Modified unbiased estimators for population variance: An application for COVID-19 deaths in Russia

Hatice Oncel Cekim

Department of Statistics, Hacettepe University, Ankara, Turkey

Correspondence
Hatice Oncel Cekim, Hacettepe University, Department of Statistics, 06800 Beytepe, Ankara, Turkey.
Email: oncelhatice@hacettepe.edu.tr

Summary
The article deals with the class unbiased forms of the variance estimators using Hartley–Rosser type in the simple random sampling. The mean squared error (since it is an unbiased estimator, variance is calculated) of the suggested, up to the first order of approximation, is derived. The proposed estimator using COVID-19 data in Russia has been proven to be more efficient than the considered estimators under the conditions. Thus, it allows us to see the variance estimator that best predicts the change according to federal regions in Russia’s number of COVID-19 deaths. This study also provides an unbiased family of estimators that are more efficient than existing estimators, which estimate the variance of the total death number of COVID-19 based on the daily new cases number. The most important difference of this article from other studies in the literature is that it is the first study to examine the variance of the COVID-19 cumulative mortality value in terms of variance estimation using the simple random sampling method with the auxiliary variable.

KEYWORDS
COVID-19, death number, Hartley–Ross type estimator, unbiased estimator, variance estimator

1 | INTRODUCTION

Since January 2020, the Coronavirus disease 2019 (COVID-19), which has spread from Wuhan, China, affecting all countries around the world, has been a serious global crisis. On March 11, WHO declared COVID-19 a "global epidemic". COVID-19 has increased rapidly since the first week it emerged in Russia. The total number of cases and the number of deaths reached approximately 11 million and 325,000 as of January 2022 (https://www.worldometers.info/coronavirus/country/Russia). However, the number of cases and mortality may differ from city to city, region to region or federal state to federal state. This may be because each part (interested unit) has different population density, health care, safeguard measures, substructure or climatic. The best parameter value that can show this difference is variance.

Variance is one of the most commonly used measures of variability to describe the change in a data set. For this reason, it is desirable to know the variance prior to many studies. In cases where the variance is not known prior, it can be estimated with minor errors with the help of sampling. For this purpose, many variance estimators are proposed in sampling methods. The superiority of the proposed estimators is evaluated with many statistical criteria. Generally, it is required to provide the features of the best estimator, such as having the minimum mean squared error (MSE) value and having a zero biased value, that is, unbiased. This study proposes a family of unbiased estimators with a minor MSE. These proposed variance estimators can be of great importance in any field where variance (change in series) is actively used. In other words, the estimators are vital for all sectors; for example, the tourism sector as a country needs to plan significant scale building works or for aircraft companies as the purchase of new aircraft would require managerial decisions to be made well in advance. It is a guide for those who want to know a predictive range of variation that can be used for the variance of death numbers in the Coronavirus pandemic that has recently affected the world. For this purpose, the effectiveness of the proposed estimators in this study is desired to be investigated using the COVID-19 data of Russia.
variance in the total number of deaths from COVID-19 by federal states of Russia provides us to see the differences in the total number of deaths between states.

Mathematical models used in studies on the COVID-19 pandemic are methods to arrive at results such as daily cases, cumulative cases, number of deaths, risk of death and incubation period. Estimators with sampling methods are one of the different mathematical models to describe COVID-19. Zaman et al. propose families of exponential estimators that yield more efficient results than existing estimators for estimating the population mean of COVID-19 risk. Shahzad et al. presented some new estimators for estimating the population variance of recovery time from COVID-19 using L-Moments and the calibration approach. It also provided new calibration estimators with L-Moments for the variance estimation of the total recovery time depending on COVID-19 data in the stratified random sampling method from January 22 to August 23, 2020, by Shahzad et al.

In sampling methodology, the accuracy of an estimator at the estimation process can be raised with an influential work of the constant term and the auxiliary information. The variance estimators are the most commonly used estimators of the population parameter. Since the classical variance estimators are biased and complex, some studies are trying to develop variance estimators. In this regard, a class of unbiased variance estimators is obtained for various distributions by Chen et al. under ranked set sampling. Haq et al. developed combined and separate estimators for estimating population variance using supplementary information in another paper. In other variance assessing studies, Sing and Solanki and Lone and Tailor obtained estimators (under conditions) with less MSE than classical unbiased estimators. work developed estimators to estimate the finite population variance of the study variable with the help of the Hartley–Ross type method. Later, In type estimators of the population variance were suggested by Cekim and Kadilar for different sampling methods.

### 2 EVALUATION OF ESTIMATORS IN LITERATURE

Many authors have defined self-developed estimators whose explanations are included in the usual population mean to improve the variance estimation. In the year 1999, Upadhyaya and Singh introduced the estimator with the help of the kurtosis \( \beta_2(x) \). Then, Kadilar and Cingi defined the ratio estimators to \( S_j^2 \), used the coefficient of variation \( C_y \) as well as additionally \( \beta_2(x) \). These estimators respectively follow as:

\[
S_j^2 = s_j^2 \frac{S_j^2}{s_j^2}, \quad j = 1, \ldots, 4. \tag{1}
\]

Here, \( S_j^2 \) and \( s_j^2 \) are the population variances, while \( s_j^2 \) and \( S_j^2 \) are the unbiased estimators. Also, expressing \( S_1 = S_2^2 + \beta_2(x), S_2 = S_2^2 + C_y, S_{13} = C_y S_{13}^2 + \beta_2(x) \) and \( S_{14} = \beta_2(x) S_{14}^2 + C_y \), it is similarly shown in \( s_j^2, j = 1, \ldots, 4 \) for samples. The bias and MSE equations of these estimators are calculated as the following equations, respectively:

\[
\begin{align*}
B(s_j^2) &= \gamma s_j^2 \Delta_1 \left[ \Delta_2 \beta_2^2(x) - \lambda_{22} \right], \quad j = 1, 2, 3, 4 \tag{2} \\
\text{MSE}(s_j^2) &= \gamma s_j^2 \left[ \Delta_1 \beta_2^2(x) - 2 \lambda_{22} \beta_2^2(y) \right], \quad j = 1, 2, 3, 4. \tag{3}
\end{align*}
\]

Here, \( \gamma = \frac{n-n-1}{n} \). \( \Delta_1 = \frac{s_j^2}{S_1^2 + s_j^2}, \Delta_2 = \frac{s_j^2}{S_2^2 + s_j^2}, \Delta_3 = C_y \frac{s_j^2}{S_3^2 + s_j^2}, \Delta_4 = \beta_2(x) \frac{s_j^2}{S_4^2 + s_j^2}, \beta_2^2(x) = \beta_2(x) - 1, \beta_2^2(y) = \beta_2(y) - 1 \) and \( \lambda_{22} = \lambda_{22} - 1 \).

Considering their estimators in (1) and biased in (2), Hartley–Ross type estimators are proposed by Kadilar and Cekim

\[
s_{\text{HC}} = \frac{s_j^2 - B(s_j^2)}{M_{j\nu}}, \quad j = 1, 2, 3, 4. \tag{4}
\]

They obtained the variance equation of \( s_{\text{HC}} \), in (5), as

\[
\begin{align*}
V(s_{\text{HC}}) &= \gamma s_j^2 \left[ \Delta_1 \beta_2^2(x) - 2 \lambda_{22} \beta_2^2(y) \right] - \gamma \Delta_1 \left[ \Delta_2 \beta_2(x) - \lambda_{22} \right]^2 \\
&- 2 \gamma \Delta_1 \left[ \Delta_2 \beta_2(x) - \lambda_{22} \right] \left[ \beta_2^2(y) - 2 \lambda_{22} \beta_2^2(y) \right] + \lambda_{22} \left( \Delta_1 \gamma_{12} - \gamma_{22} \right) \\
&+ \gamma \Delta_1 \left[ \Delta_2 \beta_2^2(x) - 2 \lambda_{22} \beta_2^2(x) \right] \left( \gamma_{22} - \gamma_{22} \right) \\
&= \gamma s_j^2 \left[ \Delta_1 \beta_2^2(x) - 2 \lambda_{22} \beta_2^2(y) \right] - \gamma \Delta_1 \left[ \Delta_2 \beta_2(x) - \lambda_{22} \right]^2 \\
&- 2 \gamma \Delta_1 \left[ \Delta_2 \beta_2(x) - \lambda_{22} \right] \left[ \beta_2^2(y) - 2 \lambda_{22} \beta_2^2(y) \right] + \lambda_{22} \left( \Delta_1 \gamma_{12} - \gamma_{22} \right) \\
&+ \gamma \Delta_1 \left[ \Delta_2 \beta_2^2(x) - 2 \lambda_{22} \beta_2^2(x) \right] \left( \gamma_{22} - \gamma_{22} \right), \quad j = 1, \ldots, 4. \tag{5}
\end{align*}
\]
Similarly, the sample parameter

\[ \lambda_{P} = \frac{1}{N-1} \sum_{i=1}^{N} (Y_{i} - \bar{Y})^{2}(X_{i} - \bar{X})' \]

where positive integers \( p \) and \( r \),

\[ \mu_{p} = \frac{1}{N-1} \sum_{i=1}^{N} (Y_{i} - \bar{Y})^{p}(X_{i} - \bar{X})' \]

\[ \nu_{r} = \frac{\mu_{p} \mu_{p}'}{\mu_{p} \mu_{p}'} \]

\[ \gamma_{22} = \frac{[\mu_{44} - \mu_{22}^2 + 4 \mu_{12} \mu_{21} (S_{x}^2 + S_{y}^2 + 2S_{xy}) - 4 (\mu_{32} \mu_{12} + \mu_{21} \mu_{23})]}{\mu_{22}^2} \]

\[ \gamma_{02} = \frac{[\mu_{22} - \mu_{02} \mu_{22} - 2 (\mu_{03} \mu_{12} + \mu_{11})]}{S_{x}^2 \mu_{22}} \]

\[ \gamma_{02} = \frac{[\mu_{24} - \mu_{02} \mu_{22} - 2 (\mu_{03} \mu_{12} + \mu_{11})]}{S_{x}^2 \mu_{22}} \]

3 SUGGESTED ESTIMATORS

Yadav et al.,\textsuperscript{16} taking into account the variance estimator proposed by using the correlation coefficient and quartiles, these values are replaced by different population parameter values, and the following estimator family is obtained:

\[ s_{m0}^2 = s_{l0}^2 \left[ \omega + (1 - \omega) \sum_{i=1}^{9} \frac{\beta_{i}^2}{\gamma_{i}} \right] \quad j = 1, \ldots, 9 \]

where \( \omega \) is a constant chosen so that the MSE of \( s_{m0} \) reaches its minimum value. Also, expressing \( S_{xy} = S_{xx}^2 + \rho, S_{yy} = \rho S_{xx}^2 + C_{x}, S_{xy} = \rho S_{xx}^2 + \beta_{x}(x), S_{ab} = C_{y} S_{xx}^2 + \rho \) and \( S_{ab} = \beta_{y}(x) S_{xx}^2 + \rho \), it is similarly shown in \( s_{m0}^2, j = 1, \ldots, 9 \) for samples. To get the bias of the estimator in (6), \( \theta \)'s term are defined as

\[ s_{l0}^2 = S_{l0}^2 (1 + \theta_{0}) \quad \text{and} \quad s_{r}^2 = S_{r}^2 (1 + \theta_{1}) \]

such that

\[ E(\theta_{0}) = E(\theta_{1}) = 0, E(\beta_{0}^2) = \gamma \beta_{0}^2 \]

\[ E(\theta_{0}^2) = \gamma \beta_{0}^2 \]

The biases for these estimators are obtained as shown:

\[ s_{m0}^2 = s_{m0}^2 \left[ \omega + (1 - \omega) \sum_{i=1}^{9} \frac{\beta_{i}^2}{\gamma_{i}} \right] \quad j = 1, \ldots, 9 \]

where \( \Delta_{x} = \frac{s_{l0}^2}{S_{l0}^2}, \Delta_{y} = \frac{s_{r}^2}{S_{r}^2}, \Delta_{x} = \frac{S_{x}^2}{S_{x}^2 + \rho}, \Delta_{y} = \frac{S_{y}^2}{S_{y}^2 + \beta_{x}(x)}, \Delta_{x} = \frac{S_{x}^2}{S_{x}^2 + \beta_{y}(x)}, \Delta_{y} = \frac{S_{y}^2}{S_{y}^2 + \beta_{y}(x)} \)

By the definition of Hartley–Ross type estimators, the amount of bias must be subtracted from the considered estimator to compute an unbiased estimator. We remove Equation (7) from Equation (6) and suggest the following estimator to derive an unbiased estimator.

\[ s_{m0}^2 = s_{m0}^2 - B(s_{m0}) \quad j = 1, \ldots, 9 \]

Similarly, the sample parameter \( (M_{22} - 1) \) is used instead of \( \lambda_{22}^2 \) in (8). To get the variance of the estimator in (8), \( \theta \)'s term are defined as

\[ M_{22} = \mu_{22} (1 + \theta_{2}), E(\theta_{2}) = 0, E(\beta_{2}^2) = D_{22}, E(\theta_{2} \beta_{2}) = D_{02} \]

The variance equation is calculated to the \( s_{m0} \) estimator according to the necessary operations as:

\[ V(s_{m0}^2) \geq s_{m0}^2 \left[ (1 - \omega)^2 A_{l} - 2 (1 - \omega) B_{l} + C \right] \quad j = 1, \ldots, 9 \]

where

\[ A_{l} = \left[ \Delta_{x} \beta_{x}^2 (x) + \gamma \Delta_{x} \beta_{x}^2 (x) - \lambda_{22}^2 \right] + 2 \Delta_{x} \beta_{x}^2 (x) \lambda_{22}^2 - 2 \lambda_{22} \Delta_{x} D_{12} \]

\[ B_{l} = \left[ \Delta_{x} \lambda_{22}^2 - \Delta_{y} \beta_{y}^2 (y) + \lambda_{22} D_{02} \right] \]

\[ C = \beta_{y}^2 (y) \]
We assumed that $|\Delta \beta_1| < 1$ can be extended $(1 + \Delta \beta_1)^{-1}$ into the operations performed to reach the variance equality in (9). The optimum value of $\omega$ is obtained as

$$\omega_{opt} = 1 - \frac{B}{A}$$

to reach the minimum variance value. When the optimum value of $\omega$ is substituted in Equation (9), the minimum variance is obtained as

$$V(s_{HRj}^2)_{min} = \gamma S^2 \left[ C - \frac{B^2}{A} \right]. \quad (10)$$

## 4 | EFFICIENCY COMPARISONS OF MENTIONED ESTIMATORS

This section shows that the proposed Hartley–Ross type unbiased estimator class is the best estimator under certain conditions. By comparing the variance or MSE values of the estimators, it can be found under which conditions they are the best. The following conditions are obtained by comparing the class of suggested estimators with the mentioned estimators in this study.

First condition:

$$V(s_{HRj}^2)_{min} - \text{MSE} \left( s_j^2 \right) < 0.$$

if

$$\left[ C - \frac{B^2}{A} \right] - \left[ \Delta_j^2 \beta_j^2(x) - 2 \Delta_j \lambda_{22}^2 + \beta_j^2(y) \right] < 0, \quad j = 1, \ldots, 9. \quad (11)$$

Second condition:

$$V(s_{HRj}^2)_{min} - V(s_{KCj}^2) < 0.$$

if

$$\left[ C - \frac{B^2}{A} \right] - \left\{ \left[ \Delta_j^2 \beta_j^2(x) - 2 \Delta_j \lambda_{22}^2 + \beta_j^2(y) \right] - \gamma \Delta_j^2 \left[ \Delta_j \beta_j^2(x) - \lambda_{22}^2 \right]^2 \\
-2 \gamma \Delta_j \left[ \Delta_j \beta_j^2(x) (\beta_j^2(y) - \Delta_j \lambda_{22}^2 + \lambda_{22}^2 (\Delta_j Y_{12} - Y_{02})) \right] \\
+ \gamma^2 \Delta_j^2 \left[ \Delta_j^2 \beta_j^2(x) (\beta_j^2(y) - 2 \Delta_j \beta_j^2(x) \lambda_{22} Y_{02} + \lambda_{22}^2 Y_{22}) \right] \right\} < 0, \quad j = 1, \ldots, 9. \quad (12)$$

These conditions can be reproduced by comparing different estimators in the literature with the proposed estimators.

## 5 | DATA DESCRIPTION AND PRACTICE

### 5.1 | Data

Since COVID-19 came to Russia, new cases, recoveries and deaths are published daily for Russia’s assessment of the COVID-19 situation by the federal region. Confirmed data used in this study were retrieved from Reference 17 in September 2021. Sampling methods use auxiliary variable information to obtain more efficient estimators. Therefore, the total number of COVID-19 deaths and daily new cases as the dependent variable ($Y$) and the auxiliary variable ($X$), respectively, are considered in the application. Descriptive statistics of the federal region in Russia are given in Table 1.

### TABLE 1  Data statistics

| $Y$ = 1337.02 | $\beta_j(x) = 67.08$ | $N = 85$ |
|-------------|-----------------|---------|
| $X = 99.59$ | $\beta_j(y) = 32.76$ | $n = 20$ |
| $C_x = 3.25$ | $S_x = 2566.00$ | $\rho = 0.90$ |
| $C_y = 1.92$ | $S_y = 324.05$ |         |
Figure 1 shows the number of daily new cases and death values for COVID-19 are highest in Moscow. Then, the areas with the highest two variables are Saint Petersburg, Moscow Oblast and Rostov Oblast, respectively. Among 85 regions, Jewish Autonomous Oblast, Nenets Autonomous Okrug and Chukotka Autonomous Okrug have the smallest new cases value. With a cumulative death value of 4, the Nenets Autonomous Okrug has the lowest among other states.

5.2 Empirical study

This section aims to find the best estimator of the variance value of COVID-19 deaths in Russia among the estimators mentioned in the manuscript. When the correlation coefficient between COVID-19 cumulative new case and death values is evaluated, these two variables show a high correlation of 0.90. In other words, the COVID-19 new case and death values are closely related. Predicting COVID-19 cumulative death can use information about daily new cases. In this context, a new family of predictors has been proposed to estimate the COVID-19 total mortality variance with this auxiliary feature. With the help of these proposed estimators, the MSE (i.e., variance as it is an unbiased estimator) estimate of the cumulative death variance of COVID-19 is calculated theoretically and numerically.

Using Equations (3), (5), and (11), the variance and MSE values of the estimators for this data set are computed in Table 2.

In Table 2 results, the variance values of the proposed estimators are smaller than both the variance and MSE values of the existing estimators in the paper. In addition, for the COVID-19 data used in the study, it is seen from Table 2 that the best estimator with the smallest variance value is

| Estimators  | Estimators  | Estimators  | MSE       |
|-------------|-------------|-------------|-----------|
| $s^2_{HR1}$ | $s^2_{KC1}$ | $s^2_1$     | 1.69565E+14 |
| $s^2_{HR2}$ | $s^2_{KC2}$ | $s^2_2$     | 1.69598E+14 |
| $s^2_{HR3}$ | $s^2_{KC3}$ | $s^2_3$     | 1.69463E+14 |
| $s^2_{HR4}$ | $s^2_{KC4}$ | $s^2_4$     | 1.69598E+14 |
| $s^2_{HR5}$ | $s^2_{KC5}$ | $s^2_5$     | 1.69595E+14 |
| $s^2_{HR6}$ | $s^2_{KC6}$ | $s^2_6$     | 1.69525E+14 |
| $s^2_{HR7}$ | $s^2_{KC7}$ | $s^2_7$     | 1.69587E+14 |
| $s^2_{HR8}$ | $s^2_{KC8}$ | $s^2_8$     | 1.69597E+14 |
| $s^2_{HR9}$ | $s^2_{KC9}$ | $s^2_9$     | 1.69568E+14 |
FIGURE 2 The variance and MSE values of the considered estimators

all suggested estimators among the mentioned estimators. The numerical results of the calculations in Table 2 are presented differently in Figure 2. Also, it is seen from Figure 2 that the proposed estimator has better values.

6 | CONCLUSION

In this study, the unbiased estimators class is proposed for variance, the most commonly used indicator of population variation. The most preferred Hartley–Ross method is used to obtain these estimators’ unbiased estimators. It is supported by a real data study that the proposed estimator is the best estimator under certain conditions. The variance of the COVID-19 death numbers, which includes 85 federal areas of Russia, is examined in the real data study. The new case numbers are considered an auxiliary variable to explore the change in the number of deaths. In line with the obtained results, it is seen that the variance estimator that best predicts the change in the number of COVID-19 deaths in Russia is the estimators suggested in the study. In future work, the class of proposed estimators can be diversified by substituting different parameters for the population parameters in $\Delta_j$. By using these proposed estimators, it will be possible to increase the variance estimates in future studies for the number of COVID-19 deaths in different countries and different periods.

CONFLICT OF INTEREST
The author declares that there are no conflicts of interest regarding the publication of this article.

DATA AVAILABILITY STATEMENT
The data used to support the findings of this study are available from the author upon request. The data that support the findings of this study are openly available in https://www.statista.com at https://www.statista.com/statistics/1102935/coronavirus-cases-by-region-in-russia/, Reference 17.

ORCID
Hatice Oncel Cekim https://orcid.org/0000-0001-8538-6296

REFERENCES
1. Fang L, Wang D, Pan G. Analysis and estimation of COVID-19 spreading in Russia based on ARIMA model. SN Compr Clin Med. 2020;2(12):2521-2527.
2. Pramanik M, Udmale P, Bisht P, Chowdhury K, Szabo S, Pal I. Climatic factors influence the spread of COVID-19 in Russia. Int J Environ Health Res. 2020;32(4):723-737. doi:10.1080/09603123.2020.1793921
3. Zaman T, Sagir M, Şahin M. A new exponential estimators for analysis of COVID-19 risk. Concurr Comput Pract Exp. 2021;34:e6806.
4. Shahzad U, Ahmad I, Almanjahie IM, Al-Noor NH. Utilizing L-moments and calibration method to estimate the variance based on COVID-19 data. Fresenius Environ Bull. 2021;30(7A):8988-8994.
5. Shahzad U, Ahmad I, Almanjahie IM, Hanif M, Al-Noor NH. L-moments and calibration based variance estimators under double stratified random sampling scheme: an application of COVID-19 pandemic. Sci Iran. 2021;1-17.
6. Chen W, Tian Y, Xie M. The global minimum variance unbiased estimator of the parameter for a truncated parameter family under the optimal ranked set sampling. J Stat Comput Simul. 2018;88(17):3399-3414. doi:10.1080/00949655.2018.1520233
7. Haq A, Usman M, Khan M. Estimation of finite population variance under stratified random sampling. Commun Stat Simul Comput. 2021;1-17.
8. Singh HP, Solanki RS. A new procedure for variance estimation in simple random sampling using auxiliary information. Stat Pap. 2013;54(2):479-497.
9. Lone HA, Tailor R. Estimation of population variance in simple random sampling. *J Stat Manag Syst.* 2017;20(1):17-38. doi:10.1080/09720510.2016.1187923
10. Cekim HO, Kadilar C. In-type estimators for the population variance in stratified random sampling. *Commun Stat Simul Comput.* 2020;49(7):1665-1677. doi:10.1080/03610918.2019.1577973
11. Hartley HO, Ross A. Unbiased ratio estimators. *Nature.* 1954;174:270-272.
12. Cekim HO, Kadilar C. In-type variance estimators in simple random sampling. *Pak J Stat Oper Res.* 2020;689-696.
13. Upadhyaya LN, Singh HP. An estimator of population variance that utilizes the kurtosis of an auxiliary variable in sample surveys. *Vikram Math J.* 1999;19:14-17.
14. Kadilar C, Cingi H. Ratio estimators for population variance in simple and stratified sampling. *Appl Math Comput.* 2006;173:1047-1058.
15. Kadilar C, Cekim HO. Hartley-Ross type variance estimators in simple random sampling. *AIP Conf Proc.* 2017;1863:560008. doi:10.1063/1.4992691
16. Yadav SK, Mishra SS, Shukla AK, Tiwari V. Improvement of estimator for population variance using correlation coefficient and quartiles of the auxiliary variable. *J Stat Appl Probab.* 2015;4(2):259.
17. https://www.statista.com/statistics/1102935/coronavirus-cases-by-region-in-russia/. Accessed September 2021.

**How to cite this article:** Oncel Cekim H. Modified unbiased estimators for population variance: An application for COVID-19 deaths in Russia. *Concurrency Computat Pract Exper.* 2022;34(22):e7169. doi: 10.1002/cpe.7169