Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 363–368
http://www.ssca.org.in/journal

Improving Data Validation

A. K. Nigam
Consultant Advisor, IASDS, Lucknow

Received: 24 March 2024; Revised: 12 June 2024; Accepted: 16 June 2024

Abstract

Data validation in Official System is usually taken as reporting at different levels, identifying the inconsistencies, and taking remedial measures. The idea of present write-up is to overhaul the concept of Validation and to increase its scope in a multi-pronged way to increase its visibility and utility to the policy makers. This would be of great help to the Official Reporting System and would revolutionize the whole approach to validation.

Key words: Official reporting system; Data validation; Improving data validation.

1. Introduction

Good quality data is a fundamental requirement for framing efficient policies. The COVID-19 experience has taught the world the importance of timely availability of reliable and relevant data for making informed decisions.

In India, data quality and reliability has long been the center of debate. Acknowledging poor data quality, several steps are now being taken at state and central levels to overcome the data quality issues. Policy framing and implementation is hugely dependent on the data. Thus, it becomes extremely important that good quality data is produced to have informed decisions on policy issues. Quality data includes factors such as accuracy, consistency, and reliability which is often lacking in the National/ State level data. One way to address this problem could be data triangulation and validation at different levels of data production and compilation through robust statistical techniques.

Data validation in Official System is perceived in a very conservative way. Validation is usually taken as reporting at different levels, identifying the inconsistencies, and taking remedial measures. For instance, for immunization coverage in children, it is usually taken as validation of reporting at different levels *viz.*, session site/ village, PHC/UPHC, CHC and district. Validation of routine immunization coverage also requires identifying errors which take place in reporting at these levels during the roll up process.

The idea of present write-up is to overhaul the concept of validation and to increase its scope in a multi-pronged way to increase its visibility and utility to the policy makers. This would help the official reporting system and would virtually revolutionize the whole approach to validation.

Corresponding Author: Arun K. Nigam
Email: dr_aknigam@yahoo.com

2. The revised validation

Good quality data is a pre-requisite to draw sound and meaningful inference from any research study involving data analysis. Poor quality data may emanate from (i) use of poor methodology, (ii) lack of sound data scrutiny and (iii) poor reporting of collected/analysed data at different stages of reporting in-built in the system. Data validation focuses on identification of the causes of poor-quality data and taking/suggesting remedial measures thereof.

2.1. Methodology

Choice of proper methodology and adequate sample size is crucial for any research study. Use of poor methodology mostly includes choosing a poor study design and inadequate sample size. Both these are widespread in not only in Indian context but are also prevalent worldwide. The book by Nigam (2016) deals these issues in detail. The discussion which follows derives heavily from this book.

Among matters of concern are choice of proper survey design including adequate sample size, clarity and coverage of questionnaires, data cleaning/handling and choice of analytical techniques for obtaining valid and efficient estimates as departure from these result in wasting precious funds employed for research programs. It also often ends up with invalid and misleading estimates, which may have strong policy implications.

2.2. Questionnaire

Besides a proper and efficient sampling design required for obtaining efficient and valid estimates, the type and coverage of questionnaires is a crucial deciding factor in obtaining quality data. Any ill-conceived questionnaire leads to substantive non-response, incorrect and evasive responses. In many surveys, questionnaires are unduly lengthy having questions not relevant to the study. On the other hand, sometimes, these are too short to provide a satisfactory coverage. A lengthy questionnaire escalates the cost of the survey and makes management and supervision work cumbersome and time consuming. It also creates problems in editing and cleaning of data and in a decrease in efficiency. A questionnaire with insufficient coverage is likely to be less efficient because of the failure to collect some vital information.

To refine the questionnaire, it is necessary to train interviewers, data editors/cleaners, and through test data analysis. Adequate time should be allotted for field practice and the training should be evaluated. There should be effective and quality monitoring during the field work and this allows for making amends for the ambiguity and inconsistencies. Proper and effective training and pre-testing allow both the project handlers and the interviewers gain insight into the spirits underlying different questions. At data entry level also, there should be data validation employing range check, valid value check as well as internal consistency checks. The follow-up checks and corrective measures improve not only the quality of data gathered but also making the resulting estimates much more relevant and consistent. This aspect, however, is usually taken rather casually in many surveys conducted in our country.

2.3. Sample size and related issues

A close look into the research studies reveals that sample size is often arbitrarily decided, without considering the extent and nature of the variability of the character being studied and even when adequate sample size is taken, there is an attempt to present analysis by sub-groups in terms of related socio-economic, demographic, housing, or household characteristics. This practice leads to decomposition of sample size according to these sub-groups. While a smaller sample size leads to invalid estimates with unduly large standard errors, a larger sample involves avoidable wasteful expenditure. In view of this, it is worthwhile to highlight some of the observations on these issues by Nigam (2006) and Nigam and Singh (2011).

2.4. Poor reporting of collected / analyzed data

In most of the large-scale surveys reporting of indicators is usually done by sub-groups like caste, religion, gender, age group, grades of nutritional status, grades of anemia *etc.* In many situations, sample size for some of these sub-groups is grossly inadequate. Examples of this can be found in the reporting of National Family Health Survey (NFHS), NNMB, Reproductive Child Health (RCH), and District Level Household Surveys (DLHS) and others. The sample size is usually ascertained for all the groups keeping in mind the precision, complexity of the design and expected non-response. Any attempt to the reporting by sub-groups makes such estimates highly imprecise. In view of this, it may be better to go for interval estimates (confidence interval) instead of point estimates. The best alternative, however, is to develop small area estimates for the sub-groups (Chapter 13 in Nigam 2016). For examples of these types of dis-aggregated reporting one may refer to Nigam (2006) and Nigam and Singh (2011). For example, in NFHS-2, nutritional status was reported only for 77 children in Hill Region, for 57 children of ST and for 65 children of Self-employed parents. The reporting has further categorization according to grades of nutritional status. The prevalence of undernutrition ranged from 40-60 percent for below-2sd and 16-30 percent for below-3sd in these groups. Any anaemia among children has been reported for 72 children in Hills, 73 in Bundelkhand and for 33 children of ST, with further division according to grades of anaemia (severe, mild *etc.*). The reported prevalence of any anaemia ranged 73-80 percent and 5-13 percent for severe anaemia. One can easily notice that sample sizes were not adequate for any of these sub-group estimates

Poor reporting of data in different stages are also widely prevalent and can be controlled through proper monitoring. Ways for controlling errors, bridging data gaps, data reduction and improving the quality of data are being discussed now. These can be applied at different stages, *viz.*, at handling of data, sample selection and estimation. Every survey, without exception, encounters the problem of missing data or data with inconsistencies. The main reasons of missing data (i) non-collection of the responses of a sample element, (ii) deletion of some responses as they fail to satisfy certain edit checks. Inconsistencies in data may be attributed to various reasons, such as errors in tabulation, data entry or even in copying from a secondary source. In all such events, it is a norm to treat it as missing data and handle it accordingly. Whereas total non-response, *i.e.*, when all of the responses on a unit are not available, can be handled by some form of weighting adjustment techniques, item non-responses are taken care of by imputation. In the sequel, we add another dimension to data validation.

Framework for data monitoring and quality assessment

Several data quality frameworks have been developed in the health sciences, most of them related to administrative data, registries and electronic health records (Mariño *et al.*, 2022, Schmidt *et al.*,2021). Although they may differ in functionality, their fundamental data quality parameters are similar. The four broad dimensions of such frameworks are - Data Integrity, Completeness, Consistency, and Accuracy (Mariño *et al.*, 2022, Schmidt *et al.*, 2021).

Data Integrity

Data integrity is the degree to which the data conforms to structural and technical requirements. This dimension includes the domains of correct structural representation of data, correct matching of records and elements across multiple datasets, and use of appropriate data value formats.

Completeness

This dimension evaluates the degree to which the required data values are present. It is evaluated with respect to two domains - crude missingness and qualified missingness. Crude missingness is the metrics of missing data values that ignore the underlying reasons for missing data. In qualified missingness assessment, metrics are developed for underlying reasons for missing data such as non-response rate, refusal rate, drop-out rate, *etc.*

Consistency

This data quality dimension encompasses the aspects of range and value violations (compliance with admissible data values or value ranges), and contradictions (presence of improbable combinations).

Accuracy

It is assessed in terms of three domains- unexpected distributions, unexpected associations, and disagreement of repeated measurements. Unexpected distribution entails presence of outliers, unexpected location, shape, scale, among other distributional discrepancies. Unexpected association may include unexpected direction or strength of association between variables. Disagreement of repeated measurements is defined in terms of intra-class reliability, inter-class reliability, or disagreement with gold standard.

Various software packages, especially on R platform, have been developed in recent years to implement data quality monitoring on the four dimensions discussed above (Mariño *et al.*, 2022). These packages offer user-friendly platform to examine data properties in an automated and efficient fashion. However, the performance of such packages depends largely on the quality of the metadata file, which is required for creating the metrics for data quality assessments. So, developing standard framework for metadata is another important aspect for data quality assessment.

2.5. Some useful statistical techniques to address data quality

As stated earlier, use of poor methodology mostly includes choosing a poor study design and inadequate sample size. We now describe some useful statistical techniques

which may go a long way in strengthening the methodology.

Imputation

Almost every large survey suffers from missing values which may also include non-response and/or outliers. Imputation technique consists of handling non-responses by replacing each missing value with a real value. Several imputation procedures are now available for assigning values for missing responses and these are deductive imputation, overall and class-mean imputations, random imputation, hot-deck imputation, and imputation based upon regression.

Principal components technique

Large scale surveys mostly have large data sets requiring reduction. Principal components technique can be used to reduce data. This technique uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

Randomized response technique

In many surveys, the intention is to seek information on sensitive characteristics to which response is either false or evasive. It is, therefore, useful to employ the randomized response technique in such cases. The technique ensures confidentiality to the respondent and has become popular in recent years. Sensitive characteristics include unsafe sexual behaviour, child abuse, drugs, *etc.* Detailed applications of this technique are discussed in Nigam (2016).

Small area estimation

One of the areas of data gaps is related to micro-level planning, which requires estimates of different activities for 'smaller' areas having inadequate sample size. This can be achieved by using small area estimation technique which is discussed in detail in Nigam (2016).

Some other useful techniques

Two other techniques need special, though brief, mention. Re-sampling inference is a technique which aims at finding the standard error of variance estimates of non-linear statistics, such as ratios, regression coefficient, index numbers, *etc.* Some other applications are the standard errors of statistics such as median (height or weight), inflation rate, wholesale price index number and the like.

Another useful technique is Snowball Sampling. This can be used in situations where large sample size is required. One such example is estimation of Maternal Mortality Ratio *etc.*

Acknowledgements

Part of this work was done exclusively by the author and included in the Draft Report- Development of Data Validation Protocol Manual, (2022) by the Institute of Applied Statistics and Development Studies (IASDS), Lucknow. An initial draft was discussed with Prof. Arvind Pandey who made several useful suggestions. The reviewer also provided additional material adding another dimension to the write-up.

References

Draft Report (2022). *Development of Data Validation Protocol Manual*. Institute of Applied Statistics and Development Studies, Lucknow.

Mariño, J., Kasbohm, E., Struckmann, S., Kapsner, L. A., and Schmidt, C. O. (2022). R packages for data quality assessments and data monitoring: a software scoping review with recommendations for future developments. *Applied Sciences*, **12**, 4238.

Nigam, A. K. (2006). *Strengthening of NNMB Surveys*. In; Arvind Pandey (Ed) Biostatistical Aspects of Health and Population, Indian Society of Medical Statistics,155-60.

Nigam, A. K. (2016). *Statistical Aspects of Community Health and Nutrition*. Woodhead Publishing India in Food Science.

Nigam, A. K. and Singh, Padam (2011). *Research Methods in Public Health Nutrition: Common Critical Factors: Chapter 11*. In Shiela C. Vir (Ed) Public Health Nutrition in Developing Countries, WPI Publication.

Schmidt, C. O., Struckmann, S., Enzenbach, C., Reineke, A., *et al*. (2021). Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Medical Research Methodology*, **21**, 1-15.