SVEIN NORDBOTTEN

ON ERRORS AND
OPTIMAL ALLOCATION IN
A CENSUS

Reprinted from the Skandinavisk Aktuarietidskrift 1957

UPPSALA 1958
ALMQVIST & WIKSELLS BOKTRYCKERI AB
On Errors and Optimal Allocation in a Census

By Svein Nordbotten (Oslo)

This paper gives an attempt to construct a model for non-random errors in a census. A simple model is established and the allocation problem in a census defined. In order to be able to handle this problem, it is transformed to a standard problem of linear programming and the principles of the simplex method for solution is described.

I. Introduction

One of the main problems in a census design is the optimal allocation of resources. Usually the solution of this problem is rather arbitrary because no “best” allocation is defined. The sampling statisticians have for a long time been working with a precise mathematical model on a special aspect of the allocation problem [6]. In terms of this model, the allocation problem is to find that allocation of a sample to strata which minimizes the cost of the investigation subject to a given condition about the sampling error. The census procedure will usually not include any random component but may lead to other types of errors. Morgenstern has called attention to the lack of a theory about these non-random errors in statistics [5].

The problem of optimal allocation has especially been studied in economic theory, and convenient allocation methods are developed. One of these is the method of linear programming [1]. The application of this method to the problem of allocation in stratified sampling is described in another article by the author [7].

In this paper a simple model for errors in census results will be presented, and the optimal allocation problem will be formulated as a linear programming problem by means of the error model.

1 – 573331 Skandinavisk Aktuarietidskrift 1957
II. A Model of Errors in Census Results

The aim of a census is to measure some characteristics. These characteristics should be defined by a set of operations such that their true values can be measured under ideal conditions [4]. To establish the ideal conditions will usually be very expensive and time-consuming. Therefore, censuses are usually conducted under less expensive conditions which may lead to errors in the results.

Consider a mass of \( N \) elements each of which is characterized by a true value \( x_i \) \((i = 1 \ldots N)\). These values can only be obtained under ideal census conditions. Practically, one obtains the following values:

\[
y_i = x_i + e_i \quad (i = 1 \ldots N).
\]

These values are called the individual measurements where \( e_i \) denotes the deviation between measurement and the true value. This deviation is called the individual error.

The \( e_i \) can be considered as the sum of error components arising in the \( K \) different stages or processes forming the census procedure:

\[
e_i = \sum_{k=1}^{K} e_{ik} \quad (i = 1 \ldots N).
\]

\( e_{ik} \) is called the individual process error of the \( k \)th process in the \( i \)th element. Some of the \( K \) processes may be planning, enumeration, interviewing, coding, editing etc. The individual error can obviously be zero at the same time as two or more individual process errors have positive and negative values.

The aim of the measurement is to estimate the true total:

\[
X = \sum_{i=1}^{N} x_i.
\]

This total is estimated by:

\[
Y = \sum_{i=1}^{N} y_i.
\]

The estimate can be considered as the sum of \( X \) and the total error \( E \):

\[
Y = X + E.
\]
On the other hand, $E$ can be divided into components called total process errors:

$$E = \sum_{k=1}^{K} E_k,$$

where $E_k$ is the sum of all individual process errors arising from the $k$th process.

The relationships among the above types of errors is systematically described in the following table.

<table>
<thead>
<tr>
<th>Classification of measurement components.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement</td>
</tr>
<tr>
<td>$y_1 = \bar{x}<em>1 + \epsilon</em>{11} + \cdots + \epsilon_{1K} = \bar{x}_1 + \epsilon_1$</td>
</tr>
<tr>
<td>(Individual measurements)</td>
</tr>
<tr>
<td>$y_N = \bar{x}<em>N + \epsilon</em>{1N} + \cdots + \epsilon_{NK} = \bar{x}_N + \epsilon_N$</td>
</tr>
<tr>
<td>(Estimate)</td>
</tr>
</tbody>
</table>

The consumers of statistics are mainly interested in the accuracy of the estimate $Y$, i.e. in the total error $E$. If possible, a statistical agency ought to give some measure of $E$ in order to guide the consumers [3]. The professional statisticians are in addition also interested in the values of the process errors and how to reduce them in order to improve the census designs in the future. A series of studies in process errors has been conducted in recent years [3, 8].

Assume that the conditions under which a census is conducted can be described by the intensities of the $K$ different statistical processes constituting the census procedure. The intensity of the $k$th process is measured by an index $t_k$ denoting the input of resources per element in this particular process. The index may for instance be number of man-hours used in processing one element at this stage of the census. This does not imply that manpower necessarily is the only input in this process. We will only require that if other kinds of input are used there must be unambiguous relations between the index input and each of the other kinds of input.
$E$ is regarded as a random variable with a conditional probability function $f(E, t_1 \ldots t_k)$. A value $E_1$ is defined by

$$
\int_{-E_1}^{E_1} f(E, t_1 \ldots t_k) \cdot dE = \varepsilon,
$$

where $\varepsilon$ is a given constant. $E_1$ may be interpreted as the maximum value of the total error subject to a confidence coefficient $\varepsilon$ and given $t_k$ ($k = 1 \ldots K$). We assume that the relationship between $E_1$ and $t_1 \ldots t_k$ is approximately linear:

$$
E_1 = A + \sum_{k=1}^{K} B_k \cdot t_k,
$$

where $A > 0$ and $B_k \geq 0$ for $(k = 1 \ldots K)$.

The parameters of (8) can be estimated through a study of the individual process errors in elements from a random sample based on the methods of experimental designs.

### III. The Allocation Problem in Census

The values of the process intensities can obviously be chosen such that the process errors and/or the total error become zero. The use of the inputs is not, however, free and must therefore be associated with a certain cost.

We assume that the relationship between the cost and the inputs of a census can be expressed by the linear relation:

$$
c = c_0 + \sum_{k=1}^{K} c_k \cdot t_k \quad c_k \geq 0 \text{ for } (k = 0, 1 \ldots K)
$$

where $c_0$ denotes the overhead costs and $c_k$ is the cost connected with the use of each unit of $t_k$.

A preliminary definition of the optimal allocation in a census can now be given:

**Definition I.** The optimal allocation in a census is that set of values $t_{opt}^k$ ($k = 1 \ldots K$) which minimizes census cost subject to a given accuracy requirement expressed by

$$
A + \sum_{k=1}^{K} B_k \cdot t_{opt}^k \leq E_1
$$

(9)
and the non-negativity restrictions

\[ e_{bc}^{\pi} \geq 0 \quad (k = 1 \ldots K). \] (11)

In the case of only two processes, this definition can be geometrically demonstrated as in Figure 1. The shaded field indicates all allocations satisfying (10) and (11), which are called admissible allocations. The line \( OQ \) is the perpendicular on the isocosts, i.e. a line through all allocations associated with the same census cost. The distance from the point \( O \) to the isocost along the perpendicular indicates the cost of the census.

We are seeking for that isocost with at least one point representing an admissible allocation and the shortest distance from \( O \) along the perpendicular. In our example, \( RP \) is that isocost, \( P \) is the optimal
allocation and the distance $S$ represents the minimum cost of the census.

The above example is unrealistic in at least two respects. First, the optimal allocation implies no intensity in process no. 1. Second, in a census there are usually more than one characteristic which are measured and to which we have to pay attention.

It is more realistic to assume that the intensity $t_k$ of a process under no circumstances can be less than a given value denoted by $t_k$ ($k = 1 \ldots K$). We define a new set of intensity indices

$$t^*_k = t_k - t_k \quad (k = 1 \ldots K). \quad (12)$$

Let us assume that we will have to pay attention to the measurement of $N$ characteristics in the design of the census. By a pilot investigation the following relations are estimated:

$$E_i^k = A_i + \sum_{k=1}^{K} B_{ik} \cdot t_k \quad (i = 1 \ldots N), \quad (13)$$

where the subscript $i$ denotes the measurement of the $i$th characteristic.

We can now give the complete definition of the optimal allocation in a census:

**Definition II.** The optimal allocation is that set of values $t_k^{opt} = t^*_k + t_k$ ($k = 1 \ldots K$) which minimizes

$$c = c_0 + \sum_{k=1}^{K} c_k \cdot t_k^{opt} + \sum_{k=1}^{K} c_k \cdot t_k \quad (14)$$

subject to a given set of accuracy requirement expressed by

$$A_i + \sum_{k=1}^{K} B_{ik} \cdot t_k^{opt} + \sum_{k=1}^{K} B_{ik} \cdot t_k \leq E_i^k \quad (i = 1 \ldots N) \quad (15)$$

and the non-negativity restrictions:

$$t_k^{opt} \geq 0 \quad (k = 1 \ldots K). \quad (16)$$

This definition is in the case of two processes and three characteristics illustrated in Figure 2. The shaded field includes all admissible allocations. $RT$ is the isocost on which the optimal allocation
Fig. 2.

(1): $E_1 = A_1 + \sum B_{1k} \cdot t_k$

(2): $E_2 = A_2 + \sum B_{2k} \cdot t_k$

(3): $E_3 = A_3 + \sum B_{3k} \cdot t_k$

$P$ is located and the distance $S$ on the perpendicular indicates the minimum cost of the census.

A geometrical solution of the allocation problem may be rather difficult as soon as the number of processes exceeds two. If the following transformations are performed

$$c^i = c_i = c_q - \sum_{k=2}^{K} c_k \cdot I_k, \quad (17)$$

$$z_i = E^i_1 - A_1 - \sum_{k=1}^{K} B_{nk} \cdot I_k \quad (i = 1 \ldots N), \quad (18)$$
the allocation problem can, however, be formulated as a standard linear programming problem, i.e.:

Minimize: \[ c^1 = \sum_{k=1}^{K} c_k \cdot t_k^i \] (20)

subject to

\[ \sum_{k=1}^{K} B_{ik} \cdot t_k^i + t_{k+i} = z_i \quad (i = 1 \ldots N), \] (21)

\[ t_k^i \geq 0 \quad (k = 1 \ldots K_N). \] (22)

The problem now comprises the intensity indices \( t_k^i \) \((k = 1 \ldots K)\) and the slack variables \( t_{k+i} \) \((i = 1 \ldots N)\). An optimal allocation will at most be composed of \( N \) positive values of these \( N + K \) variables. Let \( t^j \) \((j = 1 \ldots N)\) denote \( N \) positive values in an admissible allocation and let \( t_h \) \((h = N + 1 \ldots N + K)\) be the other \( K \) variables which are all zero.

t^j and \( c^1 \) can be expressed by

\[ t^j = \sum_{i=1}^{N} B_{ij}^t \cdot z_i - \sum_{h=N+1}^{K+N} B_{ih}^t \cdot t_h^i \quad (j = 1 \ldots N), \] (23)

\[ c^1 = \sum_{j=1}^{N} c_j \cdot t^j + \sum_{h=N+1}^{K+N} c_h \cdot t_h^i, \] (24)

where \( B_{ij}^t \) denote the coefficients in the system (21) solved with respect to \( t^j \) \((j = 1 \ldots N)\). Substituting \( t^j \) in (24) we get

\[ c^1 = \sum_{j=1}^{N} \sum_{i=1}^{N} c_j \cdot B_{ij}^t \cdot z_i + \sum_{h=N+1}^{K+N} \left( c_h - \sum_{j=1}^{N} B_{ih}^t \cdot c_j \right) \cdot t_h^i. \] (25)

If all

\[ \left( c_h - \sum_{j=1}^{N} c_j \cdot B_{ih}^t \right) > 0 \quad (h = N + 1 \ldots N + K), \] (26)

the census cost indicator \( c^1 \) could not be decreased by making any \( t_h^i \) positive. Subject to certain assumptions about the system (21), (26) will both be a necessary and a sufficient criterion on optimal allocation [2].
Several methods have been developed for finding the optimal allocation in case the criterion (26) is not satisfied for one or more \( t_i \). We shall here give a brief description of the "simplex method" [1].

Let \( t_p \) be one of the \( K \) variables which are zero, and let 
\[
(c_p - \sum_{i=1}^{N} c_i \cdot B^p_{i, p}) < 0.
\]
By increasing this intensity, the census cost can be reduced, but \( t_p \) cannot increase without decrease in \( t_i \) (\( j = 1 \ldots N \)), because (21). If \( t_0 \) is the first of the \( t^*_i \)'s to become zero, \( t_p \) can get a value
\[
t^*_p = \sum_{i=1}^{N} B^{0*}_{i, p} \cdot (c_i - B^i_{i, p}) / B^{0*}_{i, p}.
\]  
(27)

The new and lower census cost will now be indicated by
\[
c^{11} = c^1 + \left( c_p - \sum_{i=1}^{N} c_i \cdot B^r_{i, p} \right) \cdot B^{0*}_{i, p} / B^{0*}_{i, p}.
\]  
(28)

The new allocation is given by
\[
t^*_i = \sum_{i=1}^{N} B^{1*}_{i, p} \cdot (c_i - B^{0*}_{i, p}) / B^{0*}_{i, p} \cdot t^*_p, \quad (i = 1 \ldots N),
\]  
(29)

where \( t^*_p \) has been included among the \( N \) positive variables and \( t^*_0 \) has been excluded. The new coefficient \( B^{1*} \) can be found by
\[
B^{1*}_{i, j} = B^{0*}_{i, j} / B^{0*}_{i, p}, \quad (i = 1 \ldots N + K)
\]  
(30)

and the rest by
\[
B^{1*}_{i, j} = B^{1*}_{i, j} - B^{0*}_{i, j} \cdot B^{0*}_{i, p} \cdot (i = 1 \ldots N),
\]  
(31)

This procedure is repeated until the criterion (26) is satisfied.

**IV. Conclusion**

The theory of non-random errors presented in this paper is very simple. Some refinements can easily be done by assuming that the error components are multiplicative rather than additive, or/and assuming non-linear relationship between probable maximum error and process intensity. As a consequence, the allocation model may
perhaps be changed, and more advanced methods for solution must be applied.

The present model may, however, be useful as a preliminary instrument for the definition of the optimal allocation in a census.

References


(Received September 1956.)