Survey Methodology, December 2000
Vol. 26, No. 2, pp. 137-138
Statistics Canada, Catalogue No. 12-001

# An Approximate Design Effect for Unequal Weighting When Measurements May Correlate With Selection Probabilities

**Bruce D. Spencer** [1]

## Abstract

It is common practice to estimate the design effect due to weighting by 1 plus the relative variance of the weights in the sample. This formula has been justified when the selection probabilities are uncorrelated with the variable of interest. An approximation to the design effect is provided to accommodate the situation in which correlation is present.

Key Words: Weighting; Deff; Sampling variance; Complex samples.

## 1. Introduction

It is common practice to weight observations in an unequal probability sample by the reciprocals of selection probabilities. The rationale is that failure to use the weights will cause bias if the sampling weights correlate with the variable of interest. A drawback to weighting is an increase in sampling variance when the weights vary excessively in the sample. This increase may be quantified by the design effect. The design effect is the ratio of the variance of the statistic of interest under the design of interest to the variance of the statistic under simple random sampling with the same sample size (Kish 1965). Design effects are important both for approximating standard errors after the sample is in hand and for predicting standard errors ahead of time, which is critical for efficient design of samples.

Kish (1965, 1992) discussed an approximation for the design effect for weighted estimates from unequal probability samples: $1 + rvw$, with $rvw$ defined as the relative variance of the weights in the sample. Thus, if $w_i$ is the weight of unit $i$ in the sample and $\bar{w}$ is the sample mean, $rvw = n^{-1}\sum_{i=1}^{n} (w_i - \bar{w})^2 / \bar{w}^2$. Gabler, Haeder, and Lahiri (1999) used a superpopulation model to derive a design effect when clustering is present as well. Their formula, which agrees with design-based results in Kish (1965), reduces to $1 + rvw$ when there is zero intraclass correlation. The $1 + rvw$ approximation for the design effect is based on a model or design in which the weights are uncorrelated with the variable of interest (and hence an unweighted estimate would serve as well or better than the weighted estimate). Here we develop an approximation to the design effect under a model in which correlation may be present. In developing the approximation we do not assume that the population is sampled from a superpopulation. The accuracy of the approximation depends only on the characteristics of the sample design and the population of interest.

For simplicity, we will discuss single-stage unequal probability sampling with replacement. Heuristic extension of the results to sampling without replacement is indicated in section 4.

## 2. Regression Representation of Population and Sample Design

Let $y_i$ denote the measurement of interest, $P_i$ the (draw-by-draw) selection probability for a sample of size $n$, and $w_i = 1/(nP_i)$ the sampling weight for unit $i$ in a population of size $N$, $1 \le i \le N$. Observe that $\bar{P} = \sum_{i=1}^{N} P_i / N = N^{-1}$. Consider the least-squares population regression line

$$y_i = \alpha + \beta P_i + \varepsilon_i, \qquad (1)$$

with $\alpha = \bar{Y} - \beta / N, \beta = \sum_{i=1}^{N} (y_i - \bar{Y})(P_i - \bar{P}) / \sum_{i=1}^{N} (P_i - \bar{P})^2$, and $\bar{Y} = \sum_{i=1}^{N} y_i / N$. Denote the population variances of the $y$'s, the $\varepsilon$'s, the $\varepsilon^2$'s, and the $w$'s by $\sigma_y^2, \sigma_\varepsilon^2, \sigma_{\varepsilon^2}^2$, and $\sigma_w^2$, with, for example, $\sigma_y^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 / N$. Denote the population correlation between $y$ and $P$ by $\rho_{y,P}$, between $\varepsilon$ and $w$ by $\rho_{\varepsilon,w}$, and between $\varepsilon^2$ and $w$ by $\rho_{\varepsilon^2,w}$. It follows from the properties of least-squares, or equivalently from the definitions of $\alpha$ and $\beta$, that $\sum_{i=1}^{N} \varepsilon_i P_i = \sum_{i=1}^{N} \varepsilon_i / N = 0$ and $\sigma_\varepsilon^2 = (1 - \rho_{y,P}^2)\sigma_y^2$. If data are available, we can fit the regression representation (1) and estimate $\alpha$, $\beta$, $\sigma$, and $\rho$ by, say, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$, and $\hat{\rho}$.

Let $\hat{Y} = \sum_{i=1}^{n} w_i y_i$ denote the usual weighted estimator of the population total, $Y$. The variance of $\hat{Y}$ is well-known (Cochran 1977, 253) to be

$$V(\hat{Y}) = n^{-1} \sum_{i=1}^{N} P_i (y_i / P_i - Y)^2. \qquad (2)$$

Using the regression formulation (1), we may re-express the variance as

$$V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1 - \rho_{y,P}^2)\sigma_y^2 N\bar{W} + N\rho_{\varepsilon^2,w}\sigma_{\varepsilon^2}\sigma_w + 2\alpha N\rho_{\varepsilon,w}\sigma_\varepsilon\sigma_w, \qquad (3)$$

where $\bar{W} = \sum_{i=1}^{N} w_i / N$.

This expression does not rest on any assumptions about the fit of the regression model. (See section 5 for derivation).

If the regression model fits well enough so that $\rho_{\varepsilon^2,w}$ and $\rho_{\varepsilon,w}$ are zero, then the variance in (3) simplifies to $V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1 - \rho_{y,P}^2)\sigma_y^2 N\bar{W}$. If simple random sampling

[1] Bruce D. Spencer, Department of Statistics and Institute for Policy Research, 2006 Sheridan Road, Northwestern University, Evanston, IL 60208, U.S.A.

with replacement had been used, the variance would have been $n^{-1} N^2 \sigma_y^2$. Therefore, if $\rho_{\varepsilon^2, w}$ and $\rho_{\varepsilon, w}$ are negligible, the design effect is approximately

$$\text{deff} = (1 - \rho_{y, P}^2) n\bar{W} / N + (\alpha / \sigma_y)^2 (n\bar{W} / N - 1). \quad (4)$$

This approximation does not require that the residuals from the regression are negligible, and it can hold when $\sigma_\varepsilon$ is large. A referee has pointed out that the condition that $\rho_{\varepsilon^2, w}$ and $\rho_{\varepsilon, w}$ are negligible may seem unnatural in a model that regresses $y$ on $P$ rather than on $w \propto 1/P$. Note, however, that if we had not only zero correlation between $\varepsilon$ and $P$ but also independence, then we would have zero correlation between functions of $\varepsilon$ and functions of $P$, and so $\rho_{\varepsilon^2, w}$ and $\rho_{\varepsilon, w}$ would be zero as well.

## 3. Estimation of Design Effect

To estimate the design effect after the sample is in hand, we may use $1 + rvw$ to estimate $n\bar{W} / N$. To understand the rationale for this, note first that

$$1 + rvw = \frac{n^{-1} \sum_{i=1}^{n} w_i^2}{\bar{w}^2}. \quad (5)$$

The expectation of the numerator is $N\bar{W} / n$. The expectation of $\bar{w}$ is $N / n$, and so the denominator of (5) may be taken as an estimator of $(N / n)^2$. Dividing the expectation of the numerator by $(N / n)^2$, we obtain $n\bar{W} / N$. Thus the design effect may be estimated from the sample by

$$(1 - \hat{\rho}_{y, P}^2)(1 + rvw) + (\hat{\alpha} / \hat{\sigma}_y)^2 (rvw). \quad (6)$$

As a special case, note that if we set $\hat{\rho}_{y, P} = 0$, the case of "haphazard weighting" (Kish 1992), then the estimate of the design effect simplifies to

$$1 + rvw + rvw(\hat{\alpha}^2 / \hat{\sigma}_y^2). \quad (7)$$

This estimate is close to Kish's approximation when $\hat{\alpha} / \hat{\sigma}_y$ is near zero.

## 4. Sampling Without Replacement

To derive the exact design effect for sampling without replacement would be more complex, as it would require consideration of joint selection probabilities for pairs of units. A heuristic extension of the results is easy, however. Recall that the ratio of the variance of a sample mean under simple random sampling without replacement to the variance under with-replacement sampling is approximately $(1 - n / N)$.

The results we have derived for the design effect will apply to single-stage unequal probability samples of $n$ units without replacement if the variance of the Horvitz-Thompson estimator of the total is approximately $(1 - n / N)$ times the variance in (2), with $P_i$ taken as $n^{-1}$ times the overall selection probability for unit $i$ (Särndal, Swensson, and Wretman 1992, 154).

## 5. Derivation of Variance Formula (3)

From (2) we have $V(\hat{Y}) = n^{-1}(\sum_{i=1}^{N} y_i^2 / P_i - Y^2)$. Next, note that (1) implies that

$$Y^2 = (N\alpha + \beta)^2 = N^2\alpha^2 + 2N\alpha\beta + \beta^2 \quad (8)$$

and

$$\begin{aligned} \sum_{i=1}^{N} y_i^2/P_i &= \sum_{i=1}^{N} [\alpha^2/P_i + \beta^2 P_i + \varepsilon_i^2/P_i + 2\alpha\beta + 2\alpha\varepsilon_i/P_i + 2\beta\varepsilon_i] \\ &= \alpha^2 \sum_{i=1}^{N} P_i^{-1} + \beta^2 + \sum_{i=1}^{N} \varepsilon_i^2/P_i + 2N\alpha\beta \\ &\quad + 2\alpha \sum_{i=1}^{N} \varepsilon_i/P_i \\ &= \alpha^2 n \sum_{i=1}^{N} w_i + \beta^2 + n \sum_{i=1}^{N} \varepsilon_i^2 w_i + 2N\alpha\beta \\ &\quad + 2\alpha n \sum_{i=1}^{N} \varepsilon_i w_i. \end{aligned} \quad (9)$$

Subtracting (8) and (9) and dividing by $n$ yields

$$V(\hat{Y}) = \alpha^2 \left( \sum_{i=1}^{N} w_i - N^2 / n \right) + \sum_{i=1}^{N} \varepsilon_i^2 w_i + 2\alpha \sum_{i=1}^{N} \varepsilon_i w_i.$$

To obtain (3), note that

$$\begin{aligned} \sum_{i=1}^{N} \varepsilon_i^2 w_i &= N\rho_{\varepsilon^2, w}\sigma_{\varepsilon^2}\sigma_w + N\bar{W}\sigma_\varepsilon^2 \\ &= N\rho_{\varepsilon^2, w}\sigma_{\varepsilon^2}\sigma_w + (1 - \rho_{y, P}^2)\sigma_y^2 N\bar{W} \end{aligned}$$

and

$$\sum_{i=1}^{N} \varepsilon_i w_i = N\rho_{\varepsilon, w}\sigma_\varepsilon\sigma_w.$$

## References

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Ed. New York: John Wiley & Sons, Inc.

Gabler, S., Haeder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 1, 105-106.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L. (1992). Weighting for unequal $P_i$. *Journal of Official Statistics*, 8, 2, 183-200.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.