MAT02036 - Amostragem 2

Aula 20 - Amostragem Sistemática - Parâmetros e Estimação

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

Amostragem Sistemática Simples

- *O método*: selecionar cada K-ésima unidade da população;
  - N - tamanho da população;

$$N = nK + c, \quad 0 \leq c < K$$

  - K - **intervalo de selação**;
  - $n = \lfloor N/K \rfloor$ tamanho da amostra;
  - c - é o resto da divisão N/K;
  - r - **valor inicial**, número inteiro de 1 a K,

$$r \sim Uniforme - Discreta(K);$$

- Na **AS** a amostra $s_r = \{i : i = r + lK \leq N; \;\; l = 0, \ldots, n\}$, satisfaz

$$p(s) = \begin{cases} 1/K, \text{ se } \; s = s_r \text{ para } r = 1, \;\; 2, \;\; \ldots, \;\; K \\ 0, \;\; \text{caso contrário} \end{cases}$$

Aula passada

Amostragem Sistemática Simples

- A probabilidade de inclusão na amostra de uma unidade i qualquer é dada por:

$$\pi_i = \frac{1}{K}, \; i = 1, \ldots, \; N$$

- A probabilidade de inclusão das unidades $i \neq j$ na amostra é dada por:

$$\pi_{ij} = \begin{cases} 1/K, & \text{se } i \neq j \in s_r \text{ para } r = 1, \ldots, K \\ 0, & \text{caso contrário} \end{cases}$$

- As variáveis indicadoras associadas às amostras possíveis s_r:

$$I(r) = \begin{cases} 1, & \text{se a amostra é } s_r \text{ para } 1 \leq r \leq K \\ 0, & \text{caso contrário} \end{cases}$$

- O valor esperado de $I(r)$ é

$$E_{AS}[I(r)] = 1/K\,, \; r = 1, \ldots, K,$$

Estimação

Amostragem Sistemática Simples

Estimação de totais na **AS**

- O estimador tipo *Horvitz-Thompson* do total $T = \sum_{i=1}^{K} T_i$ sob **AS**,
  - o peso amostral das unidades da amostra é sempre igual a $d_i = 1/\pi_i = K$, então

$$\widehat{T}_{AS} = Kt_r = K \sum_{i \in s_r} y_i$$

  em que $t_r = \sum_{i \in s_r} y_i$ é a soma amostral dos valores observados da variável y.

- Já sabemos que este estimador é não viciado para o total populacional.

$$\begin{aligned} E_{AS}(\widehat{T}_{AS}) &= E_{AS}\left[Kt_r\right] = KE_{AS}\left[\sum_{r=1}^{K} I(r)t_r\right] \\ &= K\sum_{r=1}^{K} E_{AS}\left[I(r)\right] t_r = K\sum_{r=1}^{K} \frac{1}{K} t_r = \sum_{r=1}^{K} t_r = T \end{aligned}$$

Amostragem Sistemática Simples

Estimação de totais na **AS**

Exemplo 1

Considere a população composta de $N = 19$ unidades, cujos dados da variável de interesse y, da qual se deseja retirar uma amostra sistemática simples com intervalo de seleção com $K = 4$ para estimar o total populacional. Verifique numericamente que o estimador $\widehat{T}_{AS}$ é não viciado.

s_1	s_2	s_3	s_4
99	54	96	54
85	88	55	83
62	85	96	55
91	92	67	68
54	79	76	

(Obs. Para mostrar que o estimador é não viciado, basta verificar que a média dos seus valores possíveis é igual ao parâmetro populacional a ser estimado.)

Amostragem Sistemática Simples

Estimação de totais na AS

Exemplo 1

```r
## exemplo 1
K <- 4                              # Intervalo de seleção
pop <- matrix(c(99,54,96,54,85,88,55,83,62,85,96,55,91,92,67,68,54,79
That_r <- K * colSums(pop, na.rm = TRUE)   # estimativas para cada amo
That_r
```

```
## [1] 1564 1592 1560 1040
```

```r
EThat <- mean(That_r)               # media das estimativas de total
EThat
```

```
## [1] 1439
```

```r
T <- sum(pop,na.rm=T)               # total populacional T
T
```

Amostragem Sistemática Simples

Estimação de médias na **AS**

- Para estimar a média populacional $\overline{Y} = \frac{Y}{N} = \frac{\sum_{r=1}^{K} t_r}{\sum_{r=1}^{K} n_r}$ um estimador não viciado (❓) é dado por (quando N é conhecido)

$$\overline{y}_{AS} = \frac{\widehat{Y}_{AS}}{N} = \frac{Kt_r}{N}.$$

- $\overline{y}_{AS}$ é não viciado para $\overline{Y}$, pois $\widehat{T}_{AS}$ é não viciado para T.

- Note: $\overline{y}_{AS} \neq \overline{y}$ (média amostral), a menos que $N = nK$. (❓)

Exemplo 2

Com a mesma população do Exemplo 1, verificar que $\overline{y}_{AS}$ é não viciado para a média populacional $\overline{Y}$. *(Obs. podem ser usados os totais já estimados para cada coluna (amostra sistemática possível) da tabela, então calcular a média das estimativas e comparar com média populacional)*

Amostragem Sistemática Simples

Estimação de médias na **AS**

Exemplo 2

```
## exemplo 2
N <- sum(pop, na.rm = TRUE)    # tamanho da populacao
ybarAS_r <- That_r/N           # estimativas para cada amostra possive
ybarAS_r
```

```
## [1] 1.0868659 1.1063238 1.0840862 0.7227241
```

```
EybarAS <- mean(ybarAS_r)      # media das estimativas
EybarAS
```

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

```
YbarAS <- mean( pop, na.rm=T) # media populacional
YbarAS
```

```
## [1] 75.73684
```

Amostragem Sistemática Simples

Estimação de médias na AS

- Quando N é desconhecido, uma alternativa é o estimador do tipo razão

$$\overline{y}_{AS} = \frac{\widehat{Y}_{AS}}{\widehat{N}_{AS}} = \frac{Kt_r}{Kn_r} = \frac{t_r}{n_r} = \overline{y}_r = \overline{y}$$

Exemplo 3

Ainda com a mesma população, verificar que a média amostral $\overline{y}$ não coincide com $\overline{y}_{AS}$ e, além disso, é viciado.

```
## exemplo 3
ybar_r <- colMeans( pop, na.rm=T)  # estimativas para cada amostra po
ybar_r
```

```
## [1] 78.2 79.6 78.0 65.0
```

```
Eybar <- mean(ybar_r)                # media das estimativas
Eybar
```

Amostragem Sistemática Simples

Estimação de médias na **AS**

Verifica-se assim que a média amostral simples é um estimador para uma razão, sendo portanto viciado para estimar a média populacional. Tal estimador só será exatamente não viciado quando $N = nK$, pois:

$$\begin{aligned}
E_{AS}(\overline{y}_{AS}) = E_{AS}(\overline{y}) &= E_{AS}\left[\sum_{r=1}^{K} I(r)\overline{y}_r\right] \\
&= \frac{1}{K}\sum_{r=1}^{K}\overline{y}_r = \frac{1}{K}\sum_{r=1}^{K}\frac{t_r}{n_r} \\
&\neq \frac{\sum_{r=1}^{K} t_r}{\sum_{r=1}^{K} n_r} = \overline{Y}
\end{aligned}$$

O vício desse estimador (quando $N \neq nK$) é o preço pago quando não se conhece o tamanho N da população!

Amostragem Sistemática Simples

Estimação de uma proporção na **AS**

- Assumindo a variável indicadora

$$y_{ij} = I\left[(i,j) \in A\right] = \begin{cases} 1, & \text{se a unidade } i \text{ possui o atributo, } A \subset U; \\ 0, & \text{caso contrário.} \end{cases}$$

- Se N for conhecido, um estimador não viciado para a proporção populacional P é dado por:

$$\widehat{P}_{AS} = \frac{K}{N}\sum_{i \in s_r} y_i = \frac{K}{N} t_r = \frac{K}{N} n_a$$

onde n_a é o número de unidades na amostra com o atributo de interesse.

Exemplo 4

Verificar numericamente que o estimador $\widehat{P}_{AS}$ para a proporção P é não viciado quando N é conhecido.

Amostragem Sistemática Simples

Estimação de uma proporção na **AS**

```r
## exemplo 4
K <- 4                    # intervalo de selecao
pop <- matrix(c(0,1,0,1,1,1,0,0,0,0,1,0,0,0,1,1,1,0,1,NA),5,K,byrow=T
N <- sum(!is.na(pop))   # tamanho da populacao
PAS <- K * colSums( pop, na.rm = TRUE) / N  # estimativas para cada p
PAS
```

```
## [1] 0.4210526 0.4210526 0.6315789 0.4210526
```

```r
EPAS <- mean(PAS) # media das estimativas
EPAS
```

```
## [1] 0.4736842
```

```r
P <- mean( pop, na.rm=TRUE)    # proporcao populacional p
P
```

```
## [1] 0.4736842
```

Amostragem Sistemática Simples

Variância dos estimadores sob **AS**

A variância de $\widehat{Y}_{AS}$ sob amostragem sistemática simples é dada por:

$$
\begin{aligned}
Var_{AS}(\widehat{T}_{AS}) &= Var_{AS}\left[K\sum_{r=1}^{K} I(r)t_r\right] \\
&= K^2\left[\sum_{r=1}^{K} {t_r}^2 Var_{AS}[I(r)] + \sum_{r\neq K} COV_{AS}[I(r), I(q)]t_r t_q\right] \\
&= K^2\left[\sum_{r=1}^{K} {t_r}^2 \frac{1}{K}\left(1-\frac{1}{K}\right) + \sum_{r\neq K} t_r t_q\left(-\frac{1}{K^2}\right)\right] \\
&= K^2\left[\frac{1}{K}\sum_{r=1}^{K} {t_r}^2 - \frac{1}{K^2}\left(\sum_{r=1}^{K} {t_r}^2 + \sum_{r\neq K} t_r t_q\right)\right] \\
&= K\left[\sum_{r=1}^{K} {t_r}^2 - \left(\sum_{r=1}^{K} t_r\right)^2 / K\right] \\
&= K\sum_{r=1}^{K}(t_r - \bar{t})^2
\end{aligned}
$$

onde: $\bar{t} = \dfrac{1}{K}\displaystyle\sum_{r=1}^{K} t_r = \dfrac{T}{K}$

Amostragem Sistemática Simples

Variância dos estimadores sob **AS**

Portanto a variância é calculada a partir da soma de quadrados dos desvios entre totais das amostras possíveis em relacão à média destes totais.

Quando N é conhecido, a variância do estimador da média populacional é dada por:

$$Var_{AS}(\overline{y}_{AS}) = \frac{1}{N^2} Var_{AS}(\widehat{T}_{AS})$$

- Na **AS** ordenação da população em relação aos valores de y afeta a variância (precisão) dos estimadores.
    - O que ocorre se compararmos com uma estratégia **AAS**❓
    - E com uma estratégia de **AES** de mesmo tamanho nos estratos formados pela divisão de K intervalos de valores de y❓

Amostragem Sistemática Simples

Variância dos estimadores sob **AS**

Exemplo 5

Considere a população ordenada tal como foi apresentada no Exemplo 1:

a. calcular a variância do estimador do total considerando as possíveis amostras.

b. ordenar a população em ordem crescente (ou decrescente) dos valores de y e repetir o cálculo da variância.

c. observar que a variância do estimador do total em (a) e (b).

Obs. 1: Esse é um exemplo extremo mas ilustra o efeito da ordenação dos valores y na precisão dos estimadores na AS. Populações em que valores y seguem uma ordenação (ou aproximadamente), a AS pode ter um bom desempenho.

Obs. 2: Fica como exercício para o leitor verificar que o mesmo não ocorre quando se utiliza uma AAS. E com uma AES de mesmo tamanho em cada grupo. E com uma estratégica AES

Amostragem Sistemática Simples

Variância dos estimadores sob **AS**

Exemplo 5

```
## exemplo 5
## População na ordem natural
pop <- matrix(c(99,54,96,54,85,88,55,83,62,85,96,55,91,92,67,68,54,79
N <- sum(!is.na(pop)) # tamanho da população
tr <- colSums( pop, na.rm=T)
V_YhatAS <- K * (var(tr) * (K-1)) # variancia do estimador do total
V_YhatAS
```

```
## [1] 53219
```

```
## ordenando a populacao em ordem crescente de y
pop_ord <- matrix(sort(pop,na.last=T),5,K,byrow=T)
tr <- colSums( pop_ord, na.rm=T)
V_YhatAS_ord <- K * (var(tr) * (K-1))    # variancia do estimador do
V_YhatAS_ord
```

Amostragem Sistemática Simples

Variância dos estimadores sob **AS**

Notas:

1. Como r pode tomar apenas um valor, $Var_{AS}(\widehat{Y}_{AS})$ não pode ser diretamente estimada a partir da amostra.

2. Em @Cochran1977 é apresentada uma boa discussão sobre como a ordenação dos valores da variável de pesquisa para unidades populacionais pode afetar a eficiência de amostras sistemáticas.

3. Para populações em 'ordem aleatória', o desempenho da amostragem sistemática simples é semelhante ao da amostragem aleatória simples sem reposição (@Cochran1977, Seção 8.5).

4. Para populações com tendência linear, amostragem sistemática simples é melhor que AAS (@Cochran1977, Seção 8.6).

5. Para populações periódicas, amostragem sistemática simples com intervalo de seleção em sincronia com o período é um desastre (@Cochran1977, página 218).

Amostragem Sistemática Simples

Variância dos estimadores sob **AS**

- No caso especial onde $N = nK$, já sabemos que $\overline{y}_r = t_r/n_r$ é não viciado para $\overline{Y}$. então (@Cochran1977, página 207)

$$\sum_{r=1}^{K}\sum_{i\in s_r}(y_i-\overline{Y})^2 = n\sum_{r=1}^{K}(\overline{y}_r-\overline{Y})^2+\sum_{r=1}^{K}\sum_{i\in s_r}(y_i-\overline{y}_r)^2$$

tem-se que:

$$(N-1)S_y^2 = n\times K\times Var_{AS}(\overline{y}_{AS})+K(n-1)S_{dc}^2$$

onde S_y^2 é a variância populacional total, S_{dc}^2 é a variância *dentro* das amostras sistemáticas e $Var_{AS}(\overline{y}_{AS})$ é a variância de $\overline{y}_{AS}$ sob amostragem sistemática simples.

O estimador de média é mais eficiente sob amostragem sistemática que sob **AAS** se e somente se $S_{dc}^2 > S_y^2$ (@Cochran1977, página 208).

Amostragem Sistemática Simples

Variância dos estimadores sob **AS**

Expressão alternativa para $Var_{AS}(\overline{y}_{AS})$ quando $N = nK$: (Teorema 8.2 de @Cochran1977, página 209)

$$Var_{AS}(\overline{y}_{AS}) = \left(\frac{N-1}{N}\right)[1+(n-1)\rho]\,\frac{S_y^2}{n} \doteq [1+(n-1)\rho]\,\frac{S_y^2}{n},$$

onde

$$\rho_{int} = \frac{1}{(n-1)(N-1)S_y^2}\sum_{r=1}^{K}\sum_{i\neq j\in s_r}(y_i-\overline{Y})(y_j-\overline{Y})$$

é a *correlação intraclasse* das amostras sistemáticas possíveis.

- A correlação positiva entre unidades de uma mesma amostra aumenta a variância da média amostral na **AS** quando comparada com a **AAS**:

$$EPA(AS) = 1+(n-1)\rho_{int}\begin{cases} <1, & \text{se } \rho_{int}<0 \\ =1, & \text{se } \rho_{int}=0 \\ >1, & \text{se } \rho_{int}>0 \end{cases}$$

Amostragem Sistemática Simples

Estimação de variâncias dos estimadores sob **AS**

- Na **AS** não há um estimador não viciado para a variância dos estimadores do total e da média. O que se faz é utilizar estimadores mais adequados de acordo com a ordenação da população.

- Para N conhecido e sob a suposição de ordenação aleatória no cadastro de seleção em relação à(s) variável(eis) de interesse ($\overline{Y}_r \doteq constante$), se pode utilizar um estimador equivalente ao usado sob **AAS**,

$$\widehat{V}ar_{1AS}(\overline{y}_{AS}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i \in s_r}(y_i - \overline{y}_{AS})^2.$$

- Estimador não viciado caso a suposição de ordenação aleatória das unidades na população esteja correta.

- No caso de não se conhecer N, pode-se utilizar as alternativas dadas pelos estimadores de razão.

Amostragem Sistemática Simples

Estimação de variâncias dos estimadores sob **AS**

- No caso de N conhecido e não haer ordenação das unidades da população, um estimador para a variância do estimador do total é dado por:

$$\widehat{Var}_{1AS}(\widehat{T}_{AS}) = N^2 \widehat{Var}_{1AS}(\overline{y}_{AS})$$

- No caso em que a população esteja ordenada segundo uma "estratificação" de modo que as médias em cada intervalo de seleção variem (p.ex.: a população é ordenada segundo os valores de y), @Cochran1977 sugere, para estimar a variância do estimador da média, a expressão:

$$\widehat{V}ar_{2AS}(\overline{y}_{AS}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{2(n-1)} \sum_{i \in s_r} (y_i - y_{i+K})^2$$

Neste caso, um estimador para a variância do estimador do total é dado por:

$$\widehat{Var}_{2AS}(\widehat{T}_{AS}) = N^2 \widehat{Var}_{2AS}(\overline{y}_{AS})$$

Amostragem Sistemática Simples

Consideração finais

Alternativamente e independentemente da ordenação da população, pode-se usar um estimador do tipo replicação, onde são selecionadas q amostras sistemáticas de tamanhos n/q cada uma, tomando a variância das estimativas dadas por cada uma das amostras. Essa técnica, também chamada *amostra sistemática repetida*, está descrita em @Scheaffer2011.

Quando a seleção de uma AS for realizada a partir de um cadastro conhecido, é sempre possível reordenar as unidades aleatoriamente antes proceder a seleção. Esse é um artifício muito útil e que permite que se utilizem os estimadores equivalentes aos da AAS para estimar a variância dos estimadores. Se por um lado essa técnica viabiliza o emprego de estimadores simplificados de variância, por outro se espera que acabe resultando em menor precisão para a estimação pontual.

Pode-se encontrar boas discussões sobre a estimação da variância sob amostragem sistemática simples em @Cochran1977 ou @Thompson2012.

Amostragem Sistemática Simples

Estimadores do total, média e respectivas variâncias sob AS.

Estimador	Observação
$\widehat{T}_{AS} = Kt_r = K\sum_{r=1}^{K} I(r)t_r$	
$\overline{y}_{AS} = \frac{K}{N}t_r$	se N é conhecido
$\overline{y}_{AS} = \frac{t_r}{n_r} = \overline{y}$	se N é desconhecido
$\widehat{Var}_{1AS}(\widehat{T}_{AS}) = N^2\widehat{Var}_{1AS}(\overline{y}_{AS})$	se N é conhecido e sem ordenação
$\widehat{Var}_{2AS}(\widehat{T}_{AS}) = N^2\widehat{Var}_{2AS}(\overline{y}_{AS})$	se N é conhecido e houver ordenação
$\widehat{Var}_{1AS}(\overline{y}_{AS}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i\in s_r}(y_i - \overline{y}_{AS})^2$	se N é conhecido e sem ordenação
$\widehat{Var}_{2AS}(\overline{y}_{AS}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{2(n-1)}\sum_{i\in s_r}(y_i - y_{i+K})^2$	se N é conhecido e houver ordenação

Para casa 🏠

- Continuar exercícios.
- Ler o capítulo 3 da apostila da Profa. Vanessa.
- Ler o capítulo 8 do livro 'Amostragem: Teoria e Prática Usando R'.
- Rever os slides.

Próxima aula 📊

- Acompanhar o material no moodle.

Amostragem Sistemática

- Estimação.
- Laboratório de R

Muito obrigado!

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

Referências

- Amostragem: Teoria e Prática Usando o R
- **Elementos de Amostragem**, Bolfarine e Bussab.
- Cochran(1977)

Resumo da notação