Journal of Official Statistics, Vol. 11. No. 1, 1995, pp. 107–113

Controlling Invasion of Privacy in Surveys of Change Over Time – A Non-Technical Review[1]

Tore Dalenius[2]

1. Introduction

1.1. The notion of change over time

In many applications of survey methods to finite populations, the objective is to estimate change over time. In this paper, we will consider a finite population of N elementary units, to be referred to as "elements," at two points of time, t_1 and t_2. We will limit the discussion to the special case of a population which consists of the same N elements at both points of time. The mean of a variable at time t_1 is $\bar{X}$ and the mean of the same variable at time t_2 is $\bar{Y}$. In accordance with the non-technical aim of this review, change over time is defined simply as the difference between these two means, i.e., $D = \bar{Y} - \bar{X}$. An equivalent definition would be the mean of the N "element differences" $Y_j - X_j, j = 1 \ldots N$. By definition, D is the parameter of interest in a study of change over time. It is analytically meaningful, if the two variables reflect the same property of the elements.

1.2. Two sampling designs

We will consider two basic sampling designs for the estimation of D. Both designs call for probability sampling. For simplicity, we will assume simple random sampling without replacement.

The first design calls for selecting a sample of n elements from the population and collecting X-data from the selected elements. Now, $\bar{X}$ is estimated by $\bar{x}$. The sample is then returned to the population. At time t_2, a sample of k elements is selected. The Y-data are collected from the k elements. The mean $\bar{Y}$ is estimated by $\bar{y}$. Finally, D is estimated by $\bar{y} - \bar{x}$. The two samples are clearly (almost) independent.

The second design calls for selecting one sample of n elements, which are observed at time t_1 with respect to the X-variable, and at time t_2 with respect to the Y-variable. Again, D is estimated by $d = \bar{y} - \bar{x}$. The $\bar{y}$ and $\bar{x}$ estimates are dependent. We will refer to this second design as a regular longitudinal design.

1.3. A comparison of the two sampling designs

If applied to a population with positive correlation, R, between the two variables, the

[1] This paper makes extensive use of two research reports from the Department of Statistics, the Stockholm University, Stockholm, Sweden, which the author wrote in 1991.
[2] Division of Applied Mathematics, Brown University, Providence, RI, U.S.A.

regular longitudinal design yields a smaller variance than a design with two independent samples. More specificly, the variance of the regular longitudinal design is

$$\operatorname{Var} d = S_{\bar{y}}^2 + S_{\bar{x}}^2 - 2RS_{\bar{y}}S_{\bar{x}}$$

i.e., the sum of the variances of $\bar{y}$ and $\bar{x}$, respectively, and the corresponding covariance term. Clearly, if the correlation between X and Y and hence $\bar{y}$ and $\bar{x}$ is positive – a condition fulfilled in many social surveys – the variance of the regular longitudinal design is smaller than the variance of the first sampling design.

But the regular longitudinal design can generate greater concern among the elements for invasion of privacy than the sampling design with two independent samples. A prime source of concern is that the records with X-data are kept until time t_2.

In what follows, we will consider a modification of the regular longitudinal design to make it less prone to invasion of privacy. As a basis for considering modifications, we will elaborate on the regular longitudinal design in the next section.

2. The Regular Longitudinal Survey

2.1. *The focus*

In what follows, we will briefly review the characteristics of the regular longitudinal design which make it vulnerable to concerns about invasion of privacy over and above surveys for which the data collection is carried out at one point of time. It is these characteristics that any modifications must address.

2.2. *The population of interest*

There is a population of N elements

$$E(N) = E_1, \dots, E_j, \dots, E_N.$$

These elements may be individuals, families, etc.

With each population element, we associate two kinds of data, viz,

i. variables: X-data at the time t_1 and Y-data at the time t_2; and
ii. identifiers such as name and address: I-data.

While the data for an element may change over time (from time t_1 to time t_2), the I-data do not change over time. The data reflect some property of interest to the investigator.

2.3. *The data collection*

The investigator selects at time t_1 a sample of n elements and collects X-data. The sample is kept until time t_2, when Y-data are collected. With access to these data, linked records $[I,X,Y]$ are established and used for the estimation of D by $d = \bar{y} - \bar{x}$. Throughout this paper, we omit the indices when referring to the records.

2.4. *The specific invasion of privacy issue*

The public's concern about invasion of privacy is clearly related to the I-data and the fact that the investigator keeps the records concerning the X-data for the period t_1 to t_2.

For an illustration of the privacy issue in a specific case, reference is given to the debate reviewed in Dalenius (1988); the debate focused, to a large extent, on longitudinal surveys. It exercised a strong negative influence on the public's willingness to cooperate in interview surveys as documented in sizeable increases in non-response rates in a variety of surveys.

The modifications of the regular longitudinal design presented in Sections 3 and 4 address this specific aspect. In Section 5, finally, we present a conceivable alternative approach to data collection at two points of time.

3. A Minor Modification

3.1 *The nature of this modification*

Consider a random sample of n elements in the order they have been selected

$$E(n) = E_1, \ldots, E_j, \ldots, E_n.$$

This sample is divided into two parts, the r first

$$E(r) = E_1, \ldots, E_j, \ldots E_r$$

and the second part

$$E(m) = E_{r+1}, \ldots, E_{r+k}, \ldots, E_{r+m}.$$

For later use, we define

$$p = r/n$$

and

$$q = 1 - p = m/n.$$

3.2. *The data collection*

The r elements in $E(r)$ are subjected to the regular longitudinal design, which yields the r records $[I;X]$ and $[I;Y]$. Clearly, these records may be linked, yielding r records $[I;X;Y]$.

The m elements in $E(m)$ are used as follows:

i. At time t_1, X-data are collected and *deidentified* records $[-;\ X]$ created.
ii. At time t_2, Y-data are collected and *deidentified* records $[-;\ Y]$ created.

Clearly, the records from the time t_1 cannot be linked with the records from the time t_2. In other words, no records with X- and Y-data, as with the case of the regular design, can be created.

3.3 *Estimating change over time*

Change over time, D, is estimated by

$$d = \bar{y} - \bar{x}.$$

Denoting the sample means by $\bar{x}_r$, $\bar{x}_m$, $\bar{y}_r$, and $\bar{y}_m$ we may write

$$\bar{x} = p\bar{x}_r + q\bar{x}_m$$

with an analogous expression for $\bar{y}$.

3.4. *A potential research topic*

Clearly, the choice of p as defined above may be expected to influence the public's concern about invasion of privacy: the smaller the p, the less the concern.

But how is this relation between p and the public's concern to be formalized? The discussion of topics such as "sampling for time series" in some textbooks on sample survey methods and theory may prove helpful.

4. A Major Modification

4.1. *The nature of this modification*

The regular design as discussed in Section 2 corresponds to using $p = 1$ (and hence $q = 0$). The design as discussed in Section 3 corresponds to using $0 < p < 1$. The major modification to be discussed in this section corresponds to using $p = 0$ and hence $q = 1$.

4.2. *The data collection*

Data are collected in two rounds.

At time t_1 all m elements are observed with respect to the variable X. A set of m records is created, denoted by $[X; -]$, where "$-$" denotes no I-data recorded.

At time t_2, the same m elements are observed with respect to the variable Y. A set of m records is created, denoted by $[Y; -]$, where "$-$" denotes no I-data recorded.

Clearly, the records with X-data cannot be linked to the data with Y-data.

4.3. *Estimating change over time*

Using the two sets of data, estimates $\bar{x}$ and $\bar{y}$ are computed and D is estimated by

$$d = \bar{y} - \bar{x}.$$

This estimate has the variance

$$\operatorname{Var} d = S_{\bar{y}}^2 + S_{\bar{x}}^2 - 2RS_{\bar{y}}S_{\bar{x}}.$$

Clearly, the two first terms to the right may be estimated by standard procedures.

The estimation of R poses, however, a non-trivial problem. In what follows, we present three conceivable approaches.

4.3.1. Using prior information

If there are reasons to assume that R is (while positive) very small, this fact may be exploited. There may, for example, be reasons to assume that $0.10 < R < 0.20$, in which case estimating R by $R^* = 0.10$ would be a conservative estimate.

A similar approach may be used, if there are reasons to assume that R is very large; say $0.80 < R < 0.90$, in which case $R^* = .80$ would be a conservative estimate.

4.3.2. Using past data

In Hansen, Hurwitz, and Madow (1953), vol. 1, ch. 10, the use of "past data" for variance estimation is discussed. This approach may be applicable to the problem of estimating R.

4.3.3. Using group means

This approach exploits the following trick (D. Thorburn, personal communication, 1991). The m sample elements are grouped into G groups, with n_g elements in group $g, g = 1, \ldots, G$. To which group an element belongs must be known by the investigator, say, by placing this information on the records that could then be written as $[-; X; g]$ and $[-; Y; g]$. With access to such records, the means $\bar{y}_g$ and $\bar{x}_g$ would be computed. Then R is estimated by the correlation R^* between these group means. This amounts to using the group means as "regular" data.

4.4. Three potential research topics

The following three research topics are formulated as questions:

i. How close to 0 must R be in order to justify using $R^* = 0$ as the estimate of R?
ii. What kind of "past data" may profitably be used to estimate R?
iii. What is the best grouping of the elements in the sample and the best number of groups?

5. An Alternative Design

5.1. The nature of the alternative

The specific feature of the alternative design is that *no* data are collected at point t_1; all data are collected at time t_2. Parenthetically we mention that this feature is likely to have a favorable effect on the cost of a survey, compared with the cost of a longitudinal survey as discussed in Section 2.

5.2. The data collection

No data are collected at time t_1.

At time t_2, the investigator collects X-data in one of the following two ways:

i. The data are collected by retrospective interviews.
ii. The data are collected by the investigator from some register, for example, from the records in a previous census.

Clearly, the X-data thus collected may not be equal to the X-data of a longitudinal survey. Hence, we will denote the data collected as follows

$$X' = X + e$$

and the corresponding sample mean by $\bar{x}'$.

The Y-data are collected in the usual way.

Using the I-data and the variable data, records $[I; X'; Y]$, are now created.

5.3. *Estimating D*

Here, $D = \bar{Y} - \bar{X}$ is estimated by

$$d = \bar{y} - \bar{x}'.$$

5.4. *The invasion of privacy issue*

It appears reasonable to assume that the sample elements may view the alternative design as less threatening than the regular longitudinal design. We limit the discussion to two topics pertinent to the invasion of privacy issue.

If the design discussed in this section is to be considered an alternative to a regular longitudinal design, the X-data collected by retrospective interviews must be reasonably accurate, i.e., e as given by the formula below must be small

$$X' = X + e.$$

This suggests that the period from the point of time t_1 to t_2 must be "short." But what is a short period for various types of variables? And how does the cost of a survey enter the picture? A related research topic concerns the development of a measurement model which reflects the mechanism introducing the error, e, in the formula above. It would for sure be unrealistic to assume that e is unbiased.

6. References

6.1. *References cited in the text*

Dalenius, T. (1988). The Debate on Privacy and Surveys in Sweden. Chance, 43–47.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vol. 1, Ch. 10. New York: John Wiley and Sons.

6.2. *References not cited in the text*

Boruch, R.F. and Cecil, J.S. (1979). Assuring the Confidentiality of Social Research Data. University of Pennsylvania Press.

Boruch, R.F. and Pearson, R.W. (1988). Assessing the Quality of Longitudinal Surveys. Evaluation Review, 12, 3–59.

Dalenius, T. (1988). Controlling Invasion of Privacy in Surveys, Ch. 13. Stockholm: Statistics Sweden.

Young, C.H., Savola, K.L., and Phelps, E. (1991). Inventory of Longitudinal Studies in the Social Sciences. Sage Publications Inc., Thousands Oaks, CA.