<A Quality Assessment framework for Statistics based on data-science : Institutional Management Plan for Experimental Statistics of 'Statistics Korea'>

Bitna Kang[1]

[1]Statistics Korea(A national statistical institution of Korea)

Abstract:

Due to rapid data environment change such as big data and artificial intelligence, it is more important to produce statistics in a way that is different from the survey-oriented statistics. In order to secure the reliability and accuracy of statistics produced in a new way, it is necessary to manage them as official statistics at the national level. For this, the Statistics Korea introduced the experimental statistics system in 2021. The goal is to discover and manage statistics that are based in various data sources and that use different method. These data science-based statistics make it difficult to manage the entire statistical production process due to the diversity and incompleteness of the data. Therefore, the same quality management method as the existing one cannot be applied. However, in order to systematically manage the quality of these statistics, there must be a quality dimensions shared with other statistics. In this paper, the quality management framework was designed based on the quality evaluation dimension of Statistics Korea. The framework includes evaluation indicators, contents and procedures for each quality dimension. In order to help understand the framework, the first experimental statistics in **Korea**, the case of telecommunication mobile population movement statistics('Population Mobility Statistics'), will be presented.

Keywords:

Data-science based statistics, big data based statistics, experimental statistics, quality assessment framework

1. Introduction:

Data science is an interdisciplinary field that uses scientific methods, process, algorithms knowledge and insights from noisy, structured and unstructure data. In this paper, data-science based statistics are defined as statistics created by utilizing various data sources(big data in a broad sense including vast public big data and administration data) or data extraction and processing using artificial intelligence. In the case of statistics based on data-science, it is difficult to establish a quality management method with a single fixed standard because the development of the underlying data related technology is very fast and the methodologies are also changing rapidly. Considering this point, the quality management(assessment) framework to be presented in this paper includes a macroscopic quality process that can encompass the overall statistics based on data-science. The framework is comprehensive and flexible taking into accounts various data sources and data complexity. This framework is only at the level of presenting the idea of a formal management method for data science-based statistics from a macro perspective. In the future, it is necessary to develop detailed inspection items and to prepare quantitative quality evaluation measures such as distribution of scores for each item and selection of weights in detail.

2. Methodology:

The variety and volume of big data and the unfamiliarity of Artificial Intelligence add make it difficult to apply the quality evaluation criteria of existing statistics to the statistics based on data-science. To overcome these differences and ensure the quality of future implementations, using big data for producing official statistics needs to be based on a clear and standard business process model, similar to census or survey data and other product classes(Noviyanti et al., 2020). The first thing to consider in establishing the quality assessment framework is whether to use the statistical production process or the quality dimension as the basis. In survey statistics, the statistical production process is almost the same.

However, in data-science based statistics, the business procedure differs depending on what the data source is and how the data is extracted and processed. Therefore, it would be desirable to design the quality evaluation process that can cover the entire data-science based statistics with the quality

dimension as the basic axis. Even if the resource data and production methods change, the quality attributes that national statistics should have will be the same. However, the details that need to be checked are different. In this paper, a framework was designed, focusing on the quality dimension of the ‘Statistics Korea’ and setting sub-quality indicators reflecting the unique characteristics of data-science.

The quality dimensions covered by the quality diagnosis system of ‘Statistics Korea’ are relevance, accuracy, timeliness/punctuality, comparability/coherence, and accessibility/clarity. Relevance evaluates how meaningful and useful a statistic is from a user's point of view. Accuracy is evaluated as the difference between the unknown true value and the estimated value. Timeliness refers to how up-to-date the statistics are, and punctuality refers to the degree to which the previously announced disclosure schedule is complied with. Comparability refers to whether temporal or spatial comparison is possible, and consistency refers to how similar other analyses of the same economic and social phenomenon are. Accessibility evaluates how easily users can access statistics, and clarity is the level of information provided to help users easily understand statistics. The quality evaluation framework of this paper includes these quality dimensions and institutional evaluation process(Fig. 1).

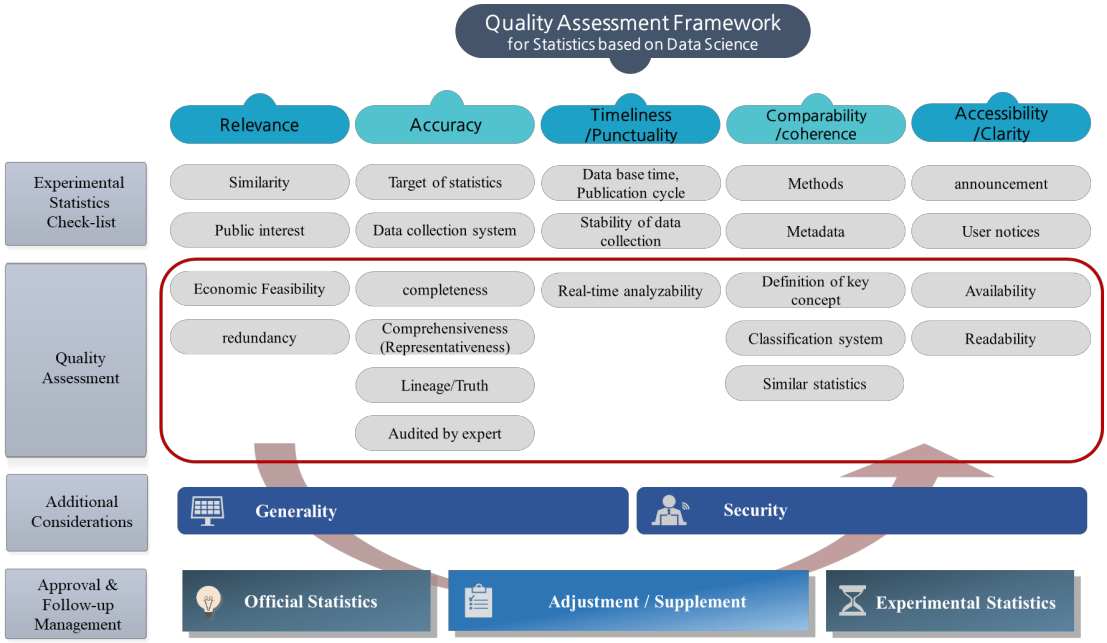
The first phase is to check whether the data science statistics can be managed as experimental statistics. Once confirmed as experimental statistics, a more in-depth quality verification procedure is performed to see if it can be formally managed by Statistics Korea. This allows you to determine whether the statistic should be managed as an official approved statistic, or if it's just experimental, or if it's not worth producing a statistic at all(Depending on the nature of the data source, the use value of statistics may have disappeared due to rapid social change). If approved as official statistics, the statistics are subject to systematic management such as regular quality evaluation in accordance with the Korean Statistical Act.

3. Result : Quality Assessment Framework for Statistics based on Data Science

3.1. Framework description

3.1.1. Experimental statistics check phase The contents evaluated at this phase are actually evaluated through the checklist by the Statistics Korea. The indicators depicted in Figure 1 have been selected and rearranged based on the quality dimension of the framework. Regarding relevance, the parts to be checked are whether there are similar or duplicate statistics and the purpose of production (public interest). However, even if it is similar to the existing statistics, it is possible to determine whether the production cycle can be shortened or whether the existing statistics can be supplemented. From the user's point of view, if the statistics can provide new utility and information to users, then there is justification in the production of statistics. The purpose of production is related to 'fitness for use'. Whether statistics are produced for the public good. Regarding accuracy, the target of statistical production and data collection system are checked. In particular, the accuracy of the data itself, which is the basis of statistics, is mainly reviewed. Therefore, it is necessary to check the reliability of the resource data(coverage, whether processing is possible in a form that guarantees representativeness, etc.) and the appropriateness of the source data provider. Regarding timeliness/punctuality, whether the data base time, the planned publication cycle is appropriate, and the stability of data collection will be important. In the comparability/consistency, production methods are reviewed. It is judged whether new types of data and methodologies different from the existing ones are used, and whether there are detailed descriptions of resource data and statistical estimation and processing methods. Regarding accessibility/clarity, it is reviewed whether press releases or reports are published, whether microdata is disclosed, and whether user notices are provided appropriately.

<Fig. 1 Quality Assessment Framework For Statistics based on Data-science>



3.1.2. Quality Assessment Phase This phase is the phase to check the quality of statistics based on data-science in earnest. Traditional data dimensions are still applicable to data science-based source data. In this step, the sub-quality dimension index was set by referring to the result of Ramasamy & Chowdhury(2020). This is a study that suggested the key quality dimensions of big data by reviewing 17 previous studies. In addition, several studies, such as Berka et al.(2010) and Smith et al.(2018), on quality evaluation methods for administrative data(the most refined form of big data) were also referenced. First, as the sub-quality dimensions of relevance, there are economic feasibility and Redundancy. Because relevance refers to how useful and meaningful the statistics are to users. In general, in the case of data-science based statistics, the amount of data to be analysed is huge, and the analysis technique will be different from the existing one, which entails a lot of cost and time. Second, reliability of data will be evaluated as the most important in accuracy. Thus, the completeness and comprehensiveness (representativeness) of the data, Lineage and truth are included. It would be an argument that many statisticians would agree with that the amount of big data is not necessarily proportional to the veracity of the results. Therefore, with regard to completeness, a relaxed standard can be applied to whether data integrity is at a identifiable level (Gartner, 2011). In addition, whether the results of data processing are within a range of known or acceptable values, and whether data errors are regularly audited by experts will also affect the reliability evaluation (Cai & Zhu, 2015). Pedigree/Lineage of the data allows us to know the source of the data so that inconsistencies in the data can be corrected. In the case of truth, it is possible to determine the reliability of the data by determining whether the data source is from a reliable source(Batini et al., 2015). Third, in timeliness/punctuality, real-time analysability can be an important evaluation criterion. This is because, in the case of certain data, the period of storage affects the quality of the results (El Alaoui, Gahi & messiussi, 2019). Fourth, in comparability/consistency, it is reviewed whether there is a clear definition of key conceptual terms, whether the classification system is appropriate, whether there is a difference from similar statistics, and if there is a difference, the content, degree, and reason for the difference are reviewed. In the absence of similar statistics to compare, expert interview may be used as an alternative. For such attributes the expert is asked on his/her assessment regarding the data accuracy of the corresponding attribute. This approach is very subjective and therefore only carried out if no external data source for the benchmark approach is available(Berka et al., 2010). Since cohesion refer to the capability if data to comply without contradictions to all properties of the reality of interest, as specified in terms if integrity constraints, data edits, business rules and other formalisms(Batini et al., 2015), this can also be considered in comparability/consistency. Lastly, accessibility and clarity are

largely reviewed for availability and understandability. Availability is related to the ability of the user to access data from his or her culture, physical status/functions, and technologies available and readability refers to ease of understanding of data by users(Batini at al., 2015). Therefore, it should be evaluated whether statistics are provided so that users can acquire and understand data without special statistical knowledge and tools.

3.1.3. Additional evaluation dimensions Generality and security are also need to be evaluated. Generality concerns whether the process of analyzing and producing statistics using data science is applicable to other data science-based sources. Considering the current status of data science utilization statistics, which are not yet mature enough in the public sector, a foundation to promote the production of these types of statistics in the future and improve their quality should be laid, so it is important to evaluate the value of statistics through generality evaluation. Security means that data is securely managed by an appropriate authority. Considering the vast amount of personal information of big data, security should also be treated as a major evaluation factor.

3.1.4. Approval and follow-up management phase According to the evaluation results in the quality evaluation phase, if the statistics are recognized as reliable statistics, they are approved and managed as official statistics. If there is a need to improve quality, it is managed with experimental statistics with a delay for a certain period of time. To this end, it is necessary to create a checklist for each quality indicator discussed above and establish a method to quantify the result. However, a contextual design is required in consideration of the diversity of data sources and processing methods of statistics based on data science, and the needs of users (Ramasamy & Chowdhury). As examples of experimental statistics accumulate, check indicators and items should be continuously revised and supplemented. Efforts to apply the framework to the field of artificial intelligence technology utilization statistics, which are still unknown in the field of official statistics, must be continued.

3.2. Case description : 'Population Mobility Statistics' in Korea

3.2.1. Experimental statistics check phase Population Mobility Statistics, the first experimental statistic by Statistics Korea, is a statistic created by using telecommunication mobile data to create domestic population movement patterns. The production contents are the amount of population movement (inside administrative districts, outside administrative districts, total) by gender, age group, location type, and province(includes city, county, district). The source data is the mobile population movement aggregate data of SKT, which has a share of the mobile communication market in Korea (42% as of 2021). As a result of examining the experimental statistics checklist, 1) there were official statistics with similar names (domestic migration statistics) with respect to the relevance dimension, but there was no similar duplication problem because the contents of the statistical table, production unit, production cycle, and production purpose were all different. This is because, in the case of domestic population movement statistics, 'movement' means a change of residence in administrative registration, but 'movement' in population mobility statistics means a case of staying in another administrative district for more than 30 minutes after leaving the place of residence. As such, it can serve as a basis for policy establishment in various fields(such as urban planning, transportation, and simulations of diseased spread) in that it provides short-term population movement information on a weekly basis by using this new data source. 2) In accuracy dimension, the most important thing to consider is whether the source data is representative of the population. Although the coverage of SKT's source data is only 42%, the total population was estimated using the population data from the registration census of the Statistics Korea to increase representativeness. 3) As for timeliness/punctuality, it is expected that stable data collection will be possible by signing an MOU with SKT, so statistics will be disclosed on a weekly basis as planned. 4) For comparability/consistency, it was confirmed that source data and detailed explanatory data on estimation and processing methods were available. 5) In terms of accessibility/clarity, precautions for use were provided appropriately, and there was

also a manual for source data. Overall, it seems to be a statistic that uses new data to aggregate the amount of movement on a weekly basis and provides useful information so that it can be used to establish related policies. However, the coverage rate of source data was incomplete, and it was difficult to say that the estimation technique was sufficiently verified. Therefore, it was decided to manage these statistics as experimental statistics that require observation for a certain period of time.

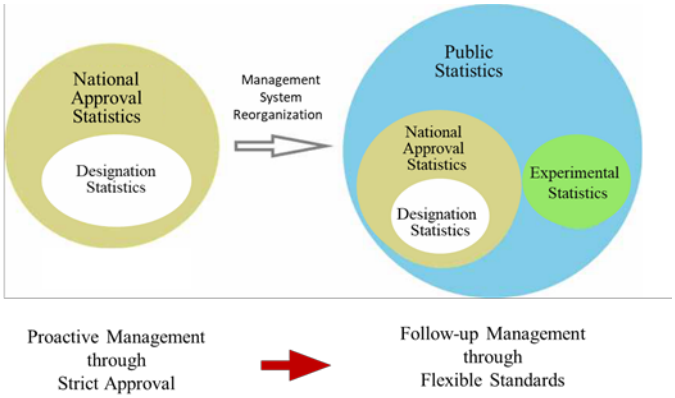
3.2.2. Quality Assessment Phase As can be seen in the Indonesian case of using mobile positioning data to delineate metropolitan areas in Indonesia(Noviyanti et al., 2020), Population mobility statistics can also be used in Korea's balanced national development, transportation, and various social and cultural policies. As supplying information for managerial decision-making is a vital feature of management accounting, using Big Data for such purposes may also be valuable and—in today's business environment—even necessary(Gartner & Hiebl, 2018). This is also applicable to the public sector. Therefore, it can be evaluated this statistics provides very useful policy information with little time and cost. In accuracy dimension, the most problematic indicator is representativeness. The population coverage of the source data is only 42%. Census used to estimate the population appears to be reliable. But the creditability of the estimation methods needs to be examined more closely. Potential bias in the data will also be an issue. While true bias is hard to establish concretely, other than through re-abstraction studies, possible biases can be detected when sampling errors occur or when coverage or responses are not complete. If necessary, bias can be evaluated as part of the research enterprise(Smith et al., 2018)

3.2.3. Additional evaluation dimensions Population mobility statistics can be examined for scalability as an example of public-private cooperation. This is because, by receiving big data from private sector, it is possible to accumulate know-how that can produce various and useful policy statistics. Meanwhile, mobility data sharing raises privacy issues, so it should be evaluated whether methods for privacy protection have been established. By referring to previous studies such as 3W model, an improved privacy-protective population mobility model(Smolak et al,, 2020), security issues should be continuously discussed and solutions should be sought.

4. Discussion and Conclusion:

In order for the public sector to effectively utilize big data and artificial intelligence technology, it is necessary to diagnose the quality and usefulness of the statistical data. However, in data science, there are many different types of generating entities (machine, human), data types (structured, semi-structured, and unstructured), and types of production institutions (public and private). Therefore, it is very difficult to establish a single quality evaluation standard that can cover all of them. The quality assessment framework discussed here did not deal with detailed quality evaluation criteria in consideration of this point. Instead, from the institutional point of view, referring to the statistical quality management procedure of the 'Statistics Korea', the macroscopic evaluation criteria were designed focusing on the quality dimension that reliable national statistics should have. For consistency in management with other statistics such as survey statistics and reporting statistics, the quality dimensions covered by the current quality evaluation system of the 'Statistics Korea' were maintained, but quality indicators reflecting the characteristics of the data science field were set for each dimension. In the institutional aspect, quality indicators and considerations were presented for each statistical management procedure (experimental statistics confirmation / quality evaluation / approval decision / follow-up management). The 'Statistics Korea' has already prepared a self-quality checklist for statistics based on big data, and also created an experimental statistics. Detailed quality diagnosis indicators for statistics based on data-science are currently under research and development. Since this framework is only at the level of presenting the ideas, it is necessary to develop detailed inspection items and to prepare quantitative quality evaluation measures such as distribution of scores for each item and selection of weights. In this regard, continuous expert advice on quality evaluation criteria should be accompanied by taking into accounts the various data sources and complex processing methods of data science.

< Fig. 2 National Statistical Management System Reorganizing Strategies in Korea >



Meanwhile, these processes should be discussed along with changes in the National Statistical Management System of the 'Statistics Korea'. Korea's national statistics were being managed through strict pre-approval screening. However, in response to changes in the statistical environment, it is trying to change little by little with a post-management strategy(Fig.2). This is because, in order to increase the satisfaction of using statistics by providing more diverse forms of statistics, it is necessary to collect, process, and analyse the necessary data among the overflowing data to capture rapidly changing social changes in a timely manner and provide useful statistical information. From this point of view, it can be said that the experimental statistics system was introduced for the purpose of extracting useful policy information more efficiently and quickly by injecting data-science based statistics that were previously used only at the internal reference level into the public sector. Considering the status and influence of national statistics, as part of an effort to guarantee the quality of these statistics, the quality evaluation framework presented in this paper should be constantly revised, supplemented, and specified.

References:

1. Isnaeni Noviyanti, Panca D. Prabawa, Dwi Puspita Sari, Ade Koswara, Titi Kanti Lestari, M. Hanif Fahyuananto and Edi Setiawan, BPS-Statistics Indonesia, Indonesia, 2020.
2. Christopher Berka, Stefan Humer, Manuela Lenk, Mathias Moser, Henrik Rechta, Eliane Schwerer. (2010), A Quality Framework for Statistics based on Administrative Data Sources using the Example of the Austrian Census 2011. AUSTRIAN JOURNAL OF STATISTICS Volume 39 (2010), Number 4, 299–308.
3. Mark Smith, Lisa M Lix, Mahmoud Azimaee, Jennifer E Enns, Justine Orr, Say Hong, and Leslie L Roos. (2018). Assessing the quality of administrative data for research:a framework from the Manitoba Centre for Health Policy. Journal of the American Medical Informatics Association, 25(3), 2018, 224–229 doi: 10.1093/jamia/ocx078.
4. Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From Data Quality to Big Data Quality. Journal of Database Management, 26(1), 60-82. https://www.igi-global.com/article/from-data-quality-to-big-dataquality/140546.
5. Cai, L. & Zhu, Y., (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(2), 1-106. El Alaoui, I., Gahi, Y., & Messoussi, R. (2019). Big Data Quality Metrics for Sentiment Analysis Approaches. In Proceedings of the 2019 International Conference on Big Data Engineering (pp. 36-43). https://dl.acm.org/citation.cfm?id=3341629.
7. Kamil Smolak, Witold Rohm, Krzysztof Knop, Katarzyna Siła-Nowicka. (2020). Population mobility modelling for mobility data simulation. Computers, Environment and Urban Systems 84 (2020) 101526.
8. Bernhard Gärtner, Martin R.W. Hiebl. (2018). Chapter 13 Issues with Big Data. Forthcoming in: Martin Quinn and Erik Strauß (Eds.): The Routledge Companion to Accounting Information Systems, Routledge, 2018.