TÉCNICAS DE AMOSTRAGEM

Ralph dos Santos Silva

https://www.ralphsilva.org/amostragem.html

Instituto de Matemática
Centro de Ciências Matemáticas e da Natureza
Universidade Federal do Rio de Janeiro

Agradecimentos ao professor Vermelho do IBGE.
Este material é uma simples atualização das notas de aula dele.

Ementa

- Amostragem aleatória simples.
- Estimadores de razão e de regressão.
- Amostragem estratificada.
- Amostragem sistemática.
- Amostragem por conglomerados.
- Métodos da solução com probabilidades desiguais.

Objetivos

- Introduzir as ideias centrais da teoria de amostragem e justificar seu uso.
- Apresentar os principais tipos básicos de desenhos de amostra, utilizando algumas formulações matemáticas e demonstrações.

Referências

- Barnett, Vic (1984). *Elements of Sampling Theory*, Holder and Stoughton, Toronto.
- Bolfarine, H e Bussab, W. O. (2005). *Elementos de Amostragem*, Edgar Blucher, São Paulo.
- Cochran, W. G. (1977). *Sampling Techniques*, 3rd Edition, John Wiley & Sons, New York.
- Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, New York.
- Rao, C. R. (2009). *Handbook of Statistics 29A - Sample Surveys: Design, Methods and Applications*, Edited by D. Pfeffermann and C. R. Rao, Elsevier, Amsterdam.
- Rao, C. R. (2009). *Handbook of Statistics 29B - Sample Surveys: Inference and Analysis*, Edited by D. Pfeffermann and C. R. Rao, Elsevier, Amsterdam.
- Stopher, P. (2012). *Collecting, Managing, and Assessing Data Using Sample Surveys*, Cambridge University Press, New York.
- Thompson, S. K. (2012). *Sampling*, 3rd Edition, John Wiley & Sons, New Jersey.

História

Segundo Leslie Kish: “O artigo de Kiar (1895) pode bem servir como data de nascimento oficial da pesquisa por amostragem, apesar de pesquisas já terem sido feitas por Laplace e Lavoisier, entre outros”.

- Portanto, pode-se dizer que a Amostragem tem pouco mais de 100 anos.
- Porém, só passa a ser aceita plenamente, segundo Kish, após a segunda guerra mundial.
- Aliás, Kish diz mais quando indica em seu artigo que a pesquisa Estatística, apesar de ser uma ferramenta que já existia no século XIX, só ao longo do século XX passou a ser economicamente viável pela introdução da amostragem probabilística.

O que é uma pesquisa?

- Coleta de informações sobre as características de interesse de unidades de uma população, usando conceitos, métodos e procedimentos bem definidos.
- Compilação dessas informações em uma forma resumida útil.

Definição de amostragem por Steven K. Thopmson

"Amostragem consiste em selecionar parte de uma população para observar, de modo que seja possível estimar alguma coisa sobre toda a população."

Pesquisa amostral

Objetivo: conhecer características sobre a população, pesquisando (estudando) a amostra.

Conceitos básicos

- Unidade (elemento; unidade elementar) é um único indivíduo ou objeto a ser medido ou observado na pesquisa.
- População (universo) é o conjunto de todas as unidades para o qual queremos obter informações ou fazer inferências.
- Amostra é o subconjunto de unidades da população que selecionamos para medir ou observar.
- Cadastro (cadastro de seleção) é a lista de unidades da população de onde a amostra é selecionada.

Pesquisa

- Pode-se examinar:
  - todas as unidades da população $\rightarrow$ CENSO ou PESQUISA EXAUSTIVA; e
  - um subconjunto selecionado de unidades da população $\rightarrow$ PESQUISA POR AMOSTRA.
- A seleção pode ser:
  - Probabilística: É A QUE VAMOS ESTUDAR!;
  - Quase probabilística: POR COTAS; e
  - Não probabilística: ESPECIALISTAS.

Pesquisa por amostra

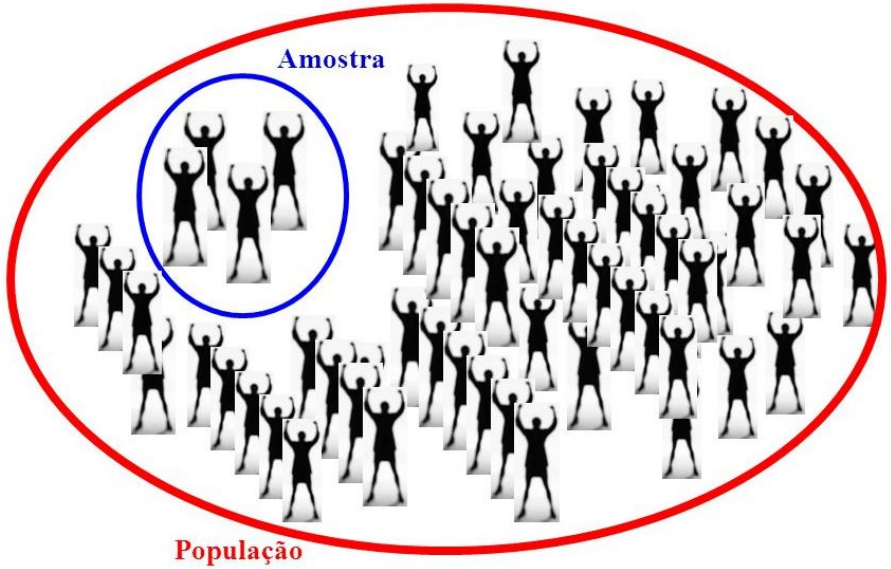

Figura 1: População e amostra.

Amostragem Probabilística

- Cada unidade da população tem probabilidade positiva de ser incluída na amostra, e esta probabilidade pode ser calculada (conhecida).
- Amostra extraída por algum método de seleção aleatória (ao acaso).
- Probabilidades de seleção incorporadas na criação de estimativas para população.

Algumas características desejáveis de uma amostra

- Capacidade de generalizar estimativas da amostra para toda a população.
- "Imparcialidade."
- Menor erro amostral possível, dado o custo, tempo e restrições operacionais (eficiência).
- Capacidade de medir a precisão das estimativas. (Estimativa: valor aproximado para o valor verdadeiro de um parâmetro da população).

Passos fundamentais

- Definição de objetivos, conceitos e recursos.
- Obtenção e avaliação do cadastro.
- Planejamento, seleção e controle da amostra.
- Estimação das quantidades de interesse.
- Avaliação da qualidade (precisão) das estimativas.

Objetivos da pesquisa

- Exemplo 1: PNAD - Pesquisa Nacional por Amostra de Domicílios.

  Objetivos: A PNAD "tem como finalidade a produção de informações básicas para o estudo do desenvolvimento sócio-econômico do país." (IBGE, PNAD 1999, volume 21, p. XIII)

- Exemplo 2: POF - Pesquisa de Orçamentos Familiares.

  Objetivos: A POF "tem como finalidade estimar o rendimento mensal bruto total dos chefes de domicílios em cada área de pesquisa com erro amostral máximo admissível inferior a 5%".

  São 11 áreas cobertas pela pesquisa: 9 regiões metropolitanas, DF e o município de Goiânia.

- Exemplo 3: LFS - *Canadian Labour Force*.

  Objetivo 1: Estimar mensalmente o número de desempregados:
  - para o Canadá com CV inferior a 2%; e
  - por província com CV entre 4% e 7%.

  Objetivo 2: Estimar trimestralmente o número de desempregados:
  - por área econômica, com CV $< 15\%$.

Todo plano amostral deve especificar:

- População alvo.
- População de pesquisa e cadastros.
- Unidade(s) de amostragem.
- Unidades de informação (de pesquisa).
- Método(s) para seleção da amostra.
- Tamanho da amostra.
- Aspecto longitudinal (pesquisas repetidas).

Definições

- População alvo: população que se desejaria atingir com a pesquisa, para a qual se gostaria de obter as informações.
- População de pesquisa ou amostrada: população a ser realmente coberta pela pesquisa.

Exemplo: PNAD 2003

- População alvo: pessoas residentes no Brasil, em uma data de referência especificada.
- População de pesquisa: pessoas residentes no Brasil, em uma data de referência especificada, menos os habitantes das áreas rurais de RO, AC, AM, RR, AP e PA.

População de pesquisa

- Unidades a serem pesquisadas;
- Características definidoras das unidades;
- Localização das unidades;
- Período de referência considerado; e
- Vinculação a cadastros.

Exemplo

Unidades	Características definidoras	Localização	Período
Pessoas	Que habitam domicílios particulares permanentes	Em Macaé	Durante a semana da pesquisa
Empresas de comércio varejista	Classificadas como supermercados	Em Recife	Em 1996
Pessoas	Maiores de 5 anos de idade	Que visitaram o museu nacional	Entre 01/07 e 30/09/1996
Alunos	Do curso de mestrado da UFRJ	No Rio de Janeiro	$1^{\underline{o}}$ trimestre de 2004
Estabelecimentos agropecuários	Produtores de café	No Paraná	No ano de 1998

Conceitos a definir em uma pesquisa por amostragem

- Unidade de referência: Unidade de observação ou sobre a qual são obtidas informações de interesse.
- Unidade informante: Unidade que fornece a informação.
- Unidade de análise: Unidade à qual a inferência é dirigida.
- Unidade de amostragem: Unidade que será selecionada para amostra.
- Domínio de análise ou de interesse: Grupo de unidades de análise agregadas para fins de tabulação, inferência e análise.

Exemplo 1: PNAD - Pesquisa Nacional por Amostra de Domicílios

- Unidade(s) de referência: Pessoas, moradora no domicílio.
- Unidade informante: Pessoa adulta, moradora do domicílio.
- Unidade(s) de análise: Pessoas, famílias, domicílios.
- Unidade(s) de amostragem: Município, setor, domicílio.
- Domínios de análise: 27 Estados, 9 Regiões Metropolitanas, País.

Exemplo 2: POF - Pesquisa de Orçamentos Familiares

- Objetivo: obter informações sobre a renda e despesa familiar.
- Unidades de referência: pessoas dentro das famílias, famílias.
- Unidade informante: pessoa de referência da família.
- Unidade de análise: família.
- Unidade amostral: setor, domicílio.

Cadastro (sistema de referência)

- Fornece meios de acesso à população de pesquisa.
- Constitui uma lista identificadora dos elementos que formam a população.
- Contém informações auxiliares para:
  - Planejar e selecionar a amostra; e
  - Utilizar na estimação dos parâmetros.

Tipos de cadastros

- Cadastro de unidades individuais:
  - Lista física ou conceitual das unidades individuais da população.
- Cadastro de áreas:
  - Lista de áreas geográficas.
- Cadastros múltiplos:
  - Dois ou mais cadastros, do mesmo tipo ou não.

Cadastro usado afeta de forma direta e irreversível:

- A definição da população de pesquisa.
- O método de coleta dos dados.
- O método de seleção da amostra.
- A qualidade dos resultados.
- O custo da pesquisa.

Um bom cadastro deve:

- Conter informação suficiente sobre cada unidade da população para identificá-la com certeza (IDENTIFICAÇÃO);
- Conter informação suficiente sobre cada unidade da população para permitir localizá-la (LOCALIZAÇÃO);
- Ser completo e sem redundâncias (duplicatas) ou omissões, preciso e atual (QUALIDADE);
- Estar disponível em um lugar central, com acesso fácil e rápido (DISPONIBILIDADE);
- Estar arranjado em uma forma adequada à amostragem; e
- Conter informação auxiliar sobre cada unidade, a fim de permitir elaborar um planejamento amostral o mais eficiente possível.

Exemplo de Cadastro no IBGE

- Cadastro de Empresas (CEMPRE).
  - Dados de empresas e unidades locais.
  - Alimentado por pesquisas próprias e registros administrativos (RAIS - Ministério do Trabalho).
  - Usado como principal cadastro para pesquisas econômicas.
  - Atualizado anualmente.
- Base Operacional Geográfica - BOG
  - Lista de unidades geográfico-estatísticas: setores, subdistritos, distritos, municípios, unidades da federação.
  - Usado no Censo Demográfico e pesquisas domiciliares (PNAD, PME,...)
  - Usado no Censo Agropecuário.

Exemplo de Base Operacional Geográfica

Figura 2: Uma foto do bairro de Copacabana na cidade do Rio de Janeiro.

Exemplo de Base Operacional Geográfica

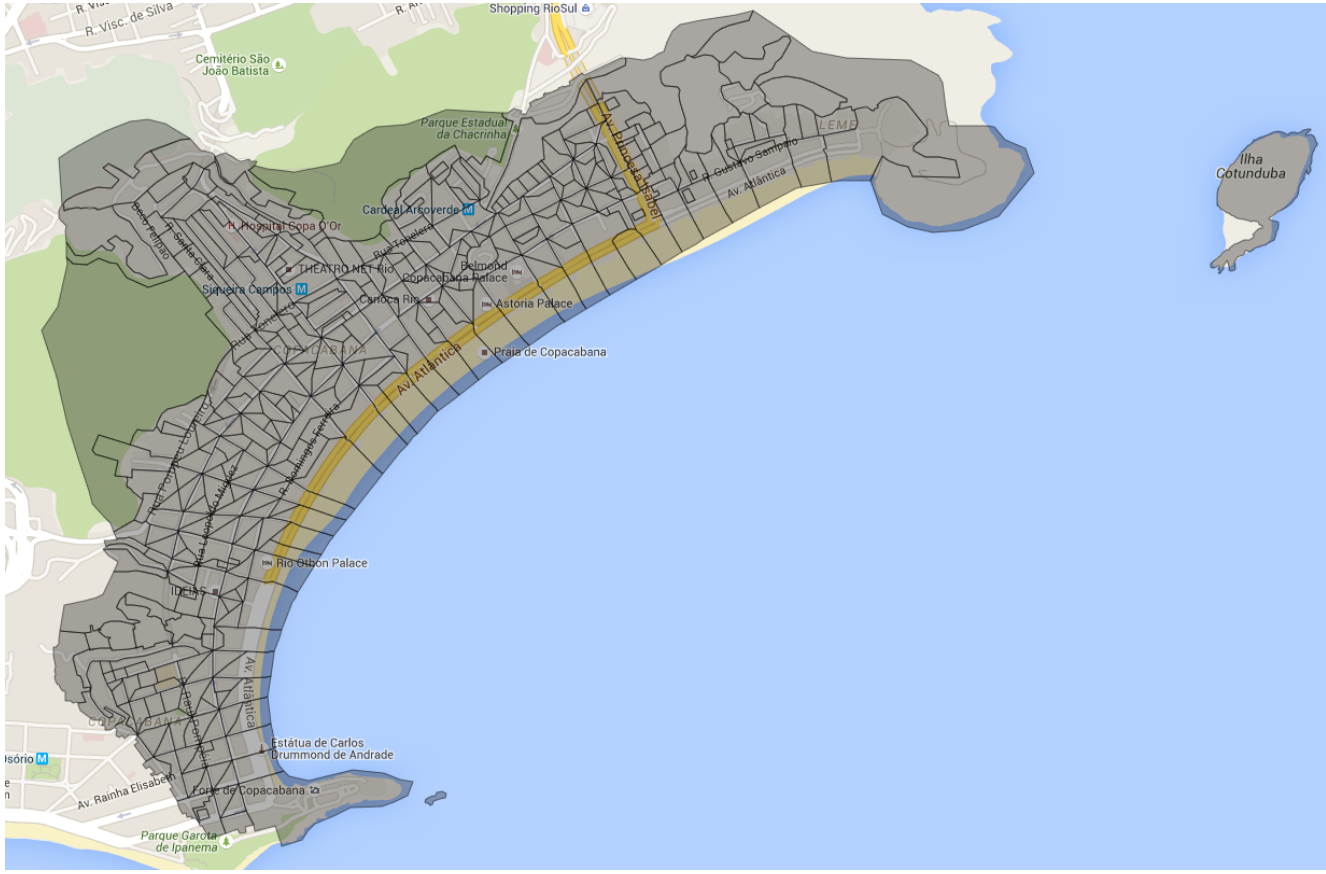

Figura 3: O cadastro de setores do IBGE para o bairro de Copacabana na cidade do Rio de Janeiro referente ao Censo de 2010.

Ver http://www.censo2010.ibge.gov.br/sinopseporsetores/

Tabela 1: Exemplo de dados do cadastro.

SETOR	Pessoa	Domic.	Rend.	PRes11	PResp15
41	348	136	1.700	36,76%	26,47%
42	584	200	1.120	25,50%	18,00%
43	548	190	1.096	21,58%	12,63%
44	302	96	1.242	27,08%	21,88%
45	280	101	1.004	17,82%	8,91%
46	370	119	614	14,29%	9,24%
47	699	215	817	14,88%	11,16%
48	455	147	1.463	33,33%	24,49%
49	584	201	1.545	35,82%	33,33%
50	788	270	1.632	30,37%	25,19%
51	624	211	1.813	31,28%	22,27%
52	349	131	1.097	25,95%	19,85%
53	599	193	926	17,62%	10,88%
54	413	123	952	7,32%	2,44%
55	510	169	1.087	28,40%	24,26%
59	662	220	1.171	26,36%	20,00%
60	838	267	847	13,86%	10,86%
61	856	292	1.073	14,04%	6,16%
62	796	255	654	11,37%	4,71%
63	324	100	556	7,00%	4,00%
64	395	149	626	10,74%	7,38%
65	442	203	790	19,21%	14,78%
69	551	175	884	18,86%	10,86%

Defeitos e soluções para um cadastro

Defeitos:

- Falta de unidades (omissão ou falha de cobertura).
- Presença de unidades estranhas à população alvo (fora do âmbito).
- Duplicação de unidades.
- Informações desatualizadas.
- Informações faltando ou incorretas.

Soluções possíveis:

- Descarte o cadastro, e crie ou use outro.
- Ajuste e corrija o cadastro mediante atualização ou ligação com outros.
- Use o cadastro existente e adote precauções contra seus defeitos.
- Use cadastros múltiplos.

Definições e Notação

Regra geral da nossa notação:

- Universo ou população:

  LETRAS MAIÚSCULAS.

- Amostra:

  letras minúsculas.

Exemplos:

$$\overline{Y} \quad \rightarrow \quad \text{média populacional;}$$
$$\overline{y} \quad \rightarrow \quad \text{média amostral.}$$

Definições e Notação: População

Seja uma população composta de N unidades elementares U_i, $i = 1, \ldots, N$.

O conjunto

$$P_N = \boldsymbol{U} = \{U_1, U_2, \ldots, U_N\}$$

define os rótulos que identificam cada unidade da população.

Temos que

$$\begin{aligned} N &\rightarrow \text{tamanho da população de pesquisa;} \\ U_i &\rightarrow \text{rótulo para uma unidade genérica } i\text{;} \\ y &\rightarrow \text{variável de pesquisa ou de interesse; e} \\ Y_i &\rightarrow \text{valor da variável } y \text{ para unidade } i\text{.} \end{aligned}$$

Alguns parâmetros de interesse

- Total populacional: $Y = \sum_{i=1}^{N} Y_i$.
- Média populacional: $\overline{Y} = \mu = \frac{1}{N} \sum_{i=1}^{N} Y_i$.
- Variância populacional: $S^2 = S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2$.
- Razão populacional: $R = R_{xy} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{\overline{Y}}{\overline{X}}$.

Alguns parâmetros de interesse

- Covariância populacional:

$$S_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})(X_i - \overline{X}).$$

- Coeficiente de correlação (de Pearson) populacional:

$$\rho = \rho_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{N} (Y_i - \overline{Y})(X_i - \overline{X})}{\sqrt{\sum_{i=1}^{N} (Y_i - \overline{Y})^2 \sum_{i=1}^{N} (X_i - \overline{X})^2}}.$$

Definições e Notação: Amostra

Qualquer subconjunto $\boldsymbol{s} \subset \boldsymbol{U}$ não vazio, selecionado para ser observado e utilizado para estimar parâmetros de $\boldsymbol{U}$:

$$\boldsymbol{s} = \{u_1, u_2, \ldots, u_n\} \subset \boldsymbol{U}.$$

Temos que

$n \rightarrow$ é o tamanho da amostra

$u_i \rightarrow$ é a unidade i da amostra, $i = 1, 2, \ldots, n$; e

$y_i \rightarrow$ é o valor da variável de interesse y para a unidade i da amostra.

Estatísticas (funções da amostra)

Seja $\{y_1, y_2, \ldots, y_n\}$ o conjunto dos dados amostrais.

- Total amostral: $t = \sum_{i=1}^{n} y_i$.
- Média amostral: $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.
- Variância amostral: $s^2 = s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$.
- Razão amostral: $r = r_{xy} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\overline{y}}{\overline{x}}$.

Estatísticas (funções da amostra)

- Covariância amostral:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}).$$

- Coeficiente de correlação (de Pearson) amostral:

$$\widehat{\rho} = \widehat{\rho}_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Exemplo de População

Seja uma população P_3, tal que uma variável de interesse y apresente o seguinte conjunto de valores: $Y_1 = 4$, $Y_2 = 5$ e $Y_3 = 3$.

$$
\begin{aligned}
Y &= \sum_{i=1}^{3} Y_i = 4 + 5 + 3 = 12 \\
\overline{Y} &= \frac{1}{3} \sum_{i=1}^{3} Y_i = \frac{4 + 5 + 3}{3} = 4 \\
S_y^2 &= \frac{1}{3 - 1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 = 1, \quad \left[\sigma^2 = \frac{2}{3} \right].
\end{aligned}
$$

Observação: na prática nunca faremos uma amostra de uma população deste tamanho.

Exemplo 1

Vamos selecionar todas as possíveis amostras de tamanho 2 dessa população, permitindo repetições.

Método:

- Selecionar uma unidade com equiprobabilidade.
- Anotar o valor de y e devolver a unidade à população.
- Selecionar a segunda unidade com equiprobabilidade.

	Amostral possível	y_1	y_2	t	$\overline{y}$	s^2
1	$\{u_1, u_2\} = \{U_1, U_1\}$	4	4	8	4	0
2	$\{u_1, u_2\} = \{U_1, U_2\}$	4	5	9	4,5	0,5
3	$\{u_1, u_2\} = \{U_1, U_3\}$	4	3	7	3,5	0,5
4	$\{u_1, u_2\} = \{U_2, U_1\}$	5	4	9	4,5	0,5
5	$\{u_1, u_2\} = \{U_2, U_2\}$	5	5	10	5	0
6	$\{u_1, u_2\} = \{U_2, U_3\}$	5	3	8	4	2
7	$\{u_1, u_2\} = \{U_3, U_1\}$	3	4	7	3,5	0,5
8	$\{u_1, u_2\} = \{U_3, U_2\}$	3	5	8	4	2
9	$\{u_1, u_2\} = \{U_3, U_3\}$	3	3	6	3	0

Exemplo 1 (continuação)

- Veja que as estatísticas (das amostras) são diferentes para cada amostra possível.
- Portanto, são variáveis aleatórias.
- Suponha que todas as amostras possíveis têm a mesma probabilidade de serem selecionadas, ou seja, $\Pr(\boldsymbol{s}_j) = 1/9$ para $j = 1, 2, \ldots, 9$.
- Vamos calcular o valor esperado do total, da média e da variância da amostra.
- O que concluímos?
  - t tendencioso para Y;
  - $\overline{y}$ não tendencioso para $\overline{Y}$; e
  - s^2 tendencioso para S_y^2.

Exemplo 2

Vamos repetir o Exemplo 1, sem admitir repetição.

Método:

- Selecionar uma unidade com equiprobabilidade.
- Selecionar, também com equiprobabilidade, a segunda unidade entre as restantes.

	Amostral possível	y_1	y_2	t	$\overline{y}$	s^2
1	$\{u_1, u_2\} = \{U_1, U_2\}$	4	5	9	4,5	0,5
2	$\{u_1, u_2\} = \{U_1, U_3\}$	4	3	7	3,5	0,5
3	$\{u_1, u_2\} = \{U_2, U_3\}$	5	3	8	4	2

- Agora temos 3 amostras possíveis com $\Pr(\boldsymbol{s}_j) = 1/3$, para $j = 1, 2, 3$.
- Vamos calcular os mesmos valores esperados do Exemplo 1.
- Quais as conclusões?
  - t tendencioso para Y;
  - $\overline{y}$ não tendencioso para $\overline{Y}$; e
  - s^2 não tendencioso para S_y^2.

Definições

- Parâmetro: é uma função dos valores da variável de interesse na população.
- Estatística: é uma função dos valores da variável de interesse na amostra.
  - O valor da estatística varia conforme a amostra selecionada, portanto é uma variável aleatória que possui um valor esperado e uma variância.
- Estimador: é uma estatística adequada para estimar o valor de um parâmetro a partir dos dados amostrais.

Estimador e Estimação

- Portanto: estimador é uma função dos dados amostrais que serve para estimar um parâmetro.
- Precisão do estimador: é dada pela variância do estimador.
- Estimativa: é o valor resultante da aplicação da função estimador aos dados da amostra.
- Todo estimador que puder ser escrito como uma combinação linear dos valores amostrais será um estimador linear.

Estimador e Estimação

Vamos considerar o problema de estimar o valor do total populacional a partir de uma amostra, ou seja:

$$\text{uma } v(\{y_1, y_2, \ldots, y_n\}) \text{ que aproxime } Y = \sum_{i=1}^{N} Y_i.$$

Será que existe um estimador linear com essa capacidade?

Ou seja, existe $\widehat{Y} = \sum_{i=1}^{n} \omega_i y_i \simeq Y = \sum_{i=1}^{N} Y_i$?

Exemplo 2 (continuação)

- Vimos que $Y = 12$.
- Mas $1,5 \times \mathrm{E}_p(t) = 12 = Y$, ou ainda $\mathrm{E}_p(1,5t) = 12 = Y$.
- Então a estatística t não é um bom estimador de Y, mas a estatística $1,5t$ pode ser!
- Se uma estatística é uma variável aleatória, ela tem uma distribuição de probabilidades (que dá a probabilidade com que a variável aleatória assume cada um dos seus valores possíveis).

Distribuição amostral

- Distribuição amostral é a distribuição de probabilidades de uma estatística.
- Distribuição amostral de t:

t	7	8	9
Pr(t)	1/3	1/3	1/3

- Distribuição amostral de $1,5t$:

1,5t	10,5	12	13,5
Pr(t)	1/3	1/3	1/3

Probabilidade de inclusão

- Define-se como probabilidade de inclusão da unidade U_i na amostra a seguinte quantidade:

$$\pi_i = \sum_{s \supset U_i} \Pr(\boldsymbol{s}).$$

- Ou seja, a probabilidade da unidade U_i da população ser incluída na amostra, é igual a probabilidade de uma das possíveis amostras que a contenha ser a amostra selecionada.

Variável indicadora de presença

- Seja a variável aleatória δ_i uma indicadora da presença da unidade U_i na amostra:

$$\delta_i = \begin{cases} 1, & \text{se a unidade } U_i \text{ percente à amostra;} \\ 0, & \text{caso contrário.} \end{cases}$$

- Portanto, essa será uma variável aleatória de Bernouli com

$$\Pr(\delta_i = 1) = \Pr(U_i \subset \boldsymbol{s}) = \pi_i.$$

- Então, temos que:

$$\mathrm{E}_p(\delta_i) = \pi_i, \qquad \mathrm{Var}_p(\delta_i) = \pi_i(1 - \pi_i).$$

- Também temos que

$$\Pr(\delta_i\delta_j = 1) = \Pr(U_i \subset \boldsymbol{s};\ U_j \subset \boldsymbol{s}) = \sum_{\boldsymbol{s} \supset U_i; U_j} \Pr(\boldsymbol{s}) = \pi_{ij},$$

  tal que

$$\mathrm{Cov}(\delta_i, \delta_j) = \pi_{ij} - \pi_i\pi_j.$$

Estimador linear (continuação)

- O estimador linear do total pode ser escrito como:

$$\widehat{Y} = \sum_{i=1}^{n} \omega_i y_i = \sum_{j=1}^{N} \omega_j Y_j \delta_j.$$

- Seu valor esperado será:

$$\mathrm{E}_p(\widehat{Y}) = \mathrm{E}_p\left(\sum_{i=1}^{n} \omega_i y_i\right) = \sum_{j=1}^{N} \omega_j Y_j \mathrm{E}_p(\delta_j) = \sum_{j=1}^{N} \omega_j \pi_j Y_j.$$

- Para que esse estimador seja não tendencioso, basta que:

$$\omega_j \pi_j = 1 \quad \Rightarrow \quad \omega_j = \frac{1}{\pi_j}.$$

Definição

- O peso amostral da unidade u_i de uma amostra probabilística é igual ao inverso de sua probabilidade de inclusão nessa amostra:

$$\omega_i = \frac{1}{\pi_i}, \quad i = 1, 2, \ldots, n.$$

- Interpretação intuitiva: o peso amostral é o número de unidades da população "representadas" pela unidade u_i da amostra.

Estimador de Horvitz-Thompson

- Então, um estimador não tendencioso para o total de uma variável de interesse será dado por:

$$\widehat{Y}_{HT} = \sum_{i=1}^{n} \omega_i y_i = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i.$$

- O estimador do total que acabamos de definir, como soma ponderada dos valores amostrais, onde o peso de cada unidade amostral é o inverso de sua probabilidade de inclusão, é chamado de estimador de Horvitz-Thompson, que foram seus formuladores.
- Esse estimador está definido para qualquer plano amostral onde todas as unidades da população tenham probabilidades positivas de serem selecionadas.

Estimador de Horvitz-Thompson (continuação)

- Variância do estimador de Horvitz-Thompson:

$$\text{Var}_p(\widehat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j.$$

- Um estimador não tendencioso para a variância do estimador de Horvitz-Thompson é:

$$\widehat{\text{Var}}_p(\widehat{Y}_{HT}) = v_p(\widehat{Y}_{HT}) = \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} y_i^2 + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j.$$

- Note a divisão por quantidades extras π_i e π_{ij} no estimador da variância do estimador de Horvitz-Thompson.

Exercícios:

1. Dado o estimador $\widehat{Y}_{HT}$, encontre a expressão para $\text{Var}_p(\widehat{Y}_{HT})$.
2. Mostre que $\widehat{\text{Var}}_p(\widehat{Y}_{HT})$ é um estimador não tendencioso de $\text{Var}_p(\widehat{Y}_{HT})$.

Amostra Aleatória Simples Sem Reposição (AAS)

- Método básico de muitos planos amostrais.
- O algoritmo base é:

  Para uma AAS de tamanho n:

  1. Selecione uma unidade da população com equiprobabilidade.
  2. Retire a unidade selecionada da população.
  3. Repita os Passos 1 e 2 até ter selecionado n unidades.
- Esse esquema garante que todas as amostras possíveis de tamanho n têm a mesma probabilidade de serem escolhidas.
- Garante que todas as unidades têm a mesma probabilidade de seleção e de inclusão.

Amostra Aleatória Simples Sem Reposição (AAS)

- A probabilidade de seleção da unidade U_i em qualquer uma das n seleções é $1/N$:

$$\begin{aligned}
\Pr(U_i \text{ ser selecionada na } 1^{\underline{a}}) &= \frac{1}{N} \\
\Pr(U_i \text{ ser selecionada na } 2^{\underline{a}}) &= \left[1 - \frac{1}{N}\right] \frac{1}{N-1} = \frac{1}{N} \\
\Pr(U_i \text{ ser selecionada na } 3^{\underline{a}}) &= \left[1 - \frac{1}{N}\right] \left[1 - \frac{1}{N-1}\right] \frac{1}{N-2} = \frac{1}{N} \\
&\vdots
\end{aligned}$$

- A probabilidade de inclusão da unidade U_i na amostra, dessa forma, será igual a probabilidade dela ser selecionada em pelo menos uma das n seleções, ou seja:

$$\pi_i = \Pr(U_i \subset \boldsymbol{s}) = \sum_{i=1}^{n} \frac{1}{N} = \frac{n}{N}.$$

Amostra Aleatória Simples Sem Reposição (AAS)

- A probabilidade de inclusão das unidades U_i e U_j na amostra, será igual a probabilidade de U_i ser selecionada em pelo menos uma das n seleções, e U_j ser selecionada em uma das $n-1$ outras seleções, ou seja:

$$\pi_{ij} = \Pr(U_i \subset \boldsymbol{s};\ U_j \subset \boldsymbol{s}) = \sum_{i=1}^{n} \frac{1}{N} \sum_{j \neq i, j=1}^{n} \frac{1}{N-1} = \frac{n}{N} \times \frac{n-1}{N-1}.$$

Amostra Aleatória Simples Sem Reposição (AAS)

- Para o estimador do total populacional, empregaremos o estimador de Horvitz-Thompson:

$$\widehat{Y}_{HT} = \widehat{Y}_{AAS} = \sum_{i=1}^{n} \omega_i y_i = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} \frac{N y_i}{n} = \frac{N}{n} \sum_{i=1}^{n} y_i = N\overline{y},$$

  sendo ω_i o peso amostral dado pelo inverso da probabilidade de inclusão.

- A variância do estimador do total é dada por

$$\text{Var}(\widehat{Y}_{AAS}) = N^2(1-f)\frac{S^2}{n}, \quad \text{sendo} \quad f = \frac{n}{N}.$$

- Um estimador não tendencioso de S^2 é dado pela variância amostral:

$$\widehat{S}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right].$$

- Consequentemente,

$$v(\widehat{Y}_{AAS}) = \widehat{\text{Var}}(\widehat{Y}_{AAS}) = N^2(1-f)\frac{s^2}{n}.$$

Amostra Aleatória Simples Sem Reposição (AAS)

- Definimos a fração amostral como $f = \frac{n}{N}$.
- O termo $(1 - f)$ é chamado de fator de correção para populações finitas.
- Note que se $N \to \infty$, o fator de correção para populações finitas será aproximadamente 1.
- Estimador da média populacional:

$$\widehat{\overline{Y}}_{AAS} = \overline{y} = \frac{\widehat{Y}_{AAS}}{N} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- Variância do estimador da média:

$$\text{Var}(\overline{y}) = (1 - f)\frac{S^2}{n} = \frac{N - n}{N} \times \frac{S^2}{n}.$$

- Estimador da variância do estimador da média:

$$v(\overline{y}) = (1 - f)\frac{s^2}{n}.$$

Exercício (Cochran)

Foram coletadas assinaturas para um abaixo assinado em 676 folhas. Cada folha tinha espaço para 42 assinaturas, mas em muitas das folhas foi coletado um número menor de assinaturas. Uma amostra aleatória simples de 50 folhas foi selecionada, e os resultados estão na tabela abaixo:

Tabela 2: Dados sobre assinaturas.

y_i	42	41	36	32	29	27	23	19	16	15
f_i	23	4	1	1	1	2	1	1	2	2
y_i	14	11	19	9	7	6	5	4	3	Total
f_i	1	1	1	1	1	3	2	1	1	50

a) Estimar o total de assinaturas do abaixo assinado e a variância do estimador.

b) Estimar o número médio de assinaturas por folha e a variância do estimador.

Seleção de uma amostral aletória simples

- Como selecionar uma amostral aletória simples de um cadastro?
- Algoritmo natural é pouco eficiente do ponto de vista computacional.
- Vamos dar 2 exemplos de Algoritmos: Hàjek, e Fan, Muller e Rezucha.

Algoritmo de Hàjek

- Selecionar um número aleatório da distribuição $\mathcal{U}(0,1)$, para cada unidade da população P_N.
- Ordenar a população segundo os valores dos aleatórios gerados.
- Selecionar as n primeiras unidades da população nessa nova ordem.

Qualquer pacote já tem pelo menos uma rotina de ordenação e geração de números pseudo-aleatórios.

Algoritmo de Fan, Muller e Rezucha

- Seja uma população com N unidades.
- Deseja-se uma amostra de tamanho n desta população.
  1. $i \leftarrow 0$
  2. $i \leftarrow i + 1$
  3. Para a unidade U_i gere um número aleatório $A_i \sim \mathcal{U}(0, 1)$.
  4. Se $A_i < \dfrac{n}{N}$, faça
     - 4.1 Inclua U_i na amostra.
     - 4.2 Faça $n \leftarrow n - 1$ e $N \leftarrow N - 1$.

     Caso contrário, se $A_i \geqslant \dfrac{n}{N}$, faça
     - 4.1 Faça $N \leftarrow N - 1$.
  5. Se $n = 0$ ou $N = 0$ pare. Caso contrário, retorne ao Passo 2.
- Processamento sequencial.
- Pode não precisar percorrer todo o cadastro.

Exercício

Suponha que exista um cadastro de 1.000.000 de unidades. Deseja-se selecionar uma amostra aleatória simples sem reposição (AAS) com 1.500 unidades desta população. Faça o que é pedido abaixo no `R` e utilize no início das rotinas `set.seed(12345)`.

1. Utilize o Algoritmo de Hàjek para selecionar esta amostra.
2. Utilize o Algoritmo de Fan, Muller e Rezucha para selecionar esta amostra.
3. Utilize a função `sample` ou `sample.int` para selecionar a amostra.
4. Compare os tempos de execução de cada algoritmo.

Amostra Aleatória Simples Com Reposição (AASc)

- Algoritmo natural da AASc:
  1. Selecione uma unidade da população com equiprobabilidade;
  2. Reponha a unidade selecionada na população;
  3. Repita os Passos 1 e 2 até ter feito n seleções.
- Para uma variável de interesse y, temos que os valores amostrais $y_1, y_2, \ldots, y_n$ serão:
  - Independentes;
  - Identicamente distribuídos; e
  - $\Pr(y_i = Y_j) = \dfrac{1}{N}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, N.$
- Distribuição amostral de y_i:

y_i	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_N
$\Pr(y_i = Y_j)$	$\frac{1}{N}$	$\frac{1}{N}$	$\cdots$	$\frac{1}{N}$	$\cdots$	$\frac{1}{N}$

- Temos que

$$\mathrm{E}_{AASc}(y_i) = \overline{Y} \quad \text{e} \quad \mathrm{Var}_{AASc}(y_i) = \sigma^2.$$

- Conclusão: y_i é um estimador não tendencioso para a média da população, mas a variância é “grande”.

Amostra Aleatória Simples Com Reposição (AASc)

- Vamos analisar a média amostral:

$$\mathrm{E}_{AASc}(\overline{y}) = \overline{Y} \quad \text{e} \quad \mathrm{Var}_{AASc}(\overline{y}) = \frac{\sigma^2}{n} = \frac{N-1}{N} \times \frac{S^2}{n}.$$

- Então, a média amostral também é estimador não tendencioso para a média da população.
- Vantagem: a variância é menor!
- Comparação com a AAS:

$$\mathrm{Var}_{AAS}(\overline{y}) \leqslant \mathrm{Var}_{AASc}(\overline{y}).$$

- Estimador da variância da média amostral

$$\widehat{\mathrm{Var}}_{AASc}(\overline{y}) = v_{AASc}(\overline{y}) = \frac{s^2}{n}.$$

- Na amostra aleatória simples com reposição, temos que

$$\mathrm{E}_{AASc}(s^2) = \sigma^2 = \frac{N-1}{N} S^2.$$

- Para o total populacional:

$$\widehat{Y}_{AASc} = N\overline{y}, \quad \mathrm{Var}(\widehat{Y}_{AASc}) = N^2 \frac{\sigma^2}{n}, \quad \text{e} \quad v(\widehat{Y}_{AASc}) = N^2 \frac{s^2}{n}.$$

Exercício

Definimos

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2.$$

Para uma amostra aleatória simples com reposição, mostre que:

1. $\mathrm{E}_{AASc}(y_i) = \overline{Y}$ e $\mathrm{Var}_{AASc}(y_i) = \sigma^2$.
2. $\mathrm{E}_{AASc}(\overline{y}) = \overline{Y}$ e $\mathrm{Var}_{AASc}(\overline{y}) = \frac{\sigma^2}{n} = \frac{N-1}{N} \times \frac{S^2}{n}$.
3. $\mathrm{Var}_{AAS}(\overline{y}) \leqslant \mathrm{Var}_{AASc}(\overline{y})$.
4. $\mathrm{E}_{AASc}(s^2) = \sigma^2 = \frac{N-1}{N}S^2$.

Amostra Aleatória Simples Sem Reposição (AAS)

- Teorema Central do Limite (TCL)
  - Se $y_1, y_2, \ldots, y_n$ são variáveis aleatórias independentes e identicamente distribuídas, vindos de uma amostra aleatória simples com reposição (AASc), tem-se que:

$$\frac{\sqrt{n}(\overline{y} - \overline{Y})}{\sigma} \to \mathcal{N}(0,1).$$

- O TCL pode ser aplicado para uma amostra aleatória simples sem reposição (AAS), sob certas condições:
  - N e n "grandes"; e
  - $f = \dfrac{n}{N}$ "pequena".
- Ou seja, sob AAS temos: $\dfrac{\overline{y} - \overline{Y}}{\sqrt{(1-f)\dfrac{S^2}{n}}} \to \mathcal{N}(0,1).$
- Veja que sob as condições anteriores, $(1-f) \to 1$ e $S^2 \to \sigma^2$.
- E a probabilidade de uma unidade ser selecionada mais de uma vez é pequena.
- Portanto, AAS e AASc se equivalem.

Amostra Aleatória Simples Sem Reposição (AAS)

- Vamos ver um exemplo de simulação.
- Seja uma P_{5000}, donde seleciona-se 1.000 amostras de tamanho 50.
- Padronizando as médias de cada amostra temos a seguinte distribuição.

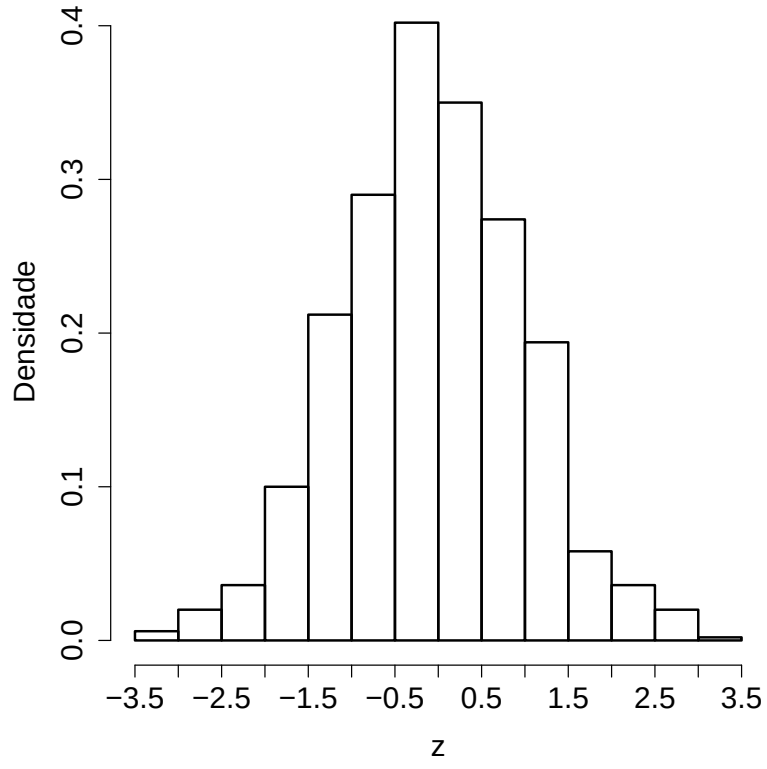

Figura 4: Frequência das médias amostrais.

Amostra Aleatória Simples Sem Reposição (AAS)

- Vê-se que a distribuição da média amostral tem uma “cara” parecida com a distribuição normal padrão.
- O número de AAS possíveis é combinação de 5.000 em grupos de 50: 2,283864431193e+120.
- Veja que em nenhum momento faz-se suposições sobre a distribuição da variável de interesse y.
- Esse é o pulo do gato: independente da distribuição da variável de interesse, a distribuição da média amostral será aproximadamente normal.
- Então, temos que $\Pr\left(\left|\frac{\overline{y}-\overline{Y}}{\sqrt{(1-f)\frac{S^2}{n}}}\right| \leqslant z_{1-\alpha/2}\right) \simeq 1-\alpha.$
- De outra forma,
  $\Pr\left(\overline{y} - z_{1-\alpha/2}\sqrt{(1-f)\frac{S^2}{n}} \leqslant \overline{Y} \leqslant \overline{y} + z_{1-\alpha/2}\sqrt{(1-f)\frac{S^2}{n}}\right) \simeq 1-\alpha,$
  sendo α o nível de significância ou $1-\alpha$ é o nível de confiança.

Amostra Aleatória Simples Sem Reposição (AAS)

- Define-se, então, um intervalo de confiança $1-\alpha$ para a média populacional, baseado no seu valor estimado pela média amostral

$$\text{IC}_{(1-\alpha)100\%}(\overline{Y}) = \left(\overline{y} - z_{1-\alpha/2}\sqrt{(1-f)\frac{S^2}{n}};\ \overline{y} + z_{1-\alpha/2}\sqrt{(1-f)\frac{S^2}{n}} \right).$$

- Como o valor da variância também não é conhecido, vamos definir um intervalo aproximado substituindo S^2 por seu estimador s^2,

$$\text{ic}_{(1-\alpha)100\%}(\overline{Y}) = \left(\overline{y} - z_{1-\alpha/2}\sqrt{(1-f)\frac{s^2}{n}};\ \overline{y} + z_{1-\alpha/2}\sqrt{(1-f)\frac{s^2}{n}} \right).$$

- Pode-se chamar de margem de erro o valor

$$D_{\overline{y}} = z_{1-\alpha/2}\sqrt{(1-f)\frac{S^2}{n}}.$$

- Para o total populacional, tem-se

$$D_{\widehat{Y}} = Nz_{1-\alpha/2}\sqrt{(1-f)\frac{S^2}{n}} = ND_{\overline{y}}.$$

Amostra Aleatória Simples Sem Reposição (AAS)

- No caso de amostras “pequenas” aconselha-se substituir $z_{1-\alpha/2}$ por $t_{1-\alpha/2;\, n-1}$, que é o valor da ordenada da distribuição t-Student, com $n-1$ graus de liberdade.
- Amostra pequena: alguns autores falam em n menor que 30, outros em n menor que 50.

Exercício

Uma AAS de 35 domicílios foi selecionada de uma população de 1.000 domicílios. O número de moradores em cada domicílio da amostra é:

3	1	4	6	6	6	1	3	6	1	2	1
1	1	2	1	2	3	4	3	3	3	6	3
3	2	6	5	6	2	6	1	5	5	6	

1. Construa um intervalo de confiança (aproximado) de 95% para o número médio de moradores por domicílio na população.
2. Construa um intervalo de confiança (aproximado) de 95% para a população total.

Amostra Aleatória Simples Sem Reposição (AAS)

- Costuma-se, também, chamar a margem de erro de erro amostral.
- Neste caso, a palavra erro não tem o significado ruim como popularmente se entende.
- O erro amostral é inerente ao processo da amostragem e existe pelo simples fato de se observar apenas uma amostra da população.
- O erro amostral é controlável e depende da combinação do tamanho da amostra e da variância populacional da variável de interesse y.

Amostra Aleatória Simples Sem Reposição (AAS)

- Então, pode-se representrar o erro amostral por:

$$E_{\widehat{\theta}} = |\widehat{\theta} - \theta|.$$

- Não se pode calcular o erro amostral, pois o parâmetro θ é desconhecido.
- Pode-se, no entanto, estabelecer um limite superior para o erro amostral e calcular a probabilidade dessa ocorrência, ou seja:

$$\Pr(E_{\widehat{\theta}} < D_{\widehat{\theta}}) = 1 - \alpha.$$

- No caso da média temos que:

$$\Pr(E_{\overline{y}} < D_{\overline{y}}) = \Pr(|\overline{y} - \overline{Y}| < D_{\overline{y}}) = 1 - \alpha.$$

Amostra Aleatória Simples Sem Reposição (AAS)

- Então, se fixarmos a margem de erro ou erro amostral máximo admitido um nível de confiança, ou seja:

$$z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{S^2}{n}} = D_{\overline{y}}.$$

- Podemos calcular, utilizando o TCL, um valor para n de maneira que o erro amostral seja menor que a margem superior fixada, com probabilidade $1-\alpha$.
- O tamanho de uma AAS será, então, função do erro amostral máximo admitido e da homogeneidade da população em relação a variável de interesse y.

Amostra Aleatória Simples Sem Reposição (AAS)

- Tamanho da amostra na amostra aleatória simples (AAS)

$$n = \frac{z^2_{1-\alpha/2} S^2}{D^2_{\overline{y}} + \frac{1}{N} z^2_{1-\alpha/2} S^2},$$

ou

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \quad \text{com} \quad n_0 = \frac{z^2_{1-\alpha/2} S^2}{D^2_{\overline{y}}}.$$

- Note que para N "grande" temos $n \simeq n_0$.
- Pode-se fixar a variância do estimador, $\text{Var}(\overline{y})$, ao invés do erro amostral.
- Neste caso,

$$n = \frac{S^2}{\text{Var}(\overline{y}) + \frac{1}{N} S^2},$$

ou

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \quad \text{com} \quad n_0 = \frac{S^2}{\text{Var}(\overline{y})}.$$

Amostra Aleatória Simples Sem Reposição (AAS)

- Outra maneira é trabalhar com o erro relativo:

$$R_{\overline{y}} = \frac{|\overline{y} - \overline{Y}|}{\overline{Y}},$$

- Assim, o erro relativo máximo é dado por

$$R_{\overline{y}} = z_{1-\alpha/2} \frac{1}{|\overline{Y}|} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}}$$

tal que

$$n = \frac{z_{1-\alpha/2}^2 S^2}{R_{\overline{y}}^2 \overline{Y}^2 + \frac{1}{N} z_{1-\alpha/2}^2 S^2},$$

ou

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \quad \text{com} \quad n_0 = \frac{z_{1-\alpha/2}^2 [\text{CV}(y)]^2}{R_{\overline{y}}^2} \quad \text{e} \quad \text{CV}(y) = \frac{S}{\overline{Y}}.$$

Exemplo

Deseja-se estimar a média salarial dos operários de uma empresa que tem 5.000 operários. Sabe-se que a variância dos salários é da ordem de 2.226.000. Qual deve ser o tamanho da amostra de maneira que o erro amostral máximo seja de R$50,00 ao nível de significância de 5%?

$$
\begin{aligned}
N &= 5.000 \\
S^2 &= 2.226.000 \\
D_{\bar{y}} &= 50 \\
z_{0,975} &= 1,96 \\
n_0 &= \frac{(1,96)^2 \times 2.226.000}{50^2} = 3420,561 \\
n &= \frac{3.420,561}{1 + 3.420,561/5.000} = 2031,077.
\end{aligned}
$$

Então, devemos tomar uma amostra de $n = 2032$ salários de operários.

AAS - Estimação de Proporções

Proporção é o número de unidades, de uma população, que possui determinada característica, dividido pelo total de unidades dessa mesma população.

- O denominador é sempre o número de unidades na população.
- Proporção é um número entre 0 e 1 (inclusive).
- Pode ser representada em porcentagem (0 e 100).

Amostragem de proporções:

- Pode-se encarar como caso particular do que foi visto anteriormente.
- O interesse é estimar a proporção de unidades da população que pertence a determinada categoria.
- Exemplos:
  - proporção de alunos do sexo masculino na UFRJ;
  - proporção de fumantes na população; e
  - peças defeituosas em um lote.

AAS - Estimação de Proporções

Notação:

- Seja a população $P_N\{U_1, U_2, \ldots, U_N\}$.
- Supõe-se P_N dividida em duas populações:

$$\begin{aligned} P_{N_C} &: \text{unidades que possuem a característica } C \\ P_{N_{NC}} &: \text{unidades que não possuem a característica } C. \end{aligned}$$

- A proporção das unidades, na população, que possuem a característica C é:

$$P = \frac{N_C}{N}.$$

- A proporção das unidades que não possuem a característica C é:

$$Q = \frac{N_{NC}}{N} = \frac{N - N_C}{N} = 1 - P.$$

AAS - Estimação de Proporções

- Seja uma característica de interesse definida da seguinte maneira:

$$Y_i = \begin{cases} 1, & \text{se } U_i \in P_{N_C}; \\ 0, & \text{se } U_i \notin P_{N_C}; \end{cases}$$

- Então, o total dessa característica será: $Y = \sum_{i=1}^{N} Y_i = N_C$.
- A média será: $\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i = \frac{N_C}{N} = P$
- A variância será: $S^2 = \frac{N}{N-1} PQ = \frac{N}{N-1} P(1-P) \qquad [\sigma^2 = PQ]$.

AAS - Estimação de Proporções

Agora, suponha uma AAS.

- Um estimador não tendencioso para P é:

$$p = \frac{n_C}{n} = \overline{y},$$

  sendo n o tamanho da amostra e n_C o número de unidades que possuem a característica na amostra.

- A variância do estimador é:

$$\text{Var}_{AAS}(p) = (1 - f)\frac{S^2}{n} = \frac{N - n}{N - 1}\frac{PQ}{n}.$$

- Podemos estimar a variância, $\text{Var}_{AAS}(p)$, por:

$$v_{AAS}(p) = \widehat{\text{Var}}_{AAS}(p) = (1 - f)\frac{pq}{n - 1} = \frac{N - n}{N}\frac{pq}{n - 1}.$$

  Este é um estimador não tendencioso?

AAS - Estimação de Proporções

- Para N grande vale a aproximação normal:

$$\Pr(|p - P| \leqslant D_p) \simeq 1 - \alpha,$$

sendo

$$D_p = z_{1-\alpha/2}\sqrt{\text{Var}_{AAS}(p)} = z_{1-\alpha/2}\sqrt{\frac{N-n}{N-1} \times \frac{PQ}{n}},$$

que pode ser aproximado por

$$d_p = z_{1-\alpha/2}\sqrt{v_{AAS}(p)} = z_{1-\alpha/2}\sqrt{\frac{N-n}{N} \times \frac{pq}{n-1}}.$$

- Temos ainda, a correção de continuidade, para n pequeno,

$$d_p = z_{1-\alpha/2}\sqrt{v_{AAS}(p)} + \frac{1}{2n} = z_{1-\alpha/2}\sqrt{\frac{N-n}{N} \times \frac{pq}{n-1}} + \frac{1}{2n}.$$

- Intervalo de confiança para P:

$$\begin{aligned} \text{IC}_{(1-\alpha)100\%}(P) &= (p - D_p;\ p + D_p) \quad \text{aproximado por} \\ \text{ic}_{(1-\alpha)100\%}(P) &= (p - d_p;\ p + d_p). \end{aligned}$$

Exercício:

Em uma eleição deseja-se estimar a votação dos candidatos *A* e *B*. Foi selecionada uma amostra (AAS) de 1.200 eleitores e destes 552 declararam seu voto em *A*, 359 em *B* e os demais se disseram indecisos. Sabendo que existem 1.800.000 eleitores:

1. Estime os votos de *A* e o respectivo $IC_9 5\%$.
2. Estime os votos de *B* e o respectivo $IC_9 5\%$.
3. Estime o número de indecisos e o respectivo $IC_9 5\%$.

Tamanho da amostra para estimar proporções

- Seja D_p a margem de erro máxima admissível. Então:

$$D_p = z_{1-\alpha/2}\sqrt{\frac{N-n}{N-1} \times \frac{PQ}{n}} \Rightarrow D_p^2 = z_{1-\alpha/2}^2 \left(\frac{N}{n} - 1\right) \frac{PQ}{N-1}.$$

- Logo,

$$n = \frac{z_{1-\alpha/2}^2 \dfrac{NPQ}{N-1}}{D_p^2 + z_{1-\alpha/2}^2 \dfrac{PQ}{N-1}} = \frac{n_0}{1 + \dfrac{n_0}{N}} \quad \text{com} \quad n_0 = \frac{z_{1-\alpha/2}^2 NPQ}{(N-1)D_p^2}.$$

- Para N “grande” tem-se:

$$n = \frac{z_{1-\alpha/2}^2 PQ}{D_p^2 + z_{1-\alpha/2}^2 \dfrac{PQ}{N}} = \frac{n_0}{1 + \dfrac{n_0}{N}} \quad \text{com} \quad n_0 = \frac{z_{1-\alpha/2}^2 PQ}{D_p^2}.$$

Tamanho da amostra para estimar proporções

- No caso de proporções a variância é limitada e é máxima quando $P = Q = 0,5$.
- Novamente temos o problema de não conhecer P e, portanto, a variância PQ.
- Como no caso das variáveis contínuas podemos:
  - Avaliar a ordem de grandeza de P em pesquisas anteriores;
  - Usar uma pesquisa piloto para ter uma estimativa de inicial de P; ou
  - Supor que $P = 0,5$ que resulta no máximo da variância ($PQ = 0,25$).
- Na terceira opção, temos

$$n = \frac{0,25 z^2_{1-\alpha/2}}{D_p^2 + \frac{0,25 z^2_{1-\alpha/2}}{N}} = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{com} \quad n_0 = \frac{z^2_{1-\alpha/2}}{4 D_p^2}.$$

Amostragem estratificada

Divisão da população em grupos chamados de estratos, motivada por:

- melhora da precisão das estimativas;
- estimativas independentes para estratos além da população como um todo;
- questões administrativas, estratos naturais.

Motivação técnica:

- $\text{Var}_{AAS}(\overline{y}) = (1 - f)\dfrac{S^2}{n}$;
- Erro diminui quando amostra cresce, mas cresce quando a variabilidade é grande;
- Uma alternativa é dividir a população em grupos (estratos) o mais homogêneos possíveis.

Exemplo:

Seja uma população P_8 sendo que se conhece as variáveis renda domiciliar (y) e bairro de moradia (x) como na tabela abaixo:

i	1	2	3	4	5	6	7	8
Y_i	13	17	6	5	10	12	19	6
X_i	B	A	B	B	B	A	A	B

$$\overline{Y} = 11; \quad \sigma^2 = 24; \quad S^2 = 27,43.$$

	Estrato A			Estrato B				
i	2	6	7	1	3	4	5	8
Y_i	17	12	19	13	6	5	10	6
X_i	A	A	A	B	B	B	B	B

$$\overline{Y}_A = 16 \quad \sigma_A^2 = 8,67 \quad S_A^2 = 13$$
$$\overline{Y}_B = 8 \quad \sigma_B^2 = 9,2 \quad S_B^2 = 11,5$$

Vamos selecionar amostras de tamanho 4 das seguintes maneiras:

- AAS na população P_N;
- AAS de tamanho 2 no estrato A e AAS de tamanho 2 no estrato B.

Então, pode-se calcular:

$$\begin{aligned} \text{Var}_{AAS}(\overline{y}) &= (1-f)\frac{S^2}{n} \\ \text{Var}_{AAS}(\overline{y}_A) &= (1-f_A)\frac{S_A^2}{n_A} \\ \text{Var}_{AAS}(\overline{y}_B) &= (1-f_B)\frac{S_B^2}{n_B}. \end{aligned}$$

Pode-se calcular a média da população como uma média ponderada das médias dos grupos A e B.

Propomos como estimador da média populacional a seguinte estatística:

$$\overline{y}_{ae} = \frac{N_A}{N}\overline{y}_A + \frac{N_B}{N}\overline{y}_B.$$

Portanto,

$$\text{Var}(\overline{y}_{ae}) = \left[\frac{N_A}{N}\right]^2 \text{Var}(\overline{y}_A) + \left[\frac{N_B}{N}\right]^2 \text{Var}(\overline{y}_B).$$

O efeito da estratificação é dado por:

$$\text{EPA} = \frac{\text{Var}(\overline{y}_{ae})}{\text{Var}(\overline{y})}.$$

Escolha do método de amostragem: melhor método é o que tem menor variância para o estimador.

Nem tudo são flores!

Vamos supor que ao invés de usar a variável bairro para estratificar, fosse usada outra variável tal que:

	Estrato A				Estrato B			
i	1	2	3	4	5	6	7	8
Y_i	13	17	6	5	10	12	19	6
X_i	B	A	B	B	B	A	A	B

$$\overline{Y}_A = 10,25 \quad \sigma_A^2 = 24,69 \quad S_A^2 = 32,92$$
$$\overline{Y}_B = 11,75 \quad \sigma_B^2 = 22,19 \quad S_B^2 = 29,58$$

- Será isto razoável?
- Qual será o valor EPA?

Formalização

Seja a população $P_N = \{U_1, U_2, \ldots, U_N\}$.

Dividida em H estratos:

$$\begin{aligned} P_{N_1} &= \{U_{11}, U_{12}, \ldots, U_{1N_1}\} \\ P_{N_2} &= \{U_{21}, U_{22}, \ldots, U_{2N_1}\} \\ \vdots \quad &\vdots \qquad\qquad \vdots \\ P_{N_H} &= \{U_{H1}, U_{H2}, \ldots, U_{HN_H}\}, \end{aligned}$$

sendo que

$$P_N = \bigcup_{h=1}^{H} P_{N_h} \quad \text{e} \quad \bigcap_{h=1}^{H} P_{N_h} = \emptyset.$$

Pode-se representar uma população estratificada por:

Estrato	Total	Média	Variância
1	τ_1	$\mu_1 = \overline{Y}_1$	σ_1^2 ou S_1^2
2	τ_2	$\mu_2 = \overline{Y}_2$	σ_2^2 ou S_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
h	τ_h	$\mu_h = \overline{Y}_h$	σ_h^2 ou S_h^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
H	τ_H	$\mu_H = \overline{Y}_H$	σ_H^2 ou S_H^2
População	τ	$\mu = \overline{Y}$	σ^2 ou S^2

Notação da população

- Tamanho do estrato h: N_h;
- Total do estrato h: $\tau_h = Y_h = \sum_{i=1}^{N_h} Y_{hi}$;
- Média do estrato h: $\mu_h = \overline{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$;
- Variância do estrato h: $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \overline{Y}_h)^2$;
- Peso do estrato h: $W_h = \frac{N_h}{N}$ tal que $\sum_{h=1}^{H} W_h = 1$;
- Tamanho da população: $N = \sum_{h=1}^{H} N_h$;
- Total da população: $\tau = \sum_{h=1}^{H} \tau_h$ ou $Y = \sum_{h=1}^{H} Y_h$;
- Média da população: $\mu = \overline{Y} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi} = \sum_{h=1}^{H} W_h \overline{Y}_h$;
- Variância populacional: $S^2 = \frac{1}{N-1} \sum_{h=1}^{H} \sum_{i=1}^{N_h} (Y_{hi} - \mu)^2$.

Notação da amostra

Considere y_{ih} como a i-ésima observação do estrato h.

- Tamanho da amostra do estrato h: n_h;
- Média amostral do estrato h: $\overline{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$;
- Variância amostral do estrato h: $s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \overline{y}_h)^2$;
- Fração amostral do estrato h: $f_h = \frac{n_h}{N_h}$;
- Tamanho da amostra: $n = \sum_{h=1}^{H} n_h$;

Resultados

Temos que $\sigma^2 = \sigma_d^2 + \sigma_e^2$.

A variância pode ser vista como uma soma das variâncias dentro dos grupos (estratos) mais uma medida da variância entre os grupos, sendo:

$$\sigma_d^2 = \sum_{h=1}^{H} W_h \sigma_h^2 \quad \text{e} \quad \sigma_e^2 = \sum_{h=1}^{H} W_h (\overline{Y}_h - \overline{Y})^2.$$

tal que

$$\sigma^2 = \sigma_d^2 + \sigma_e^2 = \sum_{h=1}^{H} W_h \sigma_h^2 + \sum_{h=1}^{H} W_h (\overline{Y}_h - \overline{Y})^2.$$

De forma análoga, tem-se

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

Estimação

O estimador da média populacional (média global) é dado por:

$$\overline{y}_{ae} = \frac{1}{N}\sum_{h=1}^{H} N_h \overline{y}_h = \sum_{h=1}^{H} W_h \overline{y}_h,$$

sendo $\overline{y}_h$ o estimador da média no estrato h.

Se todos os $\overline{y}_h$ são estimadores não tendenciosos, tem-se que:

$$\mathrm{E}(\overline{y}_{ae}) = \overline{Y}.$$

Em particular, isso ocorre se os estimadores de cada média de estrato for a correspondente média amostral.

Exemplo

Uma vila foi dividida (estratificada) em três conjuntos de domicílios segundo suas características: 1-região onde moram trabalhadores da indústria, 2-moradores mais antigos, 3-área rural. Foi selecionada uma amostra em cada estrato e investigado o número de horas em que se assiste televisão por semana em cada domicílio.

- Estimar a média de horas por semana em cada estrato e na população;
- Existe evidência que o número de horas difere em cada estrato?

Estrato 1				Estrato 2				Estrato 3			
35	28	26	41	27	4	49	10	8	15	21	7
43	29	32	37	15	41	25	30	14	30	20	11
36	25	29	31					12	32	34	24
39	38	40	45								
28	27	35	34								
$N_1 = 155$				$N_2 = 62$				$N_3 = 93$			
$n_1 = 20$				$n_2 = 8$				$n_3 = 12$			

Exemplo (continuação)

$$\begin{aligned} \overline{y}_1 &= 33,900 \quad & s_1^2 &= 33,35789 \\ \overline{y}_2 &= 25,125 \quad & s_2^2 &= 232,4107 \\ \overline{y}_3 &= 19,000 \quad & s_2^2 &= 87,63636 \end{aligned}$$

$$\overline{y}_{ae} = \frac{1}{155+62+93}[155 \times 33,9 + 62 \times 25,125 + 93 \times 19,0] = 27,675.$$

Ao nível de significância de 5%, tem-se:

$$\begin{aligned} \text{ic}_{95\%}(\overline{Y}_1) &= (31,47;\ 36,33) \\ \text{ic}_{95\%}(\overline{Y}_2) &= (15,27;\ 34,98) \\ \text{ic}_{95\%}(\overline{Y}_3) &= (14,05;\ 30,07). \end{aligned}$$

Portanto, existe diferença entre o número médio de horas em cada estrato.

Importante

O estimador sugerido para a média populacional na amostragem estratificada não é a média amostral.

A média amostral é dada por:

$$\begin{aligned}
\overline{y} &= \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} y_{hi} \\
&= \frac{1}{n}\sum_{h=1}^{H} n_h \overline{y}_h \\
&= \sum_{h=1}^{H} w_h \overline{y}_h \quad \text{com} \quad w_h = \frac{n_h}{n}, \quad h = 1, 2, \ldots, H.
\end{aligned}$$

Então, $\overline{y} = \overline{y}_{ae}$ se $w_h = W_h$ para todo $h = 1, 2, \ldots, H$.

Estimação

Um estimador para o total populacional é dado por:

$$\widehat{Y}_{ae} = \sum_{h=1}^{H} \widehat{Y}_h = \sum_{h=1}^{H} N_h \overline{y}_h = N \overline{y}_{ae} \quad \text{com} \quad \widehat{Y}_h = N_h \overline{y}_h.$$

Variância dos estimadores:

- média: $\text{Var}_{AE}(\overline{y}_{ae}) = \sum_{h=1}^{H} W_h^2 \text{Var}(\overline{y}_h)$.
- total: $\text{Var}_{AE}(\widehat{Y}_{ae}) = N^2 \sum_{h=1}^{H} W_h^2 \text{Var}(\overline{y}_h) = \sum_{h=1}^{H} N_h^2 \text{Var}(\overline{y}_n)$.

E se em todos os estratos tivermos AAS?

Alocação da amostra

Um problema na amostragem estratificada é determinar como dividir as n unidades da amostra total em cada estrato de modo que:

$$n = \sum_{h=1}^{H} n_h.$$

Chama-se este problema de alocação da amostra.

Os três tipos principais de alocação são:

- Alocação proporcional;
- Igual;
- Ótima ou de Neyman.

Alocação proporcional

Nesse tipo de alocação o número de unidades na amostra em cada estrato é proporcional ao tamanho do estrato:

$$n_h = nW_h = n\frac{N_h}{N}, \quad \text{para todo} \quad h = 1, 2, \ldots, H.$$

Portanto:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N}, \quad \text{para todo} \quad h = 1, 2, \ldots, H.$$

Alocação igual

Tem-se que

$$n_h = \frac{n}{H}, \quad \text{para todo} \quad h = 1, 2, \ldots, H.$$

Pode-se adaptar adequadamente as fórmulas das variâncias dos estimadores para cada alocação.

Exemplo

Uma região possui 60 municípios e deseja fazer uma amostragem para atualizar a estimativa do total de sua população. Para isso foi decidido pesquisar 20 cidades e deseja-se saber qual seria o mais eficiente para o caso: uma amostra aleatória simples (AAS), uma amostra aleatória estratificada (AAE) com alocação proporcional ou uma AAE com alocação igual.
As cidades foram agrupadas em dois estratos segundo a população apurada no último Censo (cidades grandes: mais de 300 mil habitantes; e cidades pequenas: menos de 300 mil habitantes). A tabela mostra essa estratificação e as populações, no censo, em milhares de habitantes.

Estrato 1														
776	622	583	502	468	468	438	437	419	416	404	382	370	346	318
Estrato 2														
297	295	294	292	290	285	270	255	250	250	244	241	238	236	234
231	220	218	215	211	204	202	201	192	190	190	188	178	178	171
167	166	163	162	157	145	141	141	139	125	122	118	112	111	110

	Estrato 1	Estrato 2
Total	6.949	9.039
Soma de Quadrado	3.417.311	1.954.179

Alocação ótima de Neyman

Sabe-se que populações (ou estratos) grandes precisam de amostras grandes.

Sabe-se que fenômenos com grande variabilidade também precisam de amostras grandes.

Suponha que o custo para pesquisar uma unidade amostral possa variar para cada estrato.

A alocação ótima de Neyman leva tudo isso em conta:

$$n_h = n \times \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}}. \tag{1}$$

O custo da pesquisa será suposto linear, ou seja:

$$C = c_0 + \sum_{h=1}^{H} n_h c_h, \tag{2}$$

sendo c_0 o "custo de escritório" que não depende de h e c_h o custo de pesquisar uma unidade do estrato h.

Alocação ótima de Neyman

Tendo o custo fixo, pode-se calcular o tamanho da amostra, substituindo (1) em (2), por:

$$n = \frac{(C - c_0)\sum_{h=1}^{H} N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h \sqrt{c_h}}.$$

Fixando a variância desejada para estimar a média como V, ou seja,

$$\begin{aligned} V &= \mathrm{Var}_{AE}(\overline{y}_{ae}) = \sum_{h=1}^{H} W_h^2 \mathrm{Var}(\overline{y}_h) = \sum_{h=1}^{H} W_h^2 (1 - f_h)\frac{S_h^2}{n_h} \\ &= \sum_{h=1}^{H} \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2, \end{aligned}$$

e substituindo (1) em V, tem-se:

$$n = \frac{\left(\sum_{h=1}^{H} W_h S_h/\sqrt{c_h}\right)\left(\sum_{h=1}^{H} W_h S_h \sqrt{c_h}\right)}{V + \frac{1}{N}\sum_{h=1}^{H} W_h S_h^2}.$$

Para o caso em que os custos para coleta dos dados independem do estrato tem-se

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} = n \times \frac{W_h S_h}{\sum_{h=1}^{H} W_h S_h}.$$

A alocação acima é a que minimiza a variância quando o tamanho total da amostra, n, é dado.

Utiliza-se também essa fórmula quando não se tem nenhuma ideia sobre o custo de coleta nos estratos.

Dado um tipo de alocação, pode-se adaptar as fórmulas das variâncias dos estimadores:

- Alocação proporcional:

$$\text{Var}(\overline{y}_{es:prop}) = \frac{1-f}{nN}\sum_{h=1}^{H} N_h S_h^2.$$

- Alocação igual ou uniforme:

$$\text{Var}(\overline{y}_{es:igual}) = \frac{1}{N^2}\left[\frac{H}{n}\sum_{h=1}^{H} N_h^2 S_h^2 - \sum_{h=1}^{H} N_h S_h^2\right].$$

- Alocação ótima ou Neyman:

$$\text{Var}(\overline{y}_{es:otima}) = \frac{1}{N^2}\left[\frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 - \sum_{h=1}^{H} N_h S_h^2\right].$$

Tamanho da amostra estratificada - caso geral

Seja V a variância mínima desejada para estimar a média da população.

Seja uma alocação qualquer $n_h = na_h$, sendo a_h a constante que define a alocação no estrato h.

Pela fórmula da variância da média temos

$$V = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{na_h}{N_h}\right) \frac{S_h^2}{na_h} = \frac{1}{n} \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{a_h} - \frac{1}{N} \sum_{h=1}^{H} W_h S_h^2.$$

Logo,

$$n = \frac{\sum_{h=1}^{H} W_h^2 \frac{S_h^2}{a_h}}{V + \frac{1}{N} \sum_{h=1}^{H} W_h S_h^2}.$$

Portanto, os a_h's vão depender da alocação escolhida.

Proporcional: $a_h = W_h$. Igual: $a_h = \frac{1}{H}$. Ótima: $a_h = \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} W_h S_h / \sqrt{c_h}}$

Pode-se usar a mesma estratégia do cálculo do tamanho da amostra na AAS, calculando n_0 e depois n, sendo:

$$n_0 = \frac{1}{V} \sum_{h=1}^{H} \frac{W_h^2 S_h^2}{a_h}, \qquad n = \frac{n_0}{1 + \frac{1}{NV} \sum_{h=1}^{H} W_h S_h^2}.$$

Fixando o erro, d, ao invés da variância V:

$$V = \left[\frac{d}{z_{1-\alpha/2}} \right]^2.$$

Intervalos de confiança

Intervalos de confiança para a média e total, supondo AAS dentro dos estratos.

- Seja a variância amostral em cada estrato:

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \overline{y}_h)^2.$$

- Um estimador não tendencioso para variância da estimativa da média é:

$$v(\overline{y}_{ae}) = \sum_{h=1}^{H} \frac{W_h^2 s_h^2}{n_h} - \sum_{h=1}^{H} \frac{W_h s_h^2}{N}.$$

- Intervalos de confiança podem ser calculados por:

$$\begin{aligned}
\text{ic}_{(1-\alpha)100\%}(\overline{Y}) &= \left(\overline{y}_{ae} - z_{1-\alpha/2} \sqrt{v(\overline{y}_{ae})}, \overline{y}_{ae} + z_{1-\alpha/2} \sqrt{v(\overline{y}_{ae})} \right) \\
\text{ic}_{(1-\alpha)100\%}(Y) &= \left(N\overline{y}_{ae} - z_{1-\alpha/2} N \sqrt{v(\overline{y}_{ae})}, N\overline{y}_{ae} + z_{1-\alpha/2} N \sqrt{v(\overline{y}_{ae})} \right).
\end{aligned}$$

Intervalos de confiança

Quando as amostras nos estratos são muito pequenas aconselha-se substituir a aproximação Normal pela t-Student com ν graus de liberdade.

Os graus de liberdade são dados por

$$\nu = \frac{\left[\sum_{h=1}^{H} g_h s_h^2\right]^2}{\sum_{h=1}^{H} \left[g_h^2 s_h^4 (n_h - 1)\right]^2} \quad \text{com} \quad g_h = \frac{N_h(N_h - n_h)}{n_h}.$$

O grau de liberdade ν pode ser arredondado para um inteiro positivo.

Estrato com tamanho unitário

Suponha que os estrato certo é $h = 1$.

Então, o estimador da média da população é

$$\overline{y}_{ae} = \sum_{h=1}^{H} W_h \overline{y}_h = W_1 \overline{Y}_1 + \sum_{h=2}^{H} W_h \overline{y}_h,$$

com variância dada por

$$\text{Var}_{AE}(\overline{y}_{ae}) = W_1^2 \text{Var}_{AE}(\overline{Y}_1) + \sum_{h=2}^{H} W_h^2 \text{Var}_{AE}(\overline{y}_h) = \sum_{h=2}^{H} W_h^2 \text{Var}_{AE}(\overline{y}_h),$$

pois

$$\text{Var}_{AE}(\overline{Y}_1) = 0.$$

Para o total populacional, tem-se:

$$\widehat{Y}_{ae} = N_1 \overline{Y}_1 + \sum_{h=2}^{H} N_h \overline{y}_h, \quad \text{e}$$

$$\text{Var}_{AE}(\widehat{Y}_{ae}) = \sum_{h=2}^{H} N_h^2 \text{Var}_{AE}(\overline{y}_h).$$

Amostragem estratificada: proporções

Em uma população com H estratos: $P_h = \dfrac{N_{ch}}{N_h}$,

sendo N_{ch} o número de unidades no estrato h com a característica.

A variância de y no estrato h é dada por: $S_h^2 = \dfrac{N_h P_h Q_h}{N_h - 1}$.

Estima-se a proporção na população por: $p_{ae} = \dfrac{1}{N} \sum_{h=1}^{H} N_h p_h = \sum_{h=1}^{H} W_h p_h$,

com $p_h = \overline{y}_{ae} = \dfrac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$.

Sua variância é dada por

$$\text{Var}_{AE}(p_{ae}) = \sum_{h=1}^{H} W_h^2 \text{Var}_{AE}(p_h) = \sum_{h=1}^{H} W_h^2 \left[\frac{N_h - n_h}{N_h - 1}\right] \frac{P_h Q_h}{n_h}.$$

Tamanho da amostragem

- Alocação proporcional: $n_h = nW_h$.
- Fixando a variância V:

$$V = \sum_{h=1}^{H} W_h^2 \left[\frac{N_h - nW_h}{N_h - 1}\right] \frac{P_h Q_h}{nW_h} = \frac{N}{n} \sum_{h=1}^{H} W_h^2 \frac{P_h Q_h}{N_h - 1} - \sum_{h=1}^{H} W_h^2 \frac{P_h Q_h}{N_h - 1} \Rightarrow$$

$$n = \frac{N \sum_{h=1}^{H} W_h^2 \frac{P_h Q_h}{N_h - 1}}{V + \sum_{h=1}^{H} W_h^2 \frac{P_h Q_h}{N_h - 1}} = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{com} \quad n_0 = \frac{N \sum_{h=1}^{H} W_h^2 \frac{P_h Q_h}{N_h - 1}}{V}.$$

- Supondo N_h "grande":

$$n = \frac{\sum_{h=1}^{H} W_h P_h Q_h}{V + \frac{1}{N} \sum_{h=1}^{H} W_h P_h Q_h} = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{com} \quad n_0 = \frac{\sum_{h=1}^{H} W_h P_h Q_h}{V}.$$

Amostragem sistemática

Definição: selecionar cada k-ésima unidade de uma população, começando de uma partida aleatória (r).

Características principais:

- As unidades são selecionadas sem reposição (sem repetição).
- As unidades têm todas a mesma chance de serem selecionadas ($1/k$).

Aspectos importantes

- Requer acesso a lista das unidades da população.
- A lista pode ser construída ao mesmo tempo em que está sendo feita a seleção da amostra e coleta dos dados.
- As unidades são selecionadas uma de cada vez.
- Deve-se especificar
  - Intervalo de seleção ou amostragem (k).
  - Partida aleatória (r).

Intervalo de seleção: parte inteira do tamanho da população dividido pelo tamanho da amostra,

$$k = \left\lfloor \frac{N}{n} \right\rfloor, \qquad \text{inteiro } (\textit{floor}).$$

Partida aleatória: número inteiro escolhido aleatoriamente entre 1 e k.

Probabilidade de seleção para todas as unidades na amostra:

$$\Pr(U_i \in \boldsymbol{s}) = \frac{1}{k}.$$

Exemplo

Temos $N = 36$ e $n = 9 \quad \Rightarrow \quad k = \left\lfloor \frac{36}{9} \right\rfloor = 4, \quad r = 2.$

X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X

Unidades selecionadas: U_2, U_6, U_{10}, U_{14}, U_{18}, U_{22}, U_{26}, U_{30} e U_{34}.

Exemplo

Temos $N = 49$ e $n = 7 \quad \Rightarrow \quad k = \left\lfloor \frac{49}{7} \right\rfloor = 7, \quad r = 5.$

+	+	+	+	+	+	+
+	+	+	+	+	+	+
+	+	+	+	+	+	+
+	+	+	+	+	+	+
+	+	+	+	+	+	+
+	+	+	+	+	+	+
+	+	+	+	+	+	+

As unidades marcadas são selecionadas para a amostra.

Vantagens

- Fácil de selecionar.
- Espalha a amostra sobre a população.
- Fácil de estimar os parâmetros.

Desvantagens

- Custo elevado (amostra espalhada).
- Influência da ordenação (periodicidade).
- Defícil estimativa da precisão.

Exemplo

Temos $N = 23$ e $n = 4 \quad \Rightarrow \quad k = 5$.

I	II	III	IV	V
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23		

Se não houver um critério de parada para o tamanho da amostra n, este pode variar de acordo com o valor de r .

Para $n > 50$ ignora-se o problema.

Alternativa sugerida por Lahiri: circular.

- Suponha que as N unidades estão dispostas em "círculo": após a última unidade vem a primeira.
- Seja k o inteiro mais próximo de N/n (*round*).
- Seja r um número aleatório entre 1 e N.
- Amostra: tome o elemento r as k-ésimas unidades seguintes ao redor do "círculo", até completar n unidades.
- Exemplo: $N = 23$, $n = 5$, $k = 5$, $r = 12$;
  Amostra: 12, 17, 22, 4 e 9.
- Exemplo: $N = 23$, $n = 3$, $k = 8$, $r = 15$;
  Amostra: 15, 23 e 8.

Tabela 3: Composição das k amostras sistemáticas.

	Número da amostra					
	1	2	$\cdots$	i	$\cdots$	k
	y_1	y_2	$\cdots$	y_i	$\cdots$	y_k
	y_{k+1}	y_{k+2}	$\cdots$	y_{k+i}	$\cdots$	y_{2k}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	$y_{(n-1)k+1}$	$y_{(n-1)k+2}$	$\cdots$	$y_{(n-1)k+i}$	$\cdots$	y_{nk}
Média	$\overline{y}_1$	$\overline{y}_2$	$\cdots$	$\overline{y}_i$	$\cdots$	$\overline{y}_k$

- Estimador do total:

$$\widehat{Y}_{AS} = k\sum_{i=1}^{n} y_i = ky_r \quad \text{com} \quad y_r = \sum_{i=1}^{n} y_i.$$

- Estimador da média:

$$\widehat{\overline{Y}}_{AS} = \overline{y}_{AS} = \frac{\widehat{Y}_{AS}}{N} = \frac{ky_r}{N}.$$

Se $N = nk$, então

$$\overline{y}_{AS} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- Variância do estimador, pela definição:

$$\text{Var}_{AS}(\overline{y}_{AS}) = \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_{i.} - \overline{Y})^2.$$

$\overline{y}_{i.}$ é a média da i-ésima amostra possível.

Estimativa da variância

Quando os valores da população estão “ordenados” aleatoriamente pode-se fazer uma analogia com AAS:

$$v(\overline{y}_{AS}) = \frac{N-n}{Nn} \times \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y}_{AS})^2.$$

Caso exista uma ordenação, as coisas podem complicar um pouco! Ver Cochran capítulo 8.

Efeito de estratificação

$$v(\overline{y}_{AS}) = \frac{N-n}{Nn} \times \frac{1}{2(n-1)} \sum (y_i - y_{i+k})^2.$$

Esta fórmula pode ser usada quando se nota uma “estratificação” na população amostrada.

Tendência linear:

$$v(\overline{y}_{AS}) = \frac{N-n}{Nn} \times \frac{1}{6(n-2)} \sum (y_i - 2y_{i+k} + y_{i+2k})^2.$$

Ambas as fórmulas podem funcionar bem para amostras razoavelmente grandes, quando as suposições forem razoáveis.

Proposta: se possível aleatorizar a população antes de selecionar (usar a ideia de Hàjek) e usar a fórmula da AAS.

Exemplos em exemplo_as.xls e exemplo_as_ordenacao.xls.

Amostragem Com Probabilidades Desiguais

- Em alguns casos na amostragem, associa-se probabilidades distintas de cada unidade da população ser incluída na amostra.
  - Unidades distintas em importância para o fenômeno estudado.
  - Unidades com tamanhos muito variados.
  - Estimativas mais precisas.
- As diferentes probabilidades de seleção serão levadas em conta na estimativa dos parâmetros e de sua respectiva precisão.

Caso de seleção com reposição

Seja uma população P_N qualquer onde a probabilidade de ser selecionada a unidade U_i para a amostra seja p_i,

$$\Pr(U_i) = p_i, \quad i = 1, 2, \ldots, N,$$

com

$$\sum_{i=1}^{N} p_i = 1, \qquad p_i > 0, \quad i = 1, 2, \ldots, N.$$

Um estimador não tendencioso para o total da população é dado por (Hansen-Hurwitz):

$$\widehat{\tau}_{HH} = \widehat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

- Variância do estimador do total: $\text{Var}(\widehat{Y}_{HH}) = \dfrac{1}{n}\sum_{i=1}^{N} p_i \left(\dfrac{Y_i}{p_i} - \tau\right)^2$.
- Pode ser estimada por: $v(\widehat{Y}_{HH}) = \dfrac{1}{n(n-1)}\sum_{i=1}^{n} \left(\dfrac{y_i}{p_i} - \widehat{\tau}_{HH}\right)^2$.
- Estimador da média: $\widehat{\mu}_{HH} = \overline{y}_{HH} = \dfrac{1}{nN}\sum_{i=1}^{n} \dfrac{y_i}{p_i} = \dfrac{\widehat{\tau}_{HH}}{N}$.
- Variância do estimador da média e seu respectivo estimador:

$$\text{Var}(\widehat{\mu}_{HH}) = \frac{1}{N^2}\text{Var}(\widehat{\tau_{HH}}) \quad \text{e} \quad v(\widehat{\mu}_{HH}) = \frac{1}{N^2}v(\widehat{\tau_{HH}}).$$

Exemplo

Estimar total de animais em uma área de 100km^2 que foi dividida em lotes com 1km de largura e com comprimento variável. Uma amostra de 4 lotes, com reposição, foi retirada. Os resultados estão na tabela abaixo:

Animais	Probabilidade
60	0,05
60	0,05
14	0,02
1	0,01

Caso Particular: Probabilidade Proporcional ao Tamanho (PPT)

- Suponha que se tenha a informação sobre uma variável, X, que possa ser entendida como tamanho da unidade amostral.
- Exemplos: valor do faturamento de uma empresa; número de pés de café de uma fazenda; área plantada de lavoura; número de alunos da escola; etc.
- Pode-se, então, associar probabilidades de seleção de cada unidade proporcionais a sua medida de tamanho (ou importância).
- No exemplo anterior tínhamos a área dos lotes.

- Dessa maneira define-se: $p_i = \dfrac{X_i}{\sum_{i=1}^{N} X_i} = \dfrac{X_i}{N\overline{X}}, \quad i = 1, 2, \ldots, N.$
- O estimador do total populacional fica: $\widehat{Y}_{PPT} = \dfrac{N\overline{X}}{n} \sum_{i=1}^{n} \dfrac{y_i}{X_i}.$
- A variância do estimador do total é dada por:

  $\text{Var}(\widehat{\tau}_{PPT}) = \dfrac{1}{nN\overline{X}} \sum_{i=1}^{N} X_i \left(\dfrac{Y_i}{X_i} N\overline{X} - \tau \right)^2.$
- A estimativa desta variância: $v(\widehat{\tau}_{PPT}) = \dfrac{1}{n(n-1)} \sum_{i=1}^{n} \left(\dfrac{y_i}{X_i} N\overline{X} - \widehat{\tau}_{PPT} \right)^2.$

Algoritmo para seleção

1. Acumular os tamanhos.
2. Determinar os intervalos de seleção.
3. Aleatório entre 1 e tamanho total.
4. Selecionar unidade correspondente ao intervalo onde caiu o aleatório.
5. Repetir 3 e 4 até selecionar n unidades.

Exemplo: Estimar a despesa total das fazendas, usando uma amostra de $n = 2$.

Fazenda	Área	Área Acumulada	Intervalo		p_i	Despesa
			Inferior	Superior		
1	50	50	1	50	1/15	50
2	500	550	51	550	2/3	700
3	200	750	551	750	4/15	180

Todas as amostras possiíveis de tamanho 2:

s	y_1	y_2	$\Pr(y_1)$	$\Pr(y_2)$	$\Pr(y_1, y_2)$	$\widehat{Y}_{PPT}$	$\widehat{Y}_{PPT} \times \Pr(y_1, y_2)$	$(\widehat{Y}_{PPT} - Y)^2 \times \Pr(y_1, y_2)$	Y_{AAS}
11	50	50	1/15	1/15	1/225	750,00	3,33	144,00	150,00
12	50	700	1/15	2/3	4/45	900,00	40,00	40,00	1.125,00
13	50	180	1/15	4/15	4/225	712,50	12,67	841,00	345,00
21	700	50	2/3	1/15	2/45	900,00	40,00	40,00	1.125,00
22	700	700	2/3	2/3	4/9	1.050,00	466,67	6.400,00	2.100,00
23	700	180	2/3	4/15	8/45	862,50	153,33	810,00	1.320,00
31	180	50	4/15	1/15	4/225	712,50	12,67	841,00	345,00
32	180	700	4/15	2/3	8/45	862,50	153,33	810,00	1.320,00
33	180	180	4/15	4/15	16/225	675,00	48,00	4.624,00	540,00
							930,00	14.550,00	930,00

- Seleção sistemática.
  1. Acumular os tamanhos.
  2. Determinar os intervalos de seleção.
  3. Intervalo de amostragem: $k = \dfrac{\text{Tamanho total acumulado}}{\text{Tamanho da amostra}} = \dfrac{N\overline{X}}{n}$.
  4. Aleatório entre 1 e k.
  5. Selecionar a unidade correspondente ao aleatório.
  6. Somar k ao aleatório e selecionar a unidade correspondente até ter a amostra desejada.
- Sistemática aleatorizada.
  1. Reordenar aleatoriamente as unidades (Hàjek).
  2. Proceder uma seleção sistemática.

Estimador de Horvitz-Thompson

- Pode ser utilizado para amostras retiradas com ou sem reposição.
- Baseado na probabilidade de inclusão da unidade na amostra.
- Seja π_i a probabilidade da unidade i ser incluída na amostra $i = 1, 2, \ldots, N$.
- Estimador não tendencioso para o total populacional dado por:

$$\widehat{\tau}_{HT} = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i},$$

sendo ν o número de unidades não repetidas na amostra.

- Será o estimador de Horvitz-Thompson não tendencioso?
- Seja a variável aleatória δ tal que:

$$\delta_i = \begin{cases} 1, & \text{se a unidade } U_i \text{ percente à amostra;} \\ 0, & \text{caso contrário.} \end{cases}$$

- Portanto, essa será uma variável aleatória de Bernouli com

$$\begin{aligned} \mathrm{E}_p(\delta_i) &= \pi_i \\ \mathrm{Var}_p(\delta_i) &= \pi_i(1-\pi_i) \\ \mathrm{Cov}(\delta_i, \delta_j) &= \pi_{ij} - \pi_i\pi_j. \end{aligned}$$

- Pode-se ver que o estimador Horvitz-Thompson é não tendencioso redefinindo-o:

$$\widehat{\tau}_{HT} = \sum_{i=1}^{N} \frac{Y_i}{\pi_i}\delta_i.$$

- Variância do estimador de Horvitz-Thompson:

$$\mathrm{Var}_p(\widehat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} Y_i Y_j.$$

- Um estimador é dado por:

$$\widehat{\mathrm{Var}}_p(\widehat{Y}_{HT}) = v_p(\widehat{Y}_{HT}) = \sum_{i=1}^{\nu} \frac{1-\pi_i}{\pi_i^2} y_i^2 + \sum_{i=1}^{\nu}\sum_{j\neq i}^{\nu} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}} y_i y_j.$$

- Problema sério: pode dar menor que zero!
- Estimador alternativo:

$$\widetilde{v}(\widehat{\tau}_{HT}) = \frac{(N-\nu)}{N}\frac{s_t^2}{\nu}, \quad s_t^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(t_i - \widehat{\tau}_{HT})^2, \quad t_i = \frac{\nu y_i}{\pi_i}.$$

Uso do estimador de Horvitz-Thompson para seleção com reposição

- Já vimos que a probabilidade de seleção da unidade i é p_i, $i = 1, 2, \ldots, N$.
- A probabilidade de inclusão da unidade i na amostra é:

$$\pi_i = 1 - (1 - p_i)^n.$$

- A probabilidade de inclusão das unidades i e j simultaneamente:

$$\pi_{ij} = \pi_i + \pi_j - [1 - (1 - p_i - p_j)^n].$$

- No caso geral (sem reposição) os valores de π_i não são simples de se obter *a priori*.
- Nos livros de Cochran e de Thompson são citadas várias alternativas (complicadas) de tratar o problema.

Exemplo

Estimar o pessoal ocupado total usando amostragem com probabilidade proporcional ao tamanho, sendo a *Receita* a variável de tamanho e o tamanho da amostra $n = 3$.

Empresa	Pessoal ocupado	Receita	p_i	π_i
1	25	800	0,0816	0,2255
2	35	1200	0,1224	0,3242
3	17	600	0,0612	0,1727
4	31	1000	0,1020	0,2759
5	16	500	0,0510	0,1454
6	35	1800	0,1837	0,4560
7	38	1800	0,1837	0,4560
8	15	400	0,0408	0,1175
9	27	900	0,0918	0,2510
10	24	800	0,0816	0,2255

Informações Auxiliares na Amostragem

O que são informações auxiliares?

- O interesse da amostragem é estimar parâmetros populacionais relativos à uma (ou mais) determinada variável y.
- A unidade i da população está associada a observação (valor) y_i da variável de interesse.
- Além da variável y pode-se ter associada à unidade i uma (ou mais) variável x correlacionada com y, que pode ser usada como variável auxiliar.

Por que usar?

- Para melhorar a precisão das estimativas!

Onde e quando?

- As variáveis auxiliares podem ser usadas na definição do plano amostral ou na definição dos estimadores.
- Já vimos dois usos: estratificação e seleção PPT (ambos na definição do plano amostral).

Exemplos:

- Variável de interesse: número de cabeças de gado.
- Auxiliares: área do estabelecimento rural, valor da produção em uma pesquisa anterior, localização, área de pastagem, etc.

- Variável de interesse: valor do faturamento da empresa.
- Auxiliares: pessoal ocupado, ramo de atividade, localização, quantidade produzida, etc.

Exemplo real do Censo Demográfico

Desde 1960 o Censo do Brasil tem dois questionários:

- BÁSICO (efetivamente um censo), aplicado a todos os domicílios com poucos quesitos (variáveis): sexo, idade, alfabetização, situação e características básicas do domicílio
- AMOSTRA, aplicado a uma amostra de domicílios (10% nas cidades grandes e 20% nas cidades pequenas, no Censo 2000): todas as variáveis do questionário básico mais outras características do domicílio, fecundidade, cor ou raça, deficiências, nupcialidade, escolaridade, religião, migração, mão-de-obra e rendimento.

As variáveis do questionário básico são usadas como variáveis auxiliares para a estimação das variáveis pesquisadas apenas no questionário da amostra.

- Calibração dos pesos amostrais.

Uso das informações auxiliares:

- No plano amostral: estratificação, seleção com PPT.
- Na estimação de parâmetros: vamos ver (ou pelo menos começar a ver):
  - Estimadores de razão.
  - Estimadores de regressão.
  - Estimadores de pós-estratificação.

Definindo uma razão

Definição de razão:

- Razão é a relação (quociente) entre duas variáveis de uma população.

Exemplos:

- Relação entre os gastos das famílias com saúde e a renda total das famílias.
- Produtividade da terra: razão entre a quantidade de soja colhida e a área plantada de soja.
- Renda *per capita*: relação entre o PIB de um país e o número de habitantes.

Não confundir com proporção que é sempre o total de uma contagem de unidades com certa característica, dividido pelo número de unidades da população.

Estimador de uma razão

- Então, uma razão será definida como:

$$R = \frac{Y}{X} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{\overline{Y}}{\overline{X}}.$$

Suponha uma AAS.

- Um estimador intuitivo para R é a razão na amostra:

$$r = \widehat{R} = \frac{\widehat{Y}}{\widehat{X}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\overline{y}}{\overline{x}}.$$

Propriedades do Estimador da Razão

- O estimador r é tendencioso.
- Exemplo prático:
  - Seja uma população P_3 com $Y = \{2, 1, 4\}$ e $X = \{4, 2, 5\}$.
  - Calculando a razão entre Y e X temos $R = 7/11$.
  - Supondo AAS, temos 3 possíveis amostras de tamanho 2 e a distribuição amostral de r é dada por:

r	1/2	2/3	5/7
Pr(r)	1/3	1/3	1/3

  - Portanto, $\mathrm{E}(r) = 79/126 \neq R = 7/11$.
- Note que os estimadores das médias de y e x são não tendenciosos, porém r é tendencioso.

Propriedades do Estimador da Razão

- O estimador r é consistente:
  - **Definição:** seja $\widehat{\Theta}$ um estimador de baseado em uma amostra de tamanho n. O estimador será dito consistente se:

$$\widehat{\Theta}_N = \theta.$$

  - Logo, r é consistente porque se o calcularmos sobre toda a população teremos $r_N = R$.
- O estimador r é dito assintoticamente não tendencioso, ou seja: a tendenciosidade do estimador tende a zero quando $n \to N$.
- Pode-se mostrar que:

$$T(r) = B(r) \simeq \frac{R(N-n)}{Nn}\left[\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right] = \frac{R(N-n)}{Nn}\left[(\mathrm{CV}_x)^2 - \rho \mathrm{CV}_x \mathrm{CV}_y\right]$$

- ρ é o coeficiente de correlação de entre as variáveis x e y.
- Então, nota-se que $T(r) \to 0$ quando $n \to N$.
- Temos que r será não tendencioso se $\overline{Y} = \rho \frac{S_y}{S_x}\overline{X}$.

Propriedades do Estimador da Razão

- Erro quadrático médio de r:

$$\text{EQM}(r) = \text{Var}(r) + [T(r)]^2 = \text{Var}(r)\left[1 + \frac{[T(r)]^2}{\text{Var}(r)}\right].$$

- Logo r será aproximadamente não tendencioso se $\dfrac{[T(r)]^2}{\text{Var}(r)} \simeq 0$.
- Na prática, admitimos $\dfrac{[T(r)]^2}{\text{Var}(r)} \leqslant 0,01$ ou $\dfrac{|T(r)|}{\sqrt{\text{Var}(r)}} \leqslant 0,10$.
- Pode-se mostrar que $\dfrac{T(r)}{\sqrt{\text{Var}(r)}} = -\rho^{\star}\dfrac{\sqrt{\text{Var}(x)}}{\overline{X}}$, sendo $\rho^{\star}$ o coeficiente de correlação entre r e $\overline{x}$.
- Como o coeficiente de correlação é no máximo 1 em valor absoluto, temos que:

$$\frac{|T(r)|}{\sqrt{\text{Var}(r)}} \leqslant \text{CV}(\overline{x}).$$

Propriedades do Estimador da Razão

- Então, para que a tendenciosidade do estimador r seja pequena devemos ter uma amostra de tamanho n tal que

$$\mathrm{CV}(\overline{x}) \leqslant 0,10 \quad \text{ou} \quad n_0 \geqslant \left[\frac{\mathrm{CV}_x}{0,10}\right]^2 \Rightarrow n \geqslant \frac{n_0}{1+(n_0/N)},$$

sendo CV_x o coeficiente de variação populacional da variável auxiliar x.

Exemplo:

Deseja-se estimar a razão entre uma variável y e outra x. Sabe-se por uma pesquisa anterior que S_x^2 é da ordem de 100 e sua média é 5. Qual deve ser o tamanho da amostra para que a tendenciosidade da estimativa seja desprezível, sabendo-se que a população tem 5.000 unidades?

$$n_0 = \left[\frac{(10/5)}{0,1}\right]^2 = 20^2 = 400 \Rightarrow n \geqslant \frac{400}{1+(400/5000)} = 370,37.$$

Variância do Estimador da Razão

- Quando pudermos supor a tendenciosidade de estimação desprezível, ou seja, $\frac{|T(r)|}{\sqrt{\text{Var}(r)}} \leqslant 0,10$, podemos calcular a variância de r por:

$$\text{Var}(r) \simeq \frac{1-f}{n\overline{X}^2}[S_y^2 + R^2 S_x^2 - 2\rho R S_x S_y] = \frac{1-f}{n\overline{X}^2} \times \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2.$$

- Se a média de x for conhecida $\text{Var}(r)$ pode ser estimada por

$$v(r) \simeq \frac{1-f}{n\overline{X}^2}[s_y^2 + r^2 s_x^2 - 2rs_{xy}] = \frac{1-f}{n\overline{X}^2} \times \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2 = \frac{1-f}{n\overline{X}^2}s_r^2.$$

Se a média de x for desconhecida, substitui-se $\overline{X}$ por $\overline{x}$ na fórmula acima.

Exemplo:

Seja a população P_3 dada na tabela:

U_i	Y_i	X_i
1	3	2
2	2	3
3	5	5

- Calcular a razão populacional entre y e x.
- Estimar R para todas as AAS possíveis de tamanho 2.
- Calcular a tendenciosidade verdadeira e a aproximada para r.
- Calcular o EQM verdadeiro e o aproximado para r.
- Calcule a variância verdadeira e aproximada para r.

Margem de Erro do Estimador da Razão

- Seja: $n_{0x} \geqslant \left[\frac{\text{CV}_x}{0,1}\right]^2$ e $n_{0y} \geqslant \left[\frac{\text{CV}_y}{0,1}\right]^2$.
- Seja: $n \geqslant \text{máx}\left\{30;\ \frac{n_{0x}}{1+(n_{0x}/N)};\ \frac{n_{0y}}{1+(n_{0y}/N)}\right\}$.
- Assim, temos que: $\frac{r-R}{\sqrt{\text{Var}(r)}} \approx \mathcal{N}(0,1)$.
- Logo: $D(r) = z_{1-\alpha/2}\sqrt{\text{Var}(r)}$ e $D_r(r) = z_{1-\alpha/2}\text{CV}(r)$.
- Estimado por: $d(r) = z_{1-\alpha/2}\sqrt{v(r)}$ e $d_r(r) = z_{1-\alpha/2}cv(r)$.

Estimadores de Razão para Total e Média

Problema (Bolfarine e Bussab):

- Deseja-se estimar a quantidade de açúcar que pode ser extraída de um caminhão de laranja. As unidades populacionais são as laranjas e a variável de interesse y é:
  - Y_i: quantidade de açúcar na laranja i.
- Acontece que não se conhece o número de laranjas, N, do caminhão para utilizar o estimador $\widehat{Y} = N\overline{y}$.
- Por outro lado, sabe-se que a quantidade de açúcar está correlacionada com o tamanho ou peso da laranja. Então, a variável x definida como
  - X_i: peso da laranja i.
- Esta pode ser usada como variável auxiliar.
- Podemos, então definir a razão ou quantidade média de açúcar por unidade de peso como $R = Y/X$, ou seja, a quantidade média de açúcar por unidade de peso é o total de açúcar pelo peso total da carga de laranja.
- Ou ainda, $Y = RX$, o total de açúcar é o peso da carga vezes a quantidade média de açúcar por unidade de peso.

- Agora ficou fácil! Basta uma amostra de n laranjas para que possamos estimar R, ou seja, calcular

$$r = \frac{\overline{y}}{\overline{x}}, \quad \text{e, consequentemente,} \quad \widehat{Y}_R = rX,$$

  sendo X o peso total da carga de laranjas!
- Portanto, podemos definir um estimador de razão para o total de uma variável de interesse y a partir de uma variável auxiliar x como:

$$\widehat{Y}_R = rX,$$

  sendo
  - r é o estimador da razão entre y e x.
  - X é o total conhecido para a variável auxiliar x.
- Para a média temos:

$$\overline{y}_R = r\overline{X} = \frac{\widehat{Y}_R}{N}.$$

- As variâncias desses estimadores são dadas por:

$$\text{Var}(\widehat{Y}_R) = X^2\text{Var}(r) \quad \text{e} \quad \text{Var}(\overline{y}_R) = \overline{X}^2\text{Var}(r).$$

  No caso de AAS, tem-se:

$$S_R^2 \simeq S_y^2 + R^2 S_x^2 - 2RS_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2.$$

$$\text{Var}(\widehat{Y}_R) \simeq N \times \frac{N-n}{n} S_R^2 \quad \text{e} \quad \text{Var}(\overline{y}_R) \simeq \frac{N-n}{nN} S_R^2 = (1-f)\frac{S_R^2}{n}.$$

- Veja que as fórmulas são parecidas com as dos estimadores naturais da AAS, a menos do S^2.
- Pode-se estimar as variâncias desses estimadores por:

$$v(\widehat{Y}_R) = X^2 v(r) \quad \text{e} \quad v(\overline{y}_R) = \overline{X}^2 v(r).$$

- No caso de AAS tem-se:

$$s_R^2 \simeq s_y^2 + r^2 s_x^2 - 2rs_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2.$$

$$v(\widehat{Y}_R) \simeq N \times \frac{N-n}{n} s_R^2 \quad \text{e} \quad v(\overline{y}_R) \simeq \frac{N-n}{nN} s_R^2 = (1-f)\frac{s_R^2}{n}.$$

Exemplo

Foi feita uma pesquisa por AAS em 100 colégios de uma população formada por 468 colégios. A tabela mostra os resultados amostrais para as variáveis número de estudantes, y, e número de professores, x, dos colégios pesquisados. Sabe-se que o total de professores para o conjunto de escolas é 15.000.

n	$\sum_{i=1}^{468} y_i$	$\sum_{i=1}^{468} y_i^2$	$\sum_{i=1}^{468} x_i$	$\sum_{i=1}^{468} x_i^2$	$\sum_{i=1}^{468} y_i x_i$
100	44988	36248004	3099	144209	2160390

1. Estimar o total de alunos usando o estimador natural.
2. Estimar o total de alunos usando estimador de razão.
3. Comparar as estimativas das variâncias dos estimadores.

Comparação do estimador de razão com o natural

- Variância do estimador natural do total na AAS:

$$\mathrm{Var}_{AAS}(\widehat{Y}) = N^2 \frac{1-f}{n} S_y^2.$$

- Variância do estimador de razão do total:

$$\mathrm{Var}_{AAS}(\widehat{Y}_R) = N^2 \frac{1-f}{n} S_R^2.$$

- Então, para que o estimador de razão seja melhor devemos ter:

$$\mathrm{Var}_{AAS}(\widehat{Y}_R) < \mathrm{Var}_{AAS}(\widehat{Y}) \Leftrightarrow \rho(x,y) > \frac{1}{2}\frac{\mathrm{CV}_x}{\mathrm{CV}_y},$$

utilizando-se o fato que

$$S_R^2 < S_y^2 \quad \Rightarrow \quad S_y^2 + R^2 S_x^2 - 2\rho(x,y) R S_x S_y < S_y^2.$$

Estimador da Razão na Amostragem Estratificada

Pode-se definir dois estimadores de razão diferentes para o total (ou média) na Amostragem Estratificada:

- Estimador de razão combinada: $\widehat{Y}_{RC}$
  - Será baseado nos estimadores estratificados dos totais populacionais de y e x, ou seja, $\widehat{Y}_{ae}$ e $\widehat{X}_{ae}$;
- Estimador de razão separada: $\widehat{Y}_{RS}$
  - Será baseado nos estimadores dos totais de x e y dentro de cada um dos estratos, ou seja, nos $\widehat{Y}_h$ e $\widehat{X}_h$.

Vamos supor uma Amostra Estratificada com AAS dentro de cada estrato, denotando por AES.

Estimador de razão combinada:

$$\widehat{Y}_{RC} = \frac{\widehat{Y}_{ae}}{\widehat{X}_{ae}} X = \frac{\sum_{h=1}^{H} \widehat{Y}_h}{\sum_{h=1}^{H} \widehat{X}_h} X = \frac{\overline{y}_{ae}}{\overline{x}_{ae}} X = r_{RC} X.$$

- Veja que o estimador da razão combinada só depende do conhecimento do total populacional, X, da variável auxiliar x e não dos totais por estrato.
- O estimador da razão combinada é consistente.
- O estimador da razão combinada é tendencioso.
- Como vimos no caso de AAS, quando temos AES também podemos ter a tendenciosidade desprezível, pois pode-se demonstrar que:

$$\frac{|T(\widehat{Y}_{RC})|}{\sqrt{\text{Var}(\widehat{Y}_{RC})}} \leqslant \text{CV}(\widehat{X}_{ae})$$

- Na prática é usual considerar a tendenciosidade desprezível se:

$$\text{CV}(\widehat{X}_{ae}) = \text{CV}(\overline{x}_{ae}) \leqslant 0,1.$$

- Então, para que a tendenciosidade seja considerada desprezível é bom que se tenha uma estimativa da média (ou total) da variável auxiliar x precisa.
- Como anteriormente, fixa-se o coeficiente de variação da estimativa da média de x em, no máximo, 10%.
- Logo, vamos calcular um tamanho de amostra tal que:

$$n \geqslant \frac{\frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h^2 S_h^2(x)}{a_h}}{0,01\overline{X}^2} + \frac{1}{N^2}\sum_{h=1}^{H} N_h S_h^2(x),$$

sendo a_h um valor que depende da alocação de amostra utilizada.

- Se o tamanho da amostra for suficiente para que a tendenciosidade seja desprezível, uma fórmula para a variância aproximada será:

$$\text{Var}(\widehat{Y}_{RC}) \simeq N^2[\text{Var}(\overline{y}_{ae}) + R^2\text{Var}(\overline{x}_{ae}) - 2R\text{Cov}(\overline{x}_{ae}, \overline{y}_{ae})].$$

- Como temos AAS dentro de cada estrato:

$$\text{Var}(\widehat{Y}_{RC}) \simeq \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h}[S_h^2(y) + R^2 S_h^2(x) - 2RS_h(x,y)],$$

- Ou ainda,

$$\text{Var}(\widehat{Y}_{RC}) \simeq \sum_{h=1}^{H} \frac{N_h}{N_h - 1} \times \frac{N_h - n_h}{n_h} \sum_{j=1}^{N_h}[(Y_{hj} - \overline{Y}_h) - R(X_{hj} - \overline{X}_h)]^2.$$

A variância pode ser estimada por:

$$v(\widehat{Y}_{RC}) \simeq \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h}[s_h^2(y) + r_{RC}^2 s_h^2(x) - 2r_{RC}s_h(x,y)],$$

sendo

$$\begin{aligned} s_h^2(y) &= \frac{1}{n_h - 1}\sum_{j=1}^{n_h}(y_{hj} - \overline{y}_h)^2 \\ s_h^2(x) &= \frac{1}{n_h - 1}\sum_{j=1}^{n_h}(x_{hj} - \overline{x}_h)^2 \\ s_h^2(x,y) &= \frac{1}{n_h - 1}\sum_{j=1}^{n_h}(x_{hj} - \overline{x}_h)(y_{hj} - \overline{y}_h). \end{aligned}$$

Para a média temos as seguintes fórmulas:

$$\overline{y}_{RC} = \frac{\widehat{Y}_{RC}}{N}, \quad \text{Var}(\overline{y}_{RC}) = \frac{1}{N^2}\text{Var}(\widehat{Y}_{RC}), \quad \text{e} \quad v(\overline{y}_{RC}) = \frac{1}{N^2}v(\widehat{Y}_{RC}).$$

Exemplo

A tabela mostra informações de todas as fazendas de uma região estratificadas segundo sua área. A variável de interesse, y, é a área com plantação de milho e a variável auxiliar, x, é a área total de cada fazenda. A ideia é comparar a precisão de diversos estimadores para uma amostra de tamanho 100, sendo 70 fazendas selecionadas no estrato 1.

Estrato	Acres	N_h	$\overline{Y}$	$\overline{X}$	S_y^2	S_x^2	S_{xy}	R
1	<160	1580	19,40	82,56	312	2055	494	0,2350
2	>160	430	51,63	244,85	922	7357	858	0,2109
Total		2010	26,30	117,28	620	7619	1453	0,2243

Estimador de razão separada:

$$\widehat{Y}_{RS} = \sum_{h=1}^{H} \widehat{Y}_{hR} = \sum_{h=1}^{H} \frac{\overline{y}_h}{\overline{x}_h} X_h \sum_{h=1}^{H} r_h X_h.$$

- É baseado nos estimadores de razão dos totais de cada estrato.
- Para este estimador é preciso conhecer os valores dos totais da variável auxiliar x para cada um dos estratos.
- É um estimador consistente, pois usa um estimador de razão para os totais dos estratos que, por sua vez, é um estimador consistente.
- Quanto à tendenciosidade, é necessário que se tenha tamanhos de amostras nos estratos, n_h, suficientemente grandes para que a tendenciosidade de cada $\widehat{Y}_{hR}$ seja desprezível.
- Se cada estrato tiver aproximadamente o mesmo nível de tendência, pode-se admitir que a tendenciosidade do estimador do total seja H vezes a tendenciosidade do estimador dos totais de cada estrato, ou seja:

$$\frac{|T(\overline{Y}_{RS})|}{\sqrt{\text{Var}(\overline{Y}_{RS})}} \leqslant \sqrt{H}\text{CV}(\overline{X}_h)$$

- Na prática, pode-se admitir:

$$\frac{|T(\overline{Y}_{RS})|}{\sqrt{\text{Var}(\overline{Y}_{RS})}} \leqslant \sqrt{H}\text{CV}(\overline{X}_h) \leqslant 0,20.$$

- A variância aproximada do estimador é dada por:

$$\text{Var}(\widehat{Y}_{RS}) \simeq \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h}[S_h^2(y) + R_h^2 S_h^2(x) - 2R_h S_h(x,y)],$$

- Ou ainda,

$$\text{Var}(\widehat{Y}_{RS}) \simeq \sum_{h=1}^{H} \frac{N_h}{N_h - 1} \times \frac{N_h - n_h}{n_h} \sum_{j=1}^{N_h}[(Y_{hj} - \overline{Y}_h) - R_h(X_{hj} - \overline{X}_h)]^2.$$

- A variância pode ser estimada por:

$$v(\widehat{Y}_{RS}) \simeq \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h}[s_h^2(y) + r_h^2 s_h^2(x) - 2r_h s_h(x,y)].$$

Comparação dos estimadores de razão combinada e separada

- ✓ Os dois estimadores são equivalentes quando $R_h = R$ para todos os estratos (pelo menos aproximadamente).
- ✓ Caso exista uma grande diferença entre os R_h será aconselhável o uso do estimador da razão separada.
- ✓ Intuitivamente, nestes casos, é preferível usar a razão separada, já que é baseado em informações mais detalhadas da variável auxiliar x.

Estimador de Regressão

- Outro tipo de estimador baseado em informações auxiliares.
- Vamos ver o caso de regressão linear simples, mas a teoria pode ser utilizada com regressão múltipla utilizando um conjunto de variáveis auxiliares.
- É também baseado em informações de uma variável auxiliar x que tenha uma correlação com a variável de interesse y.
- Muitas vezes a relação linear (ou correlação) entre y e x existe, porém a reta não passa pela origem.
- Nestes casos o estimador de regressão é preferível ao de razão.

- Vamos supor que a relação entre y e x, na população P_N pode ser expressa pela equação de regressão linear do tipo:

$$Y_i = \alpha + \beta X_i, \quad i = 1, 2, \ldots, N.$$

- De posse de uma amostra de tamanho n dessa população, pode-se definir um estimador de regressão para a média de y como:

$$\overline{y}_{reg} = \overline{y} + B(\overline{X} - \overline{x}).$$

- É fácil ver que é um estimador consistente, pois se $n = N$ temos que $\overline{x} = \overline{X}$ e $\overline{y} = \overline{Y}$ o que zera a segunda parcela da equação e faz o valor do estimador ser exatamente igual à média de y.

Estimador de regressão com B conhecido

- Seja o caso em que $B = b_0$ conhecido:

$$\overline{y}_{reg} = \overline{y} + b_0(\overline{X} - \overline{x}).$$

- Esse estimador é não tendencioso, e sua variância é dada por:

$$\text{Var}(\overline{y}_{reg}) = \frac{1-f}{n}[S_y^2 + b_0^2 S_x^2 - 2b_0 S_{xy}].$$

- O valor de b_0 que minimiza essa variância é:

$$b_0 = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{N}(X_i - \overline{X})^2}.$$

Estimador de regressão com B desconhecido

- Quando não se conhece o valor de B, é necessário estimá-lo a partir da amostra.
- O valor de b, estimador do parâmetro B, é calculado por:

$$b = \frac{s(x,y)}{s^(x)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

- b é um estimador tendencioso de B, e sua tendênciosidade, pode ser demonstrada, é da ordem de $1/n$.

Sua variância pode ser calculada por:

$$\begin{aligned}
\text{Var}(\overline{y}_{reg}) &\simeq \frac{1-f}{n} S^2(y)[1-\rho^2(x,y)] = \frac{1-f}{n} S^2_{reg} \\
&= \frac{1-f}{n}[S^2(y) + B^2 S^2(x) - 2BS(x,y)] \\
&= \frac{1-f}{n}[S^2(y) - B^2 S^2(x)].
\end{aligned}$$

E estimada por

$$\begin{aligned}
v(\overline{y}_{reg}) &\simeq \frac{1-f}{n} \times \frac{n-1}{n-2}[s^2(y) + b^2 s^2(x) - 2bs(x,y)] \\
&= \frac{1-f}{n} \times \frac{n-1}{n-2}[s^2(y) - b^2 s^2(x)] \\
&= \frac{1-f}{n} \times \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i - \overline{y})^2 - b^2 \sum_{i=1}^{n}(x_i - \overline{x})^2\right].
\end{aligned}$$

"O divisor $n-2$, no lugar de $n-1$, é sugerido porque ele é usado na teoria padrão de regressão e é sabido que dá uma estimativa não tendenciosa de S^2_{reg} se a população é infinita e a regressão é linear."

Cochran

Para o total, temos

$$\widehat{Y}_{reg} = N\overline{Y}_{reg}, \quad \text{e} \quad \text{Var}(\widehat{Y}_{reg}) = N^2\text{Var}(\overline{Y}_{reg}).$$

Comparação entre o estimador natural, o de razão e o de regressão na amostra aleatória simples:

$$\begin{aligned}
\text{Var}_{AAS}(\overline{y}) &= \frac{1-f}{n} S_y^2 \\
\text{Var}_{AAS}(\overline{y}_R) &\simeq \frac{1-f}{n} [S_y^2 + R^2 S_x^2 - 2RS_{xy}] \\
\text{Var}_{AAS}(\overline{y}_{reg}) &\simeq \frac{1-f}{n} S_y^2 [1 - \rho^2(x,y)].
\end{aligned}$$

- O estimador de regressão será equivalente ao natural somente se $\rho(x,y) = 0$.
- O estimador de regressão será melhor que o de razão a menos que a reta de regressão passe pela origem.

Exemplo

Em uma região com 256 pomares de pêssegos deseja-se verificar qual a eficiência relativa dos estimadores do total produzido de pêssegos para uma amostra aleatória simples de 100 pomares, utilizando-se como variável auxiliar o número de pés de pêssego em cada pomar.

São registradas as seguintes informações de um ano anterior sobre a produção da região:

$$\begin{aligned} S_y^2 &= 6409 \\ S_x^2 &= 3898 \\ S_{xy} &= 4434 \\ R &= 1,270. \end{aligned}$$

Pós-estratificação

- Outra maneira de utilizar variáveis auxiliares na estimação é a pós-estratificação.
- Usada quando não é possível selecionar uma amostra estratificada pela falta de um cadastro que possibilite alocar antecipadamente cada unidade ao seu estrato.
- Quando se conhece os limites dos estratos e seus tamanhos (por uma pesquisa anterior) é possível usar a ideia da estratificação na estimação dos parâmetros através da estratificação da amostra após a obtenção das informações da mesma (pós-estratificação).

Passo-a-passo da pós-estratificação

1. Selecionar uma AAS de tamanho n de P_N.
2. Para cada unidade selecionada observa-se, além da variável de interesse, o valor da variável (auxiliar) de estratificação x.
3. De acordo com os valores de x distribui-se as unidades pesquisadas em H estratos (pós-estratos) previamente definidos.
4. Considera-se a parte da amostra em cada pós estrato como uma AAS de tamanho n_h de modo que $n_1 + n_2 + \cdots + n_H = n$, sendo os n_h variáveis aleatórias **(pois dependem da amostra selecionada)**.

Estimadores para total e média populacionais:

$$\begin{aligned}
\widehat{Y}_{\text{pós}} &= \sum_{h=1}^{H} N_h \overline{y}_h = \sum_{h=1}^{H} N_h \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \\
\overline{y}_{\text{pós}} &= \sum_{h=1}^{H} W_h \overline{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \times \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{\widehat{Y}_{\text{pós}}}{N}.
\end{aligned}$$

- Esses estimadores são, do ponto de vista das fórmulas, idênticos aos estimadores da amostra estratificada.
- A diferença está na variância dos estimadores.

- Esses estimadores são não tendenciosos.
- Para mostrar isso é preciso calcular a esperança condicional do estimador dados os valores dos n_h, sabendo que

$$\mathrm{E}(X) = \mathrm{E}_Y(\mathrm{E}(X|Y)).$$

- Temos que

$$\mathrm{E}(\overline{y}_h) = \mathrm{E}_{n_1,n_2,\ldots,n_H}(\mathrm{E}(\overline{y}_h|n_1, n_2, \ldots, n_H)) = \overline{Y}_h.$$

- Portanto,

$$\mathrm{E}(\overline{y}_{\text{pós}}) = \sum_{h=1}^{H} W_h \mathrm{E}(\overline{y}_h) = \sum_{h=1}^{H} \frac{N_h}{N} \overline{Y}_h = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{Y}_h = \overline{Y}.$$

Variância dos estimadores de pós-estratificação

- Como os n_h são desconhecidos a priori, na verdade eles são variáveis aleatórias cujos valores só são conhecidos após a investigação da amostra..
- Temos que

$$\mathrm{Var}(\overline{y}_{\text{pós}}) = \sum_{h=1}^{H} W_h^2 \mathrm{E}\left(\frac{1}{n_h}\right) S_h^2 - \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{N_h}.$$

- Aproximando o valor esperado de $1/n_h$, obtemos:

$$\begin{aligned}
\mathrm{Var}(\overline{y}_{\text{pós}}) &= \frac{N-n}{Nn}\sum_{h=1}^{H} W_h S_h^2 + \frac{N-n}{Nn^2}\sum_{h=1}^{H}(1-W_h)S_h^2 \\
&= \frac{1-f}{n}\sum_{h=1}^{H} W_h S_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{H}(1-W_h)S_h^2 \\
&= \mathrm{Var}(\overline{y}_{prop}) + \frac{1-f}{n^2}\sum_{h=1}^{H}(1-W_h)S_h^2.
\end{aligned}$$

- Vê-se que a variância é igual a variância de uma amostra estratificada com alocação proporcional, mais uma correção devido a perda da proporcionalidade pela aleatoriedade da seleção de unidades pertencentes aos diversos estratos.
- Quando o tamanho da amostra, n, cresce, o segundo membro do lado direito da equação tende rapidamente a zero, então:

$$n \to N \Rightarrow \text{Var}(\overline{y}_{\text{pós}}) \to \text{Var}(\overline{y}_{prop}).$$

- Ou seja, para amostras grandes, a eficiência da pós-estratificação, em relação à amostragem aleatória simples, é semelhante a da amostragem aleatória estratificada com alocação proporcional.
- Na prática é desejável que se tenha $n_h \geqslant 20$, para uma boa performance da pós-estratificação.

Exemplo

Uma grande empresa sabe que 40% de suas contas a receber são das vendas por atacado e 60% do varejo. Todavia para identificar individualmente cada conta sem verificar o arquivo exaustivamente é muito difícil. Um consultor retira uma AAS de 100 contas com a finalidade de estimar o valor médio das contas a receber. A amostra selecionada tem 70 contas de vendas por atacado e 30 de vendas no varejo. Os dados separados por atacado e varejo estão na tabela abaixo:

	Atacado	Varejo
n_h	70	30
$\overline{y}$	520	280
s_h^2	210	90

Estimar o valor médio das contas a receber e a margem de erro.

Amostragem de Conglomerados

- Conglomerado: grupo de unidades populacionais.
- Conglomerados naturais: cacho de uvas, penca de bananas, domicílio, edifício, etc.
- Conglomerados artificiais: grupo de cinco pessoas, lote de cem parafusos, grupos compostos por cada cinco casas, etc.
- Amostragem de conglomerados: o interesse é medir alguma variável em uma amostra de unidades da população, porém, por razões operacionais (falta e cadastro, custo, etc), é selecionada uma amostra de conglomerados dessas unidades populacionais. A unidade amostral primária é o conglomerado e não a unidade populacional de interesse (onde será medida a variável de interesse).

Exemplos de Amostragem de Conglomerados

População	**Variáveis de interesse**	**Unidade de referência**	**Conglomerado**
Turmas de alunos	Alunos por turma	Turma	Escolas
Estudantes de 2^{o} grau	Aproveitamento	Estudante	Turmas
Visitantes de parques nacionais	Facilidades do parque	Visitante de parque nacional	Veículos que entram no parque
Passageiros de avião	Propósito da viagem	Passageiro de avião	Vôos
Domícilios	Características de domícilio	Domicílio	Setores

Como Fazer Amostragem de Conglomerados

- Os esquemas amostrais já estudados, AAS, AE, AS, amostragem com probabilidades desiguais, serão também usados para selecionar amostras de conglomerados.
- Amostra de conglomerados versus amostra de unidades elementares:
  - Com conglomerados o custo é mais baixo.
  - A variância amostral geralmente maior, dependendo da homogeneidade dos elementos que formam os conglomerados.
- Geralmente a opção pelo uso de conglomerados se justifica pelo menor custo e outras vantagens operacionais.