Statistics and Applications {ISSN 2452-7395 (online)}
Volume 15 Nos. 1&2, 2017 (New Series), pp 1-6

Empirical Bayes Estimation Using Quantitative Randomized Response Data

Arijit Chaudhuri and Purnima Shaw
Indian Statistical Institute, Kolkata, India

Received: August 24, 2016; Revised: February 11, 2017; Accepted: February 17, 2017

Abstract

The finite population total or mean of a sensitive characteristic is estimated by using the Randomized Response Devices I and II vide Chaudhuri and Christofides (2013). An alternative method of estimation has been provided. Empirical Bayes estimation of the population mean of the unknown prior probabilities assigned to the individuals of the finite population is presented.

Key words: Empirical Bayes estimation, General sampling scheme, Quantitative Randomized Response

---

1. Introduction

Consider a finite population denoted by $U = (1,2,\ldots,i,\ldots,N)$ of N units. A sample s is chosen from U with a pre-assigned probability $p(s)$ with first order inclusion probability $\pi_i = \sum_{s\ni i} p(s)>0 \ \forall \ i \in U$ and second order inclusion probability $\pi_{ij}=\sum_{s\ni i,j} p(s)>0 \ \ i \neq j \in U$. Let y_i be the value of the quantitative sensitive variable for the i^{th} individual in U. We wish to estimate $\bar{Y}= \frac{1}{N}\sum_{i=1}^{N} y_i$. Let E_P and E_R denote the design and RR based Expectation operators. V_P and V_R denote the design and RR based Variance operators.

Also, $E = E_R E_P = E_P E_R$ and

$$V = V_R E_P + E_R V_P = V_P E_R + E_P V_R$$

are overall Expectation and Variance operators respectively.

In Section 2 an alternative method of estimation for Device I is described. The same exercise has been performed for Device II in Section 3. Section 4 presents some numerical calculations based on simulated data.

2. Device I

Let us describe the Device I for the estimation of the finite population total or mean of a sensitive variable. A person labeled i is offered two boxes marked A and B, respectively. Identical cards bearing numbers $a_1, a_2, a_3,\ldots, a_T$ in sufficient numbers are placed in the first box and cards likewise but numbered $b_1, b_2, \ldots, b_M$ are put in the second box. The sampled

Corresponding Author: Arijit Chaudhuri
E-mail: arijitchaudhuri1@rediffmail.com

person i is requested to draw one card independently from each of the boxes, say, bearing a_j and b_k, and report

$$z_i = a_j y_i + b_k$$

to the investigator without disclosing any of the numbers of the right hand side. Then from Chaudhuri and Christofides (2013), we have,

$$r_i = \frac{z_i - \gamma}{\mu} \tag{1}$$

where $\mu = \frac{1}{T}\sum_{j=1}^{M} a_j$ and $\gamma = \frac{1}{M}\sum_{k=1}^{M} b_k$, as an unbiased estimator of y_i ·∋· $\mathrm{E_R}(r_i) = y_i$,

$$V_R(r_i) = \alpha y_i^{\,2} + \beta \text{ where } \alpha = \left(\frac{\sigma^2}{\mu^2}\right) \text{ and } \beta = \left(\frac{\psi^2}{\mu^2}\right) \text{ and} \tag{2}$$

$$\nu_{\mathrm{R}}(\mathrm{r_i}) = \frac{\alpha r_i^{\,2} + \beta}{1+\alpha} \text{ ·∋· } E_R(\nu_R(r_i)) = V_R(r_i),\ i \in U. \tag{3}$$

Now, an alternative method of estimation using Empirical Bayes approach is as follows:

Let $L_i = L(y_i)$ = Prior Probability(y-value of i is y_i). Then, using Chaudhuri and Christofides(2013), we have,

$$\begin{aligned} L(y_i|z_i) &= \text{Posterior Probability}(y - \text{value of } i \text{ is } y_i | \text{response of } i \text{ is } z_i) \\ &= \frac{L_i P(z_i|y_i)}{P(z_i)} \\ &= \frac{L_i\left(1/(TM)\right)}{\left(1/(TM)\right)} \\ &= L_i. \end{aligned}$$

Now, $E_{L_i^*}(y_i|z_i) = \frac{\sum_{j=1}^{T}\sum_{k=1}^{M} L_i y_i}{\sum_{j=1}^{T}\sum_{k=1}^{M} L_i}$, where $E_{L_i^*}(y_i|z_i)$ is the Empirical Bayes estimate for y_i.

$$\begin{aligned} &= \frac{\sum_{j=1}^{T}\sum_{k=1}^{M}\left(\frac{z_i - b_k}{a_j}\right)}{TM} \text{ [since } z_i = a_j y_i + b_k] \\ &= \frac{z_i}{T}\sum_{j=1}^{T}\frac{1}{a_j} - \frac{\sum_{j=1}^{T}\frac{1}{a_j}\sum_{k=1}^{M} b_k}{TM} \\ &= \frac{z_i - \gamma}{HM_a} = h_i \text{ (say), where } HM_a = \frac{1}{T}\sum_{j=1}^{T}\frac{1}{a_j} \end{aligned} \tag{4}$$

h_i is the Empirical Bayes estimate for y_i.

$$\begin{aligned} E_R(h_i) &= E_R\left(\frac{a_j y_i + b_k - \gamma}{HM_a}\right) \\ &= \frac{\mu y_i}{HM_a} \end{aligned}$$

$$\begin{aligned} V_R(h_i) &= \frac{V_R(z_i)}{HM_a^{\,2}} \\ &= \frac{\sigma^2 y_i^{\,2} + \psi^2}{HM_a^{\,2}} \\ &= \alpha' y_i^{\,2} + \beta', \text{ where } \alpha' = \frac{\sigma^2}{HM_a^{\,2}} \text{ and } \beta' = \frac{\psi^2}{HM_a^{\,2}} \end{aligned} \tag{5}$$

$$E_R\left[\frac{\alpha' {h_i}^2+\beta'}{1+\alpha'}\right]=\frac{\frac{\sigma^2}{HM_a{}^2}\left(\frac{\sigma^2 {y_i}^2+\psi^2+\mu^2 {y_i}^2}{HM_a{}^2}\right)+\frac{\psi^2}{HM_a{}^2}}{1+\frac{\sigma^2}{HM_a{}^2}} \text{ (since } E_R({h_i}^2)=\frac{\sigma^2 {y_i}^2+\psi^2+\mu^2 {y_i}^2}{HM_a{}^2}\text{)}$$

$$= \alpha'\left(\frac{\sigma^2+\mu^2}{\sigma^2+HM_a{}^2}\right){y_i}^2+\beta'$$

So, $\frac{\alpha' {h_i}^2+\beta'}{1+\alpha'}$ can be considered as an estimator of $V_R(h_i)$. (6)

3. Device II

The original method of estimation using Randomized Response Device II by Chaudhuri and Christofides (2013) is as follows: A sampled person labelled *i* is offered a box with a large number of identical cards such that a proportion *C* (0<*C*<1) of them is marked "True" and the remaining of them bearing real values $x_1, x_2,\ldots, x_j,\ldots, x_M$ in proportions $q_1, q_2,\ldots, q_j,\ldots, q_M$ such that $\sum_{j=1}^{M} q_j = 1 - C$. The sampled person *i* is requested to draw one of the cards and if the card is marked with "True", then he/she is to report his/her true value of *y i.e.* y_i. If instead one of the cards marked x_j is drawn, he/she is to report the value x_j and return the card back to the box. Let z_i be the value reported by *i* to the investigator. Then from Chaudhuri and Christofides (2013), we have

$$r_i = \frac{1}{C}\left(z_i - \sum_{j=1}^{M} q_j x_j\right) \quad (7)$$

as an unbiased estimator of y_i ·∋· $E_R(r_i) = y_i$,

$V_R(\mathrm{r_i}) = \alpha {y_i}^2 + \beta y_i + \psi$, where $\alpha = \frac{1}{C} - 1$, $\beta = -\frac{2}{C}\sum_{j=1}^{M} q_j x_j$ and

$$\psi = \frac{1}{C^2}\left(\sum_{j=1}^{M} q_j {x_j}^2 - \left(\sum_{j=1}^{M} q_j x_j\right)^2\right) \quad (8)$$

$$\nu_R(r_i) = \frac{\alpha {r_i}^2+\beta r_i+\psi}{1+\alpha} \text{ ·∋· } E_R(\nu_R(r_i)) = V_R(r_i),\ i \in U. \quad (9)$$

An alternative method of estimation using Empirical Bayes approach is as follows: The expression for z_i can be written as:

$$z_i = I_j y_i + (1 - I_j)x_j$$

where $I_j = 1$ if the card drawn by *i* is marked "True"
$= 0$ if the card drawn by *i* is not marked "True"

Let $L_i = L(y_i)$ = Prior Probability(*y* – value of *i* is y_i). Then, using Chaudhuri and Christofides (2013), we have

$$L(y_i|z_i) = \text{Posterior Probability}(y - \text{value of } i \text{ is } y_i | \text{response of } i \text{ is } z_i)$$
$$= \frac{CL_i}{CL_i+(1-C)(1-L_i)}$$
$$= \frac{CL_i}{(1-C)+(2C-1)L_i}$$

Now, $E_{L_i{}^*}(y_i|z_i) = \frac{\sum_{j=1}^{M} q_j \frac{CL_i}{(1-C)+(2C-1)L_i} y_i}{\sum_{j=1}^{M} q_j \frac{CL_i}{(1-C)+(2C-1)L_i}}$, where $E_{L_i{}^*}(y_i|z_i)$ is the Empirical Bayes estimate

for y_i

$$= \frac{\sum_{j=1}^{M} q_j \left[\frac{z_i - (1-I_j)x_j}{I_j}\right]}{\sum_{j=1}^{M} q_j} \text{ [since } z_i = I_j y_i + (1 - I_j)x_j]$$

$$= \sum_{j=1}^{M} q_j \left[\frac{z_i - (1-I_j)x_j}{I_j}\right],$$

which does not exist if $I_j = 0$. Hence, Empirical Bayes estimation procedure does not work out with Device II.

4. Theoretical Comparison of Efficiency of Estimates

Theoretical comparison of efficiency of estimates obtained from original method of estimation using Randomized Response Device II and estimates obtained from alternative method of estimation using Empirical Bayes approach is as follows:

$$v_R(h_i) - v_R(r_i) = \frac{\alpha' {h_i}^2 + \beta'}{1 + \alpha'} - \frac{\alpha {r_i}^2 + \beta}{1 + \alpha}$$

$$= \frac{\frac{\sigma^2}{{HM_a}^2}\left(\frac{z_i - \gamma}{HM_a}\right)^2 + \frac{\psi^2}{{HM_a}^2}}{1 + \frac{\sigma^2}{{HM_a}^2}} - \frac{\frac{\sigma^2}{\mu^2}\left(\frac{z_i - \gamma}{\mu}\right)^2 + \frac{\psi^2}{\mu^2}}{1 + \frac{\sigma^2}{\mu^2}}$$

$$= \frac{\sigma^2 (z_i - \gamma)^2 + {HM_a}^2 \psi^2}{{HM_a}^2 \left({HM_a}^2 + \sigma^2\right)} - \frac{\sigma^2 (z_i - \gamma)^2 + \mu^2 \psi^2}{\mu^2 (\mu^2 + \sigma^2)}$$

$$> 0 \text{ since Arithmetic Mean } (AM) > \text{Harmonic Mean } (HM)$$

Hence, the estimates obtained from original method of estimation using Randomized Response Device II out performs as compared to the estimates obtained from alternative method of estimation using Empirical Bayes approach in terms of efficiency.

5. Estimation

So, an Empirical Bayes approach is used to estimate $\bar{L} = \frac{1}{N}\sum_{i=1}^{N} L_i$ instead of $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ by taking h_i as an initial estimator for $L_i(y_i)$,

$\hat{\bar{L}} = \frac{1}{N}\sum_{i \in s} \frac{h_i}{\pi_i}$ $\cdot\ni\cdot$ $E(\hat{\bar{L}}) = E_P E_R(\hat{\bar{L}}) \cong \bar{L}$, vide Chaudhuri (2011). (10)

$$V(\hat{\bar{L}}) = V_P E_R(\hat{\bar{L}}) + E_P V_R(\hat{\bar{L}}) = \frac{1}{N^2}\left[\sum_{i<}^{N}\sum_{j}^{N}\left(\pi_i \pi_j - \pi_{ij}\right)\left(\frac{L_i}{\pi_i} - \frac{L_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{{L_i}^2}{\pi_i}\beta_i + \sum_{i=1}^{N}\frac{V_R(h_i)}{\pi_i}\right] \quad (11)$$

(where $\beta_i = 1 + \frac{1}{\pi_I}\sum_{j \neq i}^{N}\pi_{ij} - \sum_{i=1}^{N}\pi_i$)

If every sample s contains a common number of distinct units in it, then,

$$V(\hat{\bar{L}}) = \frac{1}{N^2}\left[\sum_{i<}^{N}\sum_{j}^{N}\left(\pi_i \pi_j - \pi_{ij}\right)\left(\frac{L_i}{\pi_i} - \frac{L_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{V_R(h_i)}{\pi_i}\right] \quad (12)$$

$$\upsilon\left(\hat{\bar{L}}\right) = \frac{1}{N^2}\left[\sum_{i<}\sum_{j\in s}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{h_i}{\pi_i}-\frac{h_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{{h_i}^2-\upsilon_R(h_i)}{\pi_i}\beta_i + \sum_{i\in s}\frac{\upsilon_R(h_i)}{\pi_i}\right]. \quad (13)$$

If every sample *s* contains a common number of distinct units in it, then,

$$\upsilon(\hat{\bar{L}}) = \frac{1}{N^2}\left[\sum_{i<}\sum_{j\in s}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{h_i}{\pi_i}-\frac{h_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{\upsilon_R(h_i)}{\pi_i}\right] \quad (14)$$

$E(\upsilon(\hat{\bar{L}})) = E_P E_R(\upsilon(\hat{\bar{L}})) \cong V(\hat{\bar{L}})$.

6. Numerical Presentation

Data for 117 households was simulated for the variable: y_i = amount of money spent in gambling (Rupees), and x_i = number of members in a household to which *i* belongs

Unit	*y*	*x*	Unit	*y*	*x*	Unit	*y*	*x*	Unit	*y*	*x*
1	2891.31	9	31	2495.64	7	61	2636.53	3	91	1940.05	4
2	4261.13	7	32	4400.64	8	62	1344.76	3	92	2724.16	8
3	2262.45	3	33	3284.96	11	63	1544.81	3	93	3199.71	3
4	2530.2	5	34	1334.98	6	64	1255.77	7	94	1241.56	10
5	2430.49	3	35	1408.34	8	65	1328.88	2	95	1173.01	3
6	4226.83	5	36	1241.83	3	66	3258.28	10	96	1435.06	10
7	3270.41	1	37	4649.75	5	67	2740.52	4	97	251.42	8
8	1179.95	1	38	2243.53	7	68	4298.5	8	98	3236.45	3
9	1902.73	9	39	1120.97	2	69	2185.7	2	99	1309.49	11
10	1482.09	8	40	1296.67	9	70	251.27	11	100	3247.36	8
11	1480.36	8	41	2878	6	71	3065.67	8	101	1271.32	10
12	250.9	1	42	1268.51	11	72	1194.98	10	102	208.24	5
13	2255.33	7	43	1258.95	1	73	179.98	8	103	246.96	8
14	2525.85	9	44	2990.47	8	74	3845.06	9	104	1474.4	2
15	1241.19	3	45	1299.93	9	75	1188.66	2	105	2430.23	5
16	1256.66	6	46	205.55	11	76	189.36	2	106	1148.49	1
17	2194.89	5	47	1245.97	3	77	1247.3	1	107	640.08	10
18	3187.48	5	48	1241.24	4	78	5004.93	6	108	3942.96	1
19	193.65	5	49	195.59	5	79	1505.03	5	109	2202.25	3
20	1669.54	5	50	2260.59	10	80	3240.26	10	110	241.63	5
21	3074.11	3	51	242.99	4	81	3254.33	5	111	4191.92	3
22	4187.81	1	52	195.08	2	82	334.97	6	112	4269.03	9
23	1264.92	6	53	3194.31	11	83	1242.27	5	113	2742.73	5
24	3196.59	7	54	2307.38	11	84	4181.9	4	114	542.3	2
25	3354.57	2	55	4842.01	3	85	187.78	7	115	1546.3	1
26	2717.12	10	56	2904.35	6	86	3242.91	8	116	1478	2
27	2927.63	1	57	3154.77	8	87	4334.62	10	117	789	6
28	4147.14	6	58	2191.78	8	88	2575.97	7			
29	3385.06	5	59	2241.53	5	89	2608.09	3			
30	2644.63	10	60	1241.82	11	90	4703.93	4			

A single sample of size 37 was drawn using the Hartley and Rao (HR) (1962) sampling scheme in which a systematic sample by Probability Proportional to Size (PPS) method is used after the random arrangement of the population units. The number of members in a household to which *i* belongs (denoted by *x*) were utilized as size measures. It is unnecessary to check if the size-measure variable is well-correlated with the *y*-variable or not as Chaudhuri (2011) has discussed. He argues that a large-scale survey is implemented on taking a single sample, which is used to estimate several parameters of which a few may relate to sensitive features. As mentioned by Chaudhuri (2011), an RR procedure is not "a sampling scheme-specific"; so an estimation method may be developed based on a general

sampling scheme and using the RR's realized on hand. So, we illustrate employing the Hartley-Rao scheme as it is so well-known and handy with several properties as are classically known.

We collect the responses taking several combinations of Device I parameters namely a_1, a_2, $a_3,\ldots, a_T$ and b_1, b_2, $b_3,\ldots, b_M$.

We calculate:

(i) the coefficient of variation,

$$CV_{Bayes} = \frac{\sqrt{v(\hat{\bar{L}})}}{\hat{\bar{L}}} 100$$

and compare this with

$$CV_{original} = \frac{\sqrt{v(e)}}{e} 100$$

where $e = \frac{1}{N}\sum_{i\in s}\frac{r_i}{\pi_i}$, $\cdot\ni\cdot\ E(e) = E_P E_R(e) = \bar{Y}$ and

$$v(e) = \frac{1}{N^2}\left[\sum_{i<}\sum_{j\in s}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{r_i}{\pi_i}-\frac{r_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{r_i{}^2-v_R(r_i)}{\pi_i}\beta_i + \sum_{i\in s}\frac{v_R(r_i)}{\pi_i}\right].$$

If every sample s contains a common number of distinct units in it, then, for the original estimator e of $\bar{Y}$, an unbiased variance estimator is

$$v(e) = \frac{1}{N^2}\left[\sum_{i<}\sum_{j\in s}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{r_i}{\pi_i}-\frac{r_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_R(r_i)}{\pi_i}\right]$$

$\cdot\ni\cdot\ E(v(e)) = E_P E_R(v(e)) = V(e)$

Bayesian approach of estimation is better than the original method *if* $CV_{Bayes} < CV_{Original}$.

(ii) The estimated efficiency ($\hat{E}$) of the Bayesian approach of estimation with respect to the original method is defined as

$$\hat{E}_{BO} = \frac{v(e)}{v(\hat{\bar{L}})} 100$$

Larger the $\hat{E}_{BO}$, the better it is to use the Bayesian approach of estimation. Performances of the procedures of 'original' and 'empirical Bayes' methods of estimation are illustrated below for few combinations of $\mu, \gamma, \sigma^2, \psi^2$ and HM_a for Device I. Other cases are not shown here.

Findings: Device I

Table 1.1: Comparison of Coefficient of Variation

μ	γ	σ^2	ψ^2	HM_a	$CV_{Original}$	CV_{Bayes}
3948.90	627.99	3085613.55	50846.93	3003.90	14.39	14.78
4211.92	569.64	3012523.03	49809.69	3187.04	12.60	13.00
4175.72	609.85	3429579.54	54945.67	2999.52	16.07	16.53
3867.73	611.71	2965948.66	51411.31	2817.27	14.00	14.46
4057.74	577.51	3159977.59	52745.65	2809.38	16.44	16.94

Table 1.2: Efficiency of estimates

μ	γ	σ^2	ψ^2	HM_a	$v(e)$	$v(\hat{\bar{L}})$	$\hat{E}_{BO}$
4211.92	569.64	3012523.03	49809.69	3187.04	87647.53	162989.69	82.35
3948.90	627.99	3085613.55	50846.93	3003.90	114983.05	209543.66	81.13
4153.47	587.89	2833768.61	50560.10	2936.82	71282.30	157226.21	78.10
3867.73	611.71	2965948.66	51411.30	2817.27	102789.05	206660.47	77.95
3959.46	653.71	2838449.49	53168.78	2809.82	64477.02	141211.84	77.16

Conclusions

Although the Empirical Bayes estimation procedure is not better than the original estimation procedure, but still the former is quite competitive.

References

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys.* Boca Raton: Chapman and Hall, CRC Press, Taylor and Francis Group.

Chaudhuri, A. and Christofides, T.C. (2013). *Indirect Questioning in Sample Surveys.* Springer-Verlag, Berlin, Heidelberg.

Warner, S.L. (1965). Randomised response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.