Capítulo 6 Estimação de Razões e Funções de Totais

6.1 Razões populacionais como parâmetros de interesse

Muitas vezes na prática de pesquisas por amostragem, o *parâmetro* populacional de interesse é uma razão entre os totais (ou médias) de duas variáveis de pesquisa definidas para cada unidade da população. Nesse caso, o *parâmetro* é denominado *razão de totais (ou de médias)* ou simplesmente *razão*. Chamando de y e x as variáveis cujos totais aparecem no numerador e denominador da razão, respectivamente, define-se a *razão* R como:

$$R = \frac{\sum_{i \in U} y_i}{\sum_{i \in U} x_i} = \frac{Y}{X} = \frac{\overline{Y}}{\overline{X}} \tag{6.1}$$

Alguns exemplos de razões de interesse incluem:

- a razão entre o total dos gastos com transporte e o total da renda das famílias;
- a produtividade da lavoura de certo produto, definida como a razão entre o total da quantidade colhida do produto pelo total da área plantada com esse produto;
- o salário médio dos trabalhadores da indústria, definido como a razão entre o total dos salários dos trabalhadores da indústria e o número total de trabalhadores da indústria; e
- a densidade de pessoas por dormitório em domicílios particulares, definida como a razão entre o número de pessoas residentes em domicílios particulares e o número total de dormitórios em domicílios particulares.

No caso da razão entre o total dos gastos com transporte (y) e o total da renda das famílias (x), a razão fica definida como:

$$\frac{\text{Total dos gastos com transporte}}{\text{Total da renda}} = \frac{\sum_{i \in U} y_i}{\sum_{i \in U} x_i} = \frac{Y}{X} = R$$

6.2 Razão de médias versus média de razões

Um cuidado importante aqui é compreender o significado do parâmetro de interesse. A razão de totais R não é a média $\left(\overline{R} = \sum_{i \in U} r_i/N\right)$ das *razões por unidade* $(r_i = y_i/x_i)$. Neste capítulo estamos interessados na razão de totais (ou médias), e não na média de razões. A média de razões $\left(\overline{R}\right)$ pode ser estimada usando os estimadores habituais de médias populacionais discutidos no capítulo 3, aplicados à variável r definida aqui.

Exemplo 6.1 Considere as densidades de habitantes por área (em km^2) das unidades da federação brasileiras apresentadas na Tabela 6.1, conforme o Censo 2010. Neste exemplo, vemos como uma discrepância grande pode ocorrer entre as duas quantidades (a razão de médias e a médias das razões), quando os valores de r_i são muito dispersos.

Tabela 6.1: Densidade demográfica por Unidade da Federação, média das razões e razão das médias para o Brasil

Unidade da Federação	Densidade demográfica (Hab/km^2)
Rondônia	6,6
Acre	4,5
Amazonas	2,2
Roraima	2,0
Pará	6,1
Amapá	4,7
Tocantins	5,0
Maranhão	19,8
Piauí	12,4
Ceará	56,8
Rio Grande do Norte	60,0
Paraíba	66,7
Pernambuco	89,6
Alagoas	112,3
Sergipe	94,4
Bahia	24,8
Minas Gerais	33,4
Espírito Santo	76,3
Rio de Janeiro	365,2
São Paulo	166,3
Paraná	52,4
Santa Catarina	65,3
Rio Grande do Sul	39,8
Mato Grosso do Sul	6,9

Unidade da Federação	**Densidade demográfica (Hab/km^2)**
Mato Grosso	3,4
Goiás	17,7
Distrito Federal	444,1
Brasil (Média das razões)	68,1
Brasil (Razão das médias)	22,4

Fonte: IBGE, Censo Demográfico 2010.

Os dados apurados nas últimas duas linhas da Tabela 6.1 ilustram bem a importância de identificar corretamente que parâmetro se deseja estimar. No exemplo aqui considerado, a densidade demográfica apurada no nível do País como um todo é 22,4 Hab/km^2, enquando a média das densidades demográficas calculadas por unidades da federação é de 68,1 Hab/km^2, mais que três vezes maior. Portanto, ao enfrentar uma situação em que o parâmetro de interesse pode ser caracterizado como uma das duas situações, cabe ao responsável por planejar a pesquisa esclarecer junto aos demandantes da mesma qual dos dois conceitos melhor descreve a quantidade de interesse. Isto permitirá selecionar adequadamente o estimador que deve ser empregado com os dados da amostra a ser coletada.

6.3 Outras funções de totais

Totais são parâmetros muito importantes, pois muitos outros parâmetros de interesse podem ser obtidos como funções de totais. Já vimos no capítulo 5 o exemplo das contagens de unidades pertencentes a determinados grupos que possuem características comuns, no caso em que as variáveis cujos totais se quer estimar são indicadoras da pertinência a tais grupos. No capítulo 3 também verificamos que as médias populacionais são razões de totais, onde no denominador aparece o número N de unidades na população, e no numerador vai o total da variável cuja média se quer definir.

Neste capítulo já definimos as *razões de totais* populacionais, e agora vamos apresentar mais alguns parâmetros que podem ser definidos como funções de totais populacionais. Para introduzir o caso mais geral, considere um parâmetro θ que pode ser definido como uma função de K totais populacionais, isto é:

$$\theta = f(Y_1, Y_2, \ldots, Y_K)$$

onde Y_k é o total populacional da variável de pesquisa y_k, para $k = 1, \ldots, K$, e f é uma função real que especifica como obter o valor de θ a partir dos totais das variáveis de pesquisa.

Um caso geral de interesse é o das funções lineares de totais populacionais, dadas por:

$$\theta = f(Y_1, Y_2, \ldots, Y_K) = \sum_{k=1}^{K} a_k \, Y_k \tag{6.2}$$

onde a_k são constantes conhecidas, para $k = 1, \ldots, K$.

Um exemplo simples de função linear é a *diferença* entre dois totais populacionais especificados, dada por:

$$D = \sum_{i \in U} y_{2i} - \sum_{i \in U} y_{1i} = Y_2 - Y_1$$

Nesse caso, temos uma situação envolvendo $K = 2$ totais populacionais, com $a_1 = 1$ e $a_2 = -1$.

Outro exemplo de parâmetro que pode ser escrito como função de totais é o caso da *função de distribuição cumulativa* num ponto, dada por:

$$F_y(a) = \frac{1}{N} \sum_{i \in U} I(y_i \leq a)$$

Nesse caso, a situação envolve apenas dois totais populacionais ($K = 2$), sendo a primeira definida como $y_{1i} \equiv I(y_i \leq a)$ e a segunda $y_{2i} \equiv 1$, e a função definida como $f(Y_1, Y_2) = Y_1 / Y_2$.

A *variância* de uma variável de pesquisa y, dada por

$$S_y^2 = \frac{1}{N-1} \left[\sum_{i \in U} {y_i}^2 - N \overline{Y}^2 \right] = \frac{1}{N-1} \left[\sum_{i \in U} {y_i}^2 - Y^2 / N \right]$$

também pode ser vista como uma função envolvendo três totais populacionais $(K = 3)$, sendo o primeiro definido como total da variável $y_{1i} \equiv y_i^2$, a segunda $y_{2i} \equiv y_i$, e a terceira como $y_{3i} \equiv 1$. A função que retorna a variância S_y^2 é definida como $f(Y_1, Y_2, Y_3) = \frac{1}{Y_3 - 1} \left[Y_1 - (Y_2^2 / Y_3) \right]$.

De maneira similar, a *covariância* e a *correlação* das variáveis y e z podem ser definidas como:

$$S_{y,z} = \frac{1}{N-1}\sum_{i\in U}\left(y_i - \overline{Y}\right)\left(z_i - \overline{Z}\right) = \frac{1}{N-1}\left[\sum_{i\in U} y_i z_i - YZ/N\right]$$

e

$$\rho_{y,z} = \frac{S_{y,z}}{S_y S_z}.$$

A covariância pode ser vista como uma função envolvendo quatro totais populacionais $(K = 4)$, sendo o primeiro definido como o total Y_1 da variável $y_{1i} \equiv y$, o segundo o total Y_2 da variável $y_{2i} \equiv z$, o terceiro o total da variável definida como $y_{3i} \equiv y_i z_i$ e o quarto o total da variável definida como $y_{4i} \equiv 1$. Por simplicidade, a função que define a correlação foi definida em função da covariância e dos desvios-padrão das duas variáveis de interesse, mas também poderia ser escrita como função de totais como já ilustrado nos demais casos.

O ponto importante a destacar nesta seção é o fato de que muitos parâmetros de interesse podem ser vistos como funções de totais populacionais, e com isso, podem também ser estimados aplicando as funções que definem os parâmetros a estimadores dos totais correspondentes. Isto corresponde essencialmente a uma espécie de *método dos momentos* para estimar parâmetros populacionais, onde cada total populacional desconhecido é substituído na expressão do parâmetro por um estimador amostral não viciado do total correspondente.

Essa ideia pode ser aplicada de maneira bastante geral para obter estimadores para grande quantidade de parâmetros populacionais que possam ser escritos em função de totais populacionais de variáveis bem definidas. Para fixar as ideias, este método é aplicado primeiramente para obter estimador para uma *razão de totais*.

6.4 Estimando razões (de totais)

Para estimar *razões de totais* (ou médias) como R, o estimador "natural" é a *razão dos estimadores Horwitz-Thompson de total*:

$$\widehat{R} = \frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i x_i} = \frac{\widehat{Y}_{HT}}{\widehat{X}_{HT}} \tag{6.3}$$

Notas

1. Tanto o numerador como o denominador do estimador $\widehat{R}$ da razão R podem variar com a amostra selecionada, s.

2. Apesar de termos estimadores não viciados para os totais populacionais $\widehat{Y}_{HT}$ e $\widehat{X}_{HT}$, em geral, $E(Z/W) \neq E(Z)E(W)$, e portanto

$$E(\widehat{R}) = E(\widehat{Y}_{HT}/\widehat{X}_{HT}) \neq E(\widehat{Y}_{HT})/E(\widehat{X}_{HT}) = Y/X = R$$

3. Logo $\widehat{R}$ é um *estimador viciado* de R.

4. O vício de $\widehat{R}$ é pequeno e chamado de *vício técnico*, para distingui-lo de outros vícios potencialmente causados por problemas tais como erros de cobertura, não resposta e outros. Para amostras grandes (isto é, com n grande), este vício é geralmente desprezível do ponto de vista prático. Veja discussão na seção 6.6.

5. Sob AAS, como os pesos amostrais são constantes (e iguais a N/n para todas as unidades da amostra), o estimador $\widehat{R}$ simplifica para:

$$\widehat{R}_{AAS} = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} = \frac{\overline{y}}{\overline{x}} \tag{6.4}$$

6.5 Estimando outras funções de totais

Quando o parâmetro de interesse é uma função de totais populacionais $\theta = f(Y_1, Y_2, \ldots, Y_K)$, o método mais simples para estimar esse parâmetro é aplicar a mesma função a estimadores não viciados dos totais, isto é, usar o estimador:

$$\hat{\theta} = f(\widehat{Y_1}, \widehat{Y_2}, \ldots, \widehat{Y_K}) \tag{6.5}$$

onde $\widehat{Y_k}$ é o estimador HT do total populacional da variável de pesquisa y_k, para $k = 1, \ldots, K$.

Para o caso de *funções lineares*, o estimador resultante será dado por:

$$\hat{\theta} = \sum_{k=1}^{K} a_k \widehat{Y_k} \tag{6.6}$$

e será um estimador não viciado do parâmetro θ.

Neste caso, a variância do estimador também é relativamente simples de obter, e dada por:

$$\begin{aligned} V_p\left(\hat{\theta}\right) &= \sum_{k=1}^{K} \sum_{j=1}^{K} a_k\, a_j Cov_p\left(\widehat{Y_k}, \widehat{Y_j}\right) \\ &= \sum_{k=1}^{K} a_k^2\, V_p\left(\widehat{Y_k}\right) + \sum_{k=1}^{K} \sum_{j \neq k=1}^{K} a_k\, a_j Cov_p\left(\widehat{Y_k}, \widehat{Y_j}\right) \end{aligned} \tag{6.7}$$

e um estimador dessa variância pode ser facilmente construído usando

$$\begin{aligned}\widehat{V}_p\left(\hat{\theta}\right) &= \sum_{k=1}^{K}\sum_{j=1}^{K} a_k\, a_j \widehat{Cov}_p\left(\widehat{Y_k}, \widehat{Y_j}\right) \\ &= \sum_{k=1}^{K} a_k^2\, \widehat{V}_p\left(\widehat{Y_k}\right) + \sum_{k=1}^{K}\sum_{j\neq k=1}^{K} a_k\, a_j \widehat{Cov}_p\left(\widehat{Y_k}, \widehat{Y_j}\right) \end{aligned} \tag{6.8}$$

Um resultado bastante útil é mostrado a seguir. Quando $\theta = \sum_{k=1}^{K} a_k\, Y_k$, então sua variância pode ser escrita como:

$$\begin{aligned} V_p\left(\hat{\theta}\right) &= E_p\Big(\hat{\theta} - \theta\Big)^2 \\ &= E_p\left(\sum_{k=1}^{K} a_k \widehat{Y_k} - \sum_{k=1}^{K} a_k Y_k\right)^2 \\ &= E_p\left(\sum_{k=1}^{K} a_k \sum_{i\in s} \pi_i^{-1} y_{ki} - \sum_{k=1}^{K} a_k \sum_{i\in U} y_{ki}\right)^2 \\ &= E_p\left[\sum_{i\in s} \pi_i^{-1}\left(\sum_{k=1}^{K} a_k y_{ki}\right) - \sum_{i\in U}\left(\sum_{k=1}^{K} a_k y_{ki}\right)\right]^2 \\ &= E_p\left(\sum_{i\in s} \pi_i^{-1} z_i - \sum_{i\in U} z_i\right)^2 \\ &= E_p\Big(\widehat{Z}_{HT} - Z\Big)^2 \\ &= V_p\left(\widehat{Z}_{HT}\right) \end{aligned} \tag{6.9}$$

onde $z_i = \sum_{k=1}^{K} a_k y_{ki}$.

Verifica-se assim que a variância do estimador para θ pode ser obtida como a variância de um estimador de total para uma *variável derivada* z definida de maneira apropriada, tornando assim a obtenção da variância do estimador do parâmetro uma tarefa mais simples, já que a variância de estimadores de total foi apresentada no capítulo 3. Isso leva também à obtenção de estimadores da variância de forma simples, usando a mesma variável derivada:

$$\widehat{V}_p\left(\hat{\theta}\right) = \widehat{V}_{HT}(\widehat{Z}_{HT}) = \sum_{i\in s}\sum_{j\in s}\left(d_i d_j - d_{ij}\right) z_i z_j \tag{6.10}$$

Este resultado tem utilidade também quando θ é uma função não linear de totais, mas essa função é contínua e diferenciável. Neste caso, é possível usar a técnica de *linearização de Taylor* para obter a variância aproximada do estimador e também um estimador para essa

variância. A ideia dessa técnica é simples: aproximar o estimador não linear por uma quantidade linearizada, obtida a partir da expansão em Série de Taylor do estimador $\hat{\theta}$ ao redor do ponto θ. A expansão é dada por:

$$\begin{aligned}\hat{\theta} &= f(\widehat{Y}_1, \widehat{Y}_2, \ldots, \widehat{Y}_K) \\ &= f(Y_1, Y_2, \ldots, Y_K) + \sum_{k=1}^{K} \left[\frac{\partial f(Y_1, Y_2, \ldots, Y_K)}{\partial Y_k}\right] \left(\widehat{Y}_k - Y_k\right) + \\ &\quad \sum_{q=2}^{\inf} \frac{1}{q!} \sum_{k=1}^{K} \left[\frac{\partial^q f(Y_1, Y_2, \ldots, Y_K)}{\partial Y_k^q}\right] \left(\widehat{Y}_k - Y_k\right)^q \\ &\doteq f(Y_1, Y_2, \ldots, Y_K) + \sum_{k=1}^{K} \left[\frac{\partial f(Y_1, Y_2, \ldots, Y_K)}{\partial Y_k}\right] \left(\widehat{Y}_k - Y_k\right) \qquad (6.11)\end{aligned}$$

onde a aproximação da última linha advém da exclusão de todos os termos de ordem igual ou superior a 2.

É comum chamar de $\hat{\theta}_L$ a quantidade

$$\begin{aligned}\hat{\theta}_L &= f(Y_1, Y_2, \ldots, Y_K) + \sum_{k=1}^{K} \left[\frac{\partial f(Y_1, Y_2, \ldots, Y_K)}{\partial Y_k}\right] \left(\widehat{Y}_k - Y_k\right) \\ &= \theta + \sum_{k=1}^{K} a_k \left(\widehat{Y}_k - Y_k\right)\end{aligned}$$

onde agora $a_k = \dfrac{\partial f(Y_1, Y_2, \ldots, Y_K)}{\partial Y_k}$, $k = 1, 2, \ldots, K$.

Assim sendo, a obtenção da variância aproximada para o estimador não linear $\hat{\theta}$ é feita calculando-se a variância do *estimador linearizado* $\hat{\theta}_L$, que é dada por:

$$\widehat{V}_p\left(\hat{\theta}\right) \doteq \widehat{V}_p\left(\hat{\theta}_L\right) = \widehat{V}_{HT}(\widehat{Z}_{HT}) = \sum_{i \in s} \sum_{j \in s} \left(d_i d_j - d_{ij}\right) z_i z_j \qquad (6.12)$$

onde $z_i = \displaystyle\sum_{k=1}^{K} \frac{\partial f(Y_1, Y_2, \ldots, Y_K)}{\partial Y_k}\, y_{ki}$.

6.6 Analisando o vício do estimador da razão

Para analisar o vício do estimador da razão, note que:

$$\begin{aligned} Cov_p\left(\widehat{R}, \widehat{X}\right) &= E_p\left(\widehat{R} \times \widehat{X}\right) - E_p\left(\widehat{R}\right) \times E_p\left(\widehat{X}\right) \\ &= E_p\left(\frac{\widehat{Y}}{\widehat{X}} \times \widehat{X}\right) - E_p\left(\widehat{R}\right) \times E_p\left(\widehat{X}\right) \\ &= E_p\left(\widehat{Y}\right) - E_p\left(\widehat{R}\right) \times E_p\left(\widehat{X}\right) \\ &= Y - E_p\left(\widehat{R}\right) \times X \end{aligned}$$

Segue-se então que:

$$\begin{aligned} E_p\left(\widehat{R}\right) &= \frac{Y}{X} - \frac{1}{X} \times Cov_p\left(\widehat{R}, \widehat{X}\right) \\ &= R - \frac{1}{X} \times Cov_p\left(\widehat{R}, \widehat{X}\right) \end{aligned}$$

Chamando agora de $B_p(\widehat{R}) = E_p\left(\widehat{R}\right) - R$, o vício do estimador $\widehat{R}$, mostrou-se que este pode ser obtido como:

$$B_p(\widehat{R}) = -\frac{1}{X} \times Cov_p\left(\widehat{R}, \widehat{X}\right) \tag{6.13}$$

Esta expressão fornece um caminho para estabelecer um limite superior para o valor absoluto do vício dividido pelo desvio padrão do estimador da razão. Notando que $Cov_p\left(\widehat{R}, \widehat{X}\right) = Corr_p\left(\widehat{R}, \widehat{X}\right) \times DP_p(\widehat{R}) \times DP_p(\widehat{X})$, segue-se que:

$$\begin{aligned} \left|\frac{B_p(\widehat{R})}{DP_p(\widehat{R})}\right| &= \frac{1}{X} \times \left|\frac{Corr_p\left(\widehat{R}, \widehat{X}\right) \times DP_p(\widehat{R}) \times DP_p(\widehat{X})}{DP_p(\widehat{R})}\right| \\ &= \left|Corr_p\left(\widehat{R}, \widehat{X}\right)\right| \times \frac{DP_p(\widehat{X})}{X} \\ &\leq CV_p(\widehat{X}) \end{aligned} \tag{6.14}$$

Assim, quando o tamanho da amostra for grande o suficiente para tornar o CV do estimador do total $\widehat{X}$ no denominador da razão $\widehat{R}$ pequeno (digamos, menor que 0,1 ou 10%), então o vício do estimador da razão será pequeno quando comparado com o desvio padrão desse estimador. Vícios desse tipo são geralmente desprezados na prática, a menos que se tenha amostra *muito pequena*.

6.7 Erro Quadrático Médio (EQM) de $\widehat{R}$

Como $\widehat{R}$ tem um pequeno vício técnico, a avaliação de sua precisão deve ser feita considerando seu *Erro Quadrático Médio (EQM)*. Entretanto, sempre que a amostra for grande o suficiente (o que pode ser avaliado calculando o coeficiente de variação da estimativa de total do denominador da razão), o EQM poderá ser bem aproximado pela variância, como indicado:

$$EQM_p(\widehat{R}) = V_p(\widehat{R}) + \left[B_p(\widehat{R})\right]^2 \doteq V_p(\widehat{R}) \tag{6.15}$$

já que o termo $\left[B_p(\widehat{R})\right]^2$ fica menor que 1% da variância $V_p(\widehat{R})$ sempre que $CV_p(\widehat{X}) \leq 0,1$.

Usando ainda o resultado da seção 6.5 quanto à linearização de estimadores que podem ser escritos como funções não lineares de totais, segue-se que podemos aproximar a variância do estimador $\widehat{R}$ pela variância do estimador linearizado, onde a razão depende dos totais populacionais de duas variàveis ($K = 2$), sendo o primeira definida como $y_{1i} \equiv y_i$ e a segunda $y_{2i} \equiv x_i$, com a função definida como $f(Y_1, Y_2) = Y_1/Y_2$.

Nesse caso, a variável linearizada z_i pode ser obtida notando que:

$$\frac{\partial f(Y_1, Y_2)}{\partial Y_1} = \frac{1}{Y_2} = \frac{1}{X}$$

e

$$\frac{\partial f(Y_1, Y_2)}{\partial Y_2} = \frac{-Y_1}{Y_2^2} = \frac{-Y}{X^2}$$

levando a

$$z_i = \frac{1}{X} y_i - \frac{Y}{X^2} x_i = \frac{1}{X} \left(y_i - R x_i\right).$$

Sendo assim, a variância aproximada do estimador $\widehat{R}$ pode ser obtida calculando a variância do estimador de total da variável linearizada z_i:

$$V_p(\widehat{R}) \doteq V_p(\widehat{Z}_{HT}) = \sum_{i \in U} \sum_{j \in U} \left(\frac{d_i d_j}{d_{ij}} - 1\right) z_i z_j \tag{6.16}$$

e um estimador dessa variância aproximada é dado por:

$$\widehat{V}_p(\widehat{R}) = \widehat{V}_p(\widehat{Z}_{HT}) = \sum_{i \in s} \sum_{j \in s} \left(d_i d_j - d_{ij}\right) z_i z_j \tag{6.17}$$

Sob AAS, as expressões acima simplificam para as expressões a seguir. Para mais detalhes, veja Cochran (1977). A variância aproximada do estimador da razão é dada por:

$$V_{AAS}(\widehat{R}) \doteq \frac{1}{\overline{X}^2}\left(\frac{1}{n}-\frac{1}{N}\right) \times \frac{1}{N-1}\sum_{i\in U}(y_i - Rx_i)^2 \tag{6.18}$$

e o estimador da variância fica igual a:

$$\widehat{V}_{AAS}(\widehat{R}) = \frac{1}{\overline{x}^2}\left(\frac{1}{n}-\frac{1}{N}\right) \times \frac{1}{n-1}\sum_{i\in s}\left(y_i - \widehat{R}x_i\right)^2 \tag{6.19}$$

A Tabela 6.2 apresenta um resumo da estimação de parâmetros de uma razão sob AAS.

Tabela 6.2: Estimadores dos parâmetros de uma razão sob AAS

Parâmetro	Estimador sob AAS
$R = \frac{Y}{X} = \frac{\overline{Y}}{\overline{X}}$	$\widehat{R} = \frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i} = \frac{\overline{y}}{\overline{x}}$
$V_{AAS}(\widehat{R}) \doteq \frac{1}{\overline{X}^2}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N-1}\sum_{i\in U}(y_i - Rx_i)^2$	$\widehat{V}_{AAS}(\widehat{R}) = \frac{1}{\overline{x}^2}\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\sum_{i\in s}\left(y_i -$

Exemplo 6.2 Estimando Razões e Médias de Razões

Considere a população de municípios brasileiros fornecida no arquivo MunicBR_dat.

1. Selecione uma AAS de $n = 200$ municípios, e use esta amostra para estimar os seguintes parâmetros populacionais:

    a. Densidade demográfica média por km^2 no Brasil;

    b. Média da variável densidade demográfica por km^2 por município.

2. Para cada uma das estimativas acima:

    a. Estime o erro padrão e o coeficiente de variação;

    b. Compare com os correspondentes parâmetros populacionais.

3. Calcule o tamanho da amostra que seria necessário para a estimar densidade demográfica média por km^2 no Brasil com erro máximo de 10 $Hab/km^2/$ com nível de confiança de 95%.

4. Selecione uma amostra do tamanho calculado em 3 e estime o parâmetro e sua margem de erro, considerando coeficiente de confiança de 95%.

Solução dos itens 1. e 2. do Exemplo 6.2 , usando R:

```
# Limpa área de trabalho
rm(list = ls())

# Carrega biblioteca(s) requerida(s)
library(sampling)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Leitura dos dados
MunicBR_dat <- readRDS(file="MunicBR_dat.rds")
str(MunicBR_dat)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    5570 obs. of  6 variables:
##  $ CodMunic : chr  "1100015" "1100023" "1100031" "1100049" ...
##  $ SiglaUF  : chr  "RO" "RO" "RO" "RO" ...
##  $ CodUF    : chr  "11" "11" "11" "11" ...
##  $ Pop      : num  25728 101269 6495 85863 18041 ...
##  $ Area     : num  7067 4427 1314 3793 2783 ...
##  $ Densidade: num  3.64 22.88 4.94 22.64 6.48 ...
```

```r
# Define semente para geração de números aleatórios para permitir repetição
set.seed(123)

# Item 1
# Tamanho da amostra
(n <- 200)
```

```
## [1] 200
```

```r
# Tamanho da população
(N <- nrow(MunicBR_dat))
```

```
## [1] 5570
```

```r
# Seleciona amostra AAS dos municípios
munic_amo <- getdata(MunicBR_dat, srswor(n,N))
str(munic_amo)
```

```
## 'data.frame':    200 obs. of  7 variables:
##  $ ID_unit  : int  4 57 137 234 253 254 332 416 496 513 ...
##  $ CodMunic : chr  "1100049" "1200179" "1400027" "1505205" ...
##  $ SiglaUF  : chr  "RO" "AC" "RR" "PA" ...
##  $ CodUF    : chr  "11" "12" "14" "15" ...
##  $ Pop      : num  85863 9836 10432 30088 17774 ...
##  $ Area     : num  3793 1703 28472 3852 4115 ...
##  $ Densidade: num  22.638 5.777 0.366 7.81 4.32 ...
```

```r
# Soluções para item 1
# a. Estima densidade demográfica média por km2 no Brasil
(r_chapeu <- munic_amo %>%
             summarise(Popm = mean(Pop),
                       Aream = mean(Area)) %>%
             mutate(Densm = Popm / Aream) %>%
             select(Densm))
```

```
##      Densm
## 1 34.74338
```

```r
# b. Estima média da densidade demográfica por km2 por município
(media.densidade <- munic_amo %>%
                    summarise(Densm = mean(Densidade)))
```

```
##      Densm
## 1 147.5197
```

```r
# Adiciona valor de r_chapeu aos dados da amostra
munic_amo <- cbind(munic_amo, r_chapeu)

# Soluções para item 2
#a.Estima o erro padrão e o coeficiente de variação
(precisao.r_chapeu <- munic_amo %>%
                      mutate(Z = Pop - Densm * Area) %>%
                      summarise(varZ = var(Z),
                                Aream = mean(Area),
                                Densm = mean(Densm)) %>%
                      mutate(dp.r_chapeu = sqrt((1/n - 1/N)*varZ)/Aream,
                             cv.r_chapeu = 100 * dp.r_chapeu / Densm)  %>%
                      select(dp.r_chapeu, cv.r_chapeu))
```

```
##   dp.r_chapeu cv.r_chapeu
## 1    8.781149    25.27431
```

```
(precisao.media.densidade <- munic_amo %>%
         summarise(Densv = var(Densidade),
                   Densm = mean(Densidade)) %>%
         mutate(dp.media.dens = sqrt((1/n - 1/N)*Densv),
                cv.media.dens = 100 * dp.media.dens / Densm) %>%
         select(dp.media.dens, cv.media.dens))
```

```
##   dp.media.dens cv.media.dens
## 1      47.43697      32.15637
```

```
# b. Calcula densidade demográfica média por km2 no Brasil
(R <- MunicBR_dat %>%
      summarise(Popm = mean(Pop),
                Aream = mean(Area)) %>%
      mutate(Densidade_pop = Popm / Aream) %>%
      select(Densidade_pop))
```

```
## # A tibble: 1 x 1
##   Densidade_pop
##           <dbl>
## 1          23.6
```

```
# Estima média da densidade demográfica por km2 por município
(densidade_media_pop <- MunicBR_dat %>%
    summarise(densidade_media_pop = mean(Densidade)))
```

```
## # A tibble: 1 x 1
##   densidade_media_pop
##                 <dbl>
## 1                114.
```

6.8 Exercícios

Exercício 6.1 Identifique e defina duas razões que você já tenha utilizado em seu trabalho.

Exercício 6.2 Para estimar o total de despesas com gastos sociais das prefeituras de uma região que abrange 281 municípios, foi selecionada uma amostra aleatória simples sem reposição de 50 municípios. Sabe-se que a população total da região é de 6.818 (em milhares de pessoas). São dadas as seguintes informações provenientes da amostra:

$$\sum_{i=1}^{50} y_i = 128.080; \quad \sum_{i=1}^{50} x_i = 1.067; \quad s_y^2 = 6.244.516; \quad s_x^2 = 454,51; \quad s_{xy} = 45.399$$

Obs.: os valores de x e y são dados em milhares.

Calcule a estimativa de total da característica y, que representa a despesa com gastos sociais, e o respectivo intervalo com 95% de confiança para essa estimativa de total baseada em cada um dos seguintes estimadores:

a. Estimador simples.
b. Estimador de razão, utilizando como variável auxiliar a população (característica x).
c. Compare e comente os resultados.

Exercício 6.3 Numa cidade com 75.000 habitantes, uma amostra aleatória simples de $n = 4$ domicílios foi selecionada dos 25.000 domicílios da cidade para estimar o custo médio de alimentação por domicílio em uma semana. A Tabela 6.3 mostra os resultados obtidos na amostra.

Tabela 6.3: Informações coletadas na amostra de domicílios

Domicílio	Moradores	Custo com alimentos
1	4	150,00
2	2	100,00
3	4	200,00
4	3	140,00

Considerando que $(N - n)/N \doteq 1$:

a. Identifique as unidades de amostragem, a variável de interesse, e alguma informação auxiliar associada com as unidades.

b. Descreva dois tipos de estimadores para estimar a despesa média por domicílio para a alimentação por uma semana na cidade. Sumarize algumas propriedades de cada estimador.
c. Estime a despesa média por domicílio usando o primeiro estimador e estime a variância do estimador.
d. Estime a despesa média por domicílio usando o segundo estimador e estime a variância do estimador.
e. Baseado nos resultados, qual estimador é preferível nesta situação? Por que?

Exercício 6.4 Num estudo para estimar a quantidade total de açúcar contida num carregamento de laranjas, uma AAS de $n = 10$ laranjas foi selecionada. Cada uma das laranjas foi pesada, retirado o suco e pesada a quantidade de açúcar no suco. A Tabela 6.4 dá os resultados para as 10 laranjas da amostra. O caminhão foi pesado carregado com as laranjas e, depois, vazio levando a conclusão que o peso total da carga de laranjas era de 1800 libras. Estime o total de açúcar contido em todas as laranjas e dê um intervalo de 95% de confiança para esse total.

Tabela 6.4: Peso e quantidade de açúcar, ambos em libras, nas laranjas da amostra

Laranja	Peso da laranja	Quantidade de açúcar
1	0,40	0,021
2	0,48	0,030
3	0,43	0,025
4	0,42	0,022
5	0,50	0,033
6	0,46	0,027
7	0,39	0,019
8	0,41	0,021
9	0,42	0,023
10	0,44	0,025

Exercício 6.5 Uma reserva florestal foi dividida em 200 áreas de mesmo tamanho. Através de aerofotogrametria foram contadas as árvores mortas de todas as áreas chegando-se a um total de 15600 árvores. Foi selecionada uma amostra aleatória de 10 áreas onde foram contadas localmente as árvores mortas. Os resultados aparecem na Tabela 6.5.

Tabela 6.5: Contagem das árvores mortas na amostra, por aerofotogrametria (x_i) e localmente(y_i)

Área	1	2	3	4	5	6	7	8	9	10
x_i	12	30	24	24	18	30	12	6	36	42
y_i	18	42	24	36	24	36	14	10	48	54

a. Supondo que a amostra é *AAS*, estime o número total de árvores mortas na reserva e dê uma estimativa para a variância.
b. Estime o vício do número de árvores mortas.
c. Recalcule as estimativas sem utilizar a informação auxiliar e compare os resultados.

Exercício 6.6 Considere uma população U com $N = 100$ unidades. Para cada uma das unidades U_i, considere o par (Y_i, X_i), onde Y_i tem distribuição $U(50, 100)$ e $x_i = 3y_i + e_i$, com $e_i \sim N(0, 16)$, $i = 1,\ 2,\ \ldots,\ 100$.

a. Usando funções do R, gere a população U com os valores de Y_i e X_i, como descrito acima.
b. Calcule os parâmetros populacionais $\overline{Y},\ \overline{X},\ S_y^2,\ S_x^2$.
c. Selecione 200 *AAS* de tamanho $n = 20$ da população gerada.
d. Para cada amostra estime $\overline{Y}$ e o respectivo intevalo de 95% de confiança, usando o estimador natural da *AAS* e o estimador de razão.
e. Verifique, para cada tipo de estimador, quantos intervalos não contém o valor verdadeiro da média populacional de y. O resultado é razoável?
f. Calcule o vício médio empírico das estimativas, utilizando

$$\widehat{\overline{B}} = \frac{1}{200}\sum|\overline{y}_i - \overline{Y}| \text{ e } \widehat{\overline{B}}_R = \frac{1}{200}\sum|\overline{y}_{Ri} - \overline{Y}|$$

g. Compare e comente os resultados.

Exercício 6.7 Foi solicitado a um grupo de 50 alunos de uma escola que cada um declarasse quanto dinheiro tinha no bolso (x_i), sem contar. A média dos valores declarados por todos os alunos foi de $\overline{X} = R\$4,93$. Em seguida foram selecionados, aleatoriamente e sem reposição, 10 desses alunos que foram solicitados a contar quanto dinheiro cada um tinha no bolso (y_i). A Tabela 6.6 mostra os dois valores para cada aluno da amostra.

Tabela 6.6: Quantias declaradas (x_i) e contadas (y_i) pelos alunos da amostra

x_i	1,91	7,41	3,03	8,63	4,76	8,12	3,89	2,75	8,66	0,85
y_i	3,65	7,56	4,24	11,93	5,86	8,05	6,55	2,95	8,63	1,44

a. Estime o total do dinheiro dos 50 alunos do grupo e a variância da estimativa.
b. Refaça as estimativas sem utilizar as informações da variável auxiliar.
c. Comente os resultados.

Exercício 6.8 Deseja-se estimar a idade média das árvores de um bosque. Determinar a idade diretamente é bastante custoso. Sabe-se, entretanto, que há uma correlação entre a idade e o diâmetro do tronco da árvore. Foi selecionada uma *AAS* de 20 árvores e para estas foi determinada a idade de cada uma. Foram medidos os diâmetros de todas as 1.132 árvores do bosque e calculado o diâmetro médio em $10,3$ polegadas.

A Tabela 6.7 apresenta os dados para a amostra selecionada.

Tabela 6.7: Diâmetro e idade das árvores da amostra

Árvore	Diâmetro	Idade	Árvore	Diâmetro	Idade
1	12,0	125	11	5,7	61
2	11,4	119	12	8,0	80
3	7,9	83	13	10,3	114
4	9.0	85	14	12,0	147
5	10,5	99	15	9,2	122
6	7,9	117	16	8,5	106
7	7,3	69	17	7,0	82
8	10,2	133	18	10,7	88
9	11,7	154	19	9,3	97
10	11,3	168	20	8,2	99

a. Estime a idade média das árvores do bosque usando o estimador não viciado e dê um intervalo de 95% de confiança.
b. Estime a idade média das árvores do bosque usando o estimador de razão e dê um intervalo de 95% de confiança.
c. Qual das estimativas você prefere e por quê?
d. Utilizando a amostra já disponível como se fosse uma pesquisa piloto, verifique se o tamanho de amostra utilizado seria suficiente para considerar o vício do estimador de razão como desprezível em termos práticos.

Exercício 6.9 Considere os dados de uma população hipotética.

A Tabela 6.8 contém os dados dessa população de pesquisa.

Tabela 6.8: População hipotética de $N = 9$ unidades

U_i	X_i	Y_i
1	13	10
2	7	7
3	11	13
4	12	17
5	4	8
6	3	1
7	11	15
8	3	7
9	5	4

a. Calcule os parâmetros populacionais: X, Y, S_x^2, S_y^2, ρ_{xy} e R.
b. Determine todas AAS possíveis de tamanho 3 e construa as distribuições amostrais dos estimadores da média utilizando o estimador não viciado e o estimador de razão.
c. Construa os histogramas para as duas distribuições e comente os resultados.
d. Calcule a média e a variância para as duas distribuições e as compare.
e. Calcule o vício do estimador de razão para a média e compare com o resultado obtido pela fórmula aproximada do vício.