Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 93–97

Comment

Erkki Pahkinen[1]

The article by Platek and Särndal offers a wide, highly relevant, and challenging perspective on whether the statisticians working in statistical agencies are able to estimate the quality of their own products and deliver this information to the users of statistics. The article ultimately deals with the question whether the statistician from a statistical agency is able to deliver relevant information from the customer's viewpoint and knows the quality of statistics and is capable of communicating it to the users. The writers' answer is that he or she cannot. This argument is based on two things. First, the statistician and the user do not have a common vocabulary concerning quality for their mutual communication. Second, the subject of communication is survey design or process, which does not have a firm background theory. It would comprise a necessary framework for the straight forward monitoring of total quality.

It appears that the scarcity and unevenness of quality estimation is based on two other things. On a more general level, it is due to the monopolistic position held by statistical agencies as the only producers of official statistics. As a competing statistical product does not exist, quality improvement is regarded as of subsidiary importance. A special level, or a level closer to the statistician, is reached when he or she starts to act as an evaluator of his or her own work. Criticising oneself, let alone one's own statistical products is difficult. At this point, we refer to an external quality monitoring system that is used elsewhere. In pharmaceutical industry, for example, a strictly normative international protocol is used for analysing and carrying out statistical studies.

In general, one single statistical agency in each nation state produces 50-100 percent of all official statistics. As a consequence, the agency gains a monopolistic position as a producer. This is confirmed in texts of law and decrees. Thus, the user of statistics is at the mercy of the producer because a competing product does not exist. This is perhaps one reason why it is preferred not to attach quality indicators to the product. They would hardly be of any use, since they could not be compared to those of a competing product. The user has to accept the product as it is delivered by the statistical agency. There are of course situations in which two competing statistical products can be compared. A good Finnish example of this is the number of unemployed (according to the ILO definition) published by the national statistical agency. This figure is based on a monthly labour force survey conducted by the agency. Another figure is the number of unemployed job seekers published monthly by the Ministry of Labour. This figure is based on an administrative register of employment authorities. It is well known that the two figures deviate from

[1] University of Jyväskylä, Department of Statistics, Jyväskylä, Finland.

each other. The public asks which one represents the truth. The statistical agency argues that the LFS figure is based on internationally agreed concepts and on a well defined production procedure of high quality. The Ministry argues that its figure reflects the real world better. Both figures are correct in their own context, but it is difficult to explain this to the public. When every statistical agency produces hundreds of statistical products, it is evident that a competing product does not exist for each of them. Thus, the responsibility for product quality rests on the shoulders of the statisticians and survey methodologists working in the agency.

An external quality monitoring system in which the actors are not the ones producing the statistics is needed in official statistics. They are incapable of evaluating their own work. The reason why Platek and Särndal have not paid attention to this in their challenging article must be an excessive trust in the statistical agency's capability of maintaining and monitoring the quality of its own products. Perhaps the reference agency, Statistics Canada, has led their thoughts astray because Statistics Canada is a) big enough and b) has nurtured its own science community. Things are not necessarily that good elsewhere.

The attachment of scientific expertise and the products of a statistical agency to the daily life of a science community, like a university, undoubtedly generates an essential element of criticism. It would be good if statistical agencies collaborated more with universities. In this respect, significant redefinitions of policy have taken place in the Nordic Countries. For instance, Statistics Finland has launched short-term and long-term scientific co-operation projects with several universities in the country. The scientific and educational co-operation network has also been documented in the latest research policy agenda of the statistical agency (Statistics Finland 2000). It consists of a permanent, highly advanced methodology network between Statistics Finland and the University Departments of Economics, Social Sciences, and Statistics. New perspectives on science policy have also opened up at Statistics Sweden. According to the new head of the agency, microfiles could be used for research purposes (Öberg 2000). This would increase the universities' prospects for conducting empirical study. Simultaneously, different research groups would be able to produce statistical products from the same raw data. The users would have a freedom of choice, and the providers of competing products would have to focus on quality issues as well.

I do not agree with the writers' view that there is a severe defect in the curriculum of an academic program if it does not include a course in National Accounts or a course whose name refers directly to a statistical product. Curricula may entail such courses without this implying any guarantees that highly qualified new statisticians will be recruited to statistical agencies. The significance of statistical agencies is important to the academic education of a statistician, but only marginally. A closer inspection of any country would reveal that the major part of statisticians defined in the article are employed outside the national statistical agencies. Such places of employment include research institutes in various fields of study, research and development departments of innovative corporations, and different sectors of public administration. By considering the matter from an academic viewpoint, one sees that statisticians can be divided into two main groups, biostatisticians being one of them. It is very unlikely that one will come across biostatisticians in statistical agencies. Thus, the basic education of a statistician varies considerably. Important qualities of this education are a good academic degree within some subject-matter or method area, for instance in statistical science, supplemented with team work skills. Learning team work skills takes place in a real learning environment if the master's thesis is written outside the university, for instance at Statistics Finland or some other producer of statistics. We have had good experiences from this procedure. (As for cooperation between statistical agencies and universities, see Pahkinen 2000; U.S. Subcommittee on Survey and Statistical Training in Federal Statistical Agencies 1998.)

Altogether, the article by Platek and Särndal offers a fine perspective on quality issues of survey production. According to them, a comprehensive theory of quality does not exist, rather it is fragmentary, built on the special measurements of each part of the production. This view is fairly unanimous around the world. The writers complain that their own educational background as statisticians in the academic community narrows their viewpoint and causes them to focus attention too heavily on the study of sampling errors and quality indicators attached to them. For instance, researchers oriented towards behavioural sciences would rather pay attention mostly to measurement error, for example the extent to which the measurement results indicate the effects of the interviewer, the interviewee, the data collection mode, and their interaction. For instance, the clustering of interviewees' responses caused by the interviewer can be measured by Kish's intra-class correlation coefficient, which should be emphasised in the same way as standard error and coefficient of variation.

Usually, survey research is conducted by a team of several statisticians. A natural division of labour is to give each member of the team the responsibility for his or her own field of expertise. One of the most valuable discussions in the article by Platek and Särndal deals with the monitoring of the quality indicators for survey research that a) would signal something concerning the total quality of survey research and b) could be interpreted by the users of the statistics and not only by the statisticians producing them. Because the writers seem to be stuck with the frills of statistical terminology, the perspective of their discussion is somewhat narrow. Thus, we shall widen ours in this respect.

The quality of survey research is clearly divided into inner and outer elements. The external quality indicators are needed for the users of statistics and the inner ones for the control of the production process. The inner measurements involve coverage rate, response rate, coefficient of variation (c.v), and mean squared error (MSE). Team members conducting survey research prefer quality indicators which they can influence by regulating the production process. This possibility is suggested in the article by Platek and Särndal. Let us take an example from a computer assisted telephone interview, where we shall examine the data collection mode. The data collection comprises the identification of the phone numbers, making contact, the actual interview, and the generation of the logfile connected with the collection process. Quality is associated with telephone coverage, questionnaire, nonresponse, and interviewer effects. Each of these four quality dimensions has its own quality indicators.

The construction of the total quality indicator of survey research turns out to be a utopian project, even though the attempt has been made. In most cases, the starting point has been the mean squared error (MSE), which includes the variance estimator of the unbiased design-based estimator added by the squared bias caused by measurement error. In a typical quality evaluation situation (nonresponse), the component of bias is based on nonresponse assumed to be nonrandom. Thus, the estimated parameter of the response group is different than that of the nonresponse group. The unknown parameter value of the nonresponse group must be estimated to allow the addition of a component of bias. As it is also an estimate, it has a variance that should be added to the formula of the mean squared error. The situation becomes even more complicated if we consider the measurement error associated with the actual measurement instrument. This indicates that the components of bias and variance are taken into account. If the error components are pairwise dependent, the computation of the mean squared error becomes a laborious and complicated process. Thus, the theoretically simple mean squared error proves to be a dissatisfactory quality indicator right at the beginning of the survey research. Let us consider the situation now from the point of view of the user, to whom the statistical agency offers the single mean squared error value to accompany the statistics. In order to get a solid grasp of the concept, the user should be backed up by the statistician's professional competence in interpreting the significance of this quality indicator.

The production of official statistics is characterised by variety, because surveys are more or less single cases to which unifying quality standards are not given. The situation is clarified by the following example taken from the pharmaceutical industry, in which a large number of statistical experiments and statistical analyses regarding the data are made. As the results need to be comparable worldwide, it is agreed that the experimental designs and statistical analyses of the data should follow a strictly normative protocol. It should be noted that the statistical analyses should be made with SAS-software. This ensures such a basically simple fact that in every analysis the variance is computed with the same algorithm. If we compare variance estimation in official statistics, different software packages – BASCULA, CALMAR, CLAN, GES, SAS, SPSS, SUDAAN, WESVAR, etc. – have been used depending on the agency and country. Thus, how can one compare even inner quality if the algorithms for the estimation of the same parameter vary?

An example of outer quality indicator and its interpretation can be taken from the practice used by the Finnish media when publishing the results of opinion polls. As late as the beginning of the 1990s, it was customary to publish the support figures of political parties with the precision of decimals. (The support of party A is 12.3 percent.) Next, ''confidence interval'' was introduced, but because of the obscurity of the term, ''the margin of error'' was taken in use instead, which is more familiar to the general public. From then on, results were published as point estimates +/- the margin of error. (The support of party A is 12.3 percent +/- 2 percent.) The adding of the margin of error indicates the randomness of the opinion poll, and also the fact, that the measurement results are ''approximate'' figures. Towards the end of the 1990s, the practice was streamlined so that decimals were left out of percentage figures because such precision was not relevant in the estimation situation. (The support of party A is 12 percent +/- 2 percent.) What does the simple example given here indicate? If we intend to construct an outer quality indicator for survey research, the user should be able to interpret it unambiguously. Also, the quality indicator or a measurement like it should be used regularly either in professional or private environment. This brings one back to the school grades of research survey: $4 = \text{fail}$, $6 = \text{passable}$, $8 = \text{good}$, etc. The grade refers to many quality indicator parts of a research survey. In the pharmaceutical industry, very simple external quality indicators are used, which are familiar to the users. The trade description reveals against which disease the drug works and which possible adverse effects can appear in certain situations. These matters are based on statistical research.

In sum, it is suggested that factors, which have an effect on the total quality of survey research are the professional competence of the statistician and the quality level set from outside the statistical agency. In this respect, measures should be taken towards the improvement of the scientific expertise of a statistical agency and towards the development of scientific networking. External quality indicators should be defined by multinational organisations like Eurostat (Europe) and the Federal Statistical System (U.S.). In this respect, the international pharmaceutical industry offers a good example of a strictly controlled statistics production. These ideas were inspired by a very interesting article by Särndal and Platek, to whom I would like to express my thanks.

Additional references

Federal Committee on Statistical Methodology (1998). Training for the Future: Addressing Tomorrow's Survey Tasks. Working Paper, 27. Statistical Policy Office, Washington, DC, U.S.A.

Pahkinen, E. (2000). Professional and Educational Qualifications as Basis for a Master's Program in Statistics. (In Finnish). Publications of Department of Statistics, University of Jyväskylä, 19/2000, Finland. Updated English version in preparation.

Statistics Finland (2000). Main Lines of Research and Development in 2000–2003. Reviews 2000/6. Statistics Finland, Helsinki.

Öberg, S. (2000). Information and Knowledge: The Role of Statistics. Proceedings of the 86th DGINS Conference. Porto, Portugal.

Received December 2000