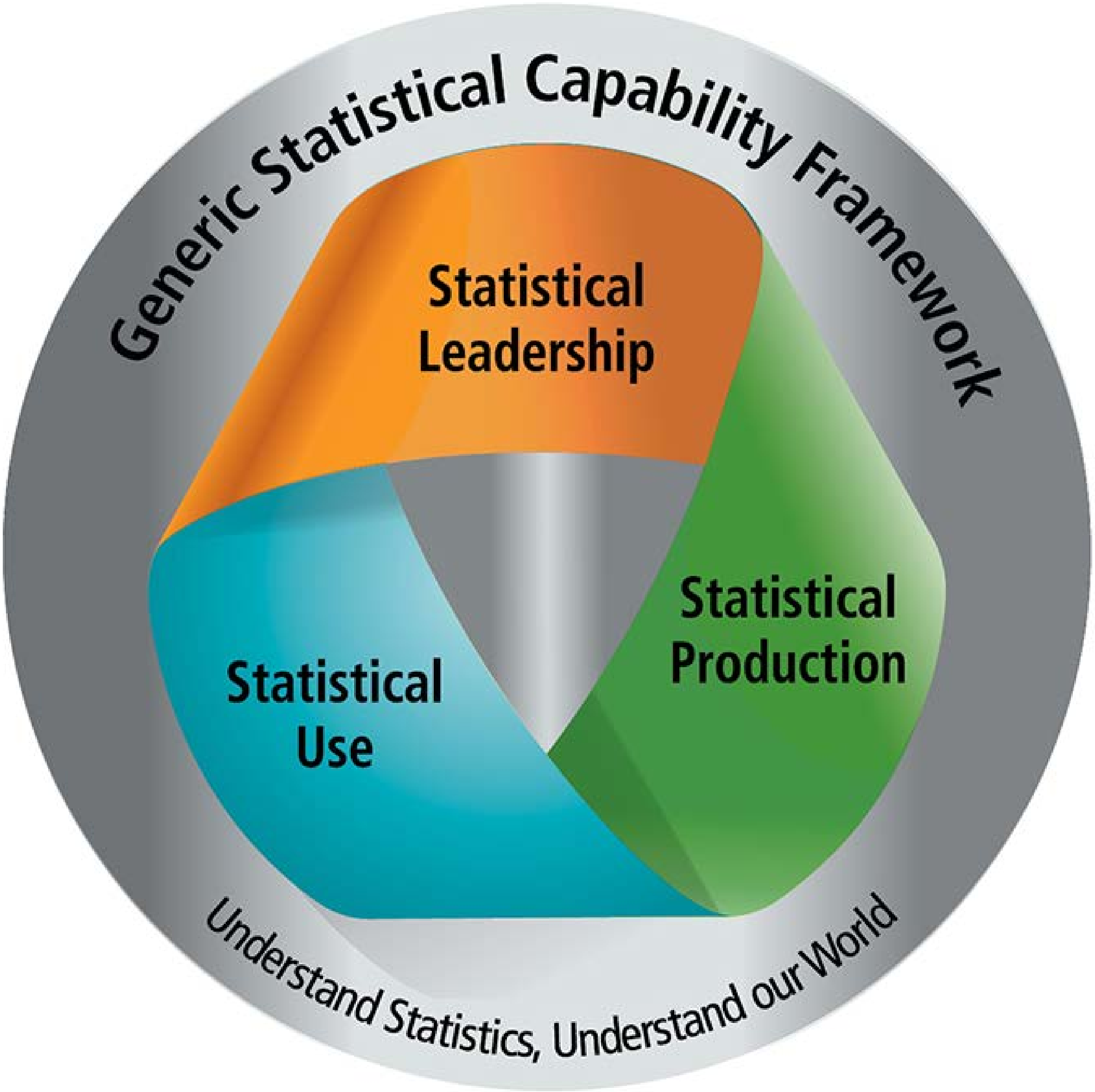
Generic Statistical Capability Framework
Statistical Leadership
Statistical Production
Statistical Use
Understand Statistics, Understand our World

“Further unleashing the potential of statistics”

Introduction

The demand for people who possess statistical capabilities across a range of disciplines continues to grow. This is a result of the increasing and ever important role of digital data and the need for evidence-based decisions. Demand is also growing due to the increasing pressure being placed on those working with data to find “innovative statistical methodologies, research techniques and software applications” (MacGillivray, 2009) in order for the potential of statistical information to be further unleashed.

Statistical capability is vital to most professions in a modern economy and is becoming increasingly important for negotiating the complexities of everyday life. But what does it mean to be statistically capable? And how can individuals and organisations build their own capability? As a starting point, it is essential to understand and clarify the capabilities required.

The Generic Statistical Capability Framework (GSCF) exists to support this understanding and clarification. It provides a common language for the broader discussion of statistical capability and is designed to be used as a guiding benchmark for organisations who are building their own statistical capability development initiatives. Like the Generic Statistical Business Process Model (GSBPM), the GSCF should be used as a guiding framework, where organisations make use of the components that are most relevant, as opposed to adopting it in its entirety or using it as a prescriptive model. The GSCF does not claim to be exhaustive list of each and every statistical capability/skill.

Purpose

The purpose of the GSCF is two-fold:

I. At the enterprise level, it provides a capability blueprint which allows organisations to assess current capability levels and develop their own enterprise wide learning initiatives. This includes capability plans aligned to the highest identified capability need/s.

II. At the individual level, it provides people with an understanding of the requirements of statistical capability. This allows staff to think more holistically about their development and career pathway.

For these reasons, the GSCF has been designed to be relevant not only to National Statistical Organisations (NSOs), but to all statistical leaders, producers, and users.

Scope

The GSCF defines and articulates a broad spectrum of capabilities required by users and producers of statistics when undertaking any kind of statistical work. It does not attempt to define the capabilities associated with statistical infrastructure or systems.

What does the Statistical Framework look like?

The GSCF comprises of three overarching dimensions:

1) **Statistical Leadership** – *Statistical Leadership refers to understanding the broader statistical environment and using this knowledge to engage and influence users and producers of statistical information to improve the quality of statistical assets and support informed decision making.*

2) **Statistical Production** – *Statistical Production refers to the activities undertaken within the statistical information system aimed at collecting data and producing statistics (OECD, 2000)*.

3) **Statistical Use** – *Statistical Use refers to the abilities required to find, evaluate, and interpret statistical information for a purpose. This includes comparing and assessing the quality of available data sources and communicating any conclusions in clear and meaningful ways that are fit for the stated purpose.*

Each dimension is then broken down according to:

- The main statistical processes within each dimension, and
- The broad set of capabilities associated with each of these statistical processes.

Dimension	**Tier 1 – Statistical Processes** This tier outlines the main statistical processes within each dimension of Statistical Leadership, Statistical Production and Statistical Use.
Statistical Leadership	**Positioning –** aligning statistical output with emerging information requirements as well as the relevant research and/or policy environment to maximise relevance.
	Influencing - shaping the local, national and international environment so that statistical information and assets are valued and used to better inform quality decisions.
	Enabling - ensuring the appropriate foundations (including capabilities) and infrastructure are in place to support the generation, management, availability, and use of high quality statistical information.
Statistical Production	**Planning** - articulating the data gap, research and/or policy need in detail. This requires scoping what information is required to address that need, scanning and assessing existing data sources and outlining how the necessary data will be obtained, collected, and/or evaluated.
	Designing and Developing - designing and developing the statistical methodologies, information technology systems, and business processes to produce the required statistical information.
	Acquiring, Processing and Data Integration - acquiring the relevant data and applying all processing requirements to it. This includes documenting associated metadata and assessing data quality at the various stages of collection and processing.
	Analysing, Validating and Evaluating - preparing the data outputs for sharing or publishing. This involves analysing and validating the accuracy of data outputs, ensuring privacy requirements are adhered to, metadata is documented, and statistical outputs are accessible in a way that takes into account the level of understanding of users.
	Disseminating – publishing the data outputs and informing relevant stakeholders of their accessibility.
	Reflecting - undertaking an evaluation of the end-to-end statistical production process to consider if it was successful, how it may be improved, and/or whether it should be undertaken again.
Statistical Use	**Defining and Articulating –** clearly defining the research and/or policy need and articulating the research questions that will be answered.
	Discovering and Assessing – finding, reconciling, and evaluating sources of information that help to understand the issue, and determining if they can appropriately inform your decision, research and/or policy need.
	Analysing and Interpreting - using appropriate techniques to find key messages in the data that will help inform your research and/or policy need.
	Communicating - presenting the key messages found in the data in a clear and accurate manner. This includes articulating the purpose of the analysis, the end decision, the research and/or policy need and any quality issues or limitations found.
	Applying – using the statistical information that has been discovered, assessed, analysed and interpreted to inform the quality decision or outcome of the research and/or policy need.

Dimension	**Tier 2 – Capabilities** This tier outlines the broad set of capabilities associated with executing the statistical processes outlined in Tier 1 relevant to the Statistical Leadership Dimension.
Statistical Leadership	**Positioning** ● Ability to: ● undertake environmental scanning to understand the broader research and/or policy context. ● identify gaps and deficiencies in available data sources, including administrative data. ● create business cases to harness critical data sets. ● set and review relevant statistical programs with key local, national and international stakeholders. ● manage, implement, and innovate within the statistical process to support improved research and/or policy decisions. ● conduct statistical work aligned to the Fundamental Principles of Official Statistics (set out by the United Nations).
	Influencing ● Ability to: ● develop systematic processes for the collection and storage of data. ● establish, maintain and leverage good working relationships with key stakeholders. ● contribute to the development and promotion of local, national and international statistical standards, classifications, frameworks and protocols. ● champion data access and facilitate data sharing initiatives across relevant statistical and non-statistical government, private and community agencies and organisations. ● negotiate agreements on the provision of data and technical services within agreed timetables and budgets. ● influence decisions on the assessment and development of statistical infrastructure and systems. ● improve data quality at the source by improving metadata standards and implementing relevant classifications to increase coherency. ● lead and connect with networks, seminars, and groups of experts to ensure consistency, coordination and collaboration of statistical activities.
	Enabling ● Ability to: ● establish, develop, and leverage relevant statistical networks. ● build the statistical understanding and knowledge of data custodians. ● assess and actively develop the statistical capability of users and producers of statistics. ● facilitate the development of necessary statistical capability learning resources. ● ensure stakeholders are aware of what data is available and how it can be accessed. ● facilitate the delivery of important local, national and international statistics to users along with related information to help them understand, use and apply the data effectively.

Dimension

Tier 2 – Capabilities

This tier outlines the broad set of capabilities associated with executing the statistical processes outlined in Tier 1 relevant to the Statistical Production dimension.

Statistical Production

Planning

- Ability to:
    - engage and partner with relevant stakeholders to understand:
        - the decision, research and/or policy aims of key data users.
        - the concept/s to be measured.
        - the variety of ways in which the outputs will be used.
        - any data gaps that need to be filled.
    - understand relevant statistical models and frameworks required to produce the statistical information (e.g. Aust. and New Zealand Standard Industrial Classification).
    - scan and assess existing data sources.
    - negotiate access to administrative or transactional data sources.
    - conduct quality assessments on datasets derived from administrative or transactional data sources.
    - develop a business case and/or project plan that includes:
        - an appropriate collection strategy,
        - prioritises the concepts/variables to be measured,
        - establishes methodological and quality parameters, and
        - a dissemination strategy.
    - identify potential quality issues and articulate any impacts on the collection process.
    - assess a planned data collection in the context of the decision, research and/or policy need to determine if the outputs will be fit for the users' purpose.
    - determine if a statistical project is necessary and/or feasible.

Designing and Developing

- Ability to:
    - Understand the relevant methodologies and processes required to:
        - develop a sampling strategy (e.g. stratification, frame, coverage, and scope of collection).
        - develop an editing strategy (e.g. micro- and macro-editing, editing methods and techniques).
        - develop an imputation strategy for the purposes of dealing with non-response or outliers (e.g. identifying data items to impute, imputation methods).
        - develop a data quality strategy (e.g. quality gates).
        - develop a data analysis plan.
        - define and document relevant metadata.
        - build and test statistical processing systems.
- In addition, when collecting data via a survey, the ability to:
    - design and build collection content (i.e. questions, data items, variable descriptions).
    - design, build, and test a collection instrument (e.g. paper form, online survey, etc.).
    - design and test training materials and instructions required for the statistical collection.
- In addition, when collecting data from administrative data sets, the ability to:
    - design and build the required database architecture.
    - design data structures, tables, dictionaries and naming conventions to ensure the accuracy and completeness of all data master files.

Acquiring, Processing and Data Integration

- Ability to:
    - match and clean administrative data records in preparation for analysis.
    - identify and resolve inconsistencies between datasets before linkage/integration.
    - extract datasets and apply appropriate data pre-processing tasks such as data cleaning, filling in missing values, or smoothing noisy data.
    - apply micro- and macro-editing techniques to identify potential errors.
    - apply weighting criteria to the sample data to create estimates for the target population.

  - apply relevant statistical models where necessary to obtain standard outputs.
  - derive new data items from existing variables.
  - confidentialise or de-identify data where appropriate.
  - document metadata and assess the quality of an administrative data set or data collection.

Analysing, Validating and Evaluating

- Ability to:
  - Effectively utilise analysis packages (e.g. SAS), and programming languages (e.g. Java) for statistical analysis and exploration purposes.
  - identify key issues in the data and dissect or isolate its components for further investigation.
  - understand and apply appropriate techniques to:
    - visualise data in tabular, graphical, or geographical format.
    - identify the main features, characteristics, and patterns in the data.
    - measure changes in key variables over time and between different subgroups.
    - measure relationships between variables.
    - create summary indicators to better communicate statistical stories present in the data.
  - confront key statistics with other data sources and identify supporting evidence to explain any differences and/or confirm similarities.
  - understand and articulate any associated data limitations (i.e. sampling and non-sampling errors, data gaps).
  - document all relevant metadata.

Disseminating

- Ability to:
  - transform output datasets, commentary, and statistics into dissemination formats.
  - publish final outputs and inform relevant stakeholders.

Reflecting

- Ability to:
  - use relevant metadata to support the evaluation of systems, processes, and procedures.
  - identify key statistical issues in the collection process and potential impacts on data quality.
  - gather feedback from stakeholders to assess the effectiveness of the statistical production process and determine if it is meeting their requirements.
  - assess the business processes of administrative data producers and identify steps in the collection and/or processing stages where data quality can be improved.
  - test new or updated processes and systems to identify and resolve potential quality issues.
  - identify and resolve inconsistencies in the acquisition and production processes.

Dimension	**Tier 2 – Capabilities** This tier outlines the broad set capabilities associated with executing the statistical processes outlined in Tier 1 relevant to the Statistical Use Dimension.
Statistical Use	**Defining and Articulating** ● Ability to: ● clearly articulate the aim and purpose of the decision, research and/or policy need. ● clearly articulate the scope of the data collection, including the population/s, geography/ies and time period/s of interest. ● formulate research questions from the aim based on the decision, research and/or policy need.
	Discovering and Assessing ● Ability to: ● locate relevant information sources and: ● understand how they differ. ● assess their ability to be linked and/or integrated. ● assess their quality in the context of the decision, research and/or policy need to determine which data sources are fit for the users' purpose.
	Analysing and Interpreting ● Ability to: ● understand relevant statistical models and frameworks used to produce statistical information (e.g. Aust. and New Zealand Standard Industrial Classification). ● understand and apply appropriate techniques to: ● link or integrate data sources. ● visualise data in tabular, graphical, or geographical format. ● identify the main features, characteristics, and patterns in the data. ● measure changes in key variables over time and between different subgroups. ● measure relationships between variables. ● create statistical models to test a hypothesis. ● identify an appropriate conclusion, finding, recommendation or prediction based on the information extracted from the data. ● find and use relevant metadata to ensure any interpretation is appropriate. ● use data quality to apply risk management principles to any interpretation.
	Communicating ● Ability to: ● describe key information using easy to understand language and visualisations. ● describe any data linking or integration that has been undertaken. ● relate key messages to the decision, research and/or policy issue under consideration. ● apply statistical reasoning and take into account the quality of the data when reaching conclusions/decisions. ● clearly articulate the limits of the data due to quality implications.
	Applying ● Ability to: ● Develop risk mitigation strategies based on data gaps or quality issues identified in the analysis.

Glossary

Note: the following definitions are explained in the context of the production or use of statistical information.

Accessibility - the ease and the conditions with which statistical information can be obtained.

Accuracy - Closeness of computations or estimates to the exact or true values that the statistics were intended to measure.

Acquisition - an asset or object bought or obtained.

Administrative data - refers to information collected primarily for administrative (not research) purposes. This type of data is collected by government departments and other organisations for the purposes of registration, transaction and record keeping, usually during the delivery of a service.

Aligning - give support to (a person, organization, or cause).

Articulating - pronounce (something) clearly and distinctly.

Business case - a justification for a proposed project or undertaking on the basis of its expected commercial benefit.

Champion (verb) - vigorously support or defend the cause of.

Characteristic - a feature or quality belonging typically to a person, place, or thing and serving to identify them.

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.

Coherence - Coherence of statistics is their adequacy to be reliably combined in different ways and for various uses.

Collaboration - the action of working with someone to produce something.

Collection instrument - the device used to collect data, such as a paper questionnaire or computer assisted interviewing system.

Commentary - a set of explanatory or critical notes that describe a dataset or set of statistics.

Component - a part or element of a larger whole.

Concept - an abstract idea or a unit of knowledge created by a unique combination of characteristics.

Conclusion - a judgement or decision reached via the process of reasoning.

Confidentialise - a set of rules or a promise that limits access or places restrictions on certain types of information.

Confidentiality - a property of data, usually resulting from legislative measures, which prevents it from unauthorized disclosure.

Consistency - consistent behaviour or treatment.

Coordination - the organization of the different elements of a complex body or activity so as to enable them to work together effectively.

Data analysis - the process of transforming raw data into usable information, often presented in the form of a published analytical article, in order to add value to the statistical output.

Data collection - the process of gathering data.

Data confrontation - the process of comparing data that has generally been derived from different surveys or other sources, especially those of different frequencies, in order to assess their coherency and the reasons for any differences identified.

Data custodians – those responsible for the safe custody, transport, and storage of data as well as any implementation of relevant business rules.

Data dictionary - a set of information describing the contents, format, and structure of a database and the relationship between its elements, used to control access to and manipulation of the database.

Data integration - combining data coming from different sources and providing users with a unified view of these data.

Data linkage - refers to the task of finding records in a data set that refer to the same entity across different data sources (e.g., data files, books, websites, databases).

Data master files - a collection of records pertaining to one of the main subjects of an information system, such as customers, employees, products and vendors. Master files contain descriptive data, such as name and address, as well as summary information, such as amount due and year-to-date sales.

Data quality - Data quality refers to the level of quality of Data. There are many definitions of data quality but data is generally considered high quality if, "they are fit for their intended uses in operations, decision making and planning." (Tom Redman<Redman, T.C. (2008).

Data set - any organised collection of data.

Data source - a specific data set, metadata set, database or metadata repository from where data or metadata are available.

Data structure - a specialized format for organizing and storing data. General data structure types include the array, the file, the record, the table, the tree, and so on. Any data structure is designed to organize data to suit a specific purpose so that it can be accessed and worked with in appropriate ways.

Database - a logical collection of information that is interrelated and that is managed and stored as a unit, for example in the same computer file. The terms database and data set are often used interchangeably.

Database architecture - database architecture focuses on the design, development, implementation and maintenance of computer programs that store and organize information for businesses, agencies and institutions.

Decision - a conclusion or resolution reached after consideration.

Deficiencies - failing or shortcoming.

De-identify - the process used to prevent a person's identity from being connected with information. Common uses of de-identification include human subject research for the sake of privacy for research participants.

Derive - to construct a new data element from other data elements using a mathematical, logical, or other type of transformation, e.g. arithmetic formula, composition, aggregation.

Dissect - analyse data in minute detail.

Dissemination - the release of information obtained through a statistical activity to users.

Editing - the activity aimed at detecting and correcting errors (logical inconsistencies) in data.

Effectiveness - the degree to which something is successful in producing a desired result; success.

Engagement - initiating a dialogue with stakeholders to find out what matters most to them in order to improve processes and decision-making.

Estimation - estimation is concerned with inference about the numerical value of unknown population values from incomplete data such as a sample

Evaluation - the making of a judgement about the amount, number, or value of something; assessment.

Facilitate - make (an action or process) easy or easier.

Feasible - possible and practical to do easily or conveniently.

Finding - information discovered as the result of an inquiry or investigation.

Geographical - information about places on the Earth's surface, knowledge about where something is, or knowledge about what is at a given location.

Graphical - relating to or in the form of a graph.

Hypothesis - a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.

Implement - put (a decision, plan, agreement, etc.) into effect.

Imputation - a procedure for entering a value for a specific data item where the response is missing or unusable.
Interpretation - the action of explaining the meaning of something.

Isolate - identify (something) and examine or deal with it separately.

Leverage - use (something) to maximum advantage.

Limitations - a limiting rule or circumstance; a restriction.

Macro-editing - a procedure for tracking suspicious data by checking aggregates or applying statistical methods on all records or on a subset of them.

Metadata - a set of data that defines and describes other data.

Methodology - a structured approach or system of methods used in a particular study or activity to solve a problem.

Micro-editing - an exhaustive check to find errors by inspecting each individual observation.

Naming convention - a set of rules for choosing the character sequence to be used for identifiers which denote variables, types, functions, and other entities in source code and documentation.

Non-response - the failure to obtain a measurement on one or more study variables for one or more elements selected for the survey.

Non-sampling error – any errors caused by factors other than those related to **sample** selection.

Outlier - a person or thing differing from all other members of a particular group or set.

Output - what is produced directly or immediately by an activity.

Parameter - unknown, quantitative measure (e.g., total revenue, mean revenue, total yield, number of unemployed) for the entire population or for specified domains which are of interest to the investigator.

Patterns - a regular and intelligible form or sequence discernible in the way in which something happens or is done.

Policy - a course or principle of action adopted or proposed by an organization or individual.

Prediction - thing predicted; a forecast.

Privacy - the state of being free from public attention.

Project plan - a discipline for stating how to complete a project within a certain timeframe, usually with defined stages, and with designated resources.

Recommendation - a suggestion or proposal as to the best course of action, especially one put forward by an authoritative body.

Research - the systematic investigation into and study of materials and sources in order to establish facts and reach new conclusions.

Risk management - the forecasting and evaluation of probable adverse conditions and/or events together with the identification of procedures to avoid or minimize their impact.

Risk mitigation - the process of taking steps to reduce adverse effects.

Sampling - the process of selecting a subset from a frame where elements are selected based on a randomised process with a known probability of selection.

Sampling error - the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a sample of values is observed.

Scope - the coverage or sphere of what is to be observed. It is the total membership or population of a defined set of people, object or events.

Stakeholders - any person or organisation with an interest or concern in something.

Statistical model - a probability distribution constructed to enable inferences to be drawn or decisions made from data.

Statistical reasoning - the action of thinking about something in a logical, sensible way.

Statistical standard - a comprehensive set of guidelines for surveys and administrative sources collecting information on a particular topic.

Subgroup - a subdivision of a group.

Summary indicator - a thing that indicates the state or level of something.

Tabular - (of data) consisting of or presented in columns or tables.

Technique - a way of carrying out a particular task, especially the execution or performance of an artistic work or a scientific procedure.

Time period - the time interval of single repetition of a varying quantity of a motion or phenomenon which repeats itself regularly.

Transactional data - data describing an event (the change as a result of a transaction). Transaction data always has a time dimension, a numerical value and refers to one or more objects (i.e. the reference data). Typical transactions are: Financial: orders, invoices, payments.

Validity - the extent to which a concept, conclusion or measurement is well-founded and corresponds accurately to the real world.

Variable - a characteristic of a unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, weight, etc., and “occupation”, “industry”, “disease”, etc.).

Weighting criteria - allowance or adjustment made in order to take account of special circumstances or compensate for a distorting factor.

Useful resources

Fundamental Principles of Official Statistics

http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

Statistical Skills for Official Statisticians

http://www.abs.gov.au/ausstats/abs@.nsf/mf/1125.0

National Statistical Service

http://www.nss.gov.au/

Stat Trek

http://stattrek.com/statistics/resources.aspx