

Unequal Probability Sampling (Chapter 6)

Unequal probability sampling is when some units in the population have probabilities of being selected from others. This handout introduces the Hansen-Hurwitz (H-H) estimator and Horvitz-Thompson (H-T) estimator, examines the properties of both types of estimators for the population total and mean, and compares the two estimators by way of an example.

The Hansen-Hurwitz (H-H) Estimator

 for random sampling with replacement.

- Suppose a sample of size n is selected randomly with replacement from a population but that on each draw, unit i has probability p_i of being selected, where $\sum_{i=1}^N p_i = 1$. The probability p_i here is called the selection probability for the i^{th} unit. Let y_i be the response variable measured on each unit selected. Note that if a unit is selected more than once, it is used as many times as it is selected. An unbiased estimator of the population total $\tau = \sum_{i=1}^N y_i$ is given by:

$$\hat{\tau}_p = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}.$$

An unbiased estimator of the population mean is $\hat{\mu}_p = (1/N)\hat{\tau}_p$.

- Dividing by p_i gives higher *weight* to units less likely to be selected.
- What happens to this estimator if $p_i = 1/N$, $i = 1, \dots, N$, so that each unit has an equal chance of selection?

Example: Consider a population of size $N = 3$, with values and corresponding selection probabilities given in the first two columns of the table to the right. Note that the true population total is $\tau = 14$. Consider taking a sample of size 1. The H-H estimates of the total for each of the 3 values (samples) are given in the third column of the table.

| Values | Probabilities | $\hat{\tau}_p$ |
|-----------|---------------|----------------|
| $y_1 = 3$ | $p_1 = .2$ | 15 |
| $y_2 = 2$ | $p_2 = .5$ | 4 |
| $y_3 = 9$ | $p_3 = .3$ | 30 |

The expected value of τ_p is:

$$E(\hat{\tau}_p) = .2(15) + .5(4) + .3(30) = \underline{14} = \tau.$$

- So, in $\hat{\tau}_p = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}$, each $\frac{y_i}{p_i}$ is unbiased for τ .

Why would you want select units with unequal probabilities?

- It may be the most convenient way to sample. Recall the example of taking sample of ponds by selecting a random point on a map. If the point lands in a pond then that pond is selected for the sample. It would require a lot more effort to enumerate all the ponds so that an SRS could be selected. See also the farm example below.

- If the response variable is positively correlated with the selection probability, then the Hansen-Hurwitz estimator can have lower variance than the estimator based on an SRS.

Properties of the Hansen-Hurwitz Estimator

$$\underline{E(\hat{\tau}_p)} =$$

$$\begin{aligned} \underline{\text{Var}(\hat{\tau}_p)} &= \text{Var} \left[\frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} \right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var} \left[\frac{y_i}{p_i} \right] \quad (\text{indep. due to sampling with replacement}) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^N p_j \left(\frac{y_j}{p_j} - \tau \right)^2 = \frac{1}{n} \sum_{j=1}^N p_j \left(\frac{y_j}{p_j} - \tau \right)^2, \end{aligned}$$

where τ is unknown, so we need to estimate it in this variance. An unbiased estimate of the variance can be computed as:

$$\widehat{\text{Var}}(\hat{\tau}_p) = \frac{1}{n(n-1)} \sum_{j=1}^n \left(\frac{y_j}{p_j} - \hat{\tau}_p \right)^2$$

Note that the properties of $\hat{\mu}_p = (1/N)\hat{\tau}_p$ follow easily:

$$E(\hat{\mu}_p) = (1/N)\tau = \mu \text{ (unbiased)}, \quad \text{Var}(\hat{\mu}_p) = (1/N)^2 \text{Var}(\hat{\tau}_p), \quad \widehat{\text{Var}}(\hat{\mu}_p) = (1/N)^2 \widehat{\text{Var}}(\hat{\tau}_p).$$

Notes on the Hansen-Hurwitz Estimator

1. We only need p_i for the units in the sample (not the whole population).
2. We need not know N in order to estimate τ .
3. If we let $y_i = 1, i = 1, \dots, N$, then $\tau = N$ and $\hat{\tau}_p = \frac{1}{n} \sum_{i=1}^n \frac{1}{p_i} = \widehat{N}$ is an estimator of N .
4. If there is low variability between the values of y_j/p_j , then the H-H estimator will have low variance, with the extreme case being when y_j and p_j are exactly proportional to each other. On the other hand, the H-H estimator will have high variance when there is high variability among the values of y_j/p_j .

Example: Consider a population of farms on a 25x25 grid of varying sizes and shapes, as given on the last page of this handout. If we randomly select a single square on this grid, then letting x_i = the area of farm i and $A = 625$ total units, the probability that farm i is selected is: $p_i = \frac{x_i}{A} = \frac{x_i}{625}$.

Let y_i = the response variable of interest (which might be x_i).

- If $y_i = x_i$, then $\tau = \sum_{i=1}^N y_i =$ the total area of all farms. In this example, the total area is known to be $A = 625$, so that $y_i = x_i$ is uninteresting here.
- If $y_i = 1, i = 1, \dots, N$, then $\tau = \sum_{i=1}^N y_i =$ the total number of farms (which is more interesting).
- The response variable y_i might also be something like the number of workers for farm i , or the income for farm i , etc.

Consider taking a sample of 5 farms with replacement with probability-proportional-to-size (PPS) and computing:

- (i) The estimated number of workers. (ii) The estimated number of farms.

How do we take a random sample of pixels?

Using the sample command in R (with `replace=T`), a random sample of size $n = 5$ pixels was taken (with replacement), as summarized in the table below. An estimate of the

| Coordinates | Farm Data | $p_i = \frac{x_i}{A} = \frac{\text{Size of Farm}}{\text{Total Area}}$ |
|-------------|-----------|---|
| 8,19 | D2 | 5/625 |
| 19,25 | C8 | 28/625 |
| 21,21 | B4 | 12/625 |
| 15,4 | A8 | 14/625 |
| 7,20 | A3 | 13/625 |

total number of workers, (where $y_i =$ the # of workers for farm i) is:

$$\hat{\tau}_p = \frac{1}{5} \left[\frac{2}{5/625} + \frac{8}{28/625} + \frac{4}{12/625} + \frac{8}{14/625} + \frac{3}{13/625} \right] = \underline{227.66 \text{ workers.}}$$

An estimate of the total number of farms (where $y_i = 1$ for all i) is:

$$\hat{\tau}_p = \frac{1}{5} \left[\frac{1}{5/625} + \frac{1}{28/625} + \frac{1}{12/625} + \frac{1}{14/625} + \frac{1}{13/625} \right] = \underline{58.42 \text{ farms.}}$$

Class Results

Truth

Workers:

$\tau = 247$ workers

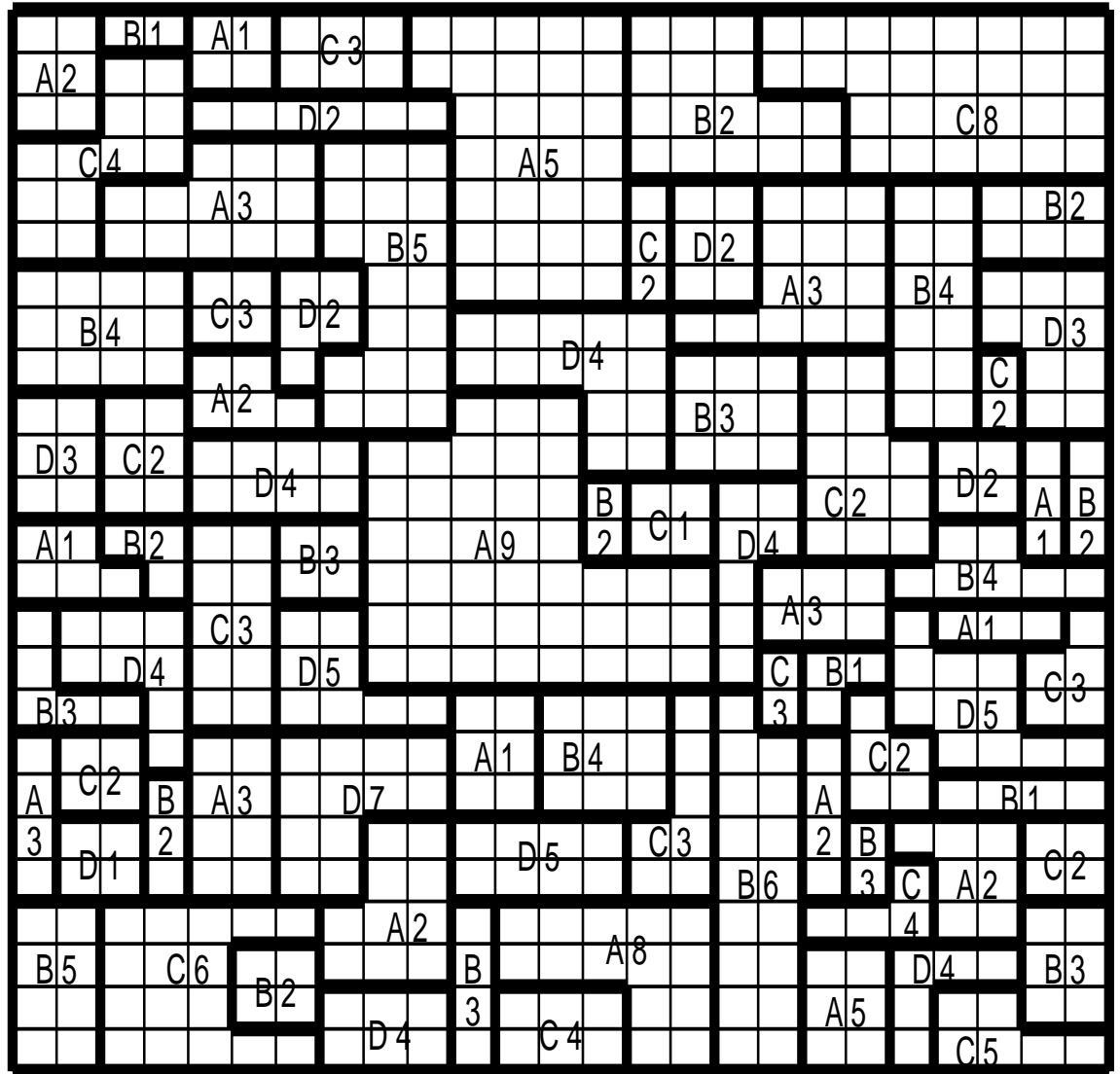
Farms:

$N = 78$ farms

Farms Map

The 25x25 grid map below represents the boundaries of farms in a certain area.

- The letters for each farm represent the “type” of farm.
- The numbers for each farm represent the number of workers on the farm.



Summary of results for PPS sampling

Here are some general results for PPS sampling with replacement. Some of these were illustrated with the farm example. This situation is discussed in Section 6.1 where the population mean and total can be estimated using the Hansen-Hurwitz estimator. Let:

$$\begin{aligned}
 N &= \text{the population size,} \\
 x_i &= \text{the size of the } i^{\text{th}} \text{ unit in the population,} \\
 \tau_x &= \sum_{i=1}^N x_i = \text{total size of all units in the population,} \\
 \mu_x &= \frac{\tau_x}{N} = \text{mean size of the units in the population,} \\
 p_i &= \frac{x_i}{\tau_x} = \text{the probability of selecting unit } i \text{ on any one draw,} \\
 n &= \text{the sample size.}
 \end{aligned}$$

1. Suppose N is unknown and we are interested in estimating it. Let $y_i = 1$ for all i in the Hansen-Hurwitz estimator:

$$\widehat{N} = \frac{1}{n} \sum_{i=1}^n \frac{1}{p_i} = \frac{\tau_x}{n} \sum_{i=1}^n \frac{1}{x_i} \quad (\text{unbiased}), \quad \widehat{\text{Var}}(\widehat{N}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{1}{p_i} - \widehat{N} \right)^2$$

2. Suppose the size variable itself is the variable of interest, that is $y_i = x_i$, and we are interested in estimating μ_x or τ_x . If N and τ_x are known, then $\mu_x = \tau_x/N$ and there is nothing to estimate. If either N or τ_x (or both) are unknown, then observe that:

$$\widehat{\mu}_x = \frac{\tau_x}{\widehat{N}} = \frac{\tau_x}{\frac{1}{n} \sum_{i=1}^n \frac{1}{p_i}} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}}$$

$\widehat{\mu}_x$ is the harmonic mean of the sizes of the units in the sample. Note that we do not need to know either N or τ_x to estimate μ_x ! This would be applicable, for example, to the pond example, where we know only the sizes of the ponds in the sample and not the total area of the ponds nor the number of ponds. There is not a closed form expression for the variance of $\widehat{\mu}_x$ and it must be approximated by either the delta method or bootstrapping (which will be discussed in the next section).

If N is known and τ_x is unknown, then we can estimate τ_x by

$$\widehat{\tau}_x = N \widehat{\mu}_x = \frac{N}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}}$$

Since $\text{Var}(\widehat{\tau}_x) = N^2 \text{Var}(\widehat{\mu}_x)$, the variance must be estimated by the delta method or bootstrapping.

3. Suppose that the variable of interest y_i is not size (e.g., number of workers in the farm example). Let

$$\tau_y = \sum_{i=1}^N y_i = \text{the population total of } y \text{ values,}$$

$$\mu_y = \frac{\tau_y}{N} = \text{the mean } y \text{ value for the population.}$$

Suppose we are just interested in estimating τ_y and/or μ_y . Then

(a) If τ_x is known (which implies that the $p_i = x_i/\tau_x$ are known), then

$$\hat{\tau}_y = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} \quad (\text{unbiased}), \quad \widehat{\text{Var}}(\hat{\tau}_y) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{p_i} - \hat{\tau}_y \right)^2.$$

If, in addition, N is known, then

$$\hat{\mu}_y = \frac{\hat{\tau}_y}{N} \quad (\text{unbiased}), \quad \widehat{\text{Var}}(\hat{\mu}_y) = \frac{1}{N^2} \widehat{\text{Var}}(\hat{\tau}_y).$$

(b) If either τ_x or N (or both) is unknown, then observe that

$$\hat{\mu}_y = \frac{\hat{\tau}_y}{\widehat{N}} = \frac{\frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}}{\frac{1}{n} \sum_{i=1}^n \frac{1}{p_i}} = \frac{\sum_{i=1}^n \frac{y_i}{x_i}}{\sum_{i=1}^n \frac{1}{x_i}}$$

Note that we do not need to know either τ_x or N to estimate μ_y . This estimator is a ratio estimator and its variance must be estimated by the delta method or bootstrapping (next section).

(c) In part (b), if N is known, then

$$\hat{\tau}_y = N \hat{\mu}_y.$$

Since $\text{Var}(\hat{\tau}_y) = N^2 \text{Var}(\hat{\mu}_y)$, the variance must be estimated by the delta method or bootstrapping.

If N and τ_x are unknown, then τ_y cannot be estimated.