Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2009-42

The privacy protection point of view of Christofides' randomized response strategy

Andreas Quatember

March 2009

1 Introduction

The presence of nonresponse and untruthful answering is natural in survey sampling. The problem increases when the subject under study is of sensitive matter. Before imputation methods were developed to compensate only for the nonresponse at the estimation stage of a survey, Warner (1965) introduced his randomized response technique. This method can be used already at the design stage of a survey with the purpose to minimize both nonresponse and incorrect answers to allow an unbiased estimation of the relative size π_A of a sensitive subset U_A of population U. Since then several randomized response methods have been proposed. Quatember (2007a) presented a standardization of such techniques for the estimation of proportions and derived the statistical properties of the standardized estimator (see: Quatember (2009)).

Christofides (2003) published a very interesting alternative procedure (=C): An unit k of a simple random sample of size n drawn with replacement is assigned an integer y_k from the values $1, 2, ..., L$ with probabilities $p_1, p_2, ..., p_L$ ($\sum p_i = 1$). The respondent k is then asked to report the random variable d with

$$d_k = \begin{cases} L + 1 - y_k & \text{if } k \in U_A \\ y_k & \text{otherwise.} \end{cases}$$

The estimator

$$\widehat{\pi}_A^C = \frac{\overline{d} - E(y)}{L + 1 - 2E(y)}$$

($L + 1 \neq 2E(y)$) is an unbiased estimator of π_A with $\overline{d}$, the sample mean of the $d_k's$ (ibid., p.197). Its variance is given by

$$V(\widehat{\pi}_A^C) = \frac{\pi_A \cdot (1 - \pi_A)}{n} + \frac{1}{n} \cdot \underbrace{\frac{V(y)}{[L + 1 - 2E(y)]^2}}_{\equiv A}. \tag{1}$$

The undeniable psychological advantage of Christofides' procedure compared to Warner's technique is that the respondent does not have to answer "yes" or "no", which might reduce the perception of the risk of being exposed even if the respondent does understand how Warner's technique allows to protect his or her privacy. Chaudhuri (2004) extends Christofides' idea to unequal probability sampling. Christofides (2005a) does this for stratified sampling and Christofides (2005b) applies his technique to the estimation of the relative size of subgroups having two sensitive attributes at the same time.

In Section 2 the performances of Christofides' and Warner's technique are compared and it is shown, that Christofides example was calculated wrongly (Christofides, 2003, p.198). Section 3 lies its focus on the perceived privacy protection – a feature, which plays a very important role in this context. In the next section it is demonstrated that Warner's method cannot be less efficient than Christofides' when this feature is taken into account.

2 Efficiency comparisons

In Section 3 of Christofides (2003) the efficiency of his technique is compared to Warner's strategy (=W). The variance of Warner's estimator $\widehat{\pi}_A^W$ for simple random sampling with replacement is given by

$$V(\widehat{\pi}_A^W) = \frac{\pi_A \cdot (1-\pi_A)}{n} + \frac{1}{n} \cdot \underbrace{\frac{1}{4} \cdot ((2p-1)^{-2} - 1)}_{\equiv B} \tag{2}$$

(Warner, 1965, p.65). For the efficiency comparison of $\widehat{\pi}_A^C$ and $\widehat{\pi}_A^W$ the term A on the right hand side of (1) has to be compared to the term B on the right hand side of (2).

For this purpose Christofides (2003) uses an example with the following "design parameters" (ibid., p.198): For Warner's design the probability p of being asked the question "Are you a member of group U_A?" is arbitrarily fixed at $p = 0.6$. This results in $B = 6$. On the other hand for Christofides own questioning design the design parameters are given there by $L = 6$ and $p_1 = 0.26, p_2 = 0.05, p_3 = 0.1, p_4 = 0.19, p_5 = 0.02$ and $p_6 = 0.38$. It is easy to show, that Christofides calculated the term A of (1) incorrectly. Because of $V(y) = 4.14$ and $(L+1-2E(y))^2 = 0.36$, term A results in 11.5 and not in 3.76. Therefore unfortunately in this example it is absolutely wrong, "that the estimator resulting from our (=Christofides'; author's note) procedure has smaller variance than the estimator [of Warner's strategy]" (ibid., p.198). Moreover if we would choose the design parameter p of Warner's strategy just as arbitrary equal to 0.8, which is proposed as sufficient for the privacy protection for almost all sensitive attributes anyhow in some publications (cf. Greenberg et al., 1969, p.526, Fidler and Kleinknecht, 1977, p.1048, or Soeken and Macready, 1982, p. 488), then term B of formula (2) would actually only equal $0.\dot{4}$!

But nevertheless it is true, that "suitably choosing" other values for L and $p_1, p_2, ...p_L$ with $L \geq 3$ can construct an estimator $\widehat{\pi}_A^C$ which will have a smaller theoretical variance than $\widehat{\pi}_A^W$ of design W with any design parameter p. For example, when $L = 6$ and $p_1, p_2, ...p_6$ are 0.5, 0.15, 0.12, 0.1, 0.08 and 0.05 than A would only equal 0.402 (rounded).

Obviously for A to be smaller than B we have to choose

a) the probability p of Warner's strategy close to 0.5,

b) the probabilities $p_1, p_2, ...p_L$ of Christofides' method in a way, that $V(y)$ is small and

c) also in a way, that $E(y)$ is far away from $(L+1)/2$.

The last two points can be summarized by the recommendation to use a highly skewed random variable y for Christofides' technique.

3 Objective measure of privacy protection

All aspects enumerated at the end of Section 2 underpin the important role of the perceived privacy protection in this context. The closer to 0.5 p of questioning design W is chosen, the higher is the level of the respondent's privacy protection. The

higher the skewness of variable y in Christofides' design is, the lower is this level. It is necessary to measure the loss of privacy which is connected with a certain randomized response strategy before comparing different strategies, because differences in the loss of privacy result in different nonresponse and untruthful answering rates (cf. Quatember, 2007b). This loss can be measured for element k in various ways (cf. Chaudhuri and Mukerjee, 1987, p.83ff) – for instance by comparing the following conditional probabilities: $P(d_k = i|k \in U_A)$ and $P(d_k = i|k \in {U_A}^c)$ with ${U_A}^c = U - U_A$ and $i = 1, 2, ...L$.

The privacy of the respondent is totally protected by a questioning design for which these two probabilities are equal for each i. Modifying the Leysieffer-Warner measures of "jeopardy" (Leysieffer and Warner, 1976, p.650) to be usable for Christofides' technique, we let (for $i \in N$ and $i \leq \frac{L+1}{2}$)

$$\lambda_i = \frac{max(P(d_k = i|k \in U_A), P(d_k = i|k \in {U_A}^c))}{min(P(d_k = i|k \in U_A), P(d_k = i|k \in {U_A}^c))} = \frac{max(p_{L+1-i}, p_i)}{min(p_{L+1-i}, p_i)} \tag{3}$$

$(1 \leq \lambda_i \leq \infty)$ and define

$$\lambda = max_i \, \lambda_i \tag{4}$$

$(i \in N$ and $i \leq \frac{L+1}{2})$. λ measures the (maximum) loss of privacy, provided by the questioning design.

4 Consequences

Using the design parameters of Christofides (2003) example (ibid., p.198), for Warner's strategy with $p = 0.6$ λ is given by $0.6/0.4 = 1.5$. On the other hand $L = 6$ and $p_1, p_2, ...p_6$ equal to 0.26, 0.05, 0.1, 0.19, 0.02, 0.38 give $\lambda = 0.05/0.02 = 2.5$. This means, that there is an objective higher loss of privacy when the new strategy is used along with these design parameters. As the understanding of the way the procedure protects the sample unit's privacy is a necessary condition (cf. Landsheer et al., 1999, p.6ff) to reduce the individuals' fear of an embarrassing "outing", the strategy with the higher loss of privacy will produce higher nonresponse and untruthful answering rates.

One can easily do calculations of all possible combinations of the design parameters p_1 to p_6 in Christofides' example, which demonstrate that for the same "λ-level" variance (1) of Christofides' randomized response strategy actually cannot be smaller than variance (2) of Warner's. This means in the example, that compared to Warner's strategy with $p = 0.6$ (resulting in $\lambda = 1.5$) one cannot find a single combination of $p_1, p_2, ...p_6$, so that the estimator $\widehat{\pi}_A^C$ has a smaller variance than π_A^W as long as the condition of the same privacy protection has to be kept. The reason is, that Warner's questioning design is actually the best performing special case of Christofides' technique (having the most skewed distribution of y), as long as we keep an eye on the privacy protection that is connected with the questioning design. And from the author's point of view there is no alternative to doing so.

5 Summary

In Christofides' procedure like in all other randomized response strategies it is not allowed to choose the design parameters "suitably" (ibid., p.198) to be able to produce a more efficient estimation of π_A. The respondent's perceived privacy protection plays a significant role in this context. When we consider the "λ-measure" of loss of privacy, Christofides' method cannot be more efficient than Warner's design.

Of course, this objective measure of loss of privacy cannot measure the undeniable psychological advantage of not having to answer "yes" or "no" but to report just an integer. Therefore the individual's subjective perception of loss of privacy might be smaller than the objectively measured one. The estimator $\widehat{\pi}_A^C$ could then be more efficient than $\widehat{\pi}_A^W$, if the design parameters are chosen in a way, which increases λ only slightly. But a "suitable" choice of them, so that λ considerably exceeds Warner's λ would certainly increase both the nonresponse as well as the untruthful answering rate. Warner's estimator would then again have a smaller MSE.

References

Chaudhuri A (2004) Christofides' randomized response technique in complex sample surveys. Metrika 60: 223–228

Chaudhuri A, Mukerjee R (1987) Randomized Response. Marcel Dekker, New York

Christofides TC (2003) A generalized randomized response technique. Metrika 57: 195–200

Christofides TC (2005a) Randomized response in stratified sampling. Journal of Statistical Planning and Inference 128: 303–310

Christofides TC (2005b) Randomized response technique for two sensitive characetristics at the same time. Metrika 62: 53–66

Fidler DS, Kleinknecht RE (1977) Randomized Response Versus Direct Questioning: Two Data Collection Methods for Sensitive Information. Psychological Bulletin 84(5): 1045–1049

Greenberg BG, Abul-Ela A-LA, Simmons WR, Horvitz DG (1969) The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association 64: 520–539

Landsheer JA, van der Heijden P, van Gils G (1999) Trust and Understanding, Two Psychological Aspects of Randomized Response. Quality and Quantity 33: 1–12

Leysieffer FW, Warner SL (1976) Respondent Jeopardy and Optimal Designs in Randomized Response Models. Journal of the American Statistical Association 71: 649–656

Quatember A (2007a) A standardized technique of randomized response. IFAS Research Paper Series 28: www.ifas.jku.at/e2550/e2756/index_ ger.html (submitted)

Quatember A (2007b) Comparing the Efficiency of Randomized Response Techniques under Uniform Conditions. IFAS Research Paper Series 23: www.ifas.jku.at/e2550/e2756/index_ ger.html (submitted)

Quatember A (2009) To cope with nonresponse and untruthful answering: Different questioning designs for different variables. In: Proceedings of the NTTS2009 – Conferences on New Techniques and Technologies for Statistics, 18-20 February, Brussels, available under: www.ntts2009.eu.

Soeken KL, Macready GB (1982) Respondents' Perceived Protection When Using Randomized Response. Psychological Bulletin 92(2): 487–489

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