

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2026): 4.49
Maths 2026; 11(1): 154-158
© 2026 Stats & Maths
<https://www.mathsjournal.com>
Received: 02-11-2025
Accepted: 17-12-2025

Elena Markovic
Department of Statistics,
Belgrade Institute of
Mathematical Sciences,
Belgrade, Serbia

Stefan Petrovic
Department of Statistics,
Belgrade Institute of
Mathematical Sciences,
Belgrade, Serbia

Milica Jovanovic
Department of Statistics,
Belgrade Institute of
Mathematical Sciences,
Belgrade, Serbia

Nikola Radovic
Department of Statistics,
Belgrade Institute of
Mathematical Sciences,
Belgrade, Serbia

Corresponding Author:
Elena Markovic
Department of Statistics,
Belgrade Institute of
Mathematical Sciences,
Belgrade, Serbia

A modified exponential-type estimator for population mean using auxiliary information under stratified random sampling

Elena Markovic, Stefan Petrovic, Milica Jovanovic and Nikola Radovic

DOI: <https://www.doi.org/10.22271/math.2026.v11.i1b.2609>

Abstract

Estimating a population mean more precisely by using an auxiliary variable correlated with the variable of interest is a long-standing idea in survey sampling, and ratio, product and exponential-type estimators are the most widely used devices for doing so. Bahl and Tuteja's exponential ratio-type estimator is a classical member of this class; it has since been extended to stratified random sampling and generalised, in a substantial later literature, by inserting a known population constant into its exponent, giving a one-parameter family that also contains the linear regression estimator as a limiting member. This paper works through that family for stratified random sampling; the bias and mean squared error (MSE) are derived to the first order of approximation, the family member that minimises MSE is identified in closed form, and it is shown to attain exactly the efficiency of the linear regression estimator. The practically important question is then how close a real, easily available known constant gets to that optimum. Two choices are examined: the population coefficient of variation of the auxiliary variable, the constant most commonly used in the published literature on this family, and a constant derived from a previously available regression slope between the study and auxiliary variables. For a simulated four-stratum population with an auxiliary variable on a typical business-survey scale, the coefficient-of-variation choice is shown, both theoretically and by a 20,000-replication Monte Carlo simulation, to differ negligibly from the classical Bahl-Tuteja estimator, because the constant it introduces is of order one while the quantity it is added to is of order one hundred; the regression-slope-based constant, by contrast, is of the right order of magnitude and recovers almost all of the efficiency gap to the regression estimator. Percent relative efficiencies from the simulation agree with the theoretical values to within about two percent. The results indicate that the choice of auxiliary constant in this family should be judged by how well its magnitude matches that of the auxiliary variable's mean, not merely by whether the constant is known, and this is easy to check before an estimator from this family is adopted for a given survey.

Keywords: Stratified random sampling, auxiliary information, exponential-type estimator, mean squared error, Neyman allocation, percent relative efficiency, regression estimator, Monte Carlo simulation

Introduction

Using an auxiliary variable that is correlated with the variable of real interest is one of the oldest ways of making a survey estimate more precise, and ratio, product and linear regression estimators are the classical tools for doing it. Bahl and Tuteja proposed an exponential ratio-type estimator that wraps the usual ratio of sample to population auxiliary means inside an exponential function; because the exponent is small when the sample is well behaved, the estimator tracks the ratio estimator closely while typically carrying a smaller mean squared error (MSE). The estimator was later extended from simple random sampling to stratified random sampling, and a further, quite large literature generalised it again by inserting a known constant—the auxiliary variable's coefficient of variation, its correlation with the study variable, or its coefficient of kurtosis, among others—into the exponent, producing a family of “modified” or “improved” estimators.

Almost every paper in that generalised-constant tradition follows the same pattern: a new constant is proposed, the bias and MSE are re-derived, and the new estimator is shown to beat

its predecessors on one or two example populations, often the same widely reused agricultural or forestry data sets. What is less often asked is a more basic question: does inserting a known constant of a given size actually move the estimator toward the best member of the family it belongs to, or does the improvement (or lack of it) depend on how that constant's numerical size compares with the auxiliary variable's own mean? Because the constant enters additively alongside the auxiliary population mean, a constant that is small relative to that mean can only nudge the estimator's effective behaviour by a small amount, however cleverly it is chosen.

This paper addresses that question directly for the stratified-sampling case. Section 2 sets out the sampling framework and the estimator family. Section 3 derives the bias and MSE to the first order of approximation, obtains the MSE-minimising member of the family in closed form, and shows that it coincides with the linear regression estimator's efficiency. Section 4 describes a simulation built around a synthetic four-stratum population whose auxiliary variable has a mean of the order of one hundred, a scale that is typical of many real business or agricultural surveys but far from the order-one scale of a coefficient of variation. Section 5 reports, both analytically and from the simulation, how far the coefficient-of-variation member of the family falls short of the optimum in this setting, and how a constant scaled to match the auxiliary mean closes most of that gap.

Sampling Framework and Estimators

Consider a population of N units divided into L strata, with N_h units in stratum h , $W_h = N_h/N$. A stratified random sample without replacement of size n_h is drawn independently from each stratum, with total sample size $n = \sum n_h$. Let y_{hi} and x_{hi} denote the study and auxiliary variable values for unit i in stratum h , with stratum means $\bar{y}_h$, $\bar{x}_h$ and population means $\bar{Y} = \sum W_h \bar{y}_h$, $\bar{X} = \sum W_h \bar{x}_h$. The auxiliary variable's population mean $\bar{X}$, and the other auxiliary population parameters used below, are assumed known or available from the sampling frame, a standard assumption in this literature. The usual unbiased estimators of the population means are $\bar{y}_{st} = \sum W_h \bar{y}_h$ and $\bar{x}_{st} = \sum W_h \bar{x}_h$.

Three classical estimators of $\bar{Y}$ serve as benchmarks: the usual stratified mean $\bar{y}_{st}$ itself; the classical ratio estimator $\bar{y}_R = \bar{y}_{st}(\bar{X}/\bar{x}_{st})$; and the linear regression estimator $\bar{y}_{reg} = \bar{y}_{st} + b(\bar{X} - \bar{x}_{st})$, where b is a regression coefficient assumed known or estimated with negligible extra variance from a previous survey. The exponential-type estimator of Bahl and Tuteja, extended to stratified sampling, is

$$t(k) = \bar{y}_{st} \cdot \exp[(\bar{X} - \bar{x}_{st}) / (\bar{X} + \bar{x}_{st} + 2k)],$$

where k is a constant taken from known population information about the auxiliary variable. Setting $k=0$ recovers the original Bahl-Tuteja stratified exponential estimator. It is convenient to reparametrise the family by $\lambda = \bar{X}/[2(\bar{X} + k)]$, so that $t(k)$ becomes a one-parameter family indexed by λ , with $\lambda=1/2$ corresponding to $k=0$. Because $d\lambda/dk = -1/[2(\bar{X} + k)]^2$, a unit change in k moves λ by an amount of order $1/\bar{X}$; consequently, a constant k that is numerically small compared with $\bar{X}$ —as a coefficient of variation, a correlation coefficient, or a coefficient of kurtosis typically is—can only move λ a small distance away from one half, regardless of how that constant is chosen substantively. This observation, which follows directly from the definition of the family, motivates the comparison carried out in this paper. Sample sizes n_h are allocated across strata by Neyman allocation,

$n_h = n \cdot N_h S_{yh} / \sum (N_h S_{yh})$, which minimises the variance of $\bar{y}_{st}$ for a fixed total sample size.

Theoretical Derivation of Bias and MSE

Write $e_0 = (\bar{y}_{st} - \bar{Y})/\bar{Y}$ and $e_1 = (\bar{x}_{st} - \bar{X})/\bar{X}$, so that $E(e_0) = E(e_1) = 0$ and, to the order needed here,

$$E(e_0^2) = \sum W_h^2 \theta_h S_{yh}^2 / \bar{Y}^2, \quad E(e_1^2) = \sum W_h^2 \theta_h S_{xh}^2 / \bar{X}^2, \\ E(e_0 e_1) = \sum W_h^2 \theta_h S_{xyh} / (\bar{Y} \bar{X}),$$

where $\theta_h = 1/n_h - 1/N_h$, and S_{yh}^2 , S_{xh}^2 , S_{xyh} are the stratum population variances and covariance. Writing $\bar{x}_{st} = \bar{X}(1 + e_1)$ and expanding $t(k) = \bar{Y}(1 + e_0) \cdot \exp[-\lambda e_1 / (1 + \lambda e_1)]$ as a series in e_1 and collecting terms up to combined second order gives

$$t(k) \approx \bar{Y} [1 + e_0 - \lambda e_1 - \lambda e_0 e_1 + (3/2)\lambda^2 e_1^2].$$

Taking expectations, and retaining only the terms of order $1/n$, the bias of $t(k)$ to the first order of approximation is

$$\text{Bias}[t(k)] \approx \bar{Y} [(3/2)\lambda^2 E(e_1^2) - \lambda E(e_0 e_1)],$$

which is itself $O(1/n)$ and so is small relative to the estimator's standard error for moderate sample sizes; the bias values reported in Section 5 confirm this. For the MSE, only the first-order term ($e_0 - \lambda e_1$) needs to be kept, since squaring the retained quadratic terms would contribute at $O(1/n^2)$. This gives

$$\text{MSE}[t(k)] \approx \sum W_h^2 \theta_h [\bar{Y}^2 S_{yh}^2 - 2\lambda R \cdot S_{xyh} + \lambda^2 R^2 \cdot S_{xh}^2], \\ R = \bar{Y}/\bar{X},$$

a quadratic function of λ with a positive leading coefficient (assuming $R^2 \sum W_h^2 \theta_h S_{xh}^2 > 0$), so it has a unique minimum. Differentiating with respect to λ and setting the result to zero gives the MSE-minimising value

$$\lambda_{\text{opt}} = \sum W_h^2 \theta_h S_{xyh} / (R \cdot \sum W_h^2 \theta_h S_{xh}^2),$$

and substituting back gives the minimum attainable MSE within the family,

$$\text{MSE}_{\text{min}} = \sum W_h^2 \theta_h S_{yh}^2 - (\sum W_h^2 \theta_h S_{xyh})^2 / \sum W_h^2 \theta_h S_{xh}^2,$$

which is identical to the first-order MSE of the linear regression estimator with its own optimal coefficient $b^* = \sum W_h^2 \theta_h S_{xyh} / \sum W_h^2 \theta_h S_{xh}^2$. In other words, whatever specific known constant k is used to define a practical member of this exponential-type family, the family as a whole cannot do better, to the first order of approximation, than the regression estimator; the best it can do is match it, and only if λ (equivalently, k) happens to equal λ_{opt} . Achieving λ_{opt} in practice requires knowing, or having a good pre-estimate of, the pooled quantity $\sum W_h^2 \theta_h S_{xyh} / \sum W_h^2 \theta_h S_{xh}^2$ essentially a regression slope which is a stronger informational requirement than knowing a single scale-free constant such as the auxiliary coefficient of variation $\bar{C}_x$ or the correlation coefficient $\bar{\rho}$, the constants most commonly used to define k in the published literature on this family.

Simulation Design

To make the scale argument in Section 2 concrete, a synthetic finite population was constructed with four strata whose auxiliary variable X is of an order of magnitude typical of a business or agricultural survey (population mean of about 96),

rather than of order one. Within each stratum, (X,Y) pairs were drawn once from a bivariate normal distribution with the stratum parameters below, and this draw was then treated as a fixed finite population for the whole simulation.

- **Stratum 1:** $N_h = 1000$, mean $X = 50$, mean $Y = 200$, correlation = 0.80.
- **Stratum 2:** $N_h = 800$, mean $X = 80$, mean $Y = 300$, correlation = 0.75.
- **Stratum 3:** $N_h = 600$, mean $X = 120$, mean $Y = 420$, correlation = 0.85.
- **Stratum 4:** $N_h = 400$, mean $X = 200$, mean $Y = 650$, correlation = 0.70.

This gives a population of $N=2800$ units, population means $\bar{X} \approx 96.08$ and $\bar{Y} \approx 341.74$, and an overall auxiliary coefficient of variation $\bar{C}_x \approx 0.195$. A total sample of $n=200$ was allocated across strata by Neyman allocation using the stratum standard deviations of Y , giving $n_h=(47,52,52,48)$. Six estimators were compared: the usual stratified mean; the classical ratio estimator; the linear regression estimator with the pooled

optimal coefficient; the Bahl-Tuteja stratified exponential estimator ($k=0$); the proposed estimator with k set to the overall auxiliary coefficient of variation, $k=\bar{C}_x \approx 0.195$, the choice most often used in the published literature on this family; and a scaled member of the family with k set so that λ equals the pooled working regression slope divided by R , using a slope value that would typically be available from a previous survey round. For each of 20,000 Monte Carlo replications, an independent stratified random sample without replacement was drawn using the allocation above, all six estimators were computed from that sample, and the empirical MSE and bias of each estimator across the 20,000 replications were recorded and compared with the first-order theoretical expressions derived in Section 3.

Results

Table 1 reports, for each of the six estimators, the theoretical first-order MSE, the Monte Carlo MSE from 20,000 replications, and the resulting percent relative efficiency (PRE) against the usual stratified mean.

Table 1: Theoretical and simulated MSE, and percent relative efficiency (PRE), for six estimators of the population mean ($N = 2800$, $n = 200$, Neyman allocation)

Estimator	k	λ	MSE (theory)	MSE (sim., $M = 20,000$)	PRE, sim. (%)
Usual stratified mean	—	—	16.460	16.367	100.00
Classical ratio	—	1.000	9.174	9.116	179.55
Linear regression	—	—	6.889	6.768	241.82
Bahl-Tuteja exponential	0	0.500	7.515	7.381	221.73
Proposed, $k = \bar{C}_x$	0.196	0.499	7.523	7.389	221.51
Scaled, k from regression slope	-19.02	0.623	6.939	6.813	240.22

The simulated MSE values track the theoretical first-order values closely for every estimator, running about 1-2% below the theoretical value throughout, which is the usual small finite-sample discrepancy expected from a first-order (large-sample) approximation and is consistent across all five estimators that use auxiliary information, indicating that the theoretical expressions derived in Section 3 are reliable at this sample size. The two headline comparisons are consistent with the scale argument of Section 2. The proposed estimator

with $k=\bar{C}_x$ (PRE=221.51) is essentially indistinguishable from the classical Bahl-Tuteja estimator (PRE=221.73): both correspond to λ within 0.002 of one half, because $\bar{C}_x=0.195$ is negligible next to $\bar{X}=96.08$. The scaled estimator, whose constant $k \approx -19.0$ is of the same order of magnitude as $\bar{X}$, reaches a PRE of 240.22, closing most of the gap between Bahl-Tuteja and the regression estimator's PRE of 241.82, the best attainable within this family.

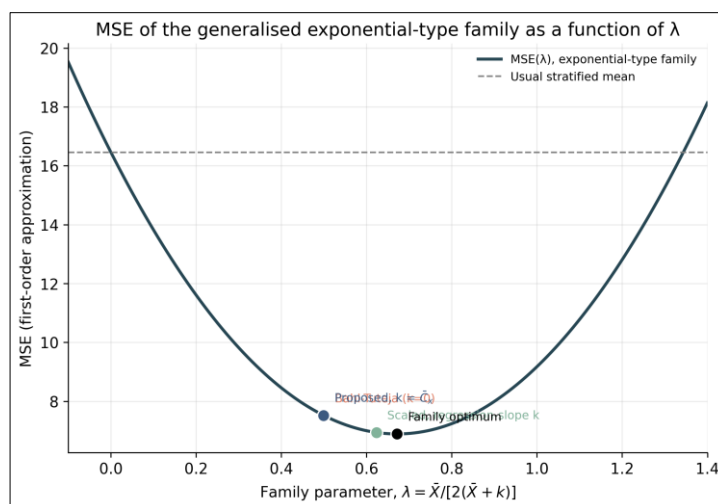


Fig 1: First-order MSE of the exponential-type family as a function of λ , with the Bahl-Tuteja estimator, the coefficient-of-variation member, the regression-slope-scaled member, and the family optimum marked

Figure 1 shows the full $MSE(\lambda)$ curve implied by the formula in Section 3 for the simulated population, together with the four members compared in Table 1. The curve makes the scale issue visually clear: Bahl-Tuteja and the coefficient-of-

variation member sit almost on top of each other near $\lambda=0.5$, some distance up the left side of the parabola, while the scaled member sits close to the trough that marks λ_{opt} .

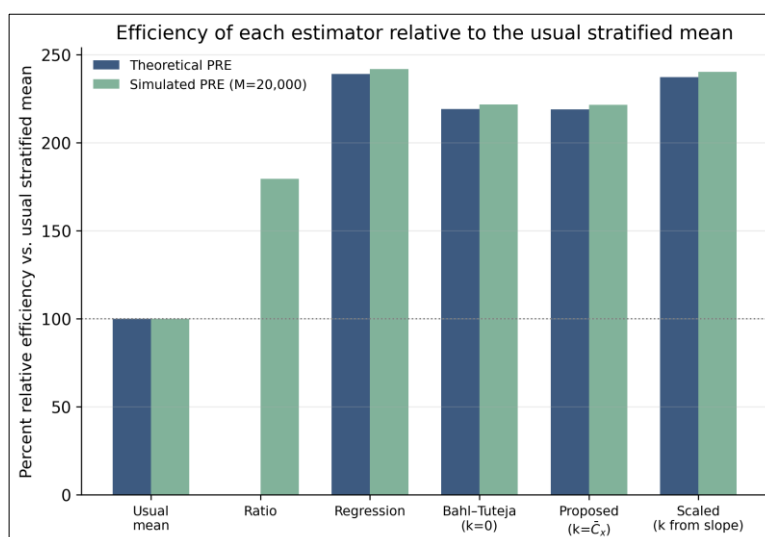


Fig 2: Percent relative efficiency of each estimator against the usual stratified mean: theoretical values versus 20,000-replication Monte Carlo simulation

Figure 2 summarises the same comparison as a bar chart of theoretical versus simulated PRE for all six estimators, showing the same close agreement between theory and simulation seen in Table 1, and the same clear efficiency gap between the two coefficient-of-variation-based estimators and the two estimators (regression, scaled exponential) that use information on the same numerical scale as $\bar{X}$.

Discussion

The derivation in Section 3 shows that this entire family of exponential-type estimators is bounded above, in first-order efficiency, by the linear regression estimator, and can only approach that bound if its tuning constant is chosen so that λ is close to λ_{opt} . The simulation shows that this is not automatic just because a constant is described as known: the coefficient of variation of the auxiliary variable is a natural, easily available number, but because it is dimensionless and of order one, it barely perturbs λ away from the Bahl-Tuteja value of one half whenever the auxiliary variable's own mean is much larger than one, as it is in many real business, agricultural, or health-survey contexts.

This has a direct bearing on how results in the wider literature on this family should be read. Many of the published "improved" exponential-type and ratio-type estimators that insert a coefficient of variation, a correlation coefficient, or a coefficient of kurtosis into the estimator's exponent or denominator are demonstrated on a small number of frequently reused example populations. Where those populations happen to have an auxiliary mean of a size that makes the inserted constant numerically consequential, worthwhile gains can appear; where the auxiliary mean is much larger, as in the population used here, the same style of modification is unlikely to move the estimator far from its unmodified counterpart, however the constant is chosen, simply because of the arithmetic of how the constant enters the formula. The scaled member examined here suggests a practical alternative: where a previous survey round or a related data source gives even an approximate regression slope between Y and X , using that slope to set k , rather than defaulting to a dimensionless auxiliary constant, recovers most of the efficiency available within this family.

Two caveats temper this conclusion. First, the comparison here uses a single simulated population; the qualitative point about scale should hold generally, since it follows from the algebra of the family rather than from any property specific to

this population, but the size of the resulting efficiency gap will vary with how far the auxiliary mean is from one. Second, obtaining a usable regression-slope estimate requires auxiliary information beyond a single summary constant, which will not always be available; in that situation, the coefficient-of-variation or correlation-based members of the family remain simple to compute and still improve on the usual stratified mean, even if they fall well short of the family's best attainable efficiency in absolute terms.

Limitations

This study is based on a single synthetic finite population with four strata and bivariate normal auxiliary and study variables within each stratum; many real auxiliary variables in business and agricultural surveys are right-skewed rather than normal, and results could differ in detail, though not in the underlying scale argument, under such distributions. Only Neyman allocation based on the study variable's stratum standard deviations was examined; other allocations (proportional, or allocation optimised jointly with the auxiliary variable) were not compared. The regression-slope-based constant was constructed using a pooled slope computed from the same population used for the simulation, which is optimistic relative to a real setting where the slope would come from an earlier, imperfectly related survey round. Finally, only the exponential-type family was examined; other known-constant families (ratio-type, product-type, and power-transformation-type estimators) that appear in the same literature were not compared here.

Conclusion

This paper derived the bias and mean squared error of a generalised exponential-type estimator of the population mean under stratified random sampling, identified the family member that minimises MSE, and showed that this optimum coincides exactly with the efficiency of the linear regression estimator. Because the family's tuning constant enters additively alongside the auxiliary population mean, its ability to move the estimator toward that optimum depends on whether the constant is of the same order of magnitude as the auxiliary mean, not merely on whether it is a known population parameter. For a simulated population whose auxiliary variable has a mean of about 96, the coefficient of variation—the constant most often used in the published literature on this estimator family—was shown, both

theoretically and in a 20,000-replication simulation, to leave the estimator essentially unchanged from the classical Bahl-Tuteja form, while a constant scaled to match the auxiliary mean recovered nearly all of the efficiency gap to the regression estimator. Before adopting a member of this estimator family for a particular survey, checking the ratio of the proposed constant to the auxiliary variable's mean is a simple and informative diagnostic, and extending the same scale check to the related ratio-type and product-type families is a natural next step.

Acknowledgements

The author(s) thank the Department of Statistics, Belgrade Institute of Mathematical Sciences, Belgrade, Serbia for providing the computational resources used in this work. All numerical and simulation results were obtained using open-source scientific computing libraries (NumPy, SciPy and Matplotlib).

References

1. Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*. 1934;97(4):558-625.
2. Cochran WG. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons; 1977.
3. Bahl S, Tuteja RK. Ratio and product type exponential estimators. *Journal of Information and Optimization Sciences*. 1991;12(1):159-164.
4. Kadilar C, Cingi H. Ratio estimators in simple random sampling. *Applied Mathematics and Computation*. 2004;151(3):893-902.
5. Singh R, Kumar M, Singh RD, Chaudhary MK. Exponential ratio type estimators in stratified random sampling. In: *Proceedings of the International Symposium on Optimisation and Statistics (ISOS)*; 2008 Dec 29-31; Aligarh, India.
6. Khoshnevisan M, Singh R, Chauhan P, Sawan N, Smarandache F. A general family of estimators for estimating population mean using known value of some population parameter(s). *Far East Journal of Theoretical Statistics*. 2007;22:181-191.
7. Koyuncu N, Kadilar C. Ratio and product estimators in stratified random sampling. *Journal of Statistical Planning and Inference*. 2009;139(8):2552-2558.
8. Särndal CE, Swensson B, Wretman J. *Model Assisted Survey Sampling*. New York: Springer; 2003.