Capítulo 4 Amostragem Aleatória Simples (AAS)

4.1 Planos amostrais e algoritmos de seleção

O planejamento da amostra de uma pesquisa requer a definição dos seguintes componentes fundamentais:

C1) Que método será usado para *seleção* da amostra;

C2) Qual será o *tamanho* da amostra;

C3) Que *estimadores* serão usados para os parâmetros de interesse; e

C4) Como será feita a *avaliação da precisão* das estimativas, isto é, como será feita a estimação da variância dos estimadores empregados.

Os componentes C1) e C2), em conjunto, dão origem à especificação do *plano amostral*. Já os componentes C3) e C4) dão origem à especificação dos *métodos de estimação* da pesquisa.

Para aplicar um plano amostral qualquer $p(s)$ (a função que associa a cada amostra possível s uma probabilidade de esta ser a amostra selecionada) precisamos contar com um *algoritmo de seleção*. Um *algoritmo de seleção* é um método que permite selecionar as unidades da amostra s de tal forma que a probabilidade de ser $s \in S$ a amostra selecionada seja igual a $p(s)$.

Há dois tipos principais de *algoritmos de seleção*: algoritmos baseados em *sequências de sorteios*, e algoritmos baseados em *processamento sequencial de listas* ou cadastros. Algoritmos baseados em sequências de sorteios são aplicados mediante realização de uma série de experimentos aleatórios, chamados *sorteios* ou extrações. Em cada *sorteio*, uma unidade é selecionada da população inteira ou de um subconjunto especificado da população, resultando em uma unidade selecionada para a amostra. Ao fim da série de sorteios, fica identificada a amostra selecionada para a pesquisa.

Algoritmos baseados em *sequências de sorteios* podem ser executados de duas formas distintas: os sorteios podem ser feitos de forma independente, o que implica que unidades já selecionadas podem eventualmente ser selecionadas mais vezes (*sorteios com reposição*), ou podem ser feitos de forma dependente ou condicional aos resultados dos sorteios antecedentes, geralmente para assegurar que unidades já selecionadas não possam ser selecionadas mais de uma vez (*sorteios sem reposição*).

Os algoritmos de seleção baseados em *processamento sequencial de listas* são aplicados mediante realização de uma série de experimentos aleatórios, executados sequencialmente para cada unidade do cadastro ou lista. Cada experimento vai resultar na inclusão ou exclusão dessa unidade da amostra s. Algumas vezes tais algoritmos podem terminar sem a necessidade de percorrer todo o cadastro ou lista.

Neste capítulo vamos ilustrar estes conceitos com um tipo de plano amostral bem simples. Mas as ideias básicas aqui introduzidas são aplicáveis de maneira geral a muitas outras situações de interesse.

4.2 Amostragem Aleatória Simples Com Reposição (AASC)

A *Amostragem Aleatória Simples Com Reposição* (*AASC*) é um plano amostral probabilístico básico, implementado por meio de um algoritmo de seleção no qual um número n prédeterminado de sorteios é feito, sendo em cada sorteio selecionada uma unidade da população (de tamanho N). Nesse plano, os sorteios são feitos de forma independente uns dos outros, isto é, *com reposição* das unidades na população antes da aplicação dos sorteios subsequentes. Cada sorteio é feito de tal forma que todas as unidades da população têm a mesma chance de ser incluídas na amostra do sorteio, e essa probabilidade é igual a $1/N$.

A forma usual de selecionar a amostra consiste em realizar n sorteios consecutivos, sendo cada seleção independente das anteriores. No primeiro passo, é selecionada a primeira unidade i_1 de U com probabilidade $1/N$. Esse processo é repetido $n-1$ vezes, sempre de forma independente, e são então selecionadas as unidades $i_2, \ldots, i_k, \ldots, i_n$ nos sorteios seguintes para compor a amostra. Cabe notar que as unidades já selecionadas podem ser repetidas na amostra. Em consequência, o número de amostras possíveis é N^n.

AASC é raramente usada na prática, pois é ineficiente em comparação com a amostragem *sem reposição* de igual tamanho inicial n, pelo fato de não incorporar nova informação quando a mesma unidade é selecionada mais de uma vez para a amostra. Na AASC o

tamanho efetivo da amostra é $m \leq n$, onde m designa o número de unidades *distintas* selecionadas.

Considere os dados amostrais para a variável y sob AASC, representados por $\{y_{i_1}, y_{i_2}, \ldots, y_{i_k}, \ldots, y_{i_n}\}$. Tais valores são observações de variáveis aleatórias $Y_1, Y_2, \ldots, Y_k, \ldots, Y_n$ independentes e identicamente distribuídas (IID), com distribuição comum dada conforme descrito na Tabela 4.1.

Tabela 4.1: Valores das variáveis aleatórias e probabilidades por unidade populacional

Unidade populacional (i)	**1**	**2**	$\ldots$	N	**Soma na linha**
Valores que Y_k pode assumir (y_i)	y_1	y_2	$\ldots$	y_N	Y
Probabilidades $[P(Y_k = y_i)]$	$1/N$	$1/N$	$\ldots$	$1/N$	1

4.2.1 Estimação do total e média populacionais sob AASC

Lembrando o princípio de estimação que está por trás do estimador tipo Horvitz-Thompson, que consiste em multiplicar cada valor observado na amostra por um peso igual ao inverso de sua probabilidade de inclusão, considere a variável aleatória igual a $Z_k = Y_k/(1/N) = N \times Y_k$. A cada sorteio de uma unidade para a amostra sob AASC, esta variável aleatória fornece um *estimador não viciado* para o total populacional Y:

$$E(Z_k) = E\left(N \times Y_k\right) = \sum_{i \in U}(N \times y_i) \times \frac{1}{N} = Y$$

É também fácil mostrar que a variância de $Z_k \ \forall \ k = 1, 2, \ldots, n$ é dada por:

$$V_{AASC}(Z_k) = V_{AASC}\left(N \times Y_k\right) = \sum_{i \in U}(N \times y_i - Y)^2 \times \frac{1}{N} = N^2 \times \sigma_y^2$$

onde $\sigma_y^2 = \dfrac{1}{N}\sum_{i \in U}(y_i - \overline{Y})^2$.

Com estes resultados, e considerando que sob AASC os dados são obtidos mediante a realização de n sorteios independentes e realizados em condições idênticas, tem-se que um *estimador não viciado* (ENV) para Y é dado por:

$$\widehat{Y}_{AASC} = \frac{1}{n}\sum_{k=1}^{n} Z_k = \frac{N}{n}\sum_{k=1}^{n} Y_k = \frac{N}{n}\sum_{k=1}^{n} y_{i_k} = \frac{N}{n}\sum_{i \in s} y_i = N \times \overline{y} \qquad (4.1)$$

onde $\overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_{i_k} = \frac{1}{n}\sum_{i \in s} y_i$ é a média dos valores observados na amostra.

Segue-se também que a variância deste estimador do total é dada por:

$$V_{AASC}(\widehat{Y}_{AASC}) = N^2 \times \sigma_y^2/n \tag{4.2}$$

A estimação dessa variância pode ser feita sem viés usando o estimador:

$$\widehat{V}_{AASC}(\widehat{Y}_{AASC}) = N^2 \times \widehat{\sigma}_y^2/n \tag{4.3}$$

onde $s_y^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \overline{y})^2$ é um ENV para a variância σ_y^2.

A estimação não viciada da média populacional $\overline{Y}$ pode ser feita dividindo o ENV do total Y por N, ou seja, $ Y_{AASC} / N$, o que resulta em usar o estimador média amostral simples $\overline{y}$ como estimador da média populacional.

A variância e estimador de variância correspondentes são também facilmente obtidos, uma vez que $V_{AASC}(\overline{y}) = V_{AASC}(\widehat{Y}_{AASC})/N^2$.

A Tabela 4.2 apresenta um resumo da estimação dos parâmetros média e total sob AASC.

Tabela 4.2: Estimadores dos parâmetros média e total sob AASC

Parâmetro	Estimador ENV sob AASC
$\overline{Y} = \sum_{i \in U} y_i/N$	$\overline{y} = \sum_{i \in s} y_i/n$
$Y = \sum_{i \in U} y_i$	$\widehat{Y}_{AASC} = N\,\overline{y}$
$\sigma_y^2 = \frac{1}{N}\sum_{i \in U}(y_i - \overline{Y})^2$	$s_y^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \overline{y})^2$
$V_{AASC}(\overline{y}) = \sigma_y^2/n$	$\widehat{V}_{AASC}(\overline{y}) = s_y^2/n$
$V_{AASC}(\widehat{Y}_{AASC}) = N^2\sigma_y^2/n$	$\widehat{V}_{AASC}(\widehat{Y}_{AASC}) = N^2 s_y^2/n$

Para provas destes resultados, veja por exemplo o Teorema 3.3 de Bolfarine e Bussab (2005).

Note que o estimador $\widehat{Y}_{AASC}$ para o total não é o estimador tipo Horvitz-Thompson para este plano amostral. Veja o Exercício 4.1 para uma discussão dessa questão.

A importância da AASC é principalmente teórica: através dela se mostra que é possível obter amostras de forma simples, cujos dados são utilizáveis mediante a aplicação de procedimentos convencionais da Inferência Estatística clássica. Por exemplo, a estimação não enviesada da média populacional $\overline{Y}$ pode ser feita simplesmente com o estimador média amostral $\overline{y}$, e a obtenção das propriedades deste estimador fica facilitada porque as variáveis aleatórias correspondentes aos valores das observações na amostra são IID, mesmo quando a população alvo tem tamanho finito. Ainda mais, o estimador $\overline{y}$ da média $\overline{Y}$ continua válido mesmo quando não se propõe um modelo estocástico para descrever a distribuição dos valores da população, sendo este estimador não viciado independente da forma que tem a distribuição dos valores da população. Tudo isso justifica a apresentação da AASC dentro do conjunto de técnicas abordadas neste livro.

Na prática, entretanto, é raro surgirem aplicações deste plano amostral. O motivo principal, como já indicado, é que AASC é ineficiente em comparação com a amostragem aleatória simples *sem reposição* de igual tamanho, como vamos mostrar na sequência.

4.3 Amostragem Aleatória Simples Sem Reposição (AAS)

A *Amostragem Aleatória Simples Sem Reposição* (*AAS*) é um plano amostral similar à *AASC*, sendo que neste caso cada unidade da população pode aparecer na amostra no máximo uma única vez, isto é, não pode haver repetição de unidades na amostra. Na verdade, *AAS* é qualquer procedimento de seleção que garanta que *todas* as amostras de tamanho n da população de tamanho N têm a *mesma probabilidade* de serem escolhidas. Como existem $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ amostras distintas em S, então $p(s) = 1/\binom{N}{n}\ \forall s \in S$, onde s é qualquer subconjunto de n inteiros distintos entre os inteiros de 1 a N.

Na *AAS* duas determinações da amostra são consideradas iguais quando constituídas das mesmas unidades da população, não importando a ordem de seleção dessas unidades.

A seleção da amostra pode ser feita realizando-se n sorteios consecutivos, de modo tal que em cada sorteio todas as unidades da população ainda não selecionadas têm igual chance de ser sorteadas, enquanto que as unidades já eventualmente selecionadas não mais participam do sorteio.

A *AAS* é um procedimento simples e básico da teoria e prática de amostragem, tendo importância não só pelas aplicações diretas como também por servir de base para muitos outros planos amostrais mais complexos. As ideias principais de amostragem podem ser com

ele desenvolvidas.

4.3.1 Algoritmo “convencional” para selecionar AAS

O algoritmo “convencional” para a seleção na AAS sugerido pelos livros-texto mais conhecidos em amostragem consiste nos seguintes passos:

1. Selecione a primeira unidade dentre as N unidades de U com probabilidades iguais a $1/N$, resultando na seleção da unidade i_1.
2. Selecione a segunda unidade dentre as $N-1$ unidades ainda não selecionadas de U com probabilidades iguais a $1/(N-1)$, resultando na seleção da unidade $i_2 \neq i_1$;

Prossiga com a seleção de uma unidade por sorteio, sempre excluindo de cada novo sorteio as unidades já selecionadas em sorteios anteriores, até a seleção da n-ésima unidade dentre as $N-n+1$ unidades de U que permanecem não selecionadas após $n-1$ sorteios, com probabilidades iguais a $1/(N-n+1)$, resultando na seleção da unidade $i_n \neq \ldots \neq i_2 \neq i_1$..

Tal algoritmo fornecia a regra para *seleção manual* com uso de *tabelas de números aleatórios* antes do aparecimento e uso de computadores para seleção de amostras. A aplicação deste algoritmo em computador é bastante ineficiente, devido à necessidade de percorrer várias vezes uma lista que pode ser bem grande: a das unidades ainda não selecionadas. Para enfrentar esta dificuldade, foram propostos diversos algoritmos mais eficientes na literatura. Apresentaremos aqui apenas dois, ambos baseados em processamento sequencial de listas, que se destacam por sua simplicidade.

4.3.2 Algoritmo de Hàjek para selecionar AAS

Este algoritmo foi proposto por Hájek (1960) e consiste nos seguintes passos:

Passo 1: Para cada $i \in U$, associe um *número pseudoaleatório* a_i, onde os a_i são determinações de variáveis aleatórias IID $A_1, A_2, \ldots, A_N$, todas com distribuição $U[0;1]$, conforme apresentado na Tabela 4.3.

Tabela 4.3: Número pseudoaleatório associado a cada unidade da população

Rótulo da unidade i	**1**	**2**	$\ldots$	N
Número pseudoaleatório (a_i)	a_1	a_2	$\ldots$	a_N

Passo 2: Ordene a população segundo os números pseudoaleatórios $a_1, a_2, \ldots, a_N$, obtendo uma *permutação aleatória* dos rótulos das unidades populacionais, conforme apresentado na Tabela 4.4.

Tabela 4.4: Número pseudoaleatório ordenado associado ao rótulo permutado de cada unidade

Rótulo permutado da unidade i	i_1	i_2	$\ldots$	i_N
Número pseudoaleatório ordenado $a_{(i)}$	$a_{(1)}$	$a_{(2)}$	$\ldots$	$a_{(N)}$

A notação $a_{(i)}$ indica o valor posicionado na i-ésima posição na sequência ordenada dos valores dos números pseudoaleatórios, isto é, corresponde à i-ésima estatística de ordem da sequência $a_1, a_2, \ldots, a_N$.

Passo 3: Para selecionar uma amostra de tamanho n, inclua na amostra uma sequência de n rótulos consecutivos quaisquer, na ordem em que aparecem nesta permutação.

Por exemplo, os rótulos $\{i_1, i_2, \ldots, i_n\}$ fornecem uma AAS.

Outro exemplo: os rótulos $\{i_{N-n+1}, i_{N-n+2}, \ldots, i_N\}$ também fornecem uma AAS de tamanho n de U.

O algoritmo de Hàjek requer duas passagens sobre a lista, mais uma operação de ordenação dos números aleatórios. Oferece grande ganho de eficiência em comparação com o algoritmo convencional, mas ainda não é o mais eficiente.

4.3.3 Algoritmo de Fan, Muller e Rezucha para selecionar AAS

Este algoritmo foi proposto por Fan, Muller, e Rezucha (1962) e consiste nos seguintes passos:

Como no algoritmo anterior, sejam a_i, $i = 1, 2, \ldots, m$, determinações de variáveis aleatórias IID $A_1, A_2, \ldots, A_m$, todas com distribuição $U[0; 1]$.

Processe sequencialmente a lista, para as unidades $i = 1, 2, 3, \ldots, m$, incluindo na amostra as unidades i tais que $a_i < \frac{n - k_{i-1}}{N - i + 1}$, onde k_{i-1} é o número de unidades selecionadas até o processamento da unidade $i - 1$. Após processar cada unidade da lista, atualize o número de unidades já incluídas na amostra. Interrompa o processamento quando $k_m = n$, o que ocorre

quando for processada a unidade de ordem m. Note que m é uma quantidade aleatória, que vai variar no intervalo $n \leq m \leq N$, e corresponde ao número de sorteios necessários para conseguir uma AAS de tamanho n da população de tamanho N.

Este algoritmo é muito eficiente em comparação com os anteriores, porque requer processar a lista no máximo uma vez, e pode nem mesmo requerer chegar ao fim da lista: a amostra pode ser selecionada por completo bem antes de chegar ao final da lista. Apesar de sua simplicidade e eficiência, há alternativas ainda mais eficientes, que entretanto não serão discutidas aqui. Caso o leitor necessite implementar um algoritmo para seleção de uma AAS, recomendamos o emprego deste algoritmo. Será suficientemente bom para a maioria das aplicações práticas.

4.3.4 Probabilidades de inclusão sob AAS

Como já indicado no capítulo 3, tratar com as distribuições de aleatorização $p(s)$ sob AAS pode ser complicado do ponto de vista prático. Särndal, Swensson, e Wretman (1992), p.29, mencionam que numa população com $N = 1.000$ unidades, o conjunto S de amostras AAS possíveis de tamanho $n = 40$ tem dimensão $\binom{N}{n} = \binom{1.000}{40} = 5,6 \times 10^{71}$. Se a população tivesse $N = 5.000$ e a amostra tamanho $n = 200$, a dimensão de S cresceria para $\binom{5.000}{200} = 1,4 \times 10^{363}$. Portanto, a enumeração de todas as amostras possíveis seria tarefa complicada, mesmo com computadores poderosos.

Note que os tamanhos de população e amostra acima são modestos do ponto de vista de aplicações práticas. Foi para eliminar essa dificuldade que introduzimos resumos simples derivados da distribuição $p(s)$. Tais resumos serão suficientes para a obtenção de propriedades de estimadores tais como valor esperado e variância, na maioria das situações de interesse prático. Esses resumos são as *probabilidades de inclusão* de unidades ou de pares de unidades na amostra.

1. Sob AAS, $\pi_i = n/N > 0, \forall i \in U$ desde que $n > 0$.
2. $f = n/N$ é chamada de *fração amostral* ou *taxa de amostragem*.
3. Estimação de variância sem vício requer $\pi_{ij} > 0, \forall i, j \in U$. Sob AAS,
   $$\pi_{ij} = \frac{n(n-1)}{N(N-1)} > 0 \;\; \forall i \neq j \in U.$$
4. Sob AAS, as probabilidades de inclusão π_i, π_{ij}, etc. não dependem de i ou j, e essa é a razão da simplicidade desse plano amostral.

Sob AAS de tamanho n de população com N, para a variável δ_i indicadora do evento ´inclusão da unidade i na amostra s´, tem-se:

$E_{AAS}[\delta_i] = \frac{n}{N}$,

$V_{AAS}[\delta_i] = \frac{n}{N}\left(1 - \frac{n}{N}\right)$ e

$COV_{AAS}[\delta_i, \delta_j] = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(1 - \frac{n}{N}\right)\left(-\frac{1}{N-1}\right)$

Assim sob AAS a correlação entre duas variáveis indicadoras de inclusão de unidades distintas na amostra é dada por:

$CORR_{AAS}[\,\delta_i\,, \delta_j] = -1/(N-1)$ se $i \neq j$

4.3.5 Estimador não viciado do total e média populacionais sob AAS

Usando o estimador não viciado de total tipo Horvitz-Thompson, e substituindo os valores das probabilidades de inclusão de primeira ordem, obtém-se:

$$\widehat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} \frac{y_i}{n/N} = N\frac{1}{n}\sum_{i \in s} y_i = N\overline{y} = \widehat{Y}_{AAS} \tag{4.4}$$

Este é um *ENV* do *total populacional* Y e em consequência, $\overline{y} = \frac{1}{n}\sum_{i \in s} y_i$ é *ENV* da *média populacional* $\overline{Y}$ sob AAS. A variância do estimador do total sob AAS é dada por:

$$V_{AAS}(\widehat{Y}_{AAS}) = N^2\left(1 - \frac{n}{N}\right)\frac{S_y^2}{n} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 \tag{4.5}$$

Um ENV da variância do estimador de total é dado por:

$$\widehat{V}_{AAS}(\widehat{Y}_{AAS}) = N^2\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n} = N^2\left(\frac{1}{n} - \frac{1}{N}\right)s_y^2 \tag{4.6}$$

onde $s_y^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \overline{y})^2$, como já definido.

A Tabela 4.5 apresenta um resumo da estimação de parâmetros média e total sob AAS.

Tabela 4.5: Estimadores dos parâmetros média e total sob AAS

Parâmetro	Estimador ENV sob AAS
$\overline{Y} = \sum_{i \in U} y_i / N$	$\overline{y} = \sum_{i \in s} y_i / n$
$Y = \sum_{i \in U} y_i$	$\widehat{Y}_{AAS} = N\overline{y} = \frac{N}{n} \sum_{i \in s} y_i$
$S_y^2 = \frac{1}{N-1} \sum_{i \in U} \left(y_i - \overline{Y}\right)^2$	$s_y^2 = \frac{1}{n-1} \sum_{i \in s} \left(y_i - \overline{y}\right)^2$
$V_{AAS}(\overline{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$	$\widehat{V}_{AAS}(\overline{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2$
$V_{AAS}(\widehat{Y}_{AAS}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$	$\widehat{V}_{AAS}(\widehat{Y}_{AAS}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2$

Notas

1. O termo $(1 - n/N) = (1 - f)$ é chamado de *fator de correção de população finita*. Quando $n/N \to 1$ então $(1 - n/N) \to 0$.
2. Se a fração amostral $f = n/N$ for pequena (tipo menor que 1% ou 2%), então a *correção de população finita* é desprezível, pois $(1 - f) \doteq 1$.
3. Quando $f \doteq 0$, a AAS (amostragem sem reposição) se comporta como se fosse AASC (com reposição).

4.3.6 Distribuição da média amostral

Sob repetições do procedimento de seleção segundo *AAS*, $\overline{y}$ tem uma distribuição de probabilidades. A distribuição exata de $\overline{y}$ depende da distribuição dos y´s na população, do tamanho da amostra n e do plano amostral $p(s)$, que neste caso, é AAS. Isto resulta numa situação complicada, que pode ser resolvida considerando a *Distribuição Assintótica da Média Amostral*.

Se n for grande e $f = n/N$ for pequena, o *Teorema Central do Limite* - ver Hájek (1960) - pode ser usado para obter a distribuição aproximada:

$$\frac{\overline{y} - E_{AAS}(\overline{y})}{\sqrt{V_{AAS}(\overline{y})}} = \frac{\overline{y} - \overline{Y}}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}} \approx N(0;1) \tag{4.7}$$

onde $N(0;1)$ denota uma variável aleatória com distribuição normal padrão com média zero e variância um. Mais detalhes podem ser obtidos em Cochran (1977), seções 2.8 e 2.15, ou em Särndal, Swensson, e Wretman (1992), seção 2.11.

É com base nessa distribuição assintótica que se pode fazer inferência por intervalos de confiança para a média populacional, e com base nesta ideia, medir a *margem de erro* de uma estimativa da média populacional. Um intervalo de confiança de nível $(1-\alpha)\%$ para a média populacional sob AAS é dado por:

$$IC_{AAS}(\overline{Y};1-\alpha) = \left[\overline{y} \mp z_{\alpha/2}\sqrt{\widehat{V}_{AAS}(\overline{y})}\right] \tag{4.8}$$

onde $z_{\alpha/2}$ é a abscissa da distribuição $N(0;1)$ que deixa à sua direita área igual a $\alpha/2$.

A *semiamplitude* do intervalo de confiança para o parâmetro nos fornece uma ideia da *margem de erro* que se tem ao estimar o parâmetro. A *margem de erro* da estimativa de média é, então, estimada por:

$$\widehat{ME}_{AAS}(\overline{y}) = z_{\alpha/2}\sqrt{\widehat{V}_{AAS}(\overline{y})} \tag{4.9}$$

Note que a *margem de erro* é também uma quantidade que se pode estimar a partir da amostra selecionada e observada. Essa é uma das vantagens importantes da *amostragem probabilística*, pois nos fornece além das estimativas pontuais diretas dos parâmetros de interesse, também indicativos da incerteza associada a tais estimativas.

4.3.7 Comparação dos planos de Amostragem Aleatória Simples Com e Sem Reposição

Do ponto de vista prático, a principal diferença entre os planos amostrais *AASC* e *AAS* é que, no primeiro, unidades populacionais podem ser selecionadas mais de uma vez para compor a amostra, o que não é possível no segundo. Ambos os planos permitem usar estimadores não viciados bem simples para o total e a média populacionais, mas o plano *AAS* fornece estimadores com menor variância para iguais tamanhos de amostra, sendo por isso mesmo preferido na prática.

Em resumo, as diferenças da *amostragem aleatória simples sem reposição* (AAS) para a *com reposição* (AASC) estão nos seguintes aspectos:

1. AAS evita repetição de seleção de unidades para a amostra.

2. AAS leva a um modelo estatístico diferente: as observações amostrais *não são independentes*.
3. AAS diminui o conjunto S de amostras possíveis.
4. AAS mantém a simplicidade dos estimadores.
5. AAS permite estimação mais eficiente da média ou total populacionais sob amostras de igual tamanho.

4.4 Determinação do tamanho da amostra

Nesta seção, procura-se responder à pergunta de que *tamanho* deve ter a amostra de uma pesquisa. A resposta a essa pergunta depende da resposta a uma de duas perguntas alternativas:

a. Quanto se pretende gastar na pesquisa? ou
b. Qual a precisão desejada (esperada) dos resultados?

A primeira decisão é qual dos dois caminhos seguir para determinar o tamanho da amostra: fixar *custo* ou *precisão*?

4.4.1 Tamanho amostral para custo fixado

Se a escolha for determinar o tamanho da amostra fixando parâmetros de *custo*, recomendamos usar como tamanho de amostra o *maior tamanho* permitido pelo orçamento (ou tempo) disponível. Nesse caso, não há uma teoria geral pronta para ser aplicada a toda e qualquer pesquisa. Há que estudar a *função de custo* de cada pesquisa e com base nela, definir o tamanho da amostra.

Exemplo 4.1 Determinando o tamanho de amostra para uma pesquisa junto a empresas

Considere um cenário em que o interesse é realizar uma pesquisa junto a empresas, para estimar alguns totais ou médias. O cliente que demanda a pesquisa informa que tem disponível um orçamento limitado, e que para a atividade de coleta da pesquisa o valor disponível é de R$ 400.000,00 (quatrocentos mil reais).

Após realizar reuniões com o cliente e ter informação mais precisa sobre o questionário e características da pesquisa, o responsável por planejar a amostra estima que coletar dados de cada empresa selecionada para a amostra terá um custo médio de R$ 200 por

questionário. Vale também comentar que é importante que, ao estimar o custo médio de coleta por questionário, o planejador da pesquisa deixe margem de segurança para cobrir eventuais dificuldades imprevistas de coleta.

Considerando o orçamento disponível para a coleta, o recomendável seria então usar uma amostra de $n = 400.000/200 = 2.000$ empresas.

Após calcular este tamanho de amostra, o responsável pelo planejamento da amostra deve comunicar ao cliente alguma ideia de que precisão seria possível alcançar com esse tamanho de amostra e orçamento, ao menos para os principais parâmetros de interesse da pesquisa. Isto ajudaria a evitar frustrações ou reclamações após a coleta dos dados e a obtenção das estimativas de interesse.

4.4.2 Tamanho amostral para precisão fixada

Se a escolha for determinar o tamanho amostral para garantir resultados com certa *precisão* (margem de erro) especificada, devemos também especificar o grau de confiança a adotar.

Exemplos:

1. “Desejamos estar 90% confiantes de que a estimativa da média está a no máximo ± 10 unidades do valor verdadeiro.”
2. “Desejamos que a estimativa da média não se afaste do valor verdadeiro mais que 10%, com probabilidade 0,95.”

Em 1) acima, estabelecemos a *margem de erro*, igual à semiamplitude do *intervalo de confiança* para $\overline{Y}$ em unidades da variável resposta, para um determinado *nível de confiança* (90% ou 0,90).

Em 2) acima, estabelecemos a *margem de erro relativa*, a semiamplitude do intervalo de confiança para $\overline{Y}$ em *termos relativos*, aceitando um *erro relativo máximo* de 10% do valor de $\overline{Y}$, para um determinado nível de confiança (95% ou 0,95).

Para determinar o tamanho amostral para precisão fixada, a ideia é usar a informação disponível sobre a distribuição do estimador e alguma informação prévia existente sobre a população.

Sabe-se que para n grande e $f = n/N$ limitada longe de 1:

$$\frac{\overline{y} - \overline{Y}}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}} \approx N(0;1)$$

Segue-se então que:

$$P\left(\frac{|\overline{y} - \overline{Y}|}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

onde $z_{\alpha/2}$ é o valor da abscissa da distribuição Normal padrão tal que $P[N(0;1) > z_{\alpha/2}] = \alpha/2$.

Segue-se então que:

$$P\left(|\overline{y} - \overline{Y}| \leq z_{\alpha/2}\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}\right) = 1 - \alpha$$

.

Logo, o erro de estimar $\overline{Y}$ usando $\overline{y}$ sob AAS é menor ou igual a $z_{\alpha/2}\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}$ com probabilidade $1 - \alpha$.

Então se desejamos estimar $\overline{Y}$ com um erro máximo de ± 10 unidades, com um nível de confiança de 90% (o que significa que o valor tabelado $z_{\alpha/2} = 1,645$), basta fazer:

$$z_{\alpha/2}\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2} = 1,645\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2} = 10$$

e resolver a equação em relação ao tamanho amostral n.

Logo:

$$1,645\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2} = 10 \Rightarrow \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 = \left(\frac{10}{1,645}\right)^2$$

Segue-se então que:

$$\frac{1}{n} = \left(\frac{10}{1,645}\right)^2 \frac{1}{S_y^2} + \frac{1}{N} \Rightarrow n = \frac{1}{\left(\frac{10}{1,645}\right)^2 \frac{1}{S_y^2} + \frac{1}{N}}$$

Para calcular o tamanho desejado da amostra precisamos conhecer N e S_y^2. Seguem algumas sugestões de como fazer para resolver a questão de que S_y^2 é também desconhecido:

1. Usar informações de *pesquisas anteriores*.
2. Fazer *amostra prévia* (*amostra piloto*) e estimar S_y^2 usando s_y^2 com os dados dessa amostra prévia.
3. Em casos especiais (proporções e outros), *usar cota superior* para o valor de S_y^2.

O caso geral

Seja d a *precisão desejada*, o *erro máximo admissível* na estimação de $\overline{Y}$, a *semiamplitude* desejada para o intervalo de confiança de $\overline{Y}$. Seja $1-\alpha$ o *coeficiente de confiança* desejado para o procedimento. Para *intervalos de confiança* de 95% usamos $z_{\alpha/2} = 1,96$.

Um intervalo de confiança não é uma especificação sobre uma particular amostra, mas sobre o desempenho do procedimento sob todas as possíveis amostras. Quando se usa um intervalo de confiança de 95% para um parâmetro, isto quer dizer que os intervalos construídos com cerca de 95 de cada 100 amostras selecionadas (sob idênticas condições) cobririam o “verdadeiro” valor do parâmetro de interesse. Para uma amostra específica, selecionada pelo método escolhido, acredita-se que é de 95% a chance que o “verdadeiro” valor seja coberto pelo intervalo:
$[\text{Estimativa} - 1,96 \times \text{desvio padrão};\ \text{Estimativa} + 1,96 \times \text{desvio padrão}]$.

Assim:

$$\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 = \left(\frac{d}{z_{\alpha/2}}\right)^2$$

Portanto, o tamanho de uma AAS que assegura precisão d com nível de confiança $1-\alpha$ é:

$$n = \frac{1}{\left(\frac{d}{z_{\alpha/2}}\right)^2 \frac{1}{S_y^2} + \frac{1}{N}} = \frac{N z_{\alpha/2}^2 S_Y^2}{N d^2 + z_{\alpha/2}^2 S_y^2} \tag{4.10}$$

Comentários

1. A expressão (4.10) só se aplica para o caso do estimador média amostral $\overline{y}$ para a média populacional $\overline{Y}$ sob AAS.

2. É possível derivar expressões similares para o caso da estimação de totais, e também de outros parâmetros.

3. Para planos amostrais mais complexos, é mais difícil resolver equações do tipo acima para determinar tamanhos amostrais, e sua alocação em estratos e conglomerados. Entretanto, a ideia de *Efeito de Plano Amostral* (EPA) vai ser útil neste contexto. Veja discussão no capítulo 13.

Exemplo 4.2 Considere a população formada pelos municípios brasileiros, conforme consta do arquivo 'MunicBR_dat.rds'. Tendo esta população em mente, imagine que seria usada para seleção de uma amostra AAS de $n = 200$ municípios. Imagine que tal amostra seria usada para estimar a *média populacional* da variável *área* dos municípios.

1. Com esta perspectiva, use os *dados populacionais* para:

    a. Calcular a *média populacional*;

    b. Calcular a *variância*, *desvio padrão* e *coeficiente de variação* do estimador usual;

    c. Avaliar a *margem de erro relativo* que a estimativa teria ao nível 95% de confiança;

    d. Determinar o *tamanho da amostra* que seria necessária para estimar a média da área com um erro máximo de $150\,km^2$ ao nível de confiança de 95%.

2. Selecione uma AAS de tamanho $n = 200$ e use os *dados amostrais* para calcular:

    a. Uma estimativa da *média populacional*;

    b. Estimativas da *variância*, *desvio padrão* e *coeficiente de variação* da média estimada;

    c. Estime a *margem de erro relativo* que a estimativa obtida em a) teria ao nível 95% de confiança;

    d. O *tamanho da amostra* que seria necessária para estimar a média da área com um erro máximo de $150\,km^2$ ao nível de confiança de 95%.

3. Compare estimativas obtidas no item 2 com os valores obtidos no item 1 e analise/comente.

Solução do Exemplo 4.2 usando R

```r
# Limpa área de trabalho
rm(list = ls())

# Carrega biblioteca(s) requerida(s)
library(sampling)

# Leitura dos dados
MunicBR_dat <- readRDS(file="MunicBR_dat.rds")
str(MunicBR_dat)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    5570 obs. of  6 variables:
##  $ CodMunic : chr  "1100015" "1100023" "1100031" "1100049" ...
##  $ SiglaUF  : chr  "RO" "RO" "RO" "RO" ...
##  $ CodUF    : chr  "11" "11" "11" "11" ...
##  $ Pop      : num  25728 101269 6495 85863 18041 ...
##  $ Area     : num  7067 4427 1314 3793 2783 ...
##  $ Densidade: num  3.64 22.88 4.94 22.64 6.48 ...
```

```r
# Item 1
# Tamanho da população
(N = nrow(MunicBR_dat))
```

```
## [1] 5570
```

```r
# Tamanho inicial da amostra
(n <- 200)
```

```
## [1] 200
```

```r
# Soluções para item 1
# a. Média populacional
(med_pop <- mean(MunicBR_dat$Area))
```

```
## [1] 1526.536
```

```
# b. Dispersão do estimador de média sob AAS
(VAR_med_amo <- ((1/n) - (1/N)) * var(MunicBR_dat$Area))
```

```
## [1] 151683.3
```

```
(DP_med_amo <- sqrt(VAR_med_amo))
```

```
## [1] 389.4654
```

```
(CV_med_amo <- 100 * DP_med_amo / med_pop)
```

```
## [1] 25.51302
```

```
# c. Margem de erro relativo do estimador de média sob AAS
(ME_med_amo <- qnorm(0.975)*CV_med_amo)
```

```
## [1] 50.00461
```

```
# d. Tamanho de amostra para obter margem de erro de 150 ao nível 95%
(d <- 150)
```

```
## [1] 150
```

```
(n_amo <- (N * qnorm(0.975)^2 * var(MunicBR_dat$Area)) /
          (N * d^2 + qnorm(0.975)^2 * var(MunicBR_dat$Area)))
```

```
## [1] 2734.689
```

```
(n_amo <- ceiling(n_amo))
```

```
## [1] 2735
```

```
# Soluções para item 2
# Seleciona amostra
munic_amo <- getdata(MunicBR_dat, srswor(n, N))

# a. Média amostral
(med_amo <- mean(munic_amo$Area))
```

```
## [1] 1587.12
```

```
# b. Estimativas da dispersão do estimador de média sob AAS
(var_med_amo <- ((1/n) - (1/N)) * var(munic_amo$Area))
```

```
## [1] 145084.7
```

```
(dp_med_amo <- sqrt(VAR_med_amo))
```

```
## [1] 389.4654
```

```
(cv_med_amo <- 100 * DP_med_amo / med_amo)
```

```
## [1] 24.53913
```

```r
# c. Margem de erro relativo do estimador de média sob AAS
(me_med_amo <- qnorm(0.975)*cv_med_amo)
```

```
## [1] 48.09581
```

```r
# d. Tamanho de amostra para obter margem de erro de 150 ao nível 95%
(d <- 150)
```

```
## [1] 150
```

```r
(n_amo_est <- (N * qnorm(0.975)^2 * var(munic_amo$Area)) /
              (N * d^2 + qnorm(0.975)^2 * var(munic_amo$Area)))
```

```
## [1] 2672.81
```

```r
(n_amo_est <- ceiling(n_amo_est))
```

```
## [1] 2673
```

4.5 Exercícios

Exercício 4.1 Mostre que o estimador $\widehat{Y}_{AASC}$ para o total não é um estimador tipo Horvitz-Thompson.

Exercício 4.2 Considere a população de 338 fazendas produtoras de cana de açúcar fornecida no arquivo "fazendas_dat.rds". Esse arquivo contém os dados de algumas variáveis econômicas medidas para cada uma das fazendas dessa população, tais como área plantada com cana de açúcar (*Area*), quantidade colhida de cana (*Quant*), receita (*Receita*) e despesa com a produção de cana (*Despesa*), e algumas variáveis de contexto sobre as fazendas, tais como região de localização (*Regiao*) e classe de tamanho da fazenda (*Classe*).

Imagine que há interesse em pesquisar por amostragem essa população de fazendas, visando estimar medidas descritivas da população, tais como os totais das variáveis Quant, Receita e Despesa. O objetivo do exercício é usar os dados fornecidos para estudar o comportamento esperado de um plano amostral. Considere a ideia de selecionar uma amostra de $n = 50$ fazendas da população usando Amostragem AAS.

1. Use os valores populacionais das variáveis de interesse (Area, Quant e Receita) para calcular os totais populacionais de interesse. Calcule também o desvio padrão (DP) e o coeficiente de variação (CV) esperados para os estimadores dos totais populacionais de interesse, supondo que o estimador Horvitz-Thompson para o total seria empregado. Compare os resultados para as diversas variáveis.

2. Selecione efetivamente uma amostra segundo o esquema amostral indicado e use essa amostra para estimar os totais populacionais de interesse, bem como os respectivos DPs e CVs. Compare os resultados com os valores obtidos no item 1 e comente aspectos dignos de nota.

3. Use a amostra selecionada por AAS para estimar a variância populacional de Quant e Receita. Use estas informações para dimensionar a amostra necessária para estimar o total com CV de 10% para cada uma das duas variáveis. Ao final, que tamanho de amostra você usaria na pesquisa para atingir o objetivo estabelecido?

4. Repita 500 vezes o item 2, e analise a distribuição resultante das estimativas de total para as variáveis *Quant* e *Receita*. Analise, comente.