniques

Sampling refers to the process of selecting units from a survey population. A well designed sample survey will provide users with accurate parameter estimates for the variables to be studied and reduce overall response burden.

The choice of sampling design has a direct impact on the estimation method and quality of the estimates. It should consider the quality of the sampling frame, the key variables to be estimated, the desired precision of the estimates, the resources and time frame for data collection and processing, and auxiliary variables that can be used for estimation.

Strongly Recommended

S1.6.1 Determine the appropriate sample size for the survey. This is dependent on relevant factors, including :

- desired level of precision for key estimates of the survey;
- sample design used for the survey;
- data collection method;
- expected non-response rate for the survey;
- whether additional information on elements of interest from the selected sample is required; and
- budget and timeframe.

S1.6.2 Conduct studies to evaluate possible sampling methods to determine the most appropriate method to be used for the survey, taking into account the data requirements to be met and method of data collection to be used for the survey.

S1.6.3 Evaluate the sample design after the completion of the survey. Compare the precision of estimates actually achieved against the planned precision during sample design and investigate the reasons for the difference to identify areas for future improvements.

S1.6.4 Conduct periodic reviews on the sample design used. Develop procedures to evaluate the quality of the sample frame and maintain the sample quality over time.

Recommended

R1.6.1 Improve the sampling efficiency using stratification, i.e. better quality of estimates with the same sample size. The choice of the stratification variables should consider the distribution and characteristics of the units in the population, availability of the stratification variables, relationship between the variables of interest and stratification variables, estimation method and desired precision of the estimates.

1.7 Questionnaire Design

Questionnaire design is one of the most critical stages in the entire survey planning process, as it has a major impact on respondent behaviour, interviewer performance, collection cost and accuracy of respondent's answers. These in turn affect the quality of data collected.

There are generally 2 types of questionnaire designs:

(i) Self-administered questionnaires, which are designed to allow respondents to complete the survey returns on their own with minimum assistance in mind; and

(ii) Interviewer-administered questionnaires, which are designed to facilitate the interviewers in capturing the information during interviews with respondents.

Good questionnaires translate the survey objectives accurately into easy to understand/administer questions that allow respondents to provide the information required for the survey. Relevance and accuracy are the two basic criteria needed when designing a questionnaire.

Strongly Recommended

S1.7.1 Document rationale for all questions, stating which survey data item each question is addressing. This is to prevent omission of any item and ensure that the question is able to draw out relevant information from respondents.

S1.7.2 State the title or subject of the survey in the introduction of the questionnaire. Also state the agency conducting or sponsoring the survey, the purpose of conducting the survey, the authority under which the survey is conducted, the confidentiality clause (if any), and request for the respondent's cooperation.

S1.7.3 Avoid statistical and technical terminology and jargon. Use language that is unambiguous and easy to understand by the target population.

S1.7.4 Design self-administered questionnaires to be user friendly, with logical flow and clear and concise instructions to guide respondents. Definitions shall be clearly highlighted.

S1.7.5 Build in editing checks for electronic questionnaires so that respondents can correct errors on the spot. This allows potential errors to be identified quickly.

Recommended

R1.7.1 Harmonise concepts and definitions with similar surveys as far as possible to facilitate data comparability over time.

R1.7.2 Encourage respondent to provide best estimates where precise responses cannot be provided.

R1.7.3 Determine whether survey should be conducted in several languages. Translate questionnaire only after the main version is finalised.

R1.7.4 Take into account the method of data collection. Electronic handheld devices and other electronic platforms have space and layout constraints. The design for internet submissions should also factor in response times.

1.8 Questionnaire Testing

Survey questionnaires have to be sufficiently and appropriately tested to identify any potential pitfalls that may be encountered so that the surveys collect the information they were designed to collect. Testing will help to identify ambiguous wording or ordering, errors in questionnaire layout and problems caused by the respondent's inability or unwillingness to answer the questions. The choice of the evaluation methods depend on the type and size of the survey, the survey content, utilization of previous survey questions or standard questions, the method of data collection, project schedule, and the budget and resources available.

Strongly Recommended

S1.8.1 Test the questionnaire at an early stage of its development. Make revisions to the questionnaire based on the findings and re-test the revised questionnaire.

Recommended

R1.8.1 Determine the appropriate method to test the questionnaire. The methods can include pilot tests or qualitative tests such as focus groups.

R1.8.2 Conduct a pilot test for new and large projects to fine-tune the questionnaire before its use in the main survey. Obtain feedback from staff involved to identify potential sources of response and non-response errors as well as areas where the questionnaire can be further improved.

R1.8.3 Check different language versions for surveys in different languages to ensure consistency.

1.9 Confidentiality

Data confidentiality refers to the legislation, regulations, methods and their application to safeguard the data given by providers. Statistical agencies using data from other sources need to be aware of the relevant legal provisions safeguarding the data. Staff should safeguard the confidentiality in the use and disclosure of information.

Strongly Recommended

S1.9.1 Ensure all officers, including temporary and contract staff, sign relevant legal confidentiality undertakings upon appointment. Officers must pledge not to disclose any individually identifiable data, even after they leave the service.

S1.9.2 Control access to confidential information on a need-to-know basis. Officers shall not discuss any confidential information obtained in the course of their work with any person other than their supervisors or the persons administering the statistical activity.

S1.9.3 Impose penalties for any wilful breaches of statistical confidentiality.

S1.9.4 Put in place the necessary technical and physical infrastructure and processes to meet the required security classification of the data. Have proper systems and procedures that grant and track all read/write access to datasets.

Recommended

R1.9.1 Be tactful when collecting data items that are perceived as sensitive (e.g. individual income and financial assets) in cases where a proxy is permitted to answer questions on behalf of a subject.

R1.9.2 Do not release data in a form that may disclose a respondent's identity if not permitted by the legislation.

R1.9.3 Use statistical disclosure control methods such as global recoding to transform data into a format that eliminates the exposure of a record or its attributes when data with risk of indirect identification of subjects involved is disseminated. Consideration should be given to the sensitivity and intended use of the data when selecting the appropriate method.

2. Data collection

After a detailed plan has been drawn up to collect the data, agencies may proceed with the data collection. The data can be obtained directly from individuals or organisations (via self-enumeration, telephone interviews or face-to-face interviews) or indirectly by extracting data from administrative sources or other sources. The method of collection would have been decided during the planning stage, taking into consideration factors like the quality of data to be collected and the cost of collection, including respondent burden.

2.1 Data collection procedures

Despite measures to maximise survey response, the number of fully completed questionnaires may fall below expectations. This may bias the survey results. It can also increase the sampling variance of estimates as the number of data points collected is lower than the original planned sample size. Proper procedures should be drawn up to encourage participation and improve the quality of data collected.

Strongly Recommended

S2.1.1 Inform respondents of the purpose of the survey, the due date for submission of returns and the appropriate channels for submission/feedback.

S2.1.2 Make efforts to reach the selected respondents. Participation by volunteers and convenience sampling shall not be used to replace non-respondents. Surveys based on those who respond voluntarily are likely to be biased towards those with strong opinions or with time on their hands while surveys based on convenience sample are likely to be biased towards those with certain characteristics.

S2.1.3 Explain to respondents that they have been selected using an established sampling method and the significance of their participation in the survey when requested. Inform respondents of the legal provisions on the confidentiality safeguards of information provided and their obligations, where relevant.

S2.1.4 Provide assistance to respondents in completing and submitting forms, e.g. providing a contact point.

S2.1.5 Monitor response rates closely and employ appropriate methods to encourage participation. For example, using a wider range of communications channels in disseminating the survey message and obtaining returns or feedback.

S2.1.6 Take appropriate and swift action to obtain information from non-respondents. Send reminder letters or call respondents if they do not submit returns after the specified due date. Record the reasons for non-response.

S2.1.7 Monitor data collection operations at regular intervals to ensure that all procedures are properly implemented.

Recommended

R2.1.1 Incentives can be considered to encourage higher response rates, especially for lengthy and complex surveys.

R2.1.2 Inform respondents of the survey date or period, rationale of study and type of data to be collected

R2.1.3 Contact the respondent or a responsible person of the household or organization at an appropriate time. For household surveys, unless specifically requested by respondents, the last fieldwork visit or telephone call of the day shall be made preferably before 9.30pm.

R2.1.4 Do not inconvenience respondents unnecessarily. Consolidate all incomplete entries and flagged responses before calling them to verify.

R2.1.5 Explore ways to improve the data collection process and response rates of recurrent surveys.

R2.1.6 Plan a follow-up strategy to prioritise between respondents. For example, large or influential units should be followed up first in business surveys and higher priority should be given to non-responding units where potential for non-response bias is high. A scoring system can be set up to aid in the prioritisation process.

R2.1.7 Establish agreements with data suppliers to provide data on a regular basis and provide updates on any data revisions, rebasing and breaks in data series.

2.2 Use of data from administrative sources and statistical databases

As administrative data have already been collected for specific purposes, using them does not increase the response burden for the public. The associated costs are also lower compared to conducting surveys. Data can also be obtained from existing statistical databases for the compilation of official statistics.

All administrative records and databases present some challenges in data quality. We need to consider the trade-offs in determining the usability of administrative records or databases in statistical production:

- Content;
- Coverage;
- Frequency;
- Accuracy;
- Timeliness;
- Cost; and
- Respondent burden.

Strongly Recommended

S2.2.1 Use source data which reasonably approximate the definitions, scope, classifications, valuation and time of recording required.

S2.2.2 Establish procedures to provide for the effective and timely flow of source data.

S2.2.3 Assess the accuracy of data from administrative sources and statistical databases regularly.

S2.2.4 Maintain contacts with other data producing agencies to better understand data requirements, avoid duplication of effort and take into account reporting burden. For example, discuss changes to administrative processes before they take place.

S2.2.5 When merging data from different sources:

- Evaluate content and coverage issues, bearing in mind that each database targets a different segment. Merging data from the different sources may result in inconsistency or record duplication.
- Establish code conversion tables for the data items obtained so that the codes can be mapped out and converted correctly, as different databases may adopt different codes and classifications.
- Take note of the reference periods for all the data items, as different databases may have different reference periods for the data collected.

S2.2.6 Assess and investigate statistical discrepancies and differences. Check with the source agency, as the database may contain outdated or missing information. For example, build in validation checks for specific data items obtained from the various agencies.

S2.2.7 Sieve out, verify and take appropriate action for rare occurrences like unique records or outliers (such as persons aged over 100 years old or unique country of birth), in consultation with the source agency to verify accuracy of data.

Recommended

R2.2.1 Study the processes by which administrative data are collected, including the scope and coverage of administrative data, to understand the data quality.

3. Data Processing

The term "data processing" encompasses a variety of activities, including:

- *administrative functions* (such as manpower and logistics planning);
- *post-collection processes* (such as coding, editing, imputation, data review and correction, data query, weighting, estimation and variance estimation, and application of disclosure limitation rules; and
- *support functions for collection technologies* (such as mail-out, check-in and follow-up).

This section provides the best practices specifically on the post-collection processes which directly affect data integrity and the results.

3.1 Data Capture and Coding

Data collected from surveys and obtained from administrative sources for statistical purposes need to be processed, and converted into an electronic format for subsequent processing and analyses. The data may need to be coded by assigning a pre-determined set of values to responses collected to facilitate compilation, aggregation and subsequent analyses.

As technology improves, data is increasingly captured directly into Information Systems with minimal human involvement. Advanced IT systems can also be developed to perform computer-assisted coding to increase productivity gains.

Strongly Recommended

S3.1.1 Design the data capture process to minimise costs and maximise timeliness and data accuracy. For example, develop appropriate IT systems to scan and capture recognised data for large surveys involving fieldwork to save time and costs.

S3.1.2 Design backup systems to prevent data loss or modifications due to system failures or human errors.

S3.1.3 Test the automated data capture systems prior to implementation to ensure reliability of the systems.

S3.1.4 Build and maintain reference files for automated coding systems.

S3.1.5 Apply consistent coding to all data items.

S3.1.6 Check survey data for accuracy. Conduct interview verification to reduce errors arising from interviewers falsifying survey responses.

S3.1.7 Establish procedures to adjust for under-coverage and population units which are out of scope. Compute grossing up factors scientifically, based on sample design.

Recommended

R3.1.1 Spell out common rules for different development teams to follow in large projects where there are different systems. This allows for greater synergies between systems.

R3.1.2 During data editing phase, gather information regarding data quality to improve future data capture and coding procedures.

3.2 Editing

Errors or inconsistencies may be introduced into the data during data collection (via surveys or administrative systems) or data entry. Data editing is a system of checks to detect potentially erroneous data. The editing process identifies inconsistent data and corrects for errors in structure and logic by checking for logical relationships and reasonableness of the responses.

Strongly Recommended

S3.2.1 Develop processes to minimise editing errors. For example, highlight doubtful cases for further examination.

S3.2.2 Validate data collected from main sources against other independent data sources where possible.

Recommended

R3.2.1 Prioritise editing, starting with more severe errors or high likelihood of influential errors.

R3.2.2 Use appropriate tools to detect outliers and provide information on how the individual responses affect the overall statistics. For example, graphical editing can be used.

R3.2.3 Monitor the editing process using indicators like editing ratios to analyse the efficiency of data processing.

R3.2.4 Learn from past experiences what type of logic and editing checks to set.

3.3 Imputation

Imputation is a process by which the missing data in a data set is replaced by known acceptable values. The process of imputation should be aimed at efficiently limiting the bias caused by missing values while maintaining consistency and objectivity in the data. However, it should not be used as a replacement for low/non response rates as it is important to strive for higher response rates to produce better quality data. The imputation process may be manual, automated, or a combination of both.

Strongly Recommended

S3.3.1 Adopt sound estimation techniques to adjust data for missing observations.

S3.3.2 Check the process using test data to detect and improve imputation method before the application to actual data.

S3.3.3 Flag all imputed values. Identify and document imputation methods and sources to facilitate monitoring and analyses.

Recommended

R3.3.1 Choose an imputation method that is stable and gives a good prediction of missing data. When choosing the imputation method, consider the type of data being imputed and the availability of historical and administrative data. In some surveys, it may be necessary to use more than one imputation method. This can be achieved through a hierarchy of imputation methods. The testing for such methods should take place at each level of the hierarchy.

R3.3.2 Monitor the imputation results using indicators like imputation ratio to study the impact of imputation.

3.4 Seasonal adjustment

A time series can generally be decomposed into three basic components:

a) a *trend-cycle* (TC) component, which is the combined long-term and growth cycle movement of the time series;

b) a *seasonal* (S) component, which is the systematic variations of the time series caused by seasonal changes, social customs or cultural festivals; and

c) an *irregular* (I) component, which comprises the random fluctuations of short-term movements of the time series.

Seasonal adjustment is the process of estimating and removing the seasonal component from a time series so as to show more clearly its underlying trend and short-term movements.

Strongly Recommended

S3.4.1 Conduct a thorough seasonal analysis to assess if seasonality is identifiable for the series intended to be adjusted. Seasonally adjust the series only if its seasonality is identifiable.

S3.4.2 Do not seasonally-adjust short time series of less than 3 years for monthly data and 4 years for quarterly data.

S3.4.3 Validate the seasonally adjusted series using the default diagnostics generated by the seasonal adjustment tool to check that the seasonally adjusted series are of acceptable quality.

Recommended

R3.4.1 Use basic graphical analysis to examine:

- Length of time series
- Presence of unusual values (zeros or outliers)
- Presence of possible breaks in the trend or seasonal pattern

- Decomposition (additive, multiplicative) scheme

R3.4.2 Use regression-based methods to estimate and remove the moving holiday and trading-day effects.

R3.4.3 Unadjusted outliers may lead to distortions in the seasonally adjusted series. Hence, the time series should be checked and pre-adjusted for outliers before the estimation and removal of seasonal variations. After the removal of seasonal variations, the outliers should be reinstated in the time series, so as to reflect the specific events (e.g. SARS) that have caused the outliers.

R3.4.4 Seasonally adjust aggregate series by either the direct method or the indirect method. In the direct method, the aggregate series is seasonally adjusted directly. Direct method is preferred if the component series show similar seasonality. In the indirect method, the seasonally adjusted aggregate series are derived by combining the seasonally adjusted component series. Indirect method is preferred if the component series show different seasonality.

4. Compilation, Analysis & Dissemination

4.1 Compilation and Analysis

Compilation of data is a process of condensing information by classifying and tabulating data into various groups or categories. This allows statistics to be produced according to a determined tabulation programme.

Data analysis involves the examination and interpretation of data. The basic steps comprise:

- Identifying issues
- Determining the availability of suitable data
- Examining the data
- Deciding which methods are appropriate to answer the questions of interest
- Applying the methods and evaluating the results
- Summarising and communicating the results

Strongly Recommended

S4.1.1 Adopt classifications which are used in the compilation of statistics that are in broad conformity with nationally/internationally recommended systems.

S4.1.2 Produce statistics that are consistent within the dataset. For example, adopt concepts and definitions for compiling quarterly estimates that are the same as those used to compile annual estimates.

S4.1.3 Produce statistics that are consistent or reconcilable over a reasonable period of time.

S4.1.4 Conduct reviews on the data prepared for dissemination to check for possible indirect disclosure of individually-identifiable data and prepare tables and outputs in a way that prevents disclosure.

S4.1.5 Undertake revision studies on a regular basis to assess the initial estimates against revised or final estimates and investigate the sources of errors, omissions and fluctuations in the data when discrepancies were found.

S4.1.6 Establish a regular and transparent schedule for data revisions. Clearly identify preliminary data from revised data.

Recommended

R4.1.1 Establish the objectives of the analysis.

R4.1.2 Determine the appropriate statistical techniques needed to achieve the objectives of the analysis.

R4.1.3 Consider the consistency and impact of data sources if more than one data source is used and integrate them into the analysis.

R4.1.4 Use summary statistics and graphs to facilitate understanding of the data and detect any unusual patterns (e.g. outliers, trend breaks).

R4.1.5 In general, use commercial software over non-commercial software as they have undergone more rigorous tests.

4.2 Dissemination

There are many channels of communication such as the internet, newspapers and publications. In all cases, the presentation should highlight the key message and relay the story behind the numbers in an informative and interesting way. The three most common forms of presentation for statistical data are textual, tabular and graphical.

Strongly Recommended

S4.2.1 Comply with the confidentiality requirements governing the data being disseminated.

S4.2.2 Create a system of checks to verify written language, accuracy of data and consistency of figures used in text, tables and charts before publishing.

S4.2.3 Present statistics that facilitates proper interpretation and meaningful comparison, e.g. through clarity of text, tables and charts.

S4.2.4 Disseminate statistics in formats that meet users' needs.

S4.2.5 Clearly identify the products derived from the statistical activity so that users are aware of the agency responsible for the products. For example, display the agency's name, logo and insignia in statistical publications. In the case of joint publications, identify the part attributable to the relevant agency. For example, clearly distinguish the statistics from policy interpretation.

S4.2.6 Provide information about the terms and conditions under which the statistics are compiled and disseminated, including safeguards on the confidentiality of data.

S4.2.7 Provide explanatory materials and briefings to minimise misinterpretation or misuse of statistics. For example, disseminate statistics in formats that facilitate the media's re-dissemination.

S4.2.8 Identify contact points for all statistical releases clearly. These contact points may include address, telephone, facsimile and email.

S4.2.9 Impose embargoes to prevent early public disclosure if the press is briefed in advance.

S4.2.10 Provide prompt service and support to users of statistics. Make available unpublished statistics that can be disseminated to users, upon request.

S4.2.11 Maintain updated catalogues of publications, documents and other services to users. Clearly disclose the prices of the statistical products and services and provide assistance in the placing of orders.

S4.2.12 Monitor and review assistance to users regularly. For example, monitor the response time to e-mail requests and adhere to the agency's public service standards when servicing requests.

S4.2.13 Request attribution when the agency's statistics are used for reproduction.

S4.2.14 Establish processes to deal with data misinterpretations or misuse of statistics.

S4.2.15 Monitor relevant media coverage of data produced by the agency. Comment and correct publicly erroneous interpretations or misuse of the statistics in the relevant media in a timely manner.

Recommended

General

R4.2.1 Ensure the headline for the statistical article or report is concise and conveys the key message.

R4.2.2 Write the opening paragraph to capture the attention of the reader and relay the key message of the data. It should not be a summary of the whole report, but rather contain the most important message.

R4.2.3 Write in simple language to cater to the masses. If the use of technical terms is needed, an explanation of the terminology should be provided.

R4.2.4 Avoid long sentences and paragraph structures. If there are too many points being discussed in a paragraph, consider using bulleted lists so that the reader can grasp the main points easily. If the text becomes too lengthy, use subheadings to break the content into different sections.

R4.2.5 Avoid fancy lettering or font effects as these may distract the reader.

R4.2.6 Round numbers in a consistent manner.

R4.2.7 Include information regarding the quality of the results. For example standard errors, confidence intervals and coefficient of variation to provide readers with information on data quality.

R4.2.8 Provide references in a systematic manner. For write-up disseminated via the Internet, provide electronic links to the references.

R4.2.9 Provide contact information for users to contact the agency for clarifications or for more information.

R4.2.10 Use correct unit when presenting statistics to avoid wrong interpretation of data. For example, take note of the difference between the terms “percentage change” and “percentage point change”. A percentage point change is the difference between two percentages expressed in percentage points while a percentage change refers to the change in data values expressed in percentages.

R4.2.11 Acknowledge and cite all sources used as far as possible.

R4.2.12 Equip staff with the relevant writing skills.

R4.2.13 Monitor readers' responses to the articles by gathering feedback, surveying the readers or observing the hit rates on the Internet. The information gathered is useful in identifying areas for improvement.

R4.2.14 Check that statistical information reported in the media is accurate and factual. Where necessary, provide clarifications and explanations to reporters and correct any reporting errors discovered.

R4.2.15 Inform users of any revision to the data, the basis and details for the revision.

R4.2.16 Benchmark the periodicity and timeliness of statistics according to dissemination standards. Release statistics based on a pre-announced schedule.

Internet articles

R4.2.17 Avoid making the reader scroll horizontally as it is not easy to read.

R4.2.18 Avoid capital letters. Use boldface for emphasis instead of underlined words, italic words or capital letters. Try to restrict underlined words to electronic links only.

R4.2.19 Include a print friendly version so that the article can be printed within the page margins.

R4.2.20 Use a strong colour contrast for background and words, such as dark lettering on a light background or vice versa.

R4.2.21 Include the date the article was published.

R4.2.22 Test all hyperlinks to ensure that they point to the correct web pages.

Graphs

R4.2.23 Avoid including too many data points as it may distract readers from the key results.

R4.2.24 Avoid using data point markers on the graphs if it makes the chart too cluttered.

R4.2.25 Avoid using 3D graphs unless the third dimension is needed to facilitate in the understanding of the subject matter.

R4.2.26 Identify the elements in the chart by using a legend or by labelling on the chart itself.

R4.2.27 Remove gridlines to reduce the clutter on the graph.

R4.2.28 Include a chart title which explains the purpose of the graph and the time periods covered.

R4.2.29 Label all axes clearly with units of measurements stated.

R4.2.30 Choose the right chart type in presenting the information:

- Vertical bar charts: They are used to show changes in value over a limited time period. They can be a useful tool to handle multiple series for comparison. A vertical bar group should not contain too many adjacent columns.
- Stacked vertical bar charts: They are similar to vertical bar charts, with an additional flexibility in displaying the sub-elements which make up the overall bar.
- Line charts: They are used for showing trends over a long time frame. They are also recommended in displaying multiple groups of data simultaneously. Avoid including too many series into one chart.

- Horizontal bar charts: They are mainly used to compare different individual values at a single point in time. The bars should be arranged in order of length, with the longest bar at the top. Similar to vertical bar groups, horizontal bar groups should not contain too many adjacent bars.
- Pie charts: They are best used to portray the contributions made by each category compared to the whole population. The categories should be arranged in order of size and placed either clockwise with the largest sector starting from 12 o'clock position or anticlockwise with the largest sector starting from 3 o'clock position.
- Histograms: They show the frequency of values in a set of data.

R4.2.31 Include a short explanation to help readers understand the significance of the data presented.

Tables

R4.2.32 Ensure that the overall format of the table is clear and concise to prevent misinterpretation. This includes spacing, the wordings used, placement and appearance of titles, and other forms of labelling.

R4.2.33 Format tables to ensure numbers are easy to find and understand.

R4.2.34 Rank the data in some form of order, such as arranging data from the largest to the smallest value.

R4.2.35 Label each table to identify the content of the table. Describe the data item and unit of measurement clearly in the table.

R4.2.36 If footnotes are used, ensure that they are marked clearly.

4.3 Documentation

Documentation refers to the recording of statistical activities, including concepts and methods used, and other details on the statistical production process. Documentation may be for internal reference or published as part of the metadata for statistical releases. Internal documentation should record not only the decisions made, but also the reasons behind these decisions. This is important in the development and implementation of future statistical activities. Maintaining comprehensive documentation will also help to:

- Facilitate the integration of datasets for cross-analyses
- Enable the preservation of knowledge and expertise, and aid in the retrieval of information when needed
- Comply with statutory requirements regarding the confidentiality of data
- Increase usability and understanding of the data

Strongly Recommended

S4.3.1 Document the entire statistical activity, including:

- adopting a metadata strategy to provide adequate information on the meaning of the data and the methodology used to collect and process them;
- citing sources and references.

S4.3.2 Perform regular updates on the documentation.

S4.3.3 Edit the documents thoroughly before publication.

Recommended

R4.3.1 Depending on the purpose of the documentation and the target audience, the details may include:

- Planning
  - Information on the objectives, uses and users of data, timeliness and data quality targets. Any changes should also be documented as they may affect the findings.
  - Concepts, definitions and classifications, as well as the type of questionnaire design used.
  - Sampling design and techniques.
  - Schedules of estimated time frame for each step in the statistical production process.
- Data Collection
  - Information about sampling, rotation and data collection.
- Quality and methodology
  - Adjustments for non-response using sample weights.
  - Calculations of variance estimates.
  - Editing and imputation methodologies applied during data processing.
  - Sources of sampling and non-sampling errors and methods of evaluation and mitigation.
  - Item response rates and non-response bias analysis.
- Systems
  - For data capture and processing systems, provide specifications of software application and update any developments in the software.
  - Information on data files, systems operations and monitoring reports.
  - Descriptions of values on the data file which are derived from other values on the file.
- Security, confidentiality and disclosure
  - Provide information on the legal, administrative and technical measures to safeguard data confidentiality, where appropriate.

- Data dissemination
  - Provide appendices on acronyms used, project team information, extracts of the relevant legislation used, samples of survey forms.

R4.3.2 For regular surveys, review the documentation regularly to ensure that it meets the needs of users.