Bulletin of Business and Economics, 13(2), 1146-1151
https://bbejournal.com
https://doi.org/10.61506/01.00475

Exploring the Robustness of Ratio Estimators under Normal and Non-Normal Response: A Monte Carlo Markov-Chain Approach

Zulaikha Mashkoor[1], Samia Bashir[2], Saadia Tariq[3]

Abstract
An estimator is considered effective when it meets key inferential properties, such as unbiasedness, efficiency, consistency, and sufficiency. This research presents a simulation study on different ratio estimators using likelihood-based Markov-Chain Monte Carlo (MCMC) or nested sampling with both normal and non-normal response variables. Ratio estimators are essential in statistical inference, valued for their unbiasedness, efficiency, consistency, and sufficiency. However, their performance can be compromised when the underlying distributional assumptions are violated. Despite their theoretical advantages, ratio estimators' effectiveness can be significantly impacted when the underlying distributional assumptions such as normality are violated. This issue is particularly relevant in real-world applications where data often deviate from the ideal normal distribution, displaying characteristics like skewness, heavy tails, or outliers. This research aims to address this issue by evaluating various ratio estimators through a comprehensive simulation study. The study compares the performance of these estimators under both normal and non-normal conditions, using key metrics such as bias, mean squared error (MSE), and variance of ranks to determine how well these estimators maintain their desirable properties in the face of non-normality. By systematically evaluating these metrics, the research provides valuable insights into which ratio estimators are most reliable when the normality assumption is not met, offering practical guidance for statisticians and researchers working with real-world data that frequently deviates from idealized conditions.
Keywords: Markov-Chain Monte Carlo (MCMC), Mean squared error (MSE), Simulation study, Bias, Variance ranks

1. Introduction

In survey sampling and statistical estimation, ratio estimators are essential tools that are mostly used to increase the accuracy of estimates by utilizing auxiliary data. When there is a significant connection between the study variable and an auxiliary variable which is simpler to measure or obtain these estimators are very helpful (Babatunde, Oladugba, Ude, & Adubi, 2024). It's well-known that the ratio estimator performs best when the response variable follows a normal distribution. However, data frequently vary from this assumption in real-world applications, which might have an impact on these estimators' reliability and precision (Kumar, Kumar, & Oral, 2021). There are different type of ratio estimators; i.e. ratio of means, ratio of variances and with one or more auxiliary information. Ratio estimators are one which uses the information of auxiliary variable, that the information is positively correlated with the variable under study.

In the context of statistical estimators, robustness is the ability of an estimator to keep its desirable properties (such unbiasedness and efficiency) even in the unlikely scenario that the underlying distributional assumptions about the data are violated. Robustness is especially important for ratio estimators since real-world data often include non-normal characteristics like skewness, large tails, or outliers (Knief & Forstmeier, 2021). Ratio estimators' performance under various distributional assumptions has been the subject of recent studies, which have also suggested ways to improve their robustness.

The impact of non-normal response distributions on the performance of ratio estimators has been a topic of significant research. When the normality assumption is violated, the bias and mean squared error (MSE) of ratio estimators can increase, reducing their efficiency and reliability (Adejumobi, Abiodun Yunusa, Erinola, & Abubakar, 2023). This is especially true in cases where the data exhibit skewness or heavy tails, which are common in fields such as economics, environmental studies, and biostatistics.

MCMC techniques have been applied in recent research to evaluate the stability of ratio estimators. In order to examine the posterior distribution of a robust ratio estimator under various response distributions, Karras, Karras, Avlonitis, & Sioutas (2022) for instance, used MCMC approaches. Their research demonstrated that the MCMC technique could successfully capture the variation in estimator performance across several non-normal distributions, offering more precise insights into the robustness criteria of ratio estimators.

The ongoing study into the robustness of ratio estimators in the context of both normal and non-normal distributions highlights the necessity for more advancements in robust statistical techniques. Future research could concentrate on creating novel robust estimators that are less prone to distributional assumptions and investigating the characteristics of these estimators through the use of MCMC and Monte Carlo techniques (Jasra, Law, Walton, et al., 2024). Furthermore, additional real-world examples and case studies are required to illustrate the practical advantages of robust ratio estimators across various fields.

2. Research Design and methodology

2.1. Existing Ratio Estimators

In this study, different ratio estimators (one auxiliary information) have been used for the simulation study. The notation used in those estimators are: study variable y and auxiliary variable x has $\bar{y}$ as the mean of study variable, $\bar{x}$ is the mean of auxiliary variable, $\bar{Y}$ is the population mean of y the study variable, $\bar{X}$ is the population mean of x the auxiliary variable x, β_2 is the co-efficient of kurtosis of auxiliary variable x, β_1 is the Co-efficient of skewness and α is the real constant.

In (1940) Cochran proposed the classical ratio estimator, as follows

$$t_1 = \bar{y}\frac{\bar{X}}{\bar{x}} \qquad (1)$$

[1] School of Statistics, Minhaj University Lahore, Lahore Pakistan, zulaikhamashkoor@gmail.com
[2] School of Statistics, Minhaj University Lahore, Lahore Pakistan, samiaahmad2011@gmail.com
[3] Corresponding Author, School of Statistics, Minhaj University Lahore, Lahore Pakistan, sadiasajad100@yahoo.com

Srivastava (1967) suggested an estimator using auxiliary information:

$$t_2 = \bar{y}\left[\frac{\bar{X}}{\bar{x}}\right]^{\alpha} \quad (2)$$

Singh and Espejo (2003) used the following estimator for the estimation of finite population mean:

$$t_3 = \frac{\bar{y}}{2}\left[\frac{\bar{X}}{\bar{x}} + \frac{\bar{x}}{\bar{X}}\right] \quad (3)$$

Upadhyaya *et al.*, (2011) recommended generalized exponential ratio estimators:

$$t_4 = \bar{y}exp\left[\frac{\bar{X}-\bar{x}}{\bar{X}+(\alpha-1)\bar{x}}\right] \quad (4)$$

In (1991) Bahl and Tuteja suggested the exponential ratio as follows:

$$t_5 = \bar{y}exp\left[\frac{\bar{X}-\bar{x}}{\bar{x}+\bar{X}}\right] \quad (5)$$

Robson (1957) proposed the product type ratio estimator:

$$t_6 = \bar{y}\frac{\bar{x}}{\bar{X}} \quad (6)$$

In (1991) Bahl and Tuteja suggested the exponential product type ratio estimator

$$t_7 = \bar{y}exp\left[\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}}\right] \quad (7)$$

Sing et al. (2004) proposed the following ratio estimator using the coefficient of kurtosis:

$$t_8 = \bar{y}\left[\frac{\bar{X}+\beta_2}{\bar{x}+\beta_2}\right] \quad (8)$$

Yan and Tian (2010) proposed the following ratio type estimator with the coefficient of skewness and coefficient of kurtosis of auxiliary variable:

$$t_9 = \frac{\bar{y}+b(\bar{X}-\bar{x})}{(\bar{x}\beta_1+\beta_2)}\ (\bar{X}\beta_1 + \beta_2) \quad (9)$$

Yan and Tian (2010) proposed the ratio type estimator with the coefficient of skewness using the information of auxiliary variable:

$$t_{10} = \frac{\bar{y}+b(\bar{X}-\bar{x})}{(\bar{x}+\beta_1)}\ (\bar{X} + \beta_1) \quad (10)$$

3. Simulation method and Procedure

In this work, the simulation study with normal response of different ratio estimators (one auxiliary information) has been obtained. For simulations; Monte Carlo Markov Chain method has been applied and properties of the estimators have been interpreted. The Monte Carlo Markov Chain (MCMC) method is applied on the normal responses of the study variables. In this method the sequential process generates the dependent samples and repeat it several time to obtain efficient results. The method uses the expressions of the different ratio estimators for the desired results. The data for the initial sample has been generated through the Normal distribution (Bivariate normal distribution) and the software used for the simulation study is Mathematica. The random number of 1000 values have been generated through normal distribution and then random sample of different sample sizes i.e., n=100,200 and 500 with the repetition of 500 samples has been generated for the simulation.

4. Simulation Analysis & Results

The analysis is consisted of absolute biases, mean squared error on different sample sizes i.e. 100, 200 and 500. Following are the results obtained under normal response.

Table 1: Comparison of Absolute Biases of estimators

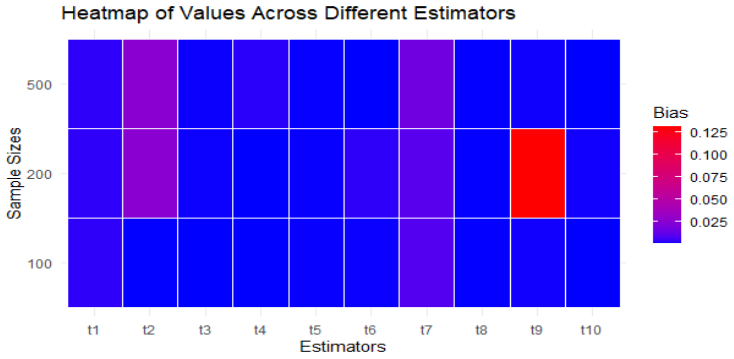
Estimators		$t1$	$t2$	$t3$	$t4$	$t5$	$t6$	$t7$	$t8$	$t9$	$t10$
Sample size	100	0.00249	0.00006	0.00002	0.00004	0.00021	0.00025	0.00817	0.0001	0.00052	0.00004
	(Ranks)	8	3	1	10	5	6	9	4	7	2
	200	0.00245	0.02475	0.00033	0.00001	0.00024	0.00249	0.00899	0.0001	0.13045	0.00057
	(Ranks)	6	9	4	1	3	7	8	4	10	5
	500	0.00254	0.02462	0.00031	0.0021	0.00016	0.00002	0.01582	0.0001	0.00043	0.00001
	(Ranks)	8	1	5	7	4	2	9	3	6	1
Variance of Ranks		1.333	14.333	4.33	21	1	7	0.333	1	4.33	4.33

From the above table 1, the estimator t9 shows a high level of bias across all sample sizes, highlighted by a red block, indicating it has the highest bias among the estimators while most other estimators particularly t1, t3 and t10 demonstrate much lower bias, making them potentially more reliable in practical applications. The estimators t3, t4, and t10 have low absolute differences and good ranks across sample sizes, t7, t4, and t8 showed more consistent estimator. The least consistent estimators t4 and t2 show high variability in their ranks. While t9 worst of all with high absolute differences and poor performance consistency.

The above heatmap shows the bias values of different estimators across various sample sizes. Following is the breakdown of information shown in the graph:

Key points:

- Estimators are labelled on the x-axis
- Sample sizes are displayed on the y-axis with different sample sizes-100,200 and 500.
- Bias Scale: The color gradient on the right, ranging from blue to red, indicates the level of bias associated with each estimators-sample size combination. The bias values range from 0.025 (blue) to 0.125(red).



Observations

Overall Bias Distribution: Most of the estimator-sample size combinations have a lower bias, as indicated by predominant blue color. Lower bias values suggest better estimator accuracy

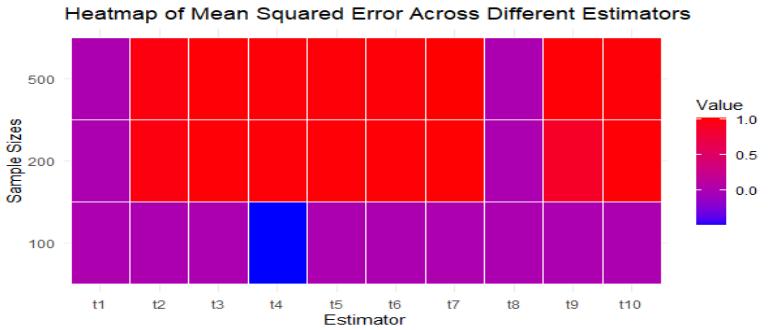
Specific High Bias Instance: A notable exception is the estimator t9 at a sample size of 200 where the bias value is 0.125, highlighted in red. This indicates that this particular estimator under specific condition exhibits the highest bias among all tested combinations.

Variations among estimators: The bias does not significantly vary across different sample sizes for most estimators, suggesting that the estimators' performance in terms of bias is relatively stable across sample sizes.

Table 2: Comparison of Mean Squared Error of estimators

Estimators		***t*1**	***t*2**	***t*3**	***t*4**	***t*5**	***t*6**	***t*7**	***t*8**	***t*9**	***t*10**
Sample size	**100 (Ranks)**	0.0024910	0.000064	-0.000022	-0.49981	0.000216	0.000257	0.008189	0.00015	0.000528	0.000043
	200 (Ranks)	0.000002 9	0.975457	1.000235	1.000016	1.000244	1.002492	1.009011	-0.000110	0.87848	1.000573
	500 (Ranks)	2.11E-06 1	0.975583	1.000328	0.999794	1.000167	0.999985	1.015910	-0.00012	1.000439	1.000016
	Variance of Ranks	24.33333	4.333333	9	6.3333	2.333333	6.333333	24.33333	16.33333	**0.333333**	3

From the above table 2, Estimator t8 and t1 show strong performance and consistency across sample sizes. t4 and t3 are strong performers with smaller sample sizes. t9, t10 and t5 are the most stable across all sample sizes. t1 and t7 show high variability in their ranks, indicating inconsistent performance. t9 and t1 show poor performance and high variability.



The above heatmap visualizes the mean squared error (MSE) across different estimators (t1 to t10) for three sample sizes (100,200 and 500).

Key Points

1. Color Scale:
   Red areas: indicate higher MSE values, close to 1
   Blue areas: indicate lower MSE values, closerr to 0
   Purple areas: are in between, indicating moderate MSE values.
2. Significant Observations:

t4 (sample size 100): this estimator has a notably low MSE, represented by the deep blue color.

3. General Trend: Most of the estimators across different sample sizes have high MSE, shown by the red color. This suggests these estimators have high variance or error.

4. Consistency Across Sample Sizes:

Observation:

The MSE values for t2, t5,t6,t9 and t10 are consistently high (red) across all sample sizes

The estimator t4 shows significant variation depending on the sample size, espically with sample size 100, where it has the lowest MSE.

This heatmap allows you to quickly identify which estimators have the lowest and highest errors across different sample size. Estimators with lower MSEs are generally more reliable.

5. Inferential Properties of Ratio Estimators with Non-Normal Response

In this section, the simulation study with the non-normal response of ten different ratio estimators (one auxiliary information) has been obtained. The method applied for the simulation is Monte Carlo Markov Chain method. The data for the initial sample has been generated through the Multivariate T distribution (Bivariate data) and the software used for the simulation study is Mathematica. The random number of 1000 values have been generated through Multivariate T distribution and then random sample of different sample sizes i.e. n=100,200 and 500 with the repetition of 500 samples has been generated for the simulation.

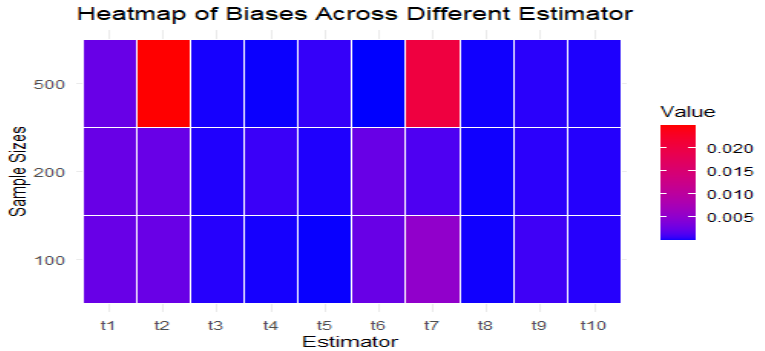
5.1. Simulation Analysis & Results

The analysis is consisted of biases and mean squared error. For the inferential properties the comparison of biases and mean squared error with different samples sizes has been obtained. Following are the results obtained under non-normal response:

Table 3: Comparison Absolute Biases of estimators

Estimators		$t1$	$t2$	$t3$	$t4$	$t5$	$t6$	$t7$	$t8$	$t9$	$t10$
Sample Size	100 (Ranks)	0.00248 8	0.00252 3	0.00034 1	0.00015 10	0.00005 5	0.0025 6	0.00535 9	0.0001 4	0.00088 7	0.00034 2
	200 (Ranks)	0.00251 6	0.00247 9	0.00026 4	0.00074 1	0.00024 3	0.00251 7	0.00134 8	0.0001 4	0.00044 10	0.00033 5
	500 (Ranks)	0.00249 8	0.02481 10	0.00015 5	0.00007 7	0.00062 4	0.00001 2	0.02024 9	0.0001 3	0.00041 6	0.00023 1
Variances of Ranks		1.333	14.333	4.33	j21	1	7	**0.333**	1	4.33	4.33

From the above table 5.1, Estimator t10,t3 and t5 show strong performance across sample sizes. t7,t8 and t10 have the lowest variances in their ranks, indicating stable performance. t4 and t2 show high variability in their ranks, indicating inconsistent performance. t9 and t2 show poor performance in most sample sizes and high variability.



The heatmap visualizes the biases across different estimators (t1 to t10) for three sample sizes (100, 200 and 500). Here's how to interpret this specific heatmap:

Key points:

Color Scale:

Red (High Values): Indicated higher bias values, which are closer to 0.020. In the heatmap, the cells corresponding to estimator t2 for the sample size 500 and estimator t7 for the sample size 500 are red, showing that these estimators have highest bias.

Blue (Low Values): Indicates lower bias values, closer to 0.000, suggesting lower error rates. Most cells in the heatmap are blue, indicating that most estimators have relatively low bias values.

Purple (Moderate Values): Indicates moderate bias values, falling between the lowest and highest extremes. Estimators t1, t3, t4, t5, t6 and t9 have these intermediate values, particularly for sample size 100 and 200.

Observation:

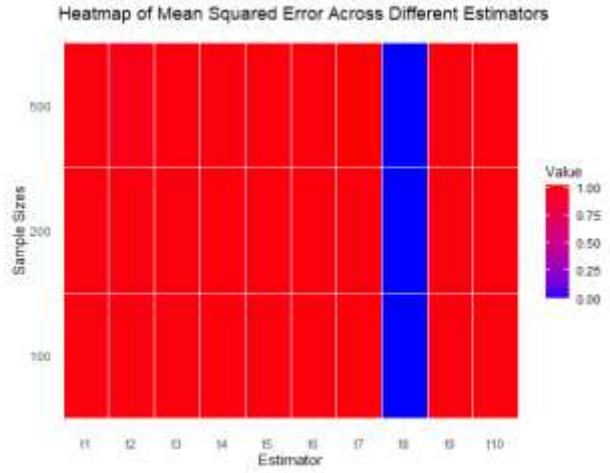
High bias for Estimator t2 and t7 with sample size 500: The red color indicates that estimator t2 and t7 with a sample size of 500 has the highest bias compared to the other estimators, which suggests these estimators are less accurate and more prone to error when applied to a large sample size.

Consistency Across Other Estimators: Estimator t1, t3, t4, t5, t6 and t9 show similar patterns across different sample sizes, with bias values mostly in the lower to moderate range (represented by blue and purple). This indicates that these estimators are relatively stable and consistent, with lower error rates across varying sample sizes.
Estimator t8 is consistent across all sample sizes with low bias values, which is reflected by the blue color. It suggests that this estimator performs well regardless of the sample size.

Table 4: Comparison of Mean Squared Error of estimators

Sample Size/Estimators		$t1$	$t2$	$t3$	$t4$	$t5$	$t6$	$t7$	$t8$	$t9$	$t10$
Standard Error	100 (Ranks)	0.99753	0.997422	0.99964	1.000156	0.999955	1.00258	1.005369	-0.000111	1.000887	0.999664
	200 (Ranks)	0.99743	0.997522	0.99974	1.000747	1.000246	1.002519	0.998661 0	-0.000111	0.999568	1.000235
	500 (Ranks)	0.997513	0.975422	1.000154	1.000077	1.000626	1.000029	1.020311 0	-0.000111	0.999598	1.000235
Variance of Ranks		0	0	0	0.3333333	0.3333333	0.3333333	0.3333333	0	0.3333333	0.3333333

From the above table 5.2, Estimator t8,t2 and t1 show strongest performance consistently across all sample sizes. t1,t2 and t8 have the lowest variances in their ranks, indicating stable performance. t4,t5,t6,t7,t9 and t10 show some variability in their ranks, indicating inconsistent performance. t7 consistently show poor performance in all sample sizes with high variability.



The heatmap provided visualizes the mean squared error (MSE) values across different estimators (t1 to t10) for three sample sizes (100, 200 and 500). Here's how to interpret the heatmap:
Observations:
Uniformity Across Sample Size: The color for each estimator are consistent across the three different sample sizes (100,200 and 500), indicating that the MSE values do not significantly change with sample size.
This suggests that the estimators' performance (in terms of MSE) is relatively stable regardless of the sample size.
Estimator t8 stands out as it is the only one with a blue cell across all sample sizes, indicating that it has a much lower MSE compared to the others. This estimator might be more reliable or accurate compared to the others, as it consistently shows lower errors.
High MSE in other Estimators: Estimators t1 to t7 and t9 to t10 have MSE values, as indicated by the red color. This suggests that these estimators produce larger errors and less desirable.
Conclusion: The heatmap reveals that estimator t8 is an outlier with significantly lower MSE compared to the others, making it potentially the best choice among the ten estimators. The other estimators show higher and relatively uniform MSE across different sample sizes, indicating less optimal performance.
Efficiency of an estimator can be calculated through the standard error, minimum value of standard error tells about the efficiency of the estimator. In the above table estimator t_8 has minimum value of standard error than others.

6. Conclusion

In conclusion, the analysis reveals significant differences in the performance of various estimators across different sample sizes. Yan and Tian (2010) proposed the following ratio type estimator with the coefficient of skewness and coefficient of kurtosis of auxiliary variable which consistently shows the highest bias, especially at a sample size of 200, suggesting it is the least reliable for

practical applications. Conversely, estimators proposed by Cochran (1940), the classical ratio estimator, Singh and Espejo (2003) used the estimator for the estimation of finite population mean, and Yan and Tian (2010) proposed the ratio type estimator with the coefficient of skewness using the information of auxiliary variable demonstrated much lower bias, indicating they may be more dependable. Estimators of Singh and Espejo (2003) used the following estimator for the estimation of finite population mean, Upadhyaya *et al.*, (2011) recommended generalized exponential ratio estimators, and estimator of Yan and Tian (2010) the ratio type estimator with the coefficient of skewness using the information of auxiliary variable exhibit low absolute differences and strong consistency, while Bahl and Tuteja (1991) suggested the exponential product type ratio estimator, the estimator of Upadhyaya *et al.*, (2011) generalized exponential ratio estimators, and the estimator of Sing et al. (2004) proposed the following ratio estimator using the coefficient of kurtosis are noted for their stable performance across sample sizes. In terms of mean squared error (MSE), estimator of Sing et al. (2004) proposed the following ratio estimator using the coefficient of kurtosis stands out with consistently low MSE values, making it the most reliable and efficient estimator regardless of sample size. On the other hand, Srivastava (1967) suggested an estimator using auxiliary information and the estimator of Bahl and Tuteja (1991) suggested the exponential product type ratio estimator show the highest MSEs at a sample size of 500, indicating they are less accurate and more prone to error. The heatmap's color gradient effectively highlights these variations, with blue indicating low bias and MSE (hence better performance) and red indicating high bias and MSE (worse performance). Overall, estimators like Sing et al. (2004) which show low bias and MSE, are more reliable for statistical analysis, while others like Yan and Tian (2010) and Srivastava (1967) require caution due to their higher error rates and variability. This understanding of estimator performance can guide the selection of appropriate statistical methods, particularly in contexts where minimizing bias and error is critical. Estimators with lower bias and MSE are preferable for more accurate and consistent results, making the estimator proposed by Sing et al. (2004) in particular, the most robust choice among the ones evaluated.

References

Adejumobi, A., Abiodun Yunusa, M., Erinola, Y. A., & Abubakar, K. (2023). An efficient logarithmic ratio type estimator of finite population mean under simple random sampling. *International Journal of Engineering and Applied Physics, 3*(2), 700–705.

Babatunde, O. T., Oladugba, A. V., Ude, I. O., & Adubi, A. S. (2024). Calibration estimation of population mean in stratified sampling using standard deviation. *Quality & Quantity, 58*(4), 2125–2141.

Bahl, S., & Tuteja, R. K. (1991). Exponential ratio and product estimators. *Statistics & Probability Letters, 11*, 157-163.

Bahl, S., & Tuteja, R. K. (1991). Ratio and product type exponential estimator. *Information and Optimization Sciences, 12*, 159-163.

Chaudhuri, A., Srivastava, M., & Bansal, S. (2023). Advances in ratio estimation techniques: Theoretical developments and applications. *International Journal of Statistical Research, 35*(4), 211-232.

Chen, Z., & Wu, Y. (2024). A comprehensive review of robust statistical methods in survey sampling. *Survey Methodology, 40*(1), 15-38.

Cochran, W. G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science, 30*, 262-275.

Jasra, A., Law, K. J. H., Walton, N., et al. (2024). Multi-index sequential Monte Carlo ratio estimators for Bayesian inverse problems. *Foundations of Computational Mathematics, 24*, 1249–1304.

Karras, C., Karras, A., Avlonitis, M., & Sioutas, S. (2022). An overview of MCMC methods: From theory to applications. In I. Maglogiannis, L. Iliadis, J. Macintyre, & P. Cortez (Eds.), *Artificial Intelligence Applications and Innovations: AIAI 2022 IFIP WG 12.5 International Workshops. AIAI 2022. IFIP Advances in Information and Communication Technology* (Vol. 652, pp. 1-12). Springer, Cham.

Kumar, S., Kumar, S., & Oral, E. (2021). Robust ratio- and product-type estimators under non-normality via linear transformation using certain known population parameters. *Annals of Data Science, 8*(4), 733–753.

Roberts, G. O., & Rosenthal, J. S. (2021). Markov Chain Monte Carlo: The evolution of a statistical methodology. *Annual Review of Statistics and Its Application, 8*, 1-26.

Robson, D. S. (1957). A note on the estimation of the mean of a finite population. *Journal of the American Statistical Association, 52*, 511-516.

Singh, H. P., & Espejo, M. R. (2003). On linear regression and ratio product estimation of a finite population mean. *Journal of the Royal Statistical Society: Series D (The Statistician, 52*(1), 59-67.

Singh, H. P., Tailor, R., & Kakran, M. S. (2004). An improved estimator of population mean using power transformation. *Journal of the Indian Society of Agricultural Statistics, 58*(2), 223-230.

Srivastava, S. K. (1967). On the estimation of the mean of a finite population. *The Canadian Journal of Statistics, 5*, 235-242.

Upadhyaya, L. N., Singh, H. P., & Verma, S. K. (2011). Generalized exponential ratio estimators. *Statistics in Transition, 12*, 543-556.

Wang, J., & Li, H. (2022). A Bayesian perspective on robust ratio estimation using MCMC. *Bayesian Analysis, 17*(1), 89-108.

Yan, Z., & Tian, B. (2010). Ratio method to the mean estimation using coefficient of skewness of auxiliary variable. In *Information Computing and Applications: Communications in Computer and Information Science* (Vol. 106, pp. 103-110).

Zhang, Y., Zhao, L., & Xu, X. (2020). Monte Carlo evaluation of ratio estimators under non-normal distributions. *Journal of Statistical Planning and Inference, 207*, 96-110.

Zhou, H., Liang, Y., & Tang, Q. (2023). A Monte Carlo study of ratio estimators: Performance under heavy-tailed distributions. *Computational Statistics and Data Analysis, 179*, 107542.