Module 2

Horwitz - Thompson Estimator, SRSWR, SRSWOR

$\underline{\mathcal{U}} = (U_1, \dots, U_N)$: a finite identifiable population

$Y : \mathcal{U} \to \mathbb{R}$: Study variable

$\underline{Y} = (Y_1, \dots, Y_N)$

Goal: To estimate $f(\underline{Y})$, for some given function f.

Estimation of Population Total

$$f(\underline{Y}) = \sum_{i=1}^{N} Y_i = Y, \text{ say.}$$

Define
$$\hat{Y}_{HT} = \sum_{i \in S} \frac{Y_i}{\pi_i},$$
where $i \in S$ means $i \in \{1, \dots, N\}$ such that $U_i \in S$

$\to$ Horvitz - Thompson Estimator (HTE)

Theorem $E(\hat{Y}_{HT}) = Y$ i.e. the HTE is an unbiased estimator of the population total Y.

Proof Define

$$\delta_i(\underline{S}) = \begin{cases} 1, & \text{if } U_i \in \underline{S} \\ 0, & \text{o.w} \end{cases}, \quad i = 1, \dots, N$$

so that $E(\delta_i(\underline{S})) = \Pr(U_i \in \underline{S}) = \pi_i, \quad i = 1, \dots, N$

Then
$$\hat{Y}_{HT} = \sum_{i=1}^{N} \delta_i(\underline{S}) \frac{Y_i}{\pi_i}.$$
$$E(\hat{Y}_{HT}) = \sum_{i=1}^{N} E(\delta_i(\underline{S})) \frac{Y_i}{\pi_i} = \sum_{i=1}^{N} Y_i = Y.$$

Theorem

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} Y_i^2 \frac{1-\pi_i}{\pi_i} + 2 \sum\sum_{1 \le i < j \le N} Y_i Y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}.$$

Proof $\hat{Y}_{HT} = \sum_{i=1}^{N} \delta_i(\underline{S}) \frac{Y_i}{\pi_i}$

$$\Rightarrow Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2} Var(\delta_i(\underline{S})) + 2 \sum\sum_{1 \le i < j \le N} \frac{Y_i}{\pi_i} \frac{Y_j}{\pi_j} Cov(\delta_i, \delta_j)$$

$E(\delta_i(\underline{S})) = \pi_i, \quad i = 1(1)N$

$Var(\delta_i(\underline{S})) = \pi_i(1-\pi_i), \quad i = 1(1)N \quad (\text{as } \delta_i^2 = \delta_i)$

$$Cov(\delta_i(\underline{S}), \delta_j(S)) = E(\delta_i(S)\, \delta_j(S)) - \pi_i \pi_j$$
$$= Pr(U_i \in S, U_j \in S) - \pi_i \pi_j$$
$$= \pi_{ij} - \pi_i \pi_j, \quad i \ne j$$

$$\Rightarrow Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}(1-\pi_i) + 2 \sum\sum_{1 \le i < j \le N} \frac{Y_i Y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j)$$

Theorem Let $(\mathcal{S}, \mathcal{S}', P)$ be FES(n) design. Then

$$V(\hat{Y}_{HT}) = \sum\sum_{1 \le i < j \le N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

Proof. We have

$$\sum_{i=1}^{N} \pi_i = n \quad \& \quad \sum_{\substack{j=1 \\ j \ne i}}^{N} \pi_{ij} = (n-1)\pi_i, \quad i = 1 \dots N$$

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2} \pi_i(1-\pi_i) + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \ne j}}^{N} Y_i Y_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}$$

$$\pi_i(1-\pi_i) = \pi_i - \pi_i^2$$
$$= \pi_i(n - \pi_i) - (n-1)\pi_i$$
$$= \pi_i \sum_{j \ne i}^{N} \pi_j - \sum_{j \ne i} \pi_{ij} = \sum_{j \ne i} (\pi_i \pi_j - \pi_{ij})$$

(FES(n))

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \frac{Y_i^2}{\pi_i^2}(\pi_i\pi_j - \pi_{ij}) + \sum\sum_{i\neq j} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j)$$

$$= \frac{1}{2}\Big[\sum\sum_{i\neq j} \frac{Y_i^2}{\pi_i^2}(\pi_i\pi_j - \pi_{ij}) + \underbrace{\sum\sum_{i\neq j} \frac{Y_j^2}{\pi_j^2}(\pi_i\pi_j - \pi_{ij})}_{\text{same quantity}}$$

$$+ 2\sum\sum_{i\neq j} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j)\Big]$$

$$= \frac{1}{2}\Big[\sum_{i\neq j}^{N}\sum^{N} (\pi_i\pi_j - \pi_{ij})\Big\{ \frac{Y_i^2}{\pi_i^2} + \frac{Y_j^2}{\pi_j^2} - 2\frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\Big\}\Big]$$

$$= \frac{1}{2}\sum_{i\neq j}^{N}\sum^{N} (\pi_i\pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

$$= \sum_{1\leq i<j\leq N} (\pi_i\pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

Theorem An unbiased estimator of $Var(\hat{Y}_{HT})$ is

$$\widehat{Var}(\hat{Y}_{HT}) = \sum_{i\in \underline{S}} \frac{Y_i^2}{\pi_i}\left(\frac{1}{\pi_i} - 1\right) + \sum\sum_{\substack{i\neq j\in S\\ i\neq j}} \frac{Y_iY_j}{\pi_{ij}}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)$$

Proof We have

$$E(\delta_i(\underline{S})) = \pi_i, \quad i = 1 \dots N$$

$$E(\delta_i(\underline{S})\,\delta_j(\underline{S})) = \pi_{ij}, \quad i, j = 1 \dots N, \quad i \neq j$$

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} Y_i^2\left(\frac{1}{\pi_i} - 1\right) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} Y_iY_j\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)$$

$$= \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} E(\delta_i)\left(\frac{1}{\pi_i} - 1\right) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N} \frac{Y_iY_j}{\pi_{ij}} E(\delta_i\delta_j)\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)$$

$$= E\Big[\sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}\left(\frac{1}{\pi_i} - 1\right)\delta_i + \sum_{i\neq j}^{N}\sum^{N} \frac{Y_iY_j}{\pi_{ij}}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)\delta_i\delta_j\Big]$$

$$= E\Big[\sum_{i\in S} \frac{Y_i^2}{\pi_i}\left(\frac{1}{\pi_i} - 1\right) + \sum_{i\in S}\sum_{\substack{j\in S\\ i\neq j}} \frac{Y_iY_j}{\pi_{ij}}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)\Big]$$

Yates-Grundy Variance Estimator

Define, for FES(n) design,

$$\widehat{Var}_{YG}(\hat{Y}_{HT}) = \sum_{\substack{i \in s \\ i<j}} \sum_{j \in s} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right)$$

$\rightarrow$ Yates-Grundy Variance Estimator

Theorem For any FES(n) design Yates-Grundy variance estimator $\widehat{Var}_{YG}(\hat{Y}_{HT})$ is an alternate unbiased estimator of $Var(\hat{Y}_{HT})$.

Proof For FES(n) design

$$Var(\hat{Y}_{HT}) = \sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \pi_{ij}$$

$$= E\left[\sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \delta_{ij} \right]$$

$$= E\left[\sum_{\substack{i \in s \\ i<j}} \sum_{j \in s} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \right]$$

Note: (I) Unbiasedness of Yates-Grundy estimator is valid only for FES designs

(II) Generally $\widehat{Var}(\hat{Y}_{HT})$ and $\hat{V}_{YG}(\hat{Y}_{HT})$ are two different estimators but for some specific designs they may be the same. For example, for SRSWOR(n) $\widehat{Var}(\hat{Y}_{HT})$ and $\hat{V}_{YG}(\hat{Y}_{HT})$ are the same.

Estimation of population total and population mean under SRSWOR(n)

$$\pi_i = \frac{n}{N} = p, \text{ say} \quad \text{and} \quad \pi_{ij} = \frac{n(n-1)}{N(N-1)} = q, \text{ say}, \; i \neq j$$

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \frac{N}{n} \sum_{i \in s} y_i = N\hat{\bar{Y}},$$

where $\hat{\bar{Y}} = \frac{1}{n} \sum_{i \in s} y_i$ is the sample mean.

Theorem. (i) $\hat{Y}_{HT} = N\hat{\bar{Y}}$ is an unbiased estimator of the population ~~tot~~ total $Y = \sum_{i=1}^{N} y_i$

(ii) $\hat{\bar{Y}}$ is an unbiased estimator of the population mean $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$

(iii) $Var(\hat{Y}_{HT}) = \frac{N(N-n)}{n} S^2 = \frac{N^2(N-n)}{n(N-1)} \sigma^2,$

where $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2$ and $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{Y})^2$ is the population variance.

(iv) $Var(\hat{\bar{Y}}) = (1-p)\frac{S^2}{n} = \frac{N(1-p)}{N-1} \frac{\sigma^2}{n}$

(v) $\hat{Var}(\hat{Y}_{HT}) = \hat{Var}_{YG}(\hat{Y}_{HT}) = \frac{N(N-n)}{n} \hat{S}^2$ is an unbiased estimator of $Var(\hat{Y}_{HT})$, where $\hat{S}^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{\bar{Y}})^2$ (sample variance) is an unbiased estimator of S^2

(vi) $\hat{Var}(\hat{\bar{Y}}) = (1-p)\frac{\hat{S}^2}{n}$ is an unbiased estimator of s^2.

Proof (i) & (ii) : Obvious

Since SRSWOR(n) is an FES(n) design,

(iii) $$\mathrm{Var}(\hat{Y}_{HT}) = \sum\sum_{1 \le i < j \le N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

$$= \frac{(p^2 - q)}{p^2} \sum\sum_{1 \le i < j \le N} (y_i - y_j)^2$$

$$= \frac{1}{2}\left[1 - \frac{q}{p^2}\right] \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i - y_j)^2$$

$$= \frac{1}{2}\left[1 - \frac{(n-1)N}{n(N-1)}\right] \left[\sum_{i=1}^{N} \sum_{j=1}^{N} (y_i - \bar{y})^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} (y_j - \bar{y})^2 - 2 \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i - \bar{y})(y_j - \bar{y}) \right]$$

$$= \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} (y_i - \bar{y})^2$$

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

(iv) $$\mathrm{Var}(\hat{\bar{y}}) = \mathrm{Var}\left(\frac{\hat{Y}_{HT}}{N}\right) = \frac{1}{N^2} \mathrm{Var}(\hat{Y}_{HT})$$

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

(v) We have

$$\widehat{\mathrm{Var}}(\hat{Y}_{HT}) = \sum_{i \in s} \frac{y_i^2}{\pi_i}\left(\frac{1}{\pi_i} - 1\right) + \sum\sum_{\substack{i,j \in s \\ i \ne j}} \frac{y_i y_j}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right)$$

$$= \frac{1-p}{p^2} \sum_{i \in s} y_i^2 + \left(\frac{1}{p^2} - \frac{1}{q} \right) \sum\sum_{\substack{i,j \in s \\ i \ne j}} y_i y_j$$

$$\frac{1-p}{p^2} = \frac{N^2}{n^2}\left(1 - \frac{n}{N}\right) = \frac{N(N-n)}{n^2}$$

$$\frac{1}{p^2} - \frac{1}{q} = \frac{N^2}{n^2} - \frac{N(N-1)}{n(n-1)} = \frac{N}{n}\left[\frac{N}{n} - \frac{N-1}{n-1}\right] = -\frac{N(N-n)}{n^2(n-1)}$$

$$\widehat{\mathrm{Var}}(\hat{Y}_{HT}) = \frac{N(N-n)}{n^2} \left[\sum_{i \in s} y_i^2 - \frac{1}{n-1} \sum\sum_{\substack{i,j \in s \\ i \ne j}} y_i y_j \right]$$

$$= \frac{N(N-n)}{n^2(n-1)} \left[n \sum_{i \in S} Y_i^2 - \left(\sum_{i \in S} Y_i^2 + \sum\sum_{\substack{i,j \in S \\ i \neq j}} Y_i Y_j \right) \right]$$

$$= \frac{N(N-n)}{n^2(n-1)} \left[n \sum_{i \in S} Y_i^2 - \left(\sum_{i \in S} Y_i \right)^2 \right]$$

$$= \frac{N(N-n)}{n(n-1)} \sum_{i \in S} (Y_i - \hat{\bar{Y}})^2$$

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

$$\hat{Var}_{YG}(\hat{Y}_{HT}) = \sum\sum_{\substack{i,j \in S \\ i<j}} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right)$$

$$= \frac{1}{p^2} \left(\frac{p^2 - q}{q} \right) \sum\sum_{\substack{i,j \in S \\ i<j}} (Y_i - Y_j)^2$$

$$= \left(\frac{1}{q} - \frac{1}{p^2} \right) \times \frac{1}{2} \sum\sum_{i,j \in S} (Y_i - Y_j)^2$$

$$= \frac{N(N-n)}{2n^2(n-1)} \sum\sum_{i,j \in S} (Y_i - Y_j)^2$$

$$= \frac{N(N-n)}{n(n-1)} \sum_{i \in S} (Y_i - \hat{\bar{Y}})^2$$

$$= \hat{Var}(\hat{Y}_{HT})$$

(VI) $$\hat{Var}(\hat{\bar{Y}}) = \frac{1}{N^2} \hat{Var}(\hat{Y}_{HT}) = \frac{N-n}{N} \frac{\hat{S}^2}{n}$$

$$= (1-f) \frac{\hat{S}^2}{n}.$$

Remark: $\hat{S}^2 =$ is an unbiased estimator of σ^2 (from (IV) and (VI))

Estimation of Population Total and Population Mean Under SRSWR(n)

$$\pi_i = 1-\left(1-\frac{1}{N}\right)^n = 1-a, \text{ say}, \quad i=1,\dots,N, \quad \text{where } a=\left(1-\frac{1}{N}\right)^n$$

$$\pi_{ij} = 1-2\left(1-\frac{1}{N}\right)^n+\left(1-\frac{2}{N}\right)^n = 1-2a+b \quad i,j=1,\dots,n, \quad i\neq j,$$

where $b=\left(1-\frac{2}{N}\right)^n$.

$$\hat{Y}_{HT} = \sum_{i\in s}\frac{Y_i}{\pi_i} = \frac{1}{1-a}\sum_{i\in s} Y_i$$

$$\hat{\bar{Y}}_{HT} = \frac{1}{N(1-a)}\sum_{i\in s} Y_i$$

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}Y_i^2 + 2\sum\sum_{1\le i<j\le N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}Y_iY_j$$

$$\boxed{Var(\hat{Y}_{HT}) = \frac{a}{1-a}\sum_{i=1}^{N}Y_i^2 + 2\frac{(b-a^2)}{(1-a)^2}\sum\sum_{1\le i<j\le N}Y_iY_j}$$

$$= \frac{a}{1-a}\sum_{i=1}^{N}Y_i^2 + \frac{b-a^2}{(1-a)^2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}Y_iY_j - \sum_{i=1}^{N}Y_i^2\right]$$

$$= \frac{a-b}{(1-a)^2}\sum_{i=1}^{N}Y_i^2 + \frac{N^2(b-a^2)}{(1-a)^2}\bar{Y}^2$$

$$= \frac{N(a-b)}{(1-a)^2}(\sigma^2+\bar{Y}^2) + \frac{N^2(b-a^2)}{(1-a)^2}\bar{Y}^2$$

$$\boxed{Var(\hat{Y}_{HT}) = \frac{N}{(1-a)^2}\left[(a-b)\sigma^2 + ((N-1)b-a(Na-1))\bar{Y}^2\right]}$$

$$\hat{Var}(\hat{Y}_{HT}) = \sum_{i\in s}\frac{1-\pi_i}{\pi_i^2}Y_i^2 + \sum\sum_{\substack{i,j\in s\\ i\neq j}}\left(\frac{1}{\pi_i\pi_j}-\frac{1}{\pi_{ij}}\right)Y_iY_j$$

$$= \frac{a}{(1-a)^2}\sum_{i\in s}Y_i^2 + \left(\frac{1}{(1-a)^2}-\frac{1}{1-2a+b}\right)\sum\sum_{\substack{i,j\in s\\ i\neq j}}Y_iY_j$$

$$= \frac{a}{(1-a)^2} \sum_{i \in S} y_i^2 + \frac{b-a^2}{(1-a)^2(1-2a+b)} \sum_{\substack{i \neq j \\ i, j \in S}} \sum y_i y_j$$

$$= \frac{a}{(1-a)^2} \sum_{i \in S} y_i^2 + \frac{b-a^2}{(1-a)^2(1-2a+b)} \left[n^2 \hat{\bar{Y}}^2 - \sum_{i \in S} y_i^2 \right)$$

$$= \frac{a-b}{(1-a)(1-2a+b)} \sum_{i \in S} y_i^2 + \frac{n^2(b-a^2)}{(1-a)^2(1-2a+b)} \hat{\bar{Y}}^2$$

$$= \frac{a-b}{(1-a)(1-2a+b)} \left[(n-1) \hat{S}^2 + n \hat{\bar{Y}}^2 \right] + \frac{n^2(b-a^2)}{(1-a)^2(1-2a+b)} \hat{\bar{Y}}^2$$

$$= \frac{(n-1)(a-b)}{(1-a)(1-2a+b)} \hat{S}^2 + \frac{n \left[a-(n-1)a^2 + (n-1)b - ab \right)}{(1-a)^2(1-2a+b)}$$

Sample Mean Based On all Observations

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{i=1}^{N} A_i Y_i$$ where A_i = # of times unit U_i appears in the sample

$(A_i, A_j) \sim Mult(n, \frac{1}{N}, \frac{1}{N})$ $A_i \sim Bin(n, \frac{1}{N})$

$A_i + A_j \sim Bin(n, \frac{2}{N})$

Define

Z_i = study variable corresponding to unit drawn at ith draw, $i = 1, 2, \ldots n$

Then $Z_1, Z_2, \ldots, Z_n$ are iid

$Pr(Z_i = Y_j) = \frac{1}{N}$, $j = 1 \ldots N$ $i = 1 \ldots N$

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{i=1}^{n} Z_i = \bar{Z}$$ $E(Z_1) = \bar{Y}$

$$E(Z_i) = \frac{1}{N} \sum_{j=1}^{N} Y_j = \bar{Y}$$

$$Var(Z_i) = E((Z_i - E(Z_i))^2) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

$$= \sigma^2 = \frac{N-1}{N} S^2$$

$$E(\hat{\bar{Y}}) = \frac{1}{n}\sum_{i=1}^{n} E(Z_i) = \bar{Y}$$

$$\boxed{E(\hat{\bar{Y}}) = \bar{Y}}$$

$$Var(\hat{\bar{Y}}) = Var(\bar{Z}) = \frac{V(Z_1)}{n} = \frac{\sigma^2}{n} = \frac{(N-1)}{N}\frac{S^2}{n}.$$

$$\boxed{Var(\hat{\bar{Y}}) = \frac{\sigma^2}{n} = \frac{N-1}{N}\frac{S^2}{n}}$$

$$Var_{SRSWOR}(\hat{\bar{Y}}) = \frac{N-n}{N}\frac{S^2}{n} < \frac{N-1}{N}\frac{S^2}{n} = Var_{SRSWR}(\hat{\bar{Y}})$$

$\Rightarrow$ SRSWOR is more efficient that SRSWR

Note: When $n << N \approx f \equiv 1$

$$Var_{SRSWOR}(\hat{\bar{Y}}) \approx \frac{S^2}{n} \approx Var_{SRSWR}(\hat{\bar{Y}})$$

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i \in S}(Y_i - \hat{\bar{Y}})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2\right]$$

$$E(\hat{S}^2) = \frac{1}{n-1}\left[n(\sigma^2 + \bar{Y}^2) - n\left(\frac{\sigma^2}{n} + \bar{Y}^2\right)\right]$$

$$= \sigma^2$$

$$\boxed{\hat{Var}(\hat{\bar{Y}}) = \frac{\hat{S}^2}{n}.}$$

Sample Mean Based on Distinct Units (SRSWR(n))

Define

$D =$ # number of distinct units in SRSWR(n)

$\hat{\bar{Y}}^* =$ sample mean of distinct units in the sample

$$= \frac{1}{D} \sum_{i=1}^{N} \delta_i Y_i,$$

where

$$\delta_i = \begin{cases} 1, & \text{if } U_i \in \underline{S} \\ 0, & \text{otherwise} \end{cases}, \quad i = 1 \ldots N$$

$E(\delta_i) = \Pr(U_i \in S) = \pi_i = 1 - a$ where $a = (1 - \frac{1}{N})^n$

$E(\delta_i \delta_j) = \Pr(U_i \in \underline{S}\ U_j \in S) = \pi_{ij} = 1 - 2a + b$, $b = (1 - \frac{2}{N})^n$

$$D = \sum_{i=1}^{N} \delta_i$$

Due to symmetry, conditional on $D = d$, $\{S_1 \ldots S_d\}$ (sample set) is uniformly distributed over all $\binom{N}{d}$ subsets of $\{U_1 \ldots, U_N\}$, i.e. $\{S_1 \ldots S_d\}$ is a SRSWOR(d) sample

$$E(\hat{\bar{Y}}^* \mid D) = \bar{Y}$$

$$Var(\hat{\bar{Y}}^* \mid D) = \left(1 - \frac{D}{N}\right) \frac{S^2}{D}$$

$$E(\hat{\bar{Y}}^*) = \bar{Y}$$

$$Var(\hat{\bar{Y}}^*) = E\left(Var(\hat{\bar{Y}}^* \mid D)\right) + Var\left(E(\hat{\bar{Y}}^* \mid D)\right)$$

$$= S^2 \left(E\left(\frac{1}{D}\right) - \frac{1}{N}\right) \geq S^2 \left(\frac{1}{n} - \frac{1}{N}\right)$$

$$= V_{SRSWR(n)}(\bar{Y})$$

$$\hat{Var}(\hat{\bar{Y}}^*) = \left(\frac{1}{D} - \frac{1}{N}\right) S_D^2 \longrightarrow \text{Not an unbiased estimator}$$

$$S_D^2 = \frac{1}{D-1} \sum_{i=1}^{D} (Y_{(i)} - D)^2 \,, \quad \text{where } Y_{(i)}: i^{th} \text{ distinct unit in the sample}$$

Note: When $n << N$, repetition of units is rare

$\Rightarrow D \simeq n \Rightarrow \hat{\bar{Y}}^*$ behaves like $\bar{Y}_{SRSWOR}$

$$E(D) = E\left(\sum_{i=1}^{N} \delta_i\right) = N(1-a)$$

$$E\left(\frac{1}{D}\right) \geq \frac{1}{E(D)} = \frac{1}{N(1-a)}$$

Distribution of D

- $D \in \{1, 2, \ldots, \min\{n, N\}\}$

$$Pr[D=d] = \sum_{i:\{c_1,\ldots,c_d\}} \sum_{\{S_1 \ldots S_n\} = \{c_1 \ldots c_d\}} \frac{1}{N^n}$$

$= \frac{1}{N^n}\binom{N}{d} \times$ # of samples of size n comprising of $\{U_1 \ldots U_d\}$ with each U_i appearing at least once

$$= \binom{N}{d} \frac{M}{N^n}$$

$M = d^n -$ # of samples of size n comprising of $\{U_1, \ldots, U_d\}$ with atleast one U_i missing

$$= d^n - \left|\bigcup_{i=1}^{d} A_i\right| ,$$

where $A_i =$ Collection of all samples of size n comprising of $\{U_1 \ldots U_d\}$ with U_i missing in it

$$\Gamma = d^n - \left[\sum_{i=1}^{d} |A_i| - \sum\sum_{1 \le i < j \le d} |A_i \cap A_j| + \dots + (-1)^{d-1} |A_1 \cap A_2 \cap \dots \cap A_d| \right]$$

$$= d^n - \left[\binom{d}{1}(d-1)^n - \binom{d}{2}(d-2)^n + \dots (-1)^{d-2}\binom{d}{d-1}(d-(d-1))^n \right]$$

$$= d^n - \sum_{j=1}^{d-1} (-1)^{j-1} \binom{d}{j} (d-j)^n$$

Stirling Number of Second Kind

$$\Pr(D=d) = \frac{\binom{N}{d}\left[d^n - \sum_{j=1}^{d-1} (-1)^{j-1}\binom{d}{j}(d-j)^n \right]}{N^n} = \frac{\binom{N}{d} \underline{d} \, S(n, d)}{N^n},$$

$d = 1, \dots, \min\{n, N\}$

$$\boxed{E(D) = E\left(\sum_{i=1}^{N} \delta_i\right) = N(1-a)}$$

$$Var(D) = Var\left(\sum_{i=1}^{N} \delta_i\right)$$

$$= \sum_{i=1}^{N} Var(\delta_i) + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ i \ne j}}^{N} Cov(\delta_i, \delta_j)$$

$$= N a(1-a) + N(N-1)\left[1 - 2a + b - (1-a)^2\right]$$

$$\boxed{Var(D) = N a(1-a) + N(N-1)(b-a^2)}$$