Wednesday, Nov 6

The Hájek Estimator

Recall that if the number of elements in the population is unknown we can use the Horvitz-Thompson estimator

$$\hat{\tau} = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i}$$

to estimate μ with the estimator

$$\hat{\mu} = \frac{\sum_{i \in \mathcal{S}} y_i/\pi_i}{\sum_{i \in \mathcal{S}} 1/\pi_i},$$

because $\sum_{i \in \mathcal{S}} 1/\pi_i$ is an estimator of the number of elements in the population. This is sometimes called the Hájek estimator.

Now suppose the number of elements in the population is *known*. This gives us *two* estimators of μ. We have the Horvitz-Thompson and Hájek estimators of μ which are

$$\hat{\mu} = \frac{1}{N} \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} \quad \text{and} \quad \hat{\mu} = \frac{\sum_{i \in \mathcal{S}} y_i/\pi_i}{\sum_{i \in \mathcal{S}} 1/\pi_i},$$

respectively. Here N denotes the number of elements in the population. We also have *two* estimators of τ. We have the Horvitz-Thompson and Hájek estimators of τ which are

$$\hat{\tau} = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} \quad \text{and} \quad \hat{\tau} = N \frac{\sum_{i \in \mathcal{S}} y_i/\pi_i}{\sum_{i \in \mathcal{S}} 1/\pi_i},$$

respectively. Which might we prefer? The Hájek estimator tends to have smaller variance than the Horvitz-Thompson estimator under the following conditions.

1. The population is relatively homogeneous meaning that σ^2 is small.
2. The sample size is random.
3. The inclusion probabilities are weakly or negatively correlated with the target variable.

Note also that some estimators require knowing N (the number of elements in the population).

Generalized Horvitz-Thompson Estimators

We can build estimators that use auxiliary variables (e.g., ratio estimators, regression estimators, and estimators used with stratification or post-stratification) using Horvitz-Thompson estimators. These are examples of what are sometimes called *generalized* Horvitz-Thompson estimators.

Example: Recall that the ratio estimator of τ_y for a simple random sampling design is

$$\hat{\tau}_y = \tau_x \bar{y}/\bar{x}.$$

Here $\bar{y}$ and $\bar{x}$ are estimators of μ_y and μ_x, respectively. But for an arbitrary sampling design the Horvitz-Thompson estimators of μ_y and μ_x are

$$\hat{\mu}_y = \frac{1}{N} \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} \quad \text{and} \quad \hat{\mu}_x = \frac{1}{N} \sum_{i \in \mathcal{S}} \frac{x_i}{\pi_i}.$$

Substituting these estimators into the ratio estimator gives us

$$\hat{\tau}_y = \tau_x \frac{\sum_{i \in \mathcal{S}} y_i/\pi_i}{\sum_{i \in \mathcal{S}} x_i/\pi_i}.$$

Example: Recall that the regression estimator of τ_y for a simple random sampling design is

$$\hat{\tau}_y = N\bar{y} + b(\tau_x - N\bar{x}).$$

Substituting the Horvitz-Thompson estimators for $\bar{y}$ and $\bar{x}$ as we did for the ratio estimator gives the regression estimator for an arbitrary design,

$$\hat{\tau}_y = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} + b\left(\tau_x - \sum_{i \in \mathcal{S}} \frac{x_i}{\pi_i}\right).$$

The slope of the line (b) that relates the target variable to the auxiliary variable is computed differently to take into account the inclusion probabilities (the details have been omitted here).

Example: For a stratified random sampling design, with simple random sampling within each stratum, the estimator of τ can be written as

$$\hat{\tau} = N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L.$$

The Horvitz-Thompson estimator of μ_j can be written as

$$\hat{\mu}_j = \frac{1}{N_j} \sum_{i \in \mathcal{S}_j} \frac{y_i}{\pi_i}.$$

So then the estimator of τ for any sampling design involving strata is

$$\hat{\tau} = \sum_{i \in \mathcal{S}_1} \frac{y_i}{\pi_i} + \sum_{i \in \mathcal{S}_2} \frac{y_i}{\pi_i} + \cdots + \sum_{i \in \mathcal{S}_L} \frac{y_i}{\pi_i}.$$

Second-Order Inclusion Probabilities

The variance of the Horvitz-Thompson estimator depends on **second-order** (or "joint") inclusion probabilities (π_{ij}) as opposed to the **first-order** inclusion probabilities (π_i). The second-order inclusion probability π_{ij} is the probability that $\mathcal{E}_i$ *and* $\mathcal{E}_j$ will *both* be included in the sample.

Note: If $i = j$ then $\pi_{ij} = \pi_i$ (i.e., the first-order inclusion probability).

Example: Suppose we have the population $\mathcal{P} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5\}$ and the following sampling design.

$$\begin{aligned}
\mathcal{S}_1 &= \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_4\},\ P(\mathcal{S}_1) = 0.1 \\
\mathcal{S}_2 &= \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_5\},\ P(\mathcal{S}_2) = 0.2 \\
\mathcal{S}_3 &= \{\mathcal{E}_1, \mathcal{E}_3, \mathcal{E}_4\},\ P(\mathcal{S}_3) = 0.3 \\
\mathcal{S}_4 &= \{\mathcal{E}_1, \mathcal{E}_3, \mathcal{E}_5\},\ P(\mathcal{S}_4) = 0.2 \\
\mathcal{S}_5 &= \{\mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\},\ P(\mathcal{S}_5) = 0.1 \\
\mathcal{S}_6 &= \{\mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_5\},\ P(\mathcal{S}_6) = 0.1
\end{aligned}$$

Based on this design we can determine the second-order inclusion probabilities for all pairs of elements.

For specific designs there is often a "shortcut" formula for the second-order inclusion probabilities.

Example: Consider the population $\mathcal{P} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\}$ and a sampling design using sampling *with replacement* with selection probabilities given below and a sample size of $n = 3$.

Elements	Second-Order Inclusion Probability
$\mathcal{E}_1, \mathcal{E}_2$	
$\mathcal{E}_1, \mathcal{E}_3$	
$\mathcal{E}_1, \mathcal{E}_4$	
$\mathcal{E}_1, \mathcal{E}_5$	
$\mathcal{E}_2, \mathcal{E}_3$	
$\mathcal{E}_2, \mathcal{E}_4$	
$\mathcal{E}_2, \mathcal{E}_5$	
$\mathcal{E}_3, \mathcal{E}_4$	
$\mathcal{E}_3, \mathcal{E}_5$	
$\mathcal{E}_4, \mathcal{E}_5$	

Element	δ_i	π_i
$\mathcal{E}_1$	0.2	0.488
$\mathcal{E}_2$	0.1	0.271
$\mathcal{E}_3$	0.2	0.488
$\mathcal{E}_4$	0.5	0.875

What are the second-order inclusion probabilities? Earlier we learned that for this design the (first-order) inclusion probabilities are

$$\pi_i = 1 - (1 - \delta_i)^n.$$

It can also be shown that the second-order inclusion probabilities are

$$\pi_{ij} = \pi_i + \pi_j - [1 - (1 - \delta_i - \delta_j)^n],$$

provided that i and j are different elements.[1]

Elements	Second-Order Inclusion Probability
$\mathcal{E}_1, \mathcal{E}_2$	0.102
$\mathcal{E}_1, \mathcal{E}_3$	0.192
$\mathcal{E}_1, \mathcal{E}_4$	0.390
$\mathcal{E}_2, \mathcal{E}_3$	0.102
$\mathcal{E}_2, \mathcal{E}_4$	0.210
$\mathcal{E}_3, \mathcal{E}_4$	0.390

Example: For simple random sampling, $\pi_i = n/N$ and

$$\pi_{ij} = \frac{n}{N} \times \frac{n-1}{N-1},$$

assuming elements i and j are distinct.

[1]The proof is fun if you know a bit of probability theory. Note that

$$\pi_{ij} = P(\mathcal{E}_i \in \mathcal{S} \cap \mathcal{E}_j \in \mathcal{S}) = P(\mathcal{E}_i \in \mathcal{S}) + P(\mathcal{E}_2 \in \mathcal{S}) - P(\mathcal{E}_i \in \mathcal{S} \cup \mathcal{E}_j \in \mathcal{S}).$$

Now $P(\mathcal{E}_i \in \mathcal{S}) = \pi_i$ and $P(\mathcal{E}_j \in \mathcal{S}) = \pi_j$, and

$$P(\mathcal{E}_i \in \mathcal{S} \cup \mathcal{E}_j \in \mathcal{S}) = 1 - P(\mathcal{E}_i \notin \mathcal{S} \cap \mathcal{E}_j \notin \mathcal{S}).$$

The probability that either element is selected on *a single draw* is $\delta_i + \delta_j$ since those events are disjoint — i.e., you cannot select both elements on a single draw. So the probability that *neither* element is selected on a single draw is $1 - (\delta_i + \delta_j)$. Then probability that neither $\mathcal{E}_i$ or $\mathcal{E}_j$ is selected on all n draws is $(1 - \delta_i - \delta_j)^n$, and thus

$$P(\mathcal{E}_i \notin \mathcal{S} \cap \mathcal{E}_j \notin \mathcal{S}) = (1 - \delta_i - \delta_j)^n.$$

Example: For stratified random sampling, if element i is in stratum k then $\pi_i = n_k/N_k$ and

$$\pi_{ij} = \begin{cases} n_k/N_k \times (n_k - 1)/(N_k - 1), & \text{if elements } i \text{ and } j \text{ are both from stratum } k, \\ \pi_i \pi_j, & \text{if elements } i \text{ and } j \text{ are not from the same stratum.} \end{cases}$$

Example: For one-stage cluster sampling with simple random sampling of clusters, $\pi_i = n/N$ and

$$\pi_{ij} = \begin{cases} n/N, & \text{if elements } i \text{ and } j \text{ are in the same cluster,} \\ n/N \times (n-1)/(N-1), & \text{if elements } i \text{ and } j \text{ are in different clusters.} \end{cases}$$

Variance of the Horvitz-Thompson Estimator

The variance of the Horvitz-Thompson estimator of τ is a function of the first-order (π_i) and second-order (π_{ij}) inclusion probabilities. It can be written as

$$V(\hat{\tau}) = \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j},$$

where it should be understood that if $i = j$ then $\pi_{ii} = \pi_i$ (i.e., the second-order inclusion probability of an element with itself is its first-order inclusion probability). There are several other ways to write this expression.

We cannot compute this variance based on only a sample of elements, but it can be estimated in various ways. One such estimator is

$$\hat{V}(\hat{\tau}) = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} y_i y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j}.$$

For certain designs (e.g., simple random sampling, stratified random sampling, cluster sampling) the (estimated) variance takes a more familiar form.

The important point to note here is that the variance of the Horvitz-Thompson estimator depends on the design through the first-order and second-order inclusion probabilities.

Specification of Inclusion Probabilities

There are two ways we specify the inclusion probabilities for a sampling design.

1. **Indirect Approach**: We specify a *sampling design* which determines the inclusion probabilities. This is usually fairly easy as we just need to derive the inclusion probabilities for the sampled elements. This is done implicitly in some designs we have already discussed such as simple random sampling, stratified random sampling, and cluster sampling. But it can also be done in other designs that we will discuss soon including *Poisson sampling*, *line-intercept sampling*, *fixed area plot sampling*, and *adaptive cluster sampling*.
2. **Direct Approach**: We specify the *inclusion probabilities* (usually just first-order) and then try to come up with a sampling design that has those inclusion probabilities. Finding a sampling design that results in desired inclusion probabilities can be quite challenging for two reasons. One is that it is computationally difficult except in very simple cases. The other is that the solution will not generally be unique — there will be multiple sampling designs that result in the *same* inclusion probabilities.

Consider the direct approach. If y_i is approximately proportional to an auxiliary variable x_i, a relatively low variance estimator of τ may be obtained if the inclusion probabilities are specified as

$$\pi_i = \frac{n x_i}{\tau_x}.$$

Here is a population of $N = 5$ elements. The inclusion probabilities are based on using the auxiliary variable a sample of $n = 3$ elements.

Element	x_i	π_i
$\mathcal{E}_1$	3	0.9
$\mathcal{E}_2$	2	0.6
$\mathcal{E}_3$	1	0.3
$\mathcal{E}_4$	1	0.3
$\mathcal{E}_5$	3	0.9

Note that $\tau_x = \sum_{i=1}^{5} x_i = 10$. Can we find a sampling design that has these inclusion probabilities? Yes, with some help from a computer, but the solution is not unique. There is more than one sampling design with the same inclusion probabilities.

	Sample Probability		
Sample	Design A	Design B	Design C
$\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$	0.03690	0.04090	0.02044
$\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_4$	0.02254	0.02486	0.00865
$\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_5$	0.46172	0.44113	0.49593
$\mathcal{E}_1$, $\mathcal{E}_3$, $\mathcal{E}_4$	0.03468	0.02704	0.05271
$\mathcal{E}_1$, $\mathcal{E}_3$, $\mathcal{E}_5$	0.17496	0.14778	0.15968
$\mathcal{E}_1$, $\mathcal{E}_4$, $\mathcal{E}_5$	0.16919	0.21829	0.16258
$\mathcal{E}_2$, $\mathcal{E}_3$, $\mathcal{E}_4$	0.00588	0.00720	0.01819
$\mathcal{E}_2$, $\mathcal{E}_3$, $\mathcal{E}_5$	0.02641	0.07019	0.02394
$\mathcal{E}_2$, $\mathcal{E}_4$, $\mathcal{E}_5$	0.04654	0.01572	0.03284
$\mathcal{E}_3$, $\mathcal{E}_4$, $\mathcal{E}_5$	0.02117	0.00689	0.02503

Typically we find that we can identify many sampling designs with the same first-order inclusion probabilities, but different second-order inclusion probabilities, so potentially different variances of the Horvitz-Thompson estimator. One strategy for dealing with this problem is to attempt to use some sort of criteria to identify to pick one sampling design among many with the same first-order inclusion probabilities using additional information. One class of solutions to this problem is what is called *balanced sampling* which we will talk about in another lecture.