CHAPTER 13

Nonresponse Errors

13.1 INTRODUCTION

In the preceding chapters, several survey designs have been discussed at length with respect to their applications. For each design, it was assumed that the true values of the variables of interest could be made available for the elements of the population under consideration. However, this is not usually the case in practice. The errors can occur at almost every stage of planning and execution of survey. These errors may be attributed to various causes right from the beginning stage, where the survey is planned and designed, to the final stage when the data are processed and analyzed. This gives rise to the following definition of nonsampling errors.

Definition 13.1 The errors arising mainly due to misleading definitions and concepts, inadequate frames, unsatisfactory questionnaire, defective methods of data collection, tabulation, coding, decoding, incomplete coverage of sample units, etc., are called *nonsampling errors.*

The incomplete coverage of units, mentioned in definition 13.1 above, occurs due to nonavailability of information from some units included in the sample. It happens if a questionnaire is mailed to a sample of units, and some respondents fail to return the completed questionnaire. If the visits are made to a sample of households, some respondents may be away from home, and others may refuse to co-operate. This leads us to definition 13.2.

Definition 13.2 The inability to collect relevant information for some of the sample units due to refusal by respondents to divulge information, their being not-at-home, sample units being inaccessible, or due to any other such reason, is termed nonresponse. The errors resulting from this incomplete coverage of the sample are called *nonresponse errors.*

Nonresponse is one of the several kinds of errors included in nonsampling errors. In this chapter, we shall only consider the problem of nonresponse. For details about other kinds of errors included in nonsampling errors, reader may refer to Sukhatme *et al.* (1984), Cochran (1977), and Zarkovich (1966).

When the respondents do not send back the required information with respect to the questionnaire mailed by the investigator, the available sample of returns is

incomplete. The resulting nonresponse is sometimes so large that it vitiates the results. While nonresponse can not be completely eliminated in practice, it could be overcome to a great extent by persuasion, or by some other methods. One way of dealing with this problem is due to Hansen and Hurwitz (1946).

13.2 HANSEN AND HURWITZ TECHNIQUE

Suppose that N units of the population can be divided into two classes. Population units (respondents) that will return the completed questionnaire, mailed to them by the investigator, without any reminders being sent, shall constitute the response class, whereas the remaining population units shall belong to the nonresponse class. Let N_1 and N_2 be the number of units in the population that form the response and nonresponse classes respectively, so that, $N = N_1+N_2$. Further, suppose that out of n units selected with equal probabilities and WOR sampling, n_1 units respond and n_2 $(= n-n_1)$ units do not respond in the first attempt. Let a WOR random subsample of h_2 units be selected from the n_2 nonresponding units, such that $n_2 = h_2 f$, for collecting information by special efforts like repeated reminders or personal interviews. Then define

$$\bar{G}_w = \frac{1}{n}\left(\sum_{i=1}^{n_1} y_i^2 + \frac{n_2}{h_2}\sum_{i=1}^{h_2} y_i^2\right) \tag{13.1}$$

$$s_{h2}^2 = \frac{1}{h_2-1}\left(\sum_{i=1}^{h_2} y_i^2 - h_2\bar{y}_{h2}^2\right) \tag{13.2}$$

where

$$\bar{y}_{h2} = \frac{1}{h_2}\sum_{i=1}^{h_2} y_i$$

We now present the following results :

Unbiased estimator of population mean $\bar{Y}$:

$$\bar{y}_w = \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}_{h2}) \tag{13.3}$$

where $\bar{y}_1$ is the sample mean based on n_1 units, and $\bar{y}_{h2}$ is defined in (13.2) above.

Variance of estimator $\bar{y}_w$:

$$V(\bar{y}_w) = \left(\frac{1}{n}-\frac{1}{N}\right)S^2 + \frac{f-1}{n}\left(\frac{N_2}{N}\right)S_2^2 \tag{13.4}$$

where S^2 and S_2^2 are the mean squares for the whole population and the nonresponse class in the population respectively.

Estimator of variance $V(\bar{y}_w)$:

$$v(\bar{y}_w) = \frac{n(N-1)}{N(n-1)}\left[\frac{N-n}{n(N-1)}(\bar{G}_w - \bar{y}_w^2) + \frac{f-1}{n}\left(\frac{n_2}{n}\right)s_{h2}^2\right] \tag{13.5}$$

The second term in (13.4) vanishes for f = 1, which should be the case if it is possible to elicit information from each of the n_2 nonrespondents.

Example 13.1

There are 570 progressive farmers in a state. The investigator is interested in estimating the average cost of herbicides used per hectare, by progressive farmers, for the paddy crop. A WOR simple random sample of 40 farmers was selected. The survey questionnaire was mailed to the selected farmers. Only 16 farmers returned the completed questionnaire. Out of the remaining 24 farmers, a further subsample of 7 farmers was selected. These selected farmers were personally contacted, and the required information obtained. The data collected are given in table 13.1.

Table 13.1 Cost (in rupees) of herbicides used per hectare

Through mail				Through personal interview	
Farmer	Cost of herbicides	Farmer	Cost of herbicides	Farmer	Cost of herbicides
1	185	9	150	1	196
2	170	10	181	2	147
3	191	11	167	3	213
4	211	12	195	4	189
5	160	13	174	5	192
6	250	14	197	6	176
7	176	15	213	7	200
8	182	16	186		

Estimate the average cost of herbicides used per hectare by the progressive farmers of the state, and work out the confidence interval for it.

Solution

From the statement of the example, we have N = 570, n = 40, n_1 = 16, n_2 = 24, and h_2 = 7. First of all, we compute means $\bar{y}_1$ and $\bar{y}_{h2}$ based on n_1 and h_2 observations respectively. Thus, we have

$$\bar{y}_1 = \frac{1}{16}(185+170+\ldots+186)$$

$$= \frac{2988}{16} \qquad \text{[from columns (2) and (4) of table 13.1]}$$

$$= 186.75$$

$$\bar{y}_{h2} = \frac{1}{7}(196+147+...+200)$$

$$= \frac{1313}{7} \qquad \text{[from column (6) of table 13.1]}$$

$$= 187.57$$

The estimate of the average cost of herbicides used per hectare is then obtained by using (13.3). This is

$$\bar{y}_w = \frac{1}{n}(n_1\bar{y}_1 + n_2\bar{y}_{h2})$$

$$= \frac{1}{40}[(16)(186.75) + (24)(187.57)]$$

$$= 187.24$$

For obtaining the estimate of variance, the expression for $\bar{G}_w$ in (13.1), and that for s^2_{h2} in (13.2), are to be evaluated. We, therefore, have

$$\bar{G}_w = \frac{1}{n}\left(\sum_{i=1}^{n_1} y_i^2 + \frac{n_2}{h_2}\sum_{i=1}^{h_2} y_i^2\right)$$

$$= \frac{1}{40}\left[(185)^2 + (170)^2 + ... + (186)^2 + \frac{24}{7}\{(196)^2 + (147)^2 + ... + (200)^2\}\right]$$

$$= \frac{1}{40}\left[566552 + \frac{24}{7}(248955)\right]$$

$$= 35502.80$$

$$s^2_{h2} = \frac{1}{h_2 - 1}\left(\sum_{i=1}^{h_2} y_i^2 - h_2\bar{y}^2_{h2}\right)$$

$$= \frac{1}{7-1}[248955 - 7(187.57)^2]$$

$$= 446.24$$

Now the variance estimator in (13.5) is given by

$$v(\bar{y}_w) = \frac{n(N-1)}{N(n-1)}\left[\frac{N-n}{n(N-1)}(\bar{G}_w - \bar{y}_w^2) + \frac{f-1}{n}\left(\frac{n_2}{n}\right)s^2_{h2}\right]$$

Substituting the values of different terms, one obtains

$$v(\bar{y}_w) = \frac{40(570-1)}{570(40-1)}\left[\frac{(570-40)}{40(570-1)}\{35502.80 - (187.24)^2\}\right.$$

$$\left. + \left\{\frac{(24/7)-1}{40}\right\}\left(\frac{24}{40}\right)(446.24)\right]$$

$$= (1.0238)[(.0233)(443.982) + (.0607)(.6)(446.24)]$$

$$= 27.2298$$

The required confidence interval is obtained by using the formula

$$\bar{y}_w \pm 2\sqrt{v(\bar{y}_w)}$$

On substituting for different terms, it reduces to

$$\begin{aligned} & 187.24 \pm 2\sqrt{27.2298} \\ = \; & 187.24 \pm 10.44 \\ = \; & 176.80,\ 197.68 \end{aligned}$$

This indicates that had all the 570 farmers been surveyed, the average cost of herbicides used per hectare would have taken a value in the closed interval [176.80, 197.68] with approximate probability .95. ■

A *cost function* appropriate for the Hansen and Hurwitz technique is given by

$$C' = c_0 n + c_1 n_1 + c_2 h_2 \tag{13.6}$$

where

c_0 = the cost of including a sample unit in the initial survey,
c_1 = the per unit cost of collecting, editing, and processing information on the study variable in the response class, and
c_2 = the per unit cost of interviewing and processing same information in the nonresponse class.

As C' varies from sample to sample, we consider expected cost

$$E(C') = C = \frac{n}{N}\left(Nc_0 + N_1 c_1 + \frac{N_2}{f} c_2\right) \tag{13.7}$$

We now determine the optimum values of n and f for which C is minimum and the variance $V(\bar{y}_w) = V_0$.

Optimum n and f for fixed variance V_0 that minimizes cost in (13.7) :

$$f = \left[\frac{c_2(S^2 - N_2 S_2^2/N)}{S_2^2 (c_0 + N_1 c_1/N)}\right]^{\frac{1}{2}} \tag{13.8}$$

$$n = \frac{S^2 + \{N_2(f-1)\, S_2^2\}/N}{V_0 + S^2/N} \tag{13.9}$$

where the symbols involved have already been defined.

Example 13.2

In a survey, the expected response rate is 40 percent, $S_2^2 = (3/4)S^2$, it costs \$ .5 to include a unit in the sample, \$2 per unit to observe study variable in the response class, and \$6 per unit to observe the study variable in the nonresponse class. Determine optimum

values of f and n if the population mean is to be estimated with a tolerable variance $V_0 = S^2/200$. Also, work out the total expected cost of the survey with the cost function considered in (13.7).

Solution

The statement of the example provides

$$N_1 = N\,\frac{40}{100},\quad N_2 = N\,\frac{60}{100},\quad S_2^2 = \frac{3}{4}\,S^2,\quad c_0 = .5,\ c_1 = 2,\ \text{and}\ c_2 = 6.$$

On making substitutions in (13.8), one arrives at the optimum value of f as

$$f = \left[\frac{6\{S^2 - (60)\,(3S^2)/400\}}{(3S^2/4)\,\{.5 + (40)\,(2)/100\}}\right]^{\frac{1}{2}}$$

$$= \left[\frac{6\{1 - (60)\,(3)/400\}}{(3/4)\,\{.5 + (40)\,(2)/100\}}\right]^{\frac{1}{2}}$$

$$= 1.84$$

The optimum value of n is given by (13.9). On substituting different values, one gets

$$n = \frac{S^2 + (60)\,(1.84 - 1)\,(3S^2)/400}{(S^2/200) + S^2/N}$$

$$= \frac{200\,[1 + (60)\,(1.84 - 1)\,(3)/400]}{1 + 200/N}$$

For large N, it yields

$$n = 275.6 \approx 276.$$

From (13.7), the total expected cost (in dollars) of the survey, is found as

$$C = 276\left[.5 + \frac{40}{100}(2) + \left(\frac{60}{100}\right)\left(\frac{1}{1.84}\right)(6)\right]$$

$$= 898.79\ \blacksquare$$

Hansen and Hurwitz technique loses its merit when nonresponse is large. Durbin (1954) observed that when $S_2^2 = S^2$, and cost of collecting data in the nonresponse class is much higher than that in the response class, it will not be worthwhile to go for this technique.

El-Badry (1956) extended Hansen and Hurwtiz technique further. He suggested to send repeated waves of questionnaires to the nonresponding units. As soon as the investigator feels that further waves will not be much effective, a subsample from the remaining nonresponding units is selected and the information is collected through personal interview. The required estimate is based on the total information collected from all the attempts.

Deming (1953) has advocated the use of *call-back technique*. According to him, if the chosen sample member is not at home or is unable to take part in an interview at the time of call, then it is advisable for the interviewer to call back. He has shown how successive recalls help in reducing the bias, and proposed a method to arrive at optimum number of recalls to achieve a given precision. He has also determined the optimum number of call-backs for the fixed total cost of survey by minimizing the variance.

An interesting plan dealing with the reduction of bias *without call-backs* is also available. This is presented in the following section.

13.3 BIAS REDUCTION WITHOUT CALL - BACKS

An ingenious technique of reducing biases present in the results of the first call was given by *Politz and Simmons* (1949), for surveys where information on study variable is collected through interview method. The proposed procedure attempts to reduce the bias, owing to incomplete sample, without resorting to successive call-backs. According to this technique, the calls are made during the evening on the six weekdays. The time of the call is assumed random within the interview hours. If the respondent is available at home, the required information is obtained. Besides, he is also asked how many times he was at home, at the time of visit, on each of the five preceding weekdays ? Suppose that n respondents were selected by using WR equal probability sampling. Further, let p_i be the probability that the i-th respondent is found at home at the time of call. Then, estimate of p_i is given by

$$\hat{p}_i = \frac{t+1}{6}, \quad t = 0, 1, \ldots, 5 \tag{13.10}$$

where t is the number of times the respondent was at home, at the time of call, during the last five evenings. Defining

$$Q_i = \sum_{t=0}^{5} \frac{1}{(t+1)} \binom{6}{t+1} p_i^{t+1} q_i^{5-t} \tag{13.11}$$

where $q_i = 1-p_i$, the estimator of population mean and the other related results, for this procedure, are then as given in (13.12) through (13.15).

Estimator of population mean $\bar{Y}$:

$$\bar{y}_{ps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\hat{p}_i} \tag{13.12}$$

where y_i assumes value zero if the respondent is not found at home at the time of call.

Bias of estimator $\bar{y}_{ps}$:

$$B(\bar{y}_{ps}) = -\frac{1}{N} \sum_{i=1}^{N} Y_i q_i^6 \tag{13.13}$$

Variance of estimator $\bar{y}_{ps}$:

$$V(\bar{y}_{ps}) = \frac{1}{n}\left[\frac{6}{N}\sum_{i=1}^{N} Y_i^2 Q_i - \left\{\frac{1}{N}\sum_{i=1}^{N} Y_i\,(1-q_i^6)\right\}^2\right] \tag{13.14}$$

where Q_i has been defined in (13.11).

Estimator of variance $V(\bar{y}_{ps})$:

$$\begin{aligned} v(\bar{y}_{ps}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{p}_i} - \bar{y}_{ps}\right)^2 \\ &= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{y_i}{\hat{p}_i}\right)^2 - n\bar{y}_{ps}^2\right] \end{aligned} \tag{13.15}$$

Example 13.3

The Wipro company is engaged in marketing of personal computers (PC). For taking certain decisions, the company needs information on the average annual maintenance cost for a personal computer purchased three to five years ago. For this purpose, a WR random sample of 28 buyers was selected from 640 buyers, to whom the company had sold personal computers during the last three to five years. The buyers were interviewed in the afternoon of the weekdays. The information in respect of annual maintenance cost (y) in rupees, and the number of days on which the respondent was available at the time of call during the last five days (denoted by t), is exhibited in table 13.2. The estimation variable assumes zero value if the respondent is not available at the time of call. Marks "—" against respondents at serial numbers 9 and 20 indicate their nonavailability.

Table 13.2 The data collected from the sampled buyers

Respondent	y_i	t	$\hat{p}_i = \frac{t+1}{6}$	Respondent	y_i	t	$\hat{p}_i = \frac{t+1}{6}$
1	500	3	.6667	15	900	5	1.0000
2	1000	5	1.0000	16	1300	4	.8333
3	3500	4	.8333	17	560	5	1.0000
4	400	4	.8333	18	700	3	.6667
5	2600	5	1.0000	19	0	3	.6667
6	1500	3	.6667	20	—	—	—
7	800	5	1.0000	21	1600	5	1.0000
8	4100	3	.6667	22	1000	4	.8333
9	—	—	—	23	520	4	.8333
10	710	3	.6667	24	1300	5	1.0000
11	2100	4	.8333	25	2500	3	.6667
12	1050	4	.8333	26	0	2	.5000
13	0	2	.5000	27	1200	4	.8333
14	1125	3	.6667	28	300	1	.3333

Estimate the average annual repair charges per PC, and obtain confidence interval for it.

Solution

From the statement of the example, we have N = 640 and n = 28. The estimate of average annual maintenance cost per PC is computed by using (13.12). The formula is

$$\bar{y}_{ps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\hat{p}_i}$$

The estimates $\hat{p}_i$ are computed and given in table 13.2. Making substitutions for n, y_i, and $\hat{p}_i$, yields

$$\bar{y}_{ps} = \frac{1}{28}\left(\frac{500}{.6667} + \frac{1000}{1.0000} + \ldots + \frac{300}{.3333}\right)$$

$$= \frac{1}{28}(39646.28)$$

$$= 1415.94$$

as an estimate of average annual maintenance cost.

The variance estimator is provided by (13.15). We, therefore, write

$$v(\bar{y}_{ps}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{\hat{p}_i} - \bar{y}_{ps}\right)^2$$

$$= \frac{1}{28(28-1)}\left[\left(\frac{500}{.6667} - 1415.94\right)^2 + \left(\frac{1000}{1.0000} - 1415.94\right)^2 + \ldots + \left(\frac{300}{.3333} - 1415.94\right)^2\right]$$

$$= \frac{1}{28(28-1)}\left[\left(\frac{500}{.6667}\right)^2 + \left(\frac{1000}{1.0000}\right)^2 + \ldots + \left(\frac{300}{.3333}\right)^2 - 28(1415.94)^2\right]$$

$$= \frac{1}{28(28-1)}[109319425 - 56136808]$$

$$= 70347.374$$

The confidence limits will be obtained by using

$$\bar{y}_{ps} \pm 2\sqrt{v(\bar{y}_{ps})}$$

$$= 1415.94 \pm 2\sqrt{70347.374}$$

$$= 1415.94 \pm 530.46$$

$$= 885.48,\ 1946.40$$

Above confidence limits indicate that the average annual maintenance cost per PC, for the population, is covered by the interval [885.48, 1946.40] with approximate probability .95. ■

13.4 WARNER'S RANDOMIZED RESPONSE MODEL

Sample surveys on human populations have established that the innocuous questions usually receive good response, whereas questions on sensitive items involving controversial assertions, stigmatizing and/or incriminating matters, which people like to hide from others, excite resistance. Direct questions about them often result in either refusal to respond, or falsification of their answers. This introduces *nonresponse error* which makes the estimation of relevant parameters, for instance, proportion of population belonging to sensitive group, unreliable. Recognizing the fact that such biases are frequent when respondents are queried directly about sensitive or embarrassing matters, Warner (1965) developed an ingenious interviewing procedure to reduce or eliminate these evasive answer biases.

This model was built on the hope that the co-operation might be better when the respondent is requested for information in anonymity, provided by randomization device, rather than on direct question basis. The randomization device creates a stochastic relationship between the question and the individual's response, and thus provides protection to confidentiality of the respondent. The information from the respondents is elicited in terms of "yes" or "no" answers without endangering their privacy. He called the procedure as *randomized response* (RR) *technique*. We first discuss the Warner's (1965) pioneer model, and then some other popular modifications and improvements.

The procedure involves the use of two statements, each of which divides the population into two mutually exclusive and complementary classes, say, A and not-A. In order to estimate π, the proportion of respondents in sensitive group A, every person selected in a WR simple random sample of n respondents is given a suitable randomization device consisting of two mutually exclusive statements of the form :

1. "I belong to sensitive group A"
2. "I belong to group not-A".

The statements (1) and (2) are represented in the randomization device with probability p and (1-p) respectively. The randomization device, for instance, may be a deck of cards of which a proportion p carries the first statement while on the remaining cards second statement is written. It could also be a spinner in which a sector formed by an angle of 360p degrees is marked with first statement, and the remaining 360(1-p) degrees sector is marked with second statement. The respondent is asked to choose one of the two statements randomly and to answer "yes" if the selected statement points to his group with respect to the attribute A, "no" otherwise. The interviewee, however, does not reveal to the interviewer as to which of the two statements has been chosen. It is assumed that these "yes" and "no" answers are reported truthfully. The formulas for estimator of proportion, its variance, and estimator of variance are given in (13.16) to (13.18).

Unbiased estimator of proportion π :

$$\hat{\pi}_w = \frac{(n'/n) - 1 + p}{2p - 1}, \quad p \neq .5, \tag{13.16}$$

where n' is the number of persons answering "yes".

Variance of estimator $\hat{\pi}_w$:

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{13.17}$$

Unbiased estimator of variance $V(\hat{\pi}_w)$:

$$v(\hat{\pi}_w) = \frac{(n'/n)(1 - n'/n)}{(n-1)(2p-1)^2} \tag{13.18}$$

The expression (13.17) reveals that it is desirable to choose p as close to 1 or 0 as possible without threatening the degree of co-operation by the respondent.

The RR technique was extended to polychotomous populations by Abul-Ela *et al.* (1967). Some test procedures for detecting untruthful answering by the respondents, for Warner's (1965) model, have been developed by Lakshmi and Raghavarao (1992), and Krishnamoorthy and Raghavarao (1993).

Example 13.4

The Mathura depot of State Transport Corporation has received complaints that some of the bus drivers resort to drunken driving. The General Manager of the depot has been asked by the state administration to start a campaign creating awareness against drunken driving among the drivers. The General Manager thought it wise to estimate the proportion of such drivers. In order to accomplish the objective, a WR simple random sample of 120 drivers, out of a total of 908 drivers, was selected. Because of the sensitive nature of inquiry, it was decided to use randomized response technique. The randomization device used consisted of a deck of 100 cards, 80 of which bore the statement "I have resorted to drunken driving". The remaining 20 cards carried the statement "I have not resorted to drunken driving".

The interviewee was asked to select a card randomly from the deck after shuffling it, and report "yes" if the statement written on the card pointed to his actual status, and "no" otherwise. He was not to tell the interviewer as to which of the two statements had been chosen. In all, there were 36 "yes" and 84 "no" responses. Using this information, estimate the proportion of bus drivers who resort to drunken driving. Also, obtain confidence limits for it.

Solution

In this case, we have n = 120, n' = 36, and p = 80/100 = .8. The required estimate is obtained by using (13.16). Therefore,

$$\hat{\pi}_w = \frac{(n'/n) - 1 + p}{2p - 1}$$

$$= \frac{(36/120) - 1 + .8}{2\ (.8) - 1}$$

$$= .167$$

We now work out the estimate of variance from (13.18). Thus, we have

$$v(\hat{\pi}_w) = \frac{(n'/n)\ (1 - n'/n)}{(n-1)\ (2p-1)^2}$$

On making substitutions, one gets

$$v(\hat{\pi}_w) = \frac{(36/120)[1 - (36/120)]}{(120-1)\ [2\ (.8) - 1]^2}$$

$$= .004902$$

The confidence limits for π are

$$\hat{\pi}_w \pm 2\sqrt{v(\hat{\pi}_w)}$$

$$= .167 \pm 2\sqrt{.004902}$$

$$= .167 \pm .140$$

$$= .027, .307$$

Thus, from the sample data, the proportion of drivers who had resorted to drunken driving is estimated as .167. Also, its population value is most likely to fall in the confidence interval [2.7, 30.7] percent. ■

The large width of the confidence interval is due to large variance. The variance of the estimator $\hat{\pi}_w$ gets inflated as it contains an additional component of variance because of the randomization of response. The method, therefore, generally requires a large sample size to obtain a reasonably small variance of the estimator.

In the Warner's model, the sample size is fixed and the number of "yes" answers is a random variable. Mangat and Singh (1991a) have given a procedure in which the number of "yes" answers is fixed but the sample size is a random variable. They have given another modification known as two-stage procedure. This technique is discussed in the next section.

13.5 MANGAT AND SINGH'S TWO-STAGE MODEL

In Mangat and Singh's (1990) method, each interviewee in the WR simple random sample of n respondents is provided with two randomization devices R_1 and R_2. The randomization device R_1 consists of two statements, namely :

1. " I belong to sensitive group A"
2. " Go to randomization device R_2",

represented with probabilities T and (1-T) respectively. The randomization device R_2 which uses two statements,

1. "I belong to sensitive group A"
2. "I belong to group not-A",

with known probabilities p and (1-p), is exactly the same as used by Warner (1965). The interviewee is instructed to experience first the randomization device R_1. One is to use R_2 only if directed by the outcome of R_1. The respondent is required to answer "yes" if the outcome points to the attribute he/she possesses, and answer "no" if the complement of his status is pointed out by the outcome. The estimator and the related results for this model are given below :

Unbiased estimator of population proportion π :

$$\hat{\pi}_{ms} = \frac{n'/n - (1-T)(1-p)}{2p-1+2T(1-p)} \tag{13.19}$$

where, as in Warner's model, n' is the number of "yes" answers.

Variance of estimator $\hat{\pi}_{ms}$:

$$V(\hat{\pi}_{ms}) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)[1-(1-T)(1-p)]}{n[2p-1+2T(1-p)]^2} \tag{13.20}$$

Estimator of variance $V(\hat{\pi}_{ms})$:

$$v(\hat{\pi}_{ms}) = \frac{(n'/n)(1-n'/n)}{(n-1)[2p-1+2T(1-p)]^2} \tag{13.21}$$

Mangat and Singh's strategy can always be made more efficient than the Warner's model by suitably choosing the value of T for any practical value p. Such a value of T is given by

$$T > \frac{1-2p}{1-p} \tag{13.22}$$

The estimator $\hat{\pi}_{ms}$ using a value of T satisfying above inequality is, therefore, expected to yield a confidence interval of smaller width than the one provided by Warner's model. However, if the probability for drawing the statement on sensitive character in Warner's model is taken equal to T+p(1-T), and it is retained at level p in the second randomization device of Mangat and Singh's model, then both the strategies are equally efficient.

Example 13.5

For example 13.4, estimate the proportion of drivers resorting to drunken driving, using Mangat and Singh's two-stage procedure, assuming that the number of "yes" answers received was 30. Take the value of T as .3.

Solution

Now we have $n = 120$, $n' = 30$, $p = .8$, and $T = .3$. The estimate of the required proportion is worked out by using (13.19). Thus,

$$\hat{\pi}_{ms} = \frac{(30/120) - (1-.3)(1-.8)}{2(.8) - 1 + 2(.3)(1-.8)}$$

$$= .153$$

The estimator of variance in this case is provided by (13.21). Hence,

$$v(\hat{\pi}_{ms}) = \frac{(n'/n)(1-n'/n)}{(n-1)[2p-1+2T(1-p)]^2}$$

$$= \frac{(.25)(1-.25)}{(120-1)[2(.8)-1+2(.3)(1-.8)]^2}$$

$$= .003039$$

The confidence limits for the population proportion are, therefore, calculated as

$$\hat{\pi}_{ms} \pm 2\sqrt{v(\hat{\pi}_{ms})}$$

$$= .153 \pm 2\sqrt{.003039}$$

$$= .153 \pm .110$$

$$= .043, .263$$

The estimate of the proportion of bus drivers who resort to drunken driving is thus found to be .153. Also, the confidence limits indicate that the population proportion is likely to vary from 4.3% to 26.3%. ■

Mangat and Singh (1991b) have extended their above procedure to the situation where the respondents are selected using SRS without replacement method. For this case also, they considered the estimator in (13.19). However, the estimator of variance is little complicated for this procedure. Several other workers have also modified the Warner's (1965) model. Franklin (1989) and Singh and Singh (1992, 1993) use continuous randomization device instead of discrete one. Mangat (1994) and Mangat *et al.* (1995) have proposed some other variants of Warner's model.

Walt R. Simmons felt that the confidence of the respondents might be enhanced further if one of the two questions refers to a nonsensitive attribute (say) y, unrelated to the stigmatized characteristic. Following his suggestion, Horvitz *et al.* (1967) developed a procedure, and called it *unrelated question randomized response model.* In short, we shall call it "U-model".

13.6 UNRELATED QUESTION MODEL

Though the *unrelated question model* was proposed by Horvitz *et al.* (1967), but the theoretical framework for it was developed by Greenberg *et al.* (1969). While developing theory, they dealt with both the situations where π_y, the proportion of innocuous character in the population, is known and when it is unknown.

13.6.1 *Case I - π_y Unknown*

For the situation when π_y is not known, two samples of sizes n_1 and n_2 respondents are drawn from the population using SRS with replacement, so that, $n_1 + n_2 = n$ is the overall sample size. Each sample is then used to collect information on both the characters. In this model, two sets of randomization device need to be used. Each device consists of the following two statements :

1. "I am a member of group A"
2. "I am a member of group Y",

where, as already mentioned, group A consists of units possessing the sensitive characteristic, and membership in group Y carries no embarrassment. Statement regarding character y in the randomization device could be -"Does your date of birth fall in the month of January or February ?". π_y, the probability for the statement to hold for any respondent, could be manipulated by increasing or decreasing the length of time period in which birth date is to fall. However, actual value of π_y may or may not be known. In the first randomization device, let group A be represented with probability p_1 and the group Y with $(1-p_1)$, whereas in the second device the groups A and Y be represented respectively with the probabilities p_2 and $(1-p_2)$. Let the first device be provided to each of the respondents in the sample of size n_1, and the second device is used for the respondents in the sample of size n_2. Each respondent is asked to select a statement randomly, and unobserved by the interviewer, from the device provided to him. He/she is required to report "yes" if the selected statement points to his/her actual status, and "no" otherwise. For i = 1, 2, let

n'_i = number of "yes" answers reported in the i-th sample, and

$$\theta_i = p_i\pi + (1-p_i)\,\pi_y \qquad (13.23)$$

be the probability that a "yes" answer will be reported by the respondents in the i-th sample. Postulating that the respondents report complete truth, the estimator and other related results are as follows :

Unbiased estimator of π when π_y is not known :

$$\hat{\pi}_g = \frac{1}{p_1 - p_2}\left[(1-p_2)\,\frac{n'_1}{n_1} - (1-p_1)\,\frac{n'_2}{n_2}\right] \qquad (13.24)$$

Variance of estimator $\hat{\pi}_g$:

$$V(\hat{\pi}_g) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2\,\frac{\theta_1(1-\theta_1)}{n_1} + (1-p_1)^2\,\frac{\theta_2(1-\theta_2)}{n_2}\right] \qquad (13.25)$$

Estimator of variance $V(\hat{\pi}_g)$:

$$v(\hat{\pi}_g) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2 \frac{(n_1'/n_1)(1-n'_1/n_1)}{n_1 - 1} + (1-p_1)^2 \frac{(n'_2/n_2)(1-n'_2/n_2)}{n_2 - 1}\right] \quad (13.26)$$

In order to minimize the variance of the estimator $\hat{\pi}_g$, Greenberg *et al.* (1969) deduced the following rules :

1. Choose one of the p_i, i = 1, 2, as close to 1 and the other as close to zero as the respondents are likely to accept, such that $p_1 + p_2 = 1$.
2. Choose π_y close to 0 or 1 according as $\pi < .5$ or $\pi > .5$. If $\pi = .5$, then $|\pi_y - .5|$ could be maximum on either side. While choosing π_y close to 0 or 1, depending on π, the investigator should take care that it should not be selected too close to 0 or 1, as it may affect the likelihood of co-operation by the respondents, and thus contradict the whole purpose of using unrelated question approach.
3. In practical situations, the total sample size $n(= n_1 + n_2)$ is fixed. One is then concerned with the choice of n_1 and n_2, so that the variance $V(\hat{\pi}_g)$ is minimized. The optimal allocation of n into n_1 and n_2 is given by

$$\frac{n_1}{n_2} = \frac{(1-p_2)\sqrt{\theta_1(1-\theta_1)}}{(1-p_1)\sqrt{\theta_2(1-\theta_2)}} \quad (13.27)$$

In application of rule (13.27), it is necessary to use a rough guess of π and π_y in order to calculate θ_1 and θ_2 defined in (13.23).

Example 13.6

A tough and neck to neck campaign is on between two candidates A and B in ward number 56 of a metropolitan city during elections for municipal corporation. Owing to surcharged and tense atmosphere, voters hesitate to divulge openly as to which candidate they would vote. In order to estimate the voting behavior in advance, two independent SRS with replacement samples - one consisting of 160 voters and other of 140 voters - were drawn from the frame (voters' list) consisting of 4160 voters.

Two decks of cards were prepared. The deck I consisted of two types of cards bearing statements :

1. "I shall vote for candidate A"
2. "I was born in the month of November",

occurring with probabilities .8 and .2 respectively. In deck II the statements (1) and (2) respectively were represented with probabilities .2 and .8 . The deck I was used for the voters in the first sample and deck II for the voters of the second sample. Each voter was asked to choose a statement randomly, unobserved by the interviewer, and say "yes"

if the selected statement points to his actual status, and "no" otherwise. On completion of survey, it was found that 55 "yes" answers had been reported by the respondents in the first sample, and 68 "yes" answers had come from the second sample.

Estimate the proportion of voters favoring candidate A, and also obtain confidence interval for it.

Solution

From the statement of the example we have $n_1 = 160$, $n_2 = 140$, so that, $n = 160 + 140 = 300$. Also, $p_1 = .8$, $p_2 = .2$, $n'_1 = 55$, and $n'_2 = 68$. The estimate of the proportion of voters who would vote for candidate A, is computed by using (13.24). Therefore,

$$\hat{\pi}_g = \frac{1}{p_1 - p_2}\left[(1-p_2)\frac{n'_1}{n_1} - (1-p_1)\frac{n'_2}{n_2}\right]$$

$$= \frac{1}{(.8-.2)}\left[(1-.2)\frac{55}{160} - (1-.8)\frac{68}{140}\right]$$

$$= .2964$$

The estimate of variance is provided by (13.26). The expression for this estimator is

$$v(\hat{\pi}_g) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2 \frac{(n'_1/n_1)(1-n_1'/n_1)}{n_1 - 1} + (1-p_1)^2 \frac{(n'_2/n_2)(1-n'_2/n_2)}{n_2 - 1}\right]$$

On making substitutions, it gives

$$v(\hat{\pi}_g) = \frac{1}{(.8-.2)^2}\left[(1-.2)^2 \frac{(55/160)(1-55/160)}{160-1} + (1-.8)^2 \frac{(68/140)(1-68/140)}{140-1}\right]$$

$$= .002722$$

The required confidence interval will, therefore, be

$$\hat{\pi}_g \pm 2\sqrt{v(\hat{\pi}_g)}$$

$$= .2964 \pm .1043$$

$$= .1921, .4007$$

The confidence limits obtained above, indicate that the candidate A is most likely to secure 19.21% to 40.07% of the total votes. The survey thus indicates a probable win for candidate B. ■

Moors (1971) and Tracy and Mangat (1995a) have advocated the estimation of proportion of the nonsensitive character by asking direct question to the respondents in one of the two samples, and collecting information on both the characters, through randomization device, from the other sample. Folsom *et al.* (1973) and Tracy and Mangat (1995b) used two unrelated questions instead of one.

13.6.2 *Case II - π_y Known*

Consider the survey where the units of the population are the employees of a university and the unrelated nonsensitive question is: "Were you born in the month of October ?". The proportion π_y of the employees born in the month of October can be had from their records available in the university. When π_y is known, the U-model gives more precise estimates. In such a case, only one sample of the respondents is required. The results for this situation are quite simple and are given below :

Unbiased estimator of population proportion π :

$$\hat{\pi}_g = \frac{(n'/n) - (1-p)\pi_y}{p} \tag{13.28}$$

Variance of estimator $\hat{\pi}_g$:

$$V(\hat{\pi}_g) = \frac{\theta(1-\theta)}{np^2} \tag{13.29}$$

where $\theta = p\pi + (1\text{-}p)\,\pi_y$.

Unbiased estimator of variance $V(\hat{\pi}_g)$:

$$v(\hat{\pi}_g) = \frac{(n'/n)(1-n'/n)}{(n-1)\,p^2} \tag{13.30}$$

The optimal p is chosen as close to 1 as it is possible and practicable. However, the choice of π_y is made as in π_y- unknown case. Some other modifications of the U-model for π_y known case, are due to Mangat *et al.* (1992) and Singh *et al.* (1993).

Example 13.7

Abernathy *et al.* (1970) conducted a survey in North Carolina to estimate the proportion of women having had an abortion during the past year, among white women of age 18-44 years. They drew a WR random sample of 782 white women. To each women, included in the sample, was provided a randomization device carrying following two statements :

1. "I was pregnant at some time during the past 12 months, and had an abortion which ended the pregnancy"
2. "I was born in the month of April".

The randomization device used consisted of a small, transparent, sealed plastic box. Inside the box, there were 35 red and 15 blue balls. The respondent was asked to shake the box of balls thoroughly, and to tip the box allowing one of the freely moving balls to appear in a window which was clearly visible to the respondent. If a red ball appeared, she was required to answer question (1), and if a blue ball appeared she answered question (2). The respondent's reply was simply "yes" or "no" without specifying to which question the answer referred. On completion of survey, it was found

that 27 "yes" answers had been reported. Estimate the proportion in question, and work out the confidence interval for it, taking the proportion of women born in April as .0826. This figure was obtained from a distribution of births occurring to North Carolina residents during 1924-1950.

Solution

From the statement of the example we have n = 782, p = 35/50 = .7, n' = 27, and π_y = .0826, so that, n'/n = 27/ 782 = .03453. The required estimate of proportion of women who had had an abortion during the past year, is provided by (13.28). Thus,

$$\hat{\pi}_g = \frac{(n'/n) - (1-p)\pi_y}{p}$$

$$= \frac{.03453 - (1-.7)\ (.0826)}{.7}$$

$$= .01393$$

We now work out the estimate of variance which is given by (13.30). This expression is

$$v(\hat{\pi}_g) = \frac{(n'/n)\ (1-n'/n)}{(n-1)\ p^2}$$

On substituting different values, one gets

$$v(\hat{\pi}_g) = \frac{(.03453)\ (1-.03453)}{(782-1)\ (.7)^2} = .0000871$$

Then, we compute confidence limits following (2.8). These limits will, therefore, be given by

$$\hat{\pi}_g \pm 2\sqrt{v(\hat{\pi}_g)}$$

$$= .01393 \pm 2\sqrt{.0000871}$$

$$= .01393 \pm .01867$$

$$= -.00474,\ .03260$$

Leaving inadmissible values attained by lower limit, the confidence interval for the proportion of women in the age group of 18-44 years who had had an abortion during the past year, will thus be [0, .03260]. ■

The foregoing RR methods are used for estimation of proportion for stigmatized attributes which are essentially qualitative in nature. In practice, however, one may also have to deal with quantitative sensitive characters. For instance, one may be interested in estimating the average number of induced abortions among the females in a region, or the amount of tax evaded by a particular section of society. The unrelated question

model, described earlier, was suitably modified by Greenberg *et al.* (1971) to deal with *quantitative sensitive variables.*

13.7 ESTIMATION OF MEAN FOR QUANTITATIVE CHARACTERS

Consider a sensitive variable x which is supposed to be continuous with true density g(.) , and y is an unrelated nonsensitive continuous variable with true density h(.) which is roughly similar to that of x. For example, the variable x may be monthly expenditure on hard liquor, whereas y is the monthly expenditure on vegetables (or milk) in the household. The problem is to estimate μ_x, the population mean of x. First, we consider the situation where the population mean of nonsensitive variable y is unknown.

13.7.1 *Case I - μ_y Not Known*

Analogous to U-model of Greenberg *et al.* (1969), two simple random samples of n_1 and n_2 respondents are drawn, so that, $n_1 + n_2 = n$ is the required sample size. In this case also, two randomization devices R_1 and R_2 are needed. Each device consists of two questions- one regarding character x, and the other regarding character y. These two questions could be :

1. "What is the amount of money spent on hard liquor per month in the household?"
2. "How much money do you spend on vegetables in the household during a month ?"

These questions are represented with probabilities p_i and $(1-p_i)$ respectively in the device R_i, i = 1, 2, such that $p_1 \neq p_2$. The device R_1 is used for the respondents in the first sample and the device R_2 for the respondents in the second sample. Each respondent is required to draw one statement randomly, unobserved by the interviewer, and report the answer concerning the variable to which the selected statement points. The respondent's reply is simply a number, without specifying to which question the answer refers.

Let the randomized responses from the first sample be denoted by z_{1j}, j = 1, 2,..., n_1, and those from the second sample by z_{2j}, j = 1, 2, ..., n_2. Define

$$\left.\begin{aligned} \bar{z}_1 &= \frac{1}{n_1}\sum_{j=1}^{n_1} z_{1j}, \qquad s_{1z}^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(z_{1j} - \bar{z}_1)^2 \\ \bar{z}_2 &= \frac{1}{n_2}\sum_{j=1}^{n_2} z_{2j}, \qquad s_{2z}^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(z_{2j} - \bar{z}_2)^2 \end{aligned}\right] \tag{13.31}$$

The population variances σ_{1z}^2 and σ_{2z}^2 for the randomized responses z_1 and z_2 are obtained as

$$\sigma_{iz}^2 = p_i \sigma_x^2 + (1-p_i)\sigma_y^2 + p_i(1-p_i)(\mu_x - \mu_y)^2, \quad i = 1, 2, \tag{13.32}$$

where σ_x^2 and σ_y^2 are the population variances of x and y respectively.

Unbiased estimator of population mean μ_x :

$$\hat{\mu}_x = \frac{(1-p_2)\,\bar{z}_1 - (1-p_1)\bar{z}_2}{p_1 - p_2}, \qquad p_1 \neq p_2 \tag{13.33}$$

Variance of estimator $\hat{\mu}_x$:

$$V(\hat{\mu}_x) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2\,\frac{\sigma_{1z}^2}{n_1} + (1-p_1)^2\,\frac{\sigma_{2z}^2}{n_2}\right] \tag{13.34}$$

Estimator of variance $V(\hat{\mu}_x)$:

$$v(\hat{\mu}_x) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2\,\frac{s_{1z}^2}{n_1} + (1-p_1)^2\,\frac{s_{2z}^2}{n_2}\right] \tag{13.35}$$

The optimal design of a randomized response survey, involving quantitative variables, also requires the appropriate choice of p_1 and p_2, wise selection of a nonsensitive variable y, and efficient allocation of total sample size into n_1 and n_2. The rules for choosing optimal values of these parameters are given below :

1. A good working rule is to select one of the p_i, i = 1, 2, close to zero and other close to 1, such that $p_1 + p_2 = 1$.
2. The unrelated character y should be chosen, such that μ_y is close to μ_x and σ_y^2 is small. However, too small a choice of σ_y^2 , compared to σ_x^2, might meet with suspicion and affect likelihood of co-operation. Therefore, keeping both efficiency and protection of privacy in mind, a choice of σ_y^2 as close as possible to σ_x^2 should be attempted.
3. Analogous to (13.27), the $V(\hat{\mu}_x)$ is minimum when the total fixed sample size is allocated according to the relation

$$\frac{n_1}{n_2} = \frac{(1-p_2)\,\sigma_{1z}}{(1-p_1)\,\sigma_{2z}} \tag{13.36}$$

with σ_{iz}^2, i = 1, 2, defined in (13.32).

Example 13.8

A survey was carried out for estimating the average number of induced abortions per woman in a small town. Two independent samples of sizes $n_1 = 24$ and $n_2 = 22$ women were drawn from the population of 1200 women in the child bearing age. The two randomization devices consisted of decks of cards bearing statements :

1. "How many abortions did you have during your lifetime ?"
2. "How many children do you think a woman should have ?"

The statements (1) and (2) occur with probabilities .8 and .2 in deck I, and with probabilities .2 and .8 in deck II.

The deck I was used for the respondents in the first sample, and deck II was employed for the second sample. Each woman in the two samples was asked to draw a statement at random, unobserved by the interviewer, and report the answer concerning the selected statement without revealing as to which question the answer referred. The information thus collected is shown in table 13.3.

Table 13.3 The responses obtained from the sampled women

Sample I				Sample II			
Woman	Z_1	Woman	Z_1	Woman	Z_2	Woman	Z_2
1	2	13	2	1	2	12	3
2	1	14	3	2	2	13	2
3	2	15	1	3	1	14	2
4	1	16	1	4	3	15	1
5	2	17	2	5	1	16	2
6	1	18	1	6	2	17	2
7	2	19	2	7	2	18	3
8	3	20	1	8	2	19	2
9	0	21	2	9	3	20	2
10	2	22	3	10	1	21	2
11	1	23	1	11	2	22	2
12	3	24	3				
Total	20		22		21		23

Estimate the average number of abortions a woman had, and construct approximately 95% level confidence interval for it.

Solution

We have $n_1 = 24$, $n_2 = 22$, so that, $n = 24 + 22 = 46$. Also, $p_1 = .8$ and $p_2 = .2$. Let us first compute

$$\bar{z}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} z_{1j} = \frac{20+22}{24} = 1.75$$

$$\bar{z}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} z_{2j} = \frac{21+23}{22} = 2.00$$

$$s_{1z}^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1} (z_{1j} - \bar{z}_1)^2$$

$$= \frac{1}{24-1}[(2-1.75)^2 + (1-1.75)^2 + \ldots + (3-1.75)^2]$$

$$= .7174$$

$$s_{2z}^2 = \frac{1}{22-1}[(2-2)^2+(2-2)^2+\ldots+(2-2)^2]$$
$$= .3810$$

The estimate of average number of induced abortions is computed by using (13.33). That gives

$$\hat{\mu}_x = \frac{(1-p_2)\bar{z}_1-(1-p_1)\bar{z}_2}{p_1-p_2}$$
$$= \frac{(1-.2)(1.75)-(1-.8)(2.00)}{.8-.2}$$
$$= 1.667$$

We now compute the estimate of variance $V(\hat{\mu}_x)$ using (13.35). Thus,

$$v(\hat{\mu}_x) = \frac{1}{(p_1-p_2)^2}\left[(1-p_2)^2\frac{s_{1z}^2}{n_1}+(1-p_1)^2\frac{s_{2z}^2}{n_2}\right]$$

On substituting different values, it becomes

$$v(\hat{\mu}_x) = \frac{1}{(.8-.2)^2}\left[(1-.2)^2\left(\frac{.7174}{24}\right)+(1-.8)^2\left(\frac{.3810}{22}\right)\right]$$
$$= .05506$$

The confidence limits within which the population average is likely to fall are worked out as

$$\hat{\mu}_x \pm 2\sqrt{v(\hat{\mu}_x)}$$
$$= 1.667 \pm 2\sqrt{.05506}$$
$$= 1.667 \pm .469$$
$$= 1.198,\ 2.136$$

Thus, the average number of induced abortions per woman, in the population of 1200 women, is likely to fall in the interval [1.198, 2.136]. ■

13.7.2 *Case II - μ_y Known*

In case of binomial responses, we have seen that the estimates from the survey could be made more efficient when the value of π_y for the neutral question was known in advance. To apply this principle to quantitative responses, one should choose a nonsensitive variable for which the population mean is known in advance. For example, the question related to nonsensitive variable might, in some cases, ask for the number of persons living in a household where average household size is known from some kind of census or previous studies.

When μ_y is known, analogous to Greenberg *et al.* (1969) model, here also only one sample is required. As in case of U-model, there is a substantial reduction in variance $V(\hat{\mu}_x)$ and the results get quite simplified.

Unbiased estimator of mean μ_x :

$$\hat{\mu}_x = \frac{\bar{z} - (1-p)\mu_y}{p} \tag{13.37}$$

Variance of estimator $\hat{\mu}_x$:

$$V(\hat{\mu}_x) = \frac{\sigma_z^2}{np^2} \tag{13.38}$$

Estimator of variance $V(\hat{\mu}_x)$:

$$v(\hat{\mu}_x) = \frac{s_z^2}{np^2} \tag{13.39}$$

Terms s_z^2 and σ_z^2 above are defined in (13.31) and (13.32).

When estimating mean for a sensitive quantitative variable x through the technique discussed in this section, it is desirable that the density functions g(.) and h(.) for the sensitive variable x and the nonsensitive variable y respectively, are similar. This helps in protecting the confidentiality of the respondents, and thus results in enhanced co-operation from them. In some cases, one may be able to find a suitable nonsensitive variable y with a density function similar to that of x, whereas in other cases no such variable can possibly be found. In such situations, one may make use of the fact that any density function can be approximated by a frequency table with k class intervals, where class frequencies are proportional to the corresponding areas obtained from the density function under consideration. One can then prepare a set of cards where on cards numbering equal to the frequency for the i-th class interval is written a number equal to the mid-point of that class interval for i = 1, 2, ..., k. These numbers on the cards will be treated as the values of the nonsensitive variable y. Thus, the total number of cards in the set will be equal to the total frequency in the frequency table. This set of cards is then mixed with another set of cards bearing the statement "What is the value of x for you ?". The two sets of cards together constitute a deck of cards which can be used as the randomization device. Total number of cards in the set bearing the statement on the sensitive variable x, is determined in such a way that their proportion in the deck is equal to p. Each sample respondent, on being instructed by the investigator, would then randomly draw a card, unobserved by the investigator, from this deck. Depending on the statement on the card drawn, the respondent will report either the value of the sensitive variable x for himself or the number written on the card.

The mean value μ_y and the variance σ_y^2 for the numbers written on the cards in the set prepared from the frequency table (which are assumed to be the values taken by the nonsensitive variable y) can be easily obtained by the investigator.

The data for μ_y known case, when the unrelated characteristic with density function similar to that of the study variable is available, can be analyzed in a straightforward manner. Below we consider an example, where searching of such an unrelated

characteristic is difficult. Instead, a density function roughly close to that of the study variable is formulated by the investigator. This density function is then used in constructing the randomization device.

Example 13.9

The Income Tax Department is interested in estimating average annual income of advocates working in a court of law. These advocates number 400. The investigator is able to formulate rough distribution of the study variable which is given in table 13.4 below.

Table 13.4 Rough distribution of the income (in '000 rupees) of advocates

Total income	Mid-points	Advocates	Relative frequency	Cards prepared
48-52	50	10	.025	5
53-57	55	24	.060	12
58-62	60	60	.150	30
63-67	65	108	.270	54
68-72	70	88	.220	44
73-77	75	64	.160	32
78-82	80	24	.060	12
83-87	85	10	.025	5
88-92	90	6	.015	3
93-97	95	4	.010	2
98-102	100	2	.005	1
Total		400		200

A set of 200 cards was then prepared by writing on them, the income equal to the mid points of the class intervals in the above frequency table. Number of cards bearing a particular number were proportional to the relative frequency in table 13.4. For instance, 5 cards carried the statement "Give your response as rupees 50 thousand", whereas 12 cards carried the statement "Give your response as rupees 55 thousand", and so on. Besides these 200 cards, 600 cards bore the statement "What is your actual income ?". Each respondent in a WR simple random sample of 28 advocates, selected a card randomly from the randomization device consisting of 800 cards, and reported a number in accordance with the statement selected (near multiple of 5 thousand if actual income is to be reported). The responses, so obtained, are given in table 13.5.

Table 13.5 The responses obtained from the selected advocates

Advocate	Response	Advocate	Response	Advocate	Response	Advocate	Response
1	70	8	60	15	80	22	65
2	80	9	55	16	75	23	60
3	50	10	80	17	65	24	65
4	55	11	60	18	100	25	90
5	100	12	95	19	90	26	100
6	75	13	60	20	50	27	65
7	50	14	70	21	85	28	70

Estimate the average annual income of an advocate, and place confidence limits on it.

Solution

The statement of the problem provides n = 28 and p = 600/800 = .75. The average annual income of an advocate from the rough distribution is computed by using columns (2) and (3) of table 13.4. Thus,

$$\mu_y = \frac{1}{400}[(10)(50) + (24)(55) + \ldots + (2)(100)]$$

$$= 68.225$$

Before proceeding to obtain the required estimate, we work out the randomized response mean $\bar{z}$ for the sample units. Thus,

$$\bar{z} = \frac{1}{n}\sum_{j=1}^{n} z_j$$

$$= \frac{1}{28}(70 + 80 + \ldots + 70)$$

$$= 72.143$$

Estimate of average annual income of an advocate is obtained by using (13.37). The expression for this estimator is

$$\hat{\mu}_x = \frac{\bar{z} - (1-p)\mu_y}{p}$$

Substituting the values of $\bar{z}$, μ_y, and p, we get

$$\hat{\mu}_x = \frac{72.143 - (1-.75)\,68.225}{.75}$$

$$= 73.449$$

In order to obtain the estimate of variance, we calculate

$$s_z^2 = \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})^2$$

$$= \frac{1}{n-1} \left(\sum_{j=1}^{n} z_j^2 - n\bar{z}^2 \right)$$

$$= \frac{1}{28-1} [(70)^2 + (80)^2 + ... + (70)^2 - (28)(72.143)^2]$$

$$= \frac{1}{27} [152450 - (28)(72.143)^2]$$

$$= 248.920$$

From (13.39), the estimator of variance is given by

$$v(\hat{\mu}_x) = \frac{s_z^2}{np^2}$$

On substituting different values, one gets

$$v(\hat{\mu}_x) = \frac{248.920}{(28)(.75)^2} = 15.8044$$

The confidence limits follow from (2.8). These are obtained as

$$\hat{\mu}_x \pm 2 \sqrt{v(\hat{\mu}_x)}$$

$$= 73.449 \pm 2 \sqrt{15.8044}$$

$$= 73.449 \pm 7.951$$

$$= 65.498, 81.400$$

It indicates that if all the 400 advocates were interviewed and had they reported their annual income (in near multiple of 5 thousand) correctly, the average annual income of an advocate, almost surely, would have taken a value in the range of 65.498 to 81.400 thousand rupees. ■

For the sake of illustration, in example 13.9 we have kept the width of class interval as 5. This could be further reduced in an actual survey so as to obtain better co-operation from the respondents, and hence elicit more accurate information. Another point to be kept in mind, while building rough distribution similar to that of the study variable, is that the end points of the frequency distribution are so chosen that possibly no value pertaining to the study variable for the survey falls outside these end points.

Remark 13.1 Several other models for estimation of mean for sensitive quantitative characters are also available. Among them, the popular ones are due to Liu *et al.* (1975) and Eichhorn and Hayre (1983).

LET US DO

13.1 What is meant by nonresponse error in sample surveys ? Which are the two methods commonly used to control its effects ?

13.2 Describe briefly the Hansen and Hurwitz technique used to reduce the nonresponse bias in mail surveys.

13.3 The interest of a car manufacturing company is to estimate the average distance run by a newly purchased car before the first free service is availed. The company's recommendation is for 500 km. The frame consists of buyers who have purchased cars during the period of past 6 months to 1 year. A simple random WOR sample of 40 buyers from a population of 780 was drawn, and the questionnaire was mailed to the sampled respondents. Of these, 24 buyers responded, and from the remaining 16 nonrespondents 10 were selected for personal interview. The information collected from the buyers regarding the distance covered by the car before they availed first free service, is given below :

Through mail				Through interview	
Buyer	Distance (in km)	Buyer	Distance (in km)	Buyer	Distance (in km)
1	609	13	463	1	636
2	836	14	768	2	885
3	450	15	740	3	490
4	490	16	621	4	510
5	670	17	550	5	712
6	860	18	462	6	806
7	1007	19	880	7	532
8	630	20	1120	8	470
9	785	21	703	9	880
10	647	22	609	10	702
11	580	23	718		
12	520	24	780		

Estimate the average distance covered by a car before first free service was availed, and place confidence limits on it.

13.4 Explain Politz and Simmons procedure for reducing the nonresponse bias without resorting to call-backs.

13.5 A private company has engaged 540 workers. It is suspected that the company is not paying the workers their due wages. The investigator working on the case managed to have a complete list of all the workers along with their residential addresses. He selected a without replacement simple random sample of 30 workers. The residences of the workers included in the sample, were visited by the investigator during late evening hours. In case the respondent was found at home, he/she was asked to report the hourly wages (in rupees) paid to him/her. The respondent was also asked the number of times he/she was at home (t), at the time of interview, during the preceding five days (excluding Sunday). The information thus collected is given below. The mark "—" indicates that the respondent was not at home at the time and date of interview, and hence no information could be collected.

Worker	Wages (in Rs)	t	Worker	Wages (in Rs)	t	Worker	Wages (in Rs)	t
1	3.50	3	11	2.75	4	21	2.25	3
2	3.24	1	12	2.50	3	22	3.00	2
3	2.00	5	13	3.00	5	23	3.00	4
4	2.00	0	14	2.00	5	24	3.25	1
5	2.25	2	15	2.75	4	25	2.00	0
6	3.00	3	16	—	—	26	3.00	2
7	2.00	5	17	3.25	5	27	2.75	4
8	3.00	4	18	2.00	4	28	2.25	5
9	3.00	3	19	—	—	29	3.00	4
10	2.00	3	20	2.50	3	30	2.00	4

Estimate average hourly wages paid to the workers, and also work out confidence interval for it.

13.6 "Randomized response technique reduces/eliminates the evasive answer bias in surveys dealing with sensitive issues". Comment on the statement.

13.7 Describe briefly the Warner's randomized response technique for estimating the prevalence of sensitive attributes.

13.8 Warner's RR technique was employed in a survey to study the prevalence of smoking among undergraduate male students in a university. A student was defined as smoker if he had consumed at least two packets of cigarettes (40 cigarettes) over the one month period preceding the interview. A WR simple random sample of 250 students was drawn from a population of 2800 students. Each interviewee, included in the sample, was provided with a randomization

device consisting of a deck of cards. This randomization device carried two mutually exclusive statements :

1. "Have you consumed at least two packets of cigarettes during the last one month period?"
2. "Did you consume less than two packets of cigarettes, or did not smoke at all during the last one month period?"

with probabilities .75 and .25 respectively. Each interviewee was required to draw a card at random after reshuffling the deck, unobserved by the interviewer, and report "yes" if the statement on the selected card points to his actual status, and "no" otherwise. There were altogether 90 "yes" and 160 "no" answers. Estimate the proportion of smokers, and set confidence limits for it.

13.9 Suppose that for the situation considered in exercise 13.8, Mangat and Singh's two-stage procedure was used. Taking $n = 250$, $p = .75$, and $T = .3$, the number of "yes" answers recorded were 84. Estimate the proportion of smokers in the population, and place confidence limits on it.

13.10 "The confidence of the respondents in the anonymity provided by Warner's pioneer RR model might be further enhanced if one of the two questions confronted by the respondents referred to a nonsensitive attribute unrelated to the sensitive attribute under study". Explain the sampling strategy based on this principle.

13.11 There are 1560 workers in a factory. The management suspects that some workers might be involved in gambling. In order to estimate the proportion of such workers, two independent samples of 70 and 65 workers were selected through SRS with replacement. Each respondent in the first sample was provided with a randomization device carrying two statements:

1. "Are you a habitual gambler ?"
2. "Are you familiar with the rules of cricket ?"

with probabilities .85 and .15. The respondents in the second sample were provided with a similar randomization device representing the statements (1) and (2) with probabilities .15 and .85 respectively. Each respondent in the first and second samples was asked to choose randomly a statement from the RR device provided to him, and to say "yes" or "no" depending on whether or not the selected statement points to his actual status. Altogether 13 "yes" answers were reported by the interviewees in the first sample, whereas 9 "yes" answers were reported from the second sample. Estimate the proportion of habitual gamblers among the workers, and also work out confidence limits for it.

13.12 The income tax department suspects that some of the teachers of a university have income from alternative sources, viz., business, tuition work, and other type of part time employment. This income is, however, not reported by teachers when the income tax is deducted at source from their salary. To estimate the proportion of such teachers, a WR simple random sample of 160 teachers was drawn from

a population of 1400. Each respondent included in the sample was provided with a randomization device. The device consisted of 200 cards of which 160 bore the statement "Do you have income from sources other than the university salary ?", and the remaining 40 cards carried the statement "Were you born in the month of October ?" The proportion of teachers born in October is known to be .0917. On completion of survey, it was found that in all there were 32 "yes" and 128 "no" responses. Estimate the proportion in question, and construct confidence interval for it.

13.13 Discuss application of RR technique in obtaining data on a quantitative sensitive variable.

13.14 A survey was undertaken to estimate the amount of money spent on alcoholic drinks by the students of a university. With this objective in view, two independent WR simple random samples of sizes 25 and 30 students were selected from a total population of 1060 students. For interviewing the respondents, two sets of randomization device were used – set 1 for the respondents in the first sample and set 2 for the second sample. Set 1 consisted of cards carrying following two questions :

1. "How much have you spent on alcoholic drinks during the last 3 months ?"
2. "How much have you spent on purchasing clothes for yourself during the last 2 months ?"

in proportions .8 and .2 respectively. The cards in set 2 carried questions (1) and (2) in proportions .2 and .8 respectively. Each respondent was required to choose randomly one of these questions and report answer with respect to the question chosen. The responses obtained, in terms of rupees, are given below :

Sample 1 :	280	170	50	0	100	40	120	80	0	0	90	190	50
	0	115	150	60	80	30	0	0	100	0	110	70	
Sample 2 :	0	110	85	220	170	60	0	150	0	170	220	0	0
	100	65	90	190	240	0	170	0	50	100	0	40	100
	60	230	0	145									

Estimate the average expenditure incurred by the students on alcoholic drinks during the last 3 months, and place confidence limits on it.

13.15 It is desired to estimate the average frequency of traffic rule violations by the university employees. To accomplish the objective, 40 employees from the total population of 950, were selected through WR simple random sampling. Each employee in the sample was provided with a randomization device carrying two statements :

1. "How many times have you been issued tickets for violating traffic rules during last five years ?"
2. "How many visits did you make to university hospital for treatment during the last two months ?"

The statements (1) and (2) were represented with probabilities .7 and .3 respectively. The average number of visits of a university employee to the hospital were worked out from the hospital records, where the file number of each employee is entered before one is treated. It came out to be 2.3. Each respondent selected a statement randomly from the device provided and answered accordingly. The responses obtained are listed below :

Responses : 2 1 0 0 3 1 2 0 4 3 0 1 4 2 5 1 3
5 1 0 2 3 1 4 5 1 0 3 2 3 1 4 0 0
5 0 1 1 0 0

Estimate average number of traffic tickets issued to a university employee, and place approximately 95% confidence level limits on it.