Carlos Tenreiro

Notas do curso de Amostragem e Sondagens

Coimbra, 2023

Maio de 2023
Versões anteriores: Mai. 2012, Mai. 2016

Nota prévia

As presentes notas foram escritas para servirem de texto de apoio às aulas de Amostragem e Sondagens, disciplina do primeiro ano, segundo semestre, do Mestrado em Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra. Apesar dos assuntos aqui tratados corresponderem, no essencial, ao lecionado, as matérias completas, que incluem os exercícios de aplicação que constam das folhas práticas da cadeira, foram expostas nas aulas.

Carlos Tenreiro
tenreiro@mat.uc.pt

Índice

Introdução

Inferência em "populações infinitas"

Nos problemas de estimação até agora estudados nas cadeiras de Estatística assumimos que a população, habitualmente idealizada, de onde recolhemos a amostra, é infinita e que as observações aí feitas são realizações de variáveis aleatórias $Y_1, \ldots, Y_n$ que admitimos serem cópias independentes de uma variável aleatória real Y. Quando pretendemos inferir sobre a média desconhecida $\mu \in \mathbb{R}$ da população, essa média é identificada com a esperança matemática da variável aleatória Y, isto é, $\mathrm{E}(Y) = \mu$, e, portanto, inferir sobre a média populacional não é mais do que inferir sobre a esperança matemática da variável Y. Tendo em conta a lei dos grandes números, somos levados a utilizar a média empírica ou média amostral definida por

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

como estimador de μ. Sabemos que $\overline{Y}$ é um estimador cêntrico de μ, isto é,

$$\mathrm{E}(\overline{Y}) = \mu,$$

com variância dada por

$$\mathrm{Var}(\overline{Y}) = \frac{\sigma^2}{n},$$

onde $\sigma^2 = \mathrm{Var}(Y)$ é interpretada como variância populacional. Sendo, em geral, esta variância populacional desconhecida, a inferência por intervalos de confiança sobre μ necessita que estimemos uma tal variância, o que podemos fazer usando a variância empírica corrigida ou variância amostral corrigida definida por

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

que sabemos ser um estimador cêntrico de σ^2. Não tendo informação adicional sobre a distribuição da variável Y que nos permita identificar a distribuição amostral de $\overline{Y}$, o teorema do limite central assegura-nos que, sendo a amostra grande, uma tal distribuição amostral é aproximadamente normal, de onde deduzimos a aproximação

$$\frac{\overline{Y}-\mu}{\sqrt{S^2/n}} \approx N(0,1),$$

ou ainda,

$$\frac{\overline{Y}-\mu}{\sqrt{\widehat{\mathrm{Var}}(\overline{Y})}} \approx N(0,1),$$

onde

$$\widehat{\mathrm{Var}}(\overline{Y}) = \frac{S^2}{n},$$

é um estimador cêntrico de $\mathrm{Var}(\overline{Y})$. O resultado anterior permite-nos finalmente concluir que um intervalo de confiança para μ, com nível aproximadamente igual a $100(1-\alpha)\%$, tem por extremidades

$$\overline{Y} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\overline{Y})},$$

onde $z_{1-\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da distribuição normal standard.

A descrição que acabámos de fazer põe em evidência os elementos essenciais à construção dum intervalo de confiança para a média desconhecida μ. Depois de identificado o estimador a utilizar para inferir sobre μ ($\overline{Y}$), precisamos que conhecer a sua variância ($\mathrm{Var}(\overline{Y})$), bem como um estimador desta ($\widehat{\mathrm{Var}}(\overline{Y})$). O conhecimentos da distribuição amostral (assintótica) de $\overline{Y}$ (normal) permite a construção do intervalo anterior, intervalo este que possui um nível de confiança aproximadamente igual ao valor $1-\alpha$ fixado à partida pelo utilizador.

Inferência em "populações finitas"

Os elementos anteriores são naturalmente comuns à construção de intervalos de confiança para uma média (ou outro parâmetro de interesse) quando a população é finita, problema que estudamos no curso de Amostragem e Sondagens. No entanto, como realçamos a seguir, o modelo estatístico subjacente ao problema de estimação em causa é distinto do anterior em que a população é assumida infinita.

Um ponto essencial e distintivo do modelo estatístico que estudamos neste curso, dito *design-based* (por oposição à abordagem *model-based* também usada na inferência estatística em populações finitas), é relativo à variável sobre a qual pretendemos obter informação, a que chamamos variável de interesse. Com efeito, admitiremos que a

variável que observamos na população não é aleatória. Quer isto dizer que os valores que essa variável toma nos vários elementos da população, população que denotamos por $\mathscr{U}$, não são considerados realizações de uma variável aleatória real, mas são considerados fixos, apesar de desconhecidos. Por isso, denotaremos essa variável por uma letra minúscula y, sendo o valor que essa variável toma na i-ésima unidade da população denotado por y_i. Sendo a população finita de tamanho $N \in \mathbb{N}$, assumiremos que cada unidade populacional está identificada por um dos primeiros N números naturais, isto é,

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

Pretendendo inferir sobre a média da população $\mathscr{U}$, sendo $\mathscr{U}$ finita esta média não é mais que

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{1}{N} \sum_{k \in \mathscr{U}} y_k.$$

A partir da extração de uma amostra s de tamanho n da população $\mathscr{U}$,

$$s = (s_1, \ldots, s_n)$$

com $s_i \in \mathscr{U}$, e da observação da variável y para os elementos da amostra

$$(y_{s_1}, \ldots, y_{s_n})$$

uma possível estimativa para a média da população pode ser dada pela média amostral

$$\hat{\bar{y}} = \frac{1}{n} \sum_{i=1}^{n} y_{s_i} = \frac{1}{n} \sum_{k \in s} y_k.$$

Como já referimos, os valores y_i que a variável de interesse toma nas várias unidades da população não são agora considerados realizações duma variável aleatória. A componente aleatória do modelo é introduzida pela escolha da amostra, que é feita por métodos aleatórios. Quer isto dizer que a amostra s será considerada uma realização de uma variável aleatória S, dita amostra aleatória, que toma valores no conjunto de todas as amostras de $\mathscr{U}$ (por simplicidade, assumimos para já que uma amostra de $\mathscr{U}$ é um subconjunto de $\mathscr{U}$), e cuja distribuição de probabilidade, fixada pelo utilizador, definirá a forma como a amostra é recolhida na população, isto é, definirá o plano de amostragem utilizado.

Notemos que o estimador $\hat{\bar{y}} = \hat{\bar{y}}(S)$ anterior, estimador natural da média populacional no modelo clássico de inferência estatística, apenas pode ser considerado natural no modelo de amostragem em populações finitas se as unidades da população tiverem igual probabilidade de serem selecionadas para a amostra, justificando a ponderação

n^{-1} atribuída a cada observação y_k, para $k \in S$. Assim, a escolha de diferentes planos de amostragem, dando origem a diferentes métodos de seleção da amostra — aí se incluindo, como veremos, os métodos clássicos de amostragem em populações finitas, como a amostragem aleatória simples, com ou sem reposição, a amostragem estratificada, a amostragem por grupos, etc. —, implicará a utilização de diferentes estimadores $\hat{\theta} = \hat{\theta}(S)$ de $\bar{y}$.

O facto da população $\mathscr{U}$ ser finita introduz outras especificidades no processo de inferência estatística que não estão presentes no modelo clássico de inferência estatística. Uma dessas especificidades diz respeito à possibilidade de possuirmos informação adicional sobre toda a população. Como veremos, essa informação pode ser usada para melhorar o processo de inferência estatística, seja pela definição de novos estimadores, seja pela introdução de novos planos de amostragem.

Independentemente do estimador $\hat{\theta} = \hat{\theta}(S)$ de $\bar{y}$ que em cada caso consideramos, a partir do momento que o podemos interpretar como uma variável aleatória, a construção de intervalos de confiança para $\bar{y}$ pode ser feita como descrevemos atrás. Para cada um dos planos de amostragem que estudamos nos vários capítulos deste curso precisamos assim de responder às seguintes questões:

— Será $\hat{\theta}$ um estimador cêntrico (ou quase cêntrico) de $\bar{y}$?

— Qual a sua variância $\mathrm{Var}(\hat{\theta})$?

— Existirá um estimador (cêntrico), $\widehat{\mathrm{Var}}(\hat{\theta})$, de $\mathrm{Var}(\hat{\theta})$?

— Será a distribuição amostral de $\hat{\theta}$ aproximadamente normal?

Sendo afirmativa a resposta às questões anteriores, um intervalo de confiança para o parâmetro de interesse $\bar{y}$ com um nível de confiança de aproximadamente $100(1-\alpha)\%$ terá por extremidades

$$\hat{\theta} \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\hat{\theta})},$$

onde $z_{1-\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da distribuição normal standard.

Bibliografia

Cochran, W.G. (1977). *Sampling techniques.* Wiley. (Capítulo 1)

Thompson, M.E. (2002). *Theory of sample surveys.* Wiley. (Capítulo 1)

1

Amostragem: conceitos básicos

População e parâmetros. Amostra e plano de amostragem. Plano de amostragem com e sem reposição. Plano de amostragem reduzido. Média, variância e erro quadrático médio dum estimador. Estimadores cêntricos. Comparação de sistemas de amostragem. Intervalos de confiança. O tamanho da amostra.

1.1 População e parâmetros

A teoria da amostragem, também conhecida com teoria das sondagens, tem como objetivo a inferência sobre caraterísticas de interesse de uma população finita, que denotaremos por $\mathscr{U}$, a partir da observação das mesmas numa (pequena) parte dessa população. Admitiremos que cada uma das N unidades da população é identificável através de um número de ordem, podendo-se assim identificar $\mathscr{U}$ com o conjunto dos N primeiros números naturais

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

A N chamamos tamanho da população. Representando por y a variável ou caraterística de interesse sobre a qual pretendemos obter informação, e por y_k o valor dessa variável para a k-ésima unidade da população, o vetor de $\mathbb{R}^N$

$$\mathbf{y} = (y_1, ..., y_N) = (y_k, k \in \mathscr{U})$$

contém o valor da caraterística y para todas as unidades da população. Quando se realiza um estudo por amostragem, o objetivo não é obter informação sobre cada unidade da população, mas sim estimar uma função de $\mathbf{y}$,

$$\theta = \theta(\mathbf{y}) = \theta(y_k, k \in \mathscr{U}),$$

a que chamamos parâmetro ou função de interesse. Este parâmetro é habitualmente a média da população

$$\theta = \bar{y} = \frac{1}{N} \sum_{k \in \mathscr{U}} y_k,$$

o total da população

$$\theta = t_y = \sum_{k \in \mathscr{U}} y_k,$$

ou um quociente de totais (ou médias) no caso de existir outra variável de interesse x

$$\theta = \frac{t_y}{t_x} = \frac{\bar{y}}{\bar{x}},$$

onde $t_y, t_x, \bar{y}, \bar{x}$ são os totais e médias da população relativamente às variáveis y e x. Reparemos que os problemas da estimação de um total e de uma média não são necessariamente equivalentes uma vez que o tamanho da população não é necessariamente conhecido, podendo nesse caso ser um parâmetro de interesse.

Outros parâmetros de interesse são a variância da população definida por

$$\sigma_y^2 = \frac{1}{N} \sum_{k \in \mathscr{U}} (y_k - \bar{y})^2$$

e a variância corrigida da população definida por

$$s_y^2 = \frac{N}{N-1} \sigma_y^2.$$

Os desvios-padrão associados são definidos por $\sigma_y = \sqrt{\sigma_y^2}$ e $s_y = \sqrt{s_y^2}$.

Nos casos em que y é uma caraterística dicotómica, isto é, $y_k = 1$ se a unidade k da população tem determinada caraterística, e $y_k = 0$ em caso contrário, $\bar{y}$ e t_y representam, respetivamente, a proporção p e o número de indivíduos na população com essa caraterística. Neste caso, a variância da população pode ser também expressa por $\sigma_y^2 = \bar{y}(1-\bar{y}) = p(1-p)$ (ver Exercício 1).

1.2 Amostra e plano de amostragem

Num estudo por amostragem, a inferência sobre um parâmetro $\theta = \theta(y_k, k \in \mathscr{U})$ é feita a partir da observação de um conjunto de unidades da população, a que chamamos genericamente amostra. A amostra é selecionada por um processo aleatório a que chamamos plano de amostragem. São estas duas noções que definimos a seguir.

Definição 1.2.1. *Chamamos amostra de tamanho $n \in \mathbb{N}_0$ de $\mathscr{U}$ a um elemento do conjunto*

$$\mathscr{U}^n = \mathscr{U} \times \cdots \times \mathscr{U},$$

onde $\mathscr{U}^0 = \{()\}$, e amostra de $\mathscr{U}$ a um elemento do conjunto

$$\tilde{\mathscr{S}} = \bigcup_{n=0}^{\infty} \mathscr{U}^n.$$

De acordo com a definição anterior, dada uma amostra s de $\mathscr{U}$ existe um inteiro $n \in \mathbb{N}_0$ tal que $s \in \mathscr{U}^n$. A um tal inteiro chamamos tamanho da amostra s e escrevemos $n(s) = n$. O conjunto das amostra de tamanho n, $\mathscr{U}^n = \{s \in \tilde{\mathscr{S}} : n(s) = n\}$, será também denotado por $\tilde{\mathscr{S}}_n$. Claramente $\#\tilde{\mathscr{S}}_n = N^n$. Assim, dada uma amostra s de $\mathscr{U}$ podemos escrever

$$s = (),$$

caso em que não é observada nenhuma unidade da população, ou

$$s = (s_1, ..., s_{n(s)}),$$

em que $s_i \in \mathscr{U}$ é dito o i-ésimo elemento da amostra s. Dizemos que uma unidade $k \in \mathscr{U}$ pertence a s (ou que s contém k), e escrevemos $k \in s$, se $s_j = k$, para algum $j \in \{1, ..., n(s)\}$. Duas amostra s e t dizem-se iguais, e escrevemos $s = t$, se forem iguais como elementos de $\tilde{\mathscr{S}}$, isto é, se $(s_1, \ldots, s_{n(s)}) = (t_1, \ldots, t_{n(t)})$.

Definição 1.2.2. *Um plano de amostragem p de suporte Q, definido numa população $\mathscr{U}$, é um par (p, Q) onde $Q \subset \tilde{\mathscr{S}}$, e p é uma probabilidade sobre Q, tal que:*

a) $p(s) > 0$, para todo o $s \in Q$;

b) Para todo o $k \in \mathscr{U}$ existe $s \in Q$ com $k \in s$.

Apesar disso não ser assumido na definição anterior, sem qualquer perda de generalidade podemos assumir que o plano de amostragem está definida em todo o $\tilde{\mathscr{S}}$ tomando $p(s) = 0$, para $s \notin Q$. O suporte Q contém assim todas as amostras que são suscetíveis de serem selecionadas, isto é, que têm probabilidade positiva de serem selecionadas. Além disso, toda a unidade da população tem uma probabilidade positiva de ser selecionada, uma vez que pertence a pelo menos uma das amostras do suporte. Esta condição é, como veremos, essencial para a existência de estimadores cêntricos (cf. Definição 1.3.3) dum parâmetro de interesse (ver Exercício 4).

Dum ponto de vista formal, um plano de amostragem pode ser sempre implementado usando o chamado método cumulativo. Para tal, começamos por considerar uma enumeração das amostras do suporte $Q = \{s_1, s_2, \ldots, s_M\}$ (estamos a supor que $\#Q = M < \infty$). Outra enumeração conduzirá, em geral, a outra implementação do mesmo plano de amostragem. Uma vez que $p(s_i) > 0$ para todo o $i = 1, \ldots, M$, e $\sum_{i=1}^{M} p(s_i) = 1$, o conjunto $\Pi = \{a_0, a_1, \ldots, a_M\}$, com $a_0 = 0$ e $a_i = \sum_{j=1}^{i} p(s_j)$, é uma partição do intervalo $[0, 1]$. Assim, para extrair uma amostra s segundo o plano p, geramos um número aleatório u no intervalo $]0, 1[$, e selecionamos a amostra $s = s_i$ caso $a_{i-1} < u \leq a_i$. Ao proceder desta forma, a probabilidade de selecionarmos a amostra s_i é exatamente $p(s_i)$. Com efeito, sendo $U \sim U]0, 1[$, a probabilidade de selecionarmos

a amostra s_i é dada por

$$P(a_{i-1} < U \leq a_i) = a_i - a_{i-1} = \sum_{j=1}^{i} p(s_j) - \sum_{j=1}^{i-1} p(s_j) = p(s_i).$$

Como podemos concluir da descrição anterior, quando o suporte Q do plano é grande, o método cumulativo pode não ser exequível.

Grande parte dos planos que consideraremos são baseados em amostras s formadas por unidades distintas, isto é, $s_i \neq s_j$, para todo o $i, j = 1, ..., n(s)$. Denotando por $\bar{\mathscr{S}} \subset \tilde{\mathscr{S}}$ o conjunto de tais amostras e por $\bar{\mathscr{S}}_n$ o subconjunto de $\bar{\mathscr{S}}$ das amostras de tamanho n, temos

$$\bar{\mathscr{S}} = \bigcup_{n=0}^{N} \bar{\mathscr{S}}_n,$$

onde

$$\bar{\mathscr{S}}_n = \{s \in \bar{\mathscr{S}} : n(s) = n\}.$$

Uma amostra em $\bar{\mathscr{S}}$ não é mais do que um subconjunto ordenado da população. Contém informação sobre a ordem mas não há repetições de unidades. Como $\#\bar{\mathscr{S}}_n = A_n^N = \frac{N!}{(N-n)!}$, onde A_n^N representa os arranjos possíveis de N elementos tomados n a n, o número total de amostras com unidades distintas é de $\sum_{j=0}^{N} \frac{N!}{(N-j)!}$.

Definição 1.2.3. *Um plano de amostragem (p, Q) diz-se* sem reposição *se $Q \subset \bar{\mathscr{S}}$. Caso contrário diz-se* com reposição.

Os planos seguintes são planos de tamanho fixo no sentido da definição seguinte.

Definição 1.2.4. *Um plano de amostragem (p, Q) diz-se de* tamanho fixo *n, com $n \in \mathbb{N}$, se $n(s) = n$ para todo o $s \in Q$. Caso contrário, o plano diz-se de tamanho* aleatório.

Exemplo 1.2.5. Um plano de amostragem simples com reposição de tamanho fixo $n \in \mathbb{N}$, que denotamos por SCR, é um plano de suporte $Q = \tilde{\mathscr{S}}_n$ em que cada amostra tem igual probabilidade de ser selecionada, isto é,

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

para todo o $s \in \tilde{\mathscr{S}}_n$. Uma vez que

$$p(s) = p(s_1, \ldots, s_n) = \frac{1}{N} \times \cdots \times \frac{1}{N},$$

um plano SCR corresponde à extração, ao acaso e com reposição, de n unidades da população com probabilidade $\frac{1}{N}$.

Exemplo 1.2.6. Um plano de amostragem simples sem reposição de tamanho fixo $n \leq N$, que denotamos por SSR, é um plano de suporte $Q = \bar{\mathscr{S}}_n$ em que cada amostra tem igual probabilidade de ser selecionada, isto é,

$$p(s) = \frac{1}{A_n^N},$$

para todo o $s \in \bar{\mathscr{S}}_n$. Uma vez que

$$p(s) = p(s_1, \ldots, s_n) = \frac{1}{N}\frac{1}{N-1} \times \cdots \times \frac{1}{N-n+1},$$

um plano SSR corresponde à extração, ao acaso e sem reposição, de n unidades da população, em que a i-ésima unidade da amostra é extraída com probabilidade $\frac{1}{N-i+1}$.

Os planos anteriores retêm informação sobre a ordem dos elementos da amostra e, no primeiro caso, sobre a multiplicidade com que cada unidade da população surge na amostra. Não pretendendo reter tal informação é possível definir a partir de um qualquer plano um novo plano a que chamamos versão reduzida do plano inicial. Representando por $\mathscr{S}$ o subconjunto de $\bar{\mathscr{S}}$ definido por

$$\mathscr{S} = \{s \in \tilde{\mathscr{S}} : s_1 < s_2 < \cdots < s_{n(s)}\},$$

vamos denotar por $r(s)$ a amostra de $\mathscr{S}$ que se obtém de s eliminando a informação relativamente à ordem e à multiplicidade dos seus elementos. Assim, se $s = (3, 6, 2, 6)$ temos $r(s) = (2, 3, 6)$, e se $s = (3, 2, 6)$ temos $r(s) = (2, 3, 6)$. A função r é uma aplicação de $\tilde{\mathscr{S}}$ em $\mathscr{S}$ chamada função de redução. O conjunto $\mathscr{S}$ pode ser identificado com as classes de equivalência determinadas em $\tilde{\mathscr{S}}$ pela relação de equivalência $\approx$ definida por $s \approx t$ sse $r(s) = r(t)$. Dito de outra forma, $\mathscr{S}$ pode ser identificado com a classe dos subconjuntos de $\mathscr{U}$.

Definição 1.2.7. *Um plano de amostragem (p, Q) diz-se reduzido se $Q \subset \mathscr{S}$. Caso contrário diz-se não reduzido.*

Um plano de amostragem reduzido é um plano sem reposição.

Como já referimos, sempre que um plano de amostragem é reduzido, as amostras podem ser identificadas com os subconjuntos de $\mathscr{U}$. Desta forma, é usual escrever $s = \{2, 3, 6\}$ em vez de $s = (2, 3, 6)$ querendo com isto afirmar que o plano em causa é reduzido.

Proposição 1.2.8. *Dado um plano de amostragem (p, Q), o plano (p^*, Q^*) definido por*

$$p^*(t) = \sum_{s \in Q:\, r(s)=t} p(s), \ \textit{para } t \in Q^*,$$

com

$$Q^* = \{r(s) : s \in Q\} = \{t \in \mathscr{S} : t = r(s) \text{ para algum } s \in Q\},$$

é um plano de amostragem reduzido dito versão reduzida do plano (p, Q).

Dem: Temos $Q^* \subset \tilde{\mathscr{S}}$, e sendo p uma probabilidade sobre Q também p^* é uma probabilidade sobre Q^*:

$$\sum_{t \in Q^*} p^*(t) = \sum_{t \in Q^*} \sum_{s \in Q:\, r(s)=t} p(s) = \sum_{s \in Q} p(s) = 1.$$

Além disso, dado $t \in Q^*$, existe $s \in Q$ tal que $r(s) = t$ e $p(s) > 0$. Por maioria de razão também $p^*(t) > 0$, verificando-se a condição a) da definição de plano de amostragem. Finalmente, dado $k \in \mathscr{U}$ sabemos existir $s \in Q$ tal que $k \in s$. Claramente $k \in r(s) \in Q^*$, existindo assim pelo menos uma amostra em Q^* à qual pertence a unidade k. ■

Atendendo a que dado um plano (p, Q) podemos sempre associar-lhe um plano reduzido (p^*, Q^*), uma questão natural é a de saber se nos podemos limitar a considerar planos reduzidos. Voltaremos a este assunto mais à frente.

Exemplo 1.2.9. A versão reduzida do plano SSR é dada por

$$p^*(s) = n! \frac{(N-n)!}{N!} = \frac{1}{C_n^N},$$

para $s \in Q^* = \mathscr{S}_n$, uma vez que o conjunto das amostras $t \in \bar{\mathscr{S}}_n$ que satisfazem $r(t) = s$ tem $n!$ elementos correspondentes às diferentes permutações que podemos obter a partir dos elementos de s.

Mais difícil é obter uma expressão para o plano reduzido associado ao plano SCR pois para $s \in \tilde{\mathscr{S}}_n$ o número de tais permutações depende das repetições existentes em s. Em particular, um tal plano não é de tamanho fixo. Outro exemplo dum plano reduzido de tamanho aleatório é o plano de Bernoulli.

Exemplo 1.2.10. Um plano de amostragem de Bernoulli de parâmetro $0 < \pi^* < 1$ é definido por

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

para todo o $s \in Q = \mathscr{S}$ (ver Exercício 7). Este plano pode ser implementado gerando para cada unidade k da população um número aleatório u_k (segundo uma distribuição uniforme sobre o intervalo $[0, 1]$), e incluindo a unidade k na amostra se e só se $u_k < \pi^*$.

1.3 Média e variância de um estimador

Fixado um plano de amostragem p de suporte Q, podemos considerar que uma amostra $s \in Q$ é a realização de uma variável aleatória S definida num espaço de probabilidade $(\Omega, \mathscr{A}, \mathrm{P})$ que toma o valor s com probabilidade $p(s)$. Assim, a distribuição de probabilidade P_S de S é definida por

$$\mathrm{P}_S(s) = \mathrm{P}(S = s) = p(s),\ s \in Q.$$

A variável S é dita amostra aleatória com plano de amostragem p.

Sendo Z uma função que a cada amostra $s \in Q$ associa o número real $Z(s)$, podemos definir a esperança matemática de Z da forma usual:

Definição 1.3.1. *Chamamos esperança matemática ou média de Z (relativamente ao plano de amostragem p de suporte Q) à quantidade*

$$\mathrm{E}(Z) = \sum_{s \in Q} Z(s)p(s),$$

sempre que

$$\sum_{s \in Q} |Z(s)|p(s) < \infty.$$

Sempre que Q é finito, caso dum plano de amostragem sem reposição, esta última condição é trivialmente verificada.

Observada uma amostra $s = (s_1, ..., s_{n(s)})$ toda a inferência sobre o parâmetro de interesse real θ deve ser baseada nos valores $y_{s_1}, ..., y_{s_{n(s)}}$, ou seja, nos valores da caraterística de interesse para as unidades que constituem a amostra recolhida.

Definição 1.3.2. *Chamamos estimador a toda a função real*

$$\hat{\theta}(s; \mathbf{y}) : \tilde{\mathscr{S}} \times \mathbb{R}^N \to \mathbb{R},$$

que, para cada amostra $s \in \tilde{\mathscr{S}}$, depende de $\mathbf{y} = (y_k, k \in \mathscr{U})$ apenas nos valores da caraterística y para as unidades de s. Por simplicidade, escreveremos habitualmente $\hat{\theta}(\mathbf{y})$ ou apenas $\hat{\theta}$.

Sendo θ o parâmetro de interesse e $\hat{\theta}$ um seu estimador, diferentes amostras s dão lugar a diferentes estimativas para θ. A média e a variância de $\hat{\theta}$, relativamente ao plano de amostragem p de suporte Q, são dadas por

$$\mathrm{E}(\hat{\theta}) = \sum_{s \in Q} \hat{\theta}(s; \mathbf{y})p(s),$$

e

$$\mathrm{Var}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2 = \sum_{s \in Q} \big(\hat{\theta}(s; \mathbf{y}) - \mathrm{E}(\hat{\theta})\big)^2 p(s) = \mathrm{E}(\hat{\theta})^2 - (\mathrm{E}(\hat{\theta}))^2.$$

Definição 1.3.3. *Chamamos viés de um estimador $\hat{\theta} = \hat{\theta}(\mathbf{y})$ do parâmetro θ a quantidade (que depende de $\mathbf{y}$)*

$$\text{Viés}(\hat{\theta}(\mathbf{y})) = \text{E}(\hat{\theta}(\mathbf{y})) - \theta(\mathbf{y}).$$

O estimador $\hat{\theta}$ diz-se cêntrico ou sem viés se

$$\text{E}(\hat{\theta}(\mathbf{y})) = \theta(\mathbf{y}), \quad \textit{para todo o } \mathbf{y} \in \mathbb{R}^N.$$

A propósito da existência de estimadores cêntricos de um parâmetro de interesse, ver o Exercício 4.

1.4 Comparação de sistemas de amostragem

Na teoria de amostragem pretendemos encontrar o plano de amostragem e o estimador que nos permite obter melhores estimativas do parâmetro de interesse θ. A este par formado pelo plano de amostragem e pelo estimador $(p, \hat{\theta})$ chamamos sistema de amostragem. A qualidade de um sistema de amostragem é normalmente avaliada através do erro quadrático médio. definido por

$$\text{EQM}(\hat{\theta}) = \text{E}(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + (\text{Viés}(\hat{\theta}))^2.$$

No caso dos estimadores cêntricos, o erro quadrático médio reduz-se à variância.

Suponhamos agora que em $\mathscr{U}$ estão definidos dois planos de amostragem p_1 e p_2 e dois estimadores $\hat{\theta}_1(\mathbf{y})$ e $\hat{\theta}_2(\mathbf{y})$, cada um deles associado ao plano de amostragem indicado. Para comparar dois sistemas de amostragem recorremos à comparação dos EQM associados a cada um dos estimadores envolvidos.

Definição 1.4.1. *Dizemos que o sistema de amostragem $(p_1, \hat{\theta}_1(\mathbf{y}))$ é pelo menos tão eficiente como $(p_2, \hat{\theta}_2(\mathbf{y}))$. em $\mathscr{Y} \subseteq \mathbb{R}^N$, se*

$$\text{EQM}(\hat{\theta}_1(\mathbf{y})) \leq \text{EQM}(\hat{\theta}_2(\mathbf{y})), \quad \textit{para todo o } \mathbf{y} \in \mathscr{Y}.$$

Se, além disso, a desigualdade estrita for verificada para algum $\mathbf{y} \in \mathscr{Y}$, então $(p_1, \hat{\theta}_1(\mathbf{y}))$ diz-se mais eficiente que $(p_2, \hat{\theta}_1(\mathbf{y}))$ em $\mathscr{Y}$. Quando $\mathscr{Y} = \mathbb{R}^N$ dizemos apenas que $(p_1, \hat{\theta}_1(\mathbf{y}))$ é pelo menos tão eficiente ou mais eficiente que $(p_2, \hat{\theta}_2(\mathbf{y}))$. Quando $p_1 = p_2$ dizemos que $\hat{\theta}_1(\mathbf{y})$ é pelo menos tão eficiente ou mais eficiente que $\hat{\theta}_2(\mathbf{y})$ em $\mathscr{Y}$.

No caso particular dos estimadores considerados $\hat{\theta}_1(\mathbf{y})$ e $\hat{\theta}_2(\mathbf{y})$ serem cêntricos, o critério de comparação anterior reduz-se à comparação das respetivas variâncias.

Reparemos que a um plano de amostragem está, em geral, associado um custo de amostragem que não é tido em conta quando a comparação de planos é exclusivamente feita a partir dos erros quadráticos médios associados.

1.5 Intervalos de confiança

Para parâmetros definidos a partir de somas, como os que considerámos atrás, é natural esperar que os estimadores respetivos, definidos também a partir de somas dos valores $y_{s_1}, ..., y_{s_{n(s)}}$ da caraterística em estudo para as unidades da amostra $s_1, ..., s_{n(s)}$, sigam aproximadamente uma distribuições normal, quando o tamanho da amostra é grande, isto é,

$$\frac{\hat{\theta} - \mathrm{E}(\hat{\theta})}{\sqrt{\mathrm{Var}(\hat{\theta})}} = \frac{\hat{\theta}(S;\mathbf{y}) - \mathrm{E}(\hat{\theta}(S;\mathbf{y}))}{\sqrt{\mathrm{Var}(\hat{\theta}(S;\mathbf{y}))}} \simeq N(0,1).$$

Resultados deste género são válidos para alguns planos de amostragem como são os casos dos planos de amostragem simples com e sem reposição. No caso do plano SSR, teoremas do limite central, isto é, resultados que estabelecem a normalidade assintótica de $\sum_{i\in S} y_i$, podem ser encontrados em Madow (1948), Erdös e Rényi (1959) e Hájek (1960) (ver também Pathak, 1988, p. 100). A prática habitual, que adotaremos neste curso, é assumir que a aproximação anterior é válida para amostras grandes (sobre a validade da aproximação normal, ver Cochran, 1977, pp. 39–44). Assim, sendo $\hat{\theta}$ um estimador cêntrico de θ baseado em somas de valores da caraterística de interesse, um intervalo de confiança para θ com nível de confiança aproximadamente igual a $100(1-\alpha)\%$ tem por extremidades

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta})}$$

onde $z_{1-\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da distribuição normal standard. Dependendo a variância de $\hat{\theta}$ de caraterísticas populacionais desconhecidas vamos substituí-la por um estimador $\widehat{\mathrm{Var}}(\hat{\theta})$ dando origem ao intervalo de confiança de extremidades

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}.$$

Este intervalo de confiança para θ pode ser interpretado da forma usual: se tomarmos sucessivas amostras segundo o plano de amostragem em causa e construirmos os respetivos intervalos com nível de confiança $100(1-\alpha)\%$, aproximadamente $100(1-\alpha)\%$ deles conterão θ.

1.6 O tamanho da amostra

Num estudo por amostragem uma questão importante a considerar é a da determinação do tamanho da amostra que devemos tomar para obter uma certa precisão para a estimativa do parâmetro em estudo. Esta precisão pode ser medida através da semiamplitude do intervalo de confiança, habitualmente designada por margem de erro.

Supondo que pretendemos um intervalo de confiança para θ com margem de erro inferior a E, o tamanho n da amostra a recolher é dado pelo menor número inteiro que satisfaz a condição

$$z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta})} \;\leq\; E$$

ou ainda,

$$\mathrm{Var}(\hat{\theta}) \;\leq\; \frac{E^2}{z^2_{1-\alpha/2}}.$$

Como a variância $\mathrm{Var}(\hat{\theta})$ depende do tamanho da amostra n e de parâmetros desconhecidos, isto é, de funções de $\mathbf{y}$, é necessário usar informação adicional sobre a variável em estudo para que possamos usar a fórmula anterior na determinação do tamanho da amostra a recolher. Tal acontece, por exemplo, quando conhecemos estimativas para os parâmetros desconhecidos que intervêm na expressão de $\mathrm{Var}(\hat{\theta})$, obtidos num estudo realizado anteriormente. Em alternativa, tais estimativas podem ser obtidas num estudo piloto, isto é, num estudo preliminar realizado sobre uma amostra de dimensão reduzida. Podemos também utilizar um processo de recolha da amostra em duas fases. A amostra recolhida na primeira fase é completada a seguir depois de ser usada para estimar a variância. Outra estratégia consiste em substituir $\mathrm{Var}(\hat{\theta})$ por um seu majorante, caso este seja conhecido. Neste caso obtém-se um valor de n superior ao estritamente necessário para obter a precisão desejada.

Depois de determinar o tamanho da amostra temos que analisar se devemos ou não considerá-lo, tendo em conta o custo de amostragem que implica. Assim, podemos ser levados a tomar para tamanho da amostra um valor que conduz a uma precisão inferior à inicialmente fixada, mas que seja economicamente mais viável.

Para o tamanho de amostra escolhido pelo critério anterior esperamos que

$$\mathrm{P}\big(|\hat{\theta} - \theta| \leq E\big) \approx 1 - \alpha.$$

Por vezes, em vez duma precisão absoluta E, pretende-se atingir uma precisão relativa e, com $0 < e < 1$, isto é, queremos escolher o tamanho da amostra de modo a termos

$$\mathrm{P}\Big(\Big|\frac{\hat{\theta}}{\theta} - 1\Big| \leq e\Big) \approx 1 - \alpha.$$

Neste caso, devemos tomar para n o menor número inteiro para o qual

$$\mathrm{Var}(\hat{\theta}) \leq \frac{e^2\theta^2}{z^2_{1-\alpha/2}}.$$

1.7 Bibliografia

Cochran, W.G. (1977). *Sampling techniques.* Wiley.

Erdös, P., Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hung. Acad. Sci. Ser. A* 4, 49–61.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci. Ser. A* 5, 361–374.

Lohr, S.L. (1999). *Sampling: design and analysis.* Duxbury Press. (Capítulo 2).

Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Ann. Math. Statist.* 19, 535–545.

Pathak, P.K. (1988). Simple random sampling. In *Handbook of Statistics, Sampling (Vol 6)*, P.R. Krishnaiah, C.R. Rao (eds.), Elsevier, 97–109.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod. (Parte inicial do Capítulo 3).

2

Amostragem aleatória simples

Amostragem simples com e sem reposição. Estimação duma média. Comparação da eficiência dos planos SCR e SSR. Estimação dum total, duma proporção e duma razão. Estimação num domínio.

2.1 Amostragem simples com reposição

Vimos no Exemplo 1.2.5 que o plano de amostragem simples com reposição (SCR) de tamanho fixo $n \in \mathbb{N}$ é definido por

$$p(s) = p(s_1, \ldots, s_n) = \frac{1}{N^n},$$

para todo o $s \in Q = \tilde{\mathscr{S}}_n = \mathscr{U}^n$.

Denotamos por $S = (S_1, \ldots, S_n)$ a amostra aleatória com plano de amostragem p.

Sendo p a distribuição de probabilidade de S, da forma produto desta distribuição podemos imediatamente deduzir o resultado seguinte que justifica a afirmação feita no Exemplo 1.2.5 de que um plano SCR corresponde à extração, ao acaso e com reposição, de n unidades da população com probabilidade $\frac{1}{N}$.

Proposição 2.1.1. *Num plano de amostragem SCR de tamanho n as variáveis $S_1, \ldots, S_n$ são independentes e possuem uma distribuição uniforme sobre $\mathscr{U}$, isto é,*

$$\mathrm{P}_{S_i}(s_i) = \frac{1}{N},$$

para $s_i \in \mathscr{U}$.

Dem: Uma vez que para $s = (s_1, \ldots, s_n) \in \tilde{\mathscr{S}}_n = \mathscr{U}^n$ se tem

$$\mathrm{P}_{(S_1,\ldots,S_n)}(s_1, \ldots, s_n) = p(s_1, \ldots, s_n) = \frac{1}{N} \times \cdots \times \frac{1}{N},$$

onde $p_i(s_i) = 1/N$, para $s_i \in \mathscr{U}$, são distribuições de probabilidade em $\mathscr{U}$, concluímos que $S_1, \ldots, S_n$ são i.i.d. com $\mathrm{P}_{S_i}(s_i) = 1/N$, para $s_i \in \mathscr{U}$. ■

Teorema 2.1.2. *Num plano de amostragem SCR de tamanho n, a média empírica*

$$\hat{\bar{y}} = \frac{1}{n}\sum_{k\in S} y_k$$

é um estimador cêntrico de $\bar{y}$ com variância

$$\mathrm{Var}(\hat{\bar{y}}) = \frac{\sigma_y^2}{n}.$$

Dem: Tendo em conta que

$$\hat{\bar{y}} = \frac{1}{n}\sum_{k\in S} y_k = \frac{1}{n}\sum_{i=1}^{n} y_{S_i},$$

e que, pela Proposição 2.1.1, as variáveis $y_{S_i}, i = 1, \ldots, n$ são i.i.d. temos

$$\mathrm{E}(\hat{\bar{y}}) = \mathrm{E}\left(\frac{1}{n}\sum_{k\in S} y_k\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_{S_i}\right) = \mathrm{E}(y_{S_1})$$

e

$$\mathrm{Var}(\hat{\bar{y}}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_{S_i}\right) = \frac{1}{n}\mathrm{Var}(y_{S_1}).$$

Usando agora o facto, estabelecido na Proposição 2.1.1, de que a variável S_1 é uniformemente distribuída sobre $\mathscr{U}$, obtemos

$$\mathrm{E}(y_{S_1}) = \sum_{k\in\mathscr{U}} y_k P(S_1 = k) = \frac{1}{N}\sum_{k\in\mathscr{U}} y_k = \bar{y}$$

e

$$\begin{aligned}\mathrm{Var}(y_{S_1}) &= \mathrm{E}(y_{S_1} - \mathrm{E}(y_{S_1}))^2 = \mathrm{E}(y_{S_1} - \bar{y})^2 \\ &= \sum_{k\in\mathscr{U}} (y_k - \bar{y})^2 P(S_1 = k) \\ &= \frac{1}{N}\sum_{k\in\mathscr{U}} (y_k - \bar{y})^2 = \sigma_y^2,\end{aligned}$$

o que conclui a demonstração. ■

A expressão para a variância do estimador $\hat{\bar{y}}$ obtida no resultado anterior é semelhante à que se obtém no caso da população infinita com observações independentes. Tal não é de admirar uma vez que, sendo a amostragem feita com reposição, estamos, de certa forma, a fazer com que a população se assemelhe a uma população infinita. Por mais que continuemos a observar, não conseguimos obter informação sobre todas as unidades da população. Uma tal expressão mostra ainda que a sua precisão depende

exclusivamente do tamanho da amostra recolhida, tamanho este considerado como uma quantidade absoluta, e não do tamanho da amostra relativamente ao tamanho da população. Este facto é observado, numa fase muito inicial do desenvolvimento da teoria estatística da amostragem por Bowley (1913, p. 673).

Teorema 2.1.3. *Num plano de amostragem SCR de tamanho $n \geq 2$, a variância empírica*

$$\hat{s}_y^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \hat{\bar{y}})^2$$

é um estimador cêntrico da variância (não corrigida) da população σ_y^2.

Dem: Tendo em conta que

$$\sum_{k\in S}(y_k - \hat{\bar{y}})^2 = \sum_{k\in S}(y_k - \bar{y})^2 - n(\hat{\bar{y}} - \bar{y})^2 = \sum_{i=1}^{n}(y_{S_i} - \bar{y})^2 - n(\hat{\bar{y}} - \bar{y})^2,$$

temos

$$\mathrm{E}(\hat{s}_y^2) = \frac{1}{n-1}\left(n\mathrm{Var}(y_{S_1}) - n\mathrm{Var}(\hat{\bar{y}})\right) = \frac{1}{n-1}\left(n\sigma_y^2 - n\frac{\sigma_y^2}{n}\right) = \sigma_y^2. \qquad \blacksquare$$

Dos resultados anteriores concluímos que num plano de amostragem SCR de tamanho n a variância de $\hat{\bar{y}}$ pode ser estimada de forma cêntrica por

$$\widehat{\mathrm{Var}}(\hat{\bar{y}}) = \frac{\hat{s}_y^2}{n}.$$

Tendo em conta os resultados anteriores, num plano de amostragem SCR de tamanho n o intervalo de confiança para a média populacional $\bar{y}$ de nível aproximado $100(1-\alpha)\%$ terá por extremidades

$$\hat{\bar{y}} \pm z_{1-\alpha/2}\frac{\hat{s}_y}{\sqrt{n}}.$$

Pretendendo um intervalo de confiança com margem de erro inferior a um valor E fixo à partida, o tamanho n da amostra a recolher deve ser tal que

$$z_{1-\alpha/2}\frac{\sigma_y}{\sqrt{n}} \leq E$$

isto é,

$$n \geq \frac{z_{1-\alpha/2}^2\sigma_y^2}{E^2}.$$

Sendo a variância σ_y^2 desconhecida, na prática esta é substituída por uma estimativa obtida num estudo anterior ou a partir de uma amostra preliminar de dimensão pequena, ou por um majorante, caso este seja conhecido. Por exemplo, quando y é uma

caraterística dicotómica, sabemos que $\sigma_y^2 = \bar{y}(1-\bar{y}) \leq \frac{1}{4}$. Usando este majorante para σ_y^2, a fórmula anterior fica

$$n \geq \frac{z_{1-\alpha/2}^2}{4E^2}.$$

Por vezes, em vez da precisão (absoluta) E, pretendermos obter uma precisão relativa e, para algum $0 < e < 1$, isto é, queremos que $P(|\hat{\bar{y}}/\bar{y} - 1| \leq e) \approx 1-\alpha$, o tamanho n da amostra deve ser escolhido de forma que

$$n \geq \frac{z_{1-\alpha/2}^2 \sigma_y^2}{e^2 \bar{y}^2} = (1 - N^{-1}) \frac{z_{1-\alpha/2}^2 \mathrm{CV}_y^2}{e^2},$$

onde

$$\mathrm{CV}_y = \sqrt{s_y^2/\bar{y}^2}$$

é o coeficiente de variação de y. Esta medida de variabilidade relativa não depende da unidade de medida. No caso em que usamos dados de estudos anteriores para aproximar a variância, verifica-se que o coeficiente de variação é, por vezes, mais estável do que a própria variância.

2.2 Amostragem simples sem reposição

Vimos no Exemplo 1.2.6 que o plano de amostragem simples sem reposição (SSR) de tamanho fixo $n \in \mathbb{N}$, com $2 \leq n \leq N$, é definido por

$$p(s) = p(s_1, \ldots, s_n) = \frac{1}{A_n^N} = \frac{(N-n)!}{N!},$$

para todo o $s \in Q = \bar{\mathscr{S}}_n$ (versão não reduzida do plano).

Denotamos por $S = (S_1, \ldots, S_n)$ a amostra aleatória com plano de amostragem p.

Proposição 2.2.1. *Num plano de amostragem SSR de tamanho n as variáveis $S_1, \ldots, S_n$ possuem um distribuição uniforme sobre $\mathscr{U}$ com*

$$\mathrm{P}_{(S_i,S_j)}(s_i, s_j) = \frac{1}{N(N-1)},$$

para $s_i, s_j \in \mathscr{U}$ com $s_i \neq s_j$ (assim $S_1, \ldots, S_n$ não são independentes).

Dem: Para $s_i, s_j \in \mathscr{U}$ com $s_i \neq s_j$ e $i < j$ temos

$$\begin{aligned}
&\mathrm{P}(S_i = s_i, S_j = s_j) \\
&= \mathrm{P}(S_1 \in \mathscr{U}, \ldots, S_i = s_i, \ldots, S_j = s_j, \ldots, S_n \in \mathscr{U}) \\
&= \sum_{s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_{j-1},s_{j+1},\ldots,s_n \in \mathscr{U}} \mathrm{P}_S(s_1, \ldots, s_i, \ldots, s_j, \ldots, s_n)
\end{aligned}$$

$$= \sum_{\substack{s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_{j-1},s_{j+1},\ldots,s_n \in \mathscr{U} \\ (s_1,\ldots,s_i,\ldots,s_j,\ldots,s_n)\in\bar{\mathscr{F}}_n}} p(s_1,\ldots,s_i,\ldots,s_j,\ldots,s_n)$$

$$= A_{n-2}^{N-2}\frac{1}{A_n^N} = \frac{(N-2)!}{(N-n)!}\frac{(N-n)!}{N!} = \frac{1}{N(N-1)}.$$

As variáveis S_i possuem um distribuição uniforme sobre $\mathscr{U}$:

$$\mathrm{P}(S_i = s_i) = \sum_{s_j\in\mathscr{U}:s_i\neq s_j} \mathrm{P}(S_i = s_i, S_j = s_j) = (N-1)\frac{1}{N(N-1)} = \frac{1}{N}. \qquad \blacksquare$$

A variância da média empírica num plano de amostragem simples sem reposição de tamanho n, que apresentamos no resultado seguinte, foi primeiramente obtida por Isserlis (1915).

Teorema 2.2.2. *Num plano de amostragem SSR de tamanho n a média empírica*

$$\hat{\bar{y}} = \frac{1}{n}\sum_{k\in S} y_k$$

é um estimador cêntrico de $\bar{y}$ com variância

$$\mathrm{Var}(\hat{\bar{y}}) = \Big(1 - \frac{n}{N}\Big)\frac{s_y^2}{n}.$$

Dem: Tendo as variáveis $S_1,\ldots,S_n$ uma distribuição uniforme sobre $\mathscr{U}$, podemos usar os argumentos utilizados no plano SCR para concluir que $\hat{\bar{y}}$ é um estimador cêntrico de $\bar{y}$ (ver demonstração do Teorema 2.1.2). Relativamente à variância do estimador, uma vez que as variáveis y_{S_i} não são independentes, esta exprime-se não só em termos das variâncias das variáveis y_{S_i} mas também em termos das respetivas covariâncias:

$$\mathrm{Var}(\hat{\bar{y}}) = \mathrm{Cov}(\hat{\bar{y}},\hat{\bar{y}}) = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathrm{Cov}(y_{S_i},y_{S_j}) = \frac{1}{n}\Big(\mathrm{Var}(y_{S_1}) + (n-1)\mathrm{Cov}(y_{S_1},y_{S_2})\Big).$$

Da Proposição 2.2.1 e da demonstração do Teorema 2.1.2, temos

$$\mathrm{E}(y_{S_1}) = \bar{y} \text{ e } \mathrm{Var}(y_{S_1}) = \sigma_y^2 = \frac{N-1}{N}s_y^2.$$

Além disso, temos também

$$\begin{aligned}\mathrm{Cov}(y_{S_1},y_{S_2}) &= \mathrm{E}((y_{S_1}-\bar{y})(y_{S_2}-\bar{y})) \\ &= \sum_{k,l\in\mathscr{U}:k\neq l}(y_k-\bar{y})(y_l-\bar{y})P(S_1=k,S_2=l) \\ &= \frac{1}{N(N-1)}\sum_{k,l\in\mathscr{U}:k\neq l}(y_k-\bar{y})(y_l-\bar{y}),\end{aligned}$$

onde

$$0 = \Big(\sum_{k \in \mathscr{U}} (y_k - \bar{y}) \Big)^2 = \sum_{k,l \in \mathscr{U}: k \neq l} (y_k - \bar{y})(y_l - \bar{y}) + \sum_{k \in \mathscr{U}} (y_k - \bar{y})^2.$$

Assim

$$\mathrm{Cov}(y_{S_1}, y_{S_2}) = \frac{-1}{N(N-1)} \sum_{k \in \mathscr{U}} (y_k - \bar{y})^2 = -\frac{s_y^2}{N}.$$

Usando os resultados anteriores obtemos finalmente

$$\mathrm{Var}(\hat{\bar{y}}) = \frac{1}{n} \Big(\frac{N-1}{N} s_y^2 - (n-1) \frac{s_y^2}{N} \Big) = \Big(1 - \frac{n}{N} \Big) \frac{s_y^2}{n}. \qquad \blacksquare$$

Do resultado anterior concluímos que a variância do estimador não depende apenas do tamanho da amostra mas também da relação entre o tamanho da amostra e o tamanho da população através do fator $1 - n/N$ dito fator de correção para populações finitas. A fracção n/N é dita taxa de sondagem ou taxa de amostragem. Contrariamente ao plano SCR, no caso do plano SSR as amostras têm no máximo tamanho N, caso em que recolhemos informação sobre todas as unidades da população. Isso explica a variância nula que obtemos para o estimador da média quando tomamos $n = N$. No entanto, quando o tamanho da população é muito grande, levando a que a taxa de amostragem seja necessariamente pequena para amostras de tamanho razoável, verificamos que, tal como acontece no plano SCR, a precisão do estimador depende principalmente do tamanho da amostra recolhida e não da taxa de amostragem.

Seguindo as linhas da demonstração do Teorema 2.1.3, podemos concluir que o estimador $\hat{s}_y^2$ é, no caso do plano SSR, um estimador cêntrico de s_y^2. Reparemos que no caso do plano SCR, mostrámos que $\hat{s}_y^2$ era estimador cêntrico, não de s_y^2, mas de σ_y^2.

Teorema 2.2.3. *Num plano de amostragem SSR de tamanho $n \geq 2$ a variância empírica*

$$\hat{s}_y^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - \hat{\bar{y}})^2$$

é um estimador cêntrico da variância corrigida da população s_y^2.

Pelo resultado anterior, concluímos que num plano de amostragem SSR de tamanho n a variância de $\hat{\bar{y}}$ pode ser estimada de forma cêntrica por

$$\widehat{\mathrm{Var}}(\hat{\bar{y}}) = \Big(1 - \frac{n}{N} \Big) \frac{\hat{s}_y^2}{n}.$$

Tendo em conta os resultados anteriores (e assumindo a normalidade assintótica da média amostral $\hat{\bar{y}}$), num plano de amostragem SSR de tamanho n o intervalo de

confiança para a média populacional $\bar{y}$ de nível aproximado $100(1-\alpha)\%$ terá por extremidades

$$\hat{\bar{y}} \pm z_{1-\alpha/2}\sqrt{1-\frac{n}{N}}\frac{\hat{s}_y}{\sqrt{n}}.$$

Pretendendo um intervalo de confiança com margem de erro inferior a um valor E fixo à partida, o tamanho n da amostra a recolher deve ser tal que

$$z_{1-\alpha/2}\sqrt{1-\frac{n}{N}}\frac{s_y}{\sqrt{n}} \leq E$$

isto é,

$$n \geq \frac{z^2_{1-\alpha/2}s^2_y}{E^2 + z^2_{1-\alpha/2}s^2_y/N}.$$

Reparemos que quando N é grande relativamente à variância populacional, a fórmula anterior reduz-se à que obtivemos para o plano SCR.

No caso de especificarmos uma precisão relativa e, com $0 < e < 1$, obtemos a fórmula

$$n \geq \frac{z^2_{1-\alpha/2}s^2_y}{e^2\bar{y}^2 + z^2_{1-\alpha/2}s^2_y/N} = \frac{z^2_{1-\alpha/2}\mathrm{CV}^2_y}{e^2 + z^2_{1-\alpha/2}\mathrm{CV}^2_y/N}. \tag{2.2.4}$$

2.3 Comparação da eficiência dos planos SCR e SSR

Representando por $\hat{\bar{y}}_{scr}$ e $\hat{\bar{y}}_{ssr}$ os estimadores da média considerados atrás em cada um dos planos SCR e SSR, dos resultados anteriores concluímos que

$$\mathrm{EQM}(\hat{\bar{y}}_{scr}) = \frac{\sigma^2_y}{n} = \Big(1-\frac{1}{N}\Big)\frac{s^2_y}{n}$$

e

$$\mathrm{EQM}(\hat{\bar{y}}_{ssr}) = \Big(1-\frac{n}{N}\Big)\frac{s^2_y}{n}.$$

Assim, o plano de amostragem SSR é mais eficiente que o plano de amostragem SCR para todo o $n \geq 2$. Quando N é grande relativamente a n, as diferenças entre os dois planos não são significativas. No entanto, quando o tamanho da população não é grande e a amostra inclui uma parte importante da população, a vantagem do plano SSR relativamente ao plano SCR pode ser muito significativa.

Reparemos que os custos dos dois planos não são necessariamente os mesmos, sendo em geral menor o do plano SCR uma vez que as unidades repetidas são apenas observadas uma vez. Por isso, a comparação anterior, exclusivamente baseada nos erros quadráticos médios dos dois sistemas de amostragem, não reflete a questão dos custos de amostragem envolvidos. Para mais informação sobre esta questão ver Pathak (1988, p. 108).

2.4 Estimação de outros parâmetros de interesse

No que se segue consideramos apenas o plano SSR. Resultados análogos podem ser deduzidos para o plano SCR.

2.4.1 Estimação dum total

Resultados análogos aos que obtivemos para o estimador da média $\bar{y}$ podem ser facilmente obtidos para o estimador do total $t_y = N\bar{y}$ definido por

$$\hat{t}_y = \frac{N}{n}\sum_{k\in S} y_k = N\hat{\bar{y}}.$$

Atendendo aos Teoremas 2.2.2 e 2.2.3, $\hat{t}_y$ é um estimador cêntrico de t_y cuja variância é dada por

$$\mathrm{Var}(\hat{t}_y) = N^2\Big(1-\frac{n}{N}\Big)\frac{s_y^2}{n},$$

e um estimador cêntrico desta variância é dado por

$$\widehat{\mathrm{Var}}(\hat{t}_y) = N^2\Big(1-\frac{n}{N}\Big)\frac{\hat{s}_y^2}{n}.$$

O intervalo de confiança para o total populacional t_y de nível aproximado $100(1-\alpha)\%$ terá por extremidades

$$\hat{t}_y \pm z_{1-\alpha/2} N\sqrt{1-\frac{n}{N}}\frac{\hat{s}_y}{\sqrt{n}}.$$

Pretendendo um intervalo de confiança com margem de erro inferior a um valor E fixo à partida, o tamanho n da amostra a recolher deve ser tal que

$$n \geq \frac{N^2 z_{1-\alpha/2}^2 s_y^2}{E^2 + N z_{1-\alpha/2}^2 s_y^2}.$$

No caso de especificarmos uma precisão relativa e, com $0 < e < 1$, o tamanho da amostra deve satisfazer (2.2.4).

2.4.2 Estimação duma proporção

Suponhamos agora que y é uma caraterística dicotómica, isto é, $y_k = 1$ se a unidade k tem a propriedade em estudo e $y_k = 0$ se tal não acontece, e que o parâmetro de interesse p é a proporção de indivíduo com tal propriedade na população. Neste caso temos

$$p = \bar{y},$$

o que justifica considerar-se o estimador de p dado por

$$\hat{p} = \hat{\bar{y}}.$$

Sabemos que $\hat{p}$ é um estimador cêntrico de p (Teorema 2.2.2),

$$\mathrm{E}(\hat{p}) = \mathrm{E}(\hat{\bar{y}}) = \bar{y} = p,$$

com variância dada por

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}(\hat{\bar{y}}) = \Big(1 - \frac{n}{N}\Big)\frac{s_y^2}{n},$$

onde

$$s_y^2 = \frac{N}{N-1}p(1-p).$$

A variância anterior pode ser estimada de forma cêntrica por (Teorema 2.2.3)

$$\widehat{\mathrm{Var}}(\hat{p}) = \Big(1 - \frac{n}{N}\Big)\frac{\hat{s}_y^2}{n},$$

onde

$$\hat{s}_y^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}).$$

O intervalo de confiança para o total populacional p de nível aproximado $100(1-\alpha)\%$ terá por extremidades

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\Big(1 - \frac{n}{N}\Big)\frac{\hat{p}(1-\hat{p})}{n-1}}.$$

Pretendendo um intervalo de confiança com margem de erro inferior a um valor E fixo à partida, o tamanho n da amostra a recolher deve ser tal que

$$n \geq \frac{z_{1-\alpha/2}^2 p(1-p)}{E^2(1-1/N) + z_{1-\alpha/2}^2 p(1-p)/N}.$$

Sendo a expressão anterior uma função crescente de $p(1-p)$ e sabendo que $p(1-p) \leq 1/4$, não havendo informação preliminar sobre p podemos substituir na fórmula anterior $p(1-p)$ por $1/4$ obtendo

$$n \geq \frac{z_{1-\alpha/2}^2}{4E^2(1-1/N) + z_{1-\alpha/2}^2/N}.$$

No caso da população ser de grande dimensão, as expressões anteriores podem ser aproximadas pelas expressões bem nossas conhecidas

$$n \geq \frac{z_{1-\alpha/2}^2 p(1-p)}{E^2}$$

e

$$n \geq \frac{z_{1-\alpha/2}^2}{4E^2}.$$

2.4.3 Estimação duma razão

Suponhamos agora que pretendemos estimar a razão entre duas média

$$r = \frac{\bar{y}}{\bar{x}},$$

onde $\bar{x} \neq 0$ e as duas variáveis y_k e x_k devem ser medidas em cada unidade da amostra. Num plano de amostragem SSR será natural tomar como estimador o quociente dos estimadores cêntricos de tais quantidades

$$\hat{r} = \frac{\hat{\bar{y}}}{\hat{\bar{x}}},$$

onde

$$\hat{\bar{y}} = \frac{1}{n}\sum_{k\in S} y_k \quad \text{e} \quad \hat{\bar{x}} = \frac{1}{n}\sum_{k\in S} x_k,$$

que é dito estimador de tipo rácio. Estimadores deste tipo podem surgir porque estamos diretamente interessados em estimar uma razão, mas podem também surgir por outras razões, algumas das quais estudaremos neste curso. Uma dessas situações ocorre quando estamos interessados em estimar o total t_y mas não conhecemos a dimensão N da população, o que nos impede de usar o estimador $\hat{t}_y = N\hat{\bar{y}}$. No caso de conhecermos o total t_x duma variável auxiliar x sobre a população (o que não quer dizer que conheçamos x_i para toda a unidade da população), será natural estimar t_y através do estimador de tipo rácio $\tilde{t}_y = t_x\hat{\bar{y}}/\hat{\bar{x}}$ uma vez que $N = t_x/\bar{x}$. Este estimador será estudado no Capítulo 4.

Sendo $\hat{r}$ um quociente de quantidades aleatórias, levanta-se naturalmente o problema do cálculo da média e da variância de $\hat{r}$. O que faremos a seguir é apresentar expressões que nos permitam o cálculo aproximado de Viés($\hat{r}$) e Var($\hat{r}$). Tal será feito a partir de desenvolvimentos de Taylor da função $h(x, y) = y/x$. Com efeito, atendendo a que

$$\frac{\partial h}{\partial x}(x,y) = -\frac{y}{x^2},\ \frac{\partial h}{\partial y}(x,y) = \frac{1}{x},\ \frac{\partial^2 h}{\partial x^2}(x,y) = \frac{2y}{x^3},\ \frac{\partial^2 h}{\partial x \partial y}(x,y) = -\frac{1}{x^2},\ \frac{\partial^2 h}{\partial y^2}(x,y) = 0,$$

então usando a fórmula de Taylor e desprezando os termos de ordem superior à segunda temos

$$\begin{aligned}\hat{r} - r &= h(\hat{\bar{x}}, \hat{\bar{y}}) - h(\bar{x}, \bar{y}) \\ &\approx -\frac{\bar{y}}{\bar{x}^2}(\hat{\bar{x}} - \bar{x}) + \frac{1}{\bar{x}}(\hat{\bar{y}} - \bar{y}) + \frac{\bar{y}}{\bar{x}^3}(\hat{\bar{x}} - \bar{x})^2 - \frac{1}{\bar{x}^2}(\hat{\bar{x}} - \bar{x})(\hat{\bar{y}} - \bar{y}).\end{aligned}$$

A partir do desenvolvimento anterior, obtemos a aproximação para o viés de $\hat{r}$ dada por

$$\begin{aligned}\text{Viés}(\hat{r}) &= \mathrm{E}(\hat{r} - r) \\ &\approx \mathrm{E}\Big(\frac{\bar{y}}{\bar{x}^3}(\hat{\bar{x}} - \bar{x})^2 - \frac{1}{\bar{x}^2}(\hat{\bar{x}} - \bar{x})(\hat{\bar{y}} - \bar{y})\Big) \\ &= \frac{\bar{y}}{\bar{x}^3}\mathrm{Var}(\hat{\bar{x}}) - \frac{1}{\bar{x}^2}\mathrm{Cov}(\hat{\bar{x}}, \hat{\bar{y}}) \\ &= \frac{r\mathrm{Var}(\hat{\bar{x}}) - \mathrm{Cov}(\hat{\bar{x}}, \hat{\bar{y}})}{\bar{x}^2}. \end{aligned} \tag{2.4.1}$$

Para obter uma aproximação para a variância do estimador consideramos a aproximação de $\hat{r} - r$ baseada nos termos de primeira ordem do desenvolvimento anterior

$$\hat{r} - r \approx -\frac{\bar{y}}{\bar{x}^2}(\hat{\bar{x}} - \bar{x}) + \frac{1}{\bar{x}}(\hat{\bar{y}} - \bar{y}),$$

o que nos permite obter

$$\begin{aligned}\mathrm{Var}(\hat{r}) &\approx \mathrm{Var}\Big(-\frac{\bar{y}}{\bar{x}^2}(\hat{\bar{x}} - \bar{x}) + \frac{1}{\bar{x}}(\hat{\bar{y}} - \bar{y})\Big) \\ &= \frac{\bar{y}^2}{\bar{x}^4}\mathrm{Var}(\hat{\bar{x}}) - \frac{2\bar{y}}{\bar{x}^3}\mathrm{Cov}(\hat{\bar{x}}, \hat{\bar{y}}) + \frac{1}{\bar{x}^2}\mathrm{Var}(\hat{\bar{y}}) \\ &= \frac{r^2\mathrm{Var}(\hat{\bar{x}}) - 2r\mathrm{Cov}(\hat{\bar{x}}, \hat{\bar{y}}) + \mathrm{Var}(\hat{\bar{y}})}{\bar{x}^2}. \end{aligned} \tag{2.4.2}$$

Proposição 2.4.3. *Num plano de amostragem SSR de tamanho $n \geq 2$ temos*

$$\mathrm{Cov}(\hat{\bar{x}}, \hat{\bar{y}}) = \Big(1 - \frac{n}{N}\Big)\frac{s_{xy}}{n}$$

onde

$$s_{xy} = \frac{1}{N-1}\sum_{k \in \mathscr{U}}(x_k - \bar{x})(y_k - \bar{y})$$

é a covariância (corrigida) entre as as variáveis x e y. Além disso, a covariância empírica entre x e y, definida por

$$\hat{s}_{xy} = \frac{1}{n-1}\sum_{k \in S}(x_k - \hat{\bar{x}})(y_k - \hat{\bar{y}}),$$

é um estimador cêntrico de s_{xy}.

Dem: A primeira parte é consequência da igualdade

$$\mathrm{Cov}(\hat{\bar{x}}, \hat{\bar{y}}) = \frac{1}{n}\,\mathrm{Cov}(x_{S_1}, y_{S_1}) + \frac{n-1}{n}\,\mathrm{Cov}(x_{S_1}, y_{S_2}),$$

onde

$$\mathrm{Cov}(x_{S_1}, y_{S_1}) = \frac{1}{N}\sum_{k \in \mathscr{U}}(x_k - \bar{x})(y_k - \bar{y})$$

e

$$\mathrm{Cov}(x_{S_1}, y_{S_2}) = \frac{1}{N(N-1)} \sum_{k,l \in \mathscr{U}: k \neq l} (x_k - \bar{x})(y_l - \bar{y}) = -\frac{1}{N(N-1)} \sum_{k \in \mathscr{U}} (x_k - \bar{x})(y_k - \bar{y}).$$

Para concluir que $\hat{s}_{xy}$ é um estimador cêntrico de s_{xy}, basta ter em conta que

$$\frac{n-1}{n}\, \hat{s}_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_{S_i} - \bar{x})(y_{S_i} - \bar{y}) - (\hat{\bar{x}} - \bar{x})(\hat{\bar{y}} - \bar{y}). \qquad \blacksquare$$

Finalmente, tendo em conta a proposição anterior e as expressões (2.4.1) e (2.4.2), obtemos:

$$\mathrm{Viés}(\hat{r}) \approx \Big(1 - \frac{n}{N}\Big) \frac{r s_x^2 - s_{xy}}{\bar{x}^2 n} \tag{2.4.4}$$

e

$$\mathrm{Var}(\hat{r}) \approx \Big(1 - \frac{n}{N}\Big) \frac{s_y^2 + r^2 s_x^2 - 2 r s_{xy}}{\bar{x}^2 n} = \Big(1 - \frac{n}{N}\Big) \frac{s_z^2}{\bar{x}^2 n}, \tag{2.4.5}$$

com z a variável auxiliar definida por

$$z_k = y_k - r x_k, k \in \mathscr{U}.$$

Estimadores das aproximações anteriores podem ser obtidos substituindo $\bar{x}, r, s_x^2, s_y^2$ e s_{xy} por $\hat{\bar{x}}, \hat{r}, \hat{s}_x^2, \hat{s}_y^2$ e $\hat{s}_{xy}$, respetivamente. O estimador de Viés$(\hat{r})$ é dado por

$$\widehat{\mathrm{Viés}}(\hat{r}) = \Big(1 - \frac{n}{N}\Big) \frac{\hat{r} \hat{s}_x^2 - \hat{s}_{xy}}{\hat{\bar{x}}^2 n}, \tag{2.4.6}$$

e o estimador de $\mathrm{Var}(\hat{r})$ é dado por

$$\widehat{\mathrm{Var}}(\hat{r}) = \Big(1 - \frac{n}{N}\Big) \frac{\hat{s}_y^2 + \hat{r}^2 \hat{s}_x^2 - 2 \hat{r} \hat{s}_{xy}}{\hat{\bar{x}}^2 n} = \Big(1 - \frac{n}{N}\Big) \frac{\hat{s}_{\hat{z}}^2}{\hat{\bar{x}}^2 n}, \tag{2.4.7}$$

onde

$$\hat{z}_k = y_k - \hat{r} x_k, k \in S.$$

Atendendo a que a aproximação obtida para o viés de $\hat{r}$ é desprezável relativamente à raiz quadrada da aproximação obtida para a sua variância, quando a amostra é grande o intervalo de confiança para r pode ser construído da forma habitual, desprezando-se o termo de viés. Neste caso um intervalo de confiança para r de nível aproximado $100(1-\alpha)\%$ é dado por

$$\hat{r} \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\hat{r})}.$$

Caso contrário, um intervalo de confiança para r de nível aproximado $100(1-\alpha)\%$ é dado por

$$\hat{r} - \widehat{\mathrm{Viés}}(\hat{r}) \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\hat{r})}.$$

2.4.4 Estimação num domínio

Recolhida uma amostra S de $\mathscr{U}$ segundo um plano SSR de tamanho n, pretendemos agora estimar totais ou médias em subpopulações de $\mathscr{U}$ também denominadas domínios. Designando tal domínio por $\mathscr{U}_0 \subset \mathscr{U}$, estamos interessados na estimação do total e média de $\mathscr{U}_0$ definidos por

$$t_{0y} = \sum_{k \in \mathscr{U}_0} y_k$$

e

$$\bar{y}_0 = \frac{1}{N_0} \sum_{k \in \mathscr{U}_0} y_k,$$

onde N_0 é o tamanho, habitualmente desconhecido, do domínio $\mathscr{U}_0$. A situação que agora consideramos é ainda mais interessante se pensarmos que não possuímos à partida qualquer listagem de $\mathscr{U}_0$, o que não nos permite extrair uma amostra SSR de $\mathscr{U}_0$ para estimar os parâmetros anteriores.

Atendendo a que a amostra S foi recolhida de $\mathscr{U}$ e não de $\mathscr{U}_0$, para apresentarmos estimadores das quantidades anteriores é necessário converter os parâmetros anteriores em parâmetros relativos à população $\mathscr{U}$. Para tal, basta definir a variável auxiliar

$$v_k = \begin{cases} y_k, & k \in \mathscr{U}_0 \\ 0, & k \notin \mathscr{U}_0, \end{cases}$$

o que permite reescrever t_{0y} e $\bar{y}_0$ como parâmetros relativos à população $\mathscr{U}$. Assim,

$$t_{0y} = \sum_{k \in \mathscr{U}} v_k = t_v \quad \text{e} \quad \bar{y}_0 = \frac{1}{N_0} t_v,$$

onde t_v e $\bar{v}$ são o total e a média da caraterística v na população $\mathscr{U}$.

Estimação do total

Tendo em conta o que vimos atrás, um estimador cêntrico de $t_{0y} = t_v$ é dado por

$$\hat{t}_{0y} = \frac{N}{n} \sum_{k \in S} v_k = \frac{N}{n} \sum_{k \in S_0} y_k,$$

onde S é um SRS de tamanho n de $\mathscr{U}$ e $S_0 = S \cap \mathscr{U}_0$. De §2.4.1 sabemos que a variância de $\hat{t}_{0y}$ é dada por

$$\mathrm{Var}(\hat{t}_{0y}) = N^2 \Big(1 - \frac{n}{N}\Big) \frac{s_v^2}{n},$$

e que esta pode ser estimada sem viés por

$$\widehat{\mathrm{Var}}(\hat{t}_{0y}) = N^2 \Big(1 - \frac{n}{N}\Big) \frac{\hat{s}_v^2}{n},$$

onde

$$\hat{s}_v^2 = \frac{1}{n-1}\sum_{k\in S}(v_k - \bar{v})^2.$$

Notemos que a variância $\hat{s}_v^2$ pode ser calculada unicamente a partir da amostra S_0. Com efeito,

$$\hat{s}_v^2 = \frac{n_0-1}{n-1}\hat{s}_{0y}^2 + \frac{n_0}{n-1}\Big(1-\frac{n_0}{n}\Big)\hat{\bar{y}}_0^2, \tag{2.4.8}$$

com $n_0 = n(S_0)$,

$$\hat{\bar{y}}_0 = \frac{1}{n_0}\sum_{k\in S_0} y_k$$

e

$$\hat{s}_{0y}^2 = \frac{1}{n_0-1}\sum_{k\in S_0}(y_k - \hat{\bar{y}}_0)^2.$$

Estimação da média

De forma análoga, caso conheçamos N_0, um estimador cêntrico de $\bar{y}_0 = \frac{1}{N_0}t_v$ é dado por

$$\tilde{\bar{y}}_0 = \frac{1}{N_0}\hat{t}_v = \frac{1}{N_0}\frac{N}{n}\sum_{k\in S} v_k = \frac{N}{N_0}\frac{1}{n}\sum_{k\in S_0} y_k.$$

A variância de $\tilde{\bar{y}}_0$ é dada por

$$\mathrm{Var}(\tilde{\bar{y}}_0) = \frac{N^2}{N_0^2}\Big(1-\frac{n}{N}\Big)\frac{s_v^2}{n},$$

e pode ser estimada sem viés por

$$\widehat{\mathrm{Var}}(\tilde{\bar{y}}_0) = \frac{N^2}{N_0^2}\Big(1-\frac{n}{N}\Big)\frac{\hat{s}_v^2}{n}.$$

Caso o tamanho N_0 do domínio $\mathscr{U}_0$ não seja conhecido, o parâmetro $\bar{y}_0$ é o quociente dos totais t_v e N_0. Com efeito, N_0 não é mais do que o total da variável auxiliar u definida por

$$u_k = \begin{cases} 1, & k \in \mathscr{U}_0 \\ 0, & k \notin \mathscr{U}_0 \end{cases}$$

isto é,

$$N_0 = \sum_{k\in\mathscr{U}} u_k.$$

Assim, um estimador cêntrico de N_0 é dado por

$$\hat{N}_0 = \frac{N}{n}\sum_{k\in S} u_k,$$

o que nos leva ao estimador de $\bar{y}_0$ dado por

$$\hat{\bar{y}}_0 = \sum_{k\in S} v_k \Big/ \sum_{k\in S} u_k = \frac{1}{n_0}\sum_{k\in S_0} y_k,$$

onde $n_0 = n(S_0)$, que não é mais do que a média amostral relativamente à amostra S_0.

Uma aproximação para o viés deste estimador de tipo rácio pode ser obtida a partir da expressão (2.4.4). Como a variável auxiliar u é dicotómica e $v_k = u_k = 0$, para todo o $k \in \mathscr{U}_0$, uma tal aproximação é sempre igual a zero, o que significa que o intervalo de confiança será construído como se se tratasse de um estimador cêntrico. O estimador (2.4.7) para a variância de $\hat{\bar{y}}_0$ pode ser escrito na forma

$$\widehat{\mathrm{Var}}(\hat{\bar{y}}_0) = \Big(1 - \frac{n}{N}\Big)\frac{n}{n-1}\frac{n_0-1}{n_0}\frac{\hat{s}^2_{0y}}{n_0}.$$

Quando o tamanho da amostra é grande, é de esperar que $n_0/n \approx N_0/N$ e nesse caso a expressão anterior para $\widehat{\mathrm{Var}}(\hat{\bar{y}}_0)$ é próxima da do estimador da variância da média amostral num plano SSR de tamanho n_0 (não aleatório) sobre $\mathscr{U}_0$:

$$\widehat{\mathrm{Var}}(\hat{\bar{y}}_0) \approx \left(1 - \frac{n_0}{N_0}\right)\frac{\hat{s}^2_{0y}}{n_0}.$$

Quando assumimos que N_0 era conhecido, vimos que o estimador $\tilde{\bar{y}}_0$ de $\bar{y}_0$ era cêntrico e que a sua variância podia ser estimada por

$$\widehat{\mathrm{Var}}(\tilde{\bar{y}}_0) = \frac{N^2}{N_0^2}\Big(1 - \frac{n}{N}\Big)\frac{\hat{s}^2_v}{n}.$$

Quando o tamanho da amostra é grande, de (2.4.8) temos também

$$\widehat{\mathrm{Var}}(\tilde{\bar{y}}_0) \geq \frac{N^2}{N_0^2}\Big(1 - \frac{n}{N}\Big)\frac{n_0-1}{n-1}\frac{\hat{s}^2_{0y}}{n} \approx \widehat{\mathrm{Var}}(\hat{\bar{y}}_0),$$

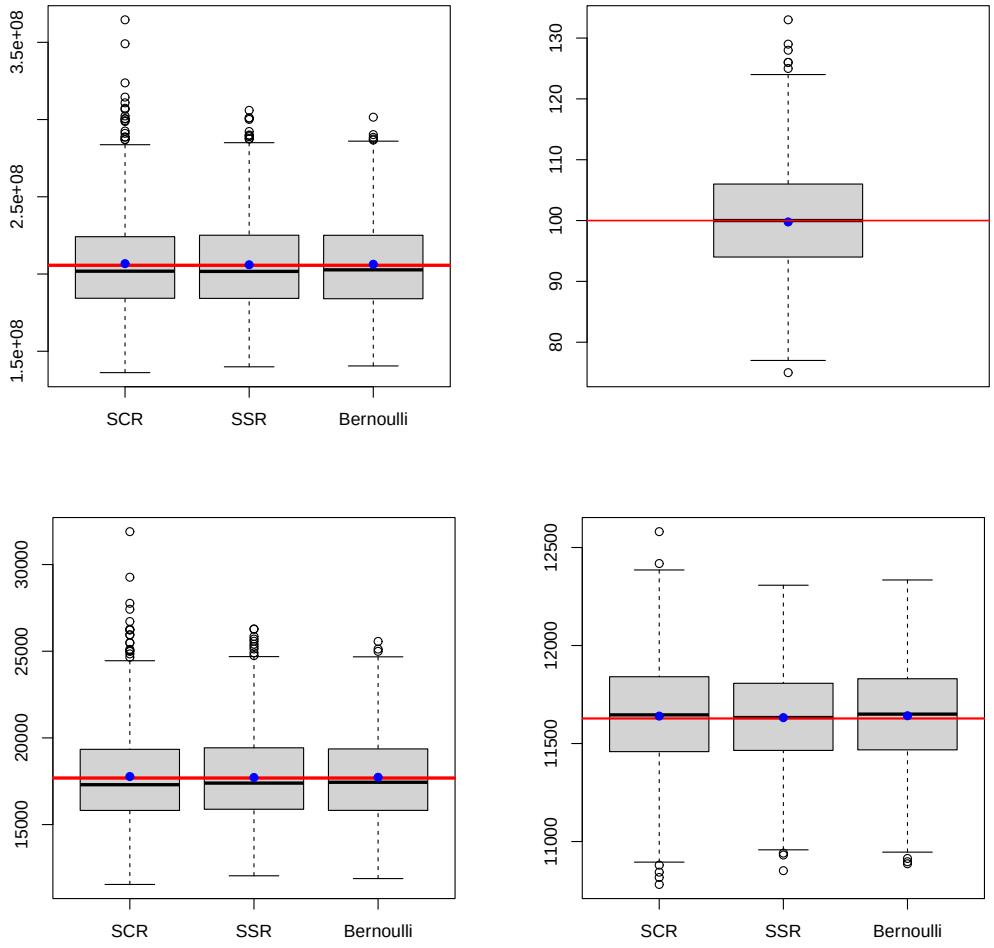
o que sugere que o estimador $\hat{\bar{y}}_0$ possa ser mais eficiente que $\tilde{\bar{y}}_0$ quando pretendemos estimar a média do domínio quando N_0 é conhecido. Nesse caso, a utilização em $\tilde{\bar{y}}_0$ da informação sobre o tamanho N_0 do domínio não traria vantagem na estimação de $\bar{y}_0$.

2.5 Alguns resultados de simulação

Para algumas das populações disponíveis na livraria *'sampling'* do R, usamos o método de Monte Carlo para descrever a distribuição de alguns dos estimadores considerados neste capítulo. Para tal, consideramos 1000 repetições do processo de amostragem para os planos SCR e SSR de tamanho n, e para o plano de Bernoulli de parâmetro

$\pi^* = n/N$. Em todos os casos tomámos $n = 100$. Nos gráficos seguintes, a linha a vermelho indica o verdadeiro valor do parâmetro de interesse, e os círculos a azul indicam a média das estimativas obtidas. No caso dos estimadores cêntricos, estes círculos devem estar sobre a linha a vermelho.

Exemplo 2.5.1. Variável de interesse: $y =$ belgianmunicipalities$TaxableIncome ($N = 589$). Parâmetro de interesse: t_y. O gráfico da direita descreve a distribuição dos tamanho das amostras obtidas pelo plano de Bernoulli.



(a) Exemplo 2.5.2

(b) Exemplo 2.5.3

Exemplo 2.5.2. Variável de interesse: $x =$ belgianmunicipalities$Tot04 ($N = 589$). Parâmetro de interesse: t_x.

Exemplo 2.5.3. Variáveis de interesse: y e x definidas nos exemplos anteriores. Parâmetro de interesse: t_y/t_x.

2.6 Bibliografia

Bowley, A.L. (1913). Working-class households in Reading. *J. Royal Statist. Soc.* 76, 672–701.

Cochran, W.G. (1977). *Sampling techniques.* Wiley. (Capítulos 3 e 4)

Isserlis, L. (1915). On the conditions under which the "probable errors" of frequency distributions have a real significance. *Proceedings of the Royal Society of London. Series A* 92, 23–41.

Lohr, S.L. (1999). *Sampling: design and analysis.* Duxbury Press. (Capítulos 2 e 3)

Thompson, M.E. (2002). *Theory of sample surveys.* Wiley. (Capítulos 2 a 5)

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod. (Capítulo 4)

Tillé, Y., Matei, A. (2021). 'sampling': survey sampling. R Package Version 2.9. http://CRAN.R-project.org/package=sampling

3

Amostragem estratificada

Estimação da média num plano de amostragem estratificada. Amostragem estratificada com afetação proporcional. Comparação da eficiência da amostragem estratificada com afetação proporcional com SSR. Afetação de Neyman. Afetação e custo. Ponderações amostrais.

3.1 Definição do plano de amostragem

Neste capítulo admitimos que a população $\mathscr{U}$ se encontra dividida em H subpopulações disjuntas, $\mathscr{U}_h$, para $h = 1, ..., H$, isto é,

$$\bigcup_{h=1}^{H} \mathscr{U}_h = \mathscr{U} \quad \text{e} \quad \mathscr{U}_i \cap \mathscr{U}_j = \emptyset, i \neq j.$$

A cada uma destas subpopulações, $\mathscr{U}_h$, chamamos estrato, e ao número, N_h, de unidades que a constituem chamamos tamanho do estrato $\mathscr{U}_h$. O tamanho da população não é mais do que a soma das dimensões dos estratos que a constituem:

$$N = \sum_{h=1}^{H} N_h.$$

Admitiremos que o tamanho de cada estrato é conhecido, podendo, assim, ser considerado como informação auxiliar sobre a população.

O plano de amostragem estratificada (simples sem reposição) consiste na extração de amostras s^h nos estratos $\mathscr{U}_h$ segundo planos de amostragem SSR de tamanho n_h, independentes entre si. A amostra final s é assim constituída por estas amostras parciais

$$s = (s^1, \ldots, s^H),$$

onde $s^h \in \bar{\mathscr{S}}_{n_h}$ para $h = 1, \ldots, H$.

O plano é assim definido por

$$p(s) = p(s^1, \ldots, s^H) = \frac{1}{A_{n_1}^{N_1}} \times ... \times \frac{1}{A_{n_H}^{N_H}},$$

(versão não reduzida) para qualquer amostra $s \in Q = \bar{\mathscr{S}}_{n_1}^1 \times \cdots \times \bar{\mathscr{S}}_{n_H}^H$, onde $\bar{\mathscr{S}}_{n_h}^h \subset \mathscr{U}_h$. Este plano de amostragem é de tamanho fixo

$$n = \sum_{h=1}^{H} n_h.$$

Representamos por $S = (S^1, ..., S^H)$ a amostra aleatória que tem p como plano de amostragem, onde as amostras aleatórias S_h possuem planos de amostragem SSR de tamanho n_h sobre $\mathscr{U}_h$. Atendendo à forma produto da distribuição de S as variáveis $S^1, \ldots, S^H$ são independentes.

Tal como acontecia no capítulo anterior o nosso objetivo continua a ser o da estimação de parâmetros (média, total, *etc*) associados a uma caraterística de interesse y definida na população $\mathscr{U}$.

3.2 Decomposições da média e da variância

Denotando, para $h = 1, \ldots, H$, por

$$\bar{y}_h = \frac{1}{N_h} \sum_{k \in \mathscr{U}_h} y_k,$$

$$\sigma_{yh}^2 = \frac{1}{N_h} \sum_{k \in \mathscr{U}_h} (y_k - \bar{y}_h)^2$$

e por

$$s_{yh}^2 = \frac{1}{N_h - 1} \sum_{k \in \mathscr{U}_h} (y_k - \bar{y}_h)^2$$

a média, a variância e a variância corrigida da caraterística de interesse y no estrato $\mathscr{U}_h$, a média $\bar{y}$ e a variância σ_y^2 podem ser escritas como

$$\bar{y} = \frac{1}{N} \sum_{k \in \mathscr{U}} y_k = \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in \mathscr{U}_h} y_k = \sum_{h=1}^{H} \frac{N_h}{N} \frac{1}{N_h} \sum_{k \in \mathscr{U}_h} y_k = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

e

$$\begin{aligned}\sigma_y^2 &= \frac{1}{N}\sum_{k\in\mathscr{U}}(y_k-\bar{y})^2 = \frac{1}{N}\sum_{h=1}^{H}\sum_{k\in\mathscr{U}_h}(y_k-\bar{y})^2\\ &= \frac{1}{N}\sum_{h=1}^{H}\sum_{k\in\mathscr{U}_h}((y_k-\bar{y}_h)+(\bar{y}_h-\bar{y}))^2\\ &= \frac{1}{N}\sum_{h=1}^{H}\sum_{k\in\mathscr{U}_h}(y_k-\bar{y}_h)^2 + \frac{1}{N}\sum_{h=1}^{H}\sum_{k\in\mathscr{U}_h}(\bar{y}_h-\bar{y})^2\\ &= \sum_{h=1}^{H}\frac{N_h}{N}\,\sigma_{yh}^2 + \sum_{h=1}^{H}\frac{N_h}{N}\,(\bar{y}_h-\bar{y})^2 = \sigma_{y,intra}^2+\sigma_{y,inter}^2, \qquad (3.2.1)\end{aligned}$$

obtendo-se a identidade usual da análise da variância que exprime a variância populacional como soma das variâncias intraestratos e interestratos.

3.3 Estimação da média

O resultado seguinte é consequência imediata dos Teoremas 2.2.2 e 2.2.3 e da independência entre as variáveis S^h, para $h = 1,\ldots,H$. A decomposição obtida para a média populacional $\bar{y}$ como média ponderada das médias dos estratos $\bar{y}_h$, explica tomarmos como estimador de $\bar{y}$ uma média ponderada, com as mesmas ponderações, das médias amostrais nos diversos estratos (ver Exercício 20).

Teorema 3.3.1. *Num plano de amostragem estratificada o estimador*

$$\hat{\bar{y}}_{est} = \sum_{h=1}^{H}\frac{N_h}{N}\,\hat{\bar{y}}_h$$

onde

$$\hat{\bar{y}}_h = \frac{1}{n_h}\sum_{k\in S^h} y_k$$

é um estimador cêntrico de $\bar{y}$ com variância

$$\mathrm{Var}(\hat{\bar{y}}_{est}) = \sum_{h=1}^{H}\frac{N_h^2}{N^2}\Big(1-\frac{n_h}{N_h}\Big)\frac{s_{yh}^2}{n_h}.$$

Além disso, se $n_h \geq 2$ para $h = 1,\ldots,H$,

$$\widehat{\mathrm{Var}}(\hat{\bar{y}}_{est}) = \sum_{h=1}^{H}\frac{N_h^2}{N^2}\Big(1-\frac{n_h}{N_h}\Big)\frac{\hat{s}_{yh}^2}{n_h}$$

com

$$\hat{s}^2_{yh} = \frac{1}{n_h - 1} \sum_{k \in S^h} (y_k - \hat{\bar{y}}_h)^2,$$

é um estimador cêntrico de $\mathrm{Var}(\hat{\bar{y}}_{est})$.

Assim, o intervalo de confiança para $\bar{y}$ de nível aproximado $100(1-\alpha)\%$ é dado por

$$\hat{\bar{y}}_{est} \pm z_{1-\alpha/2} \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{N^2} \Big(1 - \frac{n_h}{N_h}\Big) \frac{\hat{s}^2_{yh}}{n_h}}.$$

3.4 Tamanho da amostra: afetação proporcional

Atendendo à expressão da variância do estimador da média obtida no Teorema 3.3.1, o tamanho da amostra a recolher para obter um intervalo de confiança de nível aproximado $100(1-\alpha)\%$ com margem de erro não superior a um valor E fixado à partida deverá ser determinado a partir da desigualdade

$$z_{1-\alpha/2} \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{N^2} \Big(1 - \frac{n_h}{N_h}\Big) \frac{s^2_{yh}}{n_h}} \leq E.$$

Assumindo que em cada estrato recolhemos uma fração w_h de n fixada previamente, isto é, $n_h = n w_h$, obtemos

$$n \geq \frac{z^2_{1-\alpha/2} \displaystyle\sum_{h=1}^{H} \frac{N_h^2}{N^2 w_h} s^2_{yh}}{E^2 + z^2_{1-\alpha/2} \Big(\displaystyle\sum_{h=1}^{H} \frac{N_h}{N} s^2_{yh} \Big) / N}.$$

Usando o princípio da afetação de proporcional proposto por Bowley (1926), tomamos

$$w_h = N_h / N,$$

caso em que a fórmula anterior se reduz a

$$n \geq \frac{z^2_{1-\alpha/2} \displaystyle\sum_{h=1}^{H} \frac{N_h}{N} s^2_{yh}}{E^2 + z^2_{1-\alpha/2} \Big(\displaystyle\sum_{h=1}^{H} \frac{N_h}{N} s^2_{yh} \Big) / N}.$$

No caso particular em que y é uma caraterística dicotómica e não temos informação prévia sobre p_h, podemos ainda escrever

$$n \geq \frac{z^2_{1-\alpha/2} \sum_{h=1}^{H} \frac{N_h^2}{N(N_h-1)}}{4E^2 + z^2_{1-\alpha/2}\Big(\sum_{h=1}^{H} \frac{N_h^2}{N(N_h-1)}\Big)/N},$$

uma vez que $s^2_{yh} \leq N_h/(4(N_h-1))$. Se assumirmos que $N_h/(N_h-1) \approx 1$ (o que acontece quando os estratos são grandes), podemos usar a fórmula simplificada

$$n \geq \frac{z^2_{1-\alpha/2}}{4E^2 + z^2_{1-\alpha/2}/N}.$$

3.5 Eficiência relativamente ao plano SSR

Antes de apresentarmos as condições exatas sob as quais o plano de amostragem estratificada com afetação proporcional é mais eficiente que o plano SSR, começamos por mostrar que será de esperar que tal aconteça sempre que as dimensões dos estratos sejam grandes, isto é, $1/N_h \approx 0$. Com efeito, tendo em conta a decomposição (3.2.1) podemos escrever

$$s_y^2 = \frac{N}{N-1}\,\sigma_y^2 \approx \sum_{h=1}^{H} \frac{N_h}{N}\,\sigma^2_{yh} + \sigma^2_{y,inter} \approx \sum_{h=1}^{H} \frac{N_h}{N}\,s^2_{yh} + \sigma^2_{y,inter},$$

e assim do Teorema 2.2.2 obtemos a aproximação

$$\mathrm{Var}(\hat{\bar{y}}_{ssr}) = \Big(1-\frac{n}{N}\Big)\frac{s_y^2}{n} \approx \Big(1-\frac{n}{N}\Big)\sum_{h=1}^{H}\frac{N_h}{N}\,\frac{s^2_{yh}}{n} + \Big(1-\frac{n}{N}\Big)\,\frac{\sigma^2_{y,inter}}{n},$$

onde $\hat{\bar{y}}_{ssr}$ é o estimador da média por nós considerado no plano SSR (média empírica). Por outro lado, usando agora o Teorema 3.3.1 e denotando por $\hat{\bar{y}}_{pro}$ o estimador da média num plano de amostragem estratificada com afetação é proporcional, temos

$$\mathrm{Var}(\hat{\bar{y}}_{pro}) = \sum_{h=1}^{H}\frac{N_h^2}{N^2}\Big(1-\frac{n_h}{N_h}\Big)\frac{s^2_{yh}}{n_h} \approx \Big(1-\frac{n}{N}\Big)\sum_{h=1}^{H}\frac{N_h}{N}\frac{s^2_{yh}}{n},$$

uma vez que $n_h/N_h \approx n/N$. Assim

$$\mathrm{Var}(\hat{\bar{y}}_{ssr}) \approx \mathrm{Var}(\hat{\bar{y}}_{pro}) + \Big(1-\frac{n}{N}\Big)\,\frac{\sigma^2_{y,inter}}{n},$$

de onde concluímos que a vantagem da amostragem estratificada com afetação proporcional é maior quando a variância interestratos é maior (o que implica que a variância intraestratos seja menor, uma vez que a variância global σ_y^2 é constante).

A comparação exata entre os dois planos é estabelecida no resultado seguinte, onde assumimos que os efetivos dos estratos $n_h = nN_h/N$ são inteiros. Dele se deduz que a vantagem da estratificação com afetação proporcional será tanto maior quanto mais os estratos forem constituídos por unidades entre si homogéneas e os estrato forem mais heterogéneos entre si. O resultado vale também para os estimadores dum total ou duma proporção.

Teorema 3.5.1. *Se $n_h = nN_h/N$, para $h = 1, \ldots, H$, então o estimador $\hat{\bar{y}}_{pro}$ é mais eficiente que $\hat{\bar{y}}_{ssr}$ sse*

$$\sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) s_{yh}^2 \leq \sum_{h=1}^{H} N_h (\bar{y}_h - \bar{y})^2.$$

Dem: Da expressão obtida acima para a variância do estimador da média quando a afetação é proporcional deduzimos que $\mathrm{Var}(\hat{\bar{y}}_{pro}) \leq \mathrm{Var}(\hat{\bar{y}}_{ssr})$ sse

$$\sum_{h=1}^{H} \frac{N_h}{N} s_{yh}^2 \leq s_y^2. \tag{3.5.2}$$

Para concluir basta usar a decomposição (3.2.1) de onde sai a igualdade

$$s_y^2 = \frac{1}{N-1}\left(\sum_{h=1}^{H} (N_h - 1)\ s_{yh}^2 + \sum_{h=1}^{H} N_h\ (\bar{y}_h - \bar{y})^2 \right). \qquad \blacksquare$$

3.6 Afetação de Neyman

Uma abordagem distinta para a questão da afetação é proposta por Neyman (1934), que sugere escolher $n_1, \ldots, n_H$ de modo a minimizar a variância de $\hat{\bar{y}}$. Neste caso dizemos que temos um plano de amostragem estratificada com afetação óptima ou afetação de Neyman. Apesar do nome de Neyman ter ficado associado à afetação óptima, esta terá sido primeiramente considerada por Chuprov (1923). Como veremos a seguir, a minimização da variância pode ser conseguida através dum método engenhoso sugerido por Stuart (1954) que tira partido da desigualdade de Cauchy-Schwarz.

Proposição 3.6.1 (Desigualdade de Cauchy-Schwarz)**.** *Para quaisquer $a = (a_1, \ldots, a_k)$ e $b = (b_1, \ldots, b_k)$ em $\mathbb{R}^k$, com $k \in \mathbb{N}$, tem-se*

$$\left(\sum_{i=1}^{k} a_i b_i \right)^2 \leq \left(\sum_{i=1}^{k} a_i^2 \right) \left(\sum_{i=1}^{k} b_i^2 \right).$$

Além disso, vale a igualdade sse a é múltiplo escalar de b, isto é, sse existe $\lambda \in \mathbb{R}$ *tal que* $a = \lambda b$.

O resultado seguinte estabelece a forma como devemos afetar observações a cada estrato de forma a minimizar a variância $\mathrm{Var}(\hat{\bar{y}}_{est})$ para um tamanho de amostra n fixo à partida, ou como podemos minimizar o tamanho da amostra $\sum_{h=1}^{H} n_h$ para um valor fixo de $\mathrm{Var}(\hat{\bar{y}}_{est})$.

Teorema 3.6.2. *Num plano de amostragem estratificada de tamanho fixo* n*, os valores de* $n_1, ..., n_H$ *positivos, com* $\sum_{h=1}^{H} n_h = n$*, que minimizam* $\mathrm{Var}(\hat{\bar{y}}_{est})$ *são dados por*

$$n_h = nN_h s_{yh} \Big/ \sum_{l=1}^{H} N_l s_{yl},$$

para $h = 1, ..., H$*. Por outro lado, os valores positivos,* $n_1, ..., n_H$*, que minimizam* $\sum_{h=1}^{H} n_h$ *para um valor fixo* V *de* $\mathrm{Var}(\hat{\bar{y}}_{est})$ *são dados por*

$$n_h = N_h s_{yh} \left(\frac{1}{N^2} \sum_{l=1}^{H} N_l s_{yl} \right) \Bigg/ \left(\frac{1}{N^2} \sum_{l=1}^{H} N_l s_{yl}^2 + V \right),$$

para $h = 1, ..., H$.

Dem: Pelo Teorema 3.3.1, a variância de $\hat{\bar{y}}_{est}$ é dada por

$$\begin{aligned} \mathrm{Var}(\hat{\bar{y}}_{est}) &= \sum_{h=1}^{H} \frac{N_h^2}{N^2} \left(1 - \frac{n_h}{N_h}\right) \frac{s_{yh}^2}{n_h} \\ &= \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \frac{s_{yh}^2}{n_h} - \frac{1}{N^2} \sum_{h=1}^{H} N_h s_{yh}^2, \end{aligned} \tag{3.6.3}$$

onde apenas o primeiro termo depende de $n_1, \ldots, n_H$. Minimizar $\mathrm{Var}(\hat{\bar{y}}_{est})$ para um valor fixo n de $\sum_{h=1}^{H} n_h$, ou minimizar $\sum_{h=1}^{H} n_h$ para um valor fixo V de $\mathrm{Var}(\hat{\bar{y}}_{est})$ é assim equivalente a minimizar o produto

$$\left(\sum_{h=1}^{H} N_h^2 \frac{s_{yh}^2}{n_h} \right) \left(\sum_{h=1}^{H} n_h \right),$$

que, pela desigualdade de Cauchy-Schwarz, satisfaz

$$\left(\sum_{h=1}^{H} N_h^2 \frac{s_{yh}^2}{n_h} \right) \left(\sum_{h=1}^{H} n_h \right) \geq \left(\sum_{h=1}^{H} N_h s_{yh} \right)^2.$$

Atendendo a que o segundo membro não depende de n_h, o minimizante que procuramos não é mais do que o vector $(n_1, \ldots, n_h)$ que, para todo o $h = 1, \ldots, H$, satisfaz a igualdade

$$N_h \frac{s_{yh}}{\sqrt{n_h}} = \lambda \sqrt{n_h},$$

para algum $\lambda \in \mathbb{R}$, ou seja,

$$n_h = N_h s_{yh} \frac{1}{\lambda}. \tag{3.6.4}$$

No caso em que $n = \sum_{h=1}^{H} n_h$ é fixo, de (3.6.4) obtemos

$$\frac{1}{\lambda} = n \Big/ \sum_{h=1}^{H} N_h s_{yh}$$

e a primeira parte do resultado fica provada.

No caso em que $V = \mathrm{Var}(\hat{\bar{y}}_{est})$ é fixo de (3.6.3) e (3.6.4) tiramos que

$$V = \frac{\lambda}{N^2} \sum_{h=1}^{H} N_h s_{yh} - \frac{1}{N^2} \sum_{h=1}^{H} N_h s_{yh}^2$$

ou ainda

$$\frac{1}{\lambda} = \left(\frac{1}{N^2} \sum_{h=1}^{H} N_h s_{yh} \right) \Big/ \left(\frac{1}{N^2} \sum_{h=1}^{H} N_h s_{yh}^2 + V \right),$$

o que permite obter a segunda parte do resultado. ■

Uma consequência importante do resultado anterior, que historicamente implicou uma mudança na forma de pensar a amostragem, é o facto dele preconizar que em estratos mais homogéneos, isto é, com menor variabilidade, a afetação deve ser inferior à de estratos menos homogéneos. Isto implica que as probabilidades de inclusão de unidades na amostra, que são iguais para todos os elementos da população quando a afetação é proporcional, sejam diferentes para unidades pertencentes a estratos diferentes. Essa probabilidade é maior para uma unidade pertencente a um estrato com grande variabilidade (a este propósito, ver o Exercício 19).

Reparemos que os valores n_h resultantes da afetação de Neymann dados pelas expressões anteriores podem ser superiores a N_h. Tal pode acontecer, em particular, quando os estratos são pequenos mas possuem grande variabilidade. Se para um determinado estrato $\mathscr{U}_{h'}$ obtemos uma afetação superior ao tamanho do estrato será natural tomar $n_{h'} = N_{h'}$ e usar a afetação de Neymann relativamente aos restantes estratos para obter uma amostra de tamanho $n - n_{h'}$.

A segunda parte do resultado é particularmente útil quando queremos obter um intervalo de confiança de nível aproximado $100(1-\alpha)\%$ com margem de erro inferior a um valor E fixado à partida. Sendo essa margem de erro dada por

$$z_{1-\alpha/2} \sqrt{\mathrm{Var}(\hat{\bar{y}}_{est})},$$

para obter a margem de erro pretendida temos de escolher $n_1, \ldots, n_h$ de modo que

$$\mathrm{Var}(\hat{\bar{y}}_{est}) \leq \frac{E^2}{z_{1-\alpha/2}^2}.$$

Devemos assim tomar no resultado anterior

$$V = \frac{E^2}{z^2_{1-\alpha/2}}.$$

A utilização prática do resultado anterior não deixa de ser problemática uma vez que para usarmos a afetação de Neyman precisamos de conhecer as variabilidades s_{yh} de cada um dos estratos (ou aproximações destas). Se para todas as unidades populacionais conhecermos os valores de uma variável, $x_k > 0$, aproximadamente proporcional a y_k, isto é,

$$y_k \approx \lambda x_k,\ k \in \mathscr{U},$$

reparemos que a afetação resultante da primeira parte do Teorema 3.6.2 pode ser aproximada por

$$n_h \approx N_h s_{xh} \Big/ \sum_{l=1}^{H} N_l s_{xl},$$

para $h = 1, ..., H$.

3.7 Afetação e custo

Na discussão anterior o custo de amostragem não é tido em conta. Estamos sempre a admitir que o custo de amostragem não depende dos estratos. Se representarmos por c_h o custo de amostragem para cada unidade do estrato $\mathscr{U}_h$, o custo total de amostragem é dado por $\sum_{h=1}^{H} n_h c_h$. Atendendo a que no caso do custo de amostragem ser igual para todos os estratos, a função de custo anterior é proporcional ao tamanho da amostra $\sum_{h=1}^{H} n_h$, o resultado seguinte é uma generalização simples do Teorema 3.6.2.

Teorema 3.7.1. *Num plano de amostragem estratificada, os valores de $n_1, ..., n_H$ positivos, que minimizam $\mathrm{Var}(\hat{\bar{y}}_{est})$ para um valor fixo C do custo $\sum_{h=1}^{H} n_h c_h$, são dados por*

$$n_h = C(N_h s_{yh}/\sqrt{c_h}) \Big/ \sum_{l=1}^{H} N_l s_{yl} \sqrt{c_l},$$

para $h = 1, ..., H$. Por outro lado, os valores positivos, $n_1, ..., n_H$, que minimizam o custo $\sum_{h=1}^{H} n_h c_h$ para um valor fixo V de $\mathrm{Var}(\hat{\bar{y}}_{est})$ são dados por

$$n_h = (N_h s_{yh}/\sqrt{c_h}) \left(\frac{1}{N^2} \sum_{l=1}^{H} N_l s_{yl} \sqrt{c_l} \right) \Big/ \left(\frac{1}{N^2} \sum_{l=1}^{H} N_l s^2_{yl} + V \right),$$

para $h = 1, ..., H$.

Dem: Procedendo como atrás, concluímos que para todo o $h = 1, \ldots, H$ se tem

$$N_h \frac{s_{yh}}{\sqrt{n_h}} = \lambda \sqrt{n_h c_h},$$

para algum $\lambda \in \mathbb{R}$, ou seja,

$$n_h = (N_h s_{yh} / \sqrt{c_h}) \frac{1}{\lambda}. \tag{3.7.2}$$

No caso em que fixamos o custo $C = \sum_{h=1}^{N} n_h c_h$, de (3.7.2) e tiramos que

$$\frac{1}{\lambda} = C \Big/ \sum_{h=1}^{N} n_h s_{yh} \sqrt{c_h}$$

o que permite obter a primeira parte do resultado.

No caso em que a variância $V = \mathrm{Var}(\hat{\bar{y}}_{est})$ é fixada à partida, de (3.6.4) e (3.7.2) obtemos

$$V = \frac{\lambda}{N^2} \sum_{h=1}^{H} N_h s_{yh} \sqrt{c_h} - \frac{1}{N^2} \sum_{h=1}^{H} N_h s_{yh}^2,$$

ou ainda,

$$\frac{1}{\lambda} = \left(\frac{1}{N^2} \sum_{h=1}^{H} N_h s_{yh} \sqrt{c_h} \right) \Big/ \left(\frac{1}{N^2} \sum_{h=1}^{H} N_h s_{yh}^2 + V \right),$$

o que permite obter a segunda parte do resultado. ■

3.8 Ponderações amostrais

Usando um plano de amostragem estratificada, o total t_y pode ser estimado de forma cêntrica por

$$\hat{t}_{est} = N \hat{\bar{y}}_{est} = \sum_{h=1}^{H} \sum_{k \in S^h} \frac{N_h}{n_h} y_k.$$

A escrita anterior põe em evidência uma caraterística interessante deste estimador. Contrariamente aos estimadores do total em planos simples, em que cada unidade amostral recebe peso N/n, aqui cada unidade amostral pertencente ao estrato $\mathscr{U}_h$ recebe uma ponderação igual a N_h/n_h que não é mais do que o inverso da probabilidade de inclusão na amostra dessa unidade (ver Exercício 19). Isto é, para $k \in \mathscr{U}_h$,

$$\frac{N_h}{n_h} = \frac{1}{\pi_k}.$$

Além disso, a soma de todas as ponderações é igual ao tamanho da população:

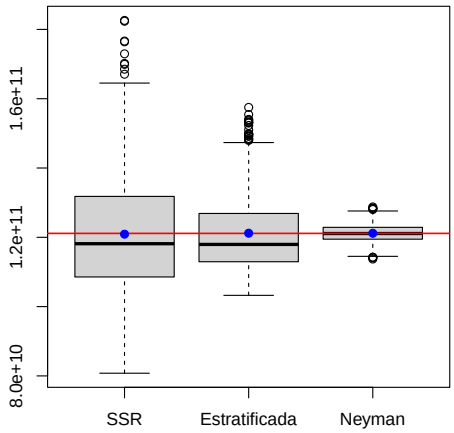
$$\sum_{h=1}^{H} \sum_{k \in S_h} \frac{N_h}{n_h} = \sum_{h=1}^{H} N_h = N.$$

Podemos interpretar este facto dizendo que cada unidade do estrato $\mathscr{U}_h$ representa N_h/n_h unidades da população e assim a amostra no seu total representa toda a população. Quando a afetação é proporcional, temos $N_h/n_h = N/n$ e, tal como no plano SSR, cada unidade da amostra representa N/n unidades da população. Quando usamos a afetação de Neyman, sabemos que a afetação n_h é pequena ou grande dependendo da variabilidade do estrato $\mathscr{U}_h$ ser pequena ou grande, respetivamente. Assim, num estrato com pequena variabilidade cada unidade da amostra representa mais unidades da população do que uma unidade pertencente a um estrato com grande variabilidade.

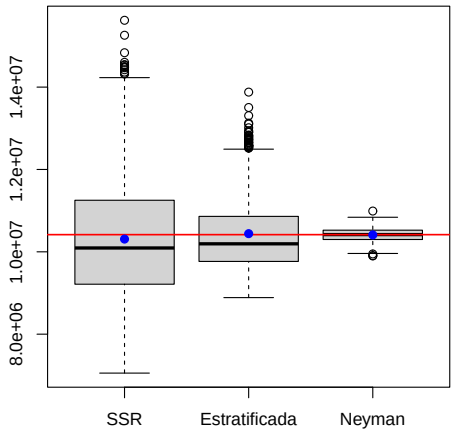
3.9 Alguns resultados de simulação

Nos exemplos seguintes usámos a afetação proporcional e a afetação de Neyman, tendo a variável $x =$ belgianmunicipalities$Tot03 sido usada para estratificar a população. Considerámos 5 estratos sendo os limites dos estratos definidos pelos valores 10000, 20000, 30000 e 50000 da variável x. Em todos os casos tomámos $n = 100$ e considerámos 1000 repetições do processo de amostragem.

Exemplo 3.9.1. Variável de interesse: $y =$ belgianmunicipalities$TaxableIncome. Parâmetro de interesse: t_y.



Exemplo 3.9.2. Variável de interesse: $y =$ belgianmunicipalities$Tot04. Parâmetro de interesse: t_y.



3.10 Bibliografia

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bull. Int. Statist. Inst.* (suppl.) 22, 6–62.

Chuprov, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* 2, 461–493, 646–683.

Cochran, W.G. (1977). *Sampling techniques.* Wiley. (Capítulo 5)

Hedayat, A.S., Sinha, B.K. (1991). *Design and inference in finite population sampling.* Wiley. (Capítulo 9)

Lohr, S.L. (1999). *Sampling: design and analysis.* Duxbury Press. (Capítulo 4)

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Royal Statist. Soc.* 97, 558–625.

Stuart, A. (1954). A simple presentation of optimum sampling results. *J. R. Stat. Soc. Ser. B* 16, 239–241.

Thompson, M.E. (2002). *Theory of sample surveys.* Wiley. (Capítulo 11)

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod. (Capítulo 7)

Tillé, Y., Matei, A. (2021). 'sampling': survey sampling. R Package Version 2.9. http://CRAN.R-project.org/package=sampling

4

Estimação com informação auxiliar

Utilização de informação auxiliar na fase de estimação. Estimadores da diferença, do quociente e da regressão. Pós-estratificação e estimador pós-estratificado.

4.1 Informação auxiliar na fase de estimação

A implementação dum plano de amostragem estratificada que seja preferível a um plano SSR necessita do conhecimento de informação auxiliar sobre a população, usada na fase da definição do plano de amostragem, que permita construir estratos homogéneos relativamente à variável de interesse cujas médias apresentem grande variabilidade entre si (ver Teorema 3.5.1). A utilização de informação auxiliar que pode ser usada na fase da definição do plano de amostragem será desenvolvida no próximo capítulo.

Para já, vamos ver como podemos usar determinado tipo de informação auxiliar na fase de estimação de forma a melhorar a qualidade da mesma. Esta ideia remonta, pelo menos, a Pierre-Simon Laplace (1749-1827) que, em 1802, estima o total da população francesa usando informação sobre o número de nascimentos ocorridos em França (Laplace, 1814, pp. 391–394, Lohr, 1999, pp. 59–62). Para uma amostra S de 30 municípios (*départements*) de todo o país, Laplace teve acesso à população total y_k do município $k \in S$:

$$\sum_{k \in S} y_k = 2037615.$$

Nos três anos anteriores a 23 de setembro de 1802, um total de 215599 nascimentos foram registados nos 30 municípios. Laplace estima em 215599/3 o número de nascimentos registados por ano nos 30 municípios. Assim, Laplace tem acesso à variável auxiliar x_k, relativa ao número de nascimentos registados no município k:

$$\sum_{k \in S} x_k = 215599/3.$$

Uma vez que

$$\sum_{k\in S} y_k \Big/ \sum_{k\in S} x_k = 28{,}352845, \tag{4.1.1}$$

Laplace estima que em cada ano ocorre um registo de nascimento por cada 28,352845 pessoas. Considerando que o número total de nascimentos registados por ano em França era de aproximadamente um milhão ($t_x = 1000000$), estima o total t_y da população francesa em

$$t_y \approx 28{,}35284486 \times 1000000 \approx 28352845.$$

Uma excelente estimativa como viriam a confirmar os censos franceses de 1806 que indicariam 29107425 para o total população francesa ([1]).

Reparemos que o quociente (4.1.1) é precisamente o quociente entre os estimadores dos totais das variáveis t_y e t_x

$$\frac{\hat{t}_y}{\hat{t}_x} = \frac{\dfrac{N}{n}\displaystyle\sum_{k\in S} y_k}{\dfrac{N}{n}\displaystyle\sum_{k\in S} x_k} = 28{,}352845,$$

e portanto o estimador de t_y considerado por Laplace é definido por

$$\hat{t}_q = \frac{\hat{t}_y}{\hat{t}_x} t_x,$$

a que chamaremos estimador do quociente.

Ao longo deste capítulo consideramos que o plano de amostragem é um plano SSR de tamanho n sobre a população $\mathscr{U}$ e o nosso objetivo é estimar o total da variável de interesse y:

$$t_y = \sum_{k\in\mathscr{U}} y_k.$$

Assumiremos que a informação auxiliar sobre a população está disponível através do conhecimento duma variável auxiliar x associada à variável de interesse y, que nos permitirá, em determinados casos, aumentar a precisão do usual estimador $\hat{t}_y$. Começaremos por estudar o caso em que a variável x é quantitativa e que conhecemos o seu total t_x. Como veremos, o conhecimento desta informação auxiliar permitir-nos-á definir estimadores alternativos de t_y: estimador da diferença, estimador do quociente e estimador da regressão.Finalmente, estudaremos o caso em que x é qualitativa permitindo associar cada indivíduo observado, e só depois de observado, a um estrato previamente definido na população, o que nos levará a definir o estimador pós-estratificado.

[1] A este propósito, ver Legoyt, A. (1987), Les premiers recensements de la population en France, jusqu'en 1856, *Journal de la société statistique de Paris* 128, 243–257.

4.2 Estimador da diferença

Como motivação para o chamado estimador da diferença, vamos assumir que a variável de interesse y_k, é, a menos de uma quantidade constante, aproximadamente igual a uma variável auxiliar x_k, isto é,

$$y_k \approx x_k + c, \quad k \in \mathscr{U},$$

para alguma constante $c \in \mathbb{R}$, temos

$$t_y \approx t_x + Nc.$$

Assim, para toda a amostra aleatória S temos

$$\hat{t}_y = \frac{N}{n}\sum_{k\in S} y_k \approx \frac{N}{n}\sum_{k\in S} x_k + Nc = \hat{t}_x + Nc,$$

de onde resulta a aproximação

$$t_y \approx t_x + \hat{t}_y - \hat{t}_x = \hat{t}_y - \hat{t}_x + t_x.$$

Assumindo que conhecemos o total t_x, a relação anterior motiva a estimação do total t_y através do estimador da diferença definido por

$$\hat{t}_d = \hat{t}_y - \hat{t}_x + t_x,$$

onde

$$\hat{t}_y = \frac{N}{n}\sum_{k\in S} y_k \quad \text{e} \quad \hat{t}_x = \frac{N}{n}\sum_{k\in S} x_k$$

são os estimadores usuais de t_y e t_x, respetivamente, num plano de amostragem SSR de tamanho n.

Atendendo a que

$$\hat{t}_d = \frac{N}{n}\sum_{k\in S}(y_k - x_k) + t_x,$$

será de esperar que $\hat{t}_d$ tenha vantagem sobre $\hat{t}_y$ se a variabilidade presente na diferença $z = y - x$ for inferior à de y. Expressões para a média e variância de $\hat{t}_d$ podem ser obtidas das expressões para a média e variância do estimador do total num plano SSR. Assim,

$$\mathrm{E}(\hat{t}_d) = \mathrm{E}(\hat{t}_y) - \mathrm{E}(\hat{t}_x) + t_x = t_y - t_x + t_x = t_y$$

e

$$\mathrm{Var}(\hat{t}_d) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2_{y-x}}{n} = N^2\left(1 - \frac{n}{N}\right)\frac{s^2_y}{n}\left(1 - \frac{2s_{yx} - s^2_x}{s^2_y}\right),$$

onde s_{yx} é a covariância entre as variáveis y e x definida na Proposição 2.4.3. Esta variância pode ser estimada por

$$\widehat{\mathrm{Var}}(\hat{t}_d) = N^2 \left(1 - \frac{n}{N}\right) \frac{\hat{s}_z^2}{n},$$

where

$$z_k = y_k - x_k,\ k \in\ S.$$

Das igualdades anteriores concluímos que a vantagem de $\hat{t}_d$ sobre $\hat{t}_y$ será máxima quando $s_{y-x}^2 = 0$, isto é, quando, para algum $b \in \mathbb{R}$, se tem $y_k = x_k + b$, para todo o $k \in \mathscr{U}$. A existir vantagem de $\hat{t}_d$ sobre $\hat{t}_y$, ela será tanto maior quanto maior for o quociente $(2s_{yx} - s_x^2)/s_y^2$. O estimador da diferença é mais eficiente que $\hat{t}_y$ sse

$$2s_{yx} - s_x^2 \geq 0,$$

ou ainda sse

$$a \geq \frac{1}{2}$$

onde

$$a = \frac{s_{yx}}{s_x^2} \tag{4.2.1}$$

é o declive da reta de regressão de y sobre x (baseada em toda a população). Em termos do coeficiente de correlação linear entre as variáveis x e y, definido por

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y},$$

podemos dizer que estimador da diferença é mais eficiente que $\hat{t}_y$ sse

$$\rho_{xy} \geq \frac{1}{2}\frac{s_x}{s_y}.$$

4.3 Estimador do quociente

Tomando como motivação para a definição deste estimador uma possível relação de proporcionalidade entre a variável de interesse y e a variável auxiliar x, da qual supomos conhecer o respetivo total t_x, suponhamos então que para algum $\lambda \in \mathbb{R}$ se tem

$$y_k \approx \lambda x_k, \quad k \in \mathscr{U}.$$

Neste caso

$$t_y \approx \lambda t_x$$

e para toda a amostra aleatória S temos

$$\hat{t}_y = \frac{N}{n}\sum_{k\in S} y_k \approx \lambda \frac{N}{n}\sum_{k\in S} x_k = \lambda \hat{t}_x.$$

Assim,

$$t_y \approx \frac{t_x}{\hat{t}_x}\,\hat{t}_y = \frac{\hat{t}_y}{\hat{t}_x}\,t_x,$$

o que motiva estimar t_y através do estimador do quociente definido por

$$\hat{t}_q = \hat{r}\,t_x,$$

com

$$\hat{r} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{N}{n}\sum_{k\in S} y_k \Big/ \frac{N}{n}\sum_{k\in S} x_k.$$

O estimador $\hat{t}_q$ é assim um estimador de tipo rácio que sabemos não ser cêntrico. Aproximações para o seu viés e variância podem ser obtidas a partir da igualdade

$$\hat{t}_q - t_y = t_x\,(\hat{r} - r)\,,$$

com

$$r = \frac{t_y}{t_x},$$

e das igualdades (2.4.4) e (2.4.5):

$$\text{Viés}(\hat{t}_q) \approx t_x\Big(1 - \frac{n}{N}\Big)\frac{rs_x^2 - s_{yx}}{\bar{x}^2 n} = N\Big(1 - \frac{n}{N}\Big)\,\frac{rs_x^2 - s_{yx}}{\bar{x}n}$$

e

$$\begin{aligned}\text{Var}(\hat{t}_q) &\approx t_x^2\Big(1 - \frac{n}{N}\Big)\frac{s_{y-rx}^2}{\bar{x}^2 n} = N^2\Big(1 - \frac{n}{N}\Big)\,\frac{s_{y-rx}^2}{n}\\ &= N^2\Big(1 - \frac{n}{N}\Big)\,\frac{s_y^2}{n}\left(1 - \frac{2rs_{yx} - r^2 s_x^2}{s_y^2}\right).\end{aligned}$$

Esta variância pode ser estimada por

$$\widehat{\text{Var}}(\hat{t}_q) = N^2\left(1 - \frac{n}{N}\right)\frac{\hat{s}_{\hat{z}}^2}{n},$$

onde

$$\hat{z}_k = y_k - \hat{r}\,x_k,\ k \in\ S.$$

Admitindo que n é grande de modo que as aproximações anteriores sejam de qualidade e possamos desprezar o viés do estimador, podemos dizer, de forma aproximada, que a vantagem do estimador do quociente sobre $\hat{t}_y$ será tanto maior quanto maior for o quociente $(2rs_{yx} - r^2 s_x^2)/s_y^2$. Essa vantagem é máxima quando a relação entre y e x pode ser aproximadamente descrita por uma linha reta que passa na origem. Este

estimador é assim especialmente aconselhado quando existe uma relação de proporcionalidade entre a variável de interesse y e a variável auxiliar x. Além disso, esperamos que $\hat{t}_q$ seja mais eficiente que o estimador $\hat{t}_y$ sse

$$r(2a - r) \geq 0$$

onde a é definido por (4.2.1). Assumindo que $r = t_y/t_x > 0$, podemos dizer que esperamos que o estimador do quociente seja mais eficiente que o estimador $\hat{t}_y$ quando

$$\rho_{xy} \geq \frac{1}{2}\frac{\mathrm{CV}_x}{\mathrm{CV}_y}.$$

Se $r = t_y/t_x < 0$ tal acontecerá quando

$$\rho_{xy} \leq -\frac{1}{2}\frac{\mathrm{CV}_x}{\mathrm{CV}_y}.$$

4.4 Estimador da regressão

Tomando como motivação para a definição do estimador da regressão uma possível relação linear entre a variável de interesse y e a variável auxiliar x, da qual supomos conhecer o respetivo total t_x, suponhamos então que

$$y_k \approx ax_k + b, \quad k \in \mathscr{U},$$

com $a, b \in \mathbb{R}$. Neste caso

$$t_y \approx a\, t_x + Nb, \tag{4.4.1}$$

e para toda a amostra aleatória S temos

$$\hat{t}_y = \frac{N}{n}\sum_{k\in S} y_k \approx a\frac{N}{n}\sum_{k\in S} x_k + Nb = a\hat{t}_x + Nb$$

e

$$\hat{s}_{yx} = \frac{1}{n-1}\sum_{k\in S}(y_k - \bar{y})(x_k - \bar{x}) \approx a\frac{1}{n-1}\sum_{k\in S}(x_k - \bar{x})^2 = a\hat{s}_x^2.$$

Assim,

$$t_y \approx \hat{a}\, t_x + (\hat{t}_y - \hat{a}\hat{t}_x) = \hat{t}_y - \hat{a}\,(\hat{t}_x - t_x),$$

com

$$\hat{a} = \frac{\hat{s}_{yx}}{\hat{s}_x^2},$$

o que motiva estimar t_y através do estimador da regressão definido por

$$\hat{t}_r = \hat{t}_y - \hat{a}\,(\hat{t}_x - t_x).$$

O estimador $\hat{t}_r$ é assim obtido da equação (4.4.1) substituindo a por $\hat{a} = \hat{s}_{yx}/\hat{s}_x^2$ e b por $\hat{b} = N^{-1}(\hat{t}_y - \hat{a}\,\hat{t}_x)$. É interessante notar que estes estimadores não são mais do que as soluções do problema de minimização

$$\min_{a,b\in\mathbb{R}} \sum_{k\in S} (y_k - ax_k - b)^2,$$

podendo, neste sentido, ser interpretados como estimadores dos mínimos quadrados do declive e a ordenada na origem da reta de regressão y sobre x, definidos como as soluções do problema de minimização

$$\min_{a,b\in\mathbb{R}} \sum_{k\in\mathscr{U}} (y_k - ax_k - b)^2.$$

Como sabemos, este problema tem solução única dada por

$$a = \frac{s_{yx}}{s_x^2} \quad \text{e} \quad b = N^{-1}(t_y - a\,t_x).$$

Uma vez que o estimador $\hat{a} = \hat{s}_{yx}/\hat{s}_x^2$ é um estimador de tipo rácio, o estimador da regressão é um estimador enviesado de t_y. Considerando o desenvolvimento

$$\hat{t}_r = \hat{t}_y - a(\hat{t}_x - t_x) - (\hat{a} - a)(\hat{t}_x - t_x),$$

concluímos que o viés de $\hat{t}_r$ é dado por

$$\text{Viés}(\hat{t}_r) = -\text{E}((\hat{a} - a)(\hat{t}_x - t_x)) = -\text{Cov}(\hat{a}, \hat{t}_x).$$

Da desigualdade de Cauchy-Schwarz temos

$$|\text{Viés}(\hat{t}_r)|^2 \leq \text{E}(\hat{a} - a)^2 \text{E}(\hat{t}_x - t_x)^2,$$

de onde se concluiu que o viés de $\hat{t}_r$ pode, em geral, ser ignorado para amostras grandes (sobre esta questão, ver Hedayat e Sinha, 1991, pp. 181–183).

Desprezando o termo $(\hat{a} - a)(\hat{t}_x - t_x)$ que surge no desenvolvimento anterior de $\hat{t}_r$, a variância de $\hat{t}_r$ poderá ser aproximada por

$$\text{Var}(\hat{t}_r) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{s_{y-ax}^2}{n} = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n} \left(1 - \rho_{xy}^2\right).$$

Esta variância pode ser estimada por

$$\widehat{\text{Var}}(\hat{t}_r) = N^2 \left(1 - \frac{n}{N}\right) \frac{\hat{s}_{\hat{z}}^2}{n},$$

onde

$$\hat{z}_k = y_k - \hat{a}\,x_k - \hat{b},\ k \in\ S.$$

Para amostras grandes, concluímos assim que o estimador da regressão é particularmente útil quando s^2_{y-ax} é próximo de zero, o que acontece quando existe uma relação aproximadamente linear entre a variável de interesse y e a variável auxiliar x. Além disso, atendendo a que

$$\mathrm{Var}(\hat{t}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2_y}{n},$$

$$\mathrm{Var}(\hat{t}_d) = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2_{y-x}}{n},$$

$$\mathrm{Var}(\hat{t}_q) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{s^2_{y-rx}}{n}$$

e

$$\mathrm{Var}(\hat{t}_r) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{s^2_{y-ax}}{n},$$

onde $s^2_{y-ax} \leq s^2_{y-\gamma x}$, para todo o $\gamma \in \mathbb{R}$, esperamos também que o estimador da regressão $\hat{t}_r$ possa ser mais eficiente que os estimadores $\hat{t}_y$, $\hat{t}_d$ e $\hat{t}_q$, quando o tamanho da amostra é grande (caso em que o viés de $\hat{t}_r$ pode ser desprezado). No entanto, é preciso não esquecer que os estimadores $\hat{t}_q$ e $\hat{t}_r$ são enviesados e que as fórmulas anteriores para as suas variâncias resultam da utilização de diversas aproximações que podem ter menor qualidade quando o tamanho da amostra não é grande.

4.5 Estimador pós-estratificado

Suponhamos que uma ou várias variáveis auxiliares permitem dividir a população em subpopulações $\mathscr{U}_h, h = 1, \ldots, H$, de tamanhos conhecidos N_h tais que

$$\bigcup_{h=1}^{H} \mathscr{U}_h = \mathscr{U} \quad \text{e} \quad \mathscr{U}_i \cap \mathscr{U}_j = \emptyset, i \neq j.$$

Apesar disso, admitimos que não temos informação *a priori* sobre as unidades da população para efetuar uma estratificação, sendo as unidades da amostra associadas à subpopulação $\mathscr{U}_h$ só depois da amostra ser recolhida. Nesta situação dizemos que efetuámos uma pós-estratificação e cada subpopulação $\mathscr{U}_h$ é dita pós-estrato.

Representando por S a amostra selecionada em $\mathscr{U}$ segundo um plano SSR de tamanho n (versão reduzida), vamos denotar por S_h a subamostra de S constituída por unidades de $\mathscr{U}_h$, para $h = 1, \ldots, H$. Assim, $S = \{S^1, \ldots, S^H\}$, com $S_h \in \mathscr{U}_h$, para $h = 1, \ldots, H$. Contrariamente à amostragem estratificada em que cada subamostra S^h é recolhida em $\mathscr{U}_h$ segundo um plano SSR de tamanho n_h, onde os tamanhos n_h das amostras de cada estrato estão fixos à partida, no caso presente os tamanhos n_h são variáveis aleatórias que podem tomar o valor 0 no caso de haver um pós-estrato

não representado na amostra. Cada uma destas variáveis n_h possui uma distribuição hipergeométrica de parâmetros N, N_h e n, isto é,

$$\mathrm{P}(n_h = r) = C_r^{N_h} C_{n-r}^{N-N_h} (C_n^N)^{-1}, \tag{4.5.1}$$

para $r = \max(0, n-(N-N_h)), \ldots, \min(n, N_h)$ (sobre a distribuição hipergeométrica, ver Cochran, 1977, pp. 55–57).

Nesta secção vamos estudar o estimador pós-estratificado do total t_y que é definido de forma muito semelhante ao estimador do total num plano de amostragem estratificada:

$$\hat{t}_p = \sum_{\substack{h=1 \\ n_h>0}}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} \hat{t}_{yh} \mathbb{1}(n_h > 0),$$

onde

$$\hat{t}_{yh} = \frac{N_h}{n_h} \sum_{k \in S^h} y_k,$$

sempre que $n_h > 0$ (quando $n_h = 0$, $\hat{t}_{yh}$ pode ser definido de forma arbitrária uma vez que $\hat{t}_{yh}\mathbb{1}(n_h > 0) = 0$).

Atendendo ao facto de $\hat{t}_{yh}$ ser um estimador de tipo rácio (n_h é uma variável aleatória), não é de esperar que este estimador seja cêntrico do total t_{yh} do estrato $\mathscr{U}_h$, onde

$$t_{yh} = \sum_{k \in \mathscr{U}_h} y_k.$$

O resultado seguinte permitir-nos-á usar os resultados estudados sobre o plano de amostragem SSR, no estudo do estimador pós-estratificado. Para tal é necessário que se trabalhe condicionalmente aos tamanhos das amostras em cada um dos estratos. Com efeito, a proposição seguinte estabelece que, condicionalmente a $n_1, \ldots, n_H$ as variáveis $S^h, h = 1, \ldots, H$, são independentes e as suas distribuições são planos SSR de tamanhos n_h sobre $\mathscr{U}_h$, para $h = 1, \ldots, H$.

Proposição 4.5.2. *Se S é uma amostra aleatória de tamanho n extraída de $\mathscr{U}$ segundo um plano de amostragem SSR reduzido então, condicionalmente a $n_1, \ldots, n_H$, as variáveis S^h, $h = 1, \ldots, H$, são independentes e as suas distribuições são planos SSR de tamanho n_h sobre $\mathscr{U}_h$, isto é, para $1 \leq r_h \leq N_h$ e $r_1 + \cdots + r_H = n$, temos*

$$\mathrm{P}(S^h = s_h | n_1 = r_1, \ldots, n_H = r_H) = \begin{cases} \dfrac{1}{C_{r_h}^{N_h}} & , n(s^h) = r_h \\ 0 & , n(s^h) \neq r_h, \end{cases}$$

para $h = 1, \ldots, H$.

Dem: Na primeira parte da demonstração determinamos a distribuição condicional das variáveis S^h, para $h = 1, \ldots, H$. A seguir estabelecemos a sua independência condicional.

Distribuição condicional de S^h, para $h = 1, \ldots, H$

Se $n(s^h) \neq r_h$ a probabilidade $\mathrm{P}(S^h = s^h | n_1 = r_1, \ldots, n_H = r_H)$ é claramente nula. Analisemos agora o caso em que $n(s^h) = r_h$. Comecemos por mostrar que

$$\mathrm{P}(S^h = s^h | n_1 = r_1, \ldots, n_H = r_H) = P(S^h = s^h | n_h = r_h). \tag{4.5.3}$$

Com efeito, temos

$$\begin{aligned}
&\mathrm{P}(S^h = s^h | n_1 = r_1, \ldots, n_H = r_H) \\
&= \frac{\mathrm{P}(S^h = s^h, n_1 = r_1, \ldots, n_H = r_H)}{P(n_1 = r_1, \ldots, n_H = r_H)} \\
&= \frac{\mathrm{P}(S^h = s^h, n_h = r_h)\mathrm{P}(n_l = r_l, l = 1, \ldots, H, l \neq h | S^h = s^h, n_h = r_h)}{\mathrm{P}(n_h = r_h)\mathrm{P}(n_l = r_l, l = 1, \ldots, H, l \neq h | n_h = r_h)} \\
&= \frac{\mathrm{P}(S^h = s^h, n_h = r_h)\mathrm{P}(n_l = r_l, l = 1, \ldots, H, l \neq h | n_h = r_h)}{\mathrm{P}(n_h = r_h)\mathrm{P}(n_l = r_l, l = 1, \ldots, H, l \neq h | n_h = r_h)} \\
&= \frac{\mathrm{P}(S^h = s^h, n_h = r_h)}{\mathrm{P}(n_h = r_h)} = \mathrm{P}(S^h = s^h | n_h = r_h).
\end{aligned}$$

Calculemos agora $\mathrm{P}(S^h = s^h | n_h = r_h)$, para $s^h \in Q_h$, onde $Q_h = \mathscr{S}_{n_h}$ é o suporte de S^h. Denotando $S = \{S^h, \bar{S}^h\}$ e representando por $\bar{Q}_h = \mathscr{S}_{n-n_h}$ o suporte de $\bar{S}^h$, temos

$$\begin{aligned}
\mathrm{P}(S^h = s^h | n_h = r_h) &= \frac{\mathrm{P}(S^h = s^h, n_h = r_h)}{\mathrm{P}(n_h = r_h)} \\
&= \frac{\mathrm{P}(S^h = s^h)\mathrm{P}(n_h = r_h | S_h = s_h)}{\mathrm{P}(n_h = r_h)} \\
&= \frac{\mathrm{P}(S^h = s^h)}{\mathrm{P}(n_h = r_h)} \\
&= \frac{\sum_{\bar{s}^h \in \bar{Q}_h} \mathrm{P}(S^h = s^h, \bar{S}^h = \bar{s}^h)}{\mathrm{P}(n_h = r_h)} \\
&= \frac{\sum_{\bar{s}^h \in \bar{Q}_h} \mathrm{P}(S = (s^h, \bar{s}^h))}{\mathrm{P}(n_h = r_h)} \\
&= \frac{C_{n-r_h}^{N-N_h}(C_n^N)^{-1}}{C_{r_h}^{N_h} C_{n-r_h}^{N-N_h}(C_n^N)^{-1}} = \frac{1}{C_{r_h}^{N_h}},
\end{aligned} \tag{4.5.4}$$

uma vez que a variável n_h possui uma distribuição hipergeométrica de parâmetros N, N_h e n.

Juntando (4.5.3) e (4.5.4), fica assim demonstrado que

$$\mathrm{P}(S^h = s^h | n_1 = r_1, \ldots, n_H = r_H) = \frac{1}{C_{r_h}^{N_h}}. \tag{4.5.5}$$

Independência condicional da família S_h, $h = 1, \ldots, H$

Por simplicidade de exposição, vamos apenas mostrar que S^h e $S^{h'}$, com $h \neq h'$, são variáveis independentes. De forma análoga se provaria a independência da família S^h, com $h = 1, \ldots, H$.

Usando o processo exposto atrás, e denotando $S = \{S^h, S^{h'}, \bar{S}^{h,h'}\}$ e $\bar{Q}_{h,h'} = \mathscr{S}_{n-n_h-n_{h'}}$ o suporte de $\bar{S}^{h,h'}$, podemos escrever

$$\begin{aligned}
&\mathrm{P}(S^h = s^h, S^{h'} = s^{h'} | n_1 = r_1, \ldots, n_H = r_H) \\
&= \mathrm{P}(S^h = s^h, S^{h'} = s^{h'} | n_h = r_h, n_{h'} = r_{h'}) \\
&= \sum_{\bar{s}_{h,h'} \in \bar{Q}_{h,h'}} \mathrm{P}(S = (s_h, s_{h'}, \bar{s}_{h,h'})) / \mathrm{P}(n_h = r_h, n_{h'} = r_{h'}) \\
&= C_{n-r_h-r_{h'}}^{N-N_h-N_{h'}} (C_n^N)^{-1} / \mathrm{P}(n_h = r_h, n_{h'} = r_{h'}).
\end{aligned}$$

Usando agora o facto de $(n_h, n_{h'})$ possuir uma distribuição hipergeométrica bivariada de parâmetros N, N_h, $N_{h'}$ e n, isto é,

$$\mathrm{P}(n_h = r, n_{h'} = r') = C_r^{N_h} C_{r'}^{N_{h'}} C_{n-r-r'}^{N-N_h-N_{h'}} (C_n^N)^{-1}, \tag{4.5.6}$$

para $r = \max(0, n - (N - N_h)), \ldots, \min(n, N_h)$ e $r' = \max(0, n - (N - N_{h'})), \ldots, \min(n, N_{h'})$, podemos escrever

$$\mathrm{P}(S^h = s^h, S^{h'} = s^{h'} | n_1 = r_1, \ldots, n_H = r_H) = \frac{1}{C_{r_h}^{N_h}} \frac{1}{C_{r_{h'}}^{N_{h'}}},$$

o que, tendo em conta (4.5.5), nos permite concluir que S^h e $S^{h'}$ são condicionalmente independentes. ■

Apesar do estimador pós-estratificado do total ser enviesado, esse viés é desprezável mesmo quando o tamanho da amostra não é muito grande.

Teorema 4.5.7. *O estimador pós-estratificado do total t_y tem por média*

$$\mathrm{E}(\hat{t}_p) = t_y - \sum_{h=1}^{H} t_{yh} \mathrm{P}(n_h = 0),$$

onde

$$\mathrm{P}(n_h = 0) \leq \exp\Big(- \frac{n N_h}{N} \Big).$$

Dem: Usando o facto de que condicionalmente a $n_\ell, \ell = 1, \ldots, H$, as amostras aleatórias S_h são planos SSR de tamanhos n_h sobre $\mathscr{U}_h$, podemos escrever

$$\begin{aligned}
\mathrm{E}(\hat{t}_p | n_\ell, \ell = 1, \ldots, H) &= \sum_{h=1}^{H} \mathrm{E}(\hat{t}_{yh} | n_\ell, \ell = 1, \ldots, H) \mathbb{1}(n_h > 0) \\
&= \sum_{h=1}^{H} t_{yh} \mathbb{1}(n_h > 0) \\
&= t_y - \sum_{h=1}^{H} t_{yh} \mathbb{1}(n_h = 0). \qquad (4.5.8)
\end{aligned}$$

o que permite obter a média de $\hat{t}_p$.

Usando agora o facto da variável n_h possuir uma distribuição hipergeométrica de parâmetros N, N_h e n, tomando $r_h = 0$ em (4.5.1) concluímos que

$$\begin{aligned}
\mathrm{P}(n_h = 0) &= \frac{C_0^{N_h} C_{n-0}^{N-N_h}}{C_n^N} = \frac{(N - N_h)!(N-n)!n!}{n!(N - N_h - n)!N!} \\
&= \frac{(N - N_h)(N - N_h - 1) \ldots (N - N_h - n + 1)}{N(N-1) \ldots (N - n + 1)} \\
&= \Big(1 - \frac{N_h}{N}\Big)\Big(1 - \frac{N_h}{N-1}\Big) \ldots \Big(1 - \frac{N_h}{N - n + 1}\Big) \qquad (4.5.9) \\
&\leq \Big(1 - \frac{N_h}{N}\Big)^n \leq \Big(\exp\Big(-\frac{N_h}{N}\Big)\Big)^n \\
&= \exp\Big(-\frac{nN_h}{N}\Big). \qquad \blacksquare
\end{aligned}$$

Teorema 4.5.10. *A variância do estimador pós-estratificado do total t_y é dada por*

$$\mathrm{Var}(\hat{t}_p) = \sum_{h=1}^{H} N_h^2 \mathrm{E}\left(\frac{1}{n_h} \mathbb{1}(n_h > 0)\right) s_{yh}^2 - \sum_{h=1}^{H} N_h s_{yh}^2 + V,$$

onde

$$0 \leq V \leq \sum_{h=1}^{H} (N_h s_{yh}^2 + t_{yh}^2) \mathrm{P}(n_h = 0) + \left(\sum_{h=1}^{H} |t_{yh}| \mathrm{P}(n_h = 0)\right)^2.$$

Dem: Usando o facto de condicionalmente a $n_\ell, \ell = 1, \ldots, H$, o plano ser estratificado, temos

$$\begin{aligned}
\mathrm{Var}(\hat{t}_p | n_\ell, \ell = 1, \ldots, H) &= \mathrm{Var}\left(\sum_{h=1}^{H} \hat{t}_{yh} \mathbb{1}(n_h > 0) \Big| n_\ell, \ell = 1, \ldots, H\right) \\
&= \sum_{h=1}^{H} \mathrm{Var}(\hat{t}_{yh} \mathbb{1}(n_h > 0) | n_\ell, \ell = 1, \ldots, H) \\
&= \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{yh}^2}{n_h} \mathbb{1}(n_h > 0),
\end{aligned}$$

e assim

$$\begin{aligned}
&\mathrm{E}(\mathrm{Var}(\hat{t}_p|n_h, h=1,\ldots,H)) \\
&= \sum_{h=1}^{H} N_h^2 \,\mathrm{E}\left(\frac{1}{n_h}\mathbb{1}(n_h>0)\right) s_{yh}^2 - \sum_{h=1}^{H} N_h s_{yh}^2 P(n_h>0) \\
&= \sum_{h=1}^{H} N_h^2 \,\mathrm{E}\left(\frac{1}{n_h}\mathbb{1}(n_h>0)\right) s_{yh}^2 - \sum_{h=1}^{H} N_h s_{yh}^2 + \sum_{h=1}^{H} N_h s_{yh}^2 \mathrm{P}(n_h=0). \qquad (4.5.11)
\end{aligned}$$

Por outro lado, de (4.5.8) tiramos que

$$\begin{aligned}
&\mathrm{Var}(\mathrm{E}(\hat{t}_p|n_\ell, \ell=1,\ldots,H)) \\
&= \sum_{h,h'=1}^{H} t_{yh} t_{yh'} \mathrm{Cov}(\mathbb{1}(n_h=0), \mathbb{1}(n_{h'}=0)) \\
&= \sum_{h,h'=1}^{H} t_{yh} t_{yh'} \Big(\mathrm{P}(n_h=0, n_{h'}=0) - \mathrm{P}(n_h=0)\mathrm{P}(n_{h'}=0)\Big). \qquad (4.5.12)
\end{aligned}$$

Para $h \neq h'$, e fazendo $r = r' = 0$ em (4.5.6), obtemos

$$\begin{aligned}
&\mathrm{P}(n_h=0, n_{h'}=0) \\
&= \frac{C_0^{N_h} C_0^{N_{h'}} C_n^{N-N_h-N_{h'}}}{C_n^N} = \frac{(N-N_h-N_{h'})!(N-n)!n!}{n!(N-N_h-N_{h'}-n)!N!} \\
&= \frac{(N-N_h-N_{h'})(N-N_h-N_{h'}-1)\ldots(N-N_h-N_{h'}-n+1)}{N(N-1)\ldots(N-n+1)}
\end{aligned}$$

Além disso, de (4.5.9) e da desigualdade $1-(a+b)/C \leq (1-a/C)(1-b/C)$, para $C, a, b > 0$, obtemos

$$\begin{aligned}
\mathrm{P}(n_h=0, n_{h'}=0) &= \Big(1-\frac{N_h+N_{h'}}{N}\Big)\Big(1-\frac{N_h+N_{h'}}{N-1}\Big)\ldots\Big(1-\frac{N_h+N_{h'}}{N-n+1}\Big) \\
&\leq \Big(1-\frac{N_h}{N}\Big)\Big(1-\frac{N_h}{N-1}\Big)\ldots\Big(1-\frac{N_h}{N-n+1}\Big) \\
&\quad \times \Big(1-\frac{N_{h'}}{N}\Big)\Big(1-\frac{N_{h'}}{N-1}\Big)\ldots\Big(1-\frac{N_{h'}}{N-n+1}\Big) \\
&= \mathrm{P}(n_h=0)\mathrm{P}(n_{h'}=0).
\end{aligned}$$

Concluímos assim que

$$\Big|\mathrm{P}(n_h=0, n_{h'}=0) - \mathrm{P}(n_h=0)\mathrm{P}(n_{h'}=0)\Big| \leq \mathrm{P}(n_h=0)P(n_{h'}=0),$$

para $h \neq h'$, e portanto de (4.5.12) obtemos

$$\begin{aligned}
&\mathrm{Var}(\mathrm{E}(\hat{t}_p|n_\ell, \ell = 1, \ldots, H)) \\
&\leq \sum_{h=1}^{H} t_{yh}^2 \mathrm{P}(n_h = 0) + \sum_{h,h'=1, h\neq h'}^{H} |t_{yh}||t_{yh'}| \mathrm{P}(n_h = 0)\mathrm{P}(n_{h'} = 0) \\
&\leq \sum_{h=1}^{H} t_{yh}^2 \mathrm{P}(n_h = 0) + \left(\sum_{h=1}^{H} |t_{yh}| \mathrm{P}(n_h = 0) \right)^2 . \qquad (4.5.13)
\end{aligned}$$

Para concluir a demonstração basta usar (4.5.11), (4.5.13) e a igualdade

$$\mathrm{Var}(\hat{t}_p) = \mathrm{Var}(\mathrm{E}(\hat{t}_p|n_\ell, \ell = 1, \ldots, H)) + \mathrm{E}(\mathrm{Var}(\hat{t}_p|n_\ell, \ell = 1, \ldots, H)). \qquad \blacksquare$$

De forma a podermos exibir uma expressão alternativa para $\mathrm{Var}(\hat{t}_p)$ que possa ser útil na prática, vamos começar por verificar que quando o tamanho da amostra é grande é válida a aproximação

$$\mathrm{E}\left(\frac{1}{n_h} \mathbb{1}(n_h > 0) \right) \approx \frac{N}{nN_h} + \frac{N - N_h}{N_h^2} \frac{N - n}{n^2}.$$

Tal como fizemos em §2.4.3, comecemos por utilizar a fórmula de Taylor de segunda ordem relativamente à função $g(x) = 1/x$ ($g'(x) = -1/x^2$, $g''(x) = 2/x^3$), para obter a aproximação

$$\begin{aligned}
\frac{1}{n_h} - \frac{1}{\mathrm{E}(n_h)} &= g(n_h) - g(\mathrm{E}(n_h)) \\
&\approx g'(\mathrm{E}(n_h))(n_h - \mathrm{E}(n_h)) + \frac{1}{2} g''(\mathrm{E}(n_h))(n_h - \mathrm{E}(n_h))^2 \\
&= -\frac{1}{\mathrm{E}(n_h)} \frac{n_h - \mathrm{E}(n_h)}{\mathrm{E}(n_h)} + \frac{1}{\mathrm{E}(n_h)} \Big(\frac{n_h - \mathrm{E}(n_h)}{\mathrm{E}(n_h)} \Big)^2,
\end{aligned}$$

ou seja,

$$\frac{1}{n_h} \approx \frac{1}{\mathrm{E}(n_h)} (1 - \epsilon + \epsilon^2),$$

e portanto

$$\mathrm{E}\left(\frac{1}{n_h} \mathbb{1}(n_h > 0) \right) \approx \frac{1}{\mathrm{E}(n_h)} \mathrm{E}\big((1 - \epsilon + \epsilon^2) \mathbb{1}(n_h > 0) \big),$$

onde

$$\epsilon = \frac{n_h - \mathrm{E}(n_h)}{\mathrm{E}(n_h)}$$

é tal que

$$\mathrm{E}(\epsilon) = 0$$

e

$$\mathrm{Var}(\epsilon) = \frac{\mathrm{Var}(n_h)}{\mathrm{E}(n_h)^2} = \frac{1}{n}\frac{N}{N_h}\frac{N-N_h}{N}\frac{N-n}{N-1},$$

uma vez que

$$\mathrm{E}(n_h) = \frac{nN_h}{N}$$

e

$$\mathrm{Var}(n_h) = \frac{nN_h}{N}\frac{N-N_h}{N}\frac{N-n}{N-1}.$$

Atendendo a que

$$\begin{aligned}
\mathrm{E}\big(\epsilon \mathbb{1}(n_h>0)\big) &= \frac{1}{\mathrm{E}(n_h)}\mathrm{E}\big(n_h\mathbb{1}(n_h>0) - \mathrm{E}(n_h)\mathbb{1}(n_h>0)\big)\\
&= \frac{1}{\mathrm{E}(n_h)}\big(\mathrm{E}(n_h) - \mathrm{E}(n_h)\mathrm{P}(n_h>0)\big)\\
&= \mathrm{P}(n_h=0),
\end{aligned}$$

e

$$\begin{aligned}
\mathrm{E}\big(\epsilon^2 \mathbb{1}(n_h>0)\big) &= \frac{1}{\mathrm{E}(n_h)^2}\mathrm{E}\big((n_h-\mathrm{E}(n_h))^2\mathbb{1}(n_h>0)\big)\\
&= \frac{1}{\mathrm{E}(n_h)^2}\big(\mathrm{Var}(n_h) - \mathrm{E}((n_h-\mathrm{E}(n_h))^2\mathbb{1}(n_h=0))\big)\\
&= \frac{1}{\mathrm{E}(n_h)^2}\big(\mathrm{Var}(n_h) - \mathrm{E}(n_h)^2\mathrm{P}(n_h=0)\big)\\
&= \frac{\mathrm{Var}(n_h)}{\mathrm{E}(n_h)^2} - \mathrm{P}(n_h=0),
\end{aligned}$$

concluímos finalmente que

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{n_h}\mathbb{1}(n_h>0)\right) &\approx \frac{1}{\mathrm{E}(n_h)}\left(1 + \frac{\mathrm{Var}(n_h)}{\mathrm{E}(n_h)^2} - 3\mathrm{P}(n_h=0)\right)\\
&= \frac{N}{nN_h}\left(1 + \frac{N^2}{n^2N_h^2}\frac{nN_h}{N}\frac{N-N_h}{N}\frac{N-n}{N-1} - 3\mathrm{P}(n_h=0)\right)\\
&\approx \frac{N}{nN_h} + \frac{N-N_h}{N_h^2}\frac{N-n}{n^2} - 3\frac{N}{nN_h}\mathrm{P}(n_h=0)\\
&\approx \frac{N}{nN_h} + \frac{N-N_h}{N_h^2}\frac{N-n}{n^2}, \qquad (4.5.14)
\end{aligned}$$

uma vez que o termo $\frac{N}{nN_h}\mathrm{P}(n_h=0)$ é desprezável quando n é grande.

Atendendo ao Teorema 4.5.10 e à aproximação (4.5.14), concluímos que a variância

do estimador pós-estratificado do total pode ser aproximada por

$$\begin{aligned}\mathrm{Var}(\hat{t}_p) &\simeq \sum_{h=1}^{H} N_h^2 \left(\frac{N}{nN_h} + \frac{N-N_h}{N_h^2}\frac{N-n}{n^2}\right) s_{yh}^2 - \sum_{h=1}^{H} N_h s_{yh}^2 \\ &= N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}\, s_{yh}^2 + \frac{1}{n^2}\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)\, s_{yh}^2\right).\end{aligned}$$

Na prática, o estimador da variância do estimador pós-estratificado do total é definido por

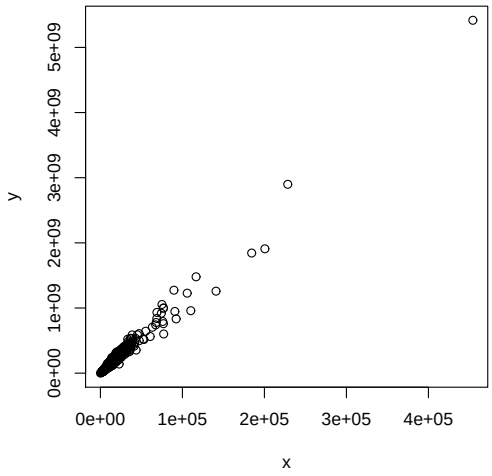
$$\widehat{\mathrm{Var}}(\hat{t}_p) = N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\sum_{\substack{h=1\\ n_h>1}}^{H}\frac{N_h}{N}\, \hat{s}_{yh}^2 + \frac{1}{n^2}\sum_{\substack{h=1\\ n_h>1}}^{H}\left(1-\frac{N_h}{N}\right)\, \hat{s}_{yh}^2\right).$$

Reparemos que o primeiro termo da aproximação obtida para a variância $\mathrm{Var}(\hat{t}_p)$ não é mais do que a variância do estimador estratificado do total quando a afetação é proporcional. Concluímos assim que é sempre preferível fazer uma estratificação *a priori*. No entanto, o último termo da expressão anterior é desprezável relativamente ao primeiro quando o tamanho da amostra é grande. Assim, não sendo possível fazer uma estratificação *a priori*, a estratégia de utilizar uma pós-estratificação é interessante, perdendo em geral pouco relativamente à estratificação com afetação proporcional. Como já sabemos, uma tal estratégia terá vantagem relativamente à amostragem aleatória simples quando os pós-estratos forem constituídos por unidades entre si homogéneas relativamente à variável de interesse e os pós-estrato forem heterogéneos entre si.

4.6 Alguns resultados de simulação

Nos exemplos seguintes a variável x =belgianmunicipalities$Tot03 foi usada como variável auxiliar. Esta variável foi também usada para definir os (pós-)estratos cujos limites são dados pelos valores 10000, 20000, 30000 e 50000 da variável x. Tomámos $n = 100$ (exceto quando indicado) e considerámos 1000 repetições do processo de amostragem.

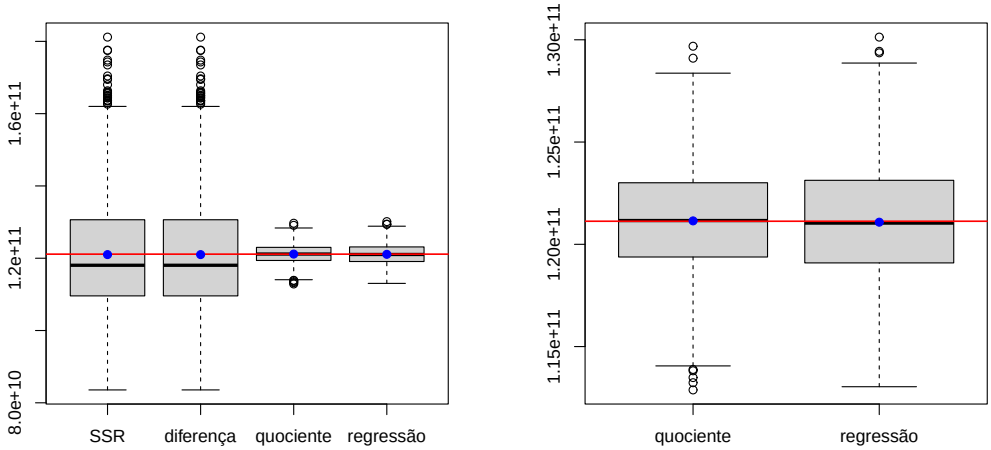
Exemplo 4.6.1. Variável de interesse: y = belgianmunicipalities$TaxableIncome. Parâmetro de interesse: t_y.



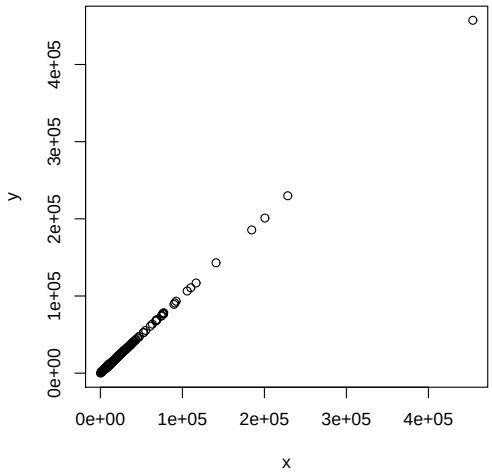
Reparemos que neste caso temos

$$\frac{\mathrm{Var}(\hat{t}_d)}{\mathrm{Var}(\hat{t}_y)} = 0{,}9998, \quad \frac{\mathrm{Var}(\hat{t}_q)}{\mathrm{Var}(\hat{t}_y)} \approx 0{,}0235, \quad \text{e} \quad \frac{\mathrm{Var}(\hat{t}_r)}{\mathrm{Var}(\hat{t}_y)} \approx 0{,}0231,$$

o que explica os resultados seguintes:



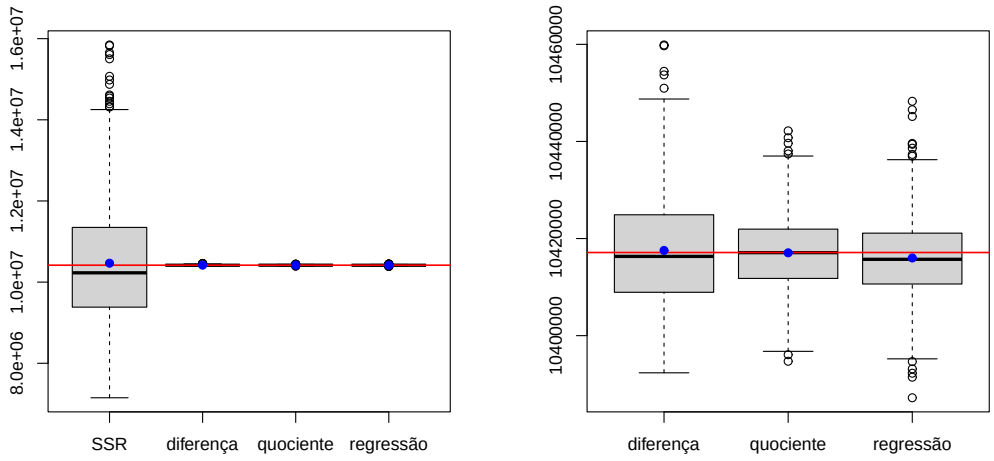
Exemplo 4.6.2. Variável de interesse: $y =$ belgianmunicipalities$Tot04. Parâmetro de interesse: t_y.



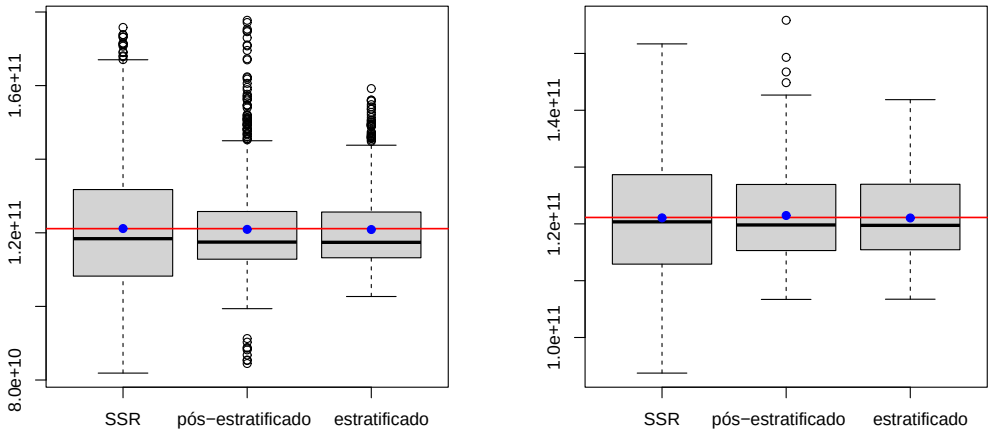
Neste caso temos

$$\frac{\mathrm{Var}(\hat{t}_d)}{\mathrm{Var}(\hat{t}_y)} = 5{,}49 \times 10^{-5}, \quad \frac{\mathrm{Var}(\hat{t}_q)}{\mathrm{Var}(\hat{t}_y)} \approx 2{,}50 \times 10^{-5} \quad \text{e} \quad \frac{\mathrm{Var}(\hat{t}_r)}{\mathrm{Var}(\hat{t}_y)} \approx 1{,}16 \times 10^{-5},$$

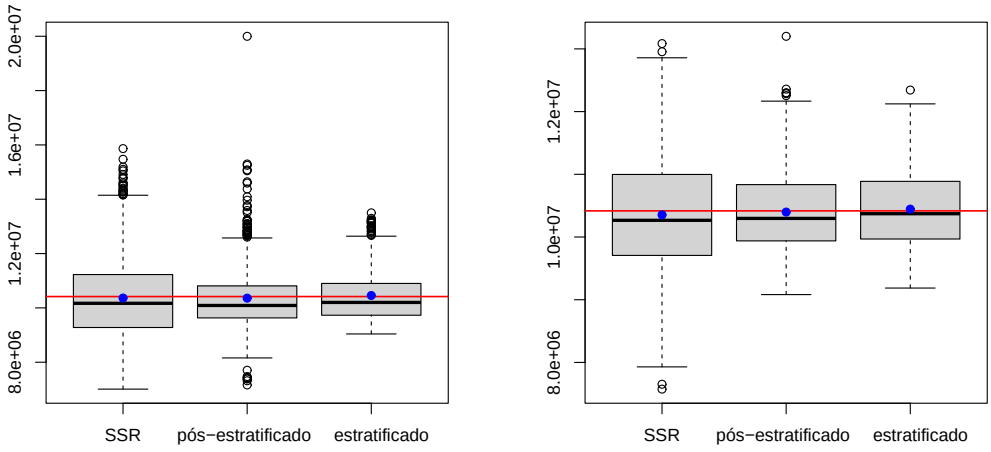
sendo de esperar um melhor desempenho dos estimadores do quociente e da regressão.



Exemplo 4.6.3. Variável de interesse: $y =$ belgianmunicipalities$TaxableIncome. Parâmetro de interesse: t_y. No gráfico da esquerda $n = 100$, enquanto que no da direita $n = 200$. Especialmente no primeiro caso vemos que a pós-estratificação perde para a estratificação. Mesmo assim, é preferível efetuar uma pós-estratificação a partir da informação disponível sobre a população do que não usar essa informação e executar um plano simples.



Exemplo 4.6.4. Variável de interesse: $y =$ belgianmunicipalities$Tot04. Parâmetro de interesse: t_y. Tomámos $n = 100$ no gráfico da esquerda e $n = 200$ no da direita. Tal como no exemplo anterior, vemos que é sempre preferível efetuar uma pós-estratificação a partir da informação disponível sobre a população do que desprezar essa informação executando um plano simples.



4.7 Bibliografia

Cochran, W.G. (1977). *Sampling techniques.* Wiley. (Capítulos 6 e 7)

Hedayat, A.S., Sinha, B.K. (1991). *Design and inference in finite population sampling.* Wiley. (Capítulo 6)

Laplace, P.S. (1814). *Théorie analytique des probabilités.* Paris: Ve. Courcier.

Lohr, S.L. (1999). *Sampling: design and analysis.* Duxbury Press. (Capítulo 3)

Thompson, M.E. (2002). *Theory of sample surveys.* Wiley. (Capítulos 7 e 8)

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod. (Capítulo 10)

Tillé, Y., Matei, A. (2021). 'sampling': survey sampling. R Package Version 2.9. http://CRAN.R-project.org/package=sampling

5

Planos de amostragem com probabilidades desiguais

Planos de amostragem com reposição. Estimador de Hansen-Hurvitz. Planos de amostragem PPS. Planos de amostragem sem reposição. Estimador de Narain-Horvitz-Thompson. O plano de Poisson. Planos de amostragem IPPS. Planos sem reposição de tamanho fixo. Estimador da variância e condições de Sen-Yates-Grundy. Os planos sistemático, de Lahiri-Midzuno e de Rao-Sampford. Normalidade assintótica e aproximação da variância.

5.1 Planos de amostragem com reposição de tamanho n

Vimos no último capítulo como podemos usar determinado tipo de informação auxiliar para construir melhores estimadores. Neste capítulo vamos estudar a possibilidade de usar informação auxiliar na fase da definição do plano de amostragem de modo a obter melhores estimadores. Em tais planos de amostragem as unidades populacionais não possuem necessariamente igual probabilidade de serem selecionadas para a amostra. Reparemos que, em geral, é esta a situação do plano de amostragem estratificada com afetação óptima.

Numa população de dimensão N, um plano de amostragem com reposição com probabilidades de inclusão desiguais de tamanho n, corresponde à extração, com reposição, de n unidades da população, em que em cada extração a probabilidade de selecionar a unidade k da população é p_k, onde $\sum_{k=1}^{N} p_k = 1$. Assim, um plano de amostragem com reposição com probabilidades de inclusão desiguais de tamanho n, que denotaremos de forma genérica por CR, é definido por

$$p(s) = p(s_1, ..., s_n) = p_{s_1} \times \cdots \times p_{s_n}$$

para todo o $s \in Q = \tilde{\mathscr{S}}_n = \mathscr{U}^n$. Representando por $S = (S_1, \ldots, S_n)$ a amostra aleatória com plano de amostragem p, da forma produto desta distribuição podemos

imediatamente deduzir o seguinte resultado.

Proposição 5.1.1. *Num plano de amostragem CR de tamanho n as variáveis $S_1, \ldots, S_n$ são independentes e identicamente distribuídas com*

$$\mathrm{P}_{S_i}(s_i) = p_{s_i}$$

para $s_i \in \mathscr{U}$.

Vamos centrar a nossa atenção na estimação do total t_y da população. Os resultados seguintes são devidos a Hansen e Hurwitz (1943).

Teorema 5.1.2. *Num plano de amostragem CR de tamanho n o estimador*

$$\hat{t}_{HH} = \sum_{k \in S} \frac{y_k}{n p_k},$$

dito estimador de Hansen-Hurvitz, é um estimador cêntrico de t_y com variância

$$\begin{aligned}
\mathrm{Var}(\hat{t}_{HH}) &= \frac{1}{n} \sum_{k \in \mathscr{U}} p_k \left(\frac{y_k}{p_k} - t_y \right)^2 \\
&= \frac{1}{n} \left(\sum_{k \in \mathscr{U}} \frac{y_k^2}{p_k} - t_y^2 \right) \\
&= \frac{1}{2n} \sum_{k,l \in \mathscr{U}} p_k p_l \left(\frac{y_k}{p_k} - \frac{y_l}{p_l} \right)^2 .
\end{aligned}$$

Dem: Atendendo a que

$$\hat{t}_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{S_i}}{p_{S_i}},$$

basta usar a Proposição 5.1.1 para concluir que $\hat{t}_{HH}$ é um estimador cêntrico de t_y uma vez que

$$\mathrm{E}\Big(\frac{y_{S_1}}{p_{S_1}}\Big) = \sum_{k \in U} \frac{y_k}{p_k} \mathrm{P}(S_1 = k) = \sum_{k \in U} \frac{y_k}{p_k} p_k = t_y.$$

As duas primeiras expressões para a variância do estimador $\hat{t}_{HH}$ resultam do facto das variáveis $\frac{y_{S_i}}{p_{S_i}}, i = 1, \ldots, n$, serem independentes e identicamente distribuídas e das duas expressões seguintes para as suas variâncias:

$$\mathrm{Var}\Big(\frac{y_{S_1}}{p_{S_1}}\Big) = \sum_{k \in U} \Big(\frac{y_k}{p_k} - t_y\Big)^2 \mathrm{P}(S_1 = k) = \sum_{k \in U} p_k \Big(\frac{y_k}{p_k} - t_y\Big)^2,$$

e

$$\mathrm{Var}\Big(\frac{y_{S_1}}{p_{S_1}}\Big) = \mathrm{E}\Big(\frac{y_{S_1}}{p_{S_1}}\Big)^2 - t_y^2 = \sum_{k\in U} \frac{y_k^2}{p_k^2}\,\mathrm{P}(S_1 = k) - t_y^2 = \sum_{k\in U} \frac{y_k^2}{p_k} - t_y^2.$$

Para obter a última das expressões para a variância do estimador de HH, basta agora ter em conta que

$$\frac{1}{2}\sum_{k,l\in\mathscr{U}} p_k p_l \left(\frac{y_k}{p_k} - \frac{y_l}{p_l}\right)^2 = \sum_{k\in U} \frac{y_k^2}{p_k} - t_y^2. \qquad \blacksquare$$

Teorema 5.1.3. *Num plano de amostragem CR de tamanho n o estimador*

$$\widehat{\mathrm{Var}}(\hat{t}_{HH}) = \frac{1}{n(n-1)} \sum_{k\in S} \left(\frac{y_k}{p_k} - \hat{t}_{HH}\right)^2$$

é um estimador cêntrico de $\mathrm{Var}(\hat{t}_{HH})$.

Dem: Tendo em conta que

$$\begin{aligned}
\sum_{k\in S} \Big(\frac{y_k}{p_k} - \hat{t}_{HH}\Big)^2 &= \sum_{k\in S} \Big(\Big(\frac{y_k}{p_k} - t_y\Big) - (\hat{t}_{HH} - t_y)\Big)^2 \\
&= \sum_{k\in S} \Big(\frac{y_k}{p_k} - t_y\Big)^2 - n(\hat{t}_{HH} - t_y)^2 \\
&= \sum_{i=1}^{n} \Big(\frac{y_{S_i}}{p_{S_i}} - t_y\Big)^2 - n(\hat{t}_{HH} - t_y)^2,
\end{aligned}$$

concluímos que

$$\mathrm{E}(\widehat{\mathrm{Var}}(\hat{t}_{HH})) = \frac{1}{n(n-1)} \left(n\mathrm{Var}\Big(\frac{y_{S_1}}{p_{S_1}}\Big) - n\mathrm{Var}(\hat{t}_{HH})\right) = \mathrm{Var}(\hat{t}_{HH}). \qquad \blacksquare$$

Se pretendermos estimar a média $\bar{y}$, o estimador

$$\hat{\bar{y}}_{HH} = \frac{1}{N} \sum_{k\in S} \frac{y_k}{np_k}$$

é um estimador cêntrico de $\bar{y}$. No caso do plano SCR temos $p_k = 1/N$ para todo o k, e os estimadores $\hat{\bar{y}}_{HH}$ e $\hat{t}_{HH}$ são precisamente os estimadores da média e do total considerados no Capítulo 2.

Atendendo à forma da variância do estimador de Hansen-Hurvitz, verificamos que a qualidade da estimação produzida pode ser melhorada através de uma escolha adequada das probabilidades p_k. Para justificar esta afirmação, admitamos que para todas as

unidades populacionais conhecemos os valores de uma variável, $x_k > 0$, que admitimos ser aproximadamente proporcional a y_k, isto é,

$$y_k \approx \lambda x_k,\ k \in \mathscr{U}.$$

Neste caso, se as probabilidades p_k forem tomadas proporcionais a x_k, isto é,

$$p_k = x_k/t_x,\ k \in \mathscr{U},$$

a variância do estimador de Hansen-Hurvitz é aproximadamente igual a zero pois neste caso

$$\frac{y_k}{p_k} \approx \lambda t_x,$$

para todo o $k \in \mathscr{U}$.

Assim, na posse de informação auxiliar sobre a população, que no presente contexto consiste no conhecimento duma variável auxiliar x aproximadamente proporcional à variável de interesse y, podemos melhorar a qualidade da estimação produzida através da utilização dum sistema de amostragem que tira partido dessa informação.

Um plano de amostragem com reposição em que as probabilidades p_k são proporcionais a uma variável auxiliar $x_k > 0$, para $k \in \mathscr{U}$, é dito plano de amostragem PPS (Probability Proportional to Size). A quantidade

$$p_k = \frac{x_k}{t_x},$$

é dita medida normalizada do tamanho da unidade k.

5.2 Planos de amostragem sem reposição

Poucos anos depois do trabalho de Hansen e Hurwitz (1943), Narain (1951) e Horvitz e Thompson (1952) propõem estimadores cêntricos do total t_y em planos de amostragem sem reposição. Sendo (p, Q) um plano de amostragem sem reposição (ver Definição 1.2.3), que denotaremos genericamente por SR, e S a amostra aleatória com distribuição de probabilidade p, sejam

$$\pi_k = \mathrm{P}(k \in S) \quad \text{e} \quad \pi_{kl} = \mathrm{P}(k, l \in S)$$

as probabilidades de inclusão de primeira e segunda ordens (ver Exercício 5). Atendendo ao papel relevante que estas probabilidades de inclusão têm, elas são também designadas por constantes estruturais do plano de amostragem. Sempre que $k = l$, da definição anterior resulta que $\pi_{kl} = \pi_k$. Sendo $\mathbb{1}_k(S)$ a variável indicatriz da unidade $k \in \mathscr{U}$ definida por

$$\mathbb{1}_k(S) = \begin{cases} 1, & \text{se } k \in S \\ 0, & \text{se } k \notin S, \end{cases}$$

as constantes estruturais do plano de amostragem são dadas por

$$\pi_k = \mathrm{E}(\mathbb{1}_k(S)) \quad \text{e} \quad \pi_{kl} = \mathrm{E}(\mathbb{1}_k(S)\mathbb{1}_l(S)).$$

Na demonstração dos resultados seguintes é utilizada uma ideia de Cornfield (1944) que observa que num plano sem reposição a soma $\sum_{k\in S} y_k$ pode ser escrita na forma $\sum_{k\in\mathscr{U}} y_k \mathbb{1}_k(S)$, o que permite que as suas média e variância se exprimam em termos das médias e das covariâncias da família de variáveis aleatórias $\mathbb{1}_k(S)$, $k \in \mathscr{U}$.

Teorema 5.2.1. *Num plano de amostragem SR o estimador*

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

dito estimador de Narain-Horvitz-Thompson *ou* π-estimador, *é um estimador cêntrico de* t_y *com variância*

$$\mathrm{Var}(\hat{t}_\pi) = \sum_{k,l\in\mathscr{U}} \frac{y_k y_l}{\pi_k \pi_l}(\pi_{kl} - \pi_k\pi_l).$$

Dem: Sendo o plano de amostragem sem reposição vale a igualdade

$$\hat{t}_\pi = \sum_{k\in S} \frac{y_k}{\pi_k} = \sum_{k\in\mathscr{U}} \frac{y_k}{\pi_k}\,\mathbb{1}_k(S),$$

e portanto

$$\mathrm{E}(\hat{t}_\pi) = \sum_{k\in\mathscr{U}} \frac{y_k}{\pi_k}\,\mathrm{E}(\mathbb{1}_k(S))$$

e

$$\mathrm{Var}(\hat{t}_\pi) = \sum_{k,l\in\mathscr{U}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\,\mathrm{Cov}(\mathbb{1}_k(S), \mathbb{1}_k(S)).$$

Para concluir basta ter em conta que $\mathrm{E}(\mathbb{1}_k(S)) = \pi_k$ e que $\mathrm{Cov}(\mathbb{1}_k(S), \mathbb{1}_l(S)) = \pi_{kl} - \pi_k\pi_l$. ■

Atendendo ao Exercício 38 podemos imediatamente apresentar um estimador cêntrico da variância anterior. No entanto, tal estimador não é necessariamente estritamente positivo.

Teorema 5.2.2. *Num plano de amostragem SR com* $\pi_{kl} > 0$, *para todo o* $k \neq l$, *o estimador*

$$\widehat{\mathrm{Var}}(\hat{t}_\pi) = \sum_{k,l\in S} \frac{y_k y_l}{\pi_k \pi_l}\,\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}},$$

é um estimador cêntrico de $\mathrm{Var}(\hat{t}_\pi)$.

Dem: Tendo em conta que

$$\mathrm{Var}(\hat{t}_\pi) = \sum_{k\in\mathscr{U}} \frac{y_k^2}{\pi_k}(1-\pi_k) + \sum_{k,l\in\mathscr{U}:k\neq l} \frac{y_k y_l}{\pi_k \pi_l}(\pi_{kl} - \pi_k\pi_l),$$

basta usar o Exercício 38 para concluir que

$$\sum_{k\in S} \frac{y_k^2}{\pi_k^2}(1-\pi_k) + \sum_{k,l\in S:k\neq l} \frac{y_k y_l}{\pi_k \pi_l} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} = \widehat{\mathrm{Var}}(\hat{t}_\pi),$$

é um estimador cêntrico de $\mathrm{Var}(\hat{t}_\pi)$ sempre que $\pi_{kl} > 0$, para todo o $k \neq l$. ■

Reparemos que o estimador anterior pode ser usado mesmo no caso em que existam probabilidades de segunda ordem nulas uma vez que tais unidades não surgem simultaneamente na amostra observada. No entanto tal estimador não é cêntrico (ver Exercício 38).

Caso estejamos interessados na estimação da média populacional $\bar{y}$, o estimador de NHT da média é naturalmente definido por

$$\hat{\bar{y}}_\pi = \frac{1}{N}\sum_{k\in S} \frac{y_k}{\pi_k}.$$

No caso do plano de amostragem ser um plano SSR de tamanho n, sabemos que $\pi_k = n/N$, caso em que $\hat{\bar{y}}_\pi$ se reduz à média amostral.

5.3 O plano de Poisson

Para um conjunto de números reais π_k^*, $k \in \mathscr{U}$, fixados à partida com $0 < \pi_k^* \leq 1$, o plano de amostragem de Poisson é definido por

$$p(s) = \prod_{k\in s} \pi_k^* \prod_{k\in\mathscr{U}\setminus s} (1-\pi_k^*),$$

para todo o $s \in Q = \mathscr{S}$. Quando $\pi_k^* = \pi$, para todo o $k \in \mathscr{U}$, o plano de Poisson reduz-se ao plano de Bernoulli (ver Exemplo 1.2.10).

Para implementar o plano de Poisson podemos proceder da forma seguinte: para cada unidade k da população geramos um número aleatório u_k segundo uma distribuição uniforme sobre o intervalo $[0,1]$ e incluímos a unidade k na amostra sempre que $u_k < \pi_k^*$ (ver Exercício 7).

Para $k, l \in \mathscr{U}$, com $k \neq l$, as probabilidade de inclusão de primeira e segunda ordens do plano de Poisson são dadas por

$$\pi_k = \pi_k^* \ \text{ e } \ \pi_{kl} = \pi_k^*\pi_l^*.$$

O resultado seguinte é consequência dos Teoremas 5.2.1 e 5.2.2.

Teorema 5.3.1. *Num plano de amostragem de Poisson, o estimador de NHT de t_y é dado por*

$$\hat{t}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k^*},$$

e tem por variância

$$\mathrm{Var}(\hat{t}_\pi) = \sum_{k \in \mathscr{U}} \frac{y_k^2}{\pi_k^*}(1 - \pi_k^*),$$

que pode ser estimada de forma cêntrica por

$$\widehat{\mathrm{Var}}(\hat{t}_\pi) = \sum_{k \in S} \frac{y_k^2}{\pi_k^{*2}}(1 - \pi_k^*).$$

A questão que naturalmente se coloca é a de saber se é possível tirar partido das probabilidades π_k^* de forma a melhorar a qualidade do estimador de NHT. Tendo em conta a forma da variância do estimador, pretendemos assim saber se é possível escolher as probabilidades de inclusão π_k^* por forma a minimizar

$$\sum_{k \in \mathscr{U}} \frac{y_k^2}{\pi_k^*}.$$

Vamos restringir a nossa análise ao caso em que fixamos um tamanho médio n para o plano de Poisson, isto é, pretendemos minimizar a expressão anterior com a restrição das probabilidades de inclusão satisfazerem a igualdade $\sum_{k \in \mathscr{U}} \pi_k^* = n$.

Usando a desigualdade de Cauchy-Schwarz (Proposição 3.6.1) e admitindo que a variável y é não negativa, sabemos que

$$\sum_{k \in \mathscr{U}} \frac{y_k^2}{\pi_k^*} = n^{-1} \sum_{k \in \mathscr{U}} \pi_k^* \sum_{k \in \mathscr{U}} \frac{y_k^2}{\pi_k^*} \geq n^{-1} \left(\sum_{k \in \mathscr{U}} y_k \right)^2,$$

onde o segundo membro não depende das probabilidades de inclusão de primeira ordem. Assim, as probabilidades de inclusão que minimizam a variância são proporcionais a y_k o que significa que

$$\pi_k^* = n y_k / t_y, \ k \in \mathscr{U},$$

com $t_y = \sum_{k \in \mathscr{U}} y_k$.

Na prática as probabilidades anteriores não podem ser calculadas pois dependem da variável de interesse y. No entanto, se, para toda a unidade da população, conhecermos os valores duma variável auxiliar $x_k > 0$ que é aproximadamente proporcional a y_k,

$$y_k \approx \lambda x_k, \ k \in \mathscr{U},$$

podemos aproximar as probabilidades de inclusão teóricas anteriores tomando

$$\pi_k^* = nx_k/t_x = np_k,\ k \in \mathscr{U},$$

onde $t_x = \sum_{k\in\mathscr{U}} x_k$ e $p_k = x_k/t_x$ é a medida normalizada do tamanho da unidade k.

Reparemos que as probabilidades de inclusão π_k^* dadas pela expressão anterior podem ser maiores que 1. Uma forma de solucionar este problema é a de selecionar para a amostra as unidades com $\pi_k^* \geq 1$, atribuindo-lhes probabilidades de inclusão iguais a 1. Sendo $\mathscr{U}_1 \subset \mathscr{U}$ o conjunto de tais unidades e n_1 o seu número, para as restantes unidades, consideradas agora como uma nova população, é necessário recalcular as probabilidades de inclusão para dela extrairmos uma amostra de tamanho médio igual a $n - n_1$. Isto significa que para $k \in \mathscr{U} \setminus \mathscr{U}_1$ devemos tomar

$$\pi_k^* = (n - n_1)x_k/t_{x,1},$$

com $t_{x,1} = \sum_{k\in\mathscr{U}\setminus\mathscr{U}_1} x_k$. Este processo deve ser repetido até que a um conjunto de unidades sejam atribuídas probabilidades de inclusão iguais a 1 e às restantes probabilidades de inclusão inferiores a 1.

Um plano de amostragem sem reposição em que as probabilidades de inclusão de primeira ordem π_k são proporcionais a uma variável $x_k > 0$, para $k \in \mathscr{U}$, é dito plano de amostragem ΠPS ou IPPS (Inclusion Probability Proportional to Size).

5.4 Planos sem reposição de tamanho fixo

No caso do plano de amostragem sem reposição ser de tamanho fixo, Sen (1953) e Yates e Grundy (1953) mostraram que a variância do estimador de Narain-Horvitz-Thompson pode ser expressa de forma diversa da apresentada atrás permitindo exibir um estimador alternativo para a variância do estimador de NHT.

Teorema 5.4.1. *Num plano de amostragem SR de tamanho fixo n a variância do estimador de Narain-Horvitz-Thompson é dada por*

$$\mathrm{Var}(\hat{t}_\pi) = -\frac{1}{2}\sum_{k,l\in\mathscr{U}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2(\pi_{kl} - \pi_k\pi_l).$$

Além disso, se $\pi_{kl} > 0$ para todo o $k \neq l$ o estimador

$$\widehat{\mathrm{Var}}_{\mathrm{SYG}}(\hat{t}_\pi) = -\frac{1}{2}\sum_{k,l\in S}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}},$$

dito estimador de Sen-Yates-Grundy, *é um estimador cêntrico de* $\mathrm{Var}(\hat{t}_\pi)$.

Dem: Atendendo ao Teorema 5.2.1, podemos escrever

$$\begin{aligned}
&-\frac{1}{2}\sum_{k,l\in\mathscr{U}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2(\pi_{kl}-\pi_k\pi_l)\\
&=-\sum_{k,l\in\mathscr{U}}\frac{y_k^2}{\pi_k^2}(\pi_{kl}-\pi_k\pi_l)+\sum_{k,l\in\mathscr{U}}\frac{y_k y_l}{\pi_k\pi_l}(\pi_{kl}-\pi_k\pi_l)\\
&=-\sum_{k\in\mathscr{U}}\frac{y_k^2}{\pi_k^2}\left(\sum_{l\in\mathscr{U}}\pi_{kl}-\pi_k\sum_{l\in\mathscr{U}}\pi_l\right)+\mathrm{Var}(\hat{t}_\pi).
\end{aligned}$$

Para concluir basta ter em conta que

$$\sum_{l\in\mathscr{U}}\pi_{kl}=\sum_{l\in\mathscr{U}}\mathrm{E}(\mathbb{I}_k(S)\mathbb{I}_l(S))=\mathrm{E}(\mathbb{I}_k(S)n)=n\pi_k,$$

e que $\sum_{\ell\in\mathscr{U}}\pi_\ell=n$ (ver Exercício 5). ■

O estimador de Sen-Yates-Grundy é não negativo sempre que $\pi_{kl}\leq\pi_k\pi_l$ para todo o $k,l\in\mathscr{U}$ com $k\neq l$. Tais condições são ditas condições de Sen-Yates-Grundy.

Tal como no caso dos planos com reposição, e atendendo à forma da variância do estimador de Narain-Horvitz-Thompson em planos de amostragem sem reposição de tamanho n, verificamos que a qualidade da estimação produzida pode ser melhorada através de uma escolha adequada das probabilidades de inclusão π_k. Com efeito, se para todas as unidades da população forem conhecidos os valores de uma variável, $x_k>0$, aproximadamente proporcional a y_k,

$$y_k\approx\lambda x_k, k\in\mathscr{U}$$

e se a probabilidade de inclusão π_k da unidade k for proporcional a x_k, isto é,

$$\pi_k=n\frac{x_k}{t_x}=np_k,$$

então a variância do estimador de Narain-Horvitz-Thompson é aproximadamente igual a zero, pois neste caso

$$\frac{y_k}{\pi_k}\approx\frac{\lambda}{n}t_x,$$

para todo o $k\in\mathscr{U}$.

Tal como no plano de Poisson reparemos que os π_k dados pela expressão anterior podem ser maiores que 1. Tal como descrevemos atrás, neste caso as probabilidades de inclusão devem ser recalculadas de modo a obtermos um conjunto de unidades com probabilidades de inclusão iguais a 1 e as restantes com probabilidades de inclusão inferiores a 1.

Apesar de ser enviesado quando pelo menos duas unidades têm probabilidades de inclusão de segunda ordem nulas, nas condições anteriores o viés do estimador de Sen-Yates-Grundy é em geral reduzido (ver Exercício 43).

5.5 Planos IPPS de tamanho n

Um problema não trivial a resolver motivado pelos resultados obtidos na secção anterior, é o da exibição de planos de amostragem sem reposição de tamanho n cujas probabilidades de inclusão π_k satisfaçam

$$\pi_k = np_k, \text{ para todo o } k \in \mathscr{U},$$

onde $p_k = x_k/t_x$ é a medida normalizada da unidade k, e que, atendendo aos resultados anteriores, satisfaçam ainda as condições

$$0 < \pi_{kl} \leq \pi_k \pi_l, \text{para todo o } k \neq l, \tag{5.5.1}$$

que garantem a existência de um estimador não negativo e cêntrico da variância do estimador de NHT. A facilidade de cálculo das probabilidades de segunda ordem é também uma propriedade que pretendemos que tais planos possuam. Como referimos a seguir, não são conhecidos exemplos de planos de amostragem que verifiquem todas as propriedades anteriores.

Antes de descrevermos alguns planos IPPS de tamanho fixo, verifiquemos que a generalização do procedimento usado para gerar planos PPS ao caso dos planos sem reposição, não produz, em geral, um plano IPPS, isto é, um plano com probabilidades de inclusão de primeira ordem iguais a np_k para $k \in \mathscr{U}$, onde $p_k = x_k/t_x$ é a medida normalizada do tamanho da unidade k.

Tal como num plano PPS de tamanho n, admitamos que uma primeira unidade é extraída com probabilidade p_k, com $k \in \mathscr{U}$. Supondo que a unidade j foi a selecionada, podemos extrair uma segunda unidade de $\mathscr{U} \setminus \{j\}$ com probabilidades proporcionais a p_k para $k \in \mathscr{U} \setminus \{j\}$:

$$p_k^j = \frac{p_k}{1 - p_j}.$$

O procedimento poderia ser continuado permitindo a extração de uma amostra de um qualquer tamanho fixo n.

Fixando-nos no caso $n = 2$ e representando por S a amostra obtida pelo plano de amostragem anterior, verifiquemos que este não possui, em geral, probabilidades de inclusão de primeira ordem iguais a np_k, para $k \in \mathscr{U}$. Com efeito, para $k \in \mathscr{U}$ temos

$$\mathrm{P}(k \in S) = p_k \left(1 + \sum_{j \in \mathscr{U}: j \neq k} \frac{p_j}{1 - p_j} \right),$$

probabilidade estas que, em geral, não são iguais a $np_k = 2p_k$.

Claro que, em alternativa, poderíamos usar outras probabilidades p_k^*, $k \in \mathscr{U}$, com $\sum_{k\in\mathscr{U}} p_k^* = 1$, que, no caso $n = 2$, teriam necessariamente que satisfazer a

$$p_k^* \left(1 + \sum_{j\in\mathscr{U}:j\neq k} \frac{p_j^*}{1-p_j^*} \right) = 2p_k,$$

para todo o $k \in \mathscr{U}$. No entanto a determinação das probabilidades p_k^* é complexa e é na prática impossível para $n > 2$. Este método, que surge descrito em Horvitz e Thompson (1952, pp. 679–681) e Yates e Grundy (1953, p. 254), é habitualmente atribuído a Narain (1951).

Há diversos planos de amostragem que permitem obter probabilidades de inclusão fixadas à partida pelo utilizador. Descrevemos a seguir três desses planos de amostragem (assumimos sempre que $N \geq 3$). Outros planos de amostragem IPPS são descritos em Hedayat e Sinha (1991, pp. 101–148), Tillé (2001, pp. 79–97) e Tillé (2006).

5.5.1 Plano sistemático com probabilidades desiguais

O plano de amostragem sistemática com probabilidades de inclusão desiguais, proposto por Madow (1949), é uma forma muito simples de selecionar amostras com probabilidades de inclusão fixas à partida. Supomos que conhecemos os np_k, para $k \in \mathscr{U}$, onde $0 < np_k < 1$ e $\sum_{k\in\mathscr{U}} p_k = 1$. Definindo $V_0 = 0$ e

$$V_k = \sum_{l=1}^{k} np_l,$$

os pontos $V_0, V_1, \ldots, V_N$ constituem uma partição do intervalo $[0, n]$ que o divide em N subintervalos de amplitudes np_k, para $k = 1, \ldots, N$.

Para selecionar uma amostra segundo um plano sistemático com probabilidades de inclusão desiguais, começamos por gerar um número aleatório u sobre o intervalo $[0, 1[$ e incluímos a unidade k na amostra sse $V_{k-1} \leq u + j - 1 < V_k$ para algum $j = 1, \ldots, n$.

As probabilidades de inclusão de segunda ordem do plano sistemático podem ser calculadas em função das probabilidades de primeira ordem mas algumas delas são nulas, sendo esta a principal limitação deste plano de amostragem (ver Tillé, 2001, pp. 91–93).

Teorema 5.5.2. *O plano de amostragem sistemático é um plano de amostragem IPPS de tamanho fixo não satisfazendo as condições* $0 < \pi_{kl} \leq \pi_k\pi_l$, *para todo o* $k \neq l$.

Dem: Para $k \in \mathscr{U}$, fixo, a probabilidade de inclusão π_k é dada por

$$\begin{aligned}\pi_k &= \mathrm{P}(k \in S) \\ &= \mathrm{P}\Big(\bigcup_{j=1}^{n} \{V_{k-1} \leq U + j - 1 < V_k\} \Big) \\ &= \sum_{j=1}^{n} \mathrm{P}\big(V_{k-1} \leq U + j - 1 < V_k\big),\end{aligned}$$

uma vez que os conjuntos $\{V_{k-1} \leq U + j - 1 < V_k\}$, para $j = 1, \ldots, n$, são disjuntos. Representando por W_j as variáveis aleatórias $U + j - 1$ que possuem distribuições uniformes sobre os intervalos $[j-1, j[$, podemos escrever

$$\pi_k = \sum_{j=1}^{n} \mathrm{P}\big(V_{k-1} \leq W_j < V_k\big).$$

Vamos em primeiro lugar analisar o caso em que o intervalo $[V_{k-1}, V_k[$ está contido em algum dos intervalos $[j-1, j[$, para $j = 1, \ldots, N$. Sendo $[j'-1, j'[$ um tal intervalo, temos

$$\pi_k = \mathrm{P}\big(V_{k-1} \leq W_{j'} < V_k\big).$$

Usando o facto de $W_{j'}$ ser uma variável uniforme sobre o intervalo $[j'-1, j'[$, concluímos que

$$\pi_k = V_k - V_{k-1} = np_k.$$

Vejamos agora o que se passa quando $[V_{k-1}, V_k[$ não está contido em nenhum dos intervalos $[j-1, j[$, para $j = 1, \ldots, N$. Neste caso existe $j' \in \{1, \ldots, N-1\}$ tal que $[V_{k-1}, V_k[\subset [j'-1, j'[\cup [j', j'+1[$ e $j' - 1 \leq V_{k-1} < j' < V_k < j' + 1$. Temos então

$$\pi_k = \mathrm{P}\big(V_{k-1} \leq W_{j'} < V_k\big) + \mathrm{P}\big(V_{k-1} \leq W_{j'+1} < V_k\big).$$

Usando o facto de $W_{j'}$ e $W_{j'+1}$ serem variáveis uniformes sobre os intervalos $[j'-1, j'[$ e $[j', j'+1[$, respetivamente, concluímos que

$$\begin{aligned}\pi_k &= \mathrm{P}\big(V_{k-1} \leq W_{j'} < j'\big) + \mathrm{P}\big(j' \leq W_{j'+1} < V_k\big) \\ &= (j' - V_{k-1}) + (V_k - j') = V_k - V_{k-1} = np_k.\end{aligned}$$ ■

5.5.2 Plano de Lahiri-Midzuno

Suponhamos que as medidas normalizadas $p_k, k \in \mathscr{U}$ satisfazem as condições

$$\frac{n-1}{N-1} \leq np_k < 1, \text{ para todo o } k \in \mathscr{U}, \tag{5.5.3}$$

onde $\sum_{k\in\mathscr{U}} p_k = 1$. O plano de Lahiri-Midzuno, proposto por Lahiri (1951) e Midzuno (1952), inicia-se com a extração de uma primeira unidade segundo a distribuição de probabilidade

$$\alpha_k = \frac{N-1}{N-n}\left(np_k - \frac{n-1}{N-1}\right),\ k \in \mathscr{U}.$$

De entre as restantes $N-1$ unidades, selecionamos $n-1$ unidades segundo um plano SSR.

Teorema 5.5.4. *Sob as condições (5.5.3), o procedimento de Lahiri-Midzuno é um plano de amostragem IPPS de tamanho fixo satisfazendo* $0 < \pi_{kl} \leq \pi_k\pi_l$, *para todo o* $k \neq l$.

Dem: Ver Exercício 45. ■

A principal limitação deste plano tem a ver com a condição preliminar imposta às medidas normalizadas $p_k,\ k \in \mathscr{U}$. Existe no entanto uma generalização deste plano que pode ser usada para quaisquer medidas normalizadas mas que não garante que as probabilidades de inclusão de segunda ordem sejam estritamente positivas (ver Tillé, 2001, pp. 108–109).

5.5.3 Plano de Rao-Sampford

Para um conjunto de medidas normalizadas p_k, $k \in \mathscr{U}$, com $0 < np_k < 1$, $k \in \mathscr{U}$, o plano de Rao-Sampford, proposto por Rao (1965) (caso $n = 2$) e Sampford (1967), inicia-se com a seleção de uma unidade com probabilidades p_k, $k \in \mathscr{U}$. As $n-1$ unidades seguintes são selecionadas, uma a uma, sempre de toda a população, com probabilidades de seleção proporcionais a $np_k/(1-np_k)$, $k \in \mathscr{U}$. Se todas as unidades selecionadas forem distintas, aceitamos essas unidades como amostra. Caso contrário, rejeitamos todas as unidades e repetimos o processo até obtermos um conjunto de n unidades distintas.

Teorema 5.5.5. *O procedimento de Rao-Sampford é um plano de amostragem IPPS de tamanho fixo satisfazendo* $0 < \pi_{kl} \leq \pi_k\pi_l$, *para todo o* $k \neq l$.

Para informação adicional sobre o plano de amostragem de Rao-Sampford, ver Hedayat e Sinha (1991, pp. 112–115) e Tillé (2006, pp. 130–136).

5.6 Normalidade do estimador de NHT

A normalidade assintótica ($n \to \infty$ e $N - n \to \infty$) do estimador de Narain-Horvitz-Thompson, que permite justificar a construção de intervalos de confiança baseados na

distribuição normal, é estudada por Hájek (1964) no caso do particular do plano de amostragem que consiste na extração com reposição de n unidades com probabilidades $\alpha_1, \ldots, \alpha_N$, com $\sum_{k \in \mathscr{U}} \alpha_k = 1$, e na seleção ou rejeição de todas as unidades selecionadas no caso delas serem, ou não, todas distintas. O processo é repetido até que uma amostra seja selecionada (*rejective sampling*). É possível provar que as probabilidades α_k podem ser escolhidas de forma que as probabilidades de inclusão de primeira ordem π_k sejam iguais a np_k, onde p_k é a medida normalizada do tamanho da unidade k. O resultado de Hájek (1964, p. 1514) é generalizado por Víšek (1979) e por Berger (1998a) a outros planos de amostragem. Em particular, sabemos que para o plano de Rao-Sampford o estimador de NHT é assintoticamente normal.

5.7 Aproximação da variância do estimador de NHT

Apesar da variância do estimador de NHT ser conhecida e ser também conhecido um estimador cêntrico desta no caso das probabilidades de inclusão de segunda ordem serem estritamente positivas, na prática tal estimador é de utilidade reduzida uma vez que as probabilidades segunda ordem são desconhecidas ou de cálculo muito difícil para muitos dos planos de amostragem conhecidos. Para que a teoria desenvolvida possa ser usada na construção de intervalos de confiança, surgem na literatura algumas propostas para aproximar e estimar a variância do estimador de NHT usando apenas as probabilidades de inclusão de primeira ordem do plano. Uma dela é devida a Hájek (1964, p. 1512) que propõe como aproximação de $\mathrm{Var}(\hat{t}_\pi)$ a quantidade (o coeficiente $N/(N-1)$ é introduzido por Berger, 1998b)

$$\sigma^2(\pi) = \frac{N}{N-1} \left(\sum_{k \in \mathscr{U}} \frac{y_k^2}{\pi_k} (1 - \pi_k) - d(\pi) G(\pi)^2 \right),$$

onde

$$d(\pi) = \sum_{k \in \mathscr{U}} \pi_k (1 - \pi_k)$$

e

$$G(\pi) = \frac{1}{d(\pi)} \sum_{k \in \mathscr{U}} y_k (1 - \pi_k).$$

Sob certas condições, que são em particular satisfeitas no caso do plano de Rao-Sampford, Berger (1998b) prova que a quantidade anterior é efetivamente uma aproximação de $\mathrm{Var}(\hat{t}_\pi)$ ($n \to \infty$, $N - n \to \infty$). Substituindo em $\sigma^2(\pi)$ cada total por estimadores de NHT, Berger (1998b) propõe o estimador de $\sigma^2(\pi)$ definido por

$$\hat{\sigma}^2(\pi) = \frac{n \hat{d}(\pi)}{(n-1) d(\pi)} \left(\sum_{k \in S} \frac{y_k^2}{\pi_k^2} (1 - \pi_k) - \hat{d}(\pi) \hat{G}(\pi)^2 \right),$$

onde

$$\hat{d}(\pi) = \sum_{k \in S} (1 - \pi_k)$$

e

$$\hat{G}(\pi) = \frac{1}{\hat{d}(\pi)} \sum_{k \in S} \frac{y_k}{\pi_k} (1 - \pi_k).$$

Os coeficientes $\frac{N}{N-1}$ e $\frac{n\hat{d}(\pi)}{(n-1)d(\pi)}$ considerados nas definições de $\sigma^2(\pi)$ e $\hat{\sigma}^2(\pi)$, respetivamente, têm como principal função ajustar a aproximação da variância e o estimador desta, de forma que no caso do plano SSR $\sigma^2(\pi)$ seja igual a $\text{Var}(\hat{t}_\pi)$ e $\hat{\sigma}^2(\pi)$ seja o usual estimador cêntrico de $\text{Var}(\hat{t}_\pi)$.

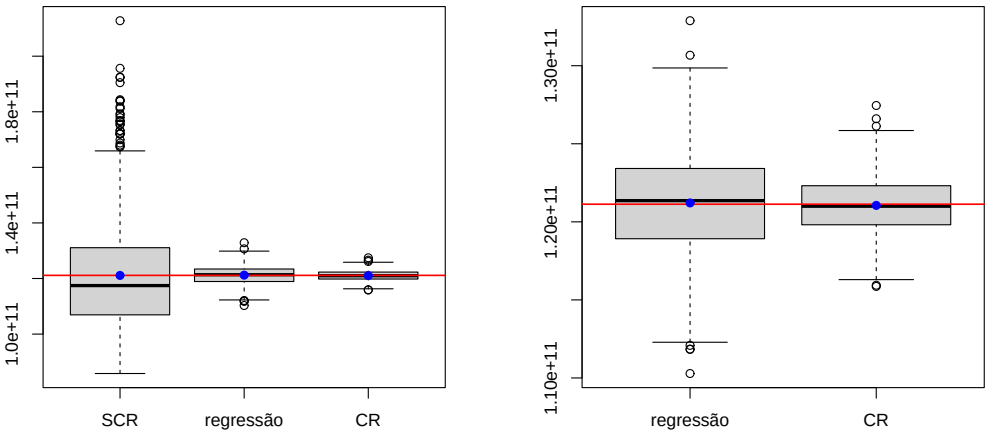
Proposição 5.7.1. *Se $\pi_k = n/N$, para todo o $k \in \mathscr{U}$, então*

$$\sigma^2(\pi) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n} \quad e \quad \hat{\sigma}^2(\pi) = N^2 \left(1 - \frac{n}{N}\right) \frac{\hat{s}_y^2}{n}.$$

5.8 Alguns resultados de simulação

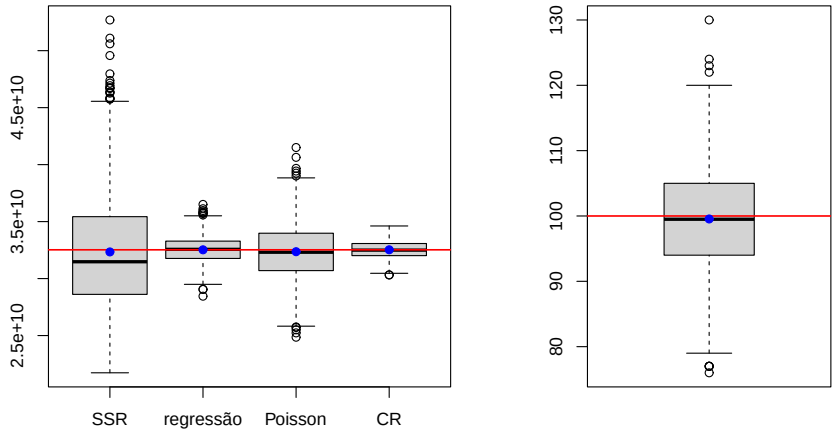
Nos exemplos seguintes a variável x =belgianmunicipalities$Tot04 foi usada como variável auxiliar. Nos dois primeiros exemplos tomámos $n = 100$ e considerámos 1000 repetições do processo de amostragem.

Exemplo 5.8.1. Variável de interesse: y = belgianmunicipalities$TaxableIncome. Parâmetro de interesse: t_y. O estimador da regressão que aqui consideramos é baseado no plano SCR.



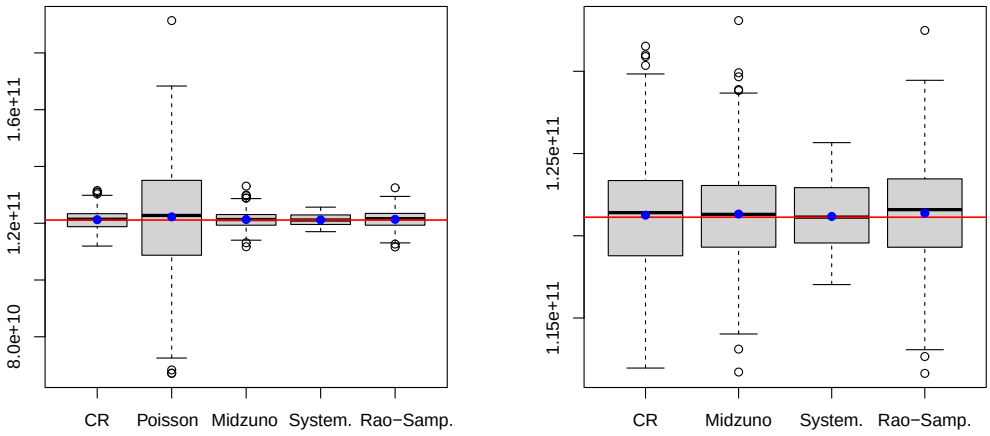
Exemplo 5.8.2. Variável de interesse: y = belgianmunicipalities$Totaltaxation. Parâmetro de interesse: t_y. O estimador da regressão que aqui consideramos é baseado

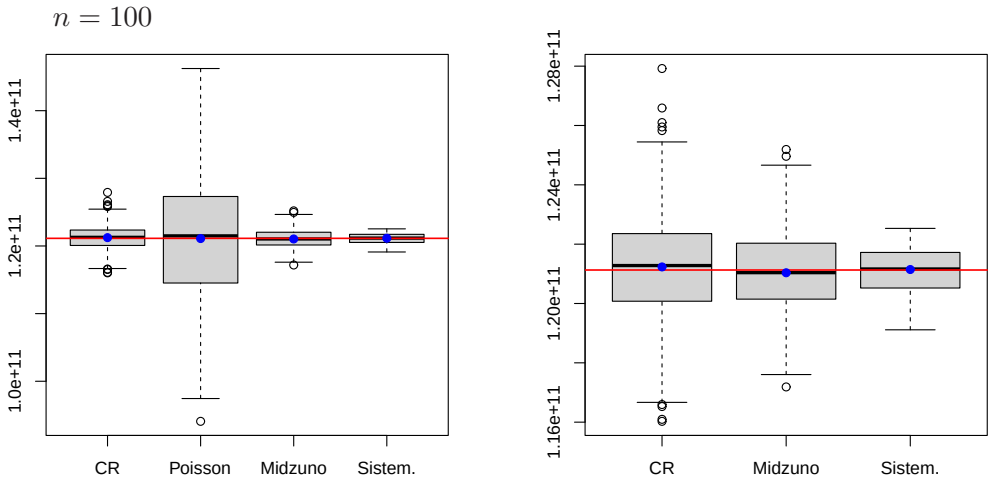
no plano SSR. O gráfico da direita descreve a distribuição dos tamanho das amostras obtidas pelo plano de Poisson.



Exemplo 5.8.3. Neste exemplo comparamos os planos de amostragem com probabilidade de inclusão desiguais com e sem reposição. Considerámos $n = 30$ e $n = 100$. Devido ao seu elevado tempo de execução, o plano de Rao-Sampford é considerado apenas quando $n = 30$. Em ambos os caso considerámos 500 repetições do processo de amostragem. Variável de interesse: $y =$ belgianmunicipalities$TaxableIncome. Parâmetro de interesse: t_y.

$n = 30$





5.9 Bibliografia

Berger, Y. (1998a). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. Statist. Plann. Inference* 67, 209–226.

Berger, Y. (1998b). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *J. Statist. Plann. Inference* 74, 149–168.

Cornfield, J. (1944). On samples from finite populations. *J. Amer. Statist. Assoc.* 39, 236–239.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* 35, 1491–1523.

Hansen, M.H., Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.* 14, 333–362.

Hedayat, A.S., Sinha, B.K. (1991). *Design and inference in finite population sampling.* Wiley. (Capítulos 3 e 5)

Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47, 663–685.

Lahiri, D.B (1951). A method of sample selection providing unbiased ratio estimates. *Bull. Int. Statist. Inst.* 33, Book 2, 133–140.

Lohr, S.L. (1999). *Sampling: design and analysis.* Duxbury Press. (Capítulo 6)

Madow, W.G. (1949). On the theory of systematic sampling, II. *Ann. Math. Statist.* 20, 333–354.

Midzuno, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Statist. Math.* 3, 99–107.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *J. Indian Soc. Agricultural Statist.* 3, 169–175.

Rao, J.N.K. (1965). On two sample schemes of unequal probability sampling without replacement. *J. Indian Statist. Assoc.* 3, 173–180.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* 54, 499–513.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *J. Indian Soc. Agric. Statist.* 5, 119–127.

Thompson, M.E. (2002). *Theory of sample surveys.* Wiley. (Capítulo 6)

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod. (Capítulo 5).

Tillé, Y. (2006). *Sampling algorithms.* Springer.

Tillé, Y., Matei, A. (2021). 'sampling': survey sampling. R Package Version 2.9. http://CRAN.R-project.org/package=sampling

Víšek, J.Á. (1979). Asymptotic distribution of simple estimate for rejective, Sampford and sucessive sampling. In: *Contribution to Statistics.* Jaroslav Hájek Memorial Volume, Jana Jureková (Ed.), D. Reidel Publishing Co., Dordrecht-Boston, Mass.-London, 263–275.

Yates, F., Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc. Ser. B* 15, 235–261.

6

Otimalidade e admissibilidade

Comparação da eficiência dos planos CR e SR. Teorema de Basu e Ghosh. Otimalidade e admissibilidade. Teoremas de Godambe e Joshi sobre a não existência de estimadores ótimos na classe dos estimadores cêntricos e sobre a admissibilidade do estimador de NHT na classe dos estimadores cêntricos do total.

6.1 Comparação da eficiência dos planos CR e SR

Do exposto no capítulo anterior ficou claro que contrariamente aos planos SR cuja implementação envolve alguma complexidade, os planos CR são extremamente fáceis de implementar. Além disso, a estimação da variância do estimador de Hansen-Hurvitz é muito mais simples do que a do estimador de Narain-Horvitz-Thompson. Para que se opte por determinado plano SR em detrimento do plano CR correspondente, é importante saber se o primeiro é mais eficiente que este último. Por outras palavras: sendo S uma amostra extraída segundo um plano SR de tamanho n com probabilidades de inclusão dadas por $\pi_k = np_k$, com $0 < np_k < 1$ e $\sum_{k\in\mathscr{U}} p_k = 1$, será o estimador de NHT definido por

$$\hat{t}_\pi = \sum_{k\in S} \frac{y_k}{\pi_k} = \sum_{k\in S} \frac{y_k}{np_k},$$

mais eficiente que o estimador de HH definido por

$$\hat{t}_{HH} = \frac{1}{n} \sum_{k\in S'} \frac{y_k}{p_k} = \sum_{k\in S'} \frac{y_k}{np_k},$$

onde S' é a amostra extraída segundo um plano CR de tamanho n em que a unidade k tem, em cada uma das n extrações, probabilidade p_k de ser selecionada?

Como sabemos, ambos os estimadores são cêntricos e as suas variâncias podem ser escritas na forma

$$\mathrm{Var}(\hat{t}_{HH}) = \sum_{k\in\mathscr{U}} \frac{z_k^2}{np_k}$$

e

$$\mathrm{Var}(\hat{t}_\pi) = \sum_{k\in\mathscr{U}} \frac{z_k^2}{np_k} + \frac{1}{n^2} \sum_{k,l\in\mathscr{U}:k\neq l} z_k z_l \frac{\pi_{kl}}{p_k p_l},$$

onde $z_k = y_k - p_k t_y$, $k \in \mathscr{U}$, é tal que $\sum_{k\in\mathscr{U}} z_k = 0$. Assim, $\hat{t}_\pi$ será mais eficiente que $\hat{t}_{HH}$ sse

$$\sum_{k,l\in\mathscr{U}:k\neq l} z_k z_l \frac{\pi_{kl}}{p_k p_l} \leq 0,$$

para todos os números reais $z_1, \ldots, z_N$ com $\sum_{k\in\mathscr{U}} z_k = 0$.

Tal é, por exemplo, o caso dos planos SCR e SSR para os quais $p_k = 1/N$, para todo $k \in \mathscr{U}$ (sobre esta propriedade ver §2.3).

Teorema 6.1.1. *No caso do plano simples sem reposição de tamanho $n \geq 2$, $\hat{t}_\pi$ é mais eficiente que $\hat{t}_{HH}$.*

Semelhante propriedade vale no caso do plano de Lahiri-Midzuno (ver §5.5.2).

Teorema 6.1.2. *No caso do plano de Lahiri-Midzuno, $\hat{t}_\pi$ é mais eficiente que $\hat{t}_{HH}$.*

No caso do plano de Rao-Sampford propriedade análoga foi estabelecida por Gabler (1981).

Teorema 6.1.3. *No caso do plano de amostragem de Rao-Sampford, $\hat{t}_\pi$ é mais eficiente que $\hat{t}_{HH}$.*

Como se ilustra no exemplo seguinte, a propriedade expressa nos resultados anteriores não é válida para um qualquer plano de amostragem.

Exemplo 6.1.4. Consideremos a variável de interesse y definida por $y_i = i$, $i \in \mathscr{U} = \{1, 2, 3, 4\}$ e o plano de amostragem SR de tamanho 2 definido por

s	(1,2)	(3,4)	(1,3)	(1,4)	(2,3)	(2,4)
$p(s)$	3/8	3/8	1/16	1/16	1/16	1/16

Para este plano temos $\pi_k = 1/2$, $k \in \mathscr{U}$, e $\mathrm{Var}(\hat{t}_\pi) = 12{,}5$ (Teorema 5.2.1). Se considerarmos o planos CR com $p_k = 1/4$, $k \in \mathscr{U}$ (trata-se dum plano SCR de tamanho 2), vemos que $np_k = 2 = \pi_k$, para $k \in \mathscr{U}$, e $\mathrm{Var}(\hat{t}_{HH}) = 10$ (Teorema 5.1.2). Isto é, o estimador de NHT não é mais eficiente que o estimador de HH.

Uma condição suficiente, sobre as probabilidades de inclusão de primeira e segunda ordens, para que $\hat{t}_\pi$ seja mais eficiente que $\hat{t}_{HH}$ é dada em Gabler (1984) no caso dos planos de amostragem conexos, isto é, planos de amostragem em que dadas duas quaisquer unidades k e l, ou a probabilidade de inclusão de segunda ordem π_{kl} é estritamente positiva, ou então existem unidades $i_1, \ldots, i_m$ tais que as probabilidades $\pi_{ki_1}, \pi_{i_1 i_2}, \ldots, \pi_{i_m l}$ são estritamente positivas.

Teorema 6.1.5. *Se o plano de amostragem SR é de tamanho fixo e conexo então* $\hat{t}_\pi$ *é mais eficiente que* $\hat{t}_{HH}$ *sempre que*

$$\sum_{k\in\mathscr{U}} \min_{l\in\mathscr{U}} \frac{\pi_{kl}}{\pi_l} > n-1.$$

6.2 O teorema de Basu e Ghosh

O principal resultado teórico que sustenta a utilização de planos sem reposição em detrimento de planos com reposição é devido a Basu e Ghosh (1967) e Basu (1969). Nele se estabelece que num qualquer plano de amostragem não há vantagem em usar a informação contida na amostra relativamente à ordem e à multiplicidade das unidades. Os primeiros resultados neste sentido são devidos a Basu (1958) e a Raj e Khamis (1958) no caso do plano SCR.

Estabelecemos a seguir o teorema de Basu e Ghosh, provando que dado um plano de amostragem e um qualquer estimador que possa depender da informação contida na amostra relativamente à ordem e à multiplicidade das unidades, este pode ser substituído, com vantagem, por outro que não usa tal informação. Uma discussão mais aprofundada sobre este assunto pode ser encontrada em Cassel et al. (1977, pp. 39–44) e Tillé (2001, pp. 27–31).

Dado um plano de amostragem geral (p, Q), vimos na Proposição 1.2.8 que lhe podemos associar uma versão reduzida (p^*, Q^*), definida, para $t \in Q^*$, por

$$p^*(t) = \sum_{s\in Q:\, r(s)=t} p(s),$$

onde $Q^* = \{r(s) : s \in Q\}$, e r é a função de redução.

Representamos por S a amostra aleatória com plano de amostragem p, e por T a amostra aleatória com plano de amostragem p^*.

Teorema 6.2.1. *Sejam* (p, Q) *um plano de amostragem e* $\hat{\theta} = \hat{\theta}(s; \mathbf{y})$ *um estimador de* θ*, definido para* $s \in Q$*, com* $\mathrm{E}(\hat{\theta}(S; \mathbf{y})^2) < \infty$*, para todo o* $\mathbf{y}$*. Então o estimador* $\tilde{\theta} = \tilde{\theta}(t; \mathbf{y})$*, definido, para* $t \in Q^*$*, por*

$$\tilde{\theta}(t; \mathbf{y}) = \frac{1}{p^*(t)} \sum_{s\in Q: r(s)=t} \hat{\theta}(s; \mathbf{y}) p(s),$$

satisfaz as seguintes propriedades:

a) $\mathrm{E}(\tilde{\theta}(T; \mathbf{y})) = \mathrm{E}(\hat{\theta}(S; \mathbf{y}))$*, para todo o* $\mathbf{y}$*.*

b) $\mathrm{Var}(\tilde{\theta}(T; \mathbf{y})) \leq \mathrm{Var}(\hat{\theta}(S; \mathbf{y}))$*, para todo o* $\mathbf{y}$*.*

Assim o sistema de amostragem $(p^*, \tilde{\theta}(\mathbf{y}))$ *é pelo menos tão eficiente como o sistema de amostragem* $(p, \hat{\theta}(\mathbf{y}))$*.*

Dem: A alínea a) decorre diretamente da definição de $\tilde{\theta}$:

$$\begin{aligned}\mathrm{E}(\tilde{\theta}(T;\mathbf{y})) &= \sum_{t\in Q^*}\tilde{\theta}(t;\mathbf{y})p^*(t) = \sum_{t\in Q^*}\sum_{s\in Q:r(s)=t}\hat{\theta}(s;\mathbf{y})p(s)\\ &= \sum_{s\in Q}\hat{\theta}(s;\mathbf{y})p(s) = \mathrm{E}(\hat{\theta}(S;\mathbf{y})).\end{aligned}$$

Para estabelecer b), comecemos por usar a desigualdade de Cauchy-Schwarz e a hipótese $\mathrm{E}(\hat{\theta}(S;\mathbf{y})^2) < \infty$ para concluir que

$$\begin{aligned}\tilde{\theta}(t;\mathbf{y})^2 &= \frac{1}{p^*(t)^2}\Big(\sum_{s\in Q:r(s)=t}\hat{\theta}(s;\mathbf{y})\sqrt{p(s)}\sqrt{p(s)}\Big)^2\\ &\le \frac{1}{p^*(t)^2}\Big(\sum_{s\in Q:r(s)=t}\hat{\theta}(s;\mathbf{y})^2p(s)\Big)\Big(\sum_{s\in Q:r(s)=t}p(s)\Big)\\ &= \frac{1}{p^*(t)}\sum_{s\in Q:r(s)=t}\hat{\theta}(s;\mathbf{y})^2p(s),\end{aligned}$$

para todo o $t \in Q^*$. Assim,

$$\mathrm{E}(\tilde{\theta}(T;\mathbf{y})^2) = \sum_{t\in Q^*}\tilde{\theta}(t;\mathbf{y})^2p^*(t) \le \sum_{t\in Q^*}\sum_{s\in Q:r(s)=t}\hat{\theta}(s;\mathbf{y})^2p(s) = \mathrm{E}(\hat{\theta}(S;\mathbf{y})^2),$$

o que, tendo em conta a alínea a), permite obter b). ■

Teorema 6.2.2. *Dado um plano de amostragem (p,Q) e um estimador $\hat{\theta} = \hat{\theta}(s;\mathbf{y})$ de θ, definido para $s \in Q$, com $\mathrm{E}(\hat{\theta}(S;\mathbf{y})^2) < \infty$, para todo o $\mathbf{y}$, seja $\tilde{\theta}$ o estimador definido no teorema anterior. Então o estimador $\bar{\theta}$ definido, para $s \in Q$, por $\bar{\theta}(s;\mathbf{y}) = \tilde{\theta}(r(s);\mathbf{y})$, é tal que:*

a) $\mathrm{E}(\bar{\theta}(S;\mathbf{y})) = \mathrm{E}(\hat{\theta}(S;\mathbf{y}))$, *para todo o* $\mathbf{y}$.

b) $\mathrm{Var}(\bar{\theta}(S;\mathbf{y})) \le \mathrm{Var}(\hat{\theta}(S;\mathbf{y}))$, *para todo o* $\mathbf{y}$.

Assim o estimador $\bar{\theta}$ é pelo menos tão eficiente como $\hat{\theta}$.

Dem: Para $k = 1, 2$ temos

$$\begin{aligned}\mathrm{E}(\bar{\theta}(S;\mathbf{y})^k) = \mathrm{E}(\tilde{\theta}(r(S);\mathbf{y})^k) &= \sum_{s\in Q}\tilde{\theta}(r(s);\mathbf{y})^kp(s) = \sum_{t\in Q^*}\sum_{s\in Q:r(s)=t}\tilde{\theta}(t;\mathbf{y})^kp(s)\\ &= \sum_{t\in Q^*}\tilde{\theta}(t;\mathbf{y})^k\sum_{s\in Q:r(s)=t}p(s) = \sum_{t\in Q^*}\tilde{\theta}(t;\mathbf{y})^kp^*(t) = \mathrm{E}(\tilde{\theta}(T;\mathbf{y})^k).\end{aligned}$$

O resultado é agora consequência do Teorema 6.2.1. ■

A não ser em casos muito particulares, não é em geral possível determinar o estimador $\bar{\theta} = \tilde{\theta}(r(s);\mathbf{y})$ definido nos resultados anteriores. No caso particular do plano

(p, Q) ser um plano SCR de tamanho n, é possível provar que tomando para $\hat{\theta}$ a média empírica

$$\hat{\theta} = \hat{\bar{y}}_{scr} = \frac{1}{n} \sum_{k \in S} y_k,$$

então o estimador $\bar{\theta}$ de Basu e Ghosh não é mais do que a média da amostra que se obtém da amostra original S depois de suprimida a informação sobre a ordem e a multiplicidade das unidades, ou seja, apenas consideramos as unidades distintas que ocorrem na amostra:

$$\bar{\theta} = \hat{\bar{y}}_{BG} = \frac{1}{n(r(S))} \sum_{k \in r(S)} y_k.$$

O Teorema 6.2.2 permite concluir que ambos os estimadores são estimadores cêntricos de $\bar{y}$ com

$$\mathrm{Var}(\hat{\bar{y}}_{BG}) \leq \mathrm{Var}(\hat{\bar{y}}_{scr}).$$

É ainda possível mostrar que

$$\mathrm{Var}(\hat{\bar{y}}_{BG}) = \left(\frac{1}{N^n} \sum_{j=1}^{N} j^{n-1} - \frac{1}{N} \right) s_y^2$$

(ver Basu, 1958, e Pathak, 1961), o que, tendo em conta o Teorema 2.1.2, permite concluir que

$$\mathrm{Var}(\hat{\bar{y}}_{BG}) < \mathrm{Var}(\hat{\bar{y}}_{scr}), \text{ para todo o } n \geq 3.$$

Apesar do estimador $\hat{\bar{y}}_{BG}$ ser mais eficiente que a média empírica quando o plano de amostragem é um SCR de tamanho n (para $n \geq 3$), ele não é mais eficiente que a média empírica $\hat{\bar{y}}_{ssr}$ num plano SSR de tamanho n. Com efeito, tendo em conta o Teorema 2.2.2 podemos concluir que

$$\mathrm{Var}(\hat{\bar{y}}_{ssr}) < \mathrm{Var}(\hat{\bar{y}}_{BG}), \text{ para todo o } n \geq 2.$$

6.3 Otimalidade

Dado um plano de amostragem, o ideal seria conseguirmos determinar o melhor de todos os possíveis estimadores do parâmetro de interesse. Como veremos de seguida, tal não é possível. Sobre esta questão ver também Hedayat e Sinha (1991, pp. 33–38), Tillé (2001, pp. 42–45) e Cassel et al. (1977, pp. 68–76).

Definição 6.3.1. *Dado um plano de amostragem (p, Q) e uma classe $\mathscr{C}$ de estimadores, dizemos que o estimador $\hat{\theta}$ é ótimo na classe $\mathscr{C}$ se $\hat{\theta}$ pertence a $\mathscr{C}$ e é pelo menos tão eficiente como todo o estimador $\tilde{\theta}(\mathbf{y})$ de $\mathscr{C}$, isto é,*

$$\mathrm{EQM}(\hat{\theta}(\mathbf{y})) \leq \mathrm{EQM}(\tilde{\theta}(\mathbf{y})), \quad \textit{para todo o } \mathbf{y} \in \mathbb{R}^N.$$

O resultado seguinte, devido a Godambe e Joshi (1965), estabelece a não existência de um estimador ótimo para qualquer plano de amostragem na classe dos estimadores cêntricos. Um resultado anterior de Godambe (1955) estabelecia uma propriedade semelhante para a classe dos estimadores da forma $\sum_{k\in S}\alpha_k(S)y_k$, ditos estimadores lineares homogéneos. A demonstração que apresentamos a seguir é devida a Basu (1971).

Definição 6.3.2. *Um plano de amostragem (p,Q) é dito um censo se $n(r(s))=N$, para todo o $s\in Q$.*

Teorema 6.3.3. *Dados um plano de amostragem que não é um censo e $\theta(\mathbf{y})$ um parâmetro que depende de todas as unidades de $\mathscr{U}$, não existe um estimador ótimo na classe dos estimadores cêntricos de $\theta(\mathbf{y})$ (classe esta que supomos não vazia).*

Dem: Consideremos $\hat{\theta}(s;\mathbf{y})$ um estimador na classe dos estimadores cêntricos de $\theta(\mathbf{y})$. Fixemos um qualquer $\mathbf{e}\in\mathbb{R}^N$ e consideremos o estimador

$$\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})=\hat{\theta}(s;\mathbf{y})-\hat{\theta}(s;\mathbf{e})+\theta(\mathbf{e}).$$

Este estimador é cêntrico

$$\mathrm{E}(\hat{\theta}_{\mathbf{e}}(S;\mathbf{y}))=\mathrm{E}(\hat{\theta}(S;\mathbf{y}))-\mathrm{E}(\hat{\theta}(S;\mathbf{e}))+\theta(\mathbf{e})=\theta(\mathbf{y})-\theta(\mathbf{e})+\theta(\mathbf{e})=\theta(\mathbf{y})$$

e

$$\hat{\theta}_{\mathbf{e}}(s;\mathbf{e})=\hat{\theta}(s;\mathbf{e})-\hat{\theta}(s;\mathbf{e})+\theta(\mathbf{e})=\theta(\mathbf{e}),$$

para todo o $s\in Q$. Assim, quando $\mathbf{y}=\mathbf{e}$ temos

$$\mathrm{EQM}(\hat{\theta}_{\mathbf{e}}(S;\mathbf{e}))=0.$$

Provámos assim que para qualquer $\mathbf{e}\in\mathbb{R}^N$ é possível construir um estimador cêntrico $\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})$ de $\theta(\mathbf{y})$ com $\mathrm{EQM}(\hat{\theta}_{\mathbf{e}}(S;\mathbf{e}))=0$.

Existindo um estimador ótimo $\tilde{\theta}(r(s);\mathbf{y})$ de $\theta(\mathbf{y})$ (pelo Teorema 6.2.2 este estimador depende de S apenas através de $r(S)$) teríamos

$$\mathrm{EQM}(\tilde{\theta}(r(S);\mathbf{y}))\leq\mathrm{EQM}(\hat{\theta}_{\mathbf{e}}(S;\mathbf{y}))$$

para todo o $\mathbf{y}\in\mathbb{R}^N$ e todo o $\mathbf{e}\in\mathbb{R}^N$, o que implicaria que

$$\mathrm{EQM}(\tilde{\theta}(r(S);\mathbf{y}))=0,\ \text{para todo o } \mathbf{y}\in\mathbb{R}^N,$$

ou seja,

$$\tilde{\theta}(r(s);\mathbf{y})=\theta(\mathbf{y}),$$

para todo o $s\in Q$ e $\mathbf{y}\in\mathbb{R}^N$. Tomando uma amostra $s\in Q$ com $n(r(s))<N$, o que é sempre possível visto o plano de amostragem não ser um censo, chegamos a uma

contradição uma vez que o primeiro membro depende de $\mathbf{y}$ apenas nas unidades que ocorrem em s enquanto que o segundo membro depende de $\mathbf{y}$ também através das unidades que não ocorrem em s. ■

Teorema 6.3.4. *Para qualquer plano de amostragem que não é um censo, não existe um estimador ótimo na classe dos estimadores cêntricos de t_y.*

Como consequência direta deste resultado concluímos que o estimador de Narain-Horvitz-Thompson não é óptimo na classe dos estimadores cêntricos de t_y.

Para outras classes de estimadores existem resultados do mesmo género estabelecendo a não existência de estimadores ótimos.

6.4 Admissibilidade

Não sendo possível encontrar estimadores ótimos, é natural que se diminua a exigência. A noção de admissibilidade no contexto da amostragem em populações finitas foi primeiramente considerada por Godambe (1960) e Roy e Chakravarti (1960) (ver Cassel et al., 1977, Capítulo 3).

Definição 6.4.1. *Dado um plano de amostragem (p, Q), um estimador $\hat{\theta}(\mathbf{y})$ pertencente a uma classe $\mathscr{C}$ de estimadores é dito* admissível na classe $\mathscr{C}$ *se não existe em $\mathscr{C}$ um estimador mais eficiente que $\hat{\theta}(\mathbf{y})$, isto é, se não existe $\tilde{\theta}(\mathbf{y}) \in \mathscr{C}$ tal que*

$$\mathrm{EQM}(\tilde{\theta}(\mathbf{y})) \leq \mathrm{EQM}(\hat{\theta}(\mathbf{y})), \text{ para todo o } \mathbf{y} \in \mathbb{R}^N$$

e

$$\mathrm{EQM}(\tilde{\theta}(\mathbf{y}_0)) < \mathrm{EQM}(\hat{\theta}(\mathbf{y}_0)), \text{ para algum } \mathbf{y}_0 \in \mathbb{R}^N.$$

Caso contrário o estimador $\hat{\theta}$ é dito não admissível na classe $\mathscr{C}$.

Dado $\hat{\theta}(s; \mathbf{y})$ um estimador na classe dos estimadores cêntricos de $\theta(\mathbf{y})$ e $\mathbf{e} = (e_k, k \in \mathscr{U}) \in \mathbb{R}^N$ fixo, definimos na demonstração do Teorema 6.3.3 o estimador cêntrico de $\theta(\mathbf{y})$ dado por

$$\hat{\theta}_{\mathbf{e}}(s; \mathbf{y}) = \hat{\theta}(s; \mathbf{y}) - \hat{\theta}(s; \mathbf{e}) + \theta(\mathbf{e}).$$

Seguindo a abordagem de Cassel et al. (1977, pp. 52–59), vamos provar que este estimador é admissível. Para tal, começamos por estabelecer o resultado auxiliar seguinte onde $\Omega_m, m = 0, 1, \ldots, N$, é a partição de $\mathbb{R}^N$ definida por

$$\Omega_m = \{\mathbf{y} \in \mathbb{R}^N : y_k \neq e_k \text{ para exatamente } m \text{ coordenadas de } \mathbf{y}\}.$$

Lema 6.4.2. *Dado um plano de amostragem sem reposição (p, Q), seja $\tilde{\theta}(s; \mathbf{y})$ um estimador cêntrico $\theta(\mathbf{y})$ com*

a) $\mathrm{EQM}(\tilde{\theta}(\mathbf{y})) \leq \mathrm{EQM}(\hat{\theta}_{\mathbf{e}}(\mathbf{y}))$*, para todo o* $\mathbf{y} \in \mathbb{R}^N$*;*

b) $\tilde{\theta}(s; \mathbf{y}) = \hat{\theta}_{\mathbf{e}}(s; \mathbf{y})$ *para todo o* $s \in Q$ *quando* $\mathbf{y} \in \Omega_m$*.*

Então $\tilde{\theta}(s; \mathbf{y}) = \hat{\theta}_{\mathbf{e}}(s; \mathbf{y})$ *para todo o* $s \in Q$ *quando* $\mathbf{y} \in \Omega_{m+1}$*.*

Dem: Seja $\mathbf{y} \in \Omega_{m+1}$ qualquer e consideremos a partição $Q_j, j = 0, 1, \ldots, m+1$ de Q definida por (estamos a usar o facto do plano ser sem reposição)

$$Q_j = \{s \in Q : y_k \neq e_k \text{ para exatamente } j \text{ coordenadas } k \in s\}.$$

1) Se $s \in \cup_{j=0}^{m} Q_j$, sabemos que $(y_k, k \in s)$ contém quando muito m coordenadas diferentes das correspondentes coordenadas de $(e_k, k \in s)$. Isto significa que podemos determinar $\mathbf{y}' \in \Omega_m$, com $\mathbf{y}'$ dependente de s, tal que

$$\hat{\theta}_{\mathbf{e}}(s; \mathbf{y}) = \hat{\theta}_{\mathbf{e}}(s; \mathbf{y}')$$

e

$$\tilde{\theta}(s; \mathbf{y}) = \tilde{\theta}(s; \mathbf{y}').$$

Usando a hipótese b) concluímos que $\hat{\theta}_{\mathbf{e}}(s; \mathbf{y}') = \tilde{\theta}(s; \mathbf{y}')$ e portanto

$$\hat{\theta}_{\mathbf{e}}(s; \mathbf{y}) = \tilde{\theta}(s; \mathbf{y}) \text{ para todo o } s \in \cup_{j=0}^{m} Q_j.$$

2) Para $s \in Q_{m+1}$ sabemos que exatamente $m+1$ coordenadas k de s são tais que $y_k \neq e_k$. Estas coordenadas são necessariamente aquelas em que $\mathbf{y}$ e $\mathbf{e}$ diferem, isto é, não dependem da amostra s. Tal facto implica que $\hat{\theta}(s; \mathbf{y})$ e $\hat{\theta}(s; \mathbf{e})$ sejam constantes para $s \in Q_{m+1}$. Assim

$$\hat{\theta}_{\mathbf{e}}(s; \mathbf{y}) = c, \text{ para todo o } s \in Q_{m+1}. \tag{6.4.3}$$

3) Sendo $\tilde{\theta}(s; \mathbf{y})$ e $\hat{\theta}_{\mathbf{e}}(s; \mathbf{y})$ estimadores cêntricos de $\theta(\mathbf{y})$, de 1) temos

$$\begin{aligned} \mathrm{E}(\tilde{\theta}(\mathbf{y})) - \mathrm{E}(\hat{\theta}_{\mathbf{e}}(\mathbf{y})) &= \sum_{s \in Q} (\tilde{\theta}(s; \mathbf{y}) - \hat{\theta}_{\mathbf{e}}(s; \mathbf{y}))p(s) \\ &= \sum_{s \in Q_{m+1}} (\tilde{\theta}(s; \mathbf{y}) - \hat{\theta}_{\mathbf{e}}(s; \mathbf{y}))p(s) = 0. \end{aligned}$$

4) Da hipótese b) e mais uma vez de 1) obtemos

$$\begin{aligned} \mathrm{Var}(\tilde{\theta}(\mathbf{y})) - \mathrm{Var}(\hat{\theta}_{\mathbf{e}}(\mathbf{y})) &= \sum_{s \in Q} (\tilde{\theta}(s; \mathbf{y})^2 - \hat{\theta}_{\mathbf{e}}(s; \mathbf{y})^2)p(s) \\ &= \sum_{s \in Q_{m+1}} (\tilde{\theta}(s; \mathbf{y})^2 - \hat{\theta}_{\mathbf{e}}(s; \mathbf{y})^2)p(s) \leq 0. \end{aligned}$$

5) Usando 1)–4) podemos escrever

$$\begin{aligned}\sum_{s\in Q}&(\tilde{\theta}(s;\mathbf{y})-\hat{\theta}_{\mathbf{e}}(s;\mathbf{y}))^2p(s)\\ &=\sum_{s\in Q_{m+1}}(\tilde{\theta}(s;\mathbf{y})^2-\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})^2)p(s)\\ &\quad+2\sum_{s\in Q_{m+1}}\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})(\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})-\tilde{\theta}(s;\mathbf{y}))p(s)\\ &\leq 2c\sum_{s\in Q_{m+1}}(\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})-\tilde{\theta}(s;\mathbf{y}))p(s)=0.\end{aligned}$$

Concluímos assim que $\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})=\tilde{\theta}(s;\mathbf{y})$ para todo o $s\in Q$. Sendo $\mathbf{y}$ qualquer em Ω_{m+1} o resultado está provado. ■

Lema 6.4.4. *Dado um plano de amostragem sem reposição (p,Q) e $\hat{\theta}(s;\mathbf{y})$ um estimador cêntrico de $\theta(\mathbf{y})$, para todo o $\mathbf{e}\in\mathbb{R}^N$ o estimador*

$$\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})=\hat{\theta}(s;\mathbf{y})-\hat{\theta}(s;\mathbf{e})+\theta(\mathbf{e})$$

é admissível na classe dos estimadores cêntricos de $\theta(\mathbf{y})$.

Dem: Suponhamos por absurdo que existe um estimador cêntrico $\tilde{\theta}(s;\mathbf{y})$ de $\theta(\mathbf{y})$ com

$$\mathrm{EQM}(\tilde{\theta}(\mathbf{y}))\leq\mathrm{EQM}(\hat{\theta}_{\mathbf{e}}(\mathbf{y})),\ \text{para todo o } \mathbf{y}\in\mathbb{R}^N, \tag{6.4.5}$$

e

$$\mathrm{EQM}(\tilde{\theta}(\mathbf{y}_0))<\mathrm{EQM}(\hat{\theta}_{\mathbf{e}}(\mathbf{y}_0)),\ \text{para algum } \mathbf{y}_0\in\mathbb{R}^N. \tag{6.4.6}$$

Ora $\hat{\theta}_{\mathbf{e}}(s;\mathbf{e})=\theta(\mathbf{e})$ e $\mathrm{EQM}(\hat{\theta}_{\mathbf{e}}(\mathbf{e}))=0$ o que, por (6.4.5), implica que $\mathrm{EQM}(\tilde{\theta}(\mathbf{e}))=0$ e portanto $\hat{\theta}_{\mathbf{e}}(s;\mathbf{e})=\tilde{\theta}(s;\mathbf{e})$ para todo o $s\in Q$. As condições do Lema 6.4.2 são assim válidas para $m=0$ o que implica que $\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})=\tilde{\theta}(s;\mathbf{y})$ para todo o $s\in Q$ e $\mathbf{y}\in\Omega_1$. Aplicações sucessivas do Lema 6.4.2 permitem concluir que $\hat{\theta}_{\mathbf{e}}(s;\mathbf{y})=\tilde{\theta}(s;\mathbf{y})$ para todo o $s\in Q$ e $\mathbf{y}\in\mathbb{R}^N$, o que contradiz (6.4.6). ■

Recorrendo ao Lema 6.4.4 e tomando para $\hat{\theta}(s;\mathbf{y})$ o estimador de Narain-Horvitz--Thompson do total e $\mathbf{e}=0$, obtemos o resultado seguinte devido a Godambe e Joshi (1965).

Teorema 6.4.7. *Dado um plano de amostragem sem reposição (p,Q), o estimador de Narain-Horvitz-Thompson de t_y é admissível na classe dos estimadores cêntricos de t_y.*

6.5 Bibliografia

Basu, D. (1958). On sampling with and without replacement. *Sankhyā* 20, 287–294.

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā Ser. A* 31, 441–454.

Basu, D. (1971). An essay on the logical foundations of survey sampling I. In V.P. Godambe e D.S. Sprott (Eds.), *Foundations of Statistical Inference*, Hold, Rinehart and Winston, Toronto, 203–242.

Basu, D., Ghosh, J.K. (1967). Sufficient statistics in sampling from a finite universe. *Proceedings of the 36th Session of International Statistical Institute*, 850–859.

Cassel, C-M., Sarndal, C-E., Wretman, J.H. (1977). *Foundations of inference in survey sampling*. Springer. (Capítulos 2 e 3)

Gabler, S. (1981). A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement. *Biometrika* 68, 725–727.

Gabler, S. (1984). On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement. *Biometrika* 71, 171–175.

Godambe, V.P. (1955). A unified theory of sampling from finite population. *J. Roy. Statist. Soc. Ser. B* 17, 269–278.

Godambe, V.P. (1960). An admissible estimate for any sampling design. *Sankhyā Ser. A* 22, 285–288

Godambe, V.P., Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations. I. *Ann. Math. Statist.* 36, 1707–1722.

Hedayat, A.S., Sinha, B.K. (1991). *Design and inference in finite population sampling*. Wiley. (Capítulo 2)

Pathak, P.K. (1961). On the evaluation of moments of distinct units in a sample. *Sankhyā Ser. A* 23, 415–420.

Raj, D., Khamis, S.H. (1958). Some remarks on sampling with and without replacement. *Ann. Math. Statist.* 29, 550–557.

Roy, J., Chakravarti, I.M. (1960). Estimating the mean of a finite population. *Ann. Math. Statist.* 31, 392–398.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies*. Dunod. (Capítulo 3)

7

Amostragem por grupos a uma e a duas etapas

Amostragem por grupos a uma etapa. Seleção dos grupos com probabilidades iguais e eficiência relativamente ao plano SSR. Amostragem sistemática. Seleção dos grupos com probabilidades proporcionais ao seu tamanho. Amostragem por grupos a duas etapas. Seleção dos grupos com probabilidades iguais e com probabilidades desiguais.

7.1 Amostragem por grupos

Tal como na amostragem estratificada, assumimos neste capítulo que a população está dividida em subpopulações ou grupos. Contrariamente àquela, na amostragem por grupos apenas em alguns dos grupos são recolhidas amostras, sendo a seleção destes grupos feita por métodos aleatórios. Vamos então admitir que a população $\mathscr{U}$ de tamanho N está dividida em M grupos $\mathscr{U}_i$ de tamanhos $N_i, i = 1, \ldots, M$, tais que

$$\bigcup_{i=1}^{M} \mathscr{U}_i = \mathscr{U} \quad \text{e} \quad \mathscr{U}_i \cap \mathscr{U}_j = \emptyset, i \neq j.$$

Claramente $N = \sum_{i=1}^{M} N_i$. Assim, a população está dividida em unidades primárias (UP), sendo cada uma destas unidades composta de unidades secundárias (US). Quando observamos todas as unidades dos grupos selecionados dizemos que temos um plano de amostragem por grupos a uma etapa ou plano de amostragem por conglomerados. Se em cada um dos grupos selecionados voltarmos a selecionar uma amostra segundo um determinado plano de amostragem dizemos que temos um plano de amostragem por grupos a duas etapas. A amostragem por grupos pode naturalmente ser generalizada a mais de duas etapas.

Os planos de amostragem por grupos são particularmente úteis quando não se dispõe de uma base de amostragem de toda a população. Em cada passo do processo precisamos de conhecer os grupos em que dividimos a população (ou subpopulações) e apenas

precisamos de construir a base de amostragem relativamente aos grupos onde finalmente vamos selecionar as unidades da população. O facto de ser por vezes impossível na prática recolher uma amostra segundo planos simples ou estratificados, em que necessitamos de listar toda a população, é a razão apresentada por Neyman (1934, pp. 568–570) para considerar planos de amostragem em várias etapas.

Neste capítulo centraremos a nossa atenção nos planos de amostragem sem reposição, mas teoria análoga poderia ser desenvolvida para planos com reposição.

7.2 Amostragem por grupos a uma etapa

Um plano de amostragem p por grupos a uma etapa consiste na extração de uma amostra de grupos segundo um plano de amostragem sem reposição p^* definido na população

$$\mathscr{U}^* = \{1, \ldots, M\} \equiv \{\mathscr{U}_1, \ldots, \mathscr{U}_M\},$$

seguindo-se a observação de todas as unidades dos grupos selecionados. Representando por S e S^* as variáveis aleatórias com distribuições p e p^*, a amostra que resulta do plano de amostragem anterior é dada por

$$S = (\mathscr{U}_i, i \in S^*).$$

Pretendendo estimar o total da população t_y, sabemos que o estimador de NHT de t_y bem com a sua variância e respetivo estimador, dependem das probabilidades de inclusão de primeira e segunda ordens π_k e π_{kl} das unidades da população $\mathscr{U}$ segundo o plano de amostragem p anterior. Sendo π_i^* e π_{ij}^* as probabilidades de inclusão de primeira e segunda ordem das unidades de $\mathscr{U}^*$ segundo o plano p^* temos:

– Se $k \in \mathscr{U}_i$, então

$$\pi_k = \mathrm{P}(k \in S) = \mathrm{P}(i \in S^*) = \pi_i^*.$$

– Se $k, l \in \mathscr{U}_i$ com $k \neq l$, então

$$\pi_{kl} = \mathrm{P}(k, l \in S) = \mathrm{P}(i \in S^*) = \pi_i^*.$$

– Se $k \in \mathscr{U}_i$ e $l \in \mathscr{U}_j$ com $i \neq j$, então

$$\pi_{kl} = \mathrm{P}(k, l \in S) = \mathrm{P}(i, j \in S^*) = \pi_{ij}^*.$$

Tendo em conta que $\pi_{ii}^* = \pi_i^*$, então $\pi_{kl} = \pi_{ij}^*$ para todo o $k \in \mathscr{U}_i$ e $l \in \mathscr{U}_j$.

É interessante verificar que, em geral, este plano não satisfaz as condições de Sen-Yates-Grundy (a menos que $\pi_i^* = 1$ para todo o $i \in \mathscr{U}^*$), uma vez que se $k, l \in \mathscr{U}_i$, $k \neq l$, então

$$\pi_{kl} - \pi_k \pi_l = \pi_i^* - \pi_i^* \pi_i^* = \pi_i^*(1 - \pi_i^*).$$

Apesar disso, o estimador cêntrico da variância poderá ser, como veremos, não negativo.

No que se segue representamos por t_{yi} o total do grupo $\mathscr{U}_i$: $t_{yi} = \sum_{k\in\mathscr{U}_i} y_k$. Naturalmente $t_y = \sum_{i=1}^{M} t_{yi}$.

Teorema 7.2.1. *Num plano de amostragem por grupos a uma etapa, o estimador de NHT do total t_y é dado por*

$$\hat{t}_\pi = \sum_{i\in S^*} \frac{t_{yi}}{\pi_i^*}$$

e tem por variância

$$\mathrm{Var}(\hat{t}_\pi) = \sum_{i,j\in\mathscr{U}^*} \frac{t_{yi}t_{yj}}{\pi_i^*\pi_j^*}\,(\pi_{ij}^* - \pi_i^*\pi_j^*).$$

Além disso, se $\pi_{ij}^* > 0$ para todo o $i \neq j$,

$$\widehat{\mathrm{Var}}(\hat{t}_\pi) = \sum_{i,j\in S^*} \frac{t_{yi}t_{yj}}{\pi_i^*\pi_j^*}\,\frac{\pi_{ij}^* - \pi_i^*\pi_j^*}{\pi_{ij}^*}$$

é um estimador cêntrico de $\mathrm{Var}(\hat{t}_\pi)$.

Dem: De acordo com o Teorema 5.2.1 o estimador de NHT do total é dado por

$$\hat{t}_\pi = \sum_{k\in S} \frac{y_k}{\pi_k} = \sum_{i\in S^*}\sum_{k\in\mathscr{U}_i} \frac{y_k}{\pi_k} = \sum_{i\in S^*} \frac{1}{\pi_i^*} \sum_{k\in\mathscr{U}_i} y_k = \sum_{i\in S^*} \frac{t_{yi}}{\pi_i^*},$$

e a sua variância é dada por

$$\begin{aligned}
\mathrm{Var}(\hat{t}_\pi) &= \sum_{k,l\in\mathscr{U}} \frac{y_k y_l}{\pi_k\pi_l}\,(\pi_{kl} - \pi_k\pi_l) \\
&= \sum_{i,j\in\mathscr{U}^*} \sum_{k\in\mathscr{U}_i, l\in\mathscr{U}_j} \frac{y_k y_l}{\pi_k\pi_l}\,(\pi_{kl} - \pi_k\pi_l) \\
&= \sum_{i,j\in\mathscr{U}^*} \sum_{k\in\mathscr{U}_i, l\in\mathscr{U}_j} \frac{y_k y_l}{\pi_i^*\pi_j^*}\,(\pi_{ij}^* - \pi_i^*\pi_j^*) \\
&= \sum_{i,j\in\mathscr{U}^*} \frac{t_{yi}t_{yj}}{\pi_i^*\pi_j^*}\,(\pi_{ij}^* - \pi_i^*\pi_j^*),
\end{aligned}$$

pois $\pi_k = \pi_i$ e $\pi_{kl} = \pi_{ij}^*$ se $k \in \mathscr{U}_i$ e $l \in \mathscr{U}_j$.

Finalmente, se $\pi_{ij}^* > 0$, para todo o $i \neq j$, temos $\pi_{kl} > 0$ para todo o $k \neq l$, e pelo

Teorema 5.2.2 um estimador cêntrico de $\mathrm{Var}(\hat{t}_\pi)$ é dado por

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{t}_\pi) &= \sum_{k,l\in S} \frac{y_k y_l}{\pi_k \pi_l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \\ &= \sum_{i,j\in S^*} \sum_{k\in\mathscr{U}_i, l\in\mathscr{U}_j} \frac{y_k y_l}{\pi_k \pi_l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \\ &= \sum_{i,j\in S^*} \sum_{k\in\mathscr{U}_i, l\in\mathscr{U}_j} \frac{y_k y_l}{\pi^*_i \pi^*_j} \frac{\pi^*_{ij} - \pi^*_i \pi^*_j}{\pi^*_{ij}} \\ &= \sum_{i,j\in S^*} \frac{t_{yi} t_{yj}}{\pi^*_i \pi^*_j} \frac{\pi^*_{ij} - \pi^*_i \pi^*_j}{\pi^*_{ij}}. \end{aligned}$$

■

Reparemos que a partir do momento em que se verifica que o estimador de NHT depende apenas do plano de amostragem p^* e da variável $z_i = t_{yi}$, para $i \in \mathscr{U}^*$, as expressões anteriores para a variância do estimador de NHT e para um estimador desta, poderiam ser obtidas diretamente a partir dos Teoremas 5.2.1 e 5.2.2 quando aplicados ao plano p^* e à variável de interesse $z_i, i \in \mathscr{U}^*$. De forma análoga, podemos obter o resultado seguinte como consequência imediata do Teorema 5.4.1.

Teorema 7.2.2. *Se o plano p^* é de tamanho fixo, a variância do estimador de NHT toma a forma*

$$\mathrm{Var}(\hat{t}_\pi) = -\frac{1}{2} \sum_{i,j\in\mathscr{U}^*} \left(\frac{t_{yi}}{\pi^*_i} - \frac{t_{yj}}{\pi^*_j} \right)^2 (\pi^*_{ij} - \pi^*_i \pi^*_j),$$

e, se $\pi^*_{ij} > 0$ para todo o $i \neq j$, a variância anterior pode ser estimada de forma cêntrica por

$$\widehat{\mathrm{Var}}_{\mathrm{SYG}}(\hat{t}_\pi) = -\frac{1}{2} \sum_{i,j\in S^*} \left(\frac{t_{yi}}{\pi^*_i} - \frac{t_{yj}}{\pi^*_j} \right)^2 \frac{\pi^*_{ij} - \pi^*_i \pi^*_j}{\pi^*_{ij}}.$$

Reparemos que mesmo que p^* tenha tamanho fixo m, o plano p não é necessariamente de tamanho fixo, a menos que os grupos tenham todos o mesmo tamanho (ver Exercício 51). Se o plano de amostragem p^* é de tamanho fixo e satisfaz as condições de SYG, isto é, $\pi^*_{ij} \leq \pi^*_i \pi^*_j$ para todo o $i \neq j$, então o estimador da variância de SYG é não negativo.

Tendo em conta a expressão anterior da variância $\mathrm{Var}(\hat{t}_\pi)$ concluímos que para tirar partido de um plano de amostragem por grupos a uma etapa com probabilidades de inclusão desiguais onde p^* tem tamanho m, devemos conhecer os valores de uma variável auxiliar $x_i > 0$ aproximadamente proporcional a t_{yi},

$$t_{yi} \approx \lambda x_i,\ i \in \mathscr{U}^*,$$

e devemos implementar um plano p^* com probabilidades de inclusão de primeira ordem dadas por

$$\pi_i^* = mx_i/t_x,\ i \in \mathscr{U}^*.$$

É comum a situação em que a variável auxiliar x_i é o tamanho N_i do grupo $\mathscr{U}_i$, isto é, $x_i = N_i$, para $i \in \mathscr{U}^*$ (ver §7.2.4 e Exercício 58).

7.2.1 Seleção dos grupos com probabilidades iguais

Quando o plano de amostragem p^* anterior é um plano SSR de tamanho m dizemos que o plano de amostragem por grupos a uma etapa é simples. Neste caso as probabilidades de inclusão de primeira e segunda ordens são dadas por

$$\pi_i^* = \frac{m}{M} \quad \text{e} \quad \pi_{ij}^* = \frac{m}{M}\frac{m-1}{M-1},$$

para $i, j \in \mathscr{U}^* = \{1, \ldots, M\}$ com $i \neq j$. Não tendo os diversos grupos a mesma dimensão, o tamanho da amostra é aleatório de tamanho médio

$$\mathrm{E}(n(S)) = \sum_{i \in \mathscr{U}^*} N_i \frac{m}{M} = \frac{Nm}{M}.$$

No caso particular dos grupos terem todos o mesmo tamanho N_0, o plano é de tamanho fixo mN_0.

De acordo com o Teorema 7.2.1, o estimador de NHT toma a forma

$$\hat{t}_\pi = \frac{M}{m} \sum_{i \in S^*} t_{yi}. \tag{7.2.3}$$

Atendendo ao facto de p^* ser um plano SSR de tamanho m e $\hat{t}_\pi$ ser o estimador usual do total $t_y = \sum_{i \in \mathscr{U}^*} t_{yi}$, em que a variável de interesse é o total do grupo $\mathscr{U}_i$, t_{yi}, os resultados do parágrafo 2.4.1 podem ser usados para obter a variância de $\hat{t}_\pi$ bem como um estimador cêntrico desta (ver Exercício 52). Em alternativa, a variância de $\hat{t}_\pi$ pode ser obtida a partir do Teorema 7.2.2:

$$\mathrm{Var}(\hat{t}_\pi) = \frac{M-m}{2m(M-1)} \sum_{i,j \in \mathscr{U}^*} (t_{yi} - t_{yj})^2 = M^2\Big(1 - \frac{m}{M}\Big)\frac{s_t^2}{m}, \tag{7.2.4}$$

(sobre esta última igualdade, ver o Exercício 39) onde s_t^2 é a variância empírica corrigida da "população» dos totais" $t_{yi}, i = 1, \ldots, M$,

$$s_t^2 = \frac{1}{M-1} \sum_{i=1}^{M} \left(t_{yi} - \frac{1}{M} \sum_{i \in \mathscr{U}^*} t_{yi} \right)^2. \tag{7.2.5}$$

Novamente pelo Teorema 7.2.2, a variância anterior pode ser estimada de forma cêntrica por

$$\widehat{\mathrm{Var}}(\hat{t}_\pi) = \frac{M(M-m)}{2m^2(m-1)} \sum_{i,j \in S^*} (t_{yi} - t_{yj})^2 = M^2 \Big(1 - \frac{m}{M}\Big) \frac{\hat{s}_t^2}{m},$$

onde

$$\hat{s}_t^2 = \frac{1}{m-1} \sum_{i \in S^*} \left(t_{yi} - \frac{1}{m} \sum_{i \in S^*} t_{yi} \right)^2. \tag{7.2.6}$$

A expressão (7.2.4) anterior põe em evidência o facto do estimador anterior poder ter grande variabilidade quando os tamanhos dos diversos grupos são muito diferentes. Tal facto é consequência da grande variabilidade que, nesse caso, pode estar presente na "população dos totais" dos diferentes grupos.

7.2.2 Eficiência relativamente ao plano SSR

No contexto do parágrafo anterior, é possível comparar o plano por grupos a uma etapa com o plano SSR no caso particular dos grupos $\mathscr{U}_i$ terem todos o mesmo tamanho N_0, o que vamos admitir neste parágrafo. Antes de efetuarmos tal comparação, comecemos por reparar que a variância do estimador $\hat{t}_\pi$ definido por (7.2.3) depende exclusivamente da variância inter-grupos. Com efeito, tendo em conta que $N = N_0 M$, $N_i = N_0$ e $t_{yi} = N_0 \bar{y}_i$, para todo o i, $\bar{y} = \frac{1}{M} \sum_{i=1}^M \bar{y}_i$, com $\bar{y}_i = \frac{1}{N_0} \sum_{k \in \mathscr{U}_i} y_k$, de (7.2.5) temos

$$s_t^2 = \frac{1}{M-1} \sum_{i=1}^M (N_0 \bar{y}_i - N_0 \bar{y})^2 = \frac{N_0^2}{M-1} \sum_{i=1}^M (\bar{y}_i - \bar{y})^2 = \frac{N_0^2 M}{M-1} \sigma_{y,inter}^2.$$

Assim, de (7.2.4) obtemos

$$\mathrm{Var}(\hat{t}_\pi) = N^2 \frac{M-m}{M-1} \frac{\sigma_{y,inter}^2}{m}. \tag{7.2.7}$$

O resultado que apresentamos a seguir estabelece que quando os grupos são internamente bastante heterogéneos (variabilidade "intra" elevada, e, por consequência, variabilidade "inter" reduzida), a amostragem por grupos é mais eficiente que a amostragem simples sem reposição. Reparemos que esta situação é oposta à observada na amostragem estratificada.

Teorema 7.2.8. *Se os grupos $\mathscr{U}_i$ têm tamanho N_0, então*

$$\frac{\mathrm{Var}(\hat{t}_\pi)}{\mathrm{Var}(\hat{t}_{ssr})} = \frac{N-1}{M-1} \left(1 - \frac{\sigma_{y,intra}^2}{\sigma_y^2} \right),$$

onde $\hat{t}_{ssr}$ representa o estimador de NHT num plano SSR de tamanho $n = mN_0$.

Dem: Basta ter em conta (3.2.1), (7.2.7) e o facto da variância do estimador de NHT num plano SSR de tamanho $n = mN_0$ poder ser escrita na forma

$$\mathrm{Var}(\hat{t}_{ssr}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n} = N^2 \frac{M-m}{N-1} \frac{\sigma_y^2}{m}. \qquad \blacksquare$$

Corolário 7.2.9. *Se os grupos $\mathscr{U}_i$ têm tamanho N_0, o plano de amostragem simples por grupos a uma etapa de tamanho mN_0 é mais eficiente que o plano SSR de tamanho mN_0 sse*

$$\frac{1}{M} \sum_{i=1}^{M} s_{yi}^2 \geq s_y^2,$$

onde s_{yi}^2 a variância empírica corrigida do grupo $\mathscr{U}_i$

$$s_{yi}^2 = \frac{1}{N_i - 1} \sum_{k \in \mathscr{U}_i} (y_k - \bar{y}_i)^2. \tag{7.2.10}$$

Os resultados anteriores podem ser também reescritos em termos do coeficiente de correlação intragrupos definido por

$$\mathrm{CCI} = \frac{\sum_{i=1}^{M} \sum_{k,l \in \mathscr{U}_i : k \neq l} (y_k - \bar{y})(y_l - \bar{y})}{(N_0 - 1) \sum_{i=1}^{M} \sum_{k \in \mathscr{U}_i} (y_k - \bar{y})^2}.$$

Este coeficiente pode ser interpretado como uma medida da homogeneidade interna dos vários grupos. Atendendo a que

$$\begin{aligned} \sum_{i=1}^{M} \sum_{k,l \in \mathscr{U}_i} (y_k - \bar{y})(y_l - \bar{y}) &= \sum_{i=1}^{M} \left(\sum_{k \in \mathscr{U}_i} (y_k - \bar{y}) \right)^2 \\ &= \sum_{i=1}^{M} (N_i \bar{y}_i - N_i \bar{y})^2 \\ &= N_0^2 \sum_{i=1}^{M} (\bar{y}_i - \bar{y})^2 = N_0 N \sigma_{y,inter}^2, \end{aligned}$$

e que

$$\sum_{i=1}^{M} \sum_{k \in \mathscr{U}_i} (y_k - \bar{y})^2 = N \sigma_y^2,$$

concluímos que

$$\mathrm{CCI} = \frac{N_0 \sigma_{y,inter}^2 - \sigma_y^2}{(N_0 - 1)\sigma_y^2} = 1 - \frac{N_0}{N_0 - 1} \frac{\sigma_{y,intra}^2}{\sigma_y^2}. \tag{7.2.11}$$

Assim, se $\sigma^2_{y,intra} = 0$, isto é, se os grupos são internamente muito homogéneos, o coeficiente anterior toma o valor máximo CCI $= 1$. O valor mínimo de CCI é atingido quando $\sigma^2_{y,inter} = 0$ e é dado por CCI $= -1/(N_0 - 1)$.

Tendo em conta (7.2.7) e (7.2.11), podemos escrever a variância do estimador de NHT em termos do coeficiente de correlação intragrupos (expressão devida a Hansen e Hurwitz, 1942)

$$\mathrm{Var}(\hat{t}_\pi) = N^2 \frac{M-m}{M-1} \frac{\sigma_y^2}{mN_0}(1 + (N_0 - 1)\mathrm{CCI}). \tag{7.2.12}$$

Tal como vimos atrás, o estimador de NHT terá uma variância tão mais pequena quando menor for CCI, ou seja, quanto mais os grupos forem internamente heterogéneos.

Do resultado seguinte concluímos que a vantagem do plano de amostragem por grupos a uma etapa relativamente ao plano SSR de tamanho mN_0 será assim tanto maior quanto menor for o coeficiente CCI, ou seja, quanto mais os grupos forem heterogéneos.

Teorema 7.2.13. *Se os grupos $\mathscr{U}_i$ têm tamanho N_0, então*

$$\frac{\mathrm{Var}(\hat{t}_\pi)}{\mathrm{Var}(\hat{t}_{ssr})} = \frac{(N-1)M}{N(M-1)}(1 + (N_0 - 1)\mathrm{CCI}).$$

Corolário 7.2.14. *Se os grupos $\mathscr{U}_i$ têm tamanho N_0, o plano de amostragem simples por grupos a uma etapa de tamanho mN_0 é mais eficiente que o plano SSR de tamanho mN_0 sse*

$$\mathrm{CCI} \leq -\frac{1}{N-1}.$$

7.2.3 Amostragem sistemática

O método de amostragem sistemática para extrair uma amostra de tamanho n de uma população de tamanho N, em que $M = N/n$ é assumido inteiro, consiste na extração ao acaso de um inteiro h entre 1 e M e na seleção dos indivíduos colocados nas posições

$$h, h + M, h + 2M, \ldots, h + (n-1)M.$$

Tal como fazem notar Madow e Madow (1944, p. 4), amostragem sistemática pode assim ser vista como um caso particular da amostragem por grupos a uma etapa em que os grupos são dados por

$$\mathscr{U}_i = \{i, ih + M, i + 2M, \ldots, i + (n-1)M\},$$

para $i = 1, \ldots, M$, e apenas um dos grupos é escolhido. Neste caso, as probabilidades de inclusão de primeira ordem são iguais às dum plano SSR de tamanho n:

$$\pi_k = \frac{1}{M} = \frac{n}{N}, \text{ para todo o } k \in \mathscr{U}.$$

Como apenas um grupo é seleccionado, temos

$$\pi_{kl} = \frac{n}{N}, \text{ para todo o } k, l \in \mathscr{U}_i$$

e

$$\pi_{kl} = 0, \text{ para todo o } k \in \mathscr{U}_i, l \in \mathscr{U}_j \text{ com } i \neq j.$$

Tendo em conta (7.2.3) com $M = N/n$ e $m = 1$ o estimador de HT de t_y é dado por

$$\hat{t}_{sis} = M t_{yS^*},$$

uma vez que S^* é uma amostra de tamanho 1 sobre $\mathscr{U}^* = \{1, \ldots, M\}$. Tendo em conta (7.2.12), a sua variância é dada por

$$\mathrm{Var}(\hat{t}_{sis}) = N^2 \frac{\sigma_y^2}{n} (1 + (n-1)\mathrm{CCI}),$$

expressão esta devida a Madow e Madow (1944).

É interessante verificar que, a menos do fator $1 + (n-1)$CCI, a expressão anterior é igual à da variância do estimador do total num plano SCR e por isso semelhante à do plano SSR. Atendendo ao Teorema 7.2.13, concluímos que a eficiência do plano de amostragem sistemática relativamente ao plano SSR de tamanho n pode ser medida por

$$\frac{\mathrm{Var}(\hat{t}_{sis})}{\mathrm{Var}(\hat{t}_{ssr})} = \frac{N-n}{N-1}(1 + (n-1)\mathrm{CCI}) \approx 1 + (n-1)\mathrm{CCI},$$

sendo esta aproximação válida quando o tamanho da população é grande comparativamente ao tamanho da amostra.

Quando a lista das unidades população exibe algum padrão periódico podemos obter grupos homogéneos o que faz com que a variabilidade do estimador seja grande. Caso contrário, quando os possíveis grupos são heterogéneos a variabilidade do estimador pode ser inferior à obtida por SSR. Isto pode acontecer quando a população está ordenada relativamente à variável de interesse (ver Levy e Lemeshow, 1999, p. 92, para um exemplo desta situação).

7.2.4 Seleção dos grupos com probabilidades desiguais

Como vimos no parágrafo 7.2.1 a seleção dos grupos com probabilidades iguais pode conduzir a um estimador com grande variabilidade quando a "população dos totais" dos grupos apresenta grande variabilidade. Isso acontece muitas vezes em situações práticas quando os tamanhos dos grupos são muito diferentes entre si. Sendo comum a situação em que a variável de interesse é proporcional ao tamanho do grupo, caso o

tamanho dos grupos seja conhecido à partida podemos efetuar uma seleção dos grupos com probabilidades proporcionais ao seu tamanho através de um plano de tamanho fixo m sem reposição. Utilizando um tal plano de amostragem com probabilidades de inclusão desiguais, esperamos que o respetivo estimador de NHT tenha uma variabilidade inferior à do estimador usado no plano com iguais probabilidades de inclusão. Assim, as probabilidades de inclusão de primeira ordem são dadas por

$$\pi_i^* = m\frac{N_i}{N},\ i = 1, \ldots, M,$$

onde admitimos que $mN_i \leq N$ para todo o $i = 1, \ldots, M$ (reparar que isto corresponde a tomar como variável auxiliar $x_i = N_i$). Admitindo que não necessitamos de recalcular as probabilidades de inclusão anteriores (por alguma delas ser superior a 1), o estimador de NHT de t_y é dado por

$$\hat{t}_\pi = \frac{N}{m}\sum_{i\in S^*}\frac{t_{yi}}{N_i} = \frac{N}{m}\sum_{i\in S^*}\bar{y}_i.$$

A variância de $\hat{t}_\pi$ e um estimador desta podem ser obtidos a partir do Teorema 7.2.2, e dependem das probabilidades de inclusão de segunda ordem do plano sem reposição utilizado.

O tamanho da amostra é aleatório sendo o seu tamanho médio dado por

$$\mathrm{E}(n(S)) = \sum_{i\in\mathscr{U}^*} N_i\pi_i^* = \sum_{i\in\mathscr{U}^*} N_i\frac{mN_i}{N} = \frac{m}{N}\sum_{i\in\mathscr{U}^*} N_i^2.$$

Para um mesmo número de grupos selecionados este plano fornece amostras com tamanho médio superior às do plano com iguais probabilidades de seleção dos grupos.

7.3 Amostragem por grupos a 2 etapas

Tal como no plano de amostragem por grupos a uma etapa, a população $\mathscr{U}$ de tamanho N está dividida em M grupos $\mathscr{U}_i$ de tamanhos $N_i, i = 1, \ldots, M$. Um plano de amostragem p por grupos a duas etapas consiste na extração de uma amostra S^* de grupos segundo um plano de amostragem sem reposição p^* definido na população

$$\mathscr{U}^* = \{1, \ldots, M\} \equiv \{\mathscr{U}_1, \ldots, \mathscr{U}_M\},$$

seguindo-se a extração de amostras S^i em cada um dos grupos selecionados segundo planos de amostragem sem reposição p^i, $i = 1, \ldots, M$, em que os diversos planos são considerados independentes entre si. As unidades de $\mathscr{U}^*$ são ditas unidades primárias

e as unidades dos grupos $\mathscr{U}_i$, $i = 1, \ldots, M$, são ditas unidades secundárias. A amostra aleatória S que resulta da utilização do plano de amostragem p é assim dada por

$$S = (S^i, i \in S^*).$$

Pretendendo estimar o total da população t_y, sabemos que o estimador de NHT de t_y bem como a sua variância e o respetivo estimador, dependem das probabilidades de inclusão de primeira e segunda ordens π_k e π_{kl} das unidades da população $\mathscr{U}$ segundo o plano de amostragem p anterior. Sendo π_i^* e π_{ij}^* as probabilidades de inclusão de primeira e segunda ordem das unidades de $\mathscr{U}^*$ segundo o plano p^*, e por π_k^i e π_{kl}^i as probabilidades de inclusão de primeira e segunda ordem das unidades de $\mathscr{U}_i$ segundo o plano p^i, temos:

– Se $k \in \mathscr{U}_i$, então

$$\pi_k = \mathrm{P}(k \in S) = \mathrm{P}(i \in S^*, k \in S^i) = \mathrm{P}(i \in S^*)\mathrm{P}(k \in S^i) = \pi_i^* \pi_k^i.$$

– Se $k, l \in \mathscr{U}_i$, então

$$\pi_{kl} = \mathrm{P}(k, l \in S) = \mathrm{P}(i \in S^*, k, l \in S^i) = \mathrm{P}(i \in S^*)\mathrm{P}(k, l \in S^i) = \pi_i^* \pi_{kl}^i.$$

– Se $k \in \mathscr{U}_i$ e $l \in \mathscr{U}_j$, com $i \neq j$, então

$$\pi_{kl} = \mathrm{P}(i, j \in S^*, k \in S^i, l \in S^j) = \mathrm{P}(i, j \in S^*)\mathrm{P}(k \in S^i)\mathrm{P}(l \in S^j) = \pi_{ij}^* \pi_k^i \pi_l^j.$$

Na posse das probabilidades de inclusão anteriores, a partir de Teorema 5.2.1 podemos determinar o estimador de NHT de t_y, a sua variância e um estimador cêntrico desta sempre que $\pi_{ij}^* > 0$ e $\pi_{kl}^i > 0$ para todo o $i, j = 1, \ldots, M$ com $i \neq j$ e todo o $k, l \in U_i$ com $k \neq l$. No entanto, na demonstração do resultado seguinte usaremos uma técnica alternativa para deduzir a variância do estimador de NHT.

Teorema 7.3.1. *Num plano de amostragem por grupos a duas etapas, o estimador de NHT do total t_y é dado por*

$$\hat{t}_\pi = \sum_{i \in S^*} \frac{\hat{t}_{yi}}{\pi_i^*}$$

onde, para $i = 1, \ldots, M$,

$$\hat{t}_{yi} = \sum_{k \in S^i} \frac{y_k}{\pi_k^i}$$

é o estimador de HT do total t_{yi} do grupo $\mathscr{U}_i$ segundo o plano de amostragem p^i. Além disso,

$$\mathrm{Var}(\hat{t}_\pi) = V_{UP} + V_{US},$$

onde V_{UP} e V_{US} são os termos de variância relativos às unidades primárias e secundárias, respetivamente, dados por

$$V_{UP} = \sum_{i,j \in \mathscr{U}^*} \frac{t_{yi} t_{yj}}{\pi_i^* \pi_j^*} \, (\pi_{ij}^* - \pi_i^* \pi_j^*),$$

e

$$V_{US} = \sum_{i \in \mathscr{U}^*} \frac{1}{\pi_i^*} \operatorname{Var}(\hat{t}_{yi}),$$

onde

$$\operatorname{Var}(\hat{t}_{yi}) = \sum_{k,l \in \mathscr{U}_i} \frac{y_k y_l}{\pi_k^i \pi_l^i} \, (\pi_{kl}^i - \pi_k^i \pi_l^i).$$

Dem: De acordo com o Teorema 5.2.1, o estimador de NHT de t_y é dado por

$$\hat{t}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{i \in S^*} \sum_{k \in S^i} \frac{y_k}{\pi_k} = \sum_{i \in S^*} \sum_{k \in S^i} \frac{y_k}{\pi_i^* \pi_k^i} = \sum_{i \in S^*} \frac{1}{\pi_i^*} \sum_{k \in S^i} \frac{y_k}{\pi_k^i},$$

onde, ainda pelo Teorema 5.2.1, $\hat{t}_{yi} = \sum_{k \in S^i} y_k / \pi_k^i$ é o estimador de NHT do total t_{yi} do grupo $\mathscr{U}_i$ segundo o plano de amostragem p^i.

De modo a simplificar o cálculo da variância do estimador, vamos lançar mão da igualdade seguinte que permite decompor o cálculo da variância em duas fases:

$$\operatorname{Var}(\hat{t}_\pi) = \operatorname{Var}(\operatorname{E}(\hat{t}_\pi | S^*)) + \operatorname{E}(\operatorname{Var}(\hat{t}_\pi | S^*)).$$

Atendendo a que $\hat{t}_{yi}$ é o estimador de NHT do total t_{yi} do grupo $\mathscr{U}_i$ e que S^* e S^i são independentes temos

$$\begin{aligned} \operatorname{E}(\hat{t}_\pi | S^*) &= \operatorname{E}\Big(\sum_{i \in \mathscr{U}^*} \frac{\hat{t}_{yi}}{\pi_i^*} \, \mathbb{1}_i(S^*) \,\Big|\, S^* \Big) = \sum_{i \in \mathscr{U}^*} \frac{\operatorname{E}(\hat{t}_{yi} | S^*)}{\pi_i^*} \, \mathbb{1}_i(S^*) \\ &= \sum_{i \in S^*} \frac{\operatorname{E}(\hat{t}_{yi} | S^*)}{\pi_i^*} = \sum_{i \in S^*} \frac{\operatorname{E}(\hat{t}_{yi})}{\pi_i^*} = \sum_{i \in S^*} \frac{t_{yi}}{\pi_i^*}. \end{aligned}$$

Usando agora o Teorema 5.2.1 obtemos

$$\operatorname{Var}(\operatorname{E}(\hat{t}_\pi | S^*)) = \operatorname{Var}\Big(\sum_{i \in S^*} \frac{t_{yi}}{\pi_i^*} \Big) = \sum_{i,j \in \mathscr{U}^*} \frac{t_{yi} t_{yj}}{\pi_i^* \pi_j^*} \, (\pi_{ij}^* - \pi_i^* \pi_j^*) = V_{UP}.$$

Por outro lado, sendo os estimadores $\hat{t}_{yi}$ independentes entre si e também independentes de S^*, temos

$$\operatorname{Var}(\hat{t}_\pi | S^*) = \operatorname{Var}\Big(\sum_{i \in \mathscr{U}^*} \frac{\hat{t}_{yi}}{\pi_i^*} \, \mathbb{1}_i(S^*) \,\Big|\, S^* \Big) = \sum_{i \in \mathscr{U}^*} \frac{\operatorname{Var}(\hat{t}_{yi} | S^*)}{\pi_i^{*2}} \, \mathbb{1}_i(S^*) = \sum_{i \in S^*} \frac{\operatorname{Var}(\hat{t}_{yi})}{\pi_i^{*2}}.$$

Assim

$$\mathrm{E}(\mathrm{Var}(\hat{t}_\pi|S^*)) = \sum_{i\in\mathscr{U}^*} \frac{\mathrm{Var}(\hat{t}_{yi})}{\pi_i^*} = V_{US},$$

onde, pelo Teorema 5.2.1,

$$\mathrm{Var}(\hat{t}_{yi}) = \mathrm{Var}\left(\sum_{k\in S_i^*} \frac{y_k}{\pi_k^i}\right) = \sum_{k,l\in\mathscr{U}_i} \frac{y_k y_l}{\pi_k^i \pi_l^i}\,(\pi_{kl}^i - \pi_k^i\pi_l^i). \qquad \blacksquare$$

Como seria de esperar o estimador de NHT num plano a duas etapas possui uma variância maior que num plano a uma etapa. O termo V_{UP} é precisamente a variância do estimador de NHT no plano de amostragem por grupos a uma etapa. O acréscimo de variabilidade é quantificado pelo termo V_{US} resultante da variabilidade da segunda etapa do processo de amostragem.

O resultado anterior generaliza os obtidos para a amostragem estratificada e para a amostragem por grupos a uma etapa. Se todos os grupos forem selecionados na primeira etapa o plano reduz-se a uma amostragem estratificada. Neste caso $\pi^i = \pi_{ij}^* = 1$ e $V_{UP} = 0$. Por outro lado, se todas as unidades dos grupos selecionados na primeira etapa forem observadas, o plano reduz-se a um plano por grupos a uma etapa e neste caso $V_{US} = 0$.

Teorema 7.3.2. *Num plano de amostragem por grupos a duas etapas, se $\pi_{ij}^* > 0$ para todo o $i \neq j$, e $\pi_{kl}^i > 0$ para todo o $k \neq l$ e $i = 1, \ldots, M$, então um estimador cêntrico de $\mathrm{Var}(\hat{t}_\pi)$ é dado por*

$$\widehat{\mathrm{Var}}(\hat{t}_\pi) = \sum_{i,j\in S^*} \frac{\hat{t}_{yi}\hat{t}_{yj}}{\pi_i^*\pi_j^*}\,\frac{\pi_{ij}^* - \pi_i^*\pi_j^*}{\pi_{ij}^*} + \sum_{i\in S^*} \frac{1}{\pi_i^*}\widehat{\mathrm{Var}}(\hat{t}_{yi}),$$

onde

$$\widehat{\mathrm{Var}}(\hat{t}_{yi}) = \sum_{k,l\in S^i} \frac{y_k y_l}{\pi_k^i\pi_l^i}\,\frac{\pi_{kl}^i - \pi_k^i\pi_l^i}{\pi_{kl}^i}.$$

Dem: Representemos por $\hat{V}_A$ e $\hat{V}_B$ as primeira e segunda parcelas de $\widehat{\mathrm{Var}}(\hat{t}_\pi)$, respetivamente. Atendendo à independência entre S^i e S^j temos

$$\mathrm{E}(\hat{t}_{yi}\hat{t}_{yj}) = \mathrm{E}(\hat{t}_{yi})^2 = \mathrm{Var}(\hat{t}_{yi}) + t_{yi}^2, \text{ se } i = j,$$

e

$$\mathrm{E}(\hat{t}_{yi}\hat{t}_{yj}) = \mathrm{E}(\hat{t}_{yi})\mathrm{E}(\hat{t}_{yj}) = t_{yi}t_{yj}, \text{ se } i \neq j.$$

Assim,

$$\begin{aligned}
\mathrm{E}(\hat{V}_A) &= \sum_{i,j\in U^*} \frac{1}{\pi_i^*\pi_j^*}\,\frac{\pi_{ij}^*-\pi_i^*\pi_j^*}{\pi_{ij}^*}\mathrm{E}(\hat{t}_{yi}\hat{t}_{yj})\mathrm{E}(\mathbb{1}_i(S^*)\mathbb{1}_j(S^*)) \\
&= \sum_{i,j\in U^*} \frac{1}{\pi_i^*\pi_j^*}\,(\pi_{ij}^*-\pi_i^*\pi_j^*)\mathrm{E}(\hat{t}_{yi}\hat{t}_{yj}) \\
&= \sum_{i,j\in \mathscr{U}^*} \frac{t_{yi}t_{yj}}{\pi_i^*\pi_j^*}\,(\pi_{ij}^*-\pi_i^*\pi_j^*) + \sum_{i\in\mathscr{U}^*}\frac{\mathrm{Var}(\hat{t}_{yi})}{\pi_i^*}\,(1-\pi_i^*) \\
&= V_{UP} + V_{US} - \sum_{i\in\mathscr{U}^*}\mathrm{Var}(\hat{t}_{yi}).
\end{aligned}$$

Por outro lado, sendo $\widehat{\mathrm{Var}}(\hat{t}_{yi})$ um estimador cêntrico de $\mathrm{Var}(\hat{t}_{yi})$ que depende de S^i, temos

$$\mathrm{E}(\hat{V}_B) = \sum_{i\in U^*}\frac{1}{\pi_i^*}\,\mathrm{E}(\widehat{\mathrm{Var}}(\hat{t}_{yi})\mathbb{1}_i(S^*)) = \sum_{i\in U^*}\frac{1}{\pi_i^*}\,\mathrm{E}(\widehat{\mathrm{Var}}(\hat{t}_{yi}))\mathrm{E}(\mathbb{1}_i(S^*)) = \sum_{i\in\mathscr{U}^*}\mathrm{Var}(\hat{t}_{yi}).$$

Finalmente

$$\mathrm{E}(\widehat{\mathrm{Var}}(\hat{t}_\pi)) = \mathrm{E}(\hat{V}_A) + \mathrm{E}(\hat{V}_B) = V_{UP} + V_{US} = \mathrm{Var}(\hat{t}_\pi). \qquad \blacksquare$$

Como podemos concluir da demonstração do resultado anterior, as duas parcelas $\hat{V}_A$ e $\hat{V}_B$, que definem o estimador $\widehat{\mathrm{Var}}(\hat{t}_\pi)$, não são estimadores cêntricos de cada uma das componentes, V_{UP} e V_{US}, da variância de $\hat{t}_\pi$. Mais precisamente, provámos que $\hat{V}_A$ é um estimador enviesado de $\mathrm{Var}(\hat{t}_\pi)$ com viés dado por

$$\mathrm{E}(\hat{V}_A) - \mathrm{Var}(\hat{t}_\pi) = -\sum_{i\in\mathscr{U}^*}\mathrm{Var}(\hat{t}_{yi}),$$

permitindo o estimador $\hat{V}_B$ corrigir tal viés. Com efeito, vimos que

$$\mathrm{E}(\hat{V}_B) = \sum_{i\in\mathscr{U}^*}\mathrm{Var}(\hat{t}_{yi}),$$

o que faz com que $\widehat{\mathrm{Var}}(\hat{t}_\pi) = \hat{V}_A + \hat{V}_B$ seja um estimador cêntrico de $\mathrm{Var}(\hat{t}_\pi)$.

Teorema 7.3.3. *Se os planos p^* e p^i são de tamanho fixo, a variância do estimador de NHT toma a forma*

$$\mathrm{Var}(\hat{t}_\pi) = -\frac{1}{2}\sum_{i,j\in\mathscr{U}^*}\left(\frac{t_{yi}}{\pi_i^*}-\frac{t_{yj}}{\pi_j^*}\right)^2(\pi_{ij}^*-\pi_i^*\pi_j^*) + \sum_{i\in\mathscr{U}^*}\frac{1}{\pi_i^*}\mathrm{Var}(\hat{t}_{yi}),$$

onde

$$\mathrm{Var}(\hat{t}_{yi}) = -\frac{1}{2}\sum_{k,l\in\mathscr{U}_i}\left(\frac{y_k}{\pi_k^i} - \frac{y_l}{\pi_l^i}\right)^2(\pi_{kl}^i - \pi_k^i\pi_l^i).$$

Além disso, se $\pi_{ij}^* > 0$ *para todo o* $i \neq j$, *e* $\pi_{kl}^i > 0$ *para todo o* $k \neq l$ *e* $i = 1, \ldots, M$, *a variância anterior pode ser estimada de forma cêntrica por*

$$\widehat{\mathrm{Var}}_{\mathrm{SYG}}(\hat{t}_\pi) = -\frac{1}{2}\sum_{i,j\in S^*}\left(\frac{\hat{t}_{yi}}{\pi_i^*} - \frac{\hat{t}_{yj}}{\pi_j^*}\right)^2\frac{\pi_{ij}^* - \pi_i^*\pi_j^*}{\pi_{ij}^*} + \sum_{i\in S^*}\frac{1}{\pi_i^*}\widehat{\mathrm{Var}}_{\mathrm{SYG}}(\hat{t}_{yi}),$$

onde

$$\widehat{\mathrm{Var}}_{\mathrm{SYG}}(\hat{t}_{yi}) = -\frac{1}{2}\sum_{k,l\in S^i}\left(\frac{y_k}{\pi_k^i} - \frac{y_l}{\pi_l^i}\right)^2\frac{\pi_{kl}^i - \pi_k^i\pi_l^i}{\pi_{kl}^i}.$$

Reparemos que nas condições anteriores, o plano p não é de tamanho fixo (ver Exercício 59). Se os planos de amostragem p^* e p^i satisfazem as condições de SYG, isto é, $\pi_{ij}^* \leq \pi_i^*\pi_j^*$ para todo o $i \neq j$, e $\pi_{kl}^i \leq \pi_k^i\pi_l^i$ para todo o $k \neq l$ e $i = 1, \ldots, M$, então o estimador $\widehat{\mathrm{Var}}_{\mathrm{SYG}}(\hat{t}_\pi)$ é não-negativo.

Tendo em conta a expressão anterior da variância $\mathrm{Var}(\hat{t}_\pi)$ concluímos que para tirar partido de um plano de amostragem por grupos a duas etapa com probabilidades de inclusão desiguais onde p^* tem tamanho fixo m e p^i é de tamanho fixo n_i, para todas as unidades primárias $i \in \mathscr{U}^*$ devemos conhecer os valores de uma variável auxiliar $x_i > 0$ aproximadamente proporcional a t_{yi},

$$t_{yi} \approx \lambda x_i,\ i \in \mathscr{U}^*,$$

e devemos implementar um plano p^* com probabilidades de inclusão de primeira ordem dadas por

$$\pi_i^* = m x_i / t_x,\ i \in \mathscr{U}^*.$$

Além disso, para $i = 1, \ldots, M$ devemos também conhecer os valores de variáveis auxiliares $z_{ik} > 0$ aproximadamente proporcionais a y_k,

$$y_k \approx \lambda_i z_{ik},\ k \in \mathscr{U}_i,$$

e devemos implementar planos p^i com probabilidades de inclusão de primeira ordem dadas por

$$\pi_k^i = n_i z_{ik} / t_{z_i},\ k \in \mathscr{U}_i.$$

Atendendo a que

$$n(S) = \sum_{i\in\mathscr{U}^*} n(S^i)1_i(S^*) = \sum_{i\in\mathscr{U}^*} n(S^i)\mathbb{1}_i(S^*) = \sum_{i\in\mathscr{U}^*}\sum_{k\in U_i}\mathbb{1}_k(S^i)\mathbb{1}_i(S^*),$$

o tamanho médio do plano é dado por

$$\mathrm{E}(n(S)) = \sum_{i\in\mathscr{U}^*}\sum_{k\in U_i}\mathrm{E}(\mathbb{1}_k(S^i))\mathrm{E}(\mathbb{1}_i(S^*)) = \sum_{i\in\mathscr{U}^*}\pi_i^*\sum_{k\in U_i}\pi_k^i.$$

7.3.1 Seleção dos grupos com probabilidades iguais

Quando o plano de amostragem p anterior é um plano SSR de tamanho m e cada um dos planos p^i é um SSR de tamanho n_i dizemos que o plano de amostragem por grupos a duas etapas é simples. Neste caso temos

$$\pi_i^* = \frac{m}{M} \quad \text{e} \quad \pi_{ij}^* = \frac{m}{M}\frac{m-1}{M-1},$$

para $i,j \in \mathscr{U}^* = \{1,\ldots,M\}$ com $i \neq j$, e

$$\pi_k^i = \frac{n_i}{N_i} \quad \text{e} \quad \pi_{kl}^i = \frac{n_i}{N_i}\frac{n_i-1}{N_i-1},$$

para $k,l \in \mathscr{U}_i$ com $k \neq l$.

Pelo Teorema 7.3.1 o estimador de NHT é dado por

$$\hat{t}_\pi = \frac{M}{m}\sum_{i\in S^*}\hat{t}_{yi}, \tag{7.3.4}$$

onde

$$\hat{t}_{yi} = \frac{N_i}{n_i}\sum_{k\in S^i} y_k.$$

A seguir, deduziremos a variância do estimador de NHT usando diretamente o Teorema 7.3.3. Um procedimento alternativo poderá ser seguido usando os resultados, bem nossos conhecidos, sobre a variância do estimador de NHT num plano SSR (ver Exercício 60). Tendo então em conta o Teorema 7.3.3, a variância de $\hat{t}_\pi$ toma a forma

$$\begin{aligned}\mathrm{Var}(\hat{t}_\pi) &= \frac{M-m}{2m(M-1)}\sum_{i,j\in\mathscr{U}^*}(t_{yi}-t_{yj})^2 + \frac{M}{m}\sum_{i\in\mathscr{U}^*}\frac{N_i-n_i}{2n_i(N_i-1)}\sum_{k,l\in\mathscr{U}_i}(y_k-y_l)^2\\ &= M^2\left(1-\frac{m}{M}\right)\frac{s_t^2}{m} + \frac{M}{m}\sum_{i=1}^{M}N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{s_{yi}^2}{n_i},\end{aligned}$$

com s_t^2 e s_{yi}^2 dados por (7.2.5) e (7.2.10), respetivamente.

Ainda pelo Teorema 7.3.3 esta variância pode ser estimada de forma cêntrica por

$$\begin{aligned}\widehat{\mathrm{Var}}_{\mathrm{SYG}}(\hat{t}_\pi) &= \frac{M(M-m)}{2m^2(m-1)}\sum_{i,j\in S^*}(\hat{t}_{yi}-\hat{t}_{yj})^2 + \frac{M}{m}\sum_{i\in S^*}\frac{N_i(N_i-n_i)}{2n_i^2(n_i-1)}\sum_{k,l\in S^i}(y_k-y_l)^2\\ &= M^2\left(1-\frac{m}{M}\right)\frac{\hat{s}_{\hat{t}}^2}{m} + \frac{M}{m}\sum_{i\in S^*}N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{\hat{s}_{yi}^2}{n_i},\end{aligned}$$

com

$$\hat{s}_{\hat{t}}^2 = \frac{1}{m-1} \sum_{i \in S^*} \left(\hat{t}_{yi} - \frac{1}{m} \sum_{j \in S^*} \hat{t}_{yj} \right)^2$$

e

$$\hat{s}_{yi}^2 = \frac{1}{n_i - 1} \sum_{k \in S^i} (y_k - \hat{\bar{y}}_i)^2.$$

O plano simples a duas etapas possui, em geral, diferentes probabilidades de inclusão para unidades populacionais pertencentes a grupos distintos. Com efeito, a probabilidade de inclusão da unidade $k \in \mathscr{U}_i$ é dada por

$$\pi_k = \frac{m n_i}{M N_i}.$$

Além disso, e a menos que $n_1 = \cdots = n_M$, este é um plano de tamanho aleatório, com tamanho médio dado por

$$\mathrm{E}(n(S)) = \frac{m}{M} \sum_{i \in \mathscr{U}^*} n_i.$$

Pretendendo obter iguais probabilidades de inclusão (ou aproximadamente iguais), podemos tomar em cada grupo amostras de tamanho proporcional ao tamanho do grupo, isto é, tomar $\frac{n_i}{N_i} \approx C$, para $i = 1, \ldots, M$, com $0 < C < 1$. No entanto, e a menos que os grupos tenham iguais tamanhos, o plano resultante é de tamanho aleatório.

7.3.2 Seleção dos grupos com probabilidades desiguais

Os inconvenientes referidos para o plano a duas etapas quando a seleção dos grupos é feita com probabilidades iguais (probabilidades de inclusão diferentes e tamanho aleatório) podem ser ultrapassados se a seleção dos m grupos for feita através de um plano de tamanho fixo com probabilidades de inclusão proporcionais ao tamanho das unidades primárias e na segunda etapa as unidades sejam selecionadas segundo um plano SSR de tamanho fixo n_0 (plano de amostragem primeiramente considerado por Hansen e Hurwitz, 1943, pp. 338–340). Este plano é de tamanho fixo $m n_0$, com

$$\pi_i^* = m \frac{N_i}{N}$$

e

$$\pi_k^i = \frac{n_0}{N_i} \quad \text{e} \quad \pi_{kl}^i = \frac{n_0}{N_i} \frac{n_0 - 1}{N_i - 1}.$$

Assim, as probabilidades de inclusão são iguais para todas as unidades da população:

$$\pi_k = \pi_i^* \pi_k^i = \frac{m n_0}{N}.$$

7.4 Bibliografia

Cochran, W.G. (1977). *Sampling techniques.* Wiley. (Capítulos 9, 9A e 10)

Hansen, M.H., Hurwitz, W.N. (1942). Relative efficiencies of various sampling units in population inquiries. *J. Amer. Statist. Assoc.* 37, 89–94.

Hansen, M.H., Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.* 14, 333–362.

Hedayat, A.S., Sinha, B.K. (1991). *Design and inference in finite population sampling.* Wiley. (Capítulo 7)

Levy, P.S., Lemeshow, S. (1999). *Sampling of populations: methods and applications.* Wiley. (Capítulos 8 a 11)

Lohr, S.L. (1999). *Sampling: design and analysis.* Duxbury Press. (Capítulo 5)

Madow, W.G., Madow, L.H. (1944). On the theory of systematic sampling, I. *Ann. Math. Statist.* 15, 1–24.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Royal Statist. Soc.* 97, 558–625.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod. (Capítulo 9)

8

Não-resposta

O problema da não-resposta. Modelo determinístico de não-resposta. Tratamento da não-resposta através dum plano de amostragem em duas fases. Tratamento da não-resposta por reponderação dos respondentes.

8.1 O problema da não-resposta

Tal como em capítulos anteriores, continuamos interessados em estimar o total t_y ou a média $\bar{y}$ de uma variável y observada numa população finita $\mathscr{U}$. Como sabemos, uma tal estimação é realizada através da implementação de um plano de amostragem p sobre $\mathscr{U}$ e da observação da variável de interesse y para os indivíduos da amostra s gerada de acordo com o plano p. Contrariamente ao que assumimos até aqui, vamos agora admitir que, por razões diversas, não conseguimos observar o valor de y para certos indivíduos de s.

Estimar t_y ou $\bar{y}$ apenas com base na subamostra dos indivíduos para os quais conseguimos observar o valor de y, a que chamaremos respondentes, poderá conduzir a um enviesamento do estimador em particular quando os não-respondentes têm um comportamento diferente do dos respondentes relativamente à variável de interesse. Dizemos então que temos um problema de não-resposta. Para informação adicional sobre este problema, em particular sobre os tipos, causas e níveis de não-resposta, ver Lohr (1999, Cap. 8) e Tillé (2001, Cap. 13).

8.2 Modelo determinístico de não-resposta

Nesta secção estudamos alguns dos resultados iniciais sobre o tratamento da não-resposta. Admitimos que da população $\mathscr{U}$ de tamanho N extraímos uma amostra segundo um plano SSR de tamanho n, e que estamos interessados em estimar a média $\bar{y}$ duma variável de interesse y. Admitiremos também que a população está dividida em duas subpopulações que vamos representar por $\mathscr{U}_1$ e por $\mathscr{U}_0$, com $\mathscr{U} = \mathscr{U}_1 \cup \mathscr{U}_0$,

constituída pelos respondentes e pelos não-respondentes, respetivamente. Assim, observada a amostra S em $\mathscr{U}$, esta pode ser decomposta em duas subamostras S^1 e S^0 em que S^h representa as unidades de s que pertencem a $\mathscr{U}_h$ (podendo uma delas ser vazia). Sendo as unidades de S^0 não-respondentes, apenas conhecemos os valores y_k da variável y para as unidades $k \in S^1$. Uma vez que as unidades da amostra só são associadas a uma das duas subpopulações após a recolha da amostra S, a situação descrita é em tudo análoga ao esquema de pós-estratificação estudado no Capítulo 4, em que os pós-estratos são, no presente contexto, as subpopulações $\mathscr{U}_1$ e $\mathscr{U}_0$. Tal como aí referimos, os tamanhos n_h das subamostras S^h são variáveis aleatórias que sabemos possuírem uma distribuição hipergeométrica de parâmetros N, N_h e n, onde N_h o tamanho de $\mathscr{U}_h$, e que, condicionalmente aos tamanhos n_1 e n_0, as variáveis S^1 e S^0 são independentes e as suas distribuições são planos SSR de tamanhos n_h sobre $\mathscr{U}_h$.

8.2.1 O estimador baseado numa subamostra dos respondentes

Comecemos por analisar o efeito de considerar a média dos respondentes como estimador de $\bar{y}$, isto é, estimar $\bar{y}$ através de

$$\hat{\bar{y}}_1 = \begin{cases} \dfrac{1}{n_1} \displaystyle\sum_{k \in S^1} y_k, & n_1 > 0 \\ 0, & n_1 = 0. \end{cases}$$

Reparemos que este estimador é precisamente o estimador considerado no parágrafo 2.4.4 quando tratámos da estimação num domínio. No caso presente o papel do domínio é desempenhado pela subpopulação dos respondentes. Para $h = 0, 1$, vamos denotar por

$$\bar{y}_h = \frac{1}{N_h} \sum_{k \in \mathscr{U}_h} y_k$$

a média da variável y na subpopulação $\mathscr{U}_h$.

Teorema 8.2.1. *Nas condições anteriores, temos*

$$\mathrm{E}(\hat{\bar{y}}_1) = \bar{y}_1(1 - \mathrm{P}(n_1 = 0)),$$

e

$$\mathrm{Viés}(\hat{\bar{y}}_1) = \mathrm{E}(\hat{\bar{y}}_1) - \bar{y} = \frac{N_0}{N}(\bar{y}_1 - \bar{y}_0) - \bar{y}\mathrm{P}(n_1 = 0).$$

Dem: Atendendo a que

$$\hat{\bar{y}}_1 = \left(\frac{1}{n_1} \sum_{k \in S^1} y_k \right) \mathbb{1}(n_1 > 0),$$

temos

$$\mathrm{E}(\hat{\bar{y}}_1|n_0,n_1)=\mathrm{E}\Big(\frac{1}{n_1}\sum_{k\in S^1}y_k\Big|n_0,n_1\Big)\mathbb{1}(n_1>0)=\bar{y}_1\mathbb{1}(n_1>0), \tag{8.2.2}$$

uma vez que condicionalmente relativamente a n_0 e n_1, S^1 é um plano SSR de tamanho n_1 sobre $\mathscr{U}_1$ (ver Proposição 4.5.2). Finalmente,

$$\mathrm{E}(\hat{\bar{y}}_1)=\mathrm{E}(\mathrm{E}(\hat{\bar{y}}_1|n_0,n_1))=\bar{y}_1\mathrm{P}(n_1>0)=\bar{y}_1(1-\mathrm{P}(n_1=0)).$$ ■

Uma vez que os termos que incluem a probabilidade $\mathrm{P}(n_1=0)$ podem ser em geral desprezados visto que, de acordo com a desigualdade (4.5.9),

$$\mathrm{P}(n_1=0)\leq\exp\Big(-\frac{nN_1}{N}\Big),$$

concluímos da proposição anterior que o estimador $\hat{\bar{y}}_1$ é na realidade um estimador da média da variável y na subpopulação dos respondentes, apresentando um viés muito reduzido. O seu viés, como estimador de $\bar{y}$, pode ser elevado se a variável observada tiver uma média muito diferentes para respondentes e para não-respondentes. No entanto, se respondentes e não-respondentes apresentarem comportamentos semelhantes relativamente à variável y, será de esperar que o estimador $\hat{\bar{y}}_1$ seja um estimador com viés muito reduzido quando considerado como estimador de $\bar{y}$. Neste último caso fará sentido considerar $\hat{\bar{y}}_1$ como estimador de $\bar{y}$ sendo a sua variância dada no resultado seguinte.

Teorema 8.2.3. *A variância de $\hat{\bar{y}}_1$ é dada por*

$$\mathrm{Var}(\hat{\bar{y}}_1)=\Big(\mathrm{E}\Big(\frac{1}{n_1}\mathbb{1}(n_1>0)\Big)-\frac{1}{N_1}\Big)s^2_{y1}+V,$$

onde s^2_{y1} designa a variância corrigida de y na população dos respondentes

$$s^2_{y1}=\frac{1}{N_1-1}\sum_{k\in\mathscr{U}_1}(y_k-\bar{y}_1)^2$$

e

$$|V|\leq\big(s^2_{y1}/N_1+\bar{y}_1^2\big)\mathrm{P}(n_1=0).$$

Dem: Tendo em conta que

$$\mathrm{Var}(\mathrm{E}(\hat{\bar{y}}_1|n_0,n_1))=\bar{y}_1^2\mathrm{P}(n_1=0)(1-\mathrm{P}(n_1=0)),$$

e que

$$\begin{aligned}\mathrm{E}(\mathrm{Var}(\hat{\bar{y}}_1|n_0,n_1))&=\mathrm{E}\Big(\Big(1-\frac{n_1}{N_1}\Big)\frac{s^2_{y1}}{n_1}\mathbb{1}(n_1>0)\Big)\\&=\mathrm{E}\Big(\frac{1}{n_1}\mathbb{1}(n_1>0)\Big)-\frac{1}{N_1}\Big)s^2_{y1}-\frac{1}{N_1}\mathrm{P}(n_1=0),\end{aligned}$$

obtemos a expressão apresentada para a variância de $\hat{\bar{y}}_1$ onde

$$V = -s_{y1}^2 \mathrm{P}(n_1 = 0)/N_1 + \bar{y}_1^2 \mathrm{P}(n_1 = 0)(1 - \mathrm{P}(n_1 = 0)). \qquad \blacksquare$$

Atendendo a (4.5.14) sabemos que

$$\mathrm{E}\Big(\frac{1}{n_1}\mathbb{I}(n_1 > 0)\Big) \approx \frac{1}{\mathrm{E}(n_1)}\left(1 + \frac{1}{n}\frac{N_0}{N_1}\left(1 - \frac{n}{N}\right)\right),$$

o que permite obter a seguinte aproximação para variância de $\hat{\bar{y}}_1$:

$$\mathrm{Var}(\hat{\bar{y}}_1) \approx \left(1 - \frac{\mathrm{E}(n_1)}{N_1} - \frac{1}{n}\frac{N_0}{N_1}\left(1 - \frac{n}{N}\right)\right)\frac{s_{y1}^2}{\mathrm{E}(n_1)},$$

onde $\mathrm{E}(n_1) = nN_1/N$. Desprezando na fórmula anterior o termo de ordem n^{-2} obtemos a expressão

$$\mathrm{Var}(\hat{\bar{y}}_1) \approx \left(1 - \frac{\mathrm{E}(n_1)}{N_1}\right)\frac{s_{y1}^2}{\mathrm{E}(n_1)} = \frac{N}{N_1}\left(1 - \frac{n}{N}\right)\frac{s_{y1}^2}{n},$$

que é habitualmente usada para motivar o estimador de $\mathrm{Var}(\hat{\bar{y}}_1)$ dado por

$$\widehat{\mathrm{Var}}(\hat{\bar{y}}_1) = \frac{N}{N_1}\left(1 - \frac{n}{N}\right)\frac{\hat{s}_{y1}^2}{n},$$

onde

$$\hat{s}_{y1}^2 = \frac{1}{n_1 - 1}\sum_{k \in S^1}(y_k - \hat{\bar{y}}_1)^2.$$

Caso o tamanho da subpopulação dos respondentes não seja conhecido, em alternativa ao estimador anterior podemos usar

$$\widehat{\mathrm{Var}}(\hat{\bar{y}}_1) = \left(1 - \frac{n}{N}\right)\frac{\hat{s}_{y1}^2}{n_1}.$$

Relativamente ao tamanho da amostra para obter um intervalo de confiança para $\bar{y}$ (não esquecer que estamos a admitir que respondentes e não-respondentes apresentarem comportamentos semelhantes relativamente à variável y) com margem de erro inferior a um valor E fixado à partida, devemos tomar

$$n \geq \frac{N}{N_1}\frac{z_{1-\alpha/2}^2 s_{y1}^2}{E^2 + z_{1-\alpha/2}^2 s_{y1}^2/N_1}.$$

Assim, para obtermos a mesma precisão que o estimador da média num plano SSR de tamanho m é necessário recolher uma amostra com tamanho aproximadamente igual a mN/N_1.

8.2.2 Tratamento da não-resposta: o plano em duas fases

No caso em que respondentes e não-respondentes apresentam comportamentos diversos relativamente à variável de interesse y, vimos que o estimador de $\bar{y}$ baseado apenas nos respondentes é enviesado. Vejamos agora como podemos utilizar um plano de amostragem em duas fases para tratar o problema da não-resposta.

Vamos admitir que um número $n_* = n_*(S^0)$ de unidades S^* de S^0 são selecionadas segundo um plano SSR, onde n_* é uma variável aleatória que depende de S^0 e toma valores em $\{1, \ldots, n_0\}$ quando $n_0 > 0$. Seguindo a abordagem proposta por Hansen e Hurwitz (1946), vamos considerar o estimador de $\bar{y}$ definido por

$$\hat{\bar{y}} = w_1 \, \hat{\bar{y}}_1 + w_0 \, \hat{\bar{y}}_0^*,$$

onde

$$w_h = \frac{n_h}{n}$$

e

$$\hat{\bar{y}}_0^* = \begin{cases} \dfrac{1}{n^*} \displaystyle\sum_{k \in S^*} y_k, & n_0 > 0 \\ 0, & n_0 = 0. \end{cases}$$

Atendendo as que as variáveis n_h possuem distribuições hipergeométricas de parâmetros N, N_h e n, temos

$$\mathrm{E}(w_h) = \frac{N_h}{N} =: W_h$$

e

$$\mathrm{Var}(w_h) = \frac{g}{n} W_h (1 - W_h),$$

onde

$$g = \frac{N - n}{N - 1}.$$

A estrutura do estimador $\hat{\bar{y}}$ merece ser realçada. Admitindo que $n_1, n_0 > 0$, $\hat{\bar{y}}$ é dado por

$$\hat{\bar{y}} = \frac{1}{n} \sum_{k \in S^1} y_k + \frac{n_0}{n_* n} \sum_{k \in S^*} y_k,$$

o que põe em evidência o facto de cada unidade de S^1 representar N/n unidades da população, enquanto que cada unidade de S^* representa $n_0 N/(n_* n)$ unidades da população (sobre a ideia de ponderação amostral, ver §3.8). Isto é, as unidades de $\mathscr{U}^0$ nas quais pudemos observar a variável de interesse são sobrevalorizadas relativamente às restantes.

Os resultados seguintes são devidos a Rao (1973).

Teorema 8.2.4. *O estimador $\hat{\bar{y}}$ é um estimador cêntrico de $\bar{y}$,*

$$\mathrm{E}(\hat{\bar{y}}) = \bar{y}.$$

Dem: Raciocinando como em (8.2.2) obtemos

$$\begin{aligned}\mathrm{E}(\hat{\bar{y}}|n_0,n_1) &= w_1\mathrm{E}(\hat{\bar{y}}_1|n_0,n_1)\mathbb{1}(n_1>0) + w_0\mathrm{E}(\hat{\bar{y}}_0^*|n_0,n_1)\mathbb{1}(n_0>0)\\ &= w_1\bar{y}_1\mathbb{1}(n_1>0) + w_0\bar{y}_0\mathbb{1}(n_0>0),\end{aligned} \tag{8.2.5}$$

uma vez que condicionalmente relativamente a n_0, S^0 é um plano SSR de tamanho n_0 sobre $\mathscr{U}_0$ e assim S^* é um plano SSR de tamanho n_* sobre $\mathscr{U}_0$ (ver Exercício 15). Finalmente temos

$$\mathrm{E}(\hat{\bar{y}}) = \bar{y}_1\mathrm{E}(w_1\mathbb{1}(n_1>0)) + \bar{y}_0\mathrm{E}(w_0\mathbb{1}(n_0>0)) = \bar{y}_1W_1 + \bar{y}_0W_0 = \bar{y}. \qquad \blacksquare$$

Teorema 8.2.6. *A variância de $\hat{\bar{y}}$ é dada por*

$$\begin{aligned}\mathrm{Var}(\hat{\bar{y}}) &= \Big(\frac{W_1}{n} - \frac{\mathrm{E}(w_1^2)}{N_1}\Big)s_{y1}^2 + \Big(\mathrm{E}\Big(\frac{w_0^2}{n_*}\Big) - \frac{\mathrm{E}(w_0^2)}{N_0}\Big)s_{y0}^2\\ &\quad + \frac{g}{n}\big(W_1(\bar{y}_1-\bar{y})^2 + W_0(\bar{y}_0-\bar{y})^2\big)\\ &= \Big(1-\frac{n}{N}\Big)\frac{s_y^2}{n} + \frac{N_0}{N}(\nu-1)\frac{s_{y0}^2}{n},\end{aligned}$$

onde

$$\nu = \frac{N}{nN_0}\mathrm{E}\Big(\frac{n_0^2}{n_*}\Big) \geq 1$$

($\nu = 1$ quando $n_* = n_0$).

Dem: Tendo em conta (8.2.5) podemos escrever

$$\begin{aligned}\mathrm{Var}(\mathrm{E}(\hat{\bar{y}}|n_0,n_1)) &= \bar{y}_1\mathrm{Var}(w_1) + \bar{y}_0\mathrm{Var}(w_0) + 2\bar{y}_1\bar{y}_0\mathrm{Cov}(w_1,w_0)\\ &= \mathrm{Var}(w_1)(\bar{y}_1-\bar{y}_0)^2\\ &= \frac{g}{n}W_1W_0(\bar{y}_1-\bar{y}_0)^2\\ &= \frac{g}{n}\big(W_1(\bar{y}_1-\bar{y})^2 + W_0(\bar{y}_0-\bar{y})^2\big).\end{aligned} \tag{8.2.7}$$

Por outro lado, condicionalmente relativamente a n_0 e n_1, S^1 e S^0 são variáveis independentes cujas distribuições são planos SSR de tamanhos n_1 e n_0 sobre $\mathscr{U}_1$ e $\mathscr{U}_0$, respetivamente. Assim

$$\begin{aligned}\mathrm{Var}(\hat{\bar{y}}|n_0,n_1) &= w_1^2\mathrm{Var}(\hat{\bar{y}}_1|n_0,n_1) + w_0^2\mathrm{Var}(\hat{\bar{y}}_0^*|n_0,n_1)\\ &= w_1^2\Big(1-\frac{n_1}{N_1}\Big)\frac{s_{y1}^2}{n_1}\mathbb{1}(n_1>0) + w_0^2\Big(1-\frac{n_*}{N_0}\Big)\frac{s_{y0}^2}{n_*}\mathbb{1}(n_0>0)\\ &= \Big(\frac{w_1}{n} - \frac{w_1^2}{N_1}\Big)s_{y1}^2\mathbb{1}(n_1>0) + \Big(\frac{w_0^2}{n_*} - \frac{w_0^2}{N_0}\Big)s_{y0}^2\mathbb{1}(n_0>0),\end{aligned}$$

o que permite escrever

$$\mathrm{E}(\mathrm{Var}(\hat{\bar{y}}|n_0,n_1)) = \left(\frac{W_1}{n} - \frac{\mathrm{E}(w_1^2)}{N_1}\right)s_{y1}^2 + \left(\mathrm{E}\Big(\frac{w_0^2}{n_*}\Big) - \frac{\mathrm{E}(w_0^2)}{N_0}\right)s_{y0}^2. \tag{8.2.8}$$

De (8.2.7) e (8.2.8) obtemos a primeira expressão dada para a variância de $\hat{\bar{y}}$.

Da decomposição (3.2.1) sabemos que

$$W_1(\bar{y}_1-\bar{y})^2 + W_0(\bar{y}_0-\bar{y})^2 = \frac{N-1}{N}s_y^2 - (W_1-N^{-1})s_{y1}^2 - (W_0-N^{-1})s_{y0}^2,$$

o que permite concluir que

$$\begin{aligned}
&\frac{g}{n}\big(W_1(\bar{y}_1-\bar{y})^2 + W_0(\bar{y}_0-\bar{y})^2\big)\\
&= \frac{g}{n}\frac{N-1}{N}s_y^2 - \frac{g}{n}(W_1-N^{-1})s_{y1}^2 - \frac{g}{n}(W_0-N^{-1})s_{y0}^2\\
&= \Big(1-\frac{n}{N}\Big)\frac{s_y^2}{n} - \frac{g}{n}(W_1-N^{-1})s_{y1}^2 - \frac{g}{n}(W_0-N^{-1})s_{y0}^2,
\end{aligned}$$

ou ainda

$$\begin{aligned}
\mathrm{Var}(\hat{\bar{y}}) = &\Big(1-\frac{n}{N}\Big)\frac{s_y^2}{n}\\
&+\left(\frac{W_1}{n} - \frac{\mathrm{E}(w_1^2)}{N_1} - \frac{g}{n}(W_1-N^{-1})\right)s_{y1}^2\\
&+\left(\mathrm{E}\Big(\frac{w_0^2}{n_*}\Big) - \frac{\mathrm{E}(w_0^2)}{N_0} - \frac{g}{n}(W_0-N^{-1})\right)s_{y0}^2.
\end{aligned}$$

Para concluir a demonstração, basta agora usar a expressão anterior e atender a que

$$\frac{\mathrm{E}(w_h^2)}{N_h} + \frac{g}{n}(W_h-N^{-1}) = \frac{W_h}{n},$$

uma vez que $\mathrm{E}(w_h) = W_h$ e $\mathrm{Var}(w_h) = gW_h(1-W_h)/n$. ■

Tomando $n_* = \max(1,[fn_0])$, com $0 < f \leq 1$ escolhido pelo utilizador e onde $[x]$ representa o menor inteiro menor ou igual que x, e admitindo válida a aproximação $n_* \simeq fn_0$, obtemos

$$\nu = \frac{N}{nN_0}\mathrm{E}\Big(\frac{n_0^2}{n_*}\Big) \simeq \frac{N}{nfN_0}\mathrm{E}(n_0) = \frac{1}{f},$$

o que permite deduzir a seguinte aproximação para a variância de $\hat{\bar{y}}$

$$\mathrm{Var}(\hat{\bar{y}}) \simeq \Big(1-\frac{n}{N}\Big)\frac{s_y^2}{n} + \frac{N_0}{N}\Big(\frac{1}{f}-1\Big)\frac{s_{y0}^2}{n}.$$

Para estimar $\mathrm{Var}(\hat{\bar{y}})$ Rao (1973, Theorem 2, p. 126) propõe o estimador

$$\widehat{\mathrm{Var}}(\hat{\bar{y}}) = \Big(1-\frac{1}{N}\Big)\Big\{\Big(\frac{n_1-1}{n-1}-\frac{n_1-1}{N-1}\Big)\frac{w_1}{n_1}\hat{s}^2_{y1}+\Big(\frac{n_0-1}{n-1}-\frac{n_*-1}{N-1}\Big)\frac{w_0}{n_*}\hat{s}^2_{y0}\Big\}$$
$$+\frac{1}{n-1}\Big(1-\frac{n}{N}\Big)\Big\{w_1(\hat{\bar{y}}_1-\hat{\bar{y}})^2+w_0(\hat{\bar{y}}^*_0-\hat{\bar{y}})^2\Big\},$$

onde

$$\hat{s}^2_{y0}=\frac{1}{n_*-1}\sum_{k\in S^*}(y_k-\hat{\bar{y}}^*_0)^2.$$

8.3 Tratamento da não-resposta por reponderação dos respondentes

No modelo determinístico de não-resposta que estudámos na secção anterior, admitimos que a população estava dividida em duas subpopulações constituídas pelos respondentes e pelos não-respondentes. Isto significa que em sucessivas amostras recolhidas da população, os sujeitos pertencentes ao primeiro destes grupos respondem sempre, isto é, possuiam uma probabilidade 1 de responder, enquanto que os do segundo grupo nunca respondem, possuindo assim uma probabilidades 0 de responder.

Vamos agora admitir que o facto do indivíduo k responder ou não, pode alterar-se em sucessivas amostras, mesmo que hipotéticas, recolhidas da população $\mathscr{U}$. Assumimos assim que conhecemos a probabilidade da unidade $k\in\mathscr{U}$ responder, que vamos denotar por ϕ_k, com $\phi_k\in\,]0,1]$, e que o facto da unidade k responder é independente das restantes unidades da população responderem, ou não, e também independente da amostra S, isto é, do plano de amostragem considerado. Estamos assim a considerar a situação em que a probabilidade duma unidade responder, ou não, depende apenas dessa unidade. Sobre o plano de amostragem, assumimos ser um plano sem reposição geral, cujas probabilidade de inclusão de primeira e segunda ordens são, como habitualmente, denotadas por π_k e por π_{kl}, para $k,l\in\mathscr{U}$.

O fenómeno de não-resposta conduz a uma amostra R contida em S, dita amostra dos respondentes, e apenas para para as unidades $k\in R$ temos acesso ao valor y_k da variável de interesse y. Atendendo às hipótese anteriores, podemos calcular a probabilidade da unidade $k\in\mathscr{U}$ pertencer à amostra dos respondentes:

$$\mathrm{P}(k\in R)=\mathrm{P}(k\in S,k\,\text{responde})=\mathrm{P}(k\in S)\mathrm{P}(k\,\text{responde})=\pi_k\phi_k.$$

De forma análoga podemos obter a probabilidade das unidade $k,l\in\mathscr{U}$, com $k\neq l$ pertencerem à amostra dos respondentes:

$$\mathrm{P}(k,l\in R)=\mathrm{P}(k,l\in S,k\,\text{e}\,l\,\text{respondem})=\mathrm{P}(k,l\in S)\mathrm{P}(k\,\text{e}\,l\,\text{respondem})=\pi_{kl}\phi_k\phi_l.$$

Nas condições anteriores, demonstramos a seguir que, se a probabilidade ϕ_k é conhecida para toda a unidade da população, é possível construir um estimador cêntrico do total da população, dito estimador de t_y por reponderação dos respondentes. Isto é, para $k \in R$ a ponderação $1/\pi_k$ que surge no estimador de NHT é substituída pela ponderação $1/(\pi_k\phi_k)$. Como é de esperar, um tal estimador possui uma variabilidade amostral superior à do estimador de NHT.

Teorema 8.3.1. *O estimador*

$$\hat{t}_\phi = \begin{cases} \displaystyle\sum_{k\in R} \frac{y_k}{\pi_k\phi_k}, & n(R) > 0 \\ 0, & n(R) = 0, \end{cases}$$

dito estimador de t_y por reponderação dos respondentes, *é um estimador cêntrico de t_y, com variância dada por*

$$\mathrm{Var}(\hat{t}_\phi) = V + \sum_{k\in\mathscr{U}} \frac{1-\phi_k}{\phi_k}\frac{y_k^2}{\pi_k},$$

onde V é a variância do estimador de NHT de t_y na ausência de não-resposta. Além disso, se $\pi_{kl} > 0$ para todo o $k \neq l$, a variância anterior pode ser estimada de forma cêntrica por

$$\widehat{\mathrm{Var}}(\hat{t}_\phi) = \sum_{k,l\in R} \frac{y_k y_l}{\pi_k\pi_l\phi_k\phi_l}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} + \sum_{k\in R}\frac{1-\phi_k}{\phi_k}\frac{y_k^2}{\pi_k\phi_k}.$$

Dem: Tendo em conta que

$$\hat{t}_\phi = \sum_{k\in\mathscr{U}} \frac{y_k}{\pi_k\phi_k}\mathbb{I}_k(R),$$

deduzimos imediatamente que

$$\mathrm{E}(\hat{t}_\phi) = \sum_{k\in\mathscr{U}} \frac{y_k}{\pi_k\phi_k}\mathrm{P}(k\in R) = \sum_{k\in\mathscr{U}} y_k = t_y,$$

e que

$$\begin{aligned}\mathrm{Var}(\hat{t}_\phi) &= \sum_{k,l\in\mathscr{U}} \frac{y_k y_l}{\pi_k\phi_k\pi_l\phi_l}\mathrm{Cov}(\mathbb{I}_k(R),\mathbb{I}_l(R)) \\ &= \sum_{k\in\mathscr{U}} \frac{y_k^2}{\pi_k\phi_k}(1-\pi_k\phi_k) + \sum_{k,l\in\mathscr{U}:k\neq l}\frac{y_k y_l}{\pi_k\pi_l}(\pi_{kl}-\pi_k\pi_l) \\ &= \sum_{k,l\in\mathscr{U}} \frac{y_k y_l}{\pi_k\pi_l}(\pi_{kl}-\pi_k\pi_l) + \sum_{k\in\mathscr{U}}\frac{1-\phi_k}{\phi_k}\frac{y_k^2}{\pi_k}.\end{aligned}$$

■

Das hipóteses feitas sobre o mecanismo de não-resposta, o único problema consiste em estimar com base na amostra S, as probabilidade de resposta ϕ_k, o que é habitualmente feito através da utilização de um modelo (a este propósito, ver Tillé, 2001, Secção 13.5).

8.3.1 Modelo de não-resposta homogénea

Uma situação em que a estimação de tais probabilidade é simples é aquela em que admitimos que a probabilidade de resposta é a mesma para todas as unidades da população, ou seja, $\phi_k = \phi$, para todo o $k \in \mathscr{U}$, com $0 < \phi \leq 1$. Neste caso podemos estimar ϕ de forma cêntrica usando o estimador

$$\hat{\phi} = \frac{1}{N} \sum_{k \in R} \frac{1}{\pi_k},$$

o que dá origem ao estimador *plug-in* de t_y definido por

$$\hat{t}_{\hat{\phi}} = \begin{cases} \dfrac{1}{\hat{\phi}} \displaystyle\sum_{k \in R} \frac{y_k}{\pi_k}, & n(R) > 0 \\ 0, & n(R) = 0. \end{cases}$$

Menos óbvia é a dedução de expressões para o viés e variância deste estimador. O facto de $\hat{t}_{\hat{\phi}}$ ser um estimador de tipo racio dificulta essa tarefa. Reparemos também que mesmo no caso em que plano de amostragem que gera S é um plano SSR, os resultados estudados na Secção 2.4.3 não podem ser aqui diretamente usados uma vez que o estimador anterior depende da amostra R e o plano de amostragem que a gera não é um plano SSR (porquê?).

Vamos de seguida analisar o viés de $\hat{t}_{\hat{\phi}}$ quando S é gerada por um plano SSR. Neste caso, sendo $n_r = n(R)$ o tamanho da amostra dos respondentes, o estimador $\hat{t}_{\hat{\phi}}$ de ϕ reduz-se a

$$\hat{\phi} = \frac{n_r}{n}$$

e o estimador *plug-in* de t_y é dado por

$$\hat{t}_{\hat{\phi}} = \begin{cases} \dfrac{N}{n_r} \displaystyle\sum_{k \in R} y_k, & n_r > 0 \\ 0, & n_r = 0. \end{cases}$$

Proposição 8.3.2. *Se S é gerada por um plano SSR de tamanho n, então*

$$\mathrm{E}(\hat{t}_{\hat{\phi}}) = t_y(1 - (1 - \phi)^n).$$

Dem: Apesar da amostra R são ser gerada por um plano SSR de tamanho n_r sobre $\mathscr{U}$, é possível mostrar que condicionalmente a n_r R é efetivamente um plano SSR de tamanho n_r sobre $\mathscr{U}$ (ver Exercício 65.(d).iii). Assim,

$$\mathrm{E}(\hat{t}_{\hat{\phi}} | n_r) = E\left(\frac{N}{n_r} \sum_{k \in R} y_k \middle| n_r \right) \mathbb{1}(n_r > 0) = t_y \mathbb{1}(n_r > 0)$$

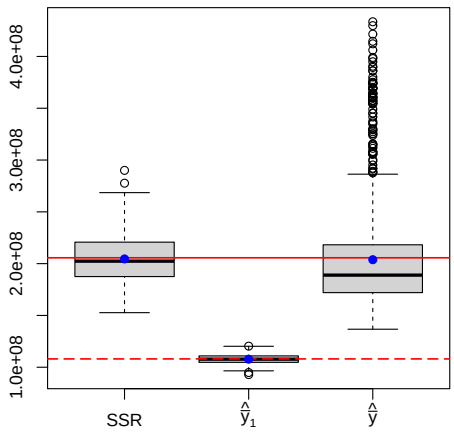
e

$$\mathrm{E}(\hat{t}_{\hat{\phi}}) = E\big(t_y \mathbb{1}(n_r > 0)\big) = t_y(1 - \mathrm{P}(n_r = 0)) = t_y(1 - (1 - \phi)^n). \qquad \blacksquare$$

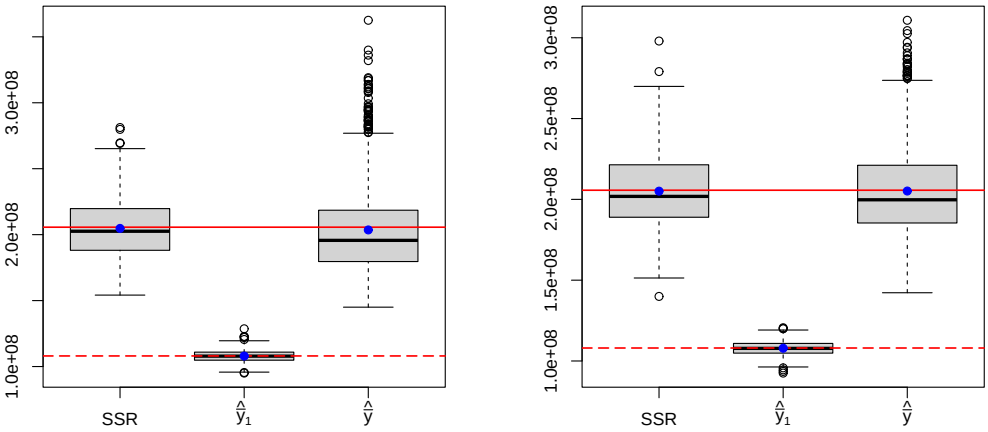
8.4 Alguns resultados de simulação

Para o modelo determinístico de não-resposta, consideramos nos exemplos seguintes a variável de interesse $y =$ belgianmunicipalities$TaxableIncome, e o parâmetro de interesse $\bar{y} = 205651072$. Como sabemos $N = 589$.

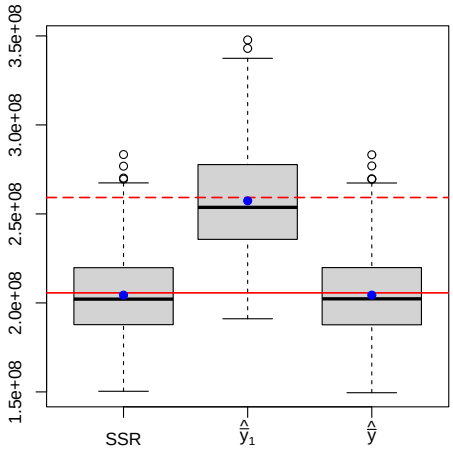
Exemplo 8.4.1. Supomos neste exemplo que os respondentes são definidos pela condição $y < quantile(y, probs = 0.75))$, isto é, não temos informação sobre o rendimento tributável dos municípios com 25% maiores rendimentos tributáveis. Neste caso, $N_1 = 441$ e $\bar{y}_1 = 108033842$. O gráfico seguinte, obtido usando 1000 repetições do processo de amostragem, compara o estimador do total num plano SSR de tamanho $n = 150$ em que todas as unidades da população são respondentes, com o estimador $\hat{\bar{y}}_1$, baseado da subamostra dos respondentes, e o estimador corrigido $\hat{\bar{y}}$ onde tomámos $n_* = \max(1, [fn_0])$, com $f = 0{,}2$.



Como podemos constatar, $\hat{\bar{y}}_1$ apresenta um forte viés como estimador de $\bar{y}$. Esse viés é corrigido pelo estimador $\hat{\bar{y}}$, que, no entanto, apresenta uma variabilidade amostral muito superior à do estimador SSR. Como se ilustra no gráfico seguinte, a situação melhora quando tomamos $f = 0{,}4$ (esquerda), e só melhora significativamente quando tomamos $f = 0{,}6$ (direita).



Exemplo 8.4.2. Supomos neste exemplo que os respondentes são definidos pela condição $y >$ quantile $(y,$ probs$=0.25)$, isto é, não temos informação sobre o rendimento tributável dos municípios com 25% menores rendimentos tributáveis. Neste caso, $N_1 = 441$ e $\bar{y}_1 = 259227154$. O gráfico seguinte, obtido usando 1000 repetições do processo de amostragem, compara o estimador do total num plano SSR de tamanho $n = 150$ em que todas as unidades da população são respondentes, com o estimador $\hat{\bar{y}}_1$, baseado da subamostra dos respondentes, e o estimador corrigido $\hat{\bar{y}}$ onde $n_* = \max(1, [fn_0])$, com $f = 0{,}2$.



Como podemos constatar, $\hat{\bar{y}}_1$ apresenta um forte viés como estimador de $\bar{y}$. Esse viés é corrigido pelo estimador $\hat{\bar{y}}$, que apresenta uma variabilidade amostral muito semelhante à do estimador SSR. Este bom comportamento de $\hat{\bar{y}}$ é consequência dos não respondentes serem os municípios com menores rendimentos tributáveis, e que, por isso, têm um menor impacto na média global $\bar{y}$.

8.5 Bibliografia

Hansen, M.H., Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *J. Amer. Statist. Assoc.* 41, 517–529.

Hedayat, A.S., Sinha, B.K. (1991). *Design and inference in finite population sampling.* Wiley. (§12.2).

Levy, P.S., Lemeshow, S. (1999). *Sampling of populations: methods and applications.* Wiley. (Capítulo 13)

Lohr, S.L. (1999). *Sampling: design and analysis.* Duxbury Press. (Capítulo 8)

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika* 60, 125–133.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod.

Bibliografia

Basu, D. (1958). On sampling with and without replacement. *Sankhyā* 20, 287–294.

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā Ser. A* 31, 441–454.

Basu, D. (1971). An essay on the logical foundations of survey sampling I. In V.P. Godambe and D.S. Sprott (Eds.), *Foundations of Statistical Inference*, Hold, Rinehart and Winston, Toronto, 203–242.

Basu, D., Ghosh, J.K. (1967). Sufficient statistics in sampling from a finite universe. *Proceedings of the 36th Session of International Statistical Institute*, 850–859.

Bellhouse, D.R. (1988). A brief history of random sampling methods. In P.R. Krishnaiah and C.R. Rao (Eds.), *Handbook of Statistics*, Vol. 6, Elsevier, 1–14.

Berger, Y. (1998a). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. Statist. Plann. Inference* 67, 209–226.

Berger, Y. (1998b). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *J. Statist. Plann. Inference* 74, 149–168.

Bowley, A.L. (1913). Working-class households in Reading. *J. Royal Statist. Soc.* 76, 672–701.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bull. Int. Statist. Inst.* (suppl.) 22, 6–62.

Cassel, C-M., Sarndal, C-E., Wretman, J.H. (1977). *Foundations of inference in survey sampling*. Springer.

Chaudhuri, A. (2014). *Modern survey sampling*. Taylor & Francis.

Chuprov, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* 2, 461–493, 646–683.

Cochran, W.G. (1946). Relative accuracy of systematic and random samples for a certain class of populations. *Ann. Math. Statist.* 17, 164–177.

Cochran, W.G. (1977). *Sampling techniques*. Wiley.

Cornfield, J. (1944). On samples from finite populations. *J. Amer. Statist. Assoc.* 39, 236–239.

Erdös, P., Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hung. Acad. Sci. Ser. A* 4, 49–61.

Gabler, S. (1981). A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement. *Biometrika* 68, 725–727.

Gabler, S. (1984). On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement. *Biometrika* 71, 171–175.

Godambe, V.P. (1955). A unified theory of sampling from finite population. *J. Roy. Statist. Soc. Ser. B* 17, 269–278.

Godambe, V.P. (1960). An admissible estimate for any sampling design. *Sankhyā Ser. A* 22, 285–288

Godambe, V.P., Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations. I. *Ann. Math. Statist.* 36, 1707–1722.

Gomes, P. (1998). Tópicos de sondagens. SPE.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci. Ser. A* 5, 361–374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* 35, 1491–1523.

Hansen, M.H., Hurwitz, W.N. (1942). Relative efficiencies of various sampling units in population inquiries. *J. Amer. Statist. Assoc.* 37, 89–94.

Hansen, M.H., Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.* 14, 333–362.

Hansen, M.H., Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *J. Amer. Statist. Assoc.* 41, 517–529.

Hedayat, A.S., Sinha, B.K. (1991). *Design and inference in finite population sampling.* Wiley.

Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47, 663–685.

Isserlis, L. (1915). On the conditions under which the "probable errors" of frequency distributions have a real significance. *Proceedings of the Royal Society of London. Series A* 92, 23–41.

Kiaer, A.N. (1905). Untitled speech with discussion. *Bull. Int. Statist. Inst.* 14, 119–134.

Kruskal, W., Mosteller, F. (1980). Representative sampling, IV: the history of the concept in statistics, 1895–1939. *Int. Stat. Rev.* 48, 169–195.

Lahiri, D.B (1951). A method of sample selection providing unbiased ratio estimates. *Bull. Int. Statist. Inst.* 33, Book 2, 133–140.

Laplace, P.S. (1814). *Théorie analytique des probabilités.* Paris: Ve. Courcier.

Levy, P.S., Lemeshow, S. (1999). *Sampling of populations: methods and applications.* Wiley.

Lohr, S.L. (1999). *Sampling: design and analysis.* Duxbury Press.

Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Ann. Math. Statist.* 19, 535–545.

Madow, W.G. (1949). On the theory of systematic sampling, II. *Ann. Math. Statist.* 20, 333–354.

Madow, W.G., Madow, L.H. (1944). On the theory of systematic sampling, I. *Ann. Math. Statist.* 15, 1–24.

McLeod, A.I., Bellhouse, D.R. (1983). A convenient algorithm for drawing a simple random sample. *J. R. Stat. Soc. Ser. C. Appl. Stat.* 32, 182–184.

Midzuno, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Statist. Math.* 3, 99–107.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *J. Indian Soc. Agricultural Statist.* 3, 169–175.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Royal Statist. Soc.* 97, 558–625.

Pathak, P.K. (1961). On the evaluation of moments of distinct units in a sample. *Sankhyā Ser. A* 23, 415–420.

Pathak, P.K. (1988). Simple random sampling. In P.R. Krishnaiah and C.R. Rao, eds., *Handbook of Statistics*, Vol 6 (Sampling), Elsevier 97–109.

Raj, D., Khamis, S.H. (1958). Some remarks on sampling with and without replacement. *Ann. Math. Statist.* 29, 550–557.

Rao, J.N.K. (1965). On two sample schemes of unequal probability sampling without replacement. *J. Indian Statist. Assoc.* 3, 173–180.

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika* 60, 125–133.

Roy, J., Chakravarti, I.M. (1960). Estimating the mean of a finite population. *Ann. Math. Statist.* 31, 392–398.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* 54, 499–513.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *J. Indian Soc. Agric. Statist.* 5, 119–127.

Singh, H.P., Solanki, R.S. (2013). An efficient class of estimators for the population mean using auxiliary information. *Comm. Statist. Theory Methods* 42, 245–163.

Sousa, M.F.C. (2002). *Amostragem, uma introdução.* Universidade Aberta.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *J. Amer. Statist. Assoc.* 43, 12–39.

Stuart, A. (1954). A simple presentation of optimum sampling results. *J. R. Stat. Soc. Ser. B* 16, 239–241.

Sukhatme, P.V. (1935). Contribution to the theory of the representative method. *Suppl. J. Royal Statist. Soc.* 2, 253–268.

Thompson, M.E. (2002). *Theory of sample surveys.* Wiley.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod.

Tillé, Y. (2006). *Sampling algorithms.* Springer.

Víšek, J.Á. (1979). Asymptotic distribution of simple estimate for rejective, Sampford and sucessive sampling. In: *Contribution to Statistics.* Jaroslav Hájek Memorial Volume, Jana Jureková (Ed.), D. Reidel Publishing Co., Dordrecht-Boston, Mass.-London, 263–275.

Yadav, S.K., Kadilar, C., Shabbir, J., Gupta, S. (2015). Improved family of estimators of population variance in simple random sampling. *J. Stat. Theory Pract.* 9, 219–226.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *J. Royal Statist. Soc.* 109, 12–30.

Yates, F., Grundy, P.M., (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc. Ser. B* 15, 235–261.

Índice Remissivo