Data Quality Management Guide in the Department of Statistics (DoS) 2017

Prepared by: The Data Quality Division

Introduction

The importance of Data Quality in statistical work stems from the importance of using the statistical data in various fields (studies and research). Therefore, many parties have advocated the need to focus on Quality Control in producing the statistical figures. This theme has been studied from different angles, some were approved but others were not. However, although there are different views, some criteria and dimensions should be taken into consideration when producing statistical data.

Although there are many statistical studies and researches that discuss data quality, it has no clear and specific concept. In any case, they all refer to a scale based on preferring statistical data compared to other statistical products and services. Data quality in statistical work refers to all areas and aspects related to the extent to which these data meet the user's need and response to his expectations in terms of performance, level, format and content, in addition to high level of information accuracy in all stages in line with the requirements of data users and decision makers. Data quality is defined by the European Statistics Office (EUROSTAT) as a set of features and characteristics in the statistical product or service that meets and satisfies the user's needs. The International Monetary Fund (IMF) defines Data Quality within the Data Quality Assessment Framework (DQAF) through a number of dimensions and levels required to achieve the data quality level.

Data quality in statistics is of great value as it is the approved basis for policy and decision making by the public and private sectors. The Statistical Institutions handbook issued by the United Nations Statistics Division (UNSD) indicated that the statistical agencies should take some significant matters into consideration in their capacity as the authorized bodies to produce and publish official statistics including giving priority to data quality, in order to gain the data users' confidence through discussing the methodology, commenting on the data quality and allowing them to examine them.

In viewing the basic and significant data quality criteria and related outputs, the main issue here is how all statistical work stages were handled and how the problems were addressed. We can say that quality of statistical work includes the theme or statistical work objective, work organization, sampling design, design and testing of the questionnaire, data collection in addition to processing, analysis and data dissemination. In other words, data quality is achieved through the following important elements (relevance to reality, confidentiality, accuracy, timeliness, accessibility, comparability, consistency and completeness).

Although the above quality elements are general elements and apply to all types of statistics, but their management by a particular agency or a certain type of statistics varies according to a set of determinants, most significant of which are the institutional and legal frameworks through which the statistical agency operates, including human resources and available techniques in data collection and dissemination and the international standards to which the statistical agency is committed in addition to financial resources.

Objectives of statistical work

Defining the objectives of statistical work and its implementation is very essential since the high quality of any work lies in its implementation … why and what is the benefit that we can draw out of it? All available options for obtaining the product of this statistical work should be taken into consideration to determine the best way of implementing this work. Planning during the implementation stage coupled with sufficient knowledge and studying of past experiences (internal or external) in order to benefit from them and to avoid any errors and obstacles encountered in these experiences. We must study the possibility of conducting the statistical work in terms of available resources and whether it is possible to obtain high-quality results in order to fulfill the purpose of carrying out such work.

Sample Design

Designing the appropriate sample is an important priority in the data quality management, since the sample survey is supposed to give a picture of the total community you will study. There are three basic elements to be met when designing the sample. What is the study community? Who are the covered elements (individuals, households or institutions)? Whether frameworks are available for this community on the basis of which the sample will be selected? It is worthy to mention that the framework for selecting the required sample should be comprehensive and modern that does not contain duplicate or missing units

because using an outdated and inaccurate framework will reflect negatively on the quality and accuracy of statistical data.

The best sampling design is to strike a balance between results' accuracy and the survey cost. The low cost surveys are possible provided that accuracy is preserved with least errors. The sampling errors occur because data are collected through a sample and not through a comprehensive survey of the community. The sampling errors are attributed to all sampling stages such as; sample design, sample size estimation and sample distribution which are measurable. The use and adherence to accurate rigorous scientific basis increases the sample effectiveness and accuracy of the resulting estimates. The calculation of sampling errors in previous surveys helps to determine the optimum design of the sample as well as optimal sample size. We should also take into account the sample distribution on various community strata and segments, so that the survey is a miniature of the study community by using the scientific methods, including the probability sampling methods where errors can be calculated and measured.

Moreover, among the errors affecting data quality which leads to biased estimates is resulting from deletion or repetition of a number of sampling units. This error is closely related to the sampling frame. As mentioned earlier, as much as the sampling frame is comprehensive and recent as much as the sample is representative and the resulting estimates are unbiased. The inclusion of error leads to biased estimates due to lack or repetition of sampling units or variance in calculating the required estimates from this survey due to adding some unnecessary items in the frame or presence of some errors in the auxiliary

information in the frame that are used to assist in sample withdrawal noting that they affect data quality and accuracy.

Questionnaire

Questionnaire design is considered as one of the most important statistical work stages. It is essential to set clear bases for the questionnaire components and the themes to be covered in the survey by developing the concepts, contents and topics to be included in the questionnaire, in addition to having clear and simple questions, selecting easy and suitable and consistent language to obtain correct and accurate responses. Questionnaire's length and large number of questions have significant impact on data quality and response rate. The questionnaire should be consistent in a way that ensures detecting the respondent's errors directly during data collection and there must be investigative questions or the so-called the "*Error Detection* or *Information Checking Questions"*, especially in the case of surveys that collect data on past topics and events where the respondent is repeatedly asked to give the correct information or some of the data collected from old people. Additionally, the questions should not have some bias to a particular answer or existence of a bias through the questions by the interviewer.

Pre- Test

Data high quality lies in establishing a pre-test work mechanism for the survey questionnaire and for fieldwork execution mechanism, especially when there is a new mechanism for implementing the work or using a new sampling method

never been implemented before. The pre-test includes testing all tools used in the survey such as; questionnaires and data entry, edit and other software, in addition to testing the work plan in terms of work mechanism, time required to conduct each interview, how to reach the respondent and so on in order to reach the acceptable and applicable situation. This pre-test is applied on all segments of respondents and should not be limited to be implemented in one area or on one type of respondents for identifying all problems and obstacles in order to avoid them while implementing the actual survey.

Non-Response

It means dropping of a number of units because they do not respond (provide information) or dropping some of the questionnaire items due to inability to complete their related data, which makes the sample or the questionnaire items incomplete. This affects the sample ratio and accuracy of the estimates and thus leads to a bias resulting from non-response. In order to minimize non-response, we try to make use of the pre-test results, especially the questionnaire's questions and to make them clear to the respondent. In order to minimize non-response, it is preferable to send messages and reminders to the selected sampling units by post, e-mail, face to face or visiting the non-responding units again at different times.

Staff Training

In order to ensure data quality, it is recommended to train the staff well on how to find the selected sampling unit and data collection methods prior to fieldwork, as

well as to train them on conducting the interviews and collecting information correctly. It is known that there are several ways to collect data, including personal interviews or data collection by telephone or by mail, so we must be careful and train the staff well according to the method to be used. The interview needs special skills in dialogue and how to earn the respondents' confidence through a friendly atmosphere by the data collector. Moreover, the staff must be trained on how to ask questions, non-bias and avoiding any hint for a specific answer that may affect the response. Staff must be trained also on how to act in case any rejection by the respondent or any other problem.

Work Edit at All Stages

The high quality and accuracy of any statistical data means checking each stage of the statistical work to ensure that data collection, edit, coding and keying are done correctly for all selected sampling units. It is necessary to control the fieldwork phase by attending the interviews and re-interviewing to make sure that data is collected correctly. All questionnaires are checked during the fieldwork phase for easy data correction again. The statistical tables and cross-tables of results should be prepared because they help in the process of data checking and non-weighted data comparisons with the weighted data.

Post Enumeration

It is an important and fundamental stage of the statistical work, because it presents an accurate picture of the accuracy and credibility of this work, especially the field work, coverage measurement and survey comprehensiveness.

We take a sample of the survey sample which was used. The work team for this operation must be independent in order to collect data of some parts of the questionnaire for being compared with the data collected previously to verify the information accuracy contained in the questionnaire and to assess the interview accuracy by the interviewer, determine the error size and identify coverage and comprehensiveness errors.

Data Evaluation and Dissemination

Data comparison is an important and fundamental stage in evaluating and improving the data quality because comparison with data from other sources, external or internal, is a significant step that must be done during the data testing, evaluation, specifying the sampling and non-sampling errors, response and coverage errors. This is done after the data processing stage is completed and prior to the dissemination stage.

Comprehensive, accurate and timely data dissemination process is an important step in data quality, as statistical data lose its relevance and quality if do not reach the decision makers, policy planners and data users in a timely manner. The data user needs information about data quality he or she will use to arrive at reliable decisions and results. It is advisable to publish comprehensive information on data quality along with the published data so that the user can provide the necessary explanations and analysis. Furthermore, the complete methodology for implementation of the statistical work should be published at all stages and the results and analysis of these data should be published objectively and

transparently. Individual confidentiality must be respected by refraining from giving any individual data for increasing the respondent's confidence in statistical work.

Jordan's Experience in Statistical Data Quality Management

The Department of Statistics (DoS) is the main and funding source of statistical data in Jordan and produces a variety of data through censuses including; population, economic and agricultural. Over time, data collection has evolved to include many important topics, including economic issues (industry, construction, tourism, unemployment and new job opportunities created by the Jordanian economy, GDP and indices) in addition to health issues (reproductive health, child health, health care, fertility and mortality). The (DoS) keeps pace with all new and scientific methods, both globally and regionally, with the aim to produce accurate and quality data that are provided periodically by data users, decision and policy makers. The (DoS) endeavors to achieve the important data quality components (i.e., relevance, confidentiality, accuracy, timeliness, accessibility, comparability, consistency and comprehensiveness).

Given the use of the statistical figure produced by the (DoS) as the official and basic figure for planners, decision-makers and policy planners to develop plans, policies and studies, therefore the quality of this figure is of great importance for arriving at conclusions, recommendations, programs, plans and policies. Thus, the (DoS) views data quality and accuracy as priority for the following factors:

- Gaining data users' confidence, increasing confidence in the (DoS) and its statistical products and increasing the use of data issued by the (DoS) as

the main reference because any doubt by the users leads to lack of credibility in the (DoS).

- Encouraging the culture of using the statistics produced by the (DoS).
- Producing error-free figures and closer to the truth.
- Increasing the data quality culture and spreading data-accuracy keenness among the (DoS) employees during all processes and stages of the statistical work.

The (DoS) achieves a high level of accuracy and quality in all stages of statistical work as follows:

1 - Preparatory stage

✓ *Planning*

During this stage, the survey planners should have sufficient knowledge of previous experiences, whether internal or external, to benefit from them and to avoid the errors and obstacles encountered in these experiments. Usually, a highly experienced team is formed and staffers who contributed to previous surveys are integrated to have more than one opinion for reaching the ideal work situation. A study then is carried out on the feasibility of carrying out the statistical work, whether it can be done in terms of the resources available in the (DoS) and whether it is possible to obtain high-quality results in order to fulfill the purpose of this work.

✓ *Users' View*

The (DoS) gives great importance to the data users' opinion to identify their views so that the statistical work meets their needs and requirements. To this end, the (DoS) forms technical committees for the surveys to be conducted and includes representatives of the main data users. These committees usually have a role in designing the survey sample and the questions of the questionnaire. In the case of new surveys, the (DoS) holds workshops for data users from different sectors to receive their views before the survey begins.

- ✓ ***Sample***

The sampling specialists coordinate with the division responsible for designing and selecting the required samples in order to study the sample required for this work in terms of accuracy level and budget availability for carrying out this work at the lowest possible costs. Sampling errors and calculation of sampling errors for previous surveys to determine the optimal sample design and size are taken into consideration. We should also take into account the sample distribution on different segments and strata of the community noting that the community was divided into four strata depending on many demographic, social and economic characteristics provided by the General Population and Housing Census-2015 so that the survey becomes a miniature of the study community. The scientific methods and techniques are adopted in the design and selection of the sample items in addition to benefiting from the international expertise in this field.

In order to ensure designing of an effective sample, the available frameworks at the (DoS) that will be used to select the sample items are updated through periodic censuses in Jordan (General Population and Housing Census, Economic

Establishments Census and the Agricultural Census) to ensure that the framework is comprehensive and recent without having repetition or lack of items. Therefore, the (DoS) makes comparisons between the framework data and any other available data whether from inside or from outside the (DoS) (i.e., other institutions that have data or information on the sampling units used, for example; the Ministry of Industry and Trade and the Social Security Corporation). An updated, comprehensive and accurate framework is being used in order to avoid coverage errors through the work of a trained and highly qualified team on all data necessary for updating the framework of some population localities or economic establishments within the selected sample shortly before implementing the survey and approving this framework to withdraw the required sampling units to ensure coverage. Therefore, the recent and comprehensive framework of all units of the study community is the basis for reducing the coverage error and improving data quality.

- ✓ ***Questionnaire***

The quality of the data collected depends mainly on questionnaire design. The (DoS) takes this criterion into account by selecting clear, simple and logical questions using clear and uncomplicated terms so that it becomes easier for the data collector to ask the respondent and get correct and accurate answers. In addition to that, lengthy questionnaire has a great impact on the quality of the data collected. A specialized technical team consisting of all related (DoS) staff, members of governmental and non-governmental organizations in addition to the

private sector is formed to check the questionnaire and give any feedback. The questionnaire is then evaluated through conducting the Pre-Test.

✓ *Pre-Test*

The Pre-Test is usually conducted to select the survey tools and to execute the field work mechanism. The questionnaire is evaluated and field tested through collecting data in certain selected areas (not within the main sample areas of the survey) to avoid non-response later. The work plan is tested in terms of field work implementation method, efficiency of data collectors, time required to conduct each interview, how to reach the respondent and so on in order to reach the acceptable and applicable situation. Thereafter, all initial survey stages (checking, coding and automatic data entry) are completed to test the data entry software in terms of the average time required for data entry of each questionnaire, testing the automated edit rules and consistency of questions. This helps in identifying all problems and obstacles to avoid and modify them in the questionnaire and in the automated programs.

✓ *Preparation of Guides*

The survey supervisors and specialists prepare the guides for all stages which will help to organize and facilitate the work. These include:

- ***Interviewer's Guide*** (*Data Collector*): It includes clear definitions and explanations of the concepts and terminology in the questionnaire in

addition to the questionnaire completing mechanism and the interview conducting instructions with the respondent and how to gain his/her confidence and selection mechanism of respondents. This guide can be used as a reference by the interviewer in case of any problem or confusion during data collection.

- ***Controller's Guide:*** It includes the tasks assigned to him, how to run the field work and the methods to be followed, in addition to checking methods of interviewers work under his supervision (usually there is a Controller supervising the work of three to four interviewers).
- ***Office Preparation Guide***: It includes clear instructions of the edit process and how it is performed by the "Edit Team" in the *Office Processing Division* noting that these are specific rules for verifying the validity and consistency of statistical data for all questionnaire items. In addition to that, it includes introduction of international standards adopted by relevant international organizations in the field of questionnaire coding, for example: *economic activity:* "main professions, academic specialization or any other questions that need to develop certain codes for them."
- ***Automated Edit Rules Guide:*** The automated edit rules are set up by the technical staff with the help of specialized programmers for detecting the errors in the questionnaire items and their consistency with each other to achieve the highest accuracy, so that the survey data becomes ready for tabulation and analysis within a short period after completion of data entry.

✓ ***Selection and Training of Employees***

The (DoS) views the interviewer as the basic element in any survey, therefore they are selected carefully noting that all of them are university graduates and their specializations are required to implement the survey. The (DoS) also trains them extensively on all questionnaire items in addition to educating them on data collection methods where he/she should create an atmosphere of mutual trust with the respondent in addition to respecting the customs and traditions. The (DoS) also emphasizes that the interviewer should maintain a decent appearance (clean and simple clothing) and should produce his/her documents issued by the (DoS) to the respondent to obtain accurate, reliable and credible data. They are also trained on how to locate the selected sampling units to ensure non-bias or working in non-required sample units as well as finding an alternative unit in some cases based on scientific method. Duration of training varies depending on the questionnaire's length and complexity but usually it takes one week and sometimes up to a month as in the *Population and Family Health Survey (PFHS)*.

The *Data Quality Division* has evaluated the training process through a questionnaire directed towards the trainees and included "*Satisfaction with the training environment, training material and satisfaction with the trainers*." This questionnaire was done electronically to save time, effort and money noting that the analytical report is in process.

2 - Data Quality during Field Work

The (DoS) endeavors to control the fieldwork stage through checking and monitoring according to job sequence of each person in the survey. The

organizational structure of each survey management consists of the following levels: interviewer*, controller, field supervisor, office checking, coding and automated preparation*. Random samples from the work of each data collector are selected to ensure that the work is done correctly. This process is performed by the controller and the supervisor, in addition to attending some data collection interviews and watching the manner of asking questions to the respondent.

✓ *Post Enumeration*

The (DoS) is always keen to conduct post-enumeration survey in many implemented statistical activities, especially in censuses, due to its importance in painting a picture related to the accuracy and credibility of the work particularly measuring the coverage and comprehensiveness.

✓ *Non-Response*

The (DoS) is working to develop a strategy to minimize the total non-response (inability to obtain all the questionnaire data) or partial non-response (inability to obtain data of one item or more of the questionnaire) through:

- Communicating with the respondents through reminder messages and explaining the importance and objectives of statistical work.
- Training the interviewers on how to deal with non-response, interview management and asking questions.
- Work flexibility through working outside the official working hours to adjust with the times of respondents.

- Using questions with easy wording that are understandable to the respondent in addition to balancing between the questionnaire length and topics covered, which reduces the burden on the interviewees.
- Studying the non-response rate and taking it into account when processing data by using the scientific method in compensation and rectification.

3 - Data Quality during the Data Processing Stage

The checking process is also important and fundamental to improve data quality during the data processing phase.

✓ *Office Processing Stage*

There is an Office Checking Team to ensure that the questionnaires are error-free noting that the team should exchange the checked questionnaires among themselves and the person in-charge of the checking process must get a sample of the checked questionnaires to make sure there are no errors made by his/her team. Likewise, there is also a Coding Team working in the light of the special guides for this process noting that the team should exchange the coded questionnaires among themselves and the person in-charge of the coding process must get a sample of the coded questionnaires to make sure there are no errors made by his/her team. In case of any unclear or vague errors, the person in-charge of the coding process must be contacted for interpretation and then all coding employees should be notified to avoid repetition of errors.

A sample of the completed questionnaires in both above stages is taken by a specialized team that has the ability to detect errors through linking the questions with each other in the checked and coded questionnaires to ensure reducing the number of errors as possible before moving to the next stage.

✓ *Automated Processing Stage*

The programmers begin to prepare the data entry software after the specialized technical staff completes preparing the survey questionnaire to be ready for field work, so that data is entered simultaneously with the data collection and office processing stages. Usually, selection of the appropriate program is taken into consideration to ensure that entry of the survey data is done with high precision and speed. The programmers test the *Automated Edit and Consistency Rules* through an experimental data entry operation. Moreover, the data entry staff is trained by explaining the questionnaire well to them in order to solve any difficulties or problems encountered during data entry and that they must refer to the Survey Liaison Officer to solve any problems that a data entry employee or programmer cannot solve. The programmer prepares the statistical tables and simple cross-tables that show the data consistency level for analysis and publication purposes.

Usually, data re-entry is made at the rate of 10% as a checking process to ensure that the data entry was done correctly with the aim to achieving accuracy and high quality of data produced by the (DoS). The criterion for data re-entry is the error ratio (if the ratio is less than 5%, only 10% of data entry is made). If the

ratio is more than 5%, the data will be re-entered at 100% to ensure data accuracy and quality.

- ✓ ***Data Storage Stage***

  After completion of the data entry process, the data is stored according to localities or serial numbers in a place that is accessible easily by employees when needed, away from any damage that may result from the storage process.

4 - Data Quality during the Evaluation and Dissemination Stage

The (DoS) carries out an assessment process of the survey results before providing them to users where data are compared with other reliable data sources because some surveys are the only data source. If data are available from another source, an assessment of the data accuracy can be done. For example, comparison of data produced by the (DoS) with data obtained from other surveys or data from administrative records such as; Civil Status Department, Ministry of Labor, legal courts, Social Security Corporation, Ministry of Industry and Trade, Central Bank and Ministry of Health. The *Variance Coefficient* and *Sampling Errors* are calculated noting that calculation of sampling errors is one of the data quality principles that are internationally recommended to be published with the results of any survey or study. One of the measures that are calculated and published by the (DoS) is the *Variance Coefficient* or the so-called *Relative Standard Error*, to ascertain the possibility of comparison among the surveys or studies, as well as the *Standard Error* and the effect of the *Sample Design* and

the *Estimation Confidence Intervals* according to many characteristics that benefit the data user to have a vision and analysis of many indicators. The survey results are published through issuance of a comprehensive report on the statistical work methodology in all its stages in addition to the sampling errors, design methodology and sample selection. The survey results or any statistical data produced by the (DoS) can be accessed on the web page which is always updated continuously.

Recent Data Quality Experience of the (DoS)

In 2007, the (DoS) has carried out the General Agricultural Census noting that for the first time, data is collected using PDAs. Data on agricultural holders are collected automatically in the fieldwork stage. This mechanism saves time and cost by excluding the employees who were engaged to check, code and enter the data of the paper questionnaires and also is supposed to ensure accuracy. The (DoS) has followed all procedures and rules that support the automated data collection process to implement the census on time and with high data quality.

In 2015, TABLETS were used for the first time for data collection in implementing the *General Population and Housing Census*. The use of information technology at all census stages had a significant impact on the data quality control and accuracy. The following are the most important data processing methods used during the census stages:

- In the *Demarcation Stage*, the GPS and the use of GIS maps helped to control and manage the field work, as these systems assist to monitor and

track enumerators directly. These systems helped also to prevent data duplication by identifying work areas on maps and linking them to enumerators.

- In the *Listing and Count Stages*, the programs and screens design had a clear impact on immediate data processing by controlling movements between questions in addition to the possibility of hiding and showing some questions in conformity with the consistency rules immediately during data collection that prevented entry of wrong data.
- The *Call Center* is one of the methods used to improve data quality and integration by calling a sample of households to verify the collected data accuracy.
- The delivery of immediate reports to the main headquarters from the field directly enabled the census administration to monitor the fieldwork and to evaluate data quality and accuracy.
- Experienced (DoS) team members were able to rectify any problem directly in a timely manner.

The (DoS) has also implemented the *Agricultural Census* in 2017 using information technology at all stages. Agricultural holders' data was collected using TABLETS for the first time after the successful implementation of the *General Population and Housing Census-2015.*

A special team has been formed to monitor the data quality of the censuses. This step is of great value in order to improve the level of cooperation and coordination within the (DoS) for achieving the common objectives. The *Data*

Quality Team is entrusted with the planning and implementation process that leads to data quality in various censuses. They must be aware of all activities and to delegate some powers to them for executing their work properly.

Joint efforts were exerted between the censuses' administration led by the Census National Director/Director General of Statistics and the Quality Team in matters covering quality assurance, performance control and implementation of performance improvement plans. Brief and regular reports with key performance indicators were prepared so that the Higher Administration is always informed of the situation. There should be an indication whether the results of the short-term plans are consistent with our goals if we want longer-term plans and strategic goals to be realized.

Expectations and Future Plans of the (DoS)

To conclude, the (DoS) is the sole officially authorized institution to produce statistical data (e.g., demographic, social, economic, health … etc.), therefore it always seeks to achieve the highest data accuracy and quality levels. Moreover, it aspires to win the King Abdullah II Award for Excellence and maintains friendly relations with all Arab and international statistical agencies and data users in Jordan.