Monday, Nov 4

Inclusion Probabilities Revisited

The **inclusion probability** of an element is the probability that it will be included in a sample.

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 inclusion probabilities for all elements.

Element ($\mathcal{E}_i$)	Inclusion Probability (π_i)
$\mathcal{E}_1$	0.8
$\mathcal{E}_2$	0.5
$\mathcal{E}_3$	0.7
$\mathcal{E}_4$	0.5
$\mathcal{E}_5$	0.5

Note that *inclusion probabilities* are not the same as *selection probabilities* when sampling with replacement. But they are related because when sampling with replacement $\pi_i = 1 - (1 - \delta_i)^n$, where π_i and δ_i are the inclusion and selection probabilities, respectively, of the i-th element.

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

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

The Horvitz-Thompson Estimator

Suppose we select a sample of elements with some arbitrary probability sampling design. The **Horvitz-Thompson estimator** of τ is

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

where $\mathcal{S}$ represents the set of *distinct* elements in the sample, and π_i is the inclusion probability of the i-th element.

Example: Consider the sampling design in the previous example with the population of five elements. Suppose that the values of the target variable are $y_1 = 1$, $y_2 = 4$, $y_3 = 2$, $y_4 = 3$, and $y_5 = 1$. What is the estimate of τ based on the Horvitz-Thompson estimator if we select, say, sample $\mathcal{S}_1$?

Example: Consider the sampling design in the previous example that used sampling with replacement. Suppose that the values of the target variable are $y_1 = 3$, $y_2 = 1$, $y_3 = 2$, and $y_4 = 5$. What would be the value of $\hat{\tau}$ based on the Horvitz-Thompson estimator if the sample was $\mathcal{S} = \{\mathcal{E}_1, \mathcal{E}_1, \mathcal{E}_2\}$?

The Horvitz-Thompson estimator of τ can be used to estimate μ.

1. If the number of elements in the population is *known*, then we divide $\hat{\tau}$ by that number.
2. If the number of elements in the population is *unknown*, it can be estimated as

$$\sum_{i \in \mathcal{S}} \frac{1}{\pi_i},$$

and then

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

Example: What would be the estimates of μ for the earlier problems if the number of elements in the population was unknown?

Some Properties of Inclusion Probabilities and the Horvitz-Thompson Estimator

1. The Horvitz-Thompson estimator is an unbiased estimator of τ for *any* sampling design provided that all $\pi_i > 0$.[1]

2. For a design with a *fixed* sample size, $\sum_{i\in\mathcal{P}} \pi_i$ (i.e., the sum of the inclusion probabilities for *all* elements in the population) equals the number of elements in the sample. For a design with a *random* sample size, $\sum_{i\in\mathcal{P}} \pi_i$ equals the *expected* number of elements in the sample.[2] An example of a design where the number of elements in the sample may be random is cluster sampling.

3. The Horvitz-Thompson estimator has zero variance if the inclusion probabilities are proportional to the target variable such that
$$\pi_i = ny_i/\tau_y.$$
This suggests that if we had the opportunity to *choose* the inclusion probabilities we might use something like
$$\pi_i = nx_i/\tau_x$$
for some auxiliary variable x_i that is assumed to be *approximately* proportional to y_i. Then the variance of $\hat{\tau}$ may not be zero, but it may be smaller than it would be had we not used the auxiliary variable to determine the inclusion probabilities. Also note that by specifying the inclusion probabilities as $\pi_i = nx_i/\tau_x$ guarantees that the sum of the inclusion probabilities in the population equals the n (if n is fixed). We will investigate how to go about specifying a design with chosen inclusion probabilities later.

[1] To prove that the Horvitz-Thompson estimator is unbiased, first note that we can write it as $\sum_{i\in\mathcal{P}} I_i y_i/\pi_i$ where $\mathcal{P}$ is the index set of *all* elements in the population, and I_i is an indicator variable such that $I_i = 1$ if the i-th element is in the sample, and $I_i = 0$ if the i-th element is not in the sample. Now to show that the estimator is unbiased we need to show that

$$E\left(\sum_{i\in\mathcal{P}} I_i \frac{y_i}{\pi_i}\right) = \tau.$$

To do this we use the fact that the expectation operator can be distributed over addition, and that $E(I_i) = 1 \times P(I_i = 1) + 0 \times P(I_i = 0) = \pi_i$ since $P(I_i) = \pi_i$ by definition. So we have that

$$E\left(\sum_{i\in\mathcal{P}} I_i \frac{y_i}{\pi_i}\right) = \sum_{i\in\mathcal{P}} E\left(I_i \frac{y_i}{\pi_i}\right) = \sum_{i\in\mathcal{P}} E(I_i)\frac{y_i}{\pi_i} = \sum_{i\in\mathcal{P}} y_i = \tau.$$

Note that we can write that $E(I_i y_i/\pi_i) = E(I_i)y_i/\pi_i$ because both y_i and π_i are considered to be fixed constants here. It is only I_i that is a random variable.

[2] To show this note that $\sum_{i\in\mathcal{P}} \pi_i = \sum_{i\in\mathcal{P}} E(I_i)$ where I_i is an *indicator variable* such that $I_i = 1$ if the i-th element is included in the sample, and $I_i = 0$ if the element is excluded from the sample. We have that $\pi_i = E(I_i)$ because $E(I_i) = 1\pi_i + 0(1-\pi_i) = \pi_i$. Now

$$\sum_{i\in\mathcal{P}} E(I_i) = E\left(\sum_{i\in\mathcal{P}} I_i\right),$$

because the expectation of a sum equals the sum of expectations, and $\sum_{i=1}^{N} I_i$ is a count of the number of elements in the sample. And if the number of elements is not random then $E\left(\sum_{i=1}^{N} I_i\right) = \sum_{i=1}^{N} I_i$.

Hansen-Hurwitz and Horvitz-Thompson

The Hansen-Hurwitz and Horvitz-Thompson estimators can be viewed as general "recipes" for estimators. But be careful not to confuse the *Hansen-Hurwitz* estimator with the *Horvitz-Thompson* estimator.

Hansen-Hurwitz Estimator

The Hansen-Hurwitz estimator can be used for any design that uses *sampling with replacement* with known *selection probabilities*. The estimator is

$$\hat{\tau} = \frac{1}{n} \sum_{i \in \mathcal{S}} \frac{y_i}{\delta_i},$$

where δ_i is the *selection probability* of the i-th element. Recall that a selection probability is the probability of selecting the element *on each draw* from the population. Also note that we may draw the same element more than once, and the *summation may involve the same element more than once.*

Example: Consider the sampling design in the previous example that used sampling with replacement. What would be the value of $\hat{\tau}$ based on the Hansen-Hurwitz estimator if the sample was $\mathcal{S} = \{\mathcal{E}_1, \mathcal{E}_1, \mathcal{E}_2\}$?

Horvitz-Thompson Estimator

The Horvitz-Thompson estimator can be used for *any* design with known *inclusion probabilities*. The estimator is

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

where π_i is the *inclusion probability* of the i-th element. Recall that an inclusion probability is the probability the element will be included in the sample. Also unlike the Hansen-Hurwitz estimator the summation is only over the *distinct* elements in the sample.

Example: What would be the value of $\hat{\tau}$ based on the Horvitz-Thompson estimator and the same sample obtained in the previous example?

Inclusion Probabilities for Some Common Designs

Inclusion probabilities can be relatively easily derived for the designs we have discussed so far.

1. For **simple random sampling** the inclusion probabilities are all

$$\pi_i = \frac{n}{N},$$

   where N is the number of elements in the population and n is the number of elements sampled.

2. For **stratified random champing** the inclusion probabilities are

$$\pi_i = n_j/N_j$$

   if the i-th element is in the j-th stratum, where N_j is the number of elements in the j-th stratum, and n_j is the number of elements sampled from the j-th stratum.

3. For **one-stage cluster sampling** the inclusion probabilities are

$$\pi_i = n/N$$

   if clusters are sampled using *simple random sampling*, where N is the number of clusters, and n is the number of sampled clusters. But if clusters are sampled with replacement with probabilities proportional to cluster size, so that the *selection probability* of the i-th cluster is $\delta_j = m_j/M$, then

$$\pi_i = 1 - (1 - m_j/M)^n,$$

   assuming that the i-th element is in the j-th cluster, where M is the number of elements in the population and m_j is the number of elements in the j-th cluster.

4. For **two-stage cluster sampling** with simple random sampling at both stages, the inclusion probabilities are

$$\pi_i = \frac{n}{N} \times \frac{m_j}{M_j}$$

   assuming that the i-th element is in the j-th cluster, where N is the number of clusters, and n is the number of sampled clusters, M_j is the number of elements in the j-th cluster, and m_j is the number of sampled elements from the j-th cluster. If clusters are sampled with replacement with probabilities proportional to cluster size, but sampling in the second stage is still simple random sampling, then

$$\pi_i = [1 - (1 - M_j/M)^n] \times \frac{m_j}{M_j}.$$

   For simple and stratified random sampling, the Horvitz-Thompson estimator for τ is the same as that we discussed in the context of those designs. For one-stage and two-stage cluster sampling with simple random sampling at both stages, the Horvitz-Thompson estimator for τ based on the inclusion probabilities shown above is the same as the unbiased estimator for τ we discussed in the context of discussing each of those designs.

Horvitz-Thompson Estimators for Some Common Designs

For simple and stratified random sampling, the Horvitz-Thompson estimator of τ is the same estimator we discussed for those designs. For one- and two-stage cluster sampling designs using simple random sampling of clusters, the Horvitz-Thompson estimators of τ are the *unbiased* estimators we discussed in class.