Metrika (2012) 75:655–672
DOI 10.1007/s00184-011-0346-8

Estimation of population proportion in randomized response sampling using weighted confidence interval construction

Horng-Jinh Chang · Mei-Pei Kuo

Received: 14 May 2010 / Published online: 11 February 2011

Abstract This paper considers the problem of estimation for binomial proportions of sensitive attributes in the population of interest. Randomized response techniques are suggested for protecting the privacy of respondents and reducing the response bias while eliciting information on sensitive attributes. By applying the Wilson (J Am Stat Assoc 22:209–212, 1927) score approach for constructing confidence intervals, various probable point estimators and confidence interval estimators are suggested for the common structures of randomized response procedures. In addition, efficiency comparisons are carried out to study the performances of the proposed estimators for both the cases of direct response surveys and randomized response surveys. Circumstances under which each proposed estimators is better are also identified.

Keywords Coverage probability · Direct response · Estimation efficiency · Privacy protection

1 Introduction

Social surveys sometimes include sensitive or threatening issues of enquiry, such as sexual behavior, drug use, and criminal behavior, that it is difficult to obtain valid and reliable information. If the respondents are asked directly about controversial topics, it often results in refusal or in untruthful responses, especially when they have

H.-J. Chang · M.-P. Kuo
Graduate Institute of Management Sciences, Tamkang University, Taipei, Taiwan
e-mail: chj@mail.tku.edu.tw

M.-P. Kuo (✉)
Department of Styling and Cosmetology, Tatung Institute of Commerce and Technology, Chiayi, Taiwan
e-mail: mpkuo.tw@gmail.com

committed sensitive behaviors. To improve respondent cooperation and to procure reliable data, Warner (1965) suggested an ingenious method called the randomized response technique. This survey technique allows the respondent to answer sensitive questions truthfully without revealing embarrassing or incriminating behavior. As a result, the technique assures a considerable degree of privacy protection in many contexts (Soeken and Macready 1982). Following the pioneering work of Warner (1965), many modifications are proposed in the literature. A good exposition of developments on randomized response techniques could refer to Chaudhuri and Mukerjee (1988). Some recent developments are Arnab and Dorffner (2007), Kim and Elam (2007), Chaudhuri and Pal (2008), Huang (2008), Pal (2008), Yu et al. (2008), Bouza (2009), Diana and Perri (2009), and Huang (2010), etc. In this paper, an attempt is made here to utilize the Wilson (1927) score approach to construct some point estimators and confidence interval estimators for the case of randomized response surveys.

Under the direct response surveys, confidence intervals are well known as an important aspect of reporting statistical results, and have been studied extensively in recent literature. The Clopper and Pearson (1934) "exact" method for constructing a confidence interval is unfortunate as it provides a coverage probability at least as large as the desired level with the tendency of providing a confidence level too large with respect to the desired level. Among approximate confidence intervals the simplest method commonly presented in elemantary statistics courses is the Wald interval, which is based on the normal approximation to the binomial distribution with variance estimated by the sample. The Wald interval has been solidly demonstrated that its coverage probability often being much lower than intended even for large sample sizes. The limits of Wald interval may be outside the parameter space [0, 1], and might result in the degenerate interval [0, 0] or [1, 1]. Without estimating the variance the Wilson (1927) score approach is a straightward application of normal approximation. The Wilson (1927) score approach has been shown to perform well, such as coverage probability close to the desired level and its limits are always belong to the parameter space [0, 1]. Therefore, Agresti and Coull (1998) comment that approximate intervals perform better than exact intervals. Bohning (1988) and Newcombe (1998) suggested the Wilson (1927) score approach as it performs well and computationally attractive. The Wilson (1927) score approach is described as follows.

Consider a dichotomous population in which every person belongs either to a group A, or to its complement $\bar{\text{A}}$. The problem of interest is to estimate the population proportion π of group A from a with-replacement simple random sample of size n. Denote by $\hat{\theta}$ the usual sample mean of the response. Then $\hat{\pi} = \hat{\theta}$ is the maximum likelihood estimator of π, which is unbiased with variance given by $Var(\hat{\pi}) = \pi(1-\pi)/n$. A two-sided $100(1-\alpha)\%$ confidence interval for π may be obtained according to the solution of the following quadratic inequality for π:

$$n(\hat{\pi}-\pi)^2 \leq z_{\alpha/2}^2 \pi(1-\pi), \tag{1}$$

where $z_{\alpha/2}^2$ denotes the upper $\alpha/2$ percentile of the standard normal distribution. If we denote by $W = n/(n+z_{\alpha/2}^2)$, on solving the inequality (1) for π, the computational formula of the confidence interval is given by

$$W\hat{\pi} + (1-W)\frac{1}{2} \pm z_{\alpha/2}\sqrt{\frac{W\hat{\pi}(1-\hat{\pi})}{n+z_{\alpha/2}^2} + \frac{1-W}{4(n+z_{\alpha/2}^2)}}, \tag{2}$$

which is recently studied by Olivier and May (2006). Instead of the regular estimator $\hat{\pi}$, the confidence interval (2) is symmetric about the point estimator

$$\frac{n}{n+z_{\alpha/2}^2}\hat{\pi} + \frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\frac{1}{2},$$

which is also suggested by Chen (1990). Obviously, the Wilson (1927) score approach provides not only a confidence interval but also a point estimator for π. It is noteworthy that the above point and interval estimations are studied only for the case of direct response surveys.

In Sect. 2, we consider the Wilson (1927) score approach under common randomized response framework. We then derive some point estimators and confidence interval estimators for some practical relevant concept. Sections 3 and 4 are devoted to empirical studies and comparisons for both the cases of direct response surveys and randomized response surveys. An operation rule for getting an effective estimator is also identified.

2 The proposed estimators

In order to estimate the proportion of a sensitive characteristic A, a randomization device is instructed to the respondents to collect sample data. Let θ denote the probability of getting a 'yes' response for a given randomized response model. In Warner (1965) model, two questions used in the randomization device are "Do you belong to group A?" and "Do you belong to group Ā?" with probabilities p and $(1-p)$ respectively. For Warner model, it is clear that $\theta = p\pi + (1-p)(1-\pi) = (2p-1)\pi + (1-p)$, $p \neq 0.5$. The randomization device used in the unrelated-question model, suggested by Horvitz et al. (1967) and then extended by Greenberg et al. (1969), consists of two questions: "Do you belong to group Ā?" and "Do you belong to group Y?", where Y denotes a neutral characteristic with known population proportion π_y. For unrelated-question model, we have $\theta = p\pi + (1-p)\pi_y$, $p \neq 0$. The Devore (1977) model is analogous to the unrelated-question model with one basic difference: The membership in group Y is certain, that is, $\pi_y = 1$. For this model, we have $\theta = p\pi + (1-p)$, $p \neq 0$. It is observed that the models mentioned above and some other models preserve a common property, which is described as follows.

Under the usual randomized response surveys, each respondent is provided with a randomization device by which he or she draws a question in a set of questions including the one relating to membership in the sensitive group A. The relationship between θ and π may be expressed in the following general form. Let v and w be nonnegative real numbers, we have $\theta = w\pi + v$, and therefore $\pi = (\theta - v)/w$, $0 \leq v < v + w \leq 1$.

Denote by $\hat{\theta}$ the sample proportion of 'yes' answers obtained from n respondents, an unbiased estimator of π is then given by

$$\hat{\pi} = \frac{\hat{\theta} - v}{w}, \tag{3}$$

with variance given by

$$Var(\hat{\pi}) = \frac{\theta(1-\theta)}{nw^2} = \frac{w^2\pi(1-\pi) + w(1-w-2v)\pi + v(1-v)}{nw^2}. \tag{4}$$

It is noted that the estimator $\hat{\pi}$ with $(v, w) = (0, 1)$ will reduce to the conventional estimator, the usual sample mean $\hat{\theta}$, which is commonly used in direct response surveys. Observe that with the generalized notation, the Warner (1965) model is obtained for $v = 1 - p$ and $w = 2p - 1$, the unrelated-question model is obtained for $v = (1-p)\pi_y$ and $w = p$, and the Devore (1977) model is obtained for $v = 1 - p$ and $w = p$. In what follows, from practical point of view, we consider the following three point estimators and confidence intervals for π.

2.1 The first estimator

Based on the unbiased estimator $\hat{\pi} = (\hat{\theta} - v)/w$, the inequality for constructing a two-sided $100(1-\alpha)\%$ confidence interval is given by

$$\left(\frac{\hat{\theta} - v}{w} - \pi\right)^2 \le z_{\alpha/2}^2 \left[\frac{w^2\pi(1-\pi) + w(1-w-2v)\pi + v(1-v)}{nw^2}\right]. \tag{5}$$

After some simple algebra, expression (5) can be expressed in the form

$$(n+z_{\alpha/2}^2)w^2\pi^2 - w[2n(\hat{\theta} - v)+z_{\alpha/2}^2(1-2v)]\pi + n(\hat{\theta} - v)^2 - z_{\alpha/2}^2 v(1-v) \le 0. \tag{6}$$

Solving (6) for π yields the following confidence interval CI_1, say, given by

$$W_1\frac{\hat{\theta} - v}{w} + (1-W_1)\frac{1-2v}{2w} \pm z_{\alpha/2}\sqrt{\frac{W_1\hat{\theta}(1-\hat{\theta})}{(n+z_{\alpha/2}^2)w^2} + \frac{1-W_1}{4(n+z_{\alpha/2}^2)w^2}}, \tag{7}$$

where $W_1 = n/(n + z_{\alpha/2}^2)$. The midpoint of interval (7) may be regarded as a point estimator $\hat{\pi}_1$, say, as

$$\hat{\pi}_1 = W_1\frac{\hat{\theta} - v}{w} + (1-W_1)\frac{1-2v}{2w}. \tag{8}$$

The mean square error of the estimator $\hat{\pi}_1$ can be obtained as

$$MSE(\hat{\pi}_1) = W_1^2 \frac{\theta(1-\theta)}{nw^2} + (1-W_1)^2 \left(\frac{1-2v}{2w} - \pi \right)^2, \tag{9}$$

where $\theta = w\pi + v$.

It is seen that the estimator $\hat{\pi}_1$ is obtained from the use of the unbiased estimator $\hat{\pi}$ in inequality (5). Another point estimator may further be established by using Wilson (1927) score approach according to the estimator $\hat{\pi}_1$. Repeating such a process sequentially, one may then interest in finding the limiting point estimator, which is studied as follows.

2.2 The second estimator

To derive the limiting point estimator, let us first consider the general linear-type estimator for π as $\hat{\pi}_l = a\hat{\theta} + b$, where a and b are known in advance. The mean square error of the general linear-type estimator $\hat{\pi}_l$ is given by

$$MSE(\hat{\pi}_l) = a^2 \frac{(w\pi + v)(1 - w\pi - v)}{n} + [(aw-1)\pi + (av+b)]^2.$$

On the basis of the general linear-type estimator, the Wilson (1927) score interval is based on the solution of the following inequality for π:

$$(a\hat{\theta} + b - \pi)^2 \le z_{\alpha/2}^2 \left\{ a^2 \frac{(w\pi + v)(1 - w\pi - v)}{n} + [(aw-1)\pi + (av+b)]^2 \right\},$$

which can be rewritten as

$$\begin{aligned} &\{n[1-(aw-1)^2 z_{\alpha/2}^2] + a^2 w^2 z_{\alpha/2}^2\}\pi^2 - [2n(a\hat{\theta}+b) + a^2 w(1-2v) z_{\alpha/2}^2 \\ &+2n(aw-1)(av+b) z_{\alpha/2}^2]\pi + n(a\hat{\theta}+b)^2 - a^2 v(1-v) z_{\alpha/2}^2 - n(av+b)^2 z_{\alpha/2}^2 \le 0, \end{aligned}$$

where $n[1-(aw-1)^2 z_{\alpha/2}^2] + a^2 w^2 z_{\alpha/2}^2 > 0$. The midpoint of the resulting confidence interval is given by

$$\frac{2na\hat{\theta} + 2n[b + (aw-1)(av+b) z_{\alpha/2}^2] + a^2 w(1-2v) z_{\alpha/2}^2}{2\left\{ n[1-(aw-1)^2 z_{\alpha/2}^2] + a^2 w^2 z_{\alpha/2}^2 \right\}}. \tag{10}$$

Suppose that the limiting estimator is of the form $\hat{\pi}_2 = \overline{a}\hat{\theta} + \overline{b}$. The limiting value $\overline{a}$ of the coefficient of $\hat{\theta}$ can be obtained by solving the following equality for $\overline{a}$:

$$\overline{a} = \frac{n\overline{a}}{n[1-(\overline{a}w-1)^2 z_{\alpha/2}^2] + \overline{a}^2 w^2 z_{\alpha/2}^2},$$

implying that the feasible solution of $\overline{a}$ is given by

$$\overline{a} = \frac{n}{(n+\sqrt{n})w}.$$

In the same way, the limiting value $\overline{b}$ can also be obtained as

$$\overline{b} = \frac{\sqrt{n}}{(n+\sqrt{n})}\frac{1-2v}{2w} - \frac{n}{(n+\sqrt{n})}\frac{v}{w}.$$

Denote by $W_2 = n/(n+\sqrt{n})$, we get the limiting point estimator as

$$\hat{\pi}_2 = W_2\frac{\hat{\theta}-v}{w} + (1-W_2)\frac{1-2v}{2w}, \tag{11}$$

with mean square error given by

$$MSE(\hat{\pi}_2) = \frac{n}{4(n+\sqrt{n})^2w^2}. \tag{12}$$

It is noted that $MSE(\hat{\pi}_2)$ is a constant as it is unrelated with the value of π. Due to the constant value of $MSE(\hat{\pi}_2)$, the confidence interval CI_2, say, is given by

$$\hat{\pi}_2 \pm z_{\alpha/2}\sqrt{\frac{n}{4(n+\sqrt{n})^2w^2}}. \tag{13}$$

2.3 The third estimator

An interesting property is observed that when $\hat{\theta} = 0.5$, the resulting values of the linear-type estimators $\hat{\pi}, \hat{\pi}_1$ and $\hat{\pi}_2$ are all equal to $(1-2v)/2w$. One may then intend to achieve the condition for such a feature. Essentially, on substituting $\hat{\theta} = 0.5$ into expression (10), the condition under which the resulting estimate being equal to $(1-2v)/2w$ can be obtained as

$$[(aw-1)z^2_{\alpha/2}+1][(aw-1)+2(v+wb)] = 0,$$

implying that

$$a = \frac{z^2_{\alpha/2}-1}{wz^2_{\alpha/2}}, \quad \text{or} \quad a+b = \frac{1-2v}{w}. \tag{14}$$

It is obvious that the regular unbiased estimator $\hat{\pi}$ and the two previous proposed estimators $\hat{\pi}_1$ and $\hat{\pi}_2$ merely satisfies the right hand side equality in (14). A worth mentioning choice of a and b satisfying both the equalities in condition (14) is that

$$a = \frac{z_{\alpha/2}^2 - 1}{wz_{\alpha/2}^2}, \quad \text{and} \quad b = \frac{1 - 2vz_{\alpha/2}^2}{2wz_{\alpha/2}^2}.$$

If we denote by $W_3 = n/(n + z_{\alpha/2}^2 - 1)$, we then obtain an estimator of π as

$$\hat{\pi}_3 = W_3 \frac{\hat{\theta} - v}{w} + (1 - W_3)\frac{1 - 2v}{2w}, \tag{15}$$

with mean square error given by

$$MSE(\hat{\pi}_3) = W_3^2 \frac{\theta(1-\theta)}{nw^2} + (1 - W_3)^2 \left(\frac{1 - 2v}{2w} - \pi\right)^2, \tag{16}$$

where $\theta = w\pi + v$. In addition, the corresponding confidence interval CI_3, say, is given by

$$\hat{\pi}_3 \pm z_{\alpha/2} \sqrt{\left(\frac{z_{\alpha/2}^2 - 1}{z_{\alpha/2}^2}\right)\left[\frac{W_3\hat{\theta}(1-\hat{\theta})}{(n + z_{\alpha/2}^2 - 1)w^2} + \frac{1 - W_3}{4(n + z_{\alpha/2}^2 - 1)w^2}\right] + K}, \tag{17}$$

where $K = W_3(1 - 2\hat{\theta})^2/4w^2 z_{\alpha/2}^4$.

An interesting property of the proposed point estimators $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ is that each estimator is a relatively simple weighted estimator incorporating sample information using $\hat{\pi} = (\hat{\theta} - v)/w$ and an uninformative prior using $(1 - 2v)/2w$, but with different weight factors W_1, W_2 and W_3. As n becomes large, each weight factor approaches unity and the corresponding estimator depends more on the sample information. In what follows, some empirical studies are carried out to study the performances of the proposed estimators. Since the proposed estimators also cover the case of direct response surveys, the performances are compared for direct response surveys and randomized response surveys separately. In particular, the randomization device is simply considered as the Warner (1965) device, so that $v = 1 - p$ and $w = 2p - 1$.

3 Comparison of point estimators

Here the efficiency aspect of the proposed point estimators is studied with respect to mean square error criterion.

3.1 Direct response surveys

It is known that when the survey attribute is not sensitive, direct response surveys may be adopted so that the choice $(v, w) = (0, 1)$ can be considered. In this case, expressions (3), (8), (11) and (15) respectively reduce to $\hat{\pi} = \hat{\theta}$; $\hat{\pi}_1 = (2n\hat{\theta} +$

$z_{\alpha/2}^2)/2(n + z_{\alpha/2}^2)$, which is also suggested by Chen (1990) and Olivier and May (2006); $\hat{\pi}_2 = (2n\hat{\theta} + \sqrt{n})/2(n + \sqrt{n})$, which is same as Casella and Berger (1990) estimator; and $\hat{\pi}_3 = (2n\hat{\theta} + z_{\alpha/2}^2 - 1)/2(n + z_{\alpha/2}^2 - 1)$. The mean square errors of the competing estimators can be obtained by substituting $(v, w) = (0, 1)$ into expressions (4), (9), (12) and (16). These are respectively given by

$$MSE(\hat{\pi}) = \frac{\pi(1-\pi)}{n}, \quad MSE(\hat{\pi}_1) = \frac{4n\pi(1-\pi) + z_{\alpha/2}^4(1-2\pi)^2}{4(n + z_{\alpha/2}^2)^2},$$

$$MSE(\hat{\pi}_2) = \frac{n}{4(n+\sqrt{n})^2}, \quad \text{and } MSE(\hat{\pi}_3) = \frac{4n\pi(1-\pi) + (z_{\alpha/2}^2 - 1)^2(1-2\pi)^2}{4(n + z_{\alpha/2}^2 - 1)^2}.$$

Since 95% confidence coefficient is commonly used, in what follows, the value of α is chosen to be 0.05 such that $z_{0.025} = 1.96$. As the graph is symmetric about 0.5, it is only required to consider $\pi \in [0, 0.5]$. To have an idea about the performances of the proposed estimators, we compute the relative efficiencies (RE) of $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ with respect to $\hat{\pi}$ by the formula

$$RE_i = \frac{MSE(\hat{\pi})}{MSE(\hat{\pi}_i)},$$

where $i = 1, 2, 3$. Relative efficiencies thus obtained are presented in Fig. 1 for $n = 30$ and 100. Note that $\hat{\pi}_i$ is more efficient than $\hat{\pi}$ if $RE_i > 1$.

From Fig. 1, it is seen that there is an interval in which one of the four competing estimators is more efficient than others. Denote by $(0, U_1)$ the interval in which $\hat{\pi}$ is better. Solving equation $MSE(\hat{\pi}) = MSE(\hat{\pi}_3)$ for π is identical to solve the following equation:

$$\pi^2 - \pi + \frac{n(z_{\alpha/2}^2 - 1)}{4(nz_{\alpha/2}^2 + n + z_{\alpha/2}^2 - 1)} = 0. \tag{18}$$

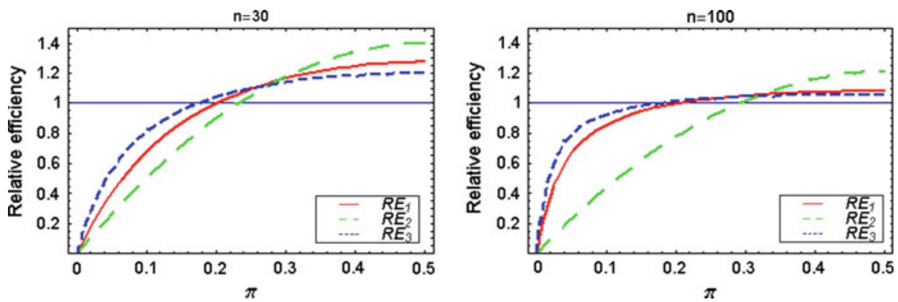

Fig. 1 Relative efficiency of the estimators $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ with respect to $\hat{\pi}$ for direct response surveys

The value of U_1 is the feasible solution of (18), which is given by

$$U_1 = \frac{1}{2}\left[1 - \frac{\sqrt{(2n + z_{\alpha/2}^2 - 1)(nz_{\alpha/2}^2 + n + z_{\alpha/2}^2 - 1)}}{nz_{\alpha/2}^2 + n + z_{\alpha/2}^2 - 1}\right]. \tag{19}$$

Similarly, on solving $MSE(\hat{\pi}_1) = MSE(\hat{\pi}_3)$ for π, and after some simple algebra, we have

$$\pi^2 - \pi + \frac{(2nz_{\alpha/2}^2 - n + 2z_{\alpha/2}^4 - 2z_{\alpha/2}^2)}{4(2nz_{\alpha/2}^2 + n + 2z_{\alpha/2}^4 - 1)} = 0.$$

The feasible solution, say U_2, for π can then be obtained as

$$U_2 = \frac{1}{2}\left[1 - \frac{\sqrt{(2n + 2z_{\alpha/2}^2 - 1)(2nz_{\alpha/2}^2 + n + 2z_{\alpha/2}^4 - 1)}}{2nz_{\alpha/2}^2 + n + 2z_{\alpha/2}^4 - 1}\right]. \tag{20}$$

On solving $MSE(\hat{\pi}_1) = MSE(\hat{\pi}_2)$ for π, we have

$$\pi^2 - \pi + \frac{n\left(n^2 + 2nz_{\alpha/2}^2 - nz_{\alpha/2}^4 - 2\sqrt{n}z_{\alpha/2}^4\right)}{4\left(n + \sqrt{n}\right)^2 (n - z_{\alpha/2}^4)} = 0.$$

The feasible solution, say U_3, for π can then be obtained as

$$U_3 = \frac{1}{2}\left[1 - \frac{\sqrt{n(n - z_{\alpha/2}^4)(2n\sqrt{n} - 2nz_{\alpha/2}^2 + n - z_{\alpha/2}^4)}}{(n + \sqrt{n})(n - z_{\alpha/2}^4)}\right]. \tag{21}$$

According to the values of U_1, U_2 and U_3 given in (19), (20) and (21), it is summarized that we should choose the estimator $\hat{\pi} = \hat{\theta}$ for $\pi \in [0, U_1] \cup [1 - U_1, 1]$; the estimator $\hat{\pi}_1 = (2n\hat{\theta} + z_{\alpha/2}^2)/2(n + z_{\alpha/2}^2)$ for $\pi \in [U_2, U_3] \cup [1 - U_3, 1 - U_2]$; the estimator $\hat{\pi}_2 = (2n\hat{\theta} + \sqrt{n})/2(n + \sqrt{n})$ for $\pi \in [U_3, 1 - U_3]$; and the estimator $\hat{\pi}_3 = (2n\hat{\theta} + z_{\alpha/2}^2 - 1)/2(n + z_{\alpha/2}^2 - 1)$ for $\pi \in [U_1, U_2] \cup [1 - U_2, 1 - U_1]$. Though we do not provide figures for larger sample size such as $n = 1000$, it is noteworthy that the results are quite similar to that given in Fig. 1. So, one may use the above conclusion to determine which estimator is better.

3.2 Randomized response surveys

Here the Warner (1965) randomization device is under consideration. For the appropriate choice of design parameter p, Greenberg et al. (1969) suggested a choice of p in the interval [0.7, 0.9]. Soeken and Macready (1982) suggested $p \in [0.7, 0.85]$, and

(Hedayat and Sinha, 1991, p. 318) suggested $p \in [0.6, 0.8]$. The values of p given in (Singh and Mangat, 1996, pp. 340–357) are chosen such that $p \in [0.7, 0.8]$. These suggest that a choice of p in the interval $[0.7, 0.8]$ would be most acceptable. The efficiency comparisons are then worked out for the cases of $p = 0.7$ and 0.8. Plots of relative efficiencies are displayed in Fig. 2 for $n = 30$ and 100.

From Fig. 2, it is seen that the estimator $\hat{\pi}$ seems to be of the least efficiency among the four competing estimators for both the cases of $p = 0.7$ and 0.8. When $p = 0.7$, the estimator $\hat{\pi}_2$ is more efficient than $\hat{\pi}$, $\hat{\pi}_1$ and $\hat{\pi}_3$ for all π when the sample size n is small. Instead, as n increases one may employ the estimator $\hat{\pi}_1$ for π being near zero. In case when $p = 0.8$, there is an interval in which one of the three proposed estimators is superior to others. It is noted that when p is fixed, the results for larger sample size are analogous to the case of $n = 100$. Relative position of relative efficiency curves remains unchanged, but all the curves approach the horizontal line $RE = 1$. In addition, though we do not provide the result for $p = 0.9$, plot of relative efficiency curves is similar to that of direct response surveys. In these regards, proceeding on the lines of Sect. 3.1, we present the following steps to determine the appropriate point estimator.

Step1. Calculate the value of U_3, where

$$U_3 = \frac{1}{2}\left[1 - \frac{\sqrt{n(n - z_{\alpha/2}^4)(2n\sqrt{n} - 2nz_{\alpha/2}^2 + n - z_{\alpha/2}^4)}}{(2p-1)(n+\sqrt{n})(n - z_{\alpha/2}^4)}\right].$$

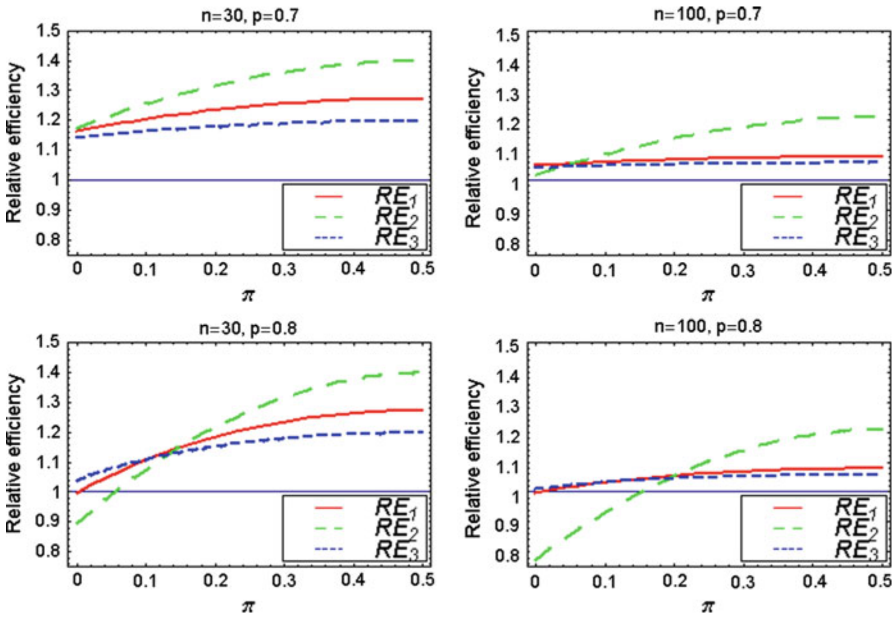

Fig. 2 Relative efficiency of the estimators $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ with respect to $\hat{\pi}$ for randomized response surveys

If $U_3 \leq 0$, one may prefer to the estimator $\hat{\pi}_2$, given in (11), for all π, and then stop. If $U_3 > 0$, one may choose the estimator $\hat{\pi}_2$ for $\pi \in [U_3, 1 - U_3]$, and then go to the Step 2.

Step2. Calculate the value of U_2, where

$$U_2 = \frac{1}{2}\left[1 - \frac{\sqrt{(2n + 2z_{\alpha/2}^2 - 1)(2nz_{\alpha/2}^2 + n + 2z_{\alpha/2}^4 - 1)}}{(2p - 1)(2nz_{\alpha/2}^2 + n + 2z_{\alpha/2}^4 - 1)}\right].$$

If $U_2 \leq 0$, one may prefer to the estimator $\hat{\pi}_1$, given in (8), for $\pi \in [0, U_3] \cup [1 - U_3, 1]$, and then stop. If $U_2 > 0$, one may choose the estimator $\hat{\pi}_1$ for $\pi \in [U_2, U_3] \cup [1 - U_3, 1 - U_2]$, and then go to the Step 3.

Step3. Calculate the value of U_1, where

$$U_1 = \frac{1}{2}\left[1 - \frac{\sqrt{(2n + z_{\alpha/2}^2 - 1)(nz_{\alpha/2}^2 + n + z_{\alpha/2}^2 - 1)}}{(2p - 1)(nz_{\alpha/2}^2 + n + z_{\alpha/2}^2 - 1)}\right].$$

If $U_1 \leq 0$, one may prefer to the estimator $\hat{\pi}_3$, given in (15), for $\pi \in [0, U_2] \cup [1 - U_2, 1]$, and then stop. If $U_1 > 0$, one may choose the estimator $\hat{\pi}_3$ for $\pi \in [U_1, U_2] \cup [1 - U_2, 1 - U_1]$, and choose the estimator $\hat{\pi}$, given in (3), for $\pi \in [0, U_1] \cup [1 - U_1, 1]$.

4 Comparison of confidence intervals

Here the performance of confidence intervals is examined by means of coverage probability and probability of generating meaningless limits.

4.1 Direct response surveys

On substituting the value $(v, w) = (0, 1)$ into expressions (7), (13) and (17), the proposed intervals CI_1, CI_2 and CI_3 are respectively given by

$$W_1\hat{\theta} + (1 - W_1)\frac{1}{2} \pm z_{\alpha/2}\sqrt{\frac{W_1\hat{\theta}(1 - \hat{\theta})}{n + z_{\alpha/2}^2} + \frac{1 - W_1}{4(n + z_{\alpha/2}^2)}},$$

$$W_2\hat{\theta} + (1 - W_2)\frac{1}{2} \pm z_{\alpha/2}\sqrt{\frac{n}{4(n + \sqrt{n})^2}},$$

$$W_3\hat{\theta} + (1 - W_3)\frac{1}{2} \pm z_{\alpha/2}\sqrt{\left(\frac{z_{\alpha/2}^2 - 1}{z_{\alpha/2}^2}\right)\left[\frac{W_3\hat{\theta}(1 - \hat{\theta})}{n + z_{\alpha/2}^2 - 1} + \frac{1 - W_3}{4(n + z_{\alpha/2}^2 - 1)}\right] + K},$$

where $K = W_3(1-2\hat{\theta})^2/4z_{\alpha/2}^4$. Except the above intervals, the Wald confidence interval is also under consideration. The Wald confidence interval CI, say, is given by

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

The first criterion used here to evaluate the performance of confidence intervals is the coverage probability. For a binomial distribution with parameter n and π, the coverage probability is given by

$$C_n(\pi) = \sum_{k=0}^{n} I(k,n)\binom{n}{k}\pi^k(1-\pi)^{n-k},$$

where $I(k,n)$ denotes the indicator function for the estimated confidence interval containing k. Mean, variance and minimum of coverage probability of the four intervals CI, CI_1, CI_2 and CI_3 are presented in Table 1 for $n = 10, 20, 30, 50, 100$. Plot of coverage probabilities for the case of $n = 30$ is illustrated in Fig. 3 for $\pi \in [0, 0.5]$. It is known that a confidence interval construction is desirable for which the coverage probability is close to the desired confidence level $100(1-\alpha)\%$. Table 1 and Fig. 3 separately shows the well known failure of Wald confidence interval CI. The confidence interval CI_1 seems to be of coverage probability closes to the desired confidence level, and of smaller variance. For the criterion of minimum value of coverage probability, the confidence interval CI_3 is superior.

Table 1 Mean, variance and minimum of coverage probability for direct response surveys

n	Measure	CI	CI_1	CI_2	CI_3
10	Mean	0.7693	0.9541	0.9535	0.9553
	Variance	0.0820	0.0005	0.0005	0.0005
	Minimum	0.0000	0.8350	0.8984	0.9020
20	Mean	0.8460	0.9530	0.9554	0.9635
	Variance	0.0420	0.0003	0.0004	0.0006
	Minimum	0.0000	0.8366	0.9165	0.9212
30	Mean	0.8749	0.9524	0.9564	0.9683
	Variance	0.0283	0.0002	0.0003	0.0008
	Minimum	0.0000	0.8371	0.9241	0.9275
50	Mean	0.9006	0.9518	0.9573	0.9740
	Variance	0.0172	0.0001	0.0003	0.0010
	Minimum	0.0000	0.8376	0.9310	0.9325
100	Mean	0.9223	0.9511	0.9583	0.9806
	Variance	0.0087	0.0001	0.0003	0.0014
	Minimum	0.0000	0.8379	0.9377	0.9380

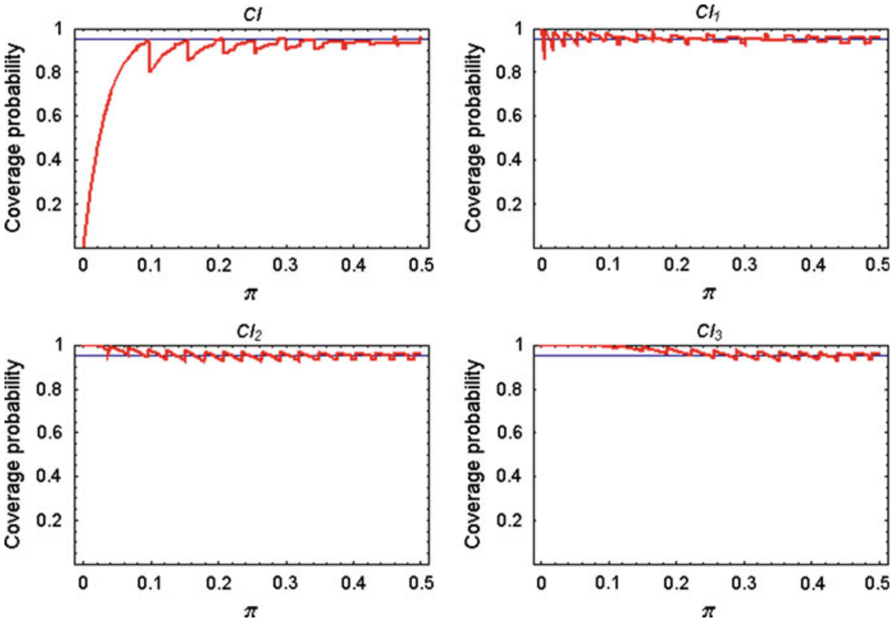

Fig. 3 Coverage probability of intervals CI, CI_1, CI_2 and CI_3 for direct response surveys

Another useful criterion is that a confidence interval should avoid generating meaningless limits such as either below zero or above unity, or both. The results are provided in Table 2. Table 2 clearly shows that the confidence interval CI_1 performs better than CI, CI_2 and CI_3 as its limits will always belong to the parameter space [0, 1].

Table 2 Probability of undesirable feature for direct response surveys

n	Event	CI	CI_1	CI_2	CI_3
10	Both limits outside [0,1]	0.0000	0.0000	0.0000	0.0000
	Limits below 0 or above 1	0.3636	0.0000	0.3636	0.3636
	Midpoint below 0 or above 1	0.0000	0.0000	0.0000	0.0000
20	Both limits outside [0,1]	0.0000	0.0000	0.0000	0.0000
	Limits below 0 or above 1	0.2857	0.0000	0.2857	0.3810
	Midpoint below 0 or above 1	0.0000	0.0000	0.0000	0.0000
30	Both limits outside [0,1]	0.0000	0.0000	0.0000	0.0000
	Limits below 0 or above 1	0.1935	0.0000	0.1935	0.3871
	Midpoint below 0 or above 1	0.0000	0.0000	0.0000	0.0000
50	Both limits outside [0,1]	0.0000	0.0000	0.0000	0.0000
	Limits below 0 or above 1	0.1176	0.0000	0.1569	0.3529
	Midpoint below 0 or above 1	0.0000	0.0000	0.0000	0.0000
100	Both limits outside [0,1]	0.0000	0.0000	0.0000	0.0000
	Limits below 0 or above 1	0.0594	0.0000	0.0990	0.3366
	Midpoint below 0 or above 1	0.0000	0.0000	0.0000	0.0000

Table 3 Mean, variance and minimum of coverage probability for $p = 0.7$ and 0.8

p	n	Measure	CI	CI_1	CI_2	CI_3
0.7	10	Mean	0.6059	0.7818	0.9454	0.9523
		Variance	0.2099	0.0809	0.0027	0.0020
		Minimum	0.0000	0.0000	0.7759	0.8113
	20	Mean	0.5350	0.6272	0.7344	0.8567
		Variance	0.2820	0.2028	0.1031	0.0226
		Minimum	0.0000	0.0000	0.0000	0.3618
	30	Mean	0.4764	0.5352	0.6116	0.7578
		Variance	0.3478	0.2934	0.2148	0.0770
		Minimum	0.0000	0.0000	0.0000	0.0000
	50	Mean	0.3996	0.4312	0.4793	0.6410
		Variance	0.4347	0.4039	0.3508	0.1758
		Minimum	0.0000	0.0000	0.0000	0.0000
	100	Mean	0.3024	0.3150	0.3395	0.4958
		Variance	0.5465	0.5336	0.5054	0.3202
		Minimum	0.0000	0.0000	0.0000	0.0000
0.8	10	Mean	0.7026	0.9415	0.9779	0.9790
		Variance	0.1088	0.0033	0.0013	0.0013
		Minimum	0.0000	0.7488	0.9349	0.9368
	20	Mean	0.6428	0.7881	0.9614	0.9778
		Variance	0.1777	0.0668	0.0008	0.0012
		Minimum	0.0000	0.0000	0.9069	0.9281
	30	Mean	0.5962	0.6941	0.9231	0.9777
		Variance	0.2264	0.1446	0.0028	0.0012
		Minimum	0.0000	0.0000	0.7677	0.9326
	50	Mean	0.5240	0.5829	0.7524	0.9783
		Variance	0.3025	0.2481	0.0880	0.0012
		Minimum	0.0000	0.0000	0.0000	0.9341
	100	Mean	0.4176	0.4442	0.5343	0.9806
		Variance	0.4182	0.3915	0.2930	0.0013
		Minimum	0.0000	0.0000	0.0000	0.9400

According to the results of Tables 1, 2 and Fig. 3, the confidence interval CI_1 is recommended for direct response surveys when constructing a 95% confidence interval for π.

4.2 Randomized response surveys

As the variance of $\hat{\pi} = (\hat{\theta} - v)/w$ is $\theta(1-\theta)/nw^2$, the Wald confidence interval CI for randomized response surveys is given by

$$\frac{\hat{\theta} - v}{w} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{nw^2}}.$$

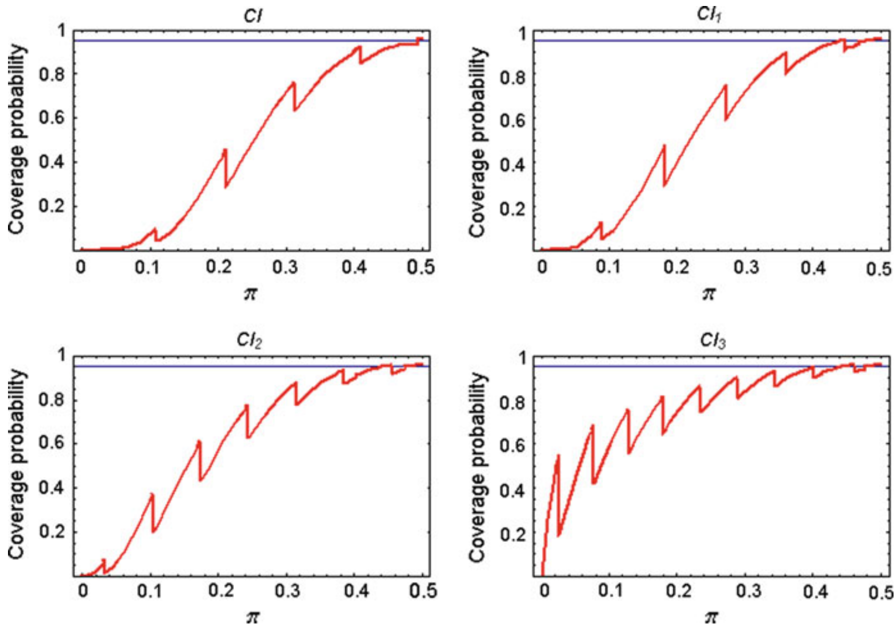

Fig. 4 Coverage probability of intervals CI, CI_1, CI_2 and CI_3 for $p = 0.7$

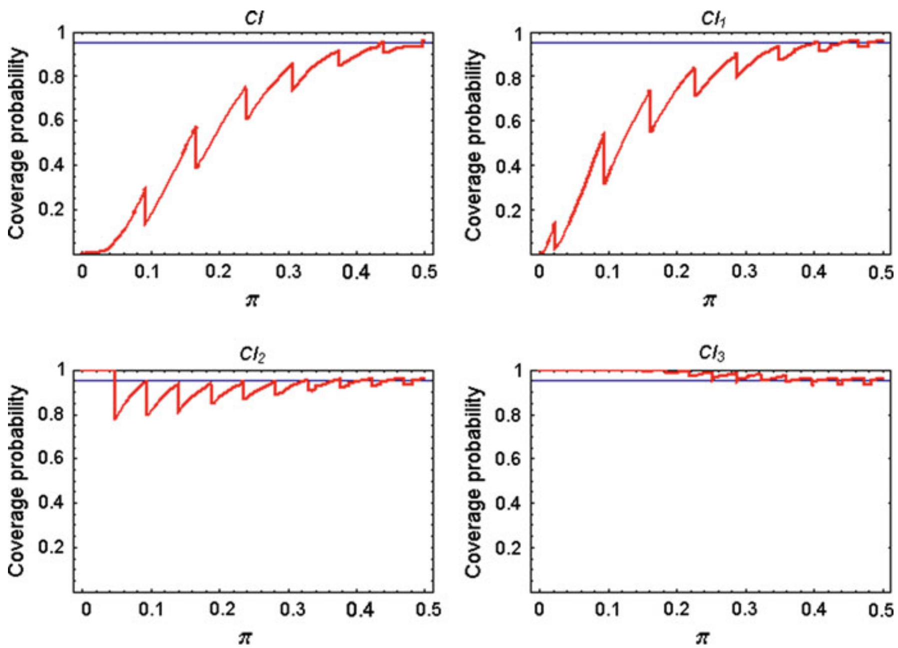

Fig. 5 Coverage probability of intervals CI, CI_1, CI_2 and CI_3 for $p = 0.8$

Table 4 Probability of undesirable feature for $p = 0.7$ and 0.8

p	n	Event	CI	CI_1	CI_2	CI_3
0.7	10	Both limits outside [0,1]	0.3636	0.1818	0.0000	0.0000
		Limits below 0 or above 1	1.0000	1.0000	1.0000	1.0000
		Midpoint below 0 or above 1	0.7273	0.5455	0.5455	0.5455
	20	Both limits outside [0,1]	0.2857	0.1905	0.0952	0.0000
		Limits below 0 or above 1	1.0000	1.0000	0.9524	0.9524
		Midpoint below 0 or above 1	0.6667	0.5714	0.5714	0.5714
	30	Both limits outside [0,1]	0.3226	0.3226	0.1935	0.0645
		Limits below 0 or above 1	0.9677	0.9032	0.9032	0.9032
		Midpoint below 0 or above 1	0.6452	0.5806	0.5161	0.5806
	50	Both limits outside [0,1]	0.3922	0.3529	0.2745	0.1176
		Limits below 0 or above 1	0.8627	0.8627	0.8235	0.8235
		Midpoint below 0 or above 1	0.6275	0.5882	0.5490	0.5882
	100	Both limits outside [0,1]	0.4356	0.4356	0.3762	0.1386
		Limits below 0 or above 1	0.7921	0.7723	0.7525	0.7921
		Midpoint below 0 or above 1	0.6139	0.5941	0.5743	0.5941
0.8	10	Both limits outside [0,1]	0.1818	0.0000	0.0000	0.0000
		Limits below 0 or above 1	1.0000	0.9091	0.9091	0.9091
		Midpoint below 0 or above 1	0.3636	0.1818	0.3636	0.3636
	20	Both limits outside [0,1]	0.1905	0.0952	0.0000	0.0000
		Limits below 0 or above 1	0.8571	0.7619	0.7619	0.7619
		Midpoint below 0 or above 1	0.3810	0.2857	0.2857	0.3810
	30	Both limits outside [0,1]	0.1935	0.1209	0.0000	0.0000
		Limits below 0 or above 1	0.7742	0.7097	0.6452	0.7097
		Midpoint below 0 or above 1	0.3871	0.3226	0.3226	0.3871
	50	Both limits outside [0,1]	0.2353	0.1961	0.0392	0.0000
		Limits below 0 or above 1	0.6667	0.6275	0.5882	0.6667
		Midpoint below 0 or above 1	0.3922	0.3529	0.3137	0.3922
	100	Both limits outside [0,1]	0.2772	0.2574	0.1584	0.0000
		Limits below 0 or above 1	0.5743	0.5545	0.5347	0.6337
		Midpoint below 0 or above 1	0.3960	0.3762	0.3366	0.3960

The three proposed confidence intervals are respectively given in (7), (13) and (17). Mean, variance and minimum of coverage probability of the four intervals CI, CI_1, CI_2 and CI_3 are presented in Table 3 for $p = 0.7$ and 0.8. Plot of coverage probabilities for the case of $n = 30$ is illustrated in Figs. 4 and 5 for $p = 0.7$ and 0.8 respectively. Table 3, Figs. 4 and 5 indicate that the Wald confidence interval CI performs worst. The confidence interval CI_3 performs well, followed by the interval CI_2. If we compare Figs. 3, 4 and 5, larger value of p will result in better performance in coverage probability. Table 4 further provides the probability of undesirable feature for the cases of $p = 0.7$ and 0.8. Table 4 clearly shows that the probability of undesirable feature for the four confidence intervals decreases as p increases. Overall, the empirical evidences of Tables 3, 4, Figs. 4 and 5 provide supports a choice of the confidence interval CI_3.

It is noted that the four confidence intervals CI, CI_1, CI_2 and CI_3 are centered at the point estimators $\hat{\pi}$, $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ respectively. Each point estimator is a relatively simple weighted mean of $\hat{\pi} = (\hat{\theta} - v)/w$ and $(1 - 2v)/2w$ with different weight factors 1, W_1, W_2 and W_3. As the sample size n increases, W_1, W_2 and W_3 approaches unity and the corresponding estimator $\hat{\pi}_1$, $\hat{\pi}_2$ and $\hat{\pi}_3$ depends more on the value of $\hat{\pi} = (\hat{\theta} - v)/w$. However, the value of $\hat{\pi} = (\hat{\theta} - v)/w$ is negative as $\hat{\theta} < v$, while greater than unity as $\hat{\theta} > w + v$. So, increasing sample size n might increase the possibility of meaningless limits, and thus decrease the coverage probabilities. Truncation does not help much, it merely improves the coverage probabilities for π close to zero or unity. To solve out this problem, one may employ some other admissible estimators such as the nonlinear estimator proposed by Raghavarao (1978). These are left for future studies.

Acknowledgments The authors are thankful to the editor and the learned referees to bring the original manuscript in the present form.

References

Agresti A, Coull BA (1998) Approximate is better than “exact” for interval estimation of binomial proportions. Am Stat 52:119–126

Arnab R, Dorffner G (2007) Randomized response techniques for complex survey designs. Stat Pap 48:131–141

Böhning D (1998) Confidence interval estimation of a rate and the choice of sample size. Stat Med 7:865–875

Bouza CN (2009) Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character. Metrika 70:267–277

Casella G, Berger RL (1990) Statistical inference. Wadsworth and Brooks/Cole, CA

Chaudhuri A, Mukerjee R (1988) Randomized response: theory and techniques. Marcel Dekker, New York

Chaudhuri A, Pal S (2008) Estimating sensitive proportions from Warner’s randomized responses in alternative ways restricting to only distinct units sampled. Metrika 68:147–156

Chen H (1990) The accuracy of approximate intervals for the binomial parameter. J Am Stat Assoc 85:514–518

Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26:404–413

Devore JL (1977) A note on the RR techniques. Commun Stat Theory Methods 6:1525–1529

Diana G, Perri PF (2009) Estimating a sensitive proportion through randomized response procedures based on auxiliary information. Stat Pap 50:661–672

Greenberg BG, Abul-Ela Abdel-Latif A, Simmons WR, Horvitz DG (1969) The unrelated question RR model: theoretical framework. J Am Stat Assoc 64:520–539

Hedayat AS, Sinha BK (1991) Design and inference in finite population sampling. Wiley, New York

Horvitz DG, Shah BV, Simmons WR (1967) The unrelated question RR model. Proc ASA Soc Stat Sec 65–72

Huang KC (2008) Estimation for sensitive characteristics using optional randomized response technique. Qual Quant 42:679–686

Huang KC (2010) Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling. Metrika 71:341–352

Kim JM, Elam ME (2007) A stratified unrelated question randomized response model. Stat Pap 48:215–233

Newcombe R (1998) Two-sided confidence intervals for the single proportion: Comparison of seven methods. Stat Med 17:857–872

Olivier J, May WL (2006) Weighted confidence interval construction for binomial parameters. Stat Methods Med Res 15:37–46

Pal S (2008) Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses. Stat Pap 49:157–164

Raghavarao D (1978) On an estimation problem in Warner's randomized response technique. Biometrics 34:87–90

Singh R, Mangat NS (1996) Elements of survey sampling. Kluwer, Dordrecht

Soeken KL, Macready GB (1982) Respondents' perceived protection when using randomized response. Psychol Bull 92:487–498

Yu JW, Tian GL, Tang ML (2008) Two new models for survey sampling with sensitive characteristic: design and analysis. Metrika 67:251–263

Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 60:63–69

Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22:209–212