# AMOSTRAGEM

## Unidade 4

## Amostragem Aleatória Simples



## Amostragem: Questões Fundamentais Para Definir

1. Métodos / esquemas para SELEÇÃO da amostra
2. TAMANHO da amostra
3. ESTIMADORES dos parâmetros de interesse
4. AVALIAÇÃO DA QUALIDADE das estimativas
   a. Variância dos estimadores
   b. Estimação da variância dos estimadores

## Amostragem Aleatória Simples Sem Reposição (AAS)

População U={ 1 , 2, ..., N }

Amostra s de tamanho fixado igual a $1 \leq n < N$ (número de unidades distintas).

**Definição:** AAS é o procedimento de seleção que garante que ***todas*** as amostras de tamanho n têm a mesma probabilidade de serem escolhidas.



## Espaço Amostral

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.

Este procedimento simples fornece a base para muitos outros mais complexos. As idéias principais de amostragem podem ser com ele desenvolvidas.

## Planos Amostrais e Esquemas de Seleção

Para implementar um plano amostral p(s) qualquer precisamos contar com um *esquema de seleção*.

**Esquema de seleção** é um mecanismo que permita 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 esquemas de seleção:

- Seqüências de sorteios;
- Processamento seqüencial da lista ou cadastro.



## Esquemas Baseados em Seqüências de Sorteios

- São implementados 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.
- Cada sorteio resulta em uma unidade selecionada para a amostra.

## Exemplo 4.1: Amostragem Aleatória Simples Com Reposição (AASC)

1) Selecionar uma unidade de U com probabilidade 1/N.

2) Repetir o passo 1) n vezes, sendo cada seleção independente das anteriores.

Unidades já selecionadas podem ser repetidas na amostra.

**Notas:**

- Procedimento gera observações que podem ser modeladas como determinações de variáveis aleatórias IID.
- Nunca usada na prática, pois não é eficiente.
- Número de amostras possíveis é $N^n$.



## Exemplo 4.2: Amostragem Aleatória Simples Sem Reposição (AAS) - Algoritmo "Convencional"

1) Selecione a primeira unidade dentre as N unidades de U com probabilidades iguais a 1/N;

2) Selecione a segunda unidade dentre as N-1 unidades ainda não selecionadas de U com probabilidades iguais a 1/(N-1);

⋮

n) Selecione a 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).

## AAS: Algoritmo “Convencional”

- ✓ Esquema fornecia a regra para uso de ‘tabelas de números aleatórios’ antes do aparecimento e uso de computadores para seleção de amostras.
- ✓ Implementação em computador é **ineficiente** para esse esquema, devido à necessidade de guardar duas listas: a das unidades já selecionadas e a das unidades ainda disponíveis.
- ✓ A cada novo sorteio, a segunda lista tem que ser percorrida para extrair uma nova unidade.





## Esquemas Baseados em Processamento de Listas

- São implementados mediante realização de uma série de experimentos aleatórios, executados seqüencialmente para cada unidade do cadastro ou lista.
- Pode não ser necessário percorrer todo o cadastro/lista.
- Para cada unidade é realizado um experimento aleatório que vai resultar na inclusão ou exclusão dessa unidade da amostra s.

# Exemplo 4.3: Amostragem de Bernoulli (AB)

As unidades aparecem no cadastro numa certa ordem, digamos igual à dos rótulos i=1,2,...,N.

Seja $\pi$ uma constante tal que $0 < \pi < 1$.

Sejam também $A_1$, $A_2$, ..., $A_N$ um conjunto de N variáveis aleatórias IID com distribuição Uniforme no intervalo [0;1], denotada $U(0;1)$.

Associamos $A_i$ com a unidade i, para todo $i \in U$.



# Exemplo 4.3: Amostragem de Bernoulli (AB)

Então processamos seqüencialmente a lista ou cadastro, testando para cada i=1,...,N a condição: $A_i < \pi$?

Quando isto ocorre, incluímos a unidade i na amostra s.

Quando a condição for falsa, a unidade não é incluída na amostra s e passamos à próxima unidade.

**<u>Exercício 4.1:</u>** Qual a distribuição de probabilidades do tamanho da amostra n sob amostragem de Bernoulli?

## Algoritmo de Hàjek para Selecionar AAS

**Passo 1**: Para cada $i \in U$, associe um **número pseudo-aleatório** $a_i$, onde os $a_i$ são determinações de variáveis aleatórias IID $A_1, A_2, ..., A_N$, todas com distribuição $U(0;1)$.

| Rótulo (i) | 1 | 2 | ... | N |
|---|---|---|---|---|
| Número aleatório ($a_i$) | $a_1$ | $a_2$ | | $a_N$ |

**Passo 2**: Reordene a população segundo os números pseudo-aleatórios $a_1, a_2, ..., a_N$, obtendo uma "**permutação aleatória**" dos rótulos.

| Rótulos | $i_1$ | $i_2$ | ... | $i_N$ |
|---|---|---|---|---|
| Número aleatório ordenado $a_{(i)}$ | $a_{(1)}$ | $a_{(2)}$ | | $a_{(N)}$ |





## Algoritmo de Hàjek para Selecionar AAS

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

Por exemplo, os rótulos $i_1, i_2, ..., i_n$ fornecem uma AAS.

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

## Algoritmo de Fan, Muller e Rezucha (1962)

Sejam $a_i$, i=1,2,...,N, determinações de variáveis aleatórias IID $A_1$, $A_2$, ..., $A_N$, todas com distribuição $U(0;1)$.

**Passo 1**: Se $a_1 < n/N$, **inclua** a unidade 1 na amostra. Caso contrário, passe à unidade 2.

**Passo 2**: Para as unidades i=2, 3, ...,N, **processe seqüencialmente a lista**, incluindo na amostra as unidades i tais que $a_i < \frac{n - n_{i-1}}{N-(i-1)} = \frac{n - n_{i-1}}{N - i + 1}$, onde $n_{i-1}$ é o número de unidades selecionadas até o processamento da unidade i-1.

**Interrompa** o processamento quando $n_{i-1} = n$.





## Probabilidades de Inclusão (Seleção)

Tratar com as distribuições de probabilidades de aleatorização p(s) 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 de amostras possíveis de tamanho n=40 sob AAS tem dimensão $\binom{N}{n} = \binom{1.000}{40} = 5,6 \times 10^{71}$.

Se a população tivesse N=5.000 e a amostra n=200, a dimensão de S cresceria para $\binom{5.000}{200} = 1,4 \times 10^{363}$.

## Probabilidades de Inclusão (Seleção)

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 na amostra** de unidades ou pares de unidades.





## Exercícios

**Exercício 4.2:** Calcule as probabilidades de inclusão de primeira e segunda ordem para uma amostra de tamanho n de uma população de tamanho N sob AAS.

**Exercício 4.3:** Calcule as probabilidades de inclusão de primeira e segunda ordem para uma amostra de Bernoulli com parâmetro $\pi$ de uma população de tamanho N.

**Exercício 4.4:** Calcule as probabilidades de inclusão de primeira ordem para uma amostra de tamanho n de uma população de tamanho N sob AASC.

# Notas

1) Sob AAS, $\pi_i = (n / N) > 0$ para todo $i \in U$ desde que $n>0$.

2) $(n / N) = f$ é chamada de *fração amostral* ou *taxa de amostragem*.

3) Estimação de variância sem vício requer $\pi_{ij} > 0$ para todo $i , j \in U$. Sob AAS, $\pi_{ij} = [n (n-1)]/ [N(N-1)] > 0 \; \forall i , j \in U$.

4) Sob AAS, as probabilidades de inclusão $\pi_i$ , $\pi_{ij}$ , etc. não dependem de i ou j, e essa é a razão da simplicidade desse plano amostral.



# Exemplo 4.5

Sob AAS de tamanho n de população com N:

$$E[\delta_i] = \frac{n}{N} \;,\; V[\delta_i] = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

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

Assim a correlação entre duas variáveis indicadoras de seleção sob AAS é $CORR[\delta_i; \delta_j] = -1/(N-1)$ se $i \neq j$.

## Estimador Não Viciado do Total Populacional Sob AAS

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} \frac{y_i}{n/N} = \sum_{i \in s} \frac{N}{n} y_i = N \frac{1}{n} \sum_{i \in s} y_i = N\,\bar{y} = \hat{Y}$$

**Variância do Estimador de Total**

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

onde $S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{Y})^2$.





## Estimador da Variância do Estimador de Total

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

onde

$$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y})^2.$$

## Estimador Não Viciado da Média Populacional Sob AAS

Sabemos que $\overline{Y} = \frac{1}{N} Y$. Logo, um estimador não viciado para a média populacional $\overline{Y}$ é dado por:

$$\overline{y}_W = \frac{1}{N} \hat{Y}_{HT} = \frac{1}{N} N \, \overline{y} = \overline{y}$$

Portanto a média amostral $\overline{y}$ é não viciada para $\overline{Y}$ sob AAS.





## Variância do Estimador de Média

$$V_{AAS}(\overline{y}) = \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$$

**Estimador da Variância do Estimador de Média**

$$\hat{V}_{AAS}(\overline{y}) = \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}$$

## Distribuição Assintótica da Média Amostral

Sob repetidas amostras (repetições do procedimento de seleção segundo AAS), $\bar{y}$ tem uma distribuição de probabilidades.

Esta distribuição depende:

- ✓ Da distribuição dos y´s na população;
- ✓ Do tamanho da amostra;
- ✓ Do plano amostral p(s).

**Resultado:** situação complicada.





## Distribuição Assintótica da Média Amostral

Se n for grande e n/N pequeno, o Teorema Central do Limite pode ser usado para obter a distribuição aproximada:

$$\frac{\bar{y} - E_{AAS}(\bar{y})}{\sqrt{\hat{V}_{AAS}(\bar{y})}} = \frac{\bar{y} - \bar{Y}}{\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}}} \approx N(0;1)$$

Ref.: Cochran(1977, seções 2.8 e 2.15)

Särndal, Swensson e Wretman (1992, seção 2.11)

## Notas

1. $S_y^2 / n$ é análogo a $\sigma_y^2 / n$ na inferência clássica.
2. O termo (1 - n/N) é chamado de <u>fator de correção de população finita</u>. Quando $n/N \to 1$, $(1 - n/N) \to 0$.
3. Se a fração amostral f = n/N for pequena, então a correção de população finita é desprezível, pois $(1 - f) \cong 1$.
4. Neste caso $(f \cong 0)$, a amostragem sem reposição se comporta como se fosse com reposição.



## RESUMO: Resultados Sob AAS

1. A média amostral $\bar{y} = \sum_{i \in s} y_i / n$ é um estimador não viciado da média populacional $\bar{Y} = \sum_{i \in U} y_i / N$.
2. A variância amostral $s_y^2 = \sum_{i \in s} (y_i - \bar{y})^2 \big/ (n\text{-}1)$ é um estimador não viciado da variância populacional $S_y^2 = \sum_{i \in U} (y_i - \bar{Y})^2 \big/ (N\text{-}1)$.
3. $V_{AAS}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$.

## RESUMO: Resultados Sob AAS

4. $\hat{V}_{AAS}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}$ é um estimador não viciado para $V_{AAS}(\bar{y})$.

5. $CV_{AAS}(\bar{y}) = CV_{AAS}(\hat{Y}_{HT})$



IBGE

## RESUMO: Resultados Sob AAS

6. A distribuição de aleatorização de $\bar{y}$ sob AAS pode ser aproximada, para n grande e n/N pequeno, pela distribuição Normal:

$$\frac{\bar{y} - E_{AAS}(\bar{y})}{\sqrt{\hat{V}_{AAS}(\bar{y})}} = \frac{\bar{y} - \bar{Y}}{\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}}} \approx N(0;1)$$

## Diferenças Para Amostragem Com Reposição

1. Evita repetição de seleção de unidades para amostra.
2. Modelo estatístico diferente: observações amostrais **não são independentes**.
3. Diminui conjunto de amostras possíveis.
4. Mantém simplicidade dos estimadores.
5. Maior eficiência na estimação da média / total populacionais para amostra de igual tamanho total.





## Dados Amostrais

$$\{y_{k_1}, y_{k_2}, \ldots, y_{k_n}\}$$

Observações de variáveis aleatórias $Y_1$, $Y_2$, ..., $Y_n$ I.D. (identicamente distribuídas) com distribuição comum dada por $p(Y_i=y_k)$ na tabela abaixo

| Unidade populacional → | 1 | 2 | ... | N | Total |
|---|---|---|---|---|---|
| Valor que $Y_i$ pode assumir | $y_1$ | $y_2$ | ... | $y_N$ | Y |
| Probabilidade $p(Y_i=y_k)$ | 1/N | 1/N | ... | 1/N | 1 |

## Valor Esperado de $Y_i$ Para Qualquer i=1,2,...,n.

$$E(Y_i) = \sum_{k \in U} y_k \times \frac{1}{N} = \overline{Y}$$

Variância de $Y_i$ para qualquer i=1,2,...,n.

$$V(Y_i) = \sum_{k \in U} (y_k - \overline{Y})^2 \times \frac{1}{N} = \frac{N-1}{N} S_y^2$$

Variáveis aleatórias $Y_1$, $Y_2$, ..., $Y_n$ NÃO são independentes.

Prova: $p\left(Y_2 = y_i \middle| Y_1 = y_i\right) = 0$ para qualquer $i \in U$.





## Determinando o Tamanho da Amostra

De que tamanho deve ser a amostra da pesquisa?

A resposta a essa pergunta depende da resposta a uma de duas perguntas alternativas:

1. Quanto se pretende gastar na pesquisa?
2. 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**?

## Tamanho Amostral Para Custo Fixado

Se a escolha for determinar o tamanho da amostra fixando parâmetros de **custo**, 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 em toda e qualquer pesquisa.

Há que estudar a função de custo de cada pesquisa e com base nela, definir o tamanho da amostra.



## 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 os resultados estão a ±10 unidades do valor verdadeiro."

2) "Desejamos que a estimativa não se afaste do valor verdadeiro mais que 10%, com probabilidade 0,95."

## Margem de Erro Absoluta e Relativa

Em 1) acima, estabelecemos a largura de um **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 largura de um 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).





## Tamanho Amostral Para Precisão Fixada

A idéia 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 n/N limitado:

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

## Tamanho Amostral Para Precisão Fixada

Segue-se então que

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

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



## Tamanho Amostral Para Precisão Fixada

Segue-se então que

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

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

## Tamanho Amostral Para Precisão Fixada

Então se desejamos estimar $\overline{Y}$ com um erro máximo de ±10 unidades, com um nível de confiança de 90%, basta fazer

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

e resolver em relação ao tamanho amostral n. Logo:

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





## Tamanho Amostral Para Precisão Fixada

Segue-se 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 resolver esta equação precisamos conhecer N e $S_y^2$.

## Tamanho Amostral Para Precisão Fixada

Mas $S_y^2$ é também desconhecido! Como fazer?

1) Usar informações de **pesquisas anteriores**.
2) Fazer **amostra prévia** / piloto e estimar $S_y^2$ usando $s_y^2$ com os dados da sua 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 **semi-amplitude** 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.

Assim:

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

## O caso geral

Portanto:

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

$$= \frac{N z_{\alpha/2}^2 S_y^2}{N d^2 + z_{\alpha/2}^2 S_y^2}$$





## Notas

1. Estas expressões só se aplicam para o caso do estimador média amostral $\bar{y}$ para a média populacional $\bar{Y}$ sob AAS.
2. Para planos amostrais mais complexos, é mais difícil resolver equações do tipo acima para determinar tamanhos amostrais, e sua alocação em estratos e conglomerados.
3. A idéia de **Efeito de Plano Amostral** (EPA) vai ser útil neste contexto.