Journal of Economics and Sustainable Development
ISSN 2222-1700 (Paper) ISSN 2222-2855 (Online)
Vol.10, No.22, 2019

Approximated Synthetic Indicator for Total Errors in Abu Dhabi Sample Surveys

Dr. Mohammed Al Rifai
Statistics Centre – Abu Dhabi, P.O. Box: 6036 Abu Dhabi, UAE

Abstract
A central role of national statistical organizations is to support the comprehensive development of their countries, through the provision of reliable high quality official statistics. There are different definitions for "quality" because of differences in the objectives sought. The concept of quality is no longer limited to the product-accuracy level, but has become more far-reaching, in a way that it comprises other quality dimensions, such as institutional arrangement supporting quality, data accuracy, relevance, clarity and interpretability, sound methodology, timeliness and punctuality, data accessibility, and consistency and coherence**.** End users mostly concentrate on the data accuracy dimension. The Statistics Centre Abu Dhabi (SCAD) has used to evaluate the data quality, including accuracy dimension in each implementation phase of the sample survey, based on a pre-determined checklist of quality standards that should be achieved. The outcome of this assessment method is a nominal measure to classify the quality in each survey phase as a Low, Medium, or High level of quality. Moderate data users and policy makers concentrate more on a quantitate measure of the total survey error, rather than a subjective or nominal measure.This paper aims to propose statistical methodology to compile approximated synthetic indicator to measure the total survey errors based on the different sources of statistical errors that could be committed in all survey phases.
Keywords: Quality, statistical error, accuracy.
DOI: 10.7176/JESD/10-22-07
Publication date: November 30th 2019

1. Introduction

1.1 Abu Dhabi

Abu Dhabi is the federal capital of the United Arab Emirates (UAE) and the largest of the seven emirates. Geographically, Abu Dhabi lies on the borders with the Kingdom of Saudi Arabia, the Sultanate of Oman, and the Arabian Gulf. Over the past 40 years, Abu Dhabi has experienced significant population growth and economic development. To manage the growth and prosperity of the Emirate, the Government of Abu Dhabi required an official agency that could provide statistics for decision-making and policy setting.

1.2 Statistics Centre – Abu Dhabi

Statistics Centre – Abu Dhabi (SCAD) was established in accordance with Law #7 for the year 2008. SCAD is responsible for the collection, classification, storage, analysis and dissemination of official statistics covering social, demographic, economic, environmental and cultural indicators.
As a young statistical office, SCAD is in the fortunate position of being able to implement best practices from international bodies and leading National Statistical Organizations (NSOs). SCAD is aiming to be a world leader in innovative and efficient methods for data collection, analysis and dissemination.

1.3 Background to the total survey errors

Data quality is a concept with many dimensions, and each dimension is linked with others. In the abstract, all dimensions of data quality are very important, but in practice, it is usually not possible to place high importance on all dimensions. Thus, with fixed financial resources, an emphasis on one dimension will result in a decrease in emphasis in another. More emphasis on accuracy can lead to less emphasis on timeliness and accessibility; or an emphasis on timeliness may result in early/preliminary release data of significantly lower accuracy. Each dimension is important to an end user, but each user may differ in terms of identifying the most important priorities for a data collection program.

This paper focus on the accuracy dimension—a dimension that has a history of measurements and reporting. The emphasis is on statistical indicators, used to describe different aspects of survey accuracy in relation to various error sources, how indicators may be measured, and whether and how they are presented to data users.

Accuracy is an important and visible aspect of quality that has been of concern to statisticians and survey methodologists for many years. It relates to the closeness between estimated and true (unknown) values. For many, accuracy means the measurement and reporting of estimates of sampling error for sample survey programs, but, in fact, the concept is much broader, taking in non-sampling error as well. Non-sampling error includes coverage error, refusal error, non-response error, and processing error.

Sampling error;
Refers to the expected variation in estimates due to the random selection scheme used to select the sample. In a random selection scheme, each unit of the population has a known, non-zero probability of being selected into the sample. The method of randomization is important, because it can be used in theory to define both the optimal estimator and the appropriate estimate for sampling error. Most surveys using random selection are designed so that sampling errors can be computed directly from the survey observations.

Most surveys are designed so that the key statistics can be precisely estimated from the sample, and the sampling error of those estimates can also be computed from the survey itself. This implies that the sample sizes are large enough that the estimates satisfy the requirements of the large-sample statistical theory developed for sample surveys. The sampling error calculation based on the standard deviation of the data in the sample, divided by the square root of the sample size, it is called the standard error. Also from the standard error and based on confidence interval estimation the marginal error, which equal the magnitude of the confidence level multiplied by the standard error, is one of the sampling error indicators.

Non-response error;
Is a well-known source of non-sampling error. It is an error of non-observation reflecting an unsuccessful attempt to obtain the desired information from an eligible unit. Non-response reduces sample size, results in increased variance, and introduces a potential for bias in the survey estimates. Non-response rates are frequently reported and are often viewed as a proxy for the quality of a survey. The complexities of the survey design often make calculation and communication of response rates confusing and potentially problematic.

Coverage error;
Is the error associated with the failure to include some population units in the frame used for sample selection (under coverage), and the error associated with the failure to identify units represented in the frame more than once (over coverage). The source of coverage error is the sampling frame itself. It is important, therefore, that information about the quality of the sampling frame and its completeness for the target population is known.

Processing error;
Occur after the survey data are collected, during the processes that convert reported data to published estimates. Each processing step, from data collection to the publication of the final survey results, can generate errors in the data or in the published statistics. Processing errors include data entry, coding, and editing and imputation errors. Imputation errors are included under processing error because many agencies treat failed edits as missing and impute values for them. Error rates are determined through quality control samples.

2. Literature Review

In the literature, no indication of any reference was found, which describes any method that combine both sampling and non-sampling errors to compile one comprehensive measure for data accuracy.

On one side, “The Subcommittee on Measuring and Reporting the Quality of Survey Data” that refers to the Executive Office of the President of the United States, indicated in its “Statistical Policy Working Paper, 2001” that the total survey error often is formulated by survey statisticians in terms of the mean squared error of the estimate, where the mean squared error is the sum of the variance and the square of the bias. However, this formulation does not capture the complexity of the problem. For example, consider the interviewer as a source of error in addition to the sampling error. Interviewer errors can result in both bias (systematic error) and variance (variable error). Since the same source of error contributes to both terms, the simple additive structure of the mean squared error may not be adequate as a model for estimating total survey error.

On the other side, an approach was presented for evaluating the accuracy of official statistics produced by Statistics Sweden ASPIRE model. This approach, is general, in that it can be applied to any specific statistical estimate to assess product quality by first decomposing the total error for a product into its major error components. The ASPIRE model then evaluates the potential for these error sources to affect data quality (referred to as “the risks of poor quality”) according to the following five quality criteria: knowledge of risks, communication, available expertise, and compliance with standards and best practices achievement towards improvement plan. For each criterion, a generic checklist could be applied to each relevant error source. A simple “yes/no” format, used for the checklists, eliminates much of the subjectivity and inter-rater variability associated with the quality assessments. In addition, the checklists incorporate an implied rating feature, so that upon completing the checklist for a criterion, the rating for that criterion is largely pre-determined based upon the last “yes”- checked item in the list. The ASPIRE model has some limitations, in that it provides a proxy measure for product quality, and it cannot provide a direct measure of the total error of a variable, estimate, or product. It relies on the assumption that reducing the risks of poor data quality and improving process quality will lead to real improvements in data quality. Also, it is subjective, in that it relies heavily on the knowledge, skill, and impartiality of the evaluators as well as the accuracy and completeness of the information available to the evaluators.

The classification of error sources in surveys, described above, provides a framework for users of statistical data to develop an understanding of the nature of the data they analyse. An understanding of the limitations of data

can assist an analyst in developing methods to compensate for the known shortcomings of their data.

3. Compilation of an Approximated Measure of Total Errors in Abu Dhabi Surveys

In addition to the sampling error, also the coverage error, processing error, non-response error, and refusal error will be included to compile an Approximated Synthetic Indicator, to measure the value of the total error in sample surveys. This value, is a complement, and reflects the accuracy level in the survey estimates. The following two concerns were undertaken upon compiling the measure in Abu Dhabi:

- In the literature review, no indication of references was found, about a method used to construct one comprehensive measure for total survey error. Mostly, the accuracy of the data is assisted through evaluating each source of error independently.
- The proposed Approximated Synthetic Indicator of total survey error in Abu Dhabi is for internal use (In Abu Dhabi), aimed to follow-up on the development plans and improvement targets in Abu Dhabi surveys program through time. It is not valid for comparing the developments of total survey error in Abu Dhabi with other regions or countries.

Based on the total survey error the following criterion can be considered for calculating the approximated accuracy level:

$$ACL\ \% = 100 - TSE\% \ \ldots\ldots\ldots\ldots (1)$$

Where:
ACL: the accuracy level in specified survey
TSE: Approximated total sum of error in the survey
The value of TSE is approximately calculated based on the survey errors from different sources as follows:
Non-response error is evaluated by the following equation:

$$NRR\% = \frac{RF + ONR}{RP + RF + ONR} \times 100 \ \ldots\ldots\ldots\ldots\ldots (2)$$

Where:
NRR%: the non-response rate
RF: Total number refused sampling units in the survey
RP: Total number of partially or completely respondent sampling units in the survey
ONR: Other response cases like inability to response, not useful respond etc.

The non-coverage rate in the Sampling Frame is given by:

$$SFR\% = \frac{NC + OS + CC}{N} \times 100 \ \ldots\ldots\ldots\ldots\ldots (3)$$

Where:
SFR%: the non-coverage rate in the sampling frame of the survey
NC: total not reachable sample units in the sampling frame
OS: total sample units which are out of scope of the survey
CC: total permanently closed sampling units in the survey
N: total selected sample survey

The processing error is decomposed into imputation errors and corrected wrong response error, which are given as follows:

$$MVR\% = \frac{mv}{RP \times r} \times 100 \ \ldots\ldots\ldots\ldots\ldots (4)$$

Where;
MVR%: is the ratio of imputed missing cases or cells related to the key questions in the survey questionnaire (These key questions are determined in advance.)
RP: total number of respond sampling units in the survey
r: Total number of key questions
mv: Total missing values imputed to the key question cells through processing phase

$$TVR\% = \frac{WN + MVN}{n \times r} \times 100 \ \ldots\ldots\ldots\ldots\ldots (5)$$

Where;
TVR%: is the ratio of total corrected wrong answers in the cells related to the key questions in the survey questionnaire. (These key questions are determined in advance.)
WN: total corrected wrong answers to the key question cells through the processing phase
MVN: total adjusted outliers to the key question cells through the processing phase

The sampling Error, represented by the average ratio of the marginal errors related to the predefined key estimates of the survey. Its formula is given by:

$$SSE\% = \frac{\sum_{i=1}^{K}(MER)_i}{K} \ \ldots\ldots\ldots\ldots\ldots (6)$$

Where:
SSE%: the average ratio of the marginal error of the key sample estimates
$(MER)_i$ The values of the marginal error related to the key estimate (i)
K: the total number of key estimates predefined in the survey
Based on the above sources, the approximated survey error is the sum of both sampling and non-sampling errors,

$$TSE = SSE\% + NSE\% \dots\dots\dots\dots\dots\dots (7)$$

Where,

$$NSE = NRR\% + SFR\% + MVR\% + TVR\% \dots\dots\dots\dots\dots\dots (8)$$

By substituting the TSE in equation (7), by its values in equation (1) above, we get the approximated accuracy level $ACL\ \%$ for the key estimated variables in the survey.

4. Conclusions

- The accuracy level in Abu Dhabi surveys, produced by the current methodology is not comparable with that in other countries or regions, because the construction of the total square errors is constructed based on sources of errors, determined according to the circumstances of the surveys implementation in Abu Dhabi.
- The produced Approximated Synthetic Indicator is for internal use in SCAD, it is to follow-up on the execution plans and targeted objectives aimed to improve the household survey program in Abu Dhabi.

References

A System for Managing the Quality of Official Statistics. Paul Biemer, Dennis Trewin, Heather Bergdah, and Lilli Japec. Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 381- 415

Measuring and Reporting Sources of Error in Surveys, Statistical Policy Working paper 31, Executive Office of the President of the United States, USA, 2001.

Manual of Quality Assurance Standards and Procedures for a Statistical Survey, Statistics Centre - Abu Dhabi, 2016.

Guideline for Measuring Statistical Quality, Version 3.1. Office for National Statistics, London, 2007.

Sources of error in Survey Researches. Available on: https://www.qualtrics.com/blog/sources-of-error-in-survey-research/