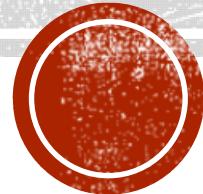


# MH4511: SAMPLING AND SURVEY

Sampling with Unequal Probability



# EXAMPLE: SIX-LEGGED PUPPY

- Suppose we want to estimate the average number of legs on the healthy puppies in Sample City puppy homes.
- Sample City has two puppy homes:
  - Puppy Palace ( $i = 1$ ) with  $30 (M_1)$  puppies, and
  - Dog's Life ( $i = 2$ ) with  $10 (M_2)$  puppies
- Let's select one puppy home with probability  $1/2$  , then select 2 puppies at random from the home.
- Use the unbiased estimator  $\hat{y}_{unb}$  to estimate the average number of legs per puppy

# EXAMPLE: SIX-LEGGED PUPPY

- We know,

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0}, \quad \text{where } \hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^n \hat{t}_i \text{ and } M_0 = \sum_{i=1}^N M_i$$

- **First scenario:** when Puppy Palace is selected (with prob=1/2). Not surprisingly, each of the two puppies sampled has 4 legs.

$$\text{So, } \hat{t}_1 = \frac{M_1}{m_1} \sum_{j=1}^{m_1} y_{ij} = \frac{30}{2}(4 + 4) = 120,$$

$$\text{And, } \hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^n \hat{t}_i = \frac{2}{1}(120) = 240$$

$$\text{Hence, } \hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{240}{30+10} = 6 \text{ (mean number of legs per puppy!)}$$

# EXAMPLE: SIX-LEGGED PUPPY

- We know,

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0}, \quad \text{where } \hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^n \hat{t}_i \text{ and } M_0 = \sum_{i=1}^N M_i$$

- **Second scenario:** when Dog's Life is selected (with prob=1/2). Suppose each of the two puppies sampled also has 4 legs.

$$\text{So, } \hat{t}_2 = \frac{M_2}{m_2} \sum_{j=1}^{m_2} y_{ij} = \frac{10}{2} (4 + 4) = 40,$$

$$\text{And, } \hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^n \hat{t}_i = \frac{2}{1} (40) = 80$$

$$\text{Hence, } \hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{80}{30+10} = 2 \text{ (mean number of legs per puppy!)}$$

- In either scenario, the unbiased estimator cannot give a reasonable estimate!!
- Although it is unbiased:

$$E(\hat{\bar{y}}_{unb}) = \frac{1}{2}(6) + \frac{1}{2}(2) = 4$$

# EXAMPLE: SIX-LEGGED PUPPY

- **First scenario:** when Puppy Palace is selected

$$\hat{t}_1 = \frac{M_1}{m_1} \sum_{j=1}^{m_1} y_{ij} = \frac{30}{2} (4 + 4) = 120,$$

What is not right?

$$\text{And, } \hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^n \hat{t}_i = \frac{2}{1} (120) = 240$$

$$\text{Hence, } \hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{240}{30+10} = 6 \text{ (mean number of legs per puppy!)}$$

- **Second scenario:** when Dog's Life is selected

$$\hat{t}_2 = \frac{M_2}{m_2} \sum_{j=1}^{m_2} y_{ij} = \frac{10}{2} (4 + 4) = 40,$$

$$\text{And, } \hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^n \hat{t}_i = \frac{2}{1} (40) = 80$$

$$\text{Hence, } \hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{80}{30+10} = 2 \text{ (mean number of legs per puppy!)}$$

# EXAMPLE: SIX-LEGGED PUPPY

- **First scenario:** when Puppy Palace is selected

$$\hat{t}_1 = \frac{M_1}{m_1} \sum_{j=1}^{m_1} y_{ij} = \frac{30}{2} (4 + 4) = 120,$$

And,  $\hat{t}_{unb} = \cancel{\frac{M_0}{m_0} \sum_{i=1}^{m_0} t_{ij}} = \cancel{\frac{4}{3}} (120) = \textcircled{160}$

Hence,  $\hat{y}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{160}{30+10} = 4$

$$\frac{4}{3} = \frac{M_0}{M_1} = \frac{1}{M_1/M_0}$$

- **Second scenario:** when Dog's Life is selected

$$\hat{t}_2 = \frac{M_2}{m_2} \sum_{j=1}^{m_2} y_{ij} = \frac{10}{2} (4 + 4) = 40,$$

And,  $\hat{t}_{unb} = \cancel{\frac{M_0}{m_0} \sum_{i=1}^{m_0} t_{ij}} = \cancel{\frac{4}{1}} (40) = \textcircled{160}$

$$\frac{4}{1} = \frac{M_0}{M_2} = \frac{1}{M_2/M_0}$$

Hence,  $\hat{y}_{unb} = \frac{\hat{t}_{unb}}{M_0} = \frac{160}{30+10} = 4$

# SAMPLING WITH EQUAL PROBABILITY

- Up till now, we have only discussed sampling schemes in which the probabilities of choosing sampling units are equal
- Such schemes are not, however, always possible or, if practicable, as efficient as schemes using unequal probabilities
- We will discuss the use of unequal probability of selection to decrease variance in this chapter

# INTRODUCTORY EXAMPLES

- Consider a survey of nursing home residents in a city to determine preferences on life-sustaining treatments
- The target population was all residents of licensed nursing homes in a city
- There were **294** such homes, with a total of **37,652** beds (number of residents not known before sampling)
- Since the survey was to be done in person, cluster sampling was essential for keeping survey costs manageable. Suppose that we choose SRS of nursing homes, then another SRS of residents within each selected home

# INTRODUCTORY EXAMPLES

- If we apply a cluster sample with equal probabilities as in last chapter, a nursing home with 20 beds is as likely to be chosen for the sample as a nursing home with 10,000 beds
- Note that the sample is only self-weighting if the subsample size for each home is proportional to the number of beds in the home, say  $m_i = M_i * 10\%$
- Recall that **self-weighting** means every unit has the same sampling weight, or same probability of being selected
- In this case,

$$\pi_{ij} = P(\text{unit } j \text{ of cluster } i \text{ is in the sample}) = \frac{n}{N} \times \frac{m_i}{M_i}$$

and,  $\omega_{ij} = \frac{1}{\pi_{ij}} = \frac{N}{n} \times \frac{M_i}{m_i}$  is constant, only if  $\frac{M_i}{m_i}$  is a constant

# INTRODUCTORY EXAMPLES

- There are 3 major shortcomings:
  - First, you would expect  $t_i$  to be proportional to the number of beds in nursing home  $i$ , so estimators from the last chapter may have large variance
  - Equal probability scheme may be difficult to administer (if the number of beds in one nursing home is small, it may require driving out to this nursing home just to interview one or two residents)
  - The cost of the sample is unknown in advance. For example, a random sample of 40 homes may consist primarily of large nursing homes, which lead to greater expense than expected  
(large nursing homes -> large subsample sizes -> greater cost;  
small nursing homes -> small subsample sizes -> less cost)

# INTRODUCTORY EXAMPLE

Sampling with unequal probabilities:

- Instead of taking a cluster sample of homes with equal probabilities, the investigators randomly drew a sample of 57 nursing homes with **probabilities proportional to the number of beds**
- Then they took an SRS of 30 beds (and their occupants) from a list of all beds within each selected nursing home.

**Equal probabilities:**

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^n \hat{t}_i = \frac{N}{n} \sum_{i=1}^n \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} \Rightarrow \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{N M_i}{n 30} y_{ij}$$

# INTRODUCTORY EXAMPLE

Sampling with unequal probabilities:

- Instead of taking a cluster sample of homes with equal probabilities, the investigators randomly drew a sample of 57 nursing homes with **probabilities proportional to the number of beds**
- Then they took an SRS of 30 beds (and their occupants) from a list of all beds within each selected nursing home.

Proportional probabilities:

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i=1}^n \hat{t}_i = \frac{N}{n} \sum_{i=1}^n \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} \Rightarrow \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{M_0}{n M_i} \frac{M_i}{30} y_{ij}$$

To select 1 cluster ( $i$ ) out of  $N$  clusters the prob is  $\frac{M_i}{M_0}$ .  
For  $n$  clusters, we have  $\frac{nM_i}{M_0}$ .

- Note that it is self-weighting if the number of beds = the number of residents; because  $\pi_{ij} = \frac{nM_i}{\sum M_i} \times \frac{30}{M_i} = \frac{30n}{\sum M_i} = \frac{30n}{M_0}$
- As a consequence, the cost is known before selecting the sample and the estimators will have smaller variance

# KEY IDEAS OF UNEQUAL PROBABILITY SAMPLING

- We deliberately select psu's with known but unequal probability
- We then compensate the unequal probabilities by suitable weights (unequal) in analysis and estimation.
- The key is that we know the selection probabilities
- For many populations with clustering, unequal probability sampling mirrors the population better
- Define

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

- And

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

# SAMPLING ONE PSU

- To illustrate the ideas, we discuss the simplest case when we select just one ( $n = 1$ ) of the  $N$  clusters to be in the sample.
- Let us start the situation in which we know the whole population.
- Example: A town has four supermarkets ranging in size from  $100m^2$  to  $1000m^2$ . We want to estimate the total amount of sales ( $t$ ) in the four stores by sampling just one store.
- We might expect that a large store would have more sales than a smaller store and the variability also varies.
- We sample only one store with probability proportional to the size of the store (unequal probabilities).

# SAMPLING ONE PSU (SAMPLING PROBABILITIES)

- Note that  $\psi_i = \pi_i$  in this example since we only sample one store, the probability that a store is selected on the first draw ( $\psi_i$ ) is the same as the probability that the store is included in the sample ( $\pi_i$ ).

$$\pi_i = \psi_i = P(\text{store } i \text{ is selected})$$

- Population data:

| Store | Size( $m^2$ ) | $\psi_i$ | $t_i$ (in thousands) |
|-------|---------------|----------|----------------------|
| A     | 100           | 1/16     | 11                   |
| B     | 200           | 2/16     | 20                   |
| C     | 300           | 3/16     | 24                   |
| D     | 1000          | 10/16    | 245                  |
| Total | 1600          | 1        | 300                  |

# SAMPLING ONE PSU (SAMPLING WEIGHT)

- Since Store A accounts for 1/16 of the total floor area of the four stores, it is sampled with probability 1/16.
- For illustrative purposes, we know the values of  $t_i$  for the whole population.
- Generally, we could select a sample of size 1 with the probabilities given above.
- Weight: we compensate for the unequal probabilities of selection by also using  $\psi_i$  in the estimator.
- As always, the sampling weight of unit  $i$  is the reciprocal of the probability of selection:  $w_i = 1/\psi_i$
- It follows that the weighted estimator of the population total is

$$\hat{t}_\psi = \sum_{i \in S} w_i t_i = \sum_{i \in S} \frac{t_i}{\psi_i}$$

Instead  
of  $N/n$  in  
SRS

# SAMPLING ONE PSU (TOTAL ESTIMATOR)

- Why do we use such a weight?

Here we select Store A with probability 1/16, so Store A's sampling weight is 16.

If the size of the store is roughly proportional to the total sales for the store, we would expect that Store A also has about 1/16 of the total sales, and that multiplying Store A's sales by 16 would estimate the total sales for all four stores.

- Samples: four sample of size 1 are possible from this sample population. (Note that  $t = 300$ )

| Sample | $\psi_i$ | $t_i$ | $\hat{t}_\psi$ | $(\hat{t}_\psi - t)^2$ |
|--------|----------|-------|----------------|------------------------|
| A      | 1/16     | 11    | 176            | 15376                  |
| B      | 2/16     | 20    | 160            | 19600                  |
| C      | 3/16     | 24    | 128            | 29584                  |
| D      | 10/16    | 245   | 392            | 8464                   |

# SAMPLING ONE PSU (PROPERTIES OF ESTIMATOR)

- Expectation:

$$E(\hat{t}_\psi) = \sum_{\text{possible samples } \mathcal{S}} P(\mathcal{S}) \hat{t}_{\psi, \mathcal{S}} = \sum_{i=1}^4 \psi_i \hat{t}_{\psi, i}$$

$$= \frac{1}{16} \times 176 + \frac{2}{16} \times 160 + \frac{3}{16} \times 128 + \frac{10}{16} \times 392 = 300$$

So,  $\hat{t}_\psi$  is unbiased. This will be always true because in general,

$$E(\hat{t}_\psi) = \sum_{\mathcal{S}} P(\mathcal{S}) \hat{t}_{\psi, \mathcal{S}} = \sum_{i=1}^N \psi_i \hat{t}_{\psi, i} = \sum_{i=1}^N \psi_i w_i t_i = \sum_{i=1}^N \psi_i \frac{t_i}{\psi_i} = \sum_{i=1}^N t_i = t$$

# SAMPLING ONE PSU (PROPERTIES OF ESTIMATOR)

- Variance:

$$Var(\hat{t}_\psi) = E(\hat{t}_\psi - t)^2 = \sum_{\text{possible samples } \mathcal{S}} P(\mathcal{S}) (\hat{t}_{\psi,\mathcal{S}} - t)^2$$

$$= \sum_{i=1}^4 \psi_i (\hat{t}_{\psi,i} - t)^2$$

$$= \frac{1}{16}(15376) + \frac{2}{16}(19600) + \frac{3}{16}(29584) + \frac{10}{16}(8464) = 14,248$$

# SAMPLING ONE PSU (COMPARE WITH SRS)

- Now, compare with SRS with size 1 ( $n = 1$ ), in which the probability if selecting each unit is  $\psi_i = \frac{1}{4}$ , so  $\frac{1}{\psi_i} = 4 = N$
- For SRS,

| Sample | $\psi_i$      | $t_i$ | $\hat{t}_{SRS} = \hat{t}_\psi$ | $(\hat{t}_{SRS} - t)^2$ |
|--------|---------------|-------|--------------------------------|-------------------------|
| A      | $\frac{1}{4}$ | 11    | 44                             | 65536                   |
| B      | $\frac{1}{4}$ | 20    | 80                             | 48400                   |
| C      | $\frac{1}{4}$ | 24    | 96                             | 41616                   |
| D      | $\frac{1}{4}$ | 245   | 980                            | 462400                  |

- As always,

$$E(\hat{t}_{SRS}) = \sum_{\text{possible samples } \mathcal{S}} P(\mathcal{S}) \hat{t}_{SRS,\mathcal{S}} = \sum_{i=1}^4 \psi_i \hat{t}_{SRS,i} = 300$$

is unbiased, but for this example the SRS variance is much larger than that from the unequal-probability design:

$$\text{Var}(\hat{t}_{SRS}) = \frac{1}{4}(65536) + \frac{1}{4}(48400) + \frac{1}{4}(41616) + \frac{1}{4}(462400) = 154,488$$

# SAMPLING ONE PSU (INTERPRETATION)

- We interpret the unequal-probability design as follows:
- Store D is the largest and we expect it to account for a large portion of the total sales.
- Therefore we give it a higher probability of being in the sample ( $10/16$ ) than it would have with an SRS ( $1/4$ ).
- If it is selected, we multiply its sales by  $(16/10)$  to estimate total sales.

# ONE-STAGE SAMPLING WITH REPLACEMENT

- Now suppose  $n > 1$ , and we sample with replacement
- In this case, probability that item  $i$  is selected on the first draw (denoted by  $\psi_i$ ) is the same as the probability that item  $i$  is selected on any other draw
- This further implies that

$$\pi_i = 1 - P(\text{unit } i \text{ is not in the sample}) = 1 - (1 - \psi_i)^n$$

Which also implies that  $\pi_i = \psi_i$  when  $n = 1$

# THE IDEA BEHIND UNEQUAL-PROBABILITY SAMPLING

- The idea behind unequal-probability sampling is simple
  - Sampling with replacement gives us  $n$  independent estimates of the population total, one for each separate psu drawn
  - Estimate the population total  $t$  by averaging those  $n$  independent estimate of  $t$
  - The estimated variance is the sample variance of the  $n$  independent estimates of  $t$ , divided by  $n$
- There are several ways to sample psus with unequal probabilities. All require that we have a measure of size for all psus in the population
- We next introduce 2 approaches to sample psus

# THE CUMULATIVE-SIZE METHOD

- Definition: the cumulative-size method generate random numbers and psus corresponding to those numbers **that** are included in the sample.
- Example: Consider the population of Introductory Statistics classes at a college show in the table below:
  - The college has 15 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 size  $M_i$  (**pps**) and then collect a questionnaire from each student in the sampled classes.
  - For this example, we have  $\psi_i = M_i/647$

| Class Number | $M_i$ | $\psi_i$ | 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    | 0.117464 | 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        |                        |     |

# HOW DOES IT WORK?

- To select the sample, generate 5 random integers with replacement between 1 and 647
- Each random number corresponds to a student and the corresponding class will be in the sample
- For example the five random numbers {553, 82, 245, 594, 150} leads to the sample {14, 3, 6, 14, 5}
- The cumulative-size method allows the same unit to appear in the sample more than once; psu 14 is included twice in the data

| Class Number | $M_i$ | $\psi_i$ | 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    | 0.117464 | 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        |                        | 594 |

# CUMULATIVE-SIZE METHOD



Each PSU is selected with probability  $\psi_i = \frac{M_i}{M_0}$

# LAHIRI'S METHOD

- Alternatively, we may use Lahiri's method which may be more tractable than the cumulative-size method when the number of psus is large
- It is a rejective method, we generate pairs of random numbers to determine whether the psus are selected or not
- The method is given as follows:
  - Draw a random number between 1 and  $N$ . This indicates which psu you are considering.
  - Suppose psu  $i$  is being considered, draw a random number between 1 and  $\max\{M_1, M_2, \dots, M_N\}$ . If this random number is less than or equal to  $M_i$ , then include psu  $i$  in the sample; otherwise go back step 1
  - Repeat until desired sample size is obtained
- Notice that  $\max\{M_1, M_2, \dots, M_N\}$  is just a constant, and the probability of selecting  $i$ -th cluster is proportional to  $M_i$ .

$$\text{Prob} \propto \frac{M_i}{\max\{M_1, M_2, \dots, M_N\}}$$

# EXAMPLE

- Let's use Lahiri's method to select a sample in the preceding example
- The largest class has  $\max\{M_1, M_2, \dots, M_N\} = 100$  students
- So, we generate pairs of random integers
  - First between 1 and 15 (to indicate which psu we are considering)  
 $F_i$  such that  $\Pr(F_i = f_i) = 1/15$  for  $f_i = 1, 2, \dots, 15$
  - Second between 1 and 100 (to decide whether to include this psu)  
 $S_i$  such that  $\Pr(S_i = s_i) = 1/100$  for  $s_i = 1, 2, \dots, 100$
- For each pair of  $(f_i, s_i)$ , we decide whether to include or not to include psu  $f_i$  in the sample, until the sample has 5 psus

# EXAMPLE

| $f_i$                 | $s_i$       | $M_{f_i}$ | Action                               |
|-----------------------|-------------|-----------|--------------------------------------|
| First r.n. (psu $i$ ) | Second r.n. |           |                                      |
| 12                    | 6           | 24        | $6 < 24$ ; include psu 12 in sample  |
| 14                    | 24          | 100       | include in sample                    |
| 1                     | 65          | 44        | $65 > 44$ ; discard and try it 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                              |

From the table we can see that the psus to be sampled are  $\{12, 14, 14, 5, 1\}$ .

# LAHIRI'S METHOD

Step 1: Each PSU is selected for consideration with probability  $\frac{1}{N}$



Step 2: For each PSU  $i$  under consideration, it is selected with probability  $\frac{M_i}{M_{max}}$



# THE THEORY

- Estimator of the population total  $t$  :

- Define the random variable  $Q_i$ ,

$Q_i$  = the number of times unit  $i$  occurs in the sample

Notice that  $Q_i \sim B(n, \psi_i)$

- For sampling with replacement, we may have a psu that appears more than once in the sample
  - $Q_i$  is a with-replacement analogue of the random variable  $Z_i$  used to indicate sample inclusion for without-replacement sampling
  - We learnt that for sample of size 1 ( $n = 1$ ),  $t_i/\psi_i$  is an unbiased estimator of the population total  $t$
  - When sampling  $n$  psus with replacement, we have  $n$  independent estimators of  $t$ , so we average them

$$\hat{t}_\psi = \frac{1}{n} \sum_{i \in S} \frac{t_i}{\psi_i} = \frac{1}{n} \sum_{i=1}^N Q_i \frac{t_i}{\psi_i}$$

# BIAS AND VARIANCE

- Unbiased:

Note that  $\sum_{i=1}^N Q_i = n$ , and  $E(Q_i) = n \cdot \psi_i$ . It follows that

$$E(\hat{t}_\psi) = \frac{1}{n} \sum_{i=1}^N E(Q_i) \frac{t_i}{\psi_i} = \sum_{i=1}^N t_i = t$$

- Variance:

The estimator  $\hat{t}_\psi$  is the average of  $n$  independent observations  $\left(\frac{t_i}{\psi_i}\right)_{i \in \mathcal{S}}$  of size 1, each with variance  $\sum_{i=1}^N \psi_i (\hat{t}_{\psi,i} - t)^2$ , (slide #19), so

$$\text{Var}(\hat{t}_\psi) = \frac{1}{n} \sum_{i=1}^N \psi_i (\hat{t}_{\psi,i} - t)^2 = \frac{1}{n} \sum_{i=1}^N \psi_i \left( \frac{t_i}{\psi_i} - t \right)^2$$

The estimated variance is

$$\widehat{\text{Var}}(\hat{t}_\psi) = \frac{1}{n(n-1)} \sum_{i \in \mathcal{S}} \left( \frac{t_i}{\psi_i} - \hat{t}_\psi \right)^2 = \frac{1}{n(n-1)} \sum_{i=1}^N q_i \left( \frac{t_i}{\psi_i} - \hat{t}_\psi \right)^2$$

# UNBIASEDNESS OF VARIANCE ESTIMATE

First, we note that

$$\begin{aligned} \sum_{i \in S} \left( \frac{t_i}{\psi_i} - \hat{t}_\psi \right)^2 &= \sum_{i \in S} \left( \left[ \frac{t_i}{\psi_i} - t \right] - [\hat{t}_\psi - t] \right)^2 \\ &= \sum_{i \in S} \left[ \frac{t_i}{\psi_i} - t \right]^2 - 2 \sum_{i \in S} \left[ \frac{t_i}{\psi_i} - t \right] [\hat{t}_\psi - t] + \sum_{i \in S} [\hat{t}_\psi - t]^2 \\ &= \sum_{i \in S} \left[ \frac{t_i}{\psi_i} - t \right]^2 - 2[n\hat{t}_\psi - nt][\hat{t}_\psi - t] + n[\hat{t}_\psi - t]^2 \\ &= \sum_{i \in S} \left[ \frac{t_i}{\psi_i} - t \right]^2 - n[\hat{t}_\psi - t]^2 \end{aligned}$$

# UNBIASEDNESS OF VARIANCE ESTIMATE

Then

$$\begin{aligned} E[\widehat{Var}(\hat{t}_\psi)] &= \frac{1}{n(n-1)} E \left[ \sum_{i \in \mathcal{S}} \left( \frac{t_i}{\psi_i} - \hat{t}_\psi \right)^2 \right] \\ &= \frac{1}{n(n-1)} E \left[ \sum_{i \in \mathcal{S}} \left[ \frac{t_i}{\psi_i} - t \right]^2 - n[\hat{t}_\psi - t]^2 \right] \\ &= \frac{1}{n(n-1)} E \left[ \sum_{i \in \mathcal{S}} \left[ \frac{t_i}{\psi_i} - t \right]^2 \right] - \frac{1}{n(n-1)} n E[\hat{t}_\psi - t]^2 \\ &= \frac{1}{n(n-1)} \sum_{i=1}^N E[Q_i] \left( \frac{t_i}{\psi_i} - t \right)^2 - \frac{1}{(n-1)} Var(\hat{t}_\psi) \end{aligned}$$

# UNBIASEDNESS OF VARIANCE ESTIMATE

So

$$\begin{aligned} E[\widehat{Var}(\hat{t}_\psi)] &= \frac{1}{n(n-1)} E \left[ \sum_{i \in \mathcal{S}} \left( \frac{t_i}{\psi_i} - \hat{t}_\psi \right)^2 \right] \\ &= \frac{1}{n(n-1)} \sum_{i=1}^N E[Q_i] \left( \frac{t_i}{\psi_i} - t \right)^2 - \frac{1}{(n-1)} Var(\hat{t}_\psi) \\ &= \frac{1}{n(n-1)} \sum_{i=1}^N n\psi_i \left( \frac{t_i}{\psi_i} - t \right)^2 - \frac{1}{(n-1)} Var(\hat{t}_\psi) \\ &= \frac{1}{n(n-1)} n nVar(\hat{t}_\psi) - \frac{1}{(n-1)} Var(\hat{t}_\psi) \\ &= \frac{n}{(n-1)} Var(\hat{t}_\psi) - \frac{1}{(n-1)} Var(\hat{t}_\psi) = Var(\hat{t}_\psi) \end{aligned}$$

# MEAN ESTIMATION

- We estimate the population mean  $\bar{y}_U$  by

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

Where  $\hat{M}_{0\psi} = \frac{1}{n} \sum_{i \in S} \frac{M_i}{\psi_i}$  estimate the total number of elements in the population, when  $M_0$  is not given

- Estimated variance:

Notice that  $\hat{\bar{y}}_\psi$  is a ratio estimator, using results in ratio estimation,

$$\widehat{Var}(\hat{\bar{y}}_\psi) = \frac{1}{(\hat{M}_{0\psi})^2} \frac{1}{n(n-1)} \sum_{i \in S} \left( \frac{t_i}{\psi_i} - \hat{\bar{y}}_\psi \frac{M_i}{\psi_i} \right)^2$$

Note that: we use  $\frac{t_i}{\psi_i}$  as  $y_i$ ,  $\frac{M_i}{\psi_i}$  as  $x_i$ ,  $\hat{\bar{y}}_\psi$  as  $\hat{B}$

# EXAMPLE

- For the Introductory Statistics classes example, suppose that we sample the psus selected by Lahiri's method, {12, 14, 14, 5, 1}
- The response is  $t_i$ , the number of hours all students in class  $i$  spent studying statistics last week, with the following data:

| Class | $\psi_i$ | $t_i$ | $t_i/\psi_i$ |
|-------|----------|-------|--------------|
| 12    | 24/647   | 75    | 2021.875     |
| 14    | 100/647  | 203   | 1313.410     |
| 14    | 100/647  | 203   | 1313.410     |
| 5     | 76/647   | 191   | 1626.013     |
| 1     | 44/647   | 168   | 2470.364     |

# EXAMPLE

- **Estimation of population total:**

The numbers in the last column of the table are the estimate of that would be obtained if the psu were the only one selected in a sample of size 1.

The population total is estimated by averaging the five values of  $t_i/\psi_i$ ,

$$\hat{t}_\psi = \frac{1}{n} \sum_{i \in S} \frac{t_i}{\psi_i} = \frac{2012.875 + 1313.410 + \dots + 2470.364}{5} = 1749.014$$

the standard error of  $\hat{t}_\psi$  is

$$SE(\hat{t}_\psi) = \sqrt{\frac{1}{n(n-1)} \sum_{i \in S} \left( \frac{t_i}{\psi_i} - \hat{t}_\psi \right)^2}$$
$$= \sqrt{\frac{(2012.875 - 1749.014)^2 + \dots + (2470.364 - 1749.014)^2}{20}} = 222.42$$

# EXAMPLE

- Estimation of population mean:

Since  $\psi_i = M_i/M_0$  for this example, we have  $\hat{M}_{0\psi} = M_0 = 647$ . That is,  $M_0$  is known.

The average amount of time a student spent studying statistics is estimated as

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{M_0} = \frac{1749.014}{647} = 2.70 \text{ (hrs)}$$

Since  $\hat{M}_{0\psi} = M_0$  is known, the  $Var(\hat{\bar{y}}_\psi)$  on slide #38 simplifies to

$$\widehat{Var}(\hat{\bar{y}}_\psi) = \frac{\widehat{Var}(\hat{t}_\psi)}{M_0^2}$$

So, the standard error of  $\hat{\bar{y}}_\psi$  is  $\frac{222.42}{647} = 0.34$ .

# DESIGNING THE SELECTION PROBABILITIES (PPS)

- We can take  $\psi_i$  proportional to  $M_i$  or some other measure of the size of psu  $i$ .
- This is called probability proportional to size (pps) sampling.
- In this case

$$\psi_i = \frac{M_i}{M_0}, \quad M_0 = \sum_{i=1}^N M_i$$

And

$$\frac{t_i}{\psi_i} = \frac{t_i M_0}{M_i} = M_0 \bar{y}_i$$

So

$$\hat{t}_\psi = \frac{1}{n} \sum_{i \in S} M_0 \bar{y}_i$$

# PPS

And

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{M_0} = \frac{1}{n} \sum_{i \in S} \bar{y}_i$$

Note that

$$\widehat{Var}(\hat{\bar{y}}_\psi) = \frac{1}{n(n-1)} \sum_{i \in S} (\bar{y}_i - \hat{\bar{y}}_\psi)^2 = \frac{1}{n} \left[ \frac{1}{(n-1)} \sum_{i \in S} (\bar{y}_i - \hat{\bar{y}}_\psi)^2 \right]$$

Which is of the form  $s^2/n$  with  $s^2$  being the sample variance of the psu means  $\bar{y}_i$ .

Note: We used pps sampling in the example about Introductory Statistics classes

# TWO-STAGE SAMPLING WITH REPLACEMENT

- The basic ideas are very similar to 1-stage sampling
- Take a sample of psus **with replacement**, choosing the  $i$ -th psu with known probability  $\psi_i$
- We then take sample of  $m_i$  ssus from each selected psu by an SRS **without replacement** of systematic sampling
- The only difference between 2-stage sampling with replacement and 1-stage sampling with replacement is that we must estimate  $t_i$  in 2-stage sampling

# SUBSAMPLING PROCEDURE

- If psu  $i$  is in the sample more than once, there are  $Q_i$  estimators of the total for psu  $i$  :  $\hat{t}_{i1}, \hat{t}_{i2}, \dots, \hat{t}_{iQ_i}$
- The subsampling procedure needs to meet two requirements:
  - If a psu is selected more than once, then a separate independent second stage sample is required.
  - The  $j$ -th subsample taken from psu  $i$  ( $j = 1, 2, \dots, Q_i$ ) is selected with
$$E(\hat{t}_{ij}) = t_i, \quad \text{and } Var(\hat{t}_{ij}) = V_i \quad \text{for all } j$$
- For example, if we take one subsample of size 5, and use it more than once for psu 1, we will not have independent subsamples

# ESTIMATING TOTAL AND MEAN

- The estimator of the population total is

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

- Estimated variance is:

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

- In a pps sample  $(\psi_i = \frac{M_i}{M_0})$  the estimator of the population mean is

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{M_0}$$

- With estimated variance

$$\widehat{Var}(\hat{\bar{y}}_\psi) = \widehat{Var}(\hat{t}_\psi) / M_0^2$$

# EXAMPLE

- Let's return to the Introductory Statistics classes examples
- Suppose we subsample 5 students in each class rather than observing  $t_i$
- The response  $y_{ij}$  is the number of hours student  $j$  in class  $i$  spent studying statistics last week

| Class | $M_i$ | $\psi_i$ | $y_{ij}$              | $\bar{y}_i$ | $\hat{t}_i = M_i \bar{y}_i$ | $\hat{t}_i/\psi_i$ |
|-------|-------|----------|-----------------------|-------------|-----------------------------|--------------------|
| 12    | 24    | 24/647   | 2, 3, 2.5, 3, 1.5     | 2.4         | 57.6                        | 1552.8             |
| 14    | 100   | 100/647  | 2.5, 2, 3, 0, 0.5     | 1.6         | 160                         | 1035.2             |
| 14    | 100   | 100/647  | 3, 0.5, 1.5, 2, 3     | 2.0         | 200                         | 1294.0             |
| 5     | 76    | 76/647   | 1, 2.5, 3.5, 2.5, 3.1 | 2.52        | 191.5                       | 1630.4             |
| 1     | 44    | 44/647   | 4, 4.5, 3, 2, 5       | 3.7         | 162.8                       | 2393.9             |
|       |       |          |                       |             | Average<br>Std Dev          | 1581.3<br>510.96   |

# EXAMPLE

- Note that class 14 appears twice in the sample; each time it appears, a different subsample of size five is collected
- Thus,

$$\hat{t}_\psi = 1581.3, \quad SE(\hat{t}_\psi) = \frac{510.96}{\sqrt{5}} = 228.51$$

- From this sample the average amount of time a student spent studying statistics is

$$\hat{\bar{y}}_\psi = \frac{\hat{t}_\psi}{M_0} = \frac{1581.3}{647} = 2.4 \text{ (hrs)}$$

- With standard error  $\frac{228.51}{647} = 0.35 \text{ (hrs)}$

# SUMMARY

Here are the steps for taking a 2-stage unequal-probability sample with replacement:

- 1) Determine  $\psi_i$ ,  $n$  and the subsampling procedure to be used within each psu (often SRS without replacement within each selected psu)
- 2) Select  $n$  psus with probabilities  $\psi_i$  and with replacement (either use cumulative-size method or Lahiri's method)
- 3) Use the procedure determined in step 1) to select subsamples from the psus chosen. Independent subsamples are used if a psu occurs more than once

# SUMMARY

- 4) Estimate the population total  $t$  from each psu in the sample as though it were the only one selected. The results is  $n$  estimates of the form  $t_{ij}/\psi_i$ , ( $i = 1, \dots, N$ ;  $j = 1, 2, \dots, Q_i$ )
- 5)  $\hat{t}_\psi$  is the average of the  $n$  estimates in the above step
- 6)  $SE(\hat{t}_\psi) = \frac{1}{\sqrt{n}} \times [\text{sample standard deviation of the } n \text{ estimates in step 4})]$