Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

COMMUNICATING VARIANCE AND SAMPLING ERRORS TO USERS

Ed Swires-Hennessy[1]

ABSTRACT

Statisticians are developing additional concepts for communicating errors associated with estimates. Many of these concepts are readily understood by statisticians but are even more difficult to explain to users than the traditional confidence interval. The proposed solution, when communicating with non-statisticians, is to improve the estimates so that the requirement for explaining the error is minimised. The user is then not confused by having too many numbers to understand.

KEY WORDS: Statistical error concept; Communication; Users.

1. DISCUSSION

The science of statistics, indeed the role of statisticians, is to derive estimates from incomplete data. Such estimates have errors. The paper by John wood summarised the sources of error and put forward a few ideas for presentation. However, such publication presents the producers with a challenge.

Two of the three papers of this session sought to demonstrate different ways of providing the user with an understanding of the error associated with data; the third provided a means of communicating variances for the interpretation of changes and turning points in repeated surveys. The paper by John wood, presented by Markus Šova, discussed the various sources of error but then concluded that the analyst should consider a trade off between the reduction in variance of estimates and the cost of achieving the reduction. Avi Singh, co-author of the paper with M. Westlake, and M. Feder, provided a new aspect of the coefficient of variation that sought to deal with problems associated with near zero estimates, introducing a new measure – the discrimination coefficient of variation.

The majority of statistical data is presented in the form of tables. Farquhar & Farquhar (1891) noted that ‘Extracting information from a table is like extracting sunlight from a cucumber’. The extraction of information may be considered almost an impossibility, but why? Basically eighty per cent of the general population are number blind: so presenting them with a data table is not providing digestible information. For the press reporters (Worcester, 2004) the situation is worse as ninety-five per cent are number blind, as a greater proportion of them have studied arts than the general population. Whilst presenting his paper, Avi Singh noted that the coefficient of variation was ‘easy to understand’: I should like to be a fly on the wall of the Prime Minister’s Office when he tries to explain this concept! The concept may be understandable by statisticians but is that the user we are trying to address?

Before addressing that question, let me look at aspects of the additional information that statisticians wish to disseminate with their estimates: I shall group sampling (and other) errors, variance estimates, confidence intervals, coefficients of variation and the new statistics proposed at the symposium, the discrimination coefficient of variation. The main purpose of this additional information is to specify accuracy: as good statisticians, we want to indicate how good (or bad!) our estimates are. The expert user may also want to consider this additional information to assess the sensitivity of any decision making that they are undertaking, and likely impact of the decision if the estimate is at the upper or lower bound of the confidence interval. Nevertheless the politician with responsibility for major spending decision, when given information that the estimate produced by statisticians has a range of possibility between 30 and 36, will respond ‘So, you mean 33’.

[1] Edward Swires-Hennessy Data Unit Wales, Columbus Walk, Cardiff, Wales, United Kingdom, CF104BY: (Ed.Swires-Hennessy@dataunitwales.gov.uk)

Significant problems confront the disseminator of such additional information. The user does not understand the intent nor, most often, the concepts. Adding more figures to the estimates simply confuses the user as few of our users are educated in the concepts behind the additional information. Providing an estimate with associated range can confuse the decision process: providing more than one estimate, each with an associated range totally flounders the general user. Further, statisticians themselves are not used to writing about the statistics and including statements on the reliability of each estimate.

Some of the problems of dissemination of error measurement arise because the concepts are not well understood by non-mathematicians. Many estimates are presented without appropriate rounding, with statisticians acting like accountants and presuming that more precision (spurious or not) is better: this just adds to the confusion. Even when diagrammatic representations of the errors are provided, the user struggles to understand the intent and interpret the information before them. It would be quite inappropriate to provide error data to an eleven year-old child who only requests one or two numbers. This leads to the question of how much additional information should be provided.

Some statisticians – and a few expert users – advocate that the additional information should be provided with all estimates. Some advocate no additional information should be provided in publications; some would relegate this information to a methodology chapter; some would not provide the information unless asked. So how do we decide what is right? Two ways of considering this problem are apparent: either starting from an analysis of the outputs or from an analysis of the users. One of the standard analyses of products disaggregates the products into three groups:

- those providing basic information which is free to all, such as 'Canada in figures', general information on the web sites etc.
- those standardised products and services which are charged at market prices, such as a statistical yearbook, social trends publications etc. and
- user-specified information such as special tabulation from the vast data resources of a national statistical institute.

One suggestion would be to provide no additional information with products in the first group; a specific chapter of additional information for products in the second group and a great deal of additional information for products in the third group – almost with each estimate produced.

It is necessary also to ask who wants the additional information. The example of the eleven year-old earlier would be a case where no additional information is wanted and thus not provided. However, an economist who is trying to model the economy may require specific additional information with all the data supplied. The difficulty for this type of user is that, often, they do not necessarily know what to do with all of the additional information!

So what should the methodologists and practitioners do? As the majority of our users do not want and do not understand the additional information we have, we should continue our research into survey methods, error sources and process impacts to reduce the many forms of error and produce better estimates. That way the need for additional information will be minimised.

The cry of the majority of users is: give me **the** number!

REFERENCES

Farquhar and Farquhar (1891), in *Economic and Industrial Delusions: A Discourse for the Case of Protection*, New York: Putnam.

Worcester, Sir R (2004), Comments following a lecture at MORI Social Research Institute, Sept 2004.