International Journal of Statistics and Probability; Vol. 6, No. 5; September 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Improving Estimation Accuracy in Nonrandomized Response Questioning Methods by Multiple Answers

Heiko Groenitz[1]

[1] Working group Statistics, School of Business and Economics, Philipps-University Marburg, Germany

Correspondence: Heiko Groenitz, Working group Statistics, School of Business and Economics, Philipps-University Marburg, Germany. E-mail: groenitz@staff.uni-marburg.de

Received: June 24, 2017 Accepted: August 8, 2017 Online Published: August 11, 2017

doi:10.5539/ijsp.v6n5p101 URL: https://doi.org/10.5539/ijsp.v6n5p101

Abstract

When private or stigmatizing characteristics are included in sample surveys, direct questions result in low cooperation of the respondents. To increase cooperation, indirect questioning procedures have been established in the literature. Nonrandomized response methods are one group of such procedures and have attracted much attention in recent years. In this article, we consider four popular nonrandomized response schemes and present a possibility to improve the estimation precision of these schemes. The basic idea is to require multiple indirect answers from each respondent. We develop a Fisher scoring algorithm for the maximum likelihood estimation in the presented new schemes and show the better efficiency of the new schemes compared with the original designs.

Keywords: Fisher scoring, indirect questioning, Löwner order, privacy of respondents, sample survey, sensitive characteristic

1. Introduction

Surveys are important tools in many disciplines of science, for instance, social science and economics. Sometimes, variables which are viewed as private or stigmatizing are involved in the survey. Examples for such sensitive variables are financial situation, political views, cheating in examinations, undeclared work, insurance fraud, and discrimination. Direct questions on such characteristics will often yield low cooperation of the respondents, i.e., answer refusal and untruthful answers will often occur. Therefor, skilful questioning procedures that protect the interviewees' privacy and deliver data enabling statistical inference were developed in the literature. One group of procedures is the class of randomized response (RR) methods. In RR techniques, the respondent conducts a random experiment and gives a certain indirect answer depending on the result of the random experiment. For example, consider the following process with the sensitive attribute undeclared work and throwing a die as random experiment: If the die shows 1 or 2, the interviewee answers the question "Have you conducted work in the last year without declaring this to the relevant public authorities?". If the die shows 3-6, the opposite question "Did you declare all your work in the last year to the relevant public authorities?" must be responded. The interviewer does not observe the random experiment and hears only yes or no, but does not know the question that is answered. This protects the privacy. Based on the indirect answers of many respondents, the distribution of the sensitive variable can be estimated. The described procedure corresponds to the RR technique by Warner (1965). Various other RR methods are available today. See, for example, Fox and Tracy (1986), Chaudhuri (2011), Chaudhuri and Christofides (2013), or Chaudhuri, Christofides and Rao (2016) for overviews.

The random experiment in RR methods is a bit cumbersome and causes doubts on the suitability of RR methods for online surveys. This motivated diverse authors to introduce nonrandomized response (NRR) methods, for example, Yu, Tian, and Tang (2008), Tan, Tian, and Tang (2009), Tang, Tian, Tang, and Liu (2009) or Groenitz (2014). In NRR schemes, an indirect answer that depends on the respondent's outcome of an auxiliary characteristic must be given. The auxiliary characteristic is defined on the same population which the sensitive characteristic is defined on. Typically, the auxiliary characteristic is independent of the sensitive attribute and possesses a known distribution. To give an example, we mention the characteristic describing whether the respondent's birthday is in January - April or not. In NRR procedures, the respondent would give the same answer if he or she is asked again.

To improve the estimation efficiency of RR methods, some authors study repeated RR methods (Eriksson, 1973; Alavi & Tajodini, 2016; Groenitz, 2016). Here, the interviewee must repeat the random experiment multiple times. Say, we have two repetitions. Depending on the sensitive characteristic and the result of the first repetition of the experiment, the first indirect answer must be given. Depending on the sensitive characteristic and the result of the second repetition, the second indirect answer must be provided. That is, two indirect answers are necessary.

In this article, we present some repeated NRR techniques. We derive inference for these procedures and show that our repeated NRR methods improve the estimation efficiency of the original NRR techniques. The basic idea for repeated NRR techniques is to involve multiple different auxiliary characteristics in the procedure. For example, one can consider the characteristic describing whether the respondent's birthday is in January to April and the characteristic describing whether the respondent's telephone number ends on 0-6.

In Section 2, we explain the NRR methods considered in this paper. In Section 3, we describe the corresponding repeated NRR designs. The maximum likelihood (ML) estimation and the estimation variance for the multiple-trial NRR schemes are addressed in Section 4. The accuracy gains of repeated NRR techniques in comparison with single-trial NRR techniques are demonstrated in Section 5.

2. NRR Methods

In this section, four NRR methods are described: The crosswise method and the triangular method (both Yu et al., 2008), the multi-category design by Tang et al. (2009), and the diagonal technique by Groenitz (2014). Let the sensitive characteristic be denoted by X. We give some concrete examples for X:

(i) $X \in \{1, 2\}$ with $X = 1$ if the person has paid the taxes for the last year correctly and $X = 2$ if he or she has evaded taxes last year.

(ii) $X \in \{1, 2\}$ with $X = 1$ if the person's annual income exceeds a certain value and $X = 2$ else.

(iii) $X \in \{1, 2, 3\}$ where $X = 1$ holds if the person never has conducted insurance fraud, $X = 2$ holds if the person has conducted insurance fraud once or twice, and $X = 3$ holds if the person has conducted insurance fraud three or more times.

(iv) $X \in \{1, 2, 3, 4\}$ where each value of X represents a certain income class.

For the crosswise and triangular method, $X \in \{1, 2\}$, i.e., X with two categories, is required. For the methods by Tang et al. (2009) and Groenitz (2014), X can have an arbitrary number of categories coded by $1, 2, ..., k$. The triangular method and the Tang et al. (2009) method demand that the category $X = 1$ is nonsensitive. The crosswise method can be applied for the examples (i) and (ii). The triangular method can handle example (i). The technique of Tang et al. (2009) is suitable for (iii) and the diagonal technique can be applied for (iii) and (iv).

For each of the considered NRR designs, a nonsensitive auxiliary variable W is necessary. The respondents' individual values of W must not be known to the interviewer or the survey agency. W and X must be independent and W must possess a known distribution. For the crosswise and triangular method, W must have the categories $W = 1$ and $W = 2$. For the Tang et al. (2009) method and the diagonal technique, W must have the k categories $W = 1, ..., W = k$. Examples for W with two categories were already given in the Introduction. A W with $k = 4$ is as follows: Let W be based on the number formed by the last three digits of the interviewee's telephone number. If this number is ≤ 624, $625 - 749$, $750 - 874$, and $875 - 999$, we define $W = 1$, $W = 2$, $W = 3$, and $W = 4$, respectively. For example, the telephone number 9478722 results in the number 722 and $W = 2$.

In the survey, the respondents provide an indirect answer A that depends on X and W. Giving an indirect answer A protects the privacy. The concrete answer schemes are:

- Crosswise method: For $X = W = 1$ or $X = W = 2$, the answer $A = 1$ must be given. For other combinations of X and W, the indirect answer is $A = 2$.
- Triangular method: For $X = W = 1$, we have $A = 1$. In the other cases, $A = 2$ is required.
- Tang et al. (2009) method: For $X = 1$, the answer is the value of the nonsensitive variable, that is, $A = W$. For $X = i$ with $i = 2, ..., k$, the answer is the value of the sensitive characteristic, that is, $A = X$.
- Diagonal technique: The answer is given by the formula $A = [(W - X) \bmod k] + 1$, however, the respondents do not receive this mathematical formula. Instead, they receive a table that illustrates the answer to give. For example, for $k = 4$, Table 1 is such a table.

Table 1. Table of required indirect answer A for diagonal technique

X/W	$W = 1$	$W = 2$	$W = 3$	$W = 4$
$X = 1$	**1**	**2**	**3**	**4**
$X = 2$	**4**	**1**	**2**	**3**
$X = 3$	**3**	**4**	**1**	**2**
$X = 4$	**2**	**3**	**4**	**1**

3. Repeated NRR Methods

In this section, we introduce a repeated version for each of the NRR methods from Section 2. Here, every respondent gives multiple indirect answers. We consider the case of two indirect answers in particular.

As preliminary consideration, let us fix some NRR scheme from Section 2 and assume that the respondent should give a first indirect answer A_1 based on the sensitive X and the nonsensitive auxiliary characteristic W and a second indirect answer A_2 also based on X and W. Then, $A_1 = A_2$ always follows. Consequently, the second indirect answer does not contain additional information. Thus, it does not work to base both indirect answers on X and W.

The solution is to utilize a separate nonsensitive auxiliary attribute for each repetition. Say, the nonsensitive auxiliary characteristic for the first and second trial is denoted by W_1 and W_2, respectively. For a fixed NRR scheme from Section 2, the interview procedure for the two-trial version is as follows. The interviewee first gives the indirect answer A_1 depending on X and W_1 according to the fixed NRR scheme. Afterward, he or she gives the second indirect answer A_2 depending on X and W_2 also according to the selected NRR scheme.

For each NRR technique from Section 2, neither the respondent's value of W_1 nor the value of W_2 must be known to the interviewer or the survey agency. For the crosswise and triangular method, $W_1, W_2 \in \{1, 2\}$ is necessary. For the Tang et al. (2009) and Groenitz (2014) method, W_1 and W_2 both must have the categories $1, ..., k$. We make three further assumptions: The vector (W_1, W_2) and X are independent, W_1 and W_2 are independent, and W_1 and W_2 possess known distributions (W_1 and W_2 are allowed to have different distributions). These three assumptions can usually be seen as fulfilled when the auxiliary characteristics are constructed, for example, from birthday periods, telephone numbers, or house numbers.

4. Statistical Inference for Repeated NRR Designs

We define π_i to be the proportion of persons in the population having X value equal to i ($i = 1, ..., k$) and set $\pi = (\pi_1, ..., \pi_{k-1})^\top$. We now develop the ML estimation for π for the repeated NRR designs and a sample of size n drawn by simple random sampling with replacement. The estimation variance is also addressed. Fix one of the repeated NRR designs and define c_{1i} and c_{2i} to be the proportion of population units with $W_1 = i$ and $W_2 = i$, respectively. Let the entry (i, j) of the $k \times k$ matrix C_1 be given by $\mathbb{P}(A_1 = i|X = j)$. Analog, let the entry (i, j) of the $k \times k$ matrix C_2 be given by $\mathbb{P}(A_2 = i|X = j)$. For the crosswise method, we have

$$C_1 = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{11} \end{pmatrix} \text{ and } C_2 = \begin{pmatrix} c_{21} & c_{22} \\ c_{22} & c_{21} \end{pmatrix}.$$

For the triangular method,

$$C_1 = \begin{pmatrix} c_{11} & 0 \\ c_{12} & 1 \end{pmatrix} \text{ and } C_2 = \begin{pmatrix} c_{21} & 0 \\ c_{22} & 1 \end{pmatrix}$$

hold. For the technique by Tang et al. (2009), the first column of C_1 equals $(c_{11}, ..., c_{1k})^\top$. The jth column of C_1 for $j = 2, ..., k$ has entry 1 as jth component while the other components are 0. In the matrix C_2, the first column is $(c_{21}, ..., c_{2k})^\top$. The jth column of C_2 for $j = 2, ..., k$ has entry 1 as jth component and entry 0 for the other components.

For the diagonal technique, each row of C_1 is a left-cyclic shift of the row above and the first row is $(c_{11}, ..., c_{1k})$. Regarding C_2, each row is again a left-cyclic shift of the row above where the first row is now $(c_{21}, ..., c_{2k})$.

Consider $a_1, a_2, x \in \{1, ..., k\}$, define

$$\begin{aligned} I_1 &= I_1(a_1, x) = \{i \in \{1, ..., k\} \,:\, W_1 = i, X = x \text{ result in answer } A_1 = a_1\}, \\ I_2 &= I_2(a_2, x) = \{j \in \{1, ..., k\} \,:\, W_2 = j, X = x \text{ result in answer } A_2 = a_2\}, \\ I &= I(a_1, a_2, x) = \{(i, j) \in \{1, ..., k\}^2 \,:\, (W_1, W_2) = (i, j), X = x \text{ yield } (A_1, A_2) = (a_1, a_2)\} \end{aligned}$$

and obtain

$$\begin{aligned}
\mathbb{P}(A_1 = a_1, A_2 = a_2 | X = x) &= \mathbb{P}(X = x)^{-1} \cdot \mathbb{P}(A_1 = a_1, A_2 = a_2, X = x) \\
&= \mathbb{P}(X = x)^{-1} \cdot \left(\sum_{(w_1,w_2)\in I} \mathbb{P}(A_1 = a_1, A_2 = a_2, W_1 = w_1, W_2 = w_2, X = x) \right. \\
&\qquad \left. + \sum_{(w_1,w_2)\notin I} \mathbb{P}(A_1 = a_1, A_2 = a_2, W_1 = w_1, W_2 = w_2, X = x) \right) \\
&= \mathbb{P}(X = x)^{-1} \cdot \left(\sum_{(w_1,w_2)\in I} \mathbb{P}(W_1 = w_1, W_2 = w_2, X = x) + 0 \right) \\
&= \mathbb{P}(X = x)^{-1} \cdot \left(\sum_{(w_1,w_2)\in I} \mathbb{P}(W_1 = w_1) \cdot \mathbb{P}(W_2 = w_2) \cdot \mathbb{P}(X = x) \right) \\
&= \sum_{(w_1,w_2)\in I} \mathbb{P}(W_1 = w_1) \cdot \mathbb{P}(W_2 = w_2) = \sum_{w_1 \in I_1} \mathbb{P}(W_1 = w_1) \cdot \sum_{w_2 \in I_2} \mathbb{P}(W_2 = w_2) \\
&= \mathbb{P}(A_1 = a_1 | X = x) \cdot \mathbb{P}(A_2 = a_2 | X = x) = C_1(a_1, x) \cdot C_2(a_2, x),
\end{aligned}$$

where entry (p, q) of the matrix C_1 and C_2 is denoted by $C_1(p, q)$ and $C_2(p, q)$, respectively. Consequently, A_1 and A_2 are conditionally independent. As next step, we define λ_{ij} to be the joint proportion of population units with $A_1 = i$ and $A_2 = j$ $(i, j = 1, ..., k)$. These joint proportions are arranged in the column vector λ of length k^2 where we first sort by the value of A_1. For example, for $k = 3$, λ is given by $\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{21}, \lambda_{22}, \lambda_{23}, \lambda_{31}, \lambda_{32}, \lambda_{33})^\top$. It follows that $\lambda = C \cdot (\pi_1, ..., \pi_k)^\top$ where C is a $k^2 \times k$ matrix and the jth column of C is given by $C_1(:, j) \otimes C_2(:, j)$. Here, $C_1(:, j)$ and $C_2(:, j)$ represents the jth column of C_1 and C_2, respectively, and the symbol $\otimes$ stands for the Kronecker matrix product. The Kronecker matrix product of two matrices $R \in \mathbb{R}^{r_1 \times r_2}$ and $S \in \mathbb{R}^{s_1 \times s_2}$ is defined as

$$R \otimes S = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1,r_2} \\ R_{21} & R_{22} & \cdots & R_{2,r_2} \\ \vdots & \vdots & & \vdots \\ R_{r_1,1} & R_{r_1,2} & \cdots & R_{r_1,r_2} \end{pmatrix} \otimes S = \left(\begin{array}{c|c|c|c} R_{11} \cdot S & R_{12} \cdot S & \cdots & R_{1,r_2} \cdot S \\ \hline R_{21} \cdot S & R_{22} \cdot S & \cdots & R_{2,r_2} \cdot S \\ \hline \vdots & \vdots & & \vdots \\ \hline R_{r_1,1} \cdot S & R_{r_1,2} \cdot S & \cdots & R_{r_1,r_2} \cdot S \end{array} \right),$$

that is, $R \otimes S$ is a matrix of size $r_1 s_1 \times r_2 s_2$. Thus, C is the columnwise Kronecker product of C_1 and C_2. To give an example, for $k = 3$, we have

$$C = \left(\begin{array}{c|c|c}
C_1(1,1) \cdot C_2(1,1) & C_1(1,2) \cdot C_2(1,2) & C_1(1,3) \cdot C_2(1,3) \\
C_1(1,1) \cdot C_2(2,1) & C_1(1,2) \cdot C_2(2,2) & C_1(1,3) \cdot C_2(2,3) \\
C_1(1,1) \cdot C_2(3,1) & C_1(1,2) \cdot C_2(3,2) & C_1(1,3) \cdot C_2(3,3) \\ \hline
C_1(2,1) \cdot C_2(1,1) & C_1(2,2) \cdot C_2(1,2) & C_1(2,3) \cdot C_2(1,3) \\
C_1(2,1) \cdot C_2(2,1) & C_1(2,2) \cdot C_2(2,2) & C_1(2,3) \cdot C_2(2,3) \\
C_1(2,1) \cdot C_2(3,1) & C_1(2,2) \cdot C_2(3,2) & C_1(2,3) \cdot C_2(3,3) \\ \hline
C_1(3,1) \cdot C_2(1,1) & C_1(3,2) \cdot C_2(1,2) & C_1(3,3) \cdot C_2(1,3) \\
C_1(3,1) \cdot C_2(2,1) & C_1(3,2) \cdot C_2(2,2) & C_1(3,3) \cdot C_2(2,3) \\
C_1(3,1) \cdot C_2(3,1) & C_1(3,2) \cdot C_2(3,2) & C_1(3,3) \cdot C_2(3,3)
\end{array} \right).$$

For the following, it is advisable to number the k^2 answer categories by $1, ..., k^2$ where we first sort by answer A_1 and then by A_2. For example, for $k = 3$, the numbering scheme is given by Table 2.

Table 2. Answer categories $1, ..., 9$ for $A_1 \in \{1, 2, 3\}$ and $A_2 \in \{1, 2, 3\}$

answer category	1	2	3	4	5	6	7	8	9
answer A_1	1	1	1	2	2	2	3	3	3
answer A_2	1	2	3	1	2	3	1	2	3

Let n_l $(l = 1, ..., k^2)$ be the observed absolute frequency of answer category l in the sample. Furthermore, let $\tilde{C}$ be the $k^2 \times (k - 1)$ matrix that arises as follows from C: The jth column of $\tilde{C}$ $(j = 1, ..., k - 1)$ is given by the difference $C(:, j) - C(:, k)$. The log-likelihood function is

$$l(\pi_1, ..., \pi_{k-1}) = (n_1, ..., n_{k^2}) \cdot \log \left[\tilde{C} \cdot (\pi_1, ..., \pi_{k-1})^\top + C(:, k) \right]$$

with componentwise application of the logarithm. The score function corresponding to l is

$$s(\pi_1,...,\pi_{k-1}) = [l'(\pi_1,...,\pi_{k-1})]^\top = \tilde{C}^\top \cdot \left(\frac{n_1}{\lambda_1},...,\frac{n_{k^2}}{\lambda_{k^2}}\right)^\top.$$

For the second derivative of l, we obtain

$$l''(\pi_1,...,\pi_{k-1}) = \quad - \quad \tilde{C}^\top \cdot diag\left(\frac{n_1}{\lambda_1^2},...,\frac{n_{k^2}}{\lambda_{k^2}^2}\right)\cdot \tilde{C}.$$

Consequently, the Fisher matrix is

$$F = F(\pi_1,...,\pi_{k-1}) = \tilde{C}^\top \cdot n \cdot diag\left(\frac{1}{\lambda_1},...,\frac{1}{\lambda_{k^2}}\right)\cdot \tilde{C}.$$

We maximize the log-likelihood by a Fisher scoring algorithm. However, other algorithms such as EM algorithm, Newton algorithm or Nelder/Mead simplex algorithm are also possible. Our Fisher scoring algorithm generates a sequence $\pi^{(t)} = (\pi_1^{(t)},...,\pi_{k-1}^{(t)})^\top$, $t = 1,2,...$, via the rule

$$\pi^{(t+1)} = \pi^{(t)} + \left[F(\pi_1^{(t)},...,\pi_{k-1}^{(t)})\right]^{-1}\cdot s(\pi_1^{(t)},...,\pi_{k-1}^{(t)})$$

until convergence. We denote the ML estimator for $\pi = (\pi_1,...,\pi_{k-1})^\top$ by $\hat{\pi} = (\hat{\pi}_1,...,\hat{\pi}_{k-1})^\top$. The asymptotic variance of this ML estimator is given by $[F(\pi_1,...,\pi_{k-1})]^{-1}$. An estimator for the asymptotic variance is $[F(\hat{\pi}_1,...,\hat{\pi}_{k-1})]^{-1}$.

5. Precision Improvement

We quantify the estimation inaccuracy by the trace of the asymptotic variance matrix of the ML estimator for $\pi = (\pi_1,...,\pi_{k-1})^\top$. For this variance matrix, we refer to the end of the previous section. We start this section with a formal proof that the estimation inaccuracy of a two-trial NRR method is always less than or equal to the estimation inaccuracy of the single-trial process.

Let A_{ij} be the jth indirect answer of respondent i ($i = 1,...,n$, $j = 1,2$). We set $f_{A_{11}}(a_{11}) = \mathbb{P}(A_{11} = a_{11})$, $f_{A_{11},A_{12}}(a_{11},a_{12}) = \mathbb{P}(A_{11} = a_{11}, A_{12} = a_{12})$, as well as $f_{A_{12}|A_{11}}(a_{12}|a_{11}) = \mathbb{P}(A_{12} = a_{12}|A_{11} = a_{11})$. The Fisher matrix can be written as

$$F = F(\pi_1,...\pi_{k-1}) = n\cdot\mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{11},A_{12}}(A_{11},A_{12})\right)^\top\cdot\frac{d}{d\pi}\log f_{A_{11},A_{12}}(A_{11},A_{12})\right].$$

We have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{11},A_{12}}(A_{11},A_{12})\right)^\top\cdot\frac{d}{d\pi}\log f_{A_{11},A_{12}}(A_{11},A_{12})\right]\\
&= \mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11}) + \frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right)^\top\right.\\
&\qquad \left.\times\left(\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11}) + \frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right)\right]\\
&= \mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right)^\top\cdot\frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right]\\
&\quad + \mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right)^\top\cdot\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right]\\
&\quad + \mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right)^\top\cdot\frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right]\\
&\quad + \mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right)^\top\cdot\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right].
\end{aligned}$$

In the following, we show that the last two summands are zero (zero matrix). We introduce the function g with

$$\begin{aligned}
g(a_{11},a_{12}) &= \left(\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(a_{12}|a_{11})\right)^\top\cdot\frac{d}{d\pi}\log f_{A_{11}}(a_{11})\\
&= \frac{1}{f_{A_{12}|A_{11}}(a_{12}|a_{11})}\cdot\left(\frac{d}{d\pi}f_{A_{12}|A_{11}}(a_{12}|a_{11})\right)^\top\cdot\frac{d}{d\pi}\log f_{A_{11}}(a_{11}).
\end{aligned}$$

It is true that

$$\begin{aligned}
&\mathbb{E}\left[g(A_{11}, A_{12})|A_{11} = a_{11}\right] \\
&= \mathbb{E}\left[\frac{1}{f_{A_{12}|A_{11}}(A_{12}|a_{11})} \cdot \left(\frac{d}{d\pi} f_{A_{12}|A_{11}}(A_{12}|a_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{11}}(a_{11}) \quad | \quad A_{11} = a_{11}\right] \\
&= \sum_{a_{12}\in\{1,..,k\}} \frac{1}{f_{A_{12}|A_{11}}(a_{12}|a_{11})} \cdot \left(\frac{d}{d\pi} f_{A_{12}|A_{11}}(a_{12}|a_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{11}}(a_{11}) \cdot f_{A_{12}|A_{11}}(a_{12}|a_{11}) \\
&= \sum_{a_{12}\in\{1,..,k\}} \left(\frac{d}{d\pi} f_{A_{12}|A_{11}}(a_{12}|a_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{11}}(a_{11}) \\
&= \left(\frac{d}{d\pi} \sum_{a_{12}\in\{1,..,k\}} f_{A_{12}|A_{11}}(a_{12}|a_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{11}}(a_{11}) \\
&= (0, ..., 0)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{11}}(a_{11}) = 0.
\end{aligned}$$

Consequently, $\mathbb{E}(\mathbb{E}(g(A_{11}, A_{12})|A_{11})) = \mathbb{E}(g(A_{11}, A_{12})) = 0$ holds. That is, the third summand is zero. Regarding the fourth summand, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right] \\
&\qquad = \left\{\mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right]\right\}^{\top} = 0.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
F &= n \cdot \mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{11}}(A_{11})\right] \\
&\quad + n \cdot \mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right] \\
&=: G + n \cdot \mathbb{E}\left[\left(\frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right)^{\top} \cdot \frac{d}{d\pi}\log f_{A_{12}|A_{11}}(A_{12}|A_{11})\right]. \qquad (1)
\end{aligned}$$

The matrix $G = G(\pi_1, ...\pi_{k-1})$ is the Fisher matrix if we only have observations on $A_{11}, ..., A_{n1}$, that is, if we only require one indirect answer per respondent. It follows from (1) that $F - G$ is positive-semidefinite. By a known property of the Löwner order (Nordström, 1989, p. 4473), we obtain that $G^{-1} - F^{-1}$ is positive-semidefinite. Thus, the trace of G^{-1} is larger than or equal to the trace of F^{-1}. G^{-1} is the asymptotic variance matrix of the ML estimator for π for one indirect answer per interviewee and F^{-1} is the asymptotic variance matrix of the ML estimator for two indirect answers per interviewee. Hence, we have shown that the estimation inaccuracy of a two-trial NRR method is always less than or equal to the estimation inaccuracy of the single-trial process.

For numerical illustration, we now compute the estimation inaccuracy of our two-trial NRR techniques for concrete parameter specifications and make comparisons to the estimation inaccuracy of the single-trial versions. For the crosswise method, we set $\pi_1 = 0.8$ and consider

$$c_{11} \in \{0.1, 0.2, ..., 0.9, 1\} \text{ and } c_{21} \in \{0.1, 0.2, ..., 0.9, 1\}. \qquad (2)$$

The quantity n times the asymptotic variance of the ML estimator for π_1 for the two-trial crosswise method is presented for any combination of c_{11} and c_{21} in the middle of Table 3. In the right column of Table 3, we provide the quantity n times the asymptotic variance of the ML estimator for π_1 for the single-trial crosswise method depending on the parameter c_{11}. Here, the asymptotic variance for the single-trial version is

$$\left(\tilde{C_1}^{\top} \cdot n \cdot diag\left(1./(C_1 \cdot (\pi_1, \pi_2)^{\top})\right) \cdot \tilde{C_1}\right)^{-1} \text{ with } \tilde{C_1} = C_1(:, 1) - C_1(:, 2)$$

and ./ symbolizing componentwise division. For the triangular method, we again consider $\pi_1 = 0.8$ and proceed analogously to the crosswise method. The computational results for the triangular method are given in Table 4. For the Tang et al. (2009) technique, we consider $k = 3$ categories, $(\pi_1, \pi_2, \pi_3) = (0.6, 0.3, 0.1)$, and 10 distributions of an auxiliary variable as follows:

$$\begin{aligned}
c^{(1)}_{Tang} &= \begin{pmatrix}0.3333 & 0.3333 & 0.3333\end{pmatrix}, & c^{(2)}_{Tang} &= \begin{pmatrix}0.4372 & 0.5123 & 0.0504\end{pmatrix},\\
c^{(3)}_{Tang} &= \begin{pmatrix}0.4153 & 0.5167 & 0.0680\end{pmatrix}, & c^{(4)}_{Tang} &= \begin{pmatrix}0.1353 & 0.7807 & 0.0841\end{pmatrix},\\
c^{(5)}_{Tang} &= \begin{pmatrix}0.0901 & 0.4471 & 0.4628\end{pmatrix}, & c^{(6)}_{Tang} &= \begin{pmatrix}0.7780 & 0.1278 & 0.0942\end{pmatrix},\\
c^{(7)}_{Tang} &= \begin{pmatrix}0.4384 & 0.5243 & 0.0373\end{pmatrix}, & c^{(8)}_{Tang} &= \begin{pmatrix}0.2212 & 0.4280 & 0.3509\end{pmatrix},\\
c^{(9)}_{Tang} &= \begin{pmatrix}0.0617 & 0.5125 & 0.4257\end{pmatrix}, & c^{(10)}_{Tang} &= \begin{pmatrix}0.3799 & 0.4751 & 0.1450\end{pmatrix}.
\end{aligned}$$

The first distribution is a uniform distribution. The other vectors $c^{(2)}_{Tang}, ..., c^{(10)}_{Tang}$ were drawn randomly. The middle of Table 5 shows n times the trace of the asymptotic variance matrix of the ML estimator for (π_1, π_2) for the Tang et al. (2009) method with two indirect answers per respondent for each combination $c^{(i)}_{Tang}$ and $c^{(j)}_{Tang}$. The right column of this table provides the quantity n times the trace of the asymptotic variance of the ML estimator for (π_1, π_2) for the single-trial method. The asymptotic variance for the single-trial Tang et al. (2009) method is

$$\left(\tilde{C_1}^{\top} \cdot n \cdot diag\left(1./(C_1 \cdot (\pi_1, \pi_2, \pi_3)^{\top})\right) \cdot \tilde{C_1}\right)^{-1}$$

with

$$\tilde{C_1} = [C_1(:,1) - C_1(:,3),\ C_1(:,2) - C_1(:,3)].$$

We finally come to the diagonal technique. Say, we have $k = 4$ categories, the vector $(\pi_1, ..., \pi_4) = (0.4,\ 0.3,\ 0.2,\ 0.1)$, and the 10 distributions of an auxiliary characteristic

$$\begin{aligned}
c^{(1)}_{DT} &= \begin{pmatrix}0.3250 & 0.2250 & 0.2250 & 0.2250\end{pmatrix}, & c^{(2)}_{DT} &= \begin{pmatrix}0.4000 & 0.2000 & 0.2000 & 0.2000\end{pmatrix},\\
c^{(3)}_{DT} &= \begin{pmatrix}0.4750 & 0.1750 & 0.1750 & 0.1750\end{pmatrix}, & c^{(4)}_{DT} &= \begin{pmatrix}0.5500 & 0.1500 & 0.1500 & 0.1500\end{pmatrix},\\
c^{(5)}_{DT} &= \begin{pmatrix}0.6250 & 0.1250 & 0.1250 & 0.1250\end{pmatrix}, & c^{(6)}_{DT} &= \begin{pmatrix}0.7000 & 0.1000 & 0.1000 & 0.1000\end{pmatrix},\\
c^{(7)}_{DT} &= \begin{pmatrix}0.7750 & 0.0750 & 0.0750 & 0.0750\end{pmatrix}, & c^{(8)}_{DT} &= \begin{pmatrix}0.8500 & 0.0500 & 0.0500 & 0.0500\end{pmatrix},\\
c^{(9)}_{DT} &= \begin{pmatrix}0.9250 & 0.0250 & 0.0250 & 0.0250\end{pmatrix}, & c^{(10)}_{DT} &= \begin{pmatrix}1.0000 & 0.0000 & 0.0000 & 0.0000\end{pmatrix}.
\end{aligned}$$

The distributions were chosen according to Groenitz (2014, p. 219) where we considered $\sigma \in \{1/20, 2/20, ..., 9/20, 10/20\}$ in Groenitz (2014, p. 219). For the two-trial and single-trial diagonal technique, the results concerning estimation inaccuracy are given in Table 6. For the right column of this table, we remark that the asymptotic variance matrix of the ML estimator for (π_1, π_2, π_3) in the single-trial diagonal technique is equal to

$$\left(\tilde{C_1}^{\top} \cdot n \cdot diag\left(1./(C_1 \cdot (\pi_1, ..., \pi_4)^{\top})\right) \cdot \tilde{C_1}\right)^{-1}$$

where

$$\tilde{C_1} = [C_1(:,1) - C_1(:,4),\ C_1(:,2) - C_1(:,4),\ C_1(:,3) - C_1(:,4)].$$

Altogether, the Tables 3-6 demonstrate that large efficiency gains are possible by two-trial NRR methods in comparison with single-trial NRR schemes.

Table 3. Inaccuracy crosswise method

	inaccuracy two-trial crosswise method										single trial
c_{11}/c_{21}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	-
0.1	0.21	0.25	0.27	0.29	0.30	0.29	0.27	0.25	0.21	0.16	0.30
0.2	0.25	0.34	0.45	0.56	0.60	0.56	0.45	0.34	0.25	0.16	0.60
0.3	0.27	0.45	0.75	1.19	1.47	1.19	0.75	0.45	0.27	0.16	1.47
0.4	0.29	0.56	1.19	3.08	6.16	3.08	1.19	0.56	0.29	0.16	6.16
0.5	0.30	0.60	1.47	6.16	—	6.16	1.47	0.60	0.30	0.16	—
0.6	0.29	0.56	1.19	3.08	6.16	3.08	1.19	0.56	0.29	0.16	6.16
0.7	0.27	0.45	0.75	1.19	1.47	1.19	0.75	0.45	0.27	0.16	1.47
0.8	0.25	0.34	0.45	0.56	0.60	0.56	0.45	0.34	0.25	0.16	0.60
0.9	0.21	0.25	0.27	0.29	0.30	0.29	0.27	0.25	0.21	0.16	0.30
1.0	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16

Note. This table shows the quantity n times the asymptotic variance of the ML estimator for π_1 for the crosswise method. For $c_{11} = 0.5$ in the single-trial procedure and $c_{11} = c_{21} = 0.5$ in the two-trial procedure, the log-likelihood does not depend on π implying that ML estimation is not adequate in these cases.

Table 4. Inaccuracy triangular method

	inaccuracy two-trial triangular method										single trial
c_{11}/c_{21}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	-
0.1	3.57	2.22	1.52	1.10	0.81	0.61	0.46	0.34	0.24	0.16	7.36
0.2	2.22	1.58	1.18	0.90	0.69	0.54	0.41	0.31	0.23	0.16	3.36
0.3	1.52	1.18	0.93	0.74	0.59	0.47	0.37	0.29	0.22	0.16	2.03
0.4	1.10	0.90	0.74	0.61	0.50	0.41	0.34	0.27	0.21	0.16	1.36
0.5	0.81	0.69	0.59	0.50	0.43	0.36	0.30	0.25	0.20	0.16	0.96
0.6	0.61	0.54	0.47	0.41	0.36	0.31	0.27	0.23	0.19	0.16	0.69
0.7	0.46	0.41	0.37	0.34	0.30	0.27	0.24	0.21	0.18	0.16	0.50
0.8	0.34	0.31	0.29	0.27	0.25	0.23	0.21	0.19	0.18	0.16	0.36
0.9	0.24	0.23	0.22	0.21	0.20	0.19	0.18	0.18	0.17	0.16	0.25
1.0	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16

Note. This table provides the quantity n times the asymptotic variance of the ML estimator for π_1 for the triangular method.

Table 5. Inaccuracy for Tang et al. (2009) design

	inaccuracy two-trial Tang et al. (2009) design										single trial
	$c^{(1)}_{Tang}$	$c^{(2)}_{Tang}$	$c^{(3)}_{Tang}$	$c^{(4)}_{Tang}$	$c^{(5)}_{Tang}$	$c^{(6)}_{Tang}$	$c^{(7)}_{Tang}$	$c^{(8)}_{Tang}$	$c^{(9)}_{Tang}$	$c^{(10)}_{Tang}$	-
$c^{(1)}_{Tang}$	0.70	0.71	0.72	0.91	0.82	0.52	0.72	0.76	0.85	0.72	2.05
$c^{(2)}_{Tang}$	0.71	0.88	0.89	1.26	0.83	0.54	0.89	0.81	0.90	0.84	1.89
$c^{(3)}_{Tang}$	0.72	0.89	0.89	1.27	0.85	0.54	0.90	0.82	0.92	0.85	1.99
$c^{(4)}_{Tang}$	0.91	1.26	1.27	2.36	1.16	0.59	1.29	1.10	1.32	1.18	7.46
$c^{(5)}_{Tang}$	0.82	0.83	0.85	1.16	1.03	0.55	0.84	0.93	1.09	0.85	8.11
$c^{(6)}_{Tang}$	0.52	0.54	0.54	0.59	0.55	0.48	0.54	0.54	0.56	0.54	0.71
$c^{(7)}_{Tang}$	0.72	0.89	0.90	1.29	0.84	0.54	0.91	0.81	0.91	0.85	1.91
$c^{(8)}_{Tang}$	0.76	0.81	0.82	1.10	0.93	0.54	0.81	0.85	0.98	0.81	3.32
$c^{(9)}_{Tang}$	0.85	0.90	0.92	1.32	1.09	0.56	0.91	0.98	1.16	0.91	12.44
$c^{(10)}_{Tang}$	0.72	0.84	0.85	1.18	0.85	0.54	0.85	0.81	0.91	0.82	2.07

Note. This table shows n times the trace of the asymptotic variance matrix of the ML estimator for (π_1, π_2) for the Tang et al. (2009) method.

Table 6. Inaccuracy for diagonal technique

	inaccuracy two-trial diagonal technique										single trial
	$c^{(1)}_{DT}$	$c^{(2)}_{DT}$	$c^{(3)}_{DT}$	$c^{(4)}_{DT}$	$c^{(5)}_{DT}$	$c^{(6)}_{DT}$	$c^{(7)}_{DT}$	$c^{(8)}_{DT}$	$c^{(9)}_{DT}$	$c^{(10)}_{DT}$	-
$c^{(1)}_{DT}$	28.60	11.60	5.88	3.50	2.31	1.64	1.22	0.94	0.75	0.61	56.97
$c^{(2)}_{DT}$	11.60	7.38	4.64	3.07	2.14	1.57	1.19	0.93	0.75	0.61	14.41
$c^{(3)}_{DT}$	5.88	4.64	3.45	2.56	1.92	1.48	1.16	0.92	0.75	0.61	6.47
$c^{(4)}_{DT}$	3.50	3.07	2.56	2.08	1.68	1.36	1.11	0.90	0.74	0.61	3.68
$c^{(5)}_{DT}$	2.31	2.14	1.92	1.68	1.45	1.24	1.05	0.88	0.74	0.61	2.37
$c^{(6)}_{DT}$	1.64	1.57	1.48	1.36	1.24	1.11	0.98	0.85	0.73	0.61	1.66
$c^{(7)}_{DT}$	1.22	1.19	1.16	1.11	1.05	0.98	0.90	0.81	0.72	0.61	1.23
$c^{(8)}_{DT}$	0.94	0.93	0.92	0.90	0.88	0.85	0.81	0.77	0.70	0.61	0.95
$c^{(9)}_{DT}$	0.75	0.75	0.75	0.74	0.74	0.73	0.72	0.70	0.67	0.61	0.75
$c^{(10)}_{DT}$	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61

Note. This table presents n times the trace of the asymptotic variance matrix of the ML estimator for (π_1, π_2, π_3) for the diagonal method by Groenitz (2014).

6. Summary

NRR designs for sensitive attributes have attracted much attention in the literature of the last years. In this article, we have considered two-trial versions of four NRR schemes. In a two-trial design, each person in the sample must provide two indirect answers. Each answer depends on a separate auxiliary characteristic. We have developed the maximum likelihood inference for the distribution of the sensitive variable and derived the asymptotic estimation variance. Moreover, we analyzed the gains in estimation precision by two indirect answers per respondent instead of one indirect answer.

Acknowledgements

The author thanks the Editor and two anonymous reviewers for their comments and suggestions on the manuscript.

References

Alavi, S. M. R., & Tajodini, M. (2016). Maximum Likelihood Estimation of Sensitive Proportion Using Repeated Randomized Response Techniques. *Journal of Applied Statistics, 43*, 563-571. https://doi.org/10.1080/02664763.2015.1070811

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys.* Chapman & Hall/CRC. https://doi.org/10.1201/b10476

Chaudhuri, A., & Christofides, T. C. (2013). *Indirect Questioning in Sample Surveys.* Springer. https://doi.org/10.1007/978-3-642-36276-7

Chaudhuri, A., Christofides, T. C., & Rao, C. R. (2016). *Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits.* Handbook of Statistics 34, North Holland. https://doi.org/10.1016/s0169-7161(16)x0002-8

Eriksson, S. A. (1973). A New Model for Randomized Response. *International Statistical Review, 41*, 101-113. https://doi.org/10.2307/1402791

Fox, J. A., & Tracy, P.E. (1986). *Randomized Response - A Method for Sensitive Surveys.* Sage. https://doi.org/10.4135/9781412985581

Groenitz, H. (2014). A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables. *Metrika, 77*, 211-224. https://doi.org/10.1007/s00184-012-0406-8

Groenitz, H. (2016). Valid Estimates for Repeated Randomized Response Methods. *Journal of Applied Statistics, in press*. https://doi.org/10.1080/02664763.2016.1267119

Nordström, K. (1989). Some Further Aspects of the Löwner-Ordering Antitonicity of the Moore-Penrose Inverse. *Communications in Statistics - Theory and Methods, 18*, 4471-4489. https://doi.org/10.1080/03610928908830167

Tan, M. T., Tian, G. L., & Tang, M. L. (2009). Sample Surveys with Sensitive Questions: A Nonrandomized Response Approach. *The American Statistician, 63*, 9-16. https://doi.org/10.1198/tast.2009.0002

Tang, M. L., Tian, G. L., Tang, N. S., & Liu, Z. (2009). A New Non-randomized Multi-category Response Model for Surveys With a Single Sensitive Question: Design and Analysis. *Journal of the Korean Statistical Society, 38*, 339-349. https://doi.org/10.1016/j.jkss.2008.12.004

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60*, 63-69. https://doi.org/10.2307/2283137

Yu, J. W., Tian, G. L., & Tang, M. L. (2008). Two New Models for Survey Sampling With Sensitive Characteristic: Design and Analysis. *Metrika, 67*, 251-263. https://doi.org/10.1007/s00184-007-0131-x

Copyrights