CHAPTER 16

Randomized Response Techniques

16.1 INTRODUCTION

In socioeconomic and biometric research, we very often gather information relating to highly sensitive issues such as induced abortion, drug addiction, HIV infection status, duration of suffering from AIDS, sexual behavior, incidence of domestic violence, and tax evasion, for example. In these situations employing the direct method of interview (asking questions directly to the respondents), the respondents often provide untrue responses or even refuse to respond because of social stigma and/or fear. Under such circumstances the randomized response (RR) techniques may be used to collect more reliable data, protect respondents' confidentiality, and avoid unacceptable rate of nonresponse. The RR technique was originated by Warner (1965).

Warner's (1965) technique was modified by Horvitz et al. (1967), Greenberg et al. (1969), Raghavrao (1971), Franklin (1989), Arnab (1990, 1996), Kuk (1990), Mangat and Singh (1990), Arcos et al. (2015), Rueda et al. (2015), among other researchers, to increase cooperation from respondents and efficiency. Applications of the RR technique in real-life situations were reported by Greenberg et al. (1969): Illegitimacy of offspring; Abernathy et al. (1970): Incidence of induced abortions; Goodstadt and Gruson (1975): Drug uses; Folsom (1973): Drinking and driving; Van der Heijden et al. (1998): Social security fraud, and Arnab and Mothupi (2015): Sexual habits of University students. Details have been given by Chaudhuri and Mukherjee (1988), Singh (2003), and Chaudhuri (2011). In this chapter, some of the RR techniques that are used for qualitative and quantitative characteristics have been described. Methods of estimation of population characteristics for various sampling designs, measures of protection of privacy, and optimality of RR strategies under various superpopulation models have also been described.

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00016-9

16.2 RANDOMIZED RESPONSE TECHNIQUES FOR QUALITATIVE CHARACTERISTICS

16.2.1 Warner's Technique: the Pioneering Method

The RR technique was introduced by Warner (1965) for estimating π_A, the proportion of units in a population possessing a certain stigmatized character A such as HIV infection status. In Warner's method, a sample s of size n is selected from a population U of size N by simple random sampling with replacement (SRSWR) procedure. Because the character A is sensitive, the respondent (unit) need not supply information directly about whether he/she possesses the character A (member of the group A), instead he/she is asked to provide an RR based on an RR device as follows:

The respondent has to select a card at random from a pack of well-scaffolded cards. The pack consists of two types of cards with known proportions, which are identical in appearance. Card type 1, with proportion $P_1(\neq 1/2)$ contains the question "Do you belong to group A?" whereas card type 2 with proportion $1 - P_1$ contains the question "Do you belong to group $\overline{A}$?" where $\overline{A}$ is the complement of group A. The respondent will supply a truthful answer "Yes" or "No" for the question mentioned in the selected card. The experiment is performed in the absence of the interviewer and hence the privacy of the respondent is maintained because the interviewer will not know which of the two questions the respondent has answered. We will call this RR model as R_w.

16.2.1.1 Estimation of Proportion

Because the sample s is selected by SRSWR method, a unit may be selected more than once in s. In particular, if a respondent is selected k (≥ 1) times, then he/she has to perform the randomized device k times independently and supplies k RRs. Let λ_{wA} be the proportion of "Yes" answers obtained from the respondents selected in the sample. Then we have the following theorem.

Theorem 16.2.1

(i) $\widehat{\pi}_{Aw} = \dfrac{\lambda_{wA} - (1 - P_1)}{2P_1 - 1}$ is an unbiased estimator of π_A

(ii) $Var(\widehat{\pi}_{Aw}) = \dfrac{\pi_A(1 - \pi_A)}{n} + \dfrac{P_1(1 - P_1)}{n(2P_1 - 1)^2}$

(iii) An unbiased estimator of $Var(\widehat{\pi}_{Aw})$ is

$$\widehat{V}ar(\widehat{\pi}_{Aw}) = \frac{\lambda_{wA}(1 - \lambda_{wA})}{(n-1)(2P_1 - 1)^2}$$

Proof

Let θ_w be the probability of getting a "Yes" response from a person selected at rth ($r = 1,\ldots, n$) draw. Then,

$\theta_w =$ Conditional probability of getting a "Yes" response from a respondent given that the person belongs to group $A\times$ unconditional probability of the selected person belonging to group $A+$ conditional probability of getting "Yes" response given that the person belongs to group $\overline{A}\times$ unconditional probability of the selected person belonging to group $\overline{A}$

$$= P_1\pi_A + (1 - P_1)(1 - \pi_A) \tag{16.2.1}$$

Because probability of getting a "Yes" answer at any draw is θ_w and the RRs are independent, the proportion λ_{wA} follows the binomial distribution with parameters and θ_w. So we get

(i) $E(\lambda_{wA}) = \theta_w = \pi_A(2P_1 - 1) + (1 - P_1)$ and hence $E(\widehat{\pi}_A) = \pi_A$

(ii) $$Var(\widehat{\pi}_{Aw}) = \frac{Var(\lambda_{wA})}{(2P_1 - 1)^2}$$

$$= \frac{\theta_w(1 - \theta_w)}{n(2P_1 - 1)^2}$$

$$= \frac{\pi_A(1 - \pi_A)}{n} + \frac{P_1(1 - P_1)}{n(2P_1 - 1)^2}$$

(iii) Since $\widehat{V}ar(\lambda_{wA}) = \dfrac{\lambda_{wA}(1 - \lambda_{wA})}{n - 1}$ is an unbiased estimator of $Var(\lambda_{wA})$, we find that

$$\widehat{V}ar(\widehat{\pi}_{Aw}) = \frac{\widehat{V}ar(\lambda_{wA})}{(2P_1 - 1)^2}$$

$$= \frac{\lambda_{WA}(1 - \lambda_{WA})}{(n - 1)(2P_1 - 1)^2}$$

16.2.1.2 Comparison With Direct Response Surveys

Suppose that the selected individual were asked directly whether he/she belongs to the sensitive group or not and $\widehat{\pi}_A$ is the proportion of "Yes" answers in the sample. In this situation we would get $E(\widehat{\pi}_A) = \pi_A$ and $Var(\widehat{\pi}_A) = \dfrac{\pi_A(1 - \pi_A)}{n}$. Clearly, $Var(\widehat{\pi}_{Aw})$ exceeds $Var(\widehat{\pi}_A)$ and the loss

of efficiency of using Warner's RR technique used for the protection of confidentiality of the respondents is

$$1 - Var(\widehat{\pi}_A)/Var(\widehat{\pi}_{Aw}) = \left[1 + \frac{\pi_A(1-\pi_A)}{P_1(1-P_1)}(2P_1-1)^2\right]^{-1} \tag{16.2.2}$$

16.2.1.3 Maximum Likelihood Estimation of Proportion

Let m be the total number of "Yes" answers and $n-m$ be the total number of "No" answers obtained from an RR survey data. The likelihood function is given by

$$L = \binom{n}{m} \theta_w^m \quad (1-\theta_w)^{n-m} \tag{16.2.3}$$

Maximizing L with respect to θ_w, Warner observed that the maximum likelihood estimator (MLE) of θ_w is $\widehat{\theta}_w = m/n = \lambda_{wA}$ and consequently the MLE of π_A is

$$\widehat{\pi}_{Aw} = \left\{\widehat{\theta}_w - (1-P_1)\right\} \Big/ (2P_1 - 1) \quad \text{with} \quad P_1 \neq 1/2 \tag{16.2.4}$$

Singh (1976, 1978) pointed out that the proposed estimators $\widehat{\theta}_w$, as well as λ_{wA}, are not the MLEs of θ_w and π_A, respectively, because they can take values outside the parametric space [0,1]. For example, $\widehat{\pi}_{Aw} = -0.25$, when $P_1 = 0.7$ and $\lambda_{wA} = 0.2$.

In case $P_1 > 1/2$, then θ_w must lie in the interval $[1-P_1, P_1]$ as $\pi_A \in [0,1]$.

Hence for $P_1 > 1/2$, the true MLE π_A is

$$\widetilde{\pi}_A = \begin{cases} \widehat{\pi}_{Aw} & \text{if } 1-P_1 < \lambda_{wA} < P_1 \\ 1 & \text{if } \lambda_{wA} \geq P_1 \\ 0 & \text{if } \lambda_{wA} \leq 1-P_1 \end{cases} \tag{16.2.5}$$

Similarly for $P_1 < 1/2$, the MLE of π_A is

$$\widetilde{\pi}_A = \begin{cases} \widehat{\pi}_{Aw} & \text{if } P_1 < \lambda_{wA} < 1-P_1 \\ 1 & \text{if } \lambda_{wA} \leq P_1 \\ 0 & \text{if } \lambda_{wA} \geq 1-P_1 \end{cases} \tag{16.2.6}$$

Singh (1978) reported that the estimator $\widehat{\pi}_{Aw}$ is inadmissible as the mean square errors of $\widehat{\pi}_{Aw}$ are larger than those of $\widetilde{\pi}_A$. For further information, readers are referred to Devore (1977) and Chaudhuri and Mukherjee (1988).

16.2.2 Greenberg et al.: Unrelated Question Method

Greenberg et al. (1969) modified Warner's method by incorporating a sensitive question (character y) along with a nonsensitive question (character x). In this method, each of the respondents selected in the sample has to pick a card at random from a pack containing two types of identical-looking cards with known proportions as in Warner's model. The type 1 cards bear the sensitive question "Do you belong to the sensitive group A?" with proportion $P_2(\neq 0)$ whereas card type 2 (with proportion $1 - P_2$) bears a question of unrelated or nonsensitive characteristic x such as "Are you an African?". We will call this RR technique as R_g. Greenberg et al. anticipated that this method may receive greater cooperation from respondents as it boosts the degree of privacy.

16.2.2.1 Estimation of Proportion

(i) π_x is known.

Let a sample of n units be selected by SRSWR method. Each of the respondents in the sample has to perform an RR independently. The probability of getting a "Yes" answer for this R_g technique is $\theta_g = P_2\pi_A + (1 - P_2)\pi_x$, where π_A and π_x are the proportions of persons in the population possessing the sensitive character y and nonsensitive character x, respectively. Let λ_g be the proportion of "Yes" answers in the sample and assume that π_x is known. Then we arrive at the following theorem.

Theorem 16.2.2

(i) $\widehat{\pi}_{Ag} = \dfrac{\lambda_g - (1 - P_2)\ \pi_x}{P_2}$ is an unbiased estimator of π_A

(ii) $Var\left(\widehat{\pi}_{Ag}\right) = \dfrac{\theta_g(1 - \theta_g)}{nP_2^2}$

where $\theta_g = P_2\pi_A + (1 - P_2)\pi_x$

(iii) An unbiased estimator of $Var\left(\widehat{\pi}_{Ag}\right)$ is

$$\widehat{Var}\left(\widehat{\pi}_{Ag}\right) = \frac{\lambda_g(1 - \lambda_g)}{(n - 1)P_2^2}$$

Proof

(i) Here we note that $E(\lambda_g) = \theta_g = P_2\pi_A + (1 - P_2)\pi_x$ giving $E\left(\widehat{\pi}_g\right) = \pi_A$

The remaining parts (ii) and (iii) follow from Theorem 16.2.1.

(ii) π_x is unknown.

In case π_x is unknown, two independent SRSWR samples s_1 and s_2 of sizes n_1 and $n_2(=n - n_1)$, respectively, are selected from the population. Each of the respondents in sample s_1 performs a randomized device R_g with $P_2 = P_{21}$ whereas the individuals in sample s_2 perform R_g with $P_2 = P_{22}$ $(\neq P_{21})$. Let $\lambda_g(1)$ and $\lambda_g(2)$ be the proportion of "Yes" answers for the samples s_1 and s_2, respectively.

Theorem 16.2.3

(i) $\widehat{\pi}^*_{Ag} = \dfrac{(1 - P_{22})\lambda_g(1) - (1 - P_{21})\lambda_g(2)}{P_{21} - P_{22}}$ is an unbiased estimator of π_A

(ii) $Var\left(\widehat{\pi}^*_{Ag}\right) = \dfrac{1}{(P_{21} - P_{22})^2}\left[\dfrac{(1 - P_{22})^2\theta_{g1}(1 - \theta_{g1})}{n_1} + \dfrac{(1 - P_{21})^2\theta_{g2}(1 - \theta_{g2})}{n_2}\right]$

where $\theta_{gi} = P_{2i}\pi_A + (1 - P_{2i})\pi_x$ for $i = 1, 2$.

(iii) An unbiased estimator of $Var\left(\widehat{\pi}^*_{Ag}\right)$ is

$$\widehat{V}ar\left(\widehat{\pi}^*_{Ag}\right) = \frac{1}{(P_{21} - P_{22})^2}\left[\frac{(1 - P_{22})^2\lambda_g(1)(1 - \lambda_g(1))}{n_1 - 1} + \frac{(1 - P_{21})^2\lambda_g(2)(1 - \lambda_g(2))}{n_2 - 1}\right]$$

Proof

The theorem can be verified easily by noting $E[\lambda_g(i)] = \theta_{gi}$, $Var[\lambda_g(i)] = \dfrac{\theta_{gi}(1 - \theta_{gi})}{n_i}$, and an unbiased estimator of $Var[\lambda_g(i)]$ is $\widehat{V}ar[\lambda_g(i)] = \dfrac{\lambda_g(i)(1 - \lambda_g(i))}{n_i - 1}$ for $i = 1, 2$.

The expression $Var\left(\widehat{\pi}^*_{Ag}\right)$ becomes small if the difference $P_{21} - P_{22}$ is large, hence for the optimal choices of P_{2i}'s, one should choose one of the P_{2i}'s to be as large as possible and the other as small as possible. The optimum value of n_i $(i = 1,2)$ obtained by minimizing $Var\left(\widehat{\pi}_{g1}\right)$ with respect to n_i, keeping $n(=n_1 + n_2)$ fixed is given by

$$opt(n_i) = n\frac{\sqrt{(1 - P_{22})^2\theta_{g1}(1 - \theta_{g1})}}{\sqrt{(1 - P_{22})^2\theta_{g1}(1 - \theta_{g1})} + \sqrt{(1 - P_{21})^2\theta_{g2}(1 - \theta_{g2})}}$$

The minimum variance of $Var\left(\widehat{\pi}^{*}_{Ag}\right)$ with $n_i = opt(n_i)$ is given by

$$V_{\min}\left(\widehat{\pi}^{*}_{Ag}\right) = \frac{\left[\sqrt{(1-P_{22})^2\theta_{g1}(1-\theta_{g1})} + \sqrt{(1-P_{21})^2\theta_{g2}(1-\theta_{g2})}\right]^2}{n(P_{21}-P_{22})^2}$$

16.2.3 Kuk's Model

In Kuk's (1990) randomized device, two boxes each containing black and white cards with known proportions $P_{31}, 1-P_{31}$ and $P_{32}, 1-P_{32} (P_{31} \neq P_{32})$ are constructed. Respondents belonging to group A should choose box 1 whereas others (those belonging to $\overline{A}$) should choose box 2 and draw c cards at random and with replacement. The respondent will then report the number of black cards drawn, as his/her RR (z). We will denote Kuk's RR technique by R_k.

Let a sample of size n be selected from the population by SRSWR and λ_{kr} be the proportion of black cards chosen by the respondent selected at the rth draw, $r = 1, \ldots, n$. Then

$$\begin{aligned} E(\lambda_{kr}) &= \text{Probability}(\text{respondent} \in A) \times E(\lambda_{kr}|\text{respondent} \in A) \\ &\quad + \text{Probability}(\text{respondent} \notin A) \times E_R(\lambda_{kr}|\text{respondent} \notin A) \\ &= \pi_A P_{31} + (1-\pi_A)P_{32} \\ &= \theta_k \end{aligned} \tag{16.2.7}$$

$$\begin{aligned} V(\lambda_{kr}) &= E\left(\lambda_{kr}^2\right) - \theta_k^2 \\ &= \text{Probability}(\text{respondent} \in A) \times E\left(\lambda_{kr}^2|\text{respondent} \in A\right) \\ &\quad + \text{Probability}(\text{respondent} \notin A) \times E\left(\lambda_{kr}^2|\text{the respondent} \notin A\right) - \theta_k^2 \\ &= \pi_A\left(\frac{P_{31}(1-P_{31})}{c} + P_{31}^2\right) + (1-\pi_A)\left(\frac{P_{32}(1-P_{32})}{c} + P_{32}^2\right) - \theta_k^2 \\ &= \frac{\pi_A P_{31} + (1-\pi_A)P_{32}}{c} + \left(1-\frac{1}{c}\right)\left\{\pi_A P_{31}^2 + (1-\pi_A)P_{32}^2\right\} - \theta_k^2 \\ &= \frac{\theta_k(1-\theta_k)}{c} + \left(1-\frac{1}{c}\right)\pi_A(1-\pi_A)(P_{31}-P_{32})^2 \\ &= \Phi_k \end{aligned} \tag{16.2.8}$$

Since λ_{kr} are iid random variables with mean θ_k and variance Φ_k, we have the following result.

Theorem 16.2.4

(i) $\widehat{\pi}_{Ak} = \dfrac{\overline{\lambda}_k - P_{32}}{P_{31} - P_{32}}$ is an unbiased estimator of π_A

where $\overline{\lambda}_k = \dfrac{1}{n}\sum_{r=1}^{n} \lambda_{kr}$ and $P_{31} \neq P_{32}$.

(ii) $V(\widehat{\pi}_{Ak}) = \dfrac{\Phi_k}{n(P_{31} - P_{32})^2}$

$$= \frac{1}{n}\left[\frac{\theta_k(1-\theta_k)}{c(P_{31} - P_{32})^2} + \pi_A(1-\pi_A)\left(1 - \frac{1}{c}\right)\right]$$

(iii) $\widehat{V}(\widehat{\pi}_{Ak}) = \dfrac{1}{n(n-1)(P_{31} - P_{32})^2}\sum_{r=1}^{n}\left(\lambda_{kr} - \overline{\lambda}_k\right)^2$ is an unbiased estimator of $V(\widehat{\pi}_{wk})$

The variance $V(\widehat{\pi}_{wk})$ decreases as c increases. Hence we can increase the efficiency of $\widehat{\pi}_{Ak}$ by increasing c. Kook's RR technique R_k reduces to Warner's technique R_w when $P_{31} = P_{32} = P$ and $c = 1$.

16.2.4 Mangat and Singh Model

Mangat and Singh (1990) proposed the two-stage RR model where in stage 1, the sampled respondent is to pick a card at random from a pack containing two types of cards with proportions T and $1 - T$ with written statements "disclose your membership to A or $\overline{A}$" and "go to randomized device R_w," respectively. If a respondent selects the card written "disclose your membership to A or $\overline{A}$", then the respondent has to reveal his/her membership to A or $\overline{A}$ truthfully, i.e., has to report "Yes" if he/she belongs to A and "No" otherwise. On the other hand if the respondent picks a card written "go to randomized device R_w", then the respondent performs Warner RR device. In this case the respondent is to pick a card from a pack containing two types of cards written "I belong to group A" with proportion P_4 and "I am a member of $\overline{A}$" with proportion $1 - P_4$, respectively, and answer "Yes" or "No." The confidentiality of the respondent is maintained because the entire experiment was performed in absence of the investigator. We call this RR technique as R_{ms}. Mangat and Singh (1990) used SRSWR method for the selection of the sample. Clearly the two-stage model reduces to Warner model if $T = 0$. For this R_{ms} model the probability of obtaining answer "Yes" from a respondent selected by SRSWR sampling procedure is

$$\theta_{ms} = \pi_A\{T + (1-T)(2P_4 - 1)\} + (1 - P_4)(1 - T) \qquad (16.2.9)$$

Let λ_{ms} be the proportion of "Yes" answers obtained from the sample of n respondents. Noting $E(\lambda_{ms}) = \theta_{ms}$, $Var(\lambda_{ms}) = \dfrac{\theta_{ms}(1-\theta_{ms})}{n}$, and $\widehat{V}ar(\lambda_{ms}) = \dfrac{\lambda_{ms}(1-\lambda_{ms})}{n-1}$, an unbiased estimator of $Var(\lambda_{ms})$, we have the following theorem.

Theorem 16.2.5

(i) $\widehat{\pi}_{Ams} = \dfrac{\lambda_{ms} - (1-P_4)(1-T)}{D}$ is an unbiased estimator of π_A
where $D = T + (1-T)(2P_4 - 1)$ and $T \neq \dfrac{1-2P_4}{2(1-P_4)}$

(ii) $V(\widehat{\pi}_{Ams}) = \dfrac{\theta_{ms}(1-\theta_{ms})}{nD^2}$

$$= \frac{1}{n}\left[\pi_A(1-\pi_A) + \frac{(1-T)(1-P_4)\{1-(1-T)(1-P_4)\}}{D^2}\right]$$

(iii) $\widehat{V}(\widehat{\pi}_{Ams}) = \dfrac{\lambda_{ms}(1-\lambda_{ms})}{(n-1)D^2}$ is an unbiased estimator of $V(\widehat{\pi}_{Ams})$

16.3 EXTENSION TO MORE THAN ONE CATEGORIES

So far, we have classified the population into two categories viz. sensitive and nonsensitive. Now, we consider situations where the population is divided into more than two mutually exclusive and exhaustive categories. For example, married women may be classified into three categories, two of which are sensitive viz. having exactly one, or more than one sexual partner other than their husbands and the nonsensitive category comprises of those who have no other sexual partner other than their husbands. To develop the theory, let us suppose that the population is classified into k mutually exclusive and exhaustive categories among which at most $k-1$ of them are designated as sensitive categories. Let the proportion of persons belonging to the jth category be π_j with $\sum \pi_j = 1$. Here we will consider the method of estimating of π_j for $j = 1,\ldots, k$.

16.3.1 Liu and Chow's Technique

In Liu and Chow's (1976) RR technique, two different colors (red and green, for example) of balls but identical in shape are placed inside a flask with a transparent narrow neck, through which only one ball can pass. Green balls are marked with the numbers 1, 2,…, k. The proportions of red balls and green balls marked with the number j are $p(>0)$ and $p_j(>0)$, where

$j = 1,\ldots, k$, respectively. Clearly, $\sum p_j = 1 - p$. Respondents are asked to close the flask, shake it thoroughly, and then turn it upside down and are told to concentrate on the bottommost ball. The respondent is required to report his/her category truthfully if the bottommost ball is red, otherwise (if it is green) he/she should report the number written on the green ball. The whole experiment is to be performed in the absence of the investigator. So the respondent is to report only one number between 1 and k. The confidentiality is maintained because the investigator will know only one number between 1 and k but will not know whether that number represents respondent's true category or the number on the bottommost green ball.

16.3.1.1 Estimation of Proportions

Let a sample of size n be selected from the population by SRSWR procedure and n_j be the number of times the RR "j" was obtained ($\sum n_j = n$). Under this RR technique the probability of getting a response "j" is

$$\begin{aligned}\lambda_j &= \text{Probability that the bottom} - \text{most ball is red and the respondent belongs to the } j\text{th group} + \text{Probability that the bottom} - \text{most ball is green and marked with } j \\ &= p\pi_j + p_j \text{ for } j = 1, \ldots, k\end{aligned}$$

Since n_j follows multinomial distribution, $\widehat{\lambda}_j = \dfrac{n_j}{n}$ becomes an unbiased estimator of λ_j and we have the following theorem.

Theorem 16.3.1

(i) An unbiased estimator of π_j is $\widehat{\pi}_j = \left(\widehat{\lambda}_j - p_j\right)/p$

(ii) Variance of $\widehat{\pi}_j$ is $V\left(\widehat{\pi}_j\right) = \lambda_j(1 - \lambda_j)/(np^2)$

(iii) An unbiased estimator of $V\left(\widehat{\pi}_j\right)$ is $\widehat{V}\left(\widehat{\pi}_j\right) = \widehat{\lambda}_j\left(1 - \widehat{\lambda}_j\right)/\{(n-1)p^2\}$

Proof

Since n_j follows multinomial distribution with parameter λ_j, we have

(i) $E\left(\widehat{\pi}_j\right) = E\left(\dfrac{\hat{\lambda}_j - p_j}{p}\right) = \dfrac{\left(\lambda_j - p_j\right)}{p} = \pi_j;$

(ii) $V\left(\widehat{\pi}_j\right) = \dfrac{V\left(\hat{\lambda}_j\right)}{p^2} = \dfrac{\lambda_j(1 - \lambda_j)}{np^2}$ and

(iii) $$\begin{aligned}E\left\{\widehat{V}\left(\widehat{\pi}_j\right)\right\} &= \frac{E\left(\hat{\lambda}_j\right) - \left\{V\left(\hat{\lambda}_j\right) + \lambda_j^2\right\}}{(n-1)p^2} \\ &= V\left(\widehat{\pi}_j\right)\end{aligned}$$

Remark 16.3.1
The estimator $\widehat{\pi}_j$ takes a negative value if $\widehat{\lambda}_j < p_j$.

16.4 RANDOMIZED RESPONSE TECHNIQUES FOR QUANTITATIVE CHARACTERISTICS

Let y be a sensitive and quantitative characteristic, such as expenditure on the prohibitive drug marijuana, and y_i be the value of the character y of the ith unit of the population and $\mathbf{y} = (y_1,\ldots, y_i,\ldots, y_N)$ be the unknown population vector.

16.4.1 Eriksson's Technique

Here we assume that y_i can take any value in the known interval (a,b). In Eriksson's (1973) RR technique, M values $Q_1(=a)$, $Q_2,\ldots, Q_M$ $(=b)$ are chosen in the interval (a,b). The vector $\mathbf{Q} = (Q_1,\ldots, Q_M)$ covers the range (a,b) and the value of M depends on the length of the interval. The respondent is supposed to report either the true value y_i with probability c or the Q_j value with probability $q_j\left(q_j > 0, \sum_j q_j = 1 - c\right)$. For example, in collecting information regarding expenditures on illegal drugs, a randomized device can be framed as follows: Take a jar containing capsules that appear identical. Each of the capsules contains exactly one ticket. Capsules of c proportions contain tickets written "Disclose your true expenditure" and the remaining capsules contain tickets marked \$0$(=Q_1)$, \$5$(=Q_2)$,\$10$(=Q_3)$,…, \$2000$(=Q_{401})$ with proportions q_1, $q_2,\ldots, q_{401}$, respectively. Here it is assumed that the minimum expenditure on drugs is zero and the maximum \$2000, and $M = 401$. The respondent is to shake the jar well and select a capsule at random. After opening the capsule, if a respondent finds the ticket "Disclose your true expenditure" then he/she has to disclose the true expenditure otherwise he/she will report the amount written in the ticket. The confidentiality of the respondent is preserved since the experiment is performed in the absence of the interviewer, and the interviewer will receive some amount as a response and the interviewer will not be able to identify whether this figure is really respondent's real expenditure or the value written on the ticket.

Let z_i be the response obtained from the ith respondent and let E_R, V_R, and C_R, respectively, denote the expectation, variance, and covariance operators with respect to the RR technique. Now, noting RRs are independent, we obtain

$$E_R(z_i) = c\ \ y_i + \sum_j q_j Q_j,\ \ E_R\left(z_i^2\right) = c\ \ y_i^2 + \sum q_j Q_j^2,\ \text{ and}$$

$$C_R(z_i, z_j) = 0\ \text{ for }\ i \neq j \tag{16.4.1}$$

Denoting

$$r_i = \left(z_i - \sum_j q_j Q_j \right) \Big/ c \tag{16.4.2}$$

as the revised RR for the ith unit, we get

$$E_R(r_i) = y_i, \quad V_R(r_i) = V_R(z_i)/c^2 = \alpha y_i^2 + \beta y_i + \gamma = \sigma_i^2, \quad \text{and} \quad C_R(r_i, r_j) = 0 \quad \text{for} \quad i \neq j \tag{16.4.3}$$

where $\alpha = \dfrac{1-c}{c}$, $\beta = -2\sum_j q_j Q_j / c$, and

$$\gamma = \left[\sum_j q_j Q_j^2 - \left(\sum_j q_j Q_j \right)^2 \right] \Big/ c^2 .$$

An unbiased estimator of σ_i^2 is

$$\widehat{\sigma}_i^2 = \frac{\alpha r_i^2 + \beta r_i + \gamma}{1 + \alpha}$$

16.4.2 Arnab's Model

Arnab (1990) proposed a more general RR model as follows:

$$\text{Model } R: E_R(r_i) = y_i, V_R(r_i) = \phi_i = \phi(y_i), \quad \text{and} \quad C_R(r_i, r_j) = 0 \quad \text{for} \quad i \neq j \tag{16.4.4}$$

where $\phi_i = \phi_i(y_i)$ is a function of y_i only.

Here we assume that a nonnegative unbiased estimator ϕ_i is available and it will be denoted by $\widehat{\phi}_i$. Most of the RR models for qualitative or quantitative characteristics satisfy the Model R and hence may be considered as a special case of the Model R given in Eq. (16.4.4).

Let us suppose that y is a qualitative variable and $y_i = 1$ if ith unit belongs to the sensitive group A and that $y_i = 0$ otherwise. In this case $\pi_A = \sum_{i \in U} y_i / N =$ proportion of individuals in the population who belong to the sensitive group A.

For a qualitative characteristic y, an unbiased estimator of $V_R(r_i) = \phi_i$ is

$$\widehat{\phi}_i = r_i(r_i - 1) \tag{16.4.5}$$

since $E_R\left(\widehat{\phi}_i\right) = E\left(r_i^2\right) - E_R(r_i) = \phi_i + \gamma_i^2 - \gamma_i = \phi_i$ as $\gamma_i^2 = \gamma_i$.

Warner's Model Let z_i be the RR obtained from the ith respondent. Here $z_i = 1$ if the ith respondent answers "Yes" and $z_i = 0$ if answers "No."

In this case we have $E_R(z_i) = \gamma_i P_1 + (1 - \gamma_i)(1 - P_1) = E_R\left(z_i^2\right)$ and $V_R(z_i) = P_1(1 - P_1)$.

Now writing $r_i = \dfrac{z_i - (1 - P_1)}{(2P_1 - 1)}$ and noting P_1 is a known constant, we find

$$E_R(r_i) = \gamma_i, \quad V_R(r_i) = \phi_i = \frac{P_1(1 - P_1)}{(2P_1 - 1)^2} = \widehat{\phi}_i = \phi, \text{ a known constant.} \tag{16.4.6}$$

Greenberg et al. Model With Known π_x In this case $\gamma_i = 1$ if the ith individual possesses the sensitive characteristic y and $\gamma_i = 0$ otherwise. Similarly, $x_i = 1$ or 0 according as to whether or not the ith unit possesses the nonsensitive characteristic x. Here also, RR $z_i = 1$ if answer is "Yes" and $z_i = 0$ if answer is "No." Thus we have $E_R(z_i) = \gamma_i P_2 + x_i(1 - P_2) = E_R\left(z_i^2\right)$ and $V_R(z_i) = P_2(1 - P_2)(\gamma_i - x_i)^2$. The revised RR $r_i = \dfrac{z_i - x_i(1 - P_2)}{P_2}$ yields

$$E_R(r_i) = \gamma_i \text{ and } V_R(r_i) = \frac{(1 - P_2)(\gamma_i - x_i)^2}{P_2} = \phi_i(\gamma_i) \text{ and } \widehat{\phi}_i = r_i(r_i - 1) \tag{16.4.7}$$

Here we note that $\phi_i(\gamma_i)$ is unknown since it involves γ_i.

Kuk's Model Let z_i be the number of black balls drawn by the ith respondent using Kuk's model R_k described in Section 16.2.3. In this case $E_R(z_i) = c[P_{31}\gamma_i + P_{32}(1 - \gamma_i)] = c[(P_{31} - P_{32})\gamma_i + P_{32}]$ and $V_R(z_i) = c[P_{31}(1 - P_{31})\gamma_i + P_{32}(1 - P_{32})(1 - \gamma_i)]$. The revised RR $r_i = \dfrac{z_i - cP_{32}}{c(P_{31} - P_{32})}$ yields

$$\begin{aligned} E_R(r_i) = \gamma_i, V_R(r_i) &= \frac{P_{31}(1 - P_{31})\gamma_i + P_{32}(1 - P_{32})(1 - \gamma_i)}{c(P_{31} - P_{32})^2} \\ &= \phi_i(\gamma_i) \text{ and } \widehat{\phi}_i = r_i(r_i - 1) \end{aligned} \tag{16.4.8}$$

Mangat and Singh (1990) Model Let $z_i = 1(0)$, if the RR response is "Yes" ("No") and $y_i = 1(0)$ if $i \in A\left(\overline{A}\right)$. Then $E_R(z_i) = y_i T + (1 - T) \times \{y_i P_4 + (1 - y_i)(1 - P_4)\} = E_R\left(z_i^2\right)$ and $r_i = \dfrac{z_i - (1 - P_4)(1 - T)}{2T(1 - P_4) + (2P_4 - 1)}$ yield

$$E_R(r_i) = y_i \text{ and } V_R(r_i) = \frac{(1 - P_4)(1 - T)\{1 - (1 - P_4)(1 - T)\}}{\{2T(1 - P_4) + (2P_4 - 1)\}^2} = \phi_i$$
$$= \widehat{\phi}_i = \phi, \text{ a known constant} \tag{16.4.9}$$

16.4.3 Christofides's Model

In Christofides (2003) randomized device, a respondent is asked to pick a card at random from a box containing M different kinds of cards marked 1, 2,..., M with proportions $p_1, p_2, \ldots, p_M$ $\left(\sum_{i=1}^{M} p_i = 1\right)$. If the respondent picks a card with number "x" ($x = 1,\ldots, M$) then he/she supplies the RR $z = M + 1 - x$ provided he/she belongs to the sensitive group "A," otherwise if he/she belongs to the nonsensitive group $\overline{A}$ then respondent reports $z = x$ as his/her RR. Hence the RR obtained from the ith respondent can be expressed as

$$z_i = (M + 1 - x)y_i + x(1 - y_i) = (M + 1)y_i + x(1 - 2y_i)$$

Now writing $E_R(x) = \sum_{k=1}^{M} kp_k = \mu$ and $V_R(x) = \sum_{k=1}^{M} k^2 p_k - \mu^2 = w^2$, we find

$$E_R(z_i) = (M + 1 - 2\mu)y_i + \mu \text{ and } V_R(z_i) = (1 - 2y_i)^2 w^2$$
$$= w^2 \text{ (since } y_i = 0 \text{ or } 1)$$

Let $r_i = \dfrac{z_i - \mu}{M + 1 - 2\mu}$, then we have

$$E_R(r_i) = y_i, \quad V_R(r_i) = \left(\frac{w}{M + 1 - 2\mu}\right)^2 = \phi_i = \widehat{\phi}_i \text{ (known constant)} \tag{16.4.10}$$

16.4.4 Eichhorn and Hayre's Model

In Eichhorn and Hayre's (1983) RR technique, the respondents selected in the sample are advised to draw a random sample from some preassigned distribution such as normal, uniform, chi-square, and so on. The mean θ and

variance γ^2 of the distribution are assumed to be known. If the ith respondent is included in the sample and selects a random sample Q_i then he/she is asked to report an RR $z_i = \gamma_i Q_i/\theta$ where γ_i is the true value of the sensitive characteristic γ. In this situation $r_i = z_i = \gamma_i Q_i/\theta$ and follows the model

$$E_R(r_i) = \gamma_i,\ V_R(r_i) = \gamma_i^2\gamma^2/\theta^2 = \phi_i \text{ and } \widehat{\phi}_i = \gamma^2 r_i^2/\left(\gamma^2+\theta^2\right) \quad (16.4.11)$$

16.4.5 Franklin's Randomized Response Technique

In Franklin's (1989) RR technique, a sample s of n units (respondents) is selected by the SRSWR method. Each of the selected respondents in s has to perform $k(\geq 1)$-independent RR trials. The ith respondent at the trial $j(=1,\ldots,\ k)$ has to draw a random sample from the density g_{ij} if he/she belongs to the sensitive group A, or if he/she belongs to $\overline{A}$, selects a random sample from the density h_{ij}. Confidentiality of the respondent is maintained because the interviewer will know only the random sample drawn but not the population from which it was selected. The random sample is selected by using some suitable randomized device such as a spinner or a random number table. In developing this theory, Franklin assumed $g_{ij} = g_j$ and $h_{ij} = h_j$ for every $i \in U$. He further assumed that the densities g_j and h_j are normal with known means μ_{1j} and μ_{2j} and known variances σ_{1j}^2 and σ_{2j}^2, respectively. In fact, Franklin made the randomized device much more interesting by using a portable electronic machine. If a respondent pushes a button on this machine, he/she gets two six-digit numbers labeled "Yes" and "No," respectively. If the respondent belongs to the group A ($\overline{A}$), he/she will supply a six-digit number labeled "Yes" ("No"). The first, second, and third two digits will correspond to $3(=k)$ independent samples from g_j and the remaining fourth, fifth, and sixth two digits represent random samples from h_j. Let z_{ij} be the RR obtained from the ith respondent at the jth trial and $\gamma_i = 1(0)$ if $i \in A\left(\in \overline{A}\right)$. Then,

$$\begin{aligned} E_R(z_{ij}) &= \gamma_i E_R(z_{ij}|i\in A) + (1-\gamma_i)E_R\left(z_{ij}\middle|i\in\overline{A}\right) \\ &= \gamma_i\mu_{1j} + (1-\gamma_i)\mu_{2j} \\ E_R\left(z_{ij}^2\right) &= \gamma_i E_R\left(z_{ij}^2|i\in A\right) + (1-\gamma_i)E_R\left(z_{ij}^2\middle|i\in\overline{A}\right) \\ &= \gamma_i\left(\sigma_{1j}^2+\mu_{1j}^2\right) + (1-\gamma_i)\left(\sigma_{2j}^2+\mu_{2j}^2\right) \end{aligned}$$

Writing $r_{ij} = (z_{ij} - \mu_{2j})/(\mu_{1j} - \mu_{2j})$, we obtain the following RR model

$$E_R(r_{ij}) = \gamma_i,\ V_R(r_{ij}) = \phi_{ij} = \frac{\gamma_i\sigma_{1j}^2 + (1-\gamma_i)\sigma_{2j}^2}{\left(\mu_{1j}-\mu_{2j}\right)^2},\ \ C_R(r_{ij}, r_{i'j'}) = 0$$
$$\text{for } (i,j) \neq (i',j') \text{ and } \widehat{\phi}_{ij} = r_{ij}(r_{ij}-1) \quad (16.4.12)$$

16.4.6 Chaudhuri's Randomized Response

In Chaudhuri's (1987) RR technique, the respondent labeled i is asked to choose independently a pair of numbers $a_i(j)$, $b_i(k)$ at random out of two sets of numbers $a_i(m)$, $b_i(r)$, $m = 1,\ldots, A_i$; $r = 1,\ldots, B_i$, given to him and to report $z_i = a_i(j)y_i + b_i(k)$ as his RR. Denoting $r_i = \left(z_i - \overline{b}_i\right)\Big/\overline{a}_i$ with $\overline{a}_i = \sum_j a_i(j)/A_i$, $\overline{b}_i = \sum_k b_i(k)/B_i$ we get,

$$E_R(r_i) = y_i, \quad V_R(r_i) = \phi_i = \alpha_i y_i^2 + \beta_i, \quad C_R(r_i, r_j) = 0 \text{ for } i \neq j \text{ and } \widehat{\phi}_i = \left(\alpha_i r_i^2 + \beta_i\right)\Big/(1 + \alpha_i) \tag{16.4.13}$$

where $\alpha_i = S_i^2(a)\Big/\overline{a}_i^2, \beta_i = S_i^2(b)\Big/\overline{a}_i^2$, $S_i^2(a) = \sum_j \{a_i(j) - \overline{a}_i\}^2/A_i$ and $S_i^2(b) = \sum_j \left\{b_i(j) - \overline{b}_i\right\}^2 \Big/ B_i$.

16.5 GENERAL METHOD OF ESTIMATION

Most surveys in practice are based on complex sampling designs and information regarding more than one character is collected at a time. Some of them are of a confidential nature whereas others are not. Furthermore, the sensitive characters need not be only of a qualitative nature, it may also be quantitative. To cope with this situation, a general method of estimating the population total has been proposed in this section following the methods proposed by Arnab (1994). Expressions for the variances of the proposed estimators and unbiased estimator of the variances have been derived. Here we suppose that a sample s of size n is selected by some arbitrary sampling design p. The inclusion probabilities for the ith unit π_i, and ith and jth units π_{ij} are assumed to be positive. Let y_i be the value of the sensitive character y for the ith unit. The value y_i cannot be directly obtained from the respondent. Hence an RR r_i is obtained by applying some suitable randomized device. The RR r_is are assumed to follow the model given in Eq. (16.4.4) viz.

$$E_R(r_i) = y_i, V_R(r_i) = \phi_i \text{ and } C_R(r_i, r_j) = 0 \text{ for } i \neq j$$

It is further assumed that a nonnegative unbiased estimator of ϕ_i is available and will be denoted by $\widehat{\phi}_i$.

16.5.1 Estimation of Total and Variance

Consider a direct method of survey, where information of y_i's is directly obtained from the respondents. In this case we propose a linear homogeneous unbiased estimator for the total Y as

$$t(s, y) = \sum_{i \in s} b_{si} y_i \tag{16.5.1}$$

where b_{si}'s are constants free from y_i's and r_i's and satisfy the unbiasedness condition

$$\sum_{s \supset i} b_{si} p(s) = 1 \quad \text{for} \quad \forall \quad i = 1, \ldots, N \tag{16.5.2}$$

Since y_i's are not directly obtained from the respondent, we replace y_i by its estimate r_i in $t(s,y)$ and obtain the following estimator for RR survey

$$t(s, r) = \sum_{i \in s} b_{si} r_i \tag{16.5.3}$$

Let $E_p(E_R)$, $V_p(V_R)$, and $C_p(C_R)$, respectively, denote the expectation, variance, and covariance operators with respect to the sampling design p (RR model). Here we note that the commutativity of the operators E_p and E_R holds in the sense $E_p E_R = E_R E_p$ (see Arnab, 1990) for any non-informative sampling design where $p(s)$ does not involve r_i's, $i \in s$.

Theorem 16.5.1

(i) $E[t(s, r)] = Y$

(ii) $V[t(s, r)] = V_p[t(s, y)] + \sum_i \alpha_{ii} \phi_i$

where

$$V_p[t(s, y)] = \sum_{i \in U} (\alpha_{ii} - 1) y_i^2 + \sum_{i \neq} \sum_{j \in U} (\alpha_{ij} - 1) y_i y_j,$$

$$\alpha_{ii} = \sum_{s \supset i} b_{si}^2 p(s) \quad \text{and} \quad \alpha_{ij} = \sum_{s \supset i,j} b_{si} b_{sj} p(s)$$

Proof

(i) $E[t(s, r)] = E_p[E_R\{t(s, r)\}]$

$$= E_p\left(\sum_{i \in s} b_{si} y_i \right)$$

$$= Y$$

$$\text{(ii)}\quad V[t(s,r)] = V_p[E_R\{t(s,r)\}] + E_p\left[V_R\left(\sum_{i\in s} b_{si} r_i\right)\right]$$

$$= V_p[t(s,\gamma)] + E_p\left(\sum_{i\in s} b_{si}^2\, \phi_i\right)$$

Now noting

$$V_p[t(s,\gamma)] = E_p\left(\sum_{i\in s} b_{si}^2\, \gamma_i + \sum_{i\neq} \sum_{j\in s} b_{si} b_{sj} \gamma_i \gamma_j - Y^2\right)$$

$$= \sum_{i\in U} \gamma_i^2 \left(\sum_{s\supset i} b_{si}^2 p(s) - 1\right) + \sum_{i\neq} \sum_{j\in U} \gamma_i \gamma_j \left(\sum_{s\supset i,j} b_{si} b_{sj} p(s) - 1\right)$$

$$= \sum_{i\in U} (\alpha_{ii} - 1)\gamma_i^2 + \sum_{i\neq} \sum_{j\in U} (\alpha_{ij} - 1)\gamma_i \gamma_j$$

and

$$E_p\left(\sum_{i\in s} b_{si}^2 \phi_i\right) = \sum_{i\in U} \phi_i \sum_{s\supset i} b_{si}^2 p(s)$$

$$= \sum_{i\in U} \alpha_{ii} \phi_i$$

we can verify the theorem.

Arnab (1994) proposed the method of unbiased estimation of $V[t(s,r)]$ as follows.

Theorem 16.5.2

Let $\widehat{V}[t(s,\gamma)]$ be a homogeneous quadratic unbiased estimator of $V[t(s,\gamma)]$. Then an unbiased estimator of $V[t(s,r)]$ is

$$\widehat{V}^*[t(s,r)] = \widehat{V}[t(s,r)] + \sum_{i\in s} b_{si}\ \widehat{\phi}_i$$

where $\widehat{V}[t(s,r)]$ is obtained by writing r_i in place of γ_i in $\widehat{V}[t(s,\gamma)]$.

Proof

A homogenous quadratic unbiased estimator $\widehat{V}[t(s,\gamma)]$ is of the form

$$\widehat{V}[t(s,\gamma)] = \sum_{i\in s} c_{ii}(s)\gamma_i^2 + \sum_{i\neq} \sum_{j\in s} c_{ij}(s)\gamma_i \gamma_j \qquad (16.5.4)$$

where the constants $c_{ii}(s)$ and $c_{ij}(s)$ are free from y_i and r_i s and chosen to make $\widehat{V}[t(s,y)]$ unbiased for $V[t(s,y)]$.

Now $E_p\left[\widehat{V}\{t(s,y)\}\right] = V[t(s,y)]$ for all possible values of $\mathbf{y} = (y_1,\ldots, y_N)$ implies

$$\sum_{s\supset i} c_{ii}(s)p(s) = \alpha_{ii} - 1 \text{ and } \sum_{s\supset i,j} c_{ij}(s)p(s) = \alpha_{ij} - 1 \tag{16.5.5}$$

Furthermore, writing r_i in place of y_i in $\widehat{V}[t(s,y)]$, we get

$$\widehat{V}[t(s,r)] = \sum_{i\in s} c_{ii}(s)\, r_i^2 + \sum_{i\neq} \sum_{j\in s} c_{ij}(s) r_i r_j$$

and

$$\begin{aligned} E\left[\widehat{V}\{t(s,r)\}\right] &= E_p\left[\sum_{i\in s} c_{ii}(s)E_R\left(r_i^2\right) + \sum_{i\neq} \sum_{j\in s} c_{ij}(s)E_R(r_i r_j)\right] \\ &= E_p\left[\widehat{V}\{t(s,y)\}\right] + E_p\left(\sum_{i\in s} c_{ii}(s)\,\phi_i\right) \\ &= E_p\left(\widehat{V}\{t(s,y)\}\right) + \sum_{i\in U}(\alpha_{ii}-1)\phi_i \\ &= V[t(s,r)] - \sum_{i\in U}\phi_i \end{aligned}$$

Now noting $E\left(\sum_{i\in s} b_{si}\ \widehat{\phi}_i\right) = E_p\left[\sum_{i\in s} b_{si}E_R\left(\widehat{\phi}_i\right)\right] = E_p\left(\sum_{i\in s} b_{si}\,\phi_i\right) = \sum_{i\in U}\phi_i$, we find

$$E\left[\widehat{V}^*\{t(s,r)\}\right] = V[t(s,r)]$$

16.5.1.1 Horvitz–Thomson Estimator

For $b_{si} = 1/\pi_i$, we have $\alpha_{ii} = 1/\pi_i$ and $\alpha_{ij} = \pi_{ij}/(\pi_i\pi_j)$. In this case the estimator (16.5.3) reduces to the Horvitz–Thomson estimator $\widehat{Y}_{ht}(rr) = \sum_{i\in s} r_i/\pi_i$. Hence substituting $\alpha_{ii} = 1/\pi_i$ and $\alpha_{ij} = \pi_{ij}/(\pi_i\pi_j)$ in Theorems 16.5.1 and 16.5.2, we get the following.

Theorem 16.5.3

For a fixed effective size sampling design

(i) $\widehat{Y}_{ht}(rr) = \sum_{i\in s} r_i/\pi_i$ is an unbiased estimator of Y.

(ii) $V\left[\widehat{Y}_{ht}(rr)\right] = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_{i\in U}\frac{\phi_i}{\pi_i}$,

and

(iii) $\widehat{V}\left[\widehat{Y}_{ht}(rr)\right] = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{\hat{\phi}_i}{\pi_i}$

is an unbiased estimator of $V\left[\widehat{Y}_{ht}(rr)\right]$.

16.5.1.2 Simple Random Sampling Without Replacement

For an simple random sampling without replacement (SRSWOR) $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/\{N(N-1)\}$. Substituting these values of π_i and π_{ij} in Theorem 16.5.3 we get the following theorem.

Theorem 16.5.4

For an SRSWOR sampling design

(i) $\overline{r}(s) = \sum_{i\in s} r_i/n$, the sample mean is an unbiased estimator of the population mean $\overline{Y}$,

(ii) $V[\overline{r}(s)] = \left[(1-f)S_y^2 + \overline{\phi}\right]\Big/n$,

and

(iii) $\widehat{V}[\overline{r}(s)] = \left[(1-f)s_r^2 + n\overline{\widehat{\phi}}_s/N\right]/n$

where $S_y^2 = \sum_{i\in U}\left(y_i - \overline{Y}\right)^2/(N-1)$, $s_r^2 = \sum_{i\in s}(r_i - \overline{r}_s)^2/(n-1)$

$f = n/N$, $\overline{\phi} = \sum_{i\in U}\phi_i/N$ and $\overline{\widehat{\phi}}_s = \sum_{i\in s}\widehat{\phi}_i/n$.

Corollary 16.5.1

Let y be a qualitative variable so that $y_i = 1$ if the ith unit possesses an attribute A, and $y_i = 0$ otherwise. Furthermore, if Warner's RR model R_w is used and $z_i = 1(0)$ when "Yes" ("No") response is obtained from the ith respondent we get

$$r_i = \frac{z_i - (1-P_1)}{(2P_1-1)},\quad V_R(r_i) = \phi_i = \frac{P_1(1-P_1)}{(2P_1-1)^2} = \phi = \widehat{\phi}_i \text{ and}$$

$$S_y^2 = N\pi_A(1-\pi_A)/(N-1)$$

where $\pi_A = \sum_{i\in U} y_i/N$ = population proportion.

An unbiased estimator for π_A, based on SRSWOR under R_w, is given by

$$\bar{r}(s) = \sum_{i \in s} r_i / n = \frac{\lambda_{WA} - (1 - P_1)}{(2P_1 - 1)} = \widehat{\pi}_A \tag{16.5.6}$$

where λ_{WA} is the proportion of "yes" answers in n RRs.

The expressions of variance of $\widehat{\pi}_A$ and its unbiased estimator are respectively given by

$$V(\widehat{\pi}_A) = \frac{(1-f)N}{n(N-1)} \pi_A(1 - \pi_A) + \frac{P_1(1 - P_1)}{n(2P_1 - 1)^2} \tag{16.5.7}$$

and

$$\widehat{V}(\widehat{\pi}_A) = \frac{(1-f)}{n-1} \frac{\lambda_{wA}(1 - \lambda_{wA})}{(2P_1 - 1)^2} + \frac{P_1(1 - P_1)}{N(2P_1 - 1)^2} \tag{16.5.8}$$

Furthermore, writing $\widehat{\pi}_A = \dfrac{\lambda_{wA} - (1 - P_1)}{2P_1 - 1}$ in Eq. (16.5.8), we get an alternative expression of $\widehat{V}(\widehat{\pi}_A)$ as

$$\widehat{V}(\widehat{\pi}_A) = \frac{1-f}{n-1} \widehat{\pi}_A(1 - \widehat{\pi}_A) + \frac{N-1}{(n-1)N} \frac{P_1(1 - P_1)}{(2P_1 - 1)^2} \tag{16.5.9}$$

16.5.1.3 Rao–Hartley–Cochran Sampling

Let a sample s of size n be selected by Rao–Hartley–Cochran (RHC, 1962) method of sampling assuming N/n is an integer. Unbiased estimators for the population total under direct and RR survey are, respectively, given by

$$t(s, y) = \widehat{Y}_{rhc} = \sum_{i \in s} \frac{y_i}{p_i} P_i \quad \text{and} \quad t(s, r) = \widehat{Y}_{rhc}(rr) = \sum_{i \in s} \frac{r_i}{p_i} P_i \tag{16.5.10}$$

where p_i is the normed size measure for the ith unit and P_i is the sum of the p_j values for the group containing the ith unit (see Section 5.6 for details).

Theorem 16.5.5

(i) $\widehat{Y}_{rhc}(rr) = \sum_{i \in s} \dfrac{r_i}{p_i} P_i$ is unbiased estimator of Y,

(ii) $V\left[\widehat{Y}_{rhc}(rr)\right] = \dfrac{N-n}{n(N-1)} \sum_{i \in U} p_i \left(\dfrac{y_i}{p_i} - Y\right)^2 + \sum_{i \in U} a_i \phi_i$

and

$$\text{(iii)}\ \ \widehat{V}\left[\widehat{Y}_{rhc}(rr)\right] = \frac{N-n}{N(n-1)}\sum_{i\in s} P_i\left(\frac{r_i}{p_i} - \widehat{Y}_{rhc}(rr)\right)^2 + \sum_{i\in s}\frac{\hat{\phi}_i}{p_i}P_i$$

$$\text{where } a_i = \frac{1}{n(N-1)}\left[N(n-1) + \frac{N-n}{p_i}\right]$$

Proof

$$\text{(i)}\ \ E\left[\widehat{Y}_{rhc}(rr)\right] = E_p\sum_{i\in s}\frac{E_R(r_i)}{p_i}P_i$$

$$= E_p\left(\sum_{i\in s}\frac{\gamma_i}{p_i}P_i\right)$$

$$= Y \ \text{(see Theorem 5.6.1.)}$$

$$\text{(ii)}\ \ V\left[\widehat{Y}_{rhc}(rr)\right] = V_p\left(\widehat{Y}_{rhc}\right) + E_p\left(\sum_{i\in s}\frac{\phi_i}{p_i^2}P_i^2\right)$$

Now

$$V_p\left(\widehat{Y}_{rhc}\right) = \frac{N-n}{n(N-1)}\sum_{i\in U} p_i\left(\frac{\gamma_i}{p_i} - Y\right)^2 \ \text{(Theorem 5.6.1)} \qquad (16.5.11)$$

and

$$E_p\left(\sum_{i\in s}\frac{\phi_i}{p_i^2}P_i^2\right) = nE_p\left(\frac{\phi_i}{p_i^2}P_i^2\right)$$

$$= nE_{G_i}\left(\sum_{j\in G_i}\frac{\phi_i}{p_j}\right)\left(\sum_{j\in G_i} p_j\right)$$

$$\text{(where } G_i \text{ is the ith group } (i = 1, \ldots, n)\text{, see Section 5.6)}$$

$$= nE_{G_i}\left(\sum_{j\in G_i}\phi_j + \sum_{j\neq}\sum_{k\in G_i}\phi_j p_k/p_j\right)$$

$$= n\left(\frac{N/n}{N}\sum_{j\in U}\phi_j + \frac{(N/n)(N/n-1)}{N(N-1)}\sum_{j\neq}\sum_{k\in U}\phi_j p_k/p_j\right)$$

$$= \sum_{j\in U} \phi_j \left\{ 1 + \frac{(N/n-1)}{(N-1)} (1-p_j)/p_j \right\}$$

$$= \sum_{j\in U} \phi_j \left[\left(1 - \frac{(N-n)}{n(N-1)} \right) + \frac{(N-n)}{n(N-1)p_j} \right]$$

$$= \sum_{j\in U} \phi_j a_j \tag{16.5.12}$$

The second part of the theorem follows from Eqs. (16.5.11) and (16.5.12).

(iii) Noting $\widehat{V}\left(\widehat{Y}_{rhc}\right) = \frac{N-n}{N(n-1)} \sum_{i\in s} P_i \left(\frac{y_i}{p_i} - \widehat{Y}_{rhc} \right)^2$ and $\sum_{i\in s} \frac{\hat{\phi}_i}{p_i} P_i$ are unbiased estimators of $V\left(\widehat{Y}_{rhc}\right)$ and $\sum_{i\in U} \phi_i$ respectively, and using Theorem 16.5.2, we find that $\widehat{V}\left[\widehat{Y}_{rhc}(rr)\right]$ is an unbiased estimator of $V\left[\widehat{Y}_{rhc}(rr)\right]$.

16.5.1.4 Probability Proportional to Aggregate Size Sampling

In Lahiri-Midzuno-Sen (1951, 1952, 1953) probability proportional to aggregate size (PPAS) sampling scheme (see Section 5.5), the probability of selecting a sample$s$ of size n is $p(s) = x_s/(M_1 X)$ where $x_s = \sum_{i\in s} x_i, X = \sum_{i\in U} x_i$, $M_1 = \binom{N-1}{n-1}$ and $x_i(>0)$ is the measure of size for the ith unit. For this sampling scheme

$$t(s,y) = \widehat{Y}_{lms} = \frac{y_s}{x_s} X \text{ and } t(s,r) = \widehat{Y}_{lms}(rr) = \frac{r_s}{x_s} X \tag{16.5.13}$$

where $y_s = \sum_{i\in s} y_i$ and $r_s = \sum_{i\in s} r_i$.

Now

$$V\left[\widehat{Y}_{lms}(rr)\right] = V_p\left(\frac{y_s}{x_s} X\right) + E_p\left\{ \sum_{i\in s} \phi_i \left(\frac{X}{x_s}\right)^2 \right\}$$

$$= -\frac{1}{2} \sum_{i\neq} \sum_{j\in U} \beta_{ij} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j} \right)^2 + \sum_{i\in U} \beta_i \phi_i \tag{16.5.14}$$

where $\beta_{ij} = X\sum_{s\supset i,j} 1/(M_1 x_s) - 1$ and $\beta_i = X\sum_{s\supset i} 1/(M_1 x_s)$(for detail, see Section 14.3.1.4).

Since $\widehat{V}\left(\widehat{Y}_{lms}\right) = B_s\frac{1}{2}\sum_{i\neq}\sum_{j\in s} x_i x_j\left(\frac{y_i}{x_i}-\frac{y_j}{x_j}\right)^2$ and $\frac{\hat{\phi}_s}{x_s}X$ are unbiased estimators of $V\left(\widehat{Y}_{lms}\right)$ and $\sum_{i\in U}\phi_i$, respectively, we obtain an unbiased estimator for $V\left[\widehat{Y}_{lms}(rr)\right]$ as

$$\widehat{V}\left[\widehat{Y}_{lms}(rr)\right] = B_s\frac{1}{2}\sum_{i\neq}\sum_{j\in s} x_i x_j\left(\frac{r_i}{x_i}-\frac{r_j}{x_j}\right)^2 + X\widehat{\phi}_s/x_s \qquad (16.5.15)$$

where $\widehat{\phi}_s = \sum_{i\in s}\widehat{\phi}_i$ and $B_s = \left(\frac{N-1}{n-1}-\frac{X}{x_s}\right)\frac{X}{x_s}$

16.5.1.5 Probability Proportional to Size With Replacement Sampling

Let a sample s of size n be selected by probability proportional to size with replacement (PPSWR) sampling scheme with the normed size measure p_i attached to the ith unit. Here we will consider the following cases:

Case I: If the ith respondent is selected in the sample $n_i(s)$ times, he/she has to perform randomized devices $n_i(s)$ times independently. Let $r(k)$, $y(k)$ and $p(k)$ be the RR, value of the variable y and the selection probability of the unit selected at the kth ($k = 1,\ldots, n$) draw. Then, $E_R\{r(k)\} = y(k) = y_j$, $V_R\{r(k)\} = \phi(k) = \phi_j$, and $p(k) = p_j$ if the rth draw produces the jth unit with probability $p_j (j = 1,\ldots, N)$. In this case we get

$$E\left(\frac{r(k)}{p(k)}\right) = E_p\left[E_R\left(\frac{r(k)}{p(k)}\right)\right] = E_p\left(\frac{y(k)}{p(k)}\right) = \sum_{j=1}^{N}\frac{y_j}{p_j}p_j = Y \qquad (16.5.16)$$

$$\begin{aligned} V\left(\frac{r(k)}{p(k)}\right) &= E_p\left[V_R\left(\frac{r(k)}{p(k)}\right)\right] + V_p\left[E_R\left(\frac{r(k)}{p(k)}\right)\right] \\ &= E_p\left[\frac{\phi(k)}{\{p(k)\}^2}\right] + V_p\left(\frac{y(k)}{p(k)}\right) \\ &= \sum_{i\in U}\frac{\phi_i}{p_i} + \left(\sum_{i\in U}\frac{y_i^2}{p_i} - Y^2\right) \end{aligned} \qquad (16.5.17)$$

Since $\frac{r(k)}{p(k)}$'s are independently and identically distributed random variables, we find from Arnab (1990) as follows:

Theorem 16.5.6

(i) $\widehat{Y}_{hh}\left(rr\right) = \frac{1}{n}\sum_{r=1}^{n}\frac{r(k)}{p(k)}$ is an unbiased estimator of Y.

(ii) $V\left[\widehat{Y}_{hh}(rr)\right] = \left[\left(\sum_{i\in U}\frac{\gamma_i^2}{p_i} - Y^2\right) + \sum_{i\in U}\frac{\phi_i}{p_i}\right]\Big/ n$

and

(iii) $\widehat{V}\left[\widehat{Y}_{hh}(rr)\right] = \frac{1}{n(n-1)}\sum_{k=1}^{n}\left\{\frac{r(k)}{p(k)} - \widehat{Y}_{hh}(rr)\right\}^2$ is an unbiased estimator of $V\left[\widehat{Y}_{hh}(rr)\right]$

Case II: Here we suppose that each of the units in the sample produces only one RR even if it is selected more than once in the sample. More specifically, suppose the ith unit is selected $n_i(s)$ times and we receive only one revised RR r_i from it. Let $\widehat{Y}^{*}_{hh}(rr) = \frac{1}{n}\sum_{i\in U} n_i(s)\frac{r_i}{p_i}$, then we have the following from Arnab (1990).

Theorem 16.5.7

(i) $E\left[\widehat{Y}^{*}_{hh}(rr)\right] = Y$

(ii) $V\left[\widehat{Y}^{*}_{hh}(rr)\right] = \left[\left(\sum_{i=1}^{N}\frac{\gamma_i^2}{p_i} - Y^2\right) + \sum_{i=1}^{N}\{1+(n-1)p_i\}\frac{\phi_i}{p_i}\right]\Big/ n$

and

(iii) $\widehat{V}\left[\widehat{Y}^{*}_{hh}(rr)\right] = \frac{1}{n(n-1)}\sum_{i\in U} n_i(s)\left\{\frac{r_i}{p_i} - \widehat{Y}^{*}_{hh}(rr)\right\}^2 + \frac{1}{n}\sum_{i\in U} n_i(s)\frac{\hat{\phi}_i}{p_i}$

Proof

The theorem can be proved by noting $E(n_i(s)) = np_i$, $V(n_i(s)) = np_i(1-p_i)$ and $Cov(n_i(s), n_j(s)) = -np_ip_j$ for $i \neq j$.

Remark 16.5.1

The estimator $\widehat{Y}_{hh}(rr)$ is more efficient than $\widehat{Y}^{*}_{hh}(rr)$ since $V\left[\widehat{Y}_{hh}(rr)\right] \leq \widehat{V}\left[\widehat{Y}^{*}_{hh}(rr)\right]$. This is because the estimator $\widehat{Y}_{hh}(rr)$ is based on $n_i(s)(\geq 1)$ RRs from the ith ($i \in s$) unit while $\widehat{Y}^{*}_{hh}(rr)$ is based on a single RR from the ith unit even it is selected $n_i(s)$ times.

16.5.1.6 Simple Random Sampling With Replacement

Substituting $p_i = 1/N$ in Theorems 16.5.6 and 16.5.7, we derive the following results.

Theorem 16.5.8

For an SRSWR sampling

(i) $\bar{r} = \frac{1}{n}\sum_{k=1}^{n} r(k)$ is an unbiased estimator for the population mean $\overline{Y}$.

(ii) $V(\bar{r}) = \left(\sigma_y^2 + \overline{\phi}\right)\Big/ n$

and

(iii) $\widehat{V}(\bar{r}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(r(k) - \bar{r})^2$

where $\overline{\phi} = \sum_{i \in U}\phi_i/N$ and $\sigma_y^2 = \sum_{i \in U}\left(y_i - \overline{Y}\right)^2/N$

Theorem 16.5.9

For an SRSWR sampling where a single RR response is obtained from the ith respondent even he/she is selected more than once.

(i) $\bar{r}^* = \frac{1}{n}\sum_{i \in U} n_i(s) r_i$ is an unbiased estimator of $\overline{Y}$.

(ii) $V(\bar{r}^*) = \left[\sigma_y^2 + \left(1 + \frac{n-1}{N}\right)\overline{\phi}\right]\Big/ n$

and

(iii) $\widehat{V}[\bar{r}^*] = \left[\frac{1}{(n-1)}\sum_{i \in U} n_i(s)(r_i - \bar{r}^*)^2 + \frac{1}{N}\sum_{i \in U} n_i(s)\widehat{\phi}_i\right]\Big/ n$

The Table 16.5.1 below shows the estimators for the population total Y, its variances, and unbiased estimators of the variances under various sampling designs.

Table 16.5.1 Unbiased estimators, variances and unbiased estimators of variances for the population total for various sampling design

Sampling design	Unbiased estimator $\widehat{Y}$	Variance $V(\widehat{Y})$	Unbiased estimator of variance $\widehat{V}(\widehat{Y})$
Arbitrary	$t(s,r)=\sum_{i\in s} b_{si} r_i$	$\sum_{i\in U}(\alpha_{ii}-1)\gamma_i^2+\sum_{i\neq}\sum_{j\in U}(\alpha_{ij}-1)\gamma_i\gamma_j+\sum_{i\in U}\alpha_{ii}\phi_i$	$\sum_{i\in s} c_{ii}(s) r_i^2+\sum_{i\neq}\sum_{j\in s} c_{ij}(s) r_i r_j+\sum_{i\in s} b_{si}\widehat{\phi}_i$
Fixed effective sample size	$\widehat{Y}_{ht}(rr)=\sum_{i\in s}\frac{r_i}{\pi_i}$	$\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{\gamma_i}{\pi_i}-\frac{\gamma_j}{\pi_j}\right)^2+\sum_{i\in U}\frac{\phi_i}{\pi_i}$	$\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{r_i}{\pi_i}-\frac{r_j}{\pi_j}\right)^2+\sum_{i\in s}\frac{\widehat{\phi}_i}{\pi_i}$
Rao–Hartley–Cochran	$\widehat{Y}_{rhc}(rr)=\sum_{i\in s}\frac{r_i}{p_i}P_i$	$\frac{N-n}{n(N-1)}\sum_{i\in U}p_i\left(\frac{\gamma_i}{p_i}-Y\right)^2+\sum_{i\in U}a_i\phi_i$	$\frac{N-n}{N(n-1)}\sum_{i\in s}P_i\left\{\frac{r_i}{p_i}-\widehat{Y}_{rhc}(rr)\right\}^2+\sum_{i\in s}\frac{\widehat{\phi}_i}{p_i}P_i$
PPAS	$\widehat{Y}_{lms}(rr)=\frac{r_s}{x_s}X$	$-\frac{1}{2}\sum_{i\neq}\sum_{j\in U}\beta_{ij}x_i x_j\left(\frac{\gamma_i}{x_i}-\frac{\gamma_j}{x_j}\right)^2+\sum_{i\in U}\beta_i\phi_i$	$B_s\frac{1}{2}\sum_{i\neq}\sum_{j\in s}x_i x_j\left(\frac{r_i}{x_i}-\frac{r_j}{x_j}\right)^2+\frac{X}{x_s}\sum_{i\in s}\widehat{\phi}_i$

Continued

Table 16.5.1 Unbiased estimators, variances and unbiased estimators of variances for the population total for various sampling design—cont'd

Sampling design	Unbiased estimator $\widehat{Y}$	Variance $V(\widehat{Y})$	Unbiased estimator of variance $\widehat{V}(\widehat{Y})$
SRSWOR	$\widehat{Y}_{wor}(rr) = N\frac{1}{n}\sum_{i\in s} r_i$	$N^2\left\{(1-f)S_y^2 + \frac{1}{N}\sum_{i\in U}\phi_i\right\}\Big/ n$	$N^2\left\{(1-f)s_r^2 + \frac{1}{N}\sum_{i\in s}\widehat{\phi}_i\right\}\Big/ n$
PPSWR	$\widehat{Y}_{hh}(rr) = \frac{1}{n}\sum_{k=1}^{n}\frac{r(k)}{p(k)}$	$\frac{1}{n}\left[\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2 + \sum_{i\in U}\frac{\phi_i}{p_i}\right]$	$\frac{1}{n(n-1)}\sum_{k=1}^{n}\left\{\frac{r(k)}{p(k)} - \widehat{Y}_{hh}(rr)\right\}^2$
PPSWR	$\widehat{Y}^*_{hh}(rr) = \frac{1}{n}\sum_{i\in U} n_i(s)\frac{r_i}{p_i}$	$\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2\Big/ n + \sum_{i\in U}\{1+(n-1)p_i\}\frac{\phi_i}{p_i}\Big/ n$	$\frac{1}{n(n-1)}\sum_{i\in U} n_i(s)\left\{\frac{r_i}{p_i} - \widehat{Y}^*_{hh}(rr)\right\}^2 + \frac{1}{n}\sum_{i\in U} n_i(s)\frac{\widehat{\phi}_i}{p_i}$
SRSWR	$\widehat{Y}_{wr}(rr) = N\frac{1}{n}\sum_{k=1}^{n} r(k)$	$N^2\left[\sigma_y^2 + \frac{1}{N}\sum_{i\in U}\phi_i\right]\Big/ n$	$\frac{N^2}{n(n-1)}\sum_{r=1}^{n}\left\{r(k) - \frac{1}{n}\sum_{j=1}^{n} r(j)\right\}^2$
SRSWR	$\widehat{Y}^*_{wr}(rr) = \frac{N}{n}\sum_{i\in U} n_i(s) r_i$	$\frac{N^2}{n}\left[\sigma_y^2 + \left\{1 + \frac{(n-1)}{N}\right\}\frac{1}{N}\sum_{i\in U}\phi_i\right]$	$N^2\left[\frac{1}{n(n-1)}\sum_{i\in U} n_i(s)\left\{r_i - \frac{1}{n}\sum_{i\in U} n_i(s) r_i\right\}^2 + \frac{1}{Nn}\sum_{i\in U} n_i(s)\widehat{\phi}_i\right]$

16.6 OPTIONAL RANDOMIZED RESPONSE TECHNIQUES

In optional RR technique (ORT) it is assumed that the aspects of inquiry are felt to be sensitive by most of the respondents but some are more willing to answer directly. In ORT, respondents are given an option either to supply RR using a specified randomized device or to respond directly according to the extent to which the respondent feels the question is sensitive or not. Most of the methods developed for ORT are limited to SRSWR sampling only. A few of the ORT techniques are available for complex surveys. ORT can be classified into two categories: Full ORT (FORT) and Partial ORT (PORT). The ORT is more efficient than compulsory RRT (CRT) because the probability of obtaining true responses in ORT is much higher than that in the CRT (Vide Arnab, 2004a).

FORT: Here respondents are given an option either to supply RR using a specified randomized device or to respond directly according to whether the respondent feels that the question is sensitive or not. In this method it is assumed that respondents who feel that the character under investigation is confidential, belong to certain group G and produce RR with probability 1, whereas the respondents who feel the character is not confidential, belong to the complementary group $\overline{G}$ and supply direct response with probability 1. FORT was proposed by Chaudhuri and Mukherjee (1988), Arnab (2004a), Chaudhuri and Saha (2005), Huang (2008), among others.

PORT: Here it is assumed that the respondents may supply direct or RR with certain probability depending on their judgment of sensitivity (mood) at the particular time of answering the question. In other words, the respondent may sometimes supply RRs and at other times direct responses during the period of the survey. Most of the researchers developed various methods of PORT, e.g., Mangat and Singh (1994), Gupta (2001), Gupta et al. (2002), Pal (2008), among others.

16.6.1 Full Optional Randomized Response Technique

Arnab (2004a) proposed a theory of estimation of population characteristics for FORT, which is applicable for both qualitative and quantitative characteristics. Let $s_G = s \bigcap G$ be the set of respondents selected in the sample s that belong to group G and $s_{\overline{G}} = s \bigcap \overline{G} = s - s_G$. In this method, respondents belonging to group s_G provide RRs using some suitable randomized device whereas respondents belonging to group $s_{\overline{G}}$ provide direct responses. Here we assume that the respondents do not disclose their membership to group G or $\overline{G}$.

16.6.1.1 Estimation of Population Total

Let r_i be the revised RR obtained from the ith respondent if he/she is included in the sample s_G and let y_i be the direct response if the ith unit belongs to $S_{\overline{G}}$. Here we suppose that the r_i's follow Model (16.4.4). Let us define

$$\tilde{r}_i = \delta_i y_i + (1 - \delta_i) r_i \tag{16.6.1}$$

where $\delta_i = 1$ if $i \in \overline{G}$ and $\delta_i = 0$ if $i \in G$.

Clearly, $\tilde{r}_i$ follows the following model

$$\begin{aligned} \text{Model } R_o: E_R(\tilde{r}_i) &= y_i,\ V_R(\tilde{r}_i) = (1 - \delta_i) V_R(r_i) \\ &= (1 - \delta_i)\phi_i \text{ and } C_R(\tilde{r}_i, \tilde{r}_j) = 0 \text{ for } i \neq j \end{aligned} \tag{16.6.2}$$

Replacing r_i by $\tilde{r}_i$ in Eq. (16.5.3), we find a linear unbiased estimator for the population total Y under ORR technique as

$$\tilde{t}(s, \tilde{r}) = \sum_{i \in s} b_{si} \tilde{r}_i \tag{16.6.3}$$

Theorem 16.6.1

(i) $\tilde{t}(s, \tilde{r})$ is an unbiased estimator for Y

(ii) $V\left[\tilde{t}(s, \tilde{r})\right] = \sum_{i \in U} (\alpha_i - 1) y_i^2 + \sum_{i \neq} \sum_{j \in U} (\alpha_{ij} - 1) y_i y_j + \sum_{i \in G} \alpha_i \ \phi_i$

(iii) An unbiased estimator of $V\left[\tilde{t}(s, \tilde{r})\right]$ is

$$\hat{V}\left[\tilde{t}(s, \tilde{r})\right] = \sum_{i \in s} c_{ii}(s) \tilde{r}_i^2 + \sum_{i \neq} \sum_{j \in s} c_{ij}(s) \tilde{r}_i \tilde{r}_j + \sum_{i \in s} (1 - \delta_i) b_{si} \hat{\phi}_i$$

where $\alpha_i = \sum_{s \supset i} b_{si}^2 p(s)$, $\alpha_{ij} = \sum_{s \supset i} b_{si} b_{sj} p(s)$, $\sum_{s \supset i} c_{ii}(s) p(s) = \alpha_i - 1$ and $\sum_{s \supset i} c_{ij}(s) p(s) = \alpha_{ij} - 1$

Proof

(i) $E\left[\tilde{t}(s, \tilde{r})\right] = E_p\left[\sum_{i \in s} b_{si} E_R(\tilde{r}_i)\right] = \sum_i y_i \sum_{s \supset i} b_{si} p(s) = Y$

(using the unbiasedness condition $\sum_{s \supset i} b_{si} p(s) = 1$)

$$\text{(ii)}\ V\left[\widetilde{t}(s,\widetilde{r})\right] = V_p\left[E_R\left\{\widetilde{t}(s,\widetilde{r})\right\}\right] + E_p\left[V_R\left\{\widetilde{t}(s,\widetilde{r})\right\}\right]$$

$$= V_p\left(\sum_{i\in s} b_{si}y_i\right) + E_p\left(\sum_{i\in s} b_{si}^2(1-\delta_i)\phi_i\right)$$

$$= \sum_{i\in U}(\alpha_i - 1)y_i^2 + \sum_{i\neq}\sum_{j\in U}(\alpha_{ij}-1)y_iy_j + \sum_{i\in U}\alpha_i(1-\delta_i)\phi_i$$

$$= \sum_{i\in U}(\alpha_i - 1)y_i^2 + \sum_{i\neq}\sum_{j\in U}(\alpha_{ij}-1)y_iy_j + \sum_{i\in G}\alpha_i\phi_i$$

$$\text{(iii)}\ E\left[\widehat{V}\left\{\widetilde{t}(s,\widetilde{r})\right\}\right] = E_p\left[\sum_{i\in s} c_{ii}(s)\left\{y_i^2 + (1-\delta_i)\phi_i\right\} + \sum_{i\neq}\sum_{j\in s} c_{ij}(s)y_iy_j + \sum_{i\in s}(1-\delta_i)b_{si}\phi_i\right]$$

$$= \sum_{i\in U}(\alpha_i - 1)\left\{y_i^2 + (1-\delta_i)\phi_i\right\} + \sum_{i\neq}\sum_{j\in U}(\alpha_{ij}-1)y_iy_j + \sum_{i\in U}(1-\delta_i)\phi_i$$

$$= V\left[\widetilde{t}(s,\widetilde{r})\right]$$

Remark 16.6.1

The variance of $\widetilde{t}(s,\widetilde{r})$ is smaller than that of $t(s,r)$ by an amount of $\sum_{i\in\overline{G}}\alpha_i\phi_i$. Hence ORR technique is more efficient than the compulsory RR technique.

16.6.1.2 Horvitz–Thompson Estimator Based on a Fixed Sample Size Design

Substituting $b_{si} = \dfrac{1}{\pi_i}$ in Eq. (16.6.3), the Horvitz–Thompson estimator for the population total Y is obtained as

$$t(s,\widetilde{r}) = t_{ht} = \sum_{i\in s}\frac{\widetilde{r}_i}{\pi_i} \tag{16.6.4}$$

The expression of variance and its unbiased estimators of t_{ht} are as follows:

$$V(t_{ht}) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{\gamma_i}{\pi_i} - \frac{\gamma_j}{\pi_j}\right)^2 + \sum_{i\in G}\frac{\phi_i}{\pi_i} \tag{16.6.5}$$

and

$$\widehat{V}(t_{ht}) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{\widetilde{r}_i}{\pi_i} - \frac{\widetilde{r}_j}{\pi_j}\right)^2 + \sum_{i\in s_G}\frac{\widehat{\phi}_i}{\pi_i} \tag{16.6.6}$$

16.6.1.3 Simple Random Sampling Without Replacement

For SRSWOR $\pi_i = n/N$ and $\pi_{ij} = \dfrac{n(n-1)}{N(N-1)}$, we get

$$t(s,\widetilde{r}) = t_{wor} = N\ \widetilde{\overline{r}}_s \tag{16.6.7}$$

$$V(t_{wor}) = \frac{N(N-n)}{n}S_\gamma^2 + \frac{N}{n}\sum_{i\in G}\phi_i \tag{16.6.8}$$

and an unbiased estimator of $V(t_{wor})$ is

$$\widehat{V}(t_{wor}) = \frac{N(N-n)}{n}s^2_{\widetilde{r}} + \frac{N}{n}\sum_{i\in s_G}\widehat{\phi}_i \tag{16.6.9}$$

where $(n-1)s^2_{\widetilde{r}} = \sum_{i\in s}(\widetilde{r}_i - \widetilde{\overline{r}}_s)^2$ and $\widetilde{\overline{r}}_s = \sum_{i\in s}\dfrac{\widetilde{r}_i}{n}$.

16.6.1.4 Rao–Hartley–Cochran Sampling

For the RHC sampling scheme given in Section 5.6, the expression for the unbiased estimator for the population total Y is

$$t(s,\widetilde{r}) = t_{rhc} = \sum_{i\in s}\frac{\widetilde{r}_i}{p_i}P_i \tag{16.6.10}$$

The expressions for the variance and its unbiased estimators are given by Arnab (2004a) as follows:

$$V(t_{rhc}) = \frac{N-n}{n(N-1)}\left(\sum_{i\in U}\frac{\gamma_i^2}{p_i} - Y^2\right) + \sum_{j\in G}a_j\phi_j \tag{16.6.11}$$

and

$$\widehat{V}(t_{rhc}) = \frac{N-n}{N(n-1)}\sum_{i\in s}\left(\frac{\widetilde{r}_i}{p_i} - t_{rhc}\right)^2 P_i + \sum_{j\in s_G}P_j\frac{\widehat{\phi}_j}{p_j} \tag{16.6.12}$$

where $a_j = \dfrac{1}{n(N-1)}\left(N(n-1) + \dfrac{N-n}{p_j}\right)$.

16.6.1.5 Probability Proportional to Size With Replacement Sampling

For PPSWR sampling, the estimator for the total Y is

$$t(s,\widetilde{r}) = t_{pps} = \frac{1}{n}\sum_{k=1}^{n}\frac{\widetilde{r}(k)}{p(k)} \tag{16.6.13}$$

where $\widetilde{r}(k) = \widetilde{r}_i, p(k) = p_i$ if the kth draw produces ith unit, $i = 1,\ldots, N$.

In this case the variance of $t(s,\widetilde{r})$ and its unbiased estimators are obtained as follows:

$$V(t_{pps}) = \frac{1}{n}\left[\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2 + \sum_{i\in G}\frac{\phi_i}{p_i}\right] \tag{16.6.14}$$

and

$$\widehat{V}(t_{pps}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}\left(\frac{\widetilde{r}(k)}{p(k)} - t_{pps}\right)^2 \tag{16.6.15}$$

16.6.1.6 Simple Random Sampling With Replacement

In case sample s is selected by SRSWR, an unbiased estimator of Y, its variance and an unbiased estimator of variance are obtained by substituting $p_i = 1/N$ in Eqs. (16.6.13)–(16.6.15) as follows:

$$t(s,\widetilde{r}) = t_{wr} = N\ \widetilde{\widetilde{r}} \tag{16.6.16}$$

$$V(t_{wr}) = \frac{N^2}{n}\left[\sigma_y^2 + \frac{1}{N}\sum_{i\in G}\phi_i\right] \tag{16.6.17}$$

and

$$\widehat{V}(t_{wr}) = \frac{N^2}{n(n-1)}\sum_{k=1}^{n}\left(r(k) - \widetilde{\widetilde{r}}\right)^2 \tag{16.6.18}$$

where $\sigma_y^2 = \dfrac{1}{N}\left(y_i - \overline{Y}\right)^2$ and $\widetilde{\widetilde{r}} = \dfrac{1}{N}\sum_{k=1}^{n} r(k)$.

16.6.2 Partial Optional Randomized Response Technique

PORT was originally proposed by Mangat and Singh (1994) for qualitative variables, and it was extended by several authors including Singh and Joarder (1997), Gupta et al. (2002, 2006, 2010, 2013), Gupta and Shabbir (2004), Huang (2004, 2008), Pal (2008), Chaudhuri and Dihidar (2009),

among others. Most of the researchers proposed PORT for SWRWR sampling whereas a few of them viz. Pal (2008), Chaudhuri and Dihidar (2009) for complex survey designs. Here we will describe PORT proposed by Gupta et al. (2002) only.

16.6.2.1 Gupta et al. Model

A sample s of size n units is selected from a population by SRSWR method. Each of the selected respondents was asked to choose one of the following two options:

(a) Report the true value of y or (b) report RR yx, where x is a random sample from a population with known mean $\mu_x = 1$ and known variance γ^2. Here also the interviewer will not know whether the respondents supplied the true response or RR. Let us denote the response obtained from the ith respondent as

$$z_i = x_i^k y_i$$

where $k = 1$, if the response is scrambled and $k = 0$, otherwise.

Gupta et al. (2002) derived the following results:

Theorem 16.6.2

(i) $\widehat{\mu}_y = \dfrac{1}{n}\sum_{i=1}^{n} z_i$ is an unbiased estimator of the μ_y

(ii) $V\left(\widehat{\mu}_y\right) = \dfrac{1}{n}\left[\sigma_y^2 + \gamma^2 W_G^2\left(\sigma_y^2 + \mu_x^2\right)\right]$

where W_G is the probability that a person will report the scramble response, which is generally unknown. W_G is called the degree of sensitivity of the attribute A.

Noting

$$W_G \cong \frac{E(\log \mathrm{z}) - E(\log \mathrm{y})}{E(\log x)},$$

an approximate estimator of W_G was obtained by Gupta et al. (2002) as

$$\widehat{W}_G = \frac{\dfrac{1}{n}\sum_{i=1}^{n} \log z_i - \log\left(\dfrac{1}{n}\sum_{i=1}^{n} z_i\right)}{\delta}$$

where $\delta = E(\log x)$.

16.7 MEASURE OF PROTECTION OF PRIVACY

To measure the success of an RR technique, a statistician's objective is to obtain efficient estimators of the parameters of interest whereas the interviewee's objective is to protect his/her privacy. It is seen that the efficiency of an estimator and maintenance of privacy are general goes in opposite directions. Hence one should compare the efficiencies of different RR techniques by keeping the degree of confidentiality fixed to a certain level. In this section, we will present a few measures of the degree of confidentiality proposed by Lanke (1975a,b, 1976), Leysieffer and Warner (1976) and Anderson (1975a,b,c), among others. More details have been given by Chaudhuri and Mukherjee (1988), Singh (2003) and Hong and Yan (2012).

16.7.1 Qualitative Characteristic With "Yes–No" Response

Suppose the units of a population are classified in two categories A (possessing a sensitive characteristic) and $\overline{A}$ (complement of A) with unknown proportions π_A and $1-\pi_A$, respectively. Each unit provides an RR response "R" either "Yes $= Y$" or "No $= N$" by using a suitable randomized device R^*. The conditional probabilities of obtaining an RR "R" from a unit, which belong to the group A and $\overline{A}$ are $P(R|A)$ and $P\left(R\middle|\overline{A}\right)$, respectively. These quantities are at the disposal of the investigator and are called design probabilities. The posterior probability of classifying an individual in group $A\left(\overline{A}\right)$ when he/she reports R is $P(A|R)\left(P\left(\overline{A}|R\right)\right)$. The probabilities $P(A|R)$ and $P\left(\overline{A}|R\right)$ are called revealing probabilities.

16.7.1.1 Leysieffer and Warner's Measure

According to Leysieffer and Warner (1976), the response "R" is said to be jeopardizing with respect to A if the posterior probability of classification increases given that a response "R," i.e.,

$$P(A|R) > \pi_A = P(A) \tag{16.7.1}$$

In this case the respondent feels exposed rather than protected.

Similarly the response "R" is jeopardized with respect to $\overline{A}$ if

$$P\left(\overline{A}|R\right) > 1-\pi_A = P\left(\overline{A}\right) \tag{16.7.2}$$

Since $P(A|R) > \pi_A$ implies $P\left(\overline{A}|R\right) < 1-\pi_A$, we find that whenever R is jeopardized with respect to A, R is not jeopardized with respect to $\overline{A}$ and vice versa.

Leysieffer and Warner (1976) provided us with an index of measure of jeopardy, which is known as the jeopardy function. Jeopardy of R with respect to A is

$$g(R,A) = P(R|A)/P\left(R\middle|\overline{A}\right) \tag{16.7.3}$$

Similarly, jeopardy of R with respect to $\overline{A}$ is

$$g\left(R,\overline{A}\right) = P\left(R\middle|\overline{A}\right)/P(R|A) = 1/g(R,A) \tag{16.7.4}$$

Now using Bayes theorem we find

$$\begin{aligned} P(A|R) &= \frac{P(R|A)\pi_A}{P(R|A)\pi_A + P\left(R\middle|\overline{A}\right)(1-\pi_A)} \quad \text{and} \\ P\left(\overline{A}|R\right) &= \frac{P\left(R\middle|\overline{A}\right)(1-\pi_A)}{P(R|A)\pi_A + P\left(R\middle|\overline{A}\right)(1-\pi_A)} \end{aligned} \tag{16.7.5}$$

Eq. (16.7.5) yields

$$g(R,A) = \frac{P(A|R)(1-\pi_A)}{P\left(\overline{A}|R\right)\pi_A} \quad \text{and} \quad g\left(R,\overline{A}\right) = \frac{P\left(\overline{A}|R\right)\pi_A}{P(A|R)(1-\pi_A)} \tag{16.7.6}$$

Hence the response R is jeopardized with respect to A if $g(R,A) > 1$ because in this case $P(A|R) > \pi_A$. Similarly R is jeopardized with respect to $\overline{A}$ if $g(R,A) < 1$. For $g(R,A) = 1$, R is nonjeopardizing with respect to A or $\overline{A}$.

Let a sample of size n be selected by SRSWR method and let each of the selected units provide an RR "R" independently using Warner's (1965) method. It is assumed that if a respondent is selected $t(\geq 1)$ times in the sample he/she supplies t RRs independently. Then the probability of getting response $R = Y$(Yes) is

$$\theta = P(Y|A)\pi_A + P\left(Y\middle|\overline{A}\right)(1-\pi_A) \tag{16.7.7}$$

Let m be the total number of "Y" answers and $\widehat{\lambda} = m/n$, then

$$\widehat{\pi}_A = \frac{\widehat{\lambda} - P\left(Y\middle|\overline{A}\right)}{P(Y|A) - P\left(Y\middle|\overline{A}\right)} \tag{16.7.8}$$

is an unbiased estimator of π_A provided $P(R = Y|A) \neq P\left(R = Y\middle|\overline{A}\right)$, i.e., $g(R,A) \neq 1$.

The variance of $\widehat{\pi}_A$ is

$$V(\widehat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n} + \frac{P(Y|A)\{1-P(Y|A)\}\pi_A + P\left(Y|\overline{A}\right)\left\{1-P\left(Y|\overline{A}\right)\right\}(1-\pi_A)}{n\left[P(Y|A)-P\left(Y|\overline{A}\right)\right]^2}$$

$$= \frac{\pi_A(1-\pi_A)}{n} + \frac{\pi_A g(Y,A) + (1-\pi_A)g\left(N,\overline{A}\right)}{n\{g(Y,A)-1\}\left\{g\left(N,\overline{A}\right)-1\right\}} \tag{16.7.9}$$

Without loss of generality, let us suppose that $P(Y|A) > P\left(Y|\overline{A}\right)$. Then $g(Y,A) > 1$ and $g\left(N,\overline{A}\right) > 1$, i.e., "$Y$" answer and "$N$" answers, are jeopardizing with respect to A and $\overline{A}$, respectively. It can be easily checked that $V(\widehat{\pi}_A)$ is a decreasing function of $g(Y,A)$ and $g\left(N,\overline{A}\right)$. Hence to minimize $V(\widehat{\pi}_A)$, one should choose $g(Y,A)$ and $g\left(N,\overline{A}\right)$ to be as large as possible. But for the purpose of protection of confidentiality, we cannot choose $g(Y,A)$ and $g\left(N,\overline{A}\right)$ more than certain level. Hence as compromise, one should minimize $V(\widehat{\pi}_A)$ keeping the jeopardy $g(Y,A)$ and $g\left(N,\overline{A}\right)$ to the maximum acceptable levels, say k_1 and k_2, respectively, which still allow cooperation. Thus one should choose the design parameters of the RR model R^*, which minimize the $V(\widehat{\pi}_A)$ subject to the constraint $1 < g(Y|A) < k_1$ and $1 < g\left(N|\overline{A}\right) < k_2$. In particular if we choose $g(Y,A) = g\left(N,\overline{A}\right) = k$, the design parameters for the optimum RR model R^* should be set so that

$$P(Y|A) = k/(k+1) \text{ and } P\left(\overline{Y}|A\right) = 1/(k+1) \tag{16.7.10}$$

Example 16.7.1 Comparison Between Warner and Mangat & Singh Model
For Warner's model R_w described in Section 16.2.1,

$$P(Y|A) = P_1, \quad P\left(Y|\overline{A}\right) = 1-P_1, P\left(N|\overline{A}\right) = P_1 \text{ and}$$
$$P(N|A) = 1-P_1 \text{ implies}$$

$$g(Y,A) = g_w(Y,A) = \frac{P(Y|A)}{P\left(Y|\overline{A}\right)} = \frac{P_1}{1-P_1}$$

$$\text{and } g\left(N,\overline{A}\right) = g_w\left(N,\overline{A}\right) = \frac{P\left(N|\overline{A}\right)}{P(N|A)} = \frac{P_1}{1-P_1} \tag{16.7.11}$$

here we choose the design parameter P_1 such that maximum allowable values of $g_w(Y,A)$ and $g_w\left(N,\overline{A}\right)$ attains to certain levels k_1 and k_2, respectively. Since $g_w(Y,A)=g_w\left(N,\overline{A}\right)$, we cannot choose k_1 value different from k_2. Hence we should choose P_1 such that $k_1=k_2=k$. So, our optimum choice of P_1 for a given value of $k(>1)$ is

$$P_1=\frac{k}{1+k} \tag{16.7.12}$$

The variance of $\widehat{\pi}_{Aw}$ under Warner model R_w with the optimal value $P_1=\dfrac{k}{1+k}$ comes out as

$$V_{opt}(\widehat{\pi}_{Aw})=\frac{1}{n}\left(\pi_A(1-\pi_A)+\frac{k}{(k-1)^2}\right) \tag{16.7.13}$$

For Mangat and Singh (1990) model R_{ms} described in Section 16.2.4.

$$P(Y|A)=T+(1-T)P_4,\quad P\left(Y\middle|\overline{A}\right)=(1-T)(1-P_4),$$
$$P\left(N\middle|\overline{A}\right)=T+(1-T)P_4,\quad P(N|A)=(1-T)(1-P_4)\ \text{implies}$$
$$g(Y,A)=g_{ms}(Y,A)=\frac{P(Y|A)}{P\left(Y\middle|\overline{A}\right)}=\frac{T+(1-T)(1-P_4)}{(1-T)(1-P_4)}\ \text{and}$$
$$g\left(N,\overline{A}\right)=g_{ms}\left(N,\overline{A}\right)=\frac{P\left(N\middle|\overline{A}\right)}{P(N|A)}=\frac{T+(1-T)P_4}{(1-T)(1-P_4)} \tag{16.7.14}$$

Here also $g_{ms}(Y,A)=g_{ms}\left(N,\overline{A}\right)=\dfrac{T+(1-T)P_4}{(1-T)(1-P_4)}$, so we choose the maximum allowable values of $g_{ms}(Y,A)$ and $g_{ms}\left(N,\overline{A}\right)$ are each equal to

$$k=\frac{T+(1-T)P_4}{(1-T)(1-P_4)} \tag{16.7.15}$$

Eq. (16.7.15) yields the optimum value of P_4 for a given value of T as

$$P_4=\frac{(1-T)k-T}{(1-T)(1+k)} \tag{16.7.16}$$

The expression for the variance of $\widehat{\pi}_{ms}$ under RR model R_{ms} with the optimum value of P_4 in Eq. (16.7.16) yields

$$V_{opt}(\widehat{\pi}_{Ams})=\frac{1}{n}\left(\pi_A(1-\pi_A)+\frac{k}{(k-1)^2}\right) \tag{16.7.17}$$

Eqs. (16.7.13) and (16.7.17) lead the following result obtained by Singh (2003).

Theorem 16.7.1

The Warner's RR technique R_w and Mangat–Singh RR technique R_{ms} are equally efficient under the same level of privacy protection measure suggested by Leysieffer and Warner (1976).

Example 16.7.2 Comparison Between Warner and Unrelated Model

For the Greenberg et al. (1969) unrelated model R_g with known π_x we have

$P(Y|A) = P_2 + (1 - P_2)\pi_x$, $P(Y|\overline{A}) = (1 - P_2)\pi_x$, $P(N|\overline{A}) = P_2 + (1 - P_2)(1 - \pi_x)$ and $P(N|A) = (1 - P_2)(1 - \pi_x)$ yields

$$\begin{aligned} g(Y, A) = g_g(Y, A) &= \frac{P_2 + (1 - P_2)\pi_x}{(1 - P_2)\pi_x} \text{ and} \\ g(N, \overline{A}) = g_g(N, \overline{A}) &= \frac{P_2 + (1 - P_2)(1 - \pi_x)}{(1 - P_2)(1 - \pi_x)} \end{aligned} \tag{16.7.18}$$

Now setting k_1 and k_2 as the maximum allowable values of $g_g(Y,A)$ and $g_g(N, \overline{A})$, the optimum choices of P_2 and π_x come out as

$$P_2 = \frac{(k_1 - 1)(k_2 - 1)}{k_1 k_2 - 1} \text{ and } \pi_x = \frac{k_2 - 1}{k_1 + k_2 - 2} \tag{16.7.19}$$

In case $\overline{A}$ is nonsensitive, we may set $k_2 = \infty$. In this case the optimum values of P_2 and π_x come out, respectively, as

$$P_2 = \frac{(k_1 - 1)}{k_1} \text{ and } \pi_x = 1 \tag{16.7.20}$$

Substituting Eq. (16.7.20) in the expression of $Var(\widehat{\pi}_{Ag})$ in Theorem 16.2.2, the optimum value of $Var(\widehat{\pi}_G)$ for a given value of $k_1 = k$ comes out as

$$V_{opt}(\widehat{\pi}_{Ag}) = \frac{1}{n}\left(\pi_A(1 - \pi_A) + \frac{(1 - \pi_A)}{k - 1}\right) \tag{16.7.21}$$

Now

$$V_{opt}(\widehat{\pi}_{ms}) - V_{opt}(\widehat{\pi}_{Ag}) = \frac{1}{n(k - 1)}\left(\frac{1}{k - 1} + \pi_A\right) \tag{16.7.22}$$

The expression (16.7.23) is positive since $k > 1$. Thus we have the following theorem.

Theorem 16.7.2
The Greenberg et al. unrelated question model R_2 with $\pi_x = 1$ is more efficient than the Warner model R_1 with the same degree of privacy protection measure suggested by Leysieffer and Warner (1976).

Remark 16.7.1
For Greenberg et al. RR technique with $\pi_x = 1$, each respondent has to draw a card from a pack of cards containing two types of card. The type 1 card bears sensitive question "Do you belong to the sensitive group A?" with proportion P_2 and the card type 2 bears the statement "Are you an African." Since $\pi_x = 1$, all the respondents are African. So, the RR "No" will come only from the respondents of the nonsensitive group $\overline{A}$. Hence if $\pi_x = 1$, respondents belong to the nonsensitive group $\overline{A}$ will be identified with probability 1.

16.7.1.2 Lanke's Measure

Lanke (1976) pointed out that the respondent possessing the attribute A (such as being HIV +ve) may feel embarrassed to disclose his/her membership to group A whereas membership of $\overline{A}$ (HIV −ve) may not be embarrassing. Hence, larger the conditional probability of becoming a member of A given a certain answer, the greater the embarrassment caused by providing that answer. Let $P(A|Y)$ $(P(A|N))$ be the conditional probability of a respondent belonging to group A given that he/she provided "Y" ("N") answer. Lanke's (1976) measure of protection based on the RR technique R^* is defined as

$$\mathcal{L}(R^*) = Max[P(A|Y), P(A|N)] \tag{16.7.23}$$

The smaller value of $\mathcal{L}(R^*)$ is more privacy protected.

Thus an RR technique R_1 is more protective than that of R_2 if $\mathcal{L}(R_1) < \mathcal{L}(R_2)$. R_1 and R_2 are equivalent if $\mathcal{L}(R_1) = \mathcal{L}(R_2)$.

Example 16.7.3 Comparison Between Warner and Mangat & Singh Model
From Eq. (16.7.5), we find that the values of $P(A|Y)$ and $P(A|N)$ for Warner's RR technique are, respectively,

$$P_w(A|Y) = \frac{\pi_A P_1}{\pi_A P_1 + (1-\pi_A)(1-P_1)} \quad \text{and}$$

$$P_w(A|N) = \frac{\pi_A (1-P_1)}{\pi_A (1-P_1) + (1-\pi_A) P_1}$$

In this case Lanke's measure is given by

$$\mathcal{L}_w = Max[P_w(A|Y), P_w(A|N)]$$

$$= \begin{cases} P_w(A|Y) = \dfrac{\pi_A P_1}{\pi_A P_1 + (1-\pi_A)(1-P_1)} & \text{if} \quad P_1 > 1/2 \\ P_w(A|Y) = \dfrac{\pi_A (1-P_1)}{\pi_A (1-P_1) + (1-\pi_A) P_1} & \text{if} \quad P_1 < 1/2 \end{cases} \tag{16.7.24}$$

Similarly for Mangat–Singh RR model, Lanke's measure is

$$\mathcal{L}_{ms} = \begin{cases} P_{ms}(A|Y) = \dfrac{\pi_A P^*}{\pi_A P^* + (1-\pi_A)(1-P^*)} & \text{if} \quad P^* > 1/2 \\ P_{ms}(A|N) = \dfrac{\pi_A (1-P^*)}{\pi_A (1-P^*) + (1-\pi_A) P^*} & \text{if} \quad P^* < 1/2 \end{cases}$$

where $P^* = T + (1-T)P_4$.

Consider the situation (i) $P_1 > 1/2$ and $P^* > 1/2$. In this case $\mathcal{L}_w = \mathcal{L}_{ms}$ yields $P_1 = P^*$, which in turn gives $Var(\widehat{\pi}_w) = Var(\widehat{\pi}_{ms})$. Similarly, for each of the other three situations viz. (ii) $P_1 > 1/2$ and $P^* < 1/2$, (iii) $P_1 < 1/2$ and $P^* > 1/2$ and (iv) $P_1 < 1/2$ and $P^* < 1/2$, we find $\mathcal{L}_w = \mathcal{L}_{ms}$, which implies $Var(\widehat{\pi}_w) = Var(\widehat{\pi}_{ms})$. Hence we have the following theorem obtained by Singh (2003).

Theorem 16.7.3

Under Lanke's measure, Warner's model and Mangat–Singh's model are equally efficient for maintaining the same level of privacy protection.

16.7.1.3 Anderson's Measure

Anderson (1975a) defined $P(A|R)$ and $P\left(\overline{A}|R\right)$ as two "risk of suspicion" corresponding to response R and suggested to restrict them such that

$$P(A|R) \le \xi_2 < 1 \quad \text{and} \quad P\left(\overline{A}|R\right) \le 1 - \xi_1 < 1 \tag{16.7.25}$$

Also since $P\left(\overline{A}|R\right) = 1 - P(A|R)$, Eq. (16.7.25) implies

$$\xi_1 \le P(A|R) \le \xi_2 \tag{16.7.26}$$

Since $P(A|R)$ depends on π_A, Eq. (16.7.26) gives

$$g(Y, A) \le \frac{1-\pi_A}{\pi_A} \frac{\xi_2}{1-\xi_2} \quad \text{and} \quad g\left(N, \overline{A}\right) \le \frac{\pi_A}{1-\pi_A} \frac{1-\xi_1}{\xi_1} \tag{16.7.27}$$

Thus, Anderson's criteria for protecting confidentiality is to set upper bounds for $g(Y,A)$ and $g\left(N,\overline{A}\right)$ and then minimize $Var(\widehat{\pi}_A)$ subject to this restrictions. For further details, readers are referred to Chaudhuri and Mukherjee (1988).

Flinger et al. (1977) provided with a measure of jeopardy as

$$\mathcal{F} = \frac{1 - \max\{P(A|Y), P(A|N)\}}{1 - \pi_A} = \frac{1 - \mathcal{L}}{1 - \pi_A} \tag{16.7.28}$$

Nayak (2007) pointed out that the respondents' protection increases as posterior probabilities $P(A|Y)$ and $P(A|N)$ decreases. Hence the RR device R_1 is better than R_2 for estimating π_A if

$$\begin{aligned} &P_{R_1}(A|Y) \le P_{R_2}(A|Y), \;\; P_{R_1}(A|N) \le P_{R_2}(A|N) \text{ and} \\ &V_{R_1}(\widehat{\pi}_A) \le V_{R_2}(\widehat{\pi}_A) \end{aligned} \tag{16.7.29}$$

where P_{Ri} and V_{Ri} denote, respectively, probability and variance with respect to the design R_i; $i = 1.2$.

16.7.2 Quantitative Characteristics

Anderson (1977) proposed a measure of privacy protection, which can be used for quantitative characteristics also. Suppose that the embarrassing characteristic y follows an unknown distribution function $F_y(t)$. Since y is a sensitive characteristic, it cannot be obtained directly from the respondent. So, an RR "R" is obtained from the respondent. The distribution of R depends on the unknown y-value of the respondent. Let the probability density of R for a given $y = t$ be $h_R(r|t)$. The density $h_R(r|t)$ is called the response density. The density $h_R(r|t)$ is generated by the RR device proposed by a statistician. The unconditional density of R

$$g_R(r) = \int h_R(r|t)dF_y(t) = E[h_R(r|t)] \tag{16.7.30}$$

is a mixture of response densities with $F_y(t)$ as mixing distribution.

The conditional density of t given $R = r$

$$f_y(t|r) = h_R(r|t)f_y(t)/g_R(r), \;\; r \in \Omega_r \tag{16.7.31}$$

where $f_y(t)$ is the marginal probability density of y and Ω_r is the set of possible RR values. The density $f_y(t|r)$ is known as revealing density. It depends both on response distributions and the unknown distribution of y.

After obtaining an RR "R = r" from a respondent, the revealing density $f_y(t|r)$ provides information about y, while the discrepancy between

$f_\gamma(t|r)$ and $f_\gamma(t)$ gives the amount of invasion of privacy caused by the RR "R = r". If the density $f_\gamma(t|r)$ has a high concentration about the true value γ the maintenance of privacy is very small, and alternatively if the spread of $f_\gamma(t|r)$ is high, privacy is well maintained. Anderson (1977) proposed a measure of privacy protection associated with the response $R = r$ as $V(\gamma|R = r)$ and the overall measure as

$$\tau = E\{V(\gamma|R)\} \tag{16.7.32}$$

Alternatively, one can use the relative measures $V(\gamma|R = r)/V(\gamma)$ and $E\{V(\gamma|R)\}/V(\gamma)$.

Example 16.7.4

Consider an RR technique where a respondent provides an RR $R = \gamma + X$ with γ as the true value of the study variable γ. Suppose that of γ is $N(\mu_\gamma,\sigma_\gamma)$, normal with unknown mean μ_γ and variance σ_γ^2, and X is $N(\mu_x,\sigma_x)$. Assuming X and γ are independent, we find $h_R(r|\gamma)$, the condition distribution of R, given γ is $N(\gamma + \mu_x,\sigma_x)$. The marginal distribution of R, $g_R(r)$, is normal with mean $\mu = \mu_x + \mu_\gamma$ and variance $\sigma^2 = \sigma_x^2 + \sigma_\gamma^2$. The conditional distribution of γ given $R = r$ is

$$\begin{aligned} \text{i.e. } f_\gamma(\gamma|r) &= h_R(r|\gamma) f_\gamma(\gamma)/g_R(r) \\ &= \frac{1}{\sqrt{2\pi}\,\widetilde{\sigma}} e^{-\frac{1}{\widetilde{\sigma}^2}(\gamma - \widetilde{\mu})^2} \end{aligned} \tag{16.7.33}$$

i.e., $f_\gamma(\gamma|r)$ is $N(\widetilde{\mu}, \widetilde{\sigma})$ where $\widetilde{\mu} = \left(\frac{r - \mu_x}{\sigma_x^2} + \frac{\mu_\gamma}{\sigma_\gamma^2}\right) \Big/ \left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_\gamma^2}\right)$ and $\widetilde{\sigma}^2 = \left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_\gamma^2}\right)^{-1}$.

Hence,

$$\tau = E[V(\gamma|R)] = \widetilde{\sigma}^2 = \left(1\Big/\sigma_x^2 + 1\Big/\sigma_\gamma^2\right)^{-1} \tag{16.7.34}$$

The larger σ_x^2 implies larger spread of the revealing distribution, i.e., the respondents are more protected.

16.8 OPTIMALITY UNDER SUPERPOPULATION MODEL

In this section, we will present few optimal strategies based on RR model. The results of this section were mainly derived by Arnab (1990, 1995a,b, 1998a, 2004a) after modification of the results stated in Chapter 6. Here we will assume that the population vector $\mathbf{y} = (y_1,\ldots, y_N)$ is a random variable that follows a superpopulation model ξ, which was described in Chapter 6. Let E_ξ, V_ξ and C_ξ denote respectively the expectation, variance, and covariance with respect to the model ξ. As stated earlier, the values of y_i's are not available directly from the respondents. If the ith unit (respondent) is included in the samples, a revised RR r_i is obtained from it by using some randomized device described in Sections 16.2 and 16.4. The responses r_i's are independent random variables satisfying the following RR model defined in Eq. (16.4.4) viz.

$$E_R(r_i) = y_i,\ V_R(r_i) = \phi_i = \phi_i(y_i) \text{ and } C_R(r_i, r_j) = 0 \text{ for } i \neq j$$

We have defined in Chapter 6 that the class C_{pu} consists of the p-unbiased estimators $t = t(s,y)$ based on y_i's for $i \in s$ satisfying the unbiasedness condition $E_p(t) = Y\ \forall \mathbf{y} \in R_N$. Replacing y_i by r_i in t, we define the pR unbiased (or simply unbiased) estimator based on the RR model as $t_r = t(s,r)$, which satisfies the unbiasedness condition

$$E_p E_R(t_r) = E_R E_p(t_r) = Y \quad \forall \quad \mathbf{y} \in R_N \tag{16.8.1}$$

The class of unbiased estimators for RR model will be denoted by C_{pur}. Similarly, C_{plr}, the class of linear unbiased estimators based on RR survey data consists of the estimators of the form

$$t_{lr} = b_s + \sum_{i \in s} b_{si} r_i \tag{16.8.2}$$

The estimator t_{lr} satisfies the unbiasedness condition

$$E_p E_R(t_{lr}) = E_R E_p(t_{lr}) = Y \quad \forall \quad \mathbf{y} \in R_N \tag{16.8.3}$$

The unbiasedness condition (16.8.3) yields

$$E_p(b_s) = \sum_s b_s p(s) = 0 \text{ and } \sum_{s \supset i} b_{si} p(s) = 1 \quad \forall i \in U \tag{16.8.4}$$

16.8.1 Product Measure Model

Model $M1$: y_i's are independently distributed with known $E_\xi(y_i) = \mu_i$ and unknown variance $V_\xi(y_i) = \sigma_i^2$. Since r_i's are independently distributed we have under model $M1$

$$\begin{aligned} E_{\xi R}(r_i) &= E_\xi\{E_R(r_i)\} = \mu_i \text{ and} \\ V_{\xi R}(r_i) &= E_\xi\{V_R(r_i)\} + V_\xi\{E_R(r_i)\} = E_\xi(\phi_i) + \sigma_i^2 \end{aligned} \tag{16.8.5}$$

where $E_{\xi R}$ and $V_{\xi R}$ denote operators of overall expectation and variance for the combination of the superpopulation model $M1$ and RR model R.

Using Theorem 6.3.2, we get the following theorem.

Theorem 16.8.1

Under the model $M1$ and a noninformative sampling design p with inclusion probability $\pi_i > 0 \ \forall \ i = 1,\ldots, N$

$$\begin{aligned} E_\xi E_R E_p(t_r - Y)^2 = E_{\xi R} V_p(t) &\geq \sum_{i \in U} \left\{E_\xi(\phi_i) + \sigma_i^2\right\}\left(\frac{1}{\pi_i} - 1\right) \\ &= E_{\xi R} V_p(t_{0r}) \quad \forall\, t_r \in C_{pur} \end{aligned}$$

where $t_{0r} = \sum_{i \in s} \dfrac{r_i - \mu_i}{\pi_i} + \sum_{i \in U} \mu_i$.

The estimator t_{0r} is analogous to the generalized difference estimator t_0 defined in Eq. 6.3.9. But t_{0r} cannot be used in practice because μ_i's are generally unknown. Consider a special case of the model $M1$ with $\mu_i = \beta\, x_i$, where β is an unknown constant but $x_i(>0)$ is a known value of an auxiliary variable x for the ith unit $i = 1,\ldots, N$. Let $p(\pi,x)$ be a fixed effective sample of size n design ($FED(n)$) with $\pi_i = np_i$, $p_i = x_i/X$, and $X = \sum_{i \in U} x_i$. For such a $p(\pi,x)$ design, t_{0r} reduces to $\widehat{Y}_{ht}(r) = \dfrac{1}{n}\sum_{i \in s} \dfrac{r_i}{p_i}$ and we get the following result similar to Theorem 6.3.4.

Theorem 16.8.2

Under the model $M1$ with $\mu_i = \beta\, x_i$,

$$E_{\xi R} V_{p(\pi,x)}(t_r) \geq \sum_{i \in U}\left\{E_\xi(\phi_i) + \sigma_i^2\right\}\left(\frac{X}{n\, x_i} - 1\right) = E_{\xi R} V_{p(\pi,x)}\left(\widehat{Y}_{ht}(r)\right)$$

$\forall\, t_r \in C_{pur}$.

Furthermore, for the model $M1$ with $\mu_i = \beta\, x_i$, $\sigma_i^2 = \sigma^2 x_i^2$, $E_\xi(\phi_i) = \lambda\, x_i^2$ and λ as a constant, Theorem 16.8.2 yields

$$E_{\xi R}\, V_p(t_r) \geq (\sigma^2 + \lambda) \sum_{i \in U} x_i^2 \left(\frac{1}{\pi_i} - 1 \right) \quad \forall\, t_r \in C_{pur} \tag{16.8.6}$$

Minimizing the right hand side of Eq. (16.8.6) with respect to π_i while keeping $\sum_{i \in U} \pi_i = n$ as fixed, we find the optimum value of $\pi_i = n\, x_i / X$ and the estimator t_{0r} reduces to $\widehat{Y}_{ht}(r) = \frac{1}{n} \sum_{i \in s} \frac{r_i}{p_i}$. Denoting the class of estimators with fixed effective size n sampling design $\mathcal{P}_n$, we derive the following result parallel to Theorem 6.3.5.

Theorem 16.8.3
Under the model $M1$ with $\mu_i = \beta\, x_i$, $\sigma_i^2 = \sigma^2 x_i^2$ and $E_\xi(\phi_i) = \lambda x_i^2$

$$E_{\xi R}\, V_p(t_r) \geq (\sigma^2 + \lambda) \left(\frac{X^2}{n} - \sum_{i \in U} x_i^2 \right)$$
$$= E_{\xi R}\, V_{p(\pi,x)} \left(\widehat{Y}_{ht}(r) \right) \quad \forall \quad t \in C_{pur}, p \in \mathcal{P}_n$$

From the theorem above, we note that the strategy $h_{0r} = \left(p_{(\pi,x)}, \widehat{Y}_{ht}(r) \right)$ is the optimum in the class of strategies $H = (p, t_r)$, $p \in \mathcal{P}_n$, $t_r \in C_{pur}$ under an RR technique with $E_\xi(\phi_i)$ proportional to x_i^2. The construction of such an optimum RR technique was provided by Arnab (1998a), and it is given in Section 16.8.3. Furthermore, if $x_i = 1 \ \forall \ i = 1, .., N$, $p_{(\pi,x)}$ reduces to p_0, where $\pi_i = \pi_0 = n/N$. In this case $\widehat{Y}_{ht}(r)$ reduces to $N\, \bar{r}_s$, where $\bar{r}_s = \sum_{i \in s} r_i / n$. The design p_0 includes SRSWOR. In this case Theorem 16.8.3 reduces to the following theorem.

Theorem 16.8.4
Under the model $M1$ with $\mu_i = \beta$, $\sigma_i^2 = \sigma^2$ and $E_\xi(\phi_i) = \lambda$

$$E_{\xi R}\, V_p(t_r) \geq (\sigma^2 + \lambda) N \left(\frac{N}{n} - 1 \right) = N^2 E_{\xi R} \left(V_{p_0}(\bar{r}_s) \right) \quad \forall\, p \in \mathcal{P}_n, t_r \in C_{pur}$$

16.8.2 Equicorrelation Model

Consider the model

$$\begin{aligned} M2\colon E_\xi(y_i) = \beta x_i,\ V_\xi(y_i) = \sigma^2 x_i^2 \ \text{ and } \ C_\xi(y_i, y_j) \\ = \rho x_i x_j \ \text{ with } \ -1/(N-1) \leq \rho \leq 1 \end{aligned} \tag{16.8.7}$$

Under the RR Model (16.4.4) we get

$$E_{\xi R}(r_i) = \beta x_i,\ V_{\xi R}(y_i) = \sigma^2 x_i^2 + \phi_i \ \text{ and } \ C_{\xi R}(r_i, r_j)$$
$$= E_\xi\{C_R(r_i, r_j)\} + C_\xi\{E_R(r_i), E_R(r_j)\} = \rho x_i, x_j \qquad (16.8.8)$$

For an estimator t_{lr} that belongs to the class C_{plr}, we have under model $M2$ given in Eq. (16.8.8)

$$\begin{aligned} E_{\xi R} V_p(t_{lr}) &= E_\xi E_R E_p (t_{lr} - Y)^2 \\ &= E_\xi E_p E_R (t_{lr} - Y)^2 \\ &= E_\xi E_p [\{E_R(t_{lr} - Y)\}^2 + V_R(t_{lr} - Y)] \\ &= E_\xi V_p(t_l) + \sum_{i \in U} \alpha_i E_\xi(\phi_i) \end{aligned} \qquad (16.8.9)$$

$$\left(\text{where } \ t_l = b_s + \sum_{i \in s} b_{si} y_i \ \text{ and } \ \alpha_i = \sum_{s \supset i} b_{si}^2 p(s) \right)$$

Using Theorem 6.3.9, we note that

$$E_\xi V_p(t_l) \geq (1 - \rho)\sigma^2 \left(\frac{X^2}{n} - \sum_{i \in U} x_i^2 \right) = E_\xi V_{p(\pi, x)} \left(\widehat{Y}_{ht}(r) \right) \ \text{ for } \ p \in \mathcal{P}_n \qquad (16.8.10)$$

The second part of Eq. (16.8.9) becomes

$$\sum_{i \in U} \alpha_i E_\xi(\phi_i) \geq \sum_{i \in U} E_\xi(\phi_i)/\pi_i \qquad (16.8.11)$$

$$\begin{aligned} \text{Since } \ \alpha_i &\geq \sum_{s \supset i} b_{si} p(s)^2 \Big/ \sum_{s \supset i} b_{si} p(s) \\ &= 1/\pi_i [\text{using unbiasedness condition } \ (16.8.4)]. \end{aligned} \qquad (16.8.12)$$

The right hand side of Eq. (16.8.12) is minimized for a fixed sample size design with $\sum_i \pi_i = n$ when

$$\pi_i = n\sqrt{E_\xi(\phi_i)} \Big/ \sum_{i \in U} \sqrt{E_\xi(\phi_i)} \qquad (16.8.13)$$

Now if we choose an RR model for which $E_\xi(\phi_i) = \lambda\, x_i^2$ with λ as a constant, then π_i becomes equal to $n\, x_i/X$. In this situation we arrive at the following theorem.

Theorem 16.8.5

Under model M_2 with $E_\xi(\phi_i) = \lambda\, x_i^2$,

$$E_{\xi R}\, V_p(t_{lr}) \geq [\lambda + (1-\rho)\sigma^2]\frac{X^2}{n} - (1-\rho)\sigma^2 \sum_{i\in U} x_i^2$$

$$= E_{\xi R}\, V_{p(\pi,x)}\left\{\widehat{Y}_{ht}(r)\right\} \text{ for } p \in \mathcal{P}_n,\quad t_{lr} \in C_{plr}$$

From Eq. (16.8.13) and Theorem 16.8.5, we note that the optimal estimators for the population total can be obtained if we can construct an RR model for which $E_\xi(\phi_i) = \lambda\, x_i^2$. We will call an RR model for which $E_\xi(\phi_i) = \lambda\, x_i^2$ is the optimal RR technique.

16.8.3 Construction of an Optimal Randomized Response Technique

Arnab (1998a) proposed modifications of Eriksson's (1973) and Chaudhuri's (1987) RR techniques so that $E_\xi(\phi_i)$ becomes proportional to x_i^2. For the model M_2 with $E_\xi(y_i) = \beta^* x_i$, we may choose constants $k_1,\ldots, k_L$, which anticipate the possible range of β^* so that $k_1 x_i,\ldots, k_L x_i$ in turn cover the range of y_i. The modification of the Eriksson RR technique is given as follows:

The respondent labeled i is to report either the true y_i with probability c or $Q_j(i) = k_j x_i$ with probability q_j for $j = 1,\ldots, L\left(q_j > 0, \sum_j q_j = 1 - c\right)$. Denoting z_i as the RR obtained from the ith respondent and $r_i = \left(z_i - x_i \sum_j k_j q_j\right)/c$, we find for the modified RR technique $E_\xi V_R(r_i) = \lambda x_i^2$, where $\lambda = \left[c(1-c)\left(\sigma^2 + \beta^{*2}\right) - 2c\beta^*\left(\sum_j k_j q_j\right) + \left(\sum_j k_j^2 q_j\right) - \left(\sum_j k_j q_j\right)^2\right]/c^2$.

Similarly we can modify Chaudhuri's (1987) RR technique by taking $A_i = A$, $B_i = B$, $a_i(j) = a(j)$, and $b_i(k) = b(k)x_i$ for $i = 1,\ldots, N$; $j = 1,\ldots, A$, and $k = 1,\ldots, B$. This modification yields $r_i = \dfrac{z_i - \overline{b}\, x_i}{\overline{a}}$ and $E_\xi(\phi_i) = \left[\left(\sigma^2 + \beta^{*2}\right)s_a^2 + s_b^2\right]x_i^2/\overline{a}^2$, where $s_a^2 = \sum_j \{a(j) - \overline{a}\}^2/A$, $s_b^2 = \sum_j \left\{b(j) - \overline{b}\right\}^2/B$, $\overline{a} = \sum_j a(j)/A$ and $\overline{b} = \sum_j b(j)/B$.

Remark 16.8.1

For the model $M2$ with $x_i = 1\, \forall\, i \in U$, $\widehat{Y}_{ht}(r)$ reduces to $N\ \overline{r}_s$ and $p_{(\pi,x)}$ reduces to p_0. Hence $N\ \overline{r}_s$ based on SRSWOR provides the optimal

strategy in the class of strategies $\mathcal{K} = (p, t)$, $p \in \mathcal{P}_n$, $t \in C_{lpr}$ when $E_\xi(\phi_i) = \phi_0$ is a constant. For the modified Eriksson's (1973) and Chaudhuri's (1987) RR techniques mentioned above $E_\xi(\phi_i) = \phi_0$ when $x_i = 1 \ \forall i \in U$.

Remark 16.8.2

Under a random permutation model, the probability of realizing the vector $y = (y_1 = Y_{i_1}, \ldots y_N = Y_{i_N}) = 1/N!$ for every permutation $(i_1, \ldots, i_N)$ of $(1, \ldots, N)$, where $Y_1, \ldots, Y_N$ are fixed numbers. In this model $E_\xi(y_i) = \overline{Y} = \sum_{i \in U} y_i/N$, $V_\xi(y_i) = \sum_{i \in U} (y_i - \overline{Y})^2/N$ and $C_\xi(y_i, y_j) = -1/(N-1)$ for $i \neq j$. Hence for the random permutation model, $\bar{r}_s$ based on an SRSWOR is the optimum strategy for estimating the population mean $\overline{Y}$ under the RR model with $E_\xi(\phi_i)$ as a constant.

16.9 EXERCISES

16.9.1 An RR survey was conducted to find out the proportion of HIV+ students in a university. A sample of 500 students was selected from 15,500 students by SRSWOR method and each student selected was asked to answer "Yes" or "No" to one of the questions "Are you HIV+ ?" and "I passed the Matriculation examination with a C grade" with a probability 0.7 and 0.3, respectively. Among the students selected in the sample, 324 students answer "Yes." Estimate the proportion of HIV+ students in the university and obtain 95% confidence interval of the proportion when the proportion of students received grade C in Matriculation examination is 20%.

16.9.2 Let a sample s of size n be selected from a finite population by varying probability sampling design with inclusion probability π_i for the ith unit. From each of the selected respondents, RRs were obtained by using Kuk's device described in Section 16.2.3. Show that $\widehat{\pi}_A = \dfrac{1}{Nc(P_{31} - P_{32})} \sum_{i \in s} \dfrac{z_i - cP_{32}}{\pi_i}$ is an unbiased estimator of the population proportion π_A. Derive the variance of $\widehat{\pi}_A$ and suggest an unbiased estimator of its variance.

16.9.3 A sample of 15 industrial workers was selected at random from 60 workers by SRSWOR method. Each of the workers was asked to select a ticket at random from a jar and multiply the number written

on the ticket by its actual expenditure on gambling as his/her RR. The responses have been given in the following table.

Workers	1	2	3	4	5	6	7	8	9	10
Randomized responses	35	95	100	125	250	375	125	100	75	60

Assuming that the number on a ticket follows Gamma distribution with mean 20 and variance 5, obtain an unbiased estimator of the average expenditure on gambling. Also, estimate the standard error of the estimator used.

16.9.4 Consider Greenberg's RR technique where a respondent answers the question "Are you HIV +ve?" with probability $p = 0.2$ and the question "Are you a black African" with probability $1 - p = 0.8$. A sample of 150 students is selected from 7500 students of a university by SRSWOR method. The proportion of "Yes" answer was 0.65. Estimate the proportion of HIV+ students and 90% confidence interval of the proportion when it is known that 60% of students come from the black African community.

16.9.5 Consider Greenberg et al. (1969)'s RR model described in Section 16.2.2 where π_x is unknown. Suppose two independent samples of sizes n_1 and n_2 are selected by SRSWOR method. Show that (i) $\widehat{\pi}_G(1) = \dfrac{(1-P_{22})\lambda_g(1) - (1-P_{21})\lambda_g(2)}{P_{21} - P_{22}}$ is an unbiased estimator of π_A, (ii) $Var(\widehat{\pi}_{G1}) = \dfrac{1}{(P_{21}-P_{22})^2}\left[\dfrac{(1-P_{22})^2\theta_{g1}(1-\theta_{g1})}{n_1} + \dfrac{(1-P_{21})^2\theta_{g2}(1-\theta_{g2})}{n_2}\right] - Q_1 - Q_2$ where

$$Q_1 = \frac{(1-P_{22})^2(n_1-1)\left\{P_{21}^2\pi_y(1-\pi_y) + (1-P_{21})^2\pi_x(1-\pi_x)\right\}}{n_1(N-1)(P_{21}-P_{22})^2} \text{ and}$$

$$Q_2 = \frac{(1-P_{21})^2(n_2-1)\left\{P_{22}^2\pi_y(1-\pi_y) + (1-P_{22})^2\pi_x(1-\pi_x)\right\}}{n_2(N-1)(P_{21}-P_{22})^2}, \text{ and}$$

(iii) find an unbiased estimator of $Var(\widehat{\pi}_{G1})$ (Kim, 1978).

16.9.6 Consider the RR technique R, proposed by Greenberg et al. (1969) where a respondent answers whether or not he/she possesses the sensitive character x with probability p and nonsensitive character y with probability $1 - p$. Suppose a sample s of size n is selected from a finite population of size N by SRSWOR method.

The sample s is portioned at random into two subsamples s_1 and s_2 of sizes n_1 and $n_2(=n-n_1)$, respectively. Respondents belonging to s_1 are asked to perform the randomized device R whereas respondents belonging to the subsample s_2 are directly asked whether or not they possess character y. Let $\widehat{\theta}_1$ and $\widehat{\pi}_y$ be the proportion of "Yes" answers in s_1 and s_2, respectively, and π_x be the proportion of respondents in the population that possesses the sensitive character x. Show that (i) $\widehat{\pi}_x = \left\{\widehat{\theta}_1 - (1-p)\widehat{\pi}_y\right\}\Big/p$ is an unbiased estimator of π_x and (ii) the minimum variance of $\widehat{\pi}_x$ with the optimum choice of n_1 and n_2 is $V_{\min}(\widehat{\pi}_x) = \left(\frac{\sqrt{A}+\sqrt{B}}{n} - C\right)\Big/p^2$, where

$$A = \theta(1-\theta) + \left[p^2\pi_x(1-\pi_x) + (1-p)^2\ \pi_y(1-\pi_y)\right]\Big/(N-1),$$

$$B = \frac{N(1-p)^2\pi_y(1-\pi_y)}{N-1},\ C = \frac{p^2\pi_x(1-\pi_x)}{N-1}, \text{ and}$$

$\theta = p\pi_x + (1-p)\pi_y$ (Arnab, 2006).

16.9.7 Consider an RR device where a person was asked to say "Yes" if he/she belongs to a certain sensitive group A. If the person does not belong to the group A, then he/she was asked to provide an RR using Warner's technique where he/she needs to select a card at random from a pack containing two different types of cards, which are identical in shape. The type I card with known proportion p bears the statement "I belong to the sensitive group A" whereas the type II card with proportion $1-p$ bares the statement "I do not belong to the group A." The respondent needs to answer "Yes" or "No". Let a sample s of size n be selected from a population by SRSWR method and $\widehat{\lambda}$ be the proportion of "Yes" answers obtained from the respondents based on the above sampling procedure. Show that

(i) $\widehat{\pi} = \left\{\widehat{\lambda} - (1-p)\right\}\Big/p$ is an unbiased estimator of π, the proportion of persons belongs to the group A in the population.

(ii) $V(\widehat{\pi}) = (1-\pi)\{\pi + (1-p)/p\}/n$

(iii) The MLE of π is $\widehat{\pi}_M = \begin{cases} 1-p & \text{if } \widehat{\lambda} \le 1-p \\ \widehat{\lambda} & \text{if } 1-p < \widehat{\lambda} \le 1 \end{cases}$

(iv) Let us further assume that π has a prior beta distribution

$$f(\pi|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma\alpha)I\beta}\ \pi^{\alpha-1}(1-\pi)^{\beta-1},\ \ 0<\pi<1$$

with known parameters α and β. Show that

(a) the posterior distribution of π, given $\widehat{\lambda}$, is

$$g(\pi|T=t,\alpha,\beta)=\frac{\sum_{j=0}^{t}\binom{t}{j}\ d^{t-j}\pi^{\alpha+j-1}(1-\pi)^{n+\beta-t-1}}{\sum_{j=0}^{t}\binom{t}{j}\ d^{t-j}B(\alpha+j,n+\beta-t)}$$

where $T=$ total number of "Yes" answers.

(b) The Bayes estimator under squared error loss is

$$\widehat{\pi}_B=\frac{\sum_{j=0}^{t}\binom{t}{j}\ d^{t-j}B(\alpha+j+1,n+\beta-t)}{\sum_{j=0}^{t}\binom{t}{j}\ d^{t-j}B(\alpha+j,n+\beta-t)}$$

(Kim et al., 2006)

16.9.8 Let a sample s of size n be selected by SRSWR method. The ith respondent in the sample s reports the true value y_i if he/she feels that the characteristic y is not confidential. Otherwise, if the respondent feels that the character y is confidential, he/she reports $z_i=s_iy_i/\theta$ as an RR where s_i is a random sample from a gamma population with known mean θ and variance γ^2. Let $r_i=y_i$ in case that the ith respondent reports the true value and $r_i=z_i$ if he/she reports a scrambled response. Show that (i) $\bar{r}=\sum_{i=1}^{n}r_i/n$ is an unbiased estimator of the population mean μ_y and (ii) $V(\bar{r})=\left[\sigma_y^2+W\gamma^2\left(\sigma_y^2+\mu_y^2\right)\right]\Big/n$, where σ_y is the population variance and W is the probability of reporting a scrambled response (Gupta et al., 2002).

16.9.9 For the RR model $E_R(r_i)=y_i$, $E_R(r_i)=\phi_i$ and $C_R(r_i,r_j)=0$ for $i\neq j$ prove the following results: (i) Under PPSWR sampling, the Hansen–Hurwitz estimator $\widehat{Y}_{hh}(r)=\dfrac{1}{n}\sum_{i\in s}\dfrac{r_i}{p_i}$ is admissible in the class of linear unbiased estimators of the population total.

(ii) Under SRSWR sampling, the sample mean $\bar{r}=\sum_{i\in s}r_i/n$ based on all the units (including repetition) is admissible in the class of linear unbiased estimators of the population mean (Arnab, 1995a).