LECTURE 20: **Chapter 16 Sample Survey Weights**

Weights are used in sample surveys to (i) adjust for different selection probabilities, and (ii) to adjust for nonresponse.

(i) Data should usually be weighted if the sample design does not give each individual an equal chance of being selected (that is, if the sample is not an SRS). For instance, when households have equal selection probabilities but one person is interviewed from within each household, this gives people from large households a smaller chance of being interviewed. This can be accounted for by using survey weights. Similarly, households with more than one telephone line have a greater chance of being selected in a random digit dialing sample, and weights can adjust for this.

DEFINITION Weights that adjust for different selection probabilities are called design weights (or base weights).

(ii) Non-response is only a problem if the non-respondents are a non-random sample of the total sample. Unfortunately, this seems almost always to be the case. In household surveys, for instance, there is lots of evidence that non-respondents are younger than respondents, and that men are harder to persuade to take part than women. Response rates also tend to be lower than average in cities and in deprived areas.

The result of these patterns is that the achieved samples for surveys often don't reflect very well the population they are meant to represent. Surveys typically overrepresent women, and those over the age of 30. And they often underrepresent those living in cities and deprived areas.

Rather than accept a poor match between the sample and the population, it is now common for survey data sets to use weights to bring the two more closely into line.

DEFINITION Weights that adjust for nonresponse are called nonresponse adjusted weights (or nonresponse weights).

(i) Design weights are used in calculating estimates for various sampling schemes. The sampling weights are the reciprocals of the selection probabilities, so that an estimator of the population total is of the form

$$x'_w = \sum_{i=1}^{n} w_i \, x_i$$

where x_i is the observed value of the characteristic, n is the sample size, and w_i is the design weight.

Example (a) For a simple random sample, the probability that element i is chosen for the sample of size n is $p_i = n/N$, and the weight is reciprocal to p_i,

$$w_i = 1/p_i = N/n, \text{ and } x' = \sum_{i=1}^{n} \frac{N}{n} x_i \,.$$

(b) Reminder: A stratified random sampling is a sampling method in which the sampling frame is partitioned into strata, and then a random sample is drawn independently within each stratum.

For a stratified random sample,

$$w_{i,h} = N_h/n_h, \text{ and } x'_{str} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{N_h}{n_h} x_{h,i}$$

where L is the number of strata, N_h is the size of stratum h, n_h is the sample size of elements from stratum h, and $x_{h,i}$ is the sample characteristic of element i from stratum h Note that the probability to select element i within stratum h is n_h/N_h, which is reciprocal to the weight $w_{i,h}$.

(c) Reminder: A simple one-stage cluster sampling is a sampling in which clusters are chosen by drawing an SRS, and within each sampled cluster, all elements are selected.

For simple one-stage cluster sample,

$$w_{ij} = \frac{M}{m}, \text{ and } x'_{clu} = \sum_{i=1}^{m} \sum_{j=1}^{N_i} \frac{M}{m} x_{ij}$$

where M is the number of clusters in the population, m is the number of clusters in the sample, N_i is the size of the ith sampled cluster, and x_{ij} is the measurement for the jth element in the ith sampled cluster. The probability to choose element j within cluster i is m/M, which is reciprocal to the weight w_{ij}.

(d) Reminder: A simple two-stage cluster sampling is a sampling in which clusters are chosen at the first stage by drawing an SRS, and at the second stage, an SRS is drawn within each selected cluster.

For a simple two-stage cluster sampling,

$$w_{ij} = \frac{N}{n}, \text{ and } x'_{clu} = \sum_{i=1}^{m} \sum_{j=1}^{\bar{n}} \frac{N}{n} x_{ij}$$

where $\bar{n} = n/m$ is the number of sampled elements in each cluster (assumed constant). The probability of selecting element j within cluster i is

$$\frac{m}{M}\frac{\bar{n}}{\bar{N}} = \frac{m}{M}\frac{n/m}{N/M} = \frac{n}{N},$$

and is reciprocal to w_{ij}.

(e) Reminder: A simple two-stage cluster sampling when clusters have different sizes is defined as follows: m clusters are chosen from among M clusters by drawing an SRS. Then within each sampled cluster i of size N_i, an SRS of n_i elements is taken, where n_i is such that n_i/N_i is as close as possible to the predetermined second-stage sampling fraction f_2.

For a simple two-stage cluster sample when cluster have different sizes,

$$w_{ij} = \frac{M}{m}\frac{N_i}{n_i}, \text{ and } x'_{clu} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{M}{m}\frac{N_i}{n_i}\,x_{ij}\,.$$

The probability of choosing element j within cluster i is $(m/M)(n_i/N_i)$, which is reciprocal to the weight w_{ij}.

(f) For a cluster sample with probabilities proportional to cluster size (the second-stage sampling can be of any type),

$$w_i = \frac{1}{\pi_i}, \text{ and } x'_{hte} = \sum_{i=1}^{m}\frac{1}{\pi_i}\,x_i$$

where π_i is the probability that cluster i appears in the sample, and x_i is the total of the ith sampled cluster.

$$w_i = \frac{1}{m\pi'_i}, \text{ and } x'_{hh} = \sum_{i=1}^{m}\frac{1}{m\pi'_i}\,x_i$$

where π'_i is the probability of drawing cluster i.

Remark 1 The weight w_i can be interpreted as the number of population elements that are represented by the sampled element i. The weights should be equal to at least 1 since each sampled element represents at least itself.

Remark 2 The sum of the design weights should be equal to population size N, that is, it must be true that $\sum_{i=1}^{n} w_i = N$.