

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 4 of the provisional agenda

METRICS FOR PREDICTING THE QUALITY OF EDITING AND IMPUTATION

Submitted by Statistics Sweden¹

¹ Prepared by Svein Nordbotten.

Abstract

The quality of statistical information will always be uncertain. To serve the users, quality declarations are desirable. The aim of the editing process is to improve the quality in statistical products. In this paper, metrics for predicting the quality of statistical products are proposed. Empirical test will be demonstrated.

Keywords: Quality prediction, quality declaration, data editing , data imputation.

I. QUALITY OF STATISTICS

1. A statistical product can be characterised by a number of properties such as its relevance for a set of applications, details, timeliness, resources used, quality, etc. In this paper the focus is on the quality aspect. Quality has been emphasised as an important goal to aim at for producers of statistics and has also been considered as an overall measure of statistical efficiency. The quality concept is, however, neither unambiguously defined nor easy to measure. The view adopted in this paper is that quality of a statistical estimate is determined by the estimate's deviation from the target value ideally to be measured. Since the target value is usually unknown, quality can only be estimated. This estimate will itself be subjected to uncertainty, and therefore the quality must be expressed in probabilistic terms.

2. Preparation of statistical products can be considered as a chain of processes each of which contributes to the total quality of the final product. The chain of processes may for example be data collection, preprocessing, editing and imputation, storing data in a statistical data base, retrieval of required data and calculation of the wanted statistics. One of the processes, the editing and imputation, has as its objective to detect and correct errors in the data in order to promote the quality of the statistical products. A number of different indicators for measuring the effect of this process have been proposed [Granquist 1997].

3. The purpose of this paper is to outline metrics predicting the quality of statistical estimates computed from data edited by means of different methods. These metrics aim at measurements for guiding the users about the sufficiency of published statistical information for their applications. The metrics do not substitute, but supplement monitoring of the editing process, the purpose of which is to give the designer of statistical surveys information about de facto changes in data due to alternative editing methods [Engström 1996, Jong 1996 Stefanowics 1997, Thomas 1996].

II. GENERAL PROBLEM

4. Consider a population of N statistical units. Without loss of generality, it can be assumed that each unit is characterised by a single y -variable value and a set of K x -variable values. The y -values are unknown target values to be observed.. The x -values are background data available from administrative registers, etc. The individual target values are denoted as

y_i , $I = 1..N$, and the values in the x -set as x_{ki} , $k = 1..K$ and $I = 1..N$.

5. The measurement of the target total, $Y = \sum_{i=1}^N y_i$, is the goal for a statistical producer. To achieve this goal, the producer collects data from each unit, but the observed values may be infected by different kinds of errors during observation and processing. The producer tries to counteract the attack of errors with the editing process. The individual data resulting from the editing process can be summarised by the expression $y_i' = y_i + e_i$ where e_i denote remaining errors.

6. It shall be assumed that $e_i, i=1..N$, are independent random variables, all with a common mean m_e and variance s_e . Without significant loss of generality, it is assumed here that $m_e=0$. The estimate $Y' = \sum y_i'$ can therefore be different from the target total value Y . The difference may vary if the same editing process was repeated. The deviation $|Y'-Y|$ is therefore an obvious candidate as a quality measure for the estimate Y' .

7. Since the quality of Y' does not manifest itself before the statistical information is used, the producer of statistics would want to issue a quality declaration for the statistical total Y' to guide the users. The declaration may be expressed by a limit D which the deviation $|Y'-Y|$ only will exceed with a specified, low probability Pr . A low limit D indicates an estimate of high quality, and vice versa. Preparation of a reliable quality declaration D for a given Pr , is a significant task for the statistical producer to satisfy the requirements of the users of his statistical products.

8. One simple approach would be to subject a small sample of the survey observations to an intensive editing using the background values. It could be assumed that this intensive editing would result in correct values, the error variance s_e^2 estimated from the sample, and this estimate used to predict the sampling error of the estimate Y' ,

$$S = s_e * \sqrt{N}$$

9. This measure expresses the uncertainty in the estimate Y' because of response and processing errors in the individual observations. It should not be confused with estimating the total and its sampling error from the sample which express the error due to the fact that only a sample of the population has been observed.

10. If the number of elements in the estimate Y' is large enough which is usually the case in statistical surveys, the distribution of the deviation $|Y' - Y|$ will be approximately normal according to the Central Limit Theorem, and a statement, for example

$$Pr(|Y'-Y| > D = 2*S) = 0.05,$$

can be expressed. This expression says the probability is only 0.05 that the deviation $|Y'-Y|$ is larger than $D=2*S$.

11. In the next section, this expression will be used and the confidence limit D adopted as an inverse quality metric in the analysis of some selected survey cases. The cases discussed are all based on the assumption that complete surveys are carried out. The analysis can however, be extended to take into account also errors due to the sample survey design. Since

this type of errors in the statistical estimates cannot, be corrected by editing , they will not be considered in the present context .

III. QUALITY METRICS

Case 1: An estimate based on unedited data

12. As a starting point, assume that data on the variable are collected for all units and that no background variables are available. By means of expert editors, errors in the individual observation values y_i can be detected and corrected. In the current case, complete editing is prohibitive because of few available expert editors, as well as the costs and the time required. However, editing of a sample of n units can be afforded. For the sample, the editing provides, in addition to the observed y_i -value , the edited value y_i' for each unit. The difference is denoted e_i and assumed to be a random variable with characteristics as specified in the previous section.

13. On the basis of the pair of values, y_i' and y_i , for each sample unit, estimates of the standard deviation s_e for the error e is computed. As an estimator for Y , the following function is used taking advantage of the edited values for the n units in the sample

$$Y' = \sum_{i=1}^n y_i + \sum_{n+1}^N y_i'$$

The sampling error for the estimate Y' will be

$$S' = s_e \sqrt{(N-n)}$$

The sampling error approaches 0 when n , the number of observations being edited approaches N , a complete edit of all units in the population observed.

The sampling error is used in the probability expression

$$Pr(|Y' - Y| > D = 2 * S') = 0.05.$$

to derive the metric D indicating the limit for the deviation $|Y' - Y|$ we will expect in 95% of the time.

Case 2: An estimate based on simple computer edited data

14. Assume that all in a population units are observed with respect to the variable and that data for a single background variable, x , are also available. With the help of editing experts, a ratio control of the observations has been designed based on knowledge from for example a previous similar survey.

15. The experts pointed out that the ratio

$$y_i / x_i = r_i,$$

where y_i and x_i are the observed variable value and the background variable value respectively, should be within a specified range to be acceptable. Observations with a ratio outside this range are considered suspect and must be examined by human editing experts.

16. The ratiocontrol was implemented in a computer programme which is used for editing the observations. NI of the N observations are rejected as suspect and passed on for subsequent expert editing. The remaining $(N-NI)$ observations are accepted by the computer editing control. From the theory of testing hypotheses, we know that two types of errors can be made in such a situation.

17. *Type I error* is rejection of a correct observation. *Type I errors* have no effects on the quality of the results in the present editing process since the suspected observations are passed on to the expert editors by whom they are identified as correct. *Type II errors*, acceptance of incorrect observations because acceptable r -values, may still, however, be among the $(N-NI)$ accepted observations and affect the quality.

18. Two sets of observations are now available for estimating Y' :

- (1) NI observations rejected by computer control and then edited by expert editors, and
- (2) $(N-NI)$ observations which were accepted by the computer control.

Based on these,

$$Y' = \sum_1^{NI} y_i + \sum_{-NI+1}^N y_j'$$

is an estimator of Y . Note that the first term will according to the design of the editing process be free from errors.

The estimator can be rewritten as

$$Y' = \sum^N y_i + \sum_{NI+1}^N e_j$$

where $e_j = y_j' - y_j$. If e_j is different from 0, an error in observation j has been accepted by the computer control. The quality of the estimate is obviously depending on the characteristics of the variable e . By drawing a small sample from the $(N-NI)$ observations and let expert editors review the sampled observations, estimates of the mean m_e and the standard deviation s_e for the variable e can be obtained. If the estimate of the mean is significantly different from 0, the editing process probably results in a biased estimate.

As a sampling error of the above estimate we can use:

$$S = s_e * \sqrt{(N-NI)}$$

In the same way as for the previous situation, a probability statement of the type $Pr(|Y'-Y| \leq D=2*S)=0.95$ permits the derivation of the quality metric

$$D=2*S$$

19. We would probably aim at designing the editing in such a way that $(N-NI)$ would represent a large fraction of the observations. On the other hand, we would also expect that the Type 2 errors were numerically small and that s_e therefore also would be small.

Case 3: An estimate based on editing with several background variables

20. Let us assume that related to each observation y' , a set of K background variables is also available. A relationship or pattern among the variables can be used for the classification of the N observations in two groups, acceptable and suspect observations based on a computerised classification model:

$\Psi_i = 0$ if y_i is not compatible with the background variables

$\Psi_i = FI(y'_i, x_{1i} \dots x_{Ki})$ where:

$\Psi_i = 1$ if y_i is compatible with the background variables

where FI denotes a mapping function from the set of patterns which may occur to the set of the two classes. The previous case was a very simple example of such a mapping function.

The results of the computer classification will be a group of NI suspect observations and another group of $(N-NI)$ accepted observations.

21. Assume that a second computerised model, $F2$, is implemented and is used to make imputations of the correct values for the set of NI suspect observations. The imputations are based on a function using the observed values and the background values as arguments

$$y_i'' = F2(y'_i, x_{1i} \dots x_{Ki})$$

The two models $F1$ and $F2$ may have been developed from a sample of edited observations, from a previous survey of the same kind or by means of expert knowledge.

An estimate of Y'' for the target total is obtained by the estimator

$$Y'' = \sum_{i=1}^{NI} y_i'' + \sum_{j=NI+1}^N y_j'$$

We denote the deviations of the individual values y'' and y' from the respective target values y as

$$y_i'' = y_i + r_i, \quad i=1 \dots NI,$$

and

$$y_j' = y_j + e_j, \quad j=(NI+1) \dots N.$$

It should be noted that we assume that the second model may introduce editing errors r_j

which we also assume are random with mean m_r and standard deviation s_r .

22. Let the expert editors examine the classifications and corrections performed by the two computer models in small samples of units from each of the two classes. The examination gives the target values of the y -variable for the sample units, and the means and standard deviations of the e - and r -variables can be estimated.

The estimate Y'' above can be rewritten as

$$Y'' = \sum_1^N y_i + \sum_1^{NI} r_i + \sum_{(NI+1)}^N e_j.$$

Subject to the assumptions made, the sampling error for Y'' is then

$$S'' = \sqrt{NI * s_r^2 + (N-NI) * s_e^2}.$$

and the confidence limit D for the deviation of the estimate from the target total is computed as above.

Case 4: An estimate based on imputed data

23. Assume that the producer of statistics can justify to collect and edit y -observations for only a sample n from the population while the background variable values are available for all the N units. From a subsample n_i of n , an imputation model

$$y_i' = f(x_{1i} \dots x_{ki})$$

of the same type as in the previous case, is developed where $y_i' = y_i + e_i$. We assume that the e -variables are random with a common finite mean m_e and standard deviation s_e .

24. Traditionally, the y -values for the sample of n units and the background values would be used in some estimator to generate estimates for the target total Y with corresponding quality measures. The quality of the estimate Y' would usually reflect the fact that uncertainty is introduced because of the sampling of n observations.

25. Alternatively, the imputation model is used to get individual predictions of y_i' for all $(N-n)$ units not observed in the sample. The estimator used is now

$$Y' = \sum_1^n y_i + \sum_{n+1}^N y_j'$$

which rewritten becomes

$$Y' = \sum_1^n y_i + \sum_{n+1}^N e_j.$$

The uncertainty is now introduced because the imputed values are assumed to be affected by random errors. The sampling error of Y' is

$$S' = s_e * \sqrt{(N-n)}$$

where s_e is the standard deviation for e . An estimate of this parameter is obtained by using a small fraction of the sample n , the $(n-n_1)$ units not used to build the imputation model. For this subsample, the imputation model is used to get imputed values in addition to the observed and edited values, and s_e is computed as the root of

$$s_e^2 = \sum (y_i' - y_i)^2 / (n - n_1 - 1)$$

As for the previous cases, the confidence limit D for the deviation of $|Y' - Y|$ is derived as for the cases discussed above.

IV. EMPIRICAL ILLUSTRATION

26. The reliability of the quality metrics presented in section 3 of this paper could be tested in a number of ways. One approach is to use a set of edited data, referred to as the target values, from a real statistical survey as the starting point for the tests of the cases described in section III. Errors could be imposed on the edited data, the editing method to be evaluated applied to identify and corrected errors, estimates of totals computed, the quality of the estimates predicted and finally the results from the method evaluated. Because the exact target values are known as well as the imposed errors, both the quality of the results of the method studied can be evaluated as well as the reliability of the quality predictions.

27. Another approach would be to find a test material for which both the original observed data including errors and the edited target data, were preserved. Unfortunately, this type of data is seldom available. When available, the quality of the edited values can frequently be questioned.

28. The metrics introduced in this paper will be illustrated with tests on empirical data of the underlying assumptions and their performances.

References

Engström, P. : Monitoring the Editing Process. Work Session on Statistical Data Editing Voorburg, November 1996.

Granquist, L.(1997): An overview of methods of evaluating data editing procedures. In Statistical Data Editing, Vol.2, Methods and Techniques. Statistical Standards and Studies No.48. UN/ECE. NY and Geneva. pp.112-122.

Jong, W. A. M. de: Designing a Complete Edit Strategy; Combining Techniques. Statistics Netherlands, 1996.

Stefanovics, B.: Selected Issues of Data Editing. In Statistical Data Editing, Vol.2, Methods and Techniques. Statistical Standards and Studies No.48. UN/ECE. NY and Geneva pp.109-111..

Thomas, J. : Statistical Measurement and Monitoring of Data Editing and Imputation in the 2001 UK Census of Population. Work Session on Statistical Data Editing. Voorburg November 1996.