Monday, Oct 28

Abundance Estimation

Abundance estimation concerns estimating (a) the number of elements in a population, or (b) the number of elements in a particular domain. Abundance estimation can often be done using designs and estimators we have already discussed if we define the target variable in a particular way.

Abundance Estimation with One-Stage Cluster Sampling

Consider a one-stage cluster sampling design and where y_{ij} is the value of the target variable for the j-th element in the i-th cluster. Suppose we define the target variable as

$$y_{ij} = 1$$

for *all* M elements in the population. Recall that we defined y_i as the sum of all elements in the i-th cluster so that

$$y_i = \sum_{j=1}^{m_i} y_{ij}.$$

But if all $y_{ij} = 1$ then $y_i = m_i$ (recall that m_i is the *number* of elements in the i-th cluster), and $\tau = M$, where M is the number of elements in the population. So here *estimation of M is the same as estimating τ when $y_i = m_i$.*

Recall that we discussed two estimators — "unbiased" and "ratio" — of τ for one-stage cluster sampling with simple random sampling of clusters:

$$\hat{\tau} = \frac{N}{n}\sum_{i \in \mathcal{S}} y_i \quad \text{and} \quad \hat{\tau} = M\frac{\sum_{i \in \mathcal{S}} y_i}{\sum_{i \in \mathcal{S}} m_i}.$$

The ratio estimator is not useful for estimating M. Why? But we could use an alternative *cluster-level* auxiliary variable x_i and use the ratio estimator

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

where x_i is the value of the auxiliary variable for the i-th *cluster*, and $\tau_x = \sum_{i=1}^{N} x_i$ is known. Why might we prefer this ratio estimator over the unbiased estimator?

Example: A warehouse contains 1000 boxes of widgets. Suppose we select a sample of five boxes using simple random sampling. The number of widgets in each of these boxes are 10, 12, 15, 9, and 14 widgets. What is our estimate of the number of widgets in the warehouse?

Example: A warehouse contains 1000 boxes of widgets Suppose we select a sample of five boxes using simple random sampling. The number of widgets in each of these boxes are 10, 12, 15, 9, and 14 widgets. We also find that the weights of these boxes are 5.5, 6.2, 7, 4, and 8.3 lbs, and it is known that the weight of all the boxes in the warehouse is 6220 lbs. Using weight as an auxiliary variable, what is our estimate of the number of widgets in the warehouse?

Now suppose we want to estimate the number of elements in a population that are in a given domain based on a one-stage cluster sampling design. Define y_{ij} as

$$y_{ij} = \begin{cases} 1, & \text{if the } j\text{-th element in the } i\text{-th cluster is in the domain,} \\ 0, & \text{otherwise.} \end{cases}$$

Now

$$y_i = \sum_{j=1}^{m_i} y_{ij}$$

is the number of elements in the i-th cluster *that are in the domain*, and

$$\tau = \sum_{i=1}^{N} y_i$$

is the number of elements in the population *that are in the domain.* So estimation of the number of elements in a domain is the same as estimating τ when y_i is the number of elements in the i-th cluster that are in the domain.

Here we could use

$$\hat{\tau} = \frac{N}{n} \sum_{i \in \mathcal{S}} y_i \quad \text{or} \quad \hat{\tau} = M \frac{\sum_{i \in \mathcal{S}} y_i}{\sum_{i \in \mathcal{S}} m_i},$$

depending on if we know M (i.e., the number of elements in the population). We also could use

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

for some other cluster-level auxiliary variable if M is unknown. Why might we prefer this ratio estimator over the unbiased estimator?

Example: A warehouse contains 1000 boxes of widgets. Suppose we select a sample of five boxes using simple random sampling. The number of widgets in each of these boxes are 10, 12, 15, 9, and 14 widgets. These contain 5, 4, 10, 3, and 7 defective widgets, respectively. What is an estimate of the number of defective widgets in the warehouse?

Example: Suppose we know that there are a total of 10000 widgets in the warehouse. Now what is our estimate if we use the number of widgets in each box as an auxiliary variable?

Example: Suppose we do not know the number of widgets in the warehouse, but we do know that the weights of these boxes are 5.5, 6.2, 7, 4, and 8.3 lbs, and that the weight of all the boxes in the warehouse is 6220 lbs. Now what is our estimate of the number of defective widgets when using box weight as an auxiliary variable?

Abundance Estimation with Simple Random Sampling

Consider a simple random sampling design and suppose we define the target variable as

$$y_i = 1$$

for *all* N elements in the population. Then $\tau = N$. How can we estimate N?

1. We cannot use $\hat{\tau} = N\bar{y}$. Why?
2. We could use the ratio estimator $\hat{\tau}_y = \tau_x \bar{y}/\bar{x}$, provided we have τ_x (and presumably not from observing $x_1, x_2, \dots, x_N$ individually as then we would know N and not need to estimate it).

Note that $\bar{y} = 1$ because all $y_i = 1$.

Example: Suppose we wish to estimate the number of fish in the hold of a trawler. The total weight of the fish in the hold is 18000 lbs. If we select a sample of 100 fish using simple random sampling and find that the mean weight in that sample is 20 lbs, what is our estimate of the number of fish in the hold using the ratio estimator with weight as an auxiliary variable?

Now suppose we want to estimate the number of elements in the population that are in a given domain. Define y_i as

$$y_i = \begin{cases} 1, & \text{if the } i\text{-th element is in the domain,} \\ 0, & \text{otherwise.} \end{cases}$$

So τ is the number of elements in the population that are in the domain.

1. We could use $\hat{\tau} = N\bar{y}$ if we know N.
2. We could use $\hat{\tau}_y = \tau_x \bar{y}/\bar{x}$ if we know τ_x for some auxiliary variable x_i.

For both estimators, note that

$$\bar{y} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i$$

is the *proportion* of elements in the sample that are in the domain.

Example: Suppose we wish to estimate the number of *defective* widgets in a box of 1000 widgets. A simple random sample of 100 widgets yields 20 defective widgets. What is our estimate of the number of defective widgets in the box?

Example: Suppose we wish to estimate the number of *female* fish in the hold of a trawler. The total weight of the fish in the hold is 18000 lbs. If we select a sample of 100 fish using simple random sampling and find that the mean weight in that sample is 20 lbs. We also find that 40 of the fish in the sample are female. What is our estimate of the number of female fish in the hold?