MAT02036 - Amostragem 2

Aula 01 - Visão Geral e Revisão

Markus Stein

Departamento de Estatística, IME/UFRGS

2022/2

Housekeeping

- Aproveitem o momento presencial para tirar dúvidas
- Se estivéssemos no ensino remoto ou à distância
  - vocês poderiam estar somente ouvindo, sem interação
  - ou assistindo vídeos e material em outro momento
- Depois das aulas, rever material da aula passada
  - fazer exercícios
  - se preparar para a próxima aula

Visão geral de amostragem e AAS

População de pesquisa

- Denominamos **população** qualquer conjunto contendo um número finito N de unidades, que compartilham alguma(s) característica(s) em comum;
    - N é o **tamanho da população**.
- As unidades deste conjunto são denominadas **unidades da população**,
    - representadas por um conjunto de N rótulos distintos denotado

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

    - i o **rótulo** para uma unidade genérica da população, $i \in U$.
- Exemplos:
    - domicílios e moradores de certa localidade;
    - indústrias instaladas num certo país;
    - fazendas situadas num certo estado;
    - alunos matriculados em uma série da rede escolar estadual em 2022.
- Definição clara e precisa da *população de pesquisa*.

População de pesquisa

- Queremos **estimar** ou **inferir** certas quantidades ou **parâmetros** de **características** (variáveis) **numéricas** medidas ou **observadas** para, em tese, toda **unidade da população**.

- *Vetor populacional* é o conjunto de valores da variável correspondentes às unidades da população.

    - Exemplo, se y é a variável de pesquisa (de interesse) e y_i é o valor dessa variável y para a unidade i, então

    $$Y_U = \{y_1,\ y_2,\ \ldots,\ y_i,\ \ldots,\ y_N\}$$

    é o *vetor populacional* gerado pela variável y.

Parâmetros-alvo (ou de interesse) podem ser quaisquer funções dos valores dos vetores populacionais: 🎯

total populacional (da variável y)	$T_y = T = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i$

População de pesquisa

mais *parâmetros-alvo*: 🎯

média populacional	$\overline{Y} = \dfrac{T_y}{N} = \dfrac{1}{N}\sum_{i \in U} y_i;$
proporção populacional P	$P = \overline{Y}$, para $y \in \{0, 1\}$.
variância populacional	$Var_y = \dfrac{1}{N}\sum_{i \in U}(y_i - \overline{Y})^2 = \dfrac{1}{N}\left[\sum_{i \in U} {y_i}^2 - N\overline{Y}^2\right]$
variância populacional	$S_y^2 = \dfrac{1}{N-1}\sum_{i \in U}(y_i - \overline{Y})^2 = \dfrac{N}{N-1}Var_y$
desvio padrão - DP populacional	$DP_y = S_y = \sqrt{S_y^2}$
coeficiente de variação - CV populacional	$CV_y = \dfrac{DP_y}{\overline{Y}} = \dfrac{S_y}{\overline{Y}}$

População de pesquisa

mais *parâmetros-alvo*: 🎯

- a *função de distribuição cumulativa empírica populacional* - *FDCEP*, por consequência, *quantil populacional*, ex. a *mediana populacional*.
  $F_y(a) = \dfrac{1}{N} \displaystyle\sum_{i \in U} I(y_i \leq a)$.

- seja z outra variável de pesquisa, tomando valores z_i, $i \in U$, a *razão de totais* das variáveis y e z como $R = \dfrac{\sum_{i \in U} y_i}{\sum_{i \in U} z_i} = Y/Z$

- a *covariância populacional* e a *correlação populacional* das variáveis y e z,
  $S_{yz} = \dfrac{1}{N-1} \displaystyle\sum_{i \in U} (y_i - \overline{Y})(z_i - \overline{Z}) = \dfrac{1}{N-1} \left[\sum_{i \in U} y_i z_i - N \overline{Y}\, \overline{Z} \right]$ e $\rho_{yz} = \dfrac{S_{yz}}{S_y S_z}$

Censo é possível?

Amostragem de populações finitas

- **pesquisa por amostragem** menor custo de obtenção dos dados, maior rapidez e redução da carga de coleta de informações.
  - conhecendo **estimativas** dos **parâmetros** e **margem de erro** podemos ter bons resultados?

- Uma **amostra** $s = \{i_1, i_2, \ldots, i_n\}$ é qualquer subconjunto não vazio de unidades selecionadas da população U, $s \subset U$ $(1 \leq n \leq N)$.

índice	soma
$i \in s$: unidade i incluída na amostra s.	$\sum_{i \in s}$: em i sobre o conjunto de rótulos de unidades em s.
$s \ni i$: a amostra s contém a unidade i.	$\sum_{s \ni i}$: em s sobre o conjunto de amostras possíveis que contêm a unidade i.

- Consideraremos somente **amostras probabilísticas**,
  - os dados amostrais para a variável y são representados por

$$Y_s = \{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\}$$

.

Amostragem probabilística

Suposições

1. S: *espaço amostral*; o conjunto de todas as amostras s possíveis, bem definido e enumerável (teoricamente), $S = \{s_1, s_2, \ldots, s_j, \ldots, s_\nu\}$;

2. $p(s)$: uma probabilidade conhecida (ou calculável) associada a cada amostra $s \in S$, denominada *plano amostral*,

$$\sum_{s \in S} p(s) = 1;$$

3. Uma única amostra $s \in S$ é selecionada com probabilidade $p(s)$.

4. Associada a cada unidade $i \in U$ existe uma probabilidade positiva, π_i, dessa ser selecionada, denominada **probabilidade de inclusão** (de primeira ordem) da unidade i, tal que

$$\pi_i = P(i \in s) = \sum_{s \ni i} p(s) > 0, \forall i \in U.$$

5. π_i das **unidades selecionadas** e outros aspectos do **plano amostral** são levados em conta ao fazer **inferência** sobre os parâmetros.

Estatísticas, estimadores e estimativas

- Uma **estatística** é uma função real dos valores observados numa amostra, qualquer $f(y_{i_1}, y_{i_2}, \ldots, y_{i_n})$. Exemplos:

total amostral ou soma amostral da variável y	$t(s) = t = \sum_{i \in s} y_i$
média amostral da variável y	$\bar{y} = \frac{t(s)}{n} = \frac{1}{n} \sum_{i \in s} y_i$

- Um **estimador** $\hat{\theta}(s)$ é uma estatística usada para estimar um certo parâmetro θ de interesse.
    - Antes de observarmos s, $\hat{\theta}(s)$ é uma variável aleatória (v.a.) cuja distribuição temos interesse em conhecer, dela derivamos propriedades importantes de $\hat{\theta}(s)$.
    - Por simplicidade, denotamos $\hat{\theta}$, sem explicitar sua dependência de s, sempre que possível.
- Após a determinação da amostra s e a coleta dos dados das unidades, o valor calculado (observado) de $\hat{\theta}$ é chamado de **estimativa** do parâmetro.

Como escolher **"bons" estimadores?**

Propriedades dos estimadores

- O *valor esperado* de $\hat{\theta}$ é denotado por $E_p(\hat{\theta})$.
  - $E_p(\cdot)$ designa o valor esperado sob a distribuição de probabilidades induzida pelo plano amostral:

$$E_p(\hat{\theta}) = \sum_{s \in S} \hat{\theta}(s) p(s).$$

- O **vício** (ou **viés** ou **tendência**) de $\hat{\theta}$ é definido como:

$$B_p(\hat{\theta}) = E_p(\hat{\theta}) - \theta.$$

- o *vício relativo* de $\hat{\theta}$ é dado por:

$$RB_p(\hat{\theta}) = \frac{B_p(\hat{\theta})}{\theta}.$$

Propriedades dos estimadores

- $\hat{\theta}$ é **não viciado** (ou **não enviesado** ou **não tendencioso**) para θ quando

$$E_p(\hat{\theta}) = \theta,$$

ou

$$B_p(\hat{\theta}) = RB_p(\hat{\theta}) = 0,$$

Primeiro critério

- Para apoiar a **escolha de estimadores** sugere então que tratemos de usar *estimadores* sem vício, ou *não viciados*, ou ao menos *aproximadamente não viciados*.

- Quando isto for possível, teremos estimadores cuja distribuição será centrada no alvo desejado da inferência.

- Um mesmo parâmetro pode ter mais de um estimador não viciado disponível. Precisamos então de um segundo critério para ajudar na escolha de estimadores.

Propriedades dos estimadores

Quando um estimador é não viciado

- Nesse caso a *variância* de $\hat{\theta}$ mede a dispersão da distribuição do estimador em torno do alvo de inferência θ.

- A *variância* do estimador $\hat{\theta}$ é definida como:

$$Var_p(\hat{\theta}) = \sum_{s \in S} [\hat{\theta}(s) - E_p(\hat{\theta})]^2 p(s).$$

- Medidas alternativas da dispersão de $\hat{\theta}$ (que dependem da variância são):

Desvio padrão - *DP* (ou erro padrão)	$DP_p(\hat{\theta}) = [Var_p(\hat{\theta})]^{1/2}$**.**
Coeficiente de variação - CV	$CV_p(\hat{\theta}) = \frac{DP_p(\hat{\theta})}{\theta}$

- O *DP* mede a dispersão em unidade de medida igual à usada na mensuração da variável de interesse; e o *CV* expressa essa medida em termos relativos.

Propriedades dos estimadores

Quando um estimador é viciado

- Uma medida mais adequada da dispersão da distribuição do estimador em torno do alvo de inferência θ é o **erro quadrático médio - EQM**:

$$EQM_p(\hat{\theta}) = \sum_{s \in S} [\hat{\theta}(s) - \theta]^2 p(s)$$

- Versões análogas ao DP e do CV para o caso de estimadores viciados:

Erro médio - EM	$EM_p(\hat{\theta}) = [EQM_p(\hat{\theta})]^{1/2}$**;**
Erro relativo médio - ERM	$ERM_p(\hat{\theta}) = \dfrac{EM_p(\hat{\theta})}{\theta}$.

Nosso segundo critério

- Queremos estimadores com os menores erros de estimação,
  - escolher sempre os estimadores com o menor *EQM*,
  - ou com a menor variância quando forem não viciados.

Como obter estimadores "ótimos"?

Obtenção de estimadores

- Qual a diferença entre "Amostragem" ou "Inferência"?
    - Ambas utlizam amostras probabilísticas e funções dos dados observados para inferir sobre parâmetros.

No **contexto da Amostragem** (de popuções finitas)

- Não se estabelece uma distribuição de probabilidade (ou modelo) para os valores da variável y na amostra (ou na população(?)).

    - diferente do "contexto usual" da **Inferência** (Estatística).

- Os parâmetros que se deseja estimar não são responsáveis pela especificação de uma tal distribuição de probabilidades (ou modelo).

    - Em geral os parâmetros de interesse são definidos como funções dos valores (considerados fixos, mas desconhecidos) da variável y na população.

- Não há um procedimento geral para gerar estimadores que sejam ótimos nalgum sentido,

    - ex., em "Inferência" temos o *método da máxima verossimilhança*.

Obtenção de estimadores

- Os princípios usados em "Amostragem" para derivar estimadores dos parâmetros de interesse são baseados na simplicidade e no *método dos momentos*.

Suponha que o parâmetro-alvo é o *total populacional* $T_y = \sum_{i \in U} y_i$, queremos:

a. usar os *dados amostrais* $\{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\}$ para *estimar* T;
b. medir ou estimar a *precisão* ou a *margem de erro* da estimativa produzida para T.

- Um **estimador linear** $\widehat{T}_w$ do total populacional T é uma combinação linear dos valores amostrais y_i com *pesos amostrais* w_i, a serem definidos,

$$\widehat{T}_w = \sum_{i \in s} w_i y_i.$$

- Podemos usar os 2 critérios sugeridos para escolha de estimadores para determinar os pesos w_i,

$$\hat{\theta}_{opt} = arg\, min_{\hat{\theta}} EQM_p(\hat{\theta}).$$

Obtenção de estimadores

Exemplo: estimação do total e AAS

- Considere uma população fictícia com $N = 4$ mulheres (unidades populacionais), de quem foi indagado o número de filhos tidos nascidos vivos (a variável y).

Valor da variável y por unidade da população de mulheres

Rótulo da unidade (i)	**1**	**2**	**3**	**4**	**Total**
Valor da variável (y_i)	0	0	2	1	3

- Existem $\binom{4}{2} = 6 = \nu$ amostras possíveis de duas **unidades distintas**.

- O conjunto de todas as amostras possíveis é dado por

$$S = \{(1;2);(1;3);(1;4);(2;3);(2;4);(3;4)\}.$$

Obtenção de estimadores

Exemplo: estimação do total e AAS

- Considere também um **plano amostral** p_1:
  - $p_1(s) = 1/6, \forall s \in S$.

```
N <- 4                    # no. elementos na pop.
i <- 1:N                  # indice dos elementos da pop.
n <- 2                    # no. elementos na amostra
nu <- choose(N, n)        # no. possíveis amostras
j <- 1:nu                 # indice dos elementos dos espaço amostral
S <- combn(N,n)           # espaço amostral
p1s <- 1/nu                # plano amostral
```

Obtenção de estimadores

Exemplo: estimação do total e AAS

- Apresentação detalhada do *plano amostral* p_1.

Informações de cada amostra possível sob plano amostral p_1

Amostra	**Unidades na Amostra** s	**Valores na Amostra** s	**Soma Amostral** (t)	**Probabilidades** $p_1(s)$
1	{1;2}	{0;0}	0	1/6
2	{1;3}	{0;2}	2	1/6
3	{1;4}	{0;1}	1	1/6
4	{2;3}	{0;2}	2	1/6
5	{2;4}	{0;1}	1	1/6
6	{3;4}	{2;1}	3	1/6
Total	_	_	_	1

Obtenção de estimadores

Exemplo: estimação do total e AAS

Probabilidade sob p_1 para cada valor de t

Valores possíveis de t	**0**	**1**	**2**	**3**
Com probabilidade $p_1(s)$	1/6	2/6	2/6	1/6

- O valor esperado de t é

$$E_{p_1}(t) = \sum_{s \in S} t(s)\, p_1(s) = 0 \times \frac{1}{6} + 1 \times \frac{2}{6} + 2 \times \frac{2}{6} + 3 \times \frac{1}{6} = \frac{9}{6} = 1,5$$

- Porém o **total populacional** é $T_y = \sum_{i \in U} y_i = 3$.

- Como $1,5 = E_{p_1}(t) \neq T = 3$, dizemos que t seria um **estimador viciado** de T sob o plano amostral p_1 adotado.

Obtenção de estimadores

Exemplo: estimação do total e AAS

- Podemos **"corrigir"** t de modo que fique **não viciado** para o total populacional;
  - Já que $Y/E_{p_1}(t) = 3/1,5 = 2$, multiplicando por 2 o valor de t resulta num estimador cujo valor esperado deve ser igual a Y.
- Considere o novo estimador de T dado por: $\widehat{T} = 2 \times t = \sum_{i \in s} 2 \times y_i = \widehat{T}_w$.

Probabilidade sob p_1 para cada valor de $2t$

Valores possíveis de $2t$	**0**	**2**	**4**	**6**
Com probabilidade $p_1(s)$	1/6	2/6	2/6	1/6

$$E_{p_1}(\widehat{T}) = \sum_{s \in S} \widehat{T}(s)\, p_1(s) = 0 \times \frac{1}{6} + 2 \times \frac{2}{6} + 4 \times \frac{2}{6} + 6 \times \frac{1}{6} = \frac{18}{6} = 3.$$

- Dizemos que $\widehat{T} = 2 \times t$ é um **estimador não viciado** de T sob o plano amostral p_1 considerado.

O método pelo qual deduzimos $\widehat{T}_w$ é viável na prática?

A distribuição de aleatorização

Plano amostral $p(s)$

- Para deduzir os pesos do estimador $\widehat{T}_w$ **supomos**:
  - conhecer todos os valores da variável de pesquisa;
  - para obter o valor esperado do estimador;
  - para então calcular pesos que levariam à obtenção do estimador ponderado não viciado.
- A função $p(s)$ definida no conjunto S de todas as amostras possíveis é uma distribuição de probabilidades.
  - É possível obter a distribuição de probabilidades de estatísticas (ou estimadores) calculadas a partir de s.
- A distribuição de probabilidades assim obtida é chamada de **distribuição de aleatorização** da estatística ou estimador.

A distribuição de aleatorização

Plano amostral $p(s)$

- Na **amostragem probabilística**, inferências são feitas considerando a **distribuição de aleatorização**.
    - A única **suposição** é considerar como fonte de variação ou incerteza a possível repetição hipotética do processo de amostragem utilizando o **plano amostral** $p(s)$, que resultaria em diferentes amostras $s_1, s_2, \ldots \in S$.
- A distribuição de $\widehat{T}_w = 2 \times t = \sum_{i \in s} 2 \times y_i$ determinada por $p(s)$ é também chamada de **distribuição amostral** do estimador.
    - Estudamos suas propriedades para avaliar se $\widehat{T}_w$ é um bom estimador para o total populacional T.

Estimadores não viciados para o total populacional

- No Exemplo vimos como ober a *distribuição amostral* de um estimador induzida pelo plano amostral $p(s)$.
    - Os tamanhos da população N e da amostra n eram muito pequenos.
    - conhecemos os valores da variável y para todas as unidades da população U.
- Na prática, trabalhar com a distribuição $p(s)$ para derivar distribuições amostrais de estimadores é complicado.

Tamanhos do espaço amostral S para valores selecionados de N e n

```r
# Gera tabela com valores de binom(N,n)
tabela <- as.data.frame(
          rbind( c(4, 2, choose(4,2)),
                 c(10, 4, choose(10,4)),
                 c(100, 10, choose(100,10)),
                 c(1000, 20, choose(1000,20)),
                 c(10000, 100, choose(10000,100))) )
colnames(tabela) <- c("N", "n", "binom(N,n)")
```

Estimadores não viciados para o total populacional

N	n	binom(N,n)
4	2	6.000000e+00
10	4	2.100000e+02
100	10	1.731031e+13
1000	20	3.394828e+41
10000	100	6.520847e+241

- Uma saída é usar propriedades simplificadoras da distribuição induzida pelo plano amostral. Tratamos disso na próxima aula.
- Por hora vamos usar uma propriedade importante que pode ser deduzida a partir da distribuição de aleatorização.

Estimadores não viciados para o total populacional

- A *probabilidade de inclusão* da unidade i na amostra é dada por:

$$P(i \in s) = \pi_i = \sum_{s \ni i} p(s).$$

- Se tomarmos o *inverso da probabilidade de inclusão* $1/\pi_i$ como peso (w_i) de uma unidade amostrada, é fácil verificar que o estimador dado por $\widehat{T}_w$ é *não viciado* para o total populacional T:

$$\widehat{T}_w = \sum_{i \in s} w_i y_i = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i \in s} {\pi_i}^{-1} y_i.$$

- Essa é propriedade importante e será demonstrada de maneira formal na próxima aula.

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- Usando a propriedade recém apresentada, os pesos amostrais nesse Exemplo são

  $$w_i = 1/\pi_i = \frac{1}{1/2} = 2, \forall i,$$

  para uma das amostras de tamanho $n = 2$.

- O estimador ponderado do total nesse caso seria dado por:

  $$\widehat{T}_w = \sum_{i \in s} w_i y_i = \sum_{i \in s} {\pi_i}^{-1} y_i = \sum_{i \in s} 2 y_i = 2t.$$

- Já sabemos que este estimador é não viciado para T.

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- Considere a mesma população fictícia do exemplo anterior.
- Mas agora adote o plano amostral p_2 (para amostras de tamanho $n = 2$).

Informações de cada amostra possível sob plano amostral p_2

Amostra	Unidades na Amostra s	Valores na Amostra s	Soma Amostral (t)	Probabilidades $p_2(s)$
1	{1;2}	{0;0}	0	0,00
2	{1;3}	{0;2}	2	0,20
3	{1;4}	{0;1}	1	0,15
4	{2;3}	{0;2}	2	0,20
5	{2;4}	{0;1}	1	0,15
6	{3;4}	{2;1}	3	0,30
Total	_	_	_	1,00

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- Vamos agora usar as informações acima para:

1.Verificar que a estatística soma amostral (t) é viciada para estimar o total populacional T.
2.Obter / definir um estimador não viciado para o total populacional T.

Probabilidade sob p_2 para cada valor de t

Valores possíveis de t	**0**	**1**	**2**	**3**
Com probabilidade $p_2(s)$	❓❓❓	❓❓❓	❓❓❓	❓❓❓

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- Vamos agora usar as informações acima para:

1.Verificar que a estatística soma amostral (t) é viciada para estimar o total populacional T.
2.Obter / definir um estimador não viciado para o total populacional T.

Probabilidade sob p_2 para cada valor de t

Valores possíveis de t	**0**	**1**	**2**	**3**
Com probabilidade $p_2(s)$	0,0	0,3	0,4	0,3

- O valor esperado de t sob o plano amostral p_2 é:

$E_{p_2}(t) =$ ❓❓❓

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- Vamos agora usar as informações acima para:

1.Verificar que a estatística soma amostral (t) é viciada para estimar o total populacional T.

2.Obter / definir um estimador não viciado para o total populacional T.

Probabilidade sob p_2 para cada valor de t

Valores possíveis de t	**0**	**1**	**2**	**3**
Com probabilidade $p_2(s)$	0,0	0,3	0,4	0,3

- O valor esperado de t sob o plano amostral p_2 é:

$$E_{p_2}(t) = \sum_{s \in S} t(s) p_2(s) = 0 \times 0,0 + 1 \times 0,3 + 2 \times 0,4 + 3 \times 0,3 = 2 < 3 = T$$

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- Para obter um estimador não viciado, devemos calcular pesos adequados para as unidades amostrais.

Probabilidade de inclusão e peso amostral de cada unidade sob o plano amostral p_2

Rótulo da unidade (i)	**1**	**2**	**3**	**4**
Probabilidade de inclusão (π_i)	❓❓❓	❓❓❓	❓❓❓	❓❓❓
Peso (w_i)	❓❓❓	❓❓❓	❓❓❓	❓❓❓

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- Para obter um estimador não viciado, devemos calcular pesos adequados para as unidades amostrais.

Probabilidade de inclusão e peso amostral de cada unidade sob o plano amostral p_2

Rótulo da unidade (i)	**1**	**2**	**3**	**4**
Probabilidade de inclusão (π_i)	7/20=0,35	7/20=0,35	7/10=0,70	3/5=0,60
Peso (w_i)	20/7=2,857	20/7=2,857	10/7=1,429	5/3=1,667

- Usando o estimador do total com os pesos adequados $\widehat{T}_w$, obtêm-se os valores das estimativas para cada amostra possível.

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

Obtenção de estimativa sob plano amostral p_2 para cada amostra possível

Amostra	**Valores na Amostra s**	**Total Amostral ponderado**	**Probabilidades $p_2(s)$**	**Total $\times$ probabilidade**
1	{0;0}	0	0,00	0
2	{0;2}	2x(10/7)	0,20	4/7
3	{0;1}	1x(5/3)	0,15	1/4
4	{0;2}	2x(10/7)	0,20	4/7
5	{0;1}	1x(5/3)	0,15	1/4
6	{2;1}	2x(10/7)+1x(5/3)	0,30	6/7+1/2
Total	-	-	1,00	3

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

1. O estimador $\widehat{T}_w$ obtido usando os pesos iguais a $1/\pi_i$ tem valor esperado igual ao total populacional T.
    - $\widehat{T}_w$ é *não viciado* também sob o plano amostral p_2.
2. Mesmo a **amostra 1** tendo **probabilidade nula** de ser selecionada **não viola** os critérios para que p_2 seja um plano de **amostragem probabilística**.
    - as unidades populacionais têm probabilidades positivas de inclusão na amostra. (Verificar❗❗❗)
3. Temos agora duas opções de plano amostral (de tamanho $n = 2$), para estimar o total populacional T.
    - para ambos os planos amostrais temos estimadores não viciados do total populacional.

Coloca-se então a pergunta: qual dos dois planos é melhor❓❓❓

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- **Estratégia 1:** seleção equiprovável de amostras com estimador de total ponderado ($\widehat{T} = 2t$).

Probabilidade de seleção sob $p_1(s)$ para cada valor do estimador ponderado

Valores possíveis de $\widehat{T} = 2t$	0	2	4	6
Com probabilidade $p(s)$ sob $p_1(s)$	1/6	2/6	2/6	1/6

- **Estratégia 2:** seleção de amostras com probabilidades desiguais e estimador de total ponderado ($\widehat{T}_w$).

Probabilidade de seleção sob $p_2(s)$ para cada valor do estimador ponderado

Valores possíveis de $\widehat{T}_w$	5/3	20/7	20/7+5/3
Com probabilidade $p(s)$ sob $p_2(s)$	0,30	0,40	0,30

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- Em ambos os casos o estimador é não viciado, então comparamos ad *variâncias dos estimadores*.

Obtenção da variância dos estimadores sob os planos amostrais p_2 e p_1

Amostra	**Valores na Amostra** s	**Estimativa sob** p_2	**Probabilidade sob** p_2	**Estimativa sob** p_1	**Probabilidade sob** p_1
1	{0;0}	0	0,00	0	1/6
2	{0;2}	2x(10/7)	0,20	4	1/6
3	{0;1}	1x(5/3)	0,15	2	1/6
4	{0;2}	2x(10/7)	0,20	4	1/6
5	{0;1}	1x(5/3)	0,15	2	1/6
6	{2;1}	2x(10/7)+1x(5/3)	0,30	6	1/6
Variância	-	1,24	-	3,67	-

Estimadores não viciados para o total populacional

(cont.) Exemplo: estimação do total e AAS

- O plano amostral p_2 fornece o *estimador não viciado com menor variância* em comparação com o plano p_1 e deve ser preferido, pois o tamanho das amostras (nossa medida de custo) é o mesmo.

- **Minimizar a variância** é o critério de desempate para escolha entre *estratégias não viciadas de amostragem e estimação* de *igual custo total*.

Este será então nosso segundo critério para escolha de estimadores. 👍

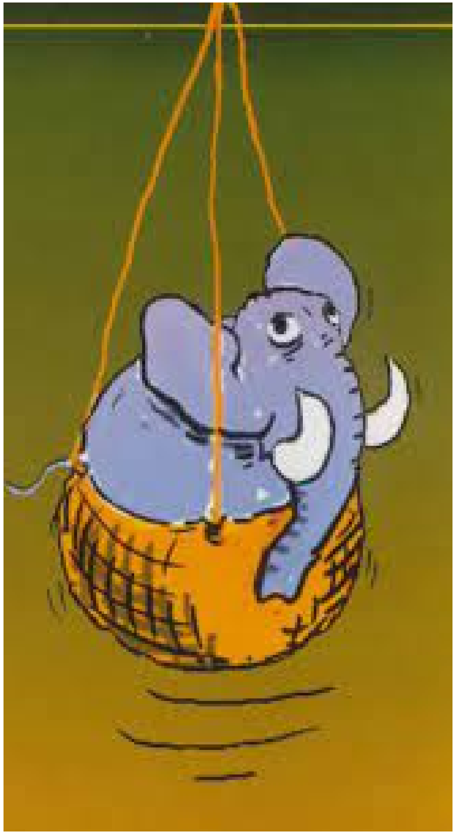
Para casa

- Compare o EQM dos estimadores t e $2t$, sob o plano p_1, no Exemplo.
- Verificar se todas unidades populacionais têm probabilidade positiva de inclusão na amostra sob p_2, no Exemplo.
- Qual o peso w_i para AASs e AASc?
- Rever os slides.
- Refazer o Exemplo.
- Ler seção 3,7 do livro 'Amostragem: Teoria e Prática Usando R'.

Próxima aula

- Teoria básica
  - Estimador linear do total
  - Propriedados do estimador Horvitz-Thompson
- Laboratório de R

Muito obrigado!



Fonte: imagem do livro *Combined Survey Sampling Inference: Weighing of Basu's Elephants: Weighing Basu's Elephants*.

Resumo da notação

Notação	População	Amostra
Índice (rótulo)	$U = \{1, 2, \ldots, i, \ldots, N\}$	$s = \{i_1, i_2, \ldots, i_n\}$
Característica	$Y_U = \{y_1,\ y_2,\ \ldots,\ y_i,\ \ldots,\ y_N\}$	$Y_s = \{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\}$
Total	$T = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i$	$\widehat{T} = t(s) = t = \sum_{i \in s} y_i$
Média	$\overline{Y} = \frac{T}{N} = \frac{1}{N} \sum_{i \in U} y_i$	$\widehat{\overline{Y}} = \overline{y} = \frac{t(s)}{n} = \frac{1}{n} \sum_{i \in s} y_i$
Variância	$Var_y = \frac{1}{N} \sum_{i \in U} (y_i - \overline{Y})^2$	$var_y = \frac{1}{n} \sum_{i \in s} (y_i - \overline{y})^2$
Variância	$S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \overline{Y})^2$	$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \overline{y})^2$

- Espaço amostral: $S = \{s_1, s_2, \ldots, s_j, \ldots, s_\nu\}$
- Plano amostral: $p(s)$, em que $\sum_{s \in S} p(s) = 1$
- Esperança em relação a $p(s)$: $E_p[t(s)] = \sum_{s \in S} t(s) p(s)$
- Variância em relação a $p(s)$: $Var_p[t(s)] = \sum_{s \in S} [t(s) - E_p(t)]^2 p(s)$