# THE DESIGN EFFECT: DO WE KNOW ALL ABOUT IT?

**Inho Park and Hyunshik Lee, Westat**
**Inho Park, Westat, 1650 Research Boulevard, Rockville, Maryland 20850**

**Key Words: Complex sample, Simple random sample, pps sampling, Point and variance estimation, Ratio mean, Total estimator**

## 1. Introduction

The design effect is widely used in survey sampling for planning a sample design and to report the effect of the sample design in estimation and analysis. It is defined as the ratio of the variance of an estimator under a sample design to that of the estimator under simple random sampling. An estimated design effect is routinely produced by sample survey software such as WesVar and SUDAAN.

This paper examines the relationship between the design effects of the total estimator and the ratio mean estimator, under unequal probability sampling. After a brief review on various definitions and practical usage of the design effect, we consider a decomposition of the design effect of the total estimator in term of that of the ratio mean. In addition, a model-assisted method is used to derive some useful approximate formulae, with which we make further comparison of the design effects of the two estimators. The approximate formulae derived here are also compared with the well-known Kish's approximate formula (Kish, 1965, 1995). Finally, we apply our formulae to an artificial population created from a complex health survey data for illustration.

## 2. Definition and Use of Design Effect in Practice

The precursor of the design effect that has been popularized by Kish (1965) was used by Cornfield (1951). He defined the efficiency of a complex sampling design as the ratio of the variance of a statistic under simple random sampling without replacement (srswor) with a sample size of *n* to the variance of the statistic under the complex design with the same sample size. The inverse of the ratio defined by Cornfield was also used by other sampling statisticians. For example, Hansen, Hurwitz, and Madow (1953, pp. 259-270) discussed the increase of the variance due to clustering effect of the cluster sampling over srswor. However, the name, the design effect, or Deff in short was coined and defined formally by Kish (1965, Section 8.2, p. 258) as "the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements." The Deff for an estimate of the population mean is given by

$$\text{Deff} = \text{Var}(\bar{y}) / \{(1 - f)\, S_y^2 / n\} \qquad (2.1)$$

where $\bar{y}$ is an estimate of the population mean $(\bar{Y})$ under a complex design with the sample size of *n*, *f* is a sampling fraction, and $S_y{}^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ is the population element variance of the *y*-variable. In general, the design effect can be defined for any meaningful statistic computed from the sample selected from the complex sample design.

The Deff is a population quantity that depends on the sample design and the statistic. The same parameter can be estimated by different estimators and their Deff's are different even under the same design. Therefore, the Deff includes not only the efficiency of the design but also the efficiency of the estimator. Särndal et al. (1992, p. 54) made this point clear by defining it as a function of the design (*p*) and the estimator $(\hat{\theta})$ for the population parameter $\theta$. Thus, we may write it as

$$\text{Deff}(p, \hat{\theta}) = \text{Var}_p(\hat{\theta}) / \text{Var}_{\text{srswor}}(\hat{\theta}') \qquad (2.2)$$

where $\hat{\theta}'$ is the usual form of estimator for $\theta$ under srswor, which is normally different from $\hat{\theta}$. For example, to estimate the population mean, one may use the ratio mean $\hat{\theta} = \sum_s w_i y_i / \sum_s w_i$ with sampling weights $w_i$ but $\hat{\theta}'$ would be the sample mean $\sum_s y_i / n$. We will see the effect of a particular statistic $\hat{\theta}$ on the design effect in Section 3. In particular, we will show that the Deff for the Hansen-Hurwitz (or Horvitz-Thompson) estimator for the population total can be very different from the ratio mean for the population mean.

Cochran (1977, p. 85) stated "The design effect has two primary uses – in sample size determination and in appraising the efficiency of more complex plans." These are still main uses of the design effect. However, other important uses have emerged; to compute the effective sample size and to modify the inferential statistic in data analysis as Scott and Rao (1981) used the average design effect to modify the Pearson type $\chi^2$-test statistic for complex samples.

Kish (1992) later advocated using a slightly different Deff, which is called Deft and uses the simple random sampling with replacement (srswr) variance in the denominator. His logic was that without-replacement sampling is a part of design and should be captured in the definition. Deft is also easier to use for making inferences. Another reason Kish quoted is that the finite population correction factor $(1-f)$ may be

difficult to compute in some situations. The new definition is given by

$$\text{Deft}(p,\hat{\theta}) = \sqrt{\text{Var}_p(\hat{\theta}) / \text{Var}_{\text{srswr}}(\hat{\theta}')} \qquad (2.3)$$

or $\text{Deft}^2(p,\hat{\theta}) = \text{Var}_p(\hat{\theta}) / \text{Var}_{\text{srswr}}(\hat{\theta})$. Survey data software such as WesVar and SUDAAN produce $\text{Deft}^2$ instead of Deff.

When the population parameter is the total (*Y*), the unbiased estimator is the (weighted) sample total, namely, $\hat{Y} = \sum_s w_i y_i$, where $w_i$ is the properly defined sampling weight and the summation is over the sample *s*. When the mean is the parameter of interest, it is usually estimated by the ratio mean, that is, $\hat{\bar{Y}} = \sum_s w_i y_i / \sum_s w_i$. It is a special case of the ratio estimator, $\sum_s w_i y_i / \sum_s w_i x_i$, where $x_i \equiv 1$ for all $i \in s$.

One common misconception about the design effects of $\hat{Y}$ and $\hat{\bar{Y}}$ is that they in values are similar. However, it has been observed that the design effect of $\hat{Y}$, $\text{deft}^2(p,\hat{Y})$, is much larger than that of the ratio mean, $\text{deft}^2(p,\hat{\bar{Y}})$ for human populations. For example, see Kish (1987) and Hansen, Hurwitz, and Madow (1953, p. 608). Some explanation can be found in Särndal et al. (1992, p. 65 and p. 133) who showed that $\text{deft}^2(p,\hat{Y})$ depends on the relative variation of the *y*-variable under Bernoulli sampling and a special case of one-stage cluster sampling. This dependence contradicts what the design effect is intended to measure as Kish (1995) explicitly described: "Deft are used to express the effects of sample design beyond the elemental variability ($S_y^2 / n$), removing both the units of measurement and sample size as nuisance parameters. With the removal of $S_y$, the units, and the sample size *n*, the design effects on the sampling errors are made generalizable (transferable) to other statistics and to other variables, within the same survey, and even to other surveys." His statement is loosely true for the ratio mean $\hat{\bar{Y}}$, as the usual approximate design effect formula is independent of a particular *y*-variable. A frequently used approximate formula for the $\text{Deft}^2$ of the ratio mean is given by Kish (1987)

$$\text{Deft}^2(p,\hat{\bar{Y}}) = \{1 + \rho(\bar{b} - 1)\}(1 + \text{cv}_w^2) \qquad (2.4)$$

where *p* is an unequal probability sampling of clusters design, $\rho$ is the intraclass correlation (often called within cluster homogeneity measure), $\bar{b}$ is the average cluster size, and $\text{cv}_w^2$ is the relative variance of the sampling weights. Gabler, Haeder, and Lahiri (1999) justified expression (2.4) using a superpopulation model. This formula is valid only when there is no correlation between the sampling weights and the survey variable (*y*). However, if the correlation is present, the formula must be modified as studied by Spencer (2000). We also examine this point further along with Spencer's approximate formulae in Section 4.

However, when one comes to the Horvitz-Thompson type of total estimator, the situation becomes very different. Particularly, when the weights are poorly correlated with the *y*-variable and the weights vary a lot, the design effect for the total estimator can be much greater than that for the ratio mean. In Section 3, we analyze this fact in detail for unequal probability sampling.

## 3. Decomposition of Design Effect of Sample Total Under Unequal Probability Sampling

Consider a sample design (denoted by $p$) with a sample size of $n$ drawn by unequal probability sampling with replacement from a finite population $U$ of size *N*. Let $y_k$ denote the *y*-value of element *k* and let $p_k$ denote the associated selection probability, where $\sum_U p_k = 1$. Note that the $n$ draws are independent since sample selection is with replacement. If $p_k$ are proportional to a positive size measure $x_k$, that is, $p_k \propto x_k$, then the sampling scheme is probability proportional to size (pps) sampling.

Let $k_i$ represent the element selected on the *i*-th draw. Then the Hansen-Hurwitz's estimator of the finite population total, $Y = \sum_U y_k$, is given as

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n} y_{k_i} / p_{k_i} \; . \qquad (3.1)$$

It is easy to see that $E_p(y_{k_i} / p_{k_i}) = Y$ and thus $\hat{Y}$ is the average of $n$ unbiased estimators of $Y$. For simplicity, we use *i* to denote $k_i$ unless any confusion arises. The variance of $\hat{Y}$ can be written as

$$\begin{aligned} \text{Var}(\hat{Y}) &= \frac{1}{n}\sum_U p_i \{(y_i / p_i) - Y\}^2 \\ &= \frac{1}{n}\sum_U p_i^{-1}(y_i - p_i Y)^2 \end{aligned} \qquad (3.2)$$

(see Särndal et al., 1992, pp. 51-52).

Now consider the estimation of the finite population mean, $\bar{Y} = Y / N$. It can be estimated by the ratio mean $\hat{\bar{Y}} = \hat{Y} / \hat{N}$ where $\hat{N} = \sum_{i=1}^{n} 1/(np_i)$ is an estimator of the population size *N*. Using Taylor linearization, as shown in Särndal et al. (1992, pp. 172-176), $\hat{\bar{Y}}$ can be approximated as

$$\hat{\bar{Y}} \doteq \bar{Y} + N^{-1}\hat{d} \, , \qquad (3.3)$$

where $\hat{d} = n^{-1}\sum_{i=1}^{n} d_i / p_i$ with $d_i = y_i - \bar{Y}$. Using expression (3.3), an approximate variance of $\hat{\bar{Y}}$ is obtained as

$$\mathrm{AVar}(\hat{\bar{Y}}) \equiv N^{-2}\mathrm{Var}(\hat{d}) = N^{-2}\frac{1}{n}\sum_U p_i^{-1} d_i^2 \quad (3.4)$$

where the equation holds since $\sum_U d_i = 0$.

Now we derive the design effects of $\hat{Y}$ and $\hat{\bar{Y}}$. Under srswr of size *n*, the total estimator is given by $\hat{Y}_s = N\bar{y}_s$ and its variance by $\mathrm{Var}_{\mathrm{srswr}}(\hat{Y}_s) = N^2 S_y^2 / n$, where $\bar{y}_s = \sum_{i=1}^{n} y_i / n$.

From (3.2) and assuming *N* is large (so that $N/(N-1) \doteq 1$), an approximate design effect of $\hat{Y}$ is obtained as

$$\mathrm{Deft}^2(\hat{Y}) \doteq \frac{\sum_U p_i^{-1}(y_i - p_i Y)^2}{\sum_U N(y_i - \bar{Y})^2}. \quad (3.5)$$

Also, since $\mathrm{Var}_{\mathrm{srswr}}(\hat{\bar{Y}}_s) = S_y^2 / n$, where $\hat{\bar{Y}}_s = \bar{y}_s$, it follows from (3.4) that

$$\mathrm{Deft}^2(\hat{\bar{Y}}) \doteq \frac{\sum_U p_i^{-1}(y_i - \bar{Y})^2}{\sum_U N(y_i - \bar{Y})^2}. \quad (3.6)$$

Note that $\mathrm{Deft}^2(\hat{Y})$ comes closer to $\mathrm{Deft}^2(\hat{\bar{Y}})$ and both approach to 1 as $p_i$ approaches to a constant, say $1/N$, for all $i \in U$. Note also that sample design *p* is omitted from expressions (3.5) and (3.6) for brevity's sake.

In addition, we see from expressions (3.5) and (3.6) that the only difference between these expressions is in the square deviations of the numerator. Using the standard ANOVA decomposition to the numerator of (3.5), we can write

$$\begin{aligned}\sum_U p_i^{-1}(y_i - p_i Y)^2 &= \sum_U p_i^{-1}(y_i - \bar{Y})^2 \\ &\quad + N^2\bar{Y}^2\sum_U p_i^{-1}(p_i - \bar{P})^2 \\ &\quad - 2N\bar{Y}\sum_U p_i^{-1}(y_i - \bar{Y})(p_i - \bar{P})\end{aligned}$$

where $\bar{P} = \sum_U p_i / N = 1/N$. Hence, we have

$$\begin{aligned}\mathrm{Deft}^2(\hat{Y}) &\doteq \mathrm{Deft}^2(\hat{\bar{Y}}) + \frac{1}{\mathrm{CV}_y^2}\sum_U p_i^{-1}(p_i - \bar{P})^2 \\ &\quad - \frac{2}{\mathrm{CV}_y^2}\frac{1}{\bar{Y}}\sum_U p_i^{-1}(y_i - \bar{Y})(p_i - \bar{P}) \\ &= \mathrm{Deft}^2(\bar{Y}) + g(\mathrm{CV}_y^2) \quad (3.7)\end{aligned}$$

where $\mathrm{CV}_y = S_y / \bar{Y}$ denotes the coefficient of variation (CV) of the *y*-variable. The last two terms are denoted by $g$ as a function of $\mathrm{CV}_y^2$.

Note that *g* is a decreasing function of $\mathrm{CV}_y^2$ and thus $\mathrm{Deft}^2(\hat{Y})$ becomes larger than $\mathrm{Deft}^2(\hat{\bar{Y}})$ as $\mathrm{CV}_y^2$ decreases. This imbalance can be dramatic when $p_i$'s vary a lot, $p_i$'s and $(y_i - \bar{Y})^2$'s are uncorrelated, and $\mathrm{CV}_y^2$ is small (see Section 5 for an example).

When $p_i$'s and $(y_i - \bar{Y})^2$'s are uncorrelated, which happens often in social or health surveys, we can approximate $\sum_U p_i^{-1}(y_i - \bar{Y})^2$ by $n\overline{W}\sum_U (y_i - \bar{Y})^2$ and then expression (3.6) can be simplified as

$$\mathrm{Deft}^2(\hat{\bar{Y}}) \doteq n\overline{W} / N,$$

where $\overline{W} = \sum_U w_i / N$ with $w_i = 1/(np_i)$. Later we will see that $n\overline{W} / N$ can be estimated by $1 + \mathrm{cv}_w^2$ where $\mathrm{cv}_w^2$ is the relative variance of the sampling weights, $w_i$'s.

**Remark 3.1** (**Binary Variable: total vs. proportion).** For a binary variable, we can show that $\mathrm{CV}_y^2 \doteq (1 - \bar{Y})/\bar{Y}$. When $p = \bar{Y}$ is close to 1, then $g(\mathrm{CV}_y^2)$ in (3.7) can be very large, while $g(\mathrm{CV}_y^2)$ becomes small and thus the two design effects are close as $\bar{Y} \to 0$.

**Remark 3.2 (Comparison with a Model-based Approach).** The difference between $\mathrm{Deft}^2(\hat{Y})$ and $\mathrm{Deft}^2(\hat{\bar{Y}})$, that is, the quantity $g(\mathrm{CV}_y^2)$ in decomposition (3.7), cannot be revealed by a model-based approach used by Gabler, Haeder, and Lahiri (1999), since $y_i$'s are treated as random variables while $w_i$'s as fixed constants. Under the model-based approach, $\mathrm{Deft}^2(\hat{Y})$ is different from $\mathrm{Deft}^2(\hat{\bar{Y}})$ by a factor of $(\hat{N}/N)^2$, which is negligible in comparison with $g(\mathrm{CV}_y^2)$. More discussion will come in Section 4.

## 4. Design Effects when the Survey Variable is Linearly Related with the Sampling Weight

Assuming a linear relation between the *y*-value and the selection probability, Spencer (2000) derived an approximate formula to the design effect of the Hansen-Hurwitz estimator under unequal probability sampling with-replacement. However, it appears that he did not recognize the difference between the design effects of the total and the ratio mean estimators, as he compared directly his approximate formula for the total estimator

with Kish's approximate formula $1+\mathrm{cv}_w^2$, which was originally derived for the ratio mean estimator. In this section, we use Spencer's set-up to derive approximate formula for the mean estimator and then we compare them with those for the total estimator in the same context of Section 3.

Consider the linear regression of selection probability $p_i$ on $y_i$ given by

(S1) $\quad y_i = A + Bp_i + e_i$.

The least-square regression coefficients of this model at the population level are given by $A = \bar{Y} - B\bar{P}$ and $B = \sum_U (y_i - \bar{Y})(p_i - \bar{P}) \big/ \sum_U (p_i - \bar{P})^2$. As before, $w_i = 1/(np_i)$ and $\bar{W} = \sum_U w_i / N$. The ratio mean estimator has an approximate variance $\mathrm{AVar}(\hat{\bar{Y}}) = N^{-2} \sum_U w_i (y_i - \bar{Y})^2$ as given in (3.4). Since $\bar{Y} = A + B\bar{P}$ and $y_i - \bar{Y} = B(p_i - \bar{P}) + e_i$, we have $\sum_U w_i \times (y_i - \bar{Y})^2 = \sum_U e_i^2 w_i - 2B\sum_U e_i w_i / N + B^2(\bar{W}/N - 1/n)$. In parallel with Spencer (2000), under (S1) we can write

$$\begin{aligned} N^2 \mathrm{AVar}(\hat{\bar{Y}} \mid \mathrm{S1}) &= (N-1) R_{e^2 w} S_{e^2} S_w + (N-1)(1 - R_{yp}^2) S_y^2 \bar{W} \\ &\quad - 2BR_{ew} S_e S_w + B^2 \left( \bar{W}/N - 1/n \right) \end{aligned} \tag{4.1}$$

where $S_e$, $S_{e^2}$, $S_w$ denote the (finite) population standard deviations of the $e_i$'s, the $e_i^2$'s, the $w_i$'s, respectively, and $R_{ew}$, $R_{e^2 w}$, $R_{yp}$ denote the population correlations of pairs $(e_i, w_i)$, $(e_i^2, w_i)$ and $(y_i, p_i)$, respectively. For example, $R_{yw} = \sum_U (y_i - \bar{Y})(w_i - \bar{W}) \big/ \{(N-1) S_y S_w\}$. To obtain (4.1), we have used the following expressions given in Spencer (2000, Section 5): $\sum_U e_i^2 w_i = (N-1) R_{e^2 w} \times S_{e^2} S_w + (N-1)\bar{W} S_y^2 (1 - R_{yp}^2)$ and $\sum_U e_i w_i = (N-1) \times R_{ew} S_e S_w$. If the regression model (S1) fits well to the population and the error variance is roughly homogeneous so that

(S2) $\quad R_{ew} \doteq 0$ and $R_{e^2 w} \doteq 0$,

then expression (4.1) further simplifies to

$$\begin{aligned} N^2 \mathrm{AVar}(\hat{\bar{Y}} \mid \mathrm{S1,S2}) &= (1 - R_{yp}^2) S_y^2 (N-1) \bar{W} + B^2 \left( \bar{W}/N - 1/n \right). \end{aligned} \tag{4.2}$$

Since $\mathrm{Var}_{\mathrm{srswr}}(\hat{\bar{Y}}_s) = S_y^2 / n$, the design effect of $\hat{\bar{Y}}$ under (S1) and (S2) is given as

$$\begin{aligned} \mathrm{Deft}^2(\hat{\bar{Y}} \mid \mathrm{S1,S2}) &\doteq (1 - R_{yp}^2) n\bar{W}/N + \frac{B^2}{S_y^2} \frac{1}{N^2} \left( n\bar{W}/N - 1 \right) \\ &= (1 - R_{yp}^2) n\bar{W}/N + \left( \frac{R_{yp}}{\mathrm{CV}_p} \right)^2 \left( n\bar{W}/N - 1 \right), \end{aligned} \tag{4.3}$$

where the second expression follows since $B = R_{yp} S_y / S_p$ and $\bar{P} = 1/N$. Note that $S_p$ denotes the population standard deviation of the $p_i$'s and $\mathrm{CV}_p$ is the coefficient of variation of $p_i$'s.

Since $E(w_i) = N/n$ and $E(w_i^2) = N\bar{W}/n$, we have $E(w_i^2)/E^2(w_i) = n\bar{W}/N$. Based on these relations, Spencer substituted $n\bar{W}/N$ in his derivation (Section 5) by an estimate

$$\frac{\sum_s w_i^2 / n}{\left( \sum_s w_i / n \right)^2}. \tag{4.4}$$

Note that expression (4.4) is equivalent to Kish's approximate formula $1+\mathrm{cv}_w^2$ for the ratio mean estimator $\hat{\bar{Y}}$. Substituting further $R_{yp}$ and $\mathrm{CV}_p$ by their respective sample estimates, say $r_{yp}$ and $\mathrm{cv}_p$, in expression (4.3), we obtain the following estimator of the design effect of the ratio mean estimator:

$$\mathrm{deft}^2(\hat{\bar{Y}} \mid \mathrm{S1,S2}) = (1 + \mathrm{cv}_w^2)(1 - r_{yp}^2) + \frac{r_{yp}^2}{\mathrm{cv}_p^2} \mathrm{cv}_w^2 \tag{4.5}$$

Kish's formula immediately follows by setting $r_{yp}$ =0, that is,

$$\mathrm{deft}^2(\hat{\bar{Y}} \mid \mathrm{S1,S2}, r_{yp} = 0) = 1 + \mathrm{cv}_w^2. \tag{4.6}$$

Meanwhile, Spencer (2000) derived the population design effect for the total estimator under (S1) and (S2) as follows:

$$\begin{aligned} \mathrm{Deft}^2(\hat{Y} \mid \mathrm{S1,S2}) &\doteq (1 - R_{yp}^2) n\bar{W}/N + (A/S_y)^2 \left( n\bar{W}/N - 1 \right) \end{aligned} \tag{4.7a}$$

which can be rewritten using $A = \bar{Y} - B\bar{P}$ as

$$\mathrm{Deft}^2(\hat{Y} \mid \mathrm{S1,S2}) \doteq (1 - R_{yp}^2) n\bar{W}/N + \left( \frac{R_{yp}}{\mathrm{CV}_p} - \frac{1}{\mathrm{CV}_y} \right)^2 \left( n\bar{W}/N - 1 \right) \tag{4.7b}$$

An estimator of this approximate design effect is then given by

$$\text{deft}^2(\hat{Y} \mid \text{S1, S2})$$

$$= (1+\text{cv}_w^2)(1-r_{yp}^2) + \left(\frac{r_{yp}}{\text{cv}_p} - \frac{1}{\text{cv}_y}\right)^2 \text{cv}_w^2 \qquad (4.8)$$

If $r_{yp} = 0$, then Spencer's formula simplifies to

$$\text{deft}^2(\hat{Y} \mid \text{S1, S2}, r_{yp} = 0) = 1 + \text{cv}_w^2 + \frac{\text{cv}_w^2}{\text{cv}_y^2}, \qquad (4.9)$$

which does not reduce to Kish's formula unless $\text{cv}_w^2 / \text{cv}_y^2 = 0$.

**Remark 4.1 (Spencer's approximate formula).** In his original derivation, Spencer proposed the following formula to estimate the design effect:

$$(1-r_{yp}^2)(1+\text{cv}_w^2) + (a/s_y)^2 \text{cv}_w^2,$$

where $a$ and $s_y$ are the sample estimates for $A$ and $S_y$, respectively or $1 + cv_w^2 + cv_w^2 a^2 / s_y^2$ as a special case for $r_{yp} = 0$ with saying that "this is close to Kish's approximation when $a/s_y$ is near zero." However, if $r_{yp} = 0$, then $a = \hat{\bar{Y}}$ and $a/s_y = 1/\text{cv}_y$, and it becomes exactly (4.9).

**Remark 4.2 (Comparison of $\text{Deft}^2(\hat{Y})$ and $\text{Deft}^2(\hat{\bar{Y}})$).** Using Cauchy-Schwarz inequality, it is easy to show that $n\overline{W}/N \geq 1$. Assume $y_i \geq 0$ so that $\text{CV}_y > 0$. From (4.3) and (4.7b), it follows that $\text{Deft}^2(\hat{Y} \mid \text{S1, S2}) \geq \text{Deft}^2(\hat{\bar{Y}} \mid \text{S1, S2})$ iff $2R_{yp} \leq \text{CV}_p / \text{CV}_y$ (the equality holds iff $2R_{yp} = \text{CV}_p / \text{CV}_y$). If $R_{yp} = 0$, $\text{Deft}^2(\hat{Y} \mid \text{S1, S2}) \geq \text{Deft}^2(\hat{\bar{Y}} \mid \text{S1, S2})$ and the equality holds under srswr, in which case $\text{Deft}^2(\hat{Y}) = \text{Deft}^2(\hat{\bar{Y}}) = 1$.

Under ppswr (i.e., $p_i \propto x_i$), $p_i = x_i / \sum_U x_i$ and thus we have $\text{CV}_p = \text{CV}_x$ and $R_{yp} = R_{yx}$ provided $p_i < 1$ for all $i \in U$. Therefore, we have $\text{Deft}^2(\hat{Y} \mid \text{S1, S2}) \geq \text{Deft}^2(\hat{\bar{Y}} \mid \text{S1, S2})$ iff $2R_{yx} \leq \text{CV}_x / \text{CV}_y$. On the other hand, the inequality can be reversed when $R_{yx}$ is large (i.e., close to 1) and even $\text{Deft}^2(\hat{Y})$ can be well below 1. This happens frequently in business surveys where the *y*-variable is heavily skewed to the right and the size measure variable ($x$) is highly correlated with the *y*-variable (see also Lehtonen and Pahkinen, 1995, p. 110).

## 5. Example

In this section we discuss comparison between the design effects of the sample total and the ratio mean estimator using an artificial dataset created from a health survey. The dataset is a subset of the adult data from the U.S. third National Health and Nutrition Examination Survey (NHANES III), which is given as a demo file in WesVar version 4. From the dataset, $N$=5,000 records were selected by simple random sampling without replacement to construct an artificial finite population. Sampled were only the records with complete responses to the four variables CIGNUM (number of cigarettes smoked per day), SYSTOLIC (average systolic blood pressure), DIASTOL (average diastolic blood pressure) and HEIGHT (height without shoes - inches). The inverse of the final weight in the demo file was used as the measure of size (MOS) for our sampling experiment. Note that the final weight in the demo file is different from the NHANES III final weight that was obtained by further adjusting the weight by poststratification.

Table 1. Parameters of the artificial population

| Variable ($y$) | $Y$ | $\bar{Y}$ | $\text{CV}_y$ | $R_{yp}$ |
|---|---|---|---|---|
| CIGNUM | 18,620 | 3.72 | 2.3781 | -.0966 |
| SYSTOLIC | 640,288 | 128.06 | .1700 | .1300 |
| DIASTOL | 377,920 | 75.58 | .1628 | .0278 |
| HEIGHT | 331,269 | 66.25 | .0606 | -.1058 |
| $p_i$ | 1 | 1/5,000 | 1.3910 | 1.0000 |
| $w_i$ | 846,946 | 169.39 | 1.2350 | -.4104 |

Table 1 presents several parameters of the artificial population on the selected four variables including the selection probabilities $p_i$'s and the sampling weights $w_i$'s. The survey variables are weakly correlated with the selection probability, where $R_{yp}$ ranges from -.1058 to .1300. $\text{CV}_y$ for SYSTOLIC, DIASTOL and HEIGHT is less than .20, while for CIGNUM it is 2.3781, which is much larger than others.

We used ppswr with $n = 100$.

Table 2 displays the decomposition of the population design effect in (3.7) for the four survey variables. Since CIGNUM has a large $\text{CV}_y$, the difference between $\text{Deft}^2(\hat{Y})$ and $\text{Deft}^2(\hat{\bar{Y}})$ is small. However, the difference is very large for the other three variables due to small $\text{CV}_y$, which causes $g(\text{CV}_y^2)$ to become dominant in $\text{Deft}^2(\hat{Y})$ for these three variables.

Table 2. Decomposition of design effect in (3.7)

| Variable | $\text{Deft}^2(\hat{\bar{Y}})$ | $\text{Deft}^2(\hat{Y})$ | $g(\text{CV}_y^2)$ |
|---|---|---|---|
| CIGNUM | 4.837 | 5.676 | .837 |
| SYSTOLIC | 2.833 | 78.205 | 75.372 |
| DIASTOL | 3.028 | 91.377 | 88.349 |
| HEIGHT | 3.488 | 668.036 | 664.548 |

Table 3 shows Pearson's correlations of the sampling weights $w_i$ with $e_i$ and $e_i^2$, and approximate values to the population design effects of the ratio mean and the sample total estimators. Although the plots of data points $(x_i, y_i)$ show less than perfect linear relationships between $x_i$ and $y_i$ (i.e., $p_i$ and $y_i$), Table 3 reveals that $R_{ew}$ and $R_{e^2w}$ are nearly zero and the approximate formulae (4.3) and (4.7a) under (S1) and (S2) trace fairly closely the true design effect given in Table 2.

Table 3. Approximate design effects under linear relation between $p_i$'s and $y_i$'s

| | Goodness of Fit | | Approximate $\text{Deft}^2$ | |
|---|---|---|---|---|
| Variable | $R_{ew}$ | $R_{e^2w}$ | $\text{Deft}^2(\hat{\bar{Y}})$ | $\text{Deft}^2(\hat{Y})$ |
| CIGNUM | 0.079 | 0.097 | 3.368 | 3.930 |
| SYSTOLIC | -0.093 | -0.078 | 3.351 | 83.293 |
| DIASTOL | -0.022 | -0.045 | 3.386 | 92.857 |
| HEIGHT | 0.058 | 0.016 | 3.364 | 660.107 |

## 6. Concluding Remarks

Kish originally intended the design effect as a measure of the effect of the sample design in parameter estimation, which is independent of the elemental variability of a particular *y*-variable, the unit of measurement, and the sample size. He is largely successful for the ratio mean estimator to achieve this goal. Probably due to this success, it is commonly misconceived that the design effect of the total estimator is similar to that of the ratio mean estimator. However, as clearly demonstrated in this paper, the design effect of the Horvitz-Thompson type estimator for the total under an unequal probability sampling design is not only dependent on the variability of *y*-variable ($\text{CV}_y$) but also can be greatly influenced by it. Another important factor in determining the design effect is the correlation ($R_{yp}$) between the selection probability and the *y*-variable as shown in Section 4. Kish's approximate design effect formulae were derived assuming tacitly that this correlation is zero. This may be the reason why the design effect is not used in business surveys, where the correlation is usually large and the design effect can be much smaller than 1. However, we can use the design effect in business surveys as well to measure the gain by the sample design (instead of the loss, which is usually the case in social surveys) over the simple random sampling. Business surveys often employ pps sampling or size-stratification. In this context, it would be worth while to pursue further on this research to take into account of without-replacement sampling and the finite population correction (fpc). The fpc can be important particularly in business surveys. We are currently working on the extension of our formulae to multi-stage sampling. We will also try to tackle these problems if time permits.

Lastly, we would like to point out the importance of an efficient estimation technique such as poststratification for estimation of the total, when $R_{yp}$ is close to zero.

## 7. References


Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New Work: Wiley.

Cornfield, J. (1951). Modern Methods in the Sampling of Human Populations. *American Journal of Public Health*, **41**, pp. 654-661.

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**, pp.105-106.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory*, Vol. I. New York: Wiley.

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

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

Kish, L. (1992). Weighting for Unequal $p_i$. *Journal of Official Statistics*, **8**, pp.183-200.

Lehtonen, R., and Pahkinen, E. J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: Wiley.

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

Scott, A. J., and Rao, J. N. K. (1981). Chi-squared Tests for Contingency Tables with Proportions estimated from Survey Data. In: D. Krewski, R. Platek, and J. N. K. Rao (eds.), *Current Topics in Survey Sampling*. New York: Academic Press, pp. 247-265.

Spencer, B. D. (2000). An Approximate Design Effect for Unequal Weighting When Measurements May Correlate With Selection Probabilities. *Survey Methodology*, **26**, pp. 137-138.

Westat (2001). *WesVar 4.0 User's Guide*. Rockville, MD: Westat, Inc.