Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

A model based justification of Kish's formula for design effects for weighting and clustering

by Siegfried Gabler, Sabine Haeder and Partha Lahiri

June 1999

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 1999
Vol. 25, No. 1, pp. 105-106
Statistics Canada, Catalogue No. 12-001-X

A model based justification of Kish's formula for design effects for weighting and clustering

Siegfried Gabler, Sabine Haeder and Partha Lahiri [1]

Abstract

In this short note, we demonstrate that the well-known formula for the design effect intuitively proposed by Kish has a model-based justification. The formula can be interpreted as a conservative value for the actual design effect.

Key Words: Cluster size; Intraclass correlation coefficient; Selection probabilities.

1. Introduction

We consider multistage, clustered, sample designs where each observation belongs to a weighting class. For example, the clusters are blocks which are selected proportional to the number of its households. Within each block the same number of households is selected with equal probabilities. A randomly chosen person of the household has to be interviewed. Then, the household sizes determine the weighting classes. Kish (1987) proposed the following formula for determining the design effect in order to incorporate the effects due to both weighting needed to counter unequal selection probabilities, and clustered selection:

$$\text{deff}_{\text{Kish}} = m \frac{\sum_{i=1}^{I} w_i^2 m_i}{\left(\sum_{i=1}^{I} w_i m_i\right)^2} [1 + (\bar{b} - 1)\rho],$$

where m_i and w_i denote the number of observations and the weight attached to the i^{th} weighting class $(i = 1, ..., I)$, $m = \sum_{i=1}^{I} m_i$, the total sample size, $\bar{b}$ is the average cluster size and ρ is the intraclass correlation coefficient. Kish's formula is very intuitive and novel, but he said that his "treatment may be incomplete and imperfect."

Kish's formula is now used by many survey samplers. In fact, the above formula will be used in the sample size determination in the European Social Surveys to be conducted by its member countries. The purpose of this note is to provide a model-based justification for using Kish's formula.

2. A model based justification of Kish's formula

Let m_{ic} be the number of observations in the c^{th} sampled cluster belonging to the I^{th} weighting class $(i = 1, ..., I; c = 1, ..., C)$. Then $m_i = \sum_{c=1}^{C} m_{ic}$, the number of observations in the I^{th} weighting class. Let $b_c = \sum_{i=1}^{I} m_{ic}$, the number of observations in the c^{th} cluster $(i = 1, ..., I; c = 1, ..., C)$ so that $\bar{b} = C^{-1} \sum_{c=1}^{C} b_c$. Let y_{cj} and w_{cj} be the observation and the weight for the j^{th} sampling unit in the c^{th} cluster $(c = 1, ..., C; j = 1, ..., b_c)$. The usual design-based estimator for the population mean is defined as

$$\bar{y}_w = \frac{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj} y_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}}.$$

To justify Kish's formula, we assume the following model:

$$\text{Var}(y_{cj}) = \sigma^2 \text{ for } c = 1, ..., C; j = 1, ..., b_c$$

$$\text{Cov}(y_{cj}, y_{c'j'}) = \begin{cases} \rho\sigma^2 \text{ if } c = c'; j \neq j' \\ 0 \qquad \text{otherwise.} \end{cases} \qquad (1)$$

The above model is appropriate to account for the cluster effect and was used earlier by others (see, *e.g.*, Skinner, Holt and Smith (1989). We shall then define design effect as $\text{deff} = \text{Var}_1(\bar{y}_w) / \text{Var}_2(\bar{y})$, where $\text{Var}_1(\bar{y}_w)$ is the variance of $\bar{y}_w$ under model (1) and $\text{Var}_2(\bar{y})$ is the variance of the overall sample mean $\bar{y}$, defined as $\sum_{c=1}^{C} \sum_{j=1}^{b_c} y_{cj}/m$, computed under the following model:

$$\text{Var}(y_{cj}) = \sigma^2 \text{ for } c = 1, ..., C; j = 1, ..., b_c$$

$$\text{Cov}(y_{cj}, y_{c'j'}) = 0 \text{ for all } (c, j) \neq (c', j'). \qquad (2)$$

Note that model (2) is appropriate under simple random sampling and provides the usual formula σ^2/m for $\text{Var}_2(\bar{y})$.

1. Siegfried Gabler and Sabine Haeder, ZUMA, B 2,1, D-68159 Mannheim; Partha Lahiri, University of Nebraska-Lincoln, NE 68588-0323.

Now, turning our attention to $\mathrm{Var}_1(\bar{y}_w)$, first note that

$$\mathrm{Var}_1\left(\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}y_{cj}\right)$$

$$=\sum_{c=1}^{C}\mathrm{Var}\left(\sum_{j=1}^{b_c} w_{cj}y_{cj}\right)$$

$$=\sum_{c=1}^{C}\left\{\sum_{j=1}^{b_c} w_{cj}^2\,\mathrm{Var}\,(y_{cj})+\sum_{j\neq j'}^{b_c} w_{cj}w_{cj'}\,\mathrm{Cov}(y_{cj},\,y_{cj'})\right\}$$

$$=\sigma^2\sum_{c=1}^{C}\left\{\sum_{j=1}^{b_c} w_{cj}^2+\rho\sum_{j\neq j'}^{b_c} w_{cj}w_{cj'}\right\}$$

$$=\sigma^2\left\{\sum_{i=1}^{I} w_i^2 m_i+\rho\sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_i m_{ic}\right)^2-\rho\sum_{i=1}^{I} w_i^2 m_i\right\} \quad (3)$$

since $\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2=\sum_{i=1}^{I} w_i^2 m_i$ and

$$\sum_{c=1}^{C}\sum_{j\neq j'}^{b_c} w_{cj}w_{cj'}=\sum_{c=1}^{C}\left\{\left(\sum_{j=1}^{b_c} w_{cj}\right)^2-\sum_{j=1}^{b_c} w_{cj}^2\right\}$$

$$=\sum_{c=1}^{C}\left\{\left(\sum_{i=1}^{I} w_i m_{ic}\right)^2-\sum_{i=1}^{I} w_i^2 m_{ic}\right\}$$

$$=\sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_i m_{ic}\right)^2-\sum_{i=1}^{I} w_i^2 m_i.$$

Noting that $\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}=\sum_{i=1}^{I} w_i m_i$, we have

$$\mathrm{Var}_1(\bar{y}_w)=\frac{\mathrm{Var}_1\left(\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}y_{cj}\right)}{\left(\sum_{i=1}^{I} m_i w_i\right)^2}$$

$$=\frac{\sigma^2\left\{\sum_{i=1}^{I} w_i^2 m_i+\rho\sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_i m_{ic}\right)^2-\rho\sum_{i=1}^{I} w_i^2 m_i\right\}}{\left(\sum_{i=1}^{I} w_i m_i\right)^2}$$

so that

$$\mathrm{deff}=m\frac{\sum_{i=1}^{I} w_i^2 m_i}{\left(\sum_{i=1}^{I} w_i m_i\right)^2}[1+(b^*-1)\rho], \quad (4)$$

where $b^*=\sum_{c=1}^{C}(\sum_{i=1}^{I} w_i m_{ic})^2/\sum_{i=1}^{I} w_i^2 m_i$.

Using the Cauchy-Schwarz inequality, we get

$$\left(\sum_{i=1}^{I} w_i m_{ic}\right)^2=b_c^2\left(\sum_{i=1}^{I} w_i\frac{m_{ic}}{b_c}\right)^2\le b_c^2\sum_{i=1}^{I} w_i^2\frac{m_{ic}}{b_c}$$

$$=b_c\sum_{i=1}^{I} w_i^2 m_{ic}$$

so that

$$b^*\le\frac{\sum_{c=1}^{C} b_c\sum_{i=1}^{I} w_i^2 m_{ic}}{\sum_{c=1}^{C}\sum_{i=1}^{I} w_i^2 m_{ic}}=\bar{b}_w,\ \text{say.} \quad (5)$$

Thus (4) and (5) yield

$$\mathrm{deff}\le m\frac{\sum_{i=1}^{I} w_i^2 m_i}{\left(\sum_{i=1}^{I} w_i m_i\right)^2}[1+(\bar{b}_w-1)\rho]. \quad (6)$$

Note that $\bar{b}_w$ can be interpreted as an average (weighted) cluster size. If $\bar{b}_w$ is equal to $\bar{b}$, *e.g.*, if all b_c are equal, the upper bound of deff is simply Kish's formula. Thus Kish's formula serves as a conservative value for the actual design effect.

Acknowledgements

The authors are thankful to the editor and the referees for their remarks which led to an improvement of the paper. The work was completed while the last author was a Guest Professor at ZUMA, the Center for Survey Research and Methodology, Mannheim, Germany.

References

Kish, L. (1987). Weighting in Deft^2. *The Survey Statistician*, June 1987.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. Chichester: Wiley.