Simple Random Sampling (SRS)

Finite Population

$$\mathcal{P} = \{x_1, x_2, ..., x_N\}$$

$$\Rightarrow \quad \mu = \frac{1}{N}\sum_{j=1}^{N} x_j, \quad \sigma^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \mu)^2, \quad \ldots$$

Sample Draw $n < N$ items from $\mathcal{P}$ without replacement $\Rightarrow \mathcal{S}$.

e.g., $\mathcal{S} = \{x_3, x_7, ..., x_{91}\}$

$$\Rightarrow \quad \widehat{\mu} = \frac{1}{n}\sum_{j\in\mathcal{S}} x_j, \quad \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{j\in\mathcal{S}}(x_j - \widehat{\mu})^2, \quad \ldots$$

Source of Randomness

$$I_j \equiv I(j \in \mathcal{S}), \quad \pi_j \equiv \mathbb{P}(j \in \mathcal{S}), \quad \pi_{j\ell} \equiv \mathbb{P}(j \in \mathcal{S} \text{ and } \ell \in \mathcal{S})$$

$$\widehat{\mu} = \frac{1}{n}\sum_{j=1}^{N} x_j I_j$$

$$\Rightarrow \mathbb{E}(\widehat{\mu}) = \frac{1}{n}\sum_{j=1}^{N} x_j \mathbb{E}(I_j) = \frac{1}{n}\sum_{j=1}^{N} x_j \pi_j$$

$$= \frac{1}{n}\sum_{j=1}^{N} x_j \frac{n}{N} = \frac{1}{N}\sum_{j=1}^{N} x_j = \mu$$

Remark For SRS,

$$\pi_j = \binom{1}{1}\binom{N-1}{n-1}\Big/\binom{N}{n} = \frac{n}{N} \quad \forall \quad j = 1, 2, ..., N.$$

Variance

$$\widehat{\mu} = \frac{1}{n}\sum_{j=1}^{N} x_j I_j$$

$$\Rightarrow \mathbb{V}\text{ar}(\widehat{\mu}) = \frac{1}{n^2}\left[\sum_{j=1}^{N} x_j^2 \mathbb{V}\text{ar}(I_j) + \sum_{j\neq\ell} x_j x_\ell \mathbb{C}\text{ov}(I_j, I_\ell)\right]$$

$$= \frac{1}{n^2}\left[\sum_{j=1}^{N} x_j^2 \pi_j(1-\pi_j) + \sum_{j\neq\ell} x_j x_\ell (\pi_{j\ell} - \pi_j\pi_\ell)\right]$$

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

Exercise Show that, for SRS, $\pi_{j\ell} = (n/N)[(n-1)/(N-1)]$ (easy) and complete the missing steps (not hard but unpleasant).

Unpleasant Details

$$\frac{1}{n^2}\left[\sum_{j=1}^{N} x_j^2 \left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right) + \sum_{j\neq\ell} x_j x_\ell \left(\frac{n}{N}\frac{n-1}{N-1} - \frac{n}{N}\frac{n}{N}\right)\right]$$

$$= \frac{1}{n^2}\left(\frac{n}{N}\right)\left[\sum_{j=1}^{N} x_j^2 \left(\frac{N-n}{N}\right) + \sum_{j\neq\ell} x_j x_\ell \underbrace{\left(\frac{n-1}{N-1} - \frac{n}{N}\right)}_{\frac{n-N}{N(N-1)}}\right]$$

$$= \frac{1}{n^2}\left(\frac{n}{N}\right)\left(\frac{N-n}{N}\right)\frac{1}{N-1}\underbrace{\left[(N-1)\sum_{j=1}^{N} x_j^2 - \sum_{j\neq\ell} x_j x_\ell\right]}_{\stackrel{(\dagger)}{=} N(N-1)\sigma^2}$$

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

Step (†)

$$\mu^2 = \left[\frac{1}{N}\sum_{j=1}^{N} x_j\right]^2 = \frac{1}{N^2}\left[\sum_{j=1}^{N} x_j^2 + \sum_{j\neq\ell} x_j x_\ell\right]$$

$$\sigma^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j-\mu)^2 = \ldots = \frac{1}{N-1}\left\{\sum_{j=1}^{N} x_j^2 - N\mu^2\right\}$$

$$= \frac{1}{N-1}\left\{\sum_{j=1}^{N} x_j^2 - \frac{1}{N}\left[\sum_{j=1}^{N} x_j^2 + \sum_{j\neq\ell} x_j x_\ell\right]\right\}$$

$$= \frac{1}{N(N-1)}\left\{(N-1)\sum_{j=1}^{N} x_j^2 - \sum_{j\neq\ell} x_j x_\ell\right\}$$

Stratified Sampling

Population

$$\mathcal{P} = \underbrace{\{x_{1,1}, ..., x_{1,N_1}\}}_{\mathcal{P}_1} \oplus \underbrace{\{x_{2,1}, ..., x_{2,N_2}\}}_{\mathcal{P}_2} \oplus \ ... \oplus \underbrace{\{x_{K,1}, ..., x_{K,N_K}\}}_{\mathcal{P}_K}$$

$$\mu_k = \frac{1}{N_k} \sum_{j=1}^{N_k} x_{k,j}, \quad \sigma_k^2 = \frac{1}{N_k - 1} \sum_{j=1}^{N_k} (x_{k,j} - \mu_k)^2, \quad k = 1, 2, ..., K$$

Sample Draw simple random sample $\mathcal{S}_k$ of size $n_k < N_k$ from $\mathcal{P}_k$, independently for $k = 1, 2, ..., K \Rightarrow \mathcal{S} = \mathcal{S}_1 \oplus \ ... \oplus \mathcal{S}_K$.

Examples Stratify by race, gender, age (18-24, 25-34, ...), etc.

Population Quantities

$$N = N_1 + N_2 + ... + N_K$$

$$\mu = \frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{N_k} x_{k,j} = \sum_{k=1}^{K}\frac{N_k}{N}\mu_k$$

$$\sigma^2 = \frac{1}{N-1}\sum_{k=1}^{K}\sum_{j=1}^{N_k}(x_{k,j}-\mu)^2 = \frac{1}{N-1}\sum_{k=1}^{K}\sum_{j=1}^{N_k}(x_{k,j}-\mu_k+\mu_k-\mu)^2$$

$$= \frac{1}{N-1}\left[\sum_{k=1}^{K}(N_k-1){\sigma_k}^2 + \sum_{k=1}^{K}N_k(\mu_k-\mu)^2\right]$$

Sample Quantities

Within Strata:

$$\widehat{\mu}_k = \frac{1}{n_k} \sum_{x_{k,j} \in \mathcal{S}_k} x_{k,j} = \frac{1}{n_k} \sum_{j=1}^{N_k} x_{k,j} I(x_{k,j} \in \mathcal{S}_k)$$

$$\mathbb{E}(\widehat{\mu}_k) = ... = \mu_k, \quad \mathbb{V}\text{ar}(\widehat{\mu}_k) = ... = \left(1 - \frac{n_k}{N_k}\right) \frac{\sigma_k^2}{n_k} \quad \text{(from SRS)}$$

Overall: $\widehat{\mu}_{\text{st}} = \sum_{k=1}^{K} \frac{N_k}{N} \widehat{\mu}_k \quad \Rightarrow$

$$\begin{cases} \mathbb{E}(\widehat{\mu}_{\text{st}}) = \sum_{k=1}^{K} \frac{N_k}{N} \mathbb{E}(\widehat{\mu}_k) = \sum_{k=1}^{K} \frac{N_k}{N} \mu_k = \mu \\ \mathbb{V}\text{ar}(\widehat{\mu}_{\text{st}}) = \sum_{k=1}^{K} \left(\frac{N_k}{N}\right)^2 \mathbb{V}\text{ar}(\widehat{\mu}_k) = \sum_{k=1}^{K} \left(\frac{N_k}{N}\right)^2 \left(1 - \frac{n_k}{N_k}\right) \frac{\sigma_k^2}{n_k} \end{cases}$$

Proportional Allocation

If $n_k = n(N_k/N)$, then

$$\mathbb{V}\text{ar}(\widehat{\mu}_{\text{st}}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\left[\sum_{k=1}^{K}\left(\frac{N_k}{N}\right)\sigma_k^2\right];$$

whereas, for SRS,

$$\begin{aligned}
\mathbb{V}\text{ar}(\widehat{\mu}) &= \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n} \\
&= \frac{1}{n}\left(1 - \frac{n}{N}\right)\left[\underbrace{\sum_{k=1}^{K}\frac{(N_k - 1)}{N-1}\sigma_k^2}_{\frac{N_k-1}{N-1}\approx\frac{N_k}{N}} + \underbrace{\sum_{k=1}^{K}\frac{N_k}{N-1}(\mu_k - \mu)^2}_{\text{main reduction here}\\ \text{between-strata variation}}\right].
\end{aligned}$$

almost always better; best if stratas different in terms of μ_k

Horvitz-Thompson Estimator

To estimate population quantity

$$\theta = \sum_{j=1}^{N} h(x_j), \quad \text{e.g.,} \quad h(x_j) = \begin{cases} x_j & (\text{total}) \\ x_j/N & (\text{mean}) \\ I(x_j \in \mathcal{A})/N & (\text{proportion}) \\ \ldots & (\ldots) \end{cases},$$

use

$$\widehat{\theta}_{\text{HT}} = \sum_{j \in \mathcal{S}} \frac{h(x_j)}{\pi_j}, \quad \pi_j = \mathbb{P}(j \in \mathcal{S})$$

where π_j depends on the specific probability sampling scheme.

Remark Principle = inverse probability weighting. Always unbiased (next slide). Generic way to derive estimators.

Horvitz-Thompson Estimator

$$\widehat{\theta}_{\text{HT}} = \sum_{j=1}^{N} \frac{h(x_j)}{\pi_j} I_j$$

$$\Rightarrow \quad \mathbb{E}(\widehat{\theta}_{\text{HT}}) = \sum_{j=1}^{N} \frac{h(x_j)}{\pi_j} \mathbb{E}(I_j) = \sum_{j=1}^{N} \frac{h(x_j)}{\pi_j} \pi_j = \sum_{j=1}^{N} h(x_j) = \theta$$

$$\Rightarrow \quad \mathbb{V}\text{ar}(\widehat{\theta}_{\text{HT}}) = \sum_{j=1}^{N} y_j^2 \mathbb{V}\text{ar}(I_j) + \sum_{j \neq \ell} y_j y_\ell \mathbb{C}\text{ov}(I_j, I_\ell)$$

$$= \sum_{j=1}^{N} y_j^2 \pi_j (1 - \pi_j) + \sum_{j \neq \ell} y_j y_\ell (\pi_{j\ell} - \pi_j \pi_\ell) = \quad \ldots$$

$$\text{where} \quad y_j = h(x_j)/\pi_j$$

Example: SRS

$$\mathbb{P}(j \in \mathcal{S}) = \frac{n}{N} \quad \forall \quad j$$

$$\mu = \frac{1}{N} \sum_{j=1}^{N} x_j \quad \Rightarrow \quad h(x_j) = x_j/N$$

$$\widehat{\mu} = \sum_{j \in \mathcal{S}} \frac{h(x_j)}{\mathbb{P}(j \in \mathcal{S})} = \sum_{j \in \mathcal{S}} \frac{x_j/N}{n/N} = \frac{1}{n} \sum_{j \in \mathcal{S}} x_j$$

Cluster Sampling

Population

$$\mathcal{P} = \underbrace{\{x_{1,1}, ..., x_{1,N_1}\}}_{\mathcal{C}_1} \oplus \underbrace{\{x_{2,1}, ..., x_{2,N_2}\}}_{\mathcal{C}_2} \oplus \ ... \ \oplus \underbrace{\{x_{M,1}, ..., x_{M,N_M}\}}_{\mathcal{C}_M}$$

$$\mu_g = \frac{1}{N_g} \sum_{j=1}^{N_g} x_{g,j}, \quad \sigma_g^2 = \frac{1}{N_g - 1} \sum_{j=1}^{N_g} (x_{g,j} - \mu_g)^2, \quad g = 1, 2, ..., M$$

Sample Draw simple random sample of $m < M$ clusters $\Rightarrow$ e.g., $\mathcal{S} = \mathcal{C}_1 \oplus \mathcal{C}_{23} \oplus \ ... \oplus \mathcal{C}_{97}$.

Example Cluster = household, classroom, postal code, etc.

Similar Analysis

(Details Omitted)

- apply HT principle to obtain unbiased estimator

$$\mu = \sum_{g=1}^{M} \frac{N_g}{N} \mu_g \quad \Rightarrow \quad \widehat{\mu}_{\text{cl}} = \sum_{\mathcal{C}_g \in \mathcal{S}} \frac{(N_g/N)\mu_g}{\mathbb{P}(\mathcal{C}_g \in \mathcal{S})} = \sum_{\mathcal{C}_g \in \mathcal{S}} \frac{(N_g/N)\mu_g}{m/M}$$

- analyze $\mathbb{V}\text{ar}(\widehat{\mu}_{\text{cl}})$

- messier algebra ... slight simplification by assuming $N_1 = N_2 = ... = N_M$ (equal cluster size)

$$\Downarrow$$

not necessarily better; best if clusters similar in terms of μ_g

Illustration

$$\mathcal{P} = \{1, 2, 3, 1, 2, 3, 1, 2, 3\}$$

ideal stratification	ideal clustering
$\mathcal{P}_1 = \{1, 1, 1\}$	$\mathcal{C}_1 = \{1, 2, 3\}$
$\mathcal{P}_2 = \{2, 2, 2\}$	$\mathcal{C}_2 = \{1, 2, 3\}$
$\mathcal{P}_3 = \{3, 3, 3\}$	$\mathcal{C}_3 = \{1, 2, 3\}$
$\sigma_k^2 = 0$	σ_g^2 large
$\sum_k (\mu_k - \mu)^2$ large	$\sum_g (\mu_g - \mu)^2 = 0$

analysis so far "merely" saying these obvious truths

Other Topics

- multi-stage sampling
  - e.g., cluster sampling, then SRS again within cluster
- network sampling (especially relevant at the moment)
  - e.g., snowball sampling (i.e., ask a friend)
- and so on ...

Key to Analysis

probability of ending up in the sample