Notas de aula de Amostragem I (EST0064)

Tópico 5
Amostragem de Conglomerados com Iguais Probabilidades

Damião Nóbrega da Silva

Departamento de Estatística/UFRN

7 de novembro de 2023

Conceitos básicos

Suponha que se queira estimar a renda per capita de uma pequena cidade. Mas:

- não se dispõe de uma lista ou cadastro dos residentes adultos da cidade (situação frequente na prática)

- construção do cadastro das unidades observacionais pode demandar bastante tempo e dinheiro

Como retirar uma amostra probabilística da população?

Conceitos básicos

Em geral, é mais fácil obter uma lista de unidades que englobem os residentes, por exemplo:

- lista de setores censitários do IBGE
- lista de domicílios

Setor censitário e domicílio são exemplos de um conglomerado.

Conglomerado = grupo de sub-unidades

Conceitos básicos

A pesquisa pode então ser realizada da seguinte forma:

- seleciona-se uma amostra de conglomerados; e
- entrevista-se todos os residentes adultos de cada conglomerado da amostra.

Amostragem de conglomerados pode seguir qualquer plano de amostragem (probabilístico ou não)

Amostragem Aleatória de Conglomerados em 1–estágio (AAC1)

1. Seleciona-se uma AASSR de conglomerados (unidades amostrais)

2. Investiga-se todas as unidades observacionais dentro de cada unidade amostral selecionada.

Porque usar AAC?

- cadastro listando unidades observacionais inexiste ou é caro ou difícil de ser construído
- custo de obtenção de observações aumenta com a distância separando as unidades observacionais

  $\Rightarrow$ AAC é mais económica que AAS.

Amostragem de conglomerados em vários estágios

AAC pode ser implementada em vários estágios.

Exemplo: Amostragem aleatória de conglomerados em 2 estágios (AAC2)

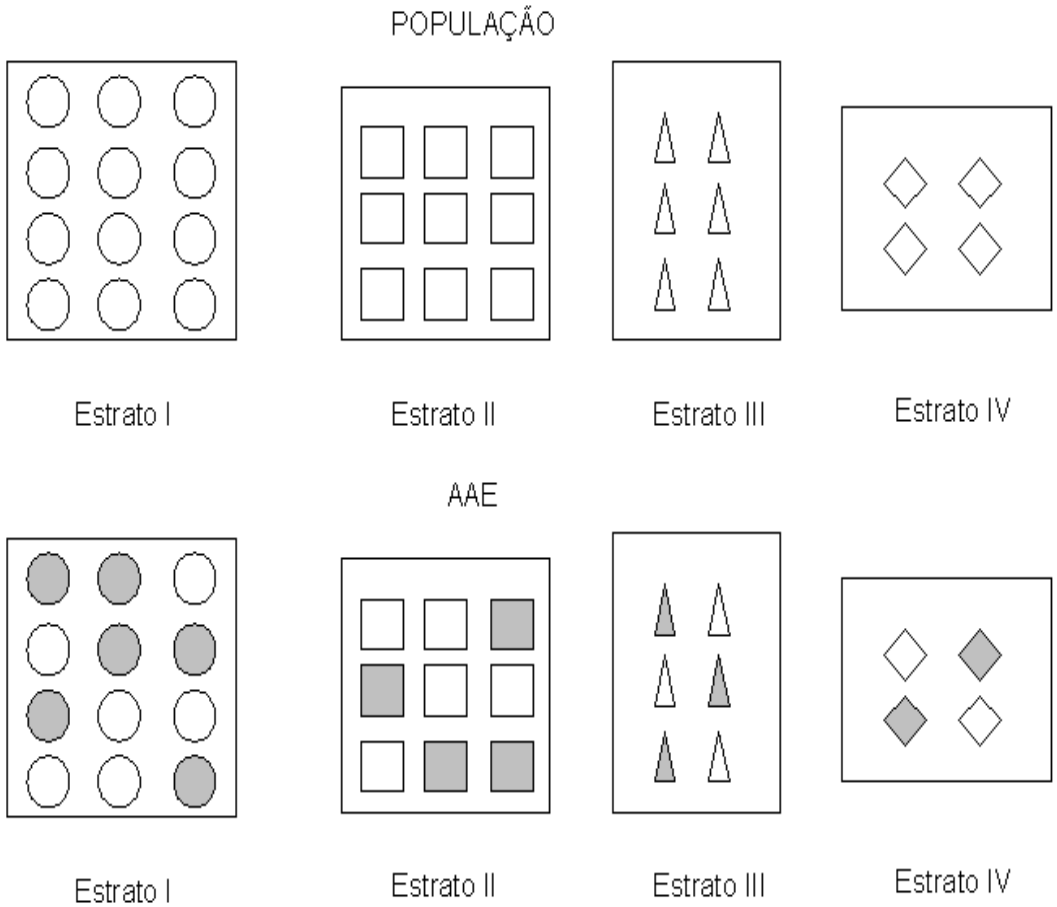
- seleção de uma AASSR de setores censitários (unidades primárias)
- de cada setor sorteado, sorteia–se uma AASSR de domicílios (unidades secundárias)
- de cada domicílio sorteado, entrevista–se todos os residentes adultos (unidades observacionais).

Amostragem de conglomerados $\times$ Amostragem estratificada

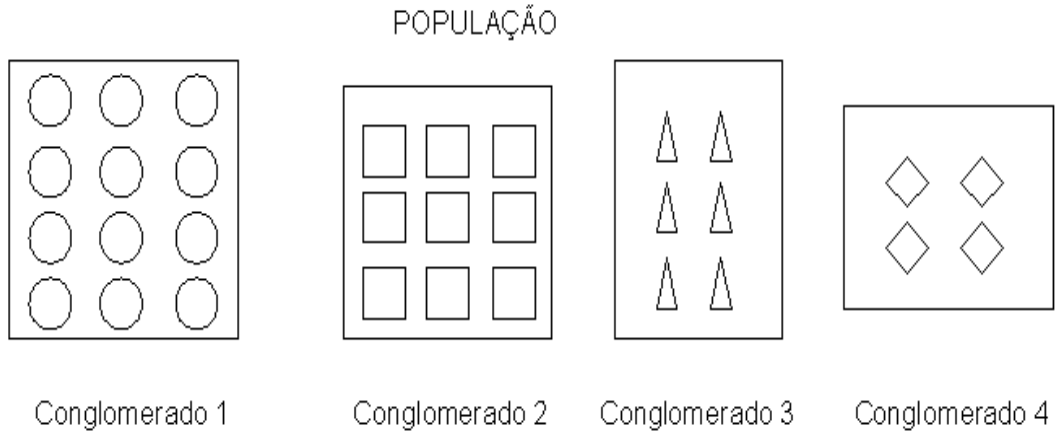
A exemplo de um estrato, um conglomerado é também um grupo de elementos da população

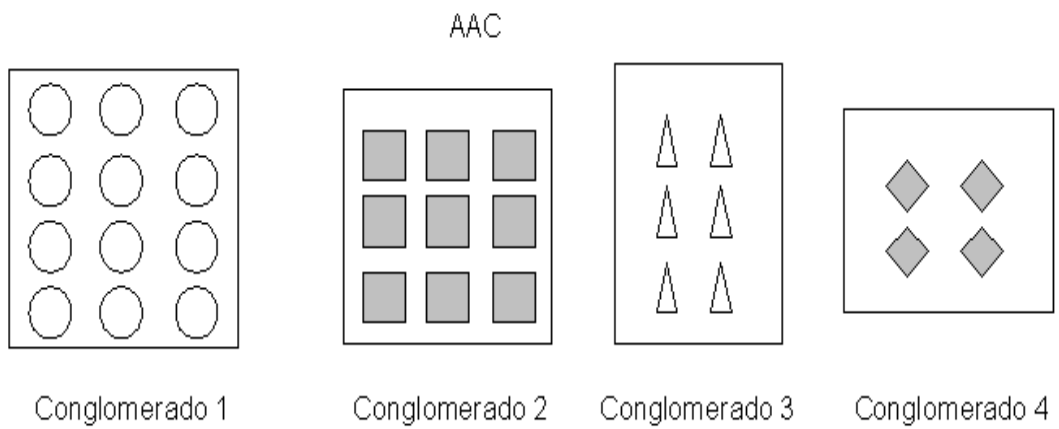
- cada elemento está em exatamente um, e somente um, estrato/conglomerado
- mas, processo de seleção dos elementos na AAE e na AAC diferem completamente
- variância da estimativa de $\bar{y}_U$ sob AAE depende da variabilidade das respostas dentro dos estratos
- variância da estimativa de $\bar{y}_U$ sob AAC depende da variabilidade entre as respostas médias dos conglomerados.

Amostragem Aleatória Estratificada



Amostragem Aleatória de Conglomerados





Amostragem de conglomerados $\times$ Amostragem estratificada

- Estratificação: geralmente aumenta a precisão das estimativas
- Amostragem de conglomerados: geralmente diminui a precisão das estimativas
  - membros de um mesmo domicílio tendem a ter respostas similares
  - peixes de um mesmo lago tendem a ter concentrações de mércurio similares,...

Notação

- Unidades de amostragem
  - UAP : unidade amostral primária
  - UAS : unidade amostral secundária,...
- $\mathcal{U} = \{1, 2, ..., N\}$: população de N conglomerados (UAPs)
- M_i : número de UAS na UAP i ($i \in \mathcal{U}$)
- $K = \sum_{i=1}^{N} M_i$: número total de UAS na população
- y_{ij} : resposta da variável y na UAS j da UAP i ($i \in \mathcal{U}$, $j = 1, \ldots, M_i$)

Totais e médias populacionais

- $t_{yi} = \sum_{j=1}^{M_i} y_{ij}$: total da variável y na UAP i $(i \in \mathcal{U})$
- $t_y = \sum_{i=1}^{N} t_{yi} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$: total populacional de y
- $\bar{y}_U = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$: média populacional de y (por UAS)
- $\bar{y}_{iU} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \frac{t_{yi}}{M_i}$: média populacional de y na UAP i $(i \in \mathcal{U})$

Variâncias populacionais

- $S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_{yi} - \frac{t_y}{N} \right)^2$: variância populacional dos totais $\{t_{yi} : i \in \mathcal{U}\}$
- $S_y^2 = \frac{1}{K-1} \sum_{i=1}^{N} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_U)^2$: variância populacional de y (por UAS)
- $S_{yi}^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{iU})^2$: variância populacional de y na UAP i ($i \in \mathcal{U}$)

Estatísticas (amostrais)

- $\mathcal{S}$: amostra de n UAP ($\mathcal{S} \subset \mathcal{U}$)
- $\mathcal{S}_i$: amostra de m_i UAS da UAP i ($i \in \mathcal{S}$)
  - $\bar{y}_i = \frac{1}{m_i} \sum_{j \in \mathcal{S}_i} y_{ij}$: média amostral (por UAS) na UAP i
  - $\widehat{t}_{yi} = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij}$: total estimado para a UAP i
  - $s^2_{yi} = \frac{1}{m_i - 1} \sum_{j \in \mathcal{S}_i} (y_{ij} - \bar{y}_i)^2$: variância amostral de y dentro da UAP i

Exemplo

Na pesquisa para estimar a renda média dos adultos residentes na pequena cidade

- t_{yi} = renda total de todos os adultos do i–ésimo domicílio
- $\bar{t}_U = \frac{1}{N}\sum_{i=1}^{N} t_{yi}$ = renda média populacional por domicílio
- $\bar{y}_U = \frac{1}{K}\sum_{i=1}^{N} t_{yi} = \frac{\sum_{i=1}^{N} t_{yi}}{\sum_{i=1}^{N} M_i}$ = renda média per capita (por residente)

Amostragem de conglomerados com iguais probabilidades
Amostragem de conglomerados em um estágio (AAC1)

Definição:

- AASSR de n *UAP* (conglomerados)
- todas as *UAS* (elementos) dentro de cada conglomerado selecionado são observadas ($m_i = M_i$)

Dois casos:

- $M_1 = M_2 = \ldots = M_N = M$
- pelo menos um M_i é diferente dos demais.

AAC1 – Conglomerado de tamanhos iguais

- $M_1 = M_2 = \ldots = M_N = M \;\Rightarrow\; m_i = M$
- AAC1 é uma AASSR de n observações: $\{t_{yi} : i \in \mathcal{S}\}$
- Estimação do total t_y

$$\widehat{t}_{y,nt} = N\bar{t}_{\mathcal{S}} = \frac{N}{n}\sum_{i\in\mathcal{S}} t_{yi}$$

$$\mathrm{Var}(\widehat{t}_{y,nt}) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}, \quad S_t^2 = \frac{\sum_{i=1}^{N}\left(t_{yi}-\frac{t_y}{N}\right)^2}{N-1}$$

$$\widehat{\mathrm{V}}(\widehat{t}_{y,nt}) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}, \quad s_t^2 = \frac{\sum_{i\in\mathcal{S}}\left(t_{yi}-\frac{\widehat{t}_{y,nt}}{N}\right)^2}{n-1}$$

AAC1 – Conglomerado de tamanhos iguais

- Estimação da média populacional $\bar{y}_U = t_y/(NM)$:

$$\widehat{\bar{y}}_{nt} = \frac{\widehat{t}_{y,nt}}{NM} = \frac{\bar{t}_S}{M}$$

$$\mathrm{Var}(\widehat{\bar{y}}_{nt}) = \frac{\mathrm{Var}(\bar{t}_S)}{M^2} = \left(1 - \frac{n}{N}\right)\frac{S_t^2}{nM^2}$$

$$\widehat{\mathrm{V}}(\widehat{\bar{y}}_{nt}) = \left(1 - \frac{n}{N}\right)\frac{s_t^2}{nM^2}$$

Exemplo

Um estudante quer estimar o IRA médio dos estudantes de uma residência universitária sem dispor de uma lista de todos os estudantes.

- residência contém 100 quartos (4 estudantes/por quarto)
- estudante escolhe 5 quartos aleatoriamente
- estudantes dos apartamentos selecionados são entrevistados para registrar o valor do IRA deles (AAC-1)

Número do estudante	Quarto (conglomerado)				
	1	2	3	4	5
1	7,7	5,9	5,0	7,5	6,7
2	6,5	7,6	6,4	7,2	4,8
3	8,6	8,2	6,3	8,6	8,2
4	7,6	6,7	4,7	9,1	8,0
Total	30,4	28,4	22,4	32,4	27,7

Exemplo

As unidades amostrais primárias (conglomerados) são os quartos

$$\Rightarrow \quad N = 100, \; n = 5 \;\; \text{e} \;\; M = 4.$$

Dos dados amostrais, tem-se que:

$$\widehat{t}_{y,nt} \;=\; \frac{100}{5}(30{,}4 + 28{,}4 + 22{,}4 + 32{,}4 + 27{,}7) \;=\; 2.826 \quad \text{e}$$

$$s_t^2 \;=\; \frac{1}{5-1}\left[\left(30{,}4 - \frac{2.826}{100}\right)^2 + \ldots + \left(27{,}7 - \frac{2.826}{100}\right)^2\right] \;=\; 14{,}0980.$$

Então, a estimativa do IRA médio (por estudante) é:

$$\widehat{\bar{y}}_{nt} \;=\; \frac{\widehat{t}_{y,nt}}{NM} \;=\; \frac{2.826}{100(4)} \;=\; 7{,}07$$

$$\mathrm{EP}(\,\widehat{\bar{y}}\,) \;=\; \frac{1}{4}\sqrt{\left(1 - \frac{5}{100}\right)\frac{14{,}0980}{5}} \;\cong\; 0{,}41.$$

AAC1 – Conglomerado de tamanhos iguais

Observação: Em uma AAC1, com uma AASSR de n UAPs, amostra é auto-ponderada, pois o peso amostral de cada observação

$$w_{ij} = \frac{1}{\Pr\{\text{UAS } j \text{ da UAP } i \text{ ser incluída na amostra}\}} = \frac{1}{n/N} = \frac{N}{n}.$$

Estes pesos amostrais podem ser usados para cálculo das estimativas pelas fórmulas usuais, isto é:

$$\widehat{t}_{y,nt} = \frac{N}{n}\sum_{i\in\mathcal{S}} t_{yi} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}$$

$$\widehat{\bar{y}}_{nt} = \frac{\widehat{t}_y}{NM} = \frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}.$$

```
### Programa para estimar o IRA médio de alunos em uma
### residência universitário de N=100 quartos de M=4 alunos
### por quarto. Dados foram coletados em uma AAC1 de
### n=5 quartos.

N  <- 100
n  <- 5
M  <- 4
quarto <- rep(1:5, rep(M, n))
aluno  <- rep(1:M, n)
ira <- c(7.7, 6.5, 8.6, 7.6,
         5.9, 7.6, 8.2, 6.7,
         5.0, 6.4, 6.3, 4.7,
         7.5, 7.2, 8.6, 9.1,
         6.7, 4.8, 8.2, 8.0)
d <- data.frame(quarto, aluno, ira, N)
d
```

```
> d
   quarto aluno ira   N
1       1     1 7.7 100
2       1     2 6.5 100
3       1     3 8.6 100
4       1     4 7.6 100
5       2     1 5.9 100
6       2     2 7.6 100
7       2     3 8.2 100
8       2     4 6.7 100
9       3     1 5.0 100
10      3     2 6.4 100
11      3     3 6.3 100
12      3     4 4.7 100
13      4     1 7.5 100
14      4     2 7.2 100
15      4     3 8.6 100
16      4     4 9.1 100
17      5     1 6.7 100
18      5     2 4.8 100
19      5     3 8.2 100
20      5     4 8.0 100
```

```
library(survey)
plano <- svydesign(~quarto, data=d, fpc=~N) # Criação do plano amostral

summary(plano)   # Sumário do plano criado
1 - level Cluster Sampling design
With (5) clusters.
svydesign(~quarto, data = d, fpc = ~N)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.05    0.05    0.05    0.05    0.05    0.05
Population size (PSUs): 100
Data variables:
[1] "quarto" "aluno"  "ira"    "N"

weights(plano)   # pesos amostrais do plano

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
```

```
# Estimativa do IRA total na população
svytotal(~ira, plano)
    total     SE
ira  2826 163.66

# Estimativa do IRA médio na população
svymean(~ira, plano)
     mean     SE
ira 7.065 0.4092

# Cálculo das estimativas do total e da média com os
# pesos amostrais
w <- weights(plano)
sum(w*ira)
[1] 2826
sum(w*ira)/sum(w)
[1] 7.065
```

Comparação de AAC1 com AASSR

Decomposição da variabilidade total entre e dentre conglomerados

Tabela de ANOVA populacional

Fonte de variação	Graus de liberdade	Somas de quadrado	Quadrado Médio
Entre congs.	$N-1$	$SQ_{entre} = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{iU} - \bar{y}_U)^2$	QM_{entre}
Dentre congs.	$N(M-1)$	$SQ_{dentre} = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{iU})^2$	QM_{dentre}
Total	$NM-1$	$SQT = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y}_U)^2$	S_y^2

$$QM_{entre} = \frac{\mathrm{SQ}_{entre}}{N-1}, \quad QM_{dentre} = \frac{\mathrm{SQ}_{dentre}}{N(M-1)} \text{ e } S_y^2 = \frac{SQT}{NM-1}$$

Comparação de AAC1 com AASSR

$$
\begin{aligned}
\mathrm{Var}(\widehat{t}_{y,nt} \mid AAC1) &= N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n} \\
&= N^2 \left(1 - \frac{n}{N}\right) \frac{M(QM_{entre})}{n}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\widehat{t}_{y,nt} \mid AASSR) &= (NM)^2 \left(1 - \frac{nM}{NM}\right) \frac{S_y^2}{nM} \\
&= N^2 \left(1 - \frac{n}{N}\right) \frac{M(S_y^2)}{n}
\end{aligned}
$$

Logo, se $QM_{entre} > S_y^2 \Rightarrow$ AAC1 é menos eficiente que AASSR.

Comparação de AAC1 com AASSR
Medidas de homogeneidade

Coeficiente de correlação intraclasse (CCI): mede homogeneidade das respostas dentro dos conglomerados [=coef. de correlação de Pearson para os $NM(M-1)$ pares (y_{ij}, y_{ik}) $(j \neq k)$]

$$CCI = 1 - \frac{M}{M-1}\frac{SQ_{dentre}}{SQT}$$

- $0 \leq SQ_{dentro}/SQT \leq 1 \Rightarrow -1/(M-1) \leq CCI \leq 1$
- se os conglomerados são perfeitamente homogêneos ($SQ_{entre} = 0$), então $CCI = 1$

R^2 ajustado: medida alternativa de homogeneidade que pode ser calculada mesmo se os conglomerados tenham tamanhos desiguais

$$R_a^2 = 1 - \frac{QM_{dentre}}{S_y^2}$$

Comparação de AAC1 com AASSR
Efeito do Plano Amostral (EPA)

Usando a relação $QM_{entre} = \frac{NM-1}{M(N-1)} S_y^2[1+(M-1)CCI]$, o EPA da AAC-1 em relação a uma AASSR com mesmo número de elementos é

$$EPA = \frac{QM_{entre}}{S_y^2} = \frac{NM-1}{M(N-1)}[1+(M-1)CCI]$$

- se $NM-1 \approx M(N-1) \quad \Rightarrow \quad EPA \approx 1+(M-1)CCI$

- $(1+(M-1)CCI)$ UASs em AAC1 dão a mesma quantidade de informação que uma UAS de uma AASSR.

Também, pode-se reescrever o EPA em função do R_a^2

$$EPA = \frac{QM_{entre}}{S_y^2} = 1 + \frac{N(M-1)}{N-1} R_a^2$$

Exemplo

Suponha que $CCI = 1/2$ e $M = 5$. Assim,

$$1 + (M-1)CCI = 1 + 4(1/2) = 3$$

"É necessário medir uma AAC1 de 300 elementos para ter a mesma precisão que uma AASSR de 100 elementos."

Observações:

- Se $CCI < 0 \Rightarrow$ AAC1 é mais eficiente que AASSR
- Se $CCI > 0 \Rightarrow$ AAC1 é menos eficiente que AASSR.

Exemplo

Considere as seguintes populações:

	População A			População B		
Conglomerado 1	10	20	30	9	10	11
Conglomerado 2	11	20	32	17	20	20
Conglomerado 3	9	17	31	31	32	30

Os elementos são os mesmos nas duas populações. Portanto, $\bar{y}_U = 20$ e $S_y^2 = 84{,}5$ para ambas A e B.

	População A		População B	
	$\bar{y}_{iU}$	S_{yi}^2	$\bar{y}_{iU}$	S_{yi}^2
Conglomerado 1	20	100	10	1
Conglomerado 2	21	111	19	3
Conglomerado 3	19	124	31	1

ANOVA para população A

Fonte	g.l.	SQ	QM	F
Entre conglomerados	2	6	3	0,03
Dentre conglomerados	6	670	111,67	
Total	8	676	84,5	

ANOVA para população B

Fonte	g.l.	SQ	QM	F
Entre conglomerados	2	666	333	199,8
Dentre conglomerados	6	10	1,67	
Total	8	676	84,5	

$$\mathrm{CCI(Pop.\ A)} = 1 - \tfrac{3}{2}\left(\tfrac{670}{676}\right) = -0{,}4867, \quad R_a^2 = 1 - \tfrac{111{,}67}{84{,}5} = -0{,}3215$$

$$\mathrm{CCI(Pop.\ B)} = 1 - \tfrac{3}{2}\left(\tfrac{10}{676}\right) = 0{,}9778, \quad R_a^2 = 1 - \tfrac{1{,}67}{84{,}5} = 0{,}9803$$

–Pop. A tem maior variação dentre conglomerados $\Rightarrow$ AAC1 é mais eficiente que AASSR (EPA $= 3/84{,}5 \approx 0{,}036$)

–Pop. B tem maior variação entre conglomerados $\Rightarrow$ AASSR é mais eficiente que AAC1 (EPA $= 333/84{,}5 \approx 3{,}941$).

AAC1 – Conglomerado de tamanhos variáveis

- Estimação não–tendenciosa de t_y

$$\widehat{t}_{y,nt} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_{yi} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij},$$

$$w_{ij} = \frac{1}{\Pr\{\text{UAS } j \text{ da UAP } i \text{ está na amostra}\}} = \frac{N}{n}$$

$$\mathrm{EP}(\widehat{t}_{y,nt}) = N \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}}$$

  - t_{yi} tendem a variar com M_i
  - $\widehat{t}_{y,nt}$ tem maior variabilidade quando M_i são variáveis

AAC1 – Conglomerado de tamanhos variáveis

- Estimação não–tendenciosa de $\bar{y}_U$

$$\widehat{\bar{y}}_{nt} = \frac{\widehat{t}_{y,nt}}{K} = \frac{\widehat{t}_{y,nt}}{\sum_{i=1}^{N} M_i}, \qquad \mathrm{EP}(\widehat{\bar{y}}_{nt}) = \frac{\mathrm{EP}(\widehat{t}_{y,nt})}{K}$$

  - requer conhecimento de $K = \sum_{i=1}^{N} M_i$
  - em muitas pesquisas M_i é conhecido apenas para $i \in \mathcal{S}$.

AAC1 – Conglomerado de tamanhos variáveis

Estimação razão de $\bar{y}_U$

- se os M_i variam, t_{yi} e M_i tem geralmente correlação positiva
- estimação razão é uma técnica conveniente com $x_i = M_i$

$$\widehat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} t_{yi}}{\sum_{i\in\mathcal{S}} M_i} = \frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}$$

$$\mathrm{EP}(\widehat{\bar{y}}_r) = \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\bar{M}_U^2}\frac{\sum_{i\in\mathcal{S}}(t_{yi}-\widehat{\bar{y}}_r M_i)^2}{n-1}}$$

$$= \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\bar{M}_U^2}\frac{\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i-\widehat{\bar{y}}_r)^2}{n-1}}$$

- $\bar{M}_U = K/N$ desconhecido, pode–se substituir $\bar{M}_U$ por $\bar{M}_{\mathcal{S}}$.

AAC1 – Conglomerado de tamanhos variáveis

Estimação razão de t_y

$$\widehat{t}_{y,r} = K\widehat{\bar{y}}_r, \qquad \mathrm{EP}(\widehat{t}_{y,r}) = K\mathrm{EP}(\widehat{\bar{y}}_r).$$

- variâncias dos estimadores razão podem ser bem menor que as dos estimadores não tendenciosos
- $\widehat{t}_{y,r}$ requer o conhecimento de $K = \sum_{i=1}^{N} M_i =$ número de elementos na população.

Amostragem de conglomerados em dois estágios com iguais probabilidades

- Quando a amostragem de conglomerados é em 1 estágio, todos os elementos dentro de um UAP selecionada são observados
- Porém, pode–se realizar sub-amostragem (em vez de um censo) dentro de cada UAP
  - elementos dentro das UAP selecionadas podem ser similares
  - custo de amostrar toda a UAP pode ser alto
  - aumentar a facilidade de obtenção da amostra
- A sub-amostragem dentro de cada UAP selecionada torna o plano amostral mais conveniente.

Amostragem de conglomerados em dois estágios com iguais probabilidades

Amostra Aleatória de Conglomerados em dois estágios (AAC2) pode ser formada da seguinte maneira:

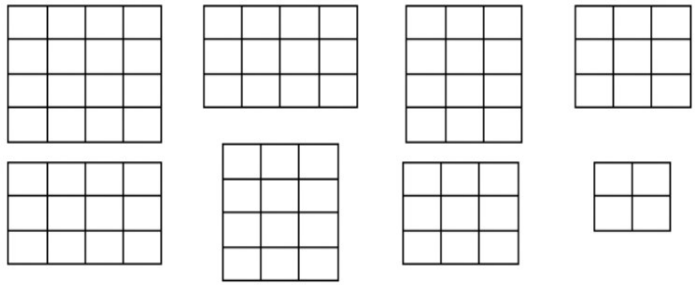
- retira–se uma AASSR de n UAPs da população de N UAPs

  Notação: Amostra primária $\rightarrow \mathcal{S}$

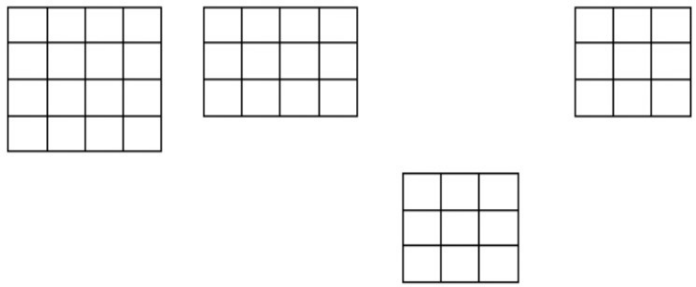
- de cada UAP i (que possui M_i elementos) na amostra primária $\mathcal{S}$, seleciona–se uma AASSR de m_i elementos

  Notação: Amostras secundárias $\rightarrow \mathcal{S}_i,\ i \in \mathcal{S}$

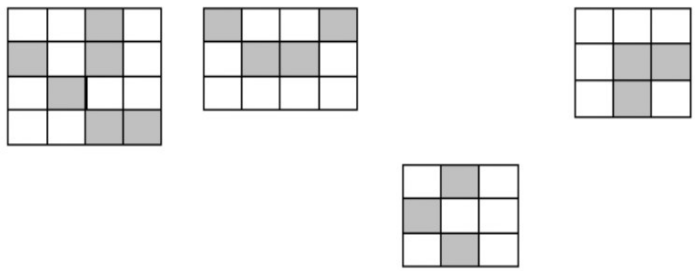
POPULAÇÃO DE N UAP



AMOSTRA de n UAP



AAC2



Estimação não tendenciosa de t_y em AAC2 com iguais probabilidades

$$\widehat{t}_{y,nt} \;=\; \frac{N}{n}\sum_{i\in\mathcal{S}}\widehat{t}_{yi} \;=\; \frac{N}{n}\sum_{i\in\mathcal{S}} M_i\bar{y}_i$$

$$\mathrm{Var}(\widehat{t}_{y,nt}) \;=\; N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\,\frac{S_{yi}^2}{m_i},$$

$$S_t^2 \;=\; \frac{\sum_{i=1}^{N}\left(t_{yi}-\frac{t_y}{N}\right)^2}{N-1} \quad \text{e} \quad S_{yi}^2 \;=\; \frac{\sum_{j=1}^{M_i}(y_{ij}-\bar{y}_{iU})^2}{M_i-1}$$

$$\widehat{\mathrm{V}}(\widehat{t}_{y,nt}) \;=\; N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in\mathcal{S}}\left(1-\frac{m_i}{M_i}\right)M_i^2\,\frac{s_{yi}^2}{m_i},$$

$$s_t^2 \;=\; \frac{\sum_{i\in\mathcal{S}}\left(\widehat{t}_{yi}-\frac{\widehat{t}_{y,nt}}{N}\right)^2}{n-1} \quad \text{e} \quad s_{yi}^2 \;=\; \frac{\sum_{j\in\mathcal{S}_i}(y_{ij}-\bar{y}_i)^2}{m_i-1}$$

Estimação não tendenciosa de $\bar{y}_U$ em AAC2 com iguais probabilidades

$$\widehat{\bar{y}}_{nt} = \frac{\widehat{t}_{y,nt}}{K},$$

$$\widehat{\mathrm{V}}(\widehat{\bar{y}}_{nt}) = \frac{\widehat{\mathrm{V}}(\widehat{t}_{y,nt})}{K^2}, \qquad \mathrm{EP}(\widehat{\bar{y}}_{nt}) = \frac{\mathrm{EP}(\widehat{t}_{y,nt})}{K}$$

Estimador $\widehat{\bar{y}}_{nt}$ requer o conhecimento do número de elementos na população K.

Estimação razão de $\bar{y}_U$ em AAC2 com iguais probabilidades

$$\widehat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} \widehat{t}_{yi}}{\sum_{i\in\mathcal{S}} M_i} = \frac{\sum_{i\in\mathcal{S}} M_i \bar{y}_i}{\sum_{i\in\mathcal{S}} M_i}$$

$$\widehat{\mathrm{V}}(\widehat{\bar{y}}_r) = \frac{1}{\bar{M}^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i}\right],$$

$$s_r^2 = \frac{\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i - \widehat{\bar{y}}_r)^2}{n-1} \quad \text{e} \quad s_{yi}^2 = \frac{\sum_{j\in\mathcal{S}_i}(y_{ij}-\bar{y}_i)^2}{m_i-1}$$

Exercício: Uma cadeia de restaurantes quer estimar o nível médio satisfação (escala de 1 a 7) de seus funcionários com o trabalho deles. A cadeia tem 120 restaurantes e um total de 6860 funcionários. Uma amostra aleatória simples de 10 restaurantes foi selecionada e, em cada um desses restaurantes, cerca de 20% dos funcionários foram entrevistados. Os dados obtidos foram os seguintes:

Restaurante	M_i	m_i	Satisfação do funcionário (y_{ij})	$\bar{y}_i$	s_{yi}
1	54	10	5, 7, 6, 5, 4, 7, 6, 6, 4, 5	5.50	1.08
2	48	10	7, 7, 7, 6, 5, 4, 7, 7, 6, 6	6.20	1.03
3	68	14	5, 6, 5, 6, 4, 5, 6, 5, 4, 5, 4, 6, 5, 6	5.14	0.77
4	70	14	6, 5, 7, 6, 7, 6, 5, 7, 5, 7, 6, 5, 7, 6	6.07	0.83
5	52	10	4, 5, 4, 5, 5, 6, 5, 4, 4, 4	4.60	0.70
6	62	12	5, 7, 6, 7, 4, 3, 1, 5, 4, 6, 4, 5	4.75	1.71
7	41	8	7, 6, 7, 7, 6, 6, 5, 7	6.38	0.74
8	53	11	6, 6, 5, 4, 6, 7, 5, 5, 7, 6, 5	5.64	0.92
9	64	12	7, 6, 5, 4, 6, 5, 7, 4, 3, 6, 5, 7	5.42	1.31
10	43	9	7, 6, 6, 5, 7, 3, 5, 4, 5	5.33	1.32

1. Como você descreveria que tipo de plano amostral foi utilizado?

2. Obtenha a estimativa do parâmetro de interesse usando um estimador não tendencioso.

3. Obtenha a estimativa do parâmetro de interesse usando o estimador razão.

1. O plano amostral adotado na pesquisa trata–se de uma amostra aleatória de conglomerados em dois estágios. No primeiro estágio, 10 restaurantes foram selecionados aleatoriamente da cadeia de 120 restaurantes. No segundo estágio, 20% dos funcionários de cada restaurante selecionado no primeiro estágio foram escolhidos aleatoriamente para serem entrevistados. A amostra contou com um total de 110 funcionários.

2. Os totais estimados dos conglomerados da amostra $\{\widehat{t}_{yi} = M_i\bar{y}_i : i \in \mathcal{S}\}$ são:

$$297.00, 297.60, 349.52, 424.90, 239.20, 294.50, 261.58, 298.92, 346.88, 229.19.$$

Uma estimativa não tendenciosa do total populacional é portanto

$$\widehat{t}_{y,nt} = \frac{N}{n}\sum_{i\in\mathcal{S}}\widehat{t}_{yi} = \frac{120}{10}\left[297.00 + 297.60 + 349.52 + \cdots + 229.19\right] = 36471.48.$$

Para estimar a variância de $\widehat{t}_{y,nt}$, é preciso calcular primeiro a variância amostral dos totais estimados $\widehat{t}_{yi}$

$$\begin{aligned} s_t^2 &= \frac{\sum_{i\in\mathcal{S}}\left(\widehat{t}_{yi} - \frac{\widehat{t}_{y,nt}}{N}\right)^2}{n-1} \\ &= \frac{\left[(297.00 - 303.929)^2 + (297.60 - 303.929)^2 + \cdots + (229.19 - 303.929)^2\right]}{10-1} \\ &= 3369.841. \end{aligned}$$

Assim, a variância estimada de $\widehat{t}_{y,nt}$ é

$$\widehat{V}(\widehat{t}_{y,nt}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in \mathcal{S}} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_{yi}^2}{m_i}$$

$$= 120^2 \left(1 - \frac{10}{120}\right) \frac{3369.841}{10} +$$

$$\frac{120}{10} \left[\left(1 - \frac{10}{54}\right) 54^2 \frac{1.08^2}{10} + \left(1 - \frac{10}{48}\right) 48^2 \frac{1.03^2}{10} + \cdots + \left(1 - \frac{9}{43}\right) 43^2 \frac{1.32^2}{9} \right]$$

$$= 4448191 + 32451.87 = 4480643$$

$$\Rightarrow \qquad EP(\widehat{t}_{y,nt}) = \sqrt{\widehat{V}(\widehat{t}_{y,nt})} = \sqrt{4480643} = 2116.753.$$

Portanto, a estimativa não tendenciosa da satisfação média de todos os funcionários da cadeira restaurante e seu erro padrão são respectivamente:

$$\widehat{\bar{y}}_{nt} = \frac{\widehat{t}_{y,nt}}{K} = \frac{36471.48}{6860} = 5.32$$

e

$$\mathrm{EP}(\widehat{\bar{y}}_{nt}) = \frac{\mathrm{EP}(\widehat{t}_{y,nt})}{K} = \frac{2116.753}{6860} = 0.3086.$$

3. A estimativa razão do total populacional é

$$\widehat{\bar{y}}_r = \frac{\sum\limits_{i\in\mathcal{S}} \widehat{t}_{yi}}{\sum\limits_{i\in\mathcal{S}} M_i} = \frac{297.00 + 297.60 + 349.52 + \cdots + 229.19}{54 + 48 + \cdots + 43} = \frac{3039.29}{555} = 5.48.$$

Para estimar a variância de $\widehat{\bar{y}}_r$, é preciso calcular primeiro a variância amostral dos totais estimados $\widehat{t}_{yi}$

$$\begin{aligned} s_r^2 &= \frac{\sum\limits_{i\in\mathcal{S}} M_i^2(\bar{y}_i - \widehat{\bar{y}}_r)^2}{n-1} = \frac{\sum\limits_{i\in\mathcal{S}} (\widehat{t}_{yi} - M_i\widehat{\bar{y}}_r)^2}{n-1} \\ &= \frac{\left[(297.00 - 54(5.48))^2 + (297.60 - 48(5.48))^2 + \cdots + (229.19 - 43(5.48))^2\right]}{10-1} \\ &= 1007.566. \end{aligned}$$

A variância estimada de $\widehat{\bar{y}}_r$ com

$$\bar{M} = \sum_{i=1}^{N} M_i/N = 6860/120 = 57.16667$$

é

$$\widehat{\mathrm{V}}_1(\widehat{\bar{y}}_r) = \frac{1}{\bar{M}^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_{yi}^2}{m_i}\right]$$

$$= \frac{1}{57.16667^2}\left(1-\frac{10}{120}\right)\frac{1007.566}{10} +$$

$$\frac{1}{57.16667^2}\frac{1}{10(120)}\left[\left(1-\frac{10}{54}\right)54^2\frac{1.08^2}{10} + \cdots + \left(1-\frac{9}{43}\right)43^2\frac{1.32^2}{9}\right]$$

$$= 0.028262 + 0.00069 = 0.028952$$

$$\Rightarrow \qquad EP_1(\widehat{\bar{y}}_r) = \sqrt{\widehat{V}(\widehat{\bar{y}}_r)} = \sqrt{0.028952} = 0.1702.$$

A variância estimada de $\widehat{\bar{y}}_r$ com

$$\bar{m} = \sum_{i \in \mathcal{S}} M_i / n = 555/10 = 55.5$$

é

$$\widehat{\mathrm{V}}_2(\widehat{\bar{y}}_r) = \frac{1}{\bar{m}^2} \left[\left(1 - \frac{n}{N}\right) \frac{s_r^2}{n} + \frac{1}{nN} \sum_{i \in \mathcal{S}} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_{yi}^2}{m_i} \right]$$

$$= \frac{1}{55.5^2} \left(1 - \frac{10}{120}\right) \frac{1007.566}{10} +$$

$$\frac{1}{55.5^2} \frac{1}{10(120)} \left[\left(1 - \frac{10}{54}\right) 54^2 \frac{1.08^2}{10} + \cdots + \left(1 - \frac{9}{43}\right) 43^2 \frac{1.32^2}{9} \right]$$

$$= 0.029985 + 0.000732 = 0.030717$$

$$\Rightarrow \qquad EP_2(\widehat{\bar{y}}_r) = \sqrt{\widehat{V}(\widehat{\bar{y}}_r)} = \sqrt{0.030717} = 0.1753.$$

Estas estimativas podem ser obtidas com o pacote survey do R com os comandos seguintes:

```
> d <- read.table("satisfacao-aac2.txt", header=TRUE)
> dim(d)
[1] 110   5
> d[1:15,]
   Restaurante Mi mi Funcionario y
1            1 54 10           1 5
2            1 54 10           2 7
3            1 54 10           3 6
4            1 54 10           4 5
5            1 54 10           5 4
6            1 54 10           6 7
7            1 54 10           7 6
8            1 54 10           8 6
9            1 54 10           9 4
10           1 54 10          10 5
11           2 48 10           1 7
12           2 48 10           2 7
13           2 48 10           3 7
14           2 48 10           4 6
15           2 48 10           5 5
...
```

```
> library(survey)
> plano <- svydesign(~Restaurante + Funcionario, data=d, fpc=~(N + Mi))
> summary(plano)

2 - level Cluster Sampling design
With (10, 110) clusters.
svydesign(~Restaurante + Funcionario, data = d, fpc = ~(N + Mi))
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01543 0.01603 0.01667 0.01655 0.01730 0.01744
Population size (PSUs): 120
Data variables:
[1] "Restaurante" "Mi"          "mi"          "Funcionario" "y"           "N"

> summary(weights(plano))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  57.33   57.82   60.00   60.55   62.40   64.80
```

```
> # Estimativas não tendenciosos do total e média
> (toty <- svytotal(~y, plano, deff=TRUE))
    total      SE   DEff
y 36469.4  2117.5 8.0866

> coef(toty)/K
       y
5.316239

> SE(toty)/K
          y
y 0.3086769

> # Estimativa razão da média
> svymean(~y, plano, deff=TRUE)
     mean      SE   DEff
y 5.47589 0.17504 2.4508
```

Note que cálculos anteriores estão sujeitos a maior erro de arrendondamento.

Uso de pesos amostrais em AAC2 com iguais probabilidades

Pesos amostrais são usualmente utilizados por institutos de estatísticas oficiais e alguns amostristas para estimar médias e totais.

Em AAC2, o cálculo do peso amostral depende da probabilidade de inclusão amostral

$$\pi_{ij} \equiv \Pr\left(\text{UAS } j \text{ da UAP } i \text{ ser selecionada para a amostra}\right).$$

Considere os eventos A_i = UAP i é selecionada para a amostra e B_j = UAS j é selecionada para a amostra.

$$\begin{aligned} \Rightarrow \pi_{ij} &= \Pr\left(A_i \cap B_j\right) \\ &= \Pr\left(A_i\right) \times \Pr\left(B_j \mid A_i\right) = \frac{n}{N}\frac{m_i}{M_i} \quad \text{para todo } i, j \end{aligned}$$

Uso de pesos amostrais em AAC2 com iguais probabilidades

Portanto, os pesos amostrais são

$$w_{ij} = \frac{1}{\pi_{ij}} = \frac{N}{n}\frac{M_i}{m_i}$$

$$\Rightarrow \quad \widehat{t}_{y,nt} = \frac{N}{n}\sum_{i\in\mathcal{S}} M_i\bar{y}_i = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}$$

$$\Rightarrow \quad \widehat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} M_i\bar{y}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}$$

- pesos fornecem procedimentos convenientes para calcular estimativas pontuais
- estimativas das variâncias requerem fórmulas específicas.

Para ilustrar o processo de cálculo de estimativas usando pesos amostrais, considere novamente a pesquisa de satisfação de funcionários da cadeia de restaurantes.

```
> w <- weights(plano)
> round(w, 1)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
64.8 64.8 64.8 64.8 64.8 64.8 64.8 64.8 64.8 64.8 57.6 57.6 57.6 57.6 57.6 57.6 57.6 57.6 57.6 57.6
  21   22   23   24   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40
58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 60.0 60.0 60.0 60.0 60.0 60.0
  41   42   43   44   45   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60
60.0 60.0 60.0 60.0 60.0 60.0 60.0 60.0 62.4 62.4 62.4 62.4 62.4 62.4 62.4 62.4 62.4 62.4 62.0 62.0
  61   62   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80
62.0 62.0 62.0 62.0 62.0 62.0 62.0 62.0 62.0 62.0 61.5 61.5 61.5 61.5 61.5 61.5 61.5 61.5 57.8 57.8
  81   82   83   84   85   86   87   88   89   90   91   92   93   94   95   96   97   98   99  100
57.8 57.8 57.8 57.8 57.8 57.8 57.8 57.8 57.8 64.0 64.0 64.0 64.0 64.0 64.0 64.0 64.0 64.0 64.0 64.0
 101  102  103  104  105  106  107  108  109  110
64.0 57.3 57.3 57.3 57.3 57.3 57.3 57.3 57.3 57.3

> round((N/n)*(d$Mi/d$mi), 1)
  [1] 64.8 64.8 64.8 64.8 64.8 64.8 64.8 64.8 64.8 64.8 57.6 57.6 57.6 57.6 57.6 57.6 57.6 57.6 57.6
 [20] 57.6 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 58.3 60.0 60.0 60.0 60.0
 [39] 60.0 60.0 60.0 60.0 60.0 60.0 60.0 60.0 60.0 60.0 62.4 62.4 62.4 62.4 62.4 62.4 62.4 62.4 62.4
 [58] 62.4 62.0 62.0 62.0 62.0 62.0 62.0 62.0 62.0 62.0 62.0 62.0 62.0 61.5 61.5 61.5 61.5 61.5 61.5
 [77] 61.5 61.5 57.8 57.8 57.8 57.8 57.8 57.8 57.8 57.8 57.8 57.8 57.8 64.0 64.0 64.0 64.0 64.0 64.0
 [96] 64.0 64.0 64.0 64.0 64.0 64.0 57.3 57.3 57.3 57.3 57.3 57.3 57.3 57.3 57.3
```

Estimativas:

```
> # Estimativa não tendenciosa do total
> sum(w*d$y)
[1] 36469.4
>
> # Estimativa razão da média
> sum(w*d$y)/sum(w)
[1] 5.475886
```

Planejando de uma AAC com iguais probabilidades

- Decisões importantes
  - precisão global
  - tamanho das UAPs
  - número de UASs a serem amostradas em cada UAP selecionada
  - número de UAPs
- Um bom sumário das técnicas para abordar estas questões é dado em Lohr (2010, seção 5.4)

Amostragem sistemática

Amostragem sistemática (AS) é um caso particular de amostragem aleatória de conglomerados (AAC).

Suponha que é desejado uma AS de tamanho 3 da população

$$\mathcal{U} = \{1 \;\; 2 \;\; 3 \;\; 4 \;\; 5 \;\; 6 \;\; 7 \;\; 8 \;\; 9 \;\; 10 \;\; 11 \;\; 12\}$$

- escolha um número aleatório entre 1 e (12/3)=4
- escolha este elemento e cada $4^{\underline{o}}$ seguinte
- as AS possíveis são

$$\{1, 5, 9\}, \; \{2, 6, 10\}, \; \{3, 7, 11\}, \; \{4, 8, 12\}.$$

Amostragem sistemática

Note que a população pode ser dividida nos seguintes conglomerados

Conglomerado			
I	II	III	IV
1	2	3	4
5	6	7	8
9	10	11	12

- população contém 4 UAPs
- 4 AAC1 são possíveis: {1,5,9}, {2,6,10}, {3,7,11}, {4,8,12}
- AS = AAC1 com $n = 1$

Amostragem sistemática - Estimação da média populacional

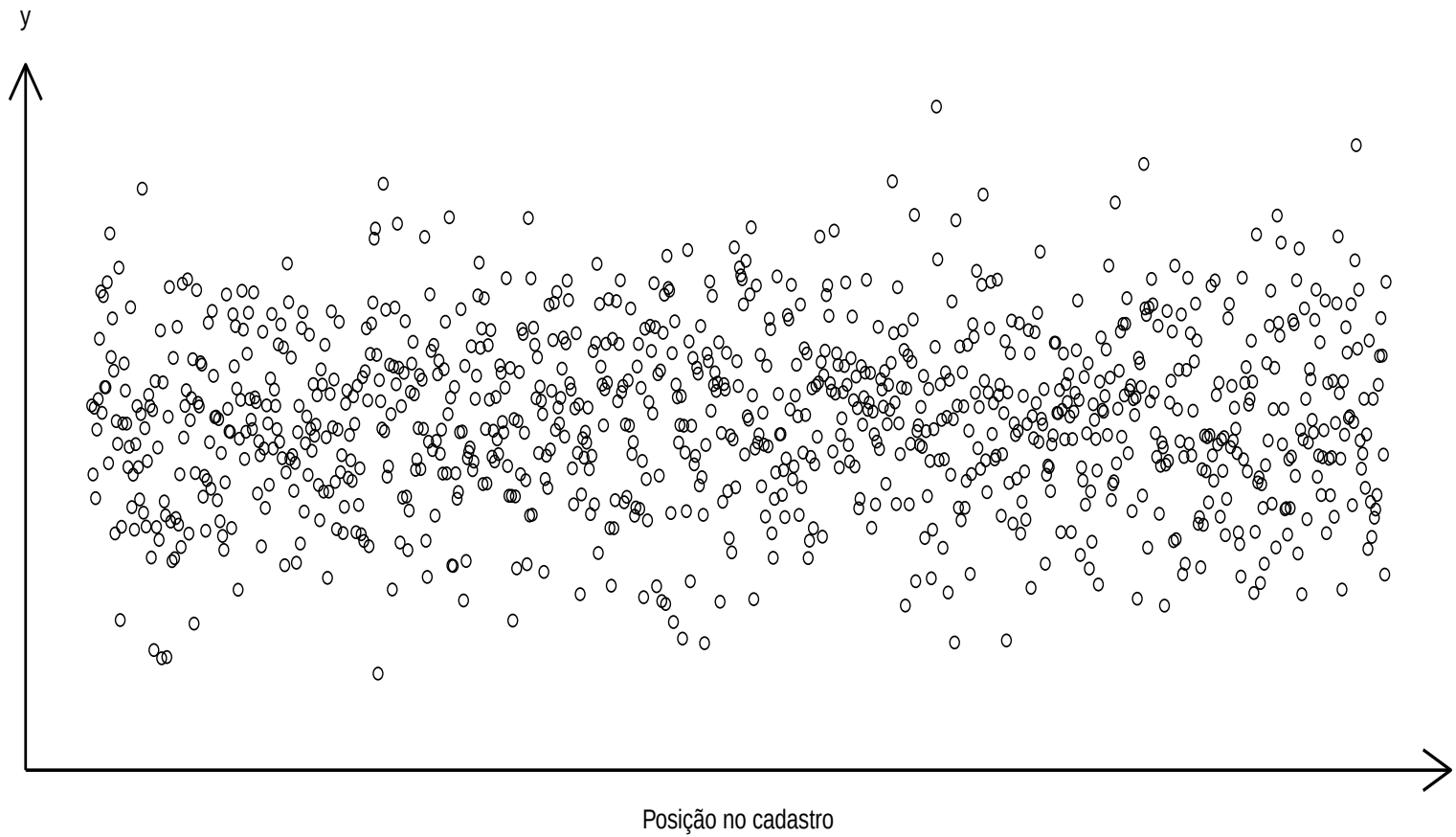
Estimador: $\bar{y}_{sis} = \bar{y}$

$$\mathrm{E}[\bar{y}_{sis}] = \bar{y}_U$$

$$\mathrm{Var}(\bar{y}_{sis}) = \left(1 - \frac{1}{N}\right)\frac{S_t^2}{M^2} = \frac{S_y^2}{M}[1 + (M-1)CCI]$$

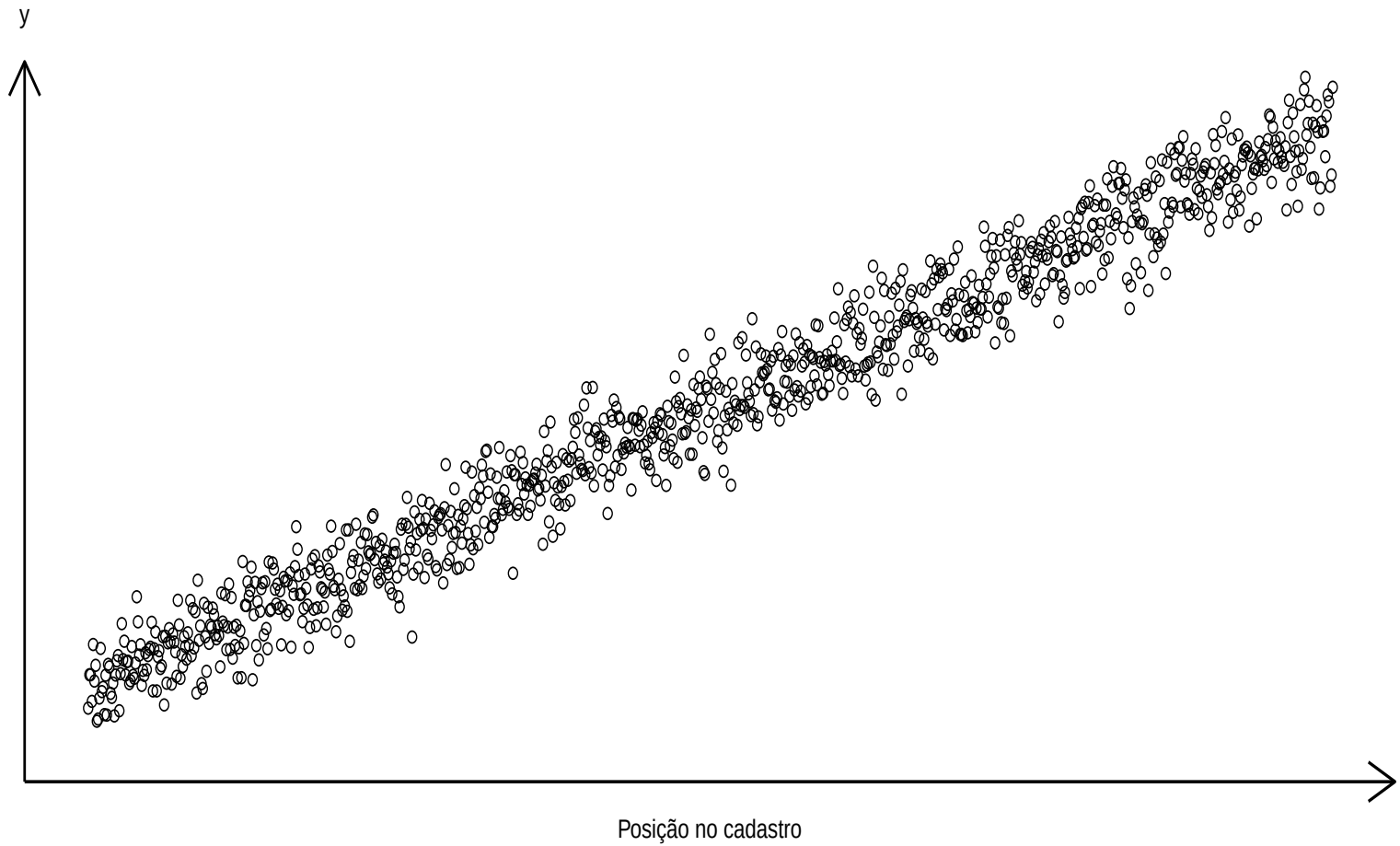
- em geral, NÃO É POSSÍVEL obter estimador não-tendencioso da $\mathrm{Var}(\bar{y}_{sis})$, pois a AS é uma AAC de tamanho $n = 1$
- estimação da variância requer maior conhecimento da estrutura da população (necessário para um uso mais efetivo da AS).

1. Lista está em ordem aleatória



$$\text{AS} \approx \text{AASSR} \quad \Rightarrow \quad \mathrm{Var}(\bar{y}_{sis}) \;\approx\; \mathrm{Var}(\bar{y} \,|\, AASSR)$$

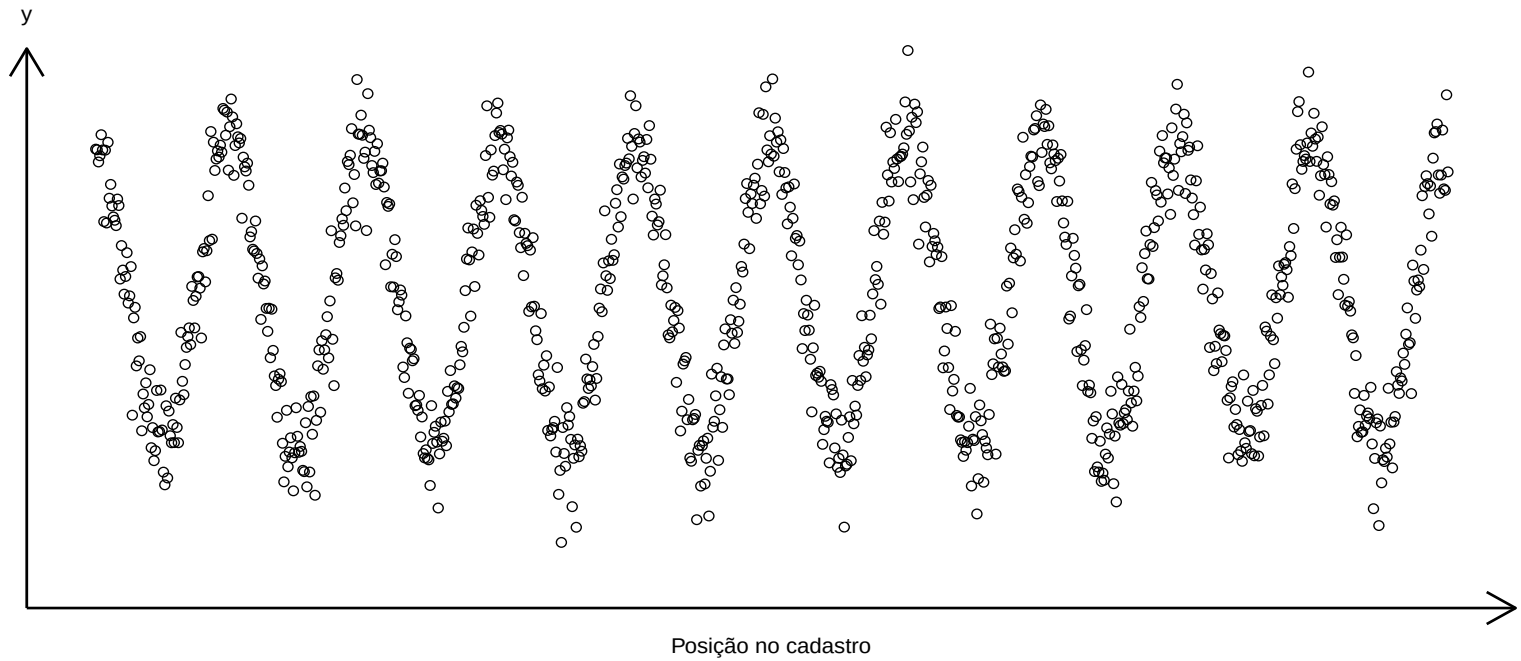
2. Lista está em ordem crescente ou decrescente



Neste caso, AS tenderá a ser mais precisa que AASSR

$$\mathrm{Var}(\bar{y}_{sis}) \; < \; \mathrm{Var}(\bar{y} \,|\, AASSR)$$

3. Lista apresentando variação periódica (cíclica)



Se intervalo de amostragem = período da tendência, a AS tenderá a ser menos precisa que AASSR

$$\mathrm{Var}(\bar{y}_{sis}) \;>\; \mathrm{Var}(\bar{y} \,|\, AASSR)$$

Deve–se usar métodos especiais de estimação de variância (por exemplo, Amostras Sistemáticas Interpenetrantes, Mahalanobis 1946).