United Nations ECE/CES/2017/24

Economic and Social Council

Distr.: General
4 April 2017

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-fifth plenary session
Geneva, 19-21 June 2017
Item 4 of the provisional agenda
The next generation of statisticians and data scientists

Statistics production system 4.0

Note by Statistics Sweden

Summary

The document discusses the skills and competence requirements for statisticians in the conditions of changing statistical production, as well as the role of official statisticians in this context. For these to be discussed and defined, the statistical production system of National Statistical Offices have to be considered. The document also takes into consideration users and their data needs, and what statistical offices can be expected to produce as a response.

The document is presented to the Conference of European Statisticians' seminar on "The next generation of statisticians and data scientists" for discussion.

GE.17-05383(E)
1705383

I. Background

1. Surveys based on the randomization theory have been and are still today a major source for the production of statistics at National Statistical Offices (NSOs). One major problem in the application of the theory is nonresponse. While the theory presupposes full response applications at NSOs turn out with some sample units not responding. The problem of nonresponse has not yet been satisfactorily resolved (e.g. Brick, 2013) and is today a major threat to the validity of sample survey statistics because of high nonresponse rates. Partly because of this problem and the high costs of sample surveys, big data is by NSOs considered as an alternative source of data for production of statistics.

2. There are many challenging problems in using big data for official statistics production, but big data has also many interesting features (e.g. Daas et al., 2015; Japec et al., 2015). Collection of data is considered as cheaper and faster in comparison with traditional sample surveys, and can in some cases be used for producing statistics almost continuously. Using big data is perceived to shift the main survey issue from data collection to analysis of data, and also, the problem of data editing becomes a more complex task (e.g. Tang, 2014).

3. Shifting the production process at an NSO from a sample survey based organization into one using alternative data sources implies the need of a different competence profile, this is obvious. Regarding the statistical competencies, focus is shifting from data collection to data analysis, and statisticians must possess skills and competencies for doing these analyses.

4. Big data has not yet been defined here. Presently some NSOs rely heavily on use of registers and administrative data. Sometimes administrative data is considered as big data, sometimes it is not. In the following, the term “alternative data sources” will be used instead of big data and is defined as other data sources than probabilistic sample surveys and registers, where registers are defined as in Wallgren and Wallgren (2014).

5. To be able to discuss skills and competence requirements and the role of statisticians, the production system at the NSOs has to be defined. There is also a need to consider users and their decision problems and what an NSO can be expected to produce.

II. Statistics production systems 1.0 – 3.0

6. The strong position of the randomization theory at NSOs can be explained by the properties of the resulting statistics. Laymen may consider statistics as homogenous and being interpretable in the same way. This is of course not true and appropriate interpretation is made with respect to how statistics are derived. A seldom posed question in practice, often not so in academia too unfortunately, is “What is the point of statistical inference?”. The answer can be formulated as interval estimates, which is “A speciality of the statistical way of induction…”, to quote Chatterjee (2003, p. 30).

7. Statistical inference provide with interval estimates which not only provide with an estimate of the parameter under study, the interval estimates carry information on the uncertainty of estimates. Interval estimates are provided in the form of confidence intervals in the frequentist interpretation of probability. Now, under the randomization theory, confidence intervals derived from sample surveys are objective in the sense they can be interpreted without reference to any assumptions made. This is especially attractive in the context of the role of NSOs in a society. In application, definitions have to be made regarding survey units, population, frame, variables, etc.

8. Statistical agencies using sample surveys as the main data source are here characterized as having a Statistical Production System 2.0 (PS 2.0). Statistics Sweden started in the 1950's to employ probability sampling and developed the production process in the following decades. The system used prior to the 1950's is called PS 1.0.

9. In the end of the 20th century, administrative data at public agencies started to earn an interest for a register based statistics production system. Administrative data have several advantages over sample surveys for statistics production (e.g. Wallgren and Wallgren, 2014), and nowadays most NSO's are working on establishing a register based statistics production system. Such a system is well developed at Statistics Sweden, and is equally, if not more, important for the statistics production today.

10. Even though Statistics Sweden have developed both a well-functioning register based statistical system and a well-functioning sample survey based system, they are not used in their full capacity as statistics are mainly produced in either of the ways, and not in combination.

11. Wallgren and Wallgren (2014, Sec. 15.1.3) provide a discussion on the view of a system approach instead of a polarization between register surveys and sample surveys. A register system can be used for producing register statistics but is also important for the design of sample surveys. They (ibid) launch the concept of "register-based sample surveys". Full utilization of available registers has the potential to improve sample survey estimates in terms of relevance and precision. Thus, a statistical PS 3.0 is here defined as a register based statistics production system with a combined utilization of register and sample survey data. Registers and sample survey are simultaneously used in survey designs.

III. Future context of operation

12. In their report, the Commission on the Future of Sweden (2013, p. 244) states: "We live in a turbulent world in which borders are becoming less important, in which countries are becoming steadily more interdependent, in which the speed of change is perhaps more rapid than ever before, in which the ever-increasing flow of information is making it increasingly difficult to determine what is most relevant and thus needs to be given priority, and in which global competition is tougher than ever." Later in the context of democratic challenges, they state (ibid, page 258): "Another challenge concerns the increasing complexity of decision-making processes. This stems mainly from the fact that many of the problems and challenges facing Sweden cut across boundaries, whether at the local, regional or national level, or between Sweden and other parts of the world. At the same time, more rather than less cross-border cooperation will be needed to solve problems in the future. There is a risk that this will cause decision-making processes to become even more complex."

13. Statistics are produced for decision-making and the picture implied by these statements is the need for faster decision making in more complex contexts, often with international aspects. The production of statistics needs to be apt for the future decision-making. Timeliness and frequency is of course one aspect, and one can also expect increased demands for comparability and coherency. New and special statistics, to shed light on new emerging but temporary problems can also be expected.

14. During the last two decades, big data has emerged and has shown to be a potential new source for producing statistics. The future can be expected to show on new additional data sources and collection techniques. One experience at Statistics Sweden is the increasing volume of administrative data resources available at different governmental agencies. This can be expected to continue and offer new survey designs. This is of course

also true regarding the private industry, holding many sources potentially useful for producing statistics. One important issue here is access to data, which may be legally problematic and owners may have no interest in delivering data.

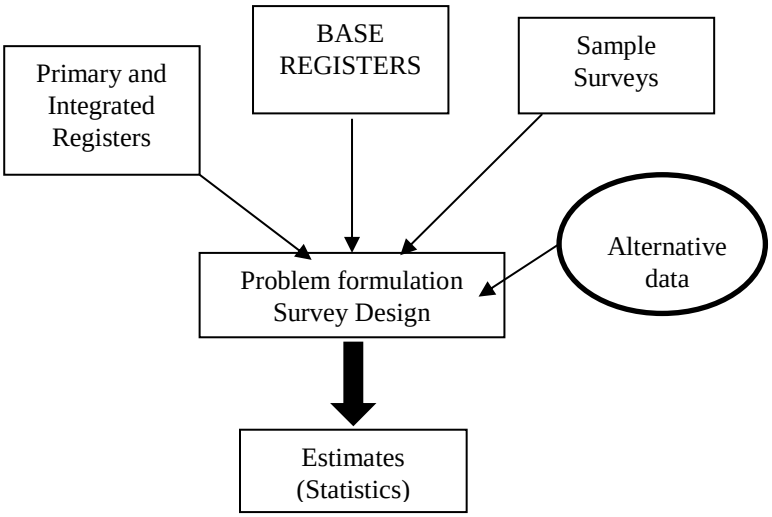
15. Data collection in sample surveys usually employs questionnaires delivered in one or the other form, or interviews over phone or face-to-face. Technology development will offer alternative data collection means. Mobile phones would today be possible to use in travel habit surveys. Credit card transactions seem possible to use in household budget surveys. Other examples are internet connections of vehicles and the Amazon Prime Air development project. The future will contribute with more technical developments, and some of them will turn out useful for data collection. So, not only will there be more data sources available, there will also be more options for collecting individual data.

IV. Statistics production system 4.0

16. With the above said, the future requirement on a statistical production system can be described with one word, flexibility. Flexible to meet new demand, flexible to produce statistics with fast dissemination of results, flexible to acquire, retrieve and process data from new sources, flexible to redesign a survey when new data sources are available or when old ones are cancelled, and flexible in communication of results. At our disposal are at least three data sources: sample surveys, registers and alternative data sources (e.g. big data).

17. A schematic picture of a system built on these data sources is given in figure 1. At centre of the system are the base registers, e.g. population and business registers. These are of outmost importance for the validity and value of the whole statistics production. Not only do they keep track of units in the population but also span the production in terms of domains and definitions, with effects on relevance, accuracy, coherence and comparability. On the left are satellite registers, keeping information on the units in the base registers, e.g. employment register and income register connected to the population register and, value added tax register and foreign trade register for the business register. These register parts constitutes a resource in the system and can be used for deriving statistics or as support to other register based surveys or sample surveys. Some statistics are derived from the base registers directly.

Figure 1
Statistics production system 4.0 in action



18. Another resource is sample surveys depicted on the right. This part has two meanings why it is depicted with two squares. One meaning is the methodology itself and the other is sample surveys earlier conducted or to be conducted. The circle on the right depicts alternative data sources, e.g. big data.

19. The basic principle of collecting data may seem simpler in case of alternative data sources compared with traditional sample survey data collection. However, in practice there are three major problems to consider. The first is the technical capture, editing and storage of data, the second problem is legal and ethical issues, and the third is data owner's willingness to share data. When alternative data has been captured, it has to be processed and edited followed by appropriate analysis. This may involve much more complex statistical modelling than is presently conducted at NSOs. These parts also involve combination and integration of data with other data sets.

20. Now, with all these resources in place, it is possible to use them in different combinations to provide statistics in a given area. It can be purely register based, purely sample survey based, or it may be utilizing all resources.

21. Unfortunately, survey sampling theory and methodology are often thought of as theory and methodology for studying the populations of objects of concern at an NSO, for example, i.e. populations of citizens, households, enterprises, etc. Survey sampling theory and methodology are general tools and can be used for special purposes in alternative survey designs involving both register and alternative data sources.

22. As an example, intercept studies are often used for tourist statistics, where foreign visitors are indirectly sampled using a sample of interception points, e.g. airports, harbours and road border crossings. These designs follow the standard format of defining units of interest, population of units, sampling design and direct data collection via e.g. interviews. If the purpose of the study is to provide data for satellite accounts it would be sufficient to estimate total number of visits and mean expenditure per visit. This opens for different survey designs where sampling plays a different role. For instance, mobile positioning data in combination with validating (calibrating) sample surveys might be used for estimation of the total number of visits. Similarly, credit card data in combination with a sample survey might be used for estimation of mean expenditures. In such a design, sample surveys play a different role bringing pieces of necessary information not available in the main alternative data sources.

23. The interesting part in this production system is the "Problem formulation/Survey design" part. As is illustrated by the example above, by reconsidering the problem formulation, alternative designs may be optional with more efficient use of available resources. Preferably but perhaps not possible, it is desirable to go one step further back and reconsider the decision problems motivating the survey. The problem formulation and survey design part will be essential for the successful use of alternative data sources.

24. A prerequisite for a capable production system fit for handling future demands and challenges is close cooperation over departments in the NSO; subject specialists, methodologists, IT-specialists, and others have to work together in finding appropriate solutions to a survey design. Combining knowledge and skills from different departments is essential for finding and testing optional survey designs. Understanding user's needs and the contexts defines what statistics to produce. Understanding available resources and how they can be utilized defines how statistics can be produced. This has to be acknowledged as an interactive process.

V. Future roles of statisticians and data scientists

25. Focusing on the title of the seminar many new roles and competencies in the future NSO are excluded. The discussion of those are not made here but are acknowledged and implicitly given by the description of SP 4.0.

26. Data science and data scientists are new concepts and it is possible to find several definitions of these terms. The terms are often associated with scientific research and in business applications. Here they also address the process of deriving new knowledge (boyd and Crawford, 2012). The main role of an NSO is not to derive new knowledge, on par with theories and models explaining real world phenomena. An NSO produces descriptions of the state of the country, e.g. unemployment rate, inflation rate, GDP, etc. Thus, an NSO need not to adapt the new approach of deriving "knowledge" or "insights", but can utilize data science tools and data scientists skills in revealing hidden structures in large and mixed data sets. These are in turn used in a survey design yielding official statistics complying with traditional statistical quality criteria. This may show official statistics production being an appropriate application area for data science, because validity of results within the model assisted survey sampling framework is not dependent on the "truth" of the model.

27. As implied in the description of SP 4.0, the key part is problem formulation and survey design. Here it will be necessary with several competencies working together in a "task force", and statisticians and data scientists play central roles. A main feature of their competencies is broad knowledge in their areas on what can be accomplished with different methods and techniques. A strategy is required securing access to appropriate competencies.

28. Statistics will also in the future be produced with sample surveys, supported by administrative data and data from alternative sources. Survey sampling will in other designs be used for supporting surveys based on administrative data and data from alternative sources.

29. In addition, statistical competencies for handling and using administrative data are also required in the future production system, and will probably be of greater importance. New administrative data sources at governmental agencies are continuously generated. There are also data available at other public authorities which are not utilized presently but may be of interest in the future. Similarly, data collected form organizations in the private sector require skills in handling register data.

30. A final aspect on the future roles of statisticians and data scientists are communication and fostering of users. The traditional way of publishing point estimates with associated margins of errors is sufficient if the user has enough training in statistics theory. This is mostly not the case and users don't know what to do with the margin of error whereby it is neglected. Not interpreting statistics within the traditional statistical accuracy framework invites alternative producers on the market, providing less costly statistics but of much less quality and potentially harming. Here it will be important to find ways of communicating statistics in such a way the user can correctly interpret statistics and at the same time gain knowledge on accuracy properties of statistics. One first and simple action is to replace the combination of point estimates and margin of errors with a confidence interval, and naming it an "interval estimate".

VI. References

Boyd, D. and Crawford K. (2012). *Critical Questions for Big Data,* Information, Communication & Society, 15:5, 662-679.

Brick, J.M. (2013). *Unit Nonresponse and Weighting Adjustments: A Critical Review*, Journal of Official Statistics, 29:3, 329-353.

Chatterjee, S.K. (2003). *Statistical Thought - A Perspective and History*, Oxford University Press, New York.

Commission on the Future of Sweden (2013). Future Challenges for Sweden, Prime Minister's Office, Sweden, Ds 2013:19 (English translation).

Daas, P.J.H., Puts, M.J., Buelens, B. and P.A.M. van den Hurk (2015). *Big Data as a Source for Official Statistics*, Journal of Official Statistics, 31:2, 249-262.

Japec, L., Kreuter, F., Berg, M., Biemer, P., et al. (2015). *Big Data in Survey Research: AAPOR Task Force Report*, Public Opinion Quarterly, 79:4, 839-880.

Tang, N. (2014). Big data cleaning. In L. Chen, Y. Jia, T. Sellis, and G. Liu (eds) *Web technologies and applications*, *Lecture notes in computer science*, 8709, 13–24. Cham: Springer International Publishing.

Wallgren and Wallgren (2014). *Register-based Statistics – Statistical Methods for Administrative Data*, *2nd Ed*, Wiley, New York.