Técnicas de Amostragem - Aula 08

Amostragem Estratificada I

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

Conteúdo

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

- definição da AASc e AASs;
- propriedades das principais estatísticas;
- normalidade e intervalo de confiança;
- tamanho da amostra;
- fizemos uma breve introdução aos geradores de números aleatórios.

Primeiras Definições e Propriedades

Definição 2.1

A amostragem **estratificada** consiste na divisão de uma população em grupos (estratos) segundo alguma(s) característica(s) conhecida(s) na população sob estudo, e de cada um desses estratos são selecionadas amostras em proporções convenientes.

Primeiras Definições e Propriedades

A estratificação é usada principalmente para resolver alguns problemas como:

- a melhoria da precisão das estimativas;
- produzir estimativas para a população toda e subpopulações;
- questões administrativas.

Na nossa disciplina focaremos no primeiro motivo: **a melhoria da precisão das estimativas**.

Foi visto (Teorema 4.3 Aula-03) que para uma amostra AASc de tamanho n, a variância do estimador média amostral $\overline{y}$, é dada por

$$\mathrm{Var}[\overline{y}] = \frac{\sigma^2}{n}.$$

Primeiras Definições e Propriedades

Se a população é muito heterogênea e as razões de custo limitam o aumento da amostra, torna-se impossível definir uma AASc da população toda com uma precisão razoável.

Uma saída para esse problema é dividir a população em subpopulações internamente mais homogêneas, ou seja, grupos com variâncias σ^2 pequenas que diminuirão o erro amostral global.

Exemplo 2.2

Considere uma pesquisa feita em uma população com $N = 8$ domicílios, onde são conhecidas as variáveis renda domiciliar (Y) e local do domicílio (W), com os códigos A para região alta e B para região baixa. Tem-se então,

$$\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8\},$$

com

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} 13 & 17 & 6 & 5 & 10 & 12 & 19 & 6 \\ B & A & B & B & B & A & A & B \end{pmatrix}$$

Vamos analisar como se comportam os parâmetros de $\boldsymbol{D}$ em função dos parâmetros das subpopulações determinadas pelos estratos A e B.

O resultado da estratificação será mais eficaz quanto maior for a habilidade do pesquisador em produzir estratos homogêneos.

O caso limite é aquele onde consegue-se a homogeneidade máxima (variância nula dentro de cada estrato) onde então a estimativa acerta o parâmetro populacional.

Primeiras Definições e Propriedades

A simples estratificação por si só não produz necessariamente estimativas mais eficientes que a AAS. O próximo exemplo ilustra tal situação.

Exemplo 2.3

Considere uma pesquisa feita em uma população com $N = 8$ domicílios, onde são conhecidas as variáveis renda domiciliar (Y) e local do domicílio (W), com os códigos A para região alta e B para região baixa. Tem-se então,

$$\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8\},$$

divididas nos estratos

$$\boldsymbol{D}_1 = (13, 17, 6, 5) \text{ e } \boldsymbol{D}_2 = (10, 12, 19, 6).$$

Vamos analisar como se comportam os parâmetros de $\boldsymbol{D}$ em função dos parâmetros das subpopulações determinadas pelos estratos 1 e 2.

A execução de um plano de amostragem estratificada (AE) exige os seguintes passos:

Passo 1

Divisão da população em subpopulações bem definidas (estratos).

Passo 2

De cada estrato retira-se uma amostra, usualmente independentes.

Primeiras Definições e Propriedades

Passo 3

Em cada amostra usam-se estimadores convenientes para os parâmetros do estrato.

Primeiras Definições e Propriedades

Passo 4

Monta-se para a população um estimador combinando os estimadores de cada estrato, e determinam-se suas propriedades.

Notações

Notações

Estrato	Dados	Total	Média	Variância
1	$\mathbf{Y}_1$ *	τ_1	$\mu_1 = \overline{Y}_1$	σ_1^2 ou S_1^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
h	$\mathbf{Y}_h$ *	τ_h	$\mu_h = \overline{Y}_h$	σ_h^2 ou S_h^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
H	$\mathbf{Y}_H$ *	τ_H	$\mu_H = \overline{Y}_H$	σ_H^2 ou S_H^2

* onde $\mathbf{Y}'_h = (Y_{h1}, \ldots, Y_{hN_h})$ é o vetor de dados no estrato h, $h = 1, \ldots, H$.

Notações

Vamos definir várias notações envolvendo os dados estratificados como acima. Para exemplificá-las, usaremos os dados à seguir:

Notações

Nome	Categoria	Nota	Nome	Categoria	Nota
Enedina	A	10	Leopoldina	A	3
Machado	A	4	Dandara	C	8
Luiz	A	5	Francisco	A	6
Marilena	B	3	Felipa	A	7
Clarice	B	9	Menininha	C	10
Heitor	B	2	Erenilton	C	9
Camargo	B	8	Vadinho	C	8
Rita	B	10	Jorge	C	3

Notações

Considere uma população bem descrita por um sistema de referências, ou seja,

$$\mathcal{U} = \{1, 2, ..., N\}$$

e que existe uma partição $\mathcal{U}_1, ..., \mathcal{U}_H$ de $\mathcal{U}$, ou seja,

$$\mathcal{U} = \bigcup_{h=1}^{H} \mathcal{U}_h \text{ e } \mathcal{U}_h \cap \mathcal{U}_l = \emptyset$$

se $h \neq l$.

Notações

Além disso, vamos supor que cada subconjunto $\mathcal{U}_h$, bem determinado, é identificado por duplas ordenadas, do seguinte modo:

$$\mathcal{U}_h = \{(h, 1), (h, 2), ..., (h, N_h)\}.$$

Notações

Assim o universo todo pode ser descrito por

$$\mathcal{U} = \{(1,1), ..., (h, N_1), ..., (h,1), ..., (h, N_h), ..., (H,2), ..., (H, N_H)\}$$

de modo a facilitar a identificação do estrato e do elemento dentro dele.

Notações

De modo análogo, as características populacionais serão identificadas por dois índices, ou seja, no caso univariado, por exemplo, tem-se o vetor de características populacionais

$$\boldsymbol{D} = (\boldsymbol{Y}_{11}, ..., \boldsymbol{Y}_{1N_1}, ..., \boldsymbol{Y}_{hi}, ..., \boldsymbol{Y}_{HN_H}),$$

ou seja, para o estrato 1 tem-se as características populacionais $\boldsymbol{Y}_{11}, ..., \boldsymbol{Y}_{1N_1}$, e assim por diante.

Notações

Eis algumas definições e relações entre os parâmetros:

Notações

- Tamanho do estrato h: N_h.

Notações

- Total do estrato h:

$$\tau_h = \sum_{i=1}^{N_h} Y_{hi}.$$

Notações

- Média do estrato h:

$$\mu_h = \overline{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}.$$

Notações

- Variância do estrato h:

$$\sigma_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h}(Y_{hi} - \mu_h)^2 \text{ ou } S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(Y_{hi} - \mu_h)^2.$$

Notações

- Tamanho do universo:
$$N = \sum_{h=1}^{H} N_h.$$

Notações

- Peso (proporção) do estrato h:

$$W_h = \frac{N_h}{N} \text{ com } \sum_{h=1}^{H} W_h = 1.$$

Notações

- Total populacional:

$$\tau = \sum_{h=1}^{H} \tau_h = \sum_{h=1}^{H} \sum_{i=1}^{N_H} Y_{hi} = \sum_{h=1}^{H} N_h \mu_h.$$

Notações

- Média populacional:

$$\mu = \overline{Y} = \frac{\tau}{N} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_H} Y_{hi} = \frac{1}{N}\sum_{h=1}^{H} N_h\mu_h = \sum_{h=1}^{H} W_h\mu_h,$$

de modo que a média global é a média ponderada dos estratos.

Notações

- Variância Populacional:

$$\sigma^2 = \sum_{h=1}^{H} W_h \sigma_h^2 + \sum_{h=1}^{H} W_h(\mu_h - \mu)^2.$$

  Também denotamos

$$\sigma^2 = \sigma_d^2 + \sigma_e^2 \text{ com}$$
$$\sigma_d^2 = \sum_{h=1}^{H} W_h \sigma_h^2 \text{ e } \sigma_e^2 = \sum_{h=1}^{H} W_h(\mu_h - \mu)^2.$$

Notações

- Variância Populacional:

$$S^2 = \sum_{h=1}^{H} \frac{N_h - 1}{N - 1} S_h^2 + \sum_{h=1}^{H} \frac{N_h}{N - 1} (\mu_h - \mu)^2.$$

Notações

- Para estratos grandes temos

$$S^2 \approx \sigma_d^2 + \sigma_e^2 \approx S_d^2 + \sigma_e^2 \text{ com}$$

$$S_d^2 = \sum_{h=1}^{H} W_h S_h^2.$$

Notações

Note que quando todos os estratos têm a mesma média, ou seja, $\mu_h = \mu$, $h = 1, ..., H$, a variância populacional σ^2 coincide com σ_d^2.

Notações

Quanto maior for σ_e^2, maior é a diferença $\sigma^2 - \sigma_d^2$.

Notações

As nomenclaturas para as estatísticas mais usadas (média, total e variância amostrais) são análogas: $\overline{y}_h, T_h, s_h^2$ respectivamente.

Notações

Lembre que, se $X_1, X_2, ..., X_H$ são variáveis aleatórias independentes, então para $X = \sum_{h=1}^{H} l_h X_h$,

$$E[X] = \sum_{h=1}^{H} l_h E[X_h] \text{ e } \mathrm{Var}[X] = \sum_{h=1}^{H} l_h^2 \mathrm{Var}[X].$$

Estimação do Total e da Média Populacional

Considere a seguinte situação:

Estimação do Total e da Média Populacional

- uma população estratificada;

- de cada estrato foi sorteada independentemente uma amostra de tamanho n_h, podendo ou não ter sido usado o mesmo plano amostral em cada estrato;

- e consideremos $\hat{\mu}_h$ um estimador não viesado para a média populacional μ_h do estrato h, ou seja, $E_A[\hat{\mu}_h] = \mu_h$, onde A é o plano amostral usado no estrato h.

Estimação do Total e da Média Populacional

Teorema 4.1

O estimador

$$\overline{y}_{es} = \frac{1}{N}\sum_{h=1}^{H} N_h\hat{\mu}_h = \sum_{h=1}^{H} W_h\hat{\mu}_h$$

é não-viesado para a média populacional μ e

$$\mathrm{Var}_A[\overline{y}_{es}] = \sum_{h=1}^{H} W_h^2 \mathrm{Var}_A[\hat{\mu}_h].$$

Corolário 4.2

Considere agora que, dentro de cada estrato, a amostra foi sorteada por um processo AASc e que $\hat{\mu}_h = \overline{y}_h$. Então

$$\overline{y}_{es} = \sum_{h=1}^{H} N_h \overline{y}_h \text{ e } \mathrm{Var}[\overline{y}_{es}] = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h},$$

com estimador não-viesado para $\mathrm{Var}[\overline{y}_{es}]$ dado por

$$\mathrm{var}[\overline{y}_{es}] = \sum_{h=1}^{H} W_h^2 \frac{s_h^2}{n_h}.$$

Estimação do Total e da Média Populacional

Este procedimento (e sua variante sem reposição) é um dos planos amostrais mais usados em problemas reais.

Alocação da Amostra pelos Estratos

Alocação da Amostra pelos Estratos

Definição 5.1

A distribuição das n unidades da amostra pelos estratos chama-se **alocação da amostra**.

Alocação da Amostra pelos Estratos

Essa distribuição é muito importante pois ela irá garantir a precisão do procedimento amostral.

Exemplo 5.2

Considere a população $\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8\}$ do Exemplo 2.1 com a estratificação

$$\mathcal{U}_1 = \{2, 4, 7\} \text{ com } \boldsymbol{D}_1 = (17, 5, 19)$$
$$\mathcal{U}_2 = \{1, 3, 5, 6, 8\} \text{ com } \boldsymbol{D}_2 = (13, 6, 10, 12, 6).$$

Vamos estudar o efeito do planejamento (EPA) para duas situações:

1. AL_1: AASs com $n_1 = 1$ e $n_2 = 2$;
2. AL_2: AASs com $n_1 = 2$ e $n_2 = 1$.

Alocação da Amostra pelos Estratos

Vamos retomar os dados da tabela que usamos para as Notações:

Alocação da Amostra pelos Estratos

Nome	Categoria	Nota	Nome	Categoria	Nota
Enedina	A	10	Leopoldina	A	3
Machado	A	4	Dandara	C	8
Luiz	A	5	Francisco	A	6
Marilena	B	3	Felipa	A	7
Clarice	B	9	Menininha	C	10
Heitor	B	2	Erenilton	C	9
Camargo	B	8	Vadinho	C	8
Rita	B	10	Jorge	C	3

Exemplo 5.3

1. Para os estratos A,B,C, realize a alocação AL_k com uma amostragem AASc com $n_A = 2$, $n_B = 3$ e $n_C = 2$. Para esta alocação, calcule $\mathrm{Var}_{\mathrm{AL}_k}[\overline{y}_{es}]$.
2. Defina uma nova alocação AL_e de sua preferência e calcule $\mathrm{Var}_{\mathrm{AL}_e}[\overline{y}_{es}]$.
3. Calcule o efeito do planejamento (EPA) entre as alocações AL_k e AL_e.

Comentários Finais

Comentários Finais

Em resumo, na aula de hoje nós:

- introduzimos a amostragem estratificada;
- listamos as notações para esse plano amostral;
- falamos da estimação do total e média populacional;
- lidamos coma a alocação do tamanho da amostra entre os estratos.

Comentários Finais

Nas próximas aulas nós vamos continuar focar em:

- tipos de alocação;
- normalidade assintótica e intervalo de confiança;
- tamanho da amostra;
- estimação da proporção.

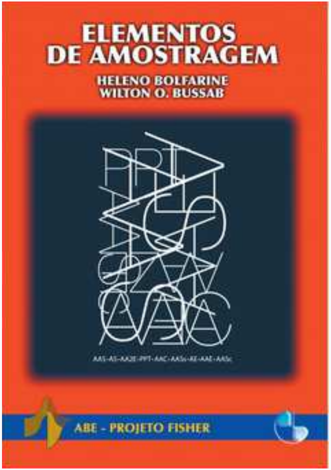
Comentários Finais

EXERCÍCIOS PARA APS (E PREPARAÇÃO PARA A N2)

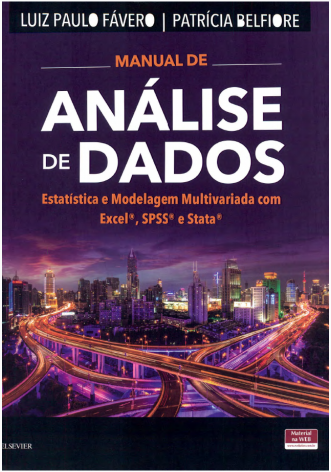
Resolva os Exercícios 8.1-8.3.

Referências

Referências



Referências



Bons Estudos!