Ch. 6: Sampling with unqual probabilities of selection

(without replacement)

Math 255, St. Clair

Horvitz-Thompson Estimator

- You take a random sample where:
  - $\mathcal{S} =$ set of sampled PSU
  - $t_i =$ "response" in PSU i
  - $n =$ PSU sample size (or unique PSU sampled)
  - $\pi_i =$ sample inclusion prob for PSU i
  - $w_i = 1/\pi_i =$ number of population units represented by PSU i
- The Horvitz-Thompson estimator is

$$\hat{t}_{HT} = \sum_{i \in \mathcal{S}} w_i t_i$$

- For any design (with or without replacement), the H-T estimator is an unbiased estimator of population total t. (HW 3 proof!)

$$E(\hat{t}_{HT}) = t = \sum_{i=1}^{N} t_i$$

Horvitz-Thompson Estimator

- All total estimates so far, except for ratio estimates, have been HT estimators
  - SRS: $w_i = N/n$
  - Stratified: $w_{hj} = N_h/n_h$
  - One-stage cluster: $w_{ij} = N/n$
- For these designs, the total estimator SE's can be derived from a general variance calculation
  - using the fact that **without replacement** designs leads to **dependence** among units being sampled

$$Var(\hat{t}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + 2 \sum_{\substack{i=1 \\ i<k}}^{N} \sum_{k=1}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k$$

Horvitz-Thompson Estimator

$$Var(\hat{t}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + 2 \sum_{\substack{i=1 \\ i<k}}^{N} \sum_{k=1}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k$$

- We need to compute the **joint** inclusion probability

$$\pi_{ik} = \pi_{ki} = P(\text{both } i, k \text{ included in the sample})$$

Horvitz-Thompson Estimator

- **SRS:** We measure $t_i = y_i$ for each unit and $\hat{t}_{HT} = N\bar{y}$.

$$\pi_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \qquad \pi_{ik} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

- **SRS:** variance is then

$$Var(\hat{t}_{HT}) = \sum_{i=1}^{N} \frac{1 - \frac{n}{N}}{\frac{n}{N}} y_i^2 + 2 \sum_{\substack{i=1 \\ i<k}}^{N} \sum_{k=1}^{N} \frac{\frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N}}{\frac{n}{N}\frac{n}{N}} y_i y_k$$

$$\ldots\text{lots of algebra}\ldots$$

$$= N^2(1 - \frac{n}{N})\frac{S^2}{n}$$

Horvitz-Thompson Estimator

- All designs covered so far have used a SRS
  - Definition: each sample of size n is equally likely
  - Implication: each PSU is equally likely
- What if we don't use a SRS?
  - Take a random sample of PSU without replacement
  - Let inclusion probs π_i vary
- Our unbiased total estimate is still $\hat{t}_{HT}$
- Our variance is computed using π_i and π_{ik}!

Example: Unequal inclusion probabilities

- **Supermarkets** estimate total sales t for $N = 4$ stores.

$$t_i = \text{total sales (thousands of dollars) at store } i$$

- **Design:** Random sample WOR with probability proportional to physical store size

$$\psi_i = \text{probability of selecting store } i \text{ on your first draw}$$

Store	**Size** m^2	ψ_i	t_i
A	100	1/16	11
B	200	2/16	20
C	300	3/16	24
D	1000	10/16	245
total	1600	1	$t = 300$

Example: Unequal inclusion probabilities

- Take a sample of $n = 2$ stores using selection probs proportional to store size.
  - but do it WOR!

- **Catch:** selection probabilities are conditional on what stores were already selected

- **Draw 1:** we sample store B

Store	**Size** m^2	$\psi_{i\|B} = P(\text{draw 2 is } i \mid \text{draw 1 is } B)$
A	100	1/16/(1-2/16) = 1/14
B	200	0
C	300	3/16/(1-2/16) = 3/14
D	1000	10/16/(1-2/16) = 10/14
total	1600	1

Example: Unequal inclusion probabilities

- Use the **individual PSU** selection probs (draw to draw) to compute the **joint inclusion** prob for each pair
- The probability that both A and B are included is

$$\begin{aligned}\pi_{AB} &= P(A_1)P(B_2 \mid A_1) + P(B_1)P(A_2 \mid B_1) \\ &= \frac{1}{16}\frac{2}{16-1} + \frac{2}{16}\frac{1}{14} \\ &\approx 0.0173\end{aligned}$$

- With $n = 2$, π_{ik} is the probability that sample $\mathcal{S} = \{i, k\}$ is our sample.
- So single PSU inclusion probabilities are the **sum of all probs of samples that contain that PSU**

$$\pi_A = \pi_{AB} + \pi_{AC} + \pi_{AD}$$

Example: Unequal inclusion probabilities

- The matrix below gives joint probs π_{ik} in the body and single PSU probs in the margins

	A	B	C	D	π_i
A	--	0.0173	0.0269	0.1458	0.1900
B	0.0173	--	0.0556	0.2976	0.3705
C	0.0269	0.0556	--	0.4567	0.5393
D	0.1458	0.2976	0.4567	--	0.9002
π_i	0.1900	0.3705	0.5393	0.9002	$n = 2$

- Note that

$$\sum_{i=1}^{N} \pi_i = n$$

Example: Unequal inclusion probabilities

- Suppose you sampled stores C and D

Store	**Size** m^2	π_i	w_i	t_i
C	300	0.5393	1.854	24
D	1000	0.9002	1.111	245

- **Estimated total:** $316.66 thousand

$$\hat{t}_{HT} \approx \frac{24}{0.5393} + \frac{245}{0.9002} = (1.854)(24) + (1.111)(245) = 316.66$$

`

Example: Unequal inclusion probabilities

- **Variance** of $\hat{t}_{HT}$ is

$$Var(\hat{t}_{HT}) = \left(\frac{1-0.1900}{0.1900} 11^2 + \cdots + \frac{1-0.9002}{0.9002} 245^2\right)$$
$$+ 2\left(\frac{0.0173-(0.1900)(0.3705)}{(0.1900)(0.3705)}(11)(20) + \right.$$
$$\left. \cdots + \frac{0.4567-(0.5393)(0.9002)}{(0.5393)(0.9002)}(24)(245)\right) = 4383.6$$

- The estimated total sales is $316.67 thousand with a SE of $66.2 thousand.
- How does this compare to a SRS of $n = 2$ stores?

Example: Unequal inclusion probabilities

- Suppose stores C and D were selected from an SRS.
- **SRS Estimated total:** \$538 thousand

$$\hat{t}_{SRS} = N\bar{y} = 4\frac{24+245}{2} = 538$$

`

- **SRS Variance** of $\hat{t}_{SRS}$ is

$$Var(\hat{t}_{HT}) = 4^2(1-\frac{2}{4})\frac{12874}{2} = 51496$$

where $S^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(t_i - \bar{t}_U) = 12874$

```
> pop <- c(11,20,24,245)
> var(pop)
[1] 12874
```


Example: Unequal inclusion probabilities

- **Prob proportional to size:** The estimated total sales is \$316.67 thousand with a SE of \$66.2 thousand.
- **SRS:** The estimated total sales is \$538 thousand with a SE of \$226.9 thousand.
- One important reason for selecting PSU with unequal probabilities:
  - can reduce SE (compared to SRS) when selection probability π_i is positively associated with the response t_i
  - called **probability proportional to size (pps)** sampling
  - most samples will contain large t_i making variation in $\hat{t}_{pps}$ less than when a small t_i is just as likely as a large

Example: Unequal inclusion probabilities

- One important reason for **not** selecting PSU with unequal probabilities:
  - if some PSU have very small π_i, then they have very high weight w_i
  - can cause imprecise **estimates of** $Var(\hat{t}_{HT})$ because of these high weights

Estimating HT variance

$$Var(\hat{t}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + 2 \sum_{\substack{i=1 \\ i<k}}^{N} \sum_{k=1}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k$$

- To estimate varaince, treat the summations as population totals
  - estimate the total with a HT-estimator!
  - weight sampled values by w_i's

Estimating HT variance

- There are three commonly used estimates of $Var(\hat{t}_{HT})$
- **Horvitz-Thompson (HT)**: unbiased and often the default software version (as in `survey`), but can be negative for samples with small inclusion probs

$$\hat{V}_{HT}(\hat{t}_{HT}) = \sum_{i \in \mathcal{S}} \frac{1-\pi_i}{\pi_i^2} t_i^2 + 2 \sum_{\substack{i \in \mathcal{S} \\ i<k}} \sum_{k \in \mathcal{S}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{t_i}{\pi_i} \frac{t_k}{\pi_k}$$

- **Sen-Yates-Grundy (SYG):** unbiased and more stable than HT version

$$\hat{V}_{SYG}(\hat{t}_{HT}) = \sum_{\substack{i \in \mathcal{S} \\ i<k}} \sum_{k \in \mathcal{S}} \frac{\pi_i \pi_k - \pi_{ik}}{\pi_{ik}} \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2$$

Estimating HT variance

- **With Replacement:** is a biased estimate that overestimates the variance, but it doesn't require joint inclusion probs!

$$\hat{V}_{WR}(\hat{t}_{HT}) = \frac{n}{n-1} \sum_{i \in \mathcal{S}} \left(\frac{t_i}{\pi_i} - \frac{\hat{t}_{HT}}{n} \right)^2$$

Example: Estimating HT variance

Sample	$P(\mathcal{S})$	$\hat{t}_{HT}$	$\hat{V}_{HT}(\hat{t}_{HT})$	$\hat{V}_{SYG}(\hat{t}_{HT})$
A,B	0.01726	111.87	-14,691.5	47.1
A,C	0.02692	102.39	-10,832.1	502.8
A,D	0.14583	330.06	4,659.3	7,939.8
B,C	0.05563	98.48	-9,705.1	232.7
B,D	0.29762	326.15	5,682.8	5,744.1
C,D	0.45673	316.67	6,782.8	3,259.8

- If we happen to sample two small stores, our HT estimate of variance is negative!
- But both are unbiased estimators of the true variance.

What about estimating population mean?

$$\sum_{\textit{all elements}} w_i$$

- Summing sampling weights over all **elements** sampled will give
  - actual population size (of elements) when weights are equal for all elements and number of elements per PSU is constant
  - an unbiased estimated population size (of elements)
- The Horvitz-Thompson estimate of population mean (per element) is

$$\hat{\bar{y}}_{HT} = \frac{\hat{t}_{HT}}{\sum_{\textit{all elements}} w_i}$$

- The `survey` package uses this when you run `svymean`
  - gives $\hat{t}_{HT}$ when you run `svytotal`