Journal of the Korean Statistical Society 43 (2014) 235–249

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Sample size determination for the parallel model in a survey with sensitive questions

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 21 May 2013
Accepted 31 August 2013
Available online 21 September 2013

AMS 2000 subject classifications:
primary 62D05
secondary 62H15

Keywords:
Crosswise model
Non-randomized response technique
Parallel model
Power function
Sample size formula
Triangular model

ABSTRACT

Recently, a new non-randomized parallel design is proposed by Tian (2013) for surveys with sensitive topics. However, the sample size formulae associated with testing hypotheses for the parallel model are not yet available. As a crucial component in surveys, the sample size formulae with the parallel design are developed in this paper by using the power analysis method for both the one- and two-sample problems. We consider both the one- and two-sample problems. The asymptotic power functions and the corresponding sample size formulae for both the one- and two-sided tests based on the large-sample normal approximation are derived. The performance is assessed through comparing the asymptotic power with the exact power and reporting the ratio of the sample sizes with the parallel model and the design of direct questioning. We numerically compare the sample sizes needed for the parallel design with those required for the crosswise and triangular models. Two theoretical justifications are also provided. An example from a survey on 'sexual practices' in San Francisco, Las Vegas and Portland is used to illustrate the proposed methods.

1. Introduction

In medical, epidemiological, public health, political, psychological, behavioral and sociological surveys, investigators would like to gather useful information on some sensitive topics or highly private questions such as sex, AIDs, abortion, drug-taking, gambling, tax evasion and so on. When such sensitive questions are asked directly, some respondents may refuse to answer or even give false answers to protect their privacy. As a result, statistical inferences based on such survey data might lead to inaccurate, unreliable or even wrong conclusions.

In order to alleviate the level of difficulty with the above-mentioned problems, Warner (1965) developed a randomized response technique which facilitates investigators to collect relatively reliable information while protecting privacy of the interviewees. The aim is to estimate the proportion of subjects with a sensitive attribute in a population. Later, some researchers extended Warner's model to other randomized response models (Mangat, 1994; Mangat & Singh, 1990; Singh & Mangat, 1996).

However, a major obstacle for the wide application of these randomized response techniques in survey practices is that some interviewees still provide untruthful answers. A possible reason is that there is a lack of trust from the interviewees because these randomization devices are totally controlled by interviewers. Thus, in order to avoid the usage of randomizing devices, recently other authors developed a non-randomized response technique (Tan, Tian, & Tang, 2009; Tang, Tian, Tang, & Liu, 2009; Tian, Tang, Liu, Tan, & Tang, 2011; Tian, Yu, Tang, & Geng, 2007; Yu, Tian, & Tang, 2008). And they have demonstrated that these non-randomized response models usually perform better than the corresponding randomized response partners in terms of efficiency and degree of privacy protection.

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

1226-3192/$ – see front matter © 2013 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jkss.2013.08.002

Table 1
The parallel model and the corresponding cell probabilities.

Category	$W=0$	$W=1$	Category	$W=0$	$W=1$	Marginal
$U=0$	○		$U=0$	$(1-q)(1-p)$		$1-q$
$U=1$	□		$U=1$	$q(1-p)$		q
$Y=0$		○	$Y=0$		$(1-\pi)p$	$1-\pi$
$Y=1$		□	$Y=1$		πp	π
			Marginal	$1-p$	p	1

Note: Please connect the two circles by a straight line if you belong to one of the two circles or connect the two squares by a straight line if you belong to one of the two squares.

More recently, Tian (2013) developed a so-called parallel design and numerically and theoretically showed that it is more efficient than the existing non-randomized crosswise and triangular designs in certain situations. However, the sample size formulae associated with testing hypotheses for the parallel model are not yet available. Since the sample size determination is a crucial step in survey practices, the main objective of this article is to develop the sample size formulae with the parallel design by using the power analysis method for both the one- and two-sample problems.

The rest of this paper is organized as follows. In Section 2, we consider the situation of one-sample problem and derive asymptotic power functions and the corresponding sample size formulae for both one- and two-sided tests based on the large-sample normal approximation. In Section 3, the performance is assessed through comparing the asymptotic power with the exact power and reporting the ratio of the sample sizes with the parallel model and the design of direct questioning. In Section 4, we numerically compare the sample sizes needed for the parallel design with those required for the crosswise model. A theoretical justification is also provided. Similar comparisons between the parallel design and the triangular design are numerically and theoretically performed in Section 5. In Section 6, we derive the sample size formula for the two-sample problem. In Section 7, an example from a survey on 'sexual practices' in San Francisco, Las Vegas and Portland is used to illustrate the proposed methods. A discussion is given in Section 8 and all technical details are put in the Appendix.

2. The non-randomized parallel model

In this section, first we briefly introduce the survey design for the non-randomized parallel model proposed by Tian (2013). Second we derive sample size formulae for both the one-sided and two-sided tests based on the power analysis method.

2.1. The survey design for the parallel model

Suppose that Y is a Bernoulli random variable corresponding to a sensitive question Q_Y (e.g., have you ever taken drugs?). Let $Y = 1$ if the answer to the question Q_Y is 'yes' and $Y = 0$ if the answer to the question Q_Y is 'no'. We are interested in estimating the unknown proportion $\pi = \Pr\{Y = 1\}$. To this end, we assume that there are two non-sensitive dichotomous variates W and U such that W, U and Y are mutually independent and $p = \Pr\{W = 1\}$ and $q = \Pr\{U = 1\}$ are known. For example, we may define $W = 1$ if the birthday of the respondent's mother is between 1 and 15 of a month and $W = 0$ otherwise. Similarly, we could define $U = 1$ if the respondent was born in the first half of a year and $U = 0$ otherwise. Thus, it is reasonable to assume that $p \approx q \approx 0.5$. More discussions on the model assumptions and the choice of W and U are given in Section 8 of Tian (2013).

Table 1 shows the survey scheme for the parallel model (Tian, 2013). The interviewer may ask the interviewee to connect the two circles by a straight line if he/she belongs to $\{U = 0,\ W = 0\}$ or $\{Y = 0,\ W = 1\}$; otherwise connect the two squares. Note that all $\{W = 0\}$, $\{W = 1\}$, $\{U = 0\}$, $\{U = 1\}$ and $\{Y = 0\}$ are non-sensitive classes, thus $\{U = 1, W = 0\} \cup \{Y = 1,\ W = 1\}$ is also a non-sensitive subclass. Therefore, whether the interviewee belongs to the sensitive class is not on record. The corresponding cell probabilities are displayed at the right-hand side of Table 1.

2.2. Sample size formulae based on the power analysis method

Following Table 1, we define a Bernoulli random variable Y^{P} as

$$Y^{\mathrm{P}} = \begin{cases} 1, & \text{if the two squares are connected,} \\ 0, & \text{if the two circles are connected,} \end{cases}$$

where the superscript 'P' indicates the Bernoulli variable for the parallel model. Thus, the probabilities of $Y^{\mathrm{P}} = 1$ and $Y^{\mathrm{P}} = 0$ are given by

$$\Pr\{Y^{\mathrm{P}} = 1\} = q(1-p) + \pi p \quad \text{and} \quad \Pr\{Y^{\mathrm{P}} = 0\} = (1-q)(1-p) + (1-\pi)p,$$

respectively.

Let $Y_{\text{obs}} = \{y_i^{\text{P}} : i = 1, \ldots, n\}$ denote the observed data for the n respondents, then the likelihood function for π is

$$L_{\text{P}}(\pi|Y_{\text{obs}}) = \prod_{i=1}^{n} \Big[q(1-p) + \pi p\Big]^{y_i^{\text{P}}} \Big[(1-q)(1-p) + (1-\pi)p\Big]^{1-y_i^{\text{P}}}.$$

Consequently, the *maximum likelihood estimate* (MLE) of π is

$$\hat{\pi}_{\text{P}} = \frac{\bar{y}^{\text{P}} - q(1-p)}{p}, \tag{2.1}$$

where $\bar{y}^{\text{P}} = (1/n)\sum_{i=1}^{n} y_i^{\text{P}}$. It is easy to verify that $\hat{\pi}_{\text{P}}$ is an unbiased estimator of π and the variance of $\hat{\pi}_{\text{P}}$ is given by

$$\text{Var}(\hat{\pi}_{\text{P}}) = \frac{\delta(1-\delta)}{np^2},$$

where $\delta \,\hat{=}\, q(1-p) + \pi p$. According to the Central Limit Theorem, $\hat{\pi}_{\text{P}}$ is asymptotically normally distributed as $n \to \infty$, i.e.,

$$\frac{\hat{\pi}_{\text{P}} - \pi}{\sqrt{\text{Var}(\hat{\pi}_{\text{P}})}} = \frac{n\hat{\pi}_{\text{P}} - n\pi}{\sqrt{n\delta(1-\delta)}/p} \mathrel{\dot{\sim}} N(0, 1). \tag{2.2}$$

2.2.1. The one-sided test

In order to test whether the population proportion (π) with the sensitive characteristic is identical to a pre-specified value (π_0), the following hypotheses are often considered,

$$H_0 : \pi = \pi_0 \quad \text{versus} \quad H_1 : \pi < \pi_0. \tag{2.3}$$

If the null hypothesis H_0 is true, from (2.2), we have

$$\frac{n\hat{\pi}_{\text{P}} - n\pi_0}{\sqrt{n\delta_0(1-\delta_0)}/p} \mathrel{\dot{\sim}} N(0, 1), \quad \text{as } n \to \infty,$$

where $\delta_0 \,\hat{=}\, q(1-p) + \pi_0 p$. Let z_α denote the upper α-th quantile of the standard normal distribution. When the event

$$\mathbb{E}_{\text{P}} = \Big\{n\hat{\pi}_{\text{P}} \le n\pi_0 - z_\alpha\sqrt{n\delta_0(1-\delta_0)}/p\Big\} \tag{2.4}$$

is observed, we should reject the null hypothesis H_0 at the α level of significance. If H_1 is true, without loss of generality, we can assume that $\pi = \pi_1$ with $\pi_1 < \pi_0$. Thus, the power of the one-sided test can be calculated approximately by

$$\begin{aligned} \text{Power (at } \pi_1) &= \Pr\{\text{rejecting } H_0|\pi = \pi_1\} \\ &= \Pr\left\{\frac{n\hat{\pi}_{\text{P}} - E_{H_1}(n\hat{\pi}_{\text{P}})}{\sqrt{\text{Var}_{H_1}(n\hat{\pi}_{\text{P}})}} \le \frac{n\pi_0 - z_\alpha\sqrt{n\delta_0(1-\delta_0)}/p - n\pi_1}{\sqrt{n\delta_1(1-\delta_1)}/p}\right\} \\ &\doteq \Phi\left(\frac{\sqrt{n}(\pi_0 - \pi_1)p - z_\alpha\sqrt{\delta_0(1-\delta_0)}}{\sqrt{\delta_1(1-\delta_1)}}\right), \end{aligned} \tag{2.5}$$

where $\delta_1 \,\hat{=}\, q(1-p) + \pi_1 p$ and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. For a given power, say, $1-\beta$, the required sample size n_{P} can be determined by solving the following equation

$$\sqrt{n_{\text{P}}}(\pi_0 - \pi_1)p - z_\alpha\sqrt{\delta_0(1-\delta_0)} = z_\beta\sqrt{\delta_1(1-\delta_1)},$$

which yields

$$n_{\text{P}} = \left[\frac{z_\alpha\sqrt{\delta_0(1-\delta_0)} + z_\beta\sqrt{\delta_1(1-\delta_1)}}{(\pi_0 - \pi_1)p}\right]^2. \tag{2.6}$$

2.2.2. The two-sided test

For a two-sided test, the two-sided hypotheses are specified by

$$H_0 : \pi = \pi_0 \quad \text{versus} \quad H_1 : \pi \neq \pi_0.$$

Given a significance level α, we only consider the equal-tailed rejection region. Note that the relationship among the power, sample size and effect size is approximately given by

$$\text{Power (at } \pi_1) \doteq \Phi\left(\frac{\sqrt{n}|\pi_0 - \pi_1|p - z_{\alpha/2}\sqrt{\delta_0(1-\delta_0)}}{\sqrt{\delta_1(1-\delta_1)}}\right).$$

In this case, the sample size is still given by (2.6) except for replacing the critical value z_α with $z_{\alpha/2}$.

3. Evaluation of performance

3.1. Comparison of the asymptotic power with the exact power

The asymptotic power function for the one-sided test is given by (2.5). To derive the exact power formula, we define a new random variable $X_{\rm P} = n\bar{y}^{\rm P} = \sum_{i=1}^{n} y_i^{\rm P}$. Then, we have $X_{\rm P} \sim \text{Binomial}(n, \delta)$ with $\delta = q(1-p)+\pi p$. The rejection region $\mathbb{E}_{\rm P}$ specified in (2.4) can be rewritten as

$$\mathbb{E}_{\rm P} = \left\{ X_{\rm P} : X_{\rm P} \le n\delta_0 - z_\alpha\sqrt{n\delta_0(1-\delta_0)} \right\}.$$

The exact power (at π_1) for any particular sample size n is determined by the following formula,

$$\begin{aligned} \text{Exact power (at } \pi_1) &= \sum_{x\in\mathbb{E}_{\rm P}} \text{Binomial } (x|n, q(1-p)+\pi_1 p) \\ &= \sum_{x\in\mathbb{E}_{\rm P}} \binom{n}{x} \delta_1^x (1-\delta_1)^{n-x}, \end{aligned} \tag{3.1}$$

where $\delta_1 = q(1-p)+\pi_1 p$. To compare the accuracy of the approximate power formula given by (2.5), in Fig. 1, we plot the exact and asymptotic powers against the sample size n for various combinations of (π_0, π_1) at $p = q = 0.5$ and $\alpha = 0.05$. Fig. 1 shows that, in general, the asymptotic power function given by (2.5) is a satisfactory approximation to the exact power defined by (3.1). Especially for large sample size n, the approximate power and the exact power are nearly the same (see in Fig. 1(iv)).

3.2. Comparison with the design of direct questioning in sample sizes

For a given pair of (π_0, π_1), we note that $n_{\rm P}$ is a decreasing function of p and an increasing function of q. It is clear that the parallel design reduces to the *design of direct questioning* (DDQ) when $p = 1$. Let $n_{\rm D}$ denote the sample size of the DDQ. In (2.6), setting $p = 1$, we obtain

$$n_{\rm D} = \left[\frac{z_\alpha\sqrt{\pi_0(1-\pi_0)} + z_\beta\sqrt{\pi_1(1-\pi_1)}}{\pi_0 - \pi_1} \right]^2. \tag{3.2}$$

Given 5% level of significance and 80% power, Table 2 reports the sample size $n_{\rm P}$ defined by (2.6) and the corresponding ratio $n_{\rm P}/n_{\rm D}$ for various combinations of (π_0, π_1, q, p). For example, when $(\pi_0, \pi_1, q, p) = (0.40, 0.25, 1/3, 0.50)$, we have $n_{\rm P}/n_{\rm D} = 4.03$, indicating that the sample size required for the parallel design is about four times of that required for the DDQ in order to achieve the identical power for the one-sided test.

In Table 2, we choose the non-sensitive dichotomous variate W to be the respondent's birthday and U to be the birthday of the respondent's mother. For example, $p = 0.42$ (i.e., $5/12$), 0.50 (i.e., $6/12$) and 0.58 (i.e., $7/12$) represent that we define $W = 1$ if the respondent was born between January to May, January to June and January to July of a year, respectively. Similarly, $q = 1/3$, $1/2$ and $2/3$ represent that we define $U = 1$ if the respondent's mother was born between the 1-st to the 10-th, the 1-st to the 15-th and the 1-st to the 20-th of a month, respectively.

4. Comparison with the Crosswise Model

In this section, we first describe the crosswise model (Yu et al., 2008) and then derive the sample size formula for the crosswise model based on the power analysis method. We next numerically compare the sample size required for the crosswise design with that needed for the parallel design. Finally, a theoretical justification is also provided.

4.1. The crosswise model

Let Y and W have the same definition as in Section 2.1, where $p = \Pr(W = 1)$ and $\pi = \Pr(Y = 1)$. The interviewer may design the questionnaire in the format as shown on the left-hand side of Table 3 and asks the interviewee to put a tick in the upper circle (i.e., $\{Y = 0, W = 0\}$) if he/she belongs to one of the two circles or put a tick in the upper square (i.e., $\{Y = 0, W = 1\}$) if he/she belongs to one of the two squares. Note that both $\{Y = 0, W = 0\}$ and $\{Y = 0, W = 1\}$ are non-sensitive. Thus, whether an interviewee belongs to the sensitive class (i.e., $\{Y = 1\}$) will not be revealed if a tick is put in the upper circle/square. Yu et al. (2008) called this the *crosswise model*.

Table 2
Sample size n_{P} for testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi = \pi_1 < \pi_0$ with 5% level of significance and 80% power and the ratio $n_{\mathrm{P}}/n_{\mathrm{D}}$.

π_0	π_1	q	$p = 1.00$	$p = 0.42$		$p = 0.50$		$p = 0.58$	
			n_{D}	n_{P}	$n_{\mathrm{P}}/n_{\mathrm{D}}$	n_{P}	$n_{\mathrm{P}}/n_{\mathrm{D}}$	n_{P}	$n_{\mathrm{P}}/n_{\mathrm{D}}$
0.50	0.40	1/3	153	832	5.44	592	3.87	444	2.90
	0.35		67	367	5.48	261	3.90	195	2.91
	0.30		37	204	5.51	145	3.92	108	2.92
0.40	0.35	1/3	583	3 206	5.50	2 273	3.90	1 697	2.91
	0.30		142	793	5.58	561	3.95	418	2.94
	0.25		61	348	5.70	246	4.03	183	3.00
0.30	0.25	1/3	501	3 009	6.01	2 108	4.21	1 555	3.10
	0.20		119	742	6.24	518	4.35	381	3.20
	0.18		81	512	6.32	357	4.41	262	3.23
0.20	0.16	1/3	584	4 333	7.42	2 972	5.09	2 142	3.67
	0.13		181	1 400	7.73	957	5.29	687	3.80
	0.10		83	678	8.17	462	5.57	330	3.98
0.10	0.08	1/3	1303	15 634	12.00	10 372	7.96	7 185	5.51
	0.06		301	3 874	12.87	2 562	8.51	1 767	5.87
	0.04		121	1 706	14.10	1 124	9.29	772	6.38
0.50	0.40	1/2	153	875	5.72	617	4.03	458	2.99
	0.35		67	388	5.79	273	4.07	203	3.03
	0.30		37	217	5.86	153	4.14	113	3.05
0.40	0.35	1/2	583	3 470	5.95	2 438	4.18	1 803	3.09
	0.30		142	864	6.08	606	4.27	447	3.15
	0.25		61	382	6.26	268	4.39	197	3.23
0.30	0.25	1/2	501	3 388	6.76	2 356	4.70	1 721	3.43
	0.20		119	841	7.07	583	4.90	425	3.57
	0.18		81	583	7.20	404	4.99	293	3.62
0.20	0.16	1/2	584	5 095	8.72	3 483	5.96	2 491	4.27
	0.13		181	1 655	9.14	1 128	6.23	804	4.44
	0.10		83	806	9.71	548	6.60	389	4.69
0.10	0.08	1/2	1303	19 347	14.85	12 898	9.90	8 928	6.85
	0.06		301	4 814	15.99	3 202	10.64	2 210	7.34
	0.04		121	2 129	17.60	1 413	11.68	972	8.03
0.50	0.40	2/3	153	851	5.56	606	3.96	454	2.97
	0.35		67	380	5.67	270	4.03	202	3.01
	0.30		37	214	5.78	152	4.11	114	3.08
0.40	0.35	2/3	583	3 472	5.96	2 466	4.23	1 837	3.15
	0.30		142	870	6.13	617	4.35	458	3.23
	0.25		61	387	6.34	274	4.49	203	3.33
0.30	0.25	2/3	501	3 504	6.99	2466	4.92	1 814	3.62
	0.20		119	875	7.35	615	5.17	451	3.79
	0.18		81	608	7.51	426	5.26	312	3.85
0.20	0.16	2/3	584	5 449	9.33	3 780	6.47	2 727	4.67
	0.13		181	1 776	9.81	1 230	6.80	885	4.89
	0.10		83	869	10.47	600	7.23	430	5.18
0.10	0.08	2/3	1303	21 422	16.44	14 564	11.18	10 221	7.84
	0.06		301	5 345	17.76	3 628	12.05	2 539	8.44
	0.04		121	2 371	19.60	1 606	13.27	1 121	9.26

Note: n_{D} denotes the sample size of the DDQ, given by (3.2).

4.2. Sample size formula for the crosswise model

Let $Y_{\mathrm{obs}} = \{y_i^{\mathrm{C}} : i = 1, \ldots, n\}$ denote the observed data for the n respondents with $y_i^{\mathrm{C}} = 1$ if the i-th respondent puts a tick in the upper circle and $y_i^{\mathrm{C}} = 0$ if the i-th respondent puts a tick in the upper square. The likelihood function for π is

$$L_{\mathrm{C}}(\pi|Y_{\mathrm{obs}}) = \prod_{i=1}^{n}\Big[(1-\pi)(1-p)+\pi p\Big]^{y_i^{\mathrm{C}}}\Big[(1-\pi)p+\pi(1-p)\Big]^{1-y_i^{\mathrm{C}}}$$

so that the MLE of π and its variance are given by

$$\hat{\pi}_{\mathrm{C}} = \frac{\bar{y}^{\mathrm{C}} - (1-p)}{2p-1} \quad \text{and} \quad \mathrm{Var}(\hat{\pi}_{\mathrm{C}}) = \frac{\gamma(1-\gamma)}{n(2p-1)^2}, \tag{4.1}$$

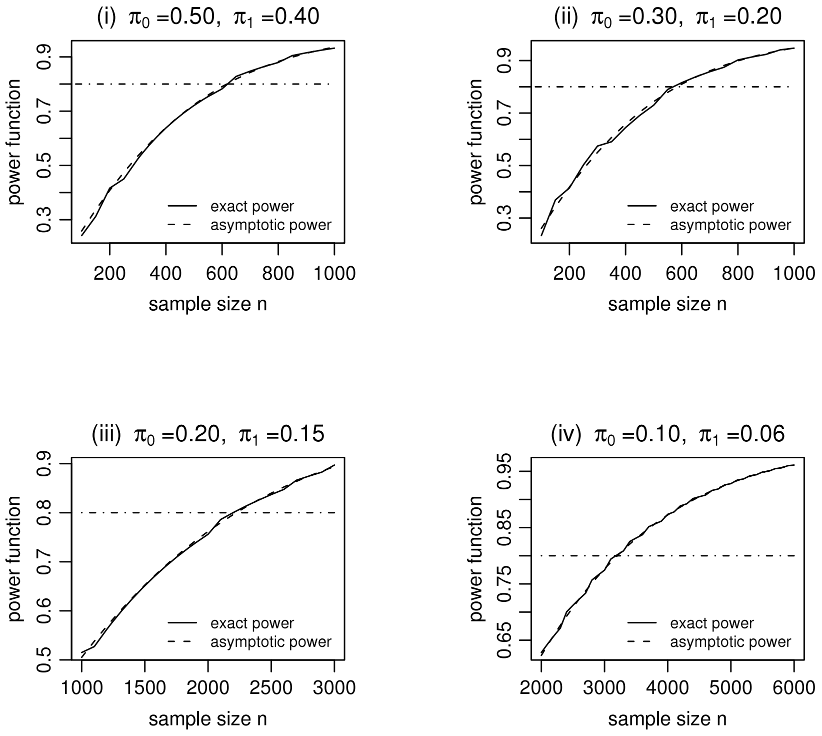

Fig. 1. Comparisons of the exact power (3.1) (denoted by solid line) with the asymptotic power (2.5) (denoted by dashed line) against the sample size n for various combinations of (π_0, π_1) at $p = q = 0.5$ and $\alpha = 0.05$. (i) $(\pi_0, \pi_1) = (0.50, 0.40)$; (ii) $(\pi_0, \pi_1) = (0.30, 0.20)$; (iii) $(\pi_0, \pi_1) = (0.20, 0.15)$; (iv) $(\pi_0, \pi_1) = (0.10, 0.06)$.

Table 3
The crosswise model and the corresponding cell probabilities.

Category	$W = 0$	$W = 1$	Category	$W = 0$	$W = 1$	Marginal
$Y = 0$	○	□	$Y = 0$	$(1-\pi)(1-p)$	$(1-\pi)p$	$1-\pi$
$Y = 1$	□	○	$Y = 1$	$\pi(1-p)$	πp	π
			Marginal	$1-p$	p	1

Note: Please put a tick in the upper circle if you belong to one of the two circles or put a tick in the upper square if you belong to one of the two squares.

respectively, where $p \neq 0.5$,

$$\bar{y}^{\mathrm{C}} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\mathrm{C}} \quad \text{and} \quad \gamma \mathrel{\hat{=}} (1-\pi)(1-p) + \pi p.$$

To derive the sample size formula for the crosswise model, we consider the same one-sided hypotheses specified in (2.3). Similar to (2.5) and (2.6), we have

$$\text{Power (at } \pi_1) \doteq \Phi\left(\frac{\sqrt{n}(\pi_0 - \pi_1)|2p-1| - z_\alpha\sqrt{\gamma_0(1-\gamma_0)}}{\sqrt{\gamma_1(1-\gamma_1)}}\right)$$

and

$$n_{\mathrm{C}} = \left[\frac{z_\alpha\sqrt{\gamma_0(1-\gamma_0)} + z_\beta\sqrt{\gamma_1(1-\gamma_1)}}{(\pi_0 - \pi_1)(2p-1)}\right]^2, \tag{4.2}$$

where $\gamma_i \mathrel{\hat{=}} (1-\pi_i)(1-p) + \pi_i p,\ i = 0, 1$ and $\pi_1 < \pi_0$.

4.3. Numerical comparisons

Intuitively, the optimal degree of privacy protection is attained at $p = 0.5$. When p is either too small or too large, the privacy of respondents cannot be protected sufficiently. Therefore, investigators should choose a p within some interval

Table 4
The ratio n_C/n_P for testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi = \pi_1 < \pi_0$ with 5% level of significance and 80% power.

π_0	π_1	q	p							
			0.42	0.45	0.49	0.51	0.55	0.58	0.6	0.65
0.50	0.40	1/3	7.25	21.26	627.29	677.91	31.41	13.59	6.88	4.83
	0.35		6.90	20.28	601.88	651.62	30.36	13.18	6.71	4.71
	0.30		7.09	20.74	612.37	662.79	30.66	13.30	6.74	4.73
0.40	0.35	1/3	7.31	21.46	633.71	684.21	31.65	13.75	6.96	4.88
	0.30		6.91	20.32	602.58	652.99	30.52	13.21	6.73	4.73
	0.25		7.06	20.69	611.16	660.52	30.66	13.28	6.73	4.73
0.30	0.25	1/3	7.39	21.70	639.74	690.01	32.19	13.96	7.04	5.08
	0.20		6.95	20.44	607.55	657.15	30.66	13.35	6.76	4.76
	0.18		7.05	20.66	611.40	661.65	30.66	13.23	6.76	4.76
0.20	0.16	1/3	7.52	22.08	653.53	707.05	32.80	14.21	7.19	5.04
	0.13		6.95	20.47	608.51	659.95	30.76	13.38	6.80	4.78
	0.10		6.95	20.37	602.34	651.88	30.26	13.13	6.66	4.68
0.10	0.08	1/3	7.60	22.32	661.64	715.55	33.21	14.42	7.28	5.10
	0.06		6.97	20.56	611.39	663.91	30.95	13.48	6.85	4.82
	0.04		6.93	20.35	601.87	651.69	30.28	13.16	6.66	4.68
0.50	0.40	1/2	7.69	22.57	670.82	727.67	33.80	14.62	7.41	5.18
	0.35		7.01	20.67	615.52	668.21	31.19	13.58	6.89	4.88
	0.30		6.91	20.30	602.56	652.97	30.36	13.18	6.69	4.70
0.40	0.35	1/2	7.99	23.61	703.47	763.60	35.59	15.46	7.83	5.49
	0.30		7.09	21.02	628.77	683.86	32.04	13.96	7.12	5.02
	0.25		6.86	20.22	601.60	652.67	30.44	13.25	6.74	4.75
0.30	0.25	1/2	8.09	23.95	715.51	777.41	36.29	15.76	8.00	5.61
	0.20		7.14	21.16	634.44	691.19	32.40	14.12	7.22	5.08
	0.18		6.86	20.24	602.77	654.87	30.60	13.31	6.79	4.78
0.20	0.16	1/2	8.14	24.12	721.27	782.26	36.55	15.90	8.06	5.67
	0.13		7.15	21.25	637.33	693.32	32.55	14.22	7.26	5.11
	0.10		6.85	20.24	604.31	656.02	30.68	13.36	6.81	4.80
0.10	0.08	1/2	8.62	25.76	777.42	848.45	39.93	17.44	8.89	6.25
	0.06		7.33	21.92	662.83	724.48	34.21	15.00	7.69	5.43
	0.04		6.86	20.37	611.78	666.67	31.33	13.70	7.01	4.95
0.50	0.40	2/3	8.71	26.07	787.67	860.76	40.58	17.74	9.06	6.37
	0.35		7.36	22.06	667.62	730.73	34.56	15.16	7.78	5.50
	0.30		6.86	20.41	614.06	669.32	31.47	13.77	7.05	4.99
0.40	0.35	2/3	8.80	26.39	799.88	874.07	41.25	18.08	9.24	6.51
	0.30		7.40	22.20	673.07	737.29	34.89	15.34	7.87	5.57
	0.25		6.87	20.46	616.17	671.90	31.66	13.88	7.10	5.02
0.30	0.25	2/3	9.49	28.82	887.00	976.81	46.74	20.65	10.7	7.57
	0.20		7.67	23.23	713.53	785.31	37.58	16.62	8.61	6.13
	0.18		6.93	20.81	633.25	694.05	32.97	14.52	7.49	5.32
0.20	0.16	2/3	9.57	29.10	897.55	989.10	47.43	20.98	10.9	7.71
	0.13		7.70	23.35	718.36	791.09	37.91	16.78	8.70	6.20
	0.10		6.94	20.86	635.37	696.80	33.13	14.60	7.54	5.36
0.10	0.08	2/3	9.65	29.40	908.60	1002.9	48.17	21.33	11.1	7.88
	0.06		7.73	23.47	723.08	797.22	38.23	16.94	8.81	6.28
	0.04		6.95	20.91	637.59	699.97	33.30	14.69	7.59	5.40

around $p = 0.5$ except for the point $p = 0.5$ at which the MLE of π does not exist. In Table 4, we select several p's within $[0.42, 0.5) \cup (0.5, 0.65]$ and report the ratio n_C/n_P for testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi = \pi_1 < \pi_0$ with 5% level of significance and 80% power. From Table 4, we can see that when p is near 0.5, the parallel model is far more efficient than the crosswise model. For example, when $p = 0.49$ or $p = 0.51$, the efficiency of the parallel model is about 601–909 or 651–1003 times of that of the crosswise model.

4.4. *A theoretical justification*

The above observations are not surprising as we have the following theoretical results. Theorem 1 below identifies some conditions under which the parallel design is more efficient than the crosswise design. The corresponding proof is given in the Appendix.

Table 5
The triangular model and the corresponding cell probabilities.

Category	$W = 0$	$W = 1$	Category	$W = 0$	$W = 1$	Marginal
$Y = 0$	○	□	$Y = 0$	$(1-\pi)(1-p)$	$(1-\pi)p$	$1-\pi$
$Y = 1$	□	□	$Y = 1$	$\pi(1-p)$	πp	π
			Marginal	$1-p$	p	1

Note: Please put a tick in the circle if you belong to this circle or put a tick in the upper square if you belong to one of the three squares.

Theorem 1. *Let $\pi,\ p,\ q \in (0,\ 1)$. For the parallel model and the crosswise model, we have*

(i) *When $p = 1/3$, the parallel model is always more efficient than the crosswise model in the sense that $n_{\mathrm{P}} \le n_{\mathrm{C}}$, if one of the following three conditions is satisfied:*
(a) *$q = 1/2$ and $\pi \in (0,\ 1)$;*
(b) *$q \in (0,\ \min\{1/2,\ 1-\pi\})$ and $\pi \in (0,\ 1)$;*
(c) *$q \in (\max\{1/2,\ 1-\pi\},\ 1)$ and $\pi \in (0,\ 1)$.*

(ii) *When $1/3 < p < 1$ and $p \neq 1/2$, the parallel model is always more efficient than the crosswise model in the sense that $n_{\mathrm{P}} \le n_{\mathrm{C}}$, if one of the following five conditions is satisfied:*
(a) *$q = 1/2$ and $\pi \in (0,\ 1)$;*
(b) *$q > 1/2,\ p > 1/2$ and $\pi \in (0,\ 1-q) \cup (H(p,q),\ 1)$;*
(c) *$q < 1/2,\ p < 1/2$ and $\pi \in (0,\ 1-q) \cup (\min\{1, H(p,q)\},\ 1)$;*
(d) *$q > 1/2,\ p < 1/2$ and $\pi \in (0,\ \max\{0, H(p,q)\}) \cup (1-q,\ 1)$;*
(e) *$q < 1/2,\ p > 1/2$ and $\pi \in (0,\ H(p,q)) \cup (1-q,\ 1)$,*
where $H(p,q)$ is given by (A.1).

5. Comparison with the triangular model

5.1. The triangular model

Let Y and W have the same definition as in Section 2.1, where $p = \Pr(W = 1)$ and $\pi = \Pr(Y = 1)$. The interviewer may design the questionnaire in the format as shown on the left-hand side of Table 5 and ask the interviewee to put a tick in the circle (i.e., $\{Y = 0, W = 0\}$) if he/she belongs to this circle or put a tick in the upper square (i.e., $\{Y = 0, W = 1\}$) if he/she belongs to one of the three squares. Note that both $\{Y = 0, W = 0\}$ and $\{Y = 0, W = 1\}$ are non-sensitive. Thus, the sensitive class (i.e., $\{Y = 1\}$) is mixed with the non-sensitive subclass (i.e., $\{Y = 0, W = 1\}$). Yu et al. (2008) called this the *triangular model.*

5.2. Sample size formula for the triangular model

Let $Y_{\mathrm{obs}} = \{y_i^{\mathrm{T}} : i = 1, \ldots, n\}$ denote the observed data for the n respondents with $y_i^{\mathrm{T}} = 1$ if the i-th respondent put a tick in the upper square and $y_i^{\mathrm{T}} = 0$ if the i-th respondent put a tick in the circle. The likelihood function for π is

$$L_{\mathrm{T}}(\pi|Y_{\mathrm{obs}}) = \prod_{i=1}^{n}\Big[\pi + (1-\pi)p\Big]^{y_i^{\mathrm{T}}}\Big[(1-\pi)(1-p)\Big]^{1-y_i^{\mathrm{T}}}$$

so that the MLE of π and its variance are given by

$$\hat{\pi}_{\mathrm{T}} = \frac{\bar{y}^{\mathrm{T}} - p}{1-p} \quad \text{and} \quad \mathrm{Var}(\hat{\pi}_{\mathrm{T}}) = \frac{\lambda(1-\lambda)}{n(1-p)^2}, \tag{5.1}$$

respectively, where

$$\bar{y}^{\mathrm{T}} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\mathrm{T}} \quad \text{and} \quad \lambda \mathrel{\hat{=}} \pi + (1-\pi)p.$$

To derive the sample size formula for the triangular model, we consider the same one-sided hypotheses specified in (2.3). Similar to (2.5) and (2.6), we have

$$\text{Power (at } \pi_1) \doteq \Phi\left(\frac{\sqrt{n}(\pi_0 - \pi_1)(1-p) - z_\alpha\sqrt{\lambda_0(1-\lambda_0)}}{\sqrt{\lambda_1(1-\lambda_1)}}\right),$$

and

$$n_{\mathrm{T}} = \left[\frac{z_\alpha\sqrt{\lambda_0(1-\lambda_0)} + z_\beta\sqrt{\lambda_1(1-\lambda_1)}}{(\pi_0 - \pi_1)(1-p)}\right]^2, \tag{5.2}$$

where $\lambda_i \mathrel{\hat{=}} \pi_i + (1-\pi_i)p,\ i = 0, 1$ and $\pi_1 < \pi_0$.

5.3. Numerical comparisons

In Table 6, we select several values of p within the interval $[0.48, 0.72]$ and report the ratio $n_{\rm T}/n_{\rm P}$ for testing $H_0 : \pi = \pi_0$ against $H_1 : \pi = \pi_1 < \pi_0$ with 5% level of significance and 80% power. From Table 6, we can see that when $p = 0.58 \approx 7/12$ or 0.72, the efficiency of the parallel model is about 1–3 or 3–10 times of that of the triangular model. In particular, when $0.54 \le p \le 0.66$ (which is the optimal range such that the privacy of respondents is protected for the triangle model), the efficiency of the parallel design is about 1–6 times of that of the triangular design.

5.4. A theoretical justification

The above observations are further confirmed by the following theoretical result. Theorem 2 below identifies the conditions under which the parallel design is more efficient than the triangular design. The corresponding proof is given in the Appendix.

Theorem 2. *Let $\pi, p, q \in (0, 1)$. For the parallel model and the triangular model, we have*

(i) *When $p = 1/2$, the parallel model is always more efficient than the triangular model (in the sense that $n_{\rm P} \le n_{\rm C}$) for any $q \in (0, 1 - 2\pi]$ and $\pi \in (0, 1/2)$.*
(ii) *When $1/2 < p < 1$, the parallel model is always more efficient than the triangular model (in the sense that $n_{\rm P} \le n_{\rm C}$) for any $q \in (0, 1)$ and $0 < \pi < (1 - p)(1 - q)$.*

6. Sample size formula for the two-sample problem

In this section, we consider two independent surveys on the same sensitive question in two different populations or regions (labeled as $k = 1, 2$) by using the parallel design. The purpose here is to determine the sample sizes in each survey in order to compare the proportions (π_k) of subjects with the sensitive characteristic. For a fixed k, we define a binary random variable $Y_k^{\rm P}$ as follows:

$$Y_k^{\rm P} = \begin{cases} 1, & \text{if the two squares are connected,} \\ 0, & \text{if the two circles are connected.} \end{cases}$$

Let π_k denote the proportion of subjects with the sensitive characteristic in population k ($k = 1, 2$), then we have

$$\Pr\{Y_k^{\rm P} = 1\} = q_k(1 - p_k) + \pi_k p_k \quad \text{and} \quad \Pr\{Y_k^{\rm P} = 0\} = (1 - q_k)(1 - p_k) + (1 - \pi_k)p_k,$$

where $p_k = \Pr\{W_k = 1\}$ and $q_k = \Pr\{U_k = 1\}$ ($k = 1, 2$) are assumed to be known but neither p_1 and p_2 nor q_1 and q_2 are necessarily the same.

Suppose that there are a total of $n_1 + n_2$ individuals taking part in the survey, where n_1 respondents participating in the survey are from the first population and n_2 respondents are from the second population. Let $Y_{\rm obs} = \{y_{ik}^{\rm P} : i = 1, \dots, n_k;\ k = 1, 2\}$ denote the observed data. The likelihood function for π_1 and π_2 is given by

$$\begin{aligned} L(\pi_0, \pi_1 | Y_{\rm obs}) &= \prod_{k=1}^{2}\prod_{i=1}^{n_k}\Big[q_k(1 - p_k) + \pi_k p_k\Big]^{y_{ik}^{\rm P}}\Big[(1 - q_k)(1 - p_k) + (1 - \pi_k)p_k\Big]^{1 - y_{ik}^{\rm P}} \\ &= \prod_{k=1}^{2}\Big[q_k(1 - p_k) + \pi_k p_k\Big]^{n_k \bar{y}_k^{\rm P}}\Big[(1 - q_k)(1 - p_k) + (1 - \pi_k)p_k\Big]^{n_k(1 - \bar{y}_k^{\rm P})}, \end{aligned}$$

where $\bar{y}_k^{\rm P} = (1/n_k)\sum_{i=1}^{n_k} y_{ik}^{\rm P}$ denote the average number of respondents connecting the two squares in the k-th population. The resulting MLE of π_k and its variance are given by

$$\hat{\pi}_k = \frac{\bar{y}_k^{\rm P} - q_k(1 - p_k)}{p_k} \quad \text{and} \quad \mathrm{Var}(\hat{\pi}_k) = \frac{\Delta_k(1 - \Delta_k)}{n_k p_k^2},$$

where $\Delta_k \mathrel{\hat{=}} q_k(1 - p_k) + \pi_k p_k$. Thus,

$$\widehat{\mathrm{Var}}(\hat{\pi}_k) = \frac{\bar{y}_k^{\rm P}(1 - \bar{y}_k^{\rm P})}{n_k p_k^2}$$

is the MLE of $\mathrm{Var}(\hat{\pi}_k)$.

Now, we consider the following two-sided hypotheses

$$H_0 : \pi_1 = \pi_2 \quad \text{versus} \quad H_1 : \pi_1 \ne \pi_2.$$

Table 6
The ratio n_T/n_P for testing $H_0 : \pi = \pi_0$ versus $H_1 : \pi = \pi_1 < \pi_0$ with 5% level of significance and 80% power.

π_0	π_1	q	p							
			0.48	0.50	0.54	0.58	0.62	0.66	0.70	0.72
0.50	0.40	1/3	0.71	0.82	1.06	1.37	1.77	2.30	2.99	3.44
	0.35		0.73	0.84	1.09	1.42	1.83	2.37	3.10	3.55
	0.30		0.75	0.86	1.12	1.46	1.89	2.46	3.21	3.69
0.40	0.35	1/3	0.81	0.93	1.22	1.59	2.06	2.69	3.53	4.06
	0.30		0.83	0.95	1.25	1.63	2.13	2.79	3.67	4.21
	0.25		0.85	0.98	1.28	1.68	2.21	2.89	3.79	4.38
0.30	0.25	1/3	0.93	1.08	1.43	1.90	2.50	3.31	4.40	5.08
	0.20		0.96	1.11	1.48	1.96	2.60	3.45	4.58	5.31
	0.18		0.96	1.12	1.50	1.99	2.63	3.51	4.68	5.42
0.20	0.16	1/3	1.07	1.25	1.70	2.30	3.10	4.18	5.67	6.62
	0.13		1.09	1.28	1.74	2.35	3.18	4.30	5.86	6.86
	0.10		1.11	1.30	1.77	2.42	3.27	4.45	6.07	7.12
0.10	0.08	1/3	1.25	1.48	2.07	2.88	4.00	5.57	7.82	9.30
	0.06		1.26	1.50	2.10	2.93	4.09	5.72	8.05	9.60
	0.04		1.28	1.52	2.13	2.99	4.19	5.87	8.31	9.93
0.50	0.40	1/2	0.68	0.78	1.02	1.33	1.73	2.25	2.94	3.37
	0.35		0.70	0.80	1.04	1.36	1.77	2.31	3.01	3.47
	0.30		0.71	0.82	1.07	1.40	1.82	2.38	3.13	3.59
0.40	0.35	1/2	0.75	0.87	1.14	1.49	1.95	2.56	3.38	3.89
	0.30		0.76	0.88	1.16	1.53	2.01	2.63	3.49	4.02
	0.25		0.78	0.90	1.19	1.56	2.05	2.72	3.59	4.17
0.30	0.25	1/2	0.83	0.97	1.29	1.71	2.28	3.03	4.05	4.71
	0.20		0.85	0.98	1.32	1.76	2.34	3.13	4.19	4.89
	0.18		0.85	0.99	1.33	1.78	2.37	3.17	4.26	4.97
0.20	0.16	1/2	0.91	1.07	1.46	1.98	2.68	3.64	4.98	5.85
	0.13		0.92	1.08	1.48	2.01	2.73	3.72	5.11	6.01
	0.10		0.93	1.09	1.50	2.05	2.79	3.81	5.25	6.18
0.10	0.08	1/2	1.00	1.19	1.66	2.31	3.22	4.52	6.38	7.63
	0.06		1.01	1.20	1.68	2.34	3.28	4.60	6.52	7.81
	0.04		1.01	1.21	1.70	2.37	3.33	4.69	6.68	8.01
0.50	0.40	2/3	0.70	0.80	1.03	1.34	1.74	2.25	2.94	3.37
	0.35		0.71	0.81	1.05	1.37	1.77	2.31	3.01	3.45
	0.30		0.72	0.82	1.07	1.39	1.82	2.35	3.09	3.54
0.40	0.35	2/3	0.75	0.86	1.12	1.46	1.91	2.50	3.30	3.80
	0.30		0.75	0.87	1.14	1.49	1.95	2.56	3.38	3.91
	0.25		0.76	0.88	1.15	1.52	1.99	2.62	3.46	4.01
0.30	0.25	2/3	0.80	0.92	1.23	1.63	2.15	2.86	3.84	4.46
	0.20		0.81	0.93	1.24	1.65	2.21	2.94	3.95	4.61
	0.18		0.81	0.94	1.25	1.67	2.22	2.97	4.00	4.66
0.20	0.16	2/3	0.85	0.99	1.34	1.81	2.44	3.32	4.54	5.34
	0.13		0.85	0.99	1.35	1.83	2.48	3.37	4.64	5.45
	0.10		0.86	1.00	1.36	1.85	2.52	3.43	4.73	5.60
0.10	0.08	2/3	0.89	1.05	1.46	2.02	2.81	3.92	5.53	6.62
	0.06		0.89	1.06	1.47	2.04	2.84	3.91	5.63	6.74
	0.04		0.90	1.06	1.48	2.06	2.87	4.03	5.73	6.87

Let $\text{SE}_+ = [\sum_{k=1}^{2} \text{Var}(\hat{\pi}_k)]^{1/2}$ and $\widehat{\text{SE}}_+ = [\sum_{k=1}^{2} \widehat{\text{Var}}(\hat{\pi}_k)]^{1/2}$ denote the MLE of SE_+. Then the null hypothesis H_0 will be rejected at the α level of significance if

$$\left| \frac{\hat{\pi}_1 - \hat{\pi}_2}{\widehat{\text{SE}}_+} \right| > z_{\alpha/2}.$$

Under the alternative hypothesis H_1, i.e., $\pi_1 - \pi_2 \neq 0$, the power of the two-sided test is approximately given by

$$\Phi\left(\frac{|\pi_1 - \pi_2| - z_{\alpha/2} \cdot \widehat{\text{SE}}_+}{\text{SE}_+} \right),$$

which can be further approximated by Chow, Shao, and Wang (2003)

$$\Phi\left(\frac{|\pi_1 - \pi_2|}{\text{SE}_+} - z_{\alpha/2} \right).$$

Table 7
MLEs of π, estimated standard errors and 95% CIs of π for three models for the sexual practice data.

Model	$\hat{\pi}$	$\widehat{\text{SE}}(\hat{\pi})$	95% CI of π	Width of the 95% CI
Parallel model	0.53049	0.020703	[0.48991, 0.57106]	0.081155
Crosswise model	0.52974	0.042233	[0.44696, 0.61251]	0.165552
Triangular model	0.52895	0.030736	[0.46870, 0.58919]	0.120485

Consequently, to achieve a desired power of $1-\beta$, we need to solve the following equation:

$$\frac{|\pi_1-\pi_2|}{\text{SE}_+} - z_{\alpha/2} = z_{\beta}. \tag{6.1}$$

Let $\rho = n_1/n_2$ be known. Then, from (6.1), we have

$$n_1 = \rho n_2, \quad \text{and} \tag{6.2}$$

$$n_2 = \frac{(z_{\alpha/2}+z_{\beta})^2}{(\pi_1-\pi_2)^2}\left[\frac{\Delta_1(1-\Delta_1)}{\rho p_1^2} + \frac{\Delta_2(1-\Delta_2)}{p_2^2}\right]. \tag{6.3}$$

7. An example

Monto (2001) reported a sexual practice study carried in three Western cities (San Francisco, Las Vegas and Portland of Oregon) of the United States. In this investigation, there are 343 individuals graduating at most from some high school and 927 individuals receiving at least some college training. In addition, it was also observed that 593 respondents have no more than one sexual partner and 668 respondents have no less than two sexual partners. The investigators would like to estimate the proportion of persons with more than one sexual partners in a population.

We first define $W = 1$ if the birthday of the respondent is between May to December and $W = 0$ if the birthday of the respondent is between January to April, and let $p = \Pr(W = 1) \approx 8/12 = 2/3$. We then define $U = 1$ if the respondent receives at least some college training and $U = 0$ if the respondent graduates at most from some high school, and let $q = \Pr(U = 1) = 927/(343+927) \approx 0.73$. Finally, we define $Y = 1$ if the respondent has at least two sexual partners and $Y = 0$ otherwise. For the purpose of illustration, we assume that W, U and Y are mutually independent although we have noted the possible association between U and Y.

With $p = 2/3$ and $q = 0.73$, the survey with the parallel design will yield 754 lines (i.e., $n\bar{y}^{\text{P}} = \sum_{i=1}^n y_i^{\text{P}} = 927 \times (1-p) + 668 \times p \approx 754$) connecting the two squares and 509 lines (i.e., $n - n\bar{y}^{\text{P}} = 343 \times (1-p) + 593 \times p \approx 509,\ n = 1263$) connecting the two circles. If the crosswise design is employed, it can be observed that 643 ticks (i.e., $n\bar{y}^{\text{C}} = \sum_{i=1}^n y_i^{\text{P}} = 593 \times (1-p) + 668 \times p \approx 643$) will be put in the upper circle and 618 ticks (i.e., $n - n\bar{y}^{\text{C}} = 593 \times p + 668 \times (1-p) \approx 618,\ n = 1261$) will be put in the upper square. Finally, the survey with the triangular design will lead to 1063 ticks (i.e., $n\bar{y}^{\text{T}} = \sum_{i=1}^n y_i^{\text{T}} = 593 \times p + 668 = 1063$) will be put in the circle and 198 ticks (i.e., $n - n\bar{y}^{\text{T}} = 593 \times (1-p) \approx 198,\ n = 1261$) will be put in the upper square.

Table 7 reports MLEs of π based on (2.1), (4.1) and (5.1), estimated standard errors and 95% *confidence intervals* (CIs) of π for the three models. From Table 7, we can see that the width of the 95% CI of π for the parallel model is the shortest among the three models.

To illustrate the proposed methods, we now determine the sample sizes required in order to guarantee 80% power with 0.05 level of significance by using the one-sided test for testing $\pi_0 = 0.65$ against $\pi_1 = 0.55$. Using the sample size formulae (2.6), (4.2) and (5.2), we obtain $n_{\text{P}} = 314$, $n_{\text{C}} = 1382$ and $n_{\text{T}} = 618$, which are required sample sizes for the parallel, crosswise and triangular designs, respectively.

Finally, we estimate how many subjects are required for comparing the proportions that people having more than one sexual partners in a population between two regions with 80% power and 0.05 level of significance using the two-sided test for testing $\pi_1 = \pi_2$ against $\pi_1 \neq \pi_2$. Assume that true proportions with sensitive character in the two regions are $\pi_1 = 0.68$ and $\pi_2 = 0.75$, respectively. Using the parallel design with $p_1 = 0.55$, $p_2 = 0.6$, $q_1 = 0.5$ and $q_2 = 0.4$, the sample sizes with $\rho = 1$ (equal allocation) are given by $n_1 = n_2 = 2306$ via (6.2) and (6.3) while the desired sample sizes are $n_1 = n_2 = 50\,054$ for the crosswise model and $n_1 = n_2 = 2849$ for the triangular model.

8. Discussion

In this paper, we derived the sample size formulae for the non-randomized parallel design based on the power analysis method for both the one- and two-sample problems. We theoretically compared the sample sizes needed for the parallel design with those required for the crosswise and triangular designs (see Theorems 1 and 2). Numerical comparisons are shown in Tables 4 and 6, from which we can observe significant improvement in efficiency.

Unlike the non-randomized crosswise design, the parallel design can be applied to the situation where $p = 0.5$ at which the privacy can be highly protected. More importantly, the parallel model can be applied to the case where both $\{Y = 0\}$

and $\{Y = 1\}$ are sensitive (cf. Table 1) while the crosswise and triangular models require that $\{Y = 0\}$ is non-sensitive. Therefore, we recommend to use the parallel design in surveys with sensitive questions.

Recently, Liu and Tian (2013a) considered multi-category parallel models in the design of surveys with sensitive questions, and Liu and Tian (2013b) proposed a variant of the parallel model for sample surveys with sensitive characteristics. Sample size determination in the two models is our possible research interest in the future.

Acknowledgment

GL Tian's research was fully supported by a grant (HKU 779210M) from the Research Grant Council of the Hong Kong Special Administrative Region.

Appendix

To prove Theorem 1, we first present a lemma.

Lemma 1. *Let* $1/3 < p < 1,\ p \neq 1/2, 0 < q < 1$ *and* $q \neq 1/2$. *Define*

$$H(p, q) \hat{=} \frac{p - q + pq}{3p - 1}. \tag{A.1}$$

We have the following conclusions:

(i) *If* $(2p-1)(2q-1) > 0$, *then* $1 - q < H(p, q)$.
(ii) *If* $(2p-1)(2q-1) < 0$, *then* $1 - q > H(p, q)$.

Proof. (i) If $(2p-1)(2q-1) > 0$, then we have $2p - 1 - 2q(2p-1) < 0$, or

$$(1-q)(3p-1) < p - q + pq.$$

Since $p > 1/3$, i.e., $3p - 1 > 0$, we immediately obtain

$$1 - q < \frac{p - q + pq}{3p - 1} = H(p, q).$$

Similarly, we can prove (ii). □

Proof of Theorem 1. From (2.6) and (4.2), we have

$$\frac{n_{\mathrm{P}}}{n_{\mathrm{C}}} = \left(\frac{2p-1}{p}\right)^2 \cdot \left[\frac{z_\alpha\sqrt{\delta_0(1-\delta_0)} + z_\beta\sqrt{\delta_1(1-\delta_1)}}{z_\alpha\sqrt{\gamma_0(1-\gamma_0)} + z_\beta\sqrt{\gamma_1(1-\gamma_1)}}\right]^2, \tag{A.2}$$

where

$$\delta_i = q(1-p) + \pi_i p \quad \text{and} \quad \gamma_i = (1-\pi_i)(1-p) + \pi_i p, \quad i = 0, 1.$$

Note that when $1/3 \le p < 1$ and $p \neq 1/2$, we always have

$$p^2 \ge (2p-1)^2,$$

i.e., the first term on the right-hand side of (A.2) is less than or equal to 1. To obtain $n_{\mathrm{P}} \le n_{\mathrm{C}}$, it suffices to show that $\delta(1-\delta) \le \gamma(1-\gamma)$ or equivalently

$$\Big[q(1-p) + \pi p\Big]\Big[(1-q)(1-p) + (1-\pi)p\Big] \le \Big[(1-\pi)(1-p) + \pi p\Big]\Big[(1-\pi)p + \pi(1-p)\Big]. \tag{A.3}$$

After some simplifications, it can be showed that (A.3) is equivalent to

$$h_{\mathrm{C}}(\pi|p, q) \hat{=} (3p-1)\pi^2 + (1 - 4p + 2pq)\pi + (1-q)(p - q + pq) \ge 0. \tag{A.4}$$

(i) When $p = 1/3$, (A.4) reduces to

$$(2q-1)(q - 1 + \pi) \ge 0. \tag{A.5}$$

(a) If $q = 1/2$, then (A.5) is always true for any $\pi \in (0,\ 1)$.
(b) If $0 < q < 1/2$, then (A.5) is equivalent to $q < 1 - \pi$. Therefore, (A.5) is always true for any $0 < q < \min\{1/2, 1-\pi\}$ and $\pi \in (0, 1)$.
(c) If $1/2 < q < 1$, then (A.5) is equivalent to $q > 1 - \pi$. Hence, (A.5) is always true for any $\max\{1/2,\ 1-\pi\} < q < 1$ and $\pi \in (0,\ 1)$.

(ii) When $1/3 < p < 1$ and $p \neq 1/2$, we always have $3p - 1 > 0$. Note that the discriminant for the quadratic function $h_{\mathrm{C}}(\pi|p, q)$ defined in (A.4) is given by

$$\begin{aligned} \Delta_{\mathrm{C}} &= (1-4p+2pq)^2 - 4(3p-1)(1-q)(p-q+pq) \\ &= (2p-1)^2(2q-1)^2. \end{aligned}$$

(a) If $q = 1/2$, then $\Delta_{\mathrm{C}} = 0$. Hence, $h_{\mathrm{C}}(\pi|p, q) \geq 0$ (i.e., (A.4)) is true for all $\pi \in (0, 1)$.
If $q \neq 1/2$, then $\Delta_{\mathrm{C}} > 0$. Hence, $h_{\mathrm{C}}(\pi|p, q) > 0$ for any $\pi \in (0, \pi_{\mathrm{C,L}}) \cup (\pi_{\mathrm{C,U}}, 1)$, where

$$\pi_{\mathrm{C,L}} = \max\left\{0, \ \frac{-(1-4p+2pq) - |(2p-1)(2q-1)|}{2(3p-1)}\right\} \tag{A.6}$$

and

$$\pi_{\mathrm{C,U}} = \min\left\{1, \ \frac{-(1-4p+2pq) + |(2p-1)(2q-1)|}{2(3p-1)}\right\}. \tag{A.7}$$

(b) If $q > 1/2$ and $p > 1/2$, then (A.6) and (A.7) can be simplified as

$$\begin{aligned} \pi_{\mathrm{C,L}} &= \max\left\{0, \ \frac{-(1-4p+2pq) - (2p-1)(2q-1)}{2(3p-1)}\right\} \\ &= \max\{0, \ 1-q\} \\ &= 1-q, \quad \text{and} \end{aligned}$$

$$\begin{aligned} \pi_{\mathrm{C,U}} &= \min\left\{1, \ \frac{-(1-4p+2pq) + (2p-1)(2q-1)}{2(3p-1)}\right\} \\ &= \min\left\{1, \ \frac{p-q+pq}{3p-1}\right\} \\ &= \frac{p-q+pq}{3p-1} \\ &\overset{(A.1)}{=} H(p,q), \end{aligned} \tag{A.8}$$

respectively, where (A.8) can be proved from

$$\begin{aligned} \frac{1}{2} < p < 1 \quad \text{and} \quad q > 0 &\Rightarrow q > 0 > \frac{1-2p}{1-p} \\ &\Rightarrow q - pq > 1 - 2p \\ &\Rightarrow 3p - 1 > p - q + pq \\ &\Rightarrow 1 > \frac{p-q+pq}{3p-1}. \end{aligned}$$

Finally, from Lemma 1(i), we have $1 - q < H(p, q)$; that is $\pi_{\mathrm{C,L}} < \pi_{\mathrm{C,U}}$.

(c) If $q < 1/2$ and $p < 1/2$, then (A.6) and (A.7) become

$$\pi_{\mathrm{C,L}} = 1-q \quad \text{and} \quad \pi_{\mathrm{C,U}} = \min\{1, \ H(p,q)\},$$

respectively. However, we now cannot simplify $\pi_{\mathrm{C,U}}$. On the one hand, from Lemma 1(i), we have $1 - q < H(p, q)$. On the other hand, $1 - q < 1$. Hence,

$$\pi_{\mathrm{C,L}} = 1-q < \min\{1, \ H(p,q)\} = \pi_{\mathrm{C,U}}.$$

(d) If $q > 1/2$ and $p < 1/2$, then (A.6) and (A.7) become

$$\pi_{\mathrm{C,L}} = \max\{0, \ H(p,q)\} \quad \text{and} \quad \pi_{\mathrm{C,U}} = 1-q,$$

respectively. However, we now cannot simplify $\pi_{\mathrm{C,L}}$. On the one hand, from Lemma 1(ii), we have $1 - q > H(p, q)$. On the other hand, $1 - q > 0$. Hence,

$$\pi_{\mathrm{C,L}} = \max\{0, \ H(p,q)\} < 1-q = \pi_{\mathrm{C,U}}.$$

(e) If $q < 1/2$ and $p > 1/2$, then (A.6) and (A.7) become

$$\pi_{\mathrm{C,L}} = \max\{0, \ H(p,q)\} \quad \text{and} \quad \pi_{\mathrm{C,U}} = 1-q,$$

respectively. Now, we have $\pi_{\mathrm{C,L}} = H(p, q)$, which can be proved from

$$\begin{aligned} p > \frac{1}{2} > q > 0 &\Rightarrow p - q > 0 \\ &\Rightarrow p - q + pq > 0 \end{aligned}$$

$$\begin{aligned}&\Rightarrow \frac{p-q+pq}{3p-1} > 0 \quad (\text{as } p > 1/3)\\ &\Rightarrow H(p,q) > 0.\end{aligned}$$

On the one hand, from Lemma 1(ii), we have $1-q > H(p,q)$. On the other hand, $1-q > 0$. Hence, $\pi_{\mathrm{C,L}} = \max\{0,\ H(p,q)\} < 1-q = \pi_{\mathrm{C,U}}$. □

Proof of Theorem 2. From (2.6) and (5.2), we have

$$\frac{n_{\mathrm{P}}}{n_{\mathrm{T}}} = \left(\frac{1-p}{p}\right)^2 \cdot \left(\frac{z_\alpha\sqrt{\delta_0(1-\delta_0)} + z_\beta\sqrt{\delta_1(1-\delta_1)}}{z_\alpha\sqrt{\lambda_0(1-\lambda_0)} + z_\beta\sqrt{\lambda_1(1-\lambda_1)}}\right)^2, \tag{A.9}$$

where

$$\delta_i = q(1-p) + \pi_i p \quad \text{and} \quad \lambda_i = \pi_i + (1-\pi_i)p, \quad i = 0, 1.$$

Note that when $1/2 \le p < 1$, we always have

$$p^2 \ge (1-p)^2,$$

i.e., the first term on the right-hand side of (A.9) is less than or equal to 1. To obtain $n_{\mathrm{P}} \le n_{\mathrm{T}}$, it suffices to show that $\delta(1-\delta) \le \lambda(1-\lambda)$, or equivalently

$$\Big[q(1-p) + \pi p\Big]\Big[(1-q)(1-p) + (1-\pi)p\Big] \le \Big[\pi + (1-\pi)p\Big](1-\pi)(1-p). \tag{A.10}$$

After some simplifications, we can show that (A.10) is equivalent to

$$h_{\mathrm{T}}(\pi|p,q) \hat{=} (2p-1)\pi^2 + [1-2p-2p(1-p)(1-q)]\pi + (1-p)(1-q)(p-q+pq) \ge 0. \tag{A.11}$$

(i) When $p = 1/2$, (A.11) reduces to

$$(1-q)(-2\pi + 1 - q)/4 \ge 0. \tag{A.12}$$

Hence, for any $q \in (0,\ 1-2\pi]$ and any $\pi \in (0,\ 1/2)$, (A.12) is true.

(ii) When $1/2 < p < 1$, we always have $2p-1 > 0$. Note that the discriminant for the quadratic function $h_{\mathrm{T}}(\pi|p,q)$ defined in (A.11) is given by

$$\begin{aligned}\Delta_{\mathrm{T}} &= [(1-2p) - 2p(1-p)(1-q)]^2 - 4(2p-1)(1-p)(1-q)(p-q+pq)\\ &= [p^2 + (1-p)^2(1-2q)]^2.\end{aligned}$$

We can show that $p^2 + (1-p)^2(1-2q) > 0$. In fact, from

$$\begin{aligned}\frac{1}{2} < p < 1 \text{ and } q < 1 &\Rightarrow p^2 > (1-p)^2 \text{ and } q < 1\\ &\Rightarrow \frac{p^2 + (1-p)^2}{2(1-p)^2} > 1 > q\\ &\Rightarrow p^2 + (1-p)^2 > 2q(1-p)^2\\ &\Rightarrow p^2 + (1-p)^2(1-2q) > 0.\end{aligned}$$

In other words, $\Delta_{\mathrm{T}} > 0$ for any $q \in (0, 1)$. Hence, $h_{\mathrm{T}}(\pi|p,q) > 0$ for any $q \in (0, 1)$ and $\pi \in (0,\ \pi_{\mathrm{T,L}}) \cup (\pi_{\mathrm{T,U}},\ 1)$, where

$$\begin{aligned}\pi_{\mathrm{T,L}} &= \max\left\{0,\ \frac{-[1-2p-2p(1-p)(1-q)] - p^2 - (1-p)^2(1-2q)}{2(2p-1)}\right\}\\ &= \max\{0,\ (1-p)(1-q)\}\\ &= (1-p)(1-q) < 1,\end{aligned}$$

and

$$\begin{aligned}\pi_{\mathrm{T,U}} &= \min\left\{1,\ \frac{-[1-2p-2p(1-p)(1-q)] + p^2 + (1-p)^2(1-2q)}{2(2p-1)}\right\}\\ &= \min\left\{1,\ \frac{p-q+pq}{2p-1}\right\}.\end{aligned}$$

In the follows, we show that $\pi_{\mathrm{T,U}} = 1$. From

$$\begin{aligned}\frac{1}{2} < p < 1 \text{ and } q < 1 &\Rightarrow 2p - 1 > 0 \text{ and } (1-p)(1-q) > 0\\ &\Rightarrow 2p - 1 > 0 \text{ and } 2p - 1 < p - q + pq\\ &\Rightarrow 1 < \frac{p - q + pq}{2p - 1},\end{aligned}$$

which implies $\pi_{\mathrm{T,U}} = 1$. In a summary, $h_{\mathrm{T}}(\pi\,|p, q) > 0$ for any $q \in (0, 1)$ and $0 < \pi < (1-p)(1-q)$. □

References

Chow, S. C., Shao, J., & Wang, H. S. (2003). *Sample size calculations in clinical research*. Boca Raton: Chapman & Hall/CRC.

Liu, Y., & Tian, G. L. (2013a). Multi-category parallel models in the design of surveys with sensitive questions. *Statistics and its Interface*, *6*(1), 137–149.

Liu, Y., & Tian, G. L. (2013b). A variant of the parallel model for sample surveys with sensitive characteristics. *Computational Statistics & Data Analysis*, *67*, 115–135.

Mangat, N. S. (1994). An improved randomized response strategy. *Journal of Royal Statistical Society. Series B*, *56*(1), 93–95.

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

Monto, M. A. (2001). Prostitution and fellatio. *The Journal of Sex Research*, *38*(2), 140–145.

Singh, R., & Mangat, N. S. (1996). *Elements of survey sampling*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Tan, M., Tian, G. L., & Tang, M. L. (2009). Sample surveys with sensitive questions: a non-randomized response approach. *The American Statistician*, *63*(1), 9–16.

Tang, M. L., Tian, G. L., Tang, N. S., & Liu, Z. Q. (2009). A new non-randomized multi-category response model for surveys with a single sensitive question: design and analysis. *Journal of the Korean Statistical Society*, *38*, 339–349.

Tian, G. L. (2013). A new non-randomized response model: the parallel model. *Statistica Neerlandica* (in revision).

Tian, G. L., Tang, M. L., Liu, Z. Q., Tan, M., & Tang, N. S. (2011). Sample size determination for the non-randomized triangular model for sensitive questions in a survey. *Statistical Methods in Medicine Research*, *20*(3), 159–173.

Tian, G. L., Yu, J. W., Tang, M. L., & Geng, Z. (2007). A new non-randomized model for analyzing sensitive questions with binary outcomes. *Statistics in Medicine*, *26*(23), 4238–4252.

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

Yu, J. W., Tian, G. L., & Tang, M. L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika*, *67*, 251–263.