STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 2
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic I: Management and evaluation of editing and imputation procedures

HOW TO MEASURE THE EFFECT OF DATA EDITING

Submitted by Statistics Denmark and Statistics Norway[1]

Invited paper

I. Introduction

1. The purpose of data editing is to correct errors before producing statistics, and essentially there is a distinction between two types of errors:

(i) Logical errors, which irrespective of size have to be corrected in order to have consistent data;
(ii) Outliers, of which the largest and most significant will have to be corrected at first.

2. In this paper we seek to describe how the effect of data editing can be measured. The two most important criteria for measuring the effect must be with respect to:

(i) Groups for which table entries are produced, i.e. groups defined by row and column variables;
(ii) Groups defined in other ways, e.g. by dividing the population into subpopulations in which we have a particular interest.

3. There is also a close connection between measuring the effect of data editing and the actual process of editing the units, whose errors are most important for the results.

4. In this paper we only consider the situation where the published statistics is either the number of units with a certain attribute (e.g. the number of employees in the age group 25-44 years) or a total (e.g. total income tax from persons in the age group 25-44 years). On the other hand, it is mainly this type of statistics or statistics derived from them (average, percentage, difference, growth rate), that are published by the National Statistical Institutes.

5. In the following paragraph we will present the notation used, and then we will present formula for measuring the effect of data editing.

II. Notation

6. The following notation will be used in this paper:

Z - statistics variable, observed value before data editing,
Y - statistics variable, observed value after data editing (i.e. value to be used in published statistics),
G - grouping variable, see above.

---

[1] Prepared by Birger Madsen, Statistics Denmark and Leiv Solheim, Statistics Norway.

Z and Y can either be values (e.g. salary, income, turnover) or a characteristic for enterprises or persons. In the latter example we can think of Z and Y as binary "characteristics", where the value 1 means having the attribute (e.g. being employed), whereas the value 0 means not having the attribute. For many such categorical (nominal) variables there are more than one category, but the point is that all such variables with more than two categories can be transformed into several binary variables, one for each category.

G is nothing but a division of the population into subpopulations, e.g. dividing enterprises according to economic activity or dividing persons according to municipality.

7. In the following we shall use z, y and g to indicate the specific values of Z, Y and G. Some further notation:

- i - index value, i.e. a numbering of the individuals in the population (e.g. persons, households or enterprises)
- w_i - weighting factor for unit no. i (in case of a sample survey)
- Σ_i - sum over all individuals in either sample (if sample survey) or the whole population (if total survey/register)
- $\Sigma_{i,g}$ - sum over groups (could be rows in the table)
- s_g - those units in the sample, who belong to group g
- U_g - those units in the population, who belong to group g

z_i, y_i - indicate values for unit i

In case of a total survey (e.g. register statistics) the weighting factor equals 1 and $s_g = U_g$ i.e. the subpopulation defined by group g.

8. The basic idea of this paper is a simple statistical model. For the population as a whole, the following model is the basis of the formula that follow:

$$Z_i = Y_i + \alpha + \varepsilon_i ,$$

where α is a parameter measuring bias and ε_i indicates deviation.

9. Looking at each subpopulation defined by G, the following model applies:

$$Z_i = Y_i + \alpha_g + \varepsilon_{i,g} ,$$

where α_g is a parameter measuring bias in each subpopulation and $\varepsilon_{i,g}$ indicates deviation in each subpopulation.

10. The published statistics can then be expressed in this way:

(1) $T_Y = \Sigma_i w_i y_i$ the number of individuals with this attribute (Y=1) in the population or a total for the whole population

(2) $T_{Y,g} = \Sigma_{i,g} w_i y_i$ the number of individuals with this attribute (Y=1) in group g or a total for group g

11. In the same way we will be able to calculate the totals before data editing:

(3) $T_Z = \Sigma_i w_i z_i$ the number of individuals with this attribute (Z=1) in the population or a total for the whole population, based on non edited data

(4) $T_{Z,g} = \Sigma_{i,g} w_i z_i$ the number of individuals with this attribute (Z=1) in group g or a total for group g, based on non edited data

III. Numeric example

12. A simple example will illustrate these concepts:

Variable	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
z_i	100	150	250	500	1 050
y_i	100	250	200	500	750
G	1	1	1	2	2
w_i	10	10	5	3	2
$w_i z_i$	1 000	1 500	1 250	1 500	2 100
$w_i y_i$	1 000	2 500	1 000	1 500	1 500

13. In this case we obtain

$T_Z = \Sigma_i w_i z_i = 1\,000 + 1\,500 + 1\,250 + 1\,500 + 2\,100 = 7\,350$
$T_Y = \Sigma_i w_i y_i = 1\,000 + 2\,500 + 1\,000 + 1\,500 + 1\,500 = 7\,500$
$\Sigma_i w_i = 10 + 10 + 5 + 3 + 2 = 30$

and correspondingly we obtain the following values within each group

G	1	2
$T_{Z,g}$	3 750	3 600
$T_{Y,g}$	4 500	3 000
$\Sigma_{i,g} w_i$	25	5

IV. Various measures of the effect of data editing

14. Accuracy is usually measured by the bias (i.e. the difference between true and observed value for a total or other published figure) and the standard deviation (i.e. how much do the observations vary on the average). In a similar way we can measure the effect of data editing. It will also appear to be natural to consider a measure, in which we calculate the effect of data editing on each single individual.

Bias in the whole population

15. This can be measured by an absolute (Bias) or relative (RelBias) measure

(5) $\text{Bias}(T_Y, T_Z) = T_Z - T_Y$

(6) $\text{RelBias}(T_Y, T_Z) = \text{Bias}(T_Z, T_Y)/T_Y = [T_Z - T_Y]/T_Y$

Bias in group g

16. This can also be measured by an absolute (Bias) or relative (RelBias) measure

(7) $\text{Bias}(T_{Y,g}, T_{Z,g}) = T_{Z,g} - T_{Y,g}$

(8) $\text{RelBias}(T_{Y,g}, T_{Z,g}) = \text{Bias}(T_{Z,g}, T_{Y,g})/T_{Y,g} = [T_{Z,g} - T_{Y,g}]/T_{Y,g}$

17. Example continued:

$\text{Bias}(T_Y, T_Z) = 7\,350 - 7\,500 = -150$
$\text{RelBias}(T_Y, T_Z) = -150/7\,500 = -0.02$

G	1	2
$\text{Bias}(T_{Y,g}, T_{Z,g})$	-750	600
$\text{RelBias}(T_{Y,g}, T_{Z,g})$	-0.167	0.200

V. Average effect of data editing

18. The average effect is defined (see below) as the deviation between non-edited and edited figures, corrected for the average difference between non-edited and edited figures.

19. First we define the weighted average of the z- and y- values with respect to the weighting factors

$\bar{z}_w = T_z/\Sigma_i w_i$
$\bar{y}_w = T_y/\Sigma_i w_i$

20. For group g we can in a similar fashion define a weighted group average of the z- and y- values

$\bar{z}_{w,g} = T_{z,g}/\Sigma_{i,g} w_i$
$\bar{y}_{w,g} = T_{y,g}/\Sigma_{i,g} w_i$

The whole population

21. We can now define the average effect of data editing as the deviation between non-edited and edited figures, corrected for the average difference between non-edited and edited figures:

(9) $\text{Dev}(T_Y, T_Z) = \{\Sigma_i w_i[z_i - \bar{z}_w - (y_i - \bar{y}_w)]^2/\Sigma_i w_i\}^{1/2}$

For group g

(10) $\text{Dev}_g(T_{Y,g}, T_{Z,g}) = \{\Sigma_{i,g} w_i[z_i - \bar{z}_{w,g} - (y_i - \bar{y}_{w,g})]^2/\Sigma_{i,g} w_i\}^{1/2}$

Relative deviation for the whole population

22. We can also define a measure for the relative average deviation, observe that this is relative to the statistics variable (i.e. after data editing):

(11) $\text{RelDev}(T_Y, T_Z) = \text{Dev}(T_Y, T_Z)/ T_Y$

Relative deviation for group g

(12) $\text{RelDev}_g(T_{Y,g}, T_{Z,g}) = \text{Dev}_g(T_{Y,g}, T_{Z,g})/ T_{Y,g}$

23. Example continued:

$\bar{z}_w$ = 7 350/30 = 245
$\bar{y}_w$ = 7 500/30 = 250

G	1	2
$\bar{z}_{w,g}$	150	720
$\bar{y}_{w,g}$	180	600

Variable	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
$z_i - \bar{z}_w - (y_i - \bar{y}_w)$	5	-95	55	5	305
$z_i - \bar{z}_{w,g} - (y_i - \bar{y}_{w,g})$	30	-70	80	-120	180
G	1	1	1	2	2
w	10	10	5	3	2
$w_i[z_i - \bar{z}_w - (y_i - \bar{y}_w)]^2$	250	90 250	15 125	75	186 050
$w_i[z_i - \bar{z}_{w,g} - (y_i - \bar{y}_{w,g})]^2$	9 000	49 000	32 000	43 200	64 800

$Dev(T_Y, T_Z) = \{[250 + 90\ 250 + 15\ 125 + 75 + 186\ 050]/30\}^{1/2} = 98.6$
$Dev_1(T_{Y,1}, T_{Z,1}) = \{[9\ 000 + 49\ 000 + 32\ 000]/25\}^{1/2} = 60.0$
$Dev_2(T_{Y,2}, T_{Z,2}) = \{[43\ 200 + 64\ 800]/5\}^{1/2} = 147.0$
$RelDev(T_Y, T_Z) = 98.6/7\ 500 = 0.013$, i.e. 1.3%
$RelDev_1(T_{Y,1}, T_{Z,1}) = 60.0/4\ 500 = 0.013$, i.e. 1.3%
$RelDev_2(T_{Y,2}, T_{Z,2}) = 147.0/3\ 000 = 0.049$, i.e. 4.9%

VI. Effect of data editing for each individual

24. For the single individual we can also define a measure for bias and deviation:

(13) $Bias_i = w_i(z_i - y_i)$

(14) $RelBias_i = Bias_i/T_Y$

(15) $Dev_i = \{w_i[z_i - \bar{z}_w - (y_i - \bar{y}_w)]^2\}^{1/2}$

(16) $RelDev_i = \{w_i[z_i - \bar{z}_w - (y_i - \bar{y}_w)]^2\}^{1/2}/T_Y$

(17) $Dev_{i,g} = \{w_i[z_i - \bar{z}_{w,g} - (y_i - \bar{y}_{w,g})]^2\}^{1/2}$

(18) $RelDev_{i,g} = \{w_i(z_i - \bar{z}_{w,g} - (y_i - \bar{y}_{w,g})]^2\}\}^{1/2}/T_{Y,g}$

25. These formula can also be used to construct all the preceding formula. Thus, formula 5 to 8 can be written as:

$Bias(T_Y, T_Z) = T_Z - T_Y = \Sigma_i\ Bias_i$
$RelBias(T_Y, T_Z) = Bias(T_Z, T_Y)/T_Y = [T_Z - T_Y]/T_Y = \Sigma_i\ RelBias_i$
$Bias(T_{Y,g}, T_{Z,g}) = T_{Z,g} - T_{Y,g} = \Sigma_{i,g}\ Bias_i$

$RelBias(T_{Y,g}, T_{Z,g}) = Bias(T_{Z,g}, T_{Y,g})/T_{Y,g} = [T_{Z,g} - T_{Y,g}]/T_{Y,g} = \Sigma_{i,g} RelBias_i$

26. Formula 9 to 12 can be written as:

$Dev^2(T_Y, T_Z) = \Sigma_i Dev_i^2/\Sigma_i w_i$
$Dev_g^2(T_{Y,g}, T_{Z,g}) = \Sigma_{i,g} Dev_{i,g}^2/\Sigma_{i,g} w_i$
$RelDev^2(T_Y, T_Z) = \Sigma_i RelDev_i^2/\Sigma_i w_i$
$RelDev_g^2(T_{Y,g}, T_{Z,g}) = \{\Sigma_{i,g} RelDev_{i,g}^2\}^{1/2}/\Sigma_{i,g} w_i$

27. Example continued

Indicator	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
$Bias_i$	0	-1 000	250	0	600
$RelBias_i$	0.0	-0.133	0.033	0.0	0.080
$RelBias_{i,g}$	0.0	-0.222	0.056	0.0	0.200
Dev_i^2	250	90 250	15 125	75	186 050
$RelDev_i^2$	0.0000	0.0016	0.0003	0.0000	0.0033
$Dev_{i,g}^2$	9 000	49 000	16 000	43 200	64 800
$RelDev_{i,g}^2$	0.0004	0.0024	0.0016	0.0048	0.0072