Planos amostrais para seleção direta de elementos de uma população.

A implementação dos planos amostrais descritos a seguir é motivada pela disponibilidade de um cadastro do tipo a).

Nesta parte, serão vistos os seguintes planos:

- Amostragem Aleatória Simples sem reposição - AAS
- Amostragem Sistemática - AS
- Amostragem de Bernoulli - BE
- Amostragem de Poisson - POI
- Amostragem com probabilidade proporcional ao tamanho - PPT
- Amostragem Estratificada - AE

7. Amostragem Aleatória Simples (sem reposição) - AAS

Plano amostral

Um plano amostral AAS atribui probabilidade igual de seleção a toda amostra de tamanho n da população (de tamanho N), a saber:

$$p(s) = \binom{N}{n}^{-1}.$$

Sob uma AAS, tem-se:

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

Esquema amostral

Fan, Muller e Rezucha (1962)

Considere uma série de realizações independetes de uma Uniforme(0,1):

$$\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots$$

Se $\varepsilon_1 < \dfrac{n}{N}$, então $k = 1 \in S$;

Senão, $k = 1 \notin S$

Para $k = 2, 3, \ldots$

Se $\varepsilon_k < \dfrac{n - n_k}{N - k + 1}$, então $k \in S$.

Senão, $k \notin S$.

n_k é o número de elementos selecionados para compor a amostra dentre os primeiros $k-1$ elementos listados no cadastro.

O processo continua até que $n_k = n$.

Estimação

Considere a estimação do total populacional de uma população finita sob um plano AAS.

<u>Resultado 7.1:</u> Sob uma AAS, o estimador de H-T assume a forma

$$\hat{t}_\pi = N\bar{y}_s$$

e tem como variância

$$Var_{AAS}(\hat{t}_\pi) = N^2(1-f)\frac{S_{yU}^2}{n}.$$

Além disso,

$$\hat{Var}_{AAS}(\hat{t}_\pi) = N^2(1-f)\frac{S_{ys}^2}{n}$$

é um estimador centrado para $Var_{AAS}(\hat{t}_\pi)$.

Prova:

Sobre a forma do estimador:

$$\hat{t}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k} = \frac{N}{n} \sum_{k \in s} y_k = N\overline{y}_s \; ;$$

Sobre sua variância:

$$Var_p(\hat{t}_\pi) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Sob uma AAS,

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N} \right)^2$$

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

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

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

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

Agora,

$$Var_{AAS}(\hat{t}_\pi) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

$$= \sum_{k\in U} \Delta_{kk}\left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k\in U}\sum_{l\neq k\in U} \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

$$
\begin{aligned}
= \; & \frac{n}{N}\left(1 - \frac{n}{N}\right)\sum_{k \in U}\left(\frac{y_k}{\pi_k}\right)^2 \\
& - \frac{1}{N-1}\left(\frac{n}{N}\right)\left(1 - \frac{n}{N}\right)\sum_{k \in U}\sum_{l \neq k \in U}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}
\end{aligned}
$$

$$
\begin{aligned}
= \; & \frac{n}{N}\left(1 - \frac{n}{N}\right)\left(\frac{N}{n}\right)^2\sum_{k \in U} y_k^2 \\
& - \frac{1}{N-1}\left(\frac{n}{N}\right)\left(1 - \frac{n}{N}\right)\left(\frac{N}{n}\right)^2\sum_{k \in U}\sum_{l \neq k \in U} y_k y_l
\end{aligned}
$$

$$
\begin{aligned}
= \; & \left(1 - \frac{n}{N}\right)\left(\frac{N}{n}\right)\frac{1}{N-1} \\
& \times \left[(N-1)\sum_{k \in U} y_k^2 - \sum_{k \in U}\sum_{l \neq \in U} y_k y_l\right]
\end{aligned}
$$

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

$$\times \left[(N-1)\sum_{k\in U} y_k^2 - \left(\sum_{k\in U} y_k\right)^2 + \sum_{k\in U} y_k^2\right]$$

$$= \left(1 - \frac{n}{N}\right)\left(\frac{N}{n}\right)\frac{1}{N-1}\left[N\sum_{k\in U} y_k^2 - \left(\sum_{k\in U} y_k\right)^2\right]$$

$$= \left(1 - \frac{n}{N}\right)\left(\frac{N^2}{n}\right)\frac{1}{N-1}\left[\sum_{k\in U} y_k^2 - N\bar{y}_U^2\right]$$

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

Sobre o estimador da variância:

A prova segue os mesmos passos do caso anterior.

■

7.1 Intervalos de confiança

Parâmetro de interesse: θ
Estimador centrado: $\hat{\theta}$

Um intervalo de confiança para θ consiste em um par de estatísticas

$$IC(s) \equiv [\hat{\theta}_L(s), \hat{\theta}_H(s)]$$

tal que $\hat{\theta}_L(s) \leq \hat{\theta}_H(s)$ para toda amostra s.

Associado a cada intervalo de confiança, existe uma probabilidade de cobertura (nível de confiança):

$$Prob[IC(s) \ni \theta] = 1 - \alpha .$$

Assim, um $IC(1-\alpha)$ para um parâmetro θ consiste em um intervalo obtido por um procedimento tal que garante uma probabilidade $(1-\alpha)$ de que o intervalo cobrirá o valor de θ.

Num contexto de população finita, a probabilidade de cobertura de um $IC(1-\alpha)$ é interpretada em função do plano amostral $p(s)$, como segue.

Imagine que o valor de θ fosse conhecido e defina os seguintes conjuntos:

$$\Im = \Im_b \cup \Im_b{}^c$$

$$\Im_b = \{ s \in \Im \ : \ IC(s) \ni \theta \}$$

Assim, a probabilidade de cobertura do intervalo é dada por:

$$Prob[IC(s) \ni \theta] = 1 - \alpha \quad ,$$

onde $\alpha = \sum_{s \in \Im_b^c} p(s)$.

Uso na prática:

Na prática, o valor verdadeiro de θ é desconhecido. O intervalo de confiança para θ é construído considerando

$$\frac{\hat{\theta}-\theta}{\sqrt{\hat{Var}_p(\hat{\theta})}} \approx N(0,1)$$

$$\therefore \quad IC(1-\alpha) \equiv \hat{\theta} \pm z_{\alpha/2}\sqrt{\hat{Var}_p(\hat{\theta})},$$

onde $z_{\alpha/2}$ é uma constante tal que
$Prob(Z > z_{\alpha/2}) = \alpha/2$, $Z \sim N(0,1)$.

Com base no procedimento acima, a probabilidade de cobertura do intervalo é, aproximadamente, $1-\alpha$.
A qualidade dessa aproximação depende de duas condições:

1. $\hat{\theta} \sim N(\theta, Var_p(\hat{\theta}))$.

2. $V\hat{a}r_p(\hat{\theta})$ ser um estimador consistente de $Var_p(\hat{\theta})$

$$\frac{\hat{\theta}-\theta}{\sqrt{V\hat{a}r_p(\hat{\theta})}} = \frac{\hat{\theta}-\theta}{\sqrt{Var_p(\hat{\theta})}} \times \sqrt{\frac{Var_p(\hat{\theta})}{V\hat{a}r_p(\hat{\theta})}}$$

7.2 O Teorema Central do Limite de Erdös, Rényi e Hájek

Denote por U_ζ o subconjunto de U onde é verificada a desigualdade

$$(y_k - \bar{y}_U) > \zeta \sqrt{Var_{AAS}(\bar{y}_s)}.$$

Estabeleça ainda a seguinte notação:

$$W_{n,N-n} = \frac{\bar{y}_s - \bar{y}_U}{\sqrt{Var_{AAS}(\bar{y}_s)}}.$$

Então, quando $n \to \infty$ e $N - n \to \infty$, tem-se que

$$W_{n,N-n} \to^D Z, \quad Z \sim N(0,1)$$

desde que $$\lim_{n, N-n \to \infty} \frac{\sum_{k \in U_\zeta} (y_k - \bar{y}_U)^2}{\sum_{k \in U} (y_k - \bar{y}_U)^2} = 0$$

para qualquer $\zeta > 0$.

Prova:

Omitida

Comentários:

- Erdös e Rényi (1959) mostraram que a condição do teorema é suficiente. Hajék (1960) mostrou que a condição é necessária.

- A respeito do problema de quão grande deve ser a amostra para que o TCL seja aplicado, Cochran (1977, pág. 42) indica como regra:

$$n \geq 25\, G_1^2 ,$$

onde $G_1 = \dfrac{\sum_{k \in U} (y_k - \bar{y}_U)^3}{N S_{yU}^3}$ e

$$S_{yU}^2 = (N-1)^{-1} \sum_{k \in U} (y_k - \bar{y}_U)^2 .$$

- Em amostras pequenas, o intervalo de confiança é calculado usando $t_{\alpha/2}$ no lugar de $z_{\alpha/2}$, onde a constante $t_{\alpha/2}$ é tal que $Prob(T > t_{\alpha/2}) = \alpha/2$, com $T \sim t_{(n-1)}$.

7.3 Definições de estimadores consistentes e assintóticamente centrados no contexto de amostragem de populações finitas.

Parâmetro de interesse: θ

Definições usadas em um curso de inferência:

Considere o problema de estimar θ via um estimador $\hat{\theta}_n$, que seja função de n variáveis aleatórias i.i.d:

$$\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$$

Estimador assintóticamente centrado:

(Definição usada em cursos de inferência estatística)

O estimador $\hat{\theta}_n$ (entenda-se a seqüência de estimadores $\hat{\theta}_1, \hat{\theta}_2, \ldots$) é dito ser um estimador assintóticamente centrado para θ, se

$$\lim_{n \to \infty} E(\hat{\theta}_n) = \theta .$$

Estimador consistente:

(Definição usada em cursos de inferência estatística)

O estimador $\hat{\theta}_n$ (entenda-se a seqüência de estimadores $\hat{\theta}_1, \hat{\theta}_2, \ldots$) é dito ser um estimador consistente para θ, se

$$\lim_{n \to \infty} Prob(|\hat{\theta}_n - \theta| > \epsilon) = 0 .$$

Essas definições não fazem sentido em um contexto de inferência para uma população finita.

Definições usadas em um curso de amostragem:

Para definir consistência e centralidade assintótica em amostragem de populações finitas, o seguinte cenário é utilizado:

Considere uma seqüência infinita de elementos $k = 1, 2, \ldots$ e de valores $y_1, y_2 \ldots$ tal que y_k represente o valor da variável de interesse y associado ao elemento k.

Considere ainda uma seqüência de populações $U_1, U_2, \ldots$, onde

$$U_\nu = \{1, 2, \ldots, N_\nu\}$$

com $N_{\nu+1} > N_\nu \quad \forall \; \nu = 1, 2, \ldots$.
Assim, note que $U_1 \subset U_2 \subset U_3 \subset \ldots$.

Defina o parâmentro de interesse

$$\theta_\nu = \theta_\nu(y_1, \ldots, y_{N_\nu})$$

Imagine agora que uma amostra de n_ν elementos seja retirada de U_ν de acordo com o plano amostral $p_\nu(.)$ (n_ν fixo e $n_{\nu+1} > n_\nu \; \forall \; \nu = 1, 2 \ldots$) .

Note que, quando $\nu \to \infty$, tem-se $n_\nu \to \infty$ e $N_\nu \to \infty$.

Seja $\hat{\theta}_\nu = \hat{\theta}_\nu(\{y_k, \forall k \in S_\nu\})$ o estimador de θ_ν em questão.

Considerando as seqüências de populações e de planos amostrais descritas, tem-se as seguintes definições:

Estimador assintóticamente centrado:

O estimador $\hat{\theta}_\nu$ (entenda-se a seqüência de estimadores $\hat{\theta}_1, \hat{\theta}_2, \ldots$) é dito ser um estimador assintóticamente centrado se

$$\lim_{\nu \to \infty} [E_{p_\nu}(\hat{\theta}_n) - \theta_\nu] = 0.$$

Estimador consistente:

O estimador $\hat{\theta}_\nu$ (entenda-se a seqüência de estimadores $\hat{\theta}_1, \hat{\theta}_2, \ldots$) é dito ser um estimador consistente se

$$\lim_{\nu \to \infty} Prob(|\hat{\theta}_\nu - \theta_\nu| > \epsilon) = 0,$$

para qualquer $\epsilon > 0$.

7.4 Cálculo de tamanho de amostra

Parâmetro de interesse: θ

Estimador: $\hat{\theta}$

Denote por d a margem de erro máxima admitida e, por α, a probabilidade dessa margem ser excedida. O objetivo é determinar o tamanho n da amostra tal que

$$Prob(|\hat{\theta} - \theta| > d) < \alpha$$

Assumindo normalidade e centralidade de $\hat{\theta}$, tem-se que

$$\frac{\hat{\theta} - \theta}{\sqrt{Var_p(\hat{\theta})}} \sim N(0,1)$$

Daí,

$$Prob\left(\frac{|\hat{\theta}-\theta|}{\sqrt{Var_p(\hat{\theta})}} > z_\alpha\right) = \alpha$$

Como a variância do estimador é função do tamanho da amostra, o valor de n é encontrado resolvendo a equação

$$z_\alpha \sqrt{Var_p(\hat{\theta})} \ \leq \ d$$

Para estimar o total populacional, por exemplo, sob uma AAS, pode-se verificar que

$$n \ = \ \left(\frac{1}{n_0} + \frac{1}{N}\right)^{-1}$$

onde $$n_0 = \frac{N^2 z_\alpha^2 S_{yU}^2}{d^2}.$$

Uso na prática

Na prática, o valor de S_{yU}^2 deve ser estimado. Isso pode ser feito por meio de uma amostra piloto ou recorrendo a informações de levantamentos passados.

8. Amostragem Sistemática - AS

Plano amostral

Um plano amostral AS atribui probabilidade igual de seleção a um número limitado a de amostras de tamanho n da população (de tamanho N), a saber:

$$p(s) = 1/a.$$

Sob uma AS, tem-se:

$$\pi_k = \frac{1}{a} \quad \text{e} \quad \pi_{kl} = \frac{1}{a}\,1_{[k,l \in s]}.$$

Esquema amostral

Fixe o valor de a de forma que n seja a parte inteira de N/a. Note que,

$$N = na + c,$$

onde $c \in [0, a)$. Se $c = 0$, o esquema amostral apresentado a seguir seleciona uma amostra de tamanho fixo, n. Do contrário, o tamanho da amostra pode vir a ser, também, $n+1$.

i. Selecione, com igual probabilidade, um inteiro r, onde $1 \leq r \leq a$.

ii. A amostra será composta da seguinte forma:

$$s_r = \{k: k = r + (j-1)a; j = 1, 2, \ldots, n_s\},$$

com $n_s = n$ quando $r \leq c$ e $n_s = n + 1$ quando $c < r \leq a$.

O espaço amostral é dado por:

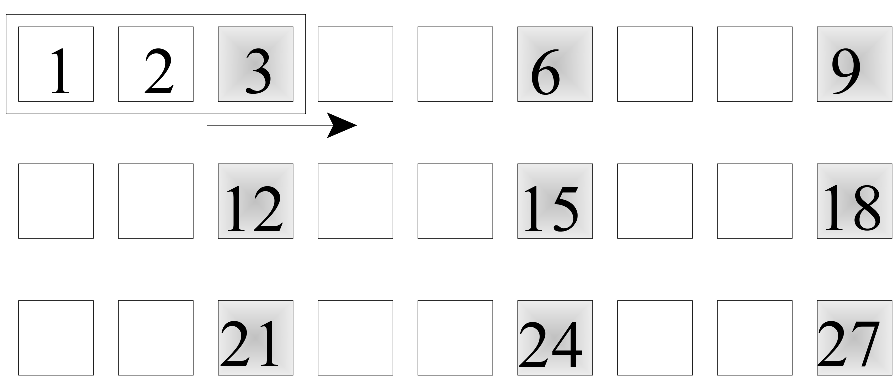
$$\mathfrak{I} = \{s_1, s_2, \ldots, s_a\}.$$

Ilustração

Considere o caso simples em que $N = 27$ e $n = 9$. Assim, $a = 3$ $(c = 0)$.

A figura a seguir ilustra o esquema amostral assumindo $r = 3$.

Cada quadrado representa um elemento, listado no cadastro, na direção esquerda-direita.



Estimação

Considere a estimação do total populacional de uma população finita, sob um plano AS.

A população tem uma estrutura peculiar evidenciada pelo esquema abaixo, para o caso em que $c = 0$:

S_1	$\cdots$	S_r	$\cdots$	S_a
y_1	$\cdots$	y_r	$\cdots$	y_a
y_{1+a}	$\cdots$	y_{r+a}	$\cdots$	y_{2a}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{1+(n-1)a}$	$\cdots$	$y_{r+(n-1)a}$	$\cdots$	y_N
t_{S_1}	$\cdots$	t_{S_r}	$\cdots$	t_{S_a}

Resultado 8.1: Sob uma AS, o estimador de H-T assume a forma

$$\hat{t}_\pi = a \sum_{k \in s} y_k = a\, t_s$$

e tem como variância

$$Var_{AS}(\hat{t}_\pi) = a\,(a-1)\,S^2_{yt},$$

onde $S^2_{yt} = \dfrac{1}{a-1} \sum_{r=1}^{a} (t_{s_r} - \bar{t})^2$ e

$$\bar{t} = \frac{1}{a} \sum_{r=1}^{a} t_{s_r}.$$

Prova:

Sobre a forma do estimador:

$$\hat{t}_\pi = \sum_{k\in S} \frac{y_k}{\pi_k} = a\sum_{k\in s} y_k = a t_s \;;$$

Sobre sua variância:

$$\begin{aligned}
Var_p(\hat{t}_\pi) &= \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \\
&= \sum_{k\in U}\sum_{l\in U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) y_k y_l \\
&= \sum_{k\in U}\sum_{l\in U} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - t^2 \\
&= \sum_{r=1}^{a} \left[\sum_{k\in s_r}\sum_{l\in s_r} \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l \right] - t^2
\end{aligned}$$

Sob uma AS,

$$\frac{\pi_{kl}}{\pi_k \pi_l} = a \text{ , para } k,l \in s_r$$

$$\begin{aligned}
\therefore\ Var_{AS}(\hat{t}_\pi) &= a \sum_{r=1}^{a} \left[\sum_{k \in s_r} \sum_{l \in s_r} y_k y_l \right] - t^2 \\
&= a \sum_{r=1}^{a} t_{s_r}^2 - t^2 \\
&= a \left[\sum_{r=1}^{a} t_{s_r}^2 - \frac{1}{a} t^2 \right] \\
&= a \sum_{r=1}^{a} (t_{s_r} - \bar{t})^2 \\
&= a(a-1) S_{yt}^2
\end{aligned}$$

■

Comentários:

- O plano AS tem a vantagem de ser de fácil implementação. Se necessário, a amostra pode ser selecionada em campo.

- O plano AS pode ser usado para garantir disperção geográfica de elementos selecionados.

- Apesar do plano AS ser de fácil implementação, o espaço amostral, comparado com o de um plano AAS, é bastante reduzido.

- Como $\pi_{kl} = 0$, para $k, l \notin s$, o plano AS não é mensurável.

- Por causa dessas limitações o plano AS é classificado, por alguns, como "quase-aleatório".

Usando o esquema amostral proposto anteriormente, quando $c>0$, o tamanho da amostra varia entre n e $n+1$ conforme $r>c$ ou $r \leq c$, respectivamente.

Além disso, quando $c=0$, em alguns casos, os tamanhos de amostra podem diferir muito daquilo que se planeja.

Exemplo: Suponha que o objetivo seja retirar uma amostra de tamanho $n=60$ de uma população com $N=149$ elementos. Usando $a=2$, tem-se $n=74$ ou $n=75$. Usando $a=3$, tem-se $n=49$

ou $n = 50$. Tamanhos de amostra entre $n = 50$ e $n = 74$ não são possíveis.

O esquema amostral a seguir, contorna esse problema.

8.1 AS pelo método do intervalo fracionado

Esquema amostral

Esse método seleciona uma AS através do uso de um valor fracionado de a, resultando em um tamanho de amostra n.

Defina $a = \dfrac{N}{n}$.

Considere ξ, uma realização de uma variável aleatória com distribuição uniforme em $(0, a)$.

A amostra será, então, composta pelos seguintes elementos:

$$S_r = \{k:\ k-1<\xi+(j-1)a \leq k; j=1,2,...,n\}$$

Exemplo:

$N = 21$; $n = 5$; $a = 4,2$;
$\xi = 2.1$.

$j =$	1	2	3	4	5
$\xi+(j-1)a =$	2.1	6.3	10.5	14.7	18.9
$k \in s$:	3	7	11	15	19

<u>Resultado 8.2:</u> Usando o método do intervalo fracionado,

$$\pi_k = \frac{n}{N}.$$

Prova:

$$\begin{aligned}\pi_k &= Prob(k \in s) \\ &= Prob(k-1 < \xi + (j-1)a \leq k) \\ &= \frac{1}{a} = \frac{n}{N}\end{aligned}$$

$\blacksquare$

O cálculo de π_{kl} requer mais trabalho. Os dois esquemas amostrais apresentados são equivalentes quando $c = 0$ e apresentam pouca diferença para valores grandes de N.

7.2 Análise da eficiência de um plano AS

Por simplicidade, considere o caso em que $c = 0$. Nessa situação, $a = N/n$. Logo,

$$\hat{t}_\pi = a\, t_s = N\bar{y}_{s_r}$$

e

$$Var_{AS}(\hat{t}_\pi) = Nn \sum_{r=1}^{a} (\bar{y}_{s_r} - \bar{y}_U)^2 .$$

Por outro lado, note que

$$\sum_{k \in U} (y_k - \bar{y}_U)^2 = \sum_{r=1}^{a} \sum_{k \in s_r} (y_k - \bar{y}_{s_r})^2 + n \sum_{r=1}^{a} (\bar{y}_{s_r} - \bar{y}_U)^2$$

$$SQT = SQD + SQE$$

Fica claro, então, que

$$Var_{AS}(\hat{t}_\pi) = N \times SQE\,.$$

Defina agora a seguinte medida de homogeneidade:

$$\delta = 1 - \frac{N-1}{N-a}\frac{SQD}{SQT}\,.$$

Os valores extremos de δ são:

$$\delta_{\min} = -\frac{a-1}{N-a}, \text{ quando } SQE = 0$$

e $\delta_{\max} = 1$, quando $SQD = 0$.

É possível mostrar que, quando $c = 0$ (e a é um inteiro),

$$Var_{AS}(\hat{t}_\pi) = \frac{N^2 S_{yU}^2}{n}[(1-f) + (n-1)\delta]$$

$$= Var_{AAS}(\hat{t}_\pi)\,[1 + \frac{(n-1)}{(1-f)}\delta]$$

$$\therefore \; deff(AS, \hat{t}_\pi) = \frac{Var_{AS}(\hat{t}_\pi)}{Var_{AAS}(\hat{t}_\pi)} = 1 + \frac{n-1}{1-f}\delta$$

Logo, AS será mais eficiente que AAS, apenas quando $\delta < 0$.

Essa situação ocorrerá dependendo da ordenação dos elementos na população.

Exemplo:

Considere o caso em que $N=20$ e $n=4$ ($a=4$). Suponha ainda que $y_k = k$. Nesse caso,

$$Var_{AAS}(\hat{t}_\pi) = 20^2(1-\frac{4}{20})\frac{35}{4} = 2.800,$$

independentemente de ordenação.

Os quadros (a), (b) e (c) a seguir ilustram o efeito da ordenação da população na eficiência de um plano AS:

Quadro (a)

s_1	s_2	s_3	s_4	s_5
1	2	3	4	5
10	9	8	7	6
11	12	13	14	15
20	19	18	17	16
42	**42**	**42**	**42**	**42**

$$Var_{AS}(\hat{t}_\pi) \;=\; N \times SQE \;=\; 20 \times 0 \;=\; 0$$

$$Var_{AAS}(\hat{t}_\pi) = 2.800$$

$$\delta \;=\; -\frac{a-1}{N-a} \;=\; -\frac{5-1}{20-5} \;=\; -0.2667$$

Quadro (b)

s_1	s_2	s_3	s_4	s_5
1	5	9	13	17
2	6	10	14	18
3	7	11	15	19
4	8	12	16	20
10	**26**	**42**	**58**	**74**

$$Var_{AS}(\hat{t}_\pi) = N \times SQE = 20 \times 640 = 12.800$$

$$Var_{AAS}(\hat{t}_\pi) = 2.800$$

$$\delta = 1 - \frac{N-1}{N-a}\frac{SQD}{SQT} \approx 0.95.$$

Quadro (c)

s_1	s_2	s_3	s_4	s_5
13	19	2	14	11
6	1	17	3	5
12	20	15	4	16
9	18	8	7	10
40	**58**	**42**	**28**	**42**

$$Var_{AS}(\hat{t}_\pi) = N \times SQE = 20 \times 114 = 2.280$$

$$Var_{AAS}(\hat{t}_\pi) = 2.800$$

$$\delta = 1 - \frac{N-1}{N-a}\frac{SQD}{SQT} \approx -0.05$$

8.3 Estimação de variância sob um plano AS

Sob o esquema amostral apresentado, não há estimador centrado para $Var_{AS}(\hat{t}_\pi)$. Uma solução por vezes empregada é a utilização de $\hat{Var}_{AAS}(\hat{t}_\pi)$ para estimar $Var_{AS}(\hat{t}_\pi)$. Porém, como visto, a menos que $\delta = 0$, este estimador pode ser bastante tendencioso.

Cochran (1977) apresenta alguns estimadores com base em modelos de superpopulação para as observações.

É possível derivar estimadores centrados para $Var_{AS}(\hat{t}_\pi)$ usando esquemas amostrais alternativos.

9. Amostragem de Bernoulli - BE

Plano amostral

Um plano amostral BE consiste em uma série de experimentos independentes, um para cada elemento da população. O plano atribui probabilidade igual de seleção (π), e de não-seleção ($1-\pi$), a cada elemento listado no cadastro. Conseqüentemente, o tamanho da amostra, denotado por n_s, é uma variável aleatória. Sob um BE, tem-se:

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s},$$

$$\pi_k = \pi \quad \text{e} \quad \pi_{kl} = \pi^2.$$

Esquema amostral

Considere um valor para π ($0<\pi<1$).

Denote por $\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots, \varepsilon_N$, uma série de N realizações de uma Uniforme(0,1).

Se $\varepsilon_k<\pi$, então, o elemento k é selecionado para compor s. Do contrário, $k \notin s$.

Estimação

Considere a estimação do total populacional de uma população finita, sob um plano BE.

<u>Resultado 9.1:</u> Sob um plano BE, o estimador de H-T assume a forma

$$\hat{t}_\pi = \pi^{-1} \sum_{k \in s} y_k$$

e tem como variância

$$Var_{BE}(\hat{t}_\pi) = \left(\frac{1}{\pi} - 1 \right) \sum_{k \in U} y_k^2 .$$

Um estimador centrado de $Var_{BE}(\hat{t}_\pi)$ é dado por

$$\hat{Var}_{BE}(\hat{t}_\pi) = \frac{1}{\pi} \left(\frac{1}{\pi} - 1 \right) \sum_{k \in U} y_k^2 .$$

Prova:

Exercício.

Os planos AAS, AS e BE, vistos até o momento, possuem a característica comum de atribuir probabilidade igual de seleção para as amostras.

Em alguns livros, os planos que compartilham essa característica são chamados de “EPSEM”, abreviação de *Equal Probability SElection Methods*.

Na prática, porém, planos que permitem probabilidades desiguais de seleção são freqüentes. A seguir, serão vistos alguns desses planos.

10. Amostragem de Poisson - POI

Plano amostral

O plano amostral POI é uma generalização do plano BE, uma vez que permite a utilização de probabilidades diferentes de seleção (e de não-seleção) para cada elemento k listado no cadastro. Denotando por π_k a probabilidade de seleção do elemento k, tem-se:

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in (U-s)} (1 - \pi_k).$$

Sob um POI,

$$\pi_{kl} = \pi_k \pi_l.$$

Esquema amostral

Considere dados os valores π_k para $k = 1, 2, \ldots, N$.

Denote por $\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots, \varepsilon_N$, uma série de N realizações de uma Uniforme(0,1).

Se $\varepsilon_k < \pi_k$, então, o elemento k é selecionado para compor s. Do contrário, $k \notin s$.

Da mesma forma que no plano BE, sob um POI, o tamanho de amostra, denotado por n_s, é uma variável aleatória.

Estimação

Considere a estimação do total populacional de uma população finita, sob um plano POI.

Resultado 10.1: Sob um plano POI, o estimador de H-T assume a sua forma geral

$$\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}$$

e tem como variância

$$Var_{POI}(\hat{t}_\pi) = \sum_{k \in U} \left(\frac{1}{\pi_k} - 1 \right) y_k^2 .$$

Um estimador centrado de $Var_{POI}(\hat{t}_\pi)$ é dado por

$$\hat{Var}_{POI}(\hat{t}_\pi) = \sum_{k \in U} \left(\frac{1}{\pi_k} - 1 \right) \left(\frac{y_k}{\pi_k} \right)^2 .$$

Prova:

Exercício.

Motivação para os próximos planos amostrais

Considere a seguinte micro-população, com $N = 20$ elementos:

k	1	2	3	4	5	6	7	8	9	10
y_k	***6***	***3***	***4***	***4***	***5***	***3***	***6***	***2***	***3***	***2***
k	11	12	13	14	15	16	17	18	19	20
y_k	***2***	***6***	***5***	***3***	***5***	***2***	***4***	***6***	***4***	***5***

Total de y_k: ***80***

Como retirar uma amostra de tamanho $n = 5$ dessa população, de forma a ter $Var_p(\hat{t}_\pi) = 0$?

11. Amostragem com probabilidade proporcional ao tamanho - PPT

Uma possível solução para o problema é a seguinte:

Särndal, Swensson e Wretman (1992, págs 91-97) apresentam esquemas amostrais para implementar amostras de tamanho $n=1$, $n=2$ e $n>2$, com probabilidade proporcional ao tamanho.