Thailand Statistician
January 2013; 11(1): 17-29
http://statassoc.or.th
Contributed paper

Bayesian Estimation of Rare Sensitive Attribute

Abdul Wakeel* [a, b] and Muhammad Aslam [b]

[a] Department of Mathematics, COMSATS Institute of Information Technology, Park Road Chak Shahzad, Islamabad, Pakistan.

[b] Department of Statistics, Faculty of Natural Sciences, Quaid-i-Azam University, Islamabad, Pakistan.

* Author for correspondence; e-mail: chabdulwakeel@gmail.com

Received: 9 October 2012
Accepted: 28 January 2013

Abstract

In this study, a Bayesian estimation of the population mean of a rare sensitive attribute has been considered and a Bayes estimator is proposed when the information from the respondent is collected through the randomized response technique (RRT) using Greenberg et al. (1969) and Land et al. (2012) models. The Gamma distribution has been used as prior information to check the behaviour of the Bayes estimator for the different values of population mean of rare sensitive and rare unrelated attribute. It is noted that Bayes estimator is efficient as compared to the Maximum Likelihood Estimator (MLE).

Keywords: Bayesian estimation, maximum likelihood estimator, mean squares error, rare sensitive attribute, simple random sampling.

1. Introduction

Collecting the information about the sensitive issue has always been a serious problem during the survey of human population. When we study the sensitive attribute and ask directly from the respondent about sensitive issue, it creates the ambiguity and evasiveness. Warner [1] was the first researcher who proposed a method of collecting the information about sensitive issue. Later on, many improvements have been made by several researchers. Some of them are Greenberg et al. [2], Chaudhuri and Mukerjee [3], Mangat and Singh [4], Mahmood et al. [5], Bhargava and Singh [6], Christofides [7], Kim and Warde [8], Diana and Perri [9], Land et al. [10].

Many researchers gave the idea of using the Bayesian estimation procedure for randomized response models. According to them, in case when the prior information is given, Bayesian estimation procedure can be used to estimate the unknown population parameter. Winkler and Franklin [11], Pitz [12], O'Hagan [13], Migon and Tachibana [14], Unnikrishnan and Kunte [15], Bar-Lev et al. [16], Barabesi and Marcheselli [17] and Kim and Tebbs [18] were some of the researchers who used the Bayesian technique for the analysis of randomized response data.

In this paper, we intend to propose an improved and efficient method for the estimation of rare sensitive attribute. We propose Bayes estimator for the population mean of rare sensitive attribute using the Greenberg et al. [2] and Land et al.'s [10] model. In Section 2, Greenberg et al. [2] and Land et al.'s [10] models are discussed. Proposed Bayesian estimation procedure is presented in Section 3. The comparative study is done in Section 4.

2. Review of Land et al. and Greenberg et al. models

Land et al. [10] proposed a method to estimate the mean number of persons possessing the rare sensitive attribute. For this purpose they used the Greenberg et al. [2] model to collect the information from the respondent. Greenberg et al.'s [2] gave an unrelated question technique, in which each individual selected in the samples are asked to reply "*yes*" or "*no*" to one of the following two statements:

(i) Do you belong to Group *A*?

(ii) Do you belong to Group *Y*?

with respective probabilities P and $1-P$.

Second question asked in the sampling does not have any effect on the first question. Greenberg et al. [2] considered "π_A" and "π_Y" the proportion of persons possessing sensitive and unrelated characteristic and discussed both the cases when

"π_Y" is known and unknown. Greenberg et al. [2] defined the probability of yes responses as:

$$Prob(yes) = \theta_0 = P\pi_A + (1-P)\pi_Y \tag{2.1}$$

Land et al. [10] gave a solution of unique problem where the estimation is done for mean number of persons having the rare sensitive characteristic. Greenberg's [2] unrelated model is used for the estimation procedure. Here the huge sample sizes are needed for the estimation procedure. They considered the rare sensitive case as the proportion of AIDS patients who continue having affairs with strangers and rare unrelated as the number of persons who have witnessed a murder. As the large sample sizes are required $n \to \infty$ and $(\pi_A, \pi_Y) \to 0$ then $n\pi_A = \lambda_A$ and $n\pi_Y = \lambda_Y$, showing the number of persons possessing rare sensitive and rare unrelated attribute. In this method each respondent selected in the sample is directed to rotate a spinner consisting of two statements:

(i) Do you belong to rare sensitive attribute A?

(ii) Do you belong to rare unrelated attribute Y?

with respective probabilities P_1 and $1-P_1$.

They have defined an unbiased estimator of λ_A, when the rare unrelated attribute is known, as:

$$\lambda_A = \frac{1}{P_1}\left[\frac{1}{n}\sum_{i=1}^{n} y_i - (1-P_1)\lambda_y\right] \tag{2.2}$$

where y_i is the response from the i^{th} respondent. The variance of the estimator $\hat{\lambda}_A$ is:

$$Var(\lambda_A) = \frac{\lambda_A}{nP_1} + \frac{(1-P_1)\lambda_y}{nP_1^2}. \tag{2.3}$$

When the rare unrelated attribute is unknown, each respondent selected in the sample is directed to rotate two spinners one after the other. Each spinner contains the same statements as in the case of known unrelated attribute, with probabilities P_1 and T_1 for the sensitive statement on the first and second spinner respectively. The unbiased estimator of λ_A is defined as:

$$\hat{\lambda}_A = \frac{1}{n(P_1 - T_1)}\left[(1-T_1)\sum_{i=1}^{n} y_{1i} - (1-P_1)\sum_{i=1}^{n} y_{2i}\right] \tag{2.4}$$

with $P_1 \neq T_1$ and having variance, where y_{1i} and y_{2i} are responses from the first and second spinner.

$$Var(\lambda_A) = C\Big[\left\{P_1(1-T_1)^2 + T_1(1-P_1)^2 - 2P_1T_1(1-P_1)(1-T_1)\right\}\lambda_A + \left\{(1-P_1)(1-T_1)(2-P_1-T_1) - 2(1-P_1)^2(1-T_1)^2\right\}\lambda_y, \tag{2.5}$$

where *C* is defined as $C = \frac{1}{n(P_1 - T_1)^2}$.

3. Bayesian estimation of λ using Land et al. [10] RRM

In the Bayesian estimation of the mean of rare sensitive attribute, we assume that the prior information about the parameter λ follows the Gamma distribution with parameter α and β, which can be written as

$$f(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda},\ \lambda > 0,\ \alpha, \beta > 0. \tag{3.1}$$

Now considering the Greenberg et al. [2] model, let X be the number of yes responses from the respondent using simple random sampling (SRS). According to them, $x_i = 1$ for the yes response with probability θ and $x_i = 0$ for no response with probability $1-\theta$, then by using the Land et al. [10] model, the conditional distribution of X given λ is as follows:

$$f_{X/\lambda}(x \mid \lambda) = \frac{e^{-\lambda}.\lambda^{x}}{x!} \quad \lambda > 0,\ x = 0,1,2,.... \tag{3.2}$$

and "λ" can be expressed by using the linear relation

$$\lambda = w\lambda_A + \upsilon, \tag{3.3}$$

where $w = P_1$ and $\upsilon = (1-P_1)\lambda_Y$.

The joint distribution of X and λ is given by the expression:

$$f(x,\lambda)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}.\frac{e^{-\lambda}.\lambda^{x}}{x!}. \tag{3.4}$$

The marginal distribution of X , by integrating the $f(X,\lambda)$ with respect to λ [19], is:

$$f(X)=\frac{\beta^{\alpha}\Gamma(\alpha+x)}{x!\times\Gamma(\alpha)(\beta+1)^{\alpha+x}}. \tag{3.5}$$

This is the probability mass function of the compound distribution to analyze the data.

Posterior distribution of $\lambda\mid x$ with its hyperparameter is given by:

$$\lambda\mid x\sim Gamma(\alpha+\sum x_i,\beta+n). \tag{3.6}$$

Bayes estimator under the squared error loss function (SELF) can be written as:

$$\hat{\lambda}_{Bayes}=\frac{\alpha+\sum x_i}{\beta+n}=\frac{\alpha'}{\beta'}. \tag{3.7}$$

The squared error loss function (SELF) is used as:

$$L(\lambda,d)=(\lambda-d)^2 \tag{3.8}$$

The estimator $\hat{\lambda}$ can also be defined as $\hat{\lambda}_{Bayes}=E(\lambda\mid x)$. Bayes posterior risk (BPR) is obtained as:

$$\underset{x,\lambda|x}{E}L(\lambda,d)=\rho^{*}(\circ)=\frac{\alpha}{\beta(\beta+n)}. \tag{3.9}$$

The total risk function is used as

$$\rho(\circ)=\rho^{*}(\circ)+nc, \tag{3.10}$$

to find the optimal sample size which increases the benefit and minimizes the cost (*say* c) of survey. The optimal sample size [See Appendix] is given by:

$$n=\left(\frac{\alpha}{\beta c}\right)^{\frac{1}{2}}-\beta. \tag{3.11}$$

We are interested in the Bayesian estimation of λ_A . From relation (3.3)

$$\lambda_A=\frac{\lambda-\upsilon}{w}. \tag{3.12}$$

As the distribution of λ is defined in (3.1), then prior for λ_A :

$$g(\lambda_A) = \frac{w\beta^{\alpha}}{\Gamma(\alpha)}(w\lambda_A + \upsilon)^{\alpha-1} e^{-\beta(w\lambda_A+\upsilon)},\ \lambda > 0, \alpha > 0, \beta > 0. \quad (3.13)$$

Gamma prior is the conjugate prior for Poisson distribution, so the posterior distribution of $\lambda_A \mid x$ can be presented as:

$$f_{\lambda_A|X}(\lambda_A \mid X) = \frac{f(\lambda_A, X)}{f(X)}. \quad (3.14)$$

This gives us the resulted posterior distribution as:

$$\lambda_A \mid x \sim Gamma(\alpha + \sum x_i, \beta + n). \quad (3.15)$$

Now the Bayes estimator for the mean number of persons possessing rare sensitive attribute is given by:

$$\hat{\lambda}_{A_{Bayes}} = E(\lambda_A \mid x) = \frac{\hat{\lambda}_{Bayes} - \upsilon}{w}. \quad (3.16)$$

The variance of the estimator is presented as:

$$Var\left(\hat{\lambda}_{Bayes}\right) = w^2 Var\left(\hat{\lambda}_{A_{Bayes}}\right). \quad (3.17)$$

When the proportion of rare unrelated attribute is known, the variance of the estimator is given by

$$Var\left(\hat{\lambda}_{A_{Bayes}}\right) = \frac{1}{w^2} Var\left(\hat{\lambda}_{Bayes}\right) = \frac{1}{w^2} \cdot \frac{\alpha + \sum x_i}{(\beta + n)^2} \quad (3.18)$$

In the case, when the proportion of rare unrelated attribute is unknown, the variance expression given in Equation (3.17) can be used to in the same way as for the first case.

4. Discussion and conclusion

In this article we have studied Land et al.'s [10] model from a Bayesian view point. With a prior distribution for λ_A, the posterior distribution is a compound of Gamma distribution. The mixture form reflects the uncertainty about how many respondents belong to Group A and Group Y.

The comparison is made using the different values of parameters. A simulation study is done by choosing the different pairs of parameters λ_A and λ_Y as (1, 10), (1, 20), (2, 10) and (5, 10), and we changed the value of P_1 from 0.6 to 0.9 with a small

difference of 0.001. In the Table-1, MSEs are presented for the specific values of P_1 and using the pairs of parameters for the fixed sample size. Complete simulated results are displayed graphically by using the scatter plot. It is observed that Bayes estimator gives the high relative efficiency for small values of the shape hyperparameter α'. From the Figures (1-4) (See Appendix), it can be easily seen that the Bayes estimator performs better than MLE. It is suggested to use the Bayesian estimates whenever to collect the information about the rare sensitive attribute.

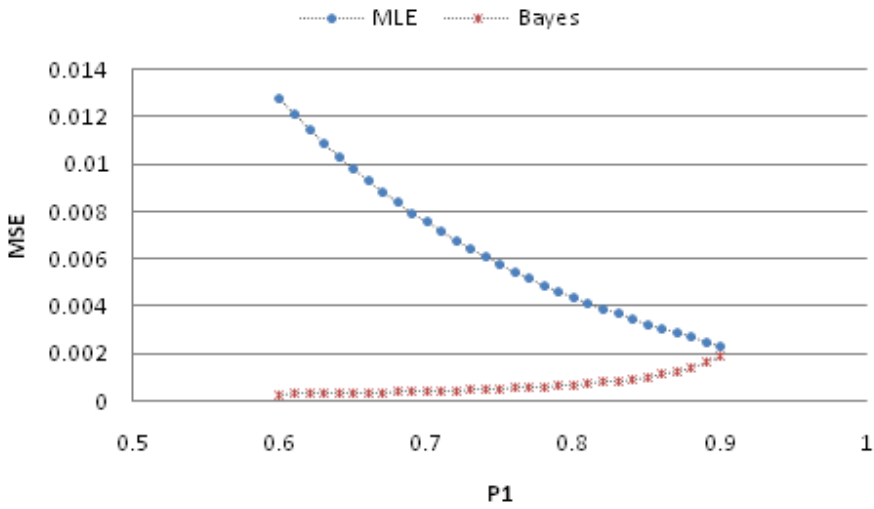

Figure 1. Graph of MSEs of $\hat{\lambda}_{A_{ML}}$ and $\hat{\lambda}_{A_{Bayes}}$ for $n = 1,000$, $\alpha = \lambda_A = 1$, $\beta = \lambda_Y = 10$.

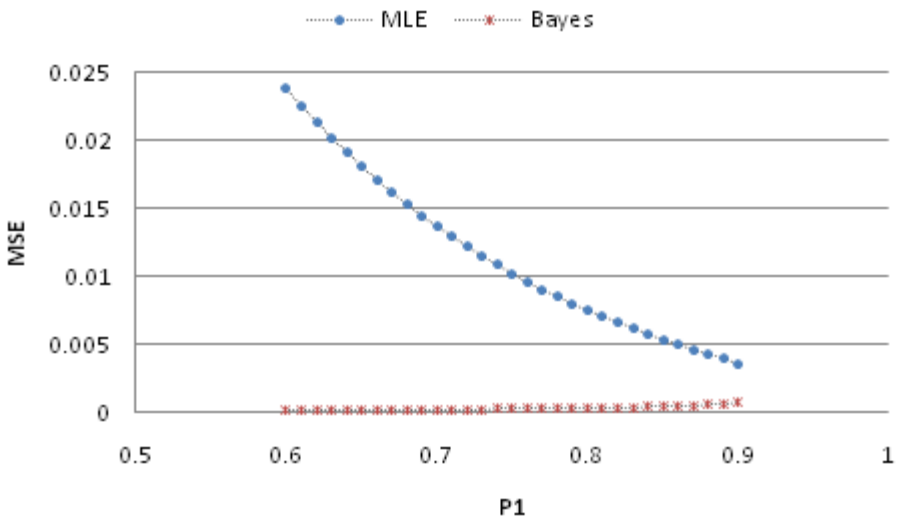

Figure 2. Graph of MSEs of $\hat{\lambda}_{A_{ML}}$ and $\hat{\lambda}_{A_{Bayes}}$ for $n = 1,000$, $\alpha = \lambda_A = 1$, $\beta = \lambda_Y = 20$.

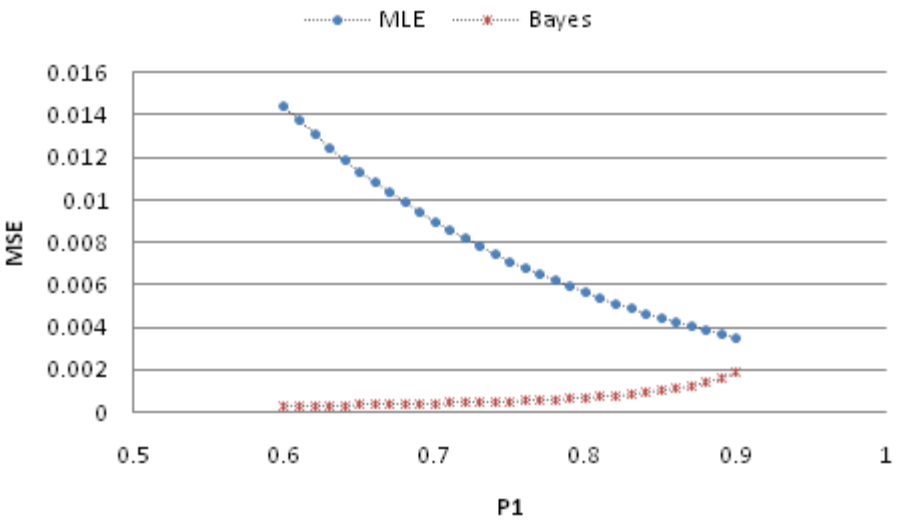

Figure 3. Graph of MSEs of $\hat{\lambda}_{A_{ML}}$ and $\hat{\lambda}_{A_{Bayes}}$ for $n = 1,000$, $\alpha = \lambda_A = 2$, $\beta = \lambda_Y = 10$.

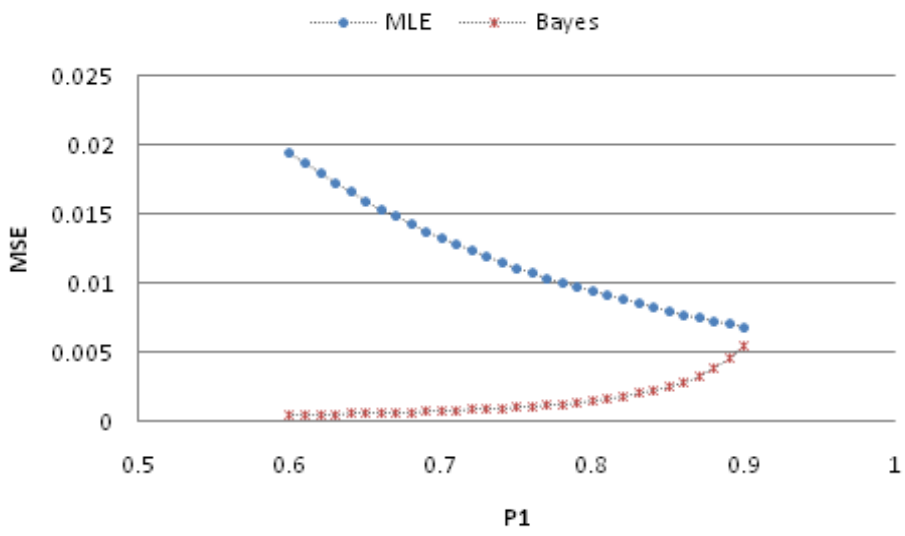

Figure 4. Graph of MSEs of $\hat{\lambda}_{A_{ML}}$ and $\hat{\lambda}_{A_{Bayes}}$ for $n = 1,000$, $\alpha = \lambda_A = 5$, $\beta = \lambda_Y = 10$.

In this paper, it is first time suggested to use the risk function for the privacy of the respondent which helps interviewer to get more reliable response. The risk function is also used to optimize the sample size and cost of the survey. To apply in practical situation the method developed in this article, one must assess the Gamma (prior) distribution for λ. The Assessment method of prior distributions is presented in Winkler [20] that can be used to assess the prior distribution of λ.

Table 1. Comparison of MLE and Bayes estimators by using different values of parameters.

n	$\lambda_A = \alpha$	$\lambda_Y = \beta$	P_1	MSE_{MLE}	MSE_{Bayes}
1,000	1	10	0.6	0.012778	0.00451
			0.7	0.007551	0.003628
			0.8	0.004375	0.002746
			0.9	0.002346	0.001864
1,000	1	20	0.6	0.023889	0.008267
			0.7	0.013673	0.006441
			0.8	0.0075	0.004615
			0.9	0.00358	0.002788
1,000	2	10	0.6	0.014444	0.0051
			0.7	0.00898	0.004315
			0.8	0.005625	0.003531
			0.9	0.003457	0.002747
1,000	5	10	0.6	0.019444	0.006867
			0.7	0.013265	0.006377
			0.8	0.009375	0.005803
			0.9	0.00679	0.005089

References

[1] Warner, S. L., Randomized response: a survey technique for eliminating evasive answer bias. Journal of American Statistical Association, 1965; 60: 63-66.

[2] Greenberg, B. G., Abul-Ela, A.L.A., Simmons, W.R. and Horvitz D. G., The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association, 1969; 64: 520-539.

[3] Chaudhuri, A. Mukerjee, R., Randomized Response: Theory and Techniques, New York: Marcel Dekker, 1988.

[4] Mangat, N. S. and Singh, R., An Alternative Randomized Response Procedure. Biometrika, 1990; **77**: 439-442.

[5] Mahmood, M. Singh, S. and Horn, S., On the Confidentiality Guaranteed under Randomized Response Sampling: A Comparison with Several New Techniques. Biometrical Journal, 1998; 40: 237-242.

[6] Bhargava, M. and Singh, R. A., modified randomization device for Warner's model. Statistica, 2000; 60: 315-321.

[7] Christofides, T.C., A generalized randomized response technique. Metrika, 2003;57: 195-200.

[8] Kim, J.M., and Warde W.D., A stratified Warner's randomized response model, Journal of Statistical Planning and Inference. 2004; 120: 155–165.

[9] Diana, G., and Perri, P. F., New Scrambled Response Models for Estimating the Mean of a Sensitive Quantitative Character. Journal Of Applied Statistics, 2010; 37: 1875-1890.

[10] Land, M., Singh, S., and Sedory, S. A., Estimation of a rare sensitive attribute using Poisson distribution, Statistics. 2012; 46: 351-360.

[11] Winkler, R.L., and Franklin, L. A., Warner's Randomized Response Model: A Bayesian Approach. Journal of the American Statistical Association, 1979; 74: 207-214.

[12] Pitz, G.F., Bayesian Analysis of Random Response Models. Psychological Bulletin, 1980; 87: 209-212.

[13] O'Hagan, A., Bayes Linear Estimators for Randomized Response Models. Journal of the American Statistical Association, 1987; 82: 580-585.

[14] Migon, H.S., and Tachibana, V. M., Bayesian approximations in randomized response model. Computational Statistics & Data Analysis, 1997; 24: 401-409.

[15] Unnikrishnan, N.K., and Kunte, S., Bayesian Analysis for Randomized Response Models, Sankhyā: The Indian Journal of Statistics, *Series B*, 1999; 61: 422-432.

[16] Bar-Lev, S., Bobovich, E., Boukai B. A., common conjugate prior structure for several randomized response models. TEST, 2003; 12: 101-113.

[17] Barabesi, L., and Marcheselli, M. A., Practical Implementation and Bayesian Estimation in Franklin's Randomized Response Procedure, Communications in Statistics-Simulation and Computation, 2006; 35: 563-573.

[18] Kim, J. M. and Tebbs, J. M., An S-W., Extensions of Mangat's randomized-response model. Journal of Statistical Planning and Inference, 2006; 136: 1554-1567.

[19] Berger, J.O., Statistical decision theory and Bayesian analysis, Second edition. New York, Springer 1985.

[20] Winkler R.L., The assessment of prior distribution in Bayesian analysis. Journal of American Statistical Association, 1967; 62, 776-800.

Appendix

Marginal distributions of the observed variable X

$$f(X)=\int_0^{\infty}\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}.\frac{e^{-\lambda}.\lambda^{x}}{x!}d\lambda\,,$$

$$f(X)=\frac{\beta^{\alpha}}{x!\Gamma(\alpha)}\int_0^{\infty}\lambda^{\alpha+x-1}e^{-\lambda(\beta+1)}d\lambda\,,$$

$$f(X)=\frac{\beta^{\alpha}\Gamma(\alpha+x)}{x!\times\Gamma(\alpha)(\beta+1)^{\alpha+x}}\,.$$

This is called the probability mass function of the compound distribution.

Posterior Distribution

Posterior distribution is defined as the mixture of prior information and likelihood information. Mathematically, it is defined as:

(Posterior Distribution) $\propto$ (Prior) . (Likelihood).

Bayes Estimator and Total Risk Function

Squared Error Loss Function (SELF) is used as $L(\lambda,d)=(\lambda-d)^2$. This can also be written as:

$$L(\lambda,d)=\left(\lambda^2+d^2-2\lambda d\right),$$

Differentiating the loss function with respect to *d* and equating to zero, we get:

$$\frac{\partial}{\partial d}\left\{L\left(\lambda,d^*\right)\right\}=\left(2d^*-2\hat{\lambda}\right)=0,\quad\Rightarrow\quad d^*=\hat{\lambda}.$$

where d^* is called the Bayes estimator which is equal to the mean of the posterior Gamma distribution.

Total Risk Function is defined as the fucntion of risk, sample size and cost for sampling.

$$\rho(\circ)=\underset{x,\lambda|x}{E}L(\lambda,d)+nc=\rho^*(\circ)+nc$$

Bayes Posterior Risk under Squared Error Loss function.

$$\underset{x,\lambda|x}{E}L(\lambda,d)=\rho^*(\circ)=\frac{\alpha}{\beta(\beta+n)}$$

Estimation of Optimal Sample Size

$$\rho(\circ)=\rho^*(\circ)+nc=\underset{x,\lambda|x}{E}L(\lambda,d)+nc=\frac{\alpha}{\beta\left(\beta+n\right)}+nc,$$

Differentiating this equation with respect to *n* and equating to zero as:

$$\frac{\partial}{\partial n}\left(\rho(\circ)\right)=\frac{\partial}{\partial n}\left\{\underset{x,\lambda|x}{E}L(\lambda,d)+nc\right\}=\frac{\partial}{\partial n}\left\{\frac{\alpha}{\beta\left(\beta+n\right)}+nc\right\}=0,$$

We get the optimal sample size as

$$n=\left(\frac{\alpha}{\beta c}\right)^{\frac{1}{2}}-\beta.$$