The Privacy Protecting Aspect of Indirect Questioning Designs

Andreas Quatember[1]

[1] *Johannes Kepler University, Linz, Austria*

Abstract
The motivation behind the application of indirect questioning designs such as randomized response (RR) techniques is the privacy protection aspect. Asking sensitive questions indirectly and not in the common direct mode has a positive effect on the answering behaviour of the respondents. Considering this aspect for RR design, on the one hand, the decision, which of the possible RR variants shall be used for a certain variable, has to take into account the type of sensitivity of the variable. On the other hand, a performance comparison between different RR designs only makes sense when it is done under the same level of privacy protection offered by them. In the paper, this aspect of indirect questioning designs is presented exemplarily for certain RR techniques, stressing the relationship between privacy protection and sampling error. Furthermore, other aspects of these techniques such as the randomization device and the practical realization of the strategy in various stages have also to be considered when talking about the privacy protection as perceived by the respondents.

Keywords: sensitivity of variables, data confidentiality, sampling theory, survey methodology.

1. Introduction

When questions on sensitive subjects, such as alcoholism, doping, illegal employment, tax evasion, harassment at work, domestic violence, and so forth, are asked in statistical surveys by the direct questioning method, the rates of nonresponse and untruthful answering will increase above the usual levels (cf. here and in the following: Quatember 2015a, ch. 6). Let U be a universe of N population units and s be a probability sample of U with sample number n and design weights d_k, defined as usual as the reciprocal of the first-order sample inclusion probabilities ($k = 1,2\ldots,N$). When nonresponse and untruthful answers occur, the well-known Horvitz-Thompson estimator t_{HT} for a total t of a study variable y, for instance, is affected by a

decomposition of the sample set *s* consisting of all sampling units into three non-overlapping sub-sets: one (s_u) consisting of all truthfully answering sampling units, another one (s_u) consisting of the untruthfully answering elements, and a third one (s_m) of the missing sampling units:

$$t_{HT} = \sum_s d_k \cdot y_k = \sum_{s_t} d_k \cdot y_k + \sum_{s_u} d_k \cdot y_k + \sum_{s_m} d_k \cdot y_k \qquad (1)$$

(the sum $\sum_s$ is the abbreviated notation for $\sum_{k \in s}$). Hence, such respondent's behaviour may cause serious problems in the estimation process because the estimators of population parameters based only on a survey's available cases in s_t and s_u might strongly be biased. It is therefore essential to not ignore nonresponse or untruthful answering. Indirect questioning (IQ) designs such as to the randomized response (RR) techniques applied at the survey's design stage aim to address this question in particular for sensitive variables. A common characteristic of these alternative questioning strategies is that, instead of directly asking the sensitive question, the variable actually asked for is randomly selected according to reasonably determined "design probabilities" applying a respective randomization device such as throwing a dice. The idea behind this approach can be described within the field of missing data: All true variable values y_k are set to missing followed by an imputation of values y_k^i ($\forall\ k \in s$). This shall reduce the individual's fear of disclosure and thus ensure respondents' cooperation, while the knowledge of the design probabilities still enables to calculate the estimator of interest.

The pioneering work in this field was presented by Warner (1965) for the estimation of the relative size of a certain subgroup of the population: Each respondent has to answer randomly either with probability p_1 the question "Are you a member of the subgroup *A* of the population *U* ($A \subseteq U$)?", or with the remaining probability $p_2 = 1 - p_1$ the alternative question "Are you a member of the complementary subgroup A^C ($A^C \subseteq U$, $A \cap A^C = \varnothing$)?" ($0 < p_1 < 1$). Since then, various RR methods have been developed, for instance, to increase the efficiency of strategies to estimate proportions, or to apply the idea to categorical and quantitative variables. Chaudhuri (2011), Chaudhuri and Christofides (2013), and Chaudhuri et al (2016) are recent books on the various aspects of these procedures. The practical use as well as their positive

effect on the response and the truthful answering rate is well documented (see, for instance, the meta-analysis in Lensvelt-Mulders et al, 2005). Warner (1971) was also the first to note that the RR strategies can also be used as methods for statistical disclosure control.

2. A Generalized RR Technique

Quatember (2009, 2014, 2015a) standardized different families of RR strategies for binary, categorical, and quantitative variables, respectively, and generalized the respective theory to general without replacement-probability sampling with arbitrary sample inclusion probabilities for the population units. This is important to overcome the limitation of the theory to simple random sampling with replacement because in these fields, in which sensitive questions are asked, sampling schemes with differing sample inclusion probabilities are often used.
The idea behind the Horvitz-Thompson based estimator t_{HT} (1) for a total t of a variable y is the model for the derivation of the theoretical properties of such a questioning design. Following this approach (cf., for instance, Quatember, 2015a, Sect. 3.3), "imputations" y_k^i of the true y_k-values can be calculated. For y_k^i with $E(y_k^i) = y_k$ and assuming cooperation, the estimator

$$t_{RR} = \sum_s d_k \cdot y_k^i \qquad (2)$$

is unbiased for the parameter t.
For example, such a standardization of RR questioning designs for the estimation of proportions can be formulated in the following way: Let A be of size N_A and let variable $y_k = I\{k \in A\}$ indicate, if a survey unit k is a member of A or not. The parameter of interest be

$$\pi_A = \frac{\sum_U y_k}{N} = \frac{N_A}{N}, \qquad (3)$$

the relative size of group A in the population. For the given estimation problem, the RR questioning design has to include the direct question on the sensitive subject of interest: "Do you belong to group A?" The second question of the RR design suggested by Warner (1965) was: "Do you belong to the complementary group A^C?" A third possible question used in RR designs is the question on membership of a group $B \subseteq U$ of known size π_B (cf. Horvitz et al., 1967). The N_B members of B are characterized by the possession of a completely innocuous

attribute (for instance, this attribute could be 'having birthday in the first nine months of a year') not related to the membership of *A*. As further possible alternatives to the direct question, the instructions just to say "yes", or "no", respectively can be part of the prcedure (cf. Fidler and Kleinknecht, 1977).

The first question is assigned a "design probability" of p_1 of being chosen, the second has one of p_2, the third has a probability of p_3. The instructions just to say "yes" or "no" are given probabilities of p_4 and p_5 ($p_1 > 0$, $0 \leq p_i < 1$ for i = 2,3,4,5, and $\sum_{i=1}^{5} p_i = 1$). This is clearly not an invitation to use all five questions/instructions in the same questioning design. Rather, the objective is to provide a unified theoretical framework that can be applied with all possible combinations of these questions. Various already existing RR procedures such as Warner's technique described above ($0 < p_i < 1$ for i = 1, 2, and $p_1 + p_2 = 1$) are included in this "standardized" design (cf. Quatember, 2009, 2012). One can choose these probabilities according to own preferences.

Let the answer z_k of a sample unit k in the RR questioning design be

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

The term

$$y_k^{\,i} = \frac{z_k - u}{v} \qquad (4)$$

with $u \equiv p_2 + p_3 \cdot \pi_B + p_4$ and $v \equiv p_1 - p_2$ yields $E(y_k^i) = y_k$.

As usual, for this linear estimator, to be able to calculate approximate confidence intervals and conduct statistical hypothesis tests, the estimation of its design-based variance is based on an exact expression $V(\hat{\pi}_A)$ of its theoretical variance and the estimation of this variance by an unbiased design-based estimator $\hat{V}(\hat{\pi}_A)$ (cf. Quatember, 2012, p.478). In this case, $V(\hat{\pi}_A)$ is given by

$$V(\hat{\pi}_A) = \frac{1}{N^2} \cdot \sum \sum_U \Delta_{kl} \cdot y_k \cdot d_k \cdot y_l \cdot d_l + \frac{1}{N^2} \cdot \sum_U \left(\frac{u(1-u)}{v^2} + \frac{1-2u-v}{v} \cdot y_k \right) \cdot d_k \qquad (5)$$

Therein, Δ_{kl} denotes the covariance of the sample inclusion indicators). According to Eq. (5), the expression

$$C \equiv \frac{1}{N^2} \cdot \sum_U \left(\frac{u(1-u)}{v^2} + \frac{1-2u-v}{v} \cdot y_k \right) \cdot d_k \qquad (6)$$

can be considered as the costs C to be paid for the higher privacy protection in terms of accuracy.

3. Objectively calculated Privacy Protection vs. Accuracy

To summarize Sect. 1, the aim of an RR questioning design is the protection of the privacy of respondents in such a way that they are willing to respond truthfully even on questions of a rather sensitive matter. Therefore, a measure of privacy protection shall be considered to describe the dependence of the efficiency of the estimation process on the level of privacy protection. For the family of RR techniques described in Sect. 2 in some detail, such a privacy protection measure with respect to a "yes"-answer ($z_k = 1$) may be given by

$$PP_1 = \frac{\min[P(z_k = 1 \,|\, k \in A), P(z_k = 1 \,|\, k \in A^C)]}{\max[P(z_k = 1 \,|\, k \in A), P(z_k = 1 \,|\, k \in A^C)]} \qquad (7)$$

(cf. Quatember 2015b, p.445). For the conditional probabilities in Eq. (7),

$$P(z_k = 1 \,|\, k \in A) = p_1 + p_3 \cdot \pi_B + p_4 = u + v$$

and

$$P(z_k = 1 \,|\, k \in A^C) = p_2 + p_3 \cdot \pi_B + p_4 = u$$

applies, respectively, with the notations u and v presented above. Therefore, PP_1 can be expressed by

$$PP_1 = \frac{u}{u+v}. \qquad (8)$$

Regarding a "no"-answer ($z_k = 0$), the measure yields

$$PP_0 = \frac{\min[P(z_k = 0 \,|\, k \in A), P(z_k = 0 \,|\, k \in A^C)]}{\max[P(z_k = 0 \,|\, k \in A), P(z_k = 0 \,|\, k \in A^C)]} = \frac{1-u-v}{1-u} \qquad (9)$$

because of

$$P(z_k = 0 \,|\, k \in A) = p_2 + p_3 \cdot (1 - \pi_B) + p_5 = 1 - u - v$$

and

$$P(z_k = 0 \,|\, k \in A^C) = p_1 + p_3 \cdot (1 - \pi_B) + p_5 = 1 - u \,.$$

The ratios PP_1 and PP_0 of two conditional probabilities equal zero when the privacy of the respondents is not protected at all by the questioning design. This applies only in the case of the direct questioning model as one more special case of this RR strategy with $p_1 = 1$. A respondent's privacy protection measures PP_1 and PP_0 reach the maximum values of one, when privacy is totally protected, meaning that answer z_k bares no information on the true value y_k of the unit. In this case, privacy is completely protected at the cost of a maximum loss of information. For a detailed discussion of optimum design parameter values for different cases of sensitive variables in the RR model for the estimation of π_A, see Quatember, 2009, pp. 147ff.

After straightforward calculations, with the measures (89) and (9), the term C in (6) can be expressed as a function f of these measures by

$$C = f(PP_1, PP_0) = \frac{1}{N^2} \cdot \frac{1}{(1 - PP_1)(1 - PP_0)} \cdot \left[PP_0 \cdot \sum_U d_k \cdot y_k + PP_1 \cdot \sum_U d_k \cdot (1 - y_k) \right] \quad (10)$$

Eq. (10) shows that under the assumption of full cooperation for a given probability sampling scheme and given y_k-values in U, the efficiency of an RR strategy depends solely on the privacy protection offered by the questioning design.

4. Subjectively perceived privacy protection

Quatember (2012) proved that in contrast to what was claimed in some publications in the past multistage RR questioning designs theoretically cannot perform better than their one-stage basic versions, when the respondents' objective privacy protection measured by (8) and (9) is also taken into consideration. Clearly, a respondent's subjective perception of privacy protection might be higher for more complicated two- or more stage designs.

Fidler and Kleinknecht (1977) showed in their study for variables of very different levels of sensitivity, that their choice of the design parameters for this RR strategy yielded nearly full and truthful response for each variable including sexual behaviour (ibd., page 1048). Inserting

these values in (9) and (9) results in privacy protection values of 3/13. This finding corresponds in the main with results that can be derived from the experiment by Soeken and Macready (1982) and with recommendations given by Greenberg et al. (1969). Therefore, choosing PP_1 and PP_0 close to a value of 0.25 should be a good choice for most variables to avoid refusals and untruthful answering of respondents in a survey with randomized responses.

Another feature of the RR questioning technique affecting the perceived privacy protection in practice is the choice of the randomization device. Such a questioning design can be implemented, for example, in the following manner: The respondent may be asked, for instance, to think of a person whose date of birth he or she knows but without delivering this information to the interviewer. Then, if, as an example, the RR technique with p_1, $p_4 > 0$ and $p_1 + p_4 = 1$ is applied, if the date of birth is within a certain interval, such as from January to September, the respondent shall answer truthfully on the sensitive question. However, if the date is, say, from October to December, the respondent shall simply answer "yes." The disjoint date groups have to cover all possible dates of birth from January to December. The chosen allocation of these dates to groups determines the design probabilities.

A mathematically more sophisticated randomizing device with no uniform distribution that requires no instrument such as a dice was suggested by Diekmann (2012) and makes use of the Newcomb–Benford distribution (cf. Newcomb, 1881). For this purpose, the respondent may be asked to think of a person of whom he or she recalls the house number. Then, if the first digit of the house number is within a certain interval as from 1 to 6, the respondent shall answer truthfully on the interesting question from above. But, if it is not, the respondent shall answer "yes." The probability of the first digit follows the Newcomb-Benford distribution. Therefore, the probability of being asked the direct question is 0.845 (cf. ibid., p. 40). The probability of the rest is 0.155. If other probabilities are needed for the respective questioning design, the grouping of the first digits should be done in another way, or the first two digits can be used to produce more possible groups. Diekmann (2012) emphasizes that for this random device there is a discrepancy between the design probabilities as perceived by the respondents and the correct probabilities. For instance, the probability of picking a house

number with first digit from 1 to 6 is believed to be around 2/3 and not 0.845. This "illusion" (ibid., p. 330) has a positive effect on the perceived privacy protection with regard to the questioning design (for a discussion on the different aspects of theoretical and perceived privacy protection, see Chaudhuri and Christofides, 2013, Chap. 7).

5. Final remarks

If the direct questioning on the sensitive variable leads to non-ignorable nonresponse and untruthful answers, as expected very often in statistical surveys, a considerably biased estimator is the consequence. For such cases, the higher complexity of the RR questioning design will surely pay off under the assumption of cooperation. The accuracy of the estimators increases although their variances exceed the (but then only) theoretical variances of the direct questioning. It is evident that the performance of an RR strategy is a function of the level of privacy it offers to the respondents. The next investigations on the privacy protection dependency of RR estimators will consider also the practically relevant case of mixing true and randomized responses (Quatember, 2016).

6. References

Chaudhuri A. (2011), Randomized response and indirect questioning techniques in surveys, CRC Press, Boca Raton.

Chaudhuri A. and Christofides T.C. (2013), Indirect questioning in sample surveys, Springer, Heidelberg.

Chaudhuri A. and Christofides T.C. and Rao C.R. (eds.) (2016), Handbook of Statistics (Volume 34): Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, Elsevier, Amsterdam (in print).

Diekmann A. (2012), Making use of "Benford's Law" for the randomized response technique, Sociological Methods & Research, 41(2), pp. 325-334.

Fidler D.S. and Kleinknecht R.E. (1977), Randomized response versus direct questioning: Two data collection methods for sensitive information, Psychological Bulletin, 84 (5), pp. 1045-1049.

Greenberg B.G. and Abul-Ela A.-L.A. and Simmons W.R. and Horvitz D.G. (1969), The unrelated question randomized response model: Theoretical framework, Journal of the American Statistical Association, 64(326), pp. 520-539.

Horvitz D.G. and Shah B.V. and Simmons W.R. (1967), The unrelated question randomized response model, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 65-72.

Lensvelt-Mulders G.J.L.M. and Hox J.J. and van der Heijden P.G.M. and Maas C.J.M. (2005), Meta-analysis of randomized response research, Sociological Methods & Research, 33(3), pp. 319-348.

Newcomb S. (1881), Note on the frequency of use of the different digits in natural numbers, American Journal of Mathematics, 4, pp. 39-40.

Quatember A. (2009), A standardization of randomized response strategies, Survey Methodology,35(2), pp. 143-152.

Quatember A. (2012), An extension of the standardized randomized response technique to a multi-stage setup, *Statistical Methods & Applications, 21*(4), pp. 475-484.

Quatember A. (2014), A randomized response design for a polychotomous sensitive population and its application to opinion polls, Model Assisted Statistics and Applications, 9, pp. 11-23.

Quatember A. (2015a), Pseudo-Populations - A Basic Concept in Statistical Surveys, Springer, Cham.

Quatember A. (2015b), Warner in practice: How to incorporate true answers in a generalized Warner model, Model Assisted Statistics and Applications, 10, pp. 441-451.

Quatember A. (2016), A Mixture of True and Randomized Responses in the Estimation of the Number of People Having a Certain Attribute, In: Chaudhuri A. and Christofides T.C. and Rao C.R. (eds.) (2016), Handbook of Statistics (Volume 34): Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, Elsevier, Amsterdam, pp. 91-103.

Soeken K.L. and Macready G.B. (1982), Respondents' perceived protection when using randomized response, Psychological Bulletin, 92 (2), pp. 487-489.

Warner S.L. (1965), Randomized response: A survey technique for eliminating evasive answer bias, Journal of the American Statistical Association, 60, pp. 63-69.

Warner S.L. (1971), The linear randomized response model, Journal of the American Statistical Association, 66, pp. 884-888.