Introduction to unequal-probability sampling; PPS sampling with replacement

Unequal probability sampling with replacement: As we saw in the jobs example, there are situations in which it is desirable to have unequal probabilities of selecting elements into the sample. In section 3.3 of the text, it is stated that for a population with elements $\{u_1, u_2, ..., u_N\}$, we might choose to sample elements **with replacement** with respective selection probabilities of $\{\delta_1, \delta_2, ..., \delta_N\}$. In that situation the estimator

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\delta_i}$$

is an unbiased estimator for τ, and an unbiased estimator of $V(\widehat{\tau})$ is given by:

$$\widehat{V}(\widehat{\tau}) = \frac{1}{n} \left[\sum_{i=1}^{n} \left(\frac{y_i}{\delta_i} - \widehat{\tau} \right)^2 / (n-1) \right] = \frac{1}{n} \left[\sum_{i=1}^{n} (\widehat{\tau}_i - \widehat{\tau})^2 / (n-1) \right].$$

Note that this variance estimator is of the form s^2/n, and is unbiased for $V(\widehat{\tau})$ because the with replacement sampling scheme yields a sample of n separate **independent** estimates of τ, namely $\widehat{\tau}_i = y_i/\delta_i$. This estimator and its' variance estimator are due to Hansen and Hurwitz (1943), and thus the estimator is called the **Hansen-Hurwitz** estimator. Sampling with replacement yields estimators whose theoretical properties are easy to understand, but sampling with replacement is often inefficient and can be impractical in some situations. Later we will discuss general approaches to unequal-probability sampling that have sampling without replacement, using estimators developed by Horvitz and Thompson (1952).

PPS sampling with replacement: In cluster sampling, it is often useful to use unequal-probability sampling with replacement with probabilities proportional to size (PPS). In this case, $\delta_i = m_i/M$, and our estimator of the total is:

$$\widehat{\tau}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\delta_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{(m_i/M)} = \frac{M}{n} \sum_{i=1}^{n} \frac{y_i}{m_i} = \frac{M}{n} \sum_{i=1}^{n} \overline{y}_i,$$

where $\overline{y}_i = y_i/m_i$ is the average of the observations in cluster i. To obtain an estimator of the mean we can just divide by M, giving

$$\widehat{\mu}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \overline{y}_i.$$

The estimated variances of these two estimators are then:

$$\widehat{V}(\widehat{\tau}_{pps}) = \frac{1}{n} \left[\sum_{i=1}^{n} \left(\widehat{\tau}_i - \widehat{\tau}_{pps} \right)^2 / (n-1) \right] = \frac{M^2}{n} \left[\sum_{i=1}^{n} \left(\overline{y}_i - \widehat{\mu}_{pps} \right)^2 / (n-1) \right]$$

and

$$\widehat{V}(\widehat{\mu}_{pps}) = \frac{1}{n} \left[\sum_{i=1}^{n} \left(\widehat{\mu}_i - \widehat{\mu}_{pps} \right)^2 / (n-1) \right] = \frac{1}{n} \left[\sum_{i=1}^{n} \left(\overline{y}_i - \widehat{\mu}_{pps} \right)^2 / (n-1) \right].$$

See the examples from lecture and in the SAS code on the web.

References:

Hansen, M.H. and Hurwitz, W.N. (1943) On the theory of sampling from a finite population. *Annals of Mathematical Statistics* 14: 333-362.

Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47: 663-685.