An Estimation of Sensitive Attribute Applying Geometric Distribution under Probability Proportional to Size Sampling

Gi-Sung Lee 1, Ki-Hak Hong 2 and Chang-Kyoon Son 3,*

1 Department of Children Welfare, Woosuk University, Wanju Jeonbuk 55338, Korea; gisung@woosuk.ac.kr
2 Department of Computer Science, Dongshin University, Naju Jeonnam 58245, Korea; khhong@dsu.ac.kr
3 Department of Applied Statistics, Dongguk University, Gyeongju Gyeongbuk 38066, Korea
* Correspondence: sonchangkyoon@gmail.com

Received: 29 August 2019; Accepted: 31 October 2019; Published: 14 November 2019

Abstract: In this paper, we extended Yennum et al.’s model, in which geometric distribution is used as a randomization device for a population that consists of different-sized clusters, and clusters are obtained by probability proportional to size (PPS) sampling. Estimators of a sensitive parameter, their variances, and their variance estimators are derived under PPS sampling and equal probability two-stage sampling, respectively. We also applied these sampling schemes to Yennum et al.’s generalized model. Numerical studies were carried out to compare the efficiencies of the proposed sampling methods for each case of Yennum et al.’s model and Yennum et al.’s generalized model.

Keywords: probability proportional to size (PPS) sampling; geometric distribution; sensitive attribute; randomization device; Yennum et al.’s model

1. Introduction

The randomized response model (RRM) was suggested by [1] to estimate the true population proportion of sensitive characteristics, such as illegal gambling, drug-abuse, tax evasion, the extent of illegal income, and the experience of abortion, among others [2–4].

Since Warner’s work, many scholars have developed the RRM in various ways. In [5,6], they arranged, summarized, and systemized various RRMs and emphasized their importance. In [7], sampling survey of sensitive attributes applied two-stage cluster sampling to RRM for a population consisting of equal-sized clusters, and [8] considered the cluster RRM for a population consisting of different-sized clusters, where the clusters are selected by probability proportional to size (PPS) sampling.

Recently, Yennum et al. [9] suggested a new randomization device to gather sensitive data in two-stages under the assumption of geometric distribution and made a generalization of their model encompassing generalized geometric distribution using [10] model.

Based on Yennum et al.’s work, it is assumed that the respondents are selected by simple random sampling with replacements, but a real survey selects respondents from various sampling schemes.

Now, we can consider a large sample of clusters. For example, to estimate the true population proportion of drug-abuse among high school students, it is possible to use a randomization device like Yennum et al.’s model via proportional sampling by considering the primary sampling unit as the school and the secondary sampling unit as the students.

From this point of view, we extend Yennum et al.’s model, in which geometric distribution is used as a randomization device based on a population that consists of different-sized clusters, and the clusters are selected by PPS sampling. Estimators of a sensitive parameter, their variances, and their variance estimators are derived by PPS sampling and equal probability two-stage sampling, respectively.
We also apply these methods to the case of Yennum et al.’s generalized model. Numerical studies are carried out to compare the efficiencies of the suggested methods in each case of Yennum et al.’s model and Yennum et al.’s generalized model.

2. An Estimation of Sensitive Attributes with Probability Proportional to Size Sampling under Yennum et al.’s Model

In Section 2, we consider a new sampling scheme to estimate sensitive attributes using Yennum et al.’s model, in which geometric distribution is used as a randomization device when \( n \) clusters are selected with proportional to size (PPS) sampling or equal probability sampling from a population that consists of \( N \) clusters of size, \( m_i(i = 1, 2, \cdots, N) \) and \( m_i(i = 1, 2, \cdots, n) \) units are selected by simple random sampling from each sampled cluster.

In Section 2.1, we consider the sampling method for the clusters via PPS sampling with replacements. Clusters by PPS sampling without replacement are considered in Section 2.2, and clusters by equal probability sampling are examined in Section 2.3.

2.1. PPS Sampling with Replacement

Let the population be composed of \( N \) clusters. In the first stage, the size of the \( n \) sample of the first sampling units (FSU) is selected with replacement by the selection probability \( p_i \) for the \( i \)th cluster. In the second stage, \( m_i \) second sampling units (SSU) are drawn by simple random sampling with replacement (SRSWR) from each FSU and are guided to carry out Yennum et al.’s randomization device.

First of all, the randomization device consists of two elements. The first randomization device for the \( i \)th cluster consists of two kinds of urns with white and black balls, where the selection probability of a white ball is \( W_i \) and the selection probability of a black ball is \( 1 - W_i \).

On the other hand, the second randomization device is composed of two kinds of urns with balls. The first device with balls contains a slip of paper including two statements, such as \("I have a sensitive attribute" with selection probability \( P_i \), and the other balls includes a statement such as \("I do not have a sensitive attribute" with selection probability \( 1 - P_i \). The second device with balls contains a slip of paper with the statement \("I do not have a sensitive attribute" with selection probability \( T_i \) and balls with the statement \("I have a sensitive attribute" with selection probability \( 1 - T_i \).

In the first stage, for the \( i \)th cluster, each interviewee draws a ball from the first randomization device, such as the urn with the white and black balls. When he or she selects a white ball, he or she is guided to pick balls from the first urn of the second randomization device, one after another, with replacement, until the first ball containing a statement matching his or her own status appears.

We assume that \( X_{i1} \) is the total number of balls drawn before he or she obtains the first ball including his or her own status in the \( i \)th cluster, and \( X_{i2} \) is the total number of balls drawn before he or she obtains the first ball with his or her own status of not having a sensitive attribute in the \( i \)th cluster. Similarly, when he or she draws a black ball, he or she is guided to pick balls from the second urn of the second randomization device, one after another, with replacement, until the first ball containing a statement matching his or her own status appears.

For the \( i \)th cluster, using the randomization device in Figure 1, the total number of balls taken by interviewees \( X_{i1}, X_{i2}, Y_{i1}, Y_{i2} \) are distributed via generalized geometric distribution. Let \( \pi_i \) and \( 1 - \pi_i \) be the true population proportion of persons who have a sensitive attribute \( A_i \) and \( A_i' \) for the \( i \)th cluster. Assume that each interviewee in the \( i \)th cluster is drawn by SRSWR.

For the \( i \)th cluster, the total number for each ball selected by interviewees through the proposed two-stage device distributes one of the following random variables: \( X_{i1} \sim Ge(P_i), X_{i2} \sim Ge(1 - P_i), Y_{i1} \sim Ge(T_i) \) and \( Y_{i2} \sim Ge(1 - T_i) \), where \( Ge(\cdot) \) represents the geometric distribution with a success probability. Let \( \pi_i \) and \( 1 - \pi_i \) be the true population proportions of persons who have a sensitive attribute \( (A_i \) and \( A_i' \) respectively) for the \( i \)th cluster. Assume that each interviewee in the \( i \)th cluster is drawn by SRSWR.
Let $Z_{ij}$ be the $j$th observed answer in the $i$th cluster; $Z_{ij}$ can be expressed as

$$Z_{ij} = \begin{cases} X_{i1}, & \text{with probability } W_i \pi_i \\ Y_{i2}, & \text{with probability } (1 - W_i)\pi_i \\ X_{i2}, & \text{with probability } W_i(1 - \pi_i) \\ Y_{i1}, & \text{with probability } (1 - W_i)(1 - \pi_i) \end{cases}$$  \quad (1)$$

Then, we can find the expected value of $Z_{ij}$ as follows:

$$E(Z_{ij}) = \pi_i \left[ \frac{W_i}{T_i} + \frac{(1-W_i)}{(1-T_i)} \right] + (1 - \pi_i) \left[ \frac{W_i}{T_i} + (1-W_i) \right] + \frac{(1-W_i)}{T_i}.$$  \quad (2)

The expected value (2) can be expressed as follows:

$$\frac{(1-T_i)P_i}{P_iT_i(1-P_i)(1-T_i)} \left[ E(Z_{ij}) (1-P_i)T_i - W_iT_i - (1-W_i)(1-P_i) \right] = \pi_i \psi_i,$$  \quad (3)

where $\psi_i = W_i(1 - 2P_i)T_i(1 - T_i) + (1 - W_i)(2T_i - 1)P_i(1 - P_i)$.

Now the estimator $\hat{\pi}_i$ for the true population proportion $\pi_i$ in the $i$th cluster is given by:

$$\hat{\pi}_i = \frac{1}{\psi_i} \left[ P_iT_i(1-P_i)(1-T_i) \frac{1}{M_i} \sum_{j=1}^{m} Z_{ij} - W_iT_iP_i(1-T_i) - P_i(1-W_i)(1-P_i)(1-T_i) \right].$$  \quad (4)

When the interviewees are drawn by SRSWR from the $i$th cluster selected with a replacement by the sampling probability $p_{ii}$, the estimator $\hat{\pi}_{ppswr}$ of the true population proportion $\pi$ for a sensitive character is given by:

$$\hat{\pi}_{ppswr} = \frac{1}{M_0} \sum_{i=1}^{n} \frac{M_i \hat{\pi}_i}{P_i},$$

$$= \frac{1}{M_0} \sum_{i=1}^{n} \frac{M_i}{P_i} \left[ P_iT_i(1-P_i)(1-T_i) \frac{1}{M_i} \sum_{j=1}^{m} Z_{ij} - W_iT_iP_i(1-T_i) - P_i(1-W_i)(1-P_i)(1-T_i) \right].$$  \quad (5)

where $M_0 = \sum_{i=1}^{N} M_i$.

**Theorem 1:** The estimator $\hat{\pi}_{ppswr}$ of the true population proportion of a sensitive attribute $\pi$ under PPS with a replacement sampling scheme is an unbiased estimator.
Proof:

\[ E_1 E_2(\hat{\pi}_{pswr}) = E_1 E_2 \left[ \frac{1}{m_0} \sum_{i=1}^{n} \frac{M_i \pi_i}{\hat{p}_i} \right] \]

and since:

\[ E_2(\hat{\pi}) = \frac{1}{\psi_n} \left[ P_i T_i (1 - P_i) (1 - T_i) \right] \]

we can obtain:

\[ E_1 E_2(\hat{\pi}_{pswr}) = E_1 \left[ \frac{1}{m_0} \sum_{i=1}^{n} \frac{M_i \pi_i}{\hat{p}_i} \right] \]

\[ = \frac{1}{n m_0} \sum_{i=1}^{n} \frac{M_i \pi_i}{\hat{p}_i} \]

\[ = \pi. \]

\[ \square \]

**Theorem 2:** The variance of \( \hat{\pi}_{pswr} \) is obtained from a two-stage procedure, such that a sample of size \( n \) FSU is selected by replacement with sampling probability \( p_j \) for the unit \( i \) from the population of \( N \) clusters with size \( M_i \) elements in the \( i \)th cluster, and the SSUs with size \( m_i \) are drawn by SRSWR from each FSU, as given by:

\[ V(\hat{\pi}_{pswr}) = \frac{1}{n m_0} \sum_{i=1}^{n} \frac{M_i \pi_i}{\hat{p}_i} - M_0 \pi_0 \]

\[ + \frac{1}{n m_0} \sum_{i=1}^{n} \frac{M_i^2 \pi_i}{m_i \hat{p}_i} \left[ \pi_i (1 - \pi_i) - \frac{\pi_i^2}{\psi_i} A_i + \frac{1}{\psi_i} B_i \right] \] (6)

where:

\[ A_i = W_i (1 - 2 \pi_i) (2 - P_i + P_i^2) T_i^2 (1 - T_i)^2 + (1 - W_i) (2 T_i - 1)(2 - T_i + T_i^2) P_i^2 (1 - P_i)^2 \]

\[ + W_i^2 T_i^2 (1 - T_i)^2 (2T_i - 1)^2 + (1 - W_i)^2 P_i^2 (1 - P_i)^2 (1 - 2T_i) \]

\[ + 2 W_i (1 - W_i) P_i T_i (1 - P_i) (1 - T_i) (P_i - T_i), \]

\[ B_i = (1 - T_i)^2 P_i^2 \left[ W_i (1 - W_i) (P_i + T_i - 1)^2 + W_i P_i T_i \right] \]

\[ + (1 - W_i) (1 - T_i) (1 - P_i)^2. \]

Proof: Given \( X_{i1} \sim G(P_i), X_{i2} \sim G(1 - P_i), Y_{i1} \sim G(T_i), Y_{i2} \sim G(1 - T_i), \) where \( G \) represents the geometric distribution with a success probability. Since the expected values of \( Z_{ij} \) and \( Z_{ij}^2 \) are

\[ E(Z_{ij}) = \pi_i \left[ \frac{W_i}{\hat{p}_i} + \frac{(1 - W_i)}{(1 - T_i)} \right] + (1 - \pi_i) \left[ \frac{W_i}{(1 - P_i)} + \frac{(1 - W_i)}{T_i} \right], \] (7)

then:

\[ E(Z_{ij}^2) = \pi_i \left[ \frac{W_i (2 - P_i)}{\hat{p}_i^2} + \frac{(1 - \pi_i)}{(1 - P_i)^2} \right] + (1 - \pi_i) \left[ \frac{W_i (1 + P_i)}{(1 - P_i)^2} \right] + \pi_i (1 - W_i) \left[ \frac{2 - T_i}{T_i^2} \right] + \pi_i (1 - W_i) \left[ \frac{1 + T_i}{1 - T_i} \right]. \] (8)

Based on (7) and (8), the variance of \( Z_{ij} \) is:

\[ \sigma_{Z_{ij}}^2 = E(Z_{ij}^2) - \left[ E(Z_{ij}) \right]^2, \] (9)
and, since $Z_{ij}$ is independent, the variance of $\hat{\pi}_i$ can be expressed by:

$$V(\hat{\pi}_i) = V\left[\frac{1}{m_i^p} \left( \sum_{j=1}^{m_i} Z_{ij} - W_i T_i (1 - T_i) - P_i (1 - W_i) (1 - P_i) (1 - T_i) \right) \right]$$

$$= \frac{p_i^2 (1 - p_i)^2 (1 - T_i)^2}{m_i^p} \sum_{j=1}^{m_i} \frac{a^2_{ij}}{\psi_i^2}$$

$$= \frac{\pi_i (1 - \pi_i)}{m_i} + \frac{\pi_i}{m_i \psi_i^2} A_i + \frac{1}{m_i \psi_i^2} B_i. \quad (10)$$

Since $V(\hat{\pi}_{ppswr}) = V_1 E_2(\hat{\pi}_{ppswr}) + E_1 V_2(\hat{\pi}_{ppswr})$, then the first and second terms are given, respectively, as:

$$V_1 E_2(\hat{\pi}_{ppswr}) = V_1 E_2\left[ \frac{1}{nM_0^p} \sum_{i=1}^{n} \frac{M_i \pi_i}{p_i} \right]$$

$$= V_1\left[ \frac{1}{nM_0^p} \sum_{i=1}^{n} \frac{M_i \pi_i}{p_i} \right]$$

and

$$E_1 V_2(\hat{\pi}_{ppswr}) = E_1 V_2\left[ \frac{1}{nM_0^p} \sum_{i=1}^{n} \frac{M_i^2 \pi_i}{p_i^2} V_2(\hat{\pi}_i) \right]$$

$$= E_1\left[ \frac{1}{nM_0^p} \sum_{i=1}^{n} \frac{M_i^2 \pi_i}{p_i^2} V_2\left( \frac{1}{M_i / p_i^2} \sum_{j=1}^{m_i} Z_{ij} - W_i T_i (1 - T_i) - P_i (1 - W_i) (1 - P_i) (1 - T_i) \right) \right]$$

Then, we can obtain the variance (10).

Moreover, an unbiased estimator of $V(\hat{\pi}_{ppswr})$ is given by

$$\hat{V}(\hat{\pi}_{ppswr}) = \frac{1}{nM_0^p} \sum_{i=1}^{n} p_i \left[ \frac{M_i \pi_i}{p_i} - M_0 \hat{\pi}_{ppswr} \right]^2$$

$$+ \frac{1}{nM_0^p} \sum_{i=1}^{n} \frac{M_i^2 \pi_i}{p_i (m_i - 1)} \left[ \hat{\pi}_i (1 - \pi_i) - \frac{\pi_i}{\psi_i^2} A_i + \frac{1}{\psi_i^2} B_i \right]. \quad (11)$$

\[\square\]

If the FSUs are selected proportional to size with $M_i$, then $p_i = M_i / M_0$. For this reason, we call this method “probability proportional to size” (PPS) sampling. When a sample of the FSU is selected by PPS sampling with replacement via sampling probability, $p_i = M_i / M_0$ for the $i$th cluster, and $m_i$ SSU are selected by SRSWRR from each FSU. The estimator $\hat{\pi}_{ppswr}$ of $\pi$ is given by:

$$\hat{\pi}_{ppswr} = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_i$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \left[ \hat{\pi}_i (1 - \pi_i) + \frac{\pi_i}{\psi_i^2} A_i + \frac{1}{\psi_i^2} B_i \right], \quad (12)$$

and the variance of $\hat{\pi}_{ppswr}$ and its estimator are as follows:

$$V(\hat{\pi}_{ppswr}) = \frac{1}{nM_0^p} \sum_{i=1}^{N} M_i (\pi_i - \hat{\pi}_i)^2$$

$$+ \frac{1}{nM_0^p} \sum_{i=1}^{N} \frac{M_i}{m_i} \left[ \hat{\pi}_i (1 - \pi_i) + \frac{\pi_i}{\psi_i^2} A_i + \frac{1}{\psi_i^2} B_i \right], \quad (13)$$
where \( \hat{\pi}_{ppswr} \) is the second inclusion probability of the \( ij \)th and \( j \)th clusters.

Furthermore, the variance estimator of \( \hat{\pi}_{ppswr} \) is as follows:

\[
V(\hat{\pi}_{ppswr}) = \frac{1}{M_0^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \frac{\theta_{ij} \theta_{ij} - \theta_{ij}}{\theta_{ij}} \right) \left[ \frac{M_i \hat{\pi}_{ij}}{\theta_{ij}} - \frac{M_j \hat{\pi}_{ij}}{\theta_{ij}} \right]^2 \\
+ \frac{1}{M_0^2} \sum_{i=1}^{N} \left[ \hat{\pi}_{ij} (1 - \hat{\pi}_{ij}) + \frac{\hat{\pi}_{ij} A_j}{\psi_j} + \frac{1}{\psi_j} B_j \right].
\]

2.3. Two-Stage Equal Probability Sampling

In this subsection, we consider a two-stage equal probability sampling design to estimate the true population proportion of a sensitive characteristic by applying Yenum et al.’s model, in which \( n \) FSUs are drawn by simple random sampling without replacement (SRSWR) from a population of \( N \) clusters with \( M_i \) elementary units for the \( i \)th cluster, and \( m_i \) SSUs are drawn by SRSWR from each FSU.

From this two-stage sampling, the estimator \( \hat{\pi}_{var} \) of \( \pi \) is given by:

\[
\hat{\pi}_{var} = \frac{N}{nM_0} \sum_{i=1}^{n} M_i \hat{\pi}_{ij},
\]

where \( \hat{\pi}_{ij} \) is an estimator of the true population proportion for a sensitive characteristic for the \( ij \)th cluster, which is the same as (4).

The variance of \( \hat{\pi}_{var} \) and its estimator are given as:

\[
V(\hat{\pi}_{var}) = \frac{N^2}{nM_0^2 (N-1)} \sum_{i=1}^{N} (M_i \pi_i - \bar{M} \pi)^2 \\
+ \frac{N}{nM_0^2} \sum_{i=1}^{N} M_i^2 \left[ \pi_i (1 - \pi_i) + \frac{\pi_i A_j}{\psi_j} + \frac{1}{\psi_j} B_j \right],
\]
where $\overline{M} = M_0/N$.

3. An Estimation of Sensitive Attributes with Probability Proportional to Size Sampling Under Yennum et al.’s Generalized Model

We consider Yennum et al.’s generalized model, in which generalized geometric distribution is used as a randomization device when $n$ clusters are sampled by PPS sampling or equal probability sampling from the population, which consists of $N$ clusters with size $M_i (i = 1, 2, \cdots, N)$, and $m_i (i = 1, 2, \cdots, n)$ units are drawn by simple random sampling from each sampled cluster.

We develop the sampling schemes for PPS sampling with replacement in Section 3.1 and those for PPS sampling without replacement in Section 3.2. Finally, equal probability sampling is presented in Section 3.3.

3.1. PPS Sampling with Replacement

Let the population be composed of $N$ clusters. In the first stage, a sample of $n$ FSUs is drawn by replacement with the sampling probability $p_i$ for the $i$th cluster. In the second stage, $m_i$ SSUs are selected by SRSWR from each FSU and guided to apply Yennum et al.’s generalized randomization device.

If the interviewees in the $i$th cluster choose a white ball during the first stage, and if they have a sensitive attribute $A$ (or $A^c$), then they are guided to pick replacement balls from the first urn of the second stage device until they take $k_{i2}$ (or $k_{i1}$) successive balls with their actual status for the first time and are then asked to determine the total number of balls as $X_{i1}$ (or $X_{i2}$).

If the interviewee in the $i$th cluster draws a black ball in the first stage, and if they have a sensitive attribute $A^c$ (or $A$), then they are guided to take replacement balls from the second urn of the second stage device until they take $k_{i2}$ (or $k_{i1}$) successive balls with their actual status for the first time and are then asked to determine the total number of balls as $Y_{i1}$ (or $Y_{i2}$).

For the $i$th cluster, using the randomization device in Figure 1, the total number of balls taken by interviewees $X_{i1}, X_{i2}, Y_{i1},$ and $Y_{i2}$ are distributed via generalized geometric distribution. Let $\pi_i$ and $1 - \pi_i$ be the true population proportion of persons who have a sensitive attribute $A$ and $A^c$ for the $i$th cluster. Assume that each interviewee in the $i$th cluster is drawn by SRSWR.

For the $j$th surveyed answer in the $i$th cluster, $Z_{ij}$ can be expressed as:

$$Z_{ij} = \begin{cases} X_{i1} & \text{with probability } W_i \pi_i, \\ Y_{i2} & \text{with probability } (1 - W_i) \pi_i, \\ X_{i2} & \text{with probability } W_i (1 - \pi_i), \\ Y_{i1} & \text{with probability } (1 - W_i) (1 - \pi_i), \end{cases}$$

(21)

The expected value of $Z_{ij}$ is given by:

$$E(Z_{ij}) = W_i \pi_i E(X_{i1}) + \pi_i (1 - W_i) E(Y_{i2}) + (1 - \pi_i) W_i E(X_{i2}) + (1 - W_i) (1 - \pi_i) E(Y_{i1})$$

$$\quad = \pi_i W_i \left[ \frac{1-p_{i1}^2}{p_{i1}(1-p_{i1})} - \frac{1-(1-p_{i2})^2}{p_{i1}(1-p_{i1})} \right] + (1 - W_i) \left[ \frac{1-(1-T_{i2})^2}{(1-T_{i2})^2} - \frac{1-T_{i2}^2}{(1-T_{i2})^2} \right]$$

$$\quad + W_i \left[ \frac{1-(1-p_{i2})^2}{p_{i1}(1-p_{i1})} \right] + (1 - W_i) \left[ \frac{1-T_{i2}^2}{(1-T_{i2})^2} \right].$$

(22)
Then, the formula (22) can be expressed as:

\[
E(Z_{ij}) - W_i \left\{ \frac{1-(1-P_i)^k_z}{p_i (1-P_i)^{2z}} \right\} - (1-W_i) \left\{ \frac{1-T_z}{(1-T_z)^{2z}} \right\} = \pi_i \left[ W_i \left\{ \frac{1-P_i}{(1-P_i)^{2z}} - \frac{1-(1-P_i)^k_z}{p_i (1-P_i)^{2z}} \right\} + (1-W_i) \left( \frac{1-(1-P_i)^k_z}{(1-T_z)^{2z}} - \frac{1-T_z}{(1-T_z)^{2z}} \right) \right].
\]  

(23)

The estimator \( \hat{\pi}_G \) of the population proportion \( \pi_i \) for the \( i \)th cluster is given by:

\[
\hat{\pi}_G = \frac{1}{m_i \phi_{12}} \left( \sum_{j=1}^{m_i} Z_{ij} - \phi_{1} \right),
\]

where:

\[
\phi_{1} = W_i \left\{ 1 - (1-P_i)^k_z \right\} \left[ \frac{1-T_z}{(1-T_z)^{2z}} + (1-W_i) \right] T_i (1-T_z)^{2z} + (1-P_i) P_i \]

and:

\[
\phi_{2} = W_i \left\{ 1 - (1-P_i)^k_z \right\} \left[ \frac{1-(1-P_i)^k_z}{p_i (1-P_i)^{2z}} \right] P_i \]  

(25)

When the interviewees are sampled by SRSWR for the \( i \)th cluster selected with a replacement by sampling probability \( p_i \), the estimator \( \hat{\pi}_{G_{ppswr}} \) of the true population proportion \( \pi \) of a sensitive attribute is:

\[
\hat{\pi}_{G_{ppswr}} = \frac{1}{n M_0} \sum_{i=1}^{n} \frac{M_i \hat{\pi}_G}{p_i} = \frac{1}{n M_0} \sum_{i=1}^{n} \frac{M_i}{p_i} \left( \frac{(1-T_z)^{k_1+1}P_i^{k_1+1}(1-P_i)^{k_2} - (1-(1-P_i)^k_z)P_i^{k_1}(1-(1-P_i)^k_z)^{k_2+1}(1-T_z)^{k_1+1}}{m_i \phi_{12}} \left( \sum_{j=1}^{m_i} Z_{ij} - \phi_{1} \right) \right),
\]

(27)

where \( M_0 = \sum_{i=1}^{N} M_i \).

**Theorem 3:** The estimator \( \hat{\pi}_{G_{ppswr}} \) of the true population proportion \( \pi \) of a sensitive character is an unbiased estimator.

**Proof:**

\[
E_1 E_2 (\hat{\pi}_{G_{ppswr}}) = E_1 \left\{ \frac{1}{n M_0} \sum_{i=1}^{n} \frac{M_i \hat{\pi}_G}{p_i} \right\} = E_1 \left\{ \frac{1}{n M_0} \sum_{i=1}^{n} \frac{M_i}{p_i} E_2 (\hat{\pi}_G) \right\},
\]

and, since:

\[
E_2 (\hat{\pi}_G) = \frac{1}{m_i \phi_{12}} \left( \sum_{j=1}^{m_i} Z_{ij} - \phi_{1} \right) = \pi_i,
\]

we can obtain:

\[
E_1 E_2 (\hat{\pi}_{G_{ppswr}}) = E_1 \left\{ \frac{1}{n M_0} \sum_{i=1}^{n} \frac{M_i \pi_i}{p_i} \right\} = \frac{1}{n M_0} \sum_{i=1}^{n} \frac{M_i \pi_i}{p_i} = \pi.
\]

\(\square\)
Theorem 4: The variance of $\hat{\tau}_{Gpsswr}$ is obtained by a two-stage sampling scheme, such that a sample of $n$ FSU is selected with replacement by sampling probability $p_i$ for the $i$th cluster from the population of $N$ clusters consisting of $M_i$ elements for the $i$th cluster, and $m_i$ SSUs are drawn by SRSWR from each FSU, as given by:

$$V(\hat{\tau}_{Gpsswr}) = \frac{1}{nM_i^2} \sum_{i=1}^{N} p_i \left[ \frac{M_i \mu_i - M_0 \pi_i}{\pi_i} \right]^2 + \frac{1}{nM_i^2} \sum_{i=1}^{N} \frac{M_i^2}{m_i p_i} \left[ \frac{(1-T_i)\pi_i^{1+p_{i2}+1}(1-p_i)^{p_i^{1}}}{q_i^2} \right]^2 \sigma_{ij}^2,$$

(28)

where:

$$\sigma_{ij}^2 = E(Z_{ij}^2) - (E(Z_{ij}))^2$$

$$= \pi_i \left[ W_i \left( \frac{1-(2k_i+1)(1-p_i)\pi_i^{1+p_{i1}+1}(1-p_i)^{p_i^{1}}}{(1-p_i)^2p_i^{2k_i}} \right) \right.\left. + (1-W_i) \left( \frac{1-(2k_i+1)T_i(1-T_i)\pi_i^{1+p_{i1}+1}(1-T_i)\pi_i^{2}}{T_i^2(1-T_i)\pi_i^{1+p_{i1}+1}} \right) \right] + (1-\pi_i) \left[ \left( \frac{1-(2k_i+1)(1-p_i)\pi_i^{1+p_{i2}}}{(1-p_i)^2p_i^{2k_i}} \right) \right.\left. + (1-W_i) \left( \frac{1-(1-T_i)\pi_i^{1+p_{i2}}}{T_i(1-T_i)\pi_i^{2}} \right) \right] - \left( \pi_i \left[ \left( \frac{1-p_i^{2k_i}}{(1-p_i)^2p_i^{2k_i}} \right) + (1-W_i) \left( \frac{1-(1-T_i)\pi_i^{1+p_{i2}}}{T_i(1-T_i)\pi_i^{2}} \right) \right] \right.\left. + (1-\pi_i) \left[ \left( \frac{1-p_i^{2k_i}}{(1-p_i)^2p_i^{2k_i}} \right) + (1-W_i) \left( \frac{1-(1-T_i)\pi_i^{1+p_{i2}}}{T_i(1-T_i)\pi_i^{2}} \right) \right] \right]^2.$$

Proof: The total number of balls taken by interviewees for the $i$th cluster, $X_{i1}, X_{i2}, Y_{i1}$ and $Y_{i2}$, are random variables with variances:

$$V(X_{i1}) = \frac{1-(2k_{i1}+1)(1-p_i)\pi_i^{1+p_{i1}+1}}{(1-p_i)^2p_i^{2k_{i1}}},$$

(30)

$$V(X_{i2}) = \frac{1-(2k_{i2}+1)(1-p_i)\pi_i^{1+p_{i2}}}{p_i^{2}(1-p_i)^{2k_{i2}}},$$

(31)

$$V(Y_{i1}) = \frac{1-(2k_{i1}+1)(1-T_i)T_i^{2k_{i1}}}{(1-T_i)^2T_i^{2k_{i1}}},$$

(32)

$$V(Y_{i2}) = \frac{1-(2k_{i2}+1)(1-T_i)T_i^{2k_{i2}}}{T_i^{2}(1-T_i)^{2k_{i2}}},$$

(33)

From (21), to drive the variance of $\hat{\tau}_{Gpsswr}$ we can obtain the expected values of $Z_{ij}$ and $Z_{ij}^2$ as follows:

$$E(Z_{ij}) = \pi_i \left[ W_i \left( \frac{1-p_i^{1+p_{i1}}}{(1-p_i)^2p_i^{2k_{i1}}} \right) + (1-W_i) \left( \frac{1-(1-T_i)\pi_i^{1+p_{i1}+1}}{T_i(1-T_i)\pi_i^{1+p_{i1}+1}} \right) \right] + (1-\pi_i) \left[ W_i \left( \frac{1-p_i^{1+p_{i2}}}{p_i(1-p_i)^{2k_{i2}}} \right) + (1-W_i) \left( \frac{1-(1-T_i)\pi_i^{1+p_{i2}}}{T_i(1-T_i)\pi_i^{2}} \right) \right].$$

(34)
\begin{align*}
E(Z_{ij}^2) &= \pi_i \left[ W_i E(X_{ij}^2) + (1 - W_i) E(Y_{ij}^2) \right] + (1 - \pi_i) \left[ W_i E(X_{ij}^2) + (1 - W_i) E(Y_{ij}^2) \right] \\
&= \pi_i \left[ W_i \left( 1 - (2k_{ij} + 1)(1 - p_i)^{\hat{\theta}_i^{k_{ij} + 1} + (1 - p_i)^{2k_{ij} + 1} + (1 - (1 - T_i)^{k_{ij}})^2} \right) \right] \\
&+ (1 - W_i) \left( 1 - (2k_{ij} + 1)T_i(1 - T_i)^{k_{ij} - 1} - (1 - T_i)^{k_{ij} + 1} + (1 - T_i)^{k_{ij}} \right) \\
&+ (1 - \pi_i) \left[ W_i \left( 1 - (2k_{ij} + 1)T_i(1 - p_i)^{k_{ij} - 1} - (1 - p_i)^{2k_{ij} + 1} + (1 - (1 - T_i)^{k_{ij}})^2 \right) \right] \\
&+ (1 - W_i) \left( 1 - (2k_{ij} + 1)(1 - T_i)^{k_{ij} - 1} + (1 - T_i)^{2k_{ij} + 1} + (1 - (1 - T_i)^{k_{ij}})^2 \right).
\end{align*}

(35)

Since \( V(\hat{\tau}_{Gppswr}) = V_1 E_2 (\hat{\tau}_{Gppswr}) + E_1 V_2 (\hat{\tau}_{Gppswr}) \),

\begin{align*}
V_1 E_2 (\hat{\tau}_{Gppswr}) &= V_1 E_2 \left[ \frac{1}{nM_0} \sum_{i=1}^{n} \frac{M_i \hat{\theta}_i}{p_i} \right] \\
&= V_1 \left[ \frac{1}{nM_0} \sum_{i=1}^{n} \frac{M_i \hat{\theta}_i}{p_i} \right] \\
&= \frac{1}{nM_0} \sum_{i=1}^{n} \frac{M_i \hat{\theta}_i}{p_i} - M_0 \pi_i^2,
\end{align*}

and:

\begin{align*}
E_1 V_2 (\hat{\tau}_{Gppswr}) &= E_1 V_2 \left[ \frac{1}{nM_0} \sum_{i=1}^{n} \frac{M_i \hat{\theta}_i}{p_i} \right] \\
&= E_1 \left[ \frac{1}{(nM_0)^2} \sum_{i=1}^{n} \frac{M_i^2}{p_i} V_2 (\hat{\theta}_G) \right] \\
&= E_1 \left[ \frac{1}{(nM_0)^2} \sum_{i=1}^{n} \frac{M_i^2}{p_i} \left( \frac{(1 - T_i)^{k_{ij} - 1} p_i^{k_{ij} + 1} (1 - p_i)^{2k_{ij} + 1} + (1 - p_i)^{k_{ij} + 1}}{\phi_{s2}^2} \right) \right] \\
&= \frac{1}{nM_0} \sum_{i=1}^{n} \frac{M_i^2}{p_i (m_i - 1)} \left[ \frac{(1 - T_i)^{k_{ij} - 1} p_i^{k_{ij} + 1} (1 - p_i)^{2k_{ij} + 1} + (1 - p_i)^{k_{ij} + 1}}{\phi_{s2}^2} \right] \sigma_{Z_i}^2.
\end{align*}

We can then obtain the variance (28). Also, an unbiased estimator of \( V(\hat{\tau}_{Gppswr}) \) is given by:

\begin{align*}
\hat{V}(\hat{\tau}_{Gppswr}) &= \frac{1}{nM_0} \sum_{i=1}^{n} \frac{M_i \hat{\theta}_i}{p_i} - M_0 \hat{\tau}_{Gppswr}^2 \\
&+ \frac{1}{nM_0} \sum_{i=1}^{n} \frac{M_i^2}{p_i (m_i - 1)} \left[ \frac{(1 - T_i)^{k_{ij} - 1} p_i^{k_{ij} + 1} (1 - p_i)^{2k_{ij} + 1} + (1 - p_i)^{k_{ij} + 1}}{\phi_{s2}^2} \right] \sigma_{Z_i}^2.
\end{align*}

(36)

\( \square \)

3.2. PPS Sampling Without Replacement

In this subsection, we consider PPS sampling without replacement to estimate the true population proportion of a sensitive characteristic by applying Yennum et al.’s generalized model, in which \( n \) FSUs are drawn by PPS sampling without replacement from a population of \( N \) clusters with \( M_i \) elementary units for the \( i \)th cluster, and \( m_i \) SSUs are drawn by SRSWR from each FSU.
From this procedure, the estimator $\hat{\pi}_{Gppswor}$ of $\pi$ is given by:

$$\hat{\pi}_{Gppswor} = \frac{1}{M_0} \sum_{i=1}^{n} M_i \hat{\pi}_{iG},$$

(37)

where $\theta_i$ is the first inclusion probability for the $i$th cluster.

The variance of $\hat{\pi}_{Gppswor}$ is given by:

$$V(\hat{\pi}_{Gppswor}) = \frac{1}{M_0^2} \sum_{i=1}^{N} \sum_{j=i}^{N} (\theta_i \theta_j - \theta_{ij}) \left[ \frac{M_i \pi_i - M_j \pi_j}{\theta_i} \right]^2 + \frac{1}{M_0^2} \sum_{i=1}^{N} M_i^2 \left[ \frac{\{1-(1-T_i)^{k_i+1}p_i^{k_i+1}(1-p)^{k_i+1}p_i^{k_i+1}\}^2}{\varphi^2_{ij}} \right].$$

(38)

where $\theta_{ij}$ is the second inclusion probability for $i$th and $j$th clusters.

Also, the variance estimator of $\hat{\pi}_{Gppswor}$ is:

$$V(\hat{\pi}_{Gppswor}) = \frac{1}{M_0^2} \sum_{i=1}^{N} \sum_{j=i}^{N} (\theta_i \theta_j - \theta_{ij}) \left[ \frac{M_i \pi_i - M_j \pi_j}{\theta_i} \right]^2 + \frac{1}{M_0^2} \sum_{i=1}^{N} M_i^2 \left[ \frac{\{1-(1-T_i)^{k_i+1}p_i^{k_i+1}(1-p)^{k_i+1}p_i^{k_i+1}\}^2}{\varphi^2_{ij}} \right].$$

(39)

3.3. Two-Stage Equal Probability Sampling

In this subsection, we consider a two-stage equal probability sampling scheme to estimate the true population proportion of a sensitive attribute by applying Yennum et al.'s generalized model, in which $n$ FSUs are drawn by SRSWOR from a population of $N$ clusters consisting of $M_i$ elementary units for the $i$th cluster, and $m_i$ SSUs are drawn by SRSWR from each FSU.

From this procedure, the estimator $\hat{\pi}_{Gwr}$ of the true population proportion $\pi$ for a sensitive attribute is given by:

$$\hat{\pi}_{Gwr} = \frac{N}{n M_0} \sum_{i=1}^{n} M_i \hat{\pi}_{iG},$$

(40)

where the estimator $\hat{\pi}_{iG}$ is the estimator of a sensitive characteristic of the $i$th cluster, which is the same as (24).

The variance and variance estimator of $\hat{\pi}_{Gwr}$ are:

$$V(\hat{\pi}_{Gwr}) = \frac{N^2}{n M_0^2} \sum_{i=1}^{n} \frac{1}{\pi_i} \left[ M_i \pi_i - \overline{M} \pi \right]^2 + \frac{N}{n M_0^2} \sum_{i=1}^{n} M_i^2 \left[ \frac{\{1-(1-T_i)^{k_i+1}p_i^{k_i+1}(1-p)^{k_i+1}p_i^{k_i+1}\}^2}{\varphi^2_{ij}} \right].$$

(41)

and:

$$\hat{V}(\hat{\pi}_{Gwr}) = \frac{N^2}{n M_0^2} \sum_{i=1}^{n} \frac{1}{\pi_i} (M_i \hat{\pi}_{iG} - \overline{M} \hat{\pi}_{Gwr})^2 + \frac{N}{n M_0^2} \sum_{i=1}^{n} M_i^2 \left[ \frac{\{1-(1-T_i)^{k_i+1}p_i^{k_i+1}(1-p)^{k_i+1}p_i^{k_i+1}\}^2}{\varphi^2_{ij}} \right].$$

(42)

respectively, where $\overline{M} = M_0/N$. 
4. Efficiency Comparisons

4.1. PPSWR Sampling versus Equal Probability Two-Stage Sampling in Yennum et al.’s Model

If we assume \( N - 1 \approx N \), then the difference between the variance of equal probability two-stage sampling, (19), and the variance of PPS with replacement sampling, (6), is given by:

\[
V(\hat{\pi}_{wr}) - V(\hat{\pi}_{ppswr}) = \frac{1}{nM_0 M} \left\{ \sum_{i=1}^{N} (M_i - \bar{M})^2 \pi^2_i + \bar{M} \left[ \sum_{i=1}^{N} (M_i - \bar{M})(\pi^2_i - \pi^2) \right] \right. \\
+ \sum_{i=1}^{N} (M_i - \bar{M})^2 \pi_i \left( \frac{1}{\psi^2_i} \left( \pi_i - \frac{\psi B}{\psi^2_i} \right) \right) \\
+ \bar{M} \left[ \sum_{i=1}^{N} (M_i - \bar{M}) \pi_i \left( \pi_i - \frac{\psi B}{\psi^2_i} \right) \right] \} \tag{43}
\]

In (43), we can see that \( V(\hat{\pi}_{wr}) = V(\hat{\pi}_{ppswr}) \) under the condition \( M_i = \bar{M} = M_0 / N \); i.e., if the cluster sizes are equal, the selection probabilities of the PPS with replacement sampling are all \( N^{-1} \) and equal to those of equal probability two-stage replacement sampling.

If the size of a cluster, \( M_i \), is significantly different, then \( \sum_{i=1}^{N} (M_i - \bar{M})^2 \pi^2_i \), the first term on the right side of (43), has large values, and the second term, \( \sum_{i=1}^{N} (M_i - \bar{M})^2 \pi^2_i \), has relatively small values. Hence, the estimation by PPS with replacement sampling is more efficient than that by equal probability two-stage replacement sampling.

We used the relative efficiency (RE) to compare the efficiency of the two sampling methods—PPS with replacement sampling and equal probability two-stage replacement sampling:

\[
RE_1 = \frac{V(\hat{\pi}_{wr})}{V(\hat{\pi}_{ppswr})} \times 100(\%).
\]

Values of \( RE_1 \) over 100% indicate that the estimator obtained by the PPS with the replacement sampling method was more efficient than the estimator obtained by the equal probability two-stage replacement sampling.

In calculating REs, we set the parameters as follows:

\[
\begin{align*}
M_0 &= 10,000, M_1 = 1,000, M_2 = 2,000, M_3 = 3,000, M_4 = 4,000, \\
m_0 &= 1,000, m_1 = 100, m_2 = 200, m_3 = 300, m_4 = 400, \\
p_1 &= 0.235, p_2 = 0.441, p_3 = 0.609, p_4 = 0.715.
\end{align*}
\]

From Table 1, when the selection probability \( W \) for the first-stage randomization device increased from 0.1 to 0.9 by 0.2 and the second stage randomization devices \( T \) increased from 0.6 to 0.8 by 0.1 and \( P \) from 0.65 to 0.90 by 0.05, REs increase under the fixed proportion of a sensitive attribute (particularly when the selection probability of the second randomization device \( T \) increased), and the RE increased according to the conditions of \( P \) and \( \pi_i \).
Table 1. The relative efficiencies (REs) of a sensitive estimator between the probability proportional to size (PPS) sampling with replacement and the equal probability two-stage sampling in Yennum et al.’s model to change $\pi_i$ and $W$.

| $\pi_i$ | $T$ | $W$ | $p$ | 0.6 | 0.7 | 0.8 | 0.9 | 0.6 | 0.7 | 0.8 | 0.9 |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0.6     | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 |
| 0.7     | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 |
| 0.8     | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| 0.9     | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |

On the other hand, RE increased when the first-stage selection probability $W$ was less than 0.5, and the values of $T$, $p$, and $\pi_i$ (from 0.1 to 0.4) decreased, but the RE decreased when the value of $W$ was greater than 0.5 under a fixed value for $T$, $p$, and $\pi_i$.

Furthermore, the greater the true population proportion of a sensitive attribute $\pi_i$, the higher the overall efficiency of Yennum et al.’s model, as shown by the values of the bottom cells in Table 1. This result agrees with the typical sampling survey methodology as the true population proportion of a sensitive attribute increases.

4.2. PPSWR Sampling versus Equal Probability Two-Stage Sampling in Yennum et al.’s Generalized Model

If we assume $N - 1 = N_i$, then the difference between the variance of equal probability two-stage sampling scheme (41) and the variance of the PPS with replacement sampling scheme (28) is given by:

$$V(\hat{\pi}_{Garr}) - V(\hat{\pi}_{Gpsw}) = \frac{1}{nM_0} \left[ \sum_{i=1}^{N} (M_i - M)^2 \pi_i^2 + M \sum_{i=1}^{N} (M_i - M)(\pi_i^2 - \pi^2) \right]$$

$$+ \sum_{i=1}^{N} (M_i - M)^2 \frac{1}{m_i} \left[ \left(1 - T_i \right)^{\pi_i + 1} p_i^{1+a} \left(1 - p_i \right)^{a + 1} p_i^{1+a} \right]^2 \sigma^2_Z$$

$$+ M \sum_{i=1}^{N} (M_i - M) \frac{1}{m_i} \left[ \left(1 - T_i \right)^{\pi_i + 1} p_i^{1+a} \left(1 - p_i \right)^{a + 1} p_i^{1+a} \right]^2 \sigma^2_Z \right].$$

(44)
In (44), we can see that \( V(\hat{\pi}_{Gwr}) = V(\hat{\pi}_{Gppswr}) \) under the condition \( M_i = \bar{M} = M_0/N \), i.e., if the cluster sizes are equal, the selection probabilities of the PPS with replacement sampling are all \( N^{-1} \) and equal to those of the equal probability two-stage replacement sampling.

If cluster sizes, \( M_i \), were significantly different, then \( \sum_{i=1}^{N} (M_i - \bar{M})^2 \pi_i^2 \), the first term of the right-hand side in (44), had large values, and the second term, \( \sum_{i=1}^{N} (M_i - \bar{M})^2 (\pi_i^2 - \bar{\pi}^2) \), had relatively small values. Hence, the estimation by PPS with replacement sampling is more efficient than that by equal probability two-stage replacement sampling.

We used the relative efficiency (RE) to compare the efficiency of the two sampling designs (PPS with replacement sampling and equal probability two-stage replacement sampling):

\[
RE_2 = \frac{V(\hat{\pi}_{Gwr})}{V(\hat{\pi}_{Gppswr})} \times 100\%
\]

Values of \( RE_2 \) over 100% indicate that the estimator obtained by PPS with the replacement sampling method was more efficient than the estimator obtained by equal probability two-stage replacement sampling.

Table 2 shows the results of the REs obtained by increasing the true population proportion \( \pi_i \) from 0.1 to 0.4 by 0.1. The selection probabilities of the randomized response model (\( W, T \) and \( P \)) are shown in Section 4.1.

**Table 2.** The REs for a sensitive estimator between the PPS with replacement sampling and equal probability two-stage sampling with replacement in Yennum et al.'s generalized model for changing \( \pi_i \) and \( W \).

| \( \pi_i \) | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|
| 0.65 | 171.5 | 172.8 | 172.8 | 163 |
| 0.7 | 167.5 | 169.7 | 170.1 | 152.7 |
| 0.75 | 162.2 | 165.9 | 167 | 135.6 |
| 0.8 | 154.2 | 160.9 | 163.3 | 106.2 |
| 0.85 | 139.8 | 136.9 | 138.5 | 54.9 |
| 0.9 | 108.4 | 136 | 150.3 | 49.4 |
| 0.6 | 180.9 | 181.1 | 181 | 177 |
| 0.7 | 178.9 | 179.5 | 179.5 | 172.1 |
| 0.75 | 176.4 | 177.5 | 177.8 | 163.1 |
| 0.8 | 172.6 | 174.9 | 175.8 | 143.4 |
| 0.85 | 165.8 | 171 | 173.1 | 91.8 |
| 0.9 | 147.3 | 162.3 | 168.5 | 55.0 |
| 0.6 | 184 | 184 | 183.9 | 181.6 |
| 0.7 | 182.8 | 183 | 182.9 | 178.5 |
| 0.75 | 181.2 | 181.7 | 182.9 | 176.2 |
| 0.8 | 178.9 | 180 | 180.4 | 159.7 |
| 0.85 | 174.8 | 177.5 | 178.6 | 114.3 |
| 0.65 | 185.5 | 185.5 | 185.4 | 183.9 |
| 0.7 | 184.7 | 184.8 | 184.7 | 181.7 |
| 0.75 | 183.5 | 183.8 | 183.9 | 177.7 |
| 0.8 | 181.9 | 182.6 | 182.9 | 168.2 |
| 0.85 | 179.2 | 180.8 | 181.6 | 130.6 |
| 0.9 | 171.7 | 177.1 | 179.3 | 71.7 |

**Table 2.** The REs for a sensitive estimator between the PPS with replacement sampling and equal probability two-stage sampling with replacement in Yennum et al.'s generalized model for changing \( \pi_i \) and \( W \).
In calculating the REs, we set the parameters as follows:

\[ M_0 = 10,000, M_1 = 1,000, M_2 = 2,000, M_3 = 3,000, M_4 = 4,000 \]

\[ m_0 = 1,000, m_1 = 100, m_2 = 200, m_3 = 300, m_4 = 400, \]

\[ p_1 = 0.235, p_2 = 0.441, p_3 = 0.609, p_4 = 0.715, \]

\[ k_1 = 2, k_2 = 1. \]

From the results of Table 2, the efficiencies vary according to changes in the probabilities of selection during the first stage \( W \) and the second stage \( T \) and \( P \) in the randomization device, but when the first-stage selection probability \( W \) is fixed, and the second-stage selection probabilities \( T \) and \( P \) increase, then the relative efficiency of the PPS sampling is better than that of the equal probability two-stage sampling in Yennum et al.’s model.

5. Conclusions

We extended Yennum et al.’s model, in which geometric distribution is used as a randomization device for a population consisting of different-sized clusters, and clusters are selected by PPS sampling. Estimators for the true population proportion of a sensitive attribute, their variances, and their variance estimators are derived under PPS sampling and equal probability two-stage sampling.

We also applied these sampling designs to the case of Yennum et al.’s generalized model. Numerical studies were carried out to compare the efficiencies of the proposed methods in each case of Yennum et al.’s model and Yennum et al.’s generalized model in cases with a replacement.

Although the experiments were assumed to use a replacement, we expected similar results for a case without replacement, as per typical sampling theory.

From the numerical study, we found that the efficiency of the two-stage sampling for probability proportional to size depends on the given parameter values, but the efficiency of Yennum et al.’s generalized model is preferred for most combinations of parameters over around 80%.

Author Contributions: Conceptualization, G.-S.L.; methodology, C.-K.S.; writing—original draft preparation, K.-H.H.; writing—review and editing, C.-K.S.; project administration and funding acquisition, G.-S.L.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1D1A3B07044007).

Acknowledgments: We would like to thank the anonymous reviewers for their very careful reading and valuable comments/suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Warner, S.L. Randomized response: A survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* **1965**, *60*, 63–69. [CrossRef] [PubMed]
2. Cochran, W.G. *Sampling Techniques*, 3rd ed.; John Wiley and Sons: New York, NY, USA, 1977.
3. Fox, J.A.; Tracy, P.E. *Randomized Response: A Method for Sensitive Survey*; Sage Publications: Newbury Park, CA, USA, 1986.
4. Kuk, A.Y.C. Asking sensitive questions indirectly. *Biometrika* **1990**, *77*, 436–438. [CrossRef]
5. Chaudhuri, A.; Mukerjee, R. *Randomized Response: Theory and Techniques*; Marcel Dekker Inc.: New York, NY, USA, 1988.
6. Ryu, J.B.; Hong, K.H.; Lee, G.S. *Randomized Response Model*; Freedom Academy: Seoul, Korean, 1993.
7. Lee, G.S.; Hong, K.H. Randomized response model by two-stage cluster sampling. *Korean Commun. Stat.* **1998**, *5*, 99–105.
8. Lee, G.S. A Study on the Randomized Response Technique by PPS Sampling. *Korean J. Appl. Stat.* **2006**, *19*, 69–80.
9. Yennum, N.Y.; Sedory, S.A.; Singh, S. Improved strategy to collect sensitive data by using geometric distribution as a randomization device. *Commun. Stat. Theory Methods* 2019, 48, 5777–5795. [CrossRef]

10. Hussain, Z.; Shabbir, J.; Pervez, Z.; Shah, S.F.; Khan, M. Generalized geometric distribution of order k: A flexible choice to randomize the response. *Commun. Stat. Simul. Comput.* 2017, 46, 4708–4721. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).