Stat 576

5-15-25

(1)

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

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

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

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

(2)

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

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

H & T showed 2 unbiased estimators of the variance

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

$$\hat{V}_2(\hat{t}_{HT}) = \frac{1}{2}\sum_{i=1}^{n} \sum_{j \neq i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$ 3

Example: Population has 3 strata

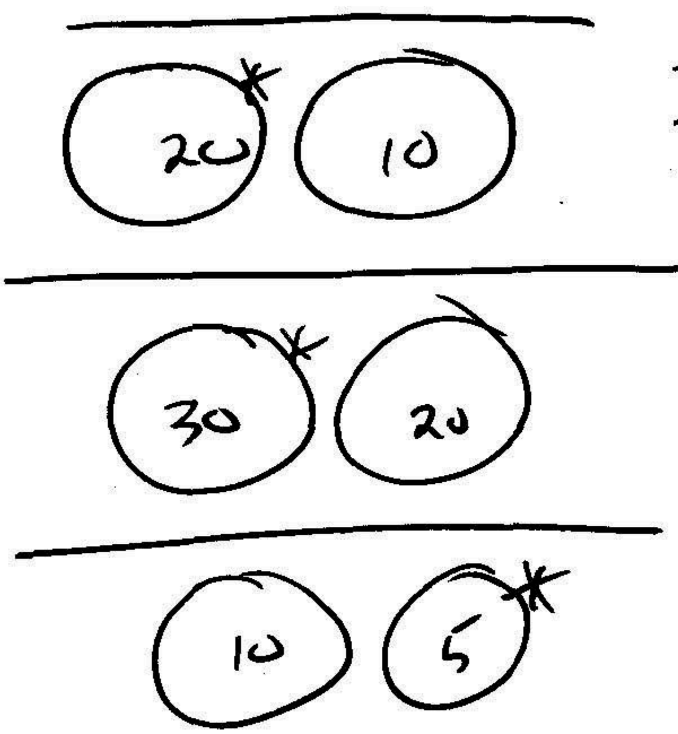
Each stratum has 2 clusters

We pick 1 cluster from each stratum

Select 2 people from each of the sampled clusters.

$n = 6$

4



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	$\frac{1}{2} \cdot 2/5 = \frac{1}{5}$

5

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

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 $\hat{V}_2[\hat{t}_{HT}]$

6

PPS sampling, cumulative range method

TABLE 6.1
Population of Introductory Statistics Classes

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		

Suppose these 3-digit random numbers are generated: {487, 369, 221, 326, 282}

Then these classes would be chosen: {13, 9, 6, 8, 7}

7

Lahiri's Method

Lahiri's (1951) method may be more tractable than the cumulative-size method when the number of psus is large. It is an example of a *rejective* method, because you generate pairs of random numbers to select psus and then reject some of them if the psu size is too small. Let N = number of psus in population and $\max\{M_i\}$ = maximum psu size.

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.

8

TABLE 6.2
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

Proof of Lahiri's Method: (9)

Let $X = \max\{M_j\}$

Let $P_1(U_i) = \text{Prob}(\text{Cluster } i \text{ is chosen on } 1^{st} \text{ draw})$

$$= \frac{1}{N} \cdot \frac{M_i}{X}$$

Prob[no cluster is selected on 1^{st} draw]

$$= 1 - P[\text{some cluster is selected on } 1^{st} \text{ draw}]$$

$$= 1 - \sum_{j=1}^{N} P[\text{cluster } j \text{ is chosen on } 1^{st} \text{ draw}]$$

$$= 1 - \sum_{j=1}^{N} \frac{1}{N}\frac{M_j}{X} = 1 - \frac{\overline{M}}{X}$$

(10)

Then $P_2(U_i) = \text{Prob}(\text{Cluster } i \text{ is chosen on } 2^{nd} \text{ draw, given that } 1^{st} \text{ draw failed})$

$$= \left(1 - \frac{\overline{M}}{X}\right)\frac{1}{N}\frac{M_i}{X}$$

Similarly, $P_K(U_i) = \left(1 - \frac{\overline{M}}{X}\right)^{K-1}\frac{1}{N}\frac{M_i}{X}$

Then $P(U_i) = \text{Prob}(U_i \text{ is } 1^{st} \text{ cluster chosen})$

$$= \sum_{K=1}^{\infty} P_K(U_i) = \frac{1}{N}\frac{M_i}{X}\sum_{K=1}^{\infty}\left(1 - \frac{\overline{M}}{X}\right)^{K-1}$$

$$= \frac{1}{N}\frac{M_i}{X}\left[1+\left(1-\frac{\bar{M}}{X}\right)+\left(1-\frac{\bar{M}}{X}\right)^2+\dots\right]$$

$$= \frac{1}{N}\frac{M_i}{X}\frac{1}{1-\left(1-\frac{\bar{M}}{X}\right)}$$

$$= \frac{1}{N}\frac{M_i}{X}\frac{X}{\bar{M}} = \frac{M_i}{K} = \psi_i.$$

Stat 576 HW#6

9 The file statepps.dat lists the number of counties, land area, and 1992 population for the 50 states plus the District of Columbia.

- **a** Use the cumulative-size method to draw a sample of size 10 with replacement, with probabilities proportional to land area. What is ψ_i for each state in your sample?
- **b** Use the cumulative-size method to draw a sample of size 10 with replacement, with probabilities proportional to population. What is ψ_i for each state in your sample?
- **c** How do the two samples differ? Which states tend to be in each sample?