Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 14(1) 2021: 1-9
http://ejournal.undip.ac.id/index.php/media_statistika

SMALL AREA ESTIMATION METHOD WITH EMPIRICAL BAYES BASED ON BETA BINOMIAL MODEL IN GENERATED DATA

Ferra Yanuar, Rahmatika Fajriyah, Dodi Devianto
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Andalas

e-mail: ferrayanuar@sci.unand.ac.id

DOI: 10.14710/medstat.14.1.1-9

Article Info:
Received: 26 August 2019
Accepted: 21 October 2020
Available Online: 30 June 2021

Keywords:
SAE; *Empirical Bayes Method*; *Beta-Binomial Model.*
.

Abstract: Small Area Estimation is one of the methods that can be used to estimate parameters in an area that has a small population. This study aims to estimate the value of the binary data parameter using the direct estimation method and an indirect estimation method by using the Empirical Bayes approach. To illustrate the method, we consider three conditions: direct estimator, empirical Bayes (EB) with auxiliary variables, and empirical Bayes without auxiliary variables. The smaller value of Mean Square Error is used to determine the better method. The results showed that the indirect estimation methods (EB method) gave the parameter value that was not much different from the direct estimation value. Then, the MSE values of indirect estimation with an auxiliary variable are smaller than the direct estimation method.

1. INTRODUCTION

Data collection is an important aspect of statistics. Data collection can be done by surveying to obtain more detailed information on the entire population's scope. Surveys are usually carried out periodically by the government or a certain agency. In a certain time scale, the government regularly conducts surveys to obtain statistical data in the national scope (Ariwijayanthi et al., 2013; Noviyanti et al., 2014).

With the implementation of regional autonomy policies implemented, local governments need data for small areas in their regions. This data is needed to take the right policies in solving problems in regional development. Still, the surveys carried out by the government are generally only at the city or provincial level. This means that if the government wants to survey in a smaller area, it requires a large sample size and requires a lot of money and time. Moreover, censuses and administrative records have limited scope. This is why local governments have difficulty obtaining data for this condition (Eka Putri et al., 2019; Hasanudin et al., 2011; Ikhsan et al., 2018).

A small area is defined as a subset of the population with a variable of interest. Small areas generally represent small geographic areas such as cities, regencies, sub-districts, villages, or *kelurahan* (Eka Putri et al., 2019; Juriah et al., 2019; Kusuma et al., 2017). A direct estimator is a classic approach to estimating small area parameters based on applying a design-based model. The problem arises when it turns out that this small area is not taken as a sample in the national survey. If the direct estimation is done, it tends to produce poor statistical values because the sample size at the regional level is usually very small. Thus,

the statistics obtained will have a large variation; even estimations cannot be done because they are not represented in the survey. Therefore, a method is needed that can provide better predictions which is more efficient, cost-effective, and can produce better statistics.

Small Area Estimation is a method used to reduce the variance value in a small area, namely by using the indirect estimation method by utilizing information from the surrounding area. Various methods of estimating small areas have been developed, especially regarding the model-based method. Several methods that are classified as model-based methods are the Empirical Bayes (EB) and Hierarchical Bayes (HB) methods (Jiming Jiang & P. Lahiri, 2001).

Binary data is data that only has two possible events, namely a successful event and a failure event. Events that are considered successful are symbolized by the number 1, while failure events are symbolized by the number 0. Binary data were chosen because there are many problems in an area that are classified as binary data. Such as literacy or non-literary rates, dropout or non-dropout rates, ownership numbers or not health insurance, etc. The number of successful events is stated to follow the binomial distribution if it meets the assumption that the response variables are independent of each other and have the same chance of success. The SAE EB method was chosen because the components of the variety were estimated from the data.

The modeling using SAE-EB approaches have been employed by many researchers, such as by Kismiantini (Kismiantini, 2010), Abadi (Abadi, 2011), and Kusuma et al. (2017). Kismiantini estimated health card ownership status in the city of Yogyakarta using the SAE EB Beta-Binomial model. Abadi estimated the proportion of poor households that were approached from per capita household expenditure using three methods: direct estimator, synthetic estimator, and SAE EB method. Kusuma et al. (2017) applied the SAE HB and SAE EB methods in estimating per capita expenditure in Banyuwangi with the SAE EB method and the SAE HB method.

This study focuses on comparing the results of estimating response data with a Binomial distribution using direct (MLE method) and indirect estimation methods (SAE-EB method). The model hypotheses are constructed in two conditions, namely with auxiliary variable and without auxiliary variable. The generated data is used in this study which is generated using RStudio version 3.6.1 (Molina & Marhuenda, 2015). Then the MSE (Mean Square Error) value is estimated as the criteria to determine the better method which produces the smallest value of MSE.

2. LITERATURE REVIEWS

2.1. Direct Estimation Based on Binomial Response Variable

Let variable z_i is a binary response variable for area *i*-th. Parameter $z_i = 1$ represents success at area *i*-th or $z_i = 0$ for fail. If the variable z_i is assumed to have a Bernoulli distribution with parameters θ_i. then the probability density function of z_i is:

$$f(z_i|\theta_i) = \theta^{z_i}(1-\theta^{z_i}) \tag{1}$$

Written as $x_i|\theta_i \overset{ind}{\sim} bernoulli(\theta_i)$ for $i = 1,2,...,n$. It is defined that $y_i = \sum_{i=1} z_i$ represents a number of successful occurrences out of n repetitions of concern in area *i*-th, so that y_i will have a Binomial distribution (n,θ) with the probability density function:

$$f(y_i|\theta_i) = \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n-y_i} \tag{2}$$

for $y_i = 1,2,\ldots,n$. $0 < \theta_i < 1$. or written as $y_i|\theta_i \overset{ind}{\sim} Binomial(n_i, \theta_i)$.

The parameter to be estimated in the small area estimation method is the proportion of a small area, symbolled by θ_i . If the sampling is carried out using the simple random method, then the proportion of estimator in area *i*-th is:

$$\hat{\theta}_i = \frac{y_i}{n_i} \tag{3}$$

The value $\hat{\theta}_i$ are obtained from the maximum likelihood estimation (MLE) method. The MLE method is one of the methods of direct estimation. This maximum likelihood estimator is unbiased because the expected value of the estimator is the same as the estimated parameter or it can be written:

$$E(\hat{\theta}_i) = \theta_i$$

So that the estimated mean square error (MSE) for $\hat{\theta}_i$ is the same as the variance for $\hat{\theta}_i$. Or it can be written with $MSE(\hat{\theta}_i) = \text{var}(\hat{\theta}_i)$.

2.2. Empirical Bayes Method Based on the Beta-Binomial Method

The estimating parameter θ_i can be done indirectly by utilizing additional information from accompanying variables through the Bayes approach. The estimated parameter in the Bayes method is assumed to have a certain distribution. There are two pieces of information in Bayes' estimation, namely prior distribution and posterior distribution (Maulina et al., 2019). For the Binomial distribution, one of the prior distributions used is the Beta distribution (Rao & Molina, 2015). The Binomial Beta Distribution begins with the assumption that a random variable Y is assumed to spread according to the Binomial distribution with the parameter (n, θ) or written as $Y \sim \text{Binomial}(n, \theta)$ with θ is the chance of success. It also assumes that the probability of success θ is spread according to the Beta distribution, or written $\theta \sim Beta(\alpha, \beta), \alpha > 0, \beta > 0$ with the probability distribution function for θ is as following (Torabi et al., 2009):

$$f(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, 0 \leq \theta \leq 1 \tag{4}$$

Meanwhile, the posterior distribution is defined as follows:

$$f(\theta|y) = \frac{f(y|\theta) f(\theta)}{f(y)} \tag{5}$$

Based on the likelihood function and prior distribution defined above, the posterior distribution for θ could be constructed using Equation (5), which then are formed as:

$$f(\theta|y|) = \frac{1}{B(y+\alpha, n-y+\beta)} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \tag{6}$$

In Bayes theorem, Bayes estimate for θ is mean of posterior distribution:

$$\hat{\theta}^{B} = E\left(\hat{\theta}\middle|y,\alpha,\beta\right) = \frac{\left(y+\hat{\alpha}\right)}{\left(n+\hat{\alpha}+\hat{\beta}\right)}$$

and variance for θ is:

$$Var\left(\hat{\theta}\middle|y,\hat{\alpha},\hat{\beta}\right) = \frac{\left(y+\hat{\alpha}\right)\left(n-y+\hat{\beta}\right)}{\left(n+\hat{\alpha}+\hat{\beta}\right)^{2}\left(1+n+\hat{\alpha}+\hat{\beta}\right)}$$

To estimate α and β, this study employ the Kleinman moment method (Rao & Molina, 2015; Torabi et al., 2009). According to Rao & Molina (2015), the Kleinman moment method is an easy and appropriate alternative to use in estimating α and β values. The form of the Kleinman moment method equation is:

$$\hat{\theta} = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} \text{ and } \frac{1}{\hat{\alpha}+\hat{\beta}+1} = \frac{n_T s_{\theta}^{2} - \hat{\theta}\left(1-\hat{\theta}\right)\left(m-1\right)}{\hat{\theta}\left(1-\hat{\theta}\right)\left[n_T - \sum_{i=1}^{m} n_i^{2} - \left(m-1\right)\right]}$$

With mean of weighted sample is:

$$\hat{\theta} = \sum_{i=1}^{m}\left(\frac{n_i}{n_T}\right)\hat{\theta}_i \; total$$

And variance of weighted sample is:

$$\widehat{\theta} = \sum_{i=1}^{m}\left(\frac{n_i}{n_T}\right)\left(\widehat{\theta}_i - \widehat{\theta}\right)^{2}$$

With $\hat{\theta}_{\hat{i}}$ is the direct estimate of the *i*-th proportion. *m* is the number of areas. n_i is the number of elements taken in area *i*-th and n_T is the total number of elements taken in area *i*-th. Thus. the value for $\widehat{\alpha}$ and $\widehat{\beta}$ are obtained (Rao & Molina, 2015; Zaja et al., 2019):

$$\hat{\alpha} = \hat{\theta}\left[\frac{\hat{\theta}\left(1-\hat{\theta}\right)\left[n_T - \sum_{i=1}^{m} n_i^{2} - \left(m-1\right)\right]}{n_T s_{\theta}^{2} - \hat{\theta}\left(1-\hat{\theta}\right)\left(m-1\right)} - 1\right] \tag{7}$$

$$\hat{\beta} = \hat{\theta}\left[\frac{\hat{\theta}\left(1-\hat{\theta}\right)\left[n_T - \sum_{i=1}^{m} n_i^{2} - \left(m-1\right)\right]}{n_T s_{\theta}^{2} - \hat{\theta}\left(1-\hat{\theta}\right)\left(m-1\right)} - 1\right]\left[\frac{1}{\hat{\theta}} - 1\right] \tag{8}$$

Based on Equation (7) and Equation (8) it is found that $\hat{\alpha}$ and $\hat{\beta}$ are the estimated value of the Beta-Binomial distribution parameter. Furthermore, the substitution of values $\hat{\alpha}$ and $\hat{\beta}$ to the following equation for θ_i^{EB}:

$$\hat{\theta}_i^{EB} = \hat{\theta}_i^{B}\left(\hat{\alpha},\hat{\beta}\right) = \hat{\gamma}_i\hat{\theta}_i + \left(1-\hat{\gamma}_i\right)\hat{\theta} \tag{9}$$

with $\hat{\gamma}_i = \frac{n_i}{n_i + \hat{\alpha} + \hat{\beta}}$. $\hat{\alpha}$ as Equation (7) and $\hat{\beta}$ is as Equation (8).

2.3. The Estimation for MSE Using Jackknife Method

The Jackknife method is one method that is often used because of its simple concept (Jiming Jiang & P. Lahiri, 2001; K. Ogundeji et al., 2013). This method was introduced by Tukey, as explained in Rao & Molina (Rao & Molina, 2015) and developed into a method that can correct an estimator's bias. The procedure is done by deleting the *i*-th observation for every *i* = 1.2.…. *m*. The estimation for MSE using the Jackknife method are as follows:

$$MSE_J(\hat{\theta}_i^{EB}) = \hat{M}_{1i} + \hat{M}_{2i} \tag{10}$$

with

$$\hat{M}_{1i} = g_{1i}(\hat{\alpha}, \hat{\beta}, y_i) - \frac{m-1}{m}\sum_{l=1}^{m}\left[g_{1i}(\hat{\alpha}_{-l}, \hat{\beta}_{-l}, y_i) - g_{1i}(\hat{\alpha}, \hat{\beta}, y_i)\right]$$

and

$$\hat{M}_{2i} = \frac{m-1}{m}\sum_{l=1}^{m}(\hat{\theta}_{i,-l}^{EB} - \hat{\theta}_i^{EB})^2$$

With θ_i^{EB} is Bayes estimates for θ_i. $\hat{\theta}_{i,-l}^{EB}$ is Bayes estimate for $\theta_{i,-l}$. $\hat{\alpha}_{-l}$ and $\hat{\beta}_{-l}$ is an estimate for *l*-th area that is deleted.

3. DATA AND METHOD

3.1. Data

The generated data are used in this study using the R i386 3.1.0. The response variable is assumed to have a binomial distribution with parameters (n_i, θ_i) that can be written by $y_i \sim Binomial(15,0.7)$ which x_i and n_i are specified. The sample sizes are 25 observations. The calculation process was carried out using Microsoft Excel and SAS 9.5.

3.2. Parameter Estimation Using SAE EB with Beta Binomial Model

Parameter estimation process based on SAE EB with Beta Binomial Model carried out in this study follow these stages:

1. Generate 15 random pieces of data that are binomial distribution with parameters (15.0.7) or can be written $y \sim Binomial(15,0.7)$.
2. Calculate the mean and variance of proportions using direct.
3. Calculate the mean and variance of proportions using the weighted mean and variance.
4. Estimate the parameters of the beta-binomial distribution, $\widehat{\alpha}$ and $\widehat{\beta}$.
5. Determine the posterior distribution.
6. Determine the Empirical Bayes estimator for $\widehat{\theta}_i^{EB}$.
7. Calculate and compare the Mean Square Error (MSE) value of the direct estimator and the indirect estimator using the Jackknife method.

4. RESULTS

It is assumed that the response variable $y_i \sim Binomial(n_i, \theta_i)$. Thus, a direct estimator for θ_i with auxiliary variable is estimated using following equation:

$$\ln\left(\frac{\hat{y}_i}{n_i}\right) = \hat{\beta}_0 + \hat{\beta}_i x_1$$

Table 1. Parameter Estimated Based on Direct Estimation Method

Parameter	Mean	Standard Error
$\widehat{\beta_0}$	6.247	43.709
$\widehat{\beta_1}$	-0.0508	0.538

Therefore. proposed model for θ_i based on direct estimation method is:

$$\ln\left(\frac{\hat{y}_i}{n_i}\right) = 6.4247 - 0.0508\mathrm{x}_1 \quad (11)$$

Furthermore. the acceptable of the model is determine based on plot of regression standardized residual and standardized predicted value. The proposed model could be accepted since the plot of data is distributed at random and no trend are detected. The corresponding plot of these results is provided in Figure 1.

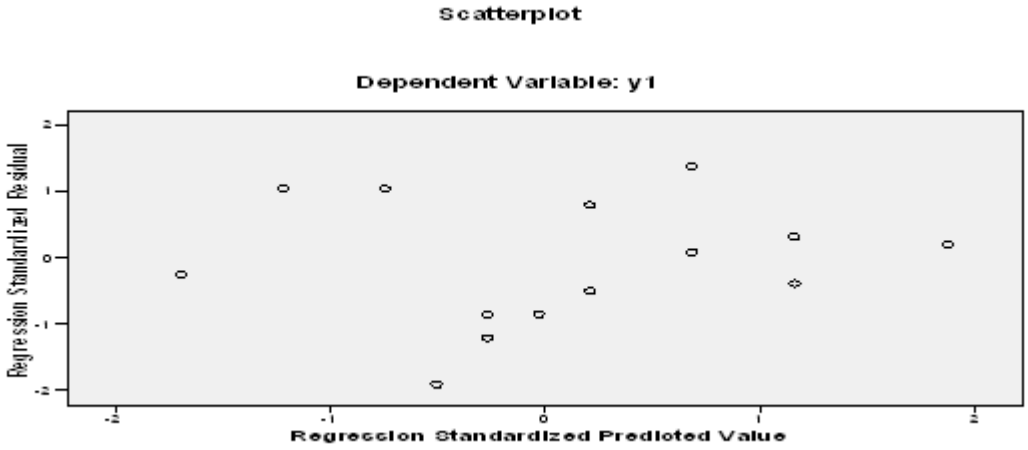

Figure 1. The Distribution of Pearson Residual

The next analysis is the parameter estimation based on indirect estimation method (SAE-EB) with auxiliary and without auxiliary variables. The results of these estimation for both direct and indirect estimation methods are presented in Table 2.

From Table 2 it can be seen that the estimated value obtained from all three estimating methods tend to produce almost the same value. Even though, it is important to determine the best estimator among these methods. The best estimator is determined based on the smallest value of MSE (Ikhsan et al., 2018). The estimated MSE values are presented in Table 3.

Table 2. Estimated Value for θ_i

No	n_i	y_i	Direct Estimation	Indirect Estimation	
				Without Auxiliary Variable	With Auxiliary Variable
1	113	92	0.8142	0.8267	0.8255
2	110	108	0.9818	0.9699	0.9690
3	102	97	0.9510	0.9432	0.9464
4	120	100	0.8333	0.8426	0.8417
5	114	108	0.9474	0.9408	0.9424
6	96	89	0.9271	0.9228	0.9250
7	105	101	0.9619	0.9525	0.9508
8	89	83	0.9326	0.9271	0.9298
9	95	92	0.9684	0.9572	0.9558
10	118	96	0.8136	0.8258	0.8258
11	122	95	0.7787	0.7952	0.7930
12	131	111	0.8473	0.8542	0.8549
13	110	108	0.9818	0.9699	0.9684
14	105	102	0.9714	0.9607	0.9587
15	121	106	0.8760	0.8795	0.8800

Table 3. Estimated Value for Mean Square Error (MSE)

No	n_i	y_i	Direct Estimation	Without Auxiliary Variable	With Auxiliary Variable
1	113	92	0.001333	0.001012	0.00083
2	110	108	0.000162	0.000161	0.00066
3	102	97	0.000457	0.000351	0.00086
4	120	100	0.001157	0.001023	0.00049
5	114	108	0.000437	0.000438	0.00044
6	96	89	0.000704	0.000504	0.00058
7	105	101	0.000349	0.000252	0.00049
8	89	83	0.000706	0.000606	0.00058
9	95	92	0.000322	0.000222	0.00050
10	118	96	0.001285	0.001185	0.00080
11	122	95	0.001413	0.001313	0.00087
12	131	111	0.000988	0.000978	0.00079
13	110	108	0.000162	0.000131	0.00045
14	105	102	0.000265	0.000246	0.00050
15	121	106	0.000898	0.000821	0.00061

According to Table above, we could see that MSE based on indirect method (with and without auxiliary variables) tend to yield smaller value than direct method. Among indirect methods, model with auxiliary variables result smaller value than without auxiliary variables. Thus, we conclude in this present study that indirect method with auxiliary variables as the best estimator. This result is similar with previous studies such as study by (Sari & Yanuar, 2020). Thus, SAE EB method could be employed to estimate parameter model for binary response variable for insufficient size sample cases.

5. CONCLUSIONS

The study concluded that the estimation of the proportion for θ_i using the EB method resulted in a better model than direct estimation method. Among EB method, model with auxiliary variable tended to result better model than without auxiliary variable. The criteria for determining better model is based on the smallest value of MSE (mean square error). The MSE value for indirect estimator (EB method) was estimated using Jackknife method since EB method is bias estimator. Meanwhile, the direct estimator is unbiased estimator thus is no need modification to estimate MSE.

REFERENCES

Abadi, S. (2011). Pendugaan Statistik Area Kecil Menggunakan Model Beta-Binomial. *Tesis*, Institut Pertanian Bogor.

Ariwijayanthi, P. E., Sumarjaya, I. W., & Bagus Oka, T. (2013). Penerapan Metode Pendugaan Area Kecil (Small Area Estimation) Pada Penentuan Proporsi Rumah Tangga Miskin Di Kabupaten Klungkung. *E-Jurnal Matematika*, *2*(3), 35. https://doi.org/10.24843/MTK.2013.v02.i03.p046

Eka Putri, N. C., Yanuar, F., & Yozza, H. (2019). Metode Bayes Empirik untuk Memodelkan Data Cacahan dengan Peubah Penyerta pada Pendugaan Area Kecil. *Jurnal Matematika UNAND*, *8*(1), 224. https://doi.org/10.25077/jmu.8.1.224-231.2019

Hasanudin, N., Padmadisastra, S., & Hajarisman, N. (2011). Pertimbangan Penting yang Mendasari Penggunaan Metode Small Area Estimation. *Prosiding Seminar Nasional Statistika*, *2*, 218–226.

Ikhsan, E., Hidayat, C. A., & Nurizza, W. A. (2018). Studi Kasus: Estimasi Persentase Penduduk Miskin di Provinsi Nusa Tenggara Timur Tahun 2017. *Jurnal Aplikasi Statistika & Komputasi Statistik*, 1–12.

Jiming Jiang & P. Lahiri. (2001). Empirical Bayes Prediction for Small Area Inference with Binary Data. *Annals of the Institute of Statistical Mathematics*, *53*(2).

Juriah, S., Yanuar, F., & Yozza, H. (2019). Pendugaan Penyebaran Penyakit Demam Berdarah Dengue (DBD) Di Kota Padang. *Jurnal Matematika UNAND*, *8*(1), 313. https://doi.org/10.25077/jmu.8.1.313-317.2019

K. Ogundeji, R., J. Adewara, A., & S. Nurudeen, T. (2013). Bayesian Sequential Estimation of Proportion of Orthopaedic Surgery Among Different Age Groups: A Case Study of National Orthopaedic Hospital, Igbobi-Nigeria. *International Journal of Statistics and Applications*, *2*(6), 108–113. https://doi.org/10.5923/j.statistics.20120206.03

Kismiantini. (2010). Penerapan Metode Bayes Empirik Pada Pendugaan Area Kecil Untuk Kasus Biner (Studi tentang Proporsi Status Kepemilikan Kartu Sehat di Kota Yogyakarta). *Seminar Nasional Penelitian, Pendidikan Dan Penerapan MIPA, FMIPA Universitas Negeri Yogyakarta*.

Kusuma, W., Iriawan, N., & Irhamah, I. (2017). Small Area Estimation of Expenditure Per-capita in Banyuwangi with Hierarchical Bayesian and Empirical Bayes Methods. *IPTEK Journal of Science*, *2*(3). https://doi.org/10.12962/j23378530.v2i3.a3185

Maulina, R. F., Djuraidah, A., & Kurnia, A. (2019). Pemodelan Kemiskinan Di Jawa Menggunakan Bayesian Spasial Probit Pendekatan Integrated Nested Laplace Approximation (INLA). *Media Statistika*, *12*(2), 140–151. https://doi.org/10.14710/medstat.12.2.140-151

Molina, I., & Marhuenda, Y. (2015). sae: An R Package for Small Area Estimation. *The R Journal*, *7*(1), 81–98. https://doi.org/10.32614/RJ-2015-007

Noviyanti, R. A., Zain, I., Sarjana, M. P., Hakim, J. A. R., & Timur, J. (2014). Pendekatan Small Area Estimation Pada Scan Statistic Untuk Pendeteksian Kantong Kemiskinan. *Prosiding Seminar Nasional Matematika, Universitas Jember*, 73–89.

Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation* (Second edition). John Wiley & Sons, Inc.

Sari, A. D., & Yanuar, F. (2020). Hierarchical Bayesian Modelling in Small Area for Estimating Binary Data. *Journal of Physics: Conference Series*, *1554*, 012049. https://doi.org/10.1088/1742-6596/1554/1/012049

Torabi, M., Datta, G. S., & Rao, J. N. K. (2009). Empirical Bayes Estimation of Small Area Means under a Nested Error Linear Regression Model with Measurement Errors in the Covariates. *Scandinavian Journal of Statistics*, *36*(2), 355–368. http://www.jstor.org/stable/41000325

Zaja, N., Yozza, H., & Yanuar, F. (2019). Small Area Estimation Dengan Pendekatan Empirical Bayes Berbasis Model Beta-Binomial Untuk Menduga Angka Pengangguran Di Sumatera Barat. *Jurnal Matematika UNAND*, *8*(1), 120. https://doi.org/10.25077/jmu.8.1.120-127.2019