

5 Sampling with Unequal Probabilities

Simple random sampling and systematic sampling are schemes where every unit in the population has the same chance of being selected. We will now consider unequal probability sampling. We have encountered an example under stratified sampling in which the units in stratum l have chance $\frac{n_l}{N_l}$ of being selected and varying such probability across strata under optimal allocation leads to increased accuracy.

5.1 Sampling with Replacement

Using with *replacement* sampling simplifies the calculations and if the sampling fraction is small this model should give a reasonable approximation to the exact behaviour of the estimators in *without replacement* sampling. Let p_j denote the probability of selecting unit y_j on the i th draw, so

$$P(Y_i = y_j) = p_j, \quad j = 1, 2, \dots, N$$

where Y_i represents a rv, not a total value. The first two moments are

$$E(Y_i) = \sum_{j=1}^N p_j y_j \quad \text{and} \quad \text{Var}(Y_i) = \sum_{j=1}^N p_j y_j^2 - \left(\sum_{j=1}^N p_j y_j \right)^2.$$

The *Hansen* and *Hurwitz* estimator for the population total Y is

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n y_i / p_i.$$

This estimator is unbiased for Y since

$$E(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^n E\left(\frac{Y_i}{p_i}\right) = \frac{n}{n} E\left(\frac{Y_i}{p_i}\right) = \frac{Y_1}{p_1} \times p_1 + \frac{Y_2}{p_2} \times p_2 + \dots + \frac{Y_N}{p_N} \times p_N = Y.$$

Under sampling *with replacement* the random variables (Y_i/p_i) are independent so

$$\text{Var}(\hat{Y}_{HH}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}\left(\frac{Y_i}{p_i}\right) = \frac{1}{n^2} \sum_{i=1}^N p_i \left(\frac{Y_i}{p_i} - Y\right)^2 = \frac{1}{n} \left(\sum_{i=1}^N \frac{Y_i^2}{p_i} - Y^2\right)$$

and is estimated by

$$\text{var}(\hat{Y}_{HH}) = \frac{1}{n} \left[\frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i}{p_i} - \hat{Y}_{HH}\right)^2 \right] = \frac{1}{n(n-1)} \left[\sum_{i=1}^n \left(\frac{y_i}{p_i}\right)^2 - n\hat{Y}_{HH}^2 \right].$$

How do we choose the p_i to minimise the variance?

The minimum is achieved if we set $p_i = Y_i/Y$, $i = 1, \dots, N$ since

$$\text{Var}(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^N p_i \left(\frac{Y_i}{p_i} - Y\right)^2 = \frac{1}{n} \sum_{i=1}^N p_i (Y - Y)^2 = 0$$

Of course we cannot use these values in practice as not all Y_i (and hence Y) are known. Instead we look for another variable, X_i , that is known and highly correlated with Y_i to construct probability estimates. Set $p_j = X_j/X$ where $X = \sum_{j=1}^N X_j$. Then

$$\hat{Y}_{HH} = \frac{X}{n} \sum_{i=1}^n \frac{y_i}{x_i}$$

the *mean-of-ratio* estimator in the *probability proportional to size* (PPS) sampling. Note

$$\text{Var}(\hat{Y}_{HH}) = \frac{X^2}{n} \text{Var}\left(\frac{y_i}{x_i}\right).$$

where $\text{Var}\left(\frac{y_i}{x_i}\right)$ is estimated by a sample variance of $\frac{y_i}{x_i}$.

When $p_i = \frac{1}{N}$, $\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i} = \frac{N}{n} \sum_{i=1}^n y_i = N\bar{y}$, the total estimator in SRS.

A natural alternative to the pps sampling here would be to use the ratio estimator

$$\hat{Y}_R = X \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

whereas

$$\hat{Y}_{HH} = \frac{X}{n} \sum_{i=1}^n \frac{y_i}{x_i}.$$

One is based on the ratio of the averages whilst the other is the average of the ratios.

Example: (Half Ackroyd case) Consider the population of firms A, B and C only with $N = 3$, $\bar{Y} = \frac{23}{3} = 7.667$ and $n = 2$. The following table shows selection probabilities of a particular sampling procedures and the sample estimates.

Sample Outcomes and HH estimators

Firm	Sales x_i	Selection probability $p_i = \frac{x_i}{X}$	Employee y_i
A	13	$13/34 = 0.3824$	9
B	12	$12/34 = 0.3529$	8
C	9	$9/34 = 0.2647$	6
	34	1.0000	23

Sample Outcomes and HH Estimates

Outcome	Probability	$\widehat{Y}_{HH} = \frac{1}{nN} \sum_{i=1}^n \frac{y_i}{p_i}$
A,A	$\frac{13}{34} \frac{13}{34} = 0.1462$	$\frac{1}{2 \cdot 3} \left(\frac{9}{13/34} + \frac{9}{13/34} \right) = 7.8461$
A,B	$\frac{13}{34} \frac{12}{34} = 0.1349$	$\frac{1}{2 \cdot 3} \left(\frac{9}{13/34} + \frac{8}{12/34} \right) = 7.7008$
A,C	$\frac{13}{34} \frac{9}{34} = 0.1012$	$\frac{1}{2 \cdot 3} \left(\frac{9}{13/34} + \frac{6}{9/34} \right) = 7.7008$
B,A	$\frac{12}{34} \frac{13}{34} = 0.1349$	$\frac{1}{2 \cdot 3} \left(\frac{8}{12/34} + \frac{9}{13/34} \right) = 7.7008$
B,B	$\frac{12}{34} \frac{12}{34} = 0.1246$	$\frac{1}{2 \cdot 3} \left(\frac{8}{12/34} + \frac{8}{12/34} \right) = 7.5555$
B,C	$\frac{12}{34} \frac{9}{34} = 0.0934$	$\frac{1}{2 \cdot 3} \left(\frac{8}{12/34} + \frac{6}{9/34} \right) = 7.5555$
C,A	$\frac{9}{34} \frac{13}{34} = 0.1012$	$\frac{1}{2 \cdot 3} \left(\frac{6}{9/34} + \frac{9}{13/34} \right) = 7.7008$
C,B	$\frac{9}{34} \frac{12}{34} = 0.0934$	$\frac{1}{2 \cdot 3} \left(\frac{6}{9/34} + \frac{8}{12/34} \right) = 7.5551$
C,C	$\frac{9}{34} \frac{9}{34} = 0.0701$	$\frac{1}{2 \cdot 3} \left(\frac{6}{9/34} + \frac{6}{9/34} \right) = 7.5551$
	1.0000	

Then the expected values and variances of these estimates are

$$\begin{aligned}
 E(\widehat{Y}_{HH}) &= (7.8461)(0.1462) + \cdots + (7.5551)(0.0701) = 7.6667 \\
 \text{Var}(\widehat{Y}_{HH}) &= (7.8461 - 7.6667)^2(0.1462) + \cdots + \\
 &\quad (7.5555 - 7.6667)^2(0.0701) \\
 &= 0.00997
 \end{aligned}$$

Note that the estimators are indeed unbiased and that they agree with the direct calculation of their variances:

$$\begin{aligned}
 \text{Var}(\widehat{Y}_{HH}) &= \frac{1}{nN^2} \left(\sum_{i=1}^N \frac{y_i^2}{p_i} - Y^2 \right) \\
 &= \frac{1}{2(3^2)} \left(\frac{9^2}{13/34} + \frac{8^2}{12/34} + \frac{6^2}{9/34} - 23^2 \right) = 0.00997
 \end{aligned}$$

If we use SRS,

$$\begin{aligned}\text{Var}(\widehat{\bar{Y}}_{srs}) &= \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n} \\ &= \left(1 - \frac{2}{3}\right) \frac{(9 - 7.\dot{6})^2 + (8 - 7.\dot{6})^2 + (6 - 7.\dot{6})^2}{2 \cdot 2} = 0.38889\end{aligned}$$

which is larger.

5.2 Inclusion PPS (IPPS) Sampling

If we are sampling without replacement the nature of the population changes after each selection and so if at each step we select using probabilities proportional to size the overall scheme will not necessarily be PPS. One way to achieve a PPS scheme is to use systematic sampling where only one random selection is necessary.

Let π_i denote the probability Y_i is selected in the *sample*, not just the j -th draw. Note that π_i is a first order *inclusion* (not selection) probability. The *Horvitz-Thompson* estimator for the population total Y is

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}$$

Let the first order inclusion indicator is

$$\begin{aligned} I_i &= 1 \text{ if } y_i \text{ is selected} \\ &= 0 \text{ otherwise.} \end{aligned}$$

Write

$$\hat{Y}_{HT} = \sum_{i=1}^N I_i \frac{y_i}{\pi_i}.$$

Using this form

$$E(\hat{Y}_{HT}) = \sum_{i=1}^N \frac{y_i}{\pi_i} E(I_i) = \sum_{i=1}^N \frac{y_i}{\pi_i} \pi_i = Y$$

so \hat{Y}_{HT} is unbiased for Y .

The second order inclusion indicator is

$$\begin{aligned} I_{ij} &= I_i I_j = 1 \text{ if both } y_i \text{ and } y_j \text{ are selected} \\ &= 0 \text{ otherwise.} \end{aligned}$$

Note $I_{ii} = I_i$. Let the second order inclusion probability $\pi_{ij} = E(I_{ij})$. Then

$$\begin{aligned}\text{Var}(\hat{Y}_{HT}) &= \text{Var}\left(\sum_{i=1}^N I_i \frac{y_i}{\pi_i}\right) \\ &= \sum_{i=1}^N \text{Var}(I_i) \frac{y_i^2}{\pi_i^2} + \sum_{i=1}^N \sum_{j=1, i \neq j}^N \text{Cov}(I_i, I_j) \frac{y_i y_j}{\pi_i \pi_j} \\ &= \sum_{i=1}^N \pi_i(1 - \pi_i) \frac{y_i^2}{\pi_i^2} + \sum_{i=1}^N \sum_{j=1, i \neq j}^N (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j}\end{aligned}$$

since $\text{Var}(I_i) = \pi_i(1 - \pi_i)$ and $\text{Cov}(I_i, I_j) = E(I_i I_j) - E(I_i)E(I_j) = \pi_{ij} - \pi_i \pi_j$. We estimate this via

$$\text{var}(\hat{Y}_{HT,1}) = \sum_{i=1}^n \frac{\pi_i(1 - \pi_i)}{\pi_i} \frac{y_i^2}{\pi_i^2} + \sum_{i=1}^n \sum_{j=1, i \neq j}^n \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j}$$

since $\frac{1}{\pi_i} = \frac{N}{n}$ say adjusts the sum of N terms to the sum of n terms and $\pi_{ij} - \pi_i \pi_j = \pi_i - \pi_i^2 = \pi_i(1 - \pi_i)$ if $i = j$. Hence

$$\text{var}(\hat{Y}_{HT,1}) = \sum_{i=1}^n \frac{1 - \pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^n \sum_{j=1, i < j}^n \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} y_i y_j$$

Note:

1. The first order inclusion probabilities π_i satisfy

$$\sum_{i=1}^N \pi_i = \sum_{i=1}^N E(I_i) = E\left(\sum_{i=1}^N I_i\right) = n \quad \text{whereas} \quad \sum_{i=1}^N p_i = 1.$$

2. The estimator in SRS is a special case of the HT estimator.

$$\begin{aligned}\hat{\bar{Y}}_{HT} &= \frac{1}{N} \left(\frac{y_1}{\pi_1} + \frac{y_2}{\pi_2} + \cdots + \frac{y_n}{\pi_n} \right) \\ &= \frac{1}{N} \left(\frac{y_1}{n/N} + \frac{y_2}{n/N} + \cdots + \frac{y_n}{n/N} \right) = \bar{y}\end{aligned}$$

3. The variance estimate can be negative (if $\pi_i\pi_j > \pi_{ij}$). If we restrict the variance to 0 in these cases, we induce bias.

Example: (Strip transect sampling) Consider a study area of 100 km² partitioned into strips 1 km wide but varying in length. A sample of $n = 4$ strips are selected by draw-by-draw with replacement. The number of animals y_i is counted in each of these 4 strips. The data are shown below:

Sample strip	Length of strip (in km)	p_i	y_i
3	2	0.02	14
7	5	0.05	60
7	5	0.05	60
56	1	0.01	1

Estimate the total number of animals in this area using the HT estimator and report its standard error.

Solution: The sample is drawn *with replacement* such that $n = 4$ and $s = \{7, 3, 56\}$.

$$\begin{aligned}\pi_3 &= 1 - (1 - p_3)^4 = 1 - (1 - 0.02)^4 = 0.0776 \\ \pi_7 &= 1 - (1 - p_7)^4 = 1 - (1 - 0.05)^4 = 0.1855 \\ \pi_{56} &= 1 - (1 - p_{56})^4 = 1 - (1 - 0.01)^4 = 0.0394 \\ \pi_{3,7} &= \pi_3 + \pi_7 - [1 - (1 - p_7 - p_3)^4] \\ &= \underbrace{0.0776 + 0.1855}_{37, 3\bar{7}, 37, \bar{3}7} - \underbrace{[1 - (1 - 0.05 - 0.02)^4]}_{37, 3\bar{7}, \bar{3}7} = 0.0112\end{aligned}$$

$$\begin{aligned}\pi_{7,56} &= \pi_7 + \pi_{56} - [1 - (1 - p_7 - p_{56})^4] \\ &= 0.1855 + 0.0394 - [1 - (1 - 0.05 - 0.01)^4] = 0.0056\end{aligned}$$

$$\begin{aligned}\pi_{3,56} &= \pi_3 + \pi_{56} - [1 - (1 - p_3 - p_{56})^4] \\ &= 0.0776 + 0.0394 - [1 - (1 - 0.02 - 0.01)^4] = 0.0023\end{aligned}$$

$$\hat{Y}_{HT} = \sum_{i \in \mathcal{S}} \frac{y_i}{\pi_i} = 2 \frac{60}{0.1855} + \frac{14}{0.0776} + \frac{1}{0.0394} = 852.6934$$

$$\begin{aligned}\text{var}(\hat{Y}_{HT,1}) &= \sum_{i \in \mathcal{S}} \frac{1 - \pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i < j} \sum \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j \\ &= \left(\frac{1 - 0.1855}{0.1855^2} 60^2 + \frac{1 - 0.0776}{0.0776^2} 14^2 + \frac{1 - 0.0394}{0.0394^2} 1^2 \right) + \\ &\quad 2 \left(\frac{0.0112 - 0.1855 \times 0.0776}{0.1855 \times 0.0776 \times 0.0112} 60 \times 14 + \right. \\ &\quad \left. \frac{0.0056 - 0.1855 \times 0.0394}{0.1855 \times 0.0394 \times 0.0056} 60 \times 1 + \right. \\ &\quad \left. \frac{0.0023 - 0.0776 \times 0.0394}{0.0776 \times 0.0394 \times 0.0023} 14 \times 1 \right) = 74,494.965 \\ \text{se}(\hat{Y}_{HT,1}) &= \sqrt{74,494.965} = 272.9\end{aligned}$$

5.3 IPPS sampling without replacement

For sampling without replacement (WOR), the inclusion probabilities are

$$\begin{aligned}\text{WOR: } \sum_{j=1, j \neq i}^N \pi_{ij} &= \sum_{j=1, j \neq i}^N E(I_i I_j) = E \left(I_i \sum_{j=1, j \neq i}^N I_j \right) \\ &= E[I_i(n - I_i)] = \pi_i n - \pi_i = (n - 1)\pi_i \\ \sum_{j \neq i}^N (\pi_{ij} - \pi_i \pi_j) &= (n - 1)\pi_i - \pi_i(n - \pi_i) = -\pi_i(1 - \pi_i)\end{aligned}$$

Then the variance for the HT estimator becomes

$$\begin{aligned}
 & \text{Var}(\hat{Y}_{HT,2}) \\
 &= \sum_{i=1}^N \pi_i(1 - \pi_i) \left(\frac{y_i}{\pi_i} \right)^2 + \sum_{i=1}^N \sum_{j=1, i \neq j}^N (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} \right) \left(\frac{y_j}{\pi_j} \right) \\
 &= \sum_{i=1}^N \left[- \sum_{j=1, i \neq j}^N (\pi_{ij} - \pi_i \pi_j) \right] \left(\frac{y_i}{\pi_i} \right)^2 - 2 \sum_{i=1}^N \sum_{j=1, i < j}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} \right) \left(\frac{y_j}{\pi_j} \right) \\
 &= \sum_{i=1}^N \sum_{j=1, i < j}^N (\pi_i \pi_j - \pi_{ij}) \left[\left(\frac{y_i}{\pi_i} \right)^2 + \left(\frac{y_j}{\pi_j} \right)^2 \right] - 2 \sum_{i=1}^N \sum_{j=1, i < j}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} \right) \left(\frac{y_j}{\pi_j} \right) \\
 &= \sum_{i=1}^N \sum_{j=i+1}^N (\pi_i \pi_j - \pi_{ij}) \left[\left(\frac{y_i}{\pi_i} \right)^2 + \left(\frac{y_j}{\pi_j} \right)^2 - 2 \left(\frac{y_i}{\pi_i} \right) \left(\frac{y_j}{\pi_j} \right) \right] \\
 &= \sum_{i=1}^N \sum_{j=i+1}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.
 \end{aligned}$$

We estimate this via

$$\text{var}(\hat{Y}_{HT,2}) = \sum_{i=1}^n \sum_{j=i+1}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

Note:

1. The variance $\text{var}(\hat{Y}_{HT,2})$ is guaranteed to be positive for any sampling WOR method satisfying $\pi_i \pi_j > \pi_{ij}$. Also $\text{var}(\hat{Y}_{HT,2})$ takes negative values less often than $\text{var}(\hat{Y}_{HT,1})$.
2. The calculating π_{ij} depends on the way the sample is selected. Note $\binom{n}{2}$ values of π_{ij} need to be calculated.
3. If Y_i are approximately proportional to an auxiliary variable X_i , i.e. $y_i \simeq r x_i$ and we set $\pi_i = \frac{n x_i}{\sum_{i=1}^N x_i} \equiv n p_i \equiv c x_i$, then for all (i, j)

$$\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \simeq \frac{r x_i}{c x_i} - \frac{r x_j}{c x_j} \simeq 0$$

or $\text{Var}(\hat{Y}_{HT,2}) \simeq 0$. The IPPS without replacement estimator is

$$\hat{Y}_{HT} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i}$$

Moreover if y_i and x_i are perfectly correlated such that $y_i = r x_i$, $i = 1, 2, \dots, N$, then $\pi_i = n y_i / Y$ and

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} = \sum_{i=1}^n \frac{y_i}{n y_i / Y} = \sum_{i=1}^n \frac{Y}{n} = Y$$

We expect good performance if $y_i \simeq r x_i$ for all i under PPS sampling.

4. Systematic IPPS is popular because of its simplicity but an unbiased estimator for $\text{Var}(\hat{Y}_{sys})$ is not available. Hurtley and Rao (1962) showed that when $n \frac{x_i}{\bar{X}} < 1$ for all i ,

$$\text{Var}(\hat{Y}_{sys,pps}) \simeq \sum_{i=1}^N \left(1 - \frac{n-1}{n} \pi_i \right) \left(\frac{y_i}{\pi_i} - \frac{Y}{n} \right)^2.$$

This expression can be estimated by

$$\text{var}(\hat{Y}_{sys,pps}) \simeq \sum_{i=1}^n \left(1 - \frac{n-1}{n} \pi_i \right) \left(\frac{y_i}{\pi_i} - \frac{\hat{Y}_{HT}}{n} \right)^2$$

Example: (Half Ackroyd case) Consider the population of firms A, B and C only with $N = 3$, $\bar{Y} = \frac{23}{3} = 7.667$. The following table shows the inclusion probabilities of a particular sampling procedures and the sample estimates.

Sample outcomes and HT estimator

Sample	2nd order Inclusion prob. π_{ij}	$\widehat{Y}_{HT} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{\pi_i}$
A,B	$\pi_{12} = 0.5$	$\widehat{Y}_{HT} = \frac{1}{3}(\frac{9}{0.8} + \frac{8}{0.7}) = 7.5595$
A,C	$\pi_{13} = 0.3$	$\widehat{Y}_{HT} = \frac{1}{3}(\frac{9}{0.8} + \frac{6}{0.5}) = 7.7500$
B,C	$\pi_{23} = 0.2$	$\widehat{Y}_{HT} = \frac{1}{3}(\frac{8}{0.7} + \frac{6}{0.5}) = 7.8095$
	1.0	

The first-order inclusion probabilities π_i are

Firm	y_i	π_i
A	9	$\pi_1 = 0.5 + 0.3 = 0.8$
B	8	$\pi_2 = 0.5 + 0.2 = 0.7$
C	6	$\pi_3 = 0.3 + 0.2 = 0.5$
Total		2.0

Note

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = \pi_1 + \pi_2 - \pi_{12} = 0.8 + 0.7 - 0.5 = 1$$

since in a sample of size $n = 2$ from $N = 3$, at least one element of the pair will be included.

The variance of the HT estimator is:

$$\begin{aligned}
& \text{Var}(\widehat{Y}_{HT,2}) \\
&= \frac{1}{N^2} \left[(\pi_1\pi_2 - \pi_{12}) \left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2} \right)^2 + (\pi_1\pi_3 - \pi_{13}) \left(\frac{y_1}{\pi_1} - \frac{y_3}{\pi_3} \right)^2 + (\pi_2\pi_3 - \pi_{23}) \left(\frac{y_2}{\pi_2} - \frac{y_3}{\pi_3} \right)^2 \right] \\
&= \frac{1}{3^2} \left[(.8 \times .7 - .5) \left(\frac{9}{.8} - \frac{8}{.7} \right)^2 + (.8 \times .5 - .3) \left(\frac{9}{.8} - \frac{6}{.5} \right)^2 + (.7 \times .5 - .2) \left(\frac{8}{.7} - \frac{6}{.5} \right)^2 \right] \\
&= 0.0119.
\end{aligned}$$

The following calculation verifies that the HT estimator is unbiased and that its variance is indeed as calculated using:

$$\begin{aligned}
E(\widehat{Y}_{HT}) &= (7.5595)(0.5) + (7.7500)(0.3) + (7.8095)(0.2) = 7.6667 \\
\text{Var}(\widehat{Y}_{HT}) &= (7.5595 - 7.6667)^2(0.5) + \dots + (7.8095 - 7.6667)^2(0.2) \\
&= 0.0119
\end{aligned}$$

The simple average estimator is biased when the inclusion probabilities are not the same for all elements.

Outcome	Inclusion probability	$\widehat{Y} = \bar{y}$
A,B	0.5	$\frac{1}{2}(9 + 8) = 8.5$
A,C	0.3	$\frac{1}{2}(9 + 6) = 7.5$
B,C	0.2	$\frac{1}{2}(8 + 6) = 7.0$
	1.0	

Note that

$$E(\bar{y}) = 0.5(8.5) + 0.3(7.5) + 0.2(7.0) = 7.9 \neq 7.667$$

Therefore, the sample mean \bar{y} is biased.

5.4 How to draw IPPS sample?

It is very difficult.

Madow (1949) method for a systematic IPPS sample

Index	x	Partial sum	Assigned interval
1	x_1	$S_1 = x_1$	$(0, S_1)$
2	x_2	$S_2 = x_1 + x_2$	(S_1, S_2)
\vdots	\vdots	\vdots	\vdots
N	x_N	$S_N = x_1 + \cdots + x_N = X$	(S_{N-1}, S_N)

Define $d = \frac{X}{n}$.

Step 1 Select a uniform random number u from a uniform dist. $U(0, 1)$ and set $u' = ud \in (0, d)$.

Step 2 Select the units whose assigned intervals contain $u', u' + d, u' + 2d, \dots, u' + (n - 1)d$.

For systematic IPPS method, it can be shown that $\pi_i = \frac{x_i}{d} = \frac{nx_i}{X} = nz_i$.

Hence $\hat{Y}_{sys} = \sum_{i=1}^n \frac{y_i}{\pi_i}$ is unbiased for the population total Y .

$$\begin{aligned}
 \pi_i &= P[u \in (S_{i-1}, S_i) \text{ or } u + d \in (S_{i-1}, S_i) \text{ or } \dots \text{ or } u + (n-1)d \in (S_{i-1}, S_i)] \\
 &= P[u \in (S_{i-1}, S_i) \text{ mod. } d] \\
 &= \frac{S_i - S_{i-1}}{d} = \frac{x_i}{d} = n \frac{x_i}{X} = nz_i
 \end{aligned}$$

Note that $x_i > d \Leftrightarrow x_i > \frac{X}{n} \Leftrightarrow n \frac{x_i}{X} > 1 \Leftrightarrow nz_i > 1$.

If $nz_i < 1$ for all i , any unit has a probability $\pi_i = nz_i$ of being selected and no unit is selected more than once.

If $nz_i > 1$ for one or more i , such units may be selected more than once in the sample but the average frequency of selection is nz_i .

Example: (IPPS sample) For a population of 7 households, the number of visits to a local supermarket last week x_i and this week y_i are given below:

Household i	No. of visit last week x_i	No. of visit this week y_i
1	3	2
2	1	1
3	11	9
4	6	5
5	4	3
6	2	1
7	3	5

Use random numbers $u = 6350$. Select a systematic IPPS sample of size $n = 3$ using the Madow (1949) method. Estimate the average number of visits to the supermarket this week and its standard error.

Solution: $d = \frac{X}{n} = \frac{30}{3} = 10$. Note that $nz_3 = 3 \times \frac{11}{30} = \frac{33}{30} > 1$. Hence unit 3 can be selected more than once.

Index i	y_i	x_i	S_i	$z_i = \frac{x_i}{X}$	Assigned interval (S_{i-1}, S_i)	Assigned interval (mod $d = 10$)	Width	Prob. π_i
1	2	3	3	$\frac{3}{30}$	(0,3)	(0,3)	3	$\frac{3}{10}$
2	1	1	4	$\frac{1}{30}$	(3,4)	(3,4)	1	$\frac{1}{10}$
3	9	11	15	$\frac{11}{30}$	(4,15)	(4,10) or (0,5)	6+5=11	$\frac{11}{10}$
4	5	6	21	$\frac{6}{30}$	(15,21)	(5,10) or (0,1)	1+5=6	$\frac{6}{10}$
5	3	4	25	$\frac{4}{30}$	(21,25)	(1,5)	4	$\frac{4}{10}$
6	1	2	27	$\frac{2}{30}$	(25,27)	(5,7)	2	$\frac{2}{10}$
7	5	3	30	$\frac{3}{30}$	(27,30)	(7,10)	3	$\frac{3}{10}$
Total	26	30		1				3

Now $u'_1 = 0.6350 \times 10 = 6.35$ from the interval (0,10). The IPPS sample contains households $\{3, 4, 6\}$. A complete list of different samples based

in $u' \in (0, 10)$ is given below.

u	$IPPS$ sample
(0,1)	$\{1,3,4\}$
(1,3)	$\{1,3,5\}$
(3,4)	$\{2,3,5\}$
(4,5)	$\{3,3,5\}$
(5,7)	$\{3,4,6\}$
(7,10)	$\{3,4,7\}$

It is difficult to estimate s.e. using the HT estimator because the second order inclusion probabilities π_{ij} are unknown. Hence the estimate of the average number of visits to the supermarket this week and its standard error using Hartley & Rao (1962) estimator for systematic sampling are

$$\begin{aligned}
 \widehat{Y}_{HT} &= \frac{1}{N} \sum_{i=1}^n \frac{y_i}{\pi_i} = \frac{1}{7} \left(\frac{9}{1.1} + \frac{5}{0.6} + \frac{1}{0.2} \right) = 3.6688 \\
 \text{var}(\widehat{Y}_{sys,ipps}) &= \frac{1}{N^2} \left[\sum_{i=1}^n \left(1 - \frac{n-1}{n} \pi_i \right) \left(\frac{y_i}{\pi_i} - \frac{\widehat{Y}_{sys}}{n} \right)^2 \right] \\
 &= \frac{1}{7^2} \left[\left(1 - \frac{2}{3} \cdot 1.1 \right) \left(\frac{9}{1.1} - \frac{3.6688}{3} \right)^2 + \right. \\
 &\quad \left(1 - \frac{2}{3} \cdot 0.6 \right) \left(\frac{5}{0.6} - \frac{3.6688}{3} \right)^2 + \\
 &\quad \left. \left(1 - \frac{2}{3} \cdot 0.2 \right) \left(\frac{1}{0.2} - \frac{3.6688}{3} \right)^2 \right] \\
 &= \frac{118.537}{7^2} = 2.4191 \\
 \text{se}(\widehat{Y}_{sys,ipps}) &= \sqrt{2.4191} = 1.5554
 \end{aligned}$$