LECTURE 21: **16.2, 16.3 Nonresponse Adjusted Weights**

A population is divided into respondent and nonrespondent subpopulations. Let $\bar{X}, \bar{X}_R, \bar{X}_{NR}$ denote respectively the entire population mean, the respondent subpopulation mean, and the nonrespondent subpopulation mean.

Definition The nonresponse bias is the difference between the respondent subpopulation mean and the entire population mean, that is, $B = \bar{X}_R - \bar{X}$.

Let N, N_R, and N_{NR} be the respective sizes of the entire population, the respondent subpopulation, and the nonrespondent subpopulation. Note that $N = N_R + N_{NR}$.

Denote by $\mathfrak{R}$ the respondent subpopulation, and by $\mathfrak{NR}$ the nonrespondent subpopulation, and let X_i be the measurement on the ith population element. Then the entire population mean can be written as

$$\bar{X} \stackrel{def}{=} \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\sum_{i\in\mathfrak{R}} X_i + \sum_{i\in\mathfrak{NR}} X_i}{N}$$

$$= \frac{\sum_{i\in\mathfrak{R}} X_i}{N_R}\frac{N_R}{N} + \frac{\sum_{i\in\mathfrak{NR}} X_i}{N_{NR}}\frac{N_{NR}}{N} = \bar{X}_R\,\gamma_R + \bar{X}_{NR}\,\gamma_{NR}$$

where

$$\gamma_R = \frac{N_R}{N} \text{ is the population response rate,}$$

and

$$\gamma_{NR} = \frac{N_{NR}}{N} \text{ is the population nonresponse rate}\,.$$

Note that

$$\gamma_R + \gamma_{NR} = \frac{N_R + N_{NR}}{N} = 1\,.$$

Therefore,

$$\bar{X} = \bar{X}_R(1-\gamma_{NR}) + \bar{X}_{NR}\gamma_{NR},$$

and thus, the nonresponse bias can be written as

$$B = \bar{X}_R - \bar{X} = \bar{X}_R - \left[\bar{X}_R(1-\gamma_{NR}) + \bar{X}_{NR}\gamma_{NR}\right]$$

$$= \bar{X}_R - \bar{X}_R + \bar{X}_R\gamma_{NR} - \bar{X}_{NR}\gamma_{NR} = \bar{X}_R\gamma_{NR} - \bar{X}_{NR}\gamma_{NR} = \gamma_{NR}(\bar{X}_R - \bar{X}_{NR})\,.$$

The estimate of the nonresponse bias is

$$\hat{B} = \hat{\gamma}_{NR}(\bar{x}_R - \bar{x}_{NR})\,.$$

Example Suppose $n = 5,000$ individuals are randomly chosen for a survey, and 3,200 persons responded. The response rate is $3,200/5,000 = 0.64$, and hence the nonresponse rate is $\hat{\gamma}_{NR} = 0.36$. Suppose the sample mean income of the respondent subpopulation is $\bar{x}_R = \$60,000$, and suppose further that according to administrative records, the mean income for the nonrespondent subpopulation is estimated to be \$51,000. The estimated nonresponse bias in this case is $\hat{B} = (0.36)(\$60,000 - \$51,000) = \$3,240$. Thus, it is estimated that the mean income of the entire population will be overestimated by \$3,240 as a result of nonresponse.

A remedy for nonresponse bias is to introduce nonresponse adjusted weights. To the ith individual in the sample, assign a probability π_{Ri} that this individual will respond. This probability is termed the response propensity for the ith individual. Thus, the probability that the ith individual in the population is selected for the sample and responds can be computed as

$$\mathbb{P}(i \text{ selected and responds}) = \mathbb{P}(i \text{ selected})\, \mathbb{P}(i \text{ responds} \,|\, i \text{ selected}) = p_i\, \pi_{Ri}$$

where p_i is the probability that the ith individual is selected.

Recall that the design weight is reciprocal to the selection probability. Thus the design weight for the ith individual is $w_i = 1/p_i$.

DEFINITION The nonresponse adjusted weight for the ith individual is

$$w_{NRi} = \frac{1}{\pi_{Ri}}.$$

DEFINITION The nonresponse adjusted design weight for the ith individual is the product of the design weight and the nonresponse adjusted weight, $w_i\, w_{NRi}$. This overall weight compensates for both the unequal selection probabilities and nonresponse.

REMARK Not adjusting for nonresponse is equivalent to setting $\pi_{Ri} = \hat{\gamma}_R$, the sample response rate. This is a very crude estimate of the response propensity. For the purposes of estimating the population mean $\bar{X}$, we may think of using $\pi_{Ri} = \hat{\gamma}_R$ as essentially assigning $\bar{x}_R$ to every nonrespondent. It is often possible to come up with much better estimate if some data on nonrespondents are available.

The simplest way to compute the response propensity is to use the weight class adjustment. The sample is divided into groups based on variables that are known for both respondents and nonrespondents and are thought to be related to response propensity. For example, response rates often vary by age and gender. If all ages and genders of nonrespondents are known, then all individuals in the sample can be classified into age by gender groups, called weighting classes.The response rate for the weighting class is taken as the response propensity for each individual in the class.

The response rate is computed according to the following formula. Suppose that in a particular class K there are n_K individuals, $n_{R,K}$ of whom are respondents. The <u>weighted response rate</u> for the class K is

$$\hat{\gamma}_{R,K} = \frac{\sum_{i=1}^{n_{R,K}} w_i}{\sum_{i=1}^{n_K} w_i}$$

where w_i is the design weight for the ith individual.

Example Suppose the sampled individuals are children. The ages are categorized as 0-2 years, 3-10 years, and 11+ years. Suppose the sum of the respondent design weights for the category "males 11 years or older" is

$$\sum_{i=1}^{n_{R,K}} w_i = 99,465.$$

This number can be interpreted as the number of males 11 years or older in the respondent subpopulation that are represented by this category of respondents in the sample.

The sum of the design weights for all (respondents and nonrespondents) males 11 years or older is

$$\sum_{i=1}^{n_K} w_i = 178,462.$$

This number can be interpreted as the total number of males 11 years or older in the entire population.

Thus, the estimate of the <u>weighted response rate</u> for the class of males 11 years or older is $\hat{\gamma}_{R,K} = 99,465/178,462 = 0.5573$, which can be interpreted as the estimated response propensity for males 11 years or older.

Finally, the nonresponse adjusted weight for each responded male 11 years or older is $w_{NRi} = 1/\hat{\gamma}_{R,K} = 1/0.5573 = 1.7942$. The nonresponse adjusted design weight for the ith responded male 11 year or older is $1.7942\, w_i$, $i = 1, \ldots, n_{R,K}$, which is interpreted as the number of individuals this individual represents in the entire population.