MAT02036 - Amostragem 2

Aula 02 - Teoria Básica

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

Aula passada 📀

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 $p(s)$: $Var_p[t(s)] = \sum_{s \in S} [t(s) - E_p(t)]^2 p(s)$

Aula passada 💿

- Vimos que trabalhar com a distribuição $p(s)$ é complicado.

  - O número total, $\nu = \binom{N}{n}$, tamanho do conjunto S cresce muito rapidamente com N e com n.
  - Então trabalhamos com a **probabilidade de inclusão** da unidade i, π_i .

- **Probabilidade de inclusão** (de primeira ordem)

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

- **Estimador linear** do total populacional (não viesado sob π):

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

Teoria Básica e AAS

Teoria básica

- Vamos olhar um pouco mais para a ideia de trabalhar com **propabilidades de inclusão** π_i.

- π_i pode ser vista como o parâmetro da distribuição de probabilidades da variável aleatória R, para a i-ésima unidade.

- Definimos a variável indicadora R_i tal que

$$R_i = \begin{cases} 1, & i \in \boldsymbol{s} \\ 0, & i \notin \boldsymbol{s} \end{cases}$$

para todo $i \in U$.

- A variável R_i é indicadora do evento 'inclusão da unidade i na amostra s'.

Teoria básica

Exemplo: estimação do total e AAS

- Para $N = 4$ e $n = 2$, as seis amostras possíveis podem ser representadas pelas indicadoras por

Representação de cada amostra possível pelas variáveis indicadoras

Amostra	Unidades na Amostra	R_1	R_2	R_3	R_4
1	$s_1 = \{1; 2\}$	1	1	0	0
2	$s_2 = \{1; 3\}$	1	0	1	0
3	$s_3 = \{1; 4\}$	1	0	0	1
4	$s_4 = \{2; 3\}$	0	1	1	0
5	$s_5 = \{2; 4\}$	0	1	0	1
6	$s_6 = \{3; 4\}$	0	0	1	1

- Cada amostra fica univocamente determinada pelas variáveis indicadoras $R_1, R_2, \ldots, R_N$ correspondentes.

Teoria básica

- As variáveis indicadoras R dependem da amostra s,
  - não indicamos explicitamente em nossa notação, mas temos que

$$\pi_i(s) = P(i \in s) = \sum_{s \ni i} p(s) = P(R_i = 1) = E_p(R_i), \forall i \in U$$

- Relembre: as **probabilidades de inclusão** π_i são ditas de **primeira ordem**.

- Sob essa ótica, precisamos também definir **probabilidades de inclusão de segunda ordem**, denotadas π_{ij}, dadas por

$$\pi_{ij} = P\left[(i,j) \in s\right] = \sum_{s \ni (i,j)} p(s) = P\left(R_{ij} = 1\right) = E_p\left(R_{ij}\right), \forall (i,j) \in U,$$

  em que $R_{ij} = R_i R_j$.

- Note que quando $i = j$, $\pi_{ij} = \pi_{ii} = \pi_i, \forall i \in U$.

Teoria básica

- Além da propriedade de valor esperado das variáveis aleatórias indicadoras R_i, pode-se também deduzir que:

  $$Var_p(R_i) = \pi_i(1 - \pi_i)$$

  e

  $$Cov_p(R_i, R_j) = \pi_{ij} - \pi_i\pi_j.$$

- Um método geral de prova em amostragem se baseia num uso inteligente das variáveis indicadoras $R_1, R_2, \ldots, R_N$.

- Uma propriedade importante dessas variáveis indicadoras é que:

  $$\sum_{i \in s} R_i = \sum_{i \in U} R_i.$$

Teoria básica

- Segue também que:

$$\sum_{i \in s} y_i = \sum_{i \in s} R_i y_i = \sum_{i \in U} R_i y_i.$$

- Convertemos a soma amostral que,
  - antes de selecionada a amostra, tem parcelas aleatórias,
  - em uma soma na população, onde as parcelas são conhecidas mas dependem das R_i.

Estimador linear do total

Estimador linear do total

- Considere o total populacional $T = \sum_{i \in U} y_i$ como parâmetro alvo;
- Um **estimador linear** de T é sempre da forma

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

onde w_i é o *peso amostral* da unidade i.

- Para que o estimador linear $\widehat{Y}_w$ de Y seja **sempre** não viciado, é preciso que:

$$E_p\left(\widehat{T}_w\right) = T \Leftrightarrow \sum_{i \in U} E_p\left(R_i\right) w_i y_i = \sum_{i \in U} y_i \Leftrightarrow \sum_{i \in U} \pi_i w_i y_i = \sum_{i \in U} y_i.$$

- Esta relação só será válida para quaisquer valores populacionais y_i da variável de pesquisa caso $\pi_i \times w_i = 1, \forall i \in U$.

Estimador linear do total

Exemplo: AAS sem reposição

Considere o plano amostral AASs para estimar o total populacional T usando o estimador $\widehat{T}_w = \widehat{T}_{AASs}$.

A. Calcule as probabilidades de inclusão de primeira ordem π_i.

B. Calcule as probabilidades de inclusão de segunda ordem π_{ij}.

C. Mostre que $\widehat{T}_{AASs}$ é não viesado para T

- C.1. usando o plano amostral $p(s)$.
- C.2. usando a probabilidade de inclusão π.

Estimador linear do total

Exemplo: AAS sem reposição

A. Temos que para a AASs $p(s) = ?$, para todo $s \in S$,

- pois o número de possíveis amostra é dado por $\nu = ?$...
- então $\pi_i = ?$

B. $\pi_{ij} = ?$

C.1. Olhar Cochran... ou slides Prof. Rodrigo.

C.2. Olhar propriedades do estimador linear acima.

Estimador linear do total

Exemplo: AAS sem reposição

- Qual a distribuição amostral do estimador do total $\widehat{T}_w$?
  - Ou da média $\overline{y}_w = \overline{\widehat{T_w}}$?
- TCL para amostras de populações finitas?
  - $\sqrt{n}(\widehat{T}_w - T) \overset{d}{\rightarrow} ?$
  - condições? quando $N \to \infty$, $n \to \infty$, $f = \frac{n}{N}$ limitado menor que 1?

Ver Bolfarine e Bussab capítulo 10, Cochran , Slides do Prof. Rodrigo, …

Quem tiver interesse em aspectos teóricos podemos revisar!

Estimador linear do total

Estimador Horvitz-Thompson

- Com pesos básicos d_i, o estimador não viciado de total fica dado pelo conhecido *estimador de Horvitz-Thompson* ou *estimador HT*:

$$\widehat{T}_{HT} = \sum_{i \in s} d_i y_i = \sum_{i \in s} {\pi_i}^{-1} y_i = \sum_{i \in s} y_i / \pi_i$$

- Assim, o estimador linear do total $\widehat{T}_w = \sum_{i \in s} w_i y_i$ será **sempre** não viciado se:
    - $w_i = {\pi_i}^{-1} = 1/\pi_i = d_i, \forall i \in U$.
    - os pesos amostrais d_i são chamados de **pesos básicos** do plano amostral.
    - outros pesos além dos definidos pelo delineamento, d_i, podem ser úteis na prática.
    - A notação w_i é reservada para designar pesos genéricos que podem ser aplicados para a obtenção de estimadores (viciados ou não).

Estimador linear do total

Estimador Horvitz-Thompson

- Este estimador está definido para qualquer
  - variável de pesquisa y e
  - para qualquer *plano amostral probabilístico* p, ou $\pi_i > 0, \forall i \in U$.
- **Amostragem probabilística** de populações finitas nos garante certa confiança de sempre dispor de estimadores não viciados como o HT.
- Lembrando: o estimador **HT** faz uso das probabilidades de inclusão π (implicadas pelo plano amostral $p(s)$),
  - mas depende através das probabilidades de inclusão de primeira ordem π_i,
  - uma condição geralmente simples de satisfazer na prática da pesquisa.

Estimador linear do total

Estimador Horvitz-Thompson

Propriedades do estimador de Horvitz-Thompson

O *estimador de Horvitz-Thompson* é *não viciado* para estimar o total, ou seja, $E_p(\widehat{T}_{HT}) = Y$.

Prova:

$$E_p(\widehat{T}_{HT}) = E_p\left[\sum_{i \in U} R_i y_i / \pi_i\right] = \sum_{i \in U} \left[E_p(R_i) y_i / \pi_i\right] = \sum_{i \in U} y_i = Y$$

Esta propriedade vale para qualquer população U, variável de interesse y e plano amostral p, desde que $\pi_i > 0, \forall i \in U$.

Estimador linear do total

Estimador Horvitz-Thompson

Propriedades do estimador de Horvitz-Thompson

A variância do estimador Horvitz-Thompson para o total é dada por:

$$
\begin{aligned}
Var_{HT}(\widehat{T}_{HT}) &= \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j \\
&= \sum_{i \in U} \sum_{j \in U} \left(\frac{d_i d_j}{d_{ij}} - 1 \right) y_i y_j
\end{aligned}
$$

onde $d_{ij} = \pi_{ij}^{-1}$.

Esta é a chamada forma de Horvitz-Thompson da variância. Existe uma outra forma para esta variância, que vamos conhecer mais adiante.

Estimador linear do total

Estimador Horvitz-Thompson

Propriedades do estimador de Horvitz-Thompson

Prova:

$$
\begin{aligned}
Var_{HT}(\widehat{T}_{HT}) &= Var_p\left(\sum_{i \in U} R_i \frac{1}{\pi_i} y_i\right) \\
&= \sum_{i \in U} \sum_{j \in U} Cov_p(R_i, R_j) \left(\frac{y_i}{\pi_i}\right) \left(\frac{y_j}{\pi_j}\right) \\
&= \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} \frac{y_j}{\pi_j}\right) \\
&= \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) y_i y_j \\
&= \sum_{i \in U} \sum_{j \in U} \left(\frac{d_i d_j}{d_{ij}} - 1\right) y_i y_j
\end{aligned}
$$

Estimador linear do total

Estimador Horvitz-Thompson

Propriedades do estimador de Horvitz-Thompson

Um estimador não viciado da variância do estimador **HT** do total é dado por:

$$\begin{aligned}\widehat{V}ar_{HT}(\widehat{T}_{HT}) &= \sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right)\\ &= \sum_{i\in s}\sum_{j\in s}\left(d_i d_j - d_{ij}\right) y_i y_j\end{aligned}$$

- Este estimador da variância foi obtido usando o princípio dos estimadores tipo Horvitz-Thompson do total.
  - Agora, como estimamos uma soma dupla na população, os pesos das parcelas nessa soma dependem das probabilidades de inclusão de **segunda ordem** π_{ij}.
  - Para ser viável, $p(s)$ tem que satisfazer a condição adicional de que $\pi_{ij} > 0 \forall i \neq j \in U$ (estritamente positivas).

Considerações

Estimador linear do total

Comentários sobre estimação de totais e respectivas variâncias em **amostragem probabilística**:

- É possível sempre **estimar sem vício um total populacional** usando uma soma amostral π-ponderada, o estimador **HT** do total.

- Expressões de variância para **avaliar a qualidade do estimador de total** sob distintas situações (população, variável) para qualquer plano amostral.

- Estimar muitos **outros parâmetros populacionais** (tais como médias, proporções e razões) com os resultados vistos na estimação de totais.

- Derivar estimadores não viciados do total populacional e da variância do estimador **HT** de total para **distintos planos amostrais** como **casos especiais** da teoria geral apresentada.

    - conveniente para a **estimação de variâncias**, cujas expressões gerais dependem de somas duplas difíceis de calcular para n grande.

    - **expressões para cada um dos planos amostrais** específicos são úteis porque permitem simplificar os cálculos da estimação de variâncias.

Estimador linear do total

- Em **planos amostrais equiponderados** (em que as probabilidades de inclusão π_i são todas iguais);
  - os pesos w_i para estimação de médias ficam todos iguais a $1/n$;
  - uma vantagem pois a tarefa de estimação fica simplificada.
- Estimadores HT do total, média e respectivas variâncias (N conhecido):

Estimadores HT	**Variâncias dos Estimadores HT**
$\widehat{T}_{HT} = \sum_{i \in s} d_i y_i = \sum_{i \in s} y_i / \pi_i$	$\widehat{Var}_{HT}(\widehat{T}_{HT}) = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \right)$
$\overline{y}_{HT} = \widehat{T}_{HT}/N = \sum_{i \in s} d_i y_i / N$	$\widehat{Var}_{HT}(\overline{y}_{HT}) = \widehat{Var}_{HT}(\widehat{T}_{HT})/N^2$

Quando N **não for conhecido**, podemos usar o **estimador de razão**
$\overline{y}^R = \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i} = \sum_{i \in s} w_i^R y_i$ e $\widehat{Var}_{HT}(\overline{y}^R) = \frac{1}{\widehat{N}_{HT}^2} \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \left(\frac{y_i - \overline{y}^R}{\pi_i} \right) \left(\frac{y_j - \overline{y}^R}{\pi_j} \right)$

Expressão alternativa para a variância - Sen-Yates-Grundy
$\widehat{Var}_{SYG}(\widehat{Y}_{HT}) = \sum_{i \in s} \sum_{j > i} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$

Laboratório de

Exercício 🏃‍♀️

Considere a população com $N = 6$ domicílios listada com os respectivos valores de variáveis de interesse.

Valores de variáveis de interesse para cada domicílio da população

Domicílio	Renda (R$)	NO. de Moradores	No. de Trabalhadores
1	800	2	2
2	4.200	4	3
3	1.600	2	1
4	500	2	1
5	900	4	2
6	2.000	1	1
Total	10.000	15	10

Exercício 🏃‍♀️

1. Para cada variável de interesse (Renda, Número de Moradores e Número de Trabalhadores), calcule os seguintes parâmetros populacionais: total, média e variância.

2. Liste o conjunto S de todas as amostras possíveis de tamanho $n = 2$ da população, considerando apenas **amostras de unidades distintas**.

3. Supondo que todas as amostras listadas no conjunto S são **equiprováveis** (Plano A), calcule:

- As probabilidades de inclusão das unidades.
- As probabilidades de inclusão dos pares de unidade.
- Os valores possíveis para o estimador Horvitz-Thompson do total populacional para a variável Renda.
- O valor esperado e a variância para o estimador Horvitz-Thompson do total populacional para a variável Renda.

Exercício

4. Considere agora que o conjunto S é formado somente pelas amostras $(1;2),(2;3),(2;4),(2;5)e(2;6)$, tendo cada uma delas probabilidade 1/5 de ser a amostra selecionada (Plano B). Repita os cálculos do item 3 para o novo plano amostral.

5. Faça gráficos dos valores possíveis do estimador de total sob os dois planos amostrais para comparar as respectivas distribuições.

6. Use os resultados obtidos em 3 e 4 para comparar os dois planos amostrais e indique qual deles seria preferível usar, caso fosse necessário amostrar duas unidades distintas da população $(n=2)$ para estimar o total da Renda. Justifique.

Exercício 🏋️‍♀️

```r
## dados da populacao e plano amostral
N <- 6                      # 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                 # AAS

## variaveis
renda <- c(800, 4200, 1600, 500, 900, 2000)
moradores <- c(2, 4, 2, 2, 4, 1)
trabalhadores <- c(2, 3, 1, 1, 2, 1)
```

Para casa

- Continuar o Exemplo
- Continuar o Exercício
- Rever os slides.
- Ler seção 11.1 a 11.3 do livro 'Amostragem: Teoria e Prática Usando R'.

Próxima aula

- Amostragem Estratificada
  - Características
  - Parâmetros
  - Estimadores

Muito obrigado!

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

Amostragem aleatória simples

Amostragem aleatória simples COM reposição

- Na **Amostragem Aleatória Simples com reposição** (AASc) as unidades da população têm a mesma chance de ser incluídas na amostra em cada sorteio, e essa probabilidade é igual a $1/N$.

- **Plano amostral**

  - Existem N^n amostras distintas em S, então

$$p(s) = 1/N^n, \forall\, s \in S.$$

- **Probabilidades de inclusão**:

  - $\pi_i \; = \; P\,(i \in s) \; = \; 1 - P\,(i \notin s) \; = \; 1 - \left(1 - \frac{1}{N}\right)^n.$

  - $\pi_{ij} = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n$, para $i, j = 1, \ldots, N.$

Amostragem aleatória simples COM reposição

- A variável Q_i denota a 'qtd. de vezes a unidade i aparece na amostra s',

$$Q_i \sim Binomial(n, 1/N)$$

  - $E_{AAS}[Q_i] = n\frac{1}{N}$
  - $Var_{AAS}[Q_i] = n\frac{1}{N}\left(1 - \frac{1}{N}\right)$
  - $Cov_{AAS}[Q_i, Q_j] = -\frac{n}{N^2}$ (*propriedade da multinomial*)

Estimador não viciado (**ENV**) para:

- o **total** populacional T: $\widehat{T}_{AASc} = N\overline{y} = N\sum_{i \in s}\frac{y_i}{n}$
- a **média** populacional $\overline{Y}$: $\widehat{\overline{Y}}_{AASc} = \frac{1}{n}\sum_{i \in s} y_i = \overline{y}$
- a **variância** populacional Var_y: $\widehat{Var}_{y,AASc} = \frac{1}{n-1}\sum_{i \in s}(y_i - \overline{y})^2 = \widehat{S}_y^2$

Amostragem aleatória simples COM reposição

- **Variância** dos **Estimadores**
  - $Var_{AASc}(\widehat{T}_{AASC}) = N^2 Var_y/n$
  - $Var_{AASc}(\overline{y}) = Var_y/n$

em que $Var_y = \frac{1}{N}\sum_{i\in U}(y_i - \overline{Y})^2 = \frac{N-1}{N}S^2$

- **ENV** da **Variância** dos **Estimadores**
  - $\widehat{Var}_{AASc}(\widehat{T}_{AASC}) = N^2 \widehat{S}_y^2/n$
  - $\widehat{Var}_{AASc}(\overline{y}) = \widehat{S}_y^2/n$

em que $\widehat{S}_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \overline{y})^2 = s_y^2$.

Amostragem aleatória simples SEM reposição

- Na **Amostragem Aleatória Simples sem reposição** (**AASs**) cada unidade da população pode aparecer na amostra no máximo uma única vez.

- **Plano amostral** sob **AASs**

  - Existem $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ amostras distintas em S, então

$$p(s) = 1/\binom{N}{n}, \forall\, s \in S.$$

- **Probabilidades de inclusão** sob **AASs**

  - $\pi_i = n/N > 0, \forall\, i \in U$, desde que $n > 0$.

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

  - Estimação de variância sem vício requer $\pi_{ij} > 0, \forall\, i, j \in U$.

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} > 0, \forall i \neq j \in U.$$

  - Sob **AASs**, as probabilidades de inclusão π_i e π_{ij} não dependem de i ou j, e essa é a razão da simplicidade desse plano amostral.

Amostragem aleatória simples SEM reposição

- A variável R_i, indicadora do evento 'inclusão da unidade i na amostra s', sob **AASs**
  - $E_{AAS}[R_i] = \frac{n}{N}$
  - $Var_{AAS}[R_i] = \frac{n}{N}\left(1 - \frac{n}{N}\right)$
  - $Cov_{AAS}[R_i, R_j] = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(1 - \frac{n}{N}\right)\left(-\frac{1}{N-1}\right)$

Estimador não viciado **(ENV)** para:

- o **total** populacional T: $\widehat{T}_{HT} = \sum_{i \in s} \frac{y_i}{n/N} = \frac{N}{n} \sum_{i \in s} y_i = N\overline{y} = \widehat{T}_{AAS}$
- a **média** populacional $\overline{Y}$: $\widehat{\overline{Y}}_{AAS} = \frac{1}{n} \sum_{i \in s} y_i = \overline{y}$
- a **variância** populacional Var_y: $\widehat{Var}_{y,AAS} = \frac{N-1}{N} \widehat{S}_y^2$,

pois $\widehat{S}_y^2$ é **ENV** de S_y^2 na $AASs$.

Amostragem aleatória simples SEM reposição

- **Variância** dos **Estimadores**
  - $Var_{AAS}(\widehat{T}_{AASs}) = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n} = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S^2$
  - $Var_{AAS}(\overline{y}) = \left(1-\frac{n}{N}\right)\frac{S^2}{n} = \left(\frac{1}{n}-\frac{1}{N}\right)S^2$

onde $S_y^2 = \frac{1}{N-1}\sum_{i\in U}(y_i-\overline{Y})^2$, como já definido.

- **ENV** da **Variância** dos **Estimadores**
  - $\widehat{Var}_{AAS}(\widehat{T}_{AAS}) = N^2\left(1-\frac{n}{N}\right)\frac{\widehat{S}_y^2}{n} = N^2\left(\frac{1}{n}-\frac{1}{N}\right)\widehat{S}_y^2$
  - $\widehat{Var}_{AAS}(\overline{y}) = \left(\frac{1}{n}-\frac{1}{N}\right)\widehat{S}_y^2$

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

Amostragem aleatória simples SEM reposição

Considerações

1. O termo $(1-n/N)=(1-f)$ é chamado de **fator de correção para população finita**.

   - Quando $n/N \to 1$, o tamanho da amostra se aproximando do tamanho da população, então $(1-n/N) \to 0$.

   - Ou seja: com amostras grandes as variâncias das estimativas tendem a ser pequenas.

2. Se a fração amostral $f=n/N$ for pequena (da ordem de 1% ou 2%), então a **correção de população finita** pode ser ignorada, pois $(1 - f) \doteq 1$.

   - Quando $f \doteq 0$, a AASse AASc (com reposição) tem comportamento semelhante em relação à precisão das estimativas.

   - *Intuitivamente*, sempre que n for muito **pequeno** em relação ao N a probabilidade de uma unidade i da população ser selecionada mais de uma vez é pequena.

Amostragem aleatória simples SEM reposição

Distribuição da média amostral

- Repetições do plano amostral $p(s)$ segundo $AASs$, $\bar{y}$ tem uma **distribuição de probabilidades exata**, que **depende**:
  - da distribuição de y na população,
  - do tamanho da amostra n e
  - do plano amostral $p(s)$, que neste caso, é AASs.
- 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* (Hajeck, 1960) sugere uma aproximação

$$\frac{\bar{y} - E_{AAS}(\bar{y})}{\sqrt{Var_{AAS}(\bar{y})}} = \frac{\bar{y} - \bar{Y}}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}} \approx Normal(0;1),$$

onde $Normal(0;1)$ denota uma variável aleatória com distribuição normal padrão.

[*] Para detalhes ver Cochran(1977), Seções 2.8 e 2.15, ou Sarndal(1992), Seção 2.11.

Amostragem aleatória simples SEM reposição

Distribuição da média amostral

- Podemos construir **intervalos de confiança**(IC) para $\overline{Y}$.
  - Um IC de nível $(1-\alpha)$ para $\overline{Y}$ é dado por

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

    onde $z_{\alpha/2}$ é o quantil $1-\frac{\alpha}{2}$, que deixa área $\alpha/2$ à sua direita.

- A **semiamplitude** do IC fornece uma ideia da **margem de erro** que se tem ao estimar o parâmetro.

$$\widehat{ME}_{AAS}(\overline{y}) = z_{\alpha/2}\sqrt{\widehat{Var}_{AAS}(\overline{y})}.$$

  - A **margem de erro** pode ser **estimada** a partir da amostra selecionada e observada.
  - **Amostragem probabilística** fornece indicativos da **incerteza associada a estimativas**, além de estimativas pontuais.

Resumo da notação

Estimadores AASc	Estimadores AASs
$\widehat{T}_{AASc} = \frac{N}{n} \sum_{i \in s} y_i = N\overline{y}$	$\widehat{T}_{AASs} = \frac{N}{n} \sum_{i \in s} y_i = N\overline{y}$
$\overline{y} = \frac{1}{n} \sum_{i \in s} y_i$	$\overline{y} = \frac{1}{n} \sum_{i \in s} y_i = \widehat{\overline{T}}_{AASs}$
$\widehat{Var}_{AASc}(\widehat{T}_{AASC}) = N^2 \widehat{S}_y^2/n$	$\widehat{Var}_{AASs}(\widehat{T}_{AASs}) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \widehat{S}_y^2$
$\widehat{Var}_{AASc}(\overline{y}) = \widehat{S}_y^2/n$	$\widehat{Var}_{AASs}(\overline{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) \widehat{S}_y^2$

em que $\widehat{S}_y^2 = \frac{1}{n-1} \sum_{i \in s}(y_i - \overline{y})^2$.

Estimadores HT	Variâncias dos Estimadores HT
$\widehat{T}_{HT} = \sum_{i \in s} d_i y_i = \sum_{i \in s} y_i/\pi_i$	$\widehat{Var}_{HT}(\widehat{T}_{HT}) = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} \frac{y_j}{\pi_j}\right)$
$\overline{y}_{HT} = \widehat{T}_{HT}/N = \sum_{i \in s} d_i y_i/N$	$\widehat{Var}_{HT}(\overline{y}_{HT}) = \widehat{Var}_{HT}(\widehat{T}_{HT})/N^2$

Referências

Slides baseados nos Capítulos 3 e 4 do livro

- Amostragem: Teoria e Prática Usando o R

Citações do Capítulo

- Cochran(1977)
- Fuller(2009)
- Hajeck(1960)
- Horvitz(1952)
- Sarndal(1992)
- Sen(1953)
- Yates(1953)
- Yates e Grundy (1953)