MAT02036 - Amostragem 2

Aula 03 - Amostragem Estratificada - Introdução 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

- Variável R_i indicadora do evento '**inclusão** da unidade i na amostra s'.

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

- Probabilidades de inclusão de **primeira ordem**.

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

- Probabilidades de inclusão de **segunda ordem**, denotadas π_{ij}, dadas por

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

- Variância e covariância de R

$$Var_p(R_i) = \pi_i(1 - \pi_i) \quad \text{e} \quad Cov_p(R_i, R_j) = \pi_{ij} - \pi_i \pi_j.$$

Aula passada

- **Estimador linear** do total T: $\widehat{Y}_w = \sum_{i \in s} w_i y_i = \sum_{i \in U} R_i w_i y_i$.
- Para que $\widehat{T}_w$ seja **sempre** não viciado:

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

  - Válida para quaisquer y_i se $\pi_i \times w_i = 1, \forall i \in U$.
- Estimador **Horvitz-Thompson**: com pesos básicos d_i, para o total Y

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

- O estimador linear do total $\widehat{T}_w = \sum_{i \in s} w_i y_i$ será **sempre não viciado** se:

$$w_i = {\pi_i}^{-1} = 1/\pi_i = d_i, \forall i \in U.$$

- **Distribuição amostral**, resultados assintóticos, ...

Aula passada

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

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

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

Amostragem Estratificada

Amostragem Estratificada

O método geral

- **Amostragem estratificada - AE** utiliza **informação auxiliar** relevante para **dividir** a população U em H **grupos** disjuntos e exaustivos,
    - geralmente mais homogêneos em relacão à(s) variável(is) de interesse, chamados *estratos*.
- Assumimos uma partição da população U (em H subconjuntos mutuamente exclusivos e exaustivos).
    - U_h: unidades pertencentes ao estrato h, para $h = 1, 2, \ldots, H$.

$$U = \bigcup_{h=1}^{H} U_h \qquad \text{e} \qquad U_h \cap U_k = \emptyset, \forall h \neq k.$$

- Seja N_h o tamanho de U_h, O tamanho total da população é dado por

$$N = N_1 + \cdots + N_h + \cdots + N_H = \sum_{h=1}^{H} N_h.$$

Amostragem Estratificada

O método geral

- Para ser viável, as **variáveis de estratificação** (usadas para dividir a população em estratos) precisam:
    - estar **disponíveis** para todas as unidades da **população** da qual se vai selecionar a amostra.
- variáveis geográficas,
    - como as **unidades da federação** ou **municípios**,
- outros tipos,
    - tais como sexo, idade,
    - número de empregados na empresa, área do estabelecimento, etc.

Amostragem Estratificada

O método geral

- Em seguida, é feita a **seleção** de amostras **dentro** de cada um dos **estratos**, de forma **independente**.
    - A amostra final é formada então pela união das amostras selecionadas em cada um dos estratos.
- Selecione uma amostra s_h de tamanho n_h, com $n_h > 0$, segundo um plano amostral $p_h(s_h)$ *independentemente* dentro de cada estrato h,
    - $n = \sum_{h=1}^{H} n_h$ é o tamanho total da amostra selecionada.
- Fica assim assegurado que cada estrato tem sua população representada na amostra completa dada por: $s = s_1 \cup \cdots \cup s_h \cup \cdots \cup s_H$.
    - Pela definição dos estratos, as amostras nos vários estratos também são conjuntos mutuamente exclusivos, $s_h \cap s_k = \emptyset$, $h \neq k$.

Exemplo 🏋️‍♀️

Considere uma pesquisa feita em uma população de 8 domicílios, onde são conhecidas as variáveis renda familiar (Y) e local do domicílio (L), com os códigos A para região alta e B para baixa (B).

```
N <- 8
domicilio <- 1:N
y <- c( 13, 17, 6, 5, 10, 12, 19, 6)
l <- c( "B", "A", "B", "B", "B", "A", "A", "B")
```

domicilio	1	2	3	4	5	6	7	8
y	13	17	6	5	10	12	19	6
l	B	A	B	B	B	A	A	B

1 Calcule $Var(\overline{y})$ sob AASc, para estimar $\overline{Y} = \frac{T}{N} = \frac{\sum_i \in U y_i}{N}$, com $n = 4$.

2 Calcule $Var(\overline{y}_{AES})$, para $\overline{y}_{AES} = \frac{N_A \, \overline{y}_A + N_B \, \overline{y}_B}{N}$ (usando L como variável estratificadora), com amostra $n_h = 2$ por estrato.

3 Compare as variâncias dos dois planos (Efeito do Plano Amostral EPA).

Exemplo 🏃

Resolução:

$N = 8$ domicílios, $N_A = 3$, $N_B = 8$;
L: localidade domicílio, $H = 2$;
Y: renda familiar.

domicilio	1	2	3	4	5	6	7	8
y	13	17	6	5	10	12	19	6
l	B	A	B	B	B	A	A	B

1 Sob AASc $Var(\bar{y}) = Var_y/n$,

```
n <- 4
Ybarra <- mean(y)
vary <- sum( (y-Ybarra)^2) / N
varybarra <- vary / n
```

2 Para $Var(\bar{y}_{AES})$, precisamos $Var(\bar{y}_A) =$ ❓ e $Var(\bar{y}_B) =$ ❓,

```
yA <- y[l=="A"]; yB <- y[l=="B"]
N_A <- length(yA); N_B <- length
nA <- nB <- 2
```

3 Então $EPA = \frac{Var(\bar{y}_{AES})}{Var(\bar{y}_{AASc})} =$ ❓

```
## ... continuar...
```

Amostragem Estratificada

O método geral

- A independência da amostragem nos diferentes **estratos** permite olhar como **populações separadas**. Uma implementação segue o algoritmo:

1. Defina a **estratificação** da população U.

2. Defina o **método para selecionar** a amostra em cada um dos estratos.

3. **Selecione a amostra** de cada estrato usando o método definido em (2).

4. **Reúna as amostras** dos vários estratos formando a amostra da pesquisa.

Amostragem Estratificada

Importância

1. Quando os **estratos** formam **domínios "naturais"** ou de interesse,
    - ex. regiões geográficas, tipos de empresas (farmácias, supermercados, lojas de departamentos) etc.,
    - a estratificação garante a seleção de amostras de tamanhos especificados em todos os estratos formados,
    - permite controlar a precisão esperada de estimativas para subgrupos da população de pesquisa definidos como estratos ou agregações de estratos.
2. Pode tornar a **amostra mais "representativa"** e
    - assegurar que todas as partes (estratos) relevantes da população sejam incluídas na amostra.

Amostragem Estratificada

Importância

3. Para **melhorar a eficiência** amostral, **reduzir a variância** dos estimadores dos parâmetros de interesse:

- quanto maior for a **homogeneidade dentro dos estratos**,
- mais a estratificação permite aumentar a precisão de estimativas, em comparação com planos de igual tamanho de amostra que não usam estratificação.

4. Se necessário usar **métodos diferentes de coleta** em diferentes subgrupos da população;

- a estratificação favorece a administração e implementação da coleta em pesquisas onde as condições de pesquisa variam entre os estratos;
- ex. estratos podem ser formados de modo a viabilizar o emprego de modos alternativos de coleta (presencial, telefone, internet, etc.).

Amostragem Estratificada

Características

- Para pesquisas em que a **estimação** é objetivo:
    - o **processo de amostragem** e **estimação** deve ser replicado separadamente em cada estrato.
- Para pesquisas onde **não são de interesse** resultados **por estrato**,
    - **parâmetros dos estratos** são estimados separadamente, de acordo com o plano amostral adotado,
    - então as **estimativas** são **agregadas** para obter estimativas referentes ao conjunto da população.
- Como **desvantagens** potenciais do método
    - requerer a **reestruturação do cadastro** antes da amostragem.
    - apenas **uma estratificação** é possível e, uma vez fixada a estratificação, a amostragem vai depender dela de forma direta.
    - subdividir em **muitos estratos** pode levar a amostras muito pequenas em cada estrato e complicações como não resposta, ou 'instabilidade nas estimativas'.

Tipos de Estratificação

Tipos de Estratificação

1. *Natural* - quando os estratos são iguais a subgrupos da população para os quais se requer estimativas com precisão controlada.

- O processo requer essencialmente ouvir os clientes ou usuários dos resultados da pesquisa.
- São eles que devem indicar que subgrupos da população requerem estimativas com precisão controlada.
- Ex. pesquisas do IBGE, que além de produzir resultados para o país como um todo, têm que produzir também resultados por unidades da federação.
- Estas são então *estratos naturais* nas pesquisas, que passam a ter suas amostras planejadas de modo a ter seleção independente em cada unidade da federação, sendo tais amostras de tamanhos suficientes para estimar com a precisão desejada em cada uma.

Tipos de Estratificação

2. *Estatística* - quando os estratos são definidos como subgrupos homogêneos da população, visando aumentar eficiência na estimação para a população como um todo.

- Não há interesse específico na estimação de parâmetros dos estratos formados.
- sua formação pode ser feita empregando métodos que visam a otimizar os efeitos da estratificação.
- Alguns destes métodos são apresentados na sequência.

Tipos de Estratificação

- Na **prática**, não é incomum encontrar aplicações onde a estratificação usada numa pesquisa **combina os dois tipos**.

    - Ex. Na **amostragem de pesquisas econômicas estruturais do IBGE** as empresas são estratificadas por **unidade da federação** e **tipo de atividade econômica**, definindo assim seus **estratos naturais**.

    - Dentro destes estratos, para **aumentar a eficiência** da amostragem, as empresas são estratificadas por faixas de tamanho, usando a variável **pessoal ocupado**, **estratificação estatística**, pois tais pesquisas não buscam estimar parâmetros populacionais nestes estratos de tamanho (dentro dos estratos naturais). Ver IBGE(2000).

Há diversos fatores que influenciam a **eficiência na AE**:

- A(s) variável(is) de estratificação.
- O número de estratos.
- A determinação dos limites dos estratos.
- A alocação da amostra nos estratos.
- O método de seleção da amostra em cada estrato.

Tipos de Estratificação

- Para **estratificação natural**, a escolha da(s) variável(is) de estratificação se dá considerando TODAS as variáveis disponíveis necessárias para definir os domínios de interesse.

- No caso da **estratificação estatística**, a escolha deve priorizar, entre as variáveis disponíveis, as que são as *melhores preditoras* da(s) variável(is) de interesse da pesquisa. Para ganhar eficiência a ideia é tornar os valores da(s) variável(is) de estudo dentro de cada estrato mais similares / homogêneos possíveis; minimizar a *variância dentro dos estratos*.

- Nos dois casos é **fundamental** ter acesso a **cadastro(s)** com informações completas sobre variáveis auxiliares que são necessárias para estratificar a população de forma eficiente.

Exemplos de possíveis planos de **AE**

Exemplos de possíveis planos

Amostragem com estratos definidos por conveniência administrativa

- Considere uma **população de passageiros** chegando a um **terminal marítimo**,
    - navios carregam tanto passageiros viajando com seus **automóveis** como passageiros que viajam **a pé**.
    - ex. *UK International Passenger Survey* (Horsfield, 2017).
- Objetivo: selecionar uma amostra de passageiros para **estimar a média dos gastos feitos na viagem** por passageiro.

Automóvel - estrato 1

- Ex. *Amostragem Sistemática* para selecionar um de cada K automóveis cruzando um ponto de fluxo na saída do navio.
- As unidades de amostragem são os automóveis (com 1 ou mais passageiros), **Amostragem conglomerada**.

Passageiros a pé - estrato 2

- Ex. *Amostragem Binomial* para selecionar passageiros ao passarem por um ponto de fluxo na saída do navio.
- As unidades de amostragem seriam os passageiros individualmente.

Exemplos de possíveis planos

Amostragem Estratificada por Corte - AEC

- Muito usado em pesquisas de **estabelecimentos** ou **instituições**, em populações onde há grande **assimetria** das principais variáveis de interesse.

- Estratos formados por **partes da população** onde existe **grande heterogeneidade** ou que **concentram grande parte do total** de uma ou mais variáveis de interesse, porém compostas de um **número relativamente pequeno** de unidades populacionais,

    - Chamado *estrato certo*, nele se faz um **censo**. Nos demais estratos são pesquisadas amostras de suas unidades.

- Ex. **pesquisas econômicas** na área da **indústria** onde um número pequeno de estabelecimentos industriais é responsável por grande parte do valor da produção.

    - Neste caso, a precisão das estimativas de totais populacionais produzidas pode ser maior com **AEC** em comparação com **AAS**.
    - Como no **estrato certo** é feito um **censo**, não há variabilidade devida ao uso de amostragem nesse estrato (variância nula do total 'estimado' nesse estrato).

Exemplos de possíveis planos

Amostragem estratificada simples

- **Diferentes planos amostrais** podem ser empregados nos diversos estratos, mas isso é **pouco comum na prática**. O mais comum é usar um mesmo tipo de amostragem nos vários estratos definidos.

- A *Amostragem Estratificada Simples - AES* é o caso mais simples, em que uma **AASs** é selecionada em cada um dos estratos.

    - Neste caso, deve estar disponível um cadastro que permita alocar as unidades da população U nos estratos definidos, conhecer os tamanhos dos estratos, e o cadastro deve estar organizado de forma a permitir a seleção da amostra em cada um dos estratos.

- Veremos agora resultados para **AES**, mas o mesmo tipo de enfoque pode ser adaptado quando for utilizado qualquer outro método de seleção nos estratos,

    - tais como *Amostragem Sistemática, Amostragem Binomial, Amostragem com PPT*, etc.
    - A **seleção** é **independente** nos diferentes estratos, basta fazer as estimativas utilizando as fórmulas adequadas ao método de seleção em cada estrato, então agregar os resultados dessas estimativas de

AES, parâmetros e estimadores

Amostragem estratificada simples

Método de seleção

- Para cada estrato $h = 1, 2, \ldots, H$, selecione por AASs uma amostra s_h de tamanho $1 \leq n_h \leq N_h$ das N_h unidades do estrato U_h (independente da seleção feita nos outros estratos).

- Qualquer algoritmo para seleção de amostras AASs pode ser empregado para a seleção das amostras nos estratos.

- O conjunto de todas as amostras possíveis S_{AES} é formado por amostras da forma:

$$s = s_1 \cup \cdots \cup s_h \cup \cdots \cup s_H =.$$

- O tamanho é dado pelo produto dos tamanhos dos conjuntos de amostras possíveis em cada um dos estratos.

$$\#S_{AES} = \#s_1 \times \ldots \times \#s_H = \prod_{h=1}^{H} \binom{N_h}{n_h} = \nu.$$

Amostragem estratificada simples

Método de seleção

- Assim, na **AASs** em cada estrato, temos o plano amostral:

$$p_h(s_h) = 1\bigg/\binom{N_h}{n_h} = \binom{N_h}{n_h}^{-1}, \forall h = 1, 2, \ldots, H.$$

- Em consequência, o plano amostral $p_{AES}(s)$ é dado por:

$$p_{AES}(s) = \prod_{h=1}^{H} p_h(s_h) = \prod_{h=1}^{H} \binom{N_h}{n_h}^{-1}$$

onde $s \in S_{AES}$.

O *tamanho total da amostra* é:

$$n = n_1 + n_2 + \cdots + n_H.$$

Amostragem estratificada simples

Parâmetros nos estratos

- **Total** populacional do **estrato** h:

$$T_h = \sum_{i \in U_h} y_i$$

- **Média** populacional do **estrato** h

$$\overline{Y_h} = T_h / N_h = \frac{1}{N_h} \sum_{i \in U_h} y_i$$

- **Variância** do **estrato** h

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_i - \overline{Y_h})^2 \quad \text{ou} \quad Var_h = \frac{N_h - 1}{N_h} S_h^2.$$

Amostragem estratificada simples

Parâmetros globais

- O **total populacional**:

$$T = \sum_{h=1}^{H} Y_h = \sum_{h=1}^{H} N_h \overline{Y_h}$$

- A **média populacional**:

$$\overline{Y} = T/N = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{Y_h} = \sum_{h=1}^{H} W_h \overline{Y}_h,$$

onde $W_h = N_h/N$ é o *peso* da média do estrato h na composição de $\overline{Y}$.

- A **variância populacional** pode ser escrita como

$$S_y^2 = \frac{1}{N-1} \sum_{h=1}^{H} \sum_{i \in U_h} \left(y_i - \overline{Y}\right)^2 \quad \text{ou} \quad Var_y = \frac{1}{N} \sum_{h=1}^{H} \sum_{i \in U_h} \left(y_i - \overline{Y}\right)^2 = \frac{N-1}{N} S_y^2.$$

Amostragem estratificada simples

Parâmetros globais

- A variância populacional S_y^2 pode ser escrita como:

$$
\begin{aligned}
S_y^2 &= \frac{1}{N-1}\sum_{h=1}^{H}\sum_{i\in U_h}\left(y_i-\overline{Y}\right)^2 \\
&= \frac{1}{N-1}\sum_{h=1}^{H}\sum_{i\in U_h}\left[\left(y_i-\overline{Y_h}\right)+\left(\overline{Y_h}-\overline{Y}\right)\right]^2 \\
&= \frac{1}{N-1}\sum_{h=1}^{H}(N_h-1)S_h^2+\frac{1}{N-1}\sum_{h=1}^{H}N_h\left(\overline{Y_h}-\overline{Y}\right)^2 \\
&= S_D^2+S_E^2,
\end{aligned}
$$

onde $S_D^2=\sum_{h=1}^{H}\frac{N_h-1}{N-1}S_h^2$ e $S_E^2=\sum_{h=1}^{H}\frac{N_h}{N-1}\left(\overline{Y_h}-\overline{Y}\right)^2$

- ou

$$
Var_y=\sum_{h=1}^{H}W_hVar_h+\sum_{h=1}^{H}W_h\left(\overline{Y_h}-\overline{Y}\right)^2=Var_D+Var_E
$$

Amostragem estratificada simples

Parâmetros globais

- $S_y^2 = S_D^2 + S_E^2$ ou $Var_y = Var_D + Var_E$ sugerem uma **decomposição** útil da variância total da variável y na população.
    - As componentes são conhecidas como **Variância Dentro** e **Variância Entre** estratos.
- Selecionar **amostras** de forma **independente** em **todos** os **estratos**, a amostragem estratificada **elimina** a componente de **variação entre** os estratos ao estimar parâmetros do conjunto da população, ex. o total ou a média.
- Para uma **variância total** S_y^2 **fixada**, **minimizar** a **Variância Dentro** S_D^2, ao definir estratos homogêneos (em relação à variável de interesse), deve reduzir grande parte da variação relevante para a estimação.

Amostragem estratificada simples

Estimação no estrato h

Temos **amostragem independentemente** em cada estrato, então os seguintes **estimadores**:

- do **total** T_h do **estrato** h,

$$\widehat{T}_h = \sum_{i \in s_h} d_i y_i = \frac{N_h}{n_h} \sum_{i \in s_h} y_i = N_h \bar{y}_h;$$

onde $d_i = N_h/n_h = \pi_i^{-1}$ é o peso das unidades i dentro do estrato h.

- da **média** $\overline{Y}_h$ do **estrato** h,

$$\bar{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_i;$$

- da **variância** S_h^2 ou Var_h do estrato h,

$$\widehat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} \left(y_i - \bar{y}_h\right)^2 = s_h^2.$$

Amostragem estratificada simples

Estimação no estrato h

- Se temos **AASs** de n_h unidades **dentro do estrato** h, **sabemos que** são válidas as seguintes propriedades:
  - $E_{AES}\left(\widehat{T}_h\right) = T_h$
  - $E_{AES}\left(\overline{y}_h\right) = \overline{Y_h}$
  - $E_{AES}\left(\widehat{S}_h^2\right) = S_h^2 \qquad \Rightarrow \widehat{S}_h^2$ é **ENV** de S_h^2 sob **AASs**
- Os resultados decorrem das **propriedades** de **estimadores** sob **AASs**.
- Sabemos mostrar esses resultados❓❓❓
- E na **AASc**❓❓❓
  - $E_{AES}\left(\widehat{S}_h^2\right) = Var_h \qquad \Rightarrow \widehat{S}_h^2$ é **ENV** de Var_h sob **AASc**

Amostragem estratificada simples

Estimação no estrato h

- **Variâncias** de estimadores de média e total por estrato na **AASs**:

Total	Média
$Var_{AES}\left(\widehat{T}_h\right) = N_h^2 Var_{AES}\left(\bar{y}_h\right)$	$Var_{AES}\left(\bar{y}_h\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$
$\widehat{Var}_{AES}\left(\widehat{T}_h\right) = N_h^2 \widehat{Var}_{AES}\left(\bar{y}_h\right)$	$\widehat{Var}_{AES}\left(\bar{y}_h\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$

- E na **AASc**❓❓❓

Total	Média
$Var_{AESc}\left(\widehat{T}_h\right) = N_h^2 Var_{AESc}\left(\bar{y}_h\right)$	$Var_{AESc}\left(\bar{y}_h\right) = \frac{1}{n_h} Var_h$
$\widehat{Var}_{AESc}\left(\widehat{T}_h\right) = N_h^2 \widehat{Var}_{AESc}\left(\bar{y}_h\right)$	$\widehat{Var}_{AESc}\left(\bar{y}_h\right) = \frac{1}{n_h} \widehat{S}_h^2$

Amostragem estratificada simples

Estimação de parâmetros globais

- O estimador do total T: $\widehat{T}_{AES} = \sum_{h=1}^{H} \widehat{T}_h = \sum_{h=1}^{H} N_h \overline{y}_h$.
- O estimador da média $\overline{Y}$: $\overline{y}_{AES} = \sum_{h=1}^{H} W_h \overline{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h$.

O **estimador** da média populacional sob **AES** é **não viciado**, isto é:

$$E_{AES}\left(\overline{y}_{AES}\right) = \overline{Y}$$

Isto segue porque $E_{AES}\left(\overline{y}_h\right) = \overline{Y_h}, \forall\, h = 1, \ldots, H$, e

$$E_{AES}\left(\sum_{h=1}^{H} W_h \overline{y}_h\right) = \sum_{h=1}^{H} W_h E_{AES}(\overline{y}_h) = \sum_{h=1}^{H} W_h \overline{Y_h} = \overline{Y}.$$

Amostragem estratificada simples

Estimação de parâmetros globais

- Para estimar a **variância do estimador** do **total** e da **média**, respectivamente, (dado **AASs** dentro dos estratos)

$$Var_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$$

e

$$Var_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 Var_{AES}\left(\overline{y}_h\right) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2,$$

- temos os **ENV**s com as expressões dadas por

$$\widehat{Var}_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$$

e

$$\widehat{Var}_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2.$$

- E quanto a **AASc**❓❓❓

Amostragem estratificada simples

Intervalo de confiança para a média $\overline{Y}$

Se $n = \sum_{h=1}^{H} n_h$ for grande, então o Teorema Central do Limite se aplica:

$$\frac{\bar{y}_{AES} - \overline{Y}}{\sqrt{\widehat{Var}_{AES}\left(\bar{y}_{AES}\right)}} \approx Normal(0;1)$$

Logo, um intervalo de confiança de nível $1-\alpha$ para $\overline{Y}$ é dado por:

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

Amostragem estratificada simples

Intervalo de confiança para a média do estrato $\overline{Y}_h$

Se os tamanhos de amostras *por estratos* n_h são suficientemente grandes, o Teorema Central do Limite também indica que:

$$\frac{\bar{y}_h - \overline{Y_h}}{\sqrt{\widehat{Var}_{AES}\left(\bar{y}_h\right)}} \approx Normal(0;1)$$

e então um intervalo de confiança de nível $1-\alpha$ para $\overline{Y_h}$ é dado por:

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

Para casa

- Continuar o Exemplo.
- Mostrar $E_{AES}\left(\widehat{T}_{AES}\right) = T$.
- Encontre $Var_{AES}\left(\widehat{T}_{AES}\right)$, tanto para **AASs** quanto para **AASc**.
- Fazer exercícios.
- Rever os slides.
- Ler a partir seção 11.3 do livro 'Amostragem: Teoria e Prática Usando R'.

Próxima aula

- Amostragem Estratificada
    - Alocação de amostras nos estratos
- Laboratório de R

Muito obrigado!

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

Resumo da notação

Parâmetros no estrato h

- Total - $T_h = \sum_{i \in U_h} y_i$
- Média - $\overline{Y_h} = T_h / N_h = \frac{1}{N_h} \sum_{i \in U_h} y_i$
- Variância - $S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_i - \overline{Y_h})^2$ ou $Var_h = Var_{h,y} = \frac{N_h - 1}{N_h} S_h^2$

Parâmetros globais

- Total - $T = \sum_{h=1}^{H} T_h = \sum_{h=1}^{H} N_h \overline{Y_h}$
- Média - $\overline{Y} = T/N = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{Y_h} = \sum_{h=1}^{H} W_h \overline{Y}_h, W_h = N_h / N$
- Variância - $S^2 = \frac{1}{N-1} \sum_{h=1}^{H} \sum_{i \in U_h} \left(y_i - \overline{Y} \right)^2$ ou $Var_y = \frac{1}{N} \sum_{h=1}^{H} (N_h - 1) S_h^2$

Decomposição da variância populacional

$$S_y^2 = \frac{1}{N-1} \sum_{h=1}^{H} (N_h - 1) S_h^2 + \frac{1}{N-1} \sum_{h=1}^{H} N_h \left(\overline{Y_h} - \overline{Y} \right)^2 = S_D^2 + S_E^2$$

ou

$$Var_y = \sum_{h=1}^{H} W_h Var_h + \sum_{h=1}^{H} W_h \left(\overline{Y_h} - \overline{Y} \right)^2 = Var_D + Var_E.$$

Resumo da notação

Estimadores no estrato h

Estimadores AASc	Estimadores AASs
$\widehat{T}_h = \frac{N_h}{n_h} \sum_{i \in s_h} y_i = N_h \overline{y}_h$	$\widehat{T}_h = \frac{N_h}{n_h} \sum_{i \in s_h} y_i = N_h \overline{y}_h$
$\overline{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_i$	$\overline{y}_h = \frac{1}{n_h} \sum_{i \in s_h} y_i$
$Var_{AESc}\left(\widehat{T}_h\right) = N_h^2 Var_{AESc}\left(\overline{y}_h\right)$	$Var_{AES}\left(\widehat{T}_h\right) = N_h^2 Var_{AES}\left(\overline{y}_h\right)$
$Var_{AESc}\left(\overline{y}_h\right) = \frac{1}{n_h} Var_h$	$Var_{AES}\left(\overline{y}_h\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$
$\widehat{Var}_{AESc}\left(\widehat{T}_h\right) = N_h^2 \widehat{Var}_{AESc}\left(\overline{y}_h\right)$	$\widehat{Var}_{AES}\left(\widehat{T}_h\right) = N_h^2 \widehat{Var}_{AES}\left(\overline{y}_h\right)$
$\widehat{Var}_{AESc}\left(\overline{y}_h\right) = \frac{1}{n_h} \widehat{S}_h^2$	$\widehat{Var}_{AES}\left(\overline{y}_h\right) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$

em que $Var_h = Var_h(y) = \frac{N_h - 1}{N_h} S_h^2$ e $\widehat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} \left(y_i - \overline{y}_h\right)^2$.

Resumo da notação

Estimadores globais

- Do total T: $\widehat{T}_{AES} = \sum_{h=1}^{H} \widehat{T}_h = \sum_{h=1}^{H} N_h \overline{y}_h.$
- Da média $\overline{Y}$: $\overline{y}_{AES} = \sum_{h=1}^{H} W_h \overline{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h.$

Variância do estimador e seu **estimador**

Sob AASc	Sob AASs
$Var_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \frac{Var_h}{n_h}$	$Var_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$
$Var_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \frac{Var_h}{n_h}$	$Var_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$
$\widehat{Var}_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \frac{\widehat{S}_h^2}{n_h}$	$\widehat{Var}_{AES}\left(\widehat{T}_{AES}\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$
$\widehat{Var}_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \frac{\widehat{S}_h^2}{n_h}$	$\widehat{Var}_{AES}\left(\overline{y}_{AES}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \widehat{S}_h^2$

Referências

Slides baseados no Capítulo 11 do livro

- Amostragem: Teoria e Prática Usando o R

Citações do Capítulo

- Horsfield(2017)
- IBGE(2000)