

MTH 432M: Introduction to Sampling Theory

Credits: 3-1-0-0 (6)

Introduction

Statistical Inference Problems

- We have a population (finite or infinite) which cannot be enumerated (due to cost, time, etc.)
- The objective is to make inferences about a characteristic of a population based on information contained in a sample selected from the population.
- The target population may be finite (e.g., average income per household in a city; eligible voters for voting on an issue in a state) or infinite (the number of germs in a living organism; outcomes of tossing a coin repeatedly).

Question: How sample should be selected?

- Each unit of the population is valuable as it contains certain amount of information about the population.
- But sampling also costs money and time.
- Trade-off Too little information may not yield good inference about a population characteristic whereas too much of information may cost money and time.

- The quantity of information contained in a sample depends on the sample size and variability in the data. This latter factor can be controlled somewhat by the method of selecting the sample (called design of sample survey).
- topic of this course

Some technical terms common to sample surveys

Census: Complete enumeration of the population (sample size = population size)

Element: An object ~~of the population~~ on which observation is made (e.g. head of a family, a voter etc.)

Population: A collection of units about which we want to make inference (e.g. collection of households; collection of voters).

Sampling Unit: Non-overlapping collections of elements from the population (e.g. household heads, voters etc.)

Frame: List of sampling units (e.g. list of all household heads; list of all voters).

- Sample: Collection of Sampling units drawn from a single or multiple frames.
- It is desirable to get a representative sample
→ the one having all salient features of the population.
- If a City comprises of 25% Young voters, 35% of middle-aged voters and 40% Old voters, then we would expect similar representation in the sample.

Probability Sampling: It is desirable that there is an element of randomness in sampling design so that properties of estimators can be studied and reliable estimates can be found.

Advantages of Sampling

- Cost and time reduction (smaller sample size)
- accuracy (desired precision can be ensured)

Framework

- $\mathcal{U} = (U_1, \dots, U_N)$: a finite population of N identifiable units, i.e. units in the population can be labeled, say from 1 to N , where the label of each unit is known (e.g. population of households, population of voters, etc.)
- N : population size, which is finite and generally known.

- Non-identifiable populations: Population of fishes in a lake (neither N is known nor units are identifiable); a population of ^{how} electric bulbs produced by a company.

- We will be interested in finite identifiable population in this course.

- Y : a study variable

$$Y: \mathcal{U} \rightarrow \mathbb{R}$$

$$y_i = Y(U_i): Y \text{ value of unit } U_i \quad (i=1, \dots, N)$$

$$\underline{y} = (y_1, \dots, y_N): \text{Vector of } Y \text{ values}$$

- $\underline{y}$ is unknown and the problem is to make inferences about $\underline{y}$.

Example: To estimate $f(\underline{y})$, for a specified real-valued function f , e.g.

Examples:

$$b(\underline{Y}) = \bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

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

Sample: $\underline{\Delta} = (u_{i1}, \dots, u_{i(n)})$: a collection of some units from $\mathcal{U}$, $u_{ij} \in \{u_1, \dots, u_N\}$,
(an ordered array) $j=1, \dots, n$

$$\underline{Y}(\underline{\Delta}) = (Y(u_{i1}), \dots, Y(u_{in})) \\ = (Y_{i1}, \dots, Y_{in}), \Delta_j \rightarrow \text{completely known}$$

Goal: To make inference about $b(\underline{Y})$ based on $\underline{Y}(\underline{\Delta})$

Note: $\underline{\Delta} = (u_{i1}, \dots, u_{in})$

• For $j \neq k$, u_{ij} and u_{ik} may be the same, i.e. repetition is allowed;

• For $j \neq k$ interchange of u_{ij} and u_{ik} in $\underline{\Delta}$

results in a different sample.

Example: For $N=3$ and $n=2$, $\{u_1, u_3\}$ and $\{u_3, u_1\}$ are considered different samples.

$n(\underline{\Delta})$ = sample size of sample $\underline{\Delta}$

$r(\underline{\Delta})$ = # of distinct units in $\underline{\Delta}$ (effective sample size)

Then

$$(1) \quad n(\underline{\Delta}) \geq r(\underline{\Delta})$$

$$(2) \quad r(\underline{\Delta}) \in \{1, \dots, N\}$$

Sampling Design

- $U = (U_1, \dots, U_N)$: a finite population of N identifiable units
- $S = \{ \underline{\Delta} = (\Delta_1, \dots, \Delta_{n(\underline{\Delta})}) : \Delta_i \in \{U_1, \dots, U_N\} \text{ } (i=1, \dots, n(\underline{\Delta})) \text{ and } n(\underline{\Delta}) \text{ is a positive integer depending on } \underline{\Delta} \}$
 = Collection of all possible samples $\underline{\Delta} = (\Delta_1, \dots, \Delta_{n(\underline{\Delta})})$ from U
 → a ^{countably} infinite set (since repetition of units is allowed)
- S' : a subset of S (finite or countably infinite)
- $\underline{S} = (S_1, \dots, S_{n(\underline{S})})$: a random variable taking values in S according to p.m.b $P(\cdot)$ that satisfies:
 - (i) $P(\underline{\Delta}) = \Pr(\underline{S} = \underline{\Delta})$, $\underline{\Delta} \in S$
 - (ii) $P(\underline{\Delta}) > 0 \quad \forall \underline{\Delta} \in S'$
 - (iii) $\sum_{\underline{\Delta} \in S'} P(\underline{\Delta}) = 1 \quad (\Rightarrow P(\underline{\Delta}) = 0 \quad \forall \underline{\Delta} \in S - S')$

Then

- (S, S', P) is called a sampling design;
- $\underline{S}$ is called a random sample
- $n(\underline{S})$ is called the sample size of random sample $\underline{S} = (S_1, \dots, S_{n(\underline{S})})$
- Note that the sample size $n(\underline{S})$ is a r.v.
- Adoption of a particular sampling design (S, S', P) means that random sample $\underline{S}$ is selected according to p.m.b. P .

- $Y(\underline{S})$ = number of distinct units in random sample $\underline{S} = (S_1, \dots, S_{Y(\underline{S})})$

Then

$$1 \leq Y(\underline{S}) \leq N(\underline{S})$$

Definition 1:

(S, S', P)

- (i) A sampling design π is said to be a fixed size (FS) design of size n (denoted by $FS(n)$) if $n(\underline{\Delta}) = n$, $\forall \underline{\Delta} \in S'$, for some fixed positive integer n .
- (ii) A sampling design (S, S', P) is said to be a fixed effective size (FES) design of size m (denoted by $FES(m)$) if $Y(\underline{\Delta}) = m$, $\forall \underline{\Delta} \in S'$, for some fixed positive integer m .

Example (Simple Random Sampling With Replacement of Sample Size n (SRSWR(n)))

$S' = \{ \underline{\Delta} = (\Delta_1, \dots, \Delta_n) : \Delta_i \in \{U_1, \dots, U_N\}, i = 1, \dots, n \}$
for some positive integer

$\rightarrow$ $FS(n)$ but not FES.

$$|S'| = N^n$$

$$P(\underline{\Delta}) = \Pr(\underline{S} = \underline{\Delta}) = k \text{ (a constant), } \forall \underline{\Delta} \in S'$$

$$\Rightarrow k = \frac{1}{N^n}$$

7A

Thus

$$P(\underline{\Lambda}) = \Pr(S = \underline{\Lambda}) = \begin{cases} \frac{1}{H^n}, & \underline{\Lambda} \in S' \\ 0, & \text{o.w.} \end{cases}$$

Fixed Sample Size = $n(N) = n$

$$P_1(U_i) = \Pr(S_1 = U_i) = \frac{1}{N}, \quad i=1, \dots, N$$

$$P_2(U_j | U_i) = \Pr(S_2 = U_j | S_1 = U_i) = \frac{1}{N}, \quad i, j=1, \dots, N, \quad U_i, U_j \in \{U_1, \dots, U_N\}$$

$$\vdots$$

$$\Pr(U_{ir} | U_{i1}, \dots, U_{ir-1}) = \frac{1}{N}, \quad i_1, \dots, i_r \in \{1, \dots, N\}$$

$$\Pr(U_{i1}, \dots, U_{ir}) = \Pr(S_1 = U_{i1}, \dots, S_r = U_{ir}) = \frac{1}{N^n}, \quad r=1, \dots, n$$

Without Replacement

Example (Simple Random Sample of Size n (SRSWOR(n)))

For some fixed positive integer n

$$S' = \{(\Lambda_1, \dots, \Lambda_n) : \Lambda_i \in \{U_1, \dots, U_N\}, \Lambda_i \neq \Lambda_j, \quad i \neq j\}$$

$$\rightarrow FS(n), FES(n); \quad r(S) = n(S) = n$$

$$|S'| = N(N-1) \dots (N-n+1)$$

$$P(\underline{\Lambda}) = k \text{ (a constant)}, \quad \forall \underline{\Lambda} \in S'$$

$$\Rightarrow P(\underline{\Lambda}) = \begin{cases} \frac{1}{N(N-1) \dots (N-n+1)}, & \underline{\Lambda} \in S' \\ 0, & \text{o.w.} \end{cases}$$

Alternatively, we may take

$$S' = \{(\Lambda_1, \dots, \Lambda_n) : \Lambda_i \in \{U_1, \dots, U_N\}\}$$

$$S' = \{(U_{c_1}, \dots, U_{c_n}) : 1 \leq c_1 < c_2 < \dots < c_n \leq N\}$$

$$\text{and } P(\underline{\Lambda}) = \frac{L_n}{N(N-1) \dots (N-n+1)} = \frac{1}{\binom{N}{n}}, \quad \underline{\Lambda} \in S'$$

→ Sampling Scheme.

7B

Have one can generate random sample $S = (S_1, \dots, S_n)$ sequentially according to above conditional distributions. (Sampling Scheme)

Under the former formulation

$$P_1(U_i) = \Pr(S_1 = U_i) = \frac{1 \times (N-1)(N-2) \dots (N-n+1)}{N(N-1) \dots (N-n+1)} = \frac{1}{N}, \quad i=1, \dots, N$$

$$\begin{aligned} P_2(U_j | U_i) &= \Pr(S_2 = U_j | S_1 = U_i) \\ &= \frac{\Pr(S_1 = U_i, S_2 = U_j)}{\Pr(S_1 = U_i)} = \frac{\frac{1 \times 1 \times (N-2)(N-3) \dots (N-n+1)}{N(N-1) \dots (N-n+1)}}{\frac{1}{N}} \\ &= \frac{1}{N-1}, \quad i \neq j \end{aligned}$$

$$\begin{aligned} P_r(U_{c_r} | U_{c_1}, \dots, U_{c_{r-1}}) &= \frac{\Pr(S_1 = U_{c_1}, \dots, S_{r-1} = U_{c_{r-1}}, S_r = U_{c_r})}{\Pr(S_1 = U_{c_1}, \dots, S_{r-1} = U_{c_{r-1}})} \\ &= \frac{\frac{1}{N(N-1) \dots (N-r+1)}}{\frac{1}{N(N-1) \dots (N-r+2)}} = \frac{1}{N-r+1}, \quad r=2, \dots, n, \\ &\quad c_1 \neq c_2 \neq \dots \neq c_r \end{aligned}$$

$$\begin{aligned} P_r(U_{c_1}, \dots, U_{c_r}) &= \Pr(S_1 = U_{c_1}, \dots, S_r = U_{c_r}) \\ &= \frac{1 \times 1 \times \dots \times 1 \times (N-r) \times (N-r+1) \dots (N-n+1)}{N(N-1) \dots (N-n+1)} \\ &= \frac{1}{N(N-1) \dots (N-r+1)}, \quad r=1, \dots, n, \\ &\quad c_1 \neq c_2 \neq \dots \neq c_r \end{aligned}$$

Here one can generate random sample $\underline{S} = (S_1, \dots, S_n)$ sequentially according to above conditional distr. (Sampling scheme)

Sampling Scheme

Systematic method of drawing a

random sample $\underline{S}$. A sampling scheme provides a sampling design. Moreover for every FS design there is a sampling scheme, which may not be unique.

Example A Sampling Scheme for $SRSWR(N)$

Draw one unit from $\{U_1, \dots, U_N\}$ randomly such that the probability of drawing each unit is the same, i.e.

Replace back the drawn unit in the collection so that all units $\{U_1, \dots, U_N\}$ are available for the 2nd draw.

Draw 2nd unit from available units $\{U_1, \dots, U_N\}$ at random such that the probability of drawing each unit is the same, i.e. $\frac{1}{N}$.

Continue this process of with replacement drawing at each stage n times so that at each stage all units $\{U_1, \dots, U_N\}$ are available for a draw and one unit is drawn at random such that the probability of drawing each unit from available units $\{U_1, \dots, U_N\}$ is the same, i.e. $\frac{1}{N}$.

Draws in different n stages are independent.

Here

$$S' = \{(\Lambda_1, \dots, \Lambda_n) : \Lambda_i \in \{U_1, \dots, U_N\}, i=1, \dots, n\}$$

$\underline{S} = (S_1, \dots, S_n)$ has p.m.b.

$$P(\underline{\Lambda}) = \begin{cases} \frac{1}{N} \cdot \frac{1}{N} \cdots \frac{1}{N}, & \text{if } \underline{\Lambda} \in S' \\ 0 & \text{o.w.} \end{cases}$$

$$= \begin{cases} \frac{1}{N^n}, & \text{if } \underline{\Lambda} \in S' \\ 0 & \text{o.w.} \end{cases}$$

Example

A Sampling Scheme for SRSWOR(N)

Draw one unit from $\{U_1, \dots, U_N\}$ at random so that the probability of drawing each unit is the same, i.e. $\frac{1}{N}$.

Do not replace back the above drawn unit in the collection so that only $N-1$ units among $\{U_1, \dots, U_N\}$ are available for the 2nd draw.

Draw 2nd unit from the available $N-1$ units in the collection at random such that the probability of drawing each unit is the same, i.e. $\frac{1}{N-1}$.

Continue this process of without replacement drawing at each stage r time so that ~~at~~ ^{for} the r th draw $N-r+1$ units are available and one unit is drawn at random from the available $N-r+1$ units (the probability of drawing each of the $N-r+1$ available units is the same, i.e. $\frac{1}{N-r+1}$).

Here

$$S' = \{(\Lambda_1, \dots, \Lambda_n) : \Lambda_i \in \{U_1, \dots, U_N\}, i=1, \dots, n \text{ are distinct}\}$$

$\underline{S} = (S_1, \dots, S_n)$ has p.m.f.

$$P(\underline{\Lambda}) = \begin{cases} \frac{1}{N} \cdot \frac{1}{N-1} \cdots \frac{1}{N-r+1}, & \text{if } \underline{\Lambda} \in S' \\ 0, & \text{o.w.} \end{cases}$$

Inclusion and Exclusion Probabilities

(S, S', P) : A sampling design

$\pi_i \equiv \pi_i(P) =$ probability that random sample $\underline{S} = (S_1, S_2, \dots, S_{n(S)})$ drawn according to P will contain U_i

$$= \sum_{\substack{\Delta = (\Delta_1, \dots, \Delta_{n(\Delta)}) \in S' \\ \Delta_l = U_i, \text{ for some } l \in \{1, \dots, n(\Delta)\}}} P(\Delta), \quad (i = 1, \dots, N)$$

$\{\pi_i : i = 1, \dots, N\}$: ^{Set of} first order inclusion probabilities

For $i \neq j$

$\pi_{ij} \equiv \pi_{ij}(P) =$ probability that random sample $\underline{S} = (S_1, S_2, \dots, S_{n(S)})$ drawn according to P will contain U_i and U_j

$$= \sum_{\substack{\Delta = (\Delta_1, \dots, \Delta_{n(\Delta)}) \in S' \\ (\Delta_l, \Delta_m) = (U_i, U_j), \text{ for some } l, m \in \{1, \dots, n(\Delta)\}}} P(\Delta)$$

$\{\pi_{ij} : 1 \leq i \neq j \leq N\}$: Set of second order inclusion probabilities.

Note:

- (i) $\pi_{ij} = \pi_{ji}$, $(i, j \in \{1, \dots, N\}, i \neq j)$
- (ii) $\pi_{ij} \leq \min\{\pi_i, \pi_j\}$, $(i, j \in \{1, \dots, N\}, i \neq j)$

$\pi_i^c \equiv \pi_i^c(P) =$ probability that random sample $\underline{S} = (S_1, S_2, \dots, S_{n(S)})$ drawn according to P does not contain U_i

$$= 1 - \pi_i$$

$$= \sum_{\substack{\Delta = (\Delta_1, \dots, \Delta_{n(\Delta)}) \in S' \\ \boxed{10} \Delta_l \neq U_i, \forall l = 1, \dots, n(\Delta)}} P(\Delta)$$

For $i \neq j$

$\pi_{ij}^c \equiv \pi_{ij}^c(P) =$ probability that random sample $\underline{S} = (S_1, S_2, \dots, S_{n(\underline{S})})$ drawn according to P neither include U_i nor U_j

$$= \sum_{\substack{\underline{\Delta} = (\Delta_1, \dots, \Delta_{n(\underline{\Delta})}) \in \mathcal{S}' \\ \Delta_l \notin \{U_i, U_j\} \ \forall l=1, \dots, n(\underline{\Delta})}} P(\underline{\Delta})$$

Theorem (i) $\pi_{ij}^c = 1 - \pi_i^c - \pi_j^c + \pi_{ij}^c$, $i \neq j$, $i, j \in \{1, \dots, N\}$

(ii) $\pi_{ij}^c = 1 - \pi_i^c - \pi_j^c + \pi_{ij}^c$

Proof $1 - \pi_{ij}^c =$ ~~the~~ probability that either $\frac{U_i}{S_i}$ or $\frac{U_j}{S_j}$ does not belong to $\underline{S}$

$$= \pi_i^c + \pi_j^c - \pi_{ij}^c \rightarrow \text{proves (i)}$$

$$\Rightarrow \pi_{ij}^c = 1 - \pi_i^c + 1 - \pi_j^c - 1 + \pi_{ij}^c = 1 - \pi_i^c - \pi_j^c + \pi_{ij}^c \rightarrow \text{proves (ii)}$$

Example SRSWR (n) design

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

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

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

$$\pi_{ij} = 1 - \pi_i^c - \pi_j^c + \pi_{ij}^c = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n$$

Example

SRSWOR(n)

$$\pi_i^c = \frac{(N-1)(N-2)\dots(N-n)}{N(N-1)\dots(N-n+1)} = \frac{N-n}{N} = 1 - \frac{n}{N}, \quad i=1, \dots, N$$

$$\pi_i = 1 - \pi_i^c = \frac{n}{N}, \quad i=1, \dots, N$$

For $i, j \in \{1, \dots, N\}$ ($i \neq j$)

$$\pi_{ij}^c = \frac{(N-2)(N-3)\dots(N-n)(N-n-1)}{N(N-1)\dots(N-n+1)} = \frac{(N-n)(N-n-1)}{N(N-1)} \\ = \left(1 - \frac{n}{N}\right)\left(1 - \frac{n}{N-1}\right)$$

$$\pi_{ij} = 1 - \pi_i^c - \pi_j^c + \pi_{ij}^c$$

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

$$= \frac{n(n-1)}{N(N-1)}, \quad i \neq j$$

Note:

For SRSWR(n) and SRSWOR(n)

$$\pi_i = \text{Constant} \quad \forall i \in \{1, \dots, N\}$$

$$\pi_{ij} = \text{Constant} \quad \forall i, j \in \{1, \dots, N\} \quad (i \neq j)$$

→ Equal probability scheme.

Notation:

For $r \in \{1, \dots, n\}$ and $i \in \{1, \dots, N\}$
 $P_r(U_i)$: probability of drawing Unit U_i at the r th draw

and, for $r, \lambda \in \{1, \dots, n\}$, $r \neq \lambda$ and $i, j \in \{1, \dots, N\}$

$P_{r,\lambda}(U_i, U_j)$ = probability of drawing Unit U_i and U_j at the r th and λ th draw respectively,

Theorem For SRSWR(n)

$$P_r(U_i) = \frac{1}{N}, \quad r \in \{1, \dots, n\}, \quad i \in \{1, \dots, N\}$$

and

$$P_{r, \lambda}(U_i, U_j) = \frac{1}{N^2}, \quad r, \lambda \in \{1, \dots, n\}, \quad r \neq \lambda, \quad i, j \in \{1, \dots, N\}$$

Proof.

$$P_r(U_i) = \frac{N \times N \times \dots \times N \times 1 \times N \times \dots \times N}{N^n} = \frac{1}{N}$$

$$P_{r, \lambda}(U_i, U_j) = \frac{N \times N \times \dots \times N \times 1 \times N \times \dots \times N \times 1 \times N \times \dots \times N}{N^n}$$

$$= \frac{N^{n-2}}{N^n} = \frac{1}{N^2}$$

Theorem

For SRSWOR(n)

$$P_r(U_i) = \frac{1}{N}, \quad r \in \{1, \dots, n\}, \quad i \in \{1, \dots, N\}$$

$$P_{r, \lambda}(U_i, U_j) = \begin{cases} \frac{1}{N(N-1)}, & r, \lambda \in \{1, \dots, n\}, \quad r \neq \lambda, \quad i, j \in \{1, \dots, N\}, \quad i \neq j \\ 0, & r, \lambda \in \{1, \dots, n\}, \quad r \neq \lambda, \quad i \in \{1, \dots, N\}, \quad i = j \end{cases}$$

Proof.

$$P_r(U_i) = \frac{(N-1)(N-2) \dots (N-r+1) \times 1 \times (N-r) \dots (N-n+1)}{N(N-1) \dots (N-n+1)}$$

$$= \frac{1}{N}, \quad r \in \{1, \dots, n\}, \quad i \in \{1, \dots, N\}$$

For $r, \lambda \in \{1, \dots, n\}, r \neq \lambda$ and $i \in \{1, \dots, N\}$ clearly

$$P_{r, \lambda}(U_i, U_i) = 0$$

and, for $r, \lambda \in \{1, \dots, n\}, r \neq \lambda$ and $i, j \in \{1, \dots, N\}, i \neq j$

~~$$P_{r, \lambda}(U_i, U_j) = \frac{(N-2)(N-3) \dots (N-r-1)(N-r) \times 1 \times (N-r-1) \dots (N-\lambda-1)(N-\lambda)}{N(N-1) \dots (N-n+1)}$$~~

$$P_{r,1}(u_i, u_j) = \frac{(N-2)(N-3) \dots (N-r) \cdot 1 \cdot (N-r-1)(N-r-2) \dots (N-r-1) \cdot (N-1)}{N(N-1) \dots (N-n+1)} \dots \frac{1}{(N-n+1)}$$

$$= \frac{1}{N(N-1)}$$

Distribution of effective sample size $r(\lambda)$

(S, S', P) : Given sampling design.
 S : Random sample

Then

$$E_p(r(\underline{S})) = \sum_{\lambda \in S'} r(\lambda) P(\lambda)$$

$$E_p(r^2(\underline{S})) = \sum_{\lambda \in S'} r^2(\lambda) P(\lambda)$$

$$V(r(\underline{S})) = E[r^2(\underline{S})] - (E[r(\underline{S})])^2$$

Definition A sampling design is said to be a fixed size design (FS design) of size n (FS(n)) if $n(\lambda) = n, \forall \lambda \in S'$.

Clearly, SRSWR(n) and SRSWOR(n) are FS(n) designs

Definition A sampling design is said to be a fixed effective size design (FES design) of size m if $r(\lambda) = m, \forall \lambda \in S'$.

- Note:
- SRSWOR(n) is a FES(n) design
 - SRSWR(n) is not a FES design
 - For FES(n) design, $E(r(\underline{S})) = n$
 $V(r(\underline{S})) = 0$

Theorem

$$(i) E[Y(S)] = \sum_{i=1}^N \pi_i$$

$$(ii) E[Y(S)(Y(S)-1)] = \sum_{i=1}^N \sum_{\substack{j=1 \\ i \neq j}}^N \pi_{ij}$$

Proof.

For $\Lambda \in S'$, defined

$$\delta_i(\Lambda) = \begin{cases} 1, & \text{if } U_i \in \Lambda \\ 0, & \text{o.w.} \end{cases}$$

(i.e. U_i in some group
of Λ)
 $i=1, \dots, N$.

Then

$$Y(S) = \sum_{i=1}^N \delta_i(S)$$

$$\begin{aligned} E(Y(S)) &= \sum_{i=1}^N E(\delta_i(S)) = \sum_{i=1}^N P(U_i \in S) \\ &= \sum_{i=1}^N \pi_i \end{aligned}$$

For $i, j \in \{1, \dots, N\}$, $i \neq j$ and $\Lambda \in S'$

$$\delta_i(\Lambda) \delta_j(\Lambda) = \begin{cases} 1, & \text{if } U_i, U_j \in \Lambda \\ 0, & \text{o.w.} \end{cases}$$

~~$$E[\delta_i(S) \delta_j(S)] = P(U_i, U_j \in \Lambda)$$~~

$$\begin{aligned} \sum_{i=1}^N \sum_{\substack{j=1 \\ i \neq j}}^N \delta_i(S) \delta_j(S) &= \sum_{i=1}^N \delta_i(S) \left(\sum_{j=1}^N \delta_j(S) - \delta_i(S) \right) \\ &= \left(\sum_{i=1}^N \delta_i(S) \right)^2 - \sum_{i=1}^N \delta_i^2(S) \\ &= [Y(S)]^2 - \sum_{i=1}^N \delta_i(S) \end{aligned}$$

$$= E[Y^2(S)] - E[Y(S)]$$

$$\begin{aligned} E(Y(S)(Y(S)-1)) &= E[Y^2(S) - Y(S)] \\ &= E[Y^2(S)] - E[Y(S)] \\ &= \sum_{i=1}^N \sum_{\substack{j=1 \\ i \neq j}}^N E[\delta_i(S) \delta_j(S)] \\ &= \sum_{i=1}^N \sum_{\substack{j=1 \\ i \neq j}}^N \pi_{ij} \end{aligned}$$

(15)

Example SRSWR(n)

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

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

$$E(Y|S) = \sum_{i=1}^N \pi_i = Na = N \left[1 - \left(1 - \frac{1}{N}\right)^n\right]$$

$$E(Y|S)(Y|S)-1) = \sum_{i=1}^N \sum_{\substack{j=1 \\ (i \neq j)}}^N \pi_{ij} = N(N-1)b$$

$$\text{Var}(Y|S) = E[Y^2|S] - [E(Y|S)]^2$$

$$= N(N-1)b + Na - N^2a^2$$

$$\begin{aligned} &= N(N-1) - 2N(N-1)\left(1 - \frac{1}{N}\right)^n + N(N-1)\left(1 - \frac{2}{N}\right)^n \\ &\quad + N - N\left(1 - \frac{1}{N}\right)^n - N^2\left[1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{1}{N}\right)^{2n}\right] \\ &= N + N\left(1 - \frac{1}{N}\right)^n + N(N-1)\left(1 - \frac{2}{N}\right)^n - N^2\left(1 - \frac{1}{N}\right)^{2n} \end{aligned}$$

Theorem

For any FES(n) design

$$(i) \quad \sum_{i=1}^N \pi_i = n$$

$$(ii) \quad \sum_{i=1}^N \sum_{\substack{j=1 \\ (i \neq j)}}^N \pi_{ij} = n(n-1)$$

Hint: For FES(n), $Y(S) \equiv n$ w.p.1

Example SRSWOR(n)

$$Y(S) = n(S) = n \Rightarrow \text{FES}(n)$$

We have seen that

$$\pi_i = \frac{n}{N}, \quad i = 1, \dots, N,$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}, \quad (i \neq j)$$

$$\Rightarrow \sum_{i=1}^N \pi_i = n$$

and

$$\sum_{i=1}^N \sum_{\substack{j=1 \\ (i \neq j)}}^N \pi_{ij} = n(n-1).$$

Theorem

For any FES(n) design

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

Proof

Define

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

Then

$$\sum_{i=1}^N \delta_i(\underline{s}) = r(\underline{s}) = n \quad \text{u.f.i.} \quad i=1, \dots, N$$

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

$$E(\delta_i(\underline{s}) \delta_j(\underline{s})) = \pi_{ij}, \quad i \neq j.$$

$$\sum_{\substack{j=1 \\ j \neq i}}^N \pi_{ij} = \sum_{\substack{j=1 \\ j \neq i}}^N E(\delta_i(\underline{s}) \delta_j(\underline{s}))$$

$$= E \left[\sum_{\substack{j=1 \\ j \neq i}}^N \delta_i(\underline{s}) \delta_j(\underline{s}) \right]$$

$$= E \left[\delta_i(\underline{s}) \left(\sum_{j=1}^N \delta_j(\underline{s}) - \delta_i(\underline{s}) \right) \right]$$

$$= E \left[\delta_i(\underline{s}) (n - \delta_i(\underline{s})) \right]$$

$$= n\pi_i - \pi_i \quad | \delta_i^2 = \delta_i$$

$$= (n-1)\pi_i.$$

Distribution of Count for SRSWR and SRSWOR.

Define

A_i = # of times unit U_i appears in the random sample, $i=1, \dots, N$

SRSWR (n)

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

$$E(A_i) = \frac{n}{N}, \quad \text{Var}(A_i) = n \times \frac{1}{N} \times (1 - \frac{1}{N}) = \frac{n(N-1)}{N^2}$$

for $i \neq j$

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

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

$$\Rightarrow \text{Var}(A_i + A_j) = n \times \frac{2}{N} \times (1 - \frac{2}{N})$$

$$\Rightarrow \frac{2n(N-1)}{N^2} + 2 \text{Cov}(A_i, A_j) = \frac{2n(N-2)}{N^2}$$

$$\Rightarrow \text{Cov}(A_i, A_j) = -\frac{n}{N^2}$$

Alt. $\pi_{A_i, A_j}(t_1, t_2) = (\frac{1}{N} e^{t_1} + \frac{1}{N} e^{t_2} + (1 - \frac{2}{N}))^n$

$$E(A_i A_j) = \left[\frac{\partial^2}{\partial t_1 \partial t_2} \pi_{A_i, A_j}(t_1, t_2) \right]_{(t_1, t_2) = (0, 0)}$$

SRSWOR (n)

$$A_i \in \{0, 1\}, \quad i = 1, \dots, n$$

$$\text{Pr}(A_i = 1) = 1 - \text{Pr}(A_i = 0) = \frac{n}{N} = \pi_i$$

$$\text{Pr}(A_i = 1, A_j = 1) = \frac{n(n-1)}{N(N-1)} = \pi_{ij}, \quad i \neq j$$

$$E(A_i) = \frac{n}{N}; \quad A_i^2 = A_i, \quad i = 1, \dots, n$$

$$E(A_i^2) = \frac{n}{N}, \quad i = 1, \dots, n$$

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

$$\text{Cov}(A_i, A_j) = E(A_i A_j) - \left(\frac{n}{N}\right)^2$$

$$= \text{Pr}(A_i = 1, A_j = 1) - \left(\frac{n}{N}\right)^2$$

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

Estimation of Population Total

$$U = (U_1, \dots, U_N)$$

$y \equiv y(U)$: Study variable

$$y = (y_1, \dots, y_N)$$

Goal:

To estimate $b(y) = b(y_1, \dots, y_N)$, where b is a real-valued function b :

Examples

(i) $b(y) = \sum_{i=1}^N y_i$ (estimation of pop. total)

(ii) $b(y) = \frac{1}{N} \sum_{i=1}^N y_i$ (estimation of pop. mean)

(iii) Let A be a certain attribute (e.g. Smoking)

Define

$$z_i = \begin{cases} 1, & \text{if } U_i \text{ possesses } A \\ 0, & \text{o.w.} \end{cases}, \quad i=1, \dots, N.$$

$$b(z) = \frac{1}{N} \sum_{i=1}^N z_i$$

= proportion of units in pop. possessing attribute A .

$$\underline{S} = (S_1, \dots, S_{n(S)})$$

a random sample drawn according to sampling design, here $S_i \in \{U_1, \dots, U_N\}$, $i=1, \dots, n(S)$

$$y(\underline{S}) = (y(S_1), \dots, y(S_{n(S)}))$$

$$= (y_1, \dots, y_{n(S)})$$

a r.v. that is exactly known after sample is taken $y(\underline{S})$

Estimator of $b(y)$

A statistic $T(S, y(\underline{S}))$ that depends on y only through $y(\underline{S})$ and is defined for each realization of $\underline{S}$.

$$|T - b(y)| \text{ or } (T - b(y))^2$$

→ error in estimating $b(y)$ by T

$$E((T - b(y))^2) = \sum_{\Delta \in S'} (T(\Delta) - b(y))^2 P(\Delta)$$

"
MSE(T)

→ mean squared error (MSE) of estimator T

Smaller value of MSE indicates that estimator is good

MSE can be used as criterion of goodness of an estimator

Unbiased Estimator A statistic $T(y)$ is said to be an unbiased estimator of $b(y)$ if

$$E(T) = b(y)$$

$$\text{i.e. } \sum_{\Delta \in S'} T(\Delta) P(\Delta) = b(y)$$

For an unbiased estimator

$$MSE(T) = \text{Var}(T) = E(T^2) - (E(T))^2$$