

Comments in Journal of Official Statistics to Richard Platek and Carl-Erik Särndal: Can a statistician deliver? October 2000.

---

## COMMENTS TO 'CAN A STATISTICIAN DELIVER?'

OR

## CAN A STATISTICIAN NAVIGATE HIS STATISTICAL PRODUCTION?

By Svein Nordbotten  
P.O.Box 309 Paradis  
5856 Bergen  
Norway

### 1. 'Can a statistician deliver?'

*Can a statistician deliver?* by Platek and Särndal (abbreviated to P&S below) is one of the important recent papers on the activities of national statistical institutes (NSIs). Their discussion covers many questions including statistics disseminated by national statistical institutes, quality concepts associated with statistics, the categories of statisticians working in a NSI, their varying views on quality aspects, the users' assumed needs for information about the statistics they receive/use, design and process characteristics, and the total survey error model. Their main conclusions are: There is no consensus about the quality concept, a comprehensive theory of statistical surveys does not exist and is not in sight, and the statisticians do not deliver satisfactory information about the statistics.

Can the statistician deliver? P&S answer honestly: "*We have no complete answer.*"

The excellent paper by P&S with all questions raised deserves discussion and comments. Based on his personal background and interests, the present commentator has selected only a few questions for discussion and suggests addition of a few others.

### 2. *Who are the users of statistics?*

P&S ask: '*Who is the statistician?*' and point out a number of different categories of specialists working in a modern NSI. Each category has often their own perception about the aim of the statistical production, but the overall justification for preparing statistics must of course be that external needs for the statistical information exist.

Without a need for statistics, there would be no external users and no justification for producing statistics. But who are these users and what are their needs? Traditionally, they were government servants each with needs for statistics to solve their respective public tasks. Later appeared representatives from local authorities with similar needs, but for different regions. It was easy for the statistician to communicate with the relatively few users. Funds were often granted directly connected to these needs, and it was assumed that the NSI prepared the statistics as well as possible within the budget and according to deadlines.

Today, the composition of users are different. Statistics are still used as an important basis for decisions in public affairs. In addition, statistics have a number of users within research, commercial firms, education, media organizations, and among individual citizens, all with different needs. The users probably have more varied views on statistics than the statisticians within a NSI. By means of modern information technology (IT), more and more users are becoming aware of official statistics as a valuable source of information. For instance, as a consequence of increasing popularity of Internet and free access the Swedish Statistical Databases (SSD), the number of registered users of the SSD increased from the end of 1999 to the end of August 2000 with almost 300 pct., and the users were mainly private companies and individuals.

*Does the statistician know the users of his statistics?* Probably not well enough. Only the most advanced NSIs have established marketing sections with responsibility for collecting and analyzing information about the users. Until recently, most statisticians seemed to loose interest in their statistics when the statistics were finished and left their professional process or were disseminated in statistical publications.

*Is it now the time to allocate more resources in the NSI for research in the composition of existing and potential users of official statistics?*

### **3. What kind of statistical data are needed?**

Until recently, perhaps still, most of the statistics were designed according to international statistical standard concepts and methods for data collection, processing and publication. These standards were created by cooperation between a few important users and the statisticians themselves. The great advantage of these standards is of course that they promote comparability in statistics from one year to year and among different countries.

But since we know only vaguely the emerging user population, *can we be certain that these standards reflect relevant concepts for the majority of statistical users? Why do more advanced users now start asking for micro data to process themselves instead of buying the standard publications and tables of their NSI? Why are the statistical*

*databases such as SSD and the accompanying programs such as PC-AXIS becoming so popular?*

IT has contributed to globalization, erased some of the obstacles of geographical distances, increased accessibility to data and improved users capability to process and analyze the data on their own desks according to their own needs. Only a decade ago, few users accessed other NSIs than that of their own country to obtain statistical information for their decisions concerning international affairs. IT has made foreign NSIs and International Statistical Agencies (ISAs) to important data sources for many users. New and needed data sets from multiple sources related to global tasks are becoming as easily accessible on World Wide Web (WWW) for the users as the NSI of their own country.

As discussed above, IT has also made statistical databases with microdata and low level statistics easily accessible for the users to process themselves to the statistical aggregates and numerical models they need to solve their particular problems. These kinds of statistical data were previously demanded mainly by a small group of researchers, but have now become requested also by private companies to get answers to current and future problems. *Will the relative demand for statistical microdata continue to increase and will the traditional statistical publications play the role of statistical catalogues? Is the statistician prepared for this kind of delivery?*

Data recordings of communications with users in the form of access and database logs are available in NSIs, which offer their data on WWW. *Can the statistician perhaps learn about and prepare himself for his electronic future by systematic analysis of these logs?*

#### **4. What is quality of statistics?**

The main theme in the paper by P&S is quality. Measures of quality have been important as a goal for the statistician to establish and provide as a guide for users who want to evaluate the usefulness of the statistics for his purpose. In *Can a statistician deliver?* it is pointed out that the perception of quality of statistics varies from one NSI to another, among categories of statisticians within the same agency and probably also within the same category over time. I believe that the perception of quality probably varies even more among the users.

P&S observe that the accuracy, and the associated total error, are vital aspects of the quality concept, but that this quality aspect is particularly difficult to measure. For this reason, the NSIs in their information about the statistical products frequently focus on other quality aspects such as timeliness. P&S indicate as an explanation that the editing operation is a neglected area from a theoretical point of view without many interactions with the sampling error specialists and their mathematical tradition.

Interaction between the editing and the sampling specialists is not strong, but the academic interest in editing and imputation has fortunately increased during the last decade. For a number of years, UN/ECE has organized workshops for data editing, and these workshops have also attracted an increasing number of theoreticians. In year 2000, a large research project on editing and imputation, EUREDIT, with financial support from the IST Programme of the European Union, was started. This project has partners from NSIs and private research companies as well as from Statistical Departments and Computer Science Departments of several universities in Europe. The results of this and similar projects will hopefully contribute to development of a theory of editing, its integration with the theories of other statistical processes and to a stronger interest in exposing the accuracy component of quality by the NSIs.

It is also interesting to note that editing, the purpose of which is to improve accuracy in statistical data, as a process is by no way neglected by the NSIs. International studies indicate that in a typical statistical survey, up to 40% of the budget may be spent on editing.

Quality is one of many survey design and process characteristics useful for both statisticians and users. Many of the problems discussed by P&S can be extended to the general class of metadata. Some NSIs are already building up special metadata systems to be able to provide their own survey designers and external users with useful information about the design and processes on which published statistics are based.

*Is one way to extend the interest of the sampling specialist to non-sampling errors to revive the total error model concept in academic teaching?*

## **5. Can the statistician declare the quality of his statistics?**

Statisticians and users agree that statistics must be disseminated with some kind of accompanying metadata. Many NSIs have had quality declarations as a goal for decades. But as explained by P&S, there is no consensus about this concept and how the measure should be published.

P&S are worried about the tendencies in NSIs to focus on the less complicated quality characteristics as timeliness, relevance, and accessibility, leaving the more difficult accuracy component in the background when publicizing their products. In contrast to accuracy, the three first mentioned characteristics do not represent any difficult measuring or declaration problems. I would like to second P&S in their worries about the missing attention to objective accuracy measures because I believe that the group of users wanting

to use an accuracy measure to choose among alternative information will be increasing in the future.

The relative importance of the different quality components varies among the user categories. If the users prefer rapid, but less accurate statistics they should get what they demand. It must still be the responsibility of the NSI to inform the users about the price in lost accuracy paid for rapid results. Imagine the users of statistics from a decennial census with highly accurate figures as to the state of affairs at the census date. Assuming the uses are about evenly distributed over the decade, they have in average to rely on more than 5 years old statistics. Some may prefer less accurate statistics reflecting the situation at the census date if they could get updated statistics every 5 years. Obviously, users with access to accuracy declared statistics, but for them irrelevant statistics, may prefer more relevant statistics without quality declarations because, as P&S point out, they trust that a NSI will not release useless statistics.

*Has time come for the NSIs to use their own tools for active collection and analysis of more data about the statistical users' views on which quality component measurements are important for them?*

## **6. Can the statistician compete?**

These comments started by the question 'who are the users?' The information processing and communication tools have developed extremely fast during the last 10-20 years and the trend promises further advancements. In some industrialized countries more than half of the population had access to Internet at the beginning of the 21<sup>st</sup> century. Predictions indicate that in 5-10 years, 15-20 % of the total world population will have access to Internet. So far, access has been limited to communication hardware getting power regularly from a plug-in power source and being connected directly by cables to the communication nets. Tomorrow, the use of mobile, wireless computers and telephones powered by solar energy will be common tools. The application potential will increase significantly.

Information exchange and intensive knowledge use seem to become necessary requirements for further development. Statistics are a special kind of compact information useful for an endless set of decisions. As such, statistical information will become demanded, evaluated and assessed as economic objects and attract the interest of the business community. So far, the statistical information production has mainly attracted opinion survey operators, but it must be expected that in few years commercial operators will appear as processors and providers of statistical information in competition with the NSI. *If the new providers are able to offer more timely, relevant, accessible and/or accurate statistical information than the NSI's, can the statistician in a NSI compete?*

The comparative advantages of the NSI are their present access to administrative data registers, compulsory collection of survey data and ability to offer statistical information as a free public service. In some countries, the NSI has developed techniques for using these data treasures in their data production. *Will the statisticians in NSIs be able to maintain their monopolies to the valuable data sources or, with the trend toward privatization, must they be prepared to share these access privileges with other operators? Can we expect that the official statistics will continue to be regarded as a free public commodity in the future?*

### **7. Can the statistician navigate his statistical production?**

I do not intend to offend friends in the NSIs by this question, but to ask *if the statistician has the necessary information about the user needs to know the 'destination' for his activities?* Without knowing the users' needs, he will not be able to decide in which direction he should aim when designing his surveys.

*Does the statistician have the necessary metadata about the statistical processes within his organization to design a cost-effective survey? Does the statistician have the tools to guide him in selecting important metadata, and to use the information on statistical needs and metadata for improving his statistical products by allocating his limited resources among different surveys in an efficient way?*

In my opinion, the most important statement in the P&S paper is *that no general theory of statistical production exists*, not even a framework for such a theory. The theory required is not simply a new statistical theory, but a synthesis of elements from statistical theory, economic theory, social sciences, psychology and management theory. P&S conclude that a complete theory is not even in sight. *Is it now time for initiating work toward a comprehensive theory for the production of official statistics? How can it be done? Who can do it?*

### **8. Final remark**

These comments started with a praise of *'Can a statistician deliver?'* At the end, I would like to express my thanks to P&S and the editor of JOS who gave me the opportunity to reflect about these questions related to official statistics. I am convinced the P&S paper will generate discussions far beyond this volume in JOS. Statisticians should be looking forward to reading the discussions.

As a final remark, I should like to join P&S in repeating that *the overall most important problem is the lack of a general theory for the statistical production. Without guidance from a complete theory, the statistician of a NSI will have problems to navigate his statistical activities successfully into the future information society.*