4. Probabilidades de inclusão na amostra

Variável aleatória indicadora de inclusão na amostra

$$I_k = I_k(S) = \begin{cases} 1 \ , \ se \ k \in S \\ 0 \ , \ c.c. \end{cases}$$

A probabilidade do elemento k ser incluído na amostra é função do plano amostral empregado:

$$\begin{aligned} \pi_k &= Prob(k \in S) \\ &= Prob(I_k = 1) \\ &= \sum_{s \ni k} p(s) \end{aligned}$$

- π_k é chamado na literatura de *probabilidade de inclusão de primeira ordem*.

- Para cada elemento da população, existe um π_k associado. Ao todo, existem N probabilidades de inclusão de primeira ordem: $\{\pi_1, \pi_2, ..., \pi_N\}$.

A probabilidade dos elementos k e l serem incluídos na amostra é dada por:

$$\begin{aligned}\pi_{kl} &= Prob(k, l \in S) \\ &= Prob(I_k I_l = 1) \\ &= \sum_{s \ni k, l} p(s)\end{aligned}$$

- π_{kl} é chamado de *probabilidade de inclusão de segunda ordem*.

- Ao todo, existem $N(N-1)/2$ probabilidades de inclusão de segunda ordem.

- Em um plano amostral probabilístico $\pi_k > 0 \;\; \forall \; k \in U$.

- Um plano amostral é dito *mensurável* se π_k e π_{kl} são estritamente positivos para todo $k \neq l \in U$.

- Em um plano amostral com tamanho de amostra fixo n, $\sum_{k \in U} \pi_k = n$.

Propriedades de I_k

$$I_k \sim Bernoulli(\pi_k)$$

$$E_p(I_k) = Prob(I_k = 1) = \pi_k$$

$$Var_p(I_k) = \pi_k\ (1 - \pi_k) = \Delta_{kk}$$

$$E_p(I_k I_l) = Prob(I_k I_l = 1) = \pi_{kl}$$

$$Cov_p(I_k, I_l) = \pi_{kl} - \pi_k \pi_l = \Delta_{kl}$$

5. O Estimador de Horvitz-Thompson (Amostragem sem reposição)

Horvitz and Thompson (1952).

$U = \{1, 2, \ldots, N\}$.

Parâmetro de interesse:

$$\theta = t = \sum_U y_k = N\bar{y}_U$$

O estimador de Horvitz-Thompson para o total populacional é dado por

$$\hat{t}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} \breve{y}_k .$$

Teorema 5.1: $\hat{t}_\pi$ é um estimador centrado para t e tem como variância

$$Var_p(\hat{t}_\pi) = \sum_{k,l \in U} \Delta_{kl}\, \breve{y}_k \breve{y}_l \ .$$

Prova:

Note que $\hat{t}_\pi = \sum_{k \in S} \breve{y}_k = \sum_{k \in U} \frac{y_k}{\pi_k} I_k.$

A única variável aleatória presente nessa expressão é a variável aleatória de inclusão na amostra.

Logo,

$$E_p(\hat{t}_\pi) = \sum_{k \in U} \frac{y_k}{\pi_k} E_p(I_k) = \sum_{k \in U} y_k = t.$$

Além disso,

$$
\begin{aligned}
Var_p(\hat{t}_\pi) &= Var_p\Big(\sum_{k \in U} \breve{y}_k I_k\Big) \\
&= \sum_{k \in U} \breve{y}_k^2 \, Var_p(I_k) + \sum_{k \neq l \in U} \breve{y}_k \breve{y}_l \, Cov_p(I_k, I_l) \\
&= \sum_{k \in U} \Delta_{kk} \breve{y}_k^2 + \sum_{k \neq l \in U} \Delta_{kl} \breve{y}_k \breve{y}_l \\
&= \sum_{kl \in U} \Delta_{kl} \breve{y}_k \breve{y}_l \,,
\end{aligned}
$$

como postulado.

■

Teorema 5.2: Em um plano mensurável, um estimador centrado para $Var_p(t_\pi)$ é dado por

$$\hat{Var}_p(t_\pi) = \sum_{kl \in S} \breve{\Delta}_{kl} \breve{y}_k \breve{y}_l .$$

Prova:

EXERCÍCIO:

Apresente o estimador de Horvitz-Thompson para a média populacional. Prove que ele é centrado e mostre sua variância. Apresente um estimador centrado para a variância do estimador.

6. O Estimador de Hansen-Hurwitz
(Amostragem com reposição)

O estimador de Horvitz-Thompson foi desenvolvido para um cenário de amostragem *sem reposição*.

É importante considerar também o caso de uma amostragem *com reposição*.

Apesar de não ser uma forma eficiente de amostragem, os resultados teóricos obtidos sob este cenário são simples e, por essa razão, podem ser úteis em planos amostrais complexos.

Considere novamente a notação

$$U = \{1,2,\ldots,N\}.$$

Uma amostra, de tamanho m, será retirada de U, usando um esquema amostral similar ao da página 27, admitindo reposição dos elementos selecionados em cada passo.

O resultado obtido é um conjunto ordenado de elementos,

$$os = \{k_1, k_2, \ldots, k_m\},$$

onde k_i representa o elemento selecionado no passo i do esquema amostral.

Admita que $p_1, p_2, \ldots, p_m$ sejam números positivos, satisfazendo a restrição

$$\sum_{k \in U} p_k = 1 ,$$

de tal forma que

p_k=*Prob(Selecionar o elemento* k*)*,

k= 1,2,...N.

Dessa forma, tem-se:

Prob($os = \{k_1, k_2, \ldots, k_m\}$ *)* = $p_{k_1} p_{k_2} \ldots p_{k_m}$.

Note que $\pi_k = 1 - (1 - p_k)^m$.

Logo, se $m > 1$ e p_k for muito pequeno, $\pi_k \approx m\, p_k$.

Parâmetro de interesse:

$$\theta = t = \sum_U y_k = N\bar{y}_U .$$

Defina:

$Z_i = \dfrac{y_k}{p_k}$, se o elemento k for selecionado no i-ésimo passo do esquema amostral.

O estimador de Hansen e Hurwitz(1943) é dado por

$$\hat{t}_{HW} = \frac{1}{m}\sum_{i=1}^{m} Z_i = \bar{Z}.$$

Teorema 6.1: $\hat{t}_{HW}$ é um estimador centrado para t e tem como variância

$$Var_p(\hat{t}_{HW}) = \frac{1}{m}\sum_{k\in U}\left(\frac{y_k}{p_k} - t\right)^2 p_k \ .$$

Prova:

Note que *Prob(* $Z_i = \frac{y_k}{p_k}$ *)=* p_k e que

$Z_1, Z_2, \ldots, Z_m$ são variáveis aleatórias independentes e idênticamente distribuídas.

Logo, os seguintes resultados são verificados:

$$E_p(\hat{t}_{HW}) = \frac{1}{m}\sum_{k\in U} E_p(Z_i)$$

$$= \frac{1}{m}\sum_{k\in U}\sum_{i=1}^{m} \frac{y_k}{p_k} p_k$$

$$= \sum_{k\in U} y_k = t.$$

e

$$Var_p(\hat{t}_{HW}) = \frac{1}{m^2}\sum_{k\in U} E_p(Z_i - t)^2$$

$$= \frac{1}{m}\sum_{k\in U}\left(\frac{y_k}{p_k} - t\right)^2 p_k.$$

■

EXERCÍCIO:

Mostre que

$$\hat{Var}_p(\hat{t}_{HW}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{HW} \right)^2$$

é um estimador centrado para

$$Var_p(\hat{t}_{HW}) = \frac{1}{m} \sum_{k \in U} \left(\frac{y_k}{p_k} - t \right)^2 p_k .$$