Distribuições amostral

Prof. Wagner Hugo Bonat

Departamento de Estatística
Universidade Federal do Paraná

Distribuição amostral de médias

Definições

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

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

onde N é o tamanho da população.

- **Unidade elementar**: refere-se a qualquer elemento $i \in \mathrm{U}$.
- **Variável**: característica a ser observada em cada unidade elementar $\rightarrow$ variável aleatória. Notação $Y_i, \quad i \in \mathrm{U}$.
- 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 $U = \{1,2,3\}$ e que 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
n^o 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.

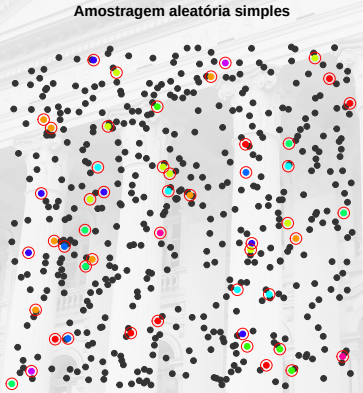


Figura 1. Amostragem aleatória simples.

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 a 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

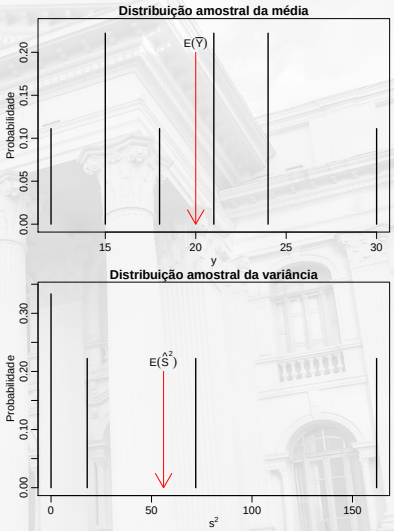


Figura 2. Distribuição amostral.

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 idênticamente 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

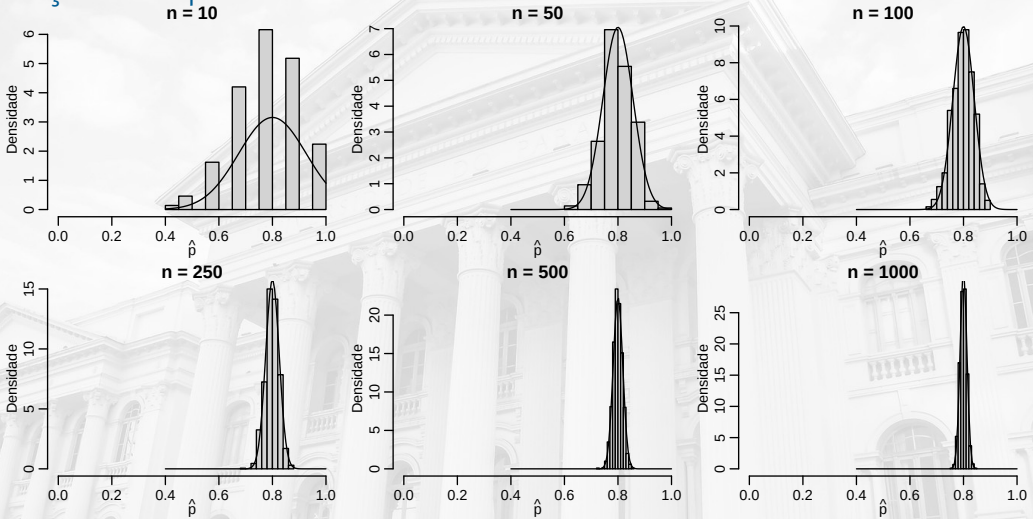


Figura 4. Distribuição amostral da proporção conforme tamanho da amostra.

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.
- $n = 600 \rightarrow \mathrm{P}(0{,}53 < \hat{p} < 0{,}59) = \mathrm{P}(-1{,}480 < Z < 1{,}480) \approx 0{,}861.$

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

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

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

Distribuição amostral de estatísticas importantes

Distribuição amostral da média

- Sejam $Y_1, \ldots, Y_n$ v.a.'s independentes e identicamente distribuídas (iid) com distribuição desconhecida, porém com média $\mathrm{E}(Y_i) = \mu$ e variância $\mathrm{V}(Y_i) = \sigma^2 < \infty$. Para amostras grandes o TLC nos diz que

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

- E para outras estatísticas de interesse?

Distribuição amostral da média

- Sejam $Y_1, \ldots, Y_n$ v.a.'s independentes e identicamente distribuídas (iid) com distribuição desconhecida, porém com média $\mathrm{E}(Y_i) = \mu$ e variância $\mathrm{V}(Y_i) = \sigma^2 < \infty$. Para amostras grandes o TLC nos diz que

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

- E para outras estatísticas de interesse?
- De forma geral é dificil obter a distribuição amostral de outras estatísticas.
- Porém para v.a.'s Normais temos alguns resultados importantes.

Amostragem de v.a.'s Normais e estatísticas relacionadas

Sejam $Y_1, \ldots, Y_n$ v.a.'s iid com distribuição $\mathrm{N}(\mu, \sigma^2)$.

Algumas estatísticas relacionadas são:

- Média amostral $\rightarrow \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- Variância amostral $\rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ ou $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.
- Estatística t-Student $\rightarrow t = \frac{\overline{Y}-\mu}{S/\sqrt{n}}$.
- Sendo duas v.a.'s Normais com variância S^2_{Y1} e S^2_{Y2}, respectivamente. A razão $\frac{S^2_{Y1}}{S^2_{Y2}}$ é chamada de estatística F.

Distribuição χ^2

- Sendo $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, então

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \text{onde} \quad n-1 \quad \text{são os graus de liberdade.}$$

- Função densidade probabilidade $Y_s \sim \chi^2_k$

$$f(y_s; k) = \frac{y_s^{\frac{k}{2}-1} e^{-\frac{y_s}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \quad \text{para} \quad k \in N \quad \text{e} \quad y_s > 0.$$

- Cálculo de probabilidades $\rightarrow$ tabelas (só para as caudas) ou softwares estatísticos.

Ilustração computacional

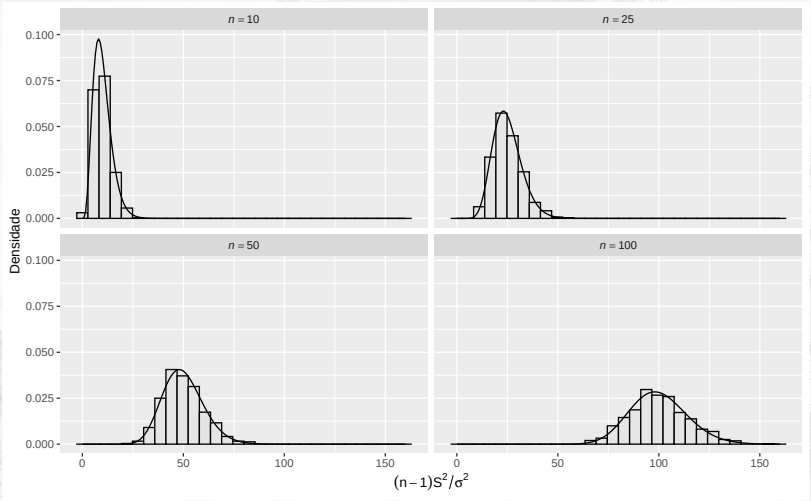


Figura 5. Distribuição amostral da estatística qui-quadrado.

Aplicações e propriedades da distribuição χ^2

- Muito comum em testes de hipóteses:
  - Independência em tabelas de contingência.
  - Bondade de ajuste.
  - Razão de verossimilhanças.
  - Log-rank.
  - Cochran-Mantel-Haenszel.
- Soma de quadrados de $n - 1$ Normais padrão independentes.
- Caso particular da distribuição Gama.

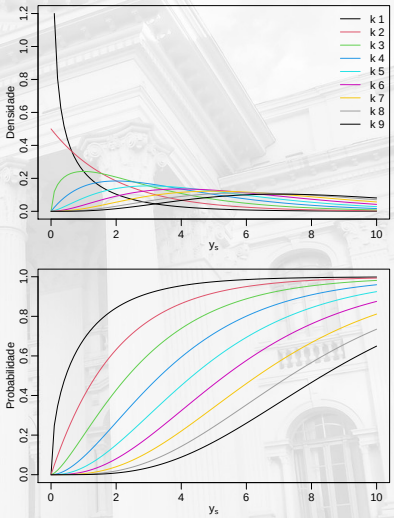


Figura 6. Distribuição qui-quadrado.

Exemplo: Bateria para celular

Uma empresa desenvolveu uma nova bateria para celular. Em média a bateria dura 60 horas com desvio-padrão de 4 horas. Suponha que o fabricante efetua um controle da qualidade das baterias onde são selecionadas aleatóriamente 7 baterias. Supondo que a duração das baterias pode ser adequadamente modelada pela distribuição Normal. Calcule

- Probabilidade da variância amostral ser maior que 16 horas.
- Probabilidade da variância amostral estar entre 4 e 36 horas.
- Probabilidade da variância amostral ser menor do que 4 horas.

Figura 7. Foto de Tyler Lastovich no Pexels.

Exemplo: Bateria para celular (cont.)

- Probabilidade da variância amostral ser maior que 16 horas.

$$\mathrm{P}(S^2 > 16) = \mathrm{P}\left((n-1)\frac{s^2}{\sigma^2} > (7-1)\frac{16}{16}\right) = P(\chi^2_{7-1} > 6) \approx 0{,}423.$$

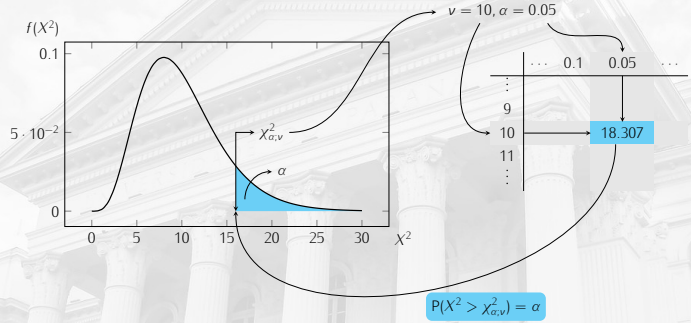
- Probabilidade da variância amostral estar entre 4 e 36 horas.

$$\mathrm{P}(4 < S^2 < 36) = \mathrm{P}\left((7-1)\frac{4}{16} < \chi^2_{7-1} < (7-1)\frac{36}{16}\right) = P(1{,}5 < \chi^2_{7-1} < 13{,}5) \approx 0{,}923.$$

- Probabilidade da variância amostral ser menor do que 4 horas.

$$\mathrm{P}(S^2 < 4) = \mathrm{P}\left(\chi^2_{7-1} < (7-1)\frac{4}{16}\right) = P(\chi^2_{7-1} < 1{,}5) \approx 0{,}040.$$

Consulta da tabela χ^2



Pontos percentuais da distribuição χ^2 com áreas na calda direita.

ν/α	$\alpha = 0.995$	0.99	0.975	0.95	0.9	0.5	0.1	0.05	0.025	0.01	0.005
$\nu = 1$	0.000	0.000	0.001	0.004	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	3.357	7.779	9.488	11.143	13.277	14.860

Figura 8. Consulta da tabela qui-quadrado.

Distribuição t-Student

- Sendo $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ e $\overline{Y}$ e S^2 a média e a variância amostral a v.a.

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

  t_{n-1} denota a distribuição t-Student com $n-1$ graus de liberdade.

- Função densidade probabilidade

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

  onde $\nu \in \mathbb{N}$ é o número de graus de liberdade e $\Gamma(\cdot)$ é a função gama.

- Cálculo de probabilidades $\rightarrow$ tabelas similares a da distribuição Normal ou softwares estatísticos.

Ilustração computacional

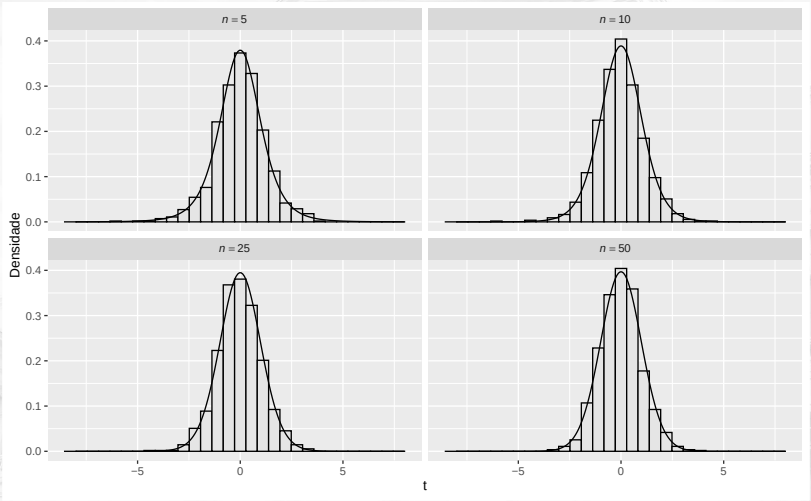


Figura 9. Distribuição amostral da estatística t-Student.

Aplicações e propriedades da distribuição *t*-Student

- Teste t e suas variações.
- Intervalo de confiança para a média.
- Simétrica em forma de sino (igual a Normal).
- Caudas mais pesadas que a Normal.
- Descreve o comportamento da razão de algumas v.a.'s.
- Desenvolvida por William Sealy Gosset (sob o pseudo nome de Student).

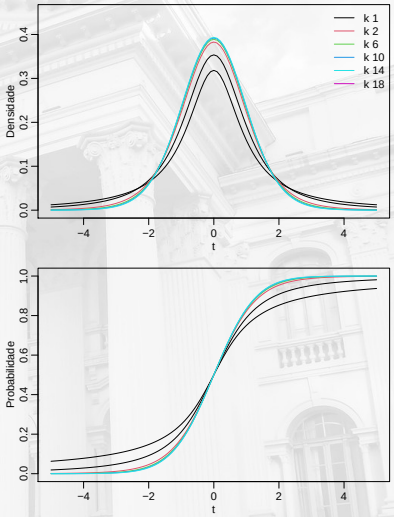


Figura 10. Distribuição t-Student.

Exemplo: Acumputura

Suponha que um experimento foi realizado para avaliar a efetividade do uso da acumputura para aliviar a dor. A taxa sensorial de 15 pacientes foi medida resultando em uma média de 8,22 e um desvio-padrão de 1,67. Supondo que a distribuição Normal é adequada para a variável de interesse obtenha um intervalo t_1 e t_2, tal que a probabilidade deste intervalo conter a média populacional seja de aproximadamente 0,95.

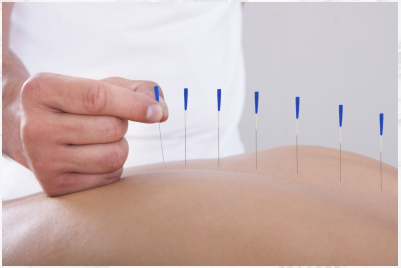


Figura 11. Extraído de www.health.harvard.edu.

Exemplo 2 (cont.)

- Precisamos obter valores $\overline{y}_{li}$ e $\overline{y}_{ls}$, tal que

$$\mathrm{P}(\overline{y}_{li} < \mu < \overline{y}_{ls}) = 0{,}95.$$

- Vamos padronizar

$$\mathrm{P}\left(\frac{\overline{y}_{li} - \mu}{s/\sqrt{n}} < \frac{\overline{y} - \mu}{s/\sqrt{n}} < \frac{\overline{y}_{ls} - \mu}{s/\sqrt{n}}\right) = \mathrm{P}\left(t_1 < \frac{\overline{y} - \mu}{s/\sqrt{n}} < t_2\right) = 0{,}95$$

- Note que a distribuição $t-$Student é simétrica, então vamos focar apenas em intervalos simétricos, o que implica que $t = t_2 = -t_1$. Assim os limites são dados por

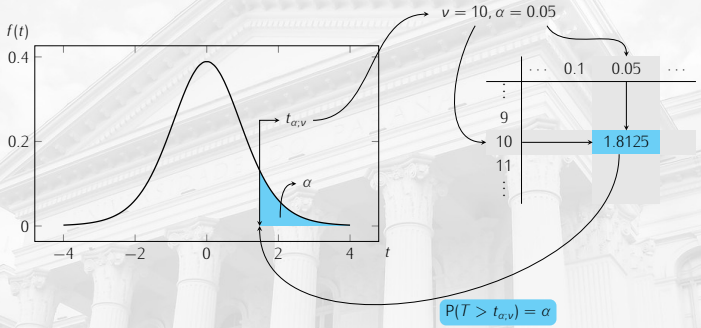
$$\overline{y} \pm t_{0{,}05/2}\frac{s}{\sqrt{n}},$$

onde $t_{0{,}05/2}$ é o valor da distribuição t-Student com $n-1$ graus de liberdade.

- Usando a tabela da distribuição t-Student com 14 graus de liberdade, temos

$$8{,}22 \pm 2{,}14\frac{1{,}67}{\sqrt{15}} \approx [7{,}30; 9{,}14].$$

Consulta da tabela t-Student



Pontos percentuais da distribuição t de Student com áreas na calda direita.

ν/α	$\alpha = 0.4$	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
$\nu = 1$	0.3249	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567	127.3213	318.3088	636.6192
2	0.2887	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248	14.0890	22.3271	31.5991
3	0.2767	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145	12.9240
4	0.2707	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041	5.5976	7.1732	8.6103
5	0.2672	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321	4.7733	5.8934	6.8688

Figura 12. Consulta da tabela t-Student.

Distribuição F de Snedecor

- Sejam $Y_{1i} \sim \mathrm{N}(\mu_1, \sigma_1^2)$ e $Y_{2i} \sim \mathrm{N}(\mu_2, \sigma_2^2)$. Com média e variância amostrais $\overline{Y}_1$, $\overline{Y}_2$, S_1^2 e S_2^2. Suponha ainda que amostras de tamanho n_1 e n_2 estão disponíveis de Y_1 e Y_2. Se $\sigma_1^2 = \sigma_2^2$, então temos que a v.a.

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1},$$

em que F_{n_1-1,n_2-1} denota a distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade.

- Função densidade probabilidade

$$f(y) = \sqrt{\frac{(d_1 y)^{d_1} d_2^{d_2}}{(d_1 y + d_2)^{d_1+d_2}}} \Big/ y\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right) \quad \text{para} \quad y > 0,$$

onde d_1 e d_2 são os graus de liberdade do numerador e denominador e $\mathrm{B}(\cdot)$ é a função beta.

- Cálculo de probabilidades $\rightarrow$ tabelas ou softwares estatísticos.

Ilustração computacional

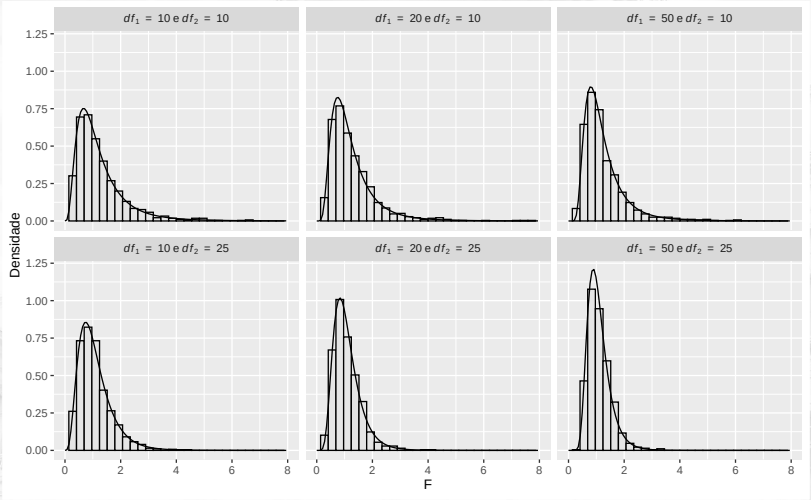


Figura 13. Distribuição amostral da estatística F.

Aplicações e propriedades da distribuição F

- Teste F para igualdade de variâncias.
- ANOVA - Analise de variância.
- Modelos de regressão.
- Razão entre v.a.'s qui-quadrado.
- Também conhecida como distribuição de Fisher-Snedecor's.
- Se $\sigma^2_{Y_1} \neq \sigma^2_{Y_2}$ a estatística F ainda tem distribuição F, porém não central.

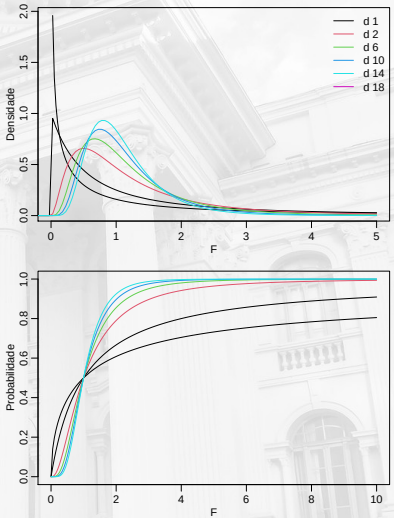
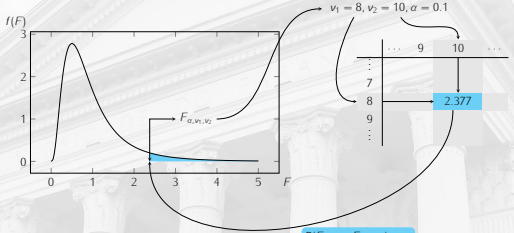


Figura 14. Distribuição F de Snedecor.

Consulta da tabela F de Snedecor



Pontos percentuais da distribuição F com áreas na calda direita para $\alpha = 0.1$.

	$\nu_2 = 1$	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
$\nu_1 = 1$	39.863	8.526	5.538	4.545	4.060	3.776	3.589	3.458	3.360	3.285	3.177	3.073	2.975	2.927	2.881	2.835	2.791	2.748
2	49.500	9.000	5.462	4.325	3.780	3.463	3.257	3.113	3.006	2.924	2.807	2.695	2.589	2.538	2.489	2.440	2.393	2.347
3	53.593	9.162	5.391	4.191	3.619	3.289	3.074	2.924	2.813	2.728	2.606	2.490	2.380	2.327	2.276	2.226	2.177	2.130
4	55.833	9.243	5.343	4.107	3.520	3.181	2.961	2.806	2.693	2.605	2.480	2.361	2.249	2.195	2.142	2.091	2.041	1.992
5	57.240	9.293	5.309	4.051	3.453	3.108	2.883	2.726	2.611	2.522	2.394	2.273	2.158	2.103	2.049	1.997	1.946	1.896
6	58.204	9.326	5.285	4.010	3.405	3.055	2.827	2.668	2.551	2.461	2.331	2.208	2.091	2.035	1.980	1.927	1.875	1.824
7	58.906	9.349	5.266	3.979	3.368	3.014	2.785	2.624	2.505	2.414	2.283	2.158	2.040	1.983	1.927	1.873	1.819	1.767

Figura 15. Consulta da tabela F de Snedecor

Exemplo: Acumputura

Suponha que um experimento foi realizado com dois grupos para avaliar a efetividade do uso da acumputura para aliviar a dor. A taxa sensorial foi medida para o grupo 1 em 5 pacientes e para o 2 em 8 pacientes. Suponha que as variâncias amostrais foram $s_1^2 = 4{,}44$ e o $s_2^2 = 1{,}5$. Assumindo que as variâncias populacionais são iguais, qual a probabilidade de ocorrer a razão $\frac{s_1^2}{s_2^2}$ ou uma mais extrema?

$$P\left(\frac{s_1^2}{s_2^2} > \frac{4{,}44}{1{,}5}\right) = \mathrm{P}(F_{5-1,8-1} > 2{,}96) \approx 0{,}1.$$

Você considera a suposição de igualdade de variâncias plausível dado o resultado do experimento?

Relações entre as distribuições

- Uma v.a. Normal padrão ao quadrado tem distribuição χ^2 com $gl = 1$.
- Uma v.a. t-Student ao quadrado tem distribuição F com $gl = 1$.
- Razão de duas v.a. χ^2 divida pelos seus gl's tem distribuição F_{n_1,n_2}.
- Distribuição F converge para a χ^2 com $n \to \infty$.
- Existem extensões não-centrais (modelo de locação e escala).
- Distribuição t-Student é uma alternativa robusta a Normal.
- Todas são relacionadas a Normal e quando gl cresce vão convergir para a Normal.

Estimação

Estimação estatística

Falar sobre **população** a partir da observação da **amostra**.

- Amostra? De qual tamanho?
- Como estimar?
- Como expressar incerteza?
- O que é “estimar bem”?

Figura 16. Analogia ao processo de estimação. Extraído de bestbowreviews.com.

Um exemplo: cardápio vegano

- Um restaurante deseja caracterizar o perfil de seus clientes.
- Questionário para uma *amostra* de clientes.
- **Q1:** Há interesse por opções veganas?
- Qual a proporção que prefere pratos veganos?

1. Dados (0/1) do questionário podem indicar um valor, por ex., 0.12 e sua incerteza: 0.12 ± 0.035 ou $(0.085, , 0.155)$.
2. Quantos questionários?

Figura 17. Foto de Pexels.

Exemplo: caracterização dos clientes

Q2: Qual será a *idade média* dos clientes?

1. Dados de idades nos questionários: distribuição normal(?)
2. Pode-se estimar, por ex., 32 anos com alguma incerteza: 32 ± 2.5 ou $(29.5\ ,\ 34.5)$.
3. Quantos questionários?
4. Diferentes opções para *estimar* o valor de idade "típica" dos clientes: *média*, *mediana*, *ponto médio*, etc. Qual as características de cada **estimador**?

Figura 18. Foto de Adrienn no Pexels.

Exemplo: tempo de refeição

Q3: Qual a *duração média* das refeições?

1. Dados do questionário: Distribuição para o tempo de permanência: (Normal(?), Gama(?))
2. Pode-se estimar, por ex., 25 min e sua incerteza: $(22, 30)$.
3. Quantos questionários?
4. Qual as características de cada **estimador**?
5. Mas, qual(ais) estimador(es)?

Figura 19. Foto de Andrea Piacquadio no Pexels.

Elementos da estimação

- Contexto do estudo: a(s) variável(eis) envolvidas.
- Comportamento (distribuição) desta variável.
- Característica (parâmetro) de interesse.
- Definição da amostra.
- Obtenção dos dados.
- *Estimação do parâmetro.*
- *Expressão da incerteza.*
- Interpretação e conclusões.

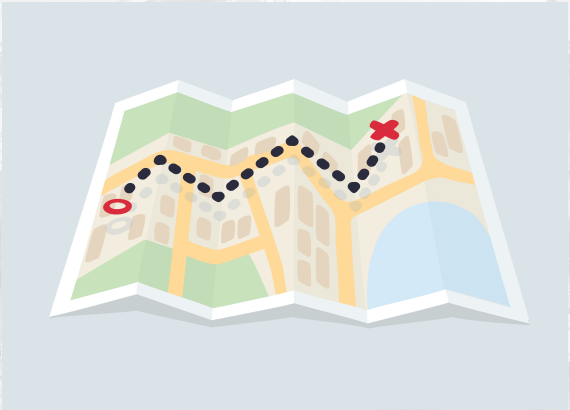


Figura 20. *Roadmap*.
https://getnave.com/blog/kanban-roadmap/

Inferência frequentista

- Objeto de inferência: **distribuição amostral**.
- A estimativa pontual é um resumo desta distribuição.
- Intervalos entre quantis representam a incerteza sobre o valor estimado.
- Compara-se estimadores concorrentes pelas características de suas distribuições amostrais.
- E para tudo isto:
  é preciso saber como estimar.

Figura 21. Distribuição amostral de diferentes estimadores de um parâmetro.

Estimação estatística

Falar sobre **população** $Y \sim \text{Dist.}_y(\theta)$ a partir da observação da **amostra** $\hat{\theta}(y_1, \ldots, y_n) \sim \text{Dist.Am.}_{\hat{\theta}}(\theta)$.

1. Como expressar incerteza?
   Estimação pontual e intervalar.
2. Amostra? De qual tamanho?
   Determinação do tamanho da amostra.
3. O que é "estimar bem"?
   Propriedades dos estimadores.
4. Como estimar?
   Métodos de estimação.

Idéias válidas em contextos mais gerais.

Figura 22. Distribuição amostral de diferentes estimadores de um parâmetro.

Estimação pontual e intervalar

Noções iniciais

Notação e definições

- $\mathbf{Y} = (Y_1, \ldots, Y_n)$ denota um vetor de v.a.'s independentes e identicamente distribuídas.
- Cada $Y_i \sim f(\boldsymbol{\theta})$ onde f denota a função densidade de probabilidade ou função de probabilidade e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ é um vetor de p parâmetros populacionais.
- $\mathbf{y} = (y_1, \ldots, y_n)$ denota o vetor de valores observados da v.a. Y.
- **Estatística**: uma estatística T é uma v.a. $T = t(\mathbf{Y})$, definida como **função da amostra**, que não depende do vetor de parâmetros $\boldsymbol{\theta}$.
- Uma **estatística** T é um **estimador** para θ se o valor realizado $t = t(\mathbf{y})$ é usado como uma **estimativa** para o valor de $\boldsymbol{\theta}$, então denotado por $\hat{\boldsymbol{\theta}}$.
- A distribuição de probabilidade de $T(\mathbf{Y}) \rightarrow$ **Distribuição amostral**.

Exemplo: idade média dos frequentadores do restaurante

Vai se tomar uma amostra de $n = 5$.

- $\mathbf{Y} = (Y_1, \ldots, Y_n)$ é definida pelas idades dos frequentadores.
- Cada idade vem de uma distribuição **da v.a. observada** $Y_i \sim f(\boldsymbol{\theta}) = \mathrm{N}(\mu, 4^2)$ com $\boldsymbol{\theta} = (\mu)$.
- A **estatística**: $T = t(\mathbf{Y}) = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y} = \hat{\mu}$ é um **estimador** da média.

Coletam-se os dados $\mathbf{y} = (y_1 = 31, y_2 = 30, y_3 = 32, y_4 = 37, y_5 = 30)$

- A **estimativa** obtida com esta amostra $\hat{\mu} = \overline{y} = 32$,
- Se a amostra é aleatória então esta estimativa é uma v.a. que tem uma A distribuição de probabilidade chamada de **distribuição amostral**.

$$\overline{Y} \sim \mathrm{N}\left(\mu, \frac{4^2}{5}\right) \text{ ou, equivalentemente, } \frac{\overline{Y} - \mu}{4/\sqrt{5}} \sim \mathrm{N}(0, 1).$$

Exemplo: estimadores para distribuição Normal

- Modelo de probabilidade: $Y_i \sim \text{N}(\mu, \sigma^2) \rightarrow \boldsymbol{\theta} = (\mu, \sigma^2)$.
- Estimadores e estimativas

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \rightarrow \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad \rightarrow \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

- Distribuições amostrais

$$\sigma^2 \text{ conhecido: } \overline{Y} \sim \text{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{ou} \quad \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \text{N}(0, 1)$$

$$\sigma^2 \text{ desconhecido: } \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim \text{t}_{n-1} \quad \text{e} \quad (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Exemplo: estimadores para distribuição de Bernoulli

- Modelo de probabilidade: $Y_i \sim \text{Ber}(p) \rightarrow \boldsymbol{\theta} = p$.
- Estimadores e estimativas

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \rightarrow \quad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- Distribuição amostral (aproximada TLC)

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

A incerteza na estimação

A estimativa pontual

- Fornece apenas **um valor** plausível de ser o verdadeiro valor do parâmetro.
- Não considera a **incerteza** devido a termos apenas uma amostra.

Como expressar a incerteza?

Baseado na **distribuição amostral** pode-se obter uma faixa de valores com determinada probabilidade de conter o parâmetro $\rightarrow$ **intervalo de confiança.**

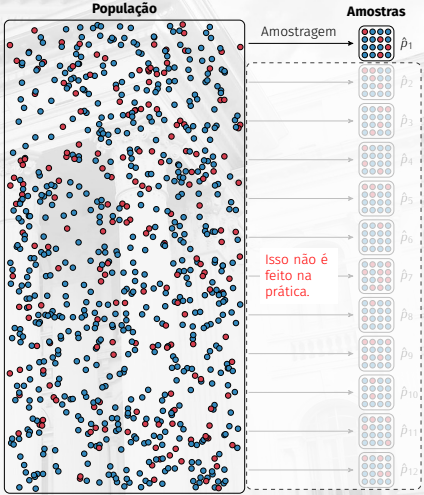


Figura 23. Processo de inferência na prática.

Intervalo de confiança para a média

Intervalos de confiança para a média quando σ^2 é conhecido

- Seja $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ e suponha que σ^2 é conhecido.
- Neste caso, temos que

$$\overline{Y} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{ou} \quad \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

- Fixando uma probabilidade $1 - \alpha$ podemos encontrar $\overline{y}_{LI}$ e $\overline{y}_{LS}$, tal que

$$\mathrm{P}(\overline{y}_{LI} < \mu < \overline{y}_{LS}) = 1 - \alpha.$$

- Vários pares $\overline{y}_{LI}$ e $\overline{y}_{LS}$ existem, então prefere-se aqueles que dão **intervalo simétrico** em relação a μ.

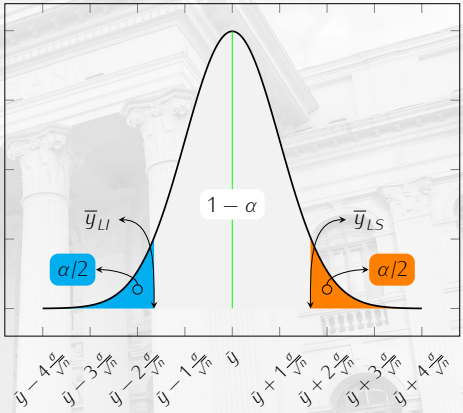


Figura 24. Intervalo de confiança para a média.

Obtenção do intervalo para μ

- Definimos limites Z na distribuição amostral padronizada

$$\mathrm{P}\left(z_{LI} < \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} < z_{LS}\right) = 1 - \alpha.$$

- Agora deixamos apenas μ no centro para obtermos,

$$\mathrm{P}\left(\overline{y} - z_{LI}\frac{\sigma}{\sqrt{n}} < \mu < \overline{y} + z_{LS}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

- Como deseja-se intervalos simétricos, então $\mathrm{abs}(z_{LI}) = \mathrm{abs}(z_{LS}) = z_{\alpha/2}$. Assim,

$$\mathrm{P}\left(\overline{y} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

- $z_{\alpha/2}$ é o quantil da distribuição Normal padrão para o valor de $1 - \alpha$ fixado.

Margem de erro e nível de confiança

- Chamamos de **erro máximo provável** ou **margem de erro** a quantidade

$$e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

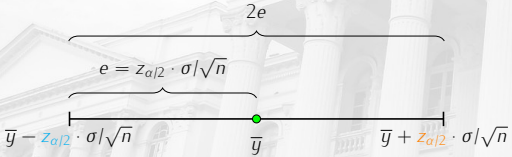
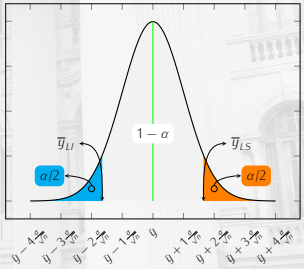


Figura 25. Margem de erro do intervalo de confiança.

- $z_{\alpha/2}$ é chamado de **valor crítico**. É o valor z que produz uma área de $\alpha/2$ na cauda superior da distribuição Normal padrão.
- Chamamos a quantidade $1 - \alpha$ de **coeficiente** de confiança ou **nível de confiança** do intervalo.



Exemplo: idade média dos frequentadores do restaurante

Y : idade dos frequentadores

$Y \sim \mathrm{N}(\mu, \sigma^2 = 4^2)$

Dados: $\mathbf{y} = (31, 30, 32, 37, 30)$

- **estimativa:** $\hat{\mu} = \overline{y} = 32$
- escolha do **nível de confiança**: 95% ($1 - \alpha = 0{,}95$ e $\alpha/2 = 0.025$)
- **valor-z**: $z_{\alpha/2} = 1.96$
- **erro máximo provável:** $e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{4}{\sqrt{5}} = 3.51$
- **intervalo de confiança (95%):** $\overline{y} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 32 \pm 3.51$

$$\mathrm{IC}_{0,95}(\mu) : (28.5 \ , \ 35.5)$$

Construção do intervalo usando a distribuição amostral

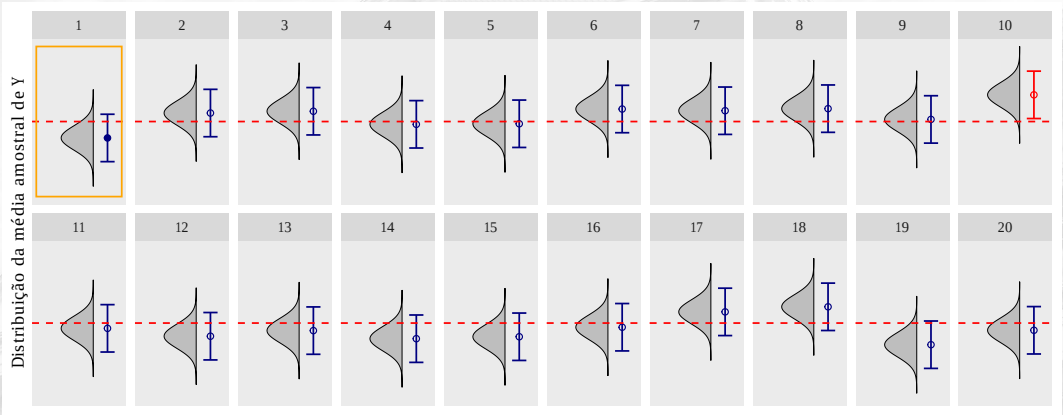


Figura 27. Construção do intervalo de confiança a partir da distribuição amostral.

Interpretação frequentista

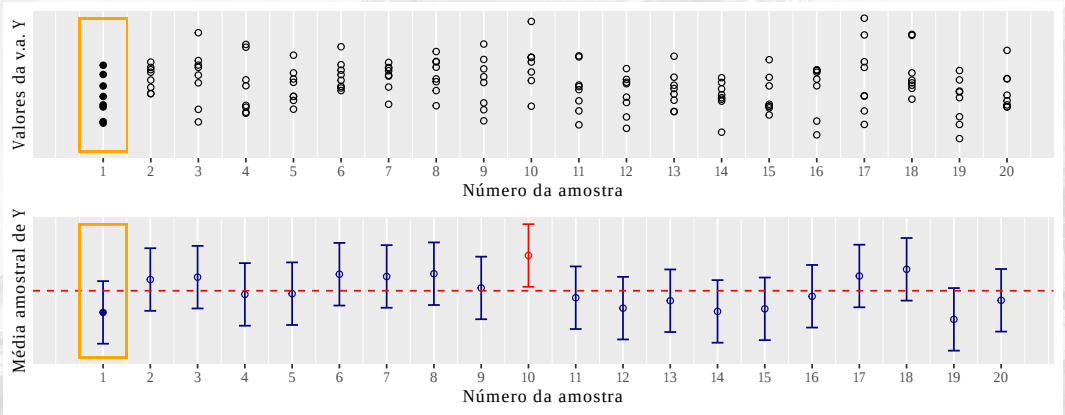


Figura 28. Interpretação frequentista do intervalo de confiança.

Interpretação do intervalo de confiança

Suponha que obtivemos um intervalo de 95% de confiança: $\text{IC}_{95\%}(\mu) = [\overline{y}_{LI}, \overline{y}_{LS}]$.

Interpretação ERRADA de IC

Temos 95% de confiança de que **a média populacional** μ se encontra entre $\overline{y}_{LI}$ e $\overline{y}_{LS}$.

Interpretação CERTA de IC

Temos 95% de confiança de que **o intervalo** entre $\overline{y}_{LI}$ e $\overline{y}_{LS}$ contém a média populacional μ.

Semanticamente as afirmações podem parecer equivalentes, mas a segunda sentença enfatiza o que é crucial: o **intervalo é aleatório** e o **parâmetro é fixo**.

Interpretação de um intervalo de confiança

- Como o intervalo de confiança é calculado a partir de uma **amostra aleatória**, este intervalo **também é aleatório**!
- Isso significa que para cada amostra aleatória que tivermos, um intervalo **diferente** será calculado.
- Como o valor de μ é fixo, é o intervalo que deve conter o valor de μ, e não o contrário.
- Isso significa que se pudessemos obter 100 amostras diferentes, e calcularmos um intervalo de confiança de 95% para cada uma das 100 amostras, esperariamos que 5 destes intervalos **não** contenham o verdadeiro valor da média populacional μ.

Exercício: performance no TOEFL

Uma escola *on-line* de idiomas preparatória para o TOEFL afirma possuir uma excelente pontuação média dos seus alunos no exame. Em uma amostra de aleatória de 50 alunos, a pontuação média foi de 560 pontos. Por estudos anteriores, sabe-se que o desvio-padrão é 25 pontos. Obtenha intervalos de confiança com 90%, 95% e 99% de confiança. Discuta as diferenças.

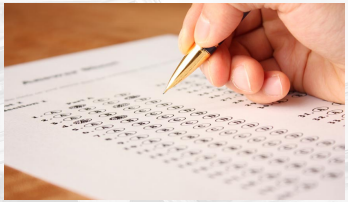


Figura 29. Extraído de elacademy.co.uk.

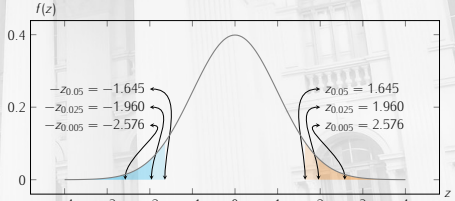


Figura 30. Quantis da distribuição Normal Padrão.

Solução

1. $1-\alpha = 0.9 \rightarrow z_{\alpha/2} = z_{0.05} = 1.645$, então

$$IC_{0.9}(\mu) = \left(560 - 1.645 \cdot \frac{25}{\sqrt{50}}, 560 + 1.645 \cdot \frac{25}{\sqrt{50}}, \right) = (554.2, 565.8).$$

2. $1-\alpha = 0.95 \rightarrow z_{\alpha/2} = z_{0.025} = 1.96$, então

$$IC_{0.95}(\mu) = \left(560 - 1.96 \cdot \frac{25}{\sqrt{50}}, 560 + 1.96 \cdot \frac{25}{\sqrt{50}}, \right) = (553.1, 566.9).$$

3. $1-\alpha = 0.99 \rightarrow z_{\alpha/2} = z_{0.005} = 2.576$, então

$$IC_{0.99}(\mu) = \left(560 - 2.576 \cdot \frac{25}{\sqrt{50}}, 560 + 2.576 \cdot \frac{25}{\sqrt{50}}, \right) = (550.9, 569.1).$$

RESUMO: Intervalos de confiança para média com σ conhecido

1. Verifique se as suposições necessárias estão satisfeitas.
   - Temos uma amostra aleatória simples.
   - σ é conhecido.
   - A população tem distribuição Normal ou $n > 30$ (regra empírica para usar o TLC).
2. Determine o nível de confiança $1 - \alpha$, e encontre o valor crítico $z_{\alpha/2}$.
3. Calcule a margem de erro $e = z_{\alpha/2} \cdot (\sigma/\sqrt{n})$.
4. Calcule $\text{IC}_{1-\alpha}(\mu)$.

Intervalos de confiança para a média quando σ^2 é desconhecido

- Seja $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ e suponha que σ^2 é desconhecido.
- Neste caso, temos que

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

em que t_{n-1} denota a distribuição t-Student com $n-1$ graus de liberdade.

- Argumentos análogos ao caso em que σ^2 é conhecido levam a

$$\mathrm{P}\left(\overline{y} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{y} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha.$$

- $t_{\alpha/2}$ é o valor da distribuição t-Student que produz uma área de $\alpha/2$ na cauda superior da distribuição.

Exercício: gastos com cartão de crédito

Um estudo foi idealizado para estimar a média anual dos débitos de cartão de crédito da população de famílias brasileiras. Uma amostra de $n = 15$ famílias forneceu os saldos de cartões de crédito. A média amostral foi de R$ 5.900,00 e o desvio padrão foi de R$ 3.058,00. Obtenha um intervalo com 95% de confiança.

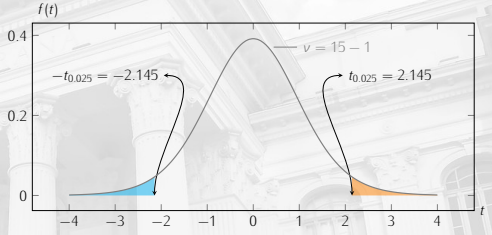


Figura 31. Quantis da distribuição t-Student.

Neste caso $t_{\alpha/2} = t_{0.025} = 2.145$ com $15-1 = 14$ graus de liberdade. Assim, o intervalo de confiança é dado por

$$\text{IC}_{1-0.95}(\mu) = \left(5900 - 2.145 \cdot \frac{3058}{\sqrt{15}}, 5900 + 2.145 \cdot \frac{3058}{\sqrt{15}}\right) \approx (4206.4, 7593.6).$$

RESUMO: Intervalos de confiança para média com σ^2 desconhecido

1. Verifique se as suposições necessárias estão satisfeitas.
   - Temos uma amostra aleatória simples.
   - Temos uma estimativa de s.
   - A população tem distribuição normal ou $n > 30$ (regra empírica para usar o TLC).
2. Determine o nível de confiança $1 - \alpha$, e encontre o valor crítico $t_{\alpha/2}$.
3. Calcule a margem de erro $e = t_{\alpha/2} \cdot (s/\sqrt{n})$.
4. Calcule $\text{IC}_{1-\alpha}(\mu)$.

Intervalo de confiança para a proporção

Intervalos de confiança para a proporção

- Seja $Y_i \sim \text{Ber}(p)$. Neste caso, temos que pelo TLC

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

- Argumentos análogos ao caso da média levam a

$$\text{P}\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha.$$

- Note que p aparece na expressão da margem de erro, o que na prática impossibilita o uso desta equação. Uma opção é substituir p por sua estimativa $\hat{p}$ e assim

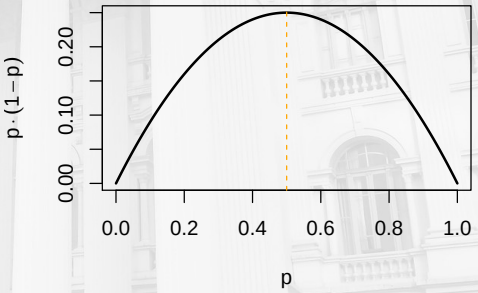
$$\text{P}\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha.$$

Intervalo de confiança para proporção

Uma possível dificuldade nessa abordagem é que em geral não conhecemos o verdadeiro valor de p para calcular o IC.
Quando **não conhecemos** a proporção populacional p, temos duas alternativas:

1. Usar $\hat{p}$ no lugar de p (**estimativa otimista**).
2. Usar $p = 0.5$ (**estimativa conservadora**). Porque quando $p = 0.5$, o termo $p(1-p)$ terá valor máximo.

p	$(1-p)$	$p(1-p)$
0.1	0.9	0.09
0.3	0.7	0.21
0.5	0.5	**0.25**
0.6	0.4	0.24
0.8	0.2	0.16



Exercício: existe aquecimento global?

Uma pesquisa realizada com 1500 adultos foram selecionados aleatoriamente para responder à pergunta se acreditam ou não no aquecimento global. 1050 entrevistados respoderam que sim. Com isso:

1. Para um nível de confiança de 95%, calcule o intervalo de confiança para a verdadeira proporção de pessoas que acreditam no aquecimento global, utilizando: i) $p = \hat{p}$ e ii) $p = 0.5$ e compare os resultados.
2. Com base nesses resultados, podemos concluir que a maioria dos adultos acredita no aquecimento global?

Figura 32. Foto de Markus Spiske no Pexels.

Solução

- Estimativa pontual: $\hat{p} = \frac{1050}{1500} = 0.7$
- Intervalo otimista

$$\text{IC}_{0.95}(p) = \left(0.7 - 1.96\sqrt{\frac{0.7(1-0.7)}{1500}}, 0.7 + 1.96\sqrt{\frac{0.7(1-0.7)}{1500}}\right) \approx (0.677, 0.723).$$

- Intervalo conservador

$$\text{IC}_{0.95}(p) = \left(0.7 - 1.96\sqrt{\frac{0.5(1-0.5)}{1500}}, 0.7 + 1.96\sqrt{\frac{0.5(1-0.5)}{1500}}\right) \approx (0.675, 0.725).$$

- Intervalo conservador será ligeiramente mais largo quando $\hat{p} \neq 0.5$.

RESUMO: Intervalo de confiança para proporção

1. Verifique se as suposições necessárias estão satisfeitas.
   - Temos uma amostra aleatória simples.
   - Há dois resultados possíveis (“sucesso”, “fracasso”).
   - As condições para a distribuição binomial são satisfeitas:
     - As tentativas são independentes.
     - A probabilidade de sucesso p permanece constante.
   - A distribuição normal pode ser usada como aproximação para a distribuição binomial, ou seja, $np \geq 5$ e $np(1-p) \geq 5$.
2. Determine o nível de confiança $1-\alpha$, e encontre o valor crítico $z_{\alpha/2}$.
3. Calcule a margem de erro $e = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$, com $p = \hat{p}$ ou $p = 0.5$.
4. Calcule $\text{IC}_{1-\alpha}(p)$.

Intervalo de confiança para a variância

Intervalo de confiança para variância

- Sendo $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, então a v.a.

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \text{em que } n-1 \text{ são os graus de liberdade.}$$

- Argumentos análogos ao caso da média, levam a

$$\mathrm{IC}_{1-\alpha}(\sigma^2) = \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \right),$$

em que $\chi^2_{\alpha/2,n-1}$ e $\chi^2_{1-\alpha/2,n-1}$ são os quantis da cauda direita e esquerda da distribuição χ^2 com $n-1$ graus de liberdade.

- Note que neste caso o intervalo **não é simétrico**.

Exercício: variabilidade no diâmetro de parafusos

Uma amostra aleatória de 20 parafusos e seus diâmetros são medidos. As medidas em milímetros foram as seguintes.

2.02	1.98	2.08	1.99	2.03
1.94	2.00	2.07	1.95	2.05
2.09	2.03	1.99	1.99	2.01
1.95	2.04	1.96	1.99	2.03

Encontre um intervalo com 90% de confiança para σ^2.

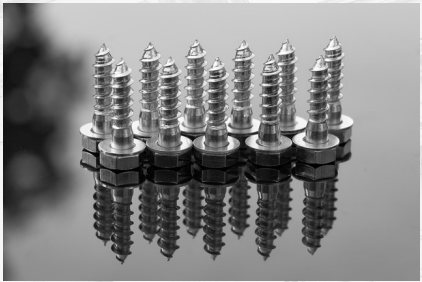


Figura 33. Foto de Pexels.

Solução

- Média e variância amostral

$$\overline{y} = 2.0095 \quad \text{e} \quad s^2 = 0.0019.$$

- Quantis da distribuição χ^2

$$\chi^2_{19,0.95} = 30.1435$$
$$\chi^2_{19,0.05} = 10.117.$$

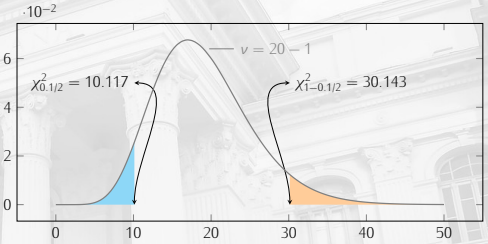


Figura 34. Quantis da distribuição χ^2.

- Assim, o intervalo de confiança é

$$\text{IC}_{0.9}(\sigma^2) = \left(\frac{(20-1)\cdot 0.0019}{30.14353}, \frac{(20-1)\cdot 0.0019}{10.11701} \right) = (0.0012, 0.0035)\,.$$

Considerações finais

Comentários finais

Em resumo

- **Intervalos de confiança** são formas de expressar incerteza.
- Os intervalos são obtidos através de quantis com base na distribuição amostral.
- Esta forma de *raciocínio* (paradigma) é chamada de **frequentista**.

Alguns tópicos adicionais

- Expressões de outros intervalos.
- Intervalos unilaterais.
- Intervalos conjuntos.
- Intervalos com diferentes probabilidades nas causas.
- Outros *paradigmas* de inferência.