The Hájek Estimator Revisited
Alan H. Dorfman, Room 4915, 2 Massachusetts Ave. NE, Washington DC 20212

Keywords: balanced sample, best linear unbiased predictor, Horvitz-Thompson estimator, ratio estimator, robustness, superpopulation model.

1. The Hájek Estimator

Consider a population \( U \) of \( N \) units indexed by \( i \). Suppose \( Y_i, i \in U \) are values in the population of interest; in particular, suppose we wish to estimate their mean \( \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \) based on a sample \( s \) taken from the population \( U \). Assume that the sample is taken according to a randomization scheme having inclusion probabilities \( \pi_i = \Pr(i \in s) \). When the \( \pi_i \) are proportional to a positive quantity \( x_i \) available over \( U \), and \( s \) has a predetermined sample size \( n \), then \( \pi_i = nx_i/N \), where \( x = \sum_{i=1}^{N} x_i/N \), and the sampling scheme is said to be probability proportional to size (pps).

Under this scheme, a well known and popular estimator attributed to Hájek (1971) is defined by

\[
\hat{Y}_{H} = \frac{\sum_{i=1}^{n} Y_i / \pi_i}{\sum_{i=1}^{n} 1/\pi_i}.
\]

He suggested this estimator in response to an observation by Basu (1971) on paradoxical behavior of the pps-unbiased Horvitz-Thompson (1952) estimator

\[
\hat{Y}_{HT} = \frac{\sum_{i=1}^{n} Y_i / \pi_i}{\sum_{i=1}^{n} 1/\pi_i}.
\]

Särndal, Swenson, and Wretman (1992, p. 182, referred to below as SSW) give several reasons for regarding the Hájek as "usually the better estimator", namely the relative behavior of \( \hat{Y}_{H} \) and \( \hat{Y}_{HT} \) when (a) the \( Y_i \) are relatively homogeneous, or (b) sample size is not fixed, or (c) the \( \pi_i \) are weakly or negatively correlated with the \( Y_i \). The Hájek estimator can be derived from the theory of optimal estimating equations if we regard \( \bar{Y} \) as the "induced finite population parameter" under the superpopulation model \( Y_i \sim (\mu, \sigma^2) \) with the \( Y_i \)'s independent (Godambe and Thompson (1986), Example 1).

Our present purpose is to examine the Hájek estimator, and the Horvitz-Thompson (HT) estimator as well, in the light of recent results using the "model-based" or prediction approach to survey sampling. In particular, we investigate consequences of a theorem connecting optimality and weighted balanced samples (Royall 1992, Theorem 2), stated below.

2. Background: Simple Balanced Samples

We continue to consider the problem of estimating the population mean \( \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \). Assume that a single auxiliary \( x_i \) is associated with and known for each unit \( i \) in the population, and that \( Y_i \) and \( x_i \) are related by the polynomial model

\[
Y_i = \sum_{j=0}^{d} \delta_j x_i^j + e_i \eta^2 \tag{3}
\]

where the errors are \( e_i \sim (0, \sigma^2) \) and uncorrelated, \( \{ \beta_j \}_{j=1}^{d} \) are a set of unknown parameters, and \( \{ \delta_j \}_{j=1}^{d} \) are 0-1 variables indicating whether the \( j \)th power term is in the model or not. Let \( M(\delta_0, \delta_1, \ldots, \delta_d; \nu) \) denote model (3) and \( \hat{Y}(\delta_0, \delta_1, \ldots, \delta_d; \nu) \) denote the BLU predictor under that model, following the convention in Royall and Herson (1973). For example, \( M(0,1:x) \) refers to the model \( Y_i = \beta x_i + e_i x_i^2 \) and \( \hat{Y}(0,1:x) \) is readily shown to be the well known ratio estimator \( \hat{Y}_R = \frac{\bar{Y}_x}{\bar{x}_x} \).

One of the principal concerns of the model-based opus is the question of robustness: how well does an estimator perform when the hypothesized model is incorrect, and what measures can be taken to guard against degradation of its performance (in terms, say, of root mean square error) under this almost inevitable circumstance? It is helpful to determine the behavior of an estimator defined in terms of one (usually fairly simple) model—the working model—and under another (usually more complicated) true model. Note; we will denote expectations with respect to a model by \( E_M \) and with respect to a sampling plan by \( E_s \).

Thus, for the ratio estimator, we ask about its bias under (3), i.e., under \( M(\delta_0, \delta_1, \ldots, \delta_d; \nu) \), and find that

\[
E_M[\hat{Y}(0,1:x) - \bar{Y}] = \bar{x} \sum_{j=1}^{d} \beta_j \left[ \frac{x_i^j}{y_s} - \frac{x_i^j}{\bar{x}} \right] \tag{4}
\]

where \( \bar{x}_s = \sum_i x_i/n \) and \( \bar{x} = \sum_{i=1}^{N} x_i/N \). Note that there is no contribution from the term \( j=1 \), a fact
not surprising, since the model underlying the ratio estimator contains \( b_1 \). If

\[
\frac{Y^{(j)}}{X_j} = \frac{X^{(j)}}{X}, \quad j = 0, \ldots, J
\]

then the ratio estimator \( \hat{Y}(0.1:x) \) is unbiased under the broad model (3). Let \( s(J) \) denote any sample satisfying the condition above, that is, \( s(J) \) is any sample for which

\[
X^{(j)} = \bar{X}^{(j)} \quad \text{for } j = 1, \ldots, J
\]

for \( j = 1, \ldots, J \). Such samples are called balanced samples (of order \( J \)) (Royall and Herson 1973). We shall refer to them as simple or unweighted balanced samples, in the light of more general results to be discussed in the next section. The main point here is that by deliberately selecting one's sample to meet the criterion (5), one protects oneself against model failure, at least of a certain (perhaps not uncommon) form.

Now a somewhat subtle observation is in order. Suppose the sampling scheme was simple random sampling (srs). Then both the Hájek and Horvitz-Thompson estimators are the sample mean \( \bar{Y} \). If the sample chosen is balanced (whether deliberately or through the chance result of our sampling scheme), then it is readily seen that the ratio estimator also reduces to the sample mean \( \bar{Y} \). Furthermore, under srs, samples can be seen to be balanced in expectation, that is

\[
E_x(\bar{X}^{(j)}) = n^{-1} \sum_{i=1}^{n} x_i X_i = \bar{X}^{(j)}, \quad j = 1, \ldots, J
\]

where \( E_x(\bullet) \) refers to expectation under the random sampling scheme. Thus the combination of srs and the Hájek/HT estimator would seem to receive added support from model-based theory.

But is this the case? To judge the relative merits of estimators, we might ask about their sensitivity to imbalance. How well does their unbiasedness hold up, if the balance aimed at by randomization is not quite achieved? It is readily shown that the ratio estimator is less sensitive to deviation from balance than the Hájek. For example, suppose the “true model” is \( M(0,1:x) \), that is a quadratic model with intercept. Suppose \( x_1 = (1 + e_1)X \) and \( x_2 = (1 + e_2)X^2 \), where, more often than not the deviations \( e_i \) will be of the same sign. Then one readily shows that the model bias of the Hájek/HT is

\[
E_x(\bar{Y} - \bar{Y}) = \beta_1 \bar{e}_1 + \beta_2 \bar{X}^2 \bar{e}_2
\]

Typically, the second term of the Hájek bias will be larger in absolute value than the corresponding term of the ratio, because of likely cancellation in the latter. Also, where \( Y \) tends to change with \( x \)—the typical circumstance in which we would be tempted to use the ratio estimator—it will be unusual for the first term of the ratio-estimator-bias to be as large in absolute value as the first term of the Hájek-bias. Clearly there are dangers in not having strict balance for either, but the Hájek will usually be more sensitive to imbalance than the ratio estimator. Thus, if one uses the Hájek, it is desirable to take a deliberately balanced sample, and not rely on srs. The frequently made claim that randomization is one’s best protection against model failure does not seem well supported when we look at the matter under the lens of alternate models.

3. Weighted Balanced Samples

By a sampling strategy we shall mean a combination of choice of sample and of estimator. Under the model \( M(0,1:x) \), it is not the case that selecting a balanced sample, and using the ratio estimator, is the most efficient procedure. The greater variance of \( Y \) at larger \( x_i \) dictates we sample units with larger \( x_i \) more heavily than simple balance will allow. The question arises whether there is a procedure that is bias-robust and most efficient under the working model.

We proceed with some degree of generality. Consider the general linear model with a diagonal covariance matrix:

\[
E_x(\bar{Y}) = X\beta, \quad \text{var}_x(\bar{Y}) = \Sigma \sigma^2
\]

which will be referred to as \( M(X:V) \), where \( \beta \) is \( p \times 1 \), \( X \) is \( N \times p \), and \( V \) is \( N \times N \). The matrix of auxiliaries can be partitioned between the sample and non-sample units as \( X = (X_s, X_n) \). The BLU predictor under this model is

\[
\hat{Y}(X:V) = N^{-1} \left( I_n Y_s + 1_n X_n \hat{\beta} \right)
\]

where \( 1_n \) and \( 1 \) are vectors of \( n \) 1's and \( N-n \) 1's, and \( \hat{\beta} = A_n^{-1} X_n^T Y, V_n \) the diagonal covariance matrix for the sample units, and \( A_n = X_n^T V_n X_n \). Let \( \tau(X) \) denote the linear manifold generated by the columns of \( X \), i.e. the vector space spanned by all linear combinations of the columns of \( X \). We will also need \( 1_n \), an \( N \)-vector of 1's.

The collection of samples that satisfy

\[
\frac{1}{n} Y_s W_s V_s X_s = \frac{1}{n} X
\]

where

\[
\frac{1}{n} Y_s W_s V_s X_s = \frac{1}{n} X
\]

761
will be denoted \( B(X; W) \) and is said to be \textit{balanced with respect to the weights root}(W) or to be \textit{root}(W) balanced. Here \( W \) is an \( N \times N \) matrix and \( W \) is the \( n \times n \) submatrix for the sample units. This form of balance turns out to be appropriate when the variance matrix of the model is given by \( V = W \sigma^2 \).

When \( W = I \), \( B(X; I) \) is the set of samples that are balanced on the columns of \( X \), i.e. \( I'X/n = I'X/n \). If the model for \( Y \) is a polynomial in \( x \), as in model (3), then \( B(X; I) \) is the set of samples satisfying \( x^{(j)} = x^{(j)} \), the simple balance conditions introduced in section 2. Thus, weighted balance contains our previous notion of balance as a special case.

**Theorem 1** (Royall 1992). Under \( M(X; V) \) if both \( V_{1_N} \) and \( V^{1/2}1_N \in \mathcal{M}(X) \),

\begin{equation}
\text{var}_M\left[ \hat{Y}(X; V) - \bar{Y} \right] \geq N^{-2}\left[ n^{-1}(1_N'V^{1/2}1_N)^2 - 1_N'V_{1_N} \right] \sigma^2.
\end{equation}

The bound is achieved if and only if \( s \in B(X; V) \), in which case, for \( W^{1/2} = n^{-1}\sum_i^{N} v_i^{1/2} \),

\begin{equation}
\hat{Y}(X; V) = n^{-1}n^{-1}(1_N'V^{1/2}1_N)(1_N'V_{1_N}^{-1}Y)
= \frac{\sigma^{(1/2)}}{n} \sum_i v_i \frac{Y_i}{v_i^{1/2}}.
\end{equation}

Note that, under the weighted balance condition, neither the estimator itself nor its variance depend explicitly on the \( X \) matrix. This has important implications. Suppose the columns of \( X \) are included in those of another matrix \( X' \). Suppose \( M(X; V) \) is the working model, and \( M(X'; V) \) the true model, that the conditions (9) on the standard deviations and variances are met, and that balance holds for the wider model, that is, \( s \in B(X'; V) \); then the estimator based on the working model \( M(X; V) \) will be BLU under \( M(X'; V) \). In other words, the estimator will still be unbiased, and nothing is lost in efficiency.

**Example 1.** Suppose the working model is the quadratic model \( M(1,1,1; x^2) \) with variance proportional to \( x^2 \). The condition \( V^{1/2}1_N \in \mathcal{M}(X) \) is met since both \( x \) and \( x^2 \) are in the model for \( E_u(Y) \). The lower bound on the variance is

\begin{equation}
\frac{1}{N^2} \left[ (N\bar{x})^2/n - \sum_{i=1}^N x_i^2 \right] \sigma^2.
\end{equation}

This bound is achieved in any balanced sample with

\begin{equation}
x_i^{(j)} = x_i^{(j)}/x
\end{equation}

for \( j = 0, 1, \text{and} 2; \) the \( j = 1 \) condition is fulfilled automatically, and (13) at \( j = 0 \) says that the harmonic mean of sample \( x \)'s equals their population arithmetic mean. Bias protection against more general polynomial models is obtained at no cost in efficiency under the working model by balancing on additional powers \( j = 3, 4, ..., J \). We refer to such balance as \( \text{root}(x^2) \) balance or just \( x \)-balance (of order \( J \)). It is also known in the literature as \( \pi \)-balance (Cumberland and Royall 1981). The BLU predictor reduces to \( \hat{\bar{Y}} = \sum_i x_i/(nx_i) \), the "mean-of-ratios estimator" — a result first derived by Kott (1984). It is also the Horvitz-Thompson estimator for a fixed size sampling design with \( \pi_i \propto x_i \). Furthermore, under the balance condition (13) corresponding to \( j = 0 \), the BLU predictor can also be written in the form of a Hájek estimator \( \hat{\bar{Y}} = \sum_i x_i/x_i \). Thus at balance, the

BLU predictor under the model coincides with both the Hájek and Horvitz-Thompson estimators corresponding to pps sampling with size variable \( x \), a sampling plan which gives (13) in design-expectation, that is, \( E_{u}(x_i^{(j)}) = x_i^{(j)}/x \). However, with only minor departures from balance, the three estimators begin to diverge, and can behave quite differently, as we show by analysis in Section 4 and by a simulation study in Section 5.

**Example 2.** Suppose the working model is the through-the-origin model \( M(0,1; x) \) with variance proportional to \( x \), which leads to the ratio estimator. The condition \( V^{1/2}1_N \in \mathcal{M}(X) \) of the Theorem is not met. This suggests (Royall 1992) that as an alternative to the ratio estimator (which is bias robust under unweighted balance) we take \( \text{root}(x) \) balance and the estimator corresponding to the minimal model that meets the conditions of the Theorem.

**Definition:** The minimal model for given variance matrix \( V \) is \( M_{\min}(V) = M(X_{1_N}, V_{1_N}) \), where \( \mathbf{X}_{1_N} = \left( v^{1/2}1_N, V_{1_N} \right) \). Given a particular variance structure, this is the smallest model to guarantee that the conditions (9) of the theorem are met.

With variance proportional to \( x \), the minimal model, given by \( E_u(Y) = \beta_0 x^{1/2} + \beta_1 x \), can be denoted by \( M_{\min}(x) = M(x^{1/2}, x; x) \). The lower bound on the variance is

\[ x_i^{(j)} = x_i^{(j)}/x \]
This bound—which is readily shown to be lower than the variance of the ratio estimator under simple
balance—is achieved under \( M_{\text{min}}(x) \) in any sample balanced in the sense that \( \bar{x}_{i}^{(2)} = \bar{x} / \bar{x}_{i}^{(2)} \). Bias protection against more general polynomial models is
obtained by balancing on additional powers:
\[
\bar{x}_{i}^{(j-2)} = \bar{x}_{i}^{(j)} / \bar{x}_{i}^{(2)} \quad \text{for} \ j = 0, 1, 2, \ldots , J. \quad (15)
\]
As before, the estimator reduces to the Horvitz-Thompson estimator for \( pps \) sampling with \( x^{1/2} \) as size
variable, and also to the Hájek if (15) is met, for \( j = 0 \).

In general, consider a sampling plan having fixed \( n \), \( pps \) sampling with size variable \( v_{i}^{1/2} \), that is,
the inclusion probability of unit \( i \) is
\[
\pi_{i} = n^{1/2} / \bar{v}_{i}^{1/2} \bar{n}^{1/2} \ . \quad (16)
\]
We shall also refer to this plan as \( pps(v^{1/2}) \) sampling. Then
(1) Weighted balance is met in design-expectation, that is \( E_{x} \left( \frac{1}{n} v_{i} I_{x}^{1/2} X_{i} \right) = \frac{1}{n} X_{x} \bar{v}^{1/2} \bar{n}^{1/2} \) for any matrix
\( X \) of regressors (even unknown ones.).
(2) The BLU predictor \( \tilde{Y}(X; V) = N^{-1} n^{-1} \left( \bar{v}_{i}^{1/2} \bar{n}^{1/2} I_{x} \right) \left( \bar{v}_{x}^{1/2} Y_{x} \right) \) in a root(\( v \))
balanced sample is just the Horvitz-Thompson estimator \( \bar{Y}_{H} = N^{-1} \sum_{i=1}^{n} Y_{i} / \pi_{i} \). The variance bound is
the one established by Godambe and Joshi (1965, Theorem 6.1) for the model-based expectation of the
design-based variance of the Horvitz-Thompson estimator.
(3) Under the condition
\[
\frac{1}{n} \bar{v}_{i}^{1/2} \bar{n}^{1/2} I_{x} = \frac{1}{N} I_{x} \bar{v}^{1/2} \bar{n}^{1/2}, \quad (17)
\]
the BLU predictor (11) can be written as the Hájek estimator \( \tilde{Y}_{H} = \sum_{i=1}^{n} \frac{Y_{i}}{\pi_{i}} \).

One method of selecting weighted balanced samples is to use the \( pps \) sampling plan above, but to
reject any sample that is insufficiently balanced on particular moments. Figure 1 depicts a weighted
balanced sample from the Hospitals population used in the simulation study described in section 5. The
population was randomly ordered and systematic \( pp(x^{1/2}) \) samples of size \( n = 30 \) were selected using the
Hartley-Rao (1962) method. The sample in the figure is approximately balanced in the sense of (15) for \( j = 0, 1, \) and 2.

4. Comparison of Estimators in Unbalanced Samples
For simplicity we focus on the situation in Example 1; the basic ideas are very general. The
minimal model when variance is proportional to \( x^{2} \) is \( M(0,1,1; x^{2}) \); the corresponding BLU estimator is, for
any sample (not just a balanced sample)
\[
\tilde{Y}_{BLU} = \frac{1}{n} \left( \bar{x}_{x}^{(2)} \bar{x}^{2} - \bar{x}_{x}^{(2)} \bar{x}^{2} \right) \left( \bar{x}^{(2)} - \bar{x}_{x}^{(2)} \right) \sum_{i=1}^{n} \frac{Y_{i}}{x_{i}} + \left( \bar{x}_{x}^{(2)} \bar{x}^{2} - \bar{x}_{x}^{(2)} \bar{x}_{x}^{2} \right) \sum_{i=1}^{n} Y_{i}
\]
The Hájek and HT estimators under \( pp(x) \) sampling are given in Example 1 above.

Now suppose that sampling yields the following “near-balance” conditions relating sample to
population \( x \)-moments:
\[
\frac{1}{n} \bar{x}_{x}^{(j-2)} = \bar{x}^{(j)} \left( 1 + e_{j} \right),
\]
for \( j = 0, 1, 2, \ldots, J \). The “errors” \( e_{j} \) represent the
distance the sample is from balance. Note that (in the present example) \( e_{0} = 0 \). Typically the \( e_{j} \) with \( j > 0 \)
will tend to have the same sign, and be opposite in
sign from \( e_{0} \).

We consider the bias \( E_{x} (\bar{Y} - \tilde{Y}) \) under a
polynomial model of order \( J \), for the Hájek estimator, Horvitz-Thompson estimator, and minimal model
BLU predictor. We find
\[
\text{Bias}_{H} = \sum_{j=0}^{J} \beta_{j} \bar{x}^{(j)} e_{j} - e_{0}
\]
(18)
(19)
\[
\text{Bias}_{BLU} = \sum_{j=0}^{J} \beta_{j} \left( \bar{x}^{(j-1)} - (\bar{x}^{(2)})^{(j-1)} / \bar{x} \right) e_{j}^{2}
\]
Notice that the multipliers of \( \beta_{0} \) in (18), of \( \beta_{1} \) in
(19), and of \( \beta_{1} \) and \( \beta_{2} \) in (20) are all 0. Thus, if all
coefficients except the intercept are 0, the Hájek can be
expected to be least affected by being away from
balance. This harmonizes with the first reason given
by SSW for preferring it (see above). Where there
exists a continuous non-constant dependency of \( Y \) on \( x \),
however, the HT will be less biased than the Hájek,
because of the likely opposition of sign of \( e_{0} \) and \( e_{j} \)
with \( j > 0 \). Examination of the second expression in
(20) suggests that in each term, both for \( j = 0 \), and \( j > 2, \) some cancellation will take place, so that the
minimal estimator will invariably be less biased than
the Horvitz-Thompson. These conjectures are borne out in a small simulation study, in the next section.

5. Simulation Study

We compared the Hájek, Horvitz-Thompson, and minimal model BLU estimators in a simulation study using three populations. Two of the populations, Hospitals and Cancer, are well-known in the survey literature (Royall and Cumberland 1981). Scatterplots reveal a strong relationship between y and x in both of these populations. The third was generated to be favorable for the Hájek estimator using a model with a common mean: \( y_i = \mu + \epsilon_i x_i \) with \( \epsilon_i \sim (0,1) \), \( \mu = 200 \), and the \( x \)'s coming from the Hospitals population. Two sets of 2,000 samples were selected from each population—one with probabilities proportional to \( x \) and the other with probabilities proportional to \( x \). In both cases, we used \( n=30 \) and the random-order/systematic-sampling method studied by Hartley and Rao (1962).

Table 1 shows the root mean square errors (rmse's) for \( \hat{Y}_{\text{H}} \), \( \hat{Y}_{\text{HT}} \), and two model-based estimators. Each rmse is computed as

\[
\text{rmse}(\hat{Y}) = \sqrt{\frac{1}{S} \sum_{s=1}^{S} (\hat{Y}_s - \bar{Y})^2 / N}
\]

where \( S=2,000 \) and \( \bar{Y} \) is one of the estimates from sample \( s \). For a model with \( \text{var}(y) \propto x' \), the minimal model is \( M_{\text{min}}(x'^2, x':x') \) with BLU predictor \( \hat{Y}_{\text{min}}(x'^2, x':x') \). The second model-based estimator was constructed by adding an intercept to the minimal model: \( \hat{Y}_{\text{min}}(1, x'^2, x':x') \) which is BLU under \( M(1, x'^2, x':x') \). When the model has \( \text{var}(y) \propto x' \), \( pp(x'^2) \) sampling produces, in design-expectation, the type of weighted balance required for optimality in Theorem 1, as noted in section 3.

As Table 1 shows, for the Hospitals and Cancer populations, the Hájek and HT estimators are far worse than either of the model-based estimators in \( pp(x'^2) \) sampling. In contrast, the Hájek is unbiased under the model used to generate the Artificial population and has the smallest rmse there. This finding verifies the analysis of the preceding section regarding the SSW observation that the Hájek will perform well when the \( y \)'s are relatively homogeneous. The minimal estimator \( \hat{Y}_{\text{min}}(x'^2, x':x') \) has the smallest rmse in both types of sampling for Hospitals and Cancer but is substantially worse in the Artificial population. Its poor showing stems from the fact that, except at balance, \( \hat{Y}_{\text{min}}(x'^2, x':x') \) is biased under a model where \( Y \) has a common mean. Note, however, that \( \hat{Y}_{\text{min}}(1, x'^2, x':x') \), which adds an intercept, has rmse's much nearer those of the Hájek in the Artificial population, even though \( \hat{Y}_{\text{min}}(1, x'^2, x':x') \) involves estimates of superfluous parameters for \( x'^2 \) and \( x' \).

The HT estimator does well in \( pp(x) \) samples in the first two populations but poorly otherwise, and, generally, shows itself to be a risky procedure. The source of the differences in the estimators is clarified by conditional analyses. We sorted each set of 2,000 samples by the sample mean \( \bar{x}'(r/2) \), which has design-expectation \( x'(r')/x'(r/2) \) in \( pp(x'^2) \) sampling. The samples were broken into 10 groups of 200 samples each and the bias and rmse calculated in each group for each estimator of the mean. The bias and rmse were plotted against the group average of \( x'(r') \) for the three populations, for both \( pp(x'^2) \) samples and \( pp(x) \) samples, producing figures of the sort laid out in Royall and Cumberland (1981). Because of space limitations of these Proceedings, the

<table>
<thead>
<tr>
<th>Population</th>
<th>( \gamma = 1 )</th>
<th>( \gamma = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitals</td>
<td>116.3</td>
<td>171.9</td>
</tr>
<tr>
<td></td>
<td>51.0</td>
<td>36.9</td>
</tr>
<tr>
<td></td>
<td>37.2</td>
<td>34.7</td>
</tr>
<tr>
<td></td>
<td>39.2</td>
<td>40.0</td>
</tr>
<tr>
<td>Cancer</td>
<td>7.2</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>3.7</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>1.8</td>
<td>1.9</td>
</tr>
<tr>
<td>Artificial</td>
<td>2.5</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>18.5</td>
<td>44.6</td>
</tr>
<tr>
<td></td>
<td>9.9</td>
<td>36.8</td>
</tr>
<tr>
<td></td>
<td>2.7</td>
<td>3.5</td>
</tr>
</tbody>
</table>
figures are not reproduced here; a fuller version of this paper which includes the figures, may be found on the Web at http://stats.bls.gov. In Hospitals and Cancer, the Hájek estimator has an egregious bias that runs from negative to positive over the range of \( \chi_{1}^{2} \), which translates to the large, unconditional rmse’s in Table 1. Only at or near weighted balance is the Hájek estimator unbiased, but, as observed in section 4, minor departures from balance lead to major biases. The Horvitz-Thompson estimator also has a substantial, systematic bias in Hospitals for both methods of sampling and in Cancer for \( pp(x^{2}) \) sampling, though the bias is smaller than that of the Hájek. The minimal estimator has uniformly small bias and rmse throughout the range of \( \chi_{1}^{2} \) in the first two populations.

In the Artificial population, on the other hand, the Hájek does well in all groups of samples as expected. The HT estimator is conditionally quite biased since it makes no allowance for an intercept. The minimal estimator is also conditionally biased in extreme samples though much less so than the Horvitz-Thompson. Examination of individual samples shows that \( \tilde{r}_{1}^{2} \) fits a nonsensical, inverted \( \omega \)-shaped curve to data that follow a horizontal straight-line model, and, thus, is a poor choice. Addition of an intercept term in \( \tilde{r}_{1}^{2} \) largely eliminates this problem. (The curves for \( \tilde{r}_{1}^{2} \) are not shown in Figure 4 to simplify the plots.) It is well to keep in mind, however, that all of these estimators are the same at strict balance on the appropriate moments.

6. Conclusions

The Hájek estimator (and to a lesser extent, the Horvitz-Thompson estimator) which corresponds to a given sampling plan has been shown analytically and empirically to be bias sensitive to deviations of the sample selected from weighted balance, even though the sampling plan achieves weighted balance in expectation. The notable exception is when the variable of interest is unrelated to the auxiliary variable.

It is preferable to use the corresponding minimal estimator. Restricting the sample to be in the class of weighted balanced samples is best. In unbalanced samples, the BLU estimator corresponding to the minimal model augmented by an intercept can be efficient both for \( Y \) variables related to the auxiliary and for \( Y \)'s having a common mean.

Acknowledgments

Any opinions expressed are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

References


