Thailand Statistician
October 2026; 24(4): 1027-1041
http://statassoc.or.th
Contributed paper

Triangular Model and Two of Its Variants Under Stratified Sampling

Eman Mansour, Reda Mazloum and Sally Abdelfatah*
Department of Statistics, Faculty of Economics and Political Science, Cairo University, Giza, Egypt.
*Corresponding author; e-mail: sally_abdelfatah@feps.edu.eg

Received: 10 August 2023
Revised: 23 June 2026
Accepted: 9 July 2026

Abstract

The non-randomized response (NRR) technique is a procedure that can be used in surveys with sensitive questions, and it has many advantages over the randomized response (RR) technique such as the ease of implementation and the ability to be used in self-administrated surveys as no random device is required. Many NRR models were developed to increase the efficiency and/or increase the respondents' cooperation such as the triangular model and the alternating triangular model. In this paper, we extended the triangular model and two of its variants; the alternating triangular model and the triangular model under non-compliance; to stratified random sampling. The maximum likelihood estimation method was employed to estimate the population proportion possessing the sensitive attribute under each of the three extended models. Unbiased estimators of the sensitive proportion were derived and their sampling variances were obtained. In addition, the optimal allocation of the sample size for the three extended models was derived.

Keywords: Estimation of proportion, maximum likelihood estimation, non-randomized response technique, stratified random sampling, triangular model

1. Introduction

Surveys are considered as very important sources of information, especially in social sciences. Some surveys may contain questions about sensitive characteristics such as abortion, plagiarism, drug addiction, family violence, tax evasion and having certain diseases like AIDs. Surveys with such direct sensitive questions may suffer from non-response and/or false responses. Thus, alternative techniques that don't directly reveal the respondent's true status relative to the sensitive question and thereby increase the respondents' cooperation with higher efficiency have appeared. The first alternative technique was introduced by Warner (1965) which is known as the randomized response (RR) technique. Warner's model is based on the existence of a random device (RD) that has two possible outcomes. Such RD can be for example a spinner or a deck of cards. The two statements that appear on Warner's RD are: I belong to the sensitive group with probability p and I do not belong to the sensitive group with probability $1-p$. If the RD is a spinner, then the respondent is asked to spin the spinner and give the answer "yes" or "no" according to his/her actual status. This is done

completely away from the interviewers. Various extensions of Warner's model have been introduced to increase the efficiency and/or increase the cooperation of the respondent.

Tan et al. (2009) pointed out that randomized response models have some drawbacks and limitations. First, they are limited to face-to-face surveys. Second, they have a lack of reproducibility. Reproducibility means getting the same conclusion (result) under different conditions. For example, in the RR models the interviewee's answer differs depending on the outcome of RD. Hence, this leads to significant loss of efficiency. Third, since the RD is under the control of the interviewers, the respondent will definitely doubt that his/her privacy is protected. Finally, the usage of RDs increases the survey costs.

To overcome the aforementioned drawbacks, an alternative technique namely the Non-Randomized Response (NRR) technique was introduced by Swensson (1974) and Takahasi and Sakasegawa (1977). The main concern of this technique is to develop models that do not depend on using RDs and therefore can be applied to both face to face and mail surveys. This technique uses a non-sensitive question as a replacement of the RD. Using the NRR technique, the respondents follow a predetermined response rule that combines information from the sensitive question of interest and information from one or more nonsensitive questions. The interviewer will only observe the final response generated by this rule but can't determine the respondent's true status from the recorded response and thereby the respondent's privacy is protected. Various NRR models were developed to increase the efficiency of existing models and /or achieve better degree of privacy protection. One model is the triangular model introduced by Yu et al. (2008). Tan et al. (2009) compared the triangular model with Warner's (1965) model and they found that the triangular model was better than Warner's model in terms of relative efficiency and degree of privacy protection. Two variants of the triangular model were then proposed. Wu and Tang (2014) developed the so-called triangular model using alternating tables. In addition, Wu and Tang (2016) extended the triangular model under non-compliance. Examples of other NRR models were developed by: Chaudhuri (2012), Groenitz (2014a), Groenitz (2014b) and Tian (2014).

In this paper, the triangular model and two of its variants, the alternating triangular model, and the triangular model under non-compliance, are extended to stratified random sampling.

The rest of this article is organized as follows: In Section 2, a review of the triangular and its two variants is introduced. The stratified extension of the triangular model is presented in Section 3. The alternating triangular model and the triangular model under non-compliance are extended to stratified random sampling in Sections 4 and 5, respectively. Finally, a conclusion is presented in Section 6.

2. Review of the Triangular and Two of Its Variants

2.1. The triangular model

For the design of the triangular model developed by Yu et al. (2008), let Y denote the variate associated with the sensitive question with probability of possessing the sensitive attribute π (i.e., $\pi = \Pr(Y = 1)$) and let W denote the variate associated with the non-sensitive question with probability of possessing the non-sensitive attribute p (i.e., $p = \Pr(W = 1)$) where $0 < p < 1$ is known, and W and Y are independent. Table 1 shows the design of the triangular model and the corresponding probability of each cell.

Table 1 The triangular model and the corresponding cell probability

Category	$W=\mathbf{0}$	$W=1$	Category	$W=\mathbf{0}$	$W=1$
$Y=\mathbf{0}$	○	•	$Y=\mathbf{0}$	$(1-\pi)(1-p)$	$(1-\pi)p$
$Y=1$	•	•	$Y=1$	$\pi(1-p)$	πp

The respondent is asked to truthfully put a tick in the circle or in the triangle formed by the three dots according to his/her actual status. As a result of this design, the subclass $\{Y=0,\ W=0\}$ represents the circle and it is a non-sensitive subclass. $\{Y=1\}\cup\{Y=0,\ W=1\}$ represents the three dots (i.e., the triangle). For a random sample of size n, let $\theta_T=(1-\pi)(1-p)$ denote the probability of putting a tick in the circle and s denote the number of ticks in the circle.

The maximum likelihood estimator (MLE) of π is given by $\hat{\pi}=1-\frac{\hat{\theta}}{1-p}$; $\hat{\theta}=\frac{s}{n}$. $\hat{\pi}$ is an unbiased estimator of π with variance: $V(\hat{\pi})=\frac{p(1-\pi)}{n(1-p)}+\frac{\pi(1-\pi)}{n}$.

Jerke and Krumpal (2013) conducted the first empirical application of the triangular model to estimate the prevalence of plagiarism in student papers in one of the courses at the University of Leipzig, Germany. They asked two sensitive questions to estimate the prevalence of partial and full plagiarism. For partial plagiarism, the sensitive question was “When writing a paper have you ever deliberately included a passage from another work without marking it as a quotation?”. For full plagiarism, the sensitive question was “Have you ever had a large part of a paper written by someone else, or passed off someone else’s paper as your own?”. The sampled students were randomly assigned to one of two survey conditions; direct questioning or Triangular model. With direct questioning, the respondents answered the two plagiarism questions directly. For the triangular model, the sensitive partial plagiarism question was combined with the non-sensitive question “Is your mother’s birthday in January, February, or March?” and the sensitive full plagiarism question was combined with the non-sensitive question “Is your father’s birthday in October, November or December?”. For the two nonsensitive questions, they approximated the probability of yes to be 0.26 for mother’s birthday and 0.23 for father’s birthday. The estimated prevalence for partial plagiarism was found to be 10% through direct questioning and 18% through the Triangular model. The estimated prevalence for full plagiarism was found to be 0% through direct questioning and 0.8% through the triangular model. The results showed that with the triangular model, a significantly higher percentage of students admitted plagiarism compared with direct questioning suggesting that the triangular model is a promising alternative to direct questioning.

2.2. The triangular model using alternating tables

Wu and Tang (2014) developed the so-called triangular model using alternating tables. They considered Yu et al. (2008) triangular model under the case where the proportion of the population possessing the non-sensitive attribute p is unknown. In the alternating triangular model, a simple random sample of n respondents is drawn from a population of size N, then it is randomly split into two groups of sizes n_1 and n_2. The table of the regular triangular model (Table 1) is assigned to the first group of participants and the alternating table (Table 2) is assigned to the second group of

participants. The alternating table has the same design of the triangular model, but the two categories of the non-sensitive question are switched (alternated).

Table 2 The alternating table assigned to the second group and the corresponding cell probability

Category	$W=1$	$W=\mathbf{0}$	Category	$W=1$	$W=\mathbf{0}$
$Y=\mathbf{0}$	○	•	$Y=\mathbf{0}$	$(1-\pi)p$	$(1-\pi)(1-p)$
$Y=1$	•	•	$Y=1$	πp	$\pi(1-p)$

Using the maximum likelihood (ML) estimation method, the estimator of π for the alternating triangular model is as follows:

$$\hat{\pi}_A = 1-\frac{s_1}{n_1}-\frac{s_2}{n_2},$$

where s_1 and s_2 are the numbers of the respondents putting a tick in the circle in the first and second groups, respectively. $\hat{\pi}_A$ is an unbiased estimator of π with variance given by

$$V\left(\hat{\pi}_A\right)=\frac{\theta_1\left(1-\theta_1\right)}{n_1}+\frac{\theta_2\left(1-\theta_2\right)}{n_2},$$

where $\theta_1=\left(1-\pi\right)\left(1-p\right)$ and $\theta_2=\left(1-\pi\right)p$ represent the probabilities of putting a tick in the circle in the first and second groups, respectively. Finally, they applied the triangular model using alternating tables on a survey about premarital sexual activity between college students in Jilin province in China.

2.3. The triangular model under non-compliance

Wu and Tang (2016) extended the triangular model to the case of incompletely truthful reporting. They presented two models for this case The dual triangular model and the triangular model under non-compliance using alternating tables. The performance of both models was good on estimating the sensitive proportion. However, they recommended the use of the triangular model under non-compliance using alternating tables as it has one non-sensitive question while, the dual triangular model has two non-sensitive questions. Hence, we will present the details of the triangular model under non-compliance using alternating tables which will be extended to stratified random sampling in our paper.

Under non-compliance, they followed procedures similar to conducting the triangular model using alternating tables (Wu and Tang 2014) with two main differences: a) they defined a new parameter γ that represents the probability of giving the non-sensitive answer (i.e., $Y=0$) while the respondent possesses the sensitive attribute and b) the probability of having the non-sensitive attribute p is known while γ is unknown. Under the non-compliance assumption, Table 3 is assigned to the first group of participants and an alternating table (Table 4) is assigned to the second group of participants.

Table 3 The table assigned to the first group and the corresponding probabilities under non-compliance

Category	$W=\mathbf{0}$	$W=1$	Probabilities
$Y=\mathbf{0}$	○	•	$\Pr(\text{Ticking circle})=(1-\pi)(1-p)+\pi\gamma$
$Y=1$	•	•	$\Pr(\text{Ticking triangle})=1-(1-\pi)(1-p)-\pi\gamma$

Table 4 The alternating table assigned to the second group and the corresponding probabilities under non-compliance

Category	$W=1$	$W=\mathbf{0}$	Probabilities
$Y=\mathbf{0}$	○	•	$\Pr(\text{Ticking the circle})=(1-\pi)p+\pi\gamma$
$Y=1$	•	•	$\Pr(\text{Ticking the triangle})=1-(1-\pi)p-\pi\gamma$

Let $\psi_1=(1-\pi)(1-p)+\pi\gamma$ and $\psi_2=(1-\pi)p+\pi\gamma$ be the probabilities of putting a tick in the circle in the first and second groups, respectively and s_1 and s_2 denote the numbers of ticks in the circle in the first and second groups, respectively.

Using the ML estimation method, the estimators of π and γ are given, respectively, by:

$$\hat{\pi}_C=\frac{\hat{\psi}_{C1}-\hat{\psi}_{C2}}{2p-1}+1;\quad p\neq\frac{1}{2},\quad \hat{\gamma}_C=\frac{\hat{\psi}_{C2}(p-1)+p\hat{\psi}_{C1}}{\hat{\psi}_{C1}-\hat{\psi}_{C2}+2p-1}$$

where $\hat{\psi}_{C1}=\frac{s_1}{n_1}$ and $\hat{\psi}_{C2}=\frac{s_2}{n_2}$.

$\hat{\pi}_C$ is an unbiased estimator of π with variance:

$$V(\hat{\pi}_C)=\frac{1}{(2p-1)^2}\left(\frac{\psi_1(1-\psi_1)}{n_1}+\frac{\psi_2(1-\psi_2)}{n_2}\right);\quad p\neq\frac{1}{2}.$$

3. The Stratified Triangular Model

In this section, the stratified extension of the triangular model is presented. Subsection 3.1 discusses the design and the ML estimation of the stratified triangular model. Subsection 3.2 presents the optimal allocation of the sample size.

3.1. Design and estimation of the stratified triangular model

Suppose that a population of size N is divided into k strata where the number of units in each stratum, N_h; $h=1,2,...,k$ is known. A simple random sample with replacement (SRSWR) of size n_h is drawn from stratum h where $n_h>1$. It is assumed that the selections in the different strata are made independently. Each respondent in the sample from stratum h is provided with the following table.

Table 5 The stratified triangular model in stratum *h* and the corresponding cell probability

Category	$W=\mathbf{0}$	$W=1$	Category	$W=\mathbf{0}$	$W=1$
$Y=\mathbf{0}$	○	•	$Y=\mathbf{0}$	$(1-\pi_h)(1-p_h)$	$(1-\pi_h)p_h$
$Y=1$	•	•	$Y=1$	$\pi_h(1-p_h)$	$\pi_h p_h$

π_h is the probability of having the sensitive attribute in stratum h and p_h is the probability of having the non-sensitive attribute in stratum h where $0 < p_h < 1$. Let $\theta_h = (1-\pi_h)(1-p_h)$ be the probability of putting a tick in the circle in stratum h,

$$Y_{ih} = \begin{cases} 1, & i^{\text{th}} \text{ respondent in stratum } h \text{ puts a tick in the circle.} \\ 0, & i^{\text{th}} \text{ respondent in stratum } h \text{ puts a tick in the triangle,} \end{cases}$$

and $s_h = \sum_{i=1}^{n_h} Y_{ih}$ denote the number of ticks in the circle in stratum h. The likelihood function is given by

$$\mathrm{L}(\pi_h \mid Y_{obs}) = \prod_{i=1}^{n_h} \theta_h^{Y_{ih}} (1-\theta_h)^{1-Y_{ih}},$$

where $Y_{obs} = \{Y_{ih} \; ; \; i = 1,2,3,...,n_h \text{ and } h = 1,2,3,...,k\}$. Setting $\frac{\partial}{\partial \theta_h} \ln \mathrm{L}(\pi_h \mid Y_{\text{obs}}) = 0,$ we get the MLE $\hat{\theta}_h$ of θ_h. Consequently, the MLE of π_h is given by

$$\hat{\pi}_h = 1 - \frac{\hat{\theta}_h}{1-p_h}, \tag{1}$$

where $\hat{\theta}_h = \frac{\sum_{i=1}^{n_h} Y_{ih}}{n_h} = \frac{s_h}{n_h}$. The estimator $\hat{\pi}_h$ given in (1) is an unbiased estimator of π_h with variance given by

$$V(\hat{\pi}_h) = \frac{p_h(1-\pi_h)}{n_h(1-p_h)} + \frac{\pi_h(1-\pi_h)}{n_h}. \tag{2}$$

An unbiased estimator of the variance of $\hat{\pi}_h$ is given by

$$\hat{V}(\hat{\pi}_h) = \frac{p_h(1-\hat{\pi}_h)}{(n_h-1)(1-p_h)} + \frac{\hat{\pi}_h(1-\hat{\pi}_h)}{(n_h-1)}.$$

The MLE of the population proportion π is

$$\hat{\pi}_S = \sum_{h=1}^{k} W_h \hat{\pi}_h = \sum_{h=1}^{k} W_h \left(1 - \frac{\hat{\theta}_h}{1-p_h}\right), \tag{3}$$

where $W_h = \frac{N_h}{N}$ is the weight of stratum h in the population, $\sum_{h=1}^{k} W_h = 1$.

Theorem 1 *The MLE $\hat{\pi}_S$ given in (3) is an unbiased estimator of the population proportion π, with variance given by*

$$V(\hat{\pi}_S) = \sum_{h=1}^{k} W_h^2 \; \frac{p_h(1-\pi_h)}{n_h(1-p_h)} + \frac{\pi_h(1-\pi_h)}{n_h}. \tag{4}$$

Proof:

The proof of unbiasedness is done directly by taking the expected values on both sides of (3). Since the selections in different strata are made independently, the variance of $\hat{\pi}_S$ is

$$V\left(\hat{\pi}_s\right)=V\left(\sum_{h=1}^{k}W_h\hat{\pi}_h\right)=\sum_{h=1}^{k}W_h^2V\left(\hat{\pi}_h\right). \tag{5}$$

Substituting $V\left(\hat{\pi}_h\right)$ of (2) into (5) gives the variance of $\hat{\pi}_S$ in (4).

Theorem 2 *An unbiased estimator of the variance of* $\hat{\pi}_S$ *is given by*

$$\hat{V}\left(\hat{\pi}_S\right)=\sum_{h=1}^{k}W_h^2\ \frac{p_h\left(1-\hat{\pi}_h\right)}{\left(n_h-1\right)\left(1-p_h\right)}+\frac{\hat{\pi}_h\left(1-\hat{\pi}_h\right)}{n_h-1}. \tag{6}$$

Proof:

The proof of unbiasedness is done directly by taking the expected values on both sides of (6).

3.2. Optimal Allocation

Under optimum allocation, the sample size for each stratum (n_h) can be chosen to minimize $V\left(\hat{\pi}_S\right)$ for a specified cost. The cost is assumed to be in the form of a linear cost function as follows:

$$C=c_o+\sum_{h=1}^{k}c_h\,n_h, \tag{7}$$

where c_o is fixed cost and c_h is the cost per unit in stratum h.

Theorem 3 *The optimal allocation of the sample size* n *is given by*

$$n_h=\frac{\frac{N_h}{\sqrt{c_h}}\left(C-c_0\right)\left[\pi_h\left(1-\pi_h\right)+\frac{p_h\left(1-\pi_h\right)}{\left(1-p_h\right)}\right]^{\frac{1}{2}}}{\sum_{h=1}^{k}\sqrt{c_h}N_h\left[\pi_h\left(1-\pi_h\right)+\frac{p_h\left(1-\pi_h\right)}{\left(1-p_h\right)}\right]^{\frac{1}{2}}},\quad h=1,2,\ldots,k, \tag{8}$$

and the minimal variance is

$$V_{Opt}\left(\hat{\pi}_S\right)=\frac{1}{\left(C-c_0\right)}\left(\sum_{h=1}^{k}\sqrt{c_h}W_h\left[\pi_h\left(1-\pi_h\right)+\frac{p_h\left(1-\pi_h\right)}{\left(1-p_h\right)}\right]^{\frac{1}{2}}\right)^2. \tag{9}$$

Proof:

Applying Cauchy-Schwarz inequality to the product $V\left(\hat{\pi}_S\right)\left(C-c_o\right)$, we get

$$n_h=n\frac{\frac{N_h}{\sqrt{c_h}}\left[\pi_h\left(1-\pi_h\right)+\frac{p_h\left(1-\pi_h\right)}{\left(1-p_h\right)}\right]^{\frac{1}{2}}}{\sum_{h=1}^{k}\frac{N_h}{\sqrt{c_h}}\left[\pi_h\left(1-\pi_h\right)+\frac{p_h\left(1-\pi_h\right)}{\left(1-p_h\right)}\right]^{\frac{1}{2}}}. \tag{10}$$

The next step to complete the proof is to find n by substituting n_h of (10) into the cost function of (7). This gives:

$$n = \frac{(C-c_0)\sum_{h=1}^{k}\frac{N_h}{\sqrt{c_h}}\left[\pi_h(1-\pi_h)+\frac{p_h(1-\pi_h)}{(1-p_h)}\right]^{\frac{1}{2}}}{\sum_{h=1}^{k}\sqrt{c_h}N_h\left[\pi_h(1-\pi_h)+\frac{p_h(1-\pi_h)}{(1-p_h)}\right]^{\frac{1}{2}}}. \quad (11)$$

First, substituting n of (11) into (10), we get the optimal sample size from stratum h (n_h) in (8). Finally, by substituting (8) in (4), we get the minimal variance of the stratified estimator $\hat{\pi}_S$ in (9). We can see that the optimal allocation in (8) depends on the values of $\pi_h;\ h=1,2,...k$ which are unknown. In applied studies, we can use values from a previous study or a good guess.

Corollary 1 *Under Neyman allocation of the sample size where the cost per unit for all strata are equal* $c_1 = c_2 = \ldots = c_k = c^*$, *the optimal allocation of the sample size* n *is given by*

$$n_h = n\frac{N_h\left[\pi_h(1-\pi_h)+\frac{p_h(1-\pi_h)}{(1-p_h)}\right]^{\frac{1}{2}}}{\sum_{h=1}^{k}N_h\left[\pi_h(1-\pi_h)+\frac{p_h(1-\pi_h)}{(1-p_h)}\right]^{\frac{1}{2}}},\quad h=1,2,\ldots,k,$$

and the minimal variance of $\hat{\pi}_S$ *is*

$$V_{\text{Ney}}(\hat{\pi}_S) = \frac{1}{n}\left(\sum_{h=1}^{k}W_h\left[\pi_h(1-\pi_h)+\frac{p_h(1-\pi_h)}{(1-p_h)}\right]^{\frac{1}{2}}\right)^2.$$

4. Stratified Alternating Triangular Model

In the present section, the alternating triangular model by Wu and Tang (2014) is extended to stratified random sampling. Accordingly, Subsection 4.1 introduces the design and the ML estimation of the stratified alternating triangular model. In Subsection 4.2, the optimal allocation of the sample size is discussed.

4.1. Design and estimation of the stratified alternating triangular model

It is assumed that the population of size N is divided into k strata where the number of units in each stratum $N_h;\ h=1,2,...,k$ is known. A SRSWR of size n_h is drawn from stratum h, then it is randomly divided into two groups of sizes n_{h1} and n_{h2} (i.e., $n_{h1}+n_{h2}=n_h$) where $n_{h1}>1$ and $n_{h2}>1$. The selections in the different strata are made independently. The first group of the sample from stratum h is provided with Table 6 and the second sampled group of stratum h is provided with Table 7.

Table 6 The table assigned to the first sample group of stratum h and the corresponding cell probability

Category	$W=\mathbf{0}$	$W=1$	Category	$W=\mathbf{0}$	$W=1$
$Y=\mathbf{0}$	○	•	$Y=\mathbf{0}$	$(1-\pi_h)(1-p_h)$	$(1-\pi_h)p_h$
$Y=1$	•	•	$Y=1$	$\pi_h(1-p_h)$	$\pi_h p_h$

Table 7 The alternating table assigned to the second sample group of stratum h and the corresponding cell probability

Category	$W=1$	$W=\mathbf{0}$	Category	$W=1$	$W=\mathbf{0}$
$Y=\mathbf{0}$	○	•	$Y=\mathbf{0}$	$(1-\pi_h)p_h$	$(1-\pi_h)(1-p_h)$
$Y=1$	•	•	$Y=1$	$\pi_h p_h$	$\pi_h(1-p_h)$

π_h is the probability of having the sensitive attribute in stratum h and p_h is the probability of having the non-sensitive attribute in stratum h.

Let $\theta_{h1}=(1-\pi_h)(1-p_h)$ and $\theta_{h2}=(1-\pi_h)p_h$ be the probabilities of putting a tick in the circle in the first and second groups of stratum h, respectively and let s_{h1} and s_{h2} denote the numbers of ticks in the circle in the first and second sample groups of stratum h, respectively.

Since the selections in the different strata are made independently and the two groups of each stratum are independent, then the likelihood function is given by

$$\mathrm{L}=(1-\theta_{h1})^{n_{h1}-s_{h1}}\theta_{h1}^{s_{h1}}(1-\theta_{h2})^{n_{h2}-s_{h2}}\theta_{h2}^{s_{h2}}.$$

Setting $\frac{\partial}{\partial\theta_{h1}}\ln L=0$ and $\frac{\partial}{\partial\theta_{h2}}\ln L=0$, we get the MLEs, $\hat{\theta}_{hA1}=\frac{s_{h1}}{n_{h1}}$ and $\hat{\theta}_{hA2}=\frac{s_{h2}}{n_{h2}}$ of θ_{h1} and θ_{h2}, respectively. Thus, the MLEs of π_h and p_h are given by

$$\hat{\pi}_{hA}=1-\hat{\theta}_{hA1}-\hat{\theta}_{hA2}\ ,\quad \hat{p}_{hA}=\frac{\hat{\theta}_{hA2}}{\hat{\theta}_{hA1}+\hat{\theta}_{hA2}}. \tag{12}$$

The estimator $\hat{\pi}_{hA}$ given in (12) is an unbiased estimator of π_h with variance:

$$V(\hat{\pi}_{hA})=\frac{\theta_{h1}(1-\theta_{h1})}{n_{h1}}+\frac{\theta_{h2}(1-\theta_{h2})}{n_{h2}}. \tag{13}$$

An unbiased estimator of the variance of $\hat{\pi}_{hA}$ is given by

$$\hat{V}(\hat{\pi}_{hA})=\frac{\hat{\theta}_{hA1}(1-\hat{\theta}_{hA1})}{n_{h1}-1}+\frac{\hat{\theta}_{hA2}(1-\hat{\theta}_{hA2})}{n_{h2}-1}.$$

Theorem 4 *The MLE of the population proportion π is*

$$\hat{\pi}_{SA}=\sum_{h=1}^{k}W_h(1-\hat{\theta}_{hA1}-\hat{\theta}_{hA2}), \tag{14}$$

and $\hat{\pi}_{SA}$ is an unbiased estimator of the population proportion π with variance:

$$V\left(\hat{\pi}_{SA}\right)=\sum_{h=1}^{k}W_h^2\left(\frac{\theta_{h1}\left(1-\theta_{h1}\right)}{n_{h1}}+\frac{\theta_{h2}\left(1-\theta_{h2}\right)}{n_{h2}}\right). \tag{15}$$

Proof:

$$\hat{\pi}_{SA}=\sum_{h=1}^{k}W_h\hat{\pi}_{hA}=\sum_{h=1}^{k}W_h\left(1-\hat{\theta}_{hA1}-\hat{\theta}_{hA2}\right),$$

where $W_h=\frac{N_h}{N}$ is the weight of stratum h in the population, $\sum_{h=1}^{k}W_h=1$. The proof of unbiasedness is done directly by taking the expected values on both sides of (14).

$$E\left(\hat{\pi}_{SA}\right)=\sum_{h=1}^{k}W_hE\left(\hat{\pi}_{hA}\right)=\sum_{h=1}^{k}W_h\pi_h=\pi.$$

Since the selections are made independently in the different strata, then the variance of $\hat{\pi}_{SA}$ is

$$V\left(\hat{\pi}_{SA}\right)=V\left(\sum_{h=1}^{k}W_h\hat{\pi}_{hA}\right)=\sum_{h=1}^{k}W_h^2V(\hat{\pi}_{hA}). \tag{16}$$

Substituting $V\left(\hat{\pi}_{hA}\right)$ from (13) into (16) gives the variance of $\hat{\pi}_{SA}$ in (15).

Theorem 5 *An unbiased estimator of the variance of* $\hat{\pi}_{SA}$ *is given by*

$$\hat{V}\left(\hat{\pi}_{SA}\right)=\sum_{h=1}^{k}W_h^2\left(\frac{\hat{\theta}_{hA1}\left(1-\hat{\theta}_{hA1}\right)}{n_{h1}-1}+\frac{\hat{\theta}_{hA2}\left(1-\hat{\theta}_{hA2}\right)}{n_{h2}-1}\right). \tag{17}$$

4.2. Optimal allocation

In this subsection, we will present the optimal allocation of n_h into the two groups of stratum h and the optimal allocation of the sample size into the different strata.

4.2.1 Optimal allocation of n_h into the two groups

Here, we will determine the optimal sizes of n_{h1} and n_{h2} that minimize the variance of the estimator $\hat{\pi}_{hA}$.

Theorem 6 *The optimal sample sizes of the two groups in stratum* h, n_{h1} *and* n_{h2}, *are given by*

$$\frac{n_{h1}}{n_{h2}}=\frac{\sqrt{\theta_{h1}\left(1-\theta_{h1}\right)}}{\sqrt{\theta_{h2}\left(1-\theta_{h2}\right)}},\quad h=1,2,\ldots,k, \tag{18}$$

and the minimal variance of $\hat{\pi}_{hA}$ *is*

$$V_{opt}\left(\hat{\pi}_{hA}\right)=\frac{\left(\sqrt{\theta_{h1}\left(1-\theta_{h1}\right)}+\sqrt{\theta_{h2}\left(1-\theta_{h2}\right)}\right)^2}{n_h}. \tag{19}$$

Proof:

The optimal sample sizes n_{h1} and n_{h2} at which $V(\hat{\pi}_{hA})$ is minimum are obtained by equating $\frac{\partial}{\partial n_{h1}} V(\hat{\pi}_{hA})$ to zero. By substituting the optimal n_{h1} and n_{h2} in (18) into $V(\hat{\pi}_{hA})$ in (13), the minimal variance of $\hat{\pi}_{hA}$ in (19) is derived.

4.2.2 Optimal allocation of the sample size into the different strata

Under the optimal allocation of n_h into the two groups, the optimal allocation of the sample size n to $n_1, n_2, \ldots, n_k$ can be chosen to minimize $V(\hat{\pi}_{SA})$ for a specified linear cost function:

$$C = c_o + \sum_{h=1}^{k} c_h n_h,$$

where c_o is fixed cost and c_h is the cost per unit for each stratum h.

Theorem 7 *In stratified random sampling with a linear cost function, the optimal allocation of the sample size n, into the different strata, is given by*

$$n_h = \frac{(C - c_0)\frac{N_h}{\sqrt{c_h}}\left[\sqrt{\theta_{h1}(1-\theta_{h1})} + \sqrt{\theta_{h2}(1-\theta_{h2})}\right]}{\sum_{h=1}^{k}\sqrt{c_h} N_h \left[\sqrt{\theta_{h1}(1-\theta_{h1})} + \sqrt{\theta_{h2}(1-\theta_{h2})}\right]}, \quad h = 1, 2, \ldots, k,$$

and the minimal variance of $\hat{\pi}_{SA}$ is

$$V_{opt}(\hat{\pi}_{SA}) = \frac{1}{(C - c_0)}\left(\sum_{h=1}^{k}\sqrt{c_h} W_h \left[\sqrt{\theta_{h1}(1-\theta_{h1})} + \sqrt{\theta_{h2}(1-\theta_{h2})}\right]\right)^2.$$

We observe that the optimal allocation of the size of each stratum into the two groups and the optimal allocation of the sample size into the different strata depend on the values of π_h and p_h; $h = 1, 2, ..., k$ which are unknown. In applied studies, we can use values of π_h and p_h from a previous study or through a good guess.

Corollary 2 *Under Neyman allocation where the cost per unit for all strata are equal $c_1 = c_2 = \ldots = c_k = c^*$, the optimal allocation of the sample size n is given by*

$$n_h = n\frac{W_h\left[\sqrt{\theta_{h1}(1-\theta_{h1})} + \sqrt{\theta_{h2}(1-\theta_{h2})}\right]}{\sum_{h=1}^{k} W_h\left[\sqrt{\theta_{h1}(1-\theta_{h1})} + \sqrt{\theta_{h2}(1-\theta_{h2})}\right]},$$

and the minimal variance of $\hat{\pi}_{SA}$ for a fixed total sample size is

$$V_{Ney}(\hat{\pi}_{SA}) = \frac{1}{n}\left(\sum_{h=1}^{k} W_h\left[\sqrt{\theta_{h1}(1-\theta_{h1})} + \sqrt{\theta_{h2}(1-\theta_{h2})}\right]\right)^2.$$

5. The Stratified Triangular Model Under Non-Compliance

In this section, the triangular model under non-compliance, using alternating tables, presented by Wu and Tang (2016) is extended to stratified random sampling. Hence, the design and the ML estimation of the stratified triangular model under non-compliance are introduced in Subsection 5.1. Subsection 5.2 considers the optimal allocation of the sample size.

5.1. Design and estimation of the stratified triangular model under non-compliance

As in the previous sections, the population is assumed to be divided into k strata where stratum h is of known size N_h, $h=1,2,...,k$. From each stratum h, a SRSWR of size n_h is drawn, then it is randomly split into two groups of sizes n_{h1} and n_{h2} (i.e. $n_{h1}+n_{h2}=n_h$) where $n_{h1}>1$ and $n_{h2}>1$. It is assumed that the selections in the different strata are made independently. Table 8 is given to the first sample group of stratum h while Table 9 is assigned to the second sample group of stratum h.

Table 8 The table assigned to the first sample group of stratum h and the corresponding probabilities

Category	$W=\mathbf{0}$	$W=1$	Probabilities
$Y=\mathbf{0}$	○	•	$\Pr(\text{Ticking the circle})=(1-\pi_h)(1-p_h)+\pi_h\gamma_h$
$Y=1$	•	•	$\Pr(\text{Ticking the triangle})=1-(1-\pi_h)(1-p_h)-\pi_h\gamma_h$

Table 9 The alternating table assigned to the second group of stratum h and the corresponding probabilities

Category	$W=1$	$W=\mathbf{0}$	Probabilities
$Y=\mathbf{0}$	○	•	$\Pr(\text{Ticking the circle})=(1-\pi_h)p_h+\pi_h\gamma_h$
$Y=1$	•	•	$\Pr(\text{Ticking the triangle})=1-(1-\pi_h)p_h-\pi_h\gamma_h$

π_h is the probability of having the sensitive attribute in stratum h, p_h is the probability of having the non-sensitive attribute in stratum h where $0<p_h<1$ and γ_h represents the probability that the respondent in stratum h provides the non-sensitive answer $(Y=0)$ while he/she possesses the sensitive attribute.

Let $\psi_{h1}=(1-\pi_h)(1-p_h)+\pi_h\gamma_h$ and $\psi_{h2}=(1-\pi_h)p_h+\pi_h\gamma_h$ be the probabilities of putting a tick in the circle in the first and second groups of stratum h, respectively and let s_{h1} and s_{h2} denote the numbers of ticks in the circle in the first and second sample groups of stratum h, respectively.

Following the same estimation method used in the previous two sections, the following ML estimators of π_h and γ_h are obtained, respectively, as follows:

$$\hat{\pi}_{hC}=\frac{\hat{\psi}_{hC1}-\hat{\psi}_{hC2}}{2p_h-1}+1;\quad p_h\neq\frac{1}{2}, \tag{20}$$

$$\hat{\gamma}_{hC}=\frac{\hat{\psi}_{hC2}(p_h-1)+p_h\hat{\psi}_{hC1}}{\hat{\psi}_{hC1}-\hat{\psi}_{hC2}+2p_h-1},$$

where $\hat{\psi}_{hC1}=\frac{s_{h1}}{n_{h1}}$ and $\hat{\psi}_{hC2}=\frac{s_{h2}}{n_{h2}}$ are the ML estimators, of ψ_{h1} and ψ_{h2} respectively. The estimator $\hat{\pi}_{hC}$ given in (20) is an unbiased estimator of π_h with variance:

$$V\left(\hat{\pi}_{hC}\right)=\frac{1}{\left(2p_h-1\right)^2}\left(\frac{\psi_{h1}\left(1-\psi_{h1}\right)}{n_{h1}}+\frac{\psi_{h2}\left(1-\psi_{h2}\right)}{n_{h2}}\right);\ p_h\neq\frac{1}{2}.$$

An unbiased estimator of the variance of $\hat{\pi}_{hC}$ is given by

$$\hat{V}\left(\hat{\pi}_{hC}\right)=\frac{1}{\left(2p_h-1\right)^2}\left(\frac{\hat{\psi}_{hC1}\left(1-\hat{\psi}_{hC1}\right)}{n_{h1}-1}+\frac{\hat{\psi}_{hC2}\left(1-\hat{\psi}_{hC2}\right)}{n_{h2}-1}\right);\ p_h\neq\frac{1}{2}.$$

Theorem 8 *The ML estimator of the population proportion* π *is*

$$\hat{\pi}_{SC}=\sum_{h=1}^{k}W_h\left(\frac{\hat{\psi}_{hC1}-\hat{\psi}_{hC2}}{2p_h-1}+1\right);\ p_h\neq\frac{1}{2}. \tag{21}$$

The estimator $\hat{\pi}_{SC}$ *given in (21) is an unbiased estimator of the population proportion* π *with variance given by*

$$V\left(\hat{\pi}_{SC}\right)=\sum_{h=1}^{k}W_h^2\ \frac{1}{\left(2p_h-1\right)^2}\left(\frac{\psi_{h1}\left(1-\psi_{h1}\right)}{n_{h1}}+\frac{\psi_{h2}\left(1-\psi_{h2}\right)}{n_{h2}}\right);\ p_h\neq\frac{1}{2}.$$

Theorem 9 *An unbiased estimator of the variance of* $\hat{\pi}_{SC}$ *is given by*

$$\hat{V}\left(\hat{\pi}_{SC}\right)=\sum_{h=1}^{k}\frac{W_h^2}{\left(2p_h-1\right)^2}\left(\frac{\hat{\psi}_{hC1}\left(1-\hat{\psi}_{hC1}\right)}{n_{h1}-1}+\frac{\hat{\psi}_{hC2}\left(1-\hat{\psi}_{hC2}\right)}{n_{h2}-1}\right);\ p_h\neq\frac{1}{2}.$$

5.2. Optimal allocation

In this subsection, we will present the optimal allocation of n_h into the two groups of stratum h and the optimal allocation of the sample size into the different strata.

5.2.1 Optimal allocation of n_h into the two groups

Theorem 10 *The optimal sample sizes of the two groups in stratum* h, n_{h1} *and* n_{h2}, *that minimize the variance of the estimator* $\hat{\pi}_{hC}$ *is given by*

$$\frac{n_{h1}}{n_{h2}}=\frac{\sqrt{\psi_{h1}\left(1-\psi_{h1}\right)}}{\sqrt{\psi_{h2}\left(1-\psi_{h2}\right)}},\quad h=1,2,\ldots,k,$$

and the minimal variance of $\hat{\pi}_{hC}$ *is*

$$V_{opt}\left(\hat{\pi}_{hC}\right)=\frac{1}{n_h\left(2p_h-1\right)^2}\left(\sqrt{\psi_{h1}\left(1-\psi_{h1}\right)}+\sqrt{\psi_{h2}\left(1-\psi_{h2}\right)}\right)^2;\ p_h\neq\frac{1}{2}.$$

5.2.2 Optimal allocation of the sample size into different strata

Under the optimal allocation of n_h into the two groups, the optimal allocation of the sample size n to $n_1,n_2,\ldots,n_k$ can be chosen to minimize $V\left(\hat{\pi}_{SC}\right)$ for a specified linear cost function:

$$C = c_o + \sum_{h=1}^{k} c_h n_h,$$

where c_o is fixed cost and c_h is the cost per unit for each stratum h.

Theorem 11 *In stratified random sampling with a linear cost function, the optimal allocation of the sample size n, into the different strata is given by*

$$n_h = \frac{(C-c_0)\frac{N_h}{\sqrt{c_h}(2p_h-1)^2}\left[\sqrt{\psi_{h1}(1-\psi_{h1})}+\sqrt{\psi_{h2}(1-\psi_{h2})}\right]}{\sum_{h=1}^{k}\frac{N_h\sqrt{c_h}}{(2p_h-1)^2}\left[\sqrt{\psi_{h1}(1-\psi_{h1})}+\sqrt{\psi_{h2}(1-\psi_{h2})}\right]}; \; p_h \neq \frac{1}{2},$$

and the minimal variance of $\hat{\pi}_{SC}$ is

$$V_{opt}(\hat{\pi}_{SC}) = \frac{1}{(C-c_0)}\left(\sum_{h=1}^{k}\frac{\sqrt{c_h}W_h}{(2p_h-1)}\left[\sqrt{\psi_{h1}(1-\psi_{h1})}+\sqrt{\psi_{h2}(1-\psi_{h2})}\right]\right)^2; \; p_h \neq \frac{1}{2}.$$

It can be noticed that the optimal allocation of the size of each stratum into the two groups and the optimal allocation of the sample size into the different strata depend on the values of π_h and γ_h; $h = 1, 2, ..., k$ which are unknown. In applied studies, we can use values of π_h and γ_h from a previous study or through a good guess.

Corollary 3 *Under Neyman allocation, the optimal allocation of the sample size n is given by*

$$n_h = n\frac{\frac{W_h}{(2p_h-1)}\left[\sqrt{\psi_{h1}(1-\psi_{h1})}+\sqrt{\psi_{h2}(1-\psi_{h2})}\right]}{\sum_{h=1}^{k}\frac{W_h}{(2p_h-1)}\left[\sqrt{\psi_{h1}(1-\psi_{h1})}+\sqrt{\psi_{h2}(1-\psi_{h2})}\right]}; \; p_h \neq \frac{1}{2},$$

and the minimal variance of $\hat{\pi}_{SC}$ for fixed total sample size is:

$$V_{Ney}(\hat{\pi}_{SC}) = \frac{1}{n}\left(\sum_{h=1}^{k}\frac{W_h}{(2p_h-1)}\left[\sqrt{\psi_{h1}(1-\psi_{h1})}+\sqrt{\psi_{h2}(1-\psi_{h2})}\right]\right)^2; \; p_h \neq \frac{1}{2}.$$

6. Conclusion

In this paper, the triangular model by Yu et al. (2008) is extended to stratified random sampling. In addition, the stratified extensions of two variants of the triangular model are presented. For the three extended models, the ML estimators of the population proportion possessing the sensitive attribute are derived and their sampling variances are obtained. For the stratified triangular model, the optimal allocation of the sample size is obtained and the corresponding minimum variance is derived. For the stratified variants of the triangular model, the optimal allocation of the stratum size, into the two groups, that minimizes the variance of the estimator in each stratum is obtained together with the optimal allocation of the sample sizes, for which the variances of the estimators are at their minimum.

Statements and Declarations

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Chaudhuri A. Unbiased estimation of a sensitive proportion in general sampling by three nonrandomized response schemes. J Stat Theory Pract. 2012; 6(2): 376-381.

Groenitz H. A new privacy-protecting survey design for multichotomous sensitive variables. Metrika. 2014a; 77(2): 211-224.

Groenitz H. Applying the nonrandomized diagonal model to estimate a sensitive distribution in complex sample surveys. J Stat Theory Pract. 2014b; 8(2): 319-342.

Jerke J, Krumpal I. Plagiarism in student papers: an empirical study using the triangular model. Methods Data Anal. 2013; 7(3): 347-368.

Swensson B. Combined questions: a new survey technique for eliminating evasive answer bias (I)—basic theory. Report No. 70 of the Errors in Surveys Research Project. Stockholm: Institute of Statistics, University of Stockholm; 1974.

Takahasi K, Sakasegawa H. A randomized response technique without making use of any randomizing device. Ann Inst Stat Math. 1977; 29(1): 1-8.

Tan M, Tian GL, Tang ML. Sample surveys with sensitive questions: a nonrandomized response approach. Am Stat. 2009; 63(1): 9-16.

Tian GL. A new non-randomized response model: the parallel model. Stat Neerl. 2014; 68(4): 293-323.

Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. J Am Stat Assoc. 1965; 60(309): 63-69.

Wu Q, Tang ML. Flexible non-randomized response models for survey with sensitive question. Stat Med. 2014; 33(6): 918-929.

Wu Q, Tang ML. Non-randomized response model for sensitive survey with noncompliance. Stat Methods Med Res. 2016; 25(6): 2827-2839.

Yu JW, Tian GL, Tang ML. Two new models for survey sampling with sensitive characteristic: design and analysis. Metrika. 2008; 67(3): 251-263.