Lecture 6: Sampling with unequal probabilities

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

One-stage sampling with replacement

Let

$$\psi_i = P(\text{select unit } i \text{ on first draw}).$$

If we sample with replacement, then ψ_i is also the probability that unit i is selected on the second draw, or the third draw, or any other given draw.

Consider the population of introductory statistics classes at a college shown in Table 6.1. The college has 15 such classes; class i has M_i students, for a total of 647 students in introductory statistics courses. We decide to sample 5 classes with replacement, with probability proportional to M_i, and then collect a questionnaire from each student in the sampled classes. For this example, then, $\psi_i = M_i/647$.

Method 1: Cumulative-size

Generate five random integers with replacement between 1 and 647. Then the psus to be chosen for the sample are those whose range cumulative M_i includes the randomly generated numbers.

Class Number	M_i	ψ_i	Cumulative M_i Range	
1	44	0.068006	1	44
2	33	0.051005	45	77
3	26	0.040185	78	103
4	22	0.034003	104	125
5	76	0.117465	126	201
6	63	0.097372	202	264
7	20	0.030912	265	284
8	44	0.068006	285	328
9	54	0.083462	329	382
10	34	0.052550	383	416
11	46	0.071097	417	462
12	24	0.037094	463	486
13	46	0.071097	487	532
14	100	0.154560	533	632
15	15	0.023184	633	647
Total	647	1		

Lahiri's Method

1. Draw a random number between 1 and N. This indicates which psu you are considering.
2. Draw a random number between 1 and $\max\{M_i\}$. If this random number is less than or equal to M_i, then include psu i in the sample; otherwise go back to step 1.
3. Repeat until desired sample size is obtained.

Lahiri's Method, for Example 6.3

First Random Number (psu i)	Second Random Number	M_i	Action
12	6	24	$6 < 24$; include psu 12 in sample
14	24	100	Include in sample
1	65	44	$65 > 44$; discard pair of numbers and try again
7	84	20	$84 > 20$; try again
10	49	34	Try again
14	47	100	Include
15	43	15	Try again
5	24	76	Include
11	87	46	Try again
1	36	44	Include

Estimates

$$\hat{t}_\psi = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i} = \frac{1}{n}\sum_{i\in\mathcal{R}} u_i = \bar{u}.$$

We estimate $V(\hat{t}_\psi)$ by

$$\hat{V}(\hat{t}_\psi) = \frac{s_u^2}{n} = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}(u_i-\bar{u})^2 = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i}-\hat{t}_\psi\right)^2.$$

Note that both of them are unbiased estimates. That is,

$$E\left(\hat{t}_\psi\right) = t,$$

and

$$E\left(\hat{V}(\hat{t}_\psi)\right) = \mathrm{Var}(\hat{t}_\psi).$$

Weights in unequal-probability sampling with replacement

$$w_i = \frac{1}{\text{expected number of hits}} = \frac{1}{E[Q_i]} = \frac{1}{n\psi_i}.$$

With this choice of weight, we have, for $\hat{t}_\psi$,

$$\hat{t}_\psi = \sum_{i \in \mathcal{R}} w_i t_i.$$

In one-stage cluster sampling with replacement, we observe all of the M_i ssus every time psu i is selected, so we define

$$w_{ij} = w_i = \frac{1}{n\psi_i}.$$

Then, in terms of the elements,

$$\hat{t}_\psi = \sum_{i \in \mathcal{R}} \sum_{j=1}^{M_i} w_{ij} y_{ij}$$

and

$$\hat{\bar{y}}_\psi = \frac{\sum_{i \in \mathcal{R}} \sum_{j=1}^{M_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{R}} \sum_{j=1}^{M_i} w_{ij}}.$$

Two-Stage Sampling with Replacement

- Step 1: Take a sample of psus with replacement, choosing the ith psu with known probability ψ_i. As in one-stage sampling with replacement, Q_i is the number of times psu i occurs in the sample.
- Step 2: Then take a probability sample of m_i subunits in the ith psu (e.g., simple random sampling without replacement).

Difference: We must estimate t_i. If psu i is in the sample more than once, there are Q_i estimators of the total for psu i:

$$\hat{t}_{i1}, \ldots, \hat{t}_{iQ_i}.$$

Two requirements for subsampling:

- Different subsamples from the same psu, though, must be sampled independently.
- The jth subsample taken from psu i (for $j = 1, \ldots, Q_i$) is selected in such a way that

$$E[\hat{t}_{ij}] = t_i.$$

Estimate

The estimate for the total is

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\hat{t}_{ij}}{\psi_i}.$$

$$\hat{V}(\hat{t}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\left(\frac{\hat{t}_{ij}}{\psi_i} - \hat{t}_\psi\right)^2.$$

The estimator of the population mean $\bar{y}_U$ is

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{\hat{M}_{0\psi}},$$

where

$$\hat{M}_{0\psi} = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{M_i}{\psi_i}$$

estimates the total number of elements in the population.

The variance estimator again uses the ratio results

$$\hat{V}(\hat{\bar{y}}_\psi) = \frac{1}{(\hat{M}_{0\psi})^2}\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\left(\frac{\hat{t}_{ij}}{\psi_i} - \frac{\hat{\bar{y}}_\psi M_i}{\psi_i}\right)^2.$$

The weights for the observation units include a factor to reflect the subsampling within each psu. If an SRS of size m_i is taken in psu i, the weight for ssu j in psu i is

$$w_{ij} = \frac{1}{n\psi_i}\frac{M_i}{m_i}.$$

Example

We subsample five students in each class

Class	M_i	ψ_i	y_{ij}	$\bar{y}_i$	$\hat{t}_i$	$\hat{t}_i/\psi_i$
12	24	0.0371	2, 3, 2.5, 3, 1.5	2.4	57.6	1552.8
14	100	0.1546	2.5, 2, 3, 0, 0.5	1.6	160.0	1035.2
14	100	0.1546	3, 0.5, 1.5, 2, 3	2.0	200.0	1294.0
5	76	0.1175	1, 2.5, 3, 5, 2.5	2.8	212.8	1811.6
1	44	0.0680	4, 4.5, 3, 2, 5	3.7	162.8	2393.9
			average			1617.5
			std. dev.			521.628

Thus, $\hat{t}_\psi = 1617.5$ and $\text{SE}(\hat{t}_\psi) = 521.628/\sqrt{5} = 233.28$. From this sample, the average amount of time a student spent studying statistics is

$$\hat{\bar{y}}_\psi = \frac{1617.5}{647} = 2.5$$

hours with standard error 233.28/647 = 0.36 hour.

Classes were selected with probability proportional to number of students in the class, so $\psi_i = M_i/M_0$. Then,

$$w_{ij} = \frac{M_0}{nM_i}\frac{M_i}{5} = \frac{647}{(5)(5)} = 25.88.$$

The population total is equivalently estimated as

$$25.88(2 + 3 + 2.5 + \cdots + 3 + 2 + 5) = 1617.5. \quad \blacksquare$$

Unequal-Probability Sampling Without Replacement

Store	Size (m^2)	ψ_i	t_i (in Thousands)
A	100	$\frac{1}{16}$	11
B	200	$\frac{2}{16}$	20
C	300	$\frac{3}{16}$	24
D	1000	$\frac{10}{16}$	245
Total	1600	1	300

Let's select two psus without replacement and with unequal probabilities.

Probabilities

$$P(\text{store A chosen on first draw}) = \psi_A = \frac{1}{16}$$

and

$$P(\text{B chosen on second draw} \mid \text{A chosen on first draw}) = \frac{\frac{2}{16}}{1 - \frac{1}{16}} = \frac{\psi_B}{1 - \psi_A}.$$

The denominator is the sum of the ψ_i's for stores B, C, and D. In general,

$$\begin{aligned} &P(\text{unit } i \text{ chosen first, unit } k \text{ chosen second}) \\ &\quad = P(\text{unit } i \text{ chosen first})\, P(\text{unit } k \text{ chosen second} \mid \text{unit } i \text{ chosen first}) \\ &\quad = \psi_i \frac{\psi_k}{1 - \psi_i}. \end{aligned}$$

Similarly,

$$P(\text{unit } k \text{ chosen first, unit } i \text{ chosen second}) = \psi_k \frac{\psi_i}{1 - \psi_k}.$$

Probabilities

Inclusion probabilities (π_i) and joint inclusion probabilities (π_{ik}) for samples of size 2. The entries of the table are the π_{ik}'s for each pair of stores (rounded to four decimal places); the margins give the π_i's for the four stores

Store i \ Store k	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_k	0.1900	0.3705	0.5393	0.9002	2.0000

size 2 consists of psus i and k:

$$\text{For } n = 2, \quad P(\text{units } i \text{ and } k \text{ in sample}) = \pi_{ik} = \psi_i \frac{\psi_k}{1 - \psi_i} + \psi_k \frac{\psi_i}{1 - \psi_k}.$$

The probability that psu i is in the sample is then

$$\pi_i = \sum_{S: i \in S} P(S).$$

Why $\sum \pi_i = 2$?

Horvitz-Thompson estimator for one-stage sampling

Assume we have a without-replacement sample of n psus, and we know the **inclusion probability**

$$\pi_i = P(\text{unit } i \text{ in sample})$$

and the **joint inclusion probability**

$$\pi_{ik} = P(\text{units } i \text{ and } k \text{ are both in the sample}).$$

The inclusion probability π_i can be calculated as the sum of the probabilities of all samples containing the ith unit and has the property that

$$\sum^{N} \pi_i = n.$$

For the π_{ik}'s,

$$\sum_{\substack{k=1 \\ k \neq i}}^{N} \pi_{ik} = (n-1)\pi_i.$$

HT estimate

Horvitz–Thompson (HT) estimator of the population total :

$$\hat{t}_{\mathrm{HT}} = \sum_{i \in \mathcal{S}} \frac{t_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{t_i}{\pi_i},$$

where $Z_i = 1$ if psu i is in the sample, and 0 otherwise.

The Horvitz–Thompson estimator is shown to be unbiased

$$E[\hat{t}_{\mathrm{HT}}] = \sum_{i=1}^{N} \pi_i \frac{t_i}{\pi_i} = t.$$

The variance of the HT estimator in one-stage sampling is

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

$$= \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2.$$

The expressions for the variance are algebraically identical.

HT estimate

Note: When the inclusion probabilities π_i or the joint inclusion probabilities π_{jk} are unequal, they are different if the sample quantities are used.

The estimator of the variance suggested by Horvitz and Thompson (1952) is

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

The SYG estimator, is

$$\hat{V}_{\mathrm{SYG}}(\hat{t}_{\mathrm{HT}}) = \frac{1}{2}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}} \frac{\pi_i\pi_k-\pi_{ik}}{\pi_{ik}}\left(\frac{t_i}{\pi_i}-\frac{t_k}{\pi_k}\right)^2.$$

Example: A sample of 2 supermarkets

- Step 1: To select the first psu, we generate a random integer from $\{1, ..., 16\}$: the random integer we generate is 12, which tells us that store D is selected on the first draw.
- Step 2: We then remove the values $\{7, ..., 16\}$ corresponding to store D, and generate a second random integer from $\{1, ..., 6\}$; we generate 6, which tells us to select store C on the second draw.

The Horvitz–Thompson estimate of the total sales for sample {C, D} is then

$$\hat{t}_{\mathrm{HT}} = \sum_{i \in S} \frac{t_i}{\pi_i} = \frac{245}{0.9002} + \frac{24}{0.5393} = 316.6639.$$

Since for this example we know the entire population, we can calculate the theoretical variance of $\hat{t}_{\mathrm{HT}}$:

$$V(\hat{t}_{\mathrm{HT}}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{k \neq i}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2 = 4383.6.$$

Example: A sample of 2 supermarkets

We have two estimates of the variance from sample {C, D}:

$$\hat{V}_{\mathrm{HT}}(\hat{t}_{\mathrm{HT}}) = \frac{(1-0.9002)(245)^2}{(0.9002)^2} + \frac{(1-0.5393)(24)^2}{(0.5393)^2} + 2\frac{0.4567-(0.9002)(0.5393)}{0.4567}\left(\frac{245}{0.9002}\right)\left(\frac{24}{0.5393}\right)$$

$$= 6782.8.$$

The SYG estimate, is

$$\hat{V}_{\mathrm{SYG}}[\hat{t}_{\mathrm{HT}}] = \frac{(0.9002)(0.5393)-0.4567}{0.4567}\left(\frac{245}{0.9002}-\frac{24}{0.5393}\right)^2 = 3259.8.$$

Variance estimates for all possible without-replacement samples of size 2, for the supermarket example

Sample, $\mathcal{S}$	$P(\mathcal{S})$	$\hat{t}_{\mathrm{HT}}$	$\hat{V}_{\mathrm{HT}}(\hat{t}_{\mathrm{HT}})$	$\hat{V}_{\mathrm{SYG}}(\hat{t}_{\mathrm{HT}})$
{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

In some designs, the estimates of the variance can be widely disparate for different samples.

Variance with replacement

An alternative suggested by Durbin (1953), which avoids some of the potential instability and computational complexity, is to pretend the units were selected with replacement and use the with-replacement variance estimator.

The with-replacement variance estimator, setting $\psi_i = \pi_i/n$, is

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

- always nonnegative
- does not require knowledge of the joint inclusion probabilities
- maybe overestimate the variance and result in conservative confidence intervals

In general, we recommend using the with-replacement variance estimator.

Horvitz-Thompson Estimator for Two-Stage Sampling

The Horvitz–Thompson estimator for two-stage sampling is similar to the estimator for one-stage sampling

$$\hat{t}_{\mathrm{HT}} = \sum_{i \in \mathcal{S}} \frac{\hat{t}_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i},$$

where $Z_i = 1$ if psu i is in the sample, and 0 otherwise.

The two-stage Horvitz–Thompson estimator is an unbiased estimator of t as long as $E[\hat{t}_i] = t_i$ for each psu i .

The variance of the HT is

$$\begin{aligned} V(\hat{t}_{\mathrm{HT}}) &= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N} \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k + \sum_{i=1}^{N} \frac{V(\hat{t}_i)}{\pi_i} \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2 + \sum_{i=1}^{N} \frac{V(\hat{t}_i)}{\pi_i}. \end{aligned}$$

Horvitz-Thompson Estimate

The Horvitz–Thompson estimator of the variance in two-stage cluster sampling is

$$\hat{V}_{\mathrm{HT}}(\hat{t}_{\mathrm{HT}}) = \sum_{i\in\mathcal{S}}(1-\pi_i)\frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{\hat{t}_i}{\pi_i}\frac{\hat{t}_k}{\pi_k} + \sum_{i\in\mathcal{S}}\frac{\hat{V}(\hat{t}_i)}{\pi_i},$$

and the SYG estimator is

$$\hat{V}_{\mathrm{SYG}}(\hat{t}_{\mathrm{HT}}) = \frac{1}{2}\sum_{i\in\mathcal{S}}\sum_{\substack{k\in\mathcal{S}\\k\neq i}}\frac{\pi_i\pi_k-\pi_{ik}}{\pi_{ik}}\left(\frac{\hat{t}_i}{\pi_i}-\frac{\hat{t}_k}{\pi_k}\right)^2 + \sum_{i\in\mathcal{S}}\frac{\hat{V}(\hat{t}_i)}{\pi_i}.$$

For most situations, we recommend using the with-replacement sampling variance estimator:

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

Weights in Unequal-Probability Samples

For a without-replacement probability sample of ssus within psus

$$\pi_{j|i} = P(j\text{th ssu in } i\text{th psu included in sample} \mid i\text{th psu is in the sample}).$$

Then,

$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{\pi_{j|i}}.$$

The overall probability that ssu j of psu i is included in the sample is $\pi_{j|i}\pi_i$. Thus, we can define the sampling weight for the (i,j)th ssu as

$$w_{ij} = \frac{1}{\pi_{j|i}\pi_i}$$

The Horvitz–Thompson estimator of the population total as

$$\hat{t}_{\text{HT}} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}.$$

The population mean is estimated by

$$\hat{\bar{y}}_{\text{HT}} = \frac{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}}.$$

Weights in Unequal-Probability Samples

Let

$$\hat{e}_i = \hat{t}_i - \hat{\bar{y}}_{\mathrm{HT}} \hat{M}_i,$$

where $\hat{M}_i = \sum_{j \in \mathcal{S}_i} (1/\pi_{j|i})$ estimates the number of ssus in psu i.

We then use the with-replacement variance to obtain:

$$\hat{V}_{\mathrm{WR}}(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{n}{n-1} \sum_{i \in \mathcal{S}} \left(\frac{\hat{e}_i}{\hat{M}_0 \pi_i} \right)^2 = \frac{n}{n-1} \sum_{i \in \mathcal{S}} \left(\frac{\sum_{j \in \mathcal{S}_i} w_{ij} (y_{ij} - \hat{\bar{y}}_{\mathrm{HT}})}{\sum_{k \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{kj}} \right)^2,$$

where $\hat{M}_0 = \sum_{i \in \mathcal{S}} \hat{M}_i = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij}$ estimates M_0, the number of ssus in the population.

Example: statistics classes

We calculate the weight for each student in the sample as

$$w_{ij} = \frac{1}{\pi_i \pi_{j|i}} = \frac{1}{\pi_i (4/M_i)}.$$

The estimated total number of hours spent studying statistics is

$$\hat{t}_{\mathrm{HT}} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij} = 2232.15.$$

This can also be calculated by $\hat{t}_{\mathrm{HT}} = \sum_{i \in \mathcal{S}} \hat{t}_i / \pi_i = 2232.15$. Using the with-replacement variance estimate

$$\hat{V}_{\mathrm{WR}}(\hat{t}_{\mathrm{HT}}) = \frac{n}{n-1} \sum_{i \in \mathcal{S}} \left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_{\mathrm{HT}}}{n} \right)^2 = \frac{5}{4} 77{,}749.9 = 97{,}187.4,$$

Example: statistics classes

We estimate the mean number of hours spent studying statistics by

$$\hat{\bar{y}}_{\mathrm{HT}} = \frac{\sum\sum_{i\in\mathcal{S}\ j\in\mathcal{S}_i} w_{ij}y_{ij}}{\sum\sum_{i\in\mathcal{S}\ j\in\mathcal{S}_i} w_{ij}} = \frac{2232.15}{647} = 3.45.$$

The estimated variance is

$$\hat{V}_{\mathrm{WR}}(\hat{\bar{y}}_{\mathrm{HT}}) = \frac{n}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{e_i}{\hat{M}_0\pi_i}\right)^2 = \frac{5}{4}(0.18574) = 0.23218,$$

so $\mathrm{SE}(\hat{\bar{y}}_{\mathrm{HT}}) = \sqrt{0.23218} = 0.482.$ ∎