ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Sample size determination for estimating prevalence and a difference between two prevalences of sensitive attributes using the non-randomized triangular design

Shi-Fang Qiu [a,*], G.Y. Zou [b], Man-Lai Tang [c]

[a] Department of Statistics, Chongqing University of Technology, Chongqing 400054, China

[b] Department of Epidemiology and Biostatistics, and Robarts Clinical Trials of Robarts Research Institute, Western University, London, Ontario, Canada N6A 5C1

[c] Department of Mathematics and Statistics, Hang Seng Management College, Hong Kong, China

ARTICLE INFO

Article history:
Received 25 April 2013
Received in revised form 5 November 2013
Accepted 26 February 2014
Available online xxxx

Keywords:
Non-randomized response
Prevalence
Confidence interval
Proportion
Sensitive

ABSTRACT

A non-randomized triangular design has been shown to be more efficient than the conventional random response model in estimating the prevalence of sensitive attributes in surveys. Since most surveys focus on estimation, herein we derive sample size formulas for estimation of prevalence and a difference between two prevalences in this design. In contrast to the conventional approach to sample size estimation, we explicitly incorporate into the formulas an assurance probability of achieving the pre-specified precision. Exact evaluation results demonstrate that these formulas perform well. The methods are illustrated using data from a real study.

1. Introduction

The success of many public health preventions and interventions relies crucially on reliable information of prevalence of sensitive attributes such as illicit drug use, risky behavior, cheating, or non-adherence to prescribed mediation or treatment. Directly asking people about such sensitive questions is generally problematic because of refusal or intentional response bias. Consequently, much effort has been made to develop methods that are effective in obtaining reliable information on sensitive attributes. The most well-known approach has been the randomized response technique suggested by Warner (1965).

In this design, an interviewee is presented with two mutually exclusive statements about the sensitive attributes, such as (a) 'I cheated' and (b) 'I never cheated', and is then instructed to provide an answer of 'Correct' or 'Incorrect' for statement (a) or (b), depending on the outcome from a randomizing device such as a dice or spinner provided by the interviewer. Without the interviewer knowing the outcome of the randomizing device, the interviewee then provides an answer. The privacy of the interviewee is protected by the fact that the interviewer only knows the answer, but does not know to which statement the interviewee is referring. While the outcome of the randomizing device for each individual is unknown to the interviewer, the chance of the randomizing device directing the interviewee to answer (a) and (b) is controlled by the interviewer. Estimation of the prevalence of the sensitive attributes can then be made at the aggregate level, using maximum likelihood theory (Warner, 1965).

* Corresponding author.
E-mail addresses: sfqiu@cqut.edu.cn, sfqiu@amss.ac.cn (S.-F. Qiu).

http://dx.doi.org/10.1016/j.csda.2014.02.019
0167-9473/

The Warner design has been applied in a wide range of contexts (see Lensvelt-Mulders et al., 2005, for a review). However, previous research has shown that the Warner model has two major drawbacks. First, the Warner model lacks of reproducibility due to the randomized device. Second, the technique results in an inefficiency when the probability of the non-sensitive binary attribute is not equal to 0.5. To overcome these limitations, Yu et al. (2008) proposed a non-randomized triangular design that uses an independent non-sensitive statement such as season of birth in the survey to indirectly obtain the answer to the sensitive question. For example, to estimate the prevalence of cheating, a respondent is asked to answer a 'Correct' or 'Incorrect' to the following statement: 'I have never cheated and my mother was born between May and August'. Previous research has shown that this non-randomized triangular design can be a viable alternative to the Warner random response design (Tan et al., 2009).

Tian et al. (2011) derived sample size formulas for the non-randomized triangular design based on the power analysis approach, while Wu et al. (in press) compared sample size formulas based on the asymptotic Wald test, score test and likelihood rate test that guarantees a nominal power of a hypothesis test at a significance level. All these formulas are based on hypothesis testing, which may not be the focus in many situations. In this paper, we present closed-form formulas for calculating sample size for the non-randomized triangular design when the objective is to estimate the prevalence of sensitive attributes and their differences. We adapted the idea used for the estimation of intraclass correlation coefficient in reliability studies (Zou, 2012). We also point out that Kelly (2007) has previously developed an R package for sample size estimation based on confidence intervals for common effect measures in behavior science, although no closed-form formula is available. In contrast to conventional sample size determination for confidence interval estimation, our formulas explicitly incorporates an assurance probability of achieving pre-specified precision, i.e., confidence interval width. For one-sample problem, we derive the sample size formulas on the basis of the Wald confidence interval and the Wilson confidence interval for the prevalence of a sensitive attribute. For a difference between two prevalences, we consider the sample size formulas based on the Wald confidence interval and the method of variance estimates recovery (MOVER) (Zou, 2008). The evaluation results show that the formulas perform well on the basis of the Wilson-type and the MOVER confidence intervals in a wide range of parameter combinations.

The rest of the paper is organized as follows. Section 2 provides a brief review of a non-randomized design. Confidence interval estimators for the prevalence of a sensitive attribute and the difference between two prevalence rates are given in Section 3. Corresponding sample size formulas are derived and illustrated in Section 4, followed by an evaluation in Section 5. The methods are illustrated using data from an induced abortion study in Taiwan. We conclude with a summary in Section 7.

2. A non-randomized triangular design for sensitive attributes

Let Y denote the variable of the sensitive attribute of interest, such as cheating, with value 1 being 'have cheated' and 0 'have not cheated'. Let W be a non-sensitive binary attribute that is independent of Y, such as 'born between May and August'. Here, W should be so chosen that the probability of $W = 1$ is known or easily estimated and we denote $\Pr(W = 1)$ by p. The aim is to estimate the probability of $Y = 1$, denoted as π.

In a face-to-face interview, the interviewer may use the format on the left-hand side of Table 1 and ask the interviewee to put a tick in the open circle or in the triangle formed by the three dots, depending on whether or not the event $\{Y = 0$ and $W = 0\}$ is true. This design has been referred to as the triangular design (Yu et al., 2008). The cell probabilities for the right side of Table 1 can be obtained by multiplying the marginal probabilities, since W and Y are independent by design. Thus, the probability of ticking the circle is given by $(1-\pi)(1-p)$, while that for ticking the triangular is $1-(1-\pi)(1-p)$, that is, $\pi+(1-\pi)p$, where p is assumed to be a known constant. Let $\Delta = \pi+(1-\pi)p$, that is, the probability of ticking the triangular is Δ, and then that of ticking the circle is $1-\Delta$. Thus, the number of respondents ticking the triangular follows a binomial distribution with parameters of n and $\Delta = \pi + (1-\pi)p$.

3. Confidence interval estimation

3.1. Confidence intervals for the prevalence of a sensitive attribute

Suppose that x of n subjects put a tick in the triangle in Table 1. Hence, the maximum likelihood estimate $\widehat{\Delta}$ of Δ (that is, the proportion of respondents ticking the triangular) is given by $\widehat{\Delta} = x/n$. Since $\widehat{\Delta} = \widehat{\pi} + (1-\widehat{\pi})p$, we have $\widehat{\pi} = (\widehat{\Delta} - p)/(1-p)$, with variance given by

$$\mathrm{var}(\widehat{\pi}) = \frac{\mathrm{var}(\widehat{\Delta})}{(1-p)^2} = \frac{\Delta(1-\Delta)}{n(1-p)^2}, \tag{1}$$

which may be consistently estimated by substituting $\widehat{\Delta}$ for Δ. A $(1-\alpha)100\%$ two-sided confidence interval for π can thus be obtained by

$$\widehat{\pi} \mp z_{\alpha/2}\left\{\widehat{\Delta}(1-\widehat{\Delta})/[n(1-p)^2]\right\}^{1/2} \tag{2}$$

where and throughout the paper $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. This procedure is referred to as the Wald method. Alternatively, one can first apply the Wilson (1927) method for constructing a confidence interval

Please cite this article in press as: Qiu, S.-F., et al., Sample size determination for estimating prevalence and a difference between two prevalences of sensitive attributes using the non-randomized triangular design. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.02.019