Estimação pontual e intervalo de confiança

Prof. Paulo Justiniano R. Jr

Departamento de Estatística
Universidade Federal do Paraná

Conteúdo

Estimação pontual e intervalar

- Terminologia.
- Intervalo de confiança para a média.
- Intervalo de confiança para a proporção.
- Intervalo de confiança para a variância.

Figura 1. Analogia ao processo de estimação. Extraído de pixabay.

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 θ, então denotado por $\hat{\theta}$.
- A distribuição de probabilidade de $T(\mathbf{Y}) \rightarrow$ **Distribuição amostral**.

Exemplo: idade média dos frequentadores do restaurante

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

- $\mathrm{Y} = (Y_1, \ldots, Y_n)$ 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**: $T = t(\mathrm{Y}) = \frac{\sum_{i=1}^n Y_i}{n} = \overline{Y} = \hat{\mu}$ 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 distribuição de probabilidade chamada de **distribuição amostral**.

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

Exemplo: estimadores para distribuição Normal

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

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

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

Exemplo: estimadores para distribuição de Bernoulli

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

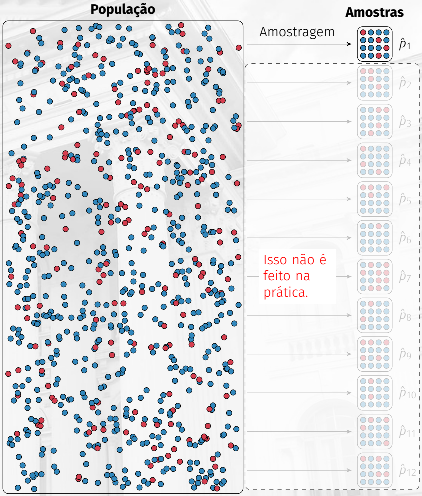


Figura 2. Processo de inferência na prática.

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

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

- 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.

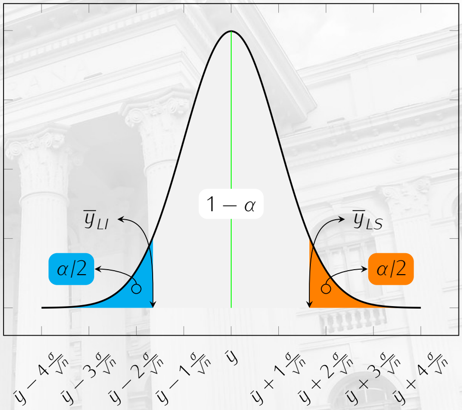


Figura 3. Intervalo de confiança para a média.

Obtenção do intervalo para μ

- Definimos limites Z na distribuição amostral padronizada

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

- Agora deixamos apenas μ no centro para obtermos,

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

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

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

- $z_{\alpha/2}$ 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

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

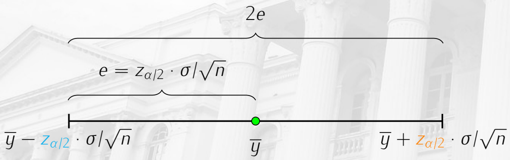
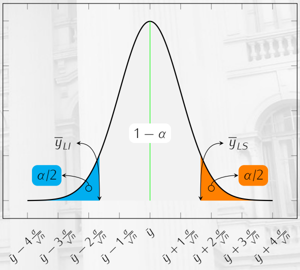


Figura 4. Margem de erro do intervalo de confiança.

- $z_{\alpha/2}$ 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

$$Y: \text{ idade dos frequentadores}$$
$$Y \sim \mathrm{N}(\mu, \sigma^2 = 4^2)$$

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%**):** $\overline{y} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 32 \pm 3.51$

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

Interpretação frequentista

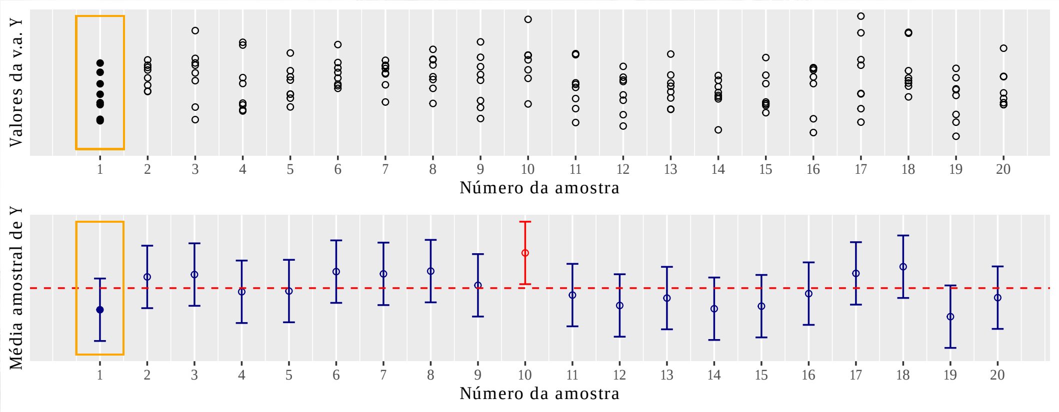


Figura 6. Interpretação frequentista do intervalo de confiança.

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.

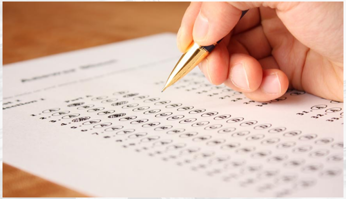


Figura 7. Extraído de elacademy.co.uk.

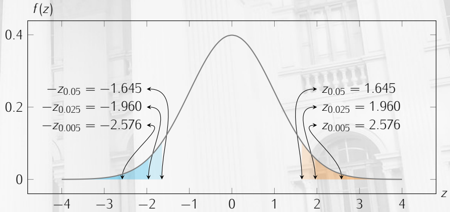


Figura 8. Quantis da distribuição Normal Padrão.

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: Intervalo 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.

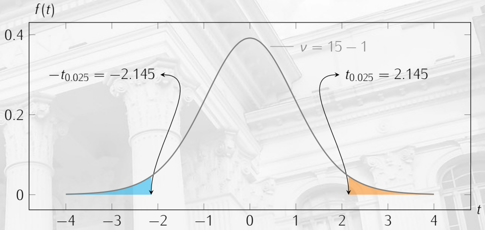


Figura 9. Quantis da distribuição t-Student.

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

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

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

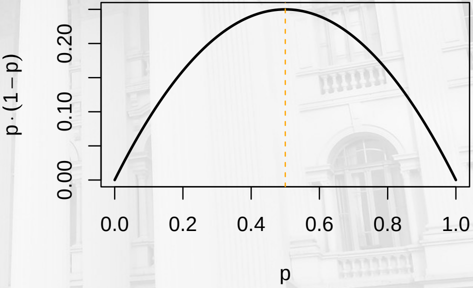
$$\text{P}\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha.$$

Intervalo de confiança para proporção

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

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.



Exercício: existe aquecimento global?

Foi realizada uma pesquisa com 1500 adultos selecionados aleatoriamente para responder à pergunta se acreditam ou não no aquecimento global. 1050 entrevistados responderam 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 10. Foto de Markus Spiske no Pexels.

Solução

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

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

- Intervalo conservador

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

- 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 $\mathrm{IC}_{1-\alpha}(p)$.

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.

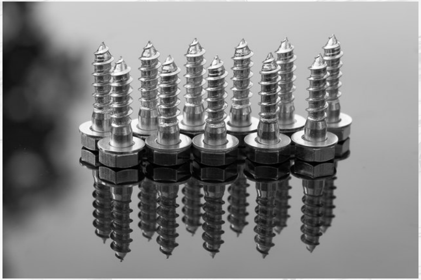


Figura 11. Foto de Pexels.

Solução

- Média e variância amostral

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

- Quantis da distribuição χ^2

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

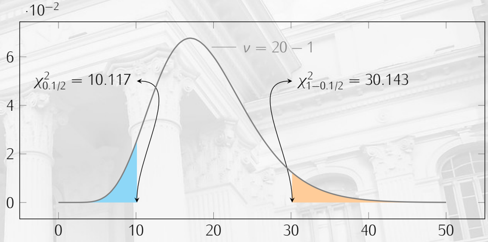


Figura 12. Quantis da distribuição χ^2.

- 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.

Resumo

Estimação pontual e intervalar

- Terminologia.
- Intervalo de confiança para a média.
- Intervalo de confiança para a proporção.
- Intervalo de confiança para a variância.

Figura 13. Foto de Karolina Grabowska no Pexels.