Lecture 9

Unequal Probability Sampling

For a SRS, each sampling unit has the same probability of being included in the sample.

The different inclusion probabilities depend on either:

1. the type of sampling procedure or

2. the probabilities may be imposed by the researcher to obtain better estimates

In either case, the unequal inclusion probabilities must be taken into account when deriving estimators of population parameters.

Example. Suppose you are interested in modeling the total amount of natural gas used across the country. There are several large distributors (e.g. Northwestern Energy) that account for a large share of the total usage; however, there are many more small distributors (e.g. municipalities) that account for a small share of the total gas used.

Hansen-Hurwitz Estimation

One method of estimating $\bar{y}_U$ and t when the probabilities of selection sampling units are not equal is

Situation:

1. A sample of size n is to be selected

2. sampling is

3. the probability of selecting the i^{th} unit equals p_i on each selection of a sampling unit.

Sampling with replacement is less precise than sampling without replacement (i.e. the variance of the estimator will be larger). However, when the sampling fraction $f = n/N$ is small, the probability that any unit will appear twice in the sample is also small.

Thus, the loss of some precision using sampling with replacement can offset the complexity of having to determine the inclusion probabilities when sampling is done without replacement.

The Hansen-Hurwitz estimator of t is:

$$\hat{t}_{hh} = \frac{1}{n} \sum_{i \in \mathcal{S}} \frac{y_i}{p_i} \tag{1}$$

where p_i is the probability of selection for unit i and $\mathcal{S}$ are the units in the sample (including repeats).

The estimated variance of $\hat{t}_{hh}$ is

$$\hat{V}(\hat{t}_{hh}) = \frac{1}{n} \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(\frac{t_i}{p_i} - \hat{t}_{hh} \right)^2 . \tag{2}$$

Because sampling is taken with replacement, each unit may be sampled more than once. Let Q_i denote the number of times unit i is sampled. Then the point estimate and variance formulas can be written as

$$\hat{t}_{hh} = \frac{1}{n} \sum_{i=1}^{N} Q_i \frac{t_i}{p_i} \tag{3}$$

$$\hat{V}(\hat{t}_{hh}) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{N} Q_i \left(\frac{t_i}{p_i} - \hat{t}_{hh} \right)^2 . \tag{4}$$

Note this assume the sampling fraction is small and that the fpc can be ignored.

The estimates for the population mean can be obtained by:

$$\hat{\bar{y}}_{hh} = \frac{1}{N} \hat{t}_{hh} \qquad\qquad \hat{V}(\hat{\bar{y}}_{hh}) = \frac{1}{N^2} \hat{V}(\hat{t}_{hh})$$

Example. Suppose $N = 5$ and we are taking $n = 2$ samples with the probability of selection below.

ID	p_i	y_i
1	1/12	9
2	2/12	22
3	2/12	19
4	4/12	42
5	3/12	28
		120

Repeat this process 4 times to illustrate how unequal probability sampling and estimation works.

```
vals <- c(9,22,19,42,28)
replicate(4,sample(1:5, 2, prob= c(1/12,2/12,2/12,4/12,3/12)))
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    5    3    5    4
## [2,]    4    4    4    5
```

Sample #	ID1	ID2	(y_1/p_1)	(y_2/p_2)	$\hat{t}_{hh}$
1					
2					
3					
4					

So how do we think about devising the selection probabilities: p_i?

The ideal case for Hansen-Hurwitz estimation occurs when each selection probability p_i

Therefore, in practice, if we believe the y_i values are nearly proportional to some known variable (like sample unit size),

Confidence intervals follow the usual prescription given the standard errors calculated above.

Horvitz-Thompson Estimation

A second method of estimating $\bar{y}_U$ and t when the probability of selecting sampling units is not equal is

Now a sample can be taken with or without replacement.

The first order inclusion probability π_i is the probability unit i will be included by a sampling design.

The second-order inclusion probability π_{ij} is the probability that

When the goal is to estimate the population total t or the mean $\bar{y}_U$, and the $\pi_i's$ are known, the Horvitz-Thompson estimators follow as:

$$\hat{t}_{ht} = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i} \qquad \text{and} \qquad \hat{\bar{y}}_{U,ht} = \frac{1}{N} \sum_{i=1}^{\nu} \frac{y_i}{\pi_i} \tag{5}$$

where ν is the effective sample size.

The effective sample size is the number of distinct units in the sample. When sampling without replacement, $\nu = n$. When sampling with replacement, $\nu \leq n$.

Because the summation is over the ν distinct units in the sample,

The variance is estimated as :

$$\hat{V}(\hat{t}_{ht}) = \sum_{i=1}^{\nu} \left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i} \right) y_i^2 + 2 \sum_{i=1}^{\nu} \sum_{j>i}^{\nu} \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right) y_i y_j \tag{6}$$

Sampling with Probabilities Proportional to Size (PPS)

Suppose that n sampling units are selected *with replacement* with selection probabilities proportional to the sizes of the units from a finite population of N units.

Let p_i be the probability that unit i is selected during the sampling with replacement process. Then $p_i = \frac{M_i}{M_T}$ where M_i is the size of unit i and $M_T = \sum_{i=1}^{N} M_i =$ the total size of the population of N units.

If sampling is done without replacement, determining inclusion probabilities is very complex. Thus, we will focus on sampling with replacement.

When sampling with replacement, the first order inclusion probability:

To find the second order inclusion probability, we use the principle of inclusion/exclusion. That is, for two events A and B, the probability that both A and B occur is:

Cluster Sampling with Unequal Cluster Sizes

Suppose the N cluster sizes $M_1, M_2, \ldots, M_N$ are not all equal and that a one-stage cluster sample of n primary sampling units ($PSUs$) is taken with the goal of estimating t and $\bar{y}_U$.

Let M_i and t_i be the sizes and totals of the n sampled PSUs. Let $m = \sum_{i=1}^{n} M_i$ be the total number of SSUs in the sample.

Primary Sampling Units Selected with PPS}

Suppose that the PSUs are selected with replacement with draw-by-draw selection probabilities p_i proportional to the sizes of the PSUs, $p_i = M_i/M_0$.

Then either Horvitz-Thompson or the Hansen -Hurwitz estimators can be used.