AMOSTRAGEM

Unidade 10

Amostragem Estratificada: Detalhamento da Implementação

Fatores Que Influenciam a Eficiência na AE

1. Escolha da(s) variável(is) de estratificação.
2. Número de estratos.
3. Determinação dos limites dos estratos.
4. Alocação da amostra nos estratos.
5. Método de seleção em cada estrato.

Escolha da(s) Variável(is) de Estratificação

1. Para **estratificação natural**, considere **TODAS** as variáveis disponíveis com as quais são definidos os estratos naturais ou domínios de interesse da pesquisa.
2. Para **estratificação estatística**, escolha entre as variáveis disponíveis as que são **melhores preditoras** das variáveis de interesse da pesquisa.
3. Quando apenas uma variável de estratificação está disponível, não há o que escolher...

Alocação Proporcional

Uma amostra representativa deveria “imitar” ou se parecer bastante com a população de onde foi extraída.

As unidades populacionais são distribuídas nos estratos segundo as proporções:

$$W_h = N_h / N \ , \ h = 1, \ldots, H \ , \ \text{com} \quad \sum_h W_h = 1,$$

As proporções amostrais nos estratos são definidas como:

$$\lambda_h = n_h / n \ , \ h = 1, \ldots, H \ , \ \text{com} \quad \sum_h \lambda_h = 1,$$

Alocação Proporcional

Então o critério acima sugeriria tentar fazer $\lambda_h = W_h$ $\forall h=1,2,...,H$.

Isto implica fazer $\frac{n_h}{n} = \frac{N_h}{N}$ ou $n_h = n\frac{N_h}{N} = nW_h$, $\forall h=1,2,...,H$.

Esta distribuição da amostra nos estratos é chamada **Alocação Proporcional**.

Comentário

Se $n_h = nW_h$, então

$$\bar{y}_{AES} = \sum_h W_h \bar{y}_h = \sum_h \frac{W_h}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

$$= \frac{1}{n} \sum_h \sum_{i=1}^{n_h} y_{hi} = \bar{y}$$

➔ Sob alocação proporcional, a média amostral simples é o estimador não viciado da média populacional.

Variância Sob Alocação Proporcional

Quando $n_h = nW_h$, a variância de $\bar{y}_{AES}$ simplifica para:

$$V_{AES/Prop}(\bar{y}_{AES}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_h W_h S_h^2$$

A expressão $\sum_h W_h S_h^2 = S_D^2$ é a **variância dentro** dos estratos, dada por uma média ponderada dos S_h^2. Então:

$$V_{AES/Prop}(\bar{y}_{AES}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_D^2$$

Variância Sob Alocação Proporcional

Esta expressão tem a mesma forma que a correspondente ao caso de AAS, com S_y^2 substituído por S_D^2.

Como a variância dentro é geralmente menor que a variância total ($S_D^2 < S_y^2$), fica evidenciado que estratificação com alocação proporcional geralmente reduz a variância do estimador quando comparada com AAS de igual tamanho.

Alocação Ótima

A maioria das pesquisas sofre restrições orçamentárias.

Se o custo total da pesquisa é fixado em C unidades monetárias, então é necessário especificar uma função custo que descreva como varia esse custo para diferentes tamanhos amostrais e alternativas de alocação.

Exemplo: **Função Custo Linear**

$$C = c_0 + \sum_h n_h c_h .$$

O Problema

Minimizar $V_{AES}(\bar{y}_{AES})$ sujeito à restrição de não ultrapassar o orçamento previsto (custo total C).

Solução:

$$V_{AES}(\bar{y}_{AES}) = \sum_{h=1}^{H} W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right)$$

$$= \sum_{h=1}^{H} W_h^2 S_h^2 / n_h - V_0$$

onde $V_0 = \sum_{h=1}^{H} W_h^2 S_h^2 / N_h .$

Alocação Ótima

Como V_0 não depende de n_h, minimizando $V_{AES}(\bar{y}_{AES})$ sujeito a $C = c_0 + \sum_h n_h c_h$ resulta em:

$$n_h \propto \left(W_h^2 S_h^2 / c_h \right)^{1/2} = W_h S_h / \sqrt{c_h}\,.$$

Isto é:

$$n_h = n \times \left[W_h S_h / \sqrt{c_h} \Big/ \textstyle\sum_{k=1}^{H} W_k S_k / \sqrt{c_k} \right]$$

Esta alocação é chamada **Alocação Ótima**.

Comentários

1. Sob a alocação ótima, selecione uma amostra maior num estrato h sempre que:
   a. o estrato tiver mais unidades (N_h grande);
   b. a variabilidade no estrato for maior (S_h grande);
   c. o custo de amostragem no estrato for menor (c_h pequeno).
2. Se $S_h = S_*$ e $c_h = c_*$ $\forall h=1,2,...,H$, ambos constantes, então $n_h \propto N_h$, isto é, a alocação ótima coincide com a alocação proporcional.

Comentários

3. Se $c_h = c_* \ \forall h=1,2,...,H$, isto é, os custos são constantes ao longo dos estratos, então $n_h \propto N_h S_h$, gerando a chamada **Alocação (Ótima) de Neyman**. Esta alocação é muito usada em pesquisas de estabelecimentos quando os desvios padrões S_h crescem com o tamanho das unidades.

4. Para um custo fixado C, assumindo função linear de custos $C = c_0 + \sum_h n_h c_h$, o tamanho total da amostra n é:

$$n = (C - c_0) \times \left[\sum_h N_h S_h / \sqrt{c_h} \Big/ \sum_h N_h S_h \sqrt{c_h} \right]$$

Comentários

5. Se **Alocação de Neyman** é usada, então o valor da variância correspondente ao mínimo é dado por

$$V_{AES/Ney}(\bar{y}_{AES}) = \frac{1}{n}\left(\sum_{h=1}^{H} W_h S_h\right)^2 - \frac{1}{N}\left(\sum_{h=1}^{H} W_h S_h^2\right)$$

O segundo termo à direita corresponde à correção de população finita.

6. As soluções acima são 'aproximadas', pois ignoram restrições do tipo $n_h \le N_h, n_h \ge 1, n_h$ inteiro $\forall h$.
   Brito (2005) oferece uma solução 'exata'.

Comparação de Alternativas de Alocação da Amostra

Usando a partição da soma de quadrados total em parcelas devidas à variação dentro e entre estratos, e ignorando termos de ordem $1/N_h$, então sob alocação de Neyman, isto é, com $n_h \propto N_h S_h$ prova-se (Cochran, 1977, p. 99) que:

$$V_{AES/Ney}(\bar{y}_{AES}) \leq V_{AES/Prop}(\bar{y}_{AES}) \leq V_{AAS}(\bar{y})$$

AES com alocação de Neyman é mais eficiente que AES com alocação proporcional, ambas superando AAS como plano amostral.

Alguns Problemas Com Alocação Ótima

(1) S_h, h=1,...,H, são desconhecidos.

(2) Pode haver muitas variáveis de pesquisa.

(3) $n_h > N_h$ em alguns casos.

(4) $n_h = 1$ em alguns casos.

(5) Ganhos de eficiência podem ser modestos, particularmente para estimação de proporções.

Soluções Possíveis

(1) S_h, h=1,...,H, são desconhecidos.

- ✓ Usar informação de variável auxiliar x.
- ✓ Usar S_{hx} para estimar S_{hy}.
- ✓ Predizer y_{hi} usando x_{hi}, então estimar S_{hy}.
- ✓ Usar a soma ou a amplitude de x_{hi} no estrato h como proxy para S_{hy}.
- ✓ Selecionar pequena amostra piloto (preliminar) e usar dados desta amostra para estimar S_{yh}.

Soluções Possíveis

(2) Muitas Variáveis de Pesquisa.
Cada variável usualmente levaria a uma alocação ótima diferente.

Qualquer método deve buscar um compromisso entre as diversas alternativas.

- ✓ Tome a média das alocações alternativas;
- ✓ Escolha uma ou duas variáveis principais;
- ✓ Use alocação proporcional.
- ✓ Construa um 'índice' das variáveis de pesquisa e use este índice para definir a alocação.

Soluções Possíveis

(3) $n_h > N_h$ para algum h →

Ponha $n_h = N_h$. (um estrato certo) e refaça a alocação ótima nos demais estratos.

(4) $n_h = 1$ →

Se estimação de variâncias for importante, então force $n_h \geq 2$. Na prática, costuma-se fazer $n_h \geq 5$ devido à não resposta.

Caso contrário, use métodos aproximados somente para estimação de variâncias, tais como agregação de estratos ou similar (ver Cochran, 1977, seção 5A.12).

Soluções Possíveis

(5) Ganhos de Eficiência

Cochran(1977, p. 99) mostra que

$$V_{AES/Ney}(\bar{y}_{AES}) \leq V_{AES/Prop}(\bar{y}_{AES}) \leq V_{AAS}(\bar{y})$$

Os ganhos possíveis de precisão dependem da relação entre a(s) variável(is) de estratificação e as variáveis de pesquisa.

Em geral, ganhos são pequenos para amostras de pessoas e variáveis ligadas a atitudes, opiniões, comportamentos, etc..

Para pesquisas amostrais de estabelecimentos ou instituições, os ganhos podem ser muito grandes.

Definição dos Limites dos Estratos

Se uma variável auxiliar x estiver disponível, seus valores podem ser usados para formar estratos.

Como devemos formar os estratos?

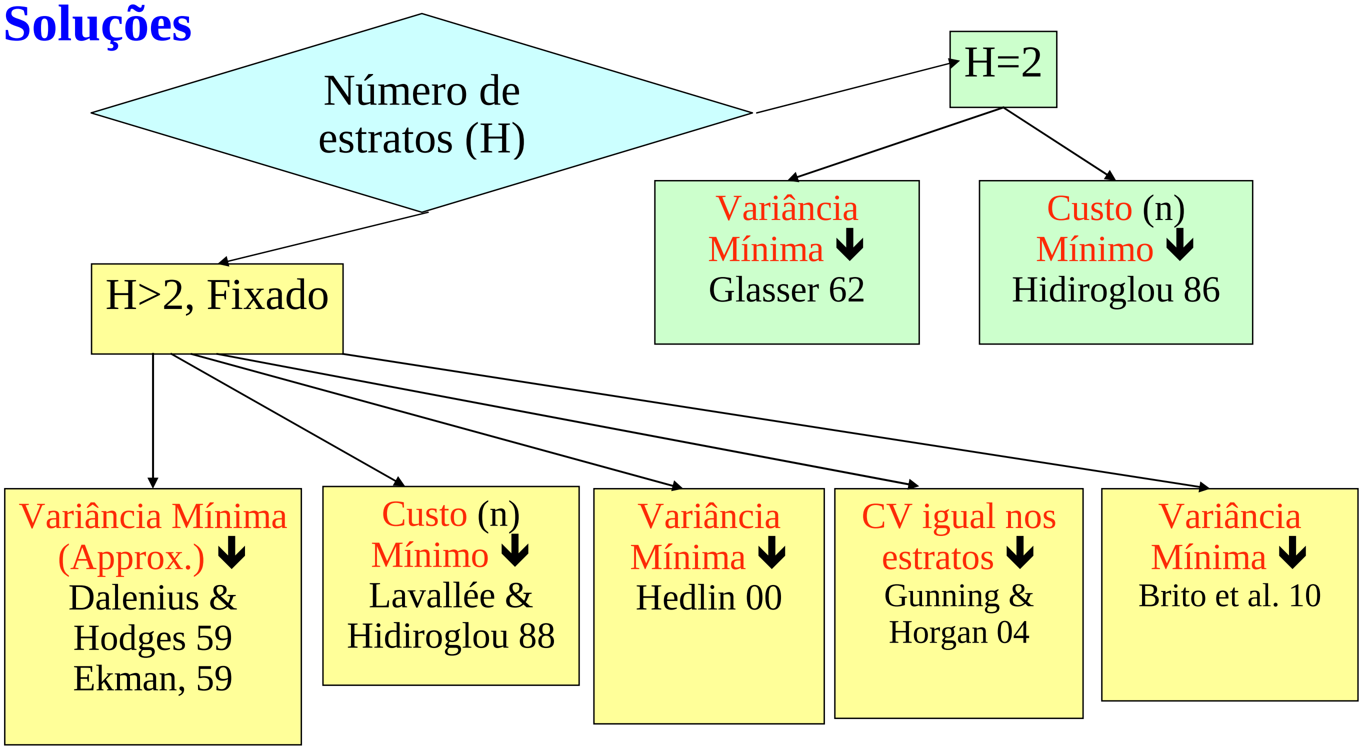
Quais os limites que devemos usar para delimitar os estratos?

Primeiro, escolha H, o número total de estratos.

Quanto maior for a correlação entre a variável de pesquisa y e a variável auxiliar x maior deve ser o número de estratos.

Evidências empíricas sugerem, entretanto, que $5 \leq H \leq 10$. Mais detalhes sobre esta escolha em seguida.

Soluções



Amostragem Estratificada Simples - Número de Estratos

- ✓ Para estimação por domínios, utilizar tantos estratos quantos sejam os domínios de interesse.
- ✓ Para estimação de total ou média global, Cochran(1977, seção 5A.8) recomenda usar até 6 (seis) estratos, se variável de estratificação for bem correlacionada com variáveis de interesse.

Justificativa

Hipóteses: N grande, n/N pequeno.

Modelo: $y_i = \alpha + \beta x_i + \varepsilon_i$ para $i \in U$.

Estratificação "ótima" em x.

Com alocação igual nos estratos ($n_h = n/H$), mostra-se que:

$$\begin{aligned} EPA(\bar{y}_{AES}) &= V_{AES}(\bar{y}_{AES})/V_{AAS}(\bar{y}) \\ &= \rho^2/H^2 + (1-\rho^2) \end{aligned}$$

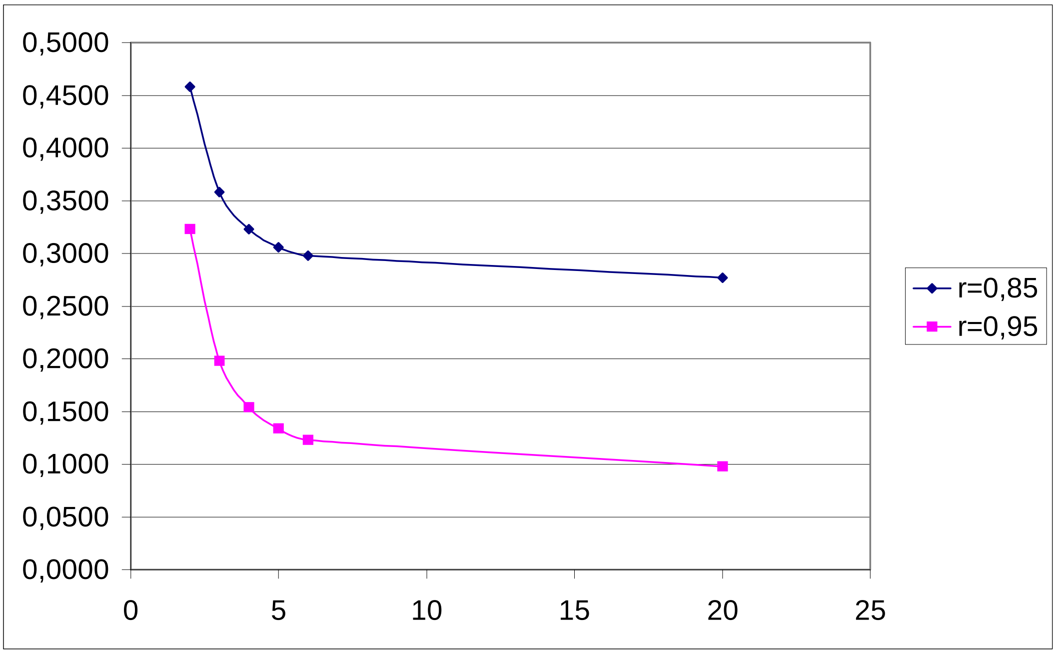
onde ρ é correlação entre x e y.

Valores de $V(\hat{Y}_{AES})/V(\hat{Y}_{AAS})$ para vários valores de H

H	2	3	4	5	6	∞
ρ=0,85	0,458	0,358	0,323	0,306	0,298	0,277
ρ=0,95	0,323	0,198	0,154	0,134	0,123	0,098

Conclusão: ganhos adicionais de eficiência com mais de seis estratos é modesto.

Figura 11.1: Ganhos de Precisão vs. Número de Estratos



Referências

Baillargeon, S. & Rivest, L. P. (2011). A General Algorithm for Univariate Stratification. Proceedings of the International Statistical Institute, Dublin.

Brito, J. A. M. (2005). Uma Formulação de Programação Inteira para o Problema de Alocação Ótima em Amostras Estratificadas. In: Anais do XXXVII Simpósio Brasileiro de Pesquisa Operacional - SOBRAPO, Gramado – RS, v. 1. p. 1851-1859.

Brito, J. A. M.; Maculan, N.; Lila, M. F. e Montenegro, F. T. (2010). An exact algorithm for the stratification problem with proportional allocation. *Optimization Letters*, v. 4, pp. 185 – 195.

Dalenius T. & Hodges Jr., Joseph L. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association*, Vol. 54, No. 285, pp. 88-101.

Glasser, G.J. (1962) On the complete coverage of large units in a statistical study. *International Statistical Review*, 30, 28-32.

Gunning, P. & Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries. *Survey Methodology*, 30, No. 2, 159-166.

Hedlin, D. (2000). A procedure for stratification based on an extended Ekman rule. *Journal of Official Statistics*, 16, 15-29.

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, n. 1, 27-31.

Lavallée, P. & Hidiroglou, M. A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

Rivest, L. P. (2002). A generalization of the Lavallée-Hidiroglou algorithm for stratification in business surveys. *Survey Methodology* 28, 191-198.