Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 33–37

Comment

Alain Desrosières[1], Jean-Claude Deville[2], and Olivier Sautory[3]

Inaccurate Measures of Fuzzy Concepts?

Among the five principal dimensions of quality referred to by Platek and Särndal, two of them seem to be orthogonal, and orthogonal to the three others: timeliness and availability/clarity. It can be pointed out that timeliness seems to be measurable, but not availability. On the contrary, accuracy (which is the major topic discussed by the authors), contents (or better still, relevance) and coherence/comparability are very related and interwoven. Among those criteria, only accuracy seems to be measurable, essentially with the use of probabilistic tools.

In the literature about quality, ‘‘relevance’’ and ‘‘accuracy’’ are often presented separately. ‘‘Relevance’’ is supposed to be appraised through customer satisfaction surveys, while ‘‘accuracy’’ is described in full detail. It would be useful to transform these usual views, through two steps:

1. to give back relevance its semantic importance: what is the meaning and what are the uses of the measured concept?
2. to make explicit the continuum of practical problems linking ‘‘relevance’’ (understood that way) and ‘‘accuracy.’’ The measurement of unemployment gives an excellent example of such a continuum.

The relevance of a statistic finds its own limitation in the fuzzy character of any statistical concept. The more elaborate is the concept, the fuzzier becomes its operational definition. Examples of statistical concepts, ordered according to ascending complexity are: number of wagons crossing the French borders (as mentioned in the French Monthly Bulletin of Statistics in the forties), people living at such and such a place, people having a specific occupation, people out of work, people presenting psychological disturbances, literacy level. Even with a more detailed and apparently more precise definition, those categories cannot be directly turned into statistics. A concept can be transformed into statistics if there is a means to define the concept by the possibility of ticking the appropriate box. This statistical activity is the transformation of a conceptual definition into an operational definition. It relies on a well defined protocol, including a questionnaire, a way to use it (face to face, postal, telephone, Internet, etc.), coding, data storage and so on. A completely formalised protocol (which is also an idealisation) is a reduction of the fuzzy character of the concept. And it is only this (ideal) reduction which allows us to speak of a

[1] INSEE, Division des Méthodes Comparées, Paris, France
[2] CREST-ENSAI, Laboratoire de Statistique d’Enquête, Paris, France
[3] INSEE, CEPE, Paris, France

true value, and therefore of an accuracy of the measure of this true value. ''The result of a test becomes the measure of intelligence.''

This reduction to a protocol is somewhat arbitrary and is always a choice between many possible solutions. Many factors come into play, such as technical feasibilities, costs, and even traditions in statistical offices.

Moreover, concepts depend on the epoch and on the area. For example, developed countries can agree on the definition of unemployment at the end of the 20th century, but this definition would make no sense in other countries and had no sense some centuries ago; for example, what does an ''active search for a job'' mean when using the ILO criteria of unemployment? It can depend on the institutional rules of the country.

From another angle, there is often an organisational problem: ''relevance'' problems are dealt with by ''subject matter'' specialists, while ''accuracy'' problems are processed by another category of professionals, the ''methodologists,'' for whom the definition of the ''concept'' is something *given*. That social division of labour reinforces the cognitive division between the semantic questions encapsulated in ''relevance,'' and the methodological ones analysed by Platek and Särndal.

Then, even if the protocol is supposed to be ideal, there remain many factors that will make the measure imprecise. Let us mention some of them:

- The protocol is not exactly applied (this leads to a large category of ''measurement errors''). It must be emphasised that the ''ideal'' protocol generally cannot be exactly followed (even in the case of the wagon crossing the border), and that it is never exactly reproducible.
- The protocol has to be applied to the whole population (which itself may be an abstract concept transformed into an operational one). Very often, the measure is performed only on a sample. To make inference to the whole population, sampling theory provides tools allowing one to measure the accuracy in some sense (standard error). However, this presupposes a completely controlled sampling procedure. In particular one has to use a perfect sampling frame allowing one to contact every member of the population. In this perfect sampling situation, one can say, in some sense, that imprecision is the same among all ''directions.''
- In the case of a census as in the case of a sampling survey, nonresponse is never avoided and makes the results more imprecise. However, the imprecision increases more or less according to the ''directions'': some categories of people are more likely to respond than others, some variables produce more difficulties than others, etc.
- Using auxiliary information at the sampling design/estimation stage can reduce the imprecision along some particular ''directions:'' variance decreases when the variable is closely related to auxiliary variables.

If we look at the possibility of measuring the accuracy and of improving it, the role of sampling theory is very particular. In the design-based framework, this theory is completely mathematised and applies to any sampling design, and moreover to any particular statistics based on any particular (set of) variable(s) (defined by the ideal protocol). This is the real strength of error theory in sampling surveys: we have a universal method to measure accuracy (even if, in practice, it is not always easy to apply).

For the effect of compensation for nonresponse, the situation tends to be nearly the same: we now have a better understanding of reweighting and imputation procedures. We can also isolate the impact of auxiliary information, especially the role played by models in nonresponse treatments. Great progress must still be made in this research field, but we can hope that satisfactory solutions are already available or will be found in the near future.

The situation is very different for inaccuracies coming from field operations. All the development in sampling theory and nonresponse treatment can be achieved at a very low cost (people thinking in their offices and using computers). By contrast, inaccuracies associated with field operations can be evaluated only by using special and generally expensive operations.

Here is an example. In France, most ''official'' surveys are mandatory. Of course, this rule (''obligation'') is not enforced in population surveys, but it can be used as an argument by the interviewers to convince people to respond to the survey. At INSEE we had to evaluate the difference between mandatory and nonmandatory surveys regarding nonresponse rates. The only means was to design a special survey (see Berthier and Dupont 1997). From a set of primary units two equivalent samples were drawn, each of them being surveyed with a well-defined protocol and in one sample the obligation argument was used but not in the other. The operation was somewhat expensive, but it was found that the proportion of refusals increased from 9 percent to 19 percent when obligation arguments were removed. At the same time, the proportion of ''not at home'' was the same (about 5 percent) in both samples.

Along the same lines, evaluation of the effect of incentives on response rates requires special experiments (see for example Singer et al. 1999).

Evaluation of measurement errors due to the absence of adherence to the ideal protocol is dramatically more complex and costly. It depends completely on the protocol and the measured variable, and can only with difficulty be applied to other settings. In particular, it is never straightforward to use knowledge coming from one survey in the planning of another one. Therefore such evaluations are very rare, because they cost very much and cannot be generalised at all. In France, we had, for instance, once the opportunity to measure an interviewer effect on a particular survey (see Berthier, Deville and Néros 1999). Unfortunately, it was not possible to perform the same operation on the Labour Force Survey, although this is the major survey, for which we could have learned something.

The summit of complexity arises when we have to compare different protocols for measuring the same concept. We can observe that it is almost always the case when we have to compare statistics elaborated at different periods (from this point of view, even the comparison between periodic population censuses is generally not very easy).

It may happen that different surveys allow one to measure the same variable with different protocols. For example, the question(s) may differ. In French surveys, the variable ''Income'' was once available in four different forms:

- as a specification of an interval to which the income belongs (Survey a).
- as the amount of total income (Survey b).
- as the amounts of a number of different components of income (Survey c).
- as the amount declared for income tax (Survey d).

For each survey, correlates of income (training, age, sex, professional classification etc.) were known and used as regressors. The four regressions gave significantly the same result, except that, from **a** to **d**, the standard deviation of the residuals was decreasing. The result could be (carefully!) interpreted as a reduction of the measurement error when the questionnaire is more precise (and also more costly).

However, every variable requires a specific methodology depending on the means used to obtain the information. A particular class of questions, for instance, appeals to the memory and is relevant to very special investigations (see Auriat 1997). More generally, questionnaire designs should appeal to many branches of, and specialists in, social and behavioural sciences such as sociology, linguistics, ethnology, and psychology. One of their contributions to statistics would consist in defining a reproducible protocol for reducing the fuzzy concepts to measurable ones.

However, the result is somewhat arbitrary and may depend on occasional factors. For example, at INSEE, an experiment was carried out to measure the difference between the unemployment rate reported by face to face interview versus telephone interview (see Lagarenne and Schuhl 1997). A first survey was carried out by telephone on 2,421 persons, 8.5 percent of them being classified as unemployed. One week later, a face to face reinterview gives 8 percent. More precisely, we have Table 1:

	Employed or inactive, telephone	Unemployed, telephone
Employed or inactive, face to face	2,201	27
Unemployed, face to face	15	178

Fortunately, the difference was not significant at the 95 percent level for a classical test. However, data can be interpreted from a Bayesian point of view. Giving a priori a 50 percent chance for one survey to produce an overestimation, a posteriori telephone interview gives a higher estimate with more than a 90 percent chance.

This experience shows that a trade-off has to be made between the fuzzy character of the concept and the never completely reproducible character of the protocol, and that there is a limit of the desired precision under which no discussion makes sense.

Is a measure of precision an impossible dream? Yes, if we have the ambition to take rigorously into account all the parameters of a statistical operation. However, we may have good hopes if we have more modest objectives and if we try to tackle the problems one after another. Ideally, it would be possible, for a particularly simple survey, to study all those parameters and to measure their impact on accuracy. Except for sampling errors, a general system, valid for all surveys, is out of scope.

Another factor of progress may come from the integration in the statistical work of some knowledge and know-how borrowed from the domain of the social and psychological sciences.

Having a reliable measure of accuracy would be a good thing. However, it would not be sufficient to get a system of optimisation of the accuracy: we are not sure that the best practices in different fields like sampling, avoiding nonresponse, decreasing measurement errors, are compatible in the same survey, because the best practices are generally also the most expensive! We are led to the need to parameterise the best practices in each field by the amount of money available for this field. And that is another story!

References

Auriat, N. (1996). Les défaillances de la mémoire humaine — Aspects cognitifs des enquêtes rétrospectives. Travaux et Documents, cahier n°136, INED/PUF, Paris. (In French).

Berthier, C., Deville J.-C., and Néros B. (1999). Une méthode de mesure de l'effet enquêteur. Actes des Journées de Méthodologie Statistique 17-18 mars 1998. Insee-Méthodes n° 84-85-86, 133-143. (In French).

Berthier, C. and Dupont, F. (1997). L'incidence du caractère obligatoire des enquêtes. Actes des Journées de Méthodologie Statistique 11-12 décembre 1996. Insee-Méthodes n° 69-70-71, 131-146. (In French).

Lagarenne, Ch. and Schuhl, P. (1997). Contrôle de qualité de l'enquête trimestrielle emploi : résultats de l'enquête Protocole. Actes des Journées de Méthodologie Statistique 18-19 octobre 1995, Insee-Méthodes n° 59-60-61, 389-413. (In French).

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., and McGonagle, K. (1999). The Effect of Incentives on Response Rates in Face-to-Face and Telephone Surveys. Journal of Official Statistics, 15, 217-230.

Received December 2000