# Amostragem por Conglomerados

Airlane P. Alencar

7 de junho de 2023

# Índice



# Referências


- Silva, Nilsa N. Amostragem Probabilística. EDUSP
- Lohr, S. Sampling.
- Bussab e Bolfarine. Elementos de Amostragem.


## *Objetivo*

- Quando as unidades amostrais estão divididas em grupos e só alguns desses grupos são selecionados para compor a amostra, temos amostragem por conglomerados.
- Podemos fazer uma lista de unidades amostrais somente para os conglomerados.

Exemplo: Sorteio escolas e suponha que cada escola tenha $M_j = 100$ alunos

Mais estágios

50 escolas $\longrightarrow$ 2 classes por escola $\longrightarrow$ 5 crianças por classe

50 escolas $\longrightarrow$ 1 classe por escola $\longrightarrow$ 10 crianças por classe

25 escolas $\longrightarrow$ 4 classes por escola $\longrightarrow$

# Vacinação

População de $K = 400$ pacientes atendidos em $N = 10$ consultórios de UBS

Para amostra de $m = 40$ pacientes, temos $f = \frac{40}{400} = 0,1$.

$N$ número de conglomerados na população.

$M_j$ é o tamanho de conglomerado $j$, $j = 1, \ldots, N$.

- Plano 1
  Sorteio em estágio único AAS de n=4 consultórios e todos os pacientes dos consultórios sorteado são amostrados.
  Temos 210 possíveis amostras (C10,4).

# Vacinação

- Plano 1
  Consultórios 1,4,5,7 $\rightarrow$ m=100
  Consultórios 2,3,6,8 $\rightarrow$ m=
- Plano 2: 2 estágios
  AAS: $a = 4$ consultórios
  $f_1 = \frac{4}{10} = 0,4$
  $f = f_1 * f_2 \rightarrow 0,1 = 0,4 f_2 \rightarrow$
  $f_2 = 1/4 = 0,25$
  Consultórios 1, 4, 5, 7 $\rightarrow$ m=26
  Consultórios 2,3, 6, 8 $\rightarrow$ m=

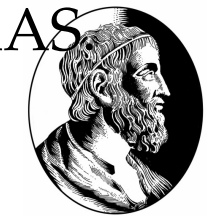
| Consult. | $M_j$ | $m_j = f_2 M_j$ | Acum $M$ |
|---|---|---|---|
| 1 | 30 | 7.5 (8) | 30 |
| 2 | 100 | 25 | 130 |
| 3 | 50 | 12.5 (13) | 180 |
| 4 | 20 | 5 | 200 |
| 5 | 30 | 7.5 (8) | 230 |
| 6 | 45 | 11.25 (11) | 275 |
| 7 | 20 | 5 | 295 |
| 8 | 40 | 10 | 335 |
| 9 | 30 | 7.5 (8) | 365 |
| 10 | 35 | 8.75 (9) | 400 |
| Soma | 400 | 100 | |

# Vacinação

- Plano 3.
  Sorteio em 2 estágios com probabilidade proporcional ao tamanho dos conglomerados (PPT).
  Mais usado, mantém tamanho de amostra desejado m.
  $f = f_1 f_2$
  $m = nb$
  $a$: tamanho da amostra no primeiro estágio= número de conglomerados sorteados
  Exemplo:
  m=40, n=4. Sorteio $b = \frac{40}{4} = 10$ elementos por consultório.
  Mantenho $f = \frac{40}{400} = f_1 f_2 = \left(\frac{4M_j}{400}\right)\left(\frac{10}{Mj}\right)$

# 1 só estágio

- Sortearemos m conglomerados e estudamos todas as suas unidades amostrais.
- É ideal que haja heterogeneidade intra conglomerado.
- Na prática, em geral os conglomerados têm unidades semelhantes, como setores censitários, escolas...
- Por exemplo, suponha que sorteamos alguns setores censitários (conglomerados) e estudamos todos os seus domicílios (unidade amostral). Queremos estimar a renda média por domicílio.
- Vale a pena ter amostra maior usando conglomerados, pois o custo em geral é bem menor do que espalhar a pesquisa usando AAS.



# 1 só estágio

A população tem N conglomerados, cada um com $M_i$ unidades amostrais, totalizando $K = \sum_{i=1}^{M} M_i$ unidades amostrais (primárias).

$$\mathcal{U} = \{(1,1), \ldots, (1,M_1), (2,1), \ldots, (2,M_2), (N,1), \ldots, (N,M_N)\}$$

Cada Conglomerado $i$ tem os elementos: $(i,1), \ldots (i,M_i)$.

Conglomerado

$$\begin{array}{c|ccccc} 1 & y_{1,1} & \cdots & y_{1,j} & \cdots & y_{1,M_1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \cdots \\ i & y_{i,1} & \cdots & y_{i,j} & \cdots & y_{i,M_i} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \cdots \\ N & y_{N,1} & \cdots & y_{N,j} & \cdots & y_{N,M_N} \end{array}$$

| Quantidade | Interpretação |
|---|---|
| $N$ | número de conglomerados (u.prim.) na população |
| $n$ | número de conglomerados (u.prim.) na amostra |
| $f = \frac{n}{N}$ | fração amostral dos conglomerados |
| $M_i$ | no. de unidades secundárias no conglomerado i |
| $K = \sum_{i=1}^{N} M_i$ | número de unidades secundárias na população |
| $y_{ij}$ | variável na unidade secundária j do cong. i |
| $\tau_i = \sum_{j}^{M_i} y_{ij} = M_i \bar{y}_i$ | total da variável y no conglomerado i |

AASs de n conglomerados de um total de N na população.
Estimador do total com sua variância

$$\widehat{\tau_{cl}} = \frac{N}{n} \sum_{i \in s}^{n} \tau_i, \qquad \tau_i \text{ conhecido} \tag{1}$$

$$Var(\widehat{\tau_{cl}}) = \frac{N^2(1-f)}{n} S_t^2 = \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N} (\tau_i - \frac{\tau}{N})^2}{N-1}$$

AASs de n conglomerados de um total de N na população. O estimador NÃO VICIADO do total com sua variância é

$$\widehat{\tau_{cl}} = \frac{N}{n}\sum_{i}^{n}\tau_i \tag{3}$$

$$Var(\widehat{\tau_{cl}}) = \frac{N^2(1-f)}{n}S_t^2 = \frac{N^2(1-f)}{n}\frac{\sum_{i=1}^{N}(\tau_i - \frac{\tau}{N})^2}{N-1} \tag{4}$$

$$\widehat{Var}(\widehat{\tau_{cl}}) = N^2(1-f)\frac{s_t^2}{n} \tag{5}$$

$$s_t^2 = \frac{1}{n-1}\sum_{i}^{n}\left(\tau_i - \frac{\widehat{\tau}_{cl}}{N}\right)^2 \tag{6}$$

Pode estimar a média por unidade secundária (domicílio):

$$\overline{y_{cl}} = \frac{\widehat{\tau}}{K}$$

$$\widehat{Var}(\overline{y_{cl}}) = \frac{\widehat{Var}(\widehat{\tau_{cl}})}{K^2}$$

# Análise de Variância Populacional - $M_i = M$

| Fonte | gl | Sum of Squares | Mean Sq=SS/gl |
|---|---|---|---|
| Entre (B) | $N-1$ | $SSB = \sum_{i=1}^{N}\sum_{j=1}^{M}(\mu_i - \mu)^2$ | $MSB$ |
| Intra (W) | $N(M-1)$ | $SSW = \sum_{i=1}^{N}\sum_{j=1}^{M}(\mu_{ij} - \mu_i)^2$ | $MSW$ |
| Total (TO) | $NM-1$ | $SST = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \mu)^2$ | $S^2$ |

# Coeficiente de Correlação Intraclasse - $M_i = M$

ICC é o coeficiente de correlação de Pearson entre todos os $NM(M-1)$ pares $(y_{ij}, y_{ik}), i = 1, \ldots, N, j \neq k = 1, \ldots, M$.

$$ICC = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\mu)(y_{ik}-\mu)}{(M-1)(NM-1)S^2}$$

$$= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\mu)(y_{ik}-\mu)}{(M-1)SSTO}$$

No ex.22 de Lohr e usaremos $SSTO = SSB + SSW$, temos:

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\mu)(y_{ik}-\mu) = M(SSB) - SSTO = M(SSTO - SSW) - SST$$

$$= (M-1)SSTO - M(SSW)$$

$$ICC = \frac{(M-1)SSTO - M.SSW}{(M-1)SSTO} = 1 - \frac{M.SSW}{(M-1)SSTO}$$

# Coeficiente de Correlação Intraclasse - $M_i = M$

$$ICC = \frac{(M-1)SSTO - M.SSW}{(M-1)SSTO} = 1 - \frac{M.SSW}{(M-1)SSTO}$$

$$\begin{aligned} 0 &\leq \frac{SSW}{SSTO} \leq 1 \\ 0 &\geq -\frac{M}{M-1}\frac{SSW}{SSTO} \geq -\frac{M}{M-1} \\ 1 - \frac{M}{M-1} &\leq 1 - \frac{M}{M-1}\frac{SSW}{SSTO} \leq 1 \\ -\frac{1}{M-1} &\leq ICC \leq 1 \end{aligned}$$

Se os conglomerados são completamente homogêneos (SSW=0), então ICC=1.

# Eficiência de estimadores não viesados

Na expressão (4), temos:

$$Var(\widehat{\tau_{cl}}) = \frac{N^2(1-f)}{n} S_t^2 = \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N}(\tau_i - \frac{\tau}{N})^2}{N-1}$$

Mas note que

$$S_t^2 = \frac{\sum_{i=1}^{N}(\tau_i - \frac{\tau}{N})^2}{N-1} = \frac{\sum_{i=1}^{N} M^2(\mu_i - \mu)^2}{N-1} = M\, MSB,$$

pois $M\, MSB = M\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(\mu_i-\mu)^2}{N-1} = M\frac{\sum_{i=1}^{N} M(\mu_i-\mu)^2}{N-1}$ então

$$Var(\widehat{\tau_{cl}}) = N^2\,(1-f)\,\frac{M\, MSB}{n} \qquad (7)$$

$$Var(\widehat{\tau_{AASs}}) = (NM)^2\left(1-\frac{nM}{NM}\right)\frac{S^2}{nM} = N^2\left(1-\frac{n}{N}\right)\frac{M\,S^2}{n}$$

# Eficiência de estimadores não viesados

Por outro lado, vamos considerar que temos AASs com NM observações.

$$Var(\widehat{\tau_{AASs}}) = (NM)^2 \left(1 - \frac{nM}{NM}\right) \frac{S^2}{nM} = N^2 \left(1 - \frac{n}{N}\right) \frac{M\,S^2}{n} \quad (9)$$

$$Var(\widehat{\tau_{cl}}) = N^2 (1 - f) \frac{M\,MSB}{n} \quad (10)$$

Se $MSB > S^2$, então a amostragem por conglomerados é menos eficiente que AASs.

Usando (7):

$$
\begin{aligned}
ICC &= 1 - \frac{M.SSW}{(M-1)SSTO} \\
(M-1)ICC &= (M-1) - \frac{M(SSTO - SSB)}{SSTO} \\
(M-1)ICC &= M - 1 - M + \frac{MSSB}{SSTO} = -1 + \frac{MSSB}{SSTO} \\
(M-1)ICC + 1 &= \frac{M(N-1)SSB}{(NM-1)S^2} \\
MSB &= (1 + (M-1)ICC)\frac{NM-1}{M(N-1)}S^2
\end{aligned}
$$

Assim, a razão entre as variâncias é

$$
\frac{Var(\widehat{\tau_{cl}})}{Var(\widehat{\tau_{AASs}})} = \frac{MSB}{S^2} = (1 + (M-1)ICC)\frac{NM-1}{(N-1)M}
$$

Usando (7):

$$MSB = (1 + (M-1)ICC)\frac{NM-1}{M(N-1)}S^2$$

Assim, a razão entre as variâncias é

$$\frac{Var(\widehat{\tau_{cl}})}{Var(\widehat{\tau_{AASs}})} = \frac{MSB}{S^2} = (1 + (M-1)ICC)\frac{NM-1}{(N-1)M} \quad (12)$$

Se N é bem grande com relação a M de modo que $\frac{NM-1}{(N-1)M} = \frac{NM-1}{NM-M} \approx 1$ então a razão entre as variâncias é dada por $(1 + (M-1)ICC)$.
Se ICC =1/2 e M=5, $(1 + (M-1)ICC) = 1 + 4/2 = 3$.

$$Var(\widehat{\tau_{cl}}) = N^2\,(1-f)\,\frac{M\,MSB}{n} \quad (13)$$

$$Var(\widehat{\tau_{AASs}}) = N^2\left(1-\frac{n}{N}\right)\frac{M\,S^2}{n}$$

Se N é bem grande com relação a M de modo que $\frac{NM-1}{(N-1)M} = \frac{NM-1}{NM-M} \approx 1$ então a razão entre as variâncias é dada por $(1 + (M-1)ICC)$.
Se ICC =1/2 e M=5, $(1 + (M-1)ICC) = 1 + 4/2 = 3$.

$$Var(\widehat{\tau_{cl}}) = N^2 (1-f) \frac{M\, MSB}{n} \longrightarrow 3 \quad (15)$$

$$Var(\widehat{\tau_{AASs}}) = N^2 \left(1 - \frac{n}{N}\right) \frac{M\, S^2}{n} \longrightarrow 1 \quad (16)$$

Precisamos pegar 300 observações usando conglomerados para ter var. equivalente para AASs com 100 observações.
Como usando conglomerados fica mais barato, vale a pena.

Pesquisador quer estimar a nota (GPA) média em seu alojamento. O alojamento tem 100 quartos (suites) com 4 estudantes cada um e 5 quartos foram sorteados.

| | Quarto | | | | |
|---|---|---|---|---|---|
| Pessoa | 1 | 2 | 3 | 4 | 5 |
| 1 | 3.08 | 2.36 | 2.00 | 3.00 | 2.68 |
| 2 | 2.60 | 3.04 | 2.56 | 2.88 | 1.92 |
| 3 | 3.44 | 3.28 | 2.52 | 3.44 | 3.28 |
| 4 | 3.04 | 2.68 | 1.88 | 3.64 | 3.20 |
| Total | 12.16 | 11.36 | 8.96 | 12.96 | 11.08 |

Estime a nota média por aluno (slide 11)

| | Quarto | | | | |
|---|---|---|---|---|---|
| Pessoa | 1 | 2 | 3 | 4 | 5 |
| 1 | 3.08 | 2.36 | 2.00 | 3.00 | 2.68 |
| 2 | 2.60 | 3.04 | 2.56 | 2.88 | 1.92 |
| 3 | 3.44 | 3.28 | 2.52 | 3.44 | 3.28 |
| 4 | 3.04 | 2.68 | 1.88 | 3.64 | 3.20 |
| Total | 12.16 | 11.36 | 8.96 | 12.96 | 11.08 |

$$\hat{\tau}_{cl} = \frac{N}{n}\sum_{i}^{n}\tau_i = \frac{100}{5}(12.16 + \ldots + 11.08) = 1130.4$$

$$s_t^2 = \frac{1}{n-1}\sum_{i}^{n}\left(\tau_i - \frac{\hat{\tau}_{cl}}{N}\right)^2 = \frac{1}{4}[(12.16 - 11.304)^2 + \ldots] = 2.256$$

$$\widehat{Var}(\hat{\tau}_{cl}) = N^2(1-f)\frac{s_t^2}{n} = 100^2\left(1 - \frac{5}{100}\right)\frac{2.256}{5} = 4285,792$$

$$\bar{y}_{cl} = \frac{\hat{\tau}_{cl}}{K} = \frac{1130.3}{400} = 2.826$$

| | Quarto | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Total | 12.16 | 11.36 | 8.96 | 12.96 | 11.08 |

$$\hat{\tau}_{cl} = \frac{N}{n}\sum_{i}^{n}\hat{\tau}_i = \frac{100}{5}(12.16 + ... + 11.08) = 1130.4$$

$$s_t^2 = \frac{1}{n-1}\sum_{i}^{n}\left(\hat{\tau}_i - \frac{\hat{\tau}_{cl}}{N}\right)^2 = \frac{1}{4}[(12.16 - 11.304)^2 + ...] = 2.256$$

$$\widehat{Var}(\hat{\tau}_{cl}) = N^2(1-f)\frac{s_t^2}{n} = 100^2\left(1 - \frac{5}{100}\right)\frac{2.256}{5} = 4285,792$$

$$\bar{y}_{cl} = \frac{\hat{\tau}_{cl}}{K} = \frac{1130.3}{400} = 2.826$$

$$\widehat{Var}(\bar{y}) = \frac{\widehat{Var}(\hat{\tau}_{cl})}{K^2} = \frac{4285,792}{400^2} = 0,0267 \quad DP = 0,164$$

| Fonte | gl | SS | MS | F |
|---|---|---|---|---|
| Entre (B) | 4 | 2.2557 | 0.56392 | 3.048 |
| Intra (W) | 15 | 2.7756 | 0.18504 | |
| Total | 19 | 5.0313 | 0.2648 | |

Para calcular ICC, precisamos estimar a ANOVA populacional, usando que MSB e MSW são estimadores não viesados de MSB e MSW pop.

| Fonte | gl | Sum of Squares | Mean Sq=SS/gl |
|---|---|---|---|
| Entre (B) | $N-1=99$ | $SSB = 55.828$ | $MSB = 0.5639$ |
| Intra (W) | $N(M-1)=300$ | $SSW = 55.512$ | $MSW = 0.18504$ |
| Total | $NM-1399$ | $SST = SSB + SSW = 111.340$ | $SST/399 = 0.279$ |

$$\widehat{ICC} = 1 - \frac{M}{M-1}\frac{\widehat{SSW}}{\widehat{SSB} + \widehat{SSW}} = 1 - \frac{4}{3}\frac{55.512}{111,34} = 0.335$$

$$\hat{S}^2 = MSTOT = \frac{SSTOT}{399} = 0.279$$

Assim, a razão entre as variâncias é

$$\frac{Var(\hat{\tau}_{cl})}{Var(\hat{\tau_{AASs}})} = \frac{0.5639}{0.279} = 2.02 \quad (17)$$

Precisamos de uma amostra de 2.02 x unidades amostrais usando conglomerados para ter variância equivalente a AASs de x elementos.

# Estimador Razão

AASs de n conglomerados de um total de N na população.
Total de cada conglomerado $\tau_i$ bem correlacionado com o tamanho do conglomerado $M_i$.
Total de unidades amostrais $K = \sum_{i=1}^{N} M_i$.
O estimador razão do total com sua variância é

$$\widehat{\tau}_r = rK, \qquad \text{est. total pop.}$$

$$r = \widehat{R} = \widehat{\mu} = \frac{\sum_i^n \tau_i}{\sum_i^n M_i} \qquad \text{est. média por unidade amostral}$$

$$\widehat{Var}(\widehat{\tau}_r) = N^2(1-f)\frac{s_r^2}{n} \qquad s_r^2 = \frac{\sum_i^n (\tau_i - rM_i)^2}{n-1}$$

$$\widehat{Var}(\widehat{R}) = (1-f)\frac{1}{\overline{M}^2}\frac{s_r^2}{n} \qquad \overline{M} = \frac{K}{N} \text{tamanho médio do cong}$$

# Amostra de só um conglomerado - Prop. ao Tamanho - p.182 - Lohr

Sorteamos 1 conglomerado.

$$\begin{aligned} \psi_i &= P(congl_i \text{ no primeiro sorteio}) \\ \pi_i &= P(congl_i \in Amostra) \\ \psi_i &= \pi_i \end{aligned}$$

Exemplo 4 lojas.

| Loja | Tamanho da loja (m2) | $\psi_i$ | $t_i$ (em milhares) |
|---|---|---|---|
| A | 100 | 1/16 | 11 |
| B | 200 | 2/16 | 20 |
| C | 300 | 3/16 | 24 |
| D | 1000 | 10/16 | 245 |
| | 1600 | 1 | 300 |

# Amostra de só um conglomerado - Prop. ao Tamanho - p.182 - Lohr

O peso de cada unidade é $w_i = \frac{1}{P(i \in amostra)} = \frac{1}{\psi_i}$

O estimador do total é

$$\widehat{t}_\psi = \sum_{i \in S} w_i t_i = \sum_{i \in S} \frac{1}{\psi_i} t_i$$

| Loja | Tamanho da loja (m2) | $\psi_i$ | $t_i$ | $\hat{t}_\psi$ | $(\hat{t}_\psi - t)^2$ |
|---|---|---|---|---|---|
| A | 100 | 1/16 | 11 | 176 | 15376 |
| B | 200 | 2/16 | 20 | 160 | 19600 |
| C | 300 | 3/16 | 24 | 128 | 29584 |
| D | 1000 | 10/16 | 245 | 392 | 8464 |
| | 1600 | 1 | 300 | | |

$$E(\widehat{t}_\psi) = \sum_s P(s)\widehat{t_{\psi,s}} = \frac{1}{16}176 + \frac{2}{16}160 + \frac{3}{16}128 + \frac{10}{16}392 = 300$$

# Amostra de só um conglomerado - Prop. ao Tamanho - p.182 - Lohr

O estimador do total é

$$\hat{t}_\psi = \sum_{i \in S} w_i t_i = \sum_{i \in S} \frac{1}{\psi_i} t_i$$

| Loja | Tamanho da loja (m2) | $\psi_i$ | $t_i$ | $\hat{t}_\psi$ | $(\hat{t}_\psi - t)^2$ |
|---|---|---|---|---|---|
| A | 100 | 1/16 | 11 | 176 | 15376 |
| B | 200 | 2/16 | 20 | 160 | 19600 |
| C | 300 | 3/16 | 24 | 128 | 29584 |
| D | 1000 | 10/16 | 245 | 392 | 8464 |
| | 1600 | 1 | 300 | | |

$$Var_{pop}(\hat{t}_\psi) = E[(\hat{t}_\psi - t)^2] = \sum_s P(s)(\widehat{t_{\psi,s}} - t)^2 = \sum_s \frac{1}{\psi_i} \left( \frac{t_i}{\psi_i} - t \right)^2$$

$$= \frac{1}{16} 15376 + \frac{2}{16} 19600 + \frac{3}{16} 29584 + \frac{10}{16} 8464 = 14248$$

# Amostra de só um conglomerado - Prop. ao Tamanho - p.182

Só para compararmos com AASc de tamanho 1, as probabilidades de cada loja ser sorteada é 1/4 e também teríamos estimador não viesado.

| Loja | Tamanho da loja (m2) | $\psi_i$ | $t_i$ | $\hat{t}_\psi$ | $(\hat{t}_\psi - t)^2$ |
|---|---|---|---|---|---|
| A | 100 | 1/4 | 11 | 44 | 65536 |
| B | 200 | 1/4 | 20 | 80 | 48400 |
| C | 300 | 1/4 | 24 | 96 | 41676 |
| D | 1000 | 1/4 | 245 | 980 | 462600 |
| | 1600 | 1 | 300 | | |

$$Var_{pop}(\hat{t}_\psi) = \sum_s \frac{1}{\psi_i}\left(\frac{t_i}{\psi_i} - t\right)^2 =$$

$$= \frac{1}{4}65536 + \frac{1}{4}48400 + \frac{1}{4}41616 + \frac{1}{4}462400 = 154488$$

# Um estágio com reposição - Prop. ao Tamanho

Sorteamos n conglomerados com reposição, então os sorteios são independentes.

$$\begin{aligned} \psi_i &= P(i \text{ no primeiro sorteio}) \\ \pi_i &= P(congl_i \in Amostra) = 1 - P(congl.i \notin Am.) = 1 - (1 - \psi_i)^n \end{aligned}$$

Como posso sortear os conglomerados de modo proporcional ao tamanho do conglomerado?
Vamos sortear 5 das 15 classes no exemplo a seguir.

# Sorteio 487, 369, 221, 326, 282 - congl 13,9,6,8,7

| Classe | $M_i$ | $\psi_i$ | Ampl. | Acumulada |
|---|---|---|---|---|
| 1 | 44 | 0.0680 | 1 | 44 |
| 2 | 33 | 0.0510 | 45 | 77 |
| 3 | 26 | 0.0402 | 78 | 103 |
| 4 | 22 | 0.0340 | 104 | 125 |
| 5 | 76 | 0.1175 | 126 | 201 |
| **6** | 63 | 0.0974 | 202 | 264 |
| **7** | 20 | 0.0309 | 265 | 284 |
| **8** | 44 | 0.0680 | 285 | 328 |
| **9** | 54 | 0.0835 | 329 | 382 |
| 10 | 34 | 0.0526 | 383 | 416 |
| 11 | 46 | 0.0711 | 417 | 462 |
| 12 | 24 | 0.0371 | 463 | 486 |
| **13** | 46 | 0.0711 | 487 | 532 |
| 14 | 100 | 0.1546 | 533 | 632 |
| 15 | 15 | 0.0232 | 633 | 647 |
| | 647 | | | |

# Um estágio com reposição - Prop. ao Tamanho

Para estudos grandes, faz amostra sistemática.
Temos 647 alunos, dividimos por 5 para obter 129,4 (arred. 129)
Sorteio um número k de 1 a 129, e pego o congl. que tem o k-ésimo aluno.
Depois pego o congl. com o aluno k+129, k+2(129), k+ 3(129), k+4(129).
ex: Sorteio 112.

# Sorteio 112 e pego 112, 241, 370, 499, 628 - congl 4, 6, 9,13,14

| Classe | M_i | Psi_i | Amplitude | Acumulada |
|---|---|---|---|---|
| 1 | 44 | 0.0680 | 1 | 44 |
| 2 | 33 | 0.0510 | 45 | 77 |
| 3 | 26 | 0.0402 | 78 | 103 |
| **4** | 22 | 0.0340 | 104 | 125 |
| 5 | 76 | 0.1175 | 126 | 201 |
| **6** | 63 | 0.0974 | 202 | 264 |
| 7 | 20 | 0.0309 | 265 | 284 |
| 8 | 44 | 0.0680 | 285 | 328 |
| **9** | 54 | 0.0835 | 329 | 382 |
| 10 | 34 | 0.0526 | 383 | 416 |
| 11 | 46 | 0.0711 | 417 | 462 |
| 12 | 24 | 0.0371 | 463 | 486 |
| **13** | 46 | 0.0711 | 487 | 532 |
| **14** | 100 | 0.1546 | 533 | 632 |
| 15 | 15 | 0.0232 | 633 | 647 |
| | 647 | | | |

# Um estágio com reposição - Prop. ao Tamanho

Considere $Q_i$ o número de vezes que o conglomerado i aparece na amostra, N é o número total de conglomerados e n é número de conglomerados sorteados. Note que $\sum_i Q_i = n$ e $E(Q_i) = n\psi_i$.
O estimador do total

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i}$$

é estimador não viesado de $t = \sum_i t_i$, total pop.

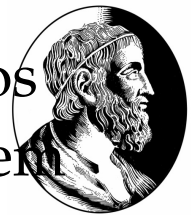
$$\begin{aligned} Var(\hat{t}_\psi) &= \frac{1}{n}\sum_{i=1}^{N} \psi_i \left(\frac{t_i}{\psi_i} - t\right)^2 \\ \widehat{Var}(\hat{t}_\psi) &= \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{\left(\frac{t_i}{\psi_i} - \widehat{t_\psi}\right)^2}{n-1} \end{aligned}$$

# Prop. ao Tamanho - Lahiri

Seja N o número de conglomerados e $max(M_i)$ o maior tamanho de conglomerado.
Lahiri(1951) propõe método de sorteio que gera amostra proporcional ao tamanho.

1. Sorteie número de 1 a N e identifique o conglomerado correspondente k.
2. Sorteie número de 1 a $max(M_i)$, se o número sorteado for menor que o tamanho $M_k$, inclua o conglomerado k na amostra, caso contrário ignore esse sorteio de k.
3. Repita os passos anteriores até ter o tamanho da amostra n.

No exemplo das classes, temos $max(M_i)$ = 100 estudantes. Sorteamos números de 1 a 15 e depois números de 1 a 100.
Sorteamos 12 e depois sorteamos 20, como 20<$M_1$2 = 24, incluímos k=12. Note que conglomerados maiores têm maior chance de serem sorteados. No final temos os conglomerados: 12, 14, 14, 5, 1.



# Prop. ao Tamanho - Lahiri

$t_i$ é o total de horas que os alunos da classe i estudaram estatística.

| Classe | $M_i$ | $\psi_i$ | $t_i$ | $t_i/\psi_i$ |
|---|---|---|---|---|
| 12 | 24 | 24/647 | 75 | 2021.875 |
| 14 | 100 | 100/647 | 203 | 1313.410 |
| 14 | 100 | 100/647 | 203 | 1313.410 |
| 5 | 76 | 67/647 | 191 | 1626.013 |
| 1 | 44 | 44/647 | 168 | 2470.364 |

O valor da última coluna é o total estimado se só temos o cong. i.

O estimador do total

$$\widehat{t}_\psi = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{t_i}{\psi_i} = \frac{2021.875 + 2(1313.41) + 1626.013 + 2470.364}{5} = 1749.014$$

$$\widehat{Var}(\widehat{t}_\psi) = \frac{1}{n}\sum_{i=1}^{N} Q_i \frac{\left(\frac{t_i}{\psi_i} - \widehat{t}_\psi\right)^2}{n-1} => EP = 222.42$$

O tempo médio estimado é $\frac{1749.014}{647} = 2.70$ com erro pad. $= \frac{222.42}{647} = 0.34$.

# 2 estágios com reposição

Pode sortear a unidade primária (setor censitário) i e depois sortear unidades secundárias (domicílios).
Se sorteou o setor 42 novamente, sorteia novamente domicílios nesse setor de modo independente.
$Q_i$ é o número de vezes que a u.primária i apareceu na amostra e teremos as estimativas do total na u.prim.i: $\widehat{t}_{i,1}, \widehat{t}_{i2}, \ldots, \widehat{t}_{iQ_i}$.

$$\widehat{t}_{\psi} = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\widehat{t}_{ij}}{\psi_i}$$

$$\widehat{Var}(\widehat{t}_{\psi}) = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{Q_i}\frac{\left(\frac{\widehat{t}_{ij}}{\psi_i} - \widehat{t_{\psi}}\right)^2}{n-1}$$

Sem reposição, usamos o estimador Horvitz-Thompson (p.196).