APPLICATIONES MATHEMATICAE
50,1 (2023), pp. 21–34

Stanisław Jaworski and Wojciech Zieliński (Warszawa)

THE OPTIMAL SAMPLE SIZE IN THE CROSSWISE MODEL FOR SENSITIVE QUESTIONS

Abstract. For interval estimation of the fraction of the population with a stigmatizing characteristic, the nonrandomized response model proposed by Tian, Yu, and Geng (2007) is considered. The most common method for constructing a confidence interval (c.i.) is through the application of the Central Limit Theorem. Unfortunately, such c.i.'s do not maintain the prescribed confidence level, in contradiction to Neyman's (1934) definition of c.i. In the present paper, the exact c.i. for this fraction is constructed, i.e., the c.i. which keeps the given confidence level. The length of the proposed c.i. depends on the given probability of a positive answer to the neutral question, and on the sample size. For such c.i.'s, the probability of a positive answer to the neutral question is established with respect to the given limit on privacy protection of the interviewee, and the optimal sample size for obtaining the c.i. of a given length is derived.

1. Introduction. We consider the estimation of the percentage of the population who have committed socially stigmatizing misdeeds such as corruption, tax frauds, illegal work (black market), drug uses, violence against children and others. The proposed c.i.'s for the percentage are asymptotic (Yu et al., 2008). These confidence intervals are not c.i. in the Neyman sense (Neyman, 1934, p. 562); they do not maintain prescribed confidence level. In what follows, the finite sample size c.i. is proposed. Its construction is based on the distribution of the Maximum Likelihood Estimator of the percentage. We consider only the crosswise model proposed by Yu et al. (2008).

2020 *Mathematics Subject Classification*: Primary 62F25; Secondary 62P20.
Key words and phrases: sensitive questions, nonrandomized response model, exact confidence interval.
Received 21 July 2023; revised 21 October 2023.
Published online 13 December 2023.

DOI: 10.4064/am2486-11-2023

An important practical problem is to determine the sample size that guarantees both a certain precision of estimation and protection of respondent privacy. To solve this issue, we impose conditions on the level of privacy protection (so that respondents can feel safe answering the questionnaire) and set a minimum sample size that meets the given precision of estimation. In this paper, we propose two criteria for the given precision of estimation based on the length of the c.i. Importantly, we do not use approximate c.i.'s, but exact ones, since we take the view that the estimation error should be completely controlled. We also take advantage of the fact that usually sensitive questions involve rare phenomena. The solution given is a natural one for the researcher who determines the confidence level, the precision of the estimate and the level of protection for the respondent.

Mathematically, let Y be a random variable such that

$$P\{Y=1\}=\pi=1-P\{Y=0\}.$$

The r.v. takes on the value 1 when the answer to the sensitive question is *YES* , and the value 0 otherwise. The number $\pi \in (0,1)$ is the probability of the positive answer to the sensitive question, i.e., $\pi \cdot 100\%$ is the percentage of interest. We want to estimate the probability π, i.e., we are going to construct a confidence interval for π.

Let $Y_1,\ldots,Y_n$ be a sample. The statistical model for the sample is

$$(\{0,1,\ldots,n\},\{\mathrm{Bin}(n,\pi),\pi\in(0,1)\}),$$

where $\mathrm{Bin}(\cdot,\cdot)$ denotes the binomial distribution.

The difficulty which arises is that the random variables $Y_1,\ldots,Y_n$ are not observable. Answers to the sensitive question are "hidden" by asking a "neutral" question, which is answered *YES* or *NO*. It is assumed that the "neutral" question is independent of the sensitive question. In a questionnaire two questions are asked: a sensitive and a neutral one. But only one answer is registered and the interviewer does not know which of the two questions the interviewee answered.

The first method of obscuring the answer to a sensitive question was proposed by Warner in 1965. His method consists in randomization of answers. This randomization is done by the respondent, and the interviewer does not know what the answer to the sensitive question is. This model has been extended in different ways (Horvitz et al., 1967; Greenberg et al., 1969; Raghavarao, 1978; Franklin, 1989; Arnab et al., 2019; Arnab, 1990, 1996; Kuk, 1990; Rueda et al., 2015).

Tian et al. 2007 proposed a nonrandomized response model (NRR). Their idea was to ask two questions simultaneously: one sensitive and one neutral. This model was extended to other, similar approaches (Yu et al., 2008; Tan et al., 2009; Tian, 2014).

In Section 2, we give the method of construction of a new confidence interval for π. In Section 3, we recall the construction of asymptotic confidence intervals. We also discuss the probability of the coverage of the confidence intervals presented. In Section 4, we present the method of sample size selection, which plays the main role in our paper. In Section 5, concluding remarks are given.

2. Confidence interval in the crosswise model.

In the crosswise model (CM), respondents are presented with two questions simultaneously, one neutral and one sensitive. They are instructed to report "1" only if the answers to both questions are the same, i.e., the observable variable in this model is Z, where

$$Z = \begin{cases} 1 & \text{if both answers are } \textit{YES} \text{ or both } \textit{NO}, \\ 0 & \text{otherwise.} \end{cases}$$

The answers of n respondents may be treated as realizations of the binomial distribution with parameters (n, ϱ), where ϱ is the probability of receiving an outcome 1 of the Z variable. Assume that the questions asked are independent and the probability of the *YES* answer to the sensitive question is π (and q for the neutral question). It is assumed that q is known. Hence, in the CM model,

$$\varrho = q\pi + (1-q)(1-\pi) = (2q-1)\pi + (1-q),$$

so

$$\pi = \frac{\varrho - (1-q)}{2q-1}.$$

Without loss of generality we can assume that $q < 0.5$.

Let $Z_1, \ldots, Z_n$ be a sample. The MLE of ϱ is $\hat{\varrho} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. The distribution of $n\hat{\varrho}$ is $\mathrm{Bin}(n;\varrho)$.

The MLE of π has the form

$$\hat{\pi}_{\mathrm{CM}} = \max\left\{\min\left\{\frac{\hat{\varrho}-(1-q)}{2q-1}, 1\right\}, 0\right\}.$$

Let $\mathrm{Bin}(\cdot, n; \varrho)$ denote the CDF of the binomial distribution with probability of success equal to ϱ, and let $B(a, b; \cdot)$ denote the CDF of the Beta distribution with parameters (a, b).

In the derivation of the pdf of $\hat{\pi}_{\mathrm{CM}}$, the following known relationship will be applied: if ξ is a binomial random variable with parameters (n, ρ) then

$$P_\rho\{\xi \le x\} = \sum_{i=0}^{x} \binom{n}{i} \rho^i (1-\rho)^{n-i} = B(n-x, x+1; 1-\rho).$$

Let $\mathcal{X}_q$ be the sample space of the estimator $\hat{\pi}_{\mathrm{CM}}$:

$$\mathcal{X}_q = \left\{x : x = \max\left\{\min\left\{\frac{i/n-(1-q)}{2q-1}, 1\right\}, 0\right\},\ i = 0, 1, \ldots, n\right\}.$$

The pdf of the distribution of $\hat{\pi}_{\mathrm{CM}}$ is

$$\begin{aligned}(1)\qquad & P_\pi\{\hat{\pi}_{\mathrm{CM}} = x\} \\ &= \begin{cases} P_\pi\{n\hat{\varrho} \geq \lceil(1-q)n\rceil\} & \text{for } x = 0, \\ \binom{n}{\lceil u\rceil}((2q-1)\pi+(1-q))^{\lceil u\rceil}(1-(2q-1)\pi-(1-q))^{n-\lceil u\rceil} & \text{for } 0 < x < 1, \\ P_\pi\{n\hat{\varrho} \leq \lfloor qn\rfloor\} & \text{for } x = 1, \end{cases} \\ &= \begin{cases} B\left(\lceil n(1-q)\rceil, n-\lceil n(1-q)\rceil+1; (2q-1)\pi+(1-q)\right) & \text{for } x = 0, \\ \binom{n}{\lceil u\rceil}((2q-1)\pi+(1-q))^{\lceil u\rceil}(1-(2q-1)\pi-(1-q))^{n-\lceil u\rceil} & \text{for } 0 < x < 1, \\ 1-B\left(\lfloor nq\rfloor+1, n-\lfloor nq\rfloor; (2q-1)\pi+(1-q)\right) & \text{for } x = 1, \end{cases}\end{aligned}$$

where $x \in \mathcal{X}_q$ and $u = n(x(2q-1)+(1-q))$. Here $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the standard ceiling and floor functions.

The CDF of $\hat{\pi}_{\mathrm{CM}}$ is

$$\begin{aligned} F_\pi(x) &= P_\pi\{\hat{\pi}_{\mathrm{CM}} \leq x\} \\ &= \begin{cases} B(\lceil u\rceil, n-\lceil u\rceil+1; (2q-1)\pi+(1-q)) & \text{for } 0 \leq x < 1, \\ 1 & \text{for } x = 1. \end{cases}\end{aligned}$$

Since the distribution of $\hat{\pi}_{\mathrm{CM}}$ is discrete, we have

$$P_\pi\{\hat{\pi}_{\mathrm{CM}} < x\} = \begin{cases} 0 & \text{for } x = 0, \\ B(\lfloor u\rfloor+1, n-\lfloor u\rfloor; (2q-1)\pi+(1-q)) & \text{for } 0 < x \leq 1. \end{cases}$$

Note that the family $\{F_\pi : \pi \in [0,1]\}$ is stochastically ordered, i.e.,

$$F_{\pi_1}(\cdot) \geq F_{\pi_2}(\cdot) \quad \text{for } \pi_1 < \pi_2.$$

Let δ be a given confidence level and let x be the observed value of the estimator $\hat{\pi}_{\mathrm{CM}}$. The equitailed confidence interval $(\pi_L(x;\delta); \pi_R(x;\delta))$ for π is defined as

$$\begin{cases} \pi_L(x;\delta) = \arg\inf_\pi P_\pi\{\hat{\pi}_{\mathrm{CM}} \leq x\} \geq \frac{1+\delta}{2}, \\ \pi_R(x;\delta) = \arg\sup_\pi P_\pi\{\hat{\pi}_{\mathrm{CM}} < x\} \leq \frac{1-\delta}{2}. \end{cases}$$

The function $\pi \mapsto F_\pi(x)$ for a given x has two jumps, at $\pi = 0$ and at $\pi = 1$. The jump at $\pi = 0$ equals $1 - B(u, n-u+1; 1-q)$, and the jump at $\pi = 1$

is $B(u, n-u+1; q)$. Hence the confidence interval for π has the form

$$\pi_L(x;\delta) = \begin{cases} 0 \text{ if } B(\lceil u\rceil + 1, n - \lceil u\rceil; 1-q) < \frac{1+\delta}{2}, \\ \dfrac{B^{-1}\big(\lceil u\rceil + 1, n - \lceil u\rceil; \frac{1+\delta}{2}\big) - (1-q)}{2q-1} & \text{otherwise,} \end{cases}$$
$$\pi_R(x;\delta) = \begin{cases} 1 \text{ if } B(\lfloor u\rfloor, n - \lfloor u\rfloor + 1; q) > \frac{1-\delta}{2}, \\ \dfrac{B^{-1}\big(\lfloor u\rfloor, n - \lfloor u\rfloor + 1; \frac{1-\delta}{2}\big) - (1-q)}{2q-1} & \text{otherwise.} \end{cases} \tag{CP}$$

The coverage probability of the above confidence interval is, by definition, greater than or equal to the given confidence level. In Figures 3.1 and 3.2, the coverage probability is presented for $n = 100$ and $n = 1000$, respectively (solid line). The confidence level is assumed to be 0.95. This coverage probability is calculated, not simulated.

3. Asymptotic c.i. The most common method of constructing a c.i. is the application of the Central Limit Theorem. This method was applied by Yu et al. (2008) and Tian (2014).

If ξ is a binomial random variable with parameters n and ϱ, then, by CLT, the distribution of ξ tends in distribution as $n \to$ to a normal variate with mean value $n\varrho$ and variance $n\varrho(1-\varrho)$. So

$$\hat{\varrho} \sim AN\bigg(\varrho, \frac{\varrho(1-\varrho)}{n}\bigg).$$

Tian (2014) considered the following estimator of π:

$$\tilde{\pi}_C = \frac{\hat{\varrho} - (1-q)}{2q-1}.$$

Its properties are as follows:

$$E_\pi \tilde{\pi}_C = \pi,$$
$$\mathrm{Var}_\pi(\tilde{\pi}_C) = \frac{\varrho(1-\varrho)}{n(2q-1)^2} = \frac{\pi(1-\pi)}{n} + \frac{q(1-q)}{n(2q-1)^2}.$$

The unbiased estimator of the variance of $\tilde{\pi}_C$ is given by the formula

$$\widehat{\mathrm{Var}\,\tilde{\pi}_C} = \frac{\hat{\varrho}(1-\hat{\varrho})}{(n-1)(2q-1)^2} = \frac{\tilde{\pi}_C(1-\tilde{\pi}_C)}{n-1} + \frac{q(1-q)}{(n-1)(2q-1)^2}.$$

By CLT we have

$$\tilde{\pi}_C \sim AN\bigg(\pi, \frac{\pi(1-\pi)}{n} + \frac{(1-q)q}{n(2q-1)^2}\bigg).$$

Hence

$$\frac{\tilde{\pi}_C - \pi}{\sqrt{\widehat{\mathrm{Var}\,\tilde{\pi}_C}}} \sim N(0,1).$$

There are two ways of constructing a c.i. on the basis of the above approximation. The first one is based on solving the inequality

$$\left|\frac{\tilde{\pi}_C-\pi}{\sqrt{\operatorname{Var}\tilde{\pi}_C}}\right|<u_{(1+\delta)/2}$$

or equivalently

$$\left(\frac{\tilde{\pi}_C-\pi}{\sqrt{\operatorname{Var}\tilde{\pi}_C}}\right)^2<u^2_{(1+\delta)/2}.$$

Solving the inequality with respect to π we obtain the following c.i.:

$$\frac{2n\tilde{\pi}_C+u^2_{(1+\delta)/2}\pm u_{(1+\delta)/2}\sqrt{4n\tilde{\pi}_C(1-\tilde{\pi}_C)+\frac{4n(1-q)q+u^2_{(1+\delta)/2}}{(1-2q)^2}}}{2(n+u^2_{(1+\delta)/2})}. \tag{WP}$$

In the second approach the variance of $\tilde{\pi}_C$ in CLT is substituted by its unbiased estimator

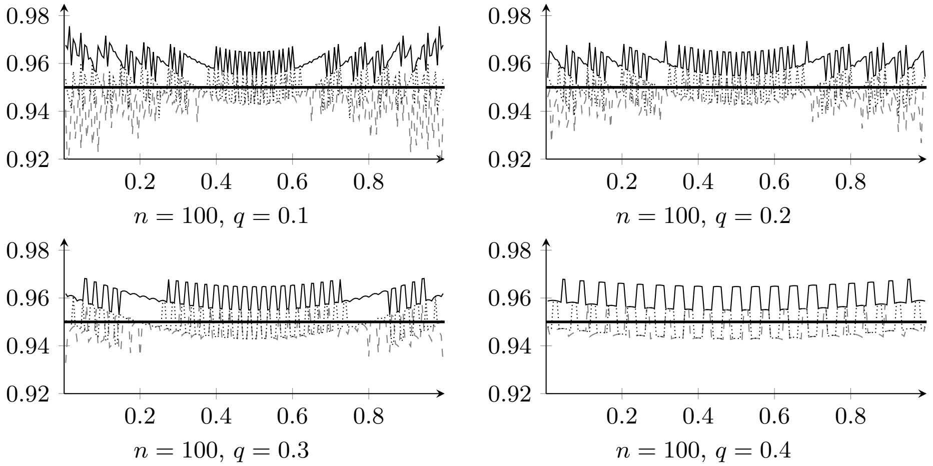
$$\frac{\tilde{\pi}_C-\pi}{\sqrt{\widehat{\operatorname{Var}\tilde{\pi}_C}}}\sim N(0,1).$$

Solving the inequality

$$\left|\frac{\tilde{\pi}_C-\pi}{\sqrt{\widehat{\operatorname{Var}\tilde{\pi}_C}}}\right|<u_{(1+\delta)/2}$$

with respect to π yields the following c.i.:

$$\left(\tilde{\pi}_C-u_{(1+\delta)/2}\sqrt{\widehat{\operatorname{Var}\tilde{\pi}_C}};\ \tilde{\pi}_C+u_{(1+\delta)/2}\sqrt{\widehat{\operatorname{Var}\tilde{\pi}_C}}\right). \tag{AP}$$



$n=100$, $q=0.1$

$n=100$, $q=0.2$

$n=100$, $q=0.3$

$n=100$, $q=0.4$

Fig. 3.1. Probability of coverage for $n=100$ and $\delta=0.95$

The coverage probabilities of the (WP) and (AP) confidence intervals are shown in Figures 3.1 and 3.2 for $n = 100$ and $n = 1000$ (dotted and dashed lines, respectively). The confidence level is assumed to be 0.95. These coverage probabilities are calculated, not simulated.

The proposed (CP) c.i. maintains the nominal confidence level, and the risk of error in the statement is not greater than $1-\gamma$. Unfortunately, asymptotic "confidence intervals" do not satisfy the Neyman (1934) definition. The probability of coverage is less than the nominal confidence level, i.e., the risk of an erroneous statement is greater than $1-\gamma$ and remains unknown. Hence in what follows we consider only (CP) confidence intervals.

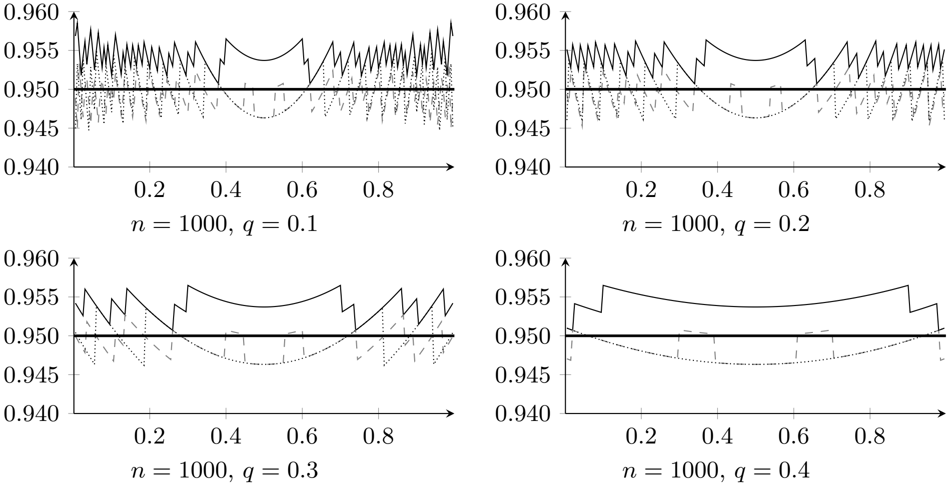


Fig. 3.2. Probability of coverage for $n = 1000$ and $\delta = 0.95$

4. The length of the exact confidence interval. Let us consider the length of the CP confidence interval. For x an observed value of the estimator $\hat{\pi}_{\mathrm{CM}}$ we have

$$l(x,q,n)$$
$$= \begin{cases} \dfrac{B^{-1}\big(\lfloor u\rfloor, n-\lfloor u\rfloor+1; \frac{1-\delta}{2}\big)-(1-q)}{2q-1} & \text{if } B(\lceil u\rceil+1, n-\lceil u\rceil; 1-q) < \frac{1+\delta}{2}, \\ \dfrac{B^{-1}\big(\lceil u\rceil+1, n-\lceil u\rceil; \frac{1+\delta}{2}\big)-(1-q)}{2q-1} & \text{if } B(\lfloor u\rfloor, n-\lfloor u\rfloor+1; q) > \frac{1-\delta}{2}, \\ \dfrac{B^{-1}\big(\lfloor u\rfloor, n-\lfloor u\rfloor+1; \frac{1-\delta}{2}\big)-B^{-1}\left(\lceil u\rceil+1, n-\lceil u\rceil; \frac{1+\delta}{2}\right)}{2q-1} & \text{otherwise.} \end{cases}$$

Recall that $u = n(x(2q-1)+(1-q))$.

The length of the c.i. is a random variable. It depends on q, n and x. There are at least three approaches to the problem of minimizing the length of the c.i.:

(a) minimizing for each x,
(b) minimizing the expected length,
(c) almost sure minimizing.

Minimizing the length for the observed x relies on changing the probabilities of over- and underestimation. This method was widely discussed by Zieliński (2010, 2017). To obtain the shortest c.i. in the Neyman sense, i.e., controlling the probability of coverage, randomization is needed. In what follows we consider equitailed c.i. and confine ourselves to the problem of minimizing the expected length and almost sure minimization.

Note that the length decreases as n increases, hence the problem of minimizing the length is equivalent to finding the appropriate q, i.e., the probability of a positive answer to the neutral question. As we proceed, the sample size n is treated as a given number.

Minimizing expected length. The problem to be solved may be written in the following way:

$$q_e^* = \arg\min_{q\in Q} \sup_{\pi\in\Pi} E_\pi^{C(\pi)} l(\hat{\pi}_{\mathrm{CM}}, q, n),$$

where Q and Π are acceptable sets for q and π and respectively,

$$E_\pi^{C(\pi)} l(\hat{\pi}_{\mathrm{CM}}, q, n) = \sum_{x\in C(\pi)} l(x, q, n) P_\pi\{\hat{\pi}_{\mathrm{CM}} = x\}$$

denotes the expected length of the c.i. covering the estimated value of π. The set $C(\pi) = \{x : \pi_L(x;\delta) < \pi < \pi_R(x;\delta)\}$ includes those values of the variable $\hat{\pi}_{\mathrm{CM}}$ for which the c.i. covers π. Without prior knowledge of q and π the set Q equals $[0, 0.5)$ (under the prior assumption that $q < 0.5$) and $\Pi = (0, 1)$.

Almost sure minimizing of length. The problem to be solved may be written in the following way:

$$q_d^* = \arg\max_{q\in Q} \inf_{\pi\in\Pi} P_\pi^{C(\pi)}\{l(\hat{\pi}_{\mathrm{CM}}, q, n) \le d\},$$

where d is a given number chosen in advance and

$$\delta \cdot P_\pi^{C(\pi)}\{l(\hat{\pi}_{\mathrm{CM}}, q, n) \le d\} = \sum_{x\in C(\pi)} P_\pi\{\hat{\pi}_{\mathrm{CM}} = x\}\mathbb{1}(l(x, q, n) \le d)$$

denotes the probability that the length of the c.i. covering the estimated value of π does not exceed d; d should be small. The function $\mathbb{1}(p)$ is equal to 1 if p is true and zero otherwise.

It is easy to see that for $Q = [0, 0.5)$, the minimal length with respect to q is obtained for $q = 0$, which is equivalent to not asking the neutral question. Such a questionnaire (without a neutral question) is useless for our purposes. Hence we have to introduce a limitation on the probability q.

Tan et al. (2009) introduced the notion of degree of privacy protection through the probabilities

$$P_\pi\{Y = 1 \mid Z = 1\} \quad \text{and} \quad P_\pi\{Y = 1 \mid Z = 0\}.$$

These probabilities are connected with the safety of the interviewee of non-discovering her/his positive answer to the sensitive question. In the CM model these probabilities are as follows (by a simple application of the Bayes theorem):

$$p_{1,1}(q) := P_\pi\{Y = 1 \mid Z = 1\} = \frac{\pi q}{\pi q + (1-\pi)(1-q)},$$
$$p_{1,0}(q) := P_\pi\{Y = 1 \mid Z = 0\} = \frac{\pi(1-q)}{\pi(1-q) + (1-\pi)q}.$$

These probabilities should be appropriately small, so that they do not exceed a given value $\gamma \in (0,1)$. Note that for all $q \in (0, 0.5)$ and all $\pi \in (0,1)$ we have $p_{1,0}(q) \geq p_{1,1}(q)$ (see Figure 4.1). Therefore we are interested in $q < 0.5$ such that

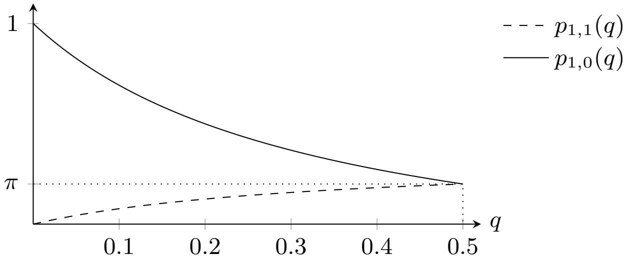


Fig. 4.1. Privacy protection versus q

$$p_{1,0}(q) \leq \gamma \quad \text{for } \pi \in \Pi. \tag{2}$$

Simple algebra gives the following condition for q:

$$q(\pi;\gamma) \leq q < 0.5 \quad \text{for } \pi \in \Pi, \tag{3}$$

where $q(\pi;\gamma) = \frac{\pi(1-\gamma)}{\gamma(1-2\pi)+\pi}$. Since $q(\gamma,\gamma) = 0.5$ for all $\pi \in \Pi$, the above condition (3) holds for $\gamma > \pi$. This means that the maximal privacy protection (i.e., minimal γ to be chosen) is limited by the percentage of the population who committed socially stigmatizing misdeeds. Hence, the problem of minimizing the length, assuming $\pi \leq \pi_0$, for a given $\pi_0 \in (0,1)$, is well defined for $q \in [q(\pi_0;\gamma), 0.5)$. The privacy protection criterion is satisfied for $\gamma > \pi_0$. Based on this, we assume that $\Pi = (0, \pi_0]$ and $Q = [q(\pi_0;\gamma), 0.5)$.

Note that Π, as well as Q, does not depend on the sample size n. So, the length of the c.i. may be minimized by choosing an appropriate sample size.

Let $d \in (0,1)$ be a given number. We would like to find a sample size that gives the c.i. of length not greater than d. In particular, we are interested in the c.i. covering the estimated value of π. There are two approaches to the problem: find minimal n such that

- $E_\pi^{C(\pi)} l(\hat{\pi}_{\rm CM}, q_e^*, n) \leq d$ for all $\pi \in \Pi$, or
- $P_\pi^{C(\pi)}\{l(\hat{\pi}_{\rm CM}, q_d^*, n) \leq d\} \geq 1-\lambda$ for a given probability $1-\lambda$ and for all $\pi \in \Pi$.

In the first approach, we want the average length of the c.i. covering the estimated value of π to be less than the given d. In the second approach, we want the length of at least $(1-\lambda)\%$ of the c.i. covering the estimated π to be less than the given d. Let us note that we have at least $\delta\%$ of intervals covering the unknown parameter π and for an infinitely large sample size n, the value $P_\pi^{C(\pi)}\{l(\hat{\pi}_{\rm CM}, q, n) \leq d\}$ is 1.

Consider the first approach. The analysis of $E_\pi^{C(\pi)} l(\hat{\pi}_{\rm CM}, q, n)$ shows that (see Figures 4.2, 4.3)

(a) for each q and n, it increases as π increases for $\pi \in (0, 0.5)$,
(b) for each π and n, it increases as q increases for $q \in (0, 0.5)$.

Hence it is enough to find the sample size n with $E_{\pi_0}^{C(\pi_0)} l(\hat{\pi}_{\rm CM}, q(\pi_0,\gamma), n) \leq d$. For a given π_0, γ, and d, the solution may be found numerically.

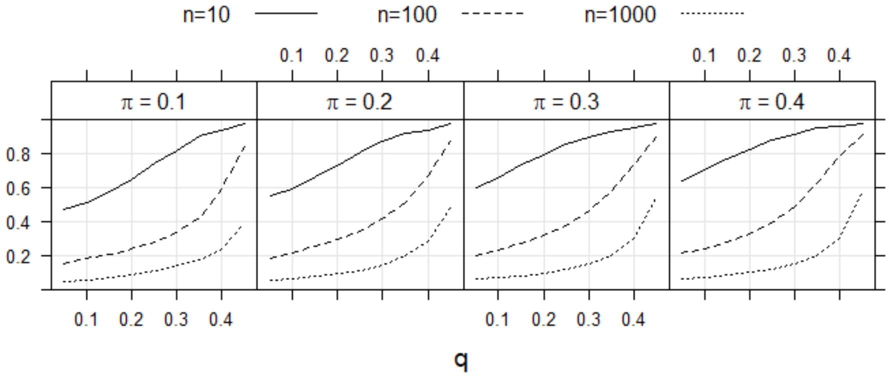


Fig. 4.2. Expected length versus q with respect to π under the condition that π is in the confidence interval

In Table 4.1 some exemplary minimal sample sizes are given for confidence level $\delta = 0.95$ and privacy protection $\gamma = 0.5$.

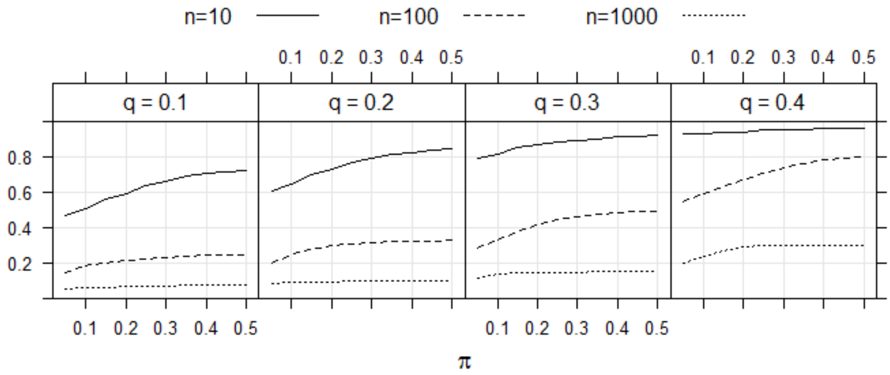


Fig. 4.3. Expected length versus π with respect to q under the condition that π is in the confidence interval

Table 4.1. The smallest n such that $E_{\pi_0}^{C(\pi_0)} l(Z_n, q(\pi_0, \gamma), n) \le d$

π_0	$d = 0.05$	$d = 0.06$
0.1	1326	929
0.2	3428	2388
0.3	8557	5955
0.4	34862	24206

Note: $\pi_0 = q(\pi_0, \gamma)$ for $\gamma = 0.5$

Now consider the second approach, i.e., we want to find a sample size n such that

$$P_{\pi}^{C(\pi)}\{l(\hat{\pi}_{\mathrm{CM}}, q, n) \le d\} \ge 1 - \lambda$$

for given d and λ. The analysis of $P_{\pi}^{C(\pi)}\{l(\hat{\pi}_{\mathrm{CM}}, q, n) \le d\}$ shows that (see Figures 4.4 and 4.5)

(a) for every d, q, and n, it decreases in π (for $\pi \in (0, 0.5)$),
(b) for every d, π, and n, it does not decrease in q (for $q \in (0, 0.5)$).

The monotonicity of $P_{\pi}^{C(\pi)}\{l(\hat{\pi}_{\mathrm{CM}}, q, n) \le d\}$ in q is disturbed due to the discreteness of the observed variable $\hat{\rho}$. However, the probability generally decreases in q (see Figure 4.5). This distinctive feature allows us to recommend that it is enough to find a sample size n such that

$$P_{\pi_0}^{C(\pi_0)}\{l(\hat{\pi}_{\mathrm{CM}}, q(\pi_0; \gamma), n) \le d\} \ge 1 - \lambda.$$

For given π_0, γ, d, and λ, the solution may be found numerically. In Table 4.2, some example minimal sample sizes are given for confidence level $\delta = 0.95$, privacy protection $\gamma = 0.5$, and $\lambda = 0.01$ and $\lambda = 0.05$.

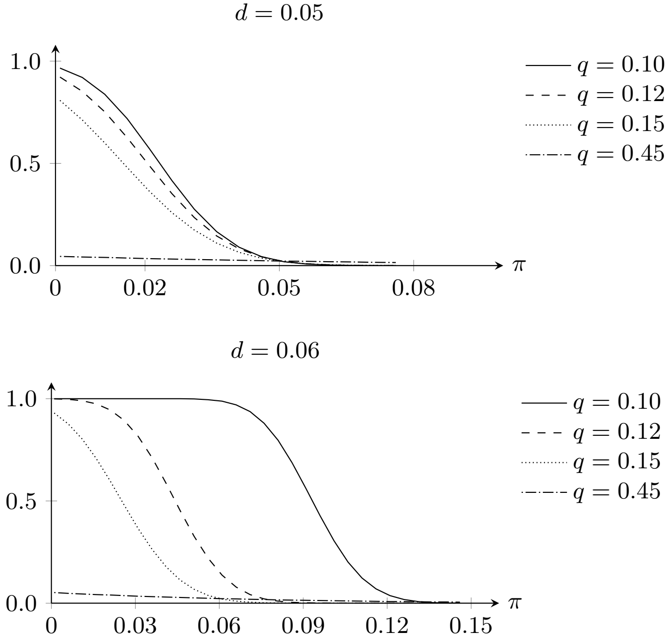


Fig. 4.4. $P_{\pi}^{C(\pi)}\{l(\hat{\pi}_{\mathrm{CM}}, q, n) \leq d\}$ versus π; the case of $n = 1000$ and $\delta = 0.95$

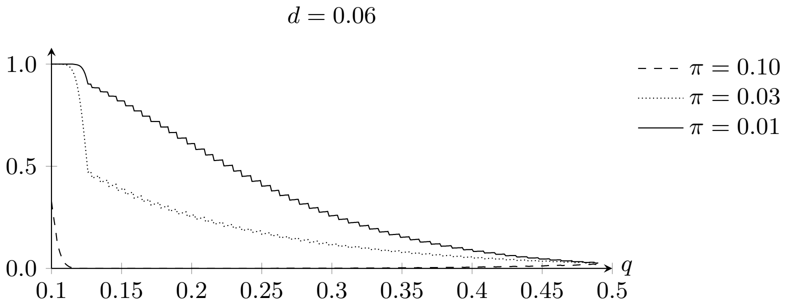


Fig. 4.5. $P_{\pi}^{C(\pi)}\{l(\hat{\pi}_{\mathrm{CM}}, q, n) \leq d\}$ versus q; the case of $n = 1000$ and $\delta = 0.95$

Table 4.2. The smallest n such that $P_{\pi_0}^{C(\pi_0)}\{l(\hat{\pi}_{\mathrm{CM}}, q(\pi_0;\gamma), n) \leq d\} \geq 1 - \lambda$

	$d = 0.05$		$d = 0.06$	
π_0	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.01$	$\lambda = 0.05$
0.1	1570	1551	1111	1094
0.2	3861	3845	2699	2686
0.3	9508	9499	6623	6615
0.4	38576	38572	26819	26815

Note: $\pi_0 = q(\pi_0, \gamma)$ for $\gamma = 0.5$

As may be expected, the sample size increases when the requirements for the length of the c.i. increase, i.e., d decreases. Also, to obtain the c.i. of a given length for π which is prior smaller, i.e., π_0 is smaller the smaller sample size is needed. It is interesting that a slight increase in sample size gives a higher percentage for c.i. for a given length covering an unknown value of π (the probability λ is smaller, i.e., $1 - \lambda$ is greater).

5. Conclusions. In the paper a new confidence interval for the fraction of sensitive questions is proposed. Recall that Neyman (1934) defined a c.i. as a method of estimation with "... the probability of an error in a statement of this sort being equal to or less than $1 - \varepsilon$, where ε is any number $0 < \varepsilon < 1$, chosen in advance. The number ε I call the confidence coefficient." (In our notation the confidence level is δ.) The new c.i. keeps the prescribed confidence level, while the very popular asymptotic c.i. does not.

An important practical problem is the sample size. We have derived the minimal sample size fulfilling two criteria: average length and almost sure length. To derive these sample sizes, we put restrictions on privacy protection, i.e., the probability of discovering the *YES* answer to the sensitive question. This probability should be appropriately small so that the interviewee can feel safe answering the questionnaire. Also, we restricted ourselves to rare phenomena, so we consider sensitive questions with a small (given in advance) probability of a positive answer.

We do not compare the length of our c.i. with asymptotic versions. The asymptotic c.i.'s must be shorter, since they do not keep a prescribed confidence level; the real probability of coverage is less than the given confidence level. Hence, the comparison of the lengths makes no sense. Note that our confidence interval is very easy to calculate; even a standard smartphone has a spreadsheet application that can calculate the quantiles of Beta distribution. Recall that asymptotic c.i.'s based on normal approximation were useful when computers were not easily available. We, therefore, recommend using our confidence interval in practice.

References

R. Arnab (1990), *On commutativity of design and model expectations in randomized response surveys*, Comm. Statist. Theory Methods 19, 3751–3757.

R. Arnab (1996), *Randomized response trials: A unified approach for qualitative data*, Comm. Statist. Theory Methods 25, 1173–1183.

R. Arnab, D. K. Shangodoyin and A. Arcos (2019), *Nonrandomized response model for complex survey designs*, Statistics in Transition N.S. 20, 67–86.

L. A. Franklin (1989), *Randomized response sampling from dichotomous populations with continuous randomization*, Survey Methodology 15, 225–235.

B. G. Greenberg, A.-L. A. Abul-Ela and D. G. Horvitz (1969), *The unrelated question randomized response model: theoretical framework*, J. Amer. Statist. Assoc. 64, 520–539.

D. G. Horvitz, B. V. Shah and W. R. Simmons (1967), *The unrelated question randomized response model*, in: Proceedings of the Social Statistics Section, Amer. Statist. Assoc., 65–72.

A. Y. C. Kuk (1990), *Asking sensitive question indirectly*, Biometrika 77, 436–438.

J. Neyman (1934), *On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection*, J. Roy. Statist. Soc. 97, 558–625.

D. Raghavarao (1978), *On an Estimation Problem in Warner's Randomized Response Technique*, Biometrics 34, 87–90.

M. Rueda, B. Cobo and A. Arcos (2015), *Package 'RRTCS': Randomized response techniques for complex surveys*, http://cran.r-project.org/web/packages/RRTCS.

M. Tan, G. L. Tian and M. L. Tang (2009), *Sample surveys with sensitive questions: a non-randomized response approach*, Amer. Statist. 63, 9–16.

G.-L. Tian (2014), *A new non-randomized response model: The parallel model*, Statistica Neerlandica 68, 293–323.

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

S. L. Warner (1965), *Randomized response: a survey technique for eliminating evasive answer bias*, J. Amer. Statist. Assoc. 60, 63–69.

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

W. Zieliński (2010), *The shortest Clopper-Pearson confidence interval for binomial probability*, Comm. Statist. Simulation Comput. 39, 188–193.

W. Zieliński (2017), *The shortest Clopper-Pearson randomized confidence interval for binomial probability*, REVSTAT J. 15, 141–153.

Stanisław Jaworski (corresponding author), Wojciech Zieliński
Department of Econometrics and Statistics
Warsaw University of Life Sciences
02-767 Warszawa, Poland
E-mail: stanislaw_jaworski@sggw.edu.pl
wojciech_zielinski@sggw.edu.pl