The Bootstrap Stratified Random Sampling in Finite Population for Traffic Survey Data

Kristiana Yunitaningtyas[1*], Indahwati[2], Muhammad Nur Aidi[3], Santi Susanti[4]
{kristiana_yunitaningtyas@apps.ipb.ac.id[1], indah.stk@gmail.com[2], Nuraidi@yahoo.com[3]}

Student of Department of Statistics, IPB University, Indonesia[1],
Lecturer of Department of Statistics, IPB University, Indonesia[2,3]
Transportation Agency, Sukabumi City, Indonesia[4]

Abstract. Traffic survey is an important technique used to measure traffic density and gas emissions produced by vehicles but generally it is carried out for a long period of time. This study aims to apply stratified random sampling to traffic survey data so as to improve the process of data collection and efficiency with a high degree of accuracy. The data is divided into strata based on traffic density and is implemented using the direct bootstrap resampling technique by paying attention to the finite population correction factor. The bootstrap in finite population is expected to resolve the overestimate variance due to the standard bootstrap. Evaluation is done by looking at the criteria of validity, reliability, and accuracy of the bootstrap statistics. The results indicated that the bias and variance decrease when bootstrap replication is large. Bootstrap sample size of 32 produced the lowest distribution of bias, adjusted variance, and MSE value.

Keywords: bootstrap, finite population, stratified random sampling, traffic survey.

1 Introduction

Traffic surveys are an important factor that forms the basis of traffic management because information can be gained from this activity regarding the number of vehicles at certain places and intervals of time. In addition, the results of the traffic surveys can be a reference for calculating emissions or exhaust gases produced by motorized vehicles. Heavy traffic will affect the quality of the surrounding air which can cause pollution and have a negative impact on public health. In general practice traffic surveys are still done manually and require a long period of time spent in data collection, at least 16 hours per day continuously. Furthermore, substantial qualified resources are needed and these two aspects can also increase the observational costs.

The results of traffic surveys from previous studies are used as a tool to improve and suggest a better method through survey sampling. The data obtained is a finite population and has small size so that a bootstrap technique is applied in the analysis. Generally the bootstrap techniques implicitly assume that the data is from an infinite population [1]. Thus, the correction factor needs to be taken into account so that the parameter estimators and the variance can capture the finiteness of the population. Various research have been carried out to look at the effect of bootstrap on sampling and to develop bootstrap applications in finite populations, such as the bootstrap on sample surveys were showed and the main developments of the bootstrap methodology for sample surveys were introduced [2], bootstrap methods in finite population sampling were reviewed and variance estimation also the constructions of confidence intervals were covered [3], and in bibliometric studies bootstrap analysis were used and applied to a finite population [1]. However, most of the research on the bootstrap

ICSA 2019, August 02-03, Bogor, Indonesia

Copyright © 2020 EAI

DOI 10.4108/eai.2-8-2019.2290544

technique applications used simple random sampling method in the analysis, little attention has been paid to the other sampling methods that can provide better results than simple random sampling.

This paper presents implementation of bootstrap on finite population using stratified random sampling as the alternative of commonly used method, simple random sampling. Stratified random sampling is chosen because the samples produced are able to represent each characteristics of the population. The data have different levels of traffic density between time intervals so simple random sampling is not appropriate to be used because the samples could only be taken from one certain density level. The aim of this study is to suggest a better approach which is more efficient but on the other hand still be able to estimate the statistical parameters with high degree of accuracy.

2 Data

The population data used in this study is a part of the research "Management of Traffic and Parking Flow in the Context of Air Quality Improvement in the Central Business District in Sukabumi City" which was conducted in March to June 2018. The data is obtained from traffic counting process in Jalan R.E Martadinata which is the busiest road in the central business district area. It consists of count data based on the types and total number of vehicles. The data was collected between 05.00 and 21.00 at 15 minute intervals. The vehicles were categorized namely into motorcycle, car, public passenger car, bus (small, medium, large), truck (medium, large, and trailer), and non–motorized. Traffic counting was carried out in two days that have highest traffic density, specifically on Monday to represent the workdays and Saturday to represent the weekends. In this paper, only data that was taken on Saturday is used because the analysis steps are the same for Monday.

3 Methods

Before the sampling method was implemented, data exploration process was done to identify the characteristics of the population data. The results are shown in tables and graphics. Strata were formed based on traffic density per 15 minutes, mainly when the density is low, medium, and high. The population data consists of intervals of every 15 minutes from 05.00-21.00, that is 64 intervals. The vehicle number is known from each interval. Then these intervals were divided into the strata according to the vehicle number or traffic density. The initial sample size of n and the bootstrap sample size $n^{'} = (n-1)/(1-f)$ [4] that is able to capture the finite population correction factor were determined.

Simple random samples were taken from each stratum without replacement and were concatenated to get bootstrap samples s^*. The allocation of samples taken from all the strata was determined by following Neyman allocation [5] :

$$n_i{}' = n'\left(\frac{N_i\sigma_i}{{}_{k=1}^{L} N_k\sigma_k}\right. \tag{1}$$

where $n_i{}'$ stands for sample size of i-th stratum, $i = 1, 2, 3. N_i$ stands for the population size of i-th stratum, σ_i is the standard deviation of i-th stratum and $\sum_{k=1}^{L} N_k\sigma_k$ is the sum between the multiplication of the population size of each stratum and its standard deviation. Statistic $\hat{\theta}^*$ was calculated from concatenated samples s^*. Sampling of all the strata was repeated B times so the bootstrap statistics were obtained $(\hat{\theta}_1^*, \dots, \hat{\theta}_b^*)$.

Bias bootstrap estimation was calculated by the equation :

$$\widehat{\text{Bias}}(\hat{\theta}) = \hat{\theta}^{*}_{(.)} - \hat{\theta}_{\text{b}} \qquad (2)$$

where$\hat{\theta}^{*}_{(.)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*}_{b}$ is the mean value of estimators from replication of bootstrap statistics B. Bias occurs when statistics from samples do not accurately reflect the true value of the population [6]. Variance of $\hat{\theta}$ was estimated by [7] :

$$\hat{V}_{boot}\left(\hat{\theta}\right) = \frac{1}{B-1}\sum\nolimits_{b=1}^{B}\left(\hat{\theta}^{*}_{b} - \hat{\theta}^{*}_{(.)}\right)^{2}. \qquad (3)$$

Adjusted variance and adjusted standard error was computed from the estimate variance to capture the existence of a finite population correction factor [1]. Adjusted variance is written :

$$V^{*'} = \frac{N-n}{N-1}\hat{V}_{boot}\left(\hat{\theta}\right) \qquad (4)$$

and adjusted standard error :

$$SE= \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad (5)$$

Based on estimation of bias andvariance, mean square error (MSE) is obtained by :

$$MSE\left(\hat{\theta}\right) = \hat{V}_{boot}\left(\hat{\theta}\right) + \left(\widehat{\text{Bias}}(\hat{\theta})\right)^{2}. \qquad (6)$$

The steps as mentioned above were repeated for different values of n and n'. The criteria used in determining the best bootstrap sample are based on validity, reliability and accuracy. The validity of an estimator can be evaluated by examining the bias of the estimator. The reliability of an estimator can be stated in terms of its sampling variance or standard error. The accuracy of an estimator is evaluted on the basis of its mean square error (MSE) [8].

4 Results and Discussion

4.1 Data Exploration

The description of population data is presented in Table 1. It describes the total, mean, variance, standard deviation, the minimum, maximum number of vehicles that pass in Jalan R.E Martadinata on Saturday besides also the information about the number of intervals in every 15 minutes, which is the population size.

Table 1. The summary of population data

Population Data	Count
Total Vehicles (τ)	79327
Mean of Vehicles (per 15 minutes) (μ)	1239
Variance of Vehicles (per 15 minutes) (σ^2)	212 023
Standard Deviation of Vehicles (per 15 minutes) (σ)	460
Minimum Number of Vehicles	340
Maximum Number of Vehicles	2826
Population Size/Total Number of Intervals (N)	64

Figure 1 shows a graph of the number of vehicles according to the time intervals of every 15 minutes on Saturday. The vehicles were counted from 05.00 to 21.00. The peak of traffic density on Saturday occured at 06.00 to 06.15. Traffic density that tends to be heavy on

Saturday also occured in the afternoon, which was at 16.45 to 17.00. From the figure it can be said that the number of vehicles that pass in Jalan R.E. Martadinata is fluctuative at known time intervals.

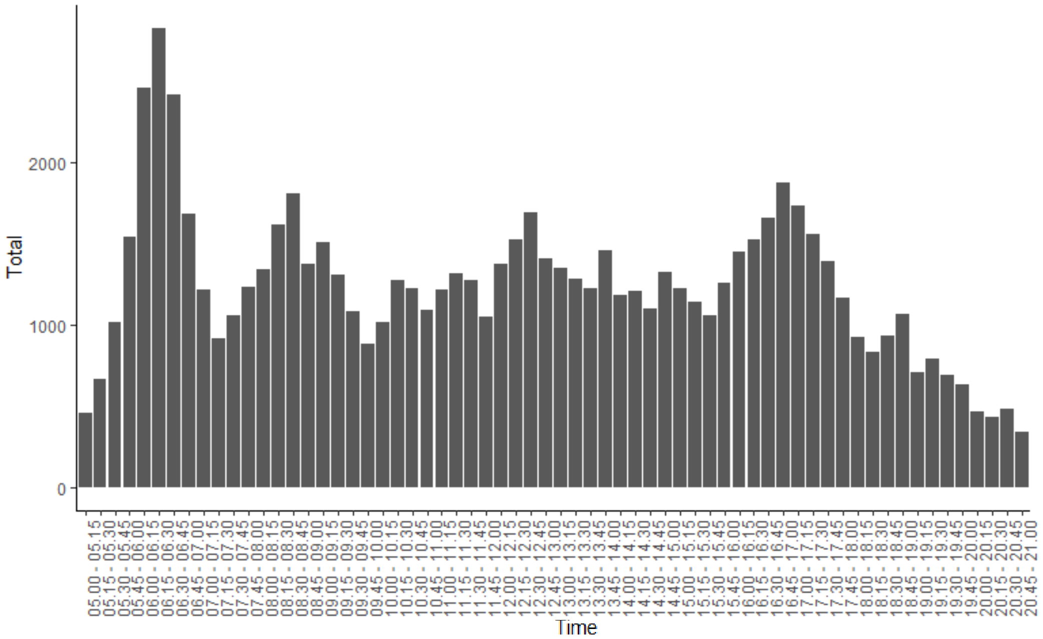

Fig. 1. Number of vehicles according to time interval

4.2 Stratification

The principle in stratified random sampling is that the population is partitioned into several groups called strata so that the units within the stratum are as similar as possible [9].The analysis was carried out by considering the finiteness of the population so that the bootstrap sample size n' was determined. The time factor was used to form strata in this study, with each interval of time consisting of observations in every 15 minutes .

Table 2 shows the grouping of every time interval according to the traffic density. Stratum 1 has the smallest mean and variance as well as the number of elements, while Stratum 2 and Stratum 3 have the same number of time intervals. Stratum 3 has the greatest mean and variance among all the strata formed. Samples of size n were taken to estimate the total number of vehicles from all strata. After n was determined then this sample size was allocated according to each stratum. This allocation is influenced by the number of elements and the costs needed to obtain observations in each stratum. In this study the costs required for all strata were assumed to be the same so Neyman allocation was used. The larger the units and the standard deviation, the more samples are taken.

Table 2. The distribution of time intervals into the strata

Strata	Time Intervals (at 15 minutes)	N_i	Number of Vehicles	Mean	Variance
1 Low Traffic	05.00-06.00 18.00-19.00 19.00-20.00 20.00-21.00	16 time intervals	39183 vehicles	748 vehicles	86642 vehicles
2 Medium Traffic	07.00-08.00 09.00-10.00 10.00-11.00 11.00-12.00 14.00-15.00 15.00-16.00	24 time intervals	37942 vehicles	1174 vehicles	141430 vehicles
3 High Traffic	06.00-07.00 08.00-09.00 12.00-13.00 13.00-14.00 16.00-17.00 17.00-18.00	24 time intervals	36032 vehicles	1633 vehicles	166437 vehicles

4.3 Bootstrap Analysis on Stratified Random Sampling

The results of different bootstrap technique replications are presented in Table 3. It can be seen that the greater the bootstrap replication, the bias, variance and adjusted variance are smaller. Bootstrap replication, B= 1000 was selected because it is able to produce smaller values of bias and variance significantly.

Table 3. Estimation of total population, bias, variance and adjusted variance from different bootstrap replication (B)

B	n	n'	$\hat{\tau}$	Bias	Variance	Adjusted Variance
10	8	8	79381	-99.88	562094414	383651743
50	10	10	79383	-91.29	103269497	70485530
100	14	16	79387	-83.32	51417465	34278310
500	17	22	79460	-67.69	9418391	6278927
1000	22	32	79433	-3.14×10^{-12}	4246502	2831002

The results of the analysis with bootstrap size, B = 1000 in estimating the number of vehicles using stratified random sampling can be seen in Table 3. Stratified random sampling method produced a bias value that tends to be in the range of 0 even though there is no tendency for the larger sample size the bias will be smaller.The smallest bias was obtained when the bootstrap sample is 16.

Table 4 presents the standard bootstrap variances without adjustment that are greater than the variances with adjustment because they cannot capture the finite population correction factor. The existence of this finiteness in population could reduce the variance and increasethe level of accuracy. Figure 2 shows the comparison between the variance and the adjusted variance of bootstrap.

Table 4. Estimation of total population, bias, variance, adjusted variance and standard error from various sample sizes with bootstrap replication of 1000

n	n'	$\hat{\tau}$	Mean Bias	Mean Variance	Mean Adjusted Variance	Mean Adjusted Standard Error
8	8	79400	0	41 117 392	36512 244	2137.43
10	10	79 276	5.53 x 10^{-13}	22802176	19525178	1398.02
14	16	79 298	-3.16 x 10^{-12}	13 246 541	10 082 510	794.22
17	22	79216	-1.35 x 10^{-12}	8 853 189	5 896 224	517.96
22	32	79433	-3.14 x 10^{-12}	4 246 502	2 154 797	259.62

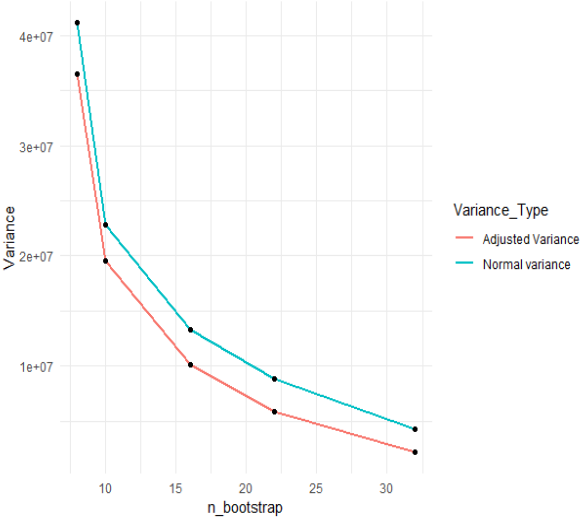

Fig. 2. Comparison of standard variance and adjusted variance of bootstrap

The smallest variance was obtained from the bootstrap sample size of 32. According to the comparison from Table 4 and Fig. 2, it can be seen that the larger the sample size, the smaller the variance. Sample size of 32 produced mean of bias value distribution that is smallest than the other sample sizes, thus it is the best sample size based on validity.. The bias and variance values obtained are then used to calculate mean square error from the method as a measure of accuracy criteria. Figure 3 shows the boxplots of bias obtained from bootstrap to see the details of bias distribution produced by every sample size.

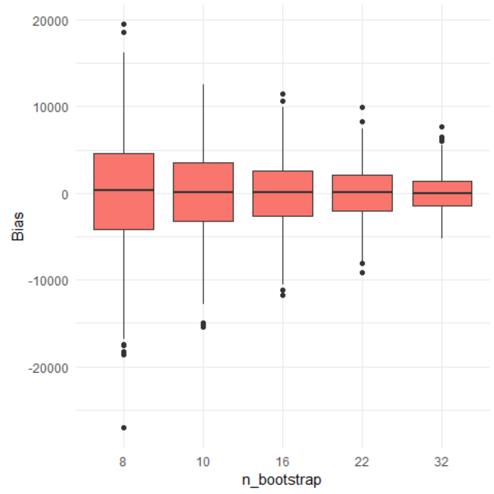

Fig. 3. Bias distribution of bootstrap sample sizes

From Fig. 3, it can be seen that for any sample size, stratified random sampling has bias distribution with the median around the value of 0 although there are outlier observations. In general, stratified random sampling method is unbiased and consistent. In terms of validity, stratified random sampling with the sample size of 32 is the best among other sizes because it can produce the smallest range of bias distribution. So sample size of 32 is chosen to estimate the number of vehicles by stratified random sampling applied for the data based on validity criteria.

Another criteria in determining the best bootstrap sample is reliability which can be seen from the standard error values. Comparison of standard errors from each sample size is shown in Figure 4.

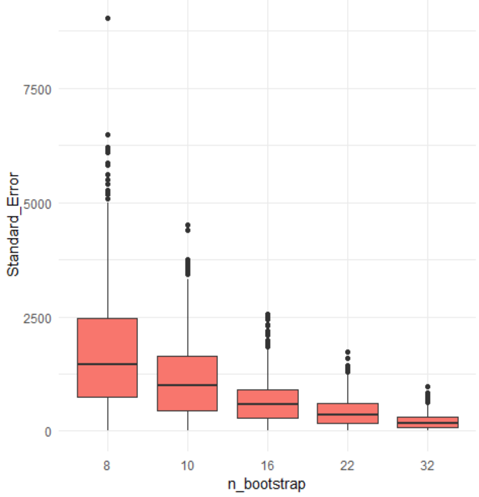

Fig. 4. Standard error distribution of bootstrap sample sizes

Sample size of 32 produced standard error distribution range that is smallest than the other sample sizes, thus it is the best sample size based on reliability. According to Fig. 4, the larger the sample size, the smaller the standard error distribution. On the other hand, the value of variance is proportional to value of mean square error. A high level of accuracy in estimating parameters is also one of the characteristics of a good estimator. The accuracy of an estimator is evaluated based on the mean square error. The smaller mean square error indicates that the accuracy is higher. Mean square error includes validity and reliability criteria. The calculation of mean square error based on adjusted variance for each sample size is described on Table 5 below.

Table 5. Mean square error for the bootstrap sample sizes

n	n'	$\hat{\tau}$	MSE
8	8	79400	36 512 244
10	10	79013	19 525 178
14	16	79239	10 082 510
17	22	79216	5 896 224
22	32	79433	2 154 797

4.4 Parameter Estimation

Estimating the number of vehicles based on the types was done from the estimated total vehicles obtained through the best sample size. The total estimation of vehicles by stratified random sampling has lowest distribution of bias and adjusted standard error as well as the lowest mean square error from the sample size of bootstrap 32 which is 79433. Table 6 shows the estimation of vehicle numbers based on the types. The estimation result shows that there is no significant difference compared to the population of each type of vehicle. This indicates that the proportion from the population can be used to estimate the number of each type of vehicle.

Table 6. Estimation of vehicles population based on vehicle types

Type of Vehicles	Number of Vehicles	Proportion	Number of Vehicles Estimation
Motorcycle	69324	0.87390	69417
Car	6558	0.08267	6567
Public Passenger Car	3264	0.04115	3268
Pick-up Car	143	0.00180	143
Small Bus	3	0.00004	3
Medium Bus	0	0.00000	0
Big Bus	8	0.00010	8
Medium Truck	10	0.00013	10
Big Truck	0	0.00000	0
Trailer Truck	4	0.00005	4
Non-motorized	13	0.00016	13
Total	79327	1	79433

5 Conclusion

In estimating parameters, it is necessary to acknowledge the characteristics of the population whether finite or infinite. In finite populations, the adjusted variance leads a smaller value than a normal bootstrap variance and it can increase the accuracy because it is able to capture the correction factor. Traffic counting process using stratified random sampling method can run more efficiently in terms of time, resources, and costs compared to the standard method, simple random sampling. By stratified random sampling, the surveys are not carried out for a long period of time throughout the day. From the simulation, bootstrap sample size of 32 produced the lowest distribution of bias, adjusted standard error, and mean square error value with an estimated number of vehicles in Saturday is 79433. The difference between the estimated values and parameters is not significant. Bootstrap technique can be applied in survey sampling methods especially when the population is finite and the sample size is small.

Acknowledgement
We thank our lecturers from Department of Statistics, IPB University for sharing insight and expertise. We would also like to show our gratitude to our colleague for providing the data that greatly assisted the research.

References

[1]Nane, T and Kooijman, K.: A bootstrap analysis for finite populations. (2018)
[2]Shao, J.: Impact of the Bootstrap on Sample Surveys 18 191–8. (2003)
[3]Mashreghi Z, Haziza D and Christian L.: A survey of bootstrap methods in finite population sampling. Statistics. Surveys. 10 1–52. (2016)
[4] McCarthy, PJ and Snowden, CB.: The bootstrap and finite population sampling. Data Evalution and Methods Research. 2 1–23.(1985)
[5]Scheaffer, RL.: Elementary survey sampling. Brooks/Cole, Cengage Learning, US. (2012)
[6]McCutcheon, AL.: Sampling Bias. Encyclopedia of Survey Research Methods.(2011)
[7]Quatember, A.:The Finite Population Bootstrap - from the Maximum Likelihood to the Horvitz-Thompson Approach. 43 93–102. (2014)
[8]Levi, PS and Lemeshow, S.: *Sampling of Populations*. John Wiley & Sons, Inc.,US. (2014)
[9]Thompson, SK.: *Sampling*. John Wiley & Sons, Inc., US. (2012)