Distribuição amostral de médias

Prof. Wagner Hugo Bonat

Departamento de Estatística
Universidade Federal do Paraná

Definições (mais detalhadas)

- **População ou Universo**: Conjunto de todos as unidades elementares.

$$\mathsf{U} = \{1, 2, \ldots, N\},$$

onde N é o tamanho da população.

- **Unidade elementar**: refere-se a qualquer elemento $i \in \mathsf{U}$.
- **Variável**: característica a ser observada em cada unidade elementar $\rightarrow$ variável aleatória. Notação $Y_i, \quad i \in \mathsf{U}$.
- Todos os valores de uma variável denotamos por $\mathbf{D} = (Y_1, \ldots, Y_N)$.
- **Função paramétrica populacional**: característica numérica qualquer da população, ou seja, uma expressão que condensa os Y_i's. Notação,

$$\theta(\mathbf{D}).$$

Exemplos: total, médias, quocientes, etc.

- É comum utilizar a expressão **parâmetro populacional**.

Exemplo: População de domicílios

Considere a **população** formada por três domicílios $U = \{1,2,3\}$ nos quais estão sendo observadas as seguintes variáveis: nome (do chefe), sexo, idade, fumante ou não, renda bruta (mensal em salários mínimos) familiar e número de trabalhadores.

Variável	Valores			Notação
unidade	1	2	3	i
nome do chefe	Ada	Beto	Ema	A_i
sexo[1]	0	1	0	X_i
idade	20	30	40	I_i
fumante	0	1	1	G_i
renda bruta	12	30	18	F_i
nº trabalhadores	1	3	2	T_i

[1] 0: feminino; 1: masculino.

[2] 0: não fumante; 1: fumante.

Exemplos de funções paramétricas populacionais

- Idade média

$$\theta(\mathbf{D}) = \frac{20 + 30 + 40}{3} = 30.$$

- Média das variáveis renda e número de trabalhadores

$$\theta(\mathbf{D}) = \begin{pmatrix} \frac{12+30+18}{3} \\ \frac{1+3+2}{3} \end{pmatrix} = \begin{pmatrix} 20 \\ 2 \end{pmatrix}.$$

- Renda média por trabalhador

$$\theta(\mathbf{D}) = \frac{12 + 30 + 18}{1 + 3 + 2} = 10.$$

Parâmetros populacionais mais usados

- Total populacional

$$\theta(\mathbf{D}) = \tau = \sum_{i=1}^{N} Y_i.$$

- Média populacional

$$\theta(\mathbf{D}) = \mu = \overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i.$$

- Variância populacional,

$$\sigma^2 = \theta(\mathbf{D}) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2,$$

ou às vezes

$$\theta(\mathbf{D}) = S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \mu)^2.$$

Amostra

- Uma **sequência** qualquer de n unidades de U é uma amostra ordenada de U,

$$\mathbf{s} = (1, \ldots, i, \ldots, n) \quad \text{tal que} \quad i \in \mathrm{U}.$$

- O **rótulo** i é chamado de i-ésimo componente de $\mathbf{s}$.
- Exemplos: Seja $\mathrm{U} = \{1,2,3\}$, os vetores $\mathbf{s}_1 = (1,2)$, $\mathbf{s}_2 = (2,1)$ e $\mathbf{s}_3 = (2,2,1,3,2)$ são amostras de U.
- Chama-se de **tamanho de amostra** o número de elementos em $\mathbf{s}$.
- Chama-se de **dados da amostra** $\mathbf{s}$ a matriz ou vetor de observações pertencentes à amostra, notação

$$\mathbf{d}_\mathbf{s} = (Y_1, \ldots, Y_n) = (Y_i, i \in \mathbf{s}).$$

Amostragem aleatória simples (AAS)

- Definição: De uma população U com N unidades elementares, sorteiam-se com **igual** probabilidade n unidades.

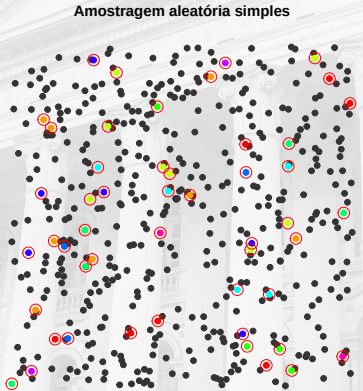

Figura 1. Amostragem aleatória simples.

Estatísticas

- Qualquer característica numérica dos dados correspondentes à **amostra** $\mathbf{s}$ é chamada de **estatística**, ou seja, qualquer função $h(\mathbf{d}_\mathbf{s})$ que relaciona as observações da amostra $\mathbf{s}$.
- Exemplo: Populações de domicílios (cont.): Considere a amostra $\mathbf{s} = (1,2)$. Para as variáveis renda bruta F e número de trabalhadores T, temos os seguintes dados da amostra:

$$\mathbf{d}_\mathbf{s} = \begin{pmatrix} 12 & 30 \\ 1 & 3 \end{pmatrix}.$$

- As médias (estatísticas) amostrais

$$\overline{f} = \frac{12+30}{2} = 21$$

e

$$\overline{t} = \frac{1+3}{2} = 2.$$

Distribuição amostral

- A **distribuição amostral** de uma **estatística** $h(\mathbf{d_s})$ é a distribuição de probabilidade da variável aleatória $H(\mathbf{d_s})$.
- Exemplo: População de domicílios (cont.): Determine a distribuição amostral da estatística $h(\mathbf{d_s})$ definida como a razão entre o total da renda familiar e o número de trabalhadores.
- População

$$\mathbf{D} = \begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} F_i \\ T_i \end{pmatrix}.$$

- Plano amostral AASc: Possíveis amostras
  $\mathbf{S} = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$.
- Calculando a estatística para a amostra $\mathbf{s} = (3,1)$,

$$r = \frac{18+12}{2+1} = 10.$$

Exemplo: População de domicílios (cont.)

- Calculando para todas as amostras temos,

$\mathbf{s}$	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
$P(\mathbf{s})$	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
$h(\mathbf{d}_\mathbf{s}) = r$	12	10,5	10	10,5	10	9,6	10	9,6	9

- Distribuição amostral de r

r	9	9,6	10	10,5	12
p_r	1/9	2/9	3/9	2/9	1/9

- Podemos resumir a distribuição amostral da v.a. R, por exemplo

$$E(R) = 9 \cdot \frac{1}{9} + 9{,}6 \cdot \frac{2}{9} + 10 \cdot \frac{3}{9} + 10{,}5 \cdot \frac{2}{9} + 12 \cdot \frac{1}{9} \approx 10{,}13.$$

$$V(R) \approx 0{,}6289.$$

Exemplo: Distribuição amostral

Considerando os dados do exemplo População de domicílios, encontre a distribuição de probabilidade das estatísticas $\overline{Y}$ e S^2 relacionadas à v.a. renda familiar para uma amostra de tamanho 2 obtida pelo plano AASc.

s	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
P(**s**)	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
$\overline{Y}$	12	21	15	21	30	24	15	24	18
s^2	0	162	18	162	0	72	18	72	0

Exemplo: Distribuição amostral (cont.)

- Distribuição amostral de $\overline{Y}$.

$\overline{Y}$	12	15	18	21	24	30
$P(\overline{y})$	1/9	2/9	1/9	2/9	2/9	1/9

- Distribuição amostral de S^2.

S^2	0	18	72	162
$P(s^2)$	3/9	2/9	2/9	2/9

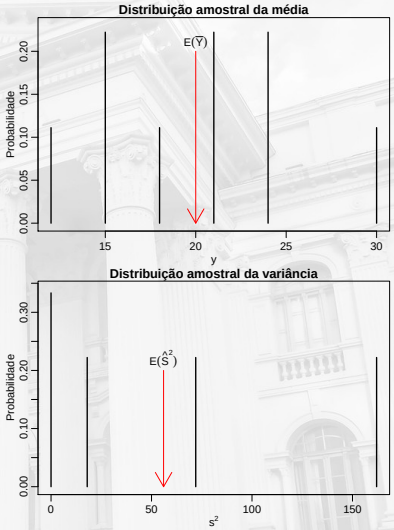

Figura 2. Distribuição amostral.

Exemplo: Distribuição amostral (cont.)

- Note que $\mathrm{E}(\overline{Y}) = 20$ e $\mathrm{Var}(\overline{Y}) = \frac{56}{2} = 28$.
- Note ainda que $\mathrm{E}(S^2) = \frac{504}{9} = 56$.
- Tanto $\mathrm{E}(\overline{Y})$ como $\mathrm{E}(S^2)$ coincidem com os parâmetros populacionais, ou seja,

$$\mu = \frac{12 + 30 + 18}{3} = 20 \quad \text{e} \quad \sigma^2 = \frac{(12-20)^2 + (30-20)^2 + (18-20)^2}{3} = 56.$$

- Esperança do estimador coincide com o valor populacional $\rightarrow$ **estimador não-viciado**.

Comentários

- A distribuição amostral caracteriza probabilisticamente a estatística de interesse.
- Pode ser resumida da mesma forma que qualquer outra distribuição de probabilidade (esperança, variância, covariância, etc).
- Para populações pequenas é fácil de ser obtida. E para populações grandes?
- Nenhuma suposição foi feita sobre a distribuição de probabilidade da v.a.
- Estratégia vista até aqui é impraticável!
- Precisamos de algo mais geral e flexível em termos práticos!!!

Figura 3. Photo by Anna Shvets from Pexels.

Distribuição amostral da média: V.a. Normal

- Seja $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ para $i = 1, \ldots, N$. Suponha que uma amostra aleatória de tamanho n, com valores observados denotados por $y_1, \ldots, y_n$ foi obtida. A distribuição amostral da média $\overline{Y}$ é dada por

$$\overline{Y} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- Segue do fato de que **combinação linear** de Normal é Normal e de que

$$\mathrm{E}(\overline{\mathrm{Y}}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\mathrm{Y_i}) = \frac{n\mu}{n} = \mu.$$

$$\mathrm{V}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{V}(Y_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Exemplo: Salário de pilotos

O salário anual médio dos pilotos de avião pode ser modelado por uma distribuição Normal com média de R$41979,00 e desvio padrão de R$5000,00. Suponha que uma amostra aleatória simples de 50 pilotos seja selecionada.

- Qual é o desvio padrão da média amostral?
- Qual é a probabilidade da média amostral ser maior que R$41979,00?
- Qual é a probabilidade da média amostral não diferir da média populacional em até R$1000,00?
- Como a probabilidade do item anterior seria alterada caso a amostra fosse de tamanho 100?

Exemplo: Salário de pilotos (cont.)

- Qual é o desvio padrão da média amostral?

  $V(\overline{Y}) = \frac{\sigma^2}{n}$. Assim, a variância é $\frac{5000^2}{50}$ e o desvio padrão da média $\frac{5000}{\sqrt{50}}$.

- Qual é a probabilidade da média amostral ser maior que R$41979,00?

$$P(\overline{Y} > 41979) = P\left(Z > \frac{41979 - 41979}{5000/\sqrt{50}}\right) = P(Z > 0) = 0{,}5.$$

- Qual é a probabilidade da média amostral não diferir da média populacional em até R$1000,00?

$$P(40979 < \overline{Y} < 42979) = P\left(\frac{40979 - 41979}{5000/\sqrt{50}} < Z < \frac{42979 - 41979}{5000/\sqrt{50}}\right)$$
$$= P(-1.414 < Z < 1.414) \approx 0{,}842.$$

Exemplo: Salário de pilotos (cont.)

- Como a probabilidade do item anterior seria alterada caso a amostra fosse de tamanho 100?

$$\begin{aligned} \mathrm{P}(40979 < \overline{Y} < 42979) &= \mathrm{P}\left(\frac{40979 - 41979}{5000/\sqrt{100}} < Z < \frac{42979 - 41979}{5000/\sqrt{100}} \right) \\ &= \mathrm{P}(-2 < Z < 2) \approx 0{,}954. \end{aligned}$$

Exemplo: Acesso à internet

Uma pesquisa divulgou que 56% das famílias brasileiras têm acesso à internet. Suponha que esta seja a verdadeira proporção populacional $p = 0{,}56$ e suponha que uma amostra de 300 famílias seja selecionada.

- Apresente a distribuição amostral de $\hat{p}$, em que $\hat{p}$ é a proporção amostral de famílias com acesso à internet.
- Qual a probabilidade de a proporção amostral não diferir da populacional em mais de $0{,}03$?
- Responda o item anterior considerando amostras de tamanho 600 e 1000.

Exemplo: Acesso à internet (comentários)

- Note que agora temos que a distribuição da v.a. **não** é Normal.
- Y - acesso à internet (SIM ou NÃO).
- $Y \sim \text{Ber}(p)$ com p sendo a probabilidade de ter acesso à internet.
- Sabemos que $\text{E}(Y) = p$ e $\text{V}(Y) = p(1-p)$.
- Sendo $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, podemos facilmente obter

$$\text{E}(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} \text{E}(Y_i) = \frac{np}{n} = p.$$

$$\text{V}(\hat{p}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{V}(Y_i) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

- Conseguimos obter a média e a variância de $\hat{p}$, mas e a distribuição?

Teorema do Limite Central

Teorema Lindeberg-Levy: Seja $Y_1, \ldots, Y_n$ uma amostra aleatória independente e identicamente distribuída com $\mathrm{E}(Y_i) = \mu$ e $\mathrm{V}(Y_i) = \sigma^2 < \infty$. Então,

$$\sqrt{n}\left(\frac{\overline{Y} - \mu}{\sigma}\right) \xrightarrow{D} Z \sim \mathrm{N}(0,1), \quad \text{para} \quad n \to \infty.$$

Forma alternativa: $\overline{Y} \sim \mathrm{N}(\mu, \sigma^2/n)$.

Ilustração computacional

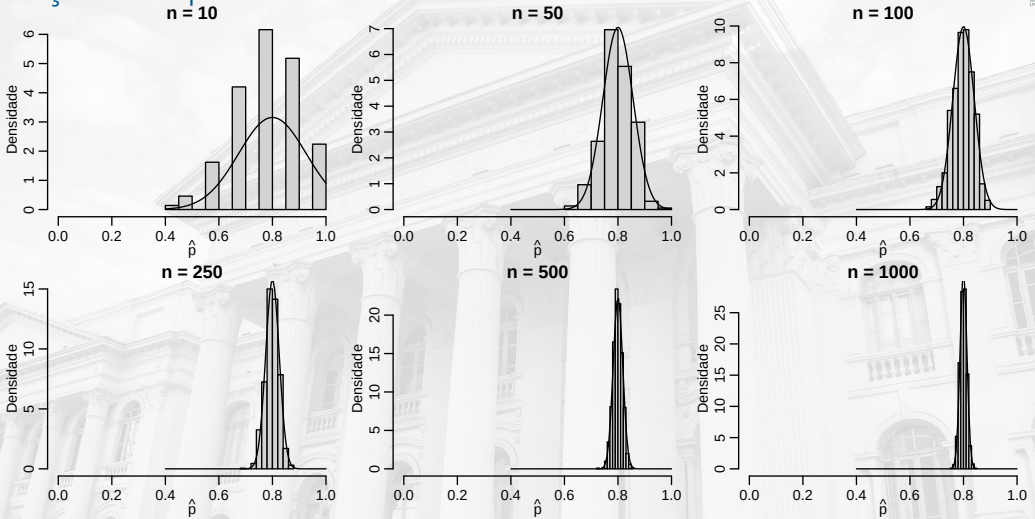

Figura 4. Distribuição amostral da proporção conforme tamanho da amostra.

Exemplo: Acesso à internet (cont.)

- Apresente a distribuição amostral de $\hat{p}$, em que $\hat{p}$ é a proporção amostral de famílias com acesso à internet.
- Usando o TLC, temos

$$\hat{p} \sim \mathrm{N}\left(p, \frac{p(1-p)}{n}\right).$$

- Qual a probabilidade de a proporção amostral não diferir da populacional em mais de 0,03?

$$\mathrm{P}(0{,}53 < \hat{p} < 0{,}59) = \mathrm{P}(-1{,}046 < Z < 1{,}046) \approx 0{,}704.$$

$$\mathrm{P}\left(\frac{0{,}53 - 0{,}56}{\sqrt{0{,}56(1-0{,}56)/300}} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < \frac{0{,}59 - 0{,}56}{\sqrt{0{,}56(1-0{,}56)/300}}\right).$$

Exemplo: Acesso à internet (cont.)

- Responda o item anterior considerando amostras de tamanho 600 e 1000.
- $n = 600 \rightarrow \mathrm{P}(0{,}53 < \hat{p} < 0{,}59) = \mathrm{P}(-1{,}480 < Z < 1{,}480) \approx 0{,}861.$

$$\mathrm{P}\left(\frac{0{,}53 - 0{,}56}{\sqrt{0{,}56(1-0{,}56)/600}} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < \frac{0{,}59 - 0{,}56}{\sqrt{0{,}56(1-0{,}56)/600}} \right).$$

- $n = 1000 \rightarrow \mathrm{P}(0{,}53 < \hat{p} < 0{,}59) = \mathrm{P}(-1{,}911 < Z < 1{,}911) \approx 0{,}944.$

$$\mathrm{P}\left(\frac{0{,}53 - 0{,}56}{\sqrt{0{,}56(1-0{,}56)/1000}} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < \frac{0{,}59 - 0{,}56}{\sqrt{0{,}56(1-0{,}56)/1000}} \right).$$

Resumo

- Distribuição amostral da média.
- Distribuição amostral aproximada da proporção.
- Teorema Central do Limite.

Figura 5. Retirada do Google imagens.