An Empirical Study of Stratified Warner's Randomized Response Model

By Souvick Bera, Sagnik Bhattacharya, Soumyadipta Das
Department of Statistics, University of Calcutta

Abstract-

This poster shows a stratified randomized response model based on Warner's model that has an optimal allocation and large gain in precision. It is shown that the proposed model is more efficient than the stratified randomized response model. Additionally, it is shown that the estimator based on the proposed method is more efficient than the Warner ,the Mangat and Singh and the Mangat estimators under the conditions presented in both the case of completely truthful reporting and that of not completely truthful reporting by the respondents.

Introduction-

Warner did the pioneering work of a randomized response technique which minimizes underreporting of a data relative to a socially undesirable or incriminating behavior questions. The Warner model required the interviewee to give a "Yes" or "No" answer either to the sensitive question or to its negative depending on the outcome of a randomizing device not reported to the interviewer. Mangat and Singh proposed a two-stage randomized response model that is a variation of the Warner model. Applicability of this model has been illustrated by Singh and Mangat. Mangat proposed another randomized response model which has the benefit of simplicity over that of Mangat and Singh. Hong et al. suggested a stratified randomized response technique that applied the same randomization device to every stratum. Stratified random sampling is generally obtained by dividing the population into non overlapping groups called strata and selecting a simple random sample from each stratum. An randomized response technique using a stratified random sampling gives the group characteristics related to each stratum estimator. Also,stratified samples protect a researcher from the possibility of obtaining a poor sample.Under Hong et al.'s proportional sampling assumption, it may be easy to derive the variance of the proposed estimator; however, it may cause a high cost because of the difficulty in obtaining a proportional sample from some stratum. To rectify this problem, we present a stratified randomized response technique using an optimal allocation which is more efficient than a stratified randomized response technique using a proportional allocation. Mahajan et al. have considered the problem of construction of optimum strata boundaries for scrambled responses.

Methodology-

In the proposed model, the population is partitioned into strata, and a sample is selected by simple random sampling with replacement in each stratum. To get the full benefit from stratification, we assume that the number of units in each stratum is known. An individual respondent in the sample of stratum 'i' is instructed to use the randomization device $\boldsymbol{R_i}$ which consists of a sensitive question $(\mathbf{S})$ card with probability $\boldsymbol{P_i}$ and its negative question $(\bar{\boldsymbol{S}})$ card with probability $\mathbf{1}-\boldsymbol{P_i}$. The respondent should answer the question by "Yes" or "No" without reporting which question card she or he has. A respondent belonging to the sample in different strata will perform different randomization devices, each having different preassigned probabilities. Let $\boldsymbol{n_i}$ denote the number of units in the sample from stratum **'i'** and **'n'** denote the total number of units in samples from all strata so that $\boldsymbol{n}=\sum_{i=1}^{k}\boldsymbol{n_i}$. Under the assumption that these "Yes" and "No" reports are made truthfully and $\boldsymbol{P_i}(\neq \mathbf{0.5})$ is set by the researcher, the probability of a "Yes" answer in a stratum 'i' for this procedure is,

$\boldsymbol{Z_i=P_i\pi_{S_i}+(1-P_i)(1-\pi_{S_i})\ \forall\ i=1,\ldots\ldots,k}$ where $\boldsymbol{Z_i}$ is the proportion of **"Yes"** answer in a stratum 'i' , $\boldsymbol{\pi_{S_i}}$ is the proportion of respondents with the sensitive trait in a stratum 'i' and $\boldsymbol{P_i}$ is the probability that a respondent in the sample stratum 'i' has a sensitive question $(\mathbf{S})$ card.

The <u>maximum likelihood estimate</u> of $\boldsymbol{\pi_{S_i}}$ is shown to be $\boldsymbol{\widehat{\pi_{S_i}}=\frac{\widehat{Z_i}-(1-P_i)}{2P_i-1}\ \forall\ i=1,\ldots\ldots,k}$, where $\boldsymbol{\widehat{Z_i}}$ is the proportion of "Yes" answer in a sample in the stratum **' i '**. Since each $\boldsymbol{\widehat{Z_i}}$ is a binomial distribution $\boldsymbol{B(n_i,Z_i)}$ and the selections in different strata are made independently, the maximum likelihood estimate of π_S is easily shown to be $\boldsymbol{\widehat{\pi_S}=\sum_{i=1}^{k}w_i\widehat{\pi_{S_i}}=\sum_{i=1}^{k}w_i\left[\frac{\widehat{Z_i}-(1-P_i)}{2P_i-1}\right]}$ where we denote **'N'** to be the number of units in the whole population, $\boldsymbol{N_i}$ to be the total number of units in the stratum 'i' and $\boldsymbol{w_i={}^{N_i}/_{N}}\ \boldsymbol{\forall i=1,\ldots\ldots,k}$ so that $\boldsymbol{w=\sum_{i=1}^{k}w_i=1}$.

As each estimator $\boldsymbol{\widehat{\pi_{S_i}}}$ is ***<u>unbiased</u>*** for $\boldsymbol{\pi_{S_i}}$, the expected value of $\boldsymbol{\widehat{\pi_S}}$ is

$$\boldsymbol{E(\widehat{\pi_S})=E\left(\sum_{i=1}^{k}w_i\widehat{\pi_{S_i}}\right)=\sum_{i=1}^{k}w_iE\left(\widehat{\pi_{S_i}}\right)=\sum_{i=1}^{k}w_i\pi_{S_i}=\pi_S}$$

Since each unbiased estimator $\widehat{\pi_{S_i}}$ has its own variance, the variance of $\widehat{\pi_S}$ is,

$$\boldsymbol{var(\widehat{\pi_S})=var\left(\sum_{i=1}^{k}w_i\widehat{\pi_{S_i}}\right)}$$

$$\boldsymbol{=\sum_{i=1}^{k}w_i^2\ var(\widehat{\pi_{S_i}})=\sum_{i=1}^{k}\frac{w_i^2}{n_i}\left[\pi_{S_i}(1-\pi_{S_i})+\frac{P_i(1-P_i)}{(2P_i-1)^2}\right]}$$

Information on $\boldsymbol{\pi_{S_i}}$ is usually unavailable. But if prior information on $\boldsymbol{\pi_{S_i}}$ is available from past experience then it helps to derive the following optimal allocation formula.

Results-

After obtaining the required estimator for stratified randomized response we arrive at the following results :

- Suppose that there are two strata in the population, $P=P_1=P_2\neq 0.5$,and $\pi_{S_1}\neq\pi_{S_2}$. The proposed estimator $\boldsymbol{\widehat{\pi_S}}$ will be more efficient than the Mangat and Singh (1990) estimator $\boldsymbol{\widehat{\pi_{ms}}}$, under the following condition, $(\pi_{S_1}-\pi_{S_2})^2+\left[\left\{\pi_{S_1}(1-\pi_{S_1})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}-\left\{\pi_{S_2}(1-\pi_{S_2})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2>\left[\left\{\frac{M(1-P)}{(2P-1)(2P-1+2M(1-P))}\right\}^2-\frac{M(1-P)}{(1-2P)(2P-1+2M(1-P))^2}\right][w_1(1-w_1)]^{-1}$
- Suppose that $\boldsymbol{\widehat{\pi_S}>1-\left[\frac{P}{2P-1}\right]^2}$ and assume that there are two strata in the population, $P=P_1=P_2\neq 0.5$, and $\pi_{S_1}\neq\pi_{S_2}$. The proposed estimator $\boldsymbol{\widehat{\pi_S}}$ will be more efficient than the Mangat (1994) estimator $\boldsymbol{\widehat{\pi_m}}$ under the following condition, $(\pi_{S_1}-\pi_{S_2})^2+\left[\left\{\pi_{S_1}(1-\pi_{S_1})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}-\left\{\pi_{S_2}(1-\pi_{S_2})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2>\frac{1-P}{w_1(1-w_1)P}\left[\left(\frac{P}{2P-1}\right)^2-\{1-(w_1\pi_{S_1}+w_2\pi_{S_2})\}\right]$
- The proposed estimator $\boldsymbol{\widehat{\pi_S}}$ is more efficient than the Warner (1965) estimator $\boldsymbol{\widehat{\pi_w}}$ in the case of two strata in the population and $P=P_1=P_2\neq 0.5$.

Conclusion-

This poster presents a stratified randomized response model which is more efficient than the Hong et al. stratified randomized response model. The proposed method is more useful than the previous methods in that a stratified randomized response method helps to solve the limitation of randomized response that is the loss of individual characteristics of the respondents. Furthermore, different choices of Pi in different strata can be more practicable with different situations.

Recommendations-

The model can be used as follows-

•A tax evasion is more sensitive for rich people and less sensitive for poor people.

•A grade of mathematics class is more sensitive for medical and science students and less sensitive for arts students.

•A drug addiction is a good example of age by sex stratum.

Therefore, the proposed method has several advantages compared to the previous randomized response methods. Researchers can apply the method to a medical related research topic or criminal related research topics with these advantages.

Remark : In Mangat and Singh model, different $M_i, i=1,\ldots,k$ can also be used. If $M_i=M,\forall i$, it will reduce to the present model.

Acknowledgements-

We are grateful to our professors for guiding us through this project. We further acknowledge the following resources :

- Jong-Min Kima, William D. Warde. A stratified Warner's randomized response model
- Cochran, W.G., 1977. Sampling Techniques, 3rd Edition. Wiley, New York.
- Hong, K., Yum, J., Lee, H., 1994. A stratified randomized response technique