Use of Non-Informative Prior in Estimating Population Proportion

Horng-Jinh Chang[1] **Mei-Pei Kuo**[2]

(Received: Jan. 13, 2011；First Revision: Feb. 21, 2011；Accepted: Mar. 1, 2011)

Abstract

Randomized response techniques are designed for protecting the privacy of respondents and reducing the response bias while eliciting information on sensitive issues. Linear function of sample proportion is commonly used for estimating population proportion. By illustrating with the Warner (1965) model, we investigate the estimation problem of choosing estimators from various perspectives including minimizing the average mean square error and squared average bias of linear estimators. Three practical estimators are proposed, and these are compared with the regular estimator with respect to mean square error criterion. The results also cover to the case of direct response surveys. In particular, one of the proposed estimators may be viewed as a generalization of Böhning and Viwatwongkasem (2005) estimator.

***Keywords*:** Direct Response, Privacy Protection, Randomized Response

[1] Graduate Institute of Management Sciences Tamkang University Professor
[2] Graduate Institute of Management Sciences Tamkang University Ph.D. candidate

1. Introduction

In some socioeconomic investigations, conducting direct response (DR) surveys on sensitive topics is likely to encounter refusals or untruthful answers. To improve on respondent cooperation and to procure reliable data, Warner (1965) proposed the following randomized response (RR) technique. Consider a dichotomous population in which every person belongs either to a sensitive group A, or to its non-sensitive complement $\overline{\mathrm{A}}$. The problem of interest is to estimate the population proportion π of group A from a with-replacement simple random sample of size n. A randomization device used to collect sample information consists of two statements: (a) I am a member of group A, and (b) I am not a member of group A, represented with probabilities p and $(1-p)$ respectively. The interviewee chooses a statement and then simply replies truthfully 'yes' or 'no' to the statement chosen. The process of selecting one of the statements is unobserved by the interviewer. Denote by $\hat{\theta}$ the sample proportion of 'yes' answers obtained from n respondents. Assuming truthful reporting, Warner (1965) suggested an estimator of π as

$$\hat{\pi}_w = \frac{\hat{\theta}-(1-p)}{2p-1}, \quad p > 0.5, \tag{1}$$

which is unbiased with variance given by

$$Var(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{(2p-1)^2 n}. \tag{2}$$

A good exposition of modifications on RR technique and other related work could be referred to Chaudhuri and Mukerjee (1988). Some recent developments are Arnab and Dorffner (2007), Chaudhuri and Pal (2008), Huang (2008), Pal (2008), Yu et al. (2008), Bouza (2009), Diana and Perri (2009), and Huang (2010), etc.

It is clear that quite a large number of RR techniques are available in the literature. In most researches, the relationship between sample proportion $\hat{\theta}$ and the estimator $\hat{\pi}$ for π can be expressed as the general linear form $\hat{\pi} = \alpha\hat{\theta} + \beta$, where α and β are known in advance. Linearity is frequently imposed in order to be able to construct unbiased estimators. For example, the values of α and β are respectively given by $\alpha = (2p-1)^{-1}$ and $\beta = (1-p)(2p-1)^{-1}$ under Warner (1965) model. Such choices of α and β yield the Warner (1965) estimator $\hat{\pi}_w$, given in (1), being unbiased for π, but the corresponding mean square error (MSE) seems not to be smallest for all α and β. One may then consider the optimal choice of α and β for which MSE attains its minimum. However, the minimum value of MSE cannot be achieved in practice due to the known value of π involved. An attempt is therefore made here to consider the average bias and average MSE with respect to a uniform prior on interval $[0,1]$, so that some practical estimators may be constructed. By illustrating with Warner (1965) RR procedure, we evaluate the average bias and average

MSE of linear estimators, and in so doing consider the conditions under which average bias is zero and/or average MSE is minimized. The proposed estimators with principal properties are presented in the following section. With respect to mean square error criterion, efficiency comparison is carried out in section 3 to study the performances of the proposed estimators.

2. The Proposed Estimators

To estimate the population proportion π, let us consider a class of estimators as $\hat{\pi} = \alpha\hat{\theta} + \beta$, where α and β are suitably chosen constants. The bias and MSE of $\hat{\pi}$ are respectively given by

$$Bias(\hat{\pi}) = [\alpha(2p-1)-1]\pi + [\alpha(1-p)+\beta], \tag{3}$$

$$MSE(\hat{\pi}) = \frac{\alpha^2[(2p-1)^2\pi(1-\pi)+p(1-p)]}{n} + \{[\alpha(2p-1)-1]\pi + [\alpha(1-p)+\beta]\}^2. \tag{4}$$

On using expressions (3) and (4), the average bias and average MSE with respect to a uniform prior on interval $[0, 1]$ can respectively be obtained as

$$\int_0^1 Bias(\hat{\pi})d\pi = \frac{\alpha-1}{2} + \beta, \tag{5}$$

$$\int_0^1 MSE(\hat{\pi})d\pi = \frac{\alpha^2(1+2p-2p^2)}{6n} + \frac{1}{3}\{[\alpha(2p-1)-1]^2 + 3[\alpha(1-p)+\beta]^2 + 3[\alpha(2p-1)-1][\alpha(1-p)+\beta]\}. \tag{6}$$

In what follows, with the appropriate values of α and β, we suggest three specific linear estimators of the population proportion π.

2.1 The First Estimator

In order for a linear estimator to be of $\int_0^1 Bias(\hat{\pi})d\pi = 0$, on using expression (5), the relationship between α and β should be chosen such that $\beta = (1-\alpha)/2$. In that case, the linear estimator reduces to

$$\hat{\pi} = \alpha\hat{\theta} + \frac{1-\alpha}{2},$$

with average MSE given by

$$\int_0^1 MSE(\hat{\pi})d\pi = \frac{\alpha^2(1+2p-2p^2)}{6n} + \frac{[\alpha(2p-1)-1]^2}{12}. \tag{7}$$

One can then construct an estimator that is of smallest average MSE for all α. Differentiating (7) with respect to α and after some simple algebra, we have

$$\frac{d}{d\alpha}\int_0^1 MSE(\hat{\pi})d\pi = \frac{\alpha[(2p-1)^2 n + 2(1+2p-2p^2)] - (2p-1)n}{6n},$$

which equals zero only for

$$\alpha = \frac{(2p-1)n}{(2p-1)^2 n + 2(1+2p-2p^2)}. \tag{8}$$

And the second derivate is given by

$$\frac{d^2}{d\alpha^2}\int_0^1 MSE(\hat{\pi})d\pi = \frac{(2p-1)^2 n + 2(1+2p-2p^2)}{6n} > 0,$$

implying that (8) is indeed a minimum point of $\int_0^1 MSE(\hat{\pi})d\pi$. The resulting linear estimator of π, say $\hat{\pi}_1$, is then given by

$$\hat{\pi}_1 = \frac{(2p-1)n\hat{\theta} - (2p-1)(1-p)n + (1+2p-2p^2)}{(2p-1)^2 n + 2(1+2p-2p^2)}. \tag{9}$$

On substituting $\beta = (1-\alpha)/2$ and (8) into expressions (3) and (4), we have the bias and MSE of the estimator $\hat{\pi}_1$ respectively given by

$$Bias(\hat{\pi}_1) = \frac{(1+2p-2p^2)(1-2\pi)}{(2p-1)^2 n + 2(1+2p-2p^2)},$$

$$MSE(\hat{\pi}_1) = \frac{(2p-1)^2 n[(2p-1)^2\pi(1-\pi) + p(1-p)] + (1+2p-2p^2)^2(1-2\pi)^2}{[(2p-1)^2 n + 2(1+2p-2p^2)]^2}. \tag{10}$$

Hence, the estimator $\hat{\pi}_1$ preserves good properties including zero average bias and minimum average MSE among estimators of form $\hat{\pi} = \alpha\hat{\theta} + \beta$. It is noted that the choice of a value near 0.5 of the design parameter p may offer the respondents an adequate sense of protection in the case of a highly 'sensitive' issue, whereas a value of p close to unity may suffice in the case of a less 'sensitive' issue. When the survey issue is not sensitive, one may choose the value of p to be unity. In that case, expression (9) reduces to $\hat{\pi}_1 = (n\hat{\theta} + 1)(n+2)^{-1}$, which is identical to the Böhning and Viwatwongkasem (2005) estimator. Thus the estimator $\hat{\pi}_1$, given in (9), may be regarded as an extension of Böhning and Viwatwongkasem (2005) estimator to cover both the cases of DR and RR sampling surveys.

2.2 The Second Estimator

It is known that the Warner (1965) estimator $\hat{\pi}_w$ is deficient estimator in the sense that it may result in negative values as $\hat{\theta} < (1-p)/(2p-1)$. Essentially, the estimator $\hat{\pi}_1$ also suffers from this drawback. The population proportion π, if estimated by $\hat{\pi}_1$, will be negative, in case when

$$\hat{\theta} < \frac{(2p-1)(1-p)n-(1+2p-2p^2)}{(2p-1)n}.$$

As an amendment to $\hat{\pi}_w$ and $\hat{\pi}_1$, we then consider a class of estimators as $\hat{\pi} = \alpha\hat{\theta}$, where α is a non-negative constant. On substituting $\beta = 0$ into expressions (3) to (6), we get the properties of $\hat{\pi} = \alpha\hat{\theta}$ as

$$Bias(\hat{\pi}) = [\alpha(2p-1)-1]\pi + \alpha(1-p), \tag{11}$$

$$MSE(\hat{\pi}) = \frac{\alpha^2[(2p-1)^2\pi(1-\pi)+p(1-p)]}{n} + \{[\alpha(2p-1)-1]\pi + \alpha(1-p)\}^2, \tag{12}$$

$$\int_0^1 Bias(\hat{\pi})d\pi = \frac{\alpha-1}{2}, \tag{13}$$

$$\int_0^1 MSE(\hat{\pi})d\pi = \frac{\alpha^2(1+2p-2p^2)}{6n} + \frac{\alpha^2(1-p+p^2)-\alpha(1+p)+1}{3}. \tag{14}$$

From (13), it is observed that the choice $\alpha = 1$ results in $\int_0^1 Bias(\hat{\pi})d\pi = 0$, and the resulting estimator of π, say $\hat{\pi}_2$, is identical to the conventional estimator $\hat{\pi}_2 = \hat{\theta}$. Substituting $\alpha = 1$ into (11) and (12) yields the bias and MSE of $\hat{\pi}_2$ as

$$Bias(\hat{\pi}_2) = (1-p)(1-2\pi),$$

$$MSE(\hat{\pi}_2) = \frac{[(2p-1)^2\pi(1-\pi)+p(1-p)]+(1-p)^2(1-2\pi)^2 n}{n}. \tag{15}$$

It is evident that the choice $\alpha = 1$ results in average unbiased estimation for π, but the corresponding MSE seems not to be smallest for all α. We then consider the optimal choice of α such that the average MSE attains its minimum, which is studied as follows.

2.3 The Third Estimator

Through a simple application of calculus computation, the minimum value of (14) occurs in case when

$$\alpha = \frac{(1+p)n}{2(1-p+p^2)n+(1+2p-2p^2)}. \tag{16}$$

Accordingly, the resulting estimator of π, say $\hat{\pi}_3$, can be obtained as

$$\hat{\pi}_3 = \frac{(1+p)n\hat{\theta}}{2(1-p+p^2)n+(1+2p-2p^2)}. \tag{17}$$

On using expressions (11), (12) and (16), we have

$$Bias(\hat{\pi}_3) = \frac{(1+p)(1-p)n-[3(1-p)n+(1+2p-2p^2)]\pi}{2(1-p+p^2)n+(1+2p-2p^2)},$$

$$MSE(\hat{\pi}_3) = \frac{(1+p)^2 n[(2p-1)^2\pi(1-\pi)+p(1-p)]}{[2(1-p+p^2)n+(1+2p-2p^2)]^2} + \frac{\{(1+p)(1-p)n-[3(1-p)n+(1+2p-2p^2)]\pi\}^2}{[2(1-p+p^2)n+(1+2p-2p^2)]^2}. \quad (18)$$

It is noted that, if the value of p is chosen to be unity, the value of (16) reduces to $\alpha = 2n(2n+1)^{-1}$, and the estimator (17) reduces to $\hat{\pi}_3 = 2n(2n+1)^{-1}\hat{\theta}$.

3. Efficiency Comparison

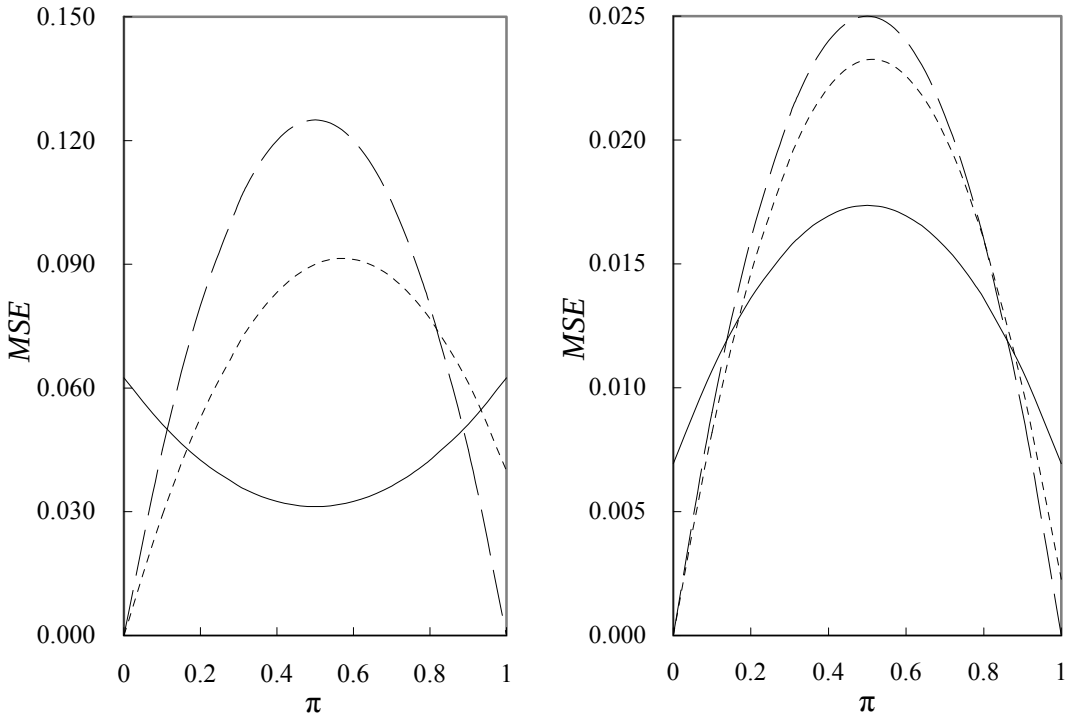
In this section, we study the efficiency aspect of the proposed estimators with respect to the mean square error criterion. In practical sampling surveys, some survey issues are sensitive but others may not be sensitive. In this regard, the efficiency comparisons are carried out for the two cases separately.

3.1 Direct Response Survey

In case when the survey issue is not sensitive, direct response surveys may be adopted. The value of p can be chosen to be unity, and the competing estimators reduce to $\hat{\pi}_1 = (n\hat{\theta}+1)(n+2)^{-1}$, $\hat{\pi}_2 = \hat{\pi}_w = \hat{\theta}$ and $\hat{\pi}_3 = 2n(2n+1)^{-1}\hat{\theta}$, with MSE respectively given by

$$MSE(\hat{\pi}_1) = \frac{n\pi(1-\pi)+(1-2\pi)^2}{(n+2)^2},\quad MSE(\hat{\pi}_2) = \frac{\pi(1-\pi)}{n},\quad MSE(\hat{\pi}_3) = \frac{4n\pi(1-\pi)+\pi^2}{(2n+1)^2}.$$

It is observed that $MSE(\hat{\pi}_1)$ is symmetric about 0.5, concave for $n > 4$, but convex otherwise. And, $MSE(\hat{\pi}_2)$ is symmetric about 0.5, and concave, while $MSE(\hat{\pi}_3)$ is symmetric about $0.5[1+(4n-1)^{-1}]$, and concave. To have some knowledge about the efficiencies, the MSEs of $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ are plotted in figure 1 for $n = 2$ and 10.



(a) $n = 2$ (b) $n = 10$

Figure 1. Mean square errors of the estimators $\hat{\pi}_1$ (solid), $\hat{\pi}_2$ (dashed) and $\hat{\pi}_3$ (dotted).

From figure 1, it is seen that there is an interval in which one of the three estimators is more efficient than others. Denote by (L, U) the interval for which $\hat{\pi}_1$ is more efficient than $\hat{\pi}_2$ and $\hat{\pi}_3$. To determine the value of L, equating $MSE(\hat{\pi}_1)$ and $MSE(\hat{\pi}_2)$, and after some algebraic simplification, we get

$$27n(n+1)\pi^2 - (28n^2 + 31n + 4)\pi + (2n+1)^2 = 0.$$

Solving the above equation for π yields the suitable solution, say L, given by

$$L = \frac{(28n^2 + 31n + 4) - \sqrt{352n^4 + 872n^3 + 645n^2 + 140n + 16}}{54n(n+1)}. \tag{19}$$

Also, equating $MSE(\hat{\pi}_1)$ and $MSE(\hat{\pi}_3)$ and after some simple algebra, we have

$$(8n+4)\pi^2 - (8n+4)\pi + n = 0.$$

The suitable solution, say U, for π can then be obtained as

$$U = \frac{(2n+1) + \sqrt{(n+1)(2n+1)}}{2(2n+1)}. \tag{20}$$

With the values of L and U given in (19) and (20), we conclude as follows.

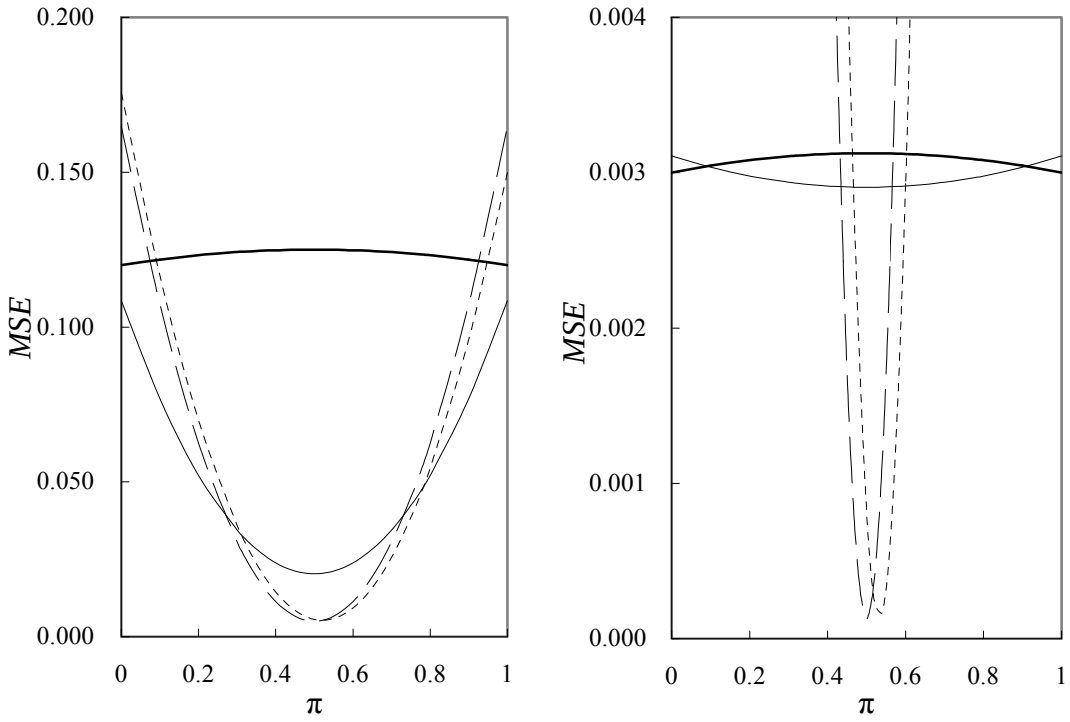
1. The estimator $(n\hat{\theta} + 1)(n+2)^{-1}$ is more efficient in case when $\pi \in (L, U)$.

2. The estimator $\hat{\theta}$ is more efficient in case when $\pi \in (U, 1]$.

3. The estimator $2n(2n+1)^{-1}\hat{\theta}$ is more efficient in case when $\pi \in [0, L)$.

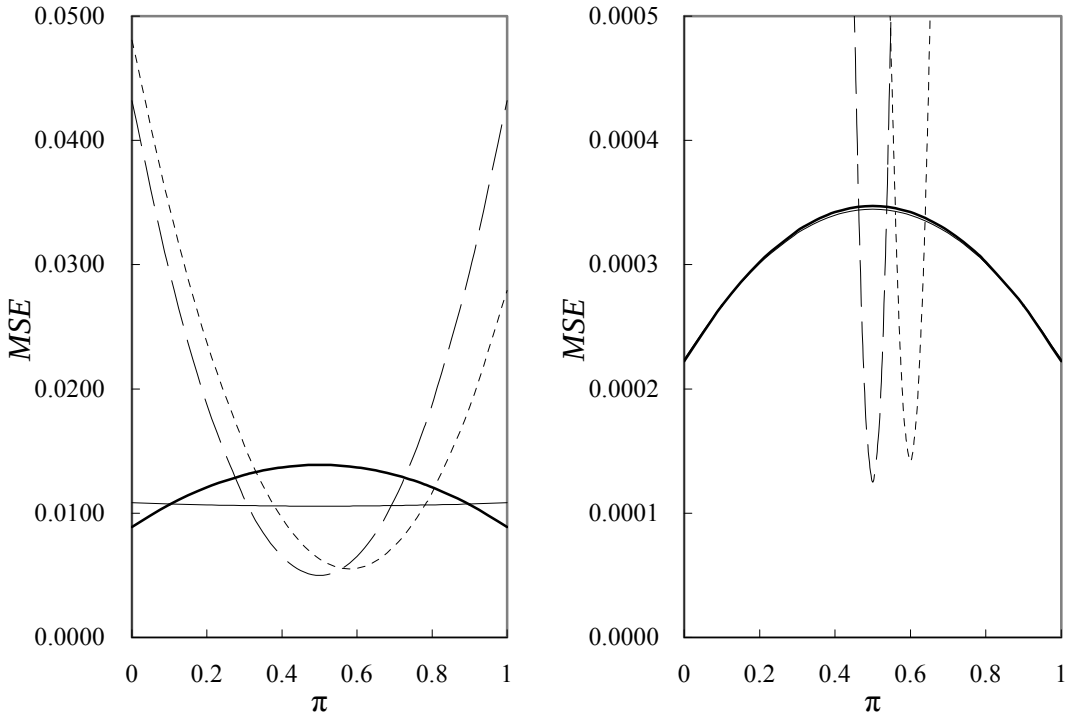
3.2 Randomized Response Survey

In what follows, an empirical study is first worked out to illustrate the possible relation of estimation efficiencies of $\hat{\pi}_w$, $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$. Since, in general, large sample size is required under randomized response sampling, without loss of generality, the sample size n is chosen to be 50 and 2000. The *MSE* s, given in (2), (10), (15) and (18), are displayed in figures 2 and 3 for $p = 0.6$ and 0.8, respectively.



(a) $n = 50$ (b) $n = 2000$

Figure 2. Mean square errors of the estimators $\hat{\pi}_w$ (boldfaced-solid), $\hat{\pi}_1$ (solid), $\hat{\pi}_2$ (dashed) and $\hat{\pi}_3$ (dotted) for $p = 0.6$.



(a) $n = 50$ (b) $n = 2000$

Figure 3. Mean square errors of the estimators $\hat{\pi}_w$ (boldfaced-solid), $\hat{\pi}_1$ (solid), $\hat{\pi}_2$ (dashed) and $\hat{\pi}_3$ (dotted) for $p = 0.8$.

As can be seen from figures 2 and 3, it seems complex to get apparent conclusion. The estimators are therefore pair-wise compared to found the conditions under which one estimator is more efficient than the other. We summarize as follows.

1. The estimator $\hat{\pi}_1$ is more efficient than $\hat{\pi}_w$ in case when $\pi \in (L_1, U_1)$, where

$$L_1 = \frac{1}{2} - \frac{\sqrt{[(2p-1)^2 n + (1+2p-2p^2)][2(1-p+p^2)n + (1+2p-2p^2)]}}{2(2p-1)[2(1-p+p^2)n + (1+2p-2p^2)]},$$

$$U_1 = \frac{1}{2} + \frac{\sqrt{[(2p-1)^2 n + (1+2p-2p^2)][2(1-p+p^2)n + (1+2p-2p^2)]}}{2(2p-1)[2(1-p+p^2)n + (1+2p-2p^2)]}.$$

2. The estimator $\hat{\pi}_2$ is more efficient than $\hat{\pi}_w$ in case when $\pi \in (L_2, U_2)$, where

$$L_2 = \frac{1}{2} - \frac{\sqrt{p[(1-p)n + p]}}{2(2p-1)[(1-p)n + p]}, \quad U_2 = \frac{1}{2} + \frac{\sqrt{p[(1-p)n + p]}}{2(2p-1)[(1-p)n + p]}.$$

3. The estimators $\hat{\pi}_3$ is more efficient than $\hat{\pi}_w$ in case when $\pi \in (L_3, U_3)$, where

$$L_3 = \frac{[2(1+p)(1-p)n^2 + (1-p+4p^2)n + (1+2p-2p^2)] - A}{2[3(1-p)n^2 + (2+p+2p^2)n + (1+2p-2p^2)]},$$

$$U_3 = \frac{[2(1+p)(1-p)n^2 + (1-p+4p^2)n + (1+2p-2p^2)] + A}{2[3(1-p)n^2 + (2+p+2p^2)n + (1+2p-2p^2)]},$$

where

$$A = \frac{\sqrt{[(1-p+4p^2)n + (1+2p-2p^2)](a_3n^3 + a_2n^2 + a_1n + a_0)}}{(2p-1)\sqrt{3(1-p)n + (1+2p-2p^2)}},$$

$$a_3 = 8(1-p)^2(1-p+p^2)(1+2p-2p^2),$$

$$a_2 = (1-p)(7+17p+4p^2-56p^3+80p^4-32p^5),$$

$$a_1 = 4(1+2p-2p^2)[1+2p^2(1-p)^2], \quad a_0 = (1+2p-2p^2)^2.$$

4. The estimators $\hat{\pi}_2$ is more efficient than $\hat{\pi}_1$ in case when $\pi \in (L_4, U_4)$, where

$$L_4 = \frac{1}{2} - \frac{\sqrt{p(2p-1)n + (1+2p-2p^2)}}{2\sqrt{(2p-1)(b_2n^2 + b_1n + b_0)}}, \quad U_4 = \frac{1}{2} + \frac{\sqrt{p(2p-1)n + (1+2p-2p^2)}}{2\sqrt{(2p-1)(b_2n^2 + b_1n + b_0)}},$$

where

$$b_2 = (2p-1)^2(1-p), \quad b_1 = 3+5p-14p^2+8p^3, \quad b_0 = (2p-1)(1+2p-2p^2).$$

5. The estimators $\hat{\pi}_3$ is more efficient than $\hat{\pi}_1$ in case when $\pi \in (L_5, U_5)$, where

$$L_5 = \frac{(c_3n^3 + c_2n^2 + c_1n + c_0) - \sqrt{d_5n^5 + d_4n^4 + d_3n^3 + d_2n^2 + d_1n + d_0}}{18(2p-1)n\{(1-p)(2p-1)^2n^2 + (2p+1)[(4-7p+4p^2)n + (1+2p-2p^2)]\}},$$

$$U_5 = \frac{(c_3n^3 + c_2n^2 + c_1n + c_0) + \sqrt{d_5n^5 + d_4n^4 + d_3n^3 + d_2n^2 + d_1n + d_0}}{18(2p-1)n\{(1-p)(2p-1)^2n^2 + (2p+1)[(4-7p+4p^2)n + (1+2p-2p^2)]\}},$$

where

$$c_3 = 6(2p-1)^3(1+p)(1-p), \quad c_2 = (2p-1)(23+25p-16p^2-76p^3+88p^4-16p^5),$$

$$c_1 = (1+2p-2p^2)(7+4p-36p^2+88p^3-32p^4), \quad c_0 = 4(1+2p-2p^2)^3,$$

$$d_5 = 24(2p-1)^4(1-p)(1-p+p^2)(1+2p-2p^2)(1-p+4p^2),$$

$$d_4 = (2p-1)^2(85-86p-135p^2+2028p^3-1428p^4-7344p^5+20352p^6 - 24576p^7+15552p^8-4352p^9+256p^{10}),$$

$$d_3 = 2(2p-1)(1+2p-2p^2)(5+213p+558p^2-1728p^3-348p^4+8424p^5 - 15072p^6+12480p^7-4608p^8+512p^9),$$

$$d_2 = 3(1+2p-2p^2)^2(39-12p-484p^2+1008p^3+1200p^4-5120p^5+6144p^6$$

$$-3072p^7+512p^8),$$

$$d_1=4(1+2p-2p^2)^4(23+8p-108p^2+176p^3-64p^4),\quad d_0=16(1+2p-2p^2)^6.$$

6. The estimators $\hat{\pi}_3$ is more efficient than $\hat{\pi}_2$ in case when $\pi\in(L_6,1]$, where

$$L_6=\frac{[(1-p)n+p][(3-p+2p^2)n+(1+2p-2p^2)]}{(1-p)(7-4p+4p^2)n^2+p(11-14p+8p^2)n+(2p-1)(1+2p-2p^2)}.$$

References

1. Arnab, R. and G. Dorffner (2007), "Randomized response techniques for complex survey designs," *Statistical Papers*, 48, pp. 131-141.
2. Böhning, D. and C. Viwatwongkasem (2005), "Revisiting proportion estimators," *Statistical Methods in Medical Research*, 14, pp. 147-169.
3. Bouza, C. N. (2009), "Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character," *Metrika*, 70, pp. 267-277.
4. Chaudhuri, A. and R. Mukerjee (1988), *Randomized Response: Theory and Techniques*, New York: Marcel Dekker.
5. Chaudhuri, A. and S. Pal (2008), "Estimating sensitive proportions from Warner's randomized responses in alternative ways restricting to only distinct units sampled," *Metrika*, 68, pp. 147-156.
6. Diana, G. and P. F. Perri (2009), "Estimating a sensitive proportion through randomized response procedures based on auxiliary information," *Statistical Papers*, 50, pp. 661-672.
7. Huang, K. C. (2008), "Estimation for sensitive characteristics using optional randomized response technique," *Quality & Quantity*, 42, pp. 679-686.
8. Huang, K. C. (2010), "Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling," *Metrika*, 71, pp. 341-352.
9. Pal, S. (2008), "Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses," *Statistical Papers*, 49, pp. 157-164.
10. Warner, S. L. (1965), "Randomized response: a survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 60, pp. 63-69.
11. Yu, J. W., G. L. Tian and M. L. Tang (2008), "Two new models for survey sampling with sensitive characteristic: design and analysis," *Metrika*, 67, pp. 251-263.