MEASURING THE ERROR OF EDITING THE QUESTIONNAIRES IN A CENSUS

By

Svein Nordbotten

Reprinted from the Journal of the American Statistical Association
June, 1955, Vol. 50, pp. 364-369
MEASURING THE ERROR OF EDITING THE QUESTIONNAIRES IN A CENSUS

SVEIN NORDBOTTEN
Central Bureau of Statistics, Oslo

This paper describes an attempt to measure the relative errors in totals and their influence on size distributions due to the questionnaire editing in a census. The experiment described was performed in connection with the 1953 Industrial Census in Norway. The results indicate that the errors are relatively small and their influence on distributions insignificant. The gains of a more thorough editing would probably be small and efforts might perhaps be better spent by improving the census procedure by other means.

INTRODUCTION

The results of a census are often published without any statement about the quality of the statistics. Both the professional statistician and the consumers of statistics would obviously profit by a greater knowledge of the accuracy of the statistics.

In a large census there may arise a great number of errors which may have an important influence on the results. Deming [2] gives an extensive list of possible errors. The professional statistician will be interested not only in a quantitative measure of the different error components, but also in their causes in order to be able to make more efficient census designs in the future.

The error components may be classified in two main groups:

(a) Random errors which arise because some element of random selection has been introduced in the census procedure.

(b) Non-random errors which are due to factors such as bad design, collection, editing, processing, etc. This group may be subdivided into subclasses.

Different approaches have been tried in evaluating total or relative non-random errors and their components in surveys and censuses. One which seems to be successful [3], is to check a small probability sample and compare census and sample results item-by-item.

In connection with the 1953 Industrial Census in Norway an experiment was done along these lines in order to try methods for evaluating quantitative measures for the editing errors and testing their influence on distributions of establishments.

364
DESCRIPTION OF THE EXPERIMENT

The purpose of the experiment presented in this paper was to investigate the relative editing error of statistics describing a mass of small manufacturing establishments, and the extent to which these errors influence certain distributions of the establishments.

The experiment was limited to Industrial Census questionnaires from six to seven thousand establishments, edited by six persons. Each day a five per cent systematic sample with random starts was drawn from the questionnaires edited that day by each person. The sample was then submitted to a second and very thorough control editing, independent of the original edit, with respect to employment, wage expenditures and value of production. The control editing was done by one specially instructed person.

The six persons performing the first editing were instructed not to contact the respondents for any supplementary information. When no answer was given or when an answer was supposed to be wrong, they were to make a rough estimate of the characteristic. In this way, 6620 questionnaires were edited. On the other hand, the one person performing the control editing was instructed to use all means to get correct information about employment, wage expenditures, and the value of production. For 103 of the 331 sample questionnaires, one or more requests for further information were sent.

These definitions were used:

(a) The individual error of a characteristic is the difference of the result obtained by the first and the second editing of a questionnaire.

(b) The relative error of a characteristic is the sum of all individual errors divided by the sum of the result obtained if all questionnaires were submitted to the control editing.

The following notation was used:

\[ n = \text{total number of establishments in the sample.} \]
\[ N = \text{total number of establishments in the population.} \]
\[ f = \text{sampling fraction.} \]
\[ L = \text{number of persons performing the editing.} \]
\[ n_h = \text{number of establishments in the sample from the } h\text{th person.} \]
\[ x_{hi} = \text{individual error on questionnaire } i \text{ in the sample from the } h\text{th person.} \]
\[ j/w = \text{value of the characteristic for questionnaire } i \text{ in the sample from the } h\text{th person after control editing.} \]
\[ r_h = \text{estimate of the relative error of the population edited by the } h\text{th person.} \]
\( r \) = estimate of the relative editing error.
\( s^2 \) = estimate of the variance of \( r \).
\( a_i(k) = 1 \) if the \( i \)th establishment in the sample is classified in class \( k \) in the first editing,
\( = 0 \) otherwise.
\( b_i(k) = 1 \) if the \( i \)th establishment in the sample is classified in class \( k \) in the control editing,
\( = 0 \) otherwise.
\( c_i(k) = 1 \) if the \( i \)th establishment in the sample is classified in class \( k \) in both editings,
\( = 0 \) otherwise.
\( M \) = number of classes.

The first aim was to estimate the relative errors of the characteristics. These errors were estimated by means of the ratio estimator

\[
(1) \quad r = \frac{\sum_{k} \sum_{i} x_{ki}}{\sum_{k} \sum_{i} y_{ki}}.
\]

This estimate is biased. On the other hand, it is consistent and the bias in this case is expected to be insignificant. When the number of questionnaires included in the population is as great as here and the sampling fraction is only five per cent, the estimate should have an approximate normal distribution if the sample is regarded as a stratified random sample. The variance of \( r \) was estimated by [1]

\[
(2) \quad s^2 = (1 - f) \sum_{k} \frac{\sum_{i} \left( \sum_{i} x_{ki}^2 + r^2 \sum_{i} y_{ki}^2 - 2r \sum_{i} x_{ki} y_{ki} \right)}{\left( \sum_{i} \sum_{i} y_{ki} \right)^2}.
\]

To test the hypothesis that the editing errors had no influence on the distributions of the establishments by employment, wage expenditures, and value of production, the following variable was formed

\[
(3) \quad K^2 = \sum_{k} \frac{(h_1(k) - h_2(k))^2}{\text{var}(h_1(k) - h_2(k))},
\]

where

\[
(4) \quad h_i(k) = \frac{N}{n} \sum_{i} a_i(k),
\]
\[ h(k) = \frac{N}{n} \sum_{i=1}^{n} b_i(k), \quad k = 1 \ldots M \]

and

\[
\text{var} [h_1(k) - h_2(k)] = \frac{N}{n} \left( \frac{N-n}{N-1} \right) L \left[ h_1(k) + h_2(k) - 2 \frac{N}{n} \sum_{i=1}^{n} c_i(k) \right].
\]

Assuming that the differences \( h_1(k) - h_2(k), k = 1 \ldots M, \) are normally distributed, the variable \( K^2 \) will be approximately chi-square distributed with \( M-1 \) degrees of freedom [4].

We chose a one per cent level of significance and defined a value \( x_0^2 \) such that the probability

\[ P(x_0^2 \geq K^2; M - 1) = 0.99. \]

The hypothesis that the editing error does not influence the distribution of the establishments should be rejected when \( K^2 > x_0^2 \).

THE RESULTS OF THE EXPERIMENT

The estimates of the relative editing errors and their standard deviations computed from formulas (1) and (2) are given in Table 1.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Relative editing error</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment</td>
<td>0.0273</td>
<td>0.0126</td>
</tr>
<tr>
<td>Wage expenditures</td>
<td>0.0248</td>
<td>0.0128</td>
</tr>
<tr>
<td>Value of production</td>
<td>0.0172</td>
<td>0.0114</td>
</tr>
</tbody>
</table>

The table indicates that the relative editing errors are rather small in spite of the imperfect editing. All of them seem to be positive. Even though there is no significant difference, it is also interesting to note that the relative error is lowest for the value of production. With a confidence coefficient of 0.99, the maximum relative error of employment is 6.5%, of wage expenditures 6.3%, and of the value of production 5.1%.

The relative distributions of the sampled establishments by employment, wage expenditures and value of production are given in Table 2.
The hypothesis that the distributions of the establishments by employment, wage expenditures, and value of production after the first editing are equal to the distributions by the same characteristics which would have been obtained if the control editing had been complete, was tested with the following results.

### TABLE 3

<table>
<thead>
<tr>
<th>Class</th>
<th>Employment</th>
<th>Wage expenditures</th>
<th>Value of prod.</th>
<th>$K^2$</th>
<th>$X^2$</th>
<th>Significant deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.5%</td>
<td>46.8%</td>
<td>27.1%</td>
<td>19.529</td>
<td>20.090</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>8.2%</td>
<td>16.6%</td>
<td>16.3%</td>
<td>4.363</td>
<td>18.475</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>34.4%</td>
<td>12.1%</td>
<td>12.6%</td>
<td>13.143</td>
<td>23.209</td>
<td>No</td>
</tr>
</tbody>
</table>

CONCLUSION

The results of the experiment described in this paper indicate that the effects of errors in editing questionnaires on statistics for a population of small establishments were small. The tests did not reject our hypothesis that there are no significant deviations between the distributions obtained by the first editing procedure and those which would have been obtained if all questionnaires were submitted to a control editing. The accuracy gained by a thorough and expensive editing of
the questionnaires is small and may perhaps be obtained cheaper by improving the census procedure by other means.

REFERENCES


