

The Horvitz-Thompson Estimator

The Horvitz-Thompson estimator is a general estimator for a population total, which can be used for any probability sampling plan. This includes both sampling with and without replacement.

- Let π_i be the probability that the i^{th} unit of the population is included in the sample (inclusion probability).
- On each unit i , we measure a response y_i , and typically seek to estimate:

$$\tau = \sum_{i=1}^N y_i \text{ (population total) } \quad \text{OR} \quad \mu = \frac{1}{N} \sum_{i=1}^N y_i \text{ (population mean).}$$

Definition: The Horvitz-Thompson (H-T) estimator of τ is given by:

$$\hat{\tau}_\pi = \sum_{i=1}^v \frac{y_i}{\pi_i} \quad \text{where the sum is taken only over the } v \text{ distinct units in the sample.}$$

- The value v is sometimes referred to as the “effective” sample size.
- The higher the probability of inclusion, π_i , of a unit i to the sample, the less weight the corresponding response y_i is given. In this way the H-T estimator, like the Hansen-Hurwitz estimator, uses probability to weight the responses in estimating the total.
- The primary difference between the H-T and H-H estimator is the fact that the former uses the inclusion probability (π_i) of the units to the sample, whereas the latter uses the probability of selection (p_i) of a unit for a single draw. The H-H estimator is restricted to random sampling with replacement while the H-T estimator can be used in much wider range of sampling plans.

Mean and Variance of the Horvitz-Thompson Estimator

Mean: $E[\hat{\tau}_\pi] = E\left[\sum_{i=1}^v \frac{y_i}{\pi_i}\right]$. Now what?

Let $z_i = \begin{cases} 1 & \text{if the } i^{th} \text{ unit is in the sample} \\ 0 & \text{otherwise} \end{cases}$, $i = 1, \dots, N$. Then:

$$E[z_i] =$$

$$\text{Var}[z_i] =$$

$$\text{Cov}(z_i, z_j) =$$

where π_{ij} is the joint inclusion probability of units i, j .

Returning to the expectation of $\hat{\tau}_\pi$, we have:

$$\begin{aligned} \underline{\text{E}}[\hat{\tau}_\pi] &= \text{E} \left[\sum_{i=1}^N z_i \frac{y_i}{\pi_i} \right] \\ &= \\ &= \end{aligned}$$

Variance:

$$\begin{aligned} \text{Var}(\hat{\tau}_\pi) &= \text{Var} \left(\sum_{i=1}^N z_i \frac{y_i}{\pi_i} \right) \\ &= \\ &= \\ &= \end{aligned}$$

- An unbiased estimator of the variance is given by:

$$\widehat{\text{Var}}(\hat{\tau}_\pi) = \sum_{i=1}^v \left(\frac{1 - \pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i=1}^v \sum_{\substack{j=1 \\ j \neq i}}^v \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{y_i y_j}{\pi_{ij}},$$

where the extra π_i in the denominator of the first term and the π_{ij} in the denominator of the second term can be attributed to the use of v sample units instead of the N population units in the theoretical variance.

Horvitz-Thompson Estimator for the Mean: To estimate the population mean μ , the corresponding Horvitz-Thompson estimator is given by:

$$\hat{\mu}_\pi = \frac{\hat{\tau}_\pi}{N}, \quad \text{with variance } \text{Var}(\hat{\mu}_\pi) = \frac{1}{N^2} \text{Var}(\hat{\tau}_\pi).$$

If N is unknown, we can estimate it (let $y_i = 1$ for all i).

Example: The H-T Estimator for SRS without replacement

Consider taking a simple random sample (SRS), without replacement, of size n from a population of size N . The inclusion and joint inclusion probabilities are:

$$\pi_i = \qquad \qquad \qquad \pi_{ij} =$$

- Note that the Horvitz-Thompson estimator for SRS without replacement becomes:

$$\hat{\tau}_\pi = \sum_{i=1}^v \frac{y_i}{\pi_i} = \sum_{i=1}^n \frac{y_i}{n/N} = N \cdot \frac{1}{n} \sum_{i=1}^n y_i = N\bar{y} = \hat{\tau},$$

the usual estimator of the population total τ for an SRS derived earlier.

- Do you think the Horvitz-Thompson variance will be $\text{Var}(\hat{\tau}) = \sqrt{N(N-n)\sigma^2/n}$ as it was before for SRS?

Example: The H-T Estimator for sampling with replacement

Reconsider the scenario for the Hansen-Hurwitz estimator, where we sample with replacement from a population such that the probabilities of selection on any given draw are unequal. These probabilities were denoted p_1, \dots, p_N for a population of size N . The inclusion and joint inclusion probabilities are:

$$\pi_i =$$

$$\pi_{ij} =$$

With these then, the Horvitz-Thompson estimator is: $\hat{\tau}_\pi = \sum_{i=1}^v \frac{y_i}{1 - (1 - p_i)^n}$, and the H-T variance is a horrendous mess.

- For sampling with replacement, it is generally easier to use the Hansen-Hurwitz estimator.
- Although the Horvitz-Thompson estimator can be used for any probability sampling plan, there is often a simpler way to derive the estimator and its variance than through inclusion probabilities.

Comparison of H-H and H-T Estimators for the Farm Example

Consider taking a random sample of size $n = 5$ (with replacement) from the $N = 625$ pixels in the map of the farms given in class, and estimating the total number of workers on all the farms. This was done earlier for the Hansen-Hurwitz estimator, and will be repeated

here for the Horvitz-Thompson estimator.

Ten samples of size $n = 5$ will be taken where the individual farms will be selected according to the “probability proportional to size” (PPS) sampling described earlier. Specifically, a pair of integers between 1 and 25 will be chosen at random and the farm with the corresponding coordinates on the map will be selected.

First, consider the Horvitz-Thompson estimator for the *single* sample of size 5 given in class to estimate the total number of workers using the Hansen-Hurwitz estimator. The sample is repeated in the table below, along with the relevant components for the H-T estimator. As the samples here were distinct, the Horvitz-Thompson estimator of the total number of

Coordinates	Data	p_i	$\pi_i = 1 - (1 - p_i)^n$
8,19	D2	$5/625 = .0080$.0394
19,25	C8	$28/625 = .0448$.2048
21,21	B4	$12/625 = .0192$.0924
15, 4	A8	$14/625 = .0224$.1071
7,20	A3	$13/625 = .0208$.0998

workers, τ_y , is:

$$\begin{aligned}\hat{\tau}_\pi &= \sum_{i=1}^n \frac{y_i}{\pi_i} = \left[\frac{2}{.0394} + \frac{8}{.2048} + \frac{4}{.0924} + \frac{8}{.1071} + \frac{3}{.0998} \right] \\ &= \underline{237.94} \text{ workers.}\end{aligned}$$

- Recall that the estimated number of workers for the Hansen-Hurwitz estimator computed earlier was 227.66 workers. Since the true total number of workers was $\tau_y = 249$, does this make the Horvitz-Thompson estimator better?

To compute the estimated variance of this estimated total number of workers, we need first to compute the joint inclusion probabilities for each pair of units in the sample. Using the formula derived in class, given as: $\pi_{ij} = \pi_i + \pi_j - (1 - (1 - p_i - p_j)^n)$, the table below gives the ten π_{ij} values corresponding to the ten pairs of units. The estimated variance is then

Unit #	Unit Number				
	1	2	3	4	5
1	-	.0066	.0029	.0034	.0032
2	-	-	.0156	.0181	.0169
3	-	-	-	.0081	.0075
4	-	-	-	-	.0087

computed as:

$$\begin{aligned}\widehat{\text{Var}}(\widehat{\tau}_\pi) &= \sum_{i=1}^v \left(\frac{1 - \pi_i}{\pi_i^2} \right) y_i^2 + \sum_{i=1}^v \sum_{\substack{j=1 \\ j \neq i}}^v \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{y_i y_j}{\pi_{ij}} \\ &= \left[\left(\frac{1 - .0394}{.0394^2} \right) 2^2 + \cdots + \left(\frac{1 - .0998}{.0998^2} \right) 3^2 \right] \\ &\quad + 2 \left[\left(\frac{.0066 - (.0394)(.2048)}{(.0394)(.2048)} \right) \frac{(2)(8)}{.0066} + \cdots + \left(\frac{.0087 - (.1071)(.0998)}{(.1071)(.0998)} \right) \frac{(8)(3)}{.0087} \right] \\ &= 11191.15 - 2(4922.037) = \underline{1347.077},\end{aligned}$$

giving a standard error of $\widehat{\text{SE}}(\widehat{\tau}_\pi) = \sqrt{1347.077} = \underline{36.70 \text{ workers}}$. This is essentially the same as that found (36.75) with the Hansen-Hurwitz estimator.

R Code for computing H-H and H-T estimates and SE's

```
> n <- 5                                # Sets the sample size
> y <- c(2,8,4,8,3)                     # Sets the vector of y-values

> # Compute H-H estimate and SE
> p <- (1/625)*c(5,28,12,14,13)         # Computes the vector of selection probs.
> tau.p <- (1/n)*sum(y/p)                # Computes the H-H estimate
> var.tau.p <- var(y/p)/n                # Computes the variance of the H-H estimate
> c(tau.p,sqrt(var.tau.p))
[1] 227.65568 36.74815

> # Compute H-T estimate and SE
> pi <- 1 - (1-p)^n                      # Computes the vector of inclusion probs.
> tau.pi <- sum(y/pi)                    # Computes the H-T estimate

> # Compute the estimated variance of the H-T estimate by computing the two terms
> # (the single sum and the double sum separately)
> var1 <- sum(y^2*(1-pi)/pi^2)           # First term of variance of H-T
> # Second term: the multiplier 2 below is because the pair i,j is the same as j,i
> var2 <- 0
> for(i in 1:(n-1)){
+   for(j in (i+1):n){
+     pi.ij <- pi[i] + pi[j] - (1-(1-p[i]-p[j])^n) #joint inclusion probability
+     var2 <- var2 + 2*(y[i]*y[j]/pi.ij)*(pi.ij - pi[i]*pi[j])/(pi[i]*pi[j])
+   }
+ }
> var.tau.pi <- var1 + var2               # Computes the H-T estimated var.
> sd.tau.pi <- sqrt(var.tau.pi)          # Computes the standard dev'n
> c(tau.pi,sd.tau.pi)
[1] 237.93735 36.70255
```