Simultaneous Estimation of Means of Two Sensitive Variables

Segun Ahmed*, Stephen A. Sedory and Sarjinder Singh
Department of Mathematics
Texas A&M University-Kingsville
Kingsville, TX 78363
E-mail*: seg3k@yahoo.com

Abstract

In this paper, we introduce a new problem of simultaneous estimation of means of two quantitative sensitive variables by using only one randomized response another pseudo response from a respondent in a sample. The proposed estimators are extended to stratified random sampling, and the relative efficiency values are computed for equal, proportional and optimum allocation with respect to the newly introduced naïve estimators.

Keywords: Two sensitive variables, estimation of means, variance, protection of respondents.

1. Introduction

The problem of estimation of population mean of a sensitive variable, such as income, underreported tax, and number of induced abortions etc. is well known in the field of randomized response sampling. Horvitz *et al.* (1967) and Greenberg *et al.* (1971) extended the Warner (1965) model to the case where the responses to the sensitive question are quantitative rather than a simple 'yes' or 'no' while estimating the proportion of a sensitive attribute. The unrelated question model can also be used to estimate correlation between two sensitive characteristics. Fox and Tracy (1984) used unrelated question model to estimate correlation between two quantitative sensitive attributes. Himmelfarb and Edgell (1980) introduced an additive model to estimate mean of a sensitive quantitative variable sensitive variable, say Y (*see* Fox, 2016). In their model each respondent scrambles in response Y by adding it to a random scrambling variable S and only then reveals the scrambled result $Z = Y + S$ to the interviewer. The mean of the response, $E(Y)$ can be estimated from a sample of Z values and the knowledge of the distribution of the scrambling variable S.
Eichhorn and Hayre (1983) introduced a new multiplicative model to estimate the population mean of a sensitive variable, say Y. In their model each respondent scrambles in response Y by multiplying it by a random scrambling variable S and only then reveals the scrambled result $Z = Y S$ to the interviewer. The mean of the response, $E(Y)$ can be estimated from a sample of Z values and the knowledge of the distribution of the scrambling variable S. Both the additive and multiplicative models have concern about the choice of use of scrambling variables. The distribution of S is assumed to be known, that is, a good guess about the maximum and minimum values of the scrambling variable S are assumed to be known. Thus it makes afraid a respondent that his/her true value of the sensitive variable can be revealed.

In this paper, we consider a different approach which can be used to estimate the means of two (or more) sensitive variables simultaneously with the help of scrambled responses. Also a respondent will be more cooperative in responding because the proposed method makes use of one scrambled response and other fake response that is free from the true sensitive variables.

2. Proposed New Randomized Response Model

Consider a population Ω consisting of finite number N of persons. Consider we selected a sample s of n persons from the population Ω by using simple random and with replacement sampling (SRSWR). Let Y_{1i} and Y_{2i} be the values of two quantitative sensitive variables associated with the ith unit in the population Ω. Let μ_{y_1} and μ_{y_2} be the true population means of the two sensitive variables Y_{1i} and Y_{2i}, respectively, which we wish to estimate. Each respondent selected in the sample is asked to generate two fake values of scrambling variables S_1 and S_2 from two known distributions. Assume S_1 and S_2 are independent, which help to maintain the protection of respondents. Let $E(S_1)=\theta_1$, $V(S_1)=\gamma_{20}$, $E(S_2)=\theta_2$ and $V(S_2)=\gamma_{02}$ are known. In the proposed randomized response model, each respondent selected in the sample is requested to report the scrambled response as:

$$Z_{1i} = S_1 Y_{1i} + S_2 Y_{2i} \tag{2.1}$$

Note here that the mixing of two sensitive variables with two scrambling variables will certainly make it difficult for an interviewer to guess the individual values of two sensitive variables. There is no restriction on the scrambling variables to take negative values, which will certainly increase respondents' cooperation while doing a face to face survey. Note that the main purpose of randomized response survey is to protect a respondent during a face to face survey, thus use of sample random sampling is highly recommended. Note that any other complex design making use of highly correlated auxiliary variable at the selection stage may threaten the privacy of a respondent.

Also each respondent is also requested to rotate a spinner which consists of two outcomes similar to Warner (1965) spinner, but has different types of outcomes. If the pointer lands in a shaded area then the respondent is asked to report the value of the scrambling variable S_1 and if the pointer lands in the non-shaded area then the respondent is asked to report the value of the scrambling variable S_2. Let P be the proportion of a shaded area and $(1-P)$ be the proportion of non-shaded area of the spinner. Thus the second response from the ith respondent is given by:

$$Z_i = \begin{cases} S_1 & \text{with probability } P \\ S_2 & \text{with probability } (1-P) \end{cases} \tag{2.2}$$

where $P \neq \dfrac{\theta_1 \gamma_{02}}{\theta_1 \gamma_{02} + \theta_2 \gamma_{20}}$.

Taking expected value on both sides of (2.1) we have

$$E(Z_{1i}) = E[S_1 Y_{1i} + S_2 Y_{2i}] = \theta_1 \mu_{y_1} + \theta_2 \mu_{y_2} \tag{2.3}$$

From (2.1) and (2.2), we generate response Z_{2i} as follows:

$$Z_{2i} = Z_i Z_{1i} = \begin{cases} S_1^2 Y_{1i} + S_1 S_2 Y_{2i} & \text{with probability } P \\ S_1 S_2 Y_{1i} + S_2^2 Y_{2i} & \text{with probabilty } (1\text{-}P) \end{cases} \tag{2.4}$$

Taking expected value on both sides of (2.4) we have

$$E(Z_{2i}) = P\left[\mu_{y_1}(\theta_1^2 + \gamma_{20}) + \theta_1\theta_2\mu_{y_2}\right] + (1-P)\left[\theta_1\theta_2\mu_{y_1} + (\theta_2^2 + \gamma_{02})\mu_{y_2}\right] \tag{2.5}$$

From (2.3) and (2.5), by the method of moments, we have:

$$\theta_1\hat{\mu}_{y_1} + \theta_2\hat{\mu}_{y_2} = \frac{1}{n}\sum_{i=1}^{n} Z_{1i} \tag{2.6}$$

and

$$\left[P(\theta_1^2 + \gamma_{20}) + (1-P)\theta_1\theta_2\right]\hat{\mu}_{y_1} + \left[P\theta_1\theta_2 + (1-P)(\theta_2^2 + \gamma_{02})\right]\hat{\mu}_{y_2} = \frac{1}{n}\sum_{i=1}^{n} Z_{2i} \tag{2.7}$$

The equations (2.6) and (2.7) can be written as:

$$\begin{bmatrix} \theta_1, & \theta_2 \\ P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2, & P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2) \end{bmatrix} \begin{bmatrix} \hat{\mu}_{y_1} \\ \hat{\mu}_{y_2} \end{bmatrix} = \begin{bmatrix} \overline{Z}_1 \\ \overline{Z}_2 \end{bmatrix} \tag{2.8}$$

where

$$\overline{Z}_1 = \frac{1}{n}\sum_{i=1}^{n} Z_{1i} \text{ and } \overline{Z}_2 = \frac{1}{n}\sum_{i=1}^{n} Z_{2i}\ .$$

Applying Cramer's rule on (2.8), we have

$$\begin{aligned} \Delta &= \begin{vmatrix} \theta_1, & \theta_2 \\ P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2, & P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2) \end{vmatrix} \\ &= P\theta_1^2\theta_2 + (1-P)\theta_1\gamma_{02} + (1-P)\theta_1\theta_2^2 - \theta_2 P\gamma_{20} - P\theta_1^2\theta_2 - (1-P)\theta_1\theta_2^2 \\ &= (1-P)\theta_1\gamma_{02} - \theta_2 P\gamma_{20}\ . \end{aligned} \tag{2.9}$$

$$\Delta_1 = \begin{vmatrix} \overline{Z}_1, & \theta_2 \\ \overline{Z}_2, & P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2) \end{vmatrix} = \left[P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\right]\overline{Z}_1 - \overline{Z}_2\theta_2 \tag{2.10}$$

and

$$\Delta_2 = \begin{vmatrix} \theta_1, & \overline{Z}_1 \\ P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2, & \overline{Z}_2 \end{vmatrix} = \theta_1\overline{Z}_2 - \left[P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\right]\overline{Z}_1 \tag{2.11}$$

From (2.9), (2.10) and (2.11), we have estimators of μ_{y_1} and μ_{y_2} are, respectively, given by

$$\hat{\mu}_{y_1} = \frac{\Delta_1}{\Delta} = \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\bar{Z}_1 - \theta_2\bar{Z}_2}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}} \tag{2.12}$$

and

$$\hat{\mu}_{y_2} = \frac{\Delta_2}{\Delta} = \frac{\theta_1\bar{Z}_2 - \{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}\bar{Z}_1}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}} \tag{2.13}$$

We have the following theorems.

Theorem 2.1. The proposed estimator $\hat{\mu}_{y1}$ is an unbiased estimator of the population mean μ_{y1}.

Proof. Taking expected value on both sides of (2.12), we have

$$E(\hat{\mu}_{y_1}) = E\left[\frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\bar{Z}_1 - \theta_2\bar{Z}_2}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}}\right]$$

$$= \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\sum_{i=1}^{n} E(Z_{1i}) - \theta_2\sum_{i=1}^{n} E(Z_{2i})}{n\{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\}}$$

$$= \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\sum_{i=1}^{n}(\theta_1\mu_{y_1} + \theta_2\mu_{y_2}) - \theta_2\sum_{i=1}^{n}\left[P\{\mu_{y_1}(\theta_1^2 + \gamma_{20}) + \theta_1\theta_2\mu_{y_2}\} + (1-P)\{\theta_1\theta_2\mu_{y_1} + (\theta_2^2 + \gamma_{02})\mu_{y_1}\right]}{n\{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\}}$$

$$= \frac{\{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\}\mu_{y_1}}{\{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\}} = \mu_{y_1}$$

which proves the theorem.

Theorem 2.2. The proposed estimator $\hat{\mu}_{y_2}$ is an unbiased estimator of the population mean μ_{y_2}.

Proof. Taking expected value on both sides of (2.13), we have

$$E(\hat{\mu}_{y_2}) = E\left[\frac{\theta_1\bar{Z}_2 - \{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}\bar{Z}_1}{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}}\right] = \mu_{y_2}$$

Hence the theorem.

Theorem 2.3. The variance of the proposed estimator $\hat{\mu}_{y_1}$ is given by

$$V(\hat{\mu}_{y_1}) = \frac{\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}^2\sigma_{Z_1}^2 + \theta_2^2\sigma_{Z_2}^2 - 2\theta_2\{P\theta_1\theta_2 + (1-P)(\gamma_{02} + \theta_2^2)\}\sigma_{Z_1Z_2}}{n\{(1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20}\}^2} \tag{2.14}$$

where

$$\sigma_{Z_1}^2 = \gamma_{20}(\sigma_{y_1}^2 + \mu_{y_1}^2) + \gamma_{02}(\sigma_{y_2}^2 + \mu_{y_2}^2) + \theta_1^2\sigma_{y_1}^2 + \theta_2^2\sigma_{y_2}^2 + 2\theta_1\theta_2\sigma_{y_1y_2} \tag{2.15}$$

$$\sigma_{Z_2}^2 = (\sigma_{y_1}^2 + \mu_{y_1}^2)\left[P(\gamma_{40} + 4\gamma_{30}\theta_1 + 6\gamma_{20}\theta_1^2 + \theta_1^4) + (1-P)(\gamma_{20} + \theta_1^2)(\gamma_{02} + \theta_2^2)\right]$$

$$+(\sigma_{y_2}^2+\mu_{y_2}^2)\left[(1-P)(\gamma_{04}+4\gamma_{03}\theta_2+6\gamma_{02}\theta_2^2+\theta_2^4)+P(\gamma_{20}+\theta_1^2)(\gamma_{02}+\theta_2^2)\right]$$
$$+2(\sigma_{y_1y_2}+\mu_{y_1}\mu_{y_2})\left[P\theta_2(\gamma_{30}-3\theta_1\gamma_{20}+\theta_1^3)+(1-P)\theta_1(\gamma_{03}-3\theta_2\gamma_{02}+\theta_2^3)\right]$$
$$-\left[\mu_{y_1}\{P(\gamma_{20}+\theta_1^2)+(1-P)\theta_1\theta_2\}+\mu_{y_2}\{P\theta_1\theta_2+(1-P)(\gamma_{02}+\theta_2^2)\}\right]^2 \quad (2.16)$$

and

$$\sigma_{Z_1Z_2}=(\sigma_{y_1}^2+\mu_{y_1}^2)\{P(\gamma_{30}-3\theta_1\gamma_{20}+\theta_1^3)+(1-P)\theta_2(\gamma_{20}+\theta_1^2)\}$$
$$+(\sigma_{y_2}^2+\mu_{y_2}^2)\{P\theta_1(\gamma_{02}+\theta_2^2)+(1-P)(\gamma_{03}-3\theta_2\gamma_{02}+\theta_2^3)\}$$
$$+2(\sigma_{y_1y_2}+\mu_{y_1}\mu_{y_2})\{P\theta_2(\gamma_{20}+\theta_1^2)+(1-P)\theta_1(\gamma_{02}+\theta_2^2)\}$$
$$-(\theta_1\mu_{y_1}+\theta_2\mu_{y_2})[\mu_{y_1}\{P(\gamma_{20}+\theta_1^2)+(1-P)\theta_1\theta_2\}+\mu_{y_2}\{P\theta_1\theta_2+(1-P)(\gamma_{02}+\theta_2^2)\}] \quad (2.17)$$

Proof. Given $E(S_1)=\theta_1$ and $E(S_2)=\theta_2$. Following Singh (2016), let us define

$$\gamma_{ab}=E[S_1-\theta_1]^a[S_2-\theta_2]^b \quad (2.18)$$

Then due to independence of the scrambling variables, we have

$$E(S_1^2)=\gamma_{20}+\theta_1^2 \quad (2.19)$$
$$E(S_1^3)=\gamma_{30}-3\theta_1\gamma_{20}+\theta_1^3 \quad (2.20)$$
$$E(S_1^4)=\gamma_{40}+4\gamma_{30}\theta_1+6\gamma_{20}\theta_1^2+\theta_1^4 \quad (2.21)$$
$$E(S_2^2)=\gamma_{02}+\theta_2^2 \quad (2.22)$$
$$E(S_2^3)=\gamma_{03}-3\theta_2\gamma_{02}+\theta_2^3 \quad (2.23)$$
$$E(S_2^4)=\gamma_{04}+4\gamma_{03}\theta_2+6\gamma_{02}\theta_2^2+\theta_2^4 \quad (2.24)$$
$$E(S_1S_2)=\theta_1\theta_2 \quad (2.25)$$
$$E(S_1^2S_2^2)=(\gamma_{20}+\theta_1^2)(\gamma_{02}+\theta_2^2) \quad (2.26)$$
$$E(S_1^3S_2)=(\gamma_{30}-3\theta_1\gamma_{20}+\theta_1^3)\theta_2 \quad (2.27)$$

and

$$E(S_1S_2^3)=\theta_1(\gamma_{03}-3\theta_2\gamma_{02}+\theta_2^3) \quad (2.28)$$

The variance $\sigma_{Z_1}^2$ is given by

$$\sigma_{Z_1}^2=E(Z_{1i}^2)-\{E(Z_{1i})\}^2$$
$$=E(S_1Y_{1i}+S_2Y_{2i})^2-\{E(S_1Y_{1i}+S_2Y_{2i})\}^2$$
$$=E[S_1^2Y_{1i}^2+S_2^2Y_{2i}^2+2S_1S_2Y_{1i}Y_{2i}]-\{E(S_1Y_{1i}+S_2Y_{2i})\}^2$$
$$=(\gamma_{20}+\theta_1^2)(\sigma_{y_1}^2+\mu_{y_1}^2)+(\gamma_{02}+\theta_2^2)(\sigma_{y_2}^2+\mu_{y_2}^2)+2\theta_1\theta_2(\sigma_{y_1y_2}+\mu_{y_1y_2})-\{\theta_1\mu_{y_1}+\theta_2\mu_{y_2}\}^2$$
$$=\gamma_{20}(\sigma_{y_1}^2+\mu_{y_1}^2)+\gamma_{02}(\sigma_{y_2}^2+\mu_{y_2}^2)+\theta_1^2\sigma_{y_1}^2+\theta_2^2\sigma_{y_2}^2+2\theta_1\theta_2\sigma_{y_1y_2} \quad (2.29)$$

The variance $\sigma_{Z_2}^2$ is given by

$$\sigma_{Z_2}^2=V(Z_{2i})=E(Z_{2i}^2)-\{E(Z_{2i})\}^2$$

$$= P\left[E(S_1^4)(\sigma_{y_1}^2 + \mu_{y_1}^2) + E(S_1^2)E(S_2^2)(\sigma_{y_2}^2 + \mu_{y_2}^2) + 2E(S_1^3)E(S_2)(\sigma_{y_1y_2} + \mu_{y_1}\mu_{y_2})\right]$$
$$+ (1-P)\left[E(S_1^2)E(S_2^2)(\sigma_{y_1}^2 + \mu_{y_1}^2) + E(S_2^4)(\sigma_{y_2}^2 + \mu_{y_2}^2) + 2E(S_1)E(S_2^3)(\sigma_{y_1y_2} + \mu_{y_1}\mu_{y_2})\right]$$
$$- \left[P\{\mu_{y_1}(\theta_1^2 + \gamma_{20}) + \theta_1\theta_2\mu_{y_2}\} + (1-P)\{\theta_1\theta_2\mu_{y_1} + (\theta_2^2 + \gamma_{02})\mu_{y_2}\right]^2 \quad (2.30)$$

On substituting the values of $E(S_1^2)$, $E(S_1^3)$, $E(S_1^4)$, $E(S_2^2)$, $E(S_2^3)$, and $E(S_2^4)$, and re-arranging the terms, we have the result.

The covariance $\sigma_{Z_1Z_2}$ between Z_{1I} and Z_{2i} is given by

$$\sigma_{Z_1Z_2} = Cov(Z_{1i}, Z_{2i}) = E(Z_{1i}Z_{2i}) - E(Z_{1i})E(Z_{2i})$$

$$= PE\left[(S_1Y_{1i} + S_2Y_{2i})(S_1^2Y_{1i} + S_1S_2Y_{2i})\right] + (1-P)E\left[(S_1Y_{1i} + S_2Y_{2i})(S_1S_2Y_{1i} + S_2^2Y_{2i})\right]$$
$$- \left[\theta_1\mu_{y_1} + \theta_2\mu_{y_2}\right]\left[P\{\mu_{y_1}(\theta_1^2 + \gamma_{20}) + \theta_1\theta_2\mu_{y_2}\} + (1-P)\{\theta_1\theta_2\mu_{y_1} + (\theta_2^2 + \gamma_{02})\mu_{y_2}\}\right]$$
$$= P\left[E(S_1^3)(\sigma_{y_1}^2 + \mu_{y_1}^2) + 2E(S_1^2)E(S_2)(\sigma_{y_1y_2} + \mu_{y_1}\mu_{y_2})\right]$$
$$+ (1-P)\left[E(S_1^2)E(S_2)(\sigma_{y_1}^2 + \mu_{y_1}^2) + 2E(S_1)E(S_2^2)(\sigma_{y_1y_2} + \mu_{y_1}\mu_{y_2}) + E(S_2^3)(\sigma_{y_2}^2 + \mu_{y_2}^2)\right]$$
$$- \left[\theta_1\mu_{y_1} + \theta_2\mu_{y_2}\right]\left[P\{\mu_{y_1}(\theta_1^2 + \gamma_{20}) + \theta_1\theta_2\mu_{y_2}\} + (1-P)\{\theta_1\theta_2\mu_{y_1} + (\theta_2^2 + \gamma_{02})\mu_{y_2}\}\right] \quad (2.31)$$

On substituting the values of $E(S_1)$, $E(S_1^2)$, $E(S_1^3)$, $E(S_2)$, $E(S_2^2)$, and $E(S_2^3)$, and rearranging the terms we have the result.

Theorem 2.4. The variance of the estimator $\hat{\mu}_{y_2}$ is given by

$$V(\hat{\mu}_{y_2}) = \frac{\theta_1^2\sigma_{Z_2}^2 + \{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}^2\sigma_{Z_1}^2 - 2\theta_1\{P(\gamma_{20} + \theta_1^2) + (1-P)\theta_1\theta_2\}\sigma_{Z_1Z_2}}{n((1-P)\theta_1\gamma_{02} - P\theta_2\gamma_{20})^2} \quad (2.32)$$

Proof. It follows from the previous theorem.

In the next section, we consider the problem of extension of the above estimators in stratified random sampling.

3. Stratified Random Sampling

The population of N units is first subdivided into L homogeneous subgroups called strata, such that the h^{th} stratum consists of N_h units, where $h = 1, 2, ..., L$ and $\sum_{h=1}^{L} N_h = N$. Consider a sample s_h of size n_h is drawn using SRSWR sampling from the h^{th} population stratum consisting of N_h units such that $\sum_{h=1}^{L} n_h = n$, the required sample size. Assume the value of the i^{th} unit of the two study variables selected from the h^{th} stratum is denoted by Y_{h1i}, and Y_{h2i} where $i = 1, 2, ..., n_h$ and $W_h = N_h/N$ is the known proportion of population units falling in the h^{th} stratum. Let μ_{hy_1} and μ_{hy_2} be the true population means of the two

sensitive variables Y_{h1i} and Y_{h2i}, respectively in the h[th] stratum. Each respondent selected in the sample is asked to generate two fake values of scrambling variables S_{h1} and S_{h2} from two known distributions. In the given h[th] stratum, assume S_{h1} and S_{h2} are independent, which help to maintain the protection of respondents. Let $E(S_{h1}) = \theta_{h1}$, $V(S_{h1}) = \gamma_{h20}$, $E(S_{h2}) = \theta_{h2}$ and $V(S_{h2}) = \gamma_{0h2}$ are known. In the proposed randomized response model, each respondent selected in the sample s_h from the h[th] stratum is requested to report the scrambled response as:

$$Z_{h1i} = S_{h1}Y_{h1i} + S_{h2}Y_{h2i} \tag{3.1}$$

Also each respondent is also requested to rotate a spinner which consists of two outcomes similar to Warner (1965) spinner, but has different types of outcomes. If the pointer lands in a shaded area then the respondent is asked to report the value of the scrambling variable S_{h1} and if the pointer lands in the non-shaded area then the respondent is asked to report the value of the scrambling variable S_{h2}. Let P_h be the proportion of a shaded area and $(1-P_h)$ be the proportion of non-shaded area of the spinner. Thus the second response from the i[th] respondent is given by:

$$Z_{hi} = \begin{cases} S_{h1} & \text{with probability } P_h \\ S_{h2} & \text{with probability } (1 - P_h) \end{cases} \tag{3.2}$$

where $P_h \neq \dfrac{\theta_{h1}\gamma_{h02}}{\theta_{h1}\gamma_{h02} + \theta_{h2}\gamma_{h20}}$.

Without loss of generality, we consider unbiased estimators of μ_{y_1} and μ_{y_2} as

$$\hat{\mu}_{y_{1(st)}} = \sum_{h=1}^{L} W_h \frac{\{P_h\theta_{h1}\theta_{h2} + (1-P_h)(\gamma_{h02} + \theta_{h2}^2)\}\bar{Z}_{h1} - \theta_{2h}\bar{Z}_{h2}}{(1-P_h)\theta_{h1}\gamma_{h02} - P_h\theta_{h2}\gamma_{h20}} \tag{3.3}$$

and

$$\hat{\mu}_{y_2(st)} = \sum_{h=1}^{L} W_h \frac{\theta_{h1}\bar{Z}_{h2} - \{P_h(\gamma_{h20} + \theta_{h1}^2) + (1-P_h)\theta_{h1}\theta_{h2}\}\bar{Z}_{h1}}{(1-P_h)\theta_{h1}\gamma_{h02} - P_h\theta_{h2}\gamma_{h20}} \tag{3.4}$$

The variance of the estimators $\hat{\mu}_{y_1(st)}$ and $\hat{\mu}_{y_2(st)}$ is given by

$$V(\hat{\mu}_{y_{1(st)}}) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_{h1}^2}{n_h} \tag{3.5}$$

and

$$V(\hat{\mu}_{y_{2(st)}}) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_{h2}^2}{n_h} \tag{3.6}$$

where

$$\sigma_{h1}^2 = \frac{\{P_h\theta_{h1}\theta_{h2} + (1-P_h)(\gamma_{h02} + \theta_{h2}^2)\}^2\sigma_{Z_{h1}}^2 + \theta_{h2}^2\sigma_{Z_{h2}}^2 - 2\theta_{h2}\{P_h\theta_{h1}\theta_{h2} + (1-P_h)(\gamma_{h02} + \theta_{h2}^2)\}\sigma_{hZ_1Z_2}}{\{(1-P_h)\theta_{h1}\gamma_{h02} - P_h\theta_{h2}\gamma_{h20}\}^2} \tag{3.7}$$

where

$$\sigma^2_{Z_{h1}} = \gamma_{h20}(\sigma^2_{hy_1} + \mu^2_{hy_1}) + \gamma_{h02}(\sigma^2_{hy_2} + \mu^2_{hy_2}) + \theta^2_{h1}\sigma^2_{hy_1} + \theta^2_{h2}\sigma^2_{hy_2} + 2\theta_{h1}\theta_{h2}\sigma_{hy_1y_2} \quad (3.8)$$

$$\sigma^2_{hZ_2} = (\sigma^2_{hy_1} + \mu^2_{hy_1})\Big[P_h(\gamma_{h40} + 4\gamma_{h30}\theta_{h1} + 6\gamma_{h20}\theta^2_{h1} + \theta^4_{h1}) + (1-P_h)(\gamma_{h20} + \theta^2_{h1})(\gamma_{h02} + \theta^2_{h2})\Big]$$
$$+ (\sigma^2_{hy_2} + \mu^2_{hy_2})\Big[(1-P_h)(\gamma_{h04} + 4\gamma_{h03}\theta_{h2} + 6\gamma_{h02}\theta^2_{h2} + \theta^4_{h2}) + P_h(\gamma_{h20} + \theta^2_{h1})(\gamma_{h02} + \theta^2_{h2})\Big]$$

$$+ 2(\sigma_{hy_1y_2} + \mu_{hy_1}\mu_{hy_2})\Big[P_h\theta_{h2}(\gamma_{h30} - 3\theta_{h1}\gamma_{h20} + \theta^3_{h1}) + (1-P_h)\theta_{h1}(\gamma_{h03} - 3\theta_{h2}\gamma_{h02} + \theta^3_{h2})\Big]$$
$$- \Big[\mu_{hy_1}\{P_h(\gamma_{h20} + \theta^2_{h1}) + (1-P_h)\theta_{h1}\theta_{h2}\} + \mu_{hy_2}\{P_h\theta_{h1}\theta_{h2} + (1-P_h)(\gamma_{h02} + \theta^2_{h2})\}\Big]^2 \quad (3.9)$$

and

$$\sigma_{hZ_1Z_2} = (\sigma^2_{hy_1} + \mu^2_{hy_1})\{P_h(\gamma_{h30} - 3\theta_{h1}\gamma_{h20} + \theta^3_{h1}) + (1-P_h)\theta_{h2}(\gamma_{h20} + \theta^2_{h1})\}$$
$$+ (\sigma^2_{hy_2} + \mu^2_{hy_2})\{P_h\theta_{h1}(\gamma_{h02} + \theta^2_{h2}) + (1-P_h)(\gamma_{h03} - 3\theta_{h2}\gamma_{h02} + \theta^3_{h2})\}$$
$$+ 2(\sigma_{hy_1y_2} + \mu_{hy_1}\mu_{hy_2})\{P_h\theta_{h2}(\gamma_{h20} + \theta^2_{h1}) + (1-P_h)\theta_{h1}(\gamma_{h02} + \theta^2_{h2})\}$$
$$- (\theta_{h1}\mu_{hy_1} + \theta_{h2}\mu_{hy_2})[\mu_{hy_1}\{P_h(\gamma_{h20} + \theta^2_{h1}) + (1-P_h)\theta_{h1}\theta_{h2}\} + \mu_{hy_2}\{P_h\theta_{h1}\theta_{h2} + (1-P_h)(\gamma_{h02} + \theta^2_{h2})\}] \quad (3.10)$$

and

$$\sigma^2_{h2} = \frac{\theta^2_{h1}\sigma^2_{hZ_2} + \{P_h(\gamma_{h20} + \theta^2_{h1}) + (1-P_h)\theta_{h1}\theta_{h2}\}^2\sigma^2_{hZ_1} - 2\theta_{h1}\{P_h(\gamma_{h20} + \theta^2_{h1}) + (1-P_h)\theta_{h1}\theta_{h2}\}\sigma_{hZ_1Z_2}}{((1-P_h)\theta_{h1}\gamma_{h02} - P_h\theta_{h2}\gamma_{h20})^2} \quad (3.11)$$

3.1 Equal Allocation

As the name of the method suggests the sub sample sizes are equal, *i.e.*, $n_h = n/L$. Under this choice of sample allocation, the variance of the estimators $\hat{\mu}_{y1(st)}$ and $\hat{\mu}_{y2(st)}$ reduces to:

$$V(\hat{\mu}_{y_{1(st)}})_E = \frac{L}{n}\sum_{h=1}^{L} W_h^2\sigma^2_{h1} \quad (3.12)$$

and

$$V(\hat{\mu}_{y_{2(st)}})_E = \frac{L}{n}\sum_{h=1}^{L} W_h^2\sigma^2_{h2} \quad (3.13)$$

3.2 Proportional Allocation

As the name of the method suggests the sub sample sizes are equal, *i.e.*, $n_h = nW_h$. Under this choice of sample allocation, the variance of the estimators $\hat{\mu}_{y1(st)}$ and $\hat{\mu}_{y2(st)}$ reduces to:

$$V(\hat{\mu}_{y_{1(st)}})_{PA} = \frac{1}{n}\sum_{h=1}^{L} W_h\sigma^2_{h1} \quad (3.14)$$

and

$$V(\hat{\mu}_{y_{2(st)}})_{PA} = \frac{1}{n}\sum_{h=1}^{L} W_h\sigma^2_{h2} \quad (3.15)$$

3.3 Optimum Allocation

In optimum allocation case, it depends on the choice of importance of characteristics among the two which we are estimating. Thus a compromised optimum allocation is considered. Let C_h be the cost of collecting information from a person (or unit) on both characteristics using the proposed randomization device in the h[th] stratum. Let C_0 be the overhead cost of the survey. Then the total cost function is given by:

$$C = C_0 + \sum_{h=1}^{L} n_h C_h \tag{3.16}$$

Let δ be the weight given to the variance of the estimator of the first mean μ_{y_1} and $(1-\delta)$ be the weight given to the estimator of the second mean μ_{y_2}. Overall we consider the minimization of the compromised variance given by:

$$V_{comp} = \delta V(\hat{\mu}_{y1(st)}) + (1-\delta)V(\hat{\mu}_{y2(st)}) = \sum_{h=1}^{L} W_h^2 \left\{ \frac{\delta\sigma_{h_1}^2 + (1-\delta)\sigma_{h2}^2}{n_h} \right\} \tag{3.17}$$

Minimization of (3.17) subject to (3.16) leads to the optimum sample sizes as:

$$n_h = n \frac{W_h \sqrt{\frac{\delta\sigma_{h1}^2 + (1-\delta)\sigma_{h_2}^2}{C_h}}}{\sum_{h=1}^{L} W_h \sqrt{\frac{\delta\sigma_{h1}^2 + (1-\delta)\sigma_{h_2}^2}{C_h}}} \tag{3.18}$$

In a particular case if $C_1 = C_2 = ... = C_L = C_c$, that is the cost of sampling in each stratum is the same then (3.16) becomes

$$n_h = n \frac{W_h \sqrt{\delta\sigma_{h1}^2 + (1-\delta)\sigma_{h_2}^2}}{\sum_{h=1}^{L} W_h \sqrt{\delta\sigma_{h1}^2 + (1-\delta)\sigma_{h_2}^2}} \tag{3.19}$$

In other words the optimum allocation reduces to the Neyman (1934) type allocation. Under optimum allocation the variances of the proposed estimators take the form:

$$V(\hat{\mu}_{y_{1(st)}})_O = \frac{1}{n} \left[\sum_{h=1}^{L} \frac{W_h \sigma_{h1}^2}{\sqrt{\frac{\delta\sigma_{h1}^2 + (1-\delta)\sigma_{h2}^2}{C_h}}} \right] \left[\sum_{h=1}^{L} W_h \sqrt{\frac{\delta\sigma_{h1}^2 + (1-\delta)\sigma_{h2}^2}{C_h}} \right] \tag{3.20}$$

and

$$V(\hat{\mu}_{y_{2(st)}})_O = \frac{1}{n} \left[\sum_{h=1}^{L} \frac{W_h \sigma_{h2}^2}{\sqrt{\frac{\delta\sigma_{h1}^2 + (1-\delta)\sigma_{h2}^2}{C_h}}} \right] \left[\sum_{h=1}^{L} W_h \sqrt{\frac{\delta\sigma_{h1}^2 + (1-\delta)\sigma_{h2}^2}{C_h}} \right] \tag{3.21}$$

In the next section, we consider different situations to investigate the performance of the proposed estimators for stratified and non-stratified sampling.

4.Numerical Comparisons

In case of use of randomized response device in a survey it is more appropriate to divide the population in two strata based on gender, that is, one stratum made of all males and another stratum made of all females. It is an observation that although the behavior (or say attitude) of males and females remains different towards the use of a randomization device while responding to survey questions especially when these questions are sensitive in nature. In this study, we are using same randomization device in both strata by assuming the behavior (or attitude) of both male and female children remain same towards the use of a proposed randomization device with the same device parameters. Here we consider the problem estimation of $Y_1 = \text{AGE}$ of a child and $Y_2 = \text{FEV}$ (forced expiratory volume). The real data is available in the dataset, FEV.DAT, available on the CD that accompanies the text by Rosner (2006) that contains data on $N = 654$ children from the Childhood Respiratory Disease Study done in Boston. Here for our numerical illustrations we consider AGE and FEV both as sensitive variable. We also consider GENDER to divide the children into two groups, viz. male children and female children. A graphical representation of AGE versus FEV for both genders (Female=0, Male=1) is shown in Fig 4.1.

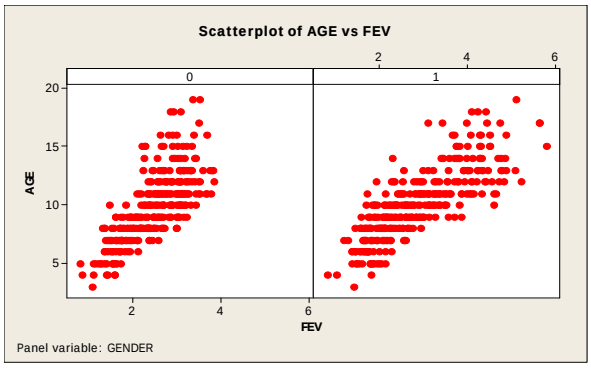

Fig. 4.1. A graphical representation of the population.

The descriptive parameters of the population for both variables AGE and FEV by Gender are given in Table 4.1.

Table 4.1. Descriptive parameters of the population.

Variable	Gender	N_h	Mean	St. Dev	Min	Med	Max
Age	0	318	9.843	2.933	3.00	10.00	19.00
	1	336	10.015	2.976	3.00	10.00	19.00
FEV	0	318	2.4512	0.6457	0.7910	2.486	3.835
	1	336	2.8124	1.0036	0.7960	2.606	5.793

Here we investigate the percent relative efficiency of the stratified random proportional allocation with respect to simple random allocation by using the same randomization device across both strata, that is, for males and females. The percent relative efficiency of the stratified random sampling estimator for the first sensitive variable is defines as:

$$\text{RE}(1) = \frac{V(\hat{\mu}_{y1})}{V(\hat{\mu}_{y_1(st)})_{PA}} \times 100\% \qquad (4.1)$$

and that for the second sensitive variable is defined as:

$$\text{RE}(2) = \frac{V(\hat{\mu}_{y2})}{V(\hat{\mu}_{y_2(st)})_{PA}} \times 100\% \qquad (4.2)$$

Note that RE(1) and RE(2) values in (4.1) and(4.2) are free from the value of the sample size n, however these depends on the values of the randomization device parameters P, θ_1, θ_2, γ_{20}, γ_{02}, γ_{30}, γ_{03}, γ_{40} and γ_{04}. Here we list a few choices of such parameters to make a randomization device which makes sure that the proposed estimator in stratified sampling can perform slightly better than the naive estimators while using sample random with replacement sampling (SRSWR). The gain in efficiency is only due to stratification, because we are using same randomization device in both strata. We wrote SAS Macro to get an idea of gain in efficiency due to stratification as given in APPENDIX.

The results obtained after executing the SAS codes are given in Table 4.2, where we fixed other parameters $\theta_1 = 2.6$, $\theta_2 = 4.5$, $\gamma_{20} = 2.3$, $\gamma_{02} = 1.2$, $\gamma_{30} = 2.2$, $\gamma_{03} = 2.4$, $\gamma_{40} = 3.2$ and $\gamma_{04} = 3.2$.

Table 4.2. Percent Relative Efficiency values.

Obs	P	RE(1)	RE(2)
1	0.48	97.34	60.97
2	0.49	94.04	105.52
3	0.50	123.28	102.56
4	0.51	103.91	101.67
5	0.52	102.12	101.24
6	0.53	101.45	100.98
7	0.55	100.89	100.69

From the population of the type shown in Fig 4.1, it is very hard to gain in efficiency of an estimator due to stratification because the variation with both strata is almost same and we are using same randomization device in both strata. It is very interesting that maximum gain of 123.28% value of RE(1) along with RE(2) value of 102.56% is attained when the value of P is 0.5, that is there is 50% chance of getting second response as S_1 or S_2. Note that one has to be very careful while making a randomization device to be used in a real survey. To our knowledge, these estimators are being first time introduced in the field of survey sampling thus a pilot survey or a simulation study based on a hypothetical population similar to the one under investigation would be helpful in making a randomization device which could ensure both privacy and efficiency.

References

Eichhorn, B.H. and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *J. Statist. Planning and Infer.*, 7, 307-316.

Fox, J. A. (2016). *Randomized Response and Related Methods*. SAGE, Los Angeles.

Fox, J.A. and Tracy, P.E. (1984). Measuring associations with randomized response. *Social Science Research*, 13, 188-197.

Greenberg, B.G.m Kuebler, R.R., Abernathy, J.R. and Horvitz, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *J. Amer. Statist. Assoc.*, 66, 243-250

Himmelfarb, S. and Edgell, S.E. (1980). Additive constants model: A randomized response technique for eliminating evasiveness to quantitative response questions. *Psychological Bulletin*, 87, 525-530.

Horvitz, D.G., Shah, B.V., and Simmons, W.R. (1967). The unrelated question randomized response model. *Proc. of Social Statistics Section, Amer. Stat. Assoc.*, 65-72.

Rosner, B. (2006). *Fundamentals of Biostatistics.* Belmont, CA: Thomson-Brooks/Cole.

Singh, S. (2016). On the estimation of correlation coefficient using scrambled responses. *Handbook of Statistics: Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, Edited by Chaudhuri, Christofides and Rao, 34, 43-90, Elsevier: North-Holand.

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

Appendix

```
*SAS MACRO USED IN THE NUMERICAL ILLUSTRATIONS;
PROC IMPORT DATAFILE = "C:\SASDATAFILES\FEV.XLS"  OUT=DATA1 DBMS=XLS
REPLACE; SHEET="FEV";
*PROC PRINT DATA=DATA1 (OBS=5);
RUN;
DATA DATA1;
SET DATA1;
DATA DATA2;
SET DATA1;
KEEP GENDER AGE FEV;
DATA DATA3;
SET DATA2;
Y1 = AGE;
Y2 = FEV;
KEEP Y1 Y2;
```

```
PROC MEANS DATA = DATA3 NOPRINT;
VAR Y1 Y2;
OUTPUT OUT = DATA4 MEAN = YM1 YM2;
DATA DATA5;
SET DATA4;
DROP _TYPE_ _FREQ_;
DATA DATA6;
SET DATA3;
IF _N_=1 THEN SET DATA5;
DY1 = Y1 - YM1;
DY2 = Y2 - YM2;
SSDY1 = DY1**2;
SSDY2 = DY2**2;
DY1DY2 = DY1*DY2;
DATA DATA7;
SET DATA6;
KEEP Y1 Y2 SSDY1 SSDY2 DY1DY2;
PROC MEANS DATA = DATA7 NOPRINT;
VAR Y1 Y2 SSDY1 SSDY2 DY1DY2;
OUTPUT OUT = DATA8 MEAN = MEANY1 MEANY2 SIG2Y1 SIG2Y2 SIGY1Y2;
DATA DATA9A;
SET DATA8;
NP = _FREQ_;
KEEP NP;
%MACRO SEGUN(KKK, PP, INTH1, INTH2, ING20, ING02);
DATA DATA9;
SET DATA8;
P = &PP;
TH1 = &INTH1;
TH2 = &INTH2;
G20 = &ING20;
G02 = &ING02;
G40 = 3.2;
G04 = 3.2;
G30 = 2.2;
G03 = 2.4;
DENO = (1-P)*TH1*G02-P*TH2*G20;
SIG2Z1 = G20*(SIG2Y1+MEANY1**2) + G02*(SIG2Y2 + MEANY2**2)+TH1**2*SIG2Y1 +
TH2**2*SIG2Y2+2*TH1*TH2*SIGY1Y2;
T1 = P*(G40+4*G30*TH1+6*G20*TH1**2) + (1-P)*(G20+TH1**2)*(G02+TH2**2);
T2 = (1-P)*(G04 +4*G03*TH2+6*G02*TH2**2+TH2**4)+P*(G20+TH1**2)*(G02+TH2**2);
T3 = P*TH2*(G30-3*TH1*G20+TH1**3) + (1-P)*TH1*(G03-3*TH2*G02+TH2**3);
T4 = MEANY1*(P*(G20+TH1**2)+(1-P)*TH1*TH2) + MEANY2*(P*TH1*TH2 +(1-
P)*(G02+TH2**2));
SIG2Z2 = (SIG2Y1+MEANY1**2)*T1 + (SIG2Y2+MEANY2**2)*T2 + 2*SIGY1Y2 - T4**2;
T5 = P*(G30-3*TH1*G20+TH1**3) + (1-P)*TH2*(G20+TH1**2);
T6 = P*TH1*(G02+TH2**2)+(1-P)*TH1*(G03-3*TH2*G02+TH2**3);
T7 = P*TH2*(G20+TH1**2) + (1-P)*TH1*(G02+TH2**2);
SIGZ1Z2 = (SIG2Y1+MEANY1**2)*T5 + (SIG2Y2+MEANY2**2)*T6 + 2
*(SIGY1Y2+MEANY1*MEANY2)*T7-(TH1*MEANY1 + TH2*MEANY2)*T4;
DATA DATA100;
SET DATA9;
KEEP P T1 T2 T3 T4 T5 T6 T7 DENO G20 G02 TH1 TH2;
DATA DATA10;
SET DATA9;
NOMIY1 = (P*TH1*TH2+(1-P)*(G02+TH2**2))**2*SIG2Z1 +TH2**2*SIG2Z2 -
2*TH2*(P*TH1*TH2+(1-P)*(G02+TH2**2))*SIGZ1Z2;
VARY1 = NOMIY1/DENO**2;
NOMIY2 = TH1**2*SIG2Z2 +(P*(G20+TH1**2)+(1-P)*TH1*TH2)**2*SIG2Z1 -
2*TH1*(P*(G20+TH1**2)+(1-P)*TH1*TH2)*SIGZ1Z2;
VARY2 = NOMIY2/DENO**2;
DATA DATA11;
```

```
SET DATA10;
KEEP VARY1 VARY2;
*PROC PRINT DATA = DATA11;
RUN;
DATA DATA12;
SET DATA2;
Y1 = AGE;
Y2 = FEV;
KEEP Y1 Y2 GENDER;
PROC SORT DATA = DATA12;
BY GENDER;
PROC MEANS DATA=DATA12 NOPRINT;
VAR Y1 Y2;
BY GENDER;
OUTPUT OUT=DATA13 MEAN = MY1GEN MY2GEN;
DATA DATA14;
SET DATA13;
DROP _TYPE_;
PROC SORT DATA=DATA14;
BY GENDER;
DATA DATA15;
MERGE DATA12 DATA14;
BY GENDER;
*PROC PRINT DATA=DATA15;
DATA DATA16;
SET DATA15;
DY1G = Y1-MY1GEN;
DY2G = Y2-MY2GEN;
SSDY1G = DY1G**2;
SSDY2G = DY2G**2;
DY1DY2G = DY1G*DY2G;
KEEP GENDER SSDY1G SSDY2G DY1DY2G Y1 Y2;
*PROC PRINT DATA=DATA16;
PROC MEANS DATA=DATA16 NOPRINT;
VAR SSDY1G SSDY2G DY1DY2G Y1 Y2;
BY GENDER;
OUTPUT OUT = DATA17 MEAN = VARY1G VARY2G CY1Y2G Y1MG Y2MG;
DATA DATA18;
SET DATA17;
DROP _TYPE_;
IF _N_=1 THEN SET DATA100;
KEEP T1 T2 T3 T4 T5 T6 T7 VARY1G VARY2G CY1Y2G GENDER Y1MG Y2MG DENO
G02 G20 TH1 TH2 P;
DATA DATA19;
SET DATA18;
SIG2Z1G = G20*(VARY1G+Y1MG**2) + G02*(VARY2G + Y2MG**2)+TH1**2*VARY1G +
TH2**2*VARY2G+2*TH1*TH2*CY1Y2G;
SIG2Z2G = (VARY1G+Y1MG**2)*T1 + (VARY2G+Y2MG**2)*T2 + 2*CY1Y2G - T4**2;
SIGZ1Z2G = (VARY1G+Y1MG**2)*T5 + (VARY2G+Y2MG**2)*T6 + 2
*(CY1Y2G+Y1MG*Y2MG)*T7-(TH1*Y1MG + TH2*Y2MG)*T4;
NOMIY1G = (P*TH1*TH2+(1-P)*(G02+TH2**2))**2*SIG2Z1G +TH2**2*SIG2Z2G -
2*TH2*(P*TH1*TH2+(1-P)*(G02+TH2**2))*SIGZ1Z2G;
VARY1G = NOMIY1G/DENO**2;
NOMIY2G = TH1**2*SIG2Z2G +(P*(G20+TH1**2)+(1-P)*TH1*TH2)**2*SIG2Z1G -
2*TH1*(P*(G20+TH1**2)+(1-P)*TH1*TH2)*SIGZ1Z2G;
VARY2G = NOMIY2G/DENO**2;
KEEP GENDER VARY1G VARY2G;
*PROC PRINT DATA=DATA19;
RUN;
PROC FREQ DATA=DATA2 NOPRINT;
TABLE GENDER/OUT=DATA200;
RUN;
```

```
DATA DATA200;
SET DATA200;
KEEP GENDER COUNT;
*PROC PRINT DATA=DATA200;
RUN;
PROC SORT DATA = DATA200;
BY GENDER;
PROC SORT DATA = DATA19;
BY GENDER;
DATA DATA20;
MERGE DATA19 DATA200;
BY GENDER;
DATA DATA21;
SET DATA20;
IF _N_=1 THEN SET DATA9A;
*PROC PRINT DATA=DATA21;
RUN;
DATA DATA22;
SET DATA21;
WI = COUNT/NP;
VARGY1 = WI*VARY1G;
VARGY2 = WI*VARY2G;
PROC MEANS DATA = DATA22 NOPRINT;
VAR VARGY1 VARGY2;
OUTPUT OUT = DATA23 SUM=VARY1ST VARY2ST;
DATA DATA23;
SET DATA23;
KEEP VARY1ST VARY2ST ;
DATA DATA24;
SET DATA23;
IF _N_ = 1 THEN SET DATA11;
RE1 = VARY1*100/VARY1ST;
RE2 = VARY2*100/VARY2ST;
DATA DATA25;
SET DATA9;
IF _N_=1 THEN SET DATA24;
KEEP P TH1 G20 G30 G40 TH2 G02 G03 G04 RE1 RE2;
PROC PRINT DATA = DATA25;
DATA DATA41&KKK;
SET DATA25;
PROC APPEND DATA=DATA41&KKK OUT=DATA2222;
RUN;
%MEND SEGUN;
%SEGUN(1, 0.50, 2.6, 4.5, 2.3, 1.2);
%SEGUN(2, 0.51, 2.6, 4.5, 2.3, 1.2);
%SEGUN(3, 0.52, 2.6, 4.5, 2.3, 1.2);
%SEGUN(4, 0.53, 2.6, 4.5, 2.3, 1.2);
%SEGUN(5, 0.55, 2.6, 4.5, 2.3, 1.2);
PROC PRINT DATA=DATA2222;
RUN;
```