Survey Statistics — Formulas

Basic concepts & terminology

Population: $U = \{1, \dots, N\}$
Target Population: all students in this high school
Element: 1-Student
Sample: A
$A = \{A : A \subset U\}$
Sample Frame: List of classes in this high school
Sample design $p(\cdot)$: distribution of A
Sampling Errors: error due to random subsampling
Non-sampling Errors:
Measurement errors: Recall bias, response error, misreporting
Selection errors: Nonresponse, improper frame coverage, non-probability sample

Estimation

Sample Inclusion Indicator: $I_k = I_k(A)$
Inclusion probability: $\pi_k = E[I_k] = P(k \in A)$
Joint inclusion probability: $\pi_{kl} = P(k, l \in A)$
Sample size: $n_s = \sum_{k \in U} I_k$
Expected sample size: $E[n_s] = \sum_{k \in U} \pi_k$
Fixed sample size properties:
1) $\sum_{k \in U} \pi_k = n$ 2) $\sum_{k \neq l \in U} \pi_{kl} = n(n-1)$

Horvitz-Thompson Estimator

$$\hat{t}_{HT} = \sum_{k \in A} \frac{y_k}{\pi_k}$$

$$E[\hat{t}_{HT}] = t_y = \sum_U y_k$$

$$V[\hat{t}_{HT}] = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l}$$

$$\hat{V}[\hat{t}_{HT}] = \sum_{k \in A} \sum_{l \in A} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k y_l}{\pi_k \pi_l}$$

Measurable design: $n_k > 0$ for all $k \in U$.
We can construct an unbiased variance estimator
SYG variance formula for a fixed sample size:

$$\hat{V}_{SYG} = -\frac{1}{2} \sum_{k \in A} \sum_{l \in A} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

Design Effect

$$\mathrm{def}(p, Y_{HT}) = \frac{V_p(\hat{V}_{HT})}{V_{SRS}(\hat{V}_{HT})}$$

Equal Probability Element Design

SRS Without Replacement

$$\pi_k = \frac{n}{N}, \quad \pi_{kl} = \frac{n(n-1)}{N(N-1)}$$

$$\hat{t}_{HT} = N\bar{y}$$

$$V[\hat{t}_{HT}] = N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

$$S^2 = \frac{1}{N-1} \sum_U (y_k - \bar{y}_U)^2$$

$$\hat{V}[\hat{t}_{HT}] = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$$

$$s^2 = \frac{1}{n-1} \sum_A (y_k - \bar{y})^2$$

Population proportion: Population mean of domain indicator variable $s_i = f[i \in U_i]$. Let $p_A = N/A^N$

$$S_i^2 = \frac{N}{N-1} p_A (1 - p_A)$$

$$S_i^2 = \frac{1}{n-1} p_A (1 - p_A)$$

$$\hat{V}[I_{HT}] \approx (1-f) \frac{p_A}{n} \approx 1/(4n)$$

Sample size determination:

$$n = \frac{z_{1-\alpha/2}^2 p(1-p)}{e^2} \quad \text{(for proportions)}$$

Finite population correction:

$$f = 1 - \frac{n}{N} \quad \text{and} \quad \hat{V}_{adj} = f \cdot \hat{V}$$

Bernoulli Sampling

$$\pi_k = \pi \quad \forall k$$

$$\pi_{kl} = \pi^2 \quad (k \neq l)$$

$$\hat{t}_{HT} = \frac{1}{\pi} \sum_A y_k$$

$$V[\hat{t}_{HT}] = \frac{1-\pi}{\pi} \sum_U y_k^2$$

Simple Random Sampling with replacement(SIR)

Draw probability $p_k = 1/N$ = probability of selecting element k in a given draw

$$\pi_k = 1 - (1 - 1/N)^m$$

$$\pi_{kl} = 1 - 2(1 - 1/N)^m + (1 - 2/N)^m$$

$$E[n_k] = N \left(1 - \left(1 - \frac{1}{N}\right)^m\right)$$

$$I_{pwr} = \frac{1}{m} \sum_{i=1}^m \frac{Z_i}{p_k} = \frac{N}{m} \sum_{i=1}^m Z_i$$

$$Z_i = y_k \text{ is selected on draw } i$$

$$Z_i \overset{i.i.d.}{\sim} (y_N, z_2^2)$$

Systemic Sampling(SY)

a: Sampling interval
$n = |N/a|$

$$\pi_k = 1/a$$

$$\pi_{kl} = \begin{cases} 1/a, & k, l \in S_i \\ 0, & \text{a.u.} \end{cases}$$

$$I_{HT} = aU_r = a \sum_{k \in U} y_k$$

$$V_{SY}[I_{HT}] = a^2 \left(1 - \frac{1}{a}\right) S_i^2$$

$$S_i^2 = \frac{1}{a-1} \sum_{r=1}^N (U_r - \hat{t})^2$$

when $N = na$,

$$V(I_{HT}) = n^2 a \sum_{r=1}^a (y_r - \hat{y}_r)^2$$

$$= N \cdot SSB = N \cdot (SST - SSW)$$

Unequal Probability Element Design

Poisson Sampling

$$I_i \overset{i.i.d.}{\sim} Bernoulli(\pi_i)$$

$$\pi_{ij} = \pi_i \pi_j$$

$$I_{HT} = \frac{N}{n} \sum_{i=1}^N \frac{I_{ij}}{n}$$

$$\hat{V}[I_{HT}] = \sum_{i=1}^N \frac{I_i}{n} \left(\frac{1}{n} - 1\right) y_i^2$$

$$\hat{V}[I_{HT}] = \sum_{i=1}^N \frac{I_i}{n} \left(\frac{1}{n} - 1\right) y_i^2$$

Optimal design: Minimize $V[I_{HT}]$ subject to $\sum_{i=1}^N \pi_i = n$: $\pi_i \propto y_i$
Hubble estimator

$$I_n = N \frac{\sum_{i=1}^N I_i y_i / n_i}{\sum_{i=1}^N I_i / n_i} = N I_{HT} / N_{HT}$$

PPS Sampling

Draw probabilities $(p_1, \cdots, p_N)$ such that $\sum_{i=1}^N p_i = 1$
Im independent selections of size 1 with replacement
Element k is selected on draw i with probability p_k

$$p_k = \frac{x_k}{\sum_{j=1}^N x_j}$$

$$\pi_k = 1 - (1 - p_k)^m$$

$$\pi_{kl} = 1 - (1 - p_k)^m - (1 - p_l)^m + (1 - p_k - p_l)^m$$

Hansen-Hurwitz estimator

$$I_{HH} = \frac{1}{m} \sum_{i=1}^N Z_i$$

$$Z_i = \frac{p_{ki}}{p_k} = \sum_{k=1}^N \frac{p_k}{p_k} [a_i = k]$$

$$Z_i \overset{i.i.d.}{\sim} (t_p, V_i)$$

π_{PBS}Sampling

Draw-by-draw method
Systemic π_{PPS} sampling
1) Choose $R \sim U(0, a_i^2)$ Unit i is selected if

$$\sum_{j=1}^{i-1} x_j < R + ka \leq \sum_{j=1}^{i} x_j$$

Stratification

$$t_y = \sum_{h=1}^H N_h \bar{y}_{U_h}$$

$$\hat{t}_{HT} = \sum_{h=1}^H \hat{t}_{HT,h}$$

$$V_{ST}[\hat{t}_{HT}] = \sum_{h=1}^H V[\hat{t}_{HT,h}]$$

$$V[\hat{t}_{str}] = \sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

Sample allocation

Minimize $\mathrm{Var}[\hat{t}_{HT}]$ subject to

- Proportional: $n_h = n\frac{N_h}{N}$
- Neyman: $n_h = n\frac{N_h S_h}{\sum N_h S_h}$
- Optimal: $n_h \propto \frac{N_h S_h}{\sqrt{c_h}}$

Coefficient of variation(CV)

$$CV = \sqrt{\hat{V}(\hat{\theta})/\hat{\theta}} = SE(\hat{\theta})/\hat{\theta}$$

Single-Stage Cluster

$U_1 = \{1, \cdots, N_I\}$: Index set of clusters in the population.
U_i: the set of elements in the i-th cluster of size M_i.
y_{ij} : measurement of item y at the j-th element $j = 1, \cdots, M_i$ in cluster i.
$Y = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \sum_{i=1}^{N} Y_i$: Population total.
$N = \sum_{i=1}^{N_I} M_i$: Population size
A_I: Index set of clusters in the sample
$n_I = |A_I|$,the number of sampled clusters
$A = \bigcup_{i\in A_I} U_i$: index set of elements in the sample
$n_A = |A| = \sum_{i\in A_I} M_i$: the number of sampled elements^2

Single-Stage

$$\hat{t} = \frac{N_I}{n_I}\sum_{i\in A_I} t_i$$

$$V[\hat{t}] = N_I^2\left(1 - \frac{n_I}{N_I}\right)\frac{S_t^2}{n_I}$$

Equal size

$$\hat{Y}_U = \frac{1}{n_I}\sum_{i\in A_I} Y_i$$

$$\mathrm{Var}[\hat{Y}_U] = \left(\frac{1}{n_I} - \frac{1}{N_I}\right)\frac{1}{N_I - 1}\sum_{i=1}^{N_I}(\hat{Y}_i - \hat{Y}_U)^2$$

$$= \frac{1}{n_I M}\left(1 - \frac{n_I}{N_I}S_U^2\right)$$

$$S_U^2 = \sum_{i=1}^{N_I} M(\hat{Y}_i - \hat{Y}_U)^2$$

$$S_w^2 = \sum_{i=1}^{N_I}\sum_{j=1}^{M}(\hat{Y}_{ij} - \hat{Y}_i)^2$$

Intracluster Correlation Coefficient

$\rho = \frac{\mathrm{Cov}[y_{ij}, y_{ik}]\ j\neq k]}{\sqrt{\hat{V}(y_{ij})\hat{V}(y_{ik})}}$
$= 1 - \frac{M}{M-1}\frac{SSW}{SST} \approx 1 - S_w^2/S^2$
$\mathrm{deff} = 1 + (M-1)\rho$
Effective sample size

$$n^* = \frac{n_A}{1 + (M-1)\rho}$$

Unequal size

Cluster Inclusion Probability: $\pi_{It} = P(i \in A_I)$
$\pi_{tij} = P(i, j \in A_I)$
Element inclusion probability: $\pi_{tk} = \pi_{It}$ if $k \in U_i$
$\pi_{it,jl} = \begin{cases}\pi_{It} & i = j\\ \pi_{tij} & i \neq j\end{cases}$
$\hat{Y}_{HT} = \sum_{i\in A_I}\frac{Y_i}{\pi_{ti}}$
$V[\hat{V}_{HT}] = \sum_{i\in U_I}\sum_{j\in U_I}\Delta_{tij}\frac{Y_i}{\pi_{ti}\,\pi_{lj}}$
$\hat{V}[\hat{V}_{HT}] = \sum_{i\in A_I}\sum_{j\in A_I}\frac{\Delta_{tij}}{\pi_{tij}}\frac{Y_i}{\pi_{ti}\,\pi_{lj}}$
Simple Random Cluster Sampling(SIC) $\hat{V}_{HT} = N_I Y_{A_I}$
$V(\hat{V}_{HT}) = N_I^2\left(1 - \frac{n_I}{N_I}\right)\frac{S_{U_I}^2}{n_I}$
$\hat{V}(\hat{V}_{HT}) = N_I^2\left(1 - \frac{n_I}{N_I}\right)\frac{S_{A_I}^2}{n_I}$
If $\pi_{It} \propto Y_i$, $\hat{Y}_{HT} = Y$
If $\pi_{It} \propto M_i$ and $\hat{Y}_i$ is constant, $\hat{Y}_{HT} = Y$

Two-Stage Cluster

n_I: Number of PSUs in the sample
m_i: Number of sampled elements in A_i
$\sum_{i\in A_I} m_i = |A|$: The number of sampled elements
Requirements: 1) Invariance: for every $i \in U_I$, A_I such that $i \in A_I$ $p_i(\cdot|A_I) = p_i(.)$
2)Independence of the second-stage design $P(\cup_{i\in A_I} A_i|A_I) = \prod_{i\in A_I} P(A_i|A_I)$
Conditional Inclusion probability $\pi_{k|i} = P(k \in A_i|i \in A_I)$
$\pi_{k|i|i} = P(k, i \in A_i|i \in A_I)$
$\Delta_{k|i|i} = \pi_{k|i|i} - \pi_{k|i|\uparrow|i|}$
Inclusion probability $\pi_{ik} = \pi_{k|i|\pi T_i}$
$\pi_{ik,jl} = \begin{cases}\pi_{It}\pi_{k|i} & i = j, k = i\\ \pi_{It}\pi_{k|i|i} & i = j, k \neq i\\ \pi_{tij}\pi_{k|i|\pi T_{i|j}} & i \neq j\end{cases}$
HT estimator $\hat{t}_{HT} = \sum_{i\in A_I}\frac{\hat{t}_i}{\pi_{It}} = \sum_{i\in A_I}\sum_{j\in A_i}\frac{y_{ij}}{\pi_{ij}}$
$\hat{t}_{t,HT} = \sum_{j\in A_i}\frac{y_{ij}}{\pi_{j|i}}$
$V_{II}[\hat{t}_{t,HT}] = \sum_{j\in U_I}\sum_{k\in U_i}\Delta_{jk|i}\frac{y_{ij}y_{kh}}{\pi_{j|i}\pi_{k|i}} \equiv V_i$
$V(\hat{t}_{t,HT}) = V_{PSU} + V_{SSU}$
$V_{PSU} = \sum_{i\in U_I}\sum_{j\in U_I}\Delta_{tij}\frac{t_i}{\pi_{ti}\,\pi_{tj}}$
$V_{SSU} = \sum_{i\in U_I}\frac{V_i}{\pi_{ti}}$
$\hat{V}_{PSU} = \sum_{i\in A_I}\sum_{j\in A_i}\frac{\Delta_{tij}}{\pi_{tij}}\frac{\hat{t}_i}{\pi_{ti}\,\pi_{tj}}$
$\hat{V}_{SSU} = \sum_{i\in A_I}\frac{\hat{V}_i}{\pi_{ti}^2}$
$E\hat{V}_{PSU} = V_{PSU} + V_{SSU} - \sum_{i\in U_I} V_i$
$E\hat{V}_{SSU} = V_{SSU} = \hat{V}_{SSU}$
$\hat{V}(\hat{t}_{HT}) = \sum_{i\in A_I}\sum_{j\in A_I}\frac{\Delta_{tij}}{\pi_{tij}}\frac{\hat{t}_i}{\pi_{ti}\,\pi_{tj}} + \sum_{i\in A_I}\frac{\hat{V}_i}{\pi_{ti}}$Compton variance estimator: $\hat{V}^* = \hat{V}_{PSU}$. If $A_I = U_I$, stratified sampling.
If $A_i = U_i$, single-stage cluster sampling
SISI design: $\pi_{It} = n_I/N_I$, $\pi_{k|i} = m_i/M_i$
$V_{SISI} = V_{PSU} + V_{SSU}$
$V_{PSU} = N_I^2\left(1 - \frac{n_I}{N_I}\right)S_{U_I}^2/n_I$
$V_{SSU} = \frac{N_I}{n_I}\sum_{i\in U_I} M_i^2\left(1 - \frac{m_i}{M_i}\right)S_{SU_i}^2/m_i$
$S_{U_I}^2 = \frac{1}{N_I - 1}\sum_{i\in U_I}(t_i - \hat{t}_{U_I})^2$
$S_{SU_i}^2 = \frac{1}{M_i - 1}\sum_{j\in U_i}(y_{ij} - \bar{y}_{U_i})^2$
$\hat{V}_{SISI}^2 = N_I^2\left(1 - \frac{n_I}{N_I}\right)S_{A_I}^2/n_I$
If $m_i/M_i = f_2$(constant), let $f_1 = n_I/N_I$
$V[\hat{y}_{HT}] = \frac{1}{n_I}(1 - f_1)B^2 + \frac{1}{n_I m}(1 - f_2)W^2$
$\approx \frac{1}{n_I m}(1 + (m-1)\delta)kS^2$
$= V_{SRS}[\hat{y}_{HT}]k(1 + (m-1)\delta)$
$k = (B^2 + W^2)/S^2$
$\delta = \frac{B^2}{B^2} + W^2$
Minimize variance subject to $C = c_0 + c_1 n_I + c_2 n_I\bar{m}$
$\bar{m}_{opt} = \sqrt{\frac{c_1}{c_2}\frac{W^2}{B^2}}$
Two-stage PPS sampling 1) PPS sampling of n_I clusters with MOS $= M_i$ 2) SISI sampling of m elements in each selected clusters.
Self-weighting design: point estimation easy; $p_i \propto M_i$: efficient; Simple variance estimation
$\hat{y}_{PPS} = \frac{1}{n_I m}\sum_{i\in A_I}\sum_{j\in A_i} y_{ij} = \frac{1}{n_I}\sum_{k=1}^{n_I} z_k$
$z_k = \hat{t}_i/M_i i$ selected in k-th PPS sampling
$\hat{t}_i = \frac{M_i}{m}\sum_{j\in A_i} y_{ij}$
$s_i^2 = \frac{1}{n_I - 1}\sum_{k=1}^{n_I}(z_k - \bar{z}_k)^2$