Department for Applied Statistics
Johannes Kepler University Linz


**IFAS Research Paper Series**
**2007-28**

# A standardized technique of randomized response

Andreas Quatember

September 2007

# 1 Introduction

Nonresponse in sample surveys may cause a biased estimation of unknown population parameters as well as an increase of the variance of this estimation. Let $U$ be the universe of $N$ population units and $U_A$ be a subset of $N_A$ elements, that belong to a class $A$ of a categorial variable under study. Moreover let $U_{A^c}$ be the group of $N_{A^c}$ elements, that do not belong to this class ($U = U_A \cup U_{A^c}$, $U_A \cap U_{A^c} = \emptyset$, $N = N_A + N_{A^c}$). Let

$$x_i = \begin{cases} 1 & \text{if unit } i \in U_A, \\ 0 & \text{otherwise.} \end{cases}$$

The parameter of interest is the relative size $\pi_A$ of the subpopulation $U_A$:

$$\pi_A = \frac{\sum_U x_k}{N} = \frac{N_A}{N} \tag{1}$$

($\sum_U x_k$ being the abbreviated notation for $\sum_{k \in U} x_k$). For a sample simple random sample $s$ of $n$ elements drawn with or without replacement the estimator of $\pi_A$ for the direct questioning ($dir$) on the subject is

$$\widehat{\pi}_A^{dir} = \frac{\sum_s x_k}{n}. \tag{2}$$

This estimator is unbiased, if fullresponse does occur. In the presence of nonresponse the sample $s$ will be divided into a "response set" $r$ ($r \subseteq s$) of size $n_r$ and a "missing set" $m$ ($m \subseteq s$) of size $n_m$ with $s = r \cup m$, $r \cap m = \emptyset$ and $n = n_r + n_m$. With these sets (2) must be rewritten as:

$$\widehat{\pi}_A^{dir} = \frac{1}{n} \cdot (\sum\nolimits_r x_k + \sum\nolimits_m x_k). \tag{3}$$

If $m$ is nonempty, the second summand in the bracket on the right hand side of (3) cannot not observed. Obviously there are two ways to cope with this problem. The first one is to estimate $\pi_A$ by using the observations of subset $r$ only. For this purpose it is necessary to raise the weights of the elements of the response set. This method is called *weighting adjustment* (see for example: Little and Rubin (2002), p.44ff). The second one is to estimate the second summand of (3) by finding substitutes for the $x_k$'s of the missing set. This procedure is called *imputation* (see also: Little and Rubin (2002), p.59ff).

But before such methods should be used on the data to compensate for nonresponse, everything should be done beforehand to keep the nonresponse rate as low as possible. If the reason for not cooperating in an interview is the refusal to answer directly on a question on a highly personal, embarrassing matter (like drug addiction, diseases, sexual behaviour, tax evasion, alcoholism or involvement in crimes) to avoid to reveal the requested information to the interviewer, *randomized response strategies* can be used.

One of the characteristics, that are common to all of these methods, is that instead of the direct questioning on the sensitive subject a questioning design is used, which does not enable the interviewer to identify the randomly selected question (or instruction) on which the respondent has given the answer. The idea is to reduce the individuals' fear of an embarrassing "outing" in front of a stranger. For this purpose the respondent has completely to understand how the design protects his or her privacy to make sure that the interviewee is willing to cooperate. Anotherone is that the construction of the questioning design does still allow the estimation of the parameter under study.

The pioneering work in this field was published by Warner (1965). In his questioning design ($W$) each respondent has to answer randomly either with probability $p_1^W$ the question "Are you a member of group $U_A$?" or with probability $p_2^W = 1 - p_1^W$ the alternative question "Are you a member of group $U_{A^c}$?" ($0 \leq p_1^W \leq 1$).

We call $p_1^W$ the *design parameter* of Warner's technique. Let

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ answers "yes",} \\ 0 & \text{otherwise.} \end{cases}$$

Then the probability $\pi_y^W \equiv P(y_i = 1)$ of a "yes"-answer wthin the questioning design $W$ is given by

$$\pi_y^W = p_1^W \cdot \pi_A + (1 - p_1^W) \cdot (1 - \pi_A). \tag{4}$$

Assuming that the randomized responding will guarantee the cooperation of all selected sample units Warner derived from (4) the following unbiased estimator of $\pi_A$:

$$\hat{\pi}_A^W = \frac{\hat{\pi}_y + p_1^W - 1}{2p_1^W - 1} \tag{5}$$

($p_1 \neq 0,5$) with $\hat{\pi}_y$, the proportion of "yes"-answers in the sample (Warner (1965), p.65).

For simple random sampling without replacement ($wor$) the variance of estimator (5) is given by

$$V_{wor}(\hat{\pi}_A^W) = \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{N - n}{N - 1} + \frac{p_1^W \cdot (1 - p_1^W)}{n \cdot (2p_1^W - 1)^2} \tag{6}$$

(Kim and Flueck (1978), p.347). For simple random sampling with replacement ($wr$) or for large populations (6) reduces to

$$V_{wr}(\hat{\pi}_A^W) = \frac{\pi_A \cdot (1 - \pi_A)}{n} + \frac{p_1^W \cdot (1 - p_1^W)}{n \cdot (2p_1^W - 1)^2} \tag{7}$$

(Warner (1965), p.65). It is the right one of the two summands in both (6) and (7) that corresponds to the increase of the variance caused by the use of Warner's randomized response technique instead of the direct questioning on the subject. A two-stage version of Warner's questioning design was presented by Mangat and Singh (1990).

Various randomized response techniques have been proposed since then. All of them make use of randomly selected questions or answers though some of them use different random devices depending on the (for the interviewer unknown) respondent's possession or nonpossession of a certain attribute (see as an example: Franklin (1989)).

# 2 The standardized randomized response technique

A standardization of randomized response strategies can be formulated in the following way: Each respondent has either to answer randomly

- with probability $p_1$ the question "Are you a member of group $U_A$?",
- with probability $p_2$ the question "Are you a member of group $U_{A^c}$?" or
- with probability $p_3$ the question "Are you a member of group $U_B$?"

or he or she has just

- to answer "yes" with probability $p_4$ or
- to answer "no" with probability $p_5$

($\sum_{i=1}^{5} p_i = 1$, $0 \leq p_i \leq 1$ for $i = 1, 2, ..., 5$). Elements of group $U_B$ should be characterized by the possession of a completely innocuous attribute $B$ (for instance the season of birth B), that should not be related to the possession of attribute $A$ (see: Horvitz et al. (1967)).

$p_1$, $p_2$, ..., $p_5$ are the design parameters of our standardized randomized response technique. Because for this design the probability $\pi_y$ of a "yes"-answer is

$$\pi_y = p_1 \cdot \pi_A + p_2 \cdot (1 - \pi_A) + p_3 \cdot \pi_B + p_4, \tag{8}$$

the unbiased estimator of $\pi_A$ is given by

$$\hat{\pi}_A = \frac{\hat{\pi}_y - p_2 - p_3 \cdot \pi_B - p_4}{p_1 - p_2} \tag{9}$$

($p_1 \neq p_2$) with $\pi_B = \frac{N_B}{N}$, the relative size of group $U_B$ in the population.

**Theorem 1:** For simple random sampling without replacement ($wor$) the variance of the standardized estimator $\hat{\pi}_A$ (9) is given by

$$V_{wor}(\hat{\pi}_A) = \frac{\pi_y \cdot (1 - \pi_y)}{n \cdot (p_1 - p_2)^2} - \frac{\pi_A \cdot (1 - \pi_A)}{n} \cdot \frac{n - 1}{N - 1}. \tag{10}$$

**Theorem 2:** For simple random sampling with replacement ($wr$) the variance of the standardized estimator $\hat{\pi}_A^{ST}$ (9) is given by

$$V_{wr}(\hat{\pi}_A) = \frac{\pi_y \cdot (1 - \pi_y)}{n \cdot (p_1 - p_2)^2}. \tag{11}$$

For the proofs of (10) and (11) see the Appendix.

To be able to calculate $\pi_A$ at all, certainly $p_1$ or at least $p_2$, a possibility on which for reasons of readability we do not point to in the following anymore, has to be larger than zero and all sample units have to cooperate. There is a total of 16 combinations of $p_1 > 0$ with other design parameters being larger than zero. These can be described as special cases of our standardized response strategy (see: Table 1). For example choosing $p_1 = 1$ gives the direct questioning on the subject. If we let $p_1 > 0$, $p_2 = 1 - p_1$ and $p_3 = p_4 = p_5 = 0$ the standardized questioning design leads to Warner's technique. With $p_1 > 0$, $p_3 = 1 - p_1$ and $p_2 = p_4 = p_5 = 0$ we get Greenberg et al.'s technique with known $\pi_B$ (The reader is referred to other questioning designs previously published as far as the author knows it, in the "References-column" of Table 1).

These considerations lead us directly to the question how to choose the design parameters of the standardized response technique to find out the strategy which performs best. This research will be done at next.

| No. | Questions/Instructions | | | | References |
|---|---|---|---|---|---|
| | $U_{A^c}$ | $U_B$ | "yes" | "no" | |
| ST1 | | | | | direct questioning |
| ST2 | • | | | | Warner (1965), Mangat and Singh (1990) |
| ST3 | | • | | | Greenberg et al. (1969), Mangat (1992) |
| ST4 | | | • | | |
| ST5 | | | | • | |
| ST6 | • | • | | | |
| ST7 | • | | • | | |
| ST8 | • | | | • | Quatember (2007), Mangat et al. (1993) |
| ST9 | | • | • | | |
| ST10 | | • | | • | Singh et al. (2003), Singh et al. (1994) |
| ST11 | | | • | • | Quatember (2007), Singh et al. (1995) |
| ST12 | • | • | • | | |
| ST13 | • | • | | • | |
| ST14 | • | | • | • | |
| ST15 | | • | • | • | |
| ST16 | • | • | • | • | |

Table 1: All 16 special cases of the standardized randomized response strategy with the question on membership of group $U_A$ being asked mandatorily (a $2^{nd}$ reference mentioned is a two-stage version of the first mentioned one-stage design)

# 3 Appendix

**Proof of (10) and (11):**
The variance of estimator (9) is given by

$$V(\hat{\pi}_A^{st}) = \frac{1}{(p_1 - p_2)^2} \cdot V(\hat{\pi}_y) = \frac{1}{n^2 \cdot (p_1 - p_2)^2} \cdot V(\sum\nolimits_s y_i). \tag{12}$$

For respondent $i$ the variables $x_i$ and $y_i$ are defined as in section 1. Furthermore let

$$a_i = \begin{cases} 1 & \text{if unit } i \text{ is asked the question on membership of } U_A, \\ 0 & \text{otherwise,} \end{cases}$$

$$b_i = \begin{cases} 1 & \text{if unit } i \text{ is asked the question on membership of } U_{A^c}, \\ 0 & \text{otherwise,} \end{cases}$$

$$c_i = \begin{cases} 1 & \text{if unit } i \text{ is asked the question on membership of } U_B, \\ 0 & \text{otherwise,} \end{cases}$$

$$d_i = \begin{cases} 1 & \text{if unit } i \in U_B, \\ 0 & \text{otherwise,} \end{cases}$$

$$e_i = \begin{cases} 1 & \text{if } i \text{ is instructed to say "yes",} \\ 0 & \text{otherwise,} \end{cases}$$

so that $y_i$ can be written as $y_i = a_i \cdot x_i + b_i \cdot (1 - x_i) + c_i \cdot d_i + e_i$. The variance of the number of "yes"-answers in the sample is

$$V(\sum\nolimits_s y_i) = E(\sum\nolimits_s y_i^2) + E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) - E^2(\sum\nolimits_s y_i). \tag{13}$$

For simple random sampling with and without replacement the first summand of (13) results in

$$E(\sum\nolimits_s y_i^2) = E(\sum\nolimits_s y_i) = n \cdot (p_1 \cdot \pi_A + p_2 \cdot (1 - \pi_A) + p_3 \cdot \pi_B + p_4). \tag{14}$$

The second summand of (13) is

$$\begin{aligned} E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot E(&a_i \cdot a_j \cdot x_i \cdot x_j + a_i \cdot x_i \cdot b_j - a_i \cdot x_i \cdot x_j \cdot b_j + \\ &+a_i \cdot x_i \cdot c_j \cdot d_j + a_i \cdot x_i \cdot e_j + b_i \cdot a_j \cdot x_j + b_i \cdot b_j - b_i \cdot b_j \cdot x_j + \\ &+b_i \cdot c_j \cdot d_j + b_i \cdot e_j - b_i \cdot x_i \cdot x_j \cdot a_j - x_i \cdot b_i \cdot b_j + \\ &+b_i \cdot x_i \cdot x_j \cdot b_j - b_i \cdot x_i \cdot c_j \cdot d_j - b_i \cdot x_i \cdot e_j + c_i \cdot d_i \cdot a_j \cdot x_j + \\ &+c_i \cdot d_i \cdot b_j - c_i \cdot d_i \cdot b_j \cdot x_j + c_i \cdot d_i \cdot c_j \cdot d_j + c_i \cdot d_i \cdot e_j + \\ &+e_i \cdot a_j \cdot x_j + e_i \cdot b_j - e_i \cdot b_j \cdot x_j + e_i \cdot c_j \cdot d_j + e_i \cdot e_j). \end{aligned}$$

Because of

$$E(x_i \cdot x_j) = \begin{cases} \pi_A^2 & \text{for simple random sampling with replacement,} \\ \frac{\pi_A \cdot (N\pi_A - 1)}{N-1} & \text{for simple random sampling without replacement} \end{cases} \tag{15}$$

we get

$$\begin{aligned} E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot (&p_1^2 \cdot \pi_A^2 + 2 \cdot p_1 \cdot p_2 \cdot \pi_A - 2 \cdot p_1 \cdot p_2 \cdot \pi_A^2 + \\ &+2 \cdot p_1 \cdot p_3 \cdot \pi_A \cdot \pi_B + 2 \cdot p_1 \cdot p_4 \cdot \pi_A + p_2^2 - 2 \cdot p_2^2 \cdot \pi_A + \\ &+2 \cdot p_2 \cdot p_3 \cdot \pi_B + 2 \cdot p_2 \cdot p_4 + p_2^2 \cdot \pi_A^2 - 2 \cdot p_2 \cdot p_3 \cdot \pi_A \cdot \pi_B - \\ &-2 \cdot p_2 \cdot p_4 \cdot \pi_A + p_3^2 \cdot \pi_B^2 + 2 \cdot p_3 \cdot p_4 \cdot \pi_B + p_4^2) \end{aligned} \tag{16}$$

for the with replacement case and for a without replacement selection of sample units we have

$$\begin{aligned} E(\sum\nolimits_{s(i\neq j)} y_i \cdot y_j) = n \cdot (n-1) \cdot (&p_1^2 \cdot \frac{\pi_A \cdot (N\pi_A - 1)}{N-1} + 2 \cdot p_1 \cdot p_2 \cdot \pi_A - \\ &-2 \cdot p_1 \cdot p_2 \cdot \frac{\pi_A \cdot (N\pi_A - 1)}{N-1} + 2 \cdot p_1 \cdot p_3 \cdot \pi_A \cdot \pi_B + \\ &+2 \cdot p_1 \cdot p_4 \cdot \pi_A + p_2^2 - 2 \cdot p_2^2 \cdot \pi_A + 2 \cdot p_2 \cdot p_3 \cdot \pi_B + 2 \cdot p_2 \cdot p_4 + \\ &+p_2^2 \cdot \frac{\pi_A \cdot (N\pi_A - 1)}{N-1} - 2 \cdot p_2 \cdot p_3 \cdot \pi_A \cdot \pi_B - 2 \cdot p_2 \cdot p_4 \cdot \pi_A + \\ &+p_3^2 \cdot \pi_B^2 + 2 \cdot p_3 \cdot p_4 \cdot \pi_B + p_4^2). \end{aligned} \tag{17}$$

The subtrahend on the right side of (13) is for both sampling methods given by

$$\begin{aligned} E^2(\sum\nolimits_s y_i) = n^2 \cdot (&p_1^2 \cdot \pi_A^2 + 2 \cdot p_1 \cdot p_2 \cdot \pi_A - 2 \cdot p_1 \cdot p_2 \cdot \pi_A^2 + \\ &+2 \cdot p_1 \cdot p_3 \cdot \pi_A \cdot \pi_B + 2 \cdot p_1 \cdot p_4 \cdot \pi_A + p_2^2 - 2 \cdot p_2^2 \cdot \pi_A + \\ &+2 \cdot p_2 \cdot p_3 \cdot \pi_B + 2 \cdot p_2 \cdot p_4 + p_2^2 \cdot \pi_A^2 - 2 \cdot p_2 \cdot p_3 \cdot \pi_A \cdot \pi_B - \\ &-2 \cdot p_2 \cdot p_4 \cdot \pi_A + p_3^2 \cdot \pi_B^2 + 2 \cdot p_3 \cdot p_4 \cdot \pi_B + p_4^2). \end{aligned} \tag{18}$$

For simple random sampling with replacement the variance (12) of $\widehat{\pi}_A^{ST}$ results in (11) by inserting (14), (16) and (18) into (13). If the sample is drawn without replacement using (17) instead of (16) results in (10).

# References


Franklin, L. (1989). Randomized Response Sampling from Dichotomous Populations with Continous Randomization. *Survey Methodology 15*(2), 225–235.

Greenberg, B., A.-L. Abul-Ela, W. Simmons, and D. Horvitz (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association 64*, 520–539.

Horvitz, D., B. Shah, and W. Simmons (1967). The unrelated question randomized response model. *1967 Social Statistics Section Proceedings of the American Statistical Association*, 65–72.

Kim, J.-I. and J. Flueck (1978). Modifications of the Randomized Response Technique for Sampling Without Replacement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 346–350.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data*. Hoboken: Wiley and Sons.

Mangat, N. (1992). Two Stage Randomized Response Sampling Procedure Using Unrelated question. *Journal of the Indian Society of Agricultural Statistics 44*, 82–87.

Mangat, N. and R. Singh (1990). An alternative randomized response procedure. *Biometrika 77*, 439–442.

Mangat, N., S. Singh, and R. Singh (1993). On the use of a modified randomization device in randomized response inquiries. *Metron 51*, 211–216.

Quatember, A. (2007). Comparing the efficiency of randomized response techniques under uniform conditions. *IFAS Research Paper Series 2007-23*, http://www.ifas.jku.at/e2550/e2756/index_ger.html.

Singh, R., S. Singh, N. Mangat, and D. Tracy (1995). An improved two stage randomized response strategy. *Statistical Papers 36*, 265–271.

Singh, S., S. Horn, R. Singh, and N. Mangat (2003). On the use of modified randomization device for estimating the prevalence of a sensitive attribute. *Statistics in transition 6*(4), 515–522.

Singh, S., R. Singh, N. Mangat, and D. Tracy (1994). An alternative device for randomized responses. *Statistica 54*, 233–243.

Warner, S. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association 60*, 63–69.