Statistical Best Practices
2020

Preface

The quality of the data produced by the Department of Statistics (DOS), research and statistics units (RSUs) and government agencies must be of a high standard to instil trust and confidence among the public and policy-makers.

The purpose of this Handbook is to provide best practices on statistical processes to achieve better data quality. It defines quality standards and the various components of quality relevant to statistical processes. The Handbook is applicable to public agencies which conduct surveys or utilise data from administrative or other sources. As agencies have the duty to ensure that the data produced by them are of high quality, agencies should refer to the *requirements* in this Handbook when producing data.

Since the first Handbook was developed in 2011, the environment has changed along with significant technological advances and availability of more data from administrative and other sources. To reflect these changes and users' evolving statistical needs, DOS has undertaken a review to update the best practices in the Handbook, incorporating the use of new technological methods (e.g. machine learning techniques, predictive analytics) and alternate data sources (e.g. big data), as well as introducing a new section on data integration in recognition of its growing importance.

The updates to the Handbook are based on inputs from various ministries and statutory boards. I would like to thank all who have contributed to the review of the Handbook.

Wong Wee Kim
Chief Statistician
Singapore

October 2020

Our Vision

National Statistical Service of Quality, Integrity and Expertise

Our Mission

We Deliver Insightful Statistics and Trusted Statistical Services that Empower Decision Making

Our Guiding Principles

Professionalism & Expertise
We adhere to professional ethics and develop statistical competency to produce quality statistics that comply with international concepts and best practices.

Relevance & Reliability
We produce statistics that users need and trust.

Accessibility
We make our statistics readily available through user-friendly platforms.

Confidentiality
We protect the confidentiality of information collected by us.

Timeliness
We disseminate statistics at the earliest possible date while maintaining data quality.

Innovation
We constantly seek ways to improve our processes, leveraging on new technology where feasible, to deliver better products and services to users.

Collaboration
We engage users on data needs, data providers and respondents on supply of data, and undertake the role of national statistical coordination.

Effectiveness
We optimise resource utilisation, leveraging on administrative and alternative data sources to ease respondent burden.

Contents

Contents *(cont'd)*

PART III: ANNEXES

I.   INTRODUCTION

To ensure reliable and trusted data are delivered to users, it is important to monitor and manage the quality of each statistical activity that an agency engages in from the development to the dissemination of statistical outputs. At the organisational or institutional level, it is also crucial to have legislation on data collection and acquisition.

A statistical activity refers to any activity in the collection, handling and processing of information to produce statistical outputs. Some examples are given below:

1. Conduct of census and survey where data are collected from all or a sample of the targeted population respectively.

2. Acquisition of information from administrative or other sources (e.g. big data), from which data are created or derived for statistical purpose.

3. Statistical activities which produce estimates by integrating data from several different sources. This may involve appending records from different datasets, or record linkage where records associated with the same person or firm from one or more databases (or time periods) are identified and combined.

Statistical Processes

This Handbook has referenced the Generic Statistical Business Process Model (GSBPM, version 5.1)[1] released in January 2019 to describe the statistical processes and sub-processes. There are five key statistical processes needed to produce data: Planning & Preparation, Survey Design, Data Collection, Data Processing, and Compilation, Analysis & Dissemination (Diagram 1). Each key process is further divided into sub-processes, which cover areas such as allocation of resources, staff training, statistical procedures and techniques, safeguard confidentiality of data, and evaluation and assessment of outputs. While the processes and sub-processes are arranged in a logical or sequential order as shown in the diagram, they can also occur in parallel and can be iterative. For instance, questionnaires may have to be redesigned after a few rounds of testing. In addition, some of the processes or sub-processes may not be applicable to every statistical activity undertaken. For example, designing a questionnaire is not necessary when data are extracted from administrative records, and some data do not need to be seasonally adjusted.

---

[1]   The GSBPM is an international standard for describing statistical processes in a coherent way which was first jointly created by the United Nations Economic Commission for Europe (UNECE), Eurostat and the Organisation for Economic Co-operation and Development (OECD).

Diagram 1 (adapted from GSBPM)
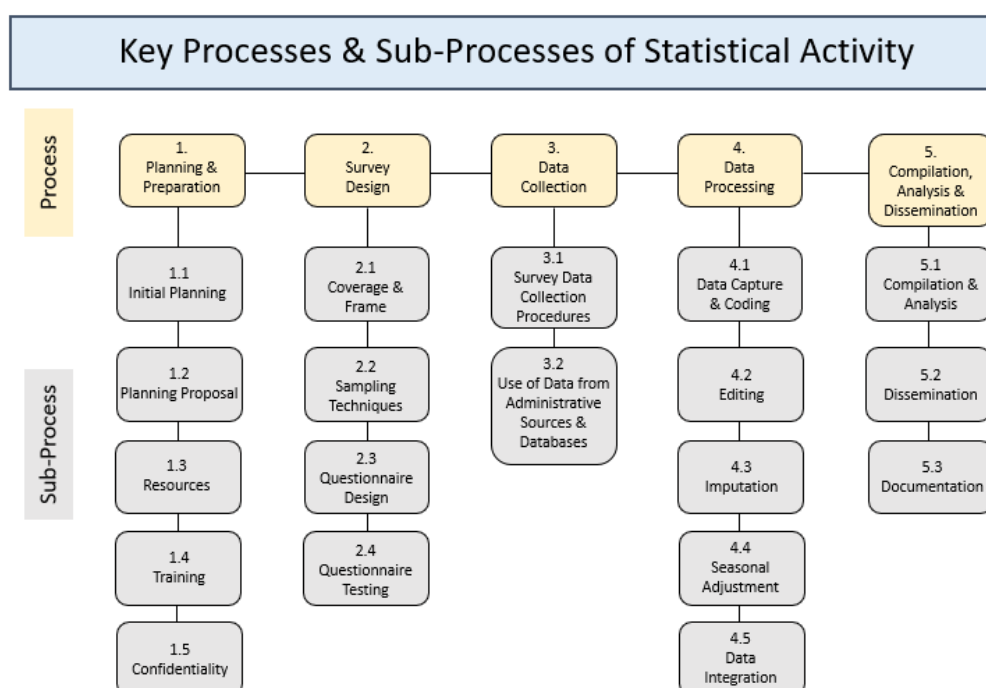
Data Quality Standards

Data quality is multi-dimensional and can be characterised by specifying a set of interrelated factors or dimensions. This Handbook identifies six data quality dimensions as guiding principles which describe what agencies must do to ensure a high level of quality in their statistical outputs. While these dimensions differ slightly across statistical agencies, the principles in this Handbook are consistent with those defined in the United Nations National Quality Assurance Framework (UN-NQAF) 2019 and Statistics Canada Quality Guidelines 2019.

1) **Relevance**: the extent to which the data meet the needs of users
   - Processes are in place to consult data users, review the usefulness of existing data and identify data needs

2) **Accuracy**: the extent to which the data correctly describe the phenomena they are designed to measure
   - Sound methodologies are applied throughout the preparation, production and dissemination of data, and the source data (such as existing surveys, administrative records and big data) are regularly assessed and validated

3) **Timeliness**: the length of time between the reference point (or the end of the reference period) to which the data pertain, and the date when the data become available
   - Data are released at the earliest possible date

4) **Accessibility**: the ease with which the data can be found, obtained and used
   - Data are disseminated in suitable and convenient formats periodically

5) **Interpretability**: availability of data and metadata needed to interpret and use the data appropriately
   - Clear information on concepts, methodology and details of the data and statistical processes is adequately documented and made available

6) **Coherence**: the extent to which the data can be reliably combined and compared with other data within a broad analytic framework over time
   - Data are logical and consistent internally as well as comparable with other data sources over time

The six dimensions of quality described are interrelated and there are trade-offs between some of them which agencies should take into consideration when developing statistical outputs. For example, if agencies wish to produce data in a timely and frequent manner, accuracy of the data may be compromised. This may arise from the complete source data from which the data are compiled being unready or from it being not cost-effective to obtain the information within the required time frame.

High data quality can be achieved not only through management of statistical outputs. It is also crucial to manage other factors at the institutional environment and statistical process levels. Factors such as ensuring professionalism and integrity through the provision of staff training (so that staff is better equipped to perform their statistical work), ensuring confidentiality (so that data providers are willing to furnish their information), ensuring cost effectiveness (so that limited resources can be put into better use to improve data quality) and managing respondent burden (so that respondents are not overly fatigued in providing accurate information) also affect the quality of data produced.

Structure

In the subsequent chapters of this Handbook, the best practices that ensure or maintain high data quality are arranged by the different key processes and sub-processes specified. Broadly, these best practices are presented as either **requirements** *for compliance* by agencies (numbered with a prefix "S", e.g. S1.1.1) or **recommendations** that agencies *can consider to adopt* (numbered with a prefix "R", e.g. R1.1.1).

Users can refer to the <u>Annexes</u> for the definitions of key statistical terms and the relevant specialised statistical manuals and guides for more information on how to carry out these statistical processes.

For specific enquiries, users may contact *singstat_ppd@singstat.gov.sg*.

II.   STATISTICAL BEST PRACTICES

1.   Planning and Preparation

This chapter presents the various stages involved in the planning and preparation for a statistical activity, viz. conducting an initial plan, writing a planning proposal, allocating resources, providing training and ensuring confidentiality of information collected.

1.1   Initial Planning

The first task in planning a statistical activity is to specify the objectives to establish the need for statistical outputs. It is important to consult potential stakeholders to align the objectives with their needs to ensure the relevance of data produced.

To meet the specified objectives, agencies should explore various data sources, such as using existing statistical or administrative data, re-designing existing surveys, undertaking new surveys, or a combination of these activities. Before designing a new survey, agencies should consider the feasibility of using available data from existing administrative sources and databases to meet the objectives which will help reduce respondent burden and cost. Agencies should also consider the relevant legislation to enable them to use existing data sources or undertake surveys. Conducting an initial planning will aid the agencies in determining the best course of action.

Requirements

S1.1.1   An authority is established for the agency to collect, process and disseminate data. This can be in the form of a legislation (e.g. Statistics Act).

S1.1.2   Objectives of the statistical activity are clearly stated.

S1.1.3   Relevant standards (concepts, definitions, classifications, etc.), good practices and recommendations by national and international bodies (e.g. DOS, United Nations or other National Statistical Offices) are adopted or adapted so that the data collected, produced and disseminated follow sound methodologies and cost-effective procedures, and are of good quality and internationally comparable.

S1.1.4   Assessments are made before launching a new survey to determine that the required data variables cannot be obtained or adequately approximated from existing data sources (e.g. administrative data, data from private sector or big data) or by having additional questions in existing surveys that are conducted or being planned by other government agencies. These include using additional

tabulations or more advanced statistical techniques such as modelling in the assessments.

Recommendations

R1.1.1   Data and indicators based on new and existing data sources are developed in response to emerging information needs, e.g. the development of nowcast indicators to meet users' need for near-present sensing.

R1.1.2   Key users and stakeholders are engaged regularly to better understand, anticipate and address their current and evolving data needs, and at the same time deprioritise outdated needs.

R1.1.3   A workgroup consisting of cross divisions/agencies is formed to pool expertise and resources together and ensure the objectivity of the data produced.

1.2   Planning Proposal

When it is determined that the necessary data are to be collected through a survey or via 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 manage the cost effectiveness of a project and assure the quality of the final outputs. It helps to determine the appropriate statistical procedures that meet users' needs while achieving a balance between the resources available and the desired quality of outputs.

Requirements

S1.2.1   The timelines of key activities, including consultation on data items with users/stakeholders, questionnaire design, data collection, data processing, data dissemination and publicity campaign, are drawn up.

S1.2.2   The budget and manpower resources required for the statistical activity are estimated.

S1.2.3   Data users / stakeholders are consulted to review the usefulness of existing data outputs and identify data requirements.

S1.2.4   The data source and methodology that best suit the purpose of the statistical activity are used, and justifications are provided. These include the proposed target population, target response rate, sample frame, sample design, sample size, timeliness and coverage of administrative data acquired, the appropriate

statistical techniques needed to achieve the objectives of the analysis, and other methodological issues.

S1.2.5   When more than one data producing agency is involved in producing parts of the data, procedures to promote consistency of methods and results are established.

S1.2.6   National and international practices in the use of key statistical concepts and definitions are followed.

S1.2.7   Data collection or acquisition procedures are planned and documented. These include the 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   ICT systems for data collection/acquisition are designed with governance and checks in place to prevent data theft, data loss or data modification, due to malicious activity from external parties, system failures, unintended access to data, and human errors during data capture, transmission and handling.

S1.2.9   Validation and consistency checks are included in the data collection/ acquisition system and process to detect any omissions and erroneous entries.

S1.2.10   Online validation checks should not significantly affect the response time of the system.

S1.2.11   Appropriate survey data collection methods are chosen to achieve the data quality objectives, such as timeliness and accuracy. These can include internet submission, and telephone and face-to-face interviews.

Recommendations

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

R1.2.2   Adopting new technology that improves the efficiency of data collection and the quality of data collected is considered. Advances in communication and computing technology offer opportunities to reduce cost and respondent burden. For example, electronic handheld devices in face-to-face interviews help reduce errors when filling up the questionnaire.

R1.2.3   A plan for quality assurance is outlined for monitoring and assessing the performance during implementation. The plan should include contingency scenarios.

R1.2.4   Indicators on data quality are defined, regularly measured, monitored, published and followed up to improve processes and outputs. Examples of quality indicators include references in media, hits on website, results from user satisfaction surveys (relevance); standard deviations and other measures of accuracy, response rates (accuracy); number and size of revisions (reliability); length of time between the end of a reference period and the dissemination of data (timeliness); and respondent burden.

1.3   Resources

Resources refer to the funding, facilities, material, staff and other assets available to the agency necessary for its effective operation. Ensuring that sufficient resources are deployed for the statistical activity is critical to achieve success of the activity. For example, providing adequate computing resources will enable staff to perform the required tasks efficiently and produce quality data.

Requirements

S1.3.1   Sufficient resources are allocated to exploit the potential of computing technology for collecting, compiling and disseminating data. For example, sufficient budget is allocated to make Application Programming Interface (API) available to enable users to access published statistics.

S1.3.2   Appropriate software is used or adapted to perform tasks for compiling and analysing data.

S1.3.3   Hardware and software are distributed adequately to facilitate the efficient collection and processing of data, and management of databases.

S1.3.4   ICT resiliency is put in place for all ICT systems to mitigate risks, respond to incidents and continue an acceptable level of service when disruptions take place.

S1.3.5   There are procedures to assess the continuing need for all data outputs (e.g. statistics, databases, publications and consultancy services) to see if any can be simplified (e.g. in terms of frequency, detailed breakdowns and coverage) or discontinued to free up resources.

1.4   Training

Training should be an on-going process to develop and equip officers, interviewers and supervisors with the skills and knowledge to minimise errors in data capture and data processing, and reduce the incidence of 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.

Requirements

S1.4.1   Basic training is provided to the staff involved in data collection, including:
- Overview of the objectives and principles behind the data collection
- Statutory obligations to the respondents (where relevant)
- Security measures to safeguard the data collected
- Definitions of the data items
- Interviewing techniques which establish trust between respondents and the 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   All training required is planned in advance. This includes training staff for contingency events such as pandemics.

S1.4.3   Training provided is monitored and evaluated on the effectiveness to achieve targeted outcomes such as improving skills and knowledge of staff.

S1.4.4   Training materials are reviewed and updated on a regular basis.

S1.4.5   Data capture and coding operators are provided with appropriate training to capture and code the data accurately.

S1.4.6   Staff is equipped with the necessary skills to use new technologies and methodologies in the editing process.

S1.4.7   Training in data collection, processing and compilation methods is provided adequately. This may include participation in seminars, courses and workshops arranged by regional and international organisations to improve the knowledge of best practices, and access to professional literature.

S1.4.8   Staff involved in the imputation process is trained on the principles behind imputation, the different imputation methods and the corresponding procedures.

S1.4.9   Staff involved in the seasonal adjustment process is trained on the principles behind seasonal adjustment, the different adjustment methods and the corresponding procedures.

S1.4.10   Interviewers are trained to take precautions to prevent contents of the survey from being disclosed during fieldwork. For example, an interviewer should not use the issued computer device in public places where there is a chance that the contents can be disclosed accidentally.

Recommendations

R1.4.1   Interviewers are provided with a script to guide survey enumeration. 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 and consistent as possible.

R1.4.2   Each interviewer is provided with a document containing all necessary procedures and instructions in handling respondents.

R1.4.3   Interviewers are instructed to be punctual and appropriately dressed for all interview appointments, and provide advance notice to the respondents if they are unable to keep to the appointment.

R1.4.4   Interviewers are instructed to complete telephone surveys or fieldwork which do not require follow-up in a single appointment wherever possible. If not, they have to remember to fix another appointment to continue the survey.

R1.4.5   Interviewers are instructed to converse in a language that the respondent is familiar with to avoid miscommunication.

R1.4.6   Arrangements are made for new interviewers to observe and learn from the more experienced interviewers for the first few visits/calls.

R1.4.7   Supervisors are equipped with the necessary management and appraisal skills.

R1.4.8   Appraisal reports on interviewers or staff are maintained to keep track of their performance and identify any need for further training.

R1.4.9   There are arrangements to review survey returns with interviewers and look for possible errors that are due to administration gaps which can be corrected on the ground. For example, checks are conducted across correlated items to ensure that the information collected is consistent and illogical answers are flagged out.

R1.4.10   Staff is provided with training on metadata management and related information and documentation systems.

R1.4.11   Staff is equipped with relevant writing skills.

1.5   Confidentiality

Data confidentiality refers to the sensitivity of the data pertaining to an individual person or firm. Staff should safeguard the confidentiality of individual information throughout all processes associated with the statistical activity in accordance with the legislation or under which the information was obtained.

Requirements

S1.5.1   All officers, including temporary and contract staff, are to sign relevant legal confidentiality undertakings upon appointment. Officers must pledge not to disclose any individually identifiable data (unless permitted by the legislation governing the data), even after they leave the service.

S1.5.2   Access to information on an individual person or firm is controlled on a need-to-know basis.

S1.5.3   Penalties are imposed for any wilful breaches of statistical confidentiality.

S1.5.4   Technical and process measures are put in place to safeguard the identifiable data.

S1.5.5   Data are not released in a form that may disclose the identity of any individual person or firm, unless permitted by the legislation governing the data.

S1.5.6 There are strict protocols to safeguard data confidentiality that apply to users with access to microdata for statistical purposes, in accordance with the relevant legislation governing the data. This may include stating the requirements to grant microdata access to users, monitoring the use of microdata, taking immediate corrective action to address any breach in data confidentiality and controlling users' access to microdata by providing them in a secure environment.

Recommendations

R1.5.1 Interviewers are instructed to 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.

R1.5.2 Statistical disclosure control methods are used to prevent the disclosure of individual information in aggregated data and to prevent re-identification of microdata. Consideration should be given to the sensitivity, intended use and relevant legislation safeguarding the data when selecting the appropriate method.

2.   Survey Design

This key statistical process includes stages relating to building and testing of survey tools (such as sampling frame and questionnaire) for the collection of survey data. Appropriate methods and procedures to control potential errors in these stages are described in this chapter. The types of errors involved are typically related to the accuracy of data collected and should be minimised. Examples of the errors can be found in Diagram 2 below:

Diagram 2
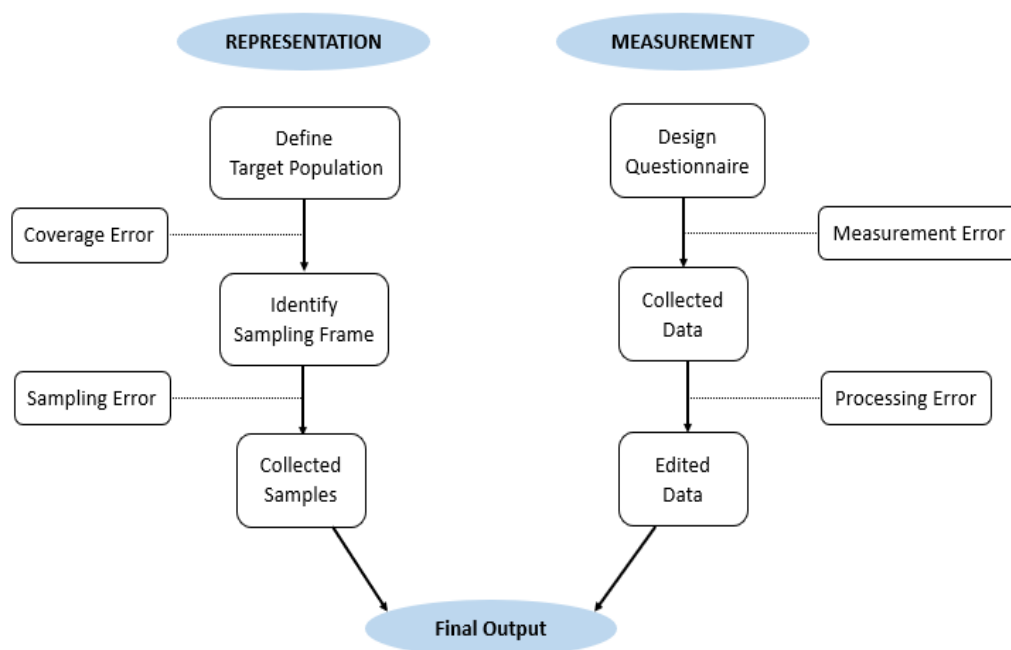
2.1   Coverage and Frame

Once the target population has been determined for a survey, a sampling frame should be identified to select the survey population. An ideal sampling frame is one that is complete, accurate and up-to-date on the condition that cost is not a constraint. When constructing the sampling frame, the relationship between the target population and the unit of selection should be considered. The frame should be chosen to conform to the target population as far as possible and contain minimal under- and over-coverage, particularly of important population sub-groups.

Requirements

S2.1.1   Assessments are made to use the most appropriate sampling frame for the survey, taking into account the cost that may be incurred.

S2.1.2   Procedures are set up to eliminate duplication, inconsistencies and errors in the sampling frame. For example, validation checks are built in to reduce such errors.

Recommendations

R2.1.1   The frames are updated on a regular basis using a systematic approach. A series of checks are put in place to preserve integrity of the frames and ensure that the data are of good quality. For example, error correction checks are built in during the updating process to ensure data quality and integrity. Records which fail the validation or edit checks should be scrutinised and corrected. Source agencies are consulted where applicable.

R2.1.2   The frame should capture the target population of the survey as far as possible.

R2.1.3   Supplementary data from other sources and sampling weights are used to minimise coverage errors of the frame.

R2.1.4   The coverage of the frame and target population are assessed to determine any coverage errors.

2.2   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 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.

Requirements

S2.2.1   The appropriate sample size is determined for the survey. Relevant factors to consider include:
- 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
- Budget and timeframe

S2.2.2   Different sampling methods are evaluated to determine the most appropriate method to be used for the survey, taking into account the data requirements and planned method of data collection for the survey.

S2.2.3   Sample design is evaluated after the completion of the survey by comparing the precision of estimates actually achieved against the planned precision during sample design. Differences in the estimates between the achieved and planned precision are investigated and reasons for the difference are identified for future improvements.

S2.2.4   There are periodic reviews on the sample design used.

Recommendations

R2.2.1   The sampling efficiency is improved by 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.

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

Good questionnaires translate survey objectives accurately into easy to understand/ administer questions that allow respondents to provide the information required for the survey.

Requirements

S2.3.1   The rationale for all questions in the questionnaire is documented with the 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.

S2.3.2   The introduction of the questionnaire provides the title/subject of the survey, name of the agency conducting/sponsoring the survey, purpose of conducting the survey, authority under which the survey is conducted, confidentiality clause (if any), and request for the respondent's cooperation.

S2.3.3   Wordings in the questionnaire are unambiguous and easily understood by the target population. For example, statistical/technical terminology and jargon, "double-barrelled" and leading questions, and questions containing double negatives are avoided. Such questions would affect the accuracy of results.

S2.3.4   Self-administered questionnaires are designed to be user friendly, with a logical flow, clear and concise instructions to guide respondents, and definitions which are highlighted clearly.

S2.3.5   Editing checks are built into electronic questionnaires so that respondents can correct errors on the spot. This allows potential errors to be identified quickly.

S2.3.6   Relevant concepts and definitions are harmonised with similar surveys to facilitate data comparability over time.

Recommendations

R2.3.1   Respondents are encouraged to provide best estimates where precise responses cannot be provided.

R2.3.2   An assessment is made on whether the survey should be conducted in several languages. The questionnaire is translated only after the main language version is finalised.

R2.3.3   The method of data collection is taken into account when designing the questionnaire. Electronic handheld devices and other electronic platforms have space and layout constraints. The design for internet submissions should also factor in response times.

2.4   Questionnaire Testing

Pre-testing the questionnaire is an important part of questionnaire design as it helps to identify any potential problems with the survey questionnaire before data collection. This includes identifying incoherent flow in questionnaire layout, ambiguous wordings and concepts that pose understanding difficulty for respondents to provide accurate information. Testing will also help to determine the duration of the surveys and estimate the non-response rate, and at the same time, it is useful to assess if questions are relevant and easily understood by respondents.

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. The choice of the evaluation methods depends on the type and size of the survey, the survey content, utilisation of previous survey questions or standard questions, the method of data collection, project schedule, and the budget and resources available.

Requirements

S2.4.1   Questionnaire is tested at the early stage of its development.

S2.4.2   Questionnaire is revised based on initial testing findings and the revised questionnaire is re-tested.

Recommendations

R2.4.1   An appropriate method is selected to test the questionnaire. The methods can include pilot tests or qualitative tests such as focus groups.

R2.4.2   A pilot test is conducted for new and large projects to fine-tune the questionnaire before its use in the main survey. Feedback from staff involved is obtained to identify potential sources of response and non-response errors as well as areas where the questionnaire can be further improved.

R2.4.3   The questionnaires in different languages are checked to ensure consistency in translations.

3. Data Collection

After a detailed plan has been drawn up to collect the data, agencies may proceed with data collection. Data can be obtained directly from individual persons or firms (via self-enumeration, telephone interviews or face-to-face interviews) or indirectly by extracting data from administrative 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.

3.1 Survey Data Collection Procedures

To ensure data collected from surveys are complete and accurate, approaches to reaching selected households and respondents, going through the survey and obtaining responses should be taken into consideration when conducting fieldwork.

Despite measures to maximise survey responses, the number of fully completed questionnaires may fall below expectations, resulting in biased 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.

Requirements

S3.1.1   Respondents are informed during data collection of the purpose of the survey, survey date or period, due date for submission of returns and appropriate channels for submission/feedback.

S3.1.2   Efforts are made to reach selected respondents, such as obtaining accurate contact information of respondents and taking appropriate and swift actions to obtain information from non-respondents. These include sending reminder letters or calling respondents if they do not submit returns after the specified due date.

S3.1.3   Convenience sampling and participation by volunteers are not used for replacing non-respondents. This is because surveys based on convenience sample are likely to be biased towards those with certain characteristics while surveys based on volunteers are likely to be biased towards those with strong opinions or with time on their hands.

S3.1.4   Explanations are provided to respondents on how they are selected based on an established sampling method and the significance of their participation in the survey when requested.

S3.1.5   Respondents are informed of the legislation on the confidentiality safeguards of information provided and their obligations.

S3.1.6   Assistance is provided to respondents in completing and submitting forms (e.g. providing a contact point).

S3.1.7   Response rates are monitored closely and appropriate methods are employed to encourage participation. For example, using a wider range of communication channels in disseminating the survey message and obtaining returns or feedback.

S3.1.8   Reasons for non-response are documented, such as refusal, no-contact and language barrier.

S3.1.9   Data collection operations are monitored at regular intervals (e.g. weekly) to ensure all procedures are properly implemented.

Recommendations

R3.1.1   Incentives can be considered to encourage higher response rates, especially for lengthy and complex surveys.

R3.1.2   The respondent or a responsible person of the household or firm is contacted at an appropriate time. Unless specifically requested by respondents, the last fieldwork visit or telephone call of the day should not be made too late into the night for household surveys.

R3.1.3   Respondents are not inconvenienced unnecessarily. For example, all incomplete entries and flagged responses are consolidated before calling them to verify.

R3.1.4   Data collection processes and response rates of recurrent surveys are reviewed regularly, such as providing multiple channels for data collection to improve response rates.

R3.1.5   There is 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.

R3.1.6   Agreements are established with data suppliers to provide data on a regular basis and provide updates on any data revisions, rebasing and breaks in data series.

R3.1.7   For large or regular surveys, an ICT system is developed to assist with data collection, for example, to send out survey forms, track survey returns, send reminders automatically, compute real-time response rates and schedule follow-ups.

3.2   Use of Data from Administrative Sources and Databases

Administrative sources refer to data collected for specific purposes under various legislation. The use of administrative data helps to reduce respondent burden and cost of conducting surveys. Data can also be obtained from existing databases comprising data that were compiled previously.

As the use of administrative sources and databases presents some challenges in data quality, we need to consider the following trade-offs in determining their usability:
- Content
- Coverage
- Frequency
- Accuracy
- Timeliness
- Cost
- Respondent burden

Requirements

S3.2.1   Source data which meet users' needs or reasonably approximate their data requirements in terms of concepts, definitions, scope, classifications, valuation and time of recording, methods and quality, are used.

S3.2.2   Procedures to provide for the effective and timely flow of data from administrative or other sources (e.g. big data) are established.

S3.2.3   Accuracy and completeness of data from administrative sources and databases are assessed regularly. For example, there are checks for missing or invalid data fields and ensure alignment with published statistics where possible.

S3.2.4   Contacts are maintained with other data producing agencies to better understand data requirements, avoid duplication of effort and take into account reporting burden. For example, changes to administrative processes are discussed before they take place.

S3.2.5   Statistical discrepancies and differences in source data are assessed and investigated which include consulting the source agencies when there is outdated or missing information in the database, and building validation checks for specific data items obtained from various agencies.

S3.2.6   Rare occurrences like unique records or outliers (such as persons aged over 100 years old or unique country of birth) are sieved out and verified for data accuracy and appropriate action is taken, in consultation with the source agencies.

4. Data Processing

4.1 Data Capture and Coding

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

As technology improves, data are increasingly captured directly into ICT Systems with minimal human involvement. Advanced ICT systems can also be developed to perform computer-assisted coding.

Requirements

S4.1.1   Automated data capture process is designed to reduce cost and maximise timeliness and data accuracy. For example, appropriate ICT systems to scan and capture data for large surveys involving fieldwork are developed to save time and cost. For large volumes of data, ICT system that can be scaled up to handle increased loads of work is used.

S4.1.2   Automated data capture systems are tested prior to implementation to ensure reliability of the systems.

S4.1.3   Reference files for automated coding systems are built and maintained.

S4.1.4   Consistent coding is applied to all data items. This includes recoding free text responses under "Others, please specify" options to pre-coded options where possible using an automated process (e.g. applying machine learning techniques).

S4.1.5   Data collected from surveys, administrative or other sources (e.g. big data) are checked for accuracy, such as checking for incomplete data and invalid data format. For survey data, these include conducting interview verification to deter, prevent and detect interviewers who falsify survey responses. For administrative data, these include checking with the source agencies to clean the data or provide usage guidelines in the metadata.

S4.1.6   Procedures to adjust for under-coverage and population units which are out of scope are established and grossing up factors are computed scientifically based on sample design.

Recommendations

R4.1.1   Common rules are spelt out for different development teams to follow in large projects where there are different systems. This allows for greater synergies between systems.

R4.1.2   Information regarding data quality is gathered to improve future data capture and coding procedures.

4.2   Editing

Data editing is a process designed to check that data collected are complete and recorded logically through detecting erroneous data and treating them consistently. Errors or inconsistencies in the data may have come from data entry or have been introduced during data collection (via surveys or administrative systems), such as respondent misunderstanding the questions, interviewer bias or non-responses. It is important to ensure errors and inconsistencies are corrected before analysing the data.

Documenting and monitoring editing process and results will help to improve overall process by providing information on errors arising from the earlier stages of data collection and processing.

Requirements

S4.2.1   Processes to minimise editing errors are developed. For example, doubtful cases are highlighted for further examination or editing checks are performed over the course of fieldwork so that prompt actions can be taken to address inconsistencies.

S4.2.2   Data collected from main sources are validated against other independent data sources where possible.

Recommendations

R4.2.1   Editing of erroneous responses is prioritised, starting with more severe errors or high likelihood of influential errors, i.e. responses that have a large impact on target survey estimates.

R4.2.2   Appropriate tools are used to detect outliers and provide information on how the individual responses affect the overall statistics. For example, graphical editing can be used.

R4.2.3   The editing process is monitored using indicators like editing ratios to analyse the efficiency of data processing.

R4.2.4   Processes to learn from past experience on the appropriate type of logic and editing checks are put in place.

4.3   Imputation

Imputation is a process by which the missing data in a dataset are 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.

Requirements

S4.3.1   Sound estimation techniques to adjust data for missing observations are adopted.

S4.3.2   Imputation process is examined by using test data to detect and improve imputation method before the application to actual data.

S4.3.3   Imputation methods, values and sources are documented to facilitate monitoring and analysis.

Recommendations

R4.3.1   An imputation method that is stable and gives a good prediction of missing data is selected. When choosing the imputation method, the type of data being imputed and the availability of historical and administrative data are considered. For 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.

R4.3.2   The imputation results are monitored using indicators like imputation ratio to study the impact of imputation.

4.4   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

    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.

Requirements

S4.4.1   Seasonal analysis is conducted thoroughly to assess if seasonality is identifiable for the series intended to be adjusted.

S4.4.2   Time series are adjusted only if the seasonality has been identified.

S4.4.3   Short time series of less than three years for monthly data and four years for quarterly data are not seasonally adjusted.

S4.4.4   Seasonally adjusted series are validated using the default diagnostics generated by the seasonal adjustment tool to check that the seasonally adjusted series are of acceptable quality.

Recommendations

R4.4.1   Basic graphical analysis is used 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

R4.4.2   Regression-based methods are used to estimate and remove the moving holiday and trading-day effects.

R4.4.3   Unadjusted outliers may lead to distortions in the seasonally adjusted series. Hence, the time series are checked and pre-adjusted for outliers before the estimation and removal of seasonal variations. After the removal of seasonal variations, the outliers are reinstated in the time series, so as to reflect the specific events (e.g. SARS, COVID-19) that have caused the outliers.

R4.4.4   Aggregate series is seasonally adjusted 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 is derived by combining the seasonally adjusted component series. Indirect method is preferred if the component series show different seasonality.

4.5   Data Integration

Data integration refers to combining data or records from separate sources which are collected for different purposes or at different times. It includes appending (or pooling) of individual records of persons or firms from different datasets and linking of individual records of persons or firms from different datasets via identifiers (e.g. identity number) or a few key variables (e.g. name, date of birth and sex of a person). Data integration enables producers to leverage on multiple data sources to provide more robust and comprehensive data, and to develop more timely and disaggregated statistics. Examples of data integration with different combinations of data sources include administrative data from different administrative sources, administrative data with data from surveys, and administrative data with new online data sources. Integrating data from multiple sources can help to reduce survey respondent burden and data collection efforts, as well as increase data usability and cost-efficiency.

Requirements

S4.5.1   The content and coverage of the data sources are evaluated during data integration. Where necessary, work with the data source agencies to resolve any data inconsistencies. Examples of checks include data consistency within and across datasets, macro-level checks against historical norms to detect potential errors, and consistency with other aggregated data.

S4.5.2   The code conversion tables for data items are established to map and convert different codes and classifications correctly when integrating data from different sources.

S4.5.3   The reference periods for all data items are considered when integrating data from different sources.

S4.5.4   For linking of records from different sources, record linkage procedures and methods are established and documented. These include the use of exact matching or probabilistic linking.

S4.5.5   Units common to all source data are identified and quality of units are examined before integrating.

S4.5.6   Apply appropriate methodology or techniques to integrate data from different sources.

Recommendations

R4.5.1   Common and unique identifiers are to be adopted as far as possible to allow for ease of data integration across multiple sources.

5. Compilation, Analysis and Dissemination

5.1 Compilation and Analysis

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

Data analysis involves the examination and interpretation of data to produce insights. 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

Requirements

S5.1.1   Classifications on the compilation of data that are in broad conformity with national/international recommendations are adopted.

S5.1.2   Data produced are consistent within the dataset. For example, concepts and definitions adopted for compiling quarterly estimates are the same as those used to compile annual estimates.

S5.1.3   Data produced are consistent or reconcilable over a reasonable period of time.

S5.1.4   Data prepared for dissemination are checked for possible indirect disclosure of individually identifiable data, and tables and outputs are prepared in a way that prevents disclosure, unless permitted by the legislation governing the data.

S5.1.5   Revision studies are undertaken on a regular basis to assess the initial estimates against revised or final estimates.

S5.1.6   The sources of errors, omissions and fluctuations in the data are investigated when discrepancies are found during revision studies. Corrective action is promptly taken to rectify any errors and omissions.

S5.1.7   Regular and transparent schedule for data revisions is established.

S5.1.8   Preliminary data are clearly distinguished from revised data.

Recommendations

R5.1.1   If more than one data source is used, the consistency and impact of data sources are considered and integrated into the analysis.

R5.1.2   Summary statistics and graphs are used to facilitate understanding of the data and detect any unusual patterns (e.g. outliers, trend breaks).

R5.1.3   In general, commercial software is used over non-commercial software as they have undergone more rigorous tests. Open-source software could be used if they are among the popular ones which have undergone rigorous testing so that the data outputs are consistent with the results obtained from established commercial software.

R5.1.4   Advanced technologies, tools and techniques such as text mining and predictive analytics are adopted in big data analysis where possible.

R5.1.5   Exploratory or preliminary data are released with appropriate communications if there are likely to be significant delays to the revised data, if the data require multiple rounds of revisions, and if the data will be revised with the availability of more complete or additional source data.

5.2   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 data are textual, tabular and graphical.

Requirements

S5.2.1   The confidentiality requirements governing the data being disseminated are complied with.

S5.2.2   There is a system of checks to verify the language, accuracy of data and consistency of figures used in text, tables and charts before publishing.

S5.2.3   The data are presented in a way that meet users' needs and facilitate proper interpretation and meaningful comparison, for example, through the use of text, customisable tabulations, infographics, interactive storyboards and videographics.

S5.2.4 The data outputs of the agency are clearly identified so that users are aware of the agency which is responsible. For example, the agency's name, logo and insignia are displayed in publications. In the case of joint publications, the part attributable to the relevant agency is identified. For example, data are clearly distinguished from policy interpretation.

S5.2.5 Information about the terms and conditions under which the data are compiled and disseminated, including safeguards on the confidentiality of data, is made available to users and stakeholders.

S5.2.6 Explanatory materials and briefings are provided to minimise misinterpretation or misuse of data. For example, a list of definitions for common terms used in the briefing materials as well as clearly labelled charts/figures are provided to minimise misinterpretation or misuse of data and to facilitate the media's reporting.

S5.2.7 Contact points for all statistical releases are clearly identified for users to contact the agency for clarifications or for more information. These contact points may include address, telephone and email.

S5.2.8 Embargoes are imposed to prevent early public disclosure. For example, impose embargoes on agencies that have early access to data before release, and the media if the press is briefed in advance.

S5.2.9 Prompt service and support is provided to users of data. This includes making available unpublished data that can be disseminated to users upon request.

S5.2.10 Catalogues of publications, documents and other services to users are updated on a regular basis.

S5.2.11 The provision of assistance to users is monitored regularly. For example, response time to email requests is monitored and the agency's public service standards are adhered to when servicing requests.

S5.2.12 Attribution is requested when the agency's data are used for reproduction.

S5.2.13 Processes to deal with data misinterpretations or misuse of data are established.

S5.2.14 Media coverage of data produced by the agency is monitored and checked for accuracy. Where necessary, comments, clarifications and explanations are provided to reporters and any reporting errors, erroneous interpretations or misuse of the data in the relevant media are corrected publicly in a timely manner.

S5.2.15   More data outputs are made available publicly where possible, and in formats that improve users' self-help experience. These include more granular data (such as via multi-dimensional data cubes), customised data and indicators.

Recommendations

General

R5.2.1   Prices of data outputs are disclosed clearly.

R5.2.2   The headline for the article or report is concise and conveys the key message.

R5.2.3   The statistical releases and documentation are written in simple language to cater to the general public. If the use of technical terms is needed, an explanation of the terminology should be provided.

R5.2.4   Fanciful lettering or font effects are avoided as these may distract the readers.

R5.2.5   Numbers are rounded in a consistent manner.

R5.2.6   The correct units are used to avoid wrong interpretation of data. For example, the difference between the terms "percentage change" and "percentage-point change" should be taken note of. 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.

R5.2.7   The opening paragraph is written to capture the attention of the reader and relay the key message. It should not be a summary of the whole report but rather contain the most important message.

R5.2.8   Long sentences and paragraph structures are avoided where possible. 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 sub-headings to break the content into different sections.

R5.2.9   Information regarding the quality of the data is included. For example, standard errors, confidence intervals and coefficient of variation are included.

R5.2.10   References are provided in a systematic manner. For example, electronic links to the references are provided for reports disseminated via the internet.

R5.2.11   All sources used are cited as far as possible.

R5.2.12　　Readers' responses to the articles are monitored by gathering feedback, surveying the readers or observing the hit rates on the internet. The information gathered would be useful in identifying areas for improvement.

R5.2.13　　Users are informed of any revision to the data, the basis and details of the revision.

R5.2.14　　The periodicity and timeliness of data are benchmarked according to dissemination standards such as Special Data Dissemination Standard (SDDS) by International Monetary Fund (IMF).

R5.2.15　　Data are released based on a pre-announced schedule and delays to the planned release dates are monitored.

R5.2.16　　There is a data dissemination strategy which is publicly available.

R5.2.17　　There are measures to promote statistical literacy. These include publishing articles on statistical issues and how statistics should be used properly.

Internet Articles

R5.2.18　　The formatting of internet articles should avoid making the reader scroll horizontally as it is not easy to read.

R5.2.19　　Boldface for emphasis is used instead of underlined words, italic words or capital letters. Underlined words are restricted to electronic links only where possible.

R5.2.20　　A print friendly version is included so that the article can be printed within the page margins.

R5.2.21　　A strong colour contrast is used for background and words such as dark lettering on a light background or vice versa.

R5.2.22　　The date the article was published is included.

R5.2.23　　All hyperlinks are tested to ensure that they point to the correct web pages.

Graphs

R5.2.24　　Too many data points should be avoided as they may distract readers from the key results.

R5.2.25　　Data point markers should not be used on the graphs if they make the chart too cluttered.

R5.2.26   Three-dimensional (3D) graphs should be avoided unless the third dimension is needed to facilitate understanding of the subject matter.

R5.2.27   The elements in the chart are identified by using a legend or by labelling on the chart itself.

R5.2.28   Gridlines are removed to reduce the clutter on the graph.

R5.2.29   A chart title which explains the purpose of the graph and the time periods covered is included.

R5.2.30   All axes are labelled clearly with units of measurements stated.

R5.2.31   The right chart type in presenting information is selected:

- 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 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.

R5.2.32   A short explanation is included to help readers understand the significance of the data presented.

R5.2.33   Appropriate, suitable and consistent colour schemes are used for the charts to enable ease of reading. For example, negative results are reflected in red while positive results are reflected in green.

<u>Tables</u>

R5.2.34   The overall format of the table should be clear and concise to prevent misinterpretation. This includes checking for spacing, the wordings used, placement and appearance of titles, and other forms of labelling.

R5.2.35   Tables are formatted to ensure numbers are easy to find and understand.

R5.2.36   The data are ranked in some form of order, such as arranging data from the largest to the smallest value.

R5.2.37   Each table is labelled to identify the content of the table. The data items and unit of measurements are described clearly in the table.

R5.2.38   Footnotes should be clearly marked.

5.3   Documentation

Documentation refers to the recording of statistical activities, including concepts and methods used, and other details on the statistical processes to describe how statistical outputs are produced. 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 information can be used as evaluation and feedback for the development and implementation of future statistical activities to improve overall quality and efficiency.

Maintaining comprehensive documentation will also help to:
• Facilitate the integration of datasets for cross-analysis
• 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 strengths and limitations of data described

Requirements

S5.3.1   The entire statistical activity is documented. For example, adequate information on the meaning of the data and the methodology used to collect and process them, the type of statistical analysis and techniques used, and sources and references are provided.

S5.3.2   Documentation is updated regularly to reflect the latest developments in the statistical activity, for example, including latest amendments in definitions, classifications and methodologies.

S5.3.3   The documents are edited thoroughly before publication.

S5.3.4   Metadata are documented, archived and disseminated according to international standards. For example, a glossary of statistical concepts is made publicly available, with definitions and descriptions standardised across different publications or releases where possible.

Recommendations

R5.3.1   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
- Specifications of software application and update any developments in the software
- Information on data files, systems operations and monitoring reports
- Description of values on the data file which are derived from other values on the file

<u>Security, Confidentiality and Disclosure</u>
- Information on the legal, administrative and technical measures to safeguard data confidentiality where appropriate

<u>Data Dissemination</u>
- Appendices on acronyms used, project team information, extracts of the relevant legislation used, samples of survey forms

R5.3.2   For regular surveys, the documentation is reviewed regularly to ensure that it meets the needs of users.

III. ANNEXES

Annex 1: Glossary of Statistical Terms

Application programming interface (API): API is a computing interface which defines interactions between multiple software intermediaries. It is a tool set that programmers use in helping them create software.

Classification: A set of discrete, exhaustive and mutually exclusive observations, which can be assigned to one or more variables to be measured in the collation and/or presentation of data.

Convenience sampling: A type of non-probability sampling which involves the sample being selected because it is readily available and convenient.

Coverage: Coverage specifies the population from which observations for a particular topic can be drawn.

Coverage error: Coverage error arises from failure to cover adequately all components of the population being studied. Incomplete sampling frames often result in coverage errors.

Data: Data can be in the form of aggregated data (i.e. statistics) or microdata (anonymised, de-identified or identifiable).

Data correction: Data correction is the activity of checking data which are possibly erroneous.

Data integrity: Data integrity is the accuracy and consistency of stored data, indicated by an absence of unauthorised access or revision to data to ensure that the information is not compromised through corruption or falsification.

Data output / Statistical output: Data output / Statistical output refers to a statistical product (such as statistics, database and publication) or statistical service (such as modelling, sampling and consultancy), where identifiable information about a person or firm is not revealed.

De-identification: De-identification refers to the removal of direct identifiers to reduce identification risk.

Identifiable seasonality: Identifiable seasonality is defined as a seasonal pattern that is not dominated by irregular fluctuations.

ICT system: A set of interacting, interrelated, or interdependent ICT hardware, software and data created, accessed, stored, processed or transmitted to serve business functions and support business operations.

Metadata: Metadata is data that define and describe other data to ensure correct and proper use and interpretation of the data by its users.

Microdata: Microdata refers to individual records which can be identifiable, de-identified or anonymised.

Non-response bias: This is the bias in survey results due to incomplete or nil returns from some respondents.

Non-response rate: In sample surveys and censuses, the failure to obtain information from a designated person or firm for any reason is often called a non-response. The proportion of such cases compared to the number surveyed is called the non-response rate.

Outlier: An outlier is a data value that lies in the tail of the statistical distribution of a set of data values.

Record linkage: Record linkage refers to linking and combining individual records or data to bring together information from two or more data sources.

Response burden: The effort, in terms of time and cost, required for respondents to provide satisfactory answers to a survey.

Response rate: The proportion of respondents completing a questionnaire compared to the number surveyed. The response rate can also apply to individual questions.

Sample design: The sample design provides information on the target and final sample sizes, strata definitions and the sample selection methodology.

Sampling frame: A list of all members of a population used as a basis for sampling.

Sample size: The number of sampling units which are to be included in the sample. In the case of a multi-stage sample, this number refers to the number of units at the final stage in the sampling.

Sampling: Sampling is a means of selecting a subset of units from a target population for the purpose of collecting information. This information is used to draw inferences about the population as a whole. The subset of units that are selected is called a sample.

Stakeholder: A stakeholder is any person or organisation that has a stake in the activity.

Standard deviation: The standard deviation is a measure of the amount of differences of a set of values relative to its means and is calculated as the square root of the variance.

Statistical disclosure control: Statistical disclosure control are techniques applied to protect statistical data in such a way that they can be released without disclosing information on specific individual persons or firms. Anonymisation and tabular protection are the techniques applied to microdata and tabular data respectively.

Stratification: Stratification consists of dividing the population into subsets (called strata) within each of which an independent sample is selected. The choice of strata is determined by the objectives of the survey, the availability of variables on the frame, the distribution of the variable of interest, and the desired precision of the estimates.

Survey population: A survey population can be defined as one which excludes some units in the target population, or which comprises differently defined units through which the target population can be accessed.

Target population: The target population is the set of units about which information is wanted and estimates are required.

Validation: A continuous monitoring of the process of compilation and of the results of this process.

Variance: Variance is the mean square deviation of the variable around the average value.

For information on other statistical terms, users may refer to the sources below:

OECD, Glossary of Statistical Terms
http://stats.oecd.org/glossary/index.htm

Eurostat, Glossary-Statistics Explained
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Category:Glossary

UN Statistical Division, NQAF Glossary, 2012
https://unstats.un.org/unsd/dnss/docs-nqaf/NQAF%20GLOSSARY.pdf

Annex 2: Statistical Manuals and Guides

Classifications

Department of Statistics, Singapore, Singapore Standard Classification by the Type of Dwelling, 2012
https://www.singstat.gov.sg/-/media/files/standards_and_classifications/sctd.pdf

Department of Statistics, Singapore, Singapore Standard Classification of Household Tenure Status, 2011
https://www.singstat.gov.sg/-/media/files/standards_and_classifications/schts.pdf

Department of Statistics, Singapore, Singapore Standard Classification of Individual Consumption According to Purpose (S-COICOP), 2016
https://www.singstat.gov.sg/-/media/files/standards_and_classifications/s-coicop/s-coicop.pdf

Department of Statistics, Singapore, Singapore Standard Classification of Marital Status, 2011
https://www.singstat.gov.sg/-/media/files/standards_and_classifications/scms.pdf

Department of Statistics, Singapore, Singapore Standard Educational Classification (SSEC), 2020
https://www.singstat.gov.sg/-/media/files/standards_and_classifications/educational_classification/ssec2020-report.pdf

Department of Statistics, Singapore, Singapore Standard Industrial Classification (SSIC), 2020
https://www.singstat.gov.sg/-/media/files/standards_and_classifications/industrial_classification/ssic2020report.pdf

Department of Statistics, Singapore, Singapore Standard Occupational Classification (SSOC), 2020
https://www.singstat.gov.sg/-/media/files/standards_and_classifications/occupational_classification/ssoc2020report.pdf

Manuals & Guides by Statistical Processes

Survey Methodology

Eurostat, Survey Sampling reference guidelines 2008, Introduction to sample design and estimation techniques
https://ec.europa.eu/eurostat/ramon/statmanuals/files/KS-RA-08-003-EN.pdf

Statistics Canada, Survey Methods and Practices, 2003
https://www150.statcan.gc.ca/n1/en/pub/12-587-x/12-587-x2003001-eng.pdf?st=3FfekFIR

UN Statistics Division, Designing Household Survey Samples: Practical Guidelines 2008
https://unstats.un.org/unsd/demographic/sources/surveys/Series_F98en.pdf

Data Editing

UN Economic Commission for Europe, Statistical Data Editing Volume No.1 Methods and Techniques, 1994
http://www.unece.org/fileadmin/DAM/stats/publications/editing/SDE1.pdf

UN Economic Commission for Europe, Statistical Data Editing Volume No.2 Methods and Techniques, 1997
http://www.unece.org/fileadmin/DAM/stats/publications/editing/SDE2.pdf

UN Economic Commission for Europe, Statistical Data Editing Volume No.3 Impact on Data Quality, 2006
http://www.unece.org/fileadmin/DAM/stats/publications/editing/SDE3.pdf

Imputation

Statistics Canada, Imputation Techniques
https://www150.statcan.gc.ca/n1/pub/12-539-x/2009001/imputation-eng.htm

UN Statistics Division, Imputation in UNECE Statistical Databases Principles and Practices
https://unstats.un.org/unsd/accsub/2008docs-CDQIO/Ses4-Pap3.pdf

Dissemination

International Monetary Fund, The Special Data Dissemination Standard (SDDS)
https://www.imf.org/external/pubs/ft/sdds/guide/2013/sddsguide13.pdf

Manuals & Guides by Statistical Domains

Economic Accounts

Organisation for Economic Co-operation and Development, OECD Benchmark Definition of Foreign Direct Investment Fourth Edition, 2008
https://www.oecd.org/daf/inv/investmentstatisticsandanalysis/40193734.pdf

Organisation for Economic Co-operation and Development, Eurostat, World Health Organization, A System of Health Accounts 2011
https://read.oecd.org/10.1787/9789264270985-en?format=pdf

International Monetary Fund, Balance of payments and international investment position manual
https://www.imf.org/external/pubs/ft/bop/2007/pdf/bpm6.pdf

International Monetary Fund, BPM6 Compilation Guide, 2014
https://www.imf.org/external/pubs/ft/bop/2014/pdf/guide.pdf

International Monetary Fund, Government Finance Statistics Manual, 2014
https://www.imf.org/external/Pubs/FT/GFS/Manual/2014/gfsfinal.pdf

UN, European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, & Word Bank, Systems of national accounts 2008
https://unstats.un.org/unsd/nationalaccount/docs/sna2008.pdf

UN Statistics Division, Manual on Statistics of International Trade in Services, 2010
https://unstats.un.org/unsd/publication/seriesm/seriesm_86rev1e.pdf

Manufacturing Survey

UN Statistics Division, Handbook on Economic Tendency Survey 2015
https://unstats.un.org/unsd/nationalaccount/docs/ETS_Handbook_wCover.pdf

UN Statistics Division, International Recommendations for Industrial Statistics 2008
https://unstats.un.org/unsd/industry/Docs/IRIS_2008_En.pdf

Research and Development

Organisation for Economic Co-operation and Development, Frascati Manual 2015
https://www.oecd-ilibrary.org/docserver/9789264239012-en.pdf?expires=1601523910&id=id&accname=guest&checksum=203C469F6CE7C559350C8B5059CDB9B7

Demographic Statistics

Department of Statistics, Singapore, Census of Population 2000 Administrative Report, 2002
https://www.singstat.gov.sg/-/media/files/publications/cop2000/census_2000_admin_report/cop2000admin.pdf

Department of Statistics, Singapore, Census of Population 2010 Administrative Report, 2011
https://www.singstat.gov.sg/-/media/files/publications/cop2010/census_2010_admin/cop2010admin.pdf

UN Statistics Division, Principles and Recommendations for Population and Housing Censuses Revision 3, 2017
https://unstats.un.org/unsd/demographic-social/Standards-and-Methodds/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf

Household Income Statistics

International Labour Organisation, Report II Household income and expenditure statistics, 2003
https://www.ilo.org/public/english/bureau/stat/download/17thicls/r2hies.pdf

UN Economic Commission for Europe, Canberra Group Handbook on Household Income Statistics
http://www.unece.org/fileadmin/DAM/stats/publications/Canberra_Group_Handbook_2nd_edition.pdf

Consumer Price Index

International Labour Organisation, Consumer Price Index Manual: Theory and practice
https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331153.pdf

International Monetary Fund, Consumer Price Index Manual: Concepts and Methods (White Cover Version)
https://www.imf.org/~/media/Files/Data/CPI/cpi-manual-concepts-and-methods-white-cover-version.ashx

UN Economic Commission for Europe, International Labour Organisation, International Monetary Fund, Organisation for Economic Co-operation and Development, Eurostat, World Bank, & UK Office for National Statistics, Practical Guide to Producing Consumer Price Indices
https://www.unece.org/fileadmin/DAM/stats/publications/Practical_Guide_to_Producing_CPI.pdf

Producer Price Index

Eurostat and Organisation for Economic Co-operation and Development, Methodological Guide for Developing Producer Price Indices for Services
https://ec.europa.eu/eurostat/documents/3859598/6485893/KS-04-14-661-EN-N.pdf

International Monetary Fund, Export and Import Price Index Manual
https://www.imf.org/external/np/sta/xipim/pdf/xipim.pdf

International Monetary Fund, Producer Price Index Manual
https://www.imf.org/~/media/Websites/IMF/imported-full-text-pdf/external/pubs/ft/ppi/2010/manual/_ppi.ashx

Voorburg Group, Services Statistics
https://www.voorburggroup.org

References

ASEAN Statistics Division, ASEAN Community Statistical System (ACSS) Code of Practice, 2012
https://www.aseanstats.org/wp-content/uploads/2017/02/Code-of-Practice-ADOPTED-CLEAN.pdf

European Statistical System and Eurostat, European Statistics Code of Practice, 2017
https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7

Eurostat, ESS Quality Glossary, 2012
https://ec.europa.eu/eurostat/ramon/coded_files/ESS_Quality_Glossary.pdf

International Monetary Fund, Data Quality Assessment Framework (DQAF), 2012
https://dsbb.imf.org/dqrs/DQAF

Organisation for Economic Co-operation and Development, Quality Framework and Guidelines for OECD Statistical Activities, Version 2011/1
http://www.oecd.org/std/qualityframeworkforoecdstatisticalactivities.htm

Statistics Canada, Quality Guidelines Sixth Edition, 2019
https://www150.statcan.gc.ca/n1/en/pub/12-539-x/12-539-x2019001-eng.pdf?st=iQ_w85BR

Statistics Finland, Quality Guidelines for Official Statistics, 2nd Revised Edition
http://www.stat.fi/meta/qg_2ed_en.pdf

Statistics New Zealand, Principles and Protocols for Producers of Tier 1 Statistics, 2007
http://archive.stats.govt.nz/~/media/Statistics/about-us/statisphere/Files/tier-1/tier1-principles-and-protocols.pdf

UK Office for National Statistics, Guidelines for Measuring Statistical Output Quality Version 4.1, 2013
https://webarchive.nationalarchives.gov.uk/20160106063521/http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/guidelines-for-measuring-statistical-output-quality.pdf

UN Economic Commission for Europe, Generic Statistical Business Process Model (GSBPM) version 5.1, 2019
https://statswiki.unece.org/download/attachments/185794796/GSBPM%20v5_1.docx?version=1&modificationDate=1554283783707&api=v2

UN Statistics Division, UN National Quality Assurance Frameworks (NQAF) Manual for Official Statistics, 2019
https://unstats.un.org/unsd/methodology/dataquality/references/1902216-UNNQAFManual-WEB.pdf

US National Center for Education Statistics, NCES Statistical Standards, 2012
https://nces.ed.gov/statprog/2012/

Project Team

Mr Tang Hsiu Chin (Director, Policy Planning Division)

Ms Wong Yuet Mei (Deputy Director, Policy Planning Division)

Mr Chan Wen Chang (Senior Assistant Director, Household, Population and Longitudinal Data Analytics Division)

Mr Ho Bing Chuen (Senior Assistant Director, Economic Accounts Division)

Ms Karen Mah (Senior Assistant Director, Chief Information Officer's Office)

Mr Roger Seen (Senior Assistant Director, Census Office)

Mr Darren Ho (Assistant Director, Income and Expenditure Division)

Ms Sarah Ng (Assistant Director, Consumer Price Indices)

Mr Shaunn Tan (Assistant Director, Producer Price Indices)

Mr Wong Wah Ming (Assistant Director, Trusted Centre for Individual and Business Data)

Mr Xie Yilin (Assistant Director, Data Visualisation)

Ms Heng Xin Yu (Statistician, Policy Planning Division)

Ms Joline Tan (Statistician, Policy Planning Division)

Ms Priscilla Lim (Senior Manager, Business Statistics Division)

Ms Wong Pui Mun (Senior Manager, Communications and Engagement Division)

Ms Hammadah Hassan (Manager, Business Statistics Division)

SINGAPORE DEPARTMENT OF STATISTICS
INFORMATION DISSEMINATION SERVICES

Statistics Singapore Website

The *Statistics Singapore Website* was launched by the Singapore Department of Statistics (DOS) in January 1995. Internet users can access the website by connecting to:

www.singstat.gov.sg

Key Singapore statistics and resources are available via the following sections:
- What's New
  Obtain latest data for key economic and social indicators, browse news releases by DOS and Research and Statistics Units (RSUs) of other public sector agencies.
- Find Data
  Choose from almost 50 topics to access the relevant statistics, press releases, infographics, charts, storyboards, videos and references.
- Publications
  Browse DOS's publications, papers and articles by topics. All publications are available for free access.
- Advance Release Calendar
  View dates of upcoming releases in the half-yearly ahead calendar.

The website also provides a convenient gateway to international statistical websites and resources:
- IMF Dissemination Standards Bulletin Board
  View metadata and latest data about Singapore's key indicators in the real, fiscal, financial and external sectors, including dissemination practices and information about pre-release access of current indicators.
- International Statistics
  Access international databases, websites of international bodies and national statistical offices.
- International Classifications
  Access quick links to international economic and social classifications.

SingStat Express

Subscribe to *SingStat Express* and receive email and SMS alerts (for local users) on the latest statistical releases of your choice. Details are available at www.singstat.gov.sg/whats-new/sign-up-for-alerts.

SingStat Mobile App

The *SingStat Mobile App* provides users with on-the-go access to latest key Singapore official statistics and ASEAN statistics via their mobile devices. The app allows visualisation of data with over 200 charts of commonly used statistics from 28 data categories. Users can also opt to receive notifications on updates to their preferred indicators, organise data categories in the order of their preference, download tables or share charts via social media, perform computation for quick analyses and bookmark indicators for quick access. The app is available for downloading in iOS and Android devices. More information is available at www.singstat.gov.sg/our-services-and-tools/singstat-mobile-app.

SingStat Table Builder

The *SingStat Table Builder* contains data series from 60 public sector agencies providing a comprehensive statistical view of Singapore economic and socio-demographic characteristics. Users may create customised data tables, and export them in different file formats and download multiple tables at one go. APIs for commonly accessed tables are provided in JSON and CSV formats for ease of data retrieval. In addition, developer APIs are available for users to customise and use the parameters available to define queries. Explore data trends with this digital service at www.singstat.gov.sg/tablebuilder.

Really Simple Syndication

Really Simple Syndication (RSS) is an easy way to stay updated on the latest statistical news released by DOS. The SingStat RSS feed delivers statistical news highlights and hyperlinks to the source documents whenever the updates are posted. More information is available at www.singstat.gov.sg/whats-new/really-simple-syndication-rss.

E-survey

The *E-survey* enables business organisations to complete and submit their survey forms through the internet. Using secured encryption protocols, the *E-survey* ensures that the information transmitted through the net is secured and protected. The system features online helps and validation checks to assist respondents in completing their survey forms. With the *E-survey*, respondents do away with the tedious paper work and manual tasks of mailing or faxing their survey returns to DOS.

Statistical Enquiries and Feedback

If you have any statistical enquiries or feedback on our services, you are welcomed to:

Email us	at	**info@singstat.gov.sg**
Fax to us	at	**(65) 6332-7689**
Call us	at	**1800-3238118* (local callers)**
		(65) 6332-7738 (overseas callers)

* Calls from mobile telephone lines to 1800 local toll free number may be subject to mobile airtime charges as imposed by the relevant mobile service provider.