# MAT02025 - Amostragem 1

## AAS: estimativa de um índice e estimativa de valores médios e totais das subpopulações

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2023

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

# Estimativa de um índice

# Relembrando

**Número índice**

No sentido mais simples do termo, podemos dizer que um **número índice** é um quociente que expressa uma dada quantidade em comparação a uma **quantidade base**. Em outras palavras, são **valores relativos**.

# Estimativa de um índice

- Frequentemente, a quantidade que deve ser estimada a partir de uma amostra aleatória simples é a **razão de duas variáveis**, ambas as quais variam de unidade para unidade.
- Em um levantamento por amostragem domiciliar, alguns exemplos são:
  - o **número de aparelhos de celular por residente** (o número de celulares e residentes variam de domicílio para domicílio);
  - a **despesa com aplicativos de transporte por residente adulto**;
  - o **número médio de horas por semana gastas assistindo programas no serviço de *streaming* por criança de 10 a 15 anos**.

# Estimativa de um índice

- A fim de estimar a primeira dessas quantidades, registraríamos para o $i$-ésimo domicílio $(i = 1, 2, \ldots, n)$ o número de residentes $X_i$ que ali vivem e o número total de aparelhos de celular $Y_i$ que eles possuem.
- O **parâmetro da população** a ser estimado é a **razão (ou índice)**

$$R = \frac{\text{número total de aparelhos de celular}}{\text{número total de residentes}} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i}.$$

# Estimativa de um índice

- A estimativa amostral correspondente é

$$\widehat{R} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\overline{y}}{\overline{x}}.$$

# Estimativa de um índice

- Exemplos dessa natureza ocorrem, frequentemente, quando a unidade de amostragem **(no caso o domicílio)** compreende um grupo ou um conjunto de elementos **(residentes)** e nosso interesse está no valor médio da população **por elemento**.
- Os índices também aparecem em muitas outras aplicações, como, por exemplo
  - o **índice de empréstimos para construções imobiliárias no total de empréstimos de um banco**;
  - ou **índice de acres plantados com trigo, no total de acres cultivados de uma fazenda**;
  - ou **índice de casos de diabetes não diagnosticado, no total de casos de diabetes**.

# Estimativa de um índice

Epidemiology/Health Services Research
ORIGINAL ARTICLE

## Prevalence of Diabetes and High Risk for Diabetes Using A1C Criteria in the U.S. Population in 1988–2006

Catherine C. Cowie, PhD[1]
Keith F. Rust, PhD[2]
Danita D. Byrd-Holt, BBA[3]
Edward W. Gregg, PhD[4]
Earl S. Ford, MD[5]
Linda S. Geiss, MS[4]
Kathleen E. Bainbridge, PhD[3]
Judith E. Fradkin, MD[1]

cal trials in type 1 and type 2 diabetic patients, which have established widely accepted A1C treatment goals for diabetes. A cut point of ≥6.5% for the diagnosis of diabetes was recommended by the

*Prevalence of diabetes using A1C*

Table 1—*Crude prevalence of diagnosed diabetes, undiagnosed diabetes (A1C ≥6.5%), total diabetes (diagnosed and undiagnosed combined), total diabetes that is undiagnosed, and at high risk for diabetes (A1C ≥6.0 to <6.5%), by age, sex, and race/ethnicity: NHANES 2003–2006 (n = 13,094)*

| | Diagnosed diabetes | Undiagnosed diabetes | Total diabetes | Total diabetes that is undiagnosed | At high risk for diabetes |
|---|---|---|---|---|---|
| Combined age-groups (years) | | | | | |
| ≥12 | 6.8 (6.1–7.5) | 1.6 (1.2–1.9) | 8.4 (7.6–9.2) | 19.0 (15.2–22.7) | 3.1 (2.7–3.4) |
| ≥20 | 7.8 (7.0–8.6) | 1.8 (1.4–2.2) | 9.6 (8.7–10.5) | 19.0 (15.2–22.7) | 3.5 (3.0–3.9) |
| ≥65 | 17.7 (15.6–19.7) | 3.5 (2.6–4.3) | 21.1 (18.7–23.5) | 16.3 (12.9–19.8) | 8.1 (6.6–9.6) |

# Estimativa de um índice

- A **distribuição amostral** de $\widehat{R}$ é mais complicada que a de $\overline{y}$, porque tanto o numerador $\overline{y}$, quanto o denominador, $\overline{x}$, variam de amostra para amostra.
- Em **pequenas amostras**, a distribuição de $\widehat{R}$ é **assimétrica**, e $\widehat{R}$ é, geralmente, uma estimativa ligeiramente **viesada** de $R$.
- Em **grandes amostras**, a distribuição de $\widehat{R}$ tende à **normalidade** e o **viés** torna-se **insignificante**.

# Estimativa de um índice

- O seguinte **resultado aproximado** servirá para a maioria dos propósitos[1].

**Teorema 14.1**

Se as variáveis $Y_i$ e $X_i$ são medidas em cada unidade de uma amostra aleatória simples de tamanho $n$, que se presume grande, a variância de $\widehat{R} = \overline{y}/\overline{x}$ é, aproximadamente[2],

$$\text{Var}\left(\widehat{R}\right) \doteq \frac{1-f}{n\overline{X}^2} \frac{\sum_{i=1}^{N}\left(Y_i - RX_i\right)^2}{N-1},$$

em que $R = \overline{Y}/\overline{X}$ é o **índice dos valores médios da população**, e $f = n/N$ é a **fração de amostragem**.

[1] A distribuição de $\widehat{R}$ é estudada com mais detalhes no Capítulo 6 de Cochran (1965) e no Capítulo 5 de Bolfarine e Bussab (2005).

[2] O símbolo $\doteq$ indica "aproximadamente igual".

# Estimativa de um índice

**Demonstração.** Note que

$$\widehat{R} - R = \frac{\overline{y}}{\overline{x}} - R = \frac{\overline{y} - R\overline{x}}{\overline{x}}.$$

- Se $n$ é grande, $\overline{x}$ não deve ser muito diferente de $\overline{X}$.
- Para evitar ter que calcular a distribuição da razão de duas variáveis aleatórias $\overline{y} - R\overline{x}$ e $\overline{x}$, substituímos $\overline{x}$ por $\overline{X}$ no denominador da expressão acima **como uma aproximação**. Isso dá

$$\widehat{R} - R \doteq \frac{\overline{y} - R\overline{x}}{\overline{X}}.$$

# Estimativa de um índice

- Agora calcule a média de todas as amostras aleatórias simples de tamanho $n$:

$$\mathrm{E}\left(\widehat{R} - R\right) \doteq \frac{\mathrm{E}\left(\bar{y} - R\bar{x}\right)}{\overline{X}} = \frac{\overline{Y} - R\overline{X}}{\overline{X}} = 0,$$

uma vez que $R = \overline{Y}/\overline{X}$.

- Isso mostra que, para a ordem de aproximação usada aqui, $\widehat{R}$ é uma estimativa não viesada de $R$.

# Estimativa de um índice

- Da expressão aproximada, também obtemos

$$\mathrm{Var}\,(\widehat{R}) = \mathrm{E}\,(\widehat{R} - R)^2 \doteq \frac{1}{\overline{X}^2}\mathrm{E}\,(\overline{y} - R\overline{x})^2.$$

- A quantidade $\overline{y} - R\overline{x}$ é a média amostral da variável $D_i = Y_i - RX_i$, cuja média populacional, $\overline{D} = \overline{Y} - R\overline{X}$, é igual a $0$.
- Portanto, podemos encontrar $\mathrm{Var}\,(\widehat{R})$ aplicando o teorema para a variância da média de uma amostra aleatória simples à variável $D_i$ e dividindo por $\overline{X}^2$.

# Estimativa de um índice

- Isso dá

$$
\begin{aligned}
\mathrm{Var}\left(\widehat{R}\right) &\doteq \frac{1}{\overline{X}^2}\mathrm{E}\left(\overline{y} - R\overline{x}\right)^2 = \frac{1}{\overline{X}^2}\frac{S_D^2}{n}(1-f) \\
&= \frac{1-f}{n\overline{X}^2}\frac{\sum_{i=1}^{N}\left(D_i - \overline{D}\right)^2}{N-1} = \frac{1-f}{n\overline{X}^2}\frac{\sum_{i=1}^{N}\left(Y_i - RX_i\right)^2}{N-1},
\end{aligned}
$$

o que completa a demonstração.

# Estimativa de um índice

- Como estimativa amostral de

$$\frac{\sum_{i=1}^{N}(Y_i - RX_i)^2}{N-1}$$

é comum tomarmos

$$\frac{\sum_{i=1}^{n}(Y_i - \widehat{R}X_i)^2}{n-1}.$$

- Pode-se demonstrar que essa estimativa tem um **viés de ordem**[3] $1/n$.

[3] Ou seja, conforme $n \to \infty$ o viés decresce a zero mais rapidamente que a sequência $1/n$. Ou, utilizando a notação $o(\cdot)$, temos que $\text{Viés}_n = o(n^{-1})$.

# Estimativa de um índice

- Para o **erro padrão estimado** de $\widehat{R}$, temos

$$s_{\widehat{R}} = \frac{\sqrt{1-f}}{\sqrt{n}\overline{X}} \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{R}X_i)^2}{n-1}}.$$

- Se $\overline{X}$ não é conhecido, a estimativa amostral $\overline{x}$ o substitui no denominador da fórmula.
- Uma fórmula prática para calcular $s_{\widehat{R}}$ é dada por

$$s_{\widehat{R}} = \frac{\sqrt{1-f}}{\sqrt{n}\overline{X}} \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - 2\widehat{R}\sum_{i=1}^{n} Y_i X_i + \widehat{R}^2 \sum_{i=1}^{n} X_i^2}{n-1}}.$$

# Estimativa do valor médio das subpopulações

# Estimativa do valor médio das subpopulações

- Em muitos levantamentos, as estimativas são feitas para cada uma das várias **classes** **(setores ou domínios)** nas quais a população está subdividida.
- Em um levantamento domiciliar, estimativas separadas podem ser necessárias para **famílias de** $0, 1, 2, \ldots$ **filhos**, para **proprietários** e **locatários**, ou para famílias em diferentes **grupos de ocupação**.

# Estimativa do valor médio das subpopulações

**Table 2**

Prevalence of diabetes mellitus in adults overall and by sex according to sociodemographic and clinical characteristics. *Brazilian National Health Survey*, 2013 and 2019.

| Characteristic | 2013 | | | | | | 2019 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total | | Men | | Women | | Total | | Men | | Women | |
| | % | 95%CI | % | 95%CI | % | 95%CI | % | 95%CI | % | 95%CI | % | 95%CI |
| **Total** | 6.2 | 5.9-6.6 | 5.3 | 4.8-5.8 | 7.0 | 6.5-7.5 | 7.7 | 7.4-8.0 | 6.9 | 6.5-7.4 | 8.4 | 8.0-8.8 |
| Age (years) | | | | | | | | | | | | |
| 18-24 | 0.5 | 0.3-0.8 | 0.4 | 0.1-0.7 | 0.6 | 0.2-1.1 | 0.7 | 0.4-1.1 | 1.0 | 0.4-1.7 | 0.4 | 0.2-0.6 |
| 25-34 | 0.8 | 0.6-1.0 | 0.7 | 0.4-1.1 | 0.9 | 0.6-1.2 | 0.9 | 0.7-1.2 | 0.7 | 0.3-1.0 | 1.1 | 0.8-1.5 |
| 35-44 | 2.9 | 2.4-3.4 | 2.4 | 1.7-3.2 | 3.4 | 2.6-4.1 | 3.1 | 2.7-3.6 | 3.1 | 2.5-3.8 | 3.2 | 2.5-3.8 |
| 45-54 | 6.6 | 5.8-7.4 | 5.8 | 4.6-6.9 | 7.3 | 6.2-8.4 | 7.7 | 6.9-8.5 | 6.9 | 5.8-8.0 | 8.4 | 7.2-9.6 |
| 55-64 | 13.5 | 12.1-14.9 | 12.1 | 9.8-14.3 | 14.7 | 12.8-16.6 | 14.8 | 13.8-15.7 | 13.6 | 12.2-15.0 | 15.8 | 14.5-17.1 |
| ≥ 65 | 19.8 | 18.2-21.3 | 17.7 | 15.1-20.3 | 21.4 | 19.3-23.5 | 21.6 | 20.5-22.7 | 20.2 | 18.6-21.9 | 22.7 | 21.2-24.1 |
| Race/Skin color | | | | | | | | | | | | |
| White | 6.7 | 6.1-7.2 | 6.0 | 5.2-6.8 | 7.3 | 6.5-8.0 | 8.0 | 7.6-8.5 | 7.9 | 7.1-8.6 | 8.2 | 7.5-8.9 |
| Black | 7.3 | 6.0-8.6 | 5.5 | 3.4-7.5 | 8.9 | 7.2-10.5 | 7.8 | 7.0-8.7 | 6.9 | 5.7-8.0 | 8.7 | 7.5-9.9 |
| Mixed-race | 5.5 | 5.0-5.9 | 4.5 | 3.8-5.1 | 6.4 | 5.7-7.0 | 7.3 | 6.9-7.7 | 5.9 | 5.3-6.4 | 8.5 | 7.9-9.1 |
| Asian | 6.3 | 3.0-9.6 | 7.3 | 0.9-13.7 | 5.6 | 2.2-9.0 | 12.8 | 7.9-17.7 | 13.6 | 6.3-21.0 | 12.0 | 5.5-18.5 |
| Indigenous | 6.9 | 2.7-11.1 | 5.4 | 1.8-18.7 * | 8.0 | 2.8-13.2 | 7.5 | 4.4-10.6 | 5.2 | 1.4-8.9 | 10.2 | 5.3-15.1 |
| Education level | | | | | | | | | | | | |
| Incomplete elementary | 9.6 | 9.0-10.3 | 6.7 | 5.8-7.5 | 12.4 | 11.3-13.4 | 12.9 | 12.3-13.5 | 10.2 | 9.4-11.0 | 15.4 | 14.4-16.3 |
| Complete elementary | 5.4 | 4.5-6.3 | 5.4 | 4.0-6.9 | 5.4 | 4.3-6.5 | 6.3 | 5.5-7.0 | 5.4 | 4.4-6.4 | 7.1 | 6.1-8.2 |
| Complete high school | 3.4 | 2.9-3.8 | 3.5 | 2.8-4.2 | 3.3 | 2.7-3.8 | 4.6 | 4.2-5.0 | 4.6 | 4.0-5.2 | 4.6 | 4.1-5.1 |
| Complete higher education | 4.1 | 3.3-5.0 | 5.5 | 3.9-7.2 | 3.1 | 2.3-4.0 | 4.7 | 4.1-5.2 | 6.0 | 5.0-7.1 | 3.6 | 3.1-4.2 |

**Fonte:** http://dx.doi.org/10.1590/0102-311X00149321.

# Estimativa do valor médio das subpopulações

- Na situação mais simples, cada unidade da população cai em um dos setors.
- Assuma que o $j$-ésimo setor contém $N_j$ unidades e seja $n_j$ o número de unidades em uma amostra aleatória simples de tamanho $n$ que por acaso caem neste setor.
- Se $Y_{jk}(k = 1, 2, \ldots, n_j)$ são as medidas nessas unidades, a média da população $\overline{Y}_j$ para o $j$-ésimo setor é estimada por

$$\overline{y}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} Y_{jk}.$$

# Estimativa do valor médio das subpopulações

- À primeira vista, $\overline{y}_j$ parece ser uma **estimativa de razão (índice)**.
- Embora $n$ seja fixo, $n_j$ variará de uma amostra de tamanho $n$ para outra.
- A complicação de uma estimativa de razão pode ser evitada considerando a distribuição de $\overline{y}_j$ sobre as amostras nas quais $n$ e $n_j$ **são fixos**.
    - Assumimos $n_j > 0$.

# Estimativa do valor médio das subpopulações

- Na totalidade das amostras, com $n$ e $n_j$ determinados, a probabilidade de que qualquer conjunto específico de $n_j$ unidades das $N_j$ unidades no setor $j$ sejam sorteadas é

$$\frac{{}_{N-N_j}C_{n-n_j}}{{}_{N-N_j}C_{n-n_j} \times {}_{N_j}C_{n_j}} = \frac{1}{{}_{N_j}C_{n_j}}.$$

- Uma vez que cada conjunto específico de $n_j$ unidades do setor $j$ pode aparecer em todas as seleções de $(n - n_j)$ unidades, dentre as $(N - N_j)$ que não estão no setor $j$, o numerador acima é o número de amostras contendo um conjunto especificado de $n_j$ e o denominador é o número total de amostras.

# Estimativa do valor médio das subpopulações

- Segue-se que os **teoremas das aulas 9, 10 e 11** se aplicam ao $Y_{jk}$ se colocarmos $n_j$ no lugar de $n$ e $N_j$ no lugar de $N$.

Do **Teorema 9.1**, $\overline{y}_j$ é um estimador **não enviesado** para $\overline{Y}_j$.

Do **Teorema 10.1**, o **erro padrão** de $\overline{y}_j$ é $\frac{S_j}{\sqrt{n_j}}\sqrt{1-(n_j/N_j)}$, em que

$$S_j^2 = \frac{1}{N_j - 1}\sum_{k=1}^{N_j}(Y_{jk} - \overline{Y}_j)^2.$$

# Estimativa do valor médio das subpopulações

De acordo com o **Teorema 11.1** e o **Corolário 11.1** , uma estimativa do erro padrão de $\overline{y}_j$ é

$$\frac{s_j}{\sqrt{n_j}}\sqrt{1-(n_j/N_j)},$$

em que

$$s_j^2 = \frac{1}{n_j-1}\sum_{k=1}^{n_j}(Y_{jk}-\overline{y}_j)^2.$$

- Se o valor de $N_j$ **não for conhecido**, a quantidade $n/N$ pode ser utilizada em lugar de $n_j/N_j$, no cálculo das **cpf**.
  - Na amostragem aleatória simples, $n_j/N_j$ é uma estimativa não enviesada de $n/N$.

# Estimativa dos valores totais das subpopulações

# Estimativa dos valores totais das subpopulações

- Suponha que em uma população (de adultos), na qual algumas pessoas são obesas (IMC $> 30$) e outras não, podemos desejar estimar, por meio de uma amostra, o ***total* de pessoas com diabetes *entre os obesos***.
- Se $N_j$ (o **número de pessoas obesas** na população) é conhecido, não há problema.
  - A estimativa a partir da amostra é $N_j\bar{y}_j$ e seu erro padrão condicional é $N_j$ vezes $\frac{S_j}{\sqrt{n_j}}\sqrt{1-(n_j/N_j)}$.

# Estimativa dos valores totais das subpopulações

- Alternativamente, se o *total de indivíduos com diabetes* for conhecido na **população**, uma estimativa de razão pode ser usada.
  - A **amostra** fornece uma estimativa da razão

$$\frac{\text{total de pessoas com diabetes entre os obesos}}{\text{total de indivíduos com diabetes}}.$$

- Isso é multiplicado pelo *total de indivíduos com diabetes* conhecido na **população**.

# Estimativa dos valores totais das subpopulações

- Se nem $N_j$, e nem o *total de indivíduos com diabetes* é conhecido, essas estimativas não podem ser feitas.
- Em vez disso, multiplicamos o *valor amostral total* das unidades $Y$ contidas no $j$-ésimo setor pelo **fator de expansão** $N/n$.
- Isso dá a estimativa

$$\hat{Y}_{T_j} = \frac{N}{n} \sum_{k=1}^{n_j} Y_{jk}.$$

- Mostraremos que $\hat{Y}_{T_j}$ **é imparcial** e obteremos seu erro padrão sobre amostras repetidas de tamanho $n$.
  - O artifício de manter $n_j$ constante, bem como $n$ não ajuda neste caso.

# Estimativa dos valores totais das subpopulações

- Ao fazermos a demonstração, voltamos à notação original, na qual $Y_i$ é a medida da $i$-ésima unidade da população.
- Defina para cada unidade na população uma nova variável $Y_i'$, em que

$$Y_i' = \begin{cases} Y_i, & \text{se a unidade pertencer ao setor } j, \\ 0, & \text{caso contrário.} \end{cases}$$

- O valor **total populacional** da variável $Y_i'$ é

$$\sum_{i=1}^{N} Y_i' = \sum_{\text{setor } j} Y_i = Y_{T_j}.$$

# Estimativa dos valores totais das subpopulações

- Em uma **amostra aleatória simples** de tamanho $n$, $Y_i' = Y_i$ para todas as $n_j$ unidades que se encontram no $j$-ésimo setor; $Y_i' = 0$ para todas as restantes $n - n_j$ unidades.
- Se $\overline{y}'$ é a média amostral de $Y_i'$, então temos

$$N\overline{y}' = \frac{N}{n}\sum_{i=1}^{n} Y_i' = \frac{N}{n}\sum_{k=1}^{n_j} Y_{jk} = \hat{Y}_{T_j}$$

- Este resultado mostra que a estimativa $\hat{Y}_{T_j}$ é $N$ vezes a média amostral de $Y_i'$.

# Estimativa dos valores totais das subpopulações

- Em repetidas amostras de tamanho $n$, podemos aplicar os **teoremas das aulas 9, 10 e 11** às variáveis $Y_i'$.
- Estes, por sua vez, mostram que $\hat{Y}_{T_j}$ é uma **estimativa imparcial** de $Y_{T_j}$ com erro padrão

$$\sigma(\hat{Y}_{T_j}) = \frac{NS'}{\sqrt{n}}\sqrt{1-(n/N)},$$

em que $S'$ é desvio padrão populacional de $Y_i'$.

# Estimativa dos valores totais das subpopulações

- Para calcular $S'$ , consideramos a população como consistindo de $N_j$ valores $Y_i$ que estão no $j$-ésimo setor e de $N - N_j$ **valores zero**. Assim

$$S'^2 = \frac{1}{N-1}\left(\sum_{\text{setor } j} Y_i^2 - \frac{Y_{T_j}^2}{N}\right).$$

# Estimativa dos valores totais das subpopulações

- Pelo teorema da **aula 11**, uma estimativa amostral do erro padrão de $\hat{Y}_{T_j}$ será

$$s(\hat{Y}_{T_j}) = \frac{Ns'}{\sqrt{n}}\sqrt{1-(n/N)}.$$

- No cálculo de $s'$, qualquer unidade que não esteja no $j$-ésimo setor recebe um valor zero.

# Comparação da eficiência dos estimadores de total no setor

- Às vezes é possível, com algum esforço, identificar e contar as unidades que não contribuem com nada, de modo que em nossa notação $(N - N_j)$, e portanto $N_j$, seja conhecido.
- Consequentemente, vale a pena examinar o quanto da $\text{Var}\left(\hat{Y}_{T_j}\right)$ é reduzido quando $N_j$ é conhecido.
- Se $N_j$ **não for conhecido**, temos (pelo **Corolário 10.2**)

$$\text{Var}\left(\hat{Y}_{T_j}\right) = \frac{N^2 S'^2}{n}\left(1 - \frac{n}{N}\right).$$

# Comparação da eficiência dos estimadores de total no setor

- Se $\overline{Y}_j$ e $S_j$ são a média e o desvio padrão no setor de interesse (ou seja, entre as unidades diferentes de zero), é possível verificar que

$$(N-1)S'^2 = (N_j - 1)S_j^2 + N_j\overline{Y}_j^2\left(1 - \frac{N_j}{N}\right).$$

- Uma vez que os termos em $1/N_j$ e $1/N$ são quase sempre insignificantes, temos

$$S'^2 \doteq P_jS_j^2 + P_jQ_j\overline{Y}_j^2,$$

em que $P_j = N_j/N$ e $Q_j = 1 - P_j$.

# Comparação da eficiência dos estimadores de total no setor

- Desta forma,

$$\mathrm{Var}\left(\hat{Y}_{T_j}\right) \doteq \frac{N^2}{n}(P_j S_j^2 + P_j Q_j \overline{Y}_j^2)\left(1 - \frac{n}{N}\right). \qquad (1)$$

# Comparação da eficiência dos estimadores de total no setor

- Se as unidades diferentes de zero forem identificadas (ou seja, se $N_j$ **for conhecido**), retiramos delas uma amostra de tamanho $n_j$. A estimativa do total do setor é $N_j\overline{y}_j$ com variância

$$\mathrm{Var}\left(N_j\overline{y}_j\right) = \frac{N_j^2}{n_j}S_j^2\left(1-\frac{n_j}{N_j}\right) = \frac{N^2}{n_j}P_j^2S_j^2\left(1-\frac{n_j}{N_j}\right). \qquad (2)$$

# Comparação da eficiência dos estimadores de total no setor

- As variâncias dadas pelas expressões (1) e (2) são comparáveis.
- Em (1), o número médio de unidades diferentes de zero na amostra de tamanho $n$ é $nP_j$.
- Se tomarmos $n_j = nP_j$ em (2), de modo que o número de valores diferentes de zero a serem medidos seja aproximadamente o mesmo com ambos os métodos, (2) torna-se

$$\mathrm{Var}\left(N_j\overline{y}_j\right) = \frac{N^2}{n}P_jS_j^2\left(1 - \frac{n}{N}\right). \qquad (3)$$

# Comparação da eficiência dos estimadores de total no setor

- A razão entre as variâncias (3) e (1) é

$$\frac{\text{Var}_{N_j \text{ conhecido}}(N_j\overline{y}_j)}{\text{Var}_{N_j \text{ desconhecido}}(\hat{Y}_{T_j})} = \frac{S_j^2}{S_j^2 + Q_j\overline{Y}_j^2} = \frac{C_j^2}{C_j^2 + Q_j} \leq 1,$$

em que $C_j = S_j/\overline{Y}_j$ é o **coeficiente de variação** entre as unidades de valores diferentes de zero.

# Comparação da eficiência dos estimadores de total no setor

**Observação**

- Como era de se esperar, a redução da variância, decorrente do conhecimento de $N_j$, é maior quando a proporção de unidades de valor nulo é grande e quando $Y_j$ varia relativamente pouco entre as unidades de valor diferente de zero.

# Para casa 1 (PQP.1)

# Para casa 1 (PQP.1)

- A tabela a seguir mostra o número de pessoas ($X_1$), a renda familiar semanal ($X_2$) e os gastos semanais com alimentação ($Y$) em uma amostra aleatória simples de 33 famílias de baixa renda.
- Como a amostra é pequena, os dados se destinam apenas a ilustrar os cálculos.

# Para casa 1 (PQP.1)

**Tabela 1:** Renda semanal e custo dos alimentos de 33 famílias

| Número da família | Tamanho ($x_1$) | Renda ($x_2$) | Custo dos alimentos ($y$) | Número da família | Tamanho ($x_1$) | Renda ($x_2$) | Custo dos alimentos ($y$) |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 62 | 14,3 | 18 | 4 | 83 | 36,0 |
| 2 | 3 | 62 | 20,8 | 19 | 2 | 85 | 20,6 |
| 3 | 3 | 87 | 22,7 | 20 | 4 | 73 | 27,7 |
| 4 | 5 | 65 | 30,5 | 21 | 2 | 66 | 25,9 |
| 5 | 4 | 58 | 41,2 | 22 | 5 | 58 | 23,3 |
| 6 | 7 | 92 | 28,2 | 23 | 3 | 77 | 39,8 |
| 7 | 2 | 88 | 24,2 | 24 | 4 | 69 | 16,8 |
| 8 | 4 | 79 | 30,0 | 25 | 7 | 65 | 37,8 |
| 9 | 2 | 83 | 24,2 | 26 | 3 | 77 | 34,8 |
| 10 | 5 | 62 | 44,4 | 27 | 3 | 69 | 28,7 |
| 11 | 3 | 63 | 13,4 | 28 | 6 | 95 | 63,0 |
| 12 | 6 | 62 | 19,8 | 29 | 2 | 77 | 19,5 |
| 13 | 4 | 60 | 29,4 | 30 | 2 | 69 | 21,6 |
| 14 | 4 | 75 | 27,1 | 31 | 6 | 69 | 18,2 |
| 15 | 2 | 90 | 22,2 | 32 | 4 | 67 | 20,1 |
| 16 | 5 | 15 | 37,7 | 33 | 2 | 63 | 20,7 |
| 17 | 3 | 69 | 22,6 | | | | |

# Para casa 1 (PQP.1)

1. Estime a partir da amostra
   a. o gasto semanal médio com comida por família;
   b. o gasto semanal médio com comida por pessoa;
   c. a porcentagem da renda que é gasta com comida.
2. Calcule os erros padrões dessas estimativas (pode ignorar as **cpf**).
3. Compartilhe os seus resultados no Fórum Geral do Moodle.

# Para casa 2 (PQP.2)

# Para casa 2 (PQP.2)

- Considere o exemplo da **aula 11 (assinaturas de uma petição)**. Depois de selecionada a amostra, o número de folhas completamente cheias (com 42 assinaturas cada) foi contado e verificou-se que eram 326. Use essa informação para fazer uma estimativa melhorada do número total de assinaturas e achar o erro padrão da sua estimativa.

# Próxima aula

- Validade da aproximação normal.

# Por hoje é só!

**Bons estudos!**