Computational Statistics and Data Analysis 67 (2013) 115–135

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A variant of the parallel model for sample surveys with sensitive characteristics

Yin Liu, Guo-Liang Tian *

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China

ARTICLE INFO

Article history:
Received 30 July 2011
Received in revised form 24 April 2013
Accepted 7 May 2013
Available online 23 May 2013

Keywords:
Asymptotic properties
Bayesian methods
Non-compliance behavior
Non-randomized response technique
The parallel model
Unmatched count technique

ABSTRACT

A new *non-randomized response* (NRR) model (called a variant of the parallel model) is proposed. The survey design and corresponding statistical inferences including likelihood-based methods, Bayesian methods and bootstrap methods are provided. Theoretical and numerical comparisons showed that the proposed variant of the parallel model over-performs two existing NRR crosswise and triangular models for most of the possible parameter ranges. An outline for handling the possible non-compliance behavior in the proposed model is provided. An illustrative example from an existing survey on 'sexual practices' in San Francisco, Las Vegas and Portland is used to demonstrate the proposed statistical analysis methods. Two real surveys on the cheating behavior in examinations at the University of Hong Kong are conducted and are used to illustrate the proposed design and analysis methods.

1. Introduction

Consider a target population which can be divided into two mutually exclusive groups: one with a sensitive attribute and the other without. Statistically, let Y be a sensitive binary variable, $\{Y = 1\}$ denote the population group that has the sensitive attribute and $\{Y = 0\}$ denote the complementary group. Usually, a well-designed survey is conducted for collecting sensitive data, which are used to estimate the proportion (denoted by $\pi = \Pr(Y = 1)$) of persons with the sensitive characteristic. Several techniques are developed to encourage truthful responses while protecting the privacy of respondents (or minimizing the interviewee's feeling of jeopardy). The first one is the *randomized response technique* (RRT), which includes Warner's design (Warner, 1965) and its improvement versions such as the unrelated question RR design (Horvitz et al., 1967; Greenberg et al., 1969). For a comprehensive review on RR designs, one is referred to Fox and Tracy (1986), Chaudhuri and Mukerjee (1988), and Chaudhuri (2011). One difficulty in implementing the RRT is the choice of an appropriate randomizing device in a self-administered setting. Another challenge in using RR models is the possible non-compliance behavior because of respondents' mistrust. A complicated or novel randomizing device may lead the interviewee to doubt the method itself or, even worst, to feel that they are being tricked by the interviewer into providing information under false pretenses. To handle non-compliance to RRT instructions, many developments on RR models were proposed by some researchers, e.g., Lakshmi and Raghavarao (1992), Clark and Desharnais (1998), Böckenholt and van der Heijden (2007), Van Den Hout and Klugkist (2009), Ostapczuk et al. (2009a), Ostapczuk et al. (2009b, 2011), Moshagen (2010), Van Den Hout et al. (2010) and Moshagen et al. (2012), and so on.

The second one is called the *unmatched count technique* (UCT), which provides absolute anonymity and confidentiality. Under UCT, two forms are needed: Form 1 contains a number of innocuous or neutral questions with answer 'yes' or 'no' and Form 2 is identical to Form 1, except for the addition of one embarrassing question of interest (Dalton et al., 1994, 1997; Coutts and Jann, 2011). The respondents of the survey are randomly assigned to one of the two groups. Participants

* Corresponding author. Tel.: +852 28591984; fax: +852 28589041.
E-mail address: gltian@hku.hk (G.-L. Tian).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.003

in group 1 (or group 2) are asked to reveal only the number of 'yes'-answer to all items listed in Form 1 (or Form 2). Since the interviewer does not know how they arrived at that number, it is safe to answer the sensitive question truthfully. One advantage of the UCT over the RRT is that no randomized device is required. The UCT is also called the item count technique (Droitcour et al., 1991; Tsuchiya et al., 2007), the unmatched block design, or block total response (Raghavarao and Federer, 1979). For more detailed description on the UCT, see Dalton et al. (1994).

The third one is called the *non-randomized response* (NRR) technique, which utilizes one or two independent non-sensitive random variates (e.g., respondent's birth date/month or the last digit of a respondent's ID card/phone number) combined with one or two sensitive random variables to form an incomplete contingency table and to indirectly obtain respondents' sensitive answers (Takahasi and Sakasegawa, 1977; Tian et al., 2007b, 2011; Yu et al., 2008; Tan et al., 2009; Tang et al., 2009). Like the UCT, the NRR designs do not require any randomizing devices.

One basic distinction between a randomized response model and a non-randomized response model is that the former usually requires a randomization device such as a coin or a die which is related to a random variable without reproducibility, while the latter requires an independent non-sensitive variate such as birth date combined with the sensitive response variable to form an incomplete contingency table, resulting in a reproducibility. That is, the same respondent may yield different answers depending on the outcome of the randomization device in repeated experiments (e.g., repeatedly flip a coin). For example, in the unrelated question design with a coin as the randomization device, if the outcome is head, the first question is answered; if the outcome is tail, the second question is answered. Suppose that the result of the first (second) flip is a tail (head), the answer is a 'yes' ('no'). As a result, interviewers do not know which answer should be collected.

It is not true that any randomized response model can be easily transformed to a non-randomized response model. Up to now, only the Warner model and the unrelated question model were successfully transformed to the non-randomized crosswise model (Yu et al., 2008) and the non-randomized parallel model (Tian, 2012), respectively. Next, although some randomized response models can be transformed to non-randomized versions, the resulting statistical analysis methods are totally different. For example, for the randomized unrelated question model with an unknown $\theta = \Pr(U = 1)$, two independent samples of sizes n_1 and n_2 and two randomization devices are required, while for its non-randomized version, i.e., the proposed variant of the parallel model in this paper, only one sample is needed without using any randomization devices and the corresponding statistical analysis methods are developed based on a trinomial distribution with two complete observations and one incomplete observation. The second example is as follows. To assess the association of two sensitive questions with binary outcomes, a randomized response model in general requires two randomization devices (Christofides, 2005), while in the non-randomized hidden sensitivity model (Tian et al., 2007b), respondents only need to answer a non-sensitive question instead of the original two sensitive questions and the corresponding analysis methods are developed based on an incomplete 4×4 contingency table. Finally, for other randomized response models (e.g., Kuk (1990)), the corresponding non-randomized partners are not yet available up to now.

Recently, Tian (2012) proposed a new NRR model, called the parallel model, to estimate the unknown proportion, $\pi = \Pr(Y = 1)$, of individuals with a sensitive characteristic. By introducing two non-sensitive dichotomous variates U and W such that Y, U and W are mutually independent, Tian (2012) developed a general framework of design and analysis for the NRR parallel model. Theoretical comparison showed that the parallel model over-performs two existing NRR crosswise and triangular models for most of the possible parameter ranges. It was noted that all these findings are based on the assumption of known proportions $\theta = \Pr(U = 1)$ and $p = \Pr(W = 1)$. However, in survey practice, it is usually difficult to choose an appropriate non-sensitive dichotomous variate U with known $\theta = \Pr(U = 1)$. Even such a binary variable U can be found and a constant θ_0 is assumed to be equal to the true value of the θ, how to test the hypothesis H_0: $\theta = \theta_0$ is still not available for the parallel design. The main goal of this paper is to propose a variant of the parallel model with unknown $\theta = \Pr(U = 1)$.

The rest of the paper is organized as follows. In Section 2, we propose the survey design for the variant of the parallel model, and discuss the estimation of the parameters, relative efficiency and the degree of privacy protection. In Section 3, three asymptotic *confidence intervals* (CIs) and the exact CI of π are derived. In addition, a modified *maximum likelihood estimate* (MLE) of π is provided and the corresponding asymptotic property is investigated. Statistical inferences on θ and two bootstrap CIs of the parameters are given in Section 4 and 5, respectively. Bayesian inferences are discussed in Section 6. Comparisons with the NRR crosswise and triangular models are conducted theoretically and numerically in Section 7. An outline for handling the possible non-compliance behavior in the proposed model is presented in Section 8. In Section 9, an illustrative example from an existing survey on 'sexual practices' in San Francisco, Las Vegas and Portland is used to demonstrate the proposed statistical analysis methods. Two real surveys on the cheating behavior in examinations at the University of Hong Kong are conducted and are used to illustrate the proposed design and analysis methods. A discussion is given in Section 10. The exact *inversion Bayesian formulas* (IBF) sampling is provided in the Appendix.

2. A new non-randomized response model: a variant of the parallel model

2.1. The survey design for the variant of the parallel model

Let $\{Y = 1\}$ denote the population class with a sensitive characteristic and $\{Y = 0\}$ denote the complementary class. The objective is to estimate the proportion $\pi = \Pr(Y = 1)$. Suppose that U and W are two non-sensitive dichotomous