

Friday, October 17

Sampling With Replacement with Unequal Selection Probabilities

This design can be described as follows:

1. Assume N sampling units, each with a specified *selection probability* δ_i .
2. Obtain a sample of n sampling units by selecting units one at a time such that on each draw the probability that the i -th unit is selected is δ_i . The selection probabilities *do not change* from draw to draw, so it is possible that we will select the same element on two or more draws (i.e., sampling *with replacement*).

Selection Versus Inclusion Probabilities

Selection probabilities (δ_i) are *not* generally equal to inclusion probabilities (π_i).

Example: Suppose we obtain a sample of n sampling units *with replacement*, where the *selection probability* of the i -th unit is δ_i . The *inclusion probability* of the i -th unit is then

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

Example: Selection probabilities are usually only useful when sampling *with replacement*. But they can be defined when sampling *without replacement* if we select the sample by selecting *one unit at a time*. Suppose we have a population with $N = 10$ elements, and a simple random sampling design *without replacement* with a sample size of $n = 5$. All elements have *inclusion probabilities* of $\pi_i = n/N = 0.5$. But the *selection probabilities* change over draws and depend on previous draws.

Here is one possible sample:

$$\mathcal{S} = \{\mathcal{E}_4, \mathcal{E}_3, \mathcal{E}_7, \mathcal{E}_1, \mathcal{E}_9\}.$$

The selection probabilities are shown in the table below.

Draw	Element										Selected
	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$	$\mathcal{E}_5$	$\mathcal{E}_6$	$\mathcal{E}_7$	$\mathcal{E}_8$	$\mathcal{E}_9$	$\mathcal{E}_{10}$	
1	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	4
2	1/9	1/9	1/9	0	1/9	1/9	1/9	1/9	1/9	1/9	3
3	1/8	1/8	0	0	1/8	1/8	1/8	1/8	1/8	1/8	7
4	1/7	1/7	0	0	1/7	1/7	0	1/7	1/7	1/7	1
5	0	1/6	0	0	1/6	1/6	0	1/6	1/6	1/6	9

Here is another possible sample:

$$\mathcal{S} = \{\mathcal{E}_2, \mathcal{E}_6, \mathcal{E}_1, \mathcal{E}_7, \mathcal{E}_3\}.$$

The selection probabilities are shown in the table below.

Note that a defining property of sampling *with replacement* is that the selection probabilities *do not change from draw to draw*. When sampling *with replacement* with *equal selection probabilities*, we might get a sample such as the following:

$$\mathcal{S} = \{\mathcal{E}_9, \mathcal{E}_9, \mathcal{E}_7, \mathcal{E}_1, \mathcal{E}_{10}\}.$$

But the selection probabilities remain constant.

They would also remain constant if the selection probabilities were *unequal*.

Draw	Element										Selected
	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$	$\mathcal{E}_5$	$\mathcal{E}_6$	$\mathcal{E}_7$	$\mathcal{E}_8$	$\mathcal{E}_9$	$\mathcal{E}_{10}$	
1	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	2
2	1/9	0	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	6
3	1/8	0	1/8	1/8	1/8	0	1/8	1/8	1/8	1/8	1
4	0	0	1/7	1/7	1/7	0	1/7	1/7	1/7	1/7	7
5	0	0	1/6	1/6	1/6	0	0	1/6	1/6	1/6	3

Draw	Element										Selected
	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$	$\mathcal{E}_5$	$\mathcal{E}_6$	$\mathcal{E}_7$	$\mathcal{E}_8$	$\mathcal{E}_9$	$\mathcal{E}_{10}$	
1	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	9
2	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	9
3	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	7
4	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1
5	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	10

The Hansen-Hurwitz Estimator

Assuming an element sampling design using sampling *with replacement*, the Hansen-Hurwitz estimator of τ is

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

The estimated variance of this estimator is

$$\hat{V}(\hat{\tau}) = \frac{1}{n(n-1)} \sum_{i \in \mathcal{S}} (y_i/\delta_i - \hat{\tau})^2.$$

Example: Suppose we select a sample of $n = 3$ units with replacement: $\mathcal{E}_5$, $\mathcal{E}_2$, and $\mathcal{E}_5$ again. The target variable values of $y_5 = 4$ and $y_2 = 1$, and selection probabilities of $\delta_5 = 0.05$ and $\delta_2 = 0.01$. What is the estimate of τ and the estimated variance of $\hat{\tau}$?

What is the Hansen-Hurwitz estimator if all selection probabilities are *equal* (i.e., all $\delta_i = 1/N$)?

Note: For the estimator of μ , we would divide $\hat{\tau}$ by the number of elements in the population (N for element sampling, and M for cluster sampling), and the estimated variance of $\hat{\tau}$ by the square of this (N^2 for element sampling, and M^2 for cluster sampling). But see the section below on estimation of μ with PPS cluster sampling.

Specification of Selection Probabilities

How should we specify the selection probabilities? The best choice would be to use

$$\delta_i = \frac{y_i}{\sum_{i=1}^N y_i}.$$

That is, $\delta_i = y_i/\tau_y$. Why would this be best?

Since we do not know y_i before we sample, instead perhaps we can find an *auxiliary variable* that is *approximately* proportional to y_i such that $y_i \approx cx_i$, then we could use

$$\delta_i = \frac{x_i}{\sum_{i=1}^N x_i}.$$

That is, $\delta_i = x_i/\tau_x$.

Note: We are implicitly assuming here that all $y_i > 0$ and all $x_i > 0$.

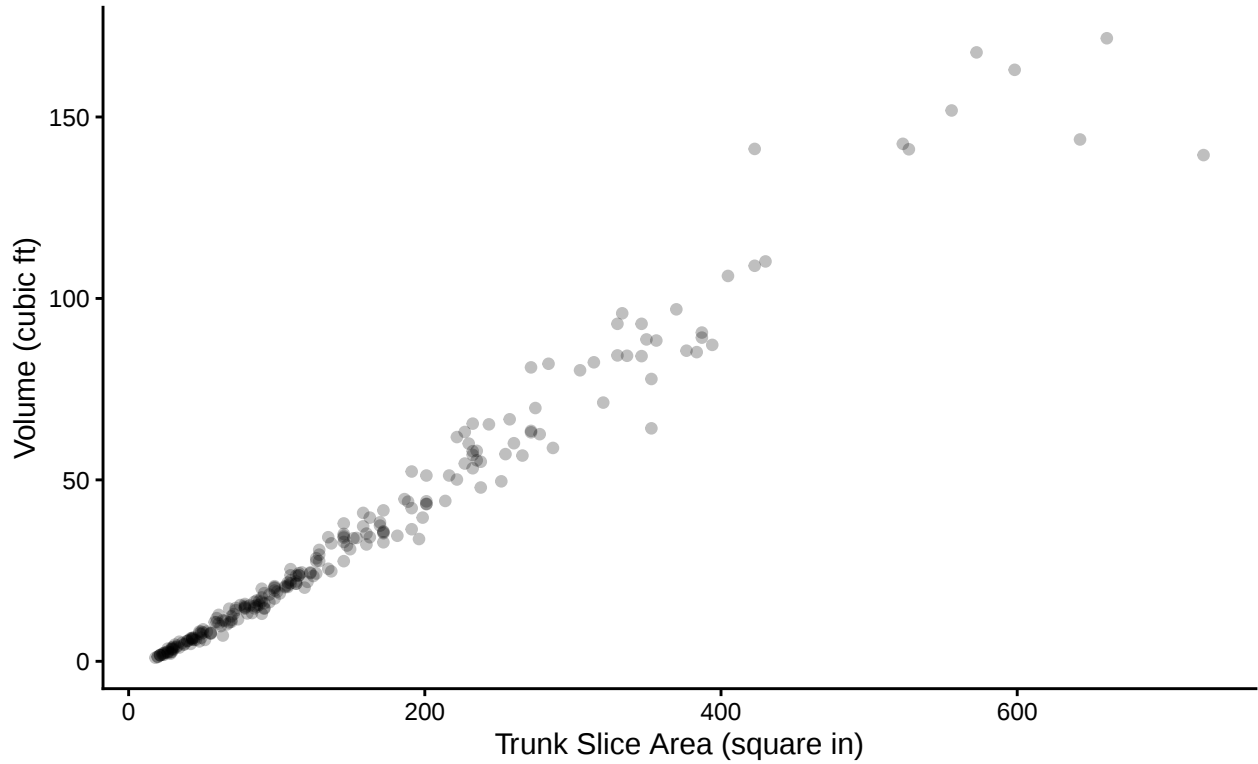
Example: Suppose we have the following $N = 5$ sampling units.

Unit	x_i	δ_i
1	1	
2	4	
3	2	
4	5	
5	8	

What would be the selection probabilities assuming that y_i is approximately proportional to x_i ?

Simulation Study

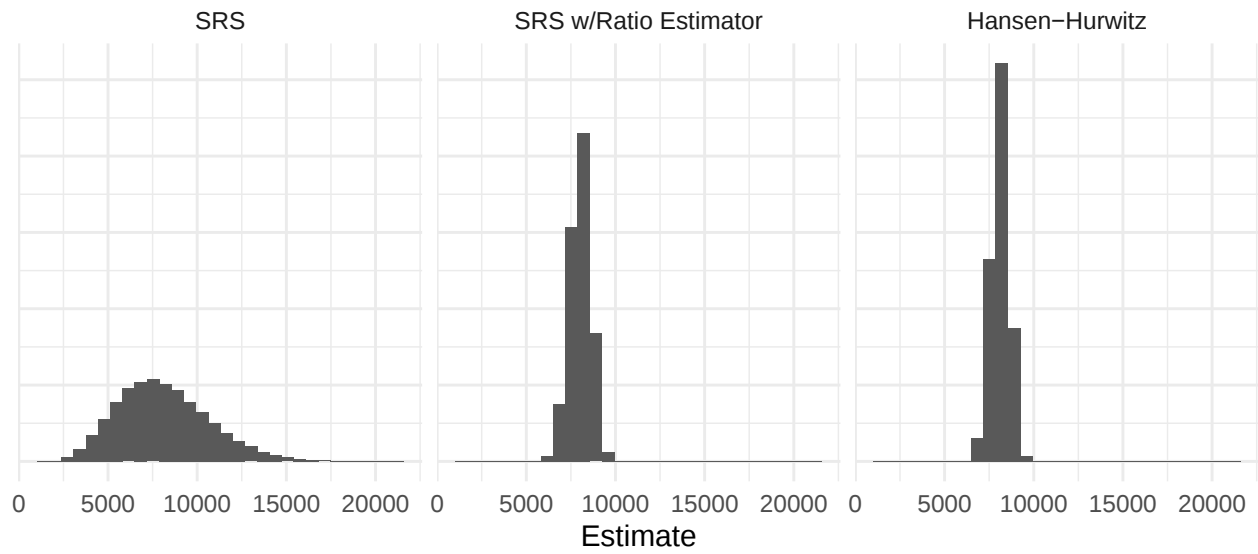
Example: Consider the observations of a target variable (volume) and an auxiliary variable (area) for a population of elements.



Let y_i and x_i be the volume and trunk slice area, respectively. Let's compare three approaches to estimating τ_y .

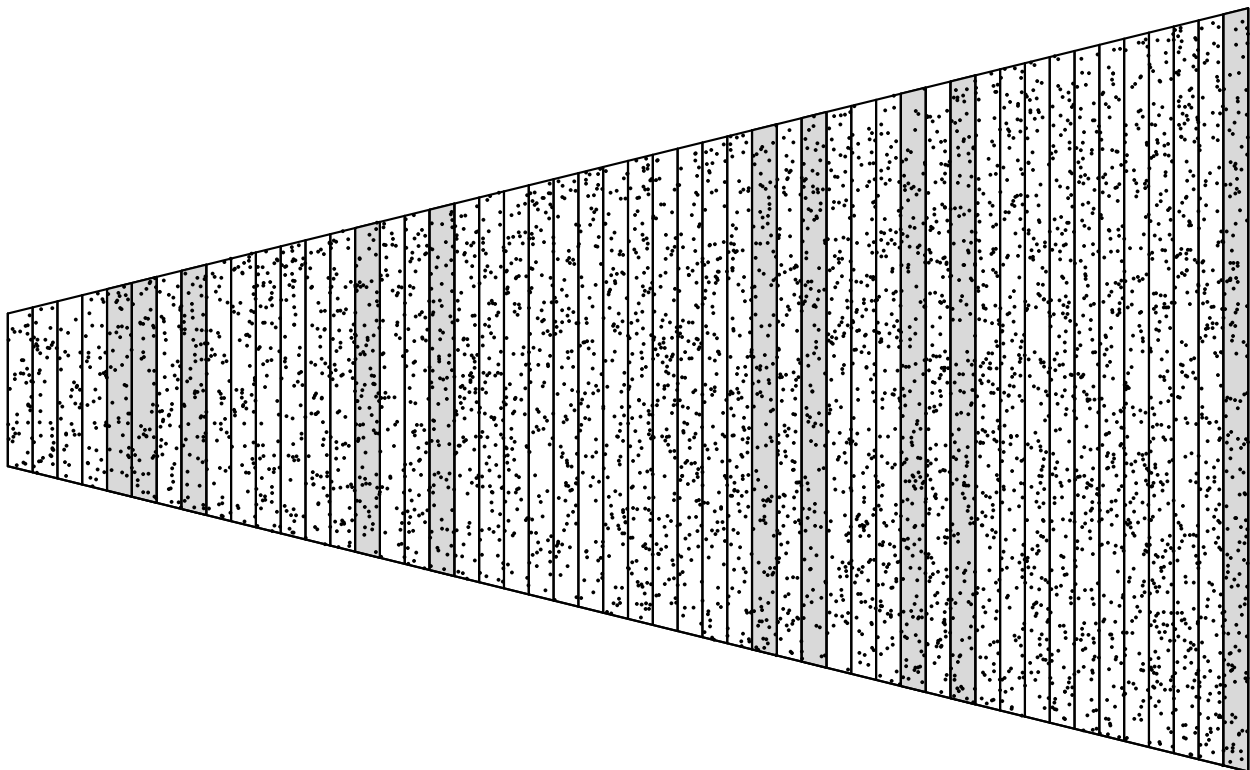
1. Simple random sampling with $\hat{\tau}_y = N\bar{y}$.
2. Simple random sampling with the ratio estimator $\hat{\tau}_y = \tau_x \bar{y} / \bar{x}$.
3. Sampling with replacement with selection probabilities of $\delta_i = x_i / \tau_x$ and using the Hansen-Hurwitz estimator.

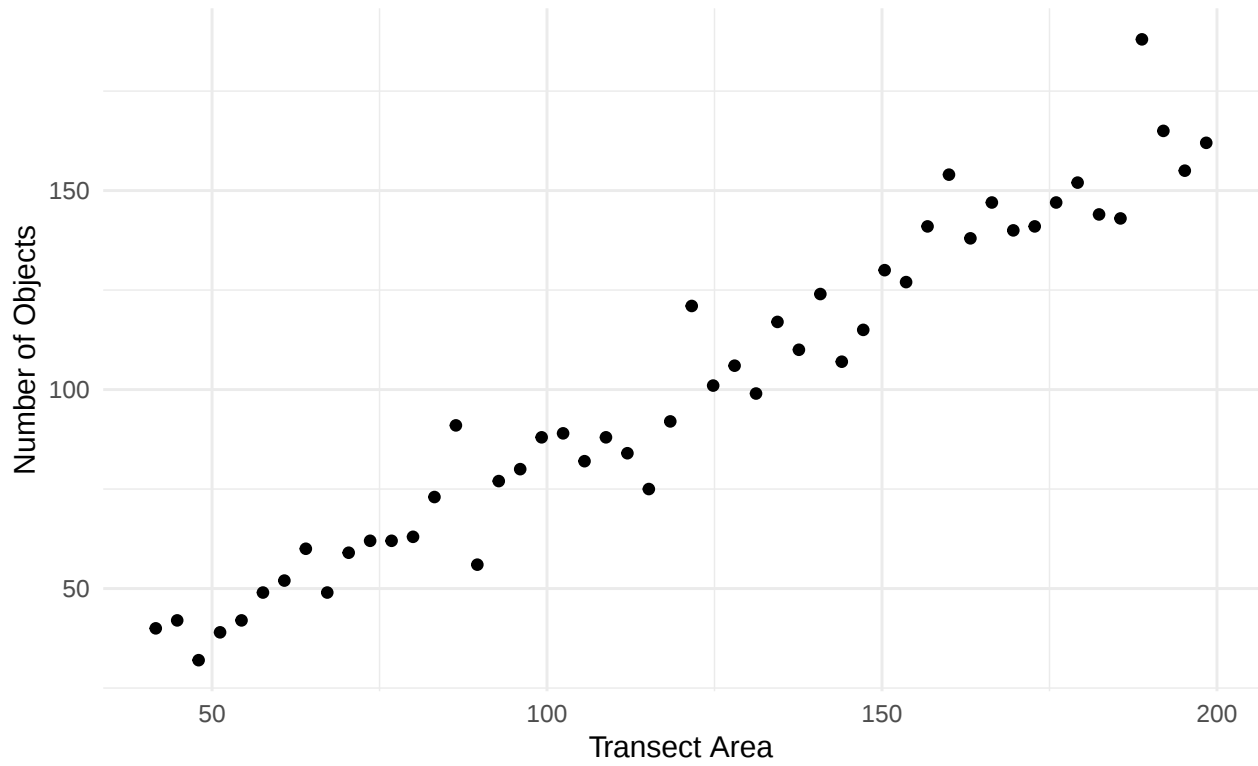
Note: The value of τ_y is 8124.9.



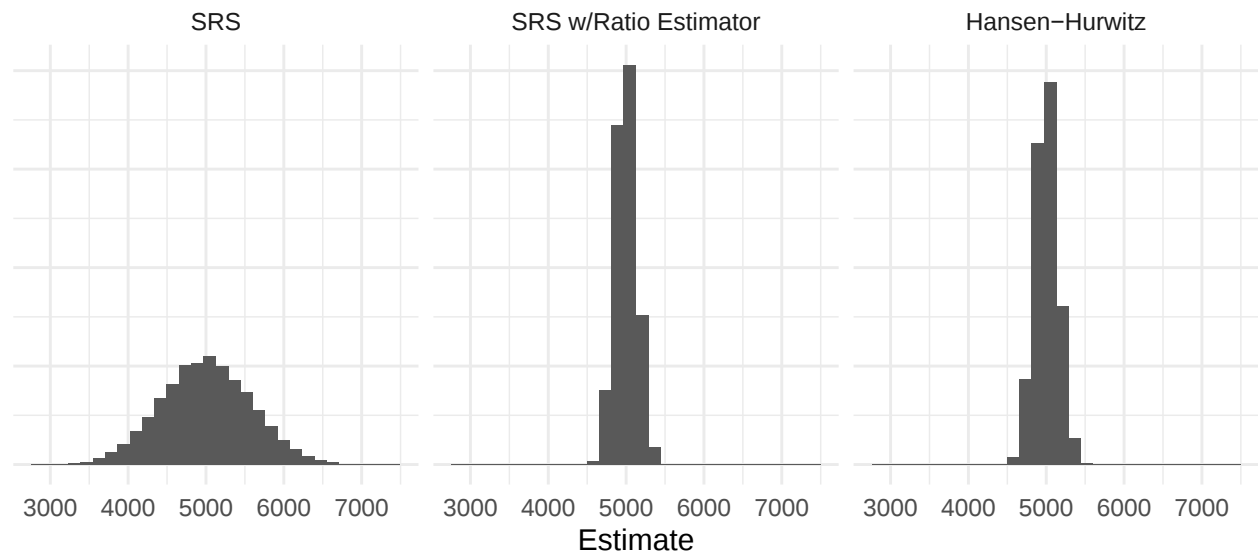
Estimator	Variance	Bound
SRS	6548857.3	5118.1470
SRS w/Ratio Estimator	342844.8	1171.0591
Hansen-Hurwitz	234438.3	968.3766

Now consider a situation of estimating the number of objects in transects where the transects vary in size. Let y_i be the number of objects, and let x_i be the size of the transect.





Note: The value of τ_y is 5000.



Estimator	Variance	Bound
SRS	331972.37	1152.3409
SRS w/Ratio Estimator	17951.49	267.9664
Hansen-Hurwitz	21302.52	291.9076

Why Ever Sample *With* Replacement

It would seem that sampling *with replacement* would be less efficient than sampling *without replacement* since there is a probability of including the same sampling unit more than once. So why would we ever use a with

replacement sampling design?

1. If n is much smaller than N (as is often the case), then the probability of selecting the same unit more than once is negligible and so the loss in efficiency is small. In such cases sampling *with* replacement is *almost* like sampling *without* replacement.
2. The mathematics of sampling with replacement is much simpler, which makes both sampling design and analysis much more tractable in practice. Later we will discuss sampling *without replacement* with *unequal inclusion probabilities*, which also can be used to improve estimation. But outside of some special cases (e.g., stratified random sampling) the actual process of selecting the sample is much more complex, requiring specialized algorithms and expertise.
3. Some specialized sampling designs/protocols cannot easily avoid selecting the same sampling unit more than once. We will discuss some of these later.

Cluster Sampling with Probabilities Proportional to Size (PPS)

Consider a cluster sampling design but where the clusters are sampled *with replacement* with (selection) probabilities proportional to *cluster size* so that we use m_i for x_i . Then

$$\delta_i = \frac{m_i}{\sum_{i=1}^N m_i} = \frac{m_i}{M}.$$

This is sometimes called sampling with **probabilities proportional to size** (PPS).

As before let y_i be the cluster total for the i -th cluster so that $y_i = \sum_{j \in S} y_{ij}$. Then the Hansen-Hurwitz estimator for τ can be written as

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

since $\delta_i = m_i/M$. The estimated variance for $\hat{\tau}$ then becomes

$$\hat{V}(\hat{\tau}) = \frac{M^2}{n(n-1)} \sum_{i \in S} (y_i/m_i - \hat{\tau}/M)^2.$$

The Hansen-Hurwitz estimator of μ is obtained as $\hat{\mu} = \hat{\tau}/M$. That is,

$$\hat{\mu} = \frac{1}{n} \sum_{i \in S} \frac{y_i}{m_i}.$$

The estimated variance of $\hat{\mu}$ is then

$$\hat{V}(\hat{\mu}) = \frac{1}{n(n-1)} \sum_{i \in S} (y_i/m_i - \hat{\mu})^2.$$

Example: A cluster sampling design selects $n = 3$ boxes using sampling with replacement with probabilities proportional to cluster size (i.e., $\delta_i = m_i/M$). The number of widgets in these boxes are $m_1 = 3$, $m_2 = 4$, and $m_3 = 5$. The total weight of the widgets in these boxes are $y_1 = 6.2$, $y_2 = 7.5$, and $y_3 = 10.3$. Assume that the population of 425 widgets is contained in 100 boxes. Assuming sampling with replacement with probabilities proportional to size, what are the estimates of τ and μ as given by the Hansen-Hurwitz estimator?