Stat 576
11-14-17

(1)

A generalization of the idea behind PPS sampling

Let π_j be the probability that Y_j is in the sample.

Let π_{ij} be the joint probability that Y_i and Y_j are both in the sample

Let $\hat{t}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ Horvitz-Thompson estimator

(2)

Horvitz-Thompson theorem:

$E(\hat{t}_{HT}) = t$

Proof: $E[\hat{t}_{HT}] = E\left[\sum_{i=1}^{n} \frac{y_i}{\pi_i}\right]$

$= E\left[\sum_{j=1}^{N} \frac{Y_j}{\pi_j} u_j\right]$ Let $u_j = \begin{cases} 1 & \text{if } Y_j \text{ is in sample} \\ 0 \end{cases}$

$u_j \sim \text{Bernoulli}(\pi_j)$

$= \sum_{j=1}^{N} \frac{Y_j}{\pi_j} E(u_j)$

$= \sum_{j=1}^{N} \frac{Y_j}{\cancel{\pi_j}} \cancel{\pi_j} = t$ //

What about the variance?

$$V[\hat{t}_{HT}] = V\left[\sum_{i=1}^{n} \frac{y_i}{\pi_i}\right]$$

$$= V\left[\sum_{j=1}^{N} \frac{y_j}{\pi_j} u_j\right]$$

$$= \sum_{j=1}^{N} \frac{y_j^2}{\pi_j^2} V[u_j] + \sum_{j=1}^{N} \sum_{k \neq j} \frac{y_j}{\pi_j} \frac{y_k}{\pi_k} Cov(u_j, u_k)$$

$$= \sum_{j=1}^{N} \frac{y_j^2}{\pi_j^2} \pi_j(1-\pi_j) + \sum_{j=1}^{N} \sum_{k \neq j} \frac{y_j y_k}{\pi_j \pi_k} \left[E(u_j u_k) - E(u_j)E(u_k)\right]$$

$$= \sum_{j=1}^{N} \frac{y_j^2}{\pi_j^2} \pi_j(1-\pi_j) + \sum_{j=1}^{N} \sum_{k \neq j} \frac{y_j y_k}{\pi_j \pi_k} \left(\pi_{jk} - \pi_j \pi_k\right)$$

This is the true variance of $\hat{t}_{HT}$.

H & T showed 2 unbiased estimators of the variance

$$\hat{V}_1(\hat{t}_{HT}) = \sum_{i=1}^{n} (1-\pi_i) \frac{y_i^2}{\pi_i^2} + \sum_{i=1}^{n} \sum_{j \neq i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j}$$

5

$$\hat{V}_2(\hat{t}_{HT}) = \sum_{i=1}^{n} \sum_{j \neq i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

<u>Example</u>: Population has 3 strata

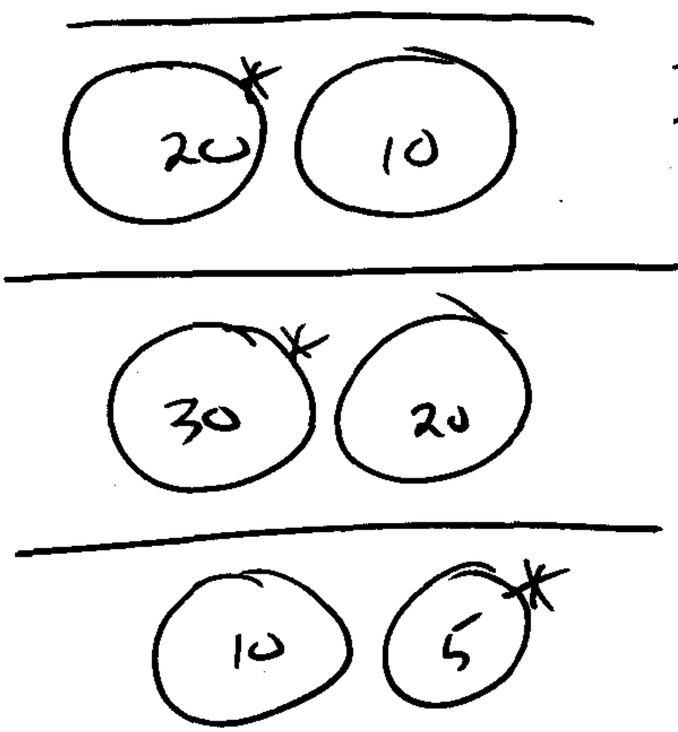
Each stratum has 2 clusters

We pick 1 cluster from each stratum

Select 2 people from each of the sampled clusters.

$n = 6$

6



Item	π_i
y_1	$\frac{1}{2} \cdot \frac{2}{20} = \frac{1}{20}$
y_2	$\frac{1}{2} \cdot 2/20 = \frac{1}{20}$
y_3	$\frac{1}{2} \cdot \frac{2}{30} = 1/30$
y_4	$\frac{1}{2} \cdot \frac{2}{30} = \frac{1}{30}$
y_5	$\frac{1}{2} \cdot \frac{2}{5} = \frac{1}{5}$
y_6	$\frac{1}{2} \cdot 2/5 = \frac{1}{5}$

(7)

$$\hat{t}_{HT} = \sum_{i=1}^{6} \frac{y_i}{\pi_i} = 20(y_1+y_2) + 30(y_3+y_4) + 5(y_5+y_6)$$

For $\hat{V}[\hat{t}_{HT}]$, we also need π_{ij} for every $i \neq j$

If y_i & y_j are from different strata, then $\pi_{ij} = \pi_i \pi_j$

$$\pi_{12} = \frac{1}{2} \frac{1}{\binom{20}{2}} = \frac{1}{2} \frac{2}{20 \cdot 19} = \frac{1}{20(19)}$$

$$\pi_{34} = \frac{1}{30(29)}$$

$$\pi_{56} = \frac{1}{5(4)}$$

Now, plug these into either $\hat{V}_1[\hat{t}_{HT}]$ or $\hat{V}_2[\hat{t}_{HT}]$

(8)

<u>Design effect</u>

$$\text{d.eff.} = \frac{V(\text{estimator} \mid \text{Complex design})}{V(\text{estimator} \mid \text{SRS of same size})}$$

<u>Example</u>: Stratified sampling with proportional allocation

$$V[\bar{y}_{str}] = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \frac{S_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

and $n_h = \frac{N_h}{N} n$

$$V[\bar{y}_{str}] = \frac{1}{N^{\cancel{2}}}\sum_{h=1}^{H} N_h^{\cancel{2}} \frac{S_h^2}{\left(\frac{\cancel{N_h} n}{\cancel{N}}\right)}\left(1 - \frac{\frac{\cancel{N_h} n}{N}}{\cancel{N_h}}\right)$$

$$= \sum_{h=1}^{H} W_h \frac{S_h^2}{n}\left(1 - \frac{n}{N}\right)$$

Also $V[\bar{y}] = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$

$$\Rightarrow \quad d.eff. = \frac{\sum_{h=1}^{H} W_h S_h^2}{S^2}$$

$$\widehat{d.eff.} = \sum_{h=1}^{H} W_h s_h^2 \Big/ s^2$$