Sampling Design

Some definitions :

① **Survey Population** >>

Let, N be a known number of units, e.g. hospitals, schools, colleges etc., each assignable identifying labels $1, 2, \ldots, N$ and taking values $Y_1, Y_2, \ldots, Y_N$ respectively of a real valued random variable Y, which are initially unknown to the investigator who is interested to estimate the total $Y = \sum_{i=1}^{N} Y_i$ or the mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$.

We call the sequence $U = \{1, 2, \ldots, N\}$ of labels a survey population. Thus, the survey population means a finite population in which the individuals are such that they are all concentrate and tangible (tagged with labels).

② **Sample**>> Selecting units from the population leads to a sequence $s = \{i_1, i_2, \ldots, i_n\}$ which is called a sample. Here $i_1, i_2, \ldots, i_n$ are the elements of U, not necessarily distinct from one another, but the order of its appearance is maintained. $n = n(s)$ is ~~refers~~ referred as the size of the sample s, while the effective sample size is denoted by ~~v~~ $v(s)$ and defined as $v(s) = |s|$ (ie. cardinality of s) ie. $v(s)$ is the no. of distinct elements in the sample s.

③ **Survey** >>

By survey, we mean the process of ascertaining the variate values. Censns or complete enumeration is carried out to help the subsequent survey.

④ **Survey Data** >> Once a specific sample is chosen, we suppose that it is possible to ascertain the values $y_{i_1}, y_{i_2}, \ldots, y_{i_n}$ of the variable Y with the respective ~~units~~ units in s. Then

$$d = \{(i_1, y_{i_1}), (i_2, y_{i_2}), \ldots, (i_n, y_{i_n})\}$$

$$= \{(i, y_i), i \in s\}$$ is called the survey data.

⑤ Estimator >>

An estimator t is a real-valued function $t(d)$, which is free of Y_i for $i \notin s$ but may or may not involve Y_i for $i \in s$.

$t(d)$ can be written alternatively as $t(d) = t(s, \underset{\sim}{Y})$

where $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_N)$

An estimator of the special importance is the sample mean $(\bar{y})$ (for estimating the population mean) defined as,

$$t(s, \underset{\sim}{Y}) = \frac{1}{n(s)} \sum_{i=1}^{N} f_{si} Y_i \qquad [\, f_{si} = 0 \text{ if } i \notin s \,]$$

where, f_{si} is the frequency of the ith unit in the sample s, such that $\sum_{i=1}^{N} f_{si} = n(s)$ & $N\bar{y}$ is called the expansion estimator of the population total.

⑥ Linear and homogeneous linear estimator >>

Consider an estimator of the form

$$t(s, \underset{\sim}{Y}) = b_s + \sum_{i=1}^{N} b_{si} Y_i$$

with $b_{si} = 0$ if $i \notin s$ and b_s and b_{si} are free of Y. It is called a linear estimator and keeping $b_s = 0$, we obtain a homogeneous linear estimator.

It should be noted that $t(s, \underset{\sim}{Y})$ is linear or homogeneously linear in Y_i, $i \in s$, but it may be a non-linear function of two random variables, e.g. taking $b_s = 0$, $b_{si} = \frac{X}{\sum_{i=1}^{N} f_{si} X_i}$ where X_i be the value of some other variable X on $i \in U$. This gives

$$t(s, \underset{\sim}{Y}) = \sum_{i=1}^{N} \left(\frac{X}{\sum_{i=1}^{N} f_{si} X_i} Y_i \right) \quad \text{(Ratio estimator)}$$

This is not linear in X and Y.

⑦ **Sampling Design >>**

A sample is chosen by assigning a selection probability to the sample s, denoted by $p(s)$. $p(s)$ satisfying the following properties.

(i) $0 \le p(s) \le 1$

(ii) $\sum_{s \in S} p(s) = 1$

where S is the totality of all samples that can be drawn from the population. p is called a sampling design. Thus, by a sampling design, we mean a probability measure defined on S, satisfying (i) and (ii).

⑧ **Informative and non-informative design >>**

A design may depend on the related variables X, Y etc. But we assume that p is free of Y unless explicitly mentioned. To emphasize this freedom, p is often referred in the literature as a non-informative design. On the other hand if p involves any component of $\underset{\sim}{Y}$, it is called an informative design.

⑨ **WOR design and WR design >>**

A design p is WOR design if no repetitions occur in any sample s with $p(s) > 0$, otherwise p is called WR design.

⑩ **FS and FES design >>**

A design p is of fixed size n (fixed effective size n) if $p(s) > 0$ implies that s is of size n (effective size n). With respect to the WOR design, there is of course no difference between FS and FES.

⑪ **SRSWOR and SRSWR design:**

A design p is called SRSWOR if $p(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \forall s \text{ with } n(s) = n \\ 0, & \text{o.w.} \end{cases}$

and it is called SRSWR if $p(s) = \begin{cases} \frac{1}{N^n} & \forall s \text{ with } n(s) = n \\ 0 & \text{, o.w.} \end{cases}$

(12) Sampling ~~Start~~ Strategy >>

The combination (p, t) denoting an estimator t based on s chosen according to the sampling design p, is called the sampling strategy.

(13) Inclusion Probability >>

For any arbitrary sampling ~~dsn~~ design p, the probability $\pi_i = \sum_{s \ni i} p(s)$ is called the first order inclusion probability of the unit i.

Again, the probability $\pi_{ij} = \sum_{s \ni \{i,j\}} p(s)$ is called the 2nd order inclusion probability of the units i and j.

$\sum_{i \in s}$:- sum over the units i belonging to s (i varies, s fixed)

$\sum_{s \ni i}$:- sum over the samples s containing the unit i (i fixed, s varies)

$s \ni i$ → s includes i

$i \in s$ → i belongs to s.

Example

$N = 5$; $n = 3$

SRSWOR

Total no. of possible samples = $\binom{5}{3} = 10$

$s_1 = \{1, 2, 3\}$ $s_4 = \{1, 3, 4\}$ $s_7 = \{2, 3, 4\}$ $s_{10} = \{3, 4, 5\}$

$s_2 = \{1, 2, 4\}$ $s_5 = \{1, 3, 5\}$ $s_8 = \{2, 3, 5\}$

$s_3 = \{1, 2, 5\}$ $s_6 = \{1, 4, 5\}$ $s_9 = \{2, 4, 5\}$

$p(s_1) = \dots = p(s_{10}) = \frac{1}{10}$

π_1 = P(unit 1 is included in the sample) = $\sum_{s \ni 1} p(s)$

$= p(s_1) + p(s_2) + \ldots + p(s_6)$

$= \frac{6}{10} = \frac{3}{5}$

Similarly, $\pi_2 = \pi_3 = \pi_4 = \pi_5 = \frac{3}{5}$

$\sum_{i=1}^{5} \pi_i = 3$ = the fixed sample size = n

The inclusion probabilities do not add up to 1.

Expectation & MSE

Whatever $\underset{\sim}{Y}$ may be, $E_p(t) = \sum_s p(s)\, t(s, \underset{\sim}{Y})$ is called the expectation of t and

$M_p(t) = E_p\left[t(s, \underset{\sim}{Y}) - Y\right]^2$ is called the MSE of t for estimating Y (the population total)

If $E_p(t(s, \underset{\sim}{Y})) = Y$, then t is called a p-unbiased estimator of Y.

Some Results Regarding Inclusion Probabilities:

(I) $E_p(v(s)) = \sum_{i=1}^{N} \pi_i$

Proof >> We know that $\pi_i = \sum_{s \ni i} p(s)$

Define, $I_{si} = \begin{cases} 1 & \text{if } s \ni i \\ 0 & \text{o.w.} \end{cases}$

$\therefore E_p(I_{si}) = \sum_s p(s) I_{si} \quad \left[\sum_s = \sum_{s \ni i} + \sum_{s \not\ni i}\right]$

$= \sum_{s \ni i} p(s) I_{si} \quad [\because I_{si} = 0 \text{ for } s \not\ni i]$

$= \sum_{s \ni i} p(s) \quad [\because I_{si} = 1 \text{ for } s \ni i]$

$= \pi_i$

$\therefore \sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} E_p(I_{si}) = E_p\left(\sum_{i=1}^{N} I_{si}\right) = E_p(v(s))$

$[\because \sum_{i=1}^{N} I_{si} = v(s)$ is the effective size of the sample

Special Case:

For a FES (n) design, $\nu(s) = n \ \forall s$

Hence, $\sum_{i=1}^{N} \pi_i = n$ (Example SRSWOR (n))

(II) $\sum_{j \neq i} \pi_{ij} = E_p(\nu(s)) - \pi_i$

Proof: Define $I_{sij} = \begin{cases} 1 & \text{if } s \ni i \ \& \ s \ni j \\ 0 & \text{o.w.} \end{cases}$

$$E_p(I_{sij}) = \sum_s p(s) I_{sij}$$

$$= \sum_{s \ni \{i,j\}} p(s) \quad \left[\because I_{sij} = 0 \text{ for } s \not\ni \{i,j\} \right.$$
$$\left. = 1 \text{ for } s \ni \{i,j\} \right]$$

$$= \pi_{ij}$$

Now $I_{sij} = \begin{cases} 1 & \text{iff } I_{si} = 1 \ \& \ I_{sj} = 1 \\ 0 & \text{o.w.} \end{cases}$

If $s \ni i$, $I_{si} = 1$

If $s \ni j$, $I_{sj} = 1$

If $s \ni i \ \& \ s \ni j$ $I_{si} I_{sj} = 1 = I_{sij}$

$$\therefore \boxed{I_{sij} = I_{si} \cdot I_{sj}}$$

Hence, $E_p(I_{si} I_{sj}) = \pi_{ij}$

or, $\sum_{j \neq i} E_p(I_{si} \cdot I_{sj}) = \sum_{j \neq i} \pi_{ij}$

or, $E_p\left(\sum_{j \neq i} I_{si} I_{sj}\right) = \sum_{j \neq i} \pi_{ij}$

or, $E_p\left(I_{si} \sum_{j \neq i} I_{sj}\right) = \sum_{j \neq i} \pi_{ij}$

or, $E_p\left[I_{si}(\nu(s) - I_{si})\right] = \sum_{j \neq i} \pi_{ij} \quad \left[\because \sum_i I_{sj} = \nu(s)\right]$

or, $E_p\left[I_{si}\, v(s) - I_{si}^2 \right] = \sum_{j \neq i} \pi_{ij}$

or, $E_p\left[v(s)\, I_{si} \right] - E_p\left[I_{si}^2 \right] = \sum_{j \neq i} \pi_{ij}$

or, $E_p\left[v(s)\, I_{si} \right] - E_p\left[I_{si} \right] = \sum_{j \neq i} \pi_{ij}$

or, $\boxed{\sum_{j \neq i} \pi_{ij} = E_p\left(v(s)\right) - \pi_i}$

72.

$$\sum_{j \neq i} \pi_{ij} = \sum_{s} p(s)\, v(s)\, I_{si} - \pi_i$$

$$= \sum_{s \ni i} p(s)\, v(s) - \pi_i$$

Special Case : $v(s) = n \quad \forall s$

Then, $\sum_{j \neq i} \pi_{ij} = n \sum_{s \ni i} p(s) - \pi_i$

$= n\pi_i - \pi_i$

$= (n-1)\pi_i$

(III) $\sum_{i=1}^{N} \sum_{j \neq i=1}^{N} \pi_{ij} = V_p\left(v(s)\right) + E_p\left(v(s)\right)\left[E_p\left(v(s)\right) - 1\right]$

Proof » Already proved,

$$\sum_{j \neq i} \pi_{ij} = \sum_{s} p(s)\, v(s)\, I_{si} - \pi_i$$

$$\therefore \sum_{i=1}^{N} \sum_{j \neq i} \pi_{ij} = \sum_{i=1}^{N} \left[\sum_{s} p(s)\, v(s)\, I_{si} \right] - \sum_{i=1}^{N} \pi_i$$

$$= \sum_{s} p(s)\, v(s) \left[\sum_{i=1}^{N} I_{si} \right] - E_p\left(v(s)\right)$$

$$= \sum_{s} p(s) \left\{ v(s) \right\}^2 - E_p\left(v(s)\right)$$

$$= V_p\left(v(s)\right) + \left[E_p\left(v(s)\right) \right]^2 - E_p\left(v(s)\right)$$

$$\therefore \sum_{i=1}^{N} \sum_{j \neq i} \pi_{ij} = V_p\left(v(s)\right) + E_p\left(v(s)\right)\left[E_p\left(v(s)\right) - 1\right]$$

SRSWR (N, n)

$$p(s) = \frac{1}{N^n} \quad \forall s$$

$$\pi_i = \sum_{s \ni i} p(s)$$

$$= \frac{1}{N^n} \sum_{s \ni i} 1$$

$$= \frac{1}{N^n} \left[\text{No. of samples including } i\right]$$

$$= \frac{1}{N^n} \left[\text{Total no. of possible samples} - \text{No. of samples excluding } i\right]$$

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

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

$$E_p\left(v(s)\right) = \sum_{i=1}^{N} \pi_i = N\left[1 - \left(1 - \frac{1}{N}\right)^n\right]$$

$$\pi_{ij} = \sum_{s \ni \{i,j\}} p(s) = P\left(s \ni i \cap s \ni j\right) = P\left(A_i \cap A_j\right)$$

$$= 1 - P\left(A_i \cap A_j\right)^c, \text{ where } A_i = \{s \ni i\}$$

$$A_j = \{s \ni j\}$$

$$= 1 - P\left(A_i^c \cup A_j^c\right)$$

$$= 1 - P\left(A_i^c\right) - P\left(A_j^c\right) + P\left(A_i^c \cap A_j^c\right)$$

$$P(A_i^c) = 1 - \pi_i = \left(1 - \frac{1}{N}\right)^n$$

$$P(A_j^c) = 1 - \pi_j = \left(1 - \frac{1}{N}\right)^n$$

$$P(A_i^c \cap A_j^c) = \frac{\text{No. of samples excluding both } i \text{ \& } j}{N^n}$$

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

$$\therefore \quad \pi_{ij} = 1 - \left(1 - \frac{1}{N}\right)^n - \left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n$$

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

$$\sum_{i \neq j} \sum \pi_{ij} = N(N-1)\left[1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n\right]$$

$$\therefore \quad V_p(v(s)) + E_p(v(s))\left[E_p(v(s)) - 1\right] = N(N-1)\left[1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n\right]$$

$$\therefore \quad V_p(v(s)) = N(N-1)\left[1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n\right] - N^2\left[1 - \left(1 - \frac{1}{N}\right)^n\right]^2 + N\left[1 - \left(1 - \frac{1}{N}\right)^n\right]$$

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

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

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

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

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

Result:

Suppose, $t(s, \underset{\sim}{Y}) = \sum_{i \in s} b_{si} Y_i$ be a homogeneous linear estimator unbiased for Y (w.r.to the sampling design p).

Then, $\boxed{\sum_{s \ni i} p(s)\, b_{si} = 1}$ $\forall i$

Proof >> Since, t is unbiased for Y,

$$E_p\left(t(s, \underset{\sim}{Y})\right) = Y$$

$$\text{or, } \sum_{s} p(s)\, t(s, \underset{\sim}{Y}) = Y$$

$$\text{or, } \sum_{s} p(s) \left[\sum_{i \in s} b_{si} Y_i \right] = Y$$

$$\text{or, } \sum_{i=1}^{N} Y_i \left[\sum_{s \ni i} p(s)\, b_{si} \right] = \sum_{i=1}^{N} Y_i$$

$$\therefore \sum_{s \ni i} p(s)\, b_{si} = 1 \quad \forall i = 1(1)N$$

Previous Example

$N = 5$, $n = 3$, SRSWOR

$\binom{5}{3} = 10$ samples

$s_1, s_2, \ldots, s_{10}$

$$\sum_{s} p(s) \left[\sum_{i \in s} b_{si} Y_i \right]$$

$$= p(s_1)\left[\sum_{i \in s_1} b_{s_1 i} Y_i\right] + p(s_2)\left[\sum_{i \in s_2} b_{s_2 i} Y_i\right] + \ldots + p(s_{10})\left[\sum_{i \in s_{10}} b_{s_{10} i} Y_i\right]$$

$$= p(s_1)\left[b_{s_1 1} Y_1 + b_{s_1 2} Y_2 + b_{s_1 3} Y_3\right] + p(s_2)\left[b_{s_2 1} Y_1 + b_{s_2 2} Y_2 + b_{s_2 4} Y_4\right]$$
$$+ p(s_3)\left[b_{s_3 1} Y_1 + b_{s_3 2} Y_2 + b_{s_3 5} Y_5\right] + p(s_4)\left[b_{s_4 1} Y_1 + b_{s_4 3} Y_3 + b_{s_4 4} Y_4\right]$$
$$+ p(s_5)\left[b_{s_5 1} Y_1 + b_{s_5 3} Y_3 + b_{s_5 5} Y_5\right] + p(s_6)\left[b_{s_6 1} Y_1 + b_{s_6 4} Y_4 + b_{s_6 5} Y_5\right]$$
$$+ p(s_7)\left[b_{s_7 2} Y_2 + b_{s_7 3} Y_3 + b_{s_7 4} Y_4\right] + p(s_8)\left[b_{s_8 2} Y_2 + b_{s_8 3} Y_3 + b_{s_8 5} Y_5\right]$$
$$+ p(s_9)\left[b_{s_9 2} Y_2 + b_{s_9 4} Y_4 + b_{s_9 5} Y_5\right] + p(s_{10})\left[b_{s_{10} 3} Y_3 + b_{s_{10} 4} Y_4 + b_{s_{10} 5} Y_5\right]$$

$$= \sum_{i=1}^{10} Y_i \Big[$$

$$= Y_1 \Big[b_{s11} p(s_1) + b_{s21} p(s_2) + b_{s31} p(s_3) + b_{s41} p(s_4) + b_{s51} p(s_5) + b_{s61} p(s_6) \Big] +$$

$$Y_2 \Big[b_{s12} p(s_1) + b_{s22} p(s_2) + b_{s32} p(s_3) + b_{s42} p(s_4) + b_{s82} p(s_8) + b_{s92} p(s_9) \Big] +$$

$$b_{s13} p(s_1) +$$

$$Y_3 \Big[b_{s13} p(s_1) + b_{s53} p(s_5) + b_{s73} p(s_7) + b_{s83} p(s_8) + b_{s103} p(s_{10}) \Big] +$$

$$Y_4 \Big[b_{s24} p(s_2) + b_{s44} p(s_4) + b_{s64} p(s_6) + b_{s74} p(s_7) + b_{s94} p(s_9) + b_{s10,4} p(s_{10}) \Big]$$

$$Y_5 \Big[b_{s35} p(s_3) + b_{s55} p(s_5) + b_{s65} p(s_6) + b_{s85} p(s_8) + b_{s95} p(s_9) + b_{s10,5} p(s_{10}) \Big]$$

$$= \sum_{i=1}^{N} Y_i \Big[\sum_{s \ni i} p(s) b_{si}$$

$$= Y_1 \Big[\sum_{s \ni 1} p(s) b_{s1} \Big] + Y_2 \Big[\sum_{s \ni 2} b_{s2} p(s) \Big] + \cdots +$$

$$Y_5 \Big[\sum_{s \ni 5} b_{s5} p(s) \Big]$$

$$= \sum_{i=1}^{85} Y_i \Big[\sum_{s \ni i} b_{si} p(s) \Big]$$

* Thus, for a HLUB of the population total,

$$\sum_{s \ni i} b_{si} p(s) = 1 \quad \forall i = 1(1)N$$

Design-Based Inference

The MSE of an estimator t of the population total Y is

$$M_p(t) = E_p\left(t(s,\underset{\sim}{Y}) - Y\right)^2$$

$$= \sum_s p(s)\left(t(s,\underset{\sim}{Y}) - Y\right)^2$$

Let, Σ_1 and Σ_2 denote respectively the sum over the samples s for which $|t(s,\underset{\sim}{Y}) - Y| \geq k > 0$ and $|t(s,\underset{\sim}{Y}) - Y| < k$

i.e. $\Sigma_1 \equiv \sum_{s:\,|t(s,\underset{\sim}{Y})-Y| \geq k}$

& $\Sigma_2 \equiv \sum_{s:\,|t(s,\underset{\sim}{Y})-Y| < k}$

Hence,

$$M_p(t) = \Sigma_1\, p(s)\left(t(s,\underset{\sim}{Y}) - Y\right)^2 + \Sigma_2\, p(s)\left(t(s,\underset{\sim}{Y}) - Y\right)^2$$

$$\geq \Sigma_1\, p(s)\left(t(s,\underset{\sim}{Y}) - Y\right)^2$$

$$\geq k^2 \Sigma_1\, p(s) \quad \left[\because \text{~~$|t(s,\underset{\sim}{Y}) - Y|$~~} \; s \in s_1 \text{ are such that } |t(s,\underset{\sim}{Y}) - Y| \geq k\right]$$

$$= k^2 \sum_{s:\,|t(s,\underset{\sim}{Y})-Y| \geq k} p(s)$$

$$= k^2\, P\left(|t(s,\underset{\sim}{Y}) - Y| \geq k\right)$$

$$\therefore P\left(|t(s,\underset{\sim}{Y}) - Y| \geq k\right) \leq \frac{M_p(t)}{k^2}$$

or, $$P\left(|t(s,\underset{\sim}{Y}) - Y| \leq k\right) \geq 1 - \frac{M_p(t)}{k^2}$$

or, $$P\left(t(s,\underset{\sim}{Y}) - k \leq Y \leq t(s,\underset{\sim}{Y}) + k\right) \geq 1 - \frac{M_p(t)}{k^2}, \; k > 0$$

This alternative version of Chebyshev's inequality is due to Neyman.

The coverage probability is inversely proportional to $M_p(t)$. Thus in order to increase the coverage probability, we need to reduce both the bias and the variance of t.

Special Case: $K = 3\sqrt{V_p(t)}$

$$P\left(t(s,Y) - 3\sqrt{V_p(t)} \le \bar{Y} \le t(s,Y) + 3\sqrt{V_p(t)}\right) \ge 1 - \frac{M_p(t)}{9V_p(t)}$$

$$= 1 - \frac{\left[V_p(t) + \{B_p(t)\}^2\right]}{9V_p(t)}$$

$$= \frac{8}{9} - \frac{\{B_p(t)\}^2}{9V_p(t)}$$

Illustration:

Suppose, Y is a binary random variables, taking values 0 and 1. We consider estimation of $\bar{Y}$.

Here, Y is binary, hence $\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} Y_i$ & $\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}^2$

$$= \frac{1}{N}\sum_{i=1}^{N} Y_i - \bar{Y}^2$$

$$= \bar{Y} - \bar{Y}^2 = \bar{Y}(1-\bar{Y})$$

If we consider SRSWR (n);

$$E_p(\bar{y}) = \bar{Y}$$

& $$V_p(\bar{y}) = \frac{\sigma_Y^2}{n} = \frac{\bar{Y}(1-\bar{Y})}{n}$$

& $$B_p(\bar{y}) = 0$$

$$P\left(t(s,Y) - 3\sqrt{V_p(t)} \le \bar{Y} \le t(s,Y) + 3\sqrt{V_p(t)}\right) \ge \frac{8}{9} - \frac{\{B_p(t)\}^2}{9V_p(t)}$$

Here, $t = \bar{y}$ & $B_p(t) = 0$

$$\therefore P\left(\bar{y} - 3\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}} \le \bar{Y} \le \bar{y} + 3\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right) \ge \frac{8}{9}$$

$V_p(t)$ is maximum at $\bar{Y} = \frac{1}{2}$ & Max $V_p(t) = \frac{1}{4n}$

Thus, replacing $V_p(t)$ by its max value, i.e. $\frac{1}{4n}$, we have,

$$P\left(\bar{y} - \frac{3}{2\sqrt{n}} \leq \bar{Y} \leq \bar{y} + \frac{3}{2\sqrt{n}}\right) \geq \frac{8}{9}.$$

⊡ Comparison between the estimators :

The estimator t of Y is said to be better than another estimator t' of Y if

$$E_p(t) = E_p(t') = Y$$

&

$$V_p(t) \leq V_p(t') \quad \forall \underset{\sim}{Y} \in \Omega \subseteq R^N$$

Where, $\Omega = \left\{ \underset{\sim}{Y} \mid -\infty < a_i < Y_i < b_i < \infty \right\}, i = 1(1)N \}$

We denote by the symbol $\{t \succ t'\}$ that t is better than t'

($\succ$: better than)

An estimator $t_0 = t_0(s, \underset{\sim}{Y})$ is called the best for Y if no other estimator t is better than t_0.

Theorem I :

No best unbiased estimator of Y exists based on any non-census design.

Proof >> A census design is defined as

$$p(s) = \begin{cases} 1 & \text{, if } s \text{ includes every element of } U \\ 0 & \text{, otherwise} \end{cases}$$

Any design other than a census design is called non-census design.

$$\Omega = \left\{ \underset{\sim}{Y} : -\infty < a_i < Y_i < b_i < \infty ; i = 1(1)N \right\}$$

If possible, let $\exists$ some unbiased estimator $t = t(s, \underset{\sim}{Y})$ of Y based on the sampling design p, with the least possible variance.

Let us choose a point $A = (A_1, A_2, \ldots, A_N) \in \Omega$ and define

$$A = \sum_{i=1}^{N} A_i$$

Define a new estimator $t_A(s, \underset{\sim}{Y}) = t(s, \underset{\sim}{Y}) - t(s, \underset{\sim}{A}) + A$

Here, $E_p\left(t_A(s, \underset{\sim}{Y})\right) = E_p\left(t(s, \underset{\sim}{Y})\right) - E_p\left(t(s, \underset{\sim}{A})\right) + A$

$$= Y - A + A$$

$$= Y$$

Thus, $t_A(s, \underset{\sim}{Y})$ is another unbiased estimator of Y.

Again,

$$V_p\left(t_A(s, \underset{\sim}{Y})\right) = E_p\left[t_A(s, \underset{\sim}{Y}) - Y\right]^2$$

$$= E_p\left[t(s, \underset{\sim}{Y}) - t(s, \underset{\sim}{A}) + A - Y\right]^2$$

$$= E_p\left[\left\{t(s, \underset{\sim}{Y}) - Y\right\} - \left\{t(s, \underset{\sim}{A}) - A\right\}\right]^2$$

$$= E_p\left[t(s, \underset{\sim}{Y}) - Y\right]^2 + E_p\left[t(s, \underset{\sim}{A}) - A\right]^2 - 2\, E_p\left[\left\{t(s, \underset{\sim}{Y}) - Y\right\}\left\{t(s, \underset{\sim}{A}) - A\right\}\right]$$

Note that,

$$V_p\left(t_A(s, \underset{\sim}{Y})\right)\Big|_{\underset{\sim}{Y} = \underset{\sim}{A}} = 0$$

But since the point $\underset{\sim}{A}$ is ~~arbitrary~~ arbitrarily chosen, it follows that

$$V_p\left(t_A(s, \underset{\sim}{Y})\right) = 0$$

$$\Rightarrow V_p\left(t(s, \underset{\sim}{Y})\right) = 0$$

$$\Rightarrow t_0(s, \underset{\sim}{Y}) = E_p\left(t(s, \underset{\sim}{Y})\right) \text{ with probability 1}$$

$$\Rightarrow t(s, \underset{\sim}{Y}) = Y \quad \text{w.p. 1.}$$

This is not possible for any ~~non-census~~ non-census design.

Hence, no best u.e. exists for any non-census design.

Theorem 2 :

A necessary and sufficient condition for the existence of an u.e. of the survey population total is

$$\pi_i > 0 \quad \forall\, i = 1(1)N.$$

Proof >> **If Part**

Suppose, $\pi_i > 0 \quad \forall\, i = 1(1)N$

Then, $\frac{Y_i}{\pi_i}$ is defined $\forall i = 1(1)N$

Define $I_{si} = \begin{cases} 1 & , \text{if } s \ni i \\ 0 & , \text{o.w.} \end{cases}$

~~$\therefore E(I_{si}) = \sum_{s \ni i} p(s) = \pi_i$~~

$$\therefore E(I_{si}) = \sum_{s} p(s)\, I_{si} = \sum_{s \ni i} p(s)\, I_{si} + \sum_{s \not\ni i} p(s)\, I_{si}$$

$$= \sum_{s \ni i} p(s) = \pi_i$$

Define, $t = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} \cdot I_{si}$

$$E_p(t) = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} E(I_{si}) = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} \cdot \pi_i = \sum_{i=1}^{N} Y_i = Y$$

$\therefore$ There exists an u.e. t of Y.

Only If Part

Suppose, $\pi_i = 0$ for some $i \in \{1, 2, \ldots, N\}$ & ~~there exists some~~ let, $t = t(s, Y)$ be an u.e. of the population total Y.

$$\therefore \pi_i = \sum_{s \ni i} p(s) = 0 \quad \text{---(1)}$$

$$E_p(t) = Y \Rightarrow \sum_{s} p(s)\, t(s, Y) = \sum_{i=1}^{N} Y_i \quad \text{---(2)}$$

From (1), it follows that $p(s) = 0 \quad \forall\, s \ni i$ since $p(s)$ are non-negative quantities.

From (2), $\sum_{s \ni i} p(s)\, t(s, Y) + \sum_{s \not\ni i} p(s)\, t(s, Y) = \sum_{i=1}^{N} Y_i$

$\Rightarrow \sum_{s \not\ni i} p(s)\, t(s, Y) = \sum_{i=1}^{N} Y_i \quad [\because p(s) = 0 \quad \forall\, s \ni i]$

This is a contradiction, since L.H.S. doesn't involve Y_i but R.H.S. involves Y_i.

Hence, $\pi_i > 0 \quad \forall i = 1(1)N$.

Theorem 3 : (Due to Godambe)

Among the class of HLUEs, no one exists with uniformly minimum variance for any usual design.

Proof >> Let, $t = t(s, Y) = \sum_{i \in s} b_{si} Y_i$ be a HLUE of population total

Since, t is unbiased for Y, it follows that $\sum_{s \ni i} p(s)\, b_{si} = 1 \quad (*)$; $\forall i = 1(1)N$

$$V_p(t) = E_p\left(t^2(s, Y)\right) - \left[E_p\left(t(s, Y)\right)\right]^2$$

$$= \sum_{s} p(s)\, t^2(s, Y) - Y^2$$

$$= \sum_{s} p(s) \left(\sum_{i \in s} b_{si} Y_i\right)^2 - Y^2$$

We minimize $V_p(t)$ subject to the constraints (*)

The Lagrangian function we minimize, is

$$V^* = V_p(t) - \sum_{i=1}^{N} \lambda_i \left(\sum_{s \ni i} p(s)\, b_{si} - 1\right), \text{ where,}$$

$\lambda_1, \lambda_2, \ldots, \lambda_N$ are the Lagrangian multipliers.

The normal equations $\frac{\partial V^*}{\partial b_{si}} = 0, \quad i = 1(1)N$

$$2\, p(s) \left(\sum_{i \in s} b_{si} Y_i\right) Y_i - \lambda_i\, p(s) = 0, \quad i = 1(1)N$$

$$\Rightarrow \lambda_i = 2 \left(\sum b_{si} Y_i\right) Y_i \Rightarrow \sum b_{si} Y_i = \frac{\lambda_i}{2 Y_i}$$

If we assume that $y_i \neq 0$ and $y_j = 0 \ \forall j \neq i$,

$$b_{si}\, y_i = \frac{\lambda_i}{2 y_i}$$

$$\Rightarrow b_{si} = \frac{\lambda_i}{2 y_i^2} \quad (\text{free of } s)$$

Putting the value of b_{si} in (*),

$$\sum_{s \ni i} b_{si}\, p(s) = 1$$

$$\Rightarrow \sum_{s \ni i} \frac{\lambda_i}{2 y_i^2}\, p(s) = 1 \quad \forall i = 1(1) N$$

$$\Rightarrow \frac{\pi_i \lambda_i}{2 y_i^2} = 1 \quad \forall i = 1(1) N$$

$$\Rightarrow \frac{\lambda_i}{2 y_i^2} = \frac{1}{\pi_i} \quad \forall i = 1(1) N$$

$$\therefore \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} y_i \frac{\lambda_i}{2 y_i^2} = \sum_{i \in s} \frac{\lambda_i}{2 y_i} = \frac{\lambda_i}{2 y_i} \left[\because y_j = 0 \ \forall j \neq i\right]$$

So, if $s_1 \ni i$, $s_2 \ni i \Rightarrow p(s_1) > 0$ & $p(s_2) > 0$

then

$$\sum_{i \in s_1} \frac{y_i}{\pi_i} = \sum_{i \in s_2} \frac{y_i}{\pi_i}$$ for the existence of UMVUE in the class of HLUE. This can't be designed unless the design p satisfies the conditions that for s_1, s_2 such that $p(s_1) > 0$, $p(s_2) > 0$, either (i) $s_1 \cap s_2 = \phi$ or (ii) $s_1 \sim s_2$ (equivalent). s_1 and s_2 are equivalent in the sense that both contain an identical set of distinct elements of U. A design p satisfying the property is called 'uni-cluster design' (UCD). For example, the design corresponding to systematic sampling. Any design other than this non-unicluster design (NUCD) (not satisfying these stringent conditions). For an UCD it is possible to realise

$$\sum_{i \in s_1} \frac{y_i}{\pi_i} = \sum_{i \in s_2} \frac{y_i}{\pi_i}$$, uniformly in $\underset{\sim}{y}$ but not for NUCD.

Hence for a NUCD no UMVUE exists.

Rao-Blackwellization:

An estimator $t = t(s, Y)$ may depend on the order in which the units appear in s and may depend on the multiplicity of the appearence of units in s.

Example: 1 Let, $P_i \left[0 < P_i < 1, \sum_{i=1}^{N} P_i = 1\right]$ be known numbers associated with the unit i of U and on the first draw a unit i be chosen with probability P_i and on the second draw from U (leaving aside unit i if chosen in the 1st draw), a unit $j (\neq i)$ be chosen with probability $\frac{P_j}{1-P_i}$. Consider

$$t = t(i,j) = \frac{1}{2}\left[\frac{Y_i}{P_i} + \frac{Y_j}{P_j/(1-P_i)}\right]$$. It can be shown that

$$E_p\left[\frac{Y_i}{P_i}\right] = Y \quad \text{and} \quad E_p\left[Y_i + \frac{Y_j}{P_j/(1-P_i)}\right] = Y$$

and hence, $E_p\left(t(i,j)\right) = Y$ but $t(i,j) \neq t(j,i)$

(Desh Raj's Estimator)

Example: 2 Consider the estimator $t = \frac{1}{n}\sum_{i=1}^{N} \frac{Y_i}{P_i} f_{si}$ (Hansen & Hurwitz's Estimator), where f_{si} is the multiplicity of the unit i in the sample s. This estimator depends upon the multiplicity of units.

With an arbitrary sample $s = \{i_1, i_2, \ldots, i_n\}$ we associate the sample $s^* = \{j_1, j_2, \ldots, j_K\}$ which is equivalent to s (denoted by $s \sim s^*$) and satisfies $j_1 < j_2 < \ldots < j_K$. Hence all the units of s are listed in s^*, without repetations and without taking into account the order in which they appear. So s^* may be called an unordered sample and should be identified without an equivalence class of ordered samples for which repetations are allowed.

Example : $s_1 = \{3, 3, 1, 5, 5\}$; $d_1 = \{(3, \overline{75}), (3, \overline{75}), (1, 50), (5, 100), (5, 100)\}$

$s_2 = \{1, 3, 1, 1, 5\}$; $d_2 = \{(1, 50), (3, \overline{75}), (1, 50), (1, 50), (5, 100)\}$

$s_3 = \{5, 3, 3, 1, 5\}$; $d_3 = \{(5, 100), (3, \overline{75}), (3, \overline{75}), (1, 50), (5, 100)\}$

Then, $s^* = \{1, 3, 5\}$

$d^* = \{(1, 50), (3, \overline{75}), (5, 100)\}$

Let, Ω = Parametric Space

Consider any design p, yielding the survey data $d = \{(i, Y_i) : i \in s\}$, compatible with the subset $\Omega_d = \{\underset{\sim}{Y} \in \Omega \mid Y_i \text{ is observed for } i \in s\}$ of the parametric space.

Define, $I_d(\underset{\sim}{Y}) = \begin{cases} 1 & \text{if } \underset{\sim}{Y} \in \Omega_d \\ 0 & \text{if } \underset{\sim}{Y} \notin \Omega_d \end{cases}$

Then the likelihood of $\underset{\sim}{Y}$ given d is ~~proportional~~ the probability of observing d when $\underset{\sim}{Y}$ is the underlying parametric point ~~$\underset{\sim}{Y}$~~ ie.

$$L_d(\underset{\sim}{Y}) = p(s)\, I_d(\underset{\sim}{Y}) = P_{\underset{\sim}{Y}}(d) \quad \text{(say)}$$

The reduced data is $d^* = \{i; Y_i \mid i \in s^*\}$

Then $\forall d$,

$$I_d(\underset{\sim}{Y}) = I_{d^*}(\underset{\sim}{Y}) \quad \& \quad L_{d^*}(\underset{\sim}{Y}) = p(s^*)\, I_{d^*}(\underset{\sim}{Y}) = P_{\underset{\sim}{Y}}(d^*)$$

For simplicity, we write $P(d)$ instead of $P_{\underset{\sim}{Y}}(d)$

$$\frac{P(d)}{P(d^*)} = \frac{P(d \cap d^*)}{P(d^*)} = P(d \mid d^*)$$

$$\frac{p(s)\, I_d(\underset{\sim}{Y})}{p(s^*)\, I_{d^*}(\underset{\sim}{Y})} = P(d \mid d^*)$$

$$\Rightarrow P(d \mid d^*) = \frac{p(s)}{p(s^*)} \quad \text{free of } \underline{d^*} (= \text{constant})$$

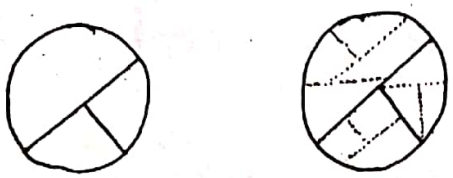
So, d^* is a sufficient statistic, given d.

Let, $\mathcal{D} = \{d\}$ = Data space

d = A specific data point

$t = t(d) = t(\lambda, \chi)$ = A statistic

Every statistic induces a partition on the data space. Partition set $t(d)$ is the collection of sets which are mutually exclusive and exhaustive. A pair of statistics t_1 and t_2 may be compared if the corresponding partitions are such that every partition set induced by t_1 is a union of more than one partition sets induced by t_2 (which means induced by t_2 is contained entirely within one of the partition sets induced by t_1). Then we say that t_1 induces a thicker partition than t_2.



t_1 t_2

If both t_1 and t_2 are sufficient statistics, then we say that t_1 induces a more effective summarization of the data than t_2, while neither sacrifices any relevant information. A sufficient statistic that induces thickest partitioning is called a minimal sufficient statistic and summarization of data is to the maximal extent, sacrificing no relevant information contained in the data.

Theorem

Given a sample $s = \{i_1, i_2, \ldots, i_n\}$ based on a non-informative design p and the data point $d = \{(i_1, y_{i_1}), \ldots, (i_n, y_{i_n})\}$, the statistic $d^* = \{(j_1, y_{j_1}), (j_2, y_{j_2}), \ldots, (j_k, y_{j_k})\}$ derived from d through the sample set $s^* = \{j_1, j_2, \ldots, j_k\}$ obtained from s by subpressing the order and ~~for~~ multiplicity of the units of U, is the minimal sufficient statistic.

Construction of a new estimator

Let, $t = t(d) = t(s, x)$ be a statistic based on d which is an unbiased estimator of γ.

s^* be the set to which one or more samples correspond $(s \sim s^*)$

$\sum_{s \sim s^*} p(s)$ = Probability of observing $s^* = p^*(s^*)$

Then the estimator $t^*(s^*, x) = \dfrac{\sum_{s \sim s^*} p(s)\, t(s, x)}{\sum_{s \sim s^*} p(s)}$ is also unbiased for γ & $V_p(t^*) \leq V_p(t)$

Proof ≫

$$E_p(t) = \sum_{s} p(s)\, t(s, x)$$

$$= \sum_{s^*} \sum_{s \sim s^*} p(s)\, t(s, x)$$

$$= \sum_{s^*} \left[t^*(s^*, x) \sum_{s \sim s^*} p(s) \right]$$

$$= \sum_{s^*} \left[t^*(s^*, x) \cdot p^*(s^*) \right]$$

$$= E_{p^*}\left(t^*(s^*, x) \right) \quad \text{—— (1)}$$

$$= \sum_{s^*} \sum_{s \sim s^*} t^*(s^*, x)\, p(s)$$

For the samples s with $s \sim s^*$, $t^*(s^*, x) = t^*(s, x)$

$$\therefore E_p(t) = \sum_{s^*} \sum_{s \sim s^*} t^*(s, x)\, p(s)$$

$$= \sum_{s} p(s)\, t^*(s, x)$$

$$= E_p(t^*)$$

Thus, t^* is also unbiased w.r.to p.

Again, $E_p(tt^*) = \sum_s t(s,x)\, t^*(s^*,x)\, p(s)$

$= \sum_{s^*} \sum_{s \sim s^*} t(s,x)\, t^*(s^*,x)\, p(s)$

$= \sum_{s^*} \left[\sum_{s \sim s^*} t(s,x)\, p(s) \right] t^*(s^*,x)$

~~$= \sum_{s^*} t^{*2}(s^*,x) \sum_{s \sim s^*} p(s)$~~

$= \sum_{s^*} t^{*2}(s^*,x)\, p^*(s^*)$ [$\because t^*(s^*,x) = \frac{\sum_{s \sim s^*} t(s,x) p(s)}{p^*(s^*)}$]

$= \sum_{s^*} t^{*2}(s^*,x) \left(\sum_{s \sim s^*} p(s) \right)$

$= \sum_{s^*} \sum_{s \sim s^*} t^{*2}(s^*,x)\, p(s)$

$= \sum_{s^*} \sum_{s \sim s^*} t^{*2}(s,x)\, p(s)$ [$\because$ for $s \sim s^*$, $t^*(s^*,x) = t^*(s,x)$]

$= \sum_{s} t^{*2}(s,x)\, p(s)$

$= E_p[t^{*2}]$ —— (2)

Now, we know that $E_p(t - t^*)^2 \geq 0$

or, $E_p(t^2) + E_p(t^{*2}) - 2E_p(tt^*) \geq 0$

or, $E_p(t^2) + E_p(t^{*2}) - 2E_p(t^{*2}) \geq 0$ [from (2)]

or, $E_p(t^2) - E_p(t^{*2}) \geq 0$

or, $E_p(t^2) \geq E_p(t^{*2})$

or, $E_p(t^2) - \{E_p(t)\}^2 \geq E_p(t^{*2}) - \{E_p(t^*)\}^2$

[$\because E_p(t) = E_p(t^*)$]

or, $V_p(t) \geq V_p(t^*)$

Horvitz- Thompson's Estimator of Population Total

Definition:

$$t_{HT} = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} I_{si} = \sum_{i \in s} \frac{Y_i}{\pi_i}$$

$$E_p\left[t_{HT}\right] = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} E_p\left(I_{si}\right) = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} \cdot \pi_i \quad \left[\because E_p\left(I_{si}\right) = \pi_i \text{ already proved}\right]$$

$$= \sum_{i=1}^{N} Y_i = Y$$

$$V_p\left(t_{HT}\right) = V_p\left[\sum_{i=1}^{N} \frac{Y_i}{\pi_i} \cdot I_{si}\right] = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2} V_p\left(I_{si}\right) + \sum\sum_{i \neq j} \frac{Y_i Y_j}{\pi_i \pi_j} + \sum\sum_{i \neq j} \frac{Y_i Y_j}{\pi_i \pi_j} \cdot \text{Cov}_p\left(I_{si}, I_{sj}\right)$$

Now, $V_p\left(I_{si}\right) = E_p\left(I_{si}^2\right) - \left[E_p\left(I_{si}\right)\right]^2$

$= E_p\left(I_{si}\right) - \left[E_p\left(I_{si}\right)\right]^2 \quad \left[\because I_{si} \text{ is an indicator variable}\right]$

$= \pi_i - \pi_i^2 = \pi_i\left(1 - \pi_i\right)$

$\text{Cov}_p\left(I_{si}, I_{sj}\right) = E_p\left(I_{si} \cdot I_{sj}\right) - E_p\left(I_{si}\right) \cdot E_p\left(I_{sj}\right)$

$= E_p\left(I_{sij}\right) - E_p\left(I_{si}\right) E_p\left(I_{sj}\right) \quad \left[\because I_{sij} = I_{si} \cdot I_{sj}\right]$

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

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

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

$= E_p\left[\sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2}\left(1 - \pi_i\right) I_{si} + \sum\sum_{i \neq j} \frac{Y_i Y_j}{\pi_i \pi_j}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}\right) I_{sij}\right]$

Thus, the unbiased estimator of $V_p(t_{HT})$ is

$$\sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2}(1-\pi_i) I_{si} + \sum_{i \neq j}\sum \frac{y_i y_j}{\pi_i \pi_j} \cdot \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} I_{sij}$$

$$= \sum_{i \in s} \frac{y_i^2}{\pi_i^2}(1-\pi_i) + \sum_{i \neq j \in s}\sum \frac{y_i y_j}{\pi_i \pi_j} \cdot \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}}$$

An alternative form of $V_p(t_{HT})$ by Sen, Yates & Grundy

If $v(s) = n$ $\forall s$ & $\pi_{ij} > 0$ $\forall i,j$

$$V_p(t_{HT}) = \sum_{i<j}\sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Proof »

$$\sum_{i<j}\sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

$$= \sum_{i<j}\sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i^2}{\pi_i^2} - 2\frac{y_i y_j}{\pi_i \pi_j} + \frac{y_j^2}{\pi_j^2}\right)$$

$$= \sum_{i<j}\sum \pi_j \frac{y_i^2}{\pi_i} + \sum_{i<j}\sum \pi_i \frac{y_j^2}{\pi_j} - 2\sum_{i<j}\sum y_i y_j$$

$$- \sum_{i<j}\sum \pi_{ij} \frac{y_i^2}{\pi_i^2} + 2\sum_{i<j}\sum \frac{y_i y_j}{\pi_i \pi_j} \pi_{ij} - \sum_{i<j}\sum \pi_{ij} \frac{y_j^2}{\pi_j^2}$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} \left(\sum_{j>i} \pi_j\right) + \sum_{j=1}^{N} \frac{y_j^2}{\pi_j} \left(\sum_{i<j} \pi_i\right) - \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} \left(\sum_{j>i} \pi_{ij}\right)$$

$$- \sum_{j=1}^{N} \frac{y_j^2}{\pi_j^2} \left(\sum_{i<j} \pi_{ij}\right) + 2\sum_{i<j}\sum \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j)$$

Interchanging the role of i and j in term 2 and 4 to have

$$\text{LHS} = \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} \left(\sum_{j>i} \pi_j\right) + \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} \left(\sum_{j<i} \pi_j\right) - \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} \left(\sum_{j>i} \pi_{ij}\right)$$

$$- \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} \left(\sum_{j<i} \pi_{ji}\right) + 2\sum_{i<j}\sum \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j)$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} \left(\sum_{j>i} \pi_j + \sum_{j<i} \pi_j\right) - \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} \left(\sum_{j>i} \pi_{ij} + \sum_{j<i} \pi_{ij}\right)$$

$$+ 2\sum_{i<j}\sum \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j)$$

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

Since, the design is fixed effective size n, hence $\sum_{j=1}^{N} \pi_j = n$ (FES (n))

$$\sum_{j \neq i} \pi_j = n - \pi_i \quad \& \quad \sum_{j \neq i} \pi_{ij} = (n-1)\pi_i$$

$$\text{LHS} = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} (n - \pi_i) - \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2} (n-1)\pi_i + 2 \sum_{i<j} \sum \frac{Y_i Y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j)$$

$$= n \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} - \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} \cdot \pi_i - n \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} + \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2} \cdot \pi_i$$

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

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

$$= V_p(\text{HT})$$