STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 18
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (ii): Measuring and evaluating data editing quality

WHEN MIGHT IT BE WORTH TRYING SELECTIVE DATA EDITING?

Contributed paper

Submitted by Statistics Sweden[1]

Abstract: This paper introduces a method to investigate the potential for applying selective data editing, i.e. "How much could possibly be gained by applying selective editing to the survey?" Raw and edited data of the survey are examined using EDA tools with SAS/Insight. The method is described step by step using data from a Swedish Industrial Survey. A study of the editing changes for one item should not take more than an hour or two for a person familiar with neither EDA tools nor SAS.

I. BACKGROUND

1. Selective editing has been shown to be an effective method for reducing editing without affecting quality. For example, Granquist and Kovar (1997) report results from a number of studies that all show savings of 50 per cent or more. The success of those methods depends on low hit rates of the original editing checks as well as on many small changes of different signs, resulting in negligible impacts on the estimates.

2. Generally, studies of selective editing methods have been carried out by applying the method to unedited data and evaluating the method using edited data from the same survey. However, such studies require time and resources and the outcome may be negative in the sense that it is hard to prove that the editing can be rationalized maintaining the quality for the particular survey. With that in mind, it may be difficult to sell selective editing methods to survey managers. A cheap and easy tool is needed to show survey managers that selective editing methods may be successfully applied to their surveys.

3. The method presented here was developed within a project at Statistics Sweden to update the Current Best Method (CBM) document on statistical data editing (Statistics Sweden 2002).

II. THE DATABASE

4. The Swedish Industrial Survey 1990 is a cut off complete enumeration survey. The database encompasses 5,540 observations including both edited and unedited data. We use the variable "turnover" in the example, where 603 changes are made.

III. INTRODUCTION

5. Figure 1 shows the frequencies of changes distributed by the size of the changes in the upper histogram, and the impact on estimates in the lower. The distribution is rather skewed. We choose log transformation of the variable to make it easier to see the distribution.

6. The smallest changes do not impact the estimates substantially. The aim is to roughly find the limit where changes are negligible and what could possibly be gained by using selective data editing. This

[1] Prepared by Gunnar Arvidson (gunnar.arvidson@scb.se) and Leopold Granquist (Leopold.granquist@scb.se).

is easily done in SAS/Insight. One only needs to change the scale (tick increment) and choose a value to test whether the calculated impact on estimates for that bar (subgroup) can be treated as negligible or not. The interaction facility in SAS/Insight makes it possible to change scale values for the two diagrams in one step.

7. The total is studied in this example. For the average the analysis would be the same, as it is the ratio of the total to a constant (the number of observations).

IV. GRAPHICAL ANALYSIS

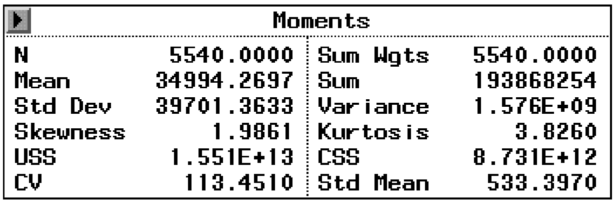
8. We calculate the following new variables in the database: the difference between the edited and the raw value (Diff); the absolute value of the Diff (DiffABS); and the impact of each change on the estimated total (Effect).

Effect is calculated as the ratio between the difference and the total:

Effect = DiffABS / Total * 10^9 + 0.5

where

> The total can be calculated from the analysis tool 'Distribution' and Sum. For sample surveys the parameter weight should be used.
>
> Distribution output:



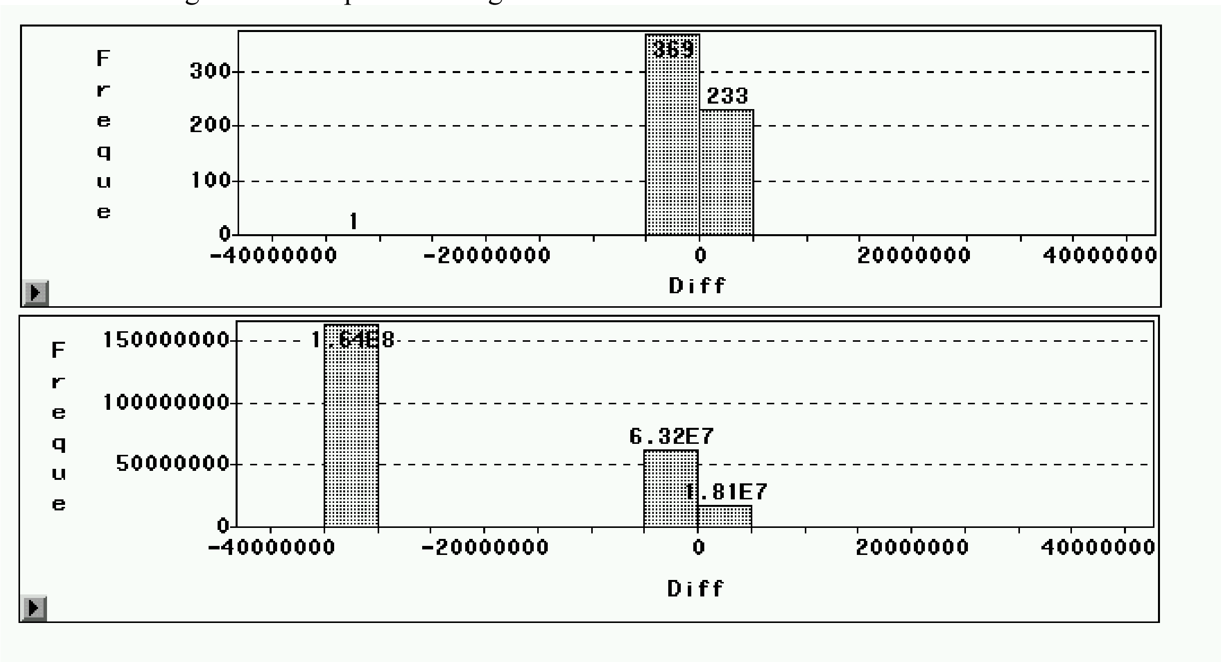
```
                     Moments
N          5540.0000  Sum Wgts   5540.0000
Mean      34994.2697  Sum        193868254
Std Dev   39701.3633  Variance   1.576E+09
Skewness      1.9861  Kurtosis      3.8260
USS        1.551E+13  CSS        8.731E+12
CV          113.4510  Std Mean    533.3970
```

> The scale factor 10^9 is used to obtain the impact values as integers for all units. This is necessary since only the integer part is used for calculations in the 'Freq' option in SAS/Insight.
>
> The term 0.5 is used to obtain correct rounding.

9. Figure 1 shows two histograms with the same x-axis, where the variable differences (Diff) is used. The upper histogram shows the number of changes, while the lower histogram shows the calculated impact, which is the sum of the Effect values (Σ Effect).

'Histogram' and the 'Freq' option on the variable Effect, are used to get the sum.

Figure 1.
The upper histogram: The number of observations by amount of changes
The lower histogram: The impact of changes on the total



Note that the charts only contain those 603 observations where the turnover variable is changed.

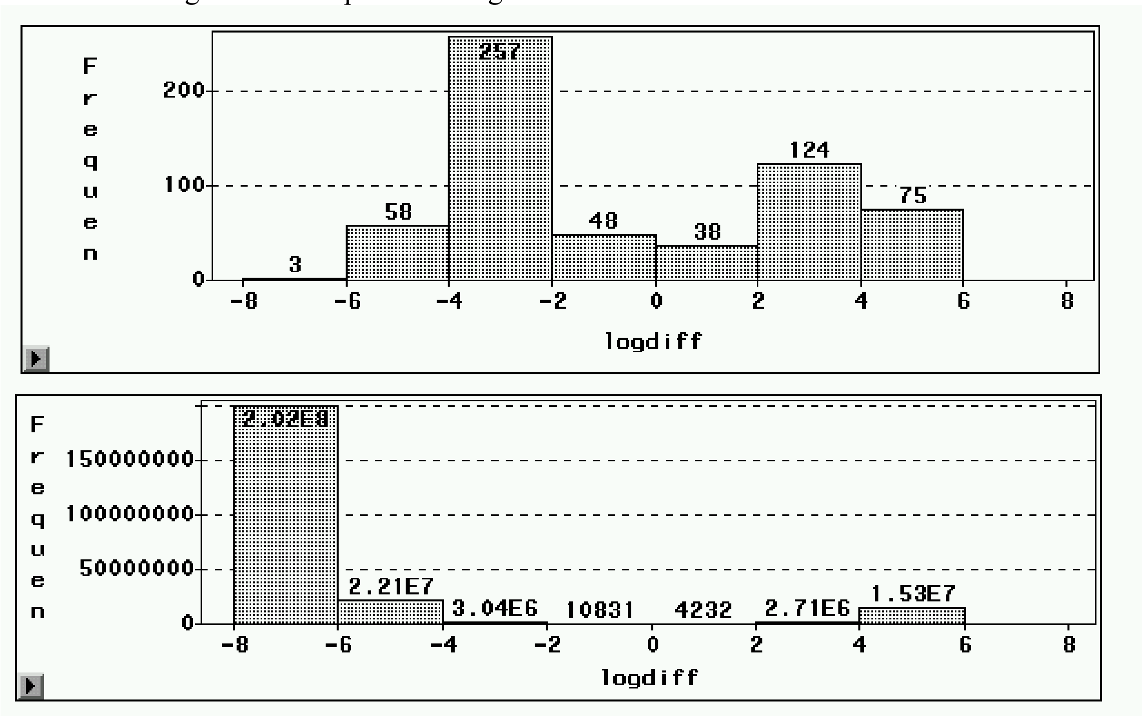
10. By changing the width of the bars, i.e. the tick increment, we can calculate the number of observations for any interval and at the same time obtain, from the height of the bar, the share of the changes to the total. The result should be interpreted as follows, including the scale factor, for the left bar: $(1.64E8 / 10^9) = (1.64 * 10^8 / 10^9) = 16.4\ \%$

11. The distribution of the changes is rather skewed with a few outliers. Therefore we produce the histograms using the log of the variable DiffABS to obtain a good overview of the changes. The variable transformation is performed in three steps:

i) calculate the absolute values (absdiff) of the differences (diff)
ii) take the log of the absolute value of differences (^{10}logabsdiff)
iii) assign the sign of the differences (diff) to the logged values (logdiff)

(The last variable can be calculated by the statement: if diff<0 then logdiff = 10logabsdiff * (–1))

Figure 2
The upper histogram: The number of observations by amount of changes (^{10}log-scale)
The lower histogram: The impact of changes on the total



12. There are 48 + 38 changes where the difference is less than ±100 (= ±2 on the log scale). For this group:
the gross impact on the total is (4 232 + 10 831) / 10^9, that is 0.015063 ‰
the net impact on the total is (4 232 – 10 831) / 10^9, that is –0.006599 ‰.

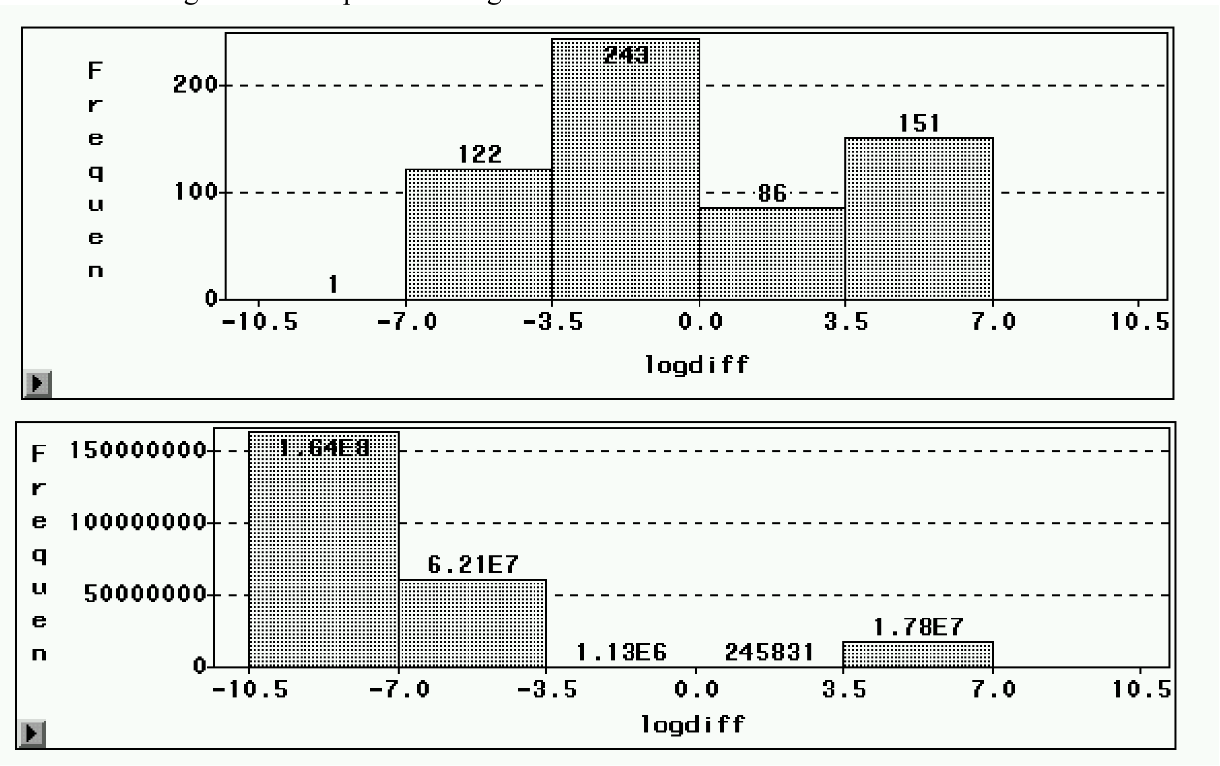
(The scale factor used in this example is 10^9.)

13. Will a survey manager consider such changes as negligible? In order to answer this question, it is necessary to have a method to assess whether the impact of small changes is negligible. One way is to look at the quality requirements and the influence of other sources of error such as undetected error of measurements, non-response bias and sampling error in sampling surveys. For this survey we consider a change in the estimate of less than 1 ‰ as negligible.

14. The aim is to calculate the number of changes that has a negligible impact on the estimate. We do this by choosing other values of the x-scale (using Ticks and Tick Increment) to see where the impact is below our limit of 1 ‰. This requires a few attempts to find the value on the x-axis that corresponds to the net impact of at most 1 ‰.

15. Our simple idea is to start around zero and extend the interval by changing the scale, until we have found the limit where the changes have a net impact less than 1 ‰. The ticks should be chosen to make the scale symmetric around zero and of equal length. Thus, for example, when trying a tick increment of 3, the first and last tick must be multiples of 3. In the example in figure 3 below with a tick increment of 3.5, the first tick is accordingly chosen as –10.5.

Figure 3
The upper histogram: The number of observations by amount of changes (^{10}log-scale)
The lower histogram: The impact of changes on the total



16. When using the value of 3.5 on the ^{10}log-scale, which is 3162 on a normal scale, the impact of the small changes on the estimate is:

gross effect = (245 831 + 1.13E6) / 10^9 = 1.38 ‰
net effect = | 245 831 – 1.13E6 | / 10^9 = 0.884 ‰

17. For the estimated total, the net effect is the relevant measure and the result is less than the limit chosen: 1 ‰. The corresponding number of units is 329, of which 243 changes are negative and 86 positive (upper histogram). There are 603 observations where the variable turnover is changed. Consequently 55% of these changes result in an impact below 1‰.

V. CONCLUDING REMARKS

18. An evaluation of an editing process should always include calculating hit rates of the edits. This should preferably be done automatically by the production system.

19. Here we have focused on the editing changes and advocate a graphical analysis using SAS/Insight. The initial aim for the method presented is to find out whether there are a substantial amount of changes, which together have a negligible impact on the estimates of important variables. If so there is a good chance that a selective editing method will reduce the editing work considerably without affecting the quality. Then a project of testing a selective editing method for the survey would likely be approved based on the results from such a study accompanied by data about the cost of handling flags of the editing system. Respondent costs could be reduced as well, in particular when respondents are recontacted to solve error flags.

20. Furthermore, using SAS/Insight or similar software for studies of editing changes can provide information of greater importance than just finding out whether selective editing methods could be successfully applied. For example, looking at editing changes in subgroups, which is necessary in our

method, can reveal that respondents in certain groups have more problems than others in furnishing acceptable data. Changing the questionnaires for such groups will be an appropriate step. Furthermore a high percentage of changes in some variables may indicate an underlying problem that respondents have severe difficulties in understanding what we are trying to study, or do not have the capacity to answer the questions. Revising the way of collecting data for those variables should be considered.

21. Just using graphics for a specific aim would certainly arouse curiosity about what is going on in the data and provoke other issues to be dealt with. The great advantage of using graphics for any task is that it forces you to see other phenomena in data that you previously were not aware of. Thus new and deeper insight into the survey could lead to new hypotheses and issues concerning the survey and serve as a basis for improvement. The described method would be an impetus for a wider use of graphics.

References:

Granquist, L and Kovar, J (1997): Editing of Survey Data: How much is enough? In L. Lyberg, P Biemer M. Collins, E. Leeuw, C. Dippo, N. Schwarz, D. Trewin (eds) Survey Measurement and Process Quality, Wiley, New York, 1997, pp. 415-436.

Statistics Sweden (2002): Guide till granskning, in Statistics Sweden series of Current Best Methods documents, March 2002 (in Swedish).

SAS/Insight, User’s Guide (1999), version 8. SAS Institute Inc.