**Examining some aspects of balanced sampling in surveys**

Guillaume Chauvet, David Haziza and Éric Lesage

*ENSAI/IRMAR*
*Département de mathématiques et de statistique, Université de Montréal*
*INSEE*

**Supplementary Material**

# S1 A numerical example

To illustrate the effect of both sources of bias in (25), we conducted a limited simulation study. We generated a finite population $P$ of size $N = 500$, consisting of two variables: an auxiliary variable $x$, whose values were available prior to sampling for all the population units and a characteristic of interest $y$. First, the $x$-values were generated from a normal distribution with mean 0 and variance equal to 1. Then, five of the $x$-values were replaced manually by the values 7.9, 8, 8.2, 8.3 and 8.5. Thus, the distribution of the $x$-values may be viewed as a mixture of two populations with 1% contamination. Then, given the $x$-values, the $y$-values were generated according to

$$y_i = 1 + 2(x_i - \bar{X})^2 + \epsilon_i,$$

where the errors $\epsilon_i$ were generated from a normal distribution with mean and variance equal to 0 and 1, respectively and $\bar{X} = N^{-1} \sum_{i \in P} x_i$. From the population, we generated $K = 10,000$ samples of size 25 according to the rejective procedure of Fuller (2009) using simple random sampling without replacement as the basic procedure; see Example 3.1 for more details. In each sample, we computed the Monte Carlo contribution (in %) to the bias of $\widehat{t}^p_{\text{reg}}$ for each of the two terms on the right hand-side of (25), referred below as term 1 and term 2, respectively, for several values of the balancing tolerance $\gamma$. In addition, for each value of $\gamma$, we computed the percent Monte Carlo relative bias of $\widehat{t}^p_{\text{reg}}$. The results are shown in Table 1.

From Table 1, we note that, for small values of $\gamma$, the contribution of term 2 was small. For example for $\gamma \leq 0.32$, the contribution of Term 2 was ranging from 0% to 5%. This can be explained by the fact that a small value of $\gamma$ corresponds to a high the rejection rate, which in turns

Table 1: Monte Carlo contribution (in %) of each term on the right hand-side of (25) and percent Monte Carlo relative bias of $\widehat{t}^p_{\text{reg}}$

| $\gamma$ | 0.01 | 0.06 | 0.12 | 0.32 | 0.67 | 1.15 | 1.65 | 2.24 | 2.83 |
|---|---|---|---|---|---|---|---|---|---|
| Rejection rate (in %) | 95.0 | 84.0 | 74.0 | 54.0 | 30.0 | 13.0 | 5.0 | 1.0 | 0.3 |
| Contribution (in %) Term 1 | 100 | 100 | 99 | 95 | 77 | 51 | 29 | 9 | 2 |
| Contribution (in %) Term 2 | 0 | 0 | 1 | 5 | 23 | 49 | 71 | 91 | 98 |
| Monte Carlo relative bias of $\widehat{t}^p_{\text{reg}}$ (in %) | -8.8 | -8.1 | -9.1 | -9.1 | -8.8 | -9.9 | -11.2 | -12.4 | -12.3 |

implies that the term $\mathbf{t_z} - \widehat{\mathbf{t}}^p_{\mathbf{z}}$ is close to zero; that is, the resulting estimator does not suffer from a small sample bias. On the other hand, a small value of $\gamma$ tends to lead to inclusion probabilities $\pi_i$ significantly different from the basic inclusion probabilities $p_i$, especially for units exhibiting a large $x$-value. As a result, the contribution of term 1 is expected to be large. In contrast, large values of $\gamma$ correspond to a low rejection rate, in which case there is not much difference between a rejective sampling procedure and a non-rejective one. In this case, the $\pi$'s are expected to be close to the $p_i$'s and the resulting estimator would suffer from small sample bias only; that is, the contribution of term 2 is expected to be large. This is confirmed by the results shown in Table 1.

## S2 Other simulation results

Tables 3-11 below show the Monte Carlo results corresponding to a 50% rejection rate. The results are relatively similar to those obtained for a 90% rejection rate, except for Basic and Fuller. The latter showed some differences, especially in terms of RE. For example, when the $x$-values were generated according to a mixture distribution, Basic showed a value of RE equal to 57% for $n = 25$ when the $y$-values were generated according to the linear model; see Table 4. The same was true, regardless of the distribution of the $x$-value. The differences between Basic and Fuller can be explained by the fact that, with a 50% rejection rate, the term $\mathbf{t_z} - \widehat{\mathbf{t}}^p_{\mathbf{z}}$ on the right

| | | | Linear | | | Quadratic | | | Exponential | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Var. | Mean | Var. | Corr. | Mean | Var. | Corr. | Mean | Var. | Corr. |
| Normal | 1.96 | 1.03 | 1.04 | 5.09 | 0.90 | 3.16 | 8.10 | -0.01 | 4.54 | 28.0 | 0.82 |
| Mixture | 2.08 | 1.41 | 0.96 | 6.65 | 0.92 | 3.89 | 62.0 | -0.39 | 16.5 | 15 900 | 0.53 |
| Lognormal | 1.52 | 2.80 | 1.05 | 12.5 | 0.96 | 6.62 | 6.27 | 0.77 | 5 570 | $8.7\ 10^9$ | 0.47 |
| | | | Bump | | | Anova | | | Logistic | | |
| | | | Mean | Var. | Corr. | Mean | Var. | Corr. | Mean | Var. | Corr. |
| Normal | | | 1.65 | 21.5 | -0.01 | 16.0 | 64.6 | 0.70 | 0.48 | 0.25 | 0.62 |
| Mixture | | | 2.59 | 77.0 | 0.36 | 16.2 | 61.8 | 0.47 | 0.53 | 0.25 | 0.56 |
| Lognormal | | | 5.47 | 647 | 0.76 | 9.18 | 800 | 0.46 | 0.30 | 0.21 | 0.57 |

Table 2: Mean and variance of the auxiliary variable for three distributions, and mean, variance and coefficient of correlation with the auxiliary variable of the variable of interest for six models

hand-side of (24) may have been large for some samples since a low rejection rate corresponds to a large value of the balancing tolerance $\gamma$. In this case, using a calibration strategy such as Fuller generally "pays off" in terms of efficiency.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| Linear | RB | -0 | -0 | -0 | -0 | -0 | -0 | -0 | -1 |
| | CV | 23 | 19 | 24 | 19 | 20 | 19 | 42 | 19 |
| | RE | 63 | 100 | 62 | 100 | 85 | 98 | 20 | 95 |
| Quadratic | RB | -2 | -2 | -0 | -1 | -0 | -0 | -0 | -4 |
| | CV | 17 | 18 | 18 | 18 | 18 | 18 | 18 | 19 |
| | RE | 103 | 100 | 93 | 95 | 95 | 95 | 99 | 84 |
| Exponential | RB | -1 | -1 | 0 | -0 | -0 | -0 | 0 | -3 |
| | CV | 15 | 13 | 16 | 14 | 14 | 13 | 23 | 14 |
| | RE | 79 | 100 | 69 | 93 | 87 | 95 | 32 | 88 |
| Bump | RB | -5 | -7 | -1 | -3 | -0 | -1 | -0 | -12 |
| | CV | 55 | 56 | 56 | 56 | 56 | 56 | 56 | 59 |
| | RE | 102 | 100 | 100 | 99 | 101 | 101 | 101 | 85 |
| Stratification | RB | 0 | 1 | -0 | 0 | 0 | 0 | 0 | 1 |
| | CV | 4 | 8 | -2 | 2 | 0 | 1 | 1 | 16 |
| | RE | 90 | 100 | 95 | 100 | 95 | 98 | 52 | 88 |
| Logistic | RB | -0 | -0 | -0 | -0 | 0 | 0 | -0 | -0 |
| | CV | 17 | 16 | 17 | 16 | 16 | 16 | 20 | 16 |
| | RE | 92 | 100 | 93 | 101 | 101 | 103 | 63 | 97 |

Table 3: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 25$, with a rejection rate equal to 50% and for a normal distribution of $x$.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| Linear | RB | -1 | -0 | 0 | 0 | 0 | -0 | 0 | -0 |
| | CV | 28 | 21 | 31 | 21 | 26 | 21 | 53 | 21 |
| | RE | 57 | 100 | 45 | 100 | 67 | 100 | 16 | 101 |
| Quadratic | RB | -8 | -10 | -0 | -6 | 0 | -2 | -0 | -14 |
| | CV | 34 | 33 | 49 | 39 | 39 | 37 | 39 | 32 |
| | RE | 97 | 100 | 50 | 75 | 77 | 88 | 76 | 95 |
| Exponential | RB | -25 | -29 | 0 | -20 | 2 | -8 | 0 | -38 |
| | CV | 115 | 108 | 177 | 130 | 139 | 124 | 148 | 101 |
| | RE | 90 | 100 | 40 | 72 | 65 | 81 | 57 | 107 |
| Bump | RB | -13 | -16 | -0 | -10 | 1 | -4 | -0 | -23 |
| | CV | 59 | 58 | 80 | 66 | 66 | 63 | 66 | 58 |
| | RE | 99 | 100 | 57 | 80 | 81 | 90 | 82 | 91 |
| Stratification | RB | 1 | 2 | -0 | 1 | 0 | 0 | 0 | 2 |
| | CV | 14 | 20 | -0 | 11 | 1 | 6 | 1 | 29 |
| | RE | 97 | 100 | 101 | 89 | 95 | 94 | 76 | 91 |
| Logistic | RB | 1 | 1 | 0 | 1 | -0 | 0 | 0 | 1 |
| | CV | 16 | 15 | 15 | 15 | 15 | 15 | 18 | 16 |
| | RE | 94 | 100 | 99 | 99 | 99 | 100 | 70 | 97 |

Table 4: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 25$, with a rejection rate equal to 50% and for a mixture distribution of $x$.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| Linear | RB | -3 | -0 | -0 | -0 | 0 | 0 | -0 | -0 |
| | CV | 30 | 19 | 51 | 19 | 33 | 19 | 65 | 19 |
| | RE | 38 | 100 | 13 | 100 | 32 | 98 | 8 | 94 |
| Quadratic | RB | -15 | -17 | 0 | -14 | -0 | -8 | -0 | -21 |
| | CV | 42 | 36 | 108 | 39 | 60 | 41 | 73 | 37 |
| | RE | 79 | 100 | 13 | 91 | 43 | 89 | 29 | 85 |
| Exponential | RB | -66 | -72 | -0 | -69 | -1 | -39 | -3 | -67 |
| | CV | 188 | 156 | 574 | 171 | 319 | 207 | 320 | 152 |
| | RE | 74 | 100 | 9 | 87 | 29 | 67 | 29 | 106 |
| Bump | RB | -18 | -21 | 0 | -18 | -0 | -10 | -0 | -27 |
| | CV | 53 | 46 | 132 | 49 | 75 | 52 | 90 | 49 |
| | RE | 81 | 100 | 15 | 92 | 45 | 90 | 31 | 81 |
| Stratification | RB | 3 | 5 | 0 | 4 | 0 | 3 | 0 | 9 |
| | CV | 17 | 28 | 1 | 20 | 0 | 13 | 0 | 40 |
| | RE | 108 | 100 | 116 | 96 | 117 | 103 | 90 | 73 |
| Logistic | RB | 2 | 4 | -0 | 3 | -0 | 2 | -0 | 5 |
| | CV | 25 | 25 | 24 | 25 | 24 | 25 | 30 | 26 |
| | RE | 99 | 100 | 107 | 100 | 104 | 101 | 70 | 90 |

Table 5: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 25$, with a rejection rate equal to 50% and for a log-normal distribution of $x$.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| Linear | RB | 0 | -0 | 0 | 0 | -0 | -0 | 0 | -0 |
| | CV | 16 | 13 | 16 | 13 | 14 | 13 | 29 | 13 |
| | RE | 64 | 100 | 64 | 100 | 91 | 100 | 20 | 100 |
| Quadratic | RB | -1 | -1 | 0 | -0 | -0 | -0 | 0 | -2 |
| | CV | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 13 |
| | RE | 101 | 100 | 97 | 98 | 96 | 96 | 99 | 91 |
| Exponential | RB | -0 | -1 | 0 | -0 | -0 | -0 | 0 | -1 |
| | CV | 10 | 9 | 11 | 9 | 10 | 9 | 16 | 9 |
| | RE | 78 | 100 | 73 | 97 | 89 | 94 | 33 | 96 |
| Bump | RB | -2 | -3 | 0 | -1 | -1 | -1 | 0 | -5 |
| | CV | 38 | 38 | 38 | 38 | 38 | 38 | 38 | 39 |
| | RE | 101 | 100 | 100 | 99 | 100 | 100 | 101 | 93 |
| Stratification | RB | 0 | 0 | -0 | 0 | 0 | 0 | -0 | 1 |
| | CV | 3 | 5 | -1 | 1 | 0 | 1 | -0 | 10 |
| | RE | 89 | 100 | 91 | 100 | 97 | 98 | 51 | 93 |
| Logistic | RB | 0 | 0 | 0 | 0 | -0 | -0 | -0 | -0 |
| | CV | 11 | 11 | 11 | 11 | 11 | 11 | 14 | 11 |
| | RE | 92 | 100 | 93 | 101 | 97 | 98 | 61 | 98 |

Table 6: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 50$, with a rejection rate equal to 50% and for a normal distribution of $x$.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| Linear | RB | -0 | -0 | 0 | 0 | 0 | 0 | 0 | -0 |
| | CV | 19 | 14 | 19 | 14 | 16 | 14 | 36 | 14 |
| | RE | 58 | 100 | 54 | 100 | 80 | 103 | 16 | 99 |
| Quadratic | RB | -3 | -4 | 0 | -2 | 0 | -1 | 1 | -7 |
| | CV | 25 | 25 | 29 | 27 | 26 | 26 | 27 | 25 |
| | RE | 99 | 100 | 76 | 85 | 91 | 96 | 85 | 94 |
| Exponential | RB | -10 | -13 | 1 | -6 | 0 | -3 | 1 | -22 |
| | CV | 88 | 85 | 104 | 95 | 93 | 89 | 103 | 82 |
| | RE | 95 | 100 | 69 | 83 | 86 | 94 | 70 | 103 |
| Bump | RB | -5 | -7 | 1 | -3 | 0 | -1 | 1 | -12 |
| | CV | 43 | 42 | 48 | 46 | 45 | 44 | 46 | 43 |
| | RE | 100 | 100 | 80 | 87 | 92 | 96 | 89 | 93 |
| Stratification | RB | 0 | 1 | -0 | 0 | -0 | 0 | -0 | 1 |
| | CV | 8 | 12 | -1 | 5 | -0 | 2 | -1 | 22 |
| | RE | 97 | 100 | 99 | 93 | 97 | 96 | 77 | 94 |
| Logistic | RB | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | CV | 11 | 10 | 11 | 10 | 11 | 11 | 13 | 11 |
| | RE | 93 | 100 | 95 | 99 | 98 | 98 | 67 | 97 |

Table 7: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 50$, with a rejection rate equal to 50% and for a mixture distribution of $x$.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| Linear | RB | -1 | 0 | 0 | 0 | 0 | 0 | -0 | 0 |
| | CV | 21 | 13 | 24 | 13 | 19 | 13 | 45 | 13 |
| | RE | 38 | 100 | 30 | 100 | 46 | 100 | 8 | 97 |
| Quadratic | RB | -7 | -8 | 0 | -5 | -0 | -3 | -0 | -12 |
| | CV | 34 | 31 | 49 | 34 | 39 | 32 | 51 | 30 |
| | RE | 83 | 100 | 42 | 87 | 69 | 100 | 39 | 97 |
| Exponential | RB | -31 | -37 | 1 | -29 | -0 | -17 | -1 | -48 |
| | CV | 184 | 172 | 274 | 193 | 218 | 187 | 225 | 149 |
| | RE | 89 | 100 | 41 | 82 | 65 | 88 | 61 | 127 |
| Bump | RB | -8 | -10 | 0 | -7 | -0 | -4 | -0 | -15 |
| | CV | 43 | 39 | 60 | 42 | 48 | 40 | 62 | 38 |
| | RE | 84 | 100 | 44 | 87 | 71 | 100 | 41 | 96 |
| Stratification | RB | 1 | 2 | 0 | 1 | -0 | 1 | -0 | 4 |
| | CV | 11 | 17 | 1 | 11 | -0 | 6 | -1 | 28 |
| | RE | 103 | 100 | 107 | 96 | 113 | 106 | 85 | 83 |
| Logistic | RB | 1 | 2 | 0 | 1 | 0 | 1 | -0 | 3 |
| | CV | 17 | 17 | 17 | 17 | 17 | 17 | 21 | 17 |
| | RE | 96 | 100 | 100 | 99 | 104 | 103 | 70 | 95 |

Table 8: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 50$, with a rejection rate equal to 50% and for a log-normal distribution of $x$.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| Linear | RB | -0 | -0 | 0 | 0 | -0 | -0 | 0 | -0 |
| | CV | 11 | 9 | 11 | 9 | 9 | 9 | 19 | 9 |
| | RE | 66 | 100 | 65 | 100 | 100 | 105 | 20 | 102 |
| Quadratic | RB | -0 | -0 | -0 | -0 | -0 | -0 | 0 | -1 |
| | CV | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| | RE | 101 | 100 | 99 | 99 | 102 | 102 | 103 | 98 |
| Exponential | RB | -0 | -0 | -0 | -0 | -0 | -0 | 0 | -1 |
| | CV | 7 | 6 | 7 | 6 | 6 | 6 | 10 | 6 |
| | RE | 79 | 100 | 77 | 99 | 99 | 102 | 35 | 100 |
| Bump | RB | -1 | -1 | -0 | -0 | 0 | 0 | 0 | -2 |
| | CV | 25 | 25 | 25 | 25 | 25 | 25 | 25 | 26 |
| | RE | 101 | 100 | 100 | 100 | 103 | 103 | 101 | 97 |
| Stratification | RB | 0 | 0 | -0 | 0 | 0 | 0 | -0 | 0 |
| | CV | 1 | 2 | -1 | 0 | 0 | 1 | -1 | 6 |
| | RE | 89 | 100 | 90 | 100 | 99 | 100 | 52 | 99 |
| Logistic | RB | -0 | -0 | -0 | -0 | -0 | -0 | -0 | -0 |
| | CV | 8 | 7 | 8 | 7 | 7 | 7 | 9 | 7 |
| | RE | 92 | 100 | 92 | 100 | 100 | 101 | 63 | 102 |

Table 9: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 100$, with a rejection rate equal to 50% and for a normal distribution of $x$.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| Linear | RB | -0 | -0 | 0 | -0 | 0 | 0 | 0 | 0 |
| | CV | 13 | 10 | 13 | 10 | 10 | 10 | 24 | 10 |
| | RE | 57 | 100 | 56 | 100 | 87 | 100 | 16 | 98 |
| Quadratic | RB | -1 | -2 | -0 | -1 | 0 | -0 | -0 | -4 |
| | CV | 17 | 17 | 18 | 18 | 17 | 17 | 18 | 17 |
| | RE | 99 | 100 | 90 | 94 | 97 | 99 | 89 | 93 |
| Exponential | RB | -5 | -6 | -0 | -3 | 0 | -1 | -0 | -11 |
| | CV | 60 | 59 | 64 | 62 | 61 | 60 | 68 | 60 |
| | RE | 96 | 100 | 85 | 92 | 95 | 98 | 75 | 95 |
| Bump | RB | -2 | -3 | -0 | -2 | 0 | -0 | -0 | -6 |
| | CV | 29 | 29 | 30 | 30 | 29 | 29 | 30 | 30 |
| | RE | 100 | 100 | 92 | 95 | 98 | 100 | 92 | 93 |
| Stratification | RB | 0 | 0 | 0 | 0 | -0 | -0 | 0 | 1 |
| | CV | 7 | 9 | 1 | 4 | -1 | -0 | 0 | 16 |
| | RE | 97 | 100 | 97 | 97 | 100 | 98 | 78 | 95 |
| Logistic | RB | 0 | 0 | -0 | 0 | 0 | 0 | -0 | 0 |
| | CV | 7 | 7 | 7 | 7 | 7 | 7 | 8 | 7 |
| | RE | 94 | 100 | 95 | 100 | 103 | 103 | 69 | 99 |

Table 10: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 100$, with a rejection rate equal to 50% and for a mixture distribution of $x$.

| Model | | Basic | *Fuller* | MC | MC-Reg | Cube | Cube-Reg | SRS | SRS-Reg |
|---|---|---|---|---|---|---|---|---|---|
| | RB | -0 | -0 | 0 | -0 | -0 | -0 | 0 | 0 |
| Linear | CV | 14 | 9 | 14 | 9 | 11 | 9 | 30 | 9 |
| | RE | 39 | 100 | 36 | 100 | 61 | 102 | 8 | 100 |
| | RB | -2 | -3 | 0 | -2 | -0 | -1 | 0 | -6 |
| Quadratic | CV | 24 | 22 | 27 | 23 | 24 | 21 | 34 | 22 |
| | RE | 84 | 100 | 67 | 91 | 88 | 107 | 43 | 93 |
| | RB | -10 | -14 | 2 | -8 | -2 | -8 | -1 | -28 |
| Exponential | CV | 138 | 134 | 157 | 144 | 141 | 134 | 149 | 125 |
| | RE | 96 | 100 | 74 | 88 | 92 | 101 | 82 | 111 |
| | RB | -2 | -4 | 1 | -2 | -0 | -2 | 0 | -8 |
| Bump | CV | 30 | 28 | 34 | 29 | 29 | 27 | 42 | 28 |
| | RE | 85 | 100 | 68 | 92 | 89 | 107 | 45 | 92 |
| | RB | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| Stratification | CV | 6 | 9 | 1 | 5 | 0 | 2 | 2 | 20 |
| | RE | 99 | 100 | 100 | 98 | 110 | 107 | 81 | 88 |
| | RB | 0 | 1 | 0 | 0 | 0 | 0 | -0 | 1 |
| Logistic | CV | 12 | 11 | 11 | 11 | 11 | 11 | 14 | 11 |
| | RE | 95 | 100 | 96 | 99 | 101 | 101 | 68 | 96 |

Table 11: Monte Carlo percent relative bias, percent coefficient of variation, percent bias ratio and percent relative efficiency of several estimators under three sampling designs of size $n = 100$, with a rejection rate equal to 50% and for a log-normal distribution of $x$.

## S3 Proof of equation (30)

We first compute an expansion of $pr(-\gamma \le X \le \gamma) = F_X(\gamma) - F_X(-\gamma)$. From the formal Edgeworth expansion (see Thompson, 1997, equation (3.41)), we get

$$\begin{aligned} F_X(\gamma) &= \psi(\gamma) - \phi(\gamma)\left\{\frac{\kappa_3}{6}(\gamma^2-1)\right\} \\ &\quad - \phi(\gamma)\left\{\frac{\kappa_4}{24}(\gamma^3-3\gamma) + \frac{\kappa_3^2}{72}(\gamma^5 - 10\gamma^3 + 15\gamma)\right\} + o(d^{-1}), \\ F_X(-\gamma) &= 1 - \psi(\gamma) - \phi(\gamma)\left\{\frac{\kappa_3}{6}(\gamma^2-1)\right\} \qquad \text{(S3.1)} \\ &\quad + \phi(\gamma)\left\{\frac{\kappa_4}{12}(\gamma^3-3\gamma) + \frac{\kappa_3^2}{36}(\gamma^5 - 10\gamma^3 + 15\gamma)\right\} + o(d^{-1}), \end{aligned}$$

where $\psi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and the probability density function of a standard normal distribution with

$$\begin{aligned} \kappa_3 &\equiv \mu_3(X) = d^{-3/2}\sum_{i\in P}\tilde{x}_i^3\ p_i(1-p_i)(1-2p_i), \\ \kappa_4 &\equiv \mu_4(X) - 3\{\mu_2(X)\}^2 = d^{-2}\sum_{i\in P}\tilde{x}_i^4\ p_i(1-p_i)\{1-6p_i(1-p_i)\}, \end{aligned}$$

and $\mu_m(X)$ denoting the centered moment of order $m$ of the random variable $X$. Equations (S3.1) lead to

$$\begin{aligned} pr(-\gamma \le X \le \gamma) &= \{2\psi(\gamma) - 1\} \qquad \text{(S3.2)} \\ &\quad -\phi(\gamma)\left\{\frac{\kappa_4}{12}(\gamma^3-3\gamma) + \frac{\kappa_3^2}{36}(\gamma^5 - 10\gamma^3 + 15\gamma)\right\} + o(d^{-1}). \end{aligned}$$

We now obtain an expansion of $pr(-\gamma \le X \le \gamma | I_{bi} = 1)$. We have

$$pr(X \le \gamma | I_{bi} = 1) = pr(X_i \le \gamma_i),$$

where

$$X_i = \frac{1}{\sqrt{d - \tilde{x}_i^2\ p_i(1-p_i)}}\sum_{j\neq i\in U}\tilde{x}_j(I_{bj} - p_j)$$

and

$$\gamma_i = \frac{1}{\sqrt{1 - d^{-1}\tilde{x}_i^2\ p_i(1-p_i)}}\left\{\gamma - d^{-1/2}\tilde{x}_i\ (1-p_i)\right\}.$$

Applying once again the formal Edgeworth expansion to $X_i$ leads to

$$\begin{aligned} F_{X_i}(\gamma_i) &= \psi(\gamma_i) - \phi(\gamma_i)\left\{\frac{\kappa_{3i}}{6}(\gamma_i^2-1)\right\} \\ &\quad - \phi(\gamma_i)\left\{\frac{\kappa_{4i}}{24}(\gamma_i^3-3\gamma_i) + \frac{\kappa_{3i}^2}{72}(\gamma_i^5-10\gamma_i^3+15\gamma_i)\right\} + o(d^{-1}), \end{aligned} \tag{S3.3}$$

where

$$\kappa_{3i} \equiv \mu_3(X_i) = \left\{d - \tilde{x}_i^2\ p_i(1-p_i)\right\}^{-3/2} \sum_{j\neq i\in P} \tilde{x}_j^3\ p_j(1-p_j)(1-2p_j)$$

and

$$\kappa_{4i} \equiv \mu_4(X_i) - 3\{\mu_2(X_i)\}^2 = \left\{d - \tilde{x}_i^2\ p_i(1-p_i)\right\}^{-2} \sum_{j\neq i\in P} \tilde{x}_j^4\ p_j(1-p_j)\{1-6p_j(1-p_j)\}.$$

By neglecting terms of smaller order than $d^{-1}$, we obtain after some algebra

$$\begin{aligned} F_{X_i}(\gamma_i) &= \psi(\gamma) - \phi(\gamma)\left\{\frac{1}{\sqrt{d}}\tilde{x}_i(1-p_i) + \frac{1}{6}\kappa_3(\gamma^2-1)\right\} \\ &\quad + \left[\frac{1}{2d}\tilde{x}_i^2(1-p_i)\left(\gamma p_i\phi(\gamma) + (1-p_i)\phi'(\gamma)\right)\right. \\ &\qquad + \frac{\kappa_3\gamma\phi(\gamma)}{3\sqrt{d}}\tilde{x}_i(1-p_i) - \frac{\kappa_4\phi(\gamma)}{24}(\gamma^3-3\gamma) \\ &\qquad \left. -\phi(\gamma)\frac{\kappa_3^2}{72}(\gamma^5-10\gamma^3+15\gamma) + \frac{1}{6}\frac{\kappa_3(\gamma^2-1)}{\sqrt{d}}\tilde{x}_i(1-p_i)\phi'(\gamma)\right] \\ &+ o(d^{-1}). \end{aligned} \tag{S3.4}$$

Also, $pr(X \leq -\gamma | I_{bi} = 1) = F_{X_i}(-\tilde{\gamma}_i)$, where

$$\tilde{\gamma}_i = \frac{1}{\sqrt{1 - d^{-1}\tilde{x}_i^2\ p_i(1-p_i)}}\left\{\gamma + d^{-1/2}\tilde{x}_i\ (1-p_i)\right\}.$$

Similar arguments lead to

$$\begin{aligned} F_{X_i}(-\tilde{\gamma}_i) &= \{1-\psi(\gamma)\} - \phi(\gamma)\left\{\frac{1}{\sqrt{d}}\tilde{x}_i(1-p_i) + \frac{1}{6}\kappa_3(\gamma^2-1)\right\} \\ &\quad + \left[\frac{1}{2d}\tilde{x}_i^2(1-p_i)\left\{-\gamma p_i\phi(\gamma) - (1-p_i)\phi'(\gamma)\right\}\right. \\ &\quad - \frac{\kappa_3\gamma\phi(\gamma)}{3\sqrt{d}}\tilde{x}_i(1-p_i) + \frac{\kappa_4\phi(\gamma)}{24}(\gamma^3-3\gamma) \\ &\quad \left. +\phi(\gamma)\frac{\kappa_3^2}{72}(\gamma^5-10\gamma^3+15\gamma) - \frac{1}{6}\frac{\kappa_3(\gamma^2-1)}{\sqrt{d}}\tilde{x}_i(1-p_i)\phi'(\gamma)\right] \\ &+ o(d^{-1}). \end{aligned} \tag{S3.5}$$

We obtain

$$\begin{aligned} pr(-\gamma \leq X \leq \gamma | I_i = 1) &= \{2\psi(\gamma)-1\} + \frac{1}{d}\tilde{x}_i^2(1-p_i)\left\{\gamma p_i\phi(\gamma) + (1-p_i)\phi'(\gamma)\right\} \\ &+ 2\left\{\frac{\kappa_3\gamma\phi(\gamma)}{3\sqrt{d}}\tilde{x}_i(1-p_i) - \frac{\kappa_4\phi(\gamma)}{24}(\gamma^3-3\gamma)\right. \\ &\quad \left. -\phi(\gamma)\frac{\kappa_3^2}{72}(\gamma^5-10\gamma^3+15\gamma) + \frac{1}{6}\frac{\kappa_3(\gamma^2-1)}{\sqrt{d}}\tilde{x}_i(1-p_i)\phi'(\gamma)\right\} \\ &+ o(d^{-1}). \end{aligned} \tag{S3.6}$$

Plugging (S3.2) and (S3.6) into (29) and making use of the identity $\phi'(\gamma) = -\gamma\phi(\gamma)$, we obtain (30).

## References

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika* **96**, 933–944.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman and Hall.