Available online at http://www.journalcra.com

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

International Journal of Current Research
Vol. 11, Issue, 08, pp.6039-6044, August, 2019

DOI: https://doi.org/10.24941/ijcr.36306.08.2019

ISSN: 0975-833X

RESEARCH ARTICLE

ON COMPARISON SOME ESTIMATORS IN SMALL AREA STUDY

*William W.S. Chen

Department of Statistics, the George Washington University, Washington D.C. 20013

ARTICLE INFO

Article History:
Received 24th May, 2019
Received in revised form
16th June, 2019
Accepted 20th July, 2019
Published online 31st August, 2019

Corresponding author:
William W.S. Chen

ABSTRACT

Efron and Morris (1975) gave an amusing example of batting averages of major league baseball players in the United States, to illustrate the superiority of James Stein estimators over direct estimators. In this paper, we include eight more competitors and attempts to seek a best estimator of all. These new estimators include the overall sample proportion, a synthetic estimator using the previous year batting average, two composite estimators that we mentioned, the Bayes estimators using either noninformative or informative prior distribution. We use about 370 more times at batting average was taken as the true value. Since the true values are assumed to be known, we can compute the relative overall accuracies. Four more criteria have been included to increase the selection results.

Key Words:

Basyes estimator, Composite estimator, Direct estimator, James-Stein estimator, Major league baseball players, Noninformative prior distribution, Optimal weight, Overall sample proportion, Previous year batting average, Synthetic estimator, True value.

***Citation: William W.S. Chen,* 2019.** "On Comparison Some Estimators in Small Area Study", *International Journal of Current Research*, 11, (08), 6039-6044.

INTRODUCTION

We define "small area" as a population for which inadequate or even no direct reliable information is available for the variable of interest. This could happen for a variety of reason. For example, in intercensal years, direct population counts are often not available for many small areas. In the census, population counts are frequently not accurate for certain minority groups or for illegal immigrants. Statistics on rare events obtainable from registries could provide misleading information simply due to small population sizes. Small area estimation has been of interest for a long time, especially among demographers to estimate small area population counts and other characteristics of interest in the postcensal years. Small area statistics were used as early as 11th century England and 17th century Canada. In those days, census, special surveys, or administrative records were used to obtain small area statistics. There is an increasing demand for diverse, rich and current statistics for small domains. Such statistics are needed for the planning of reform, welfare and administration in many fields and allocation of federal funds to local governments. For example, the "Improving America`s Schools Act" requires SAIPE estimates of poor school age children for counties, as well as, school districts in order to allocate more than $7 billion annually for educationally disadvantaged students. In this paper, we use an amusing example, based on Efron (1975), dated April 26, 1970 from the New York Times, of batting averages of major league baseball players to compare some estimators. Except their proposed superiority James Stein estimator over direct estimator, we also include some more competitor estimators. The new estimators include the overall sample proportion, a synthetic estimator using the previous year's batting average, two composite estimators that we mentioned, the Bayes estimators using either noninformative or informative prior distribution. The comparison based on the relative overall accuracies of the following ratio: $R=\sum(\hat{p}_i-p_i)^2/\sum(\hat{p}_{i,c}-p_i)^2$ where $\hat{p}_i$ is a direct estimator, and p_i is true value, $\hat{p}_{i,c}$ is a competitor estimator. Since the numerator is a fixed amount, larger R values mean better competitor estimators. We also include four more criterions to compare those estimators. Thus, it is more accurate to select the best estimators from different views.

Estimators: Refer to the baseball data estimate the true season batting average of each player using the following methods. Direct estimate: When the characteristic x being measured represents the presence or absence of some dichotomous attribute, the sample proportion is generally denoted by p_x and is given by $\hat{p}_x = x/n$ Where x is the number of sample elements having the dichotomous attribute.

Overall sample proportion: Gelman (1995) considered the problem of predicting the batting averages of all 18 players for the entire 1970 season, and added their career batting averages up to the 1969 season, x_1, and the number of previous times at bat for each player, x_2, We calculate the sum of the products as follow: $\sum_1^{18} x_1 x_2 = 13344.418$ and overall sample proportion

$$\frac{\sum_1^{18} x_1 x_2}{\sum x_2} = \frac{13344.418}{48327} = 0.2761 \,. \tag{2.1}$$

Synthetic Estimation: Gonzales(1973) describes synthetic estimates as follows: An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas under the assumption that the small areas have the same characteristics as the large area, we identify these estimates as synthetic estimates. In 1968, the National Institute for Health first used synthetic methods to estimate state long and short-term disabilities from the National Health interview survey data. Using synthetic estimation usually has the following advantages. It is simple and intuitive. It could apply to general sampling designs. It could borrow strength from similar events. It will provide estimates for area with no sample from the sample survey. For the current application, we use the previous year's batting average, i.e.

$$\frac{\sum_1^{18} x_1}{18} = \frac{4.564}{18} = 0.25356 \tag{2.2}$$

Two composite estimators: The reason to use composite estimator is to balance the potential bias of the synthetic estimator against the instability of the design based direct estimator.

$\hat{y}_{ic} = \phi_i \hat{y}_{i1} + (1-\phi_i)\hat{y}_{i2}$ where $\hat{y}_{i1}$: direct estimator for the ith small area $\hat{y}_{i2}$: synthetic estimator for the ith small area. ϕ_i : suitably chosen weight, $0 \le \phi_i \le 1$.

Our next objective is aim the optimal ϕ_i : subject to minimize $Mse(\hat{y}_{ic})$ with respect to ϕ_i, We assume that $Cov(\hat{y}_{i1}, \hat{y}_{i2}) \approx 0$ Consider

$$Mse(\hat{y}_{ic}) = E[\phi_i \hat{y}_{i1} + (1-\phi_i)\hat{y}_{i2} - y_i]^2 = E[\phi_i(\hat{y}_{i1} - y_i) + (1-\phi_i)(\hat{y}_{i2} - y_i)]^2$$

$$\approx \phi_i^2 Var(\hat{y}_{i1}) + (1-\phi_i)^2 MSE(\hat{y}_{i2}) = f(\phi_i)$$

$$Since \ \mathrm{E}[(\hat{y}_{i1} - y_i)(\hat{y}_{i2} - y_i)] = \mathrm{E}(\hat{y}_{i1} - y_i)[(\hat{y}_{i2} - Ey_{i2}) + \mathrm{E}(\hat{y}_{i2} - y_i)]$$

$$= \mathrm{Cov}(\hat{y}_{i1} - y_{i2}) + (E\hat{y}_{i2} - y_i)(\mathrm{E}\hat{y}_{i1} - y_i) \approx 0$$

In above derivation, we used

$$\mathrm{E}\,\hat{\mathrm{y}}_{\mathrm{i1}} \approx y_i \text{ and } \mathrm{Cov}(\hat{y}_{i1} - y_{i2}) \approx 0$$

$$f^{/}(\phi_i) = 2\phi_i Var(\hat{y}_{i1}) - 2(1-\phi_i) MSE(\hat{y}_{i2})$$

Therefore, the optimal $\hat{\phi}_i^*$ is given by

$$\phi_i^* = \frac{MSE(\hat{y}_{i2})}{MSE(\hat{y}_{i2}) + Var(\hat{y}_{i1})} = \frac{1}{1+F_i} \quad \text{where } \mathrm{F_i} = \frac{Var(\hat{y}_{i1})}{MSE((\hat{y}_{i2})}$$

The parameter $\hat{\phi}_i^*$ can be estimated by

$$\phi_i^* = \frac{MSE(\hat{y}_{i2})}{(\hat{y}_{i2}-\hat{y}_{i1})^2} = \frac{(\hat{y}_{i2}-\hat{y}_{i1})^2 - v(\hat{y}_{i1})}{(\hat{y}_{i2}-\hat{y}_{i1})^2} == 1 - \frac{v(\hat{y}_{i1})}{(\hat{y}_{i2}-\hat{y}_{i1})^2}$$

$\hat{\phi}_i^*$ is usually very unstable. Applying the above theory to our baseball data, we could obtain two more composite estimators as follow:

$\hat{p}_{i1}$: direct design based estimator, sample proportion =0.265389;

$\hat{p}_{i2}$: Overall sample proportion, synthetic estimator =0.2761;

$\hat{p}_{i3}$: A synthetic estimator using the previous hatting Average = 0.25356;

$$Var(\hat{p}_{i1}) = \frac{\hat{p}_i(1-p_i)}{n} = \frac{0.265389(1-0.265389)}{18} = 0.010831$$

$$\hat{\phi}_i^* = 1 - \frac{Var(\hat{p}_{i1})}{(\hat{p}_{i2}-\hat{p}_{i1})^2} = 1 - \frac{0.010831}{(0.2761-0.2654)^2} = -93.6021$$

D1) 1st composite estimator is

$$\begin{aligned}\hat{p}_{ic1} &= \hat{\phi}_i^* \hat{p}_{i1} + (1-\hat{\phi}_i^*)\hat{p}_{i2} \\ &= -93.6021 * 0.2654 + (1+93.6021) * 0.2761 = 1.27764 \end{aligned} \qquad (2.3)$$

(D2) 2nd composite estimator is

$$\begin{aligned}\hat{p}_{ic2} &= \hat{\phi}_i^* \hat{p}_{i3} + (1-\hat{\phi}_i^*)\hat{p}_{i1} \\ &= -76.2619 * 0.25356 + (1+76.2619) * 0.2654 = 1.1683 \end{aligned} \qquad (2.4)$$

The Bayes estimator using uniform prior, noninformative prior, on the true proportion. There has been a desire for prior distributions that can be guaranteed to play a minimal role in the posterior distribution. Such distributions are sometimes called "reference prior distribution," and the prior density is described as vague, flat, diffuse or no ninformative. The rationale for using noninformative prior distributions is often said to be to let the data speak for themselves, so that inferences are unaffected by information external to the current data. In the current case, we may assume prior density as Beta distribution.

$$f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad \alpha > 0, \quad \beta > 0.$$

and let sampling distribution, y_i / p ,have Bernoulli distribution, i=1,2,…n, then the posterior distribution of p is given by

$$\mathrm{f(p1y_i)} = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\propto [p^{\sum y_i}(1-p)^{n-\sum y_i}]p^{\alpha-1}(1-p)^{\beta-1} = p^{\sum y_i+\alpha-1}(1-p)^{n-\sum y_i+\beta-1}$$

A beta $(\alpha + \sum y_i,\ \beta{+}n - \sum y_i)$. The posterior mean is given by

$E(p1\sum y_i = n\hat{p}_i) = \dfrac{\alpha + n\hat{p}_i}{\alpha+\beta+n}$. If we consider a special case of beta Distribution, i.e. prior is uniform distribution

(noninformative prior) it means $(\alpha = 1\ ,\ \beta{=}1)$ and rewrite the posterior mean in the composite estimator form as follows:

$$E(p1\sum y_i = np_i) = \frac{\alpha + np_i}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right) + \frac{n}{\alpha+\beta+n} p_i$$

Suppose we choose $\hat{\phi} = \frac{\alpha+\beta}{\alpha+\beta+n} = \frac{2}{2+n}$, $1-\hat{\phi} = \frac{n}{\alpha+\beta+n} = \frac{n}{2+n}$, then $E(p1\sum y_i = np_i) =$

$$\hat{\phi}\left(\frac{\alpha}{\alpha+\beta}\right) + (1-\phi)p_i = 0.0425x0.5 + 0.9574 * p_i \tag{2.5}$$

Where p_i is direct estimator.

For some other values of beta (α,β) distributed and say,

$\alpha = \beta = \frac{\sqrt{n}}{2}$ and let $\hat{\phi} = \frac{n}{\alpha+\beta+n}$, $1\text{-}\hat{\phi} = \frac{\alpha+\beta}{\alpha+\beta+n}$, then again this conditional mean can be expressed as a composite estimator of (μ, p_i) $E(p1y=\sum y_i) = (1-\phi)\mu + \hat{\phi}p_i$ where $\mu = \frac{\alpha}{\alpha+\beta}$ with the given

$\alpha = \beta = \frac{\sqrt{n}}{2}$, we can compute $\hat{\phi} = \frac{n}{n+\sqrt{n}} = \frac{1}{1+\frac{1}{\sqrt{n}}}$ so, with

$n = 45$, $\hat{\phi} = 0.87$ and $1\text{-}\hat{\phi} = 0.13$, μ=0.5, we finally get conditional expectation

$$E(p1np_i) = 0.13 * 0.5 + 0.87 * p_i, \tag{2.6}$$

For different direct estimator p_i we can easily compute the posterior mean.

G. Suppose that we choose $\alpha = \beta = n$, Again choose

$\phi = \frac{n}{\alpha+\beta+n} = \frac{1}{3}$; $1\text{-}\hat{\phi} = \frac{\alpha+\beta}{\alpha+\beta+n} = \frac{2}{3}$. and the posterior mean can be expressed as

$$E(p1np_i) == (1-\hat{\phi}\}\mu + \hat{\phi}p_i = 0.66 * 0.5 + 0.33 * p_i \tag{2.7}$$

H. If we repeat the previous process and let $\alpha = \beta = \frac{n}{2}$.

Define $\phi = \frac{n}{\alpha+\beta+n} = \frac{1}{2}$; $1\text{-}\hat{\phi} = \frac{\alpha+\beta}{\alpha+\beta+n} = \frac{1}{2}$. and the posterior mean can be expressed as

$$E(p1np_i) == (1-\hat{\phi}\}\mu + \hat{\phi}p_i = 0.5 * 0.5 + 0.5 * p_i \tag{2.8}$$

SELECTION CRITERION: In this section we list five criterions that we will be used for the comparison of all estimators.

$$R = \frac{\sum_i (\hat{P}_i - P_i)^2}{\sum_i (\hat{P}_{i,sel} - P_i)^2} \tag{3.1}$$

where $\hat{P}_i$ is direct estimator, P_i *is* true value, $\hat{P}_{i,sel}$ is selected estimator

Average Square deviation (ASD)

$$ASD = \frac{1}{18}\sum_{1}^{18} (\hat{P}_{i,sel} - P_i)^2 \tag{3.2}$$

Average ratio square deviation (ARSD)

$$ARSD = \frac{1}{18}\sum_{1}^{18} \frac{(\hat{P}_{i,sel} - P_i)^2}{p_i} \tag{3.3}$$

Absolute values average deviation (AAD)

$$AAD = \frac{1}{18}\sum_{1}^{18} \left|\hat{P}_{i,sel} - P_i\right| \tag{3.4}$$

Absolute values ratio average deviation (ARAD)

$$ARAD = \frac{1}{18}\sum_{1}^{18} \frac{\left|\hat{P}_{i,sel} - P_i\right|}{P_i} \tag{3.5}$$

Criterion (3.1) is useful to us. If we assume the data came from the normal distribution then the numerator and denominator sum square of deviation from the mean has both $\chi^2_{(n-1)}$ distribution. This leads to the ratio has F distribution. We have the table values for this probability distribution. We prefer criterion (3.1) than others. However, the other four criterions still useful for a good reference for comparison purposes.

Summary

Table 1 Comparison of estimators VS five criterionl

	Direct	J-S Estimator	Compromise JS Estimator	model [2.1]	model [2.2]
$\Sigma(\hat{P}_{i,sel}-P)$	0.003	-0.002	0.043	0.1958	-0.2092
$\Sigma(\hat{P}_{i,sel}-P)^2$	0.07660	0.02188	0.01873	0.02686	0.02716
R	1.00	3.50091	4.08970	2.85182	2.82032
C3.2	0.00426	0.00122	0.00104	0.00149	0.00151
C3.3	0.01556	0.00465	0.00405	0.00622	0.00538
C3.4	0.0556	0.02656	0.02394	0.02993	0.03173
C3.5	0.20750	0.10333	0.09375	0.12113	0.11883

Table 1(Continue)Comparison of estimators VS five criterion

	Model [2.3]	model [2.4]	model [2.5]	model [2.6]	model [2.7]	model [2.8]
$\Sigma(\hat{P}_{i,sel}-P)$	18.222	16.2554	0.18289	0.5520	2.7424	2.1145
$\Sigma(\hat{P}_{i,sel}-P)^2$	18.47	14.7046	0.07288	0.0774	0.44142	0.2784
R	0.0041	0.0052	1.0510	0.9897	0.1735	0.2751
C3.2	1.0263	0.8169	0.0040	0.0043	0.0245	0.0155
C3.3	3.9879	3.1783	0.0150	0.0165	0.0992	0.0626
C3.4	1.0124	0.9031	0.0541	0.0532	0.1524	0.1174
C3.5	3.9138	3.4934	0.2027	0.2015	0.6020	0.4657

Concluding Remarks: Using the R-Criterion, we can compare the James-Stein estimator, 3.50091, with our best competitor, overall sample proportion estimator, 2.85182, and next one, previous year batting average estimator, 2.82032. Their deviations are mild. If we apply(3.4), absolute value average deviation, we found the difference between these estimators are 0.00337 and 0.00517. In percent, it has only 0.337% and 0.517%. We can conclude that the James-Stein estimator is even better than summarize overall past experiences together to get the best estimator.

Or superior to use previous year experience. If we use the same criterions and apply to model (2.3) and (2.4), the results are poor due to the R-values are small and the deviations are large. This result is not surprising as we already pointed out that the unstable weight value of ϕ would cause this. Compare this with the Bayes estimator, model [2.5] and [2.6], the R-value for this model is close to 1. We can conclude that these estimators have similar efficiency as direct estimators but with smaller absolute average deviation. Based on these evaluations we recommend use these two models, models [2.5] and [2.6], as our selected models. While the model [2.7] or [2.8] will be less interest if the same criterions used. This causes by composite estimators have heavier weight on direct estimator side. For other criterions, average square deviation [3.2], and average ratio square deviation [3.3], are consistent with the other criterion. The difference between James-Stein estimator and model [2.1], [2.2] are not significant, while with model [2.5] and [2.6] are larger. The "James-Stein estimator" is much superior to other estimators. From table 1, we can clearly see that compromised
J-S estimator is the best.

REFERENCES

Efron, B 1975. Biased Versus Unbiased Estimation, Advances in Mathematics, 16, 259-277.

Efron, B., and Morris, C.E. 1972a. Limiting the Risk of Bayes and Empirical Bayes Estimators, Part II: The Empirical Bayes Case, Journal of the American Statistical Association, 67, 130-139.

Efron, B., and Morris, C.E. 1972b. Empirical Bayes on Vector Observations: An Extension of Stein`s Method, Biometrika, 59, 335-347.

Efron, B., and Morris, C.E. 1973. Stein`s Estimation Rule and its Competitors- An Empirical Bayes Approach, Journal of the American Statistical Association, 68, 117-130.

Efron, B., and Morris, C.E. 1975. Data Analysis Using Stein`s Estimate and its Generalizations, Journal of the American Statistical Association, 70, 311-319.

Gelman, A. Carlin, J.B. Stern H.S. and Rubin, D.B. (2004), Bayesian Data Analysis, second edition, Chapman & Hall CRC Press company.

Gonzalez, M.E. 1973. Use and evaluation of synthetic estimators. In the proceedings of the Social Statistics Section 33-36. American statistical association, Washington D.C.

Rao, J.N.K. 2003. Small Area Estimation. Wiley Series in Survey Methodology. John Wiley & Sons, Inc.

APPENDIX

player	pi_hat	true p	J-S est	c J-S est	X1	X2	x1*x2
Clemente	0.4	0.346	0.293	0.334	0.314	8142	2556.588
Robinson	0.378	0.298	0.289	0.312	0.303	7542	2285.226
Howard	0.356	0.276	0.284	0.29	0.256	86	22.016
Johnston	0.333	0.221	0.279	0.279	0.25	2065	516.25
Berry	0.311	0.273	0.275	0.275	0.275	4826	1327.15
Spencer	0.311	0.27	0.275	0.275	0.264	3210	847.44
Kessinge	0.289	0.263	0.27	0.27	0.246	2244	552.024
Alvarado	0.267	0.21	0.265	0.265	0.244	454	110.776
Santo	0.244	0.269	0.261	0.261	0.281	5658	1589.898
Swoboda	0.244	0.23	0.261	0.261	0.248	2753	682.744
Unser	0.222	0.264	0.256	0.256	0.255	2281	581.655
Williams	0.222	0.256	0.256	0.256	0.257	1216	312.512
Scott	0.222	0.304	0.256	0.256	0.271	888	240.648
Petrocel	0.222	0.264	0.256	0.256	0.255	1139	290.445
Rodriguez	0.222	0.226	0.256	0.256	0.244	1967	479.948
Campaneris	0.2	0.285	0.251	0.251	0.234	291	68.094
Munson	0.178	0.319	0.247	0.243	0.118	51	6.018
Alvis	0.156	0.2	0.242	0.221	0.249	3514	874.986

Overall Model [2.1]	Previous Model [2.2]	Model [2.3]	Model [2.4]	Model [2.5]	Model [2.6]	Model [2.7]	Model [2.8]
0.2761	0.2536	1.2776	1.1683	0.40426	0.413	0.462	0.45
0.2761	0.2536	1.2776	1.1683	0.38319	0.39386	0.45474	0.439
0.2761	0.2536	1.2776	1.1683	0.36213	0.37472	0.44748	0.428
0.2761	0.2536	1.2776	1.1683	0.34011	0.35471	0.43989	0.4165
0.2761	0.2536	1.2776	1.1683	0.31905	0.33557	0.43263	0.4055
0.2761	0.2536	1.2776	1.1683	0.31905	0.33557	0.43263	0.4055
0.2761	0.2536	1.2776	1.1683	0.29798	0.31643	0.42537	0.3945
0.2761	0.2536	1.2776	1.1683	0.27692	0.29729	0.41811	0.3835
0.2761	0.2536	1.2776	1.1683	0.25490	0.27728	0.41052	0.372
0.2761	0.2536	1.2776	1.1683	0.25490	0.27728	0.41052	0.372
0.2761	0.2536	1.2776	1.1683	0.23384	0.25814	0.40326	0.361
0.2761	0.2536	1.2776	1.1683	0.23384	0.25814	0.40326	0.361
0.2761	0.2536	1.2776	1.1683	0.23384	0.25814	0.40326	0.361
0.2761	0.2536	1.2776	1.1683	0.23384	0.25814	0.40326	0.361
0.2761	0.2536	1.2776	1.1683	0.23384	0.25814	0.40326	0.361
0.2761	0.2536	1.2776	1.1683	0.21278	0.239	0.396	0.35
0.2761	0.2536	1.2776	1.1683	0.19171	0.21986	0.38874	0.339
0.2761	0.2536	1.2776	1.1683	0.17065	0.20072	0.38148	0.328

****