Inferência Estatística I

Intervalos de confiança

Prof. Paulo Cerqueira Jr
Faculdade de Estatística - FAEST
Instituto de Ciências Exatas e Naturais - ICEN

https://github.com/paulocerqueirajr

Introdução

Introdução

- Tudo que vimos até agora estava relacionado a `estimação pontual`.
- Dessa forma a inferência é baseada somente em um valor.
- E como cada `estimador pontual` é uma variável aleatória e uma distribuição de probabilidade, podemos obter uma estimativa mais informativa para o parâmetro de interesse.
- Que implica em encontrar maior precisão para o valor obtido, dando uma ideia de variabilidade do estimador.

Quantidade Pivotal

Definição 1 Uma v.a. $Q(X_1, X_2, \ldots, X_n; \theta) = Q(\mathbf{X}; \theta)$ é dita ser uma quantidade pivotal para o parâmetro μ se sua distribuição for independente de θ.

- Observe que $Q(\mathbf{X}; \theta)$ não é uma estatística, pois seu valor depende de θ.
- Podemos, para cada $0 < \gamma < 1$ fixado, encontrar λ_1 e λ_2 na distribuição de $Q(\mathbf{X}; \theta)$ de modo que

$$P(\lambda_1 \leq Q(\mathbf{X}; \theta) \leq \lambda_2) = \gamma.$$

Quantidade pivotal

- Se para cada $\mathbf{X}$ existirem $t_1(\mathbf{X})$ e $t_2(\mathbf{X})$

$$\lambda_1 \leq Q(\mathbf{X};\theta) \leq \lambda_2 \Leftrightarrow t_1(\mathbf{X}) \leq \theta \leq t_2(\mathbf{X})$$

então

$$P(t_1(\mathbf{X}) \leq \theta \leq t_2(\mathbf{X})) = \gamma.$$

- Assim, $[t_1(\mathbf{X}); t_2(\mathbf{X})]$ é um intervalo aleatório que contém θ com coeficiente de confiança γ, chamado de Intervalo de Confiança para θ.

Importante:

A notação usada para referenciar intervalos de confiança é:

$$IC(\theta, \gamma\%).$$

Quantidade pivotal

Observação 1:

Em geral, quando a v.a. X é discreta não conseguimos determinar λ_1 e λ_2 que satisfazem a expressão exatamente. Neste caso, os valores de λ_1 e λ_2 devem satisfazer a expressão para um coeficiente de confiança maior ou igual a γ (o mais próximo possível).

Observação 2:

Existem muitos pares $(\lambda_1; \lambda_2)$ satisfazendo a expressão, devemos escolher aquele que produz o intervalo de menor comprimento.

Exemplo:

Exercício 1 Seja $X_1, X_2, \ldots, X_n$ uma a.a. de $X \sim Exp(\theta)$. Construa um intervalo de confiança (I.C.) para θ.

Podemos mostrar que $T = \sum_{i=1}^{n} X_i$ é suficiente para θ, e $T \sim Gama(n; \theta)$. Logo, T não pode ser considerada uma quantidade pivotal para θ. Por outro lado, também podemos mostrar que

$$Q(\mathbf{X}; \theta) = 2\theta \sum_{i=1}^{n} X_i \sim \chi^2_{2n}.$$

Agora Q é uma quantidade pivotal. Então,

$$P(\lambda_1 \leq 2\theta \sum_{i=1}^{n} X_i \leq \lambda_2) = \gamma.$$

Exemplo:

Isolando o θ temos,

$$\lambda_1 \leq 2\theta \sum_{i=1}^{n} X_i \leq \lambda_2 \Leftrightarrow \frac{\lambda_1}{2\sum_{i=1}^{n} X_i} \leq \theta \leq \frac{\lambda_2}{2\sum_{i=1}^{n} X_i}$$

Logo,

$$IC(\theta, \gamma\%) = \left[\frac{\lambda_1}{2\sum_{i=1}^{n} X_i}; \frac{\lambda_2}{2\sum_{i=1}^{n} X_i}\right].$$

Exemplo

Supondo que temos uma amostra aleatória de tempos de vida de aparelhos eletrônicos, e queremos calcular o $IC(\theta, 90\%)$.

$$\begin{cases} n = 20 \\ \bar{X} = 3,975 \Leftrightarrow \sum_{i=1}^{n} X_i = n * \bar{X} = 20 * 3,975 = 79,5 \\ \chi^2_{2*20} = \chi^2_{40} \end{cases}$$

Os valores de λ_1 e λ_2 são tais que $P(\lambda_1 \leq \chi^2_{40} \leq \lambda_2) = 90\%$.

Exemplo

Como existirão vários pares de valores $(\lambda_1; \lambda_2)$ na distribuição χ^2_{40} satisfazendo a condição acima, tomaremos aquele par que satisfaz também

$$P(\chi^2_{40} \leq \lambda_1) = \frac{1-0,9}{2} = 0,05 = P(\chi^2_{40} \geq \lambda_2)$$

Logo, $\lambda_1 = 26,51$ e $\lambda_2 = 55,76$,

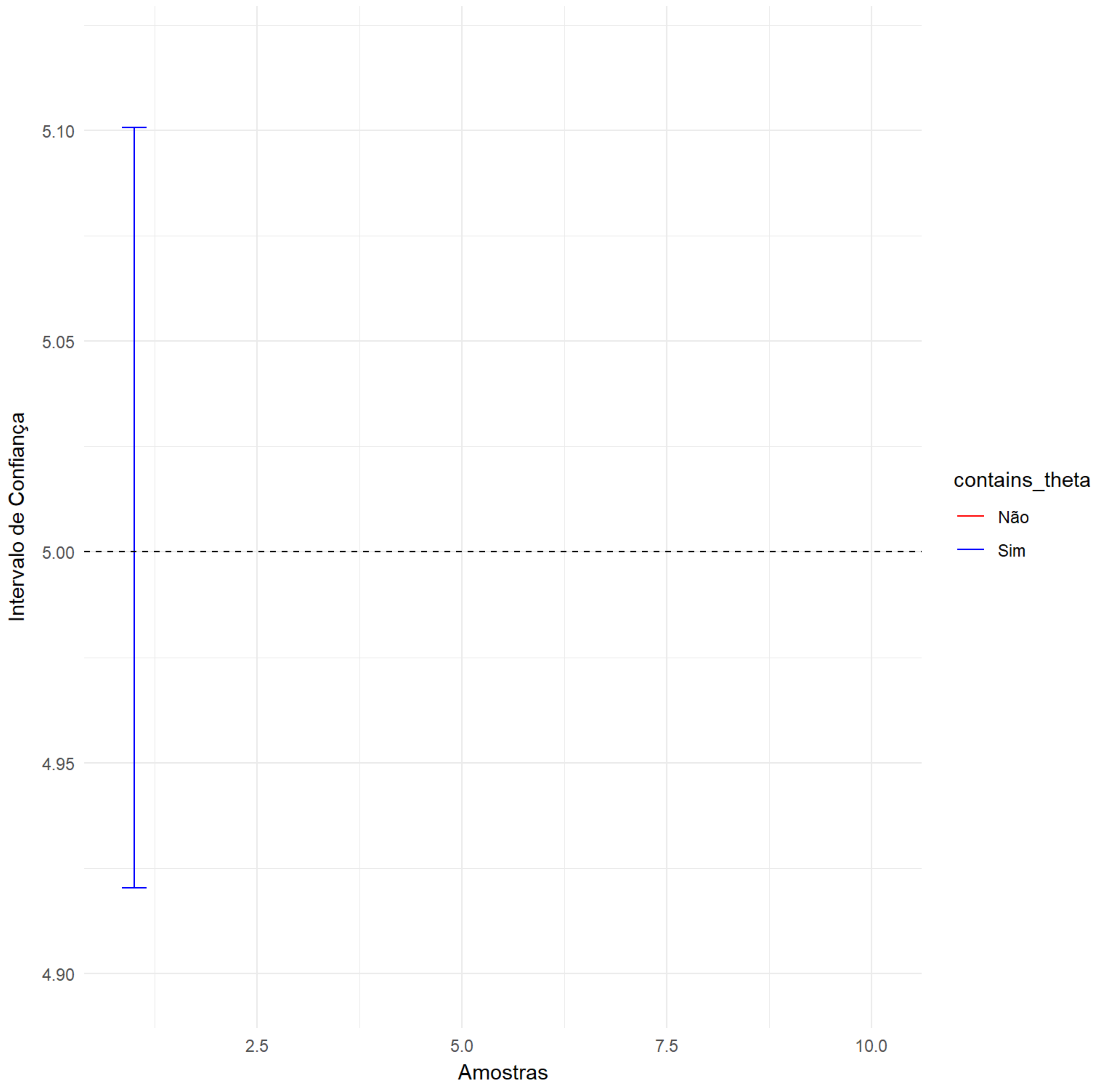
$$IC(\theta, 90\%) = \left[\frac{26,51}{2*79,5}; \frac{55,76}{2*79,5}\right] = [0,17; 0,35]\,.$$

Interpretação:

Este intervalo contém o verdadeiro valor de θ com probabilidade de 90%, ou seja, de cada 100 intervalos numéricos construídos desta forma, aproximadamente 90% deles vão conter o valor verdadeiro de θ.

Graficamente

Intervalos de confiança com $\gamma = 0,9$.



https://github.com/paulocerqueirajr

Intervalo de confiança para populações normais

Caso de uma única

- Considere inicialmente uma o caso de somente uma amostra.
- Neste caso é necessário adotar alguma suposições sobre a distribuição de probabilidade da característica X.
- Dessa forma, assuma agora que a amostra aleatória tenha a seguinte suposição

$$X_1, \ldots, X_n \text{ uma a.a. de } X \sim N(\mu, \sigma^2).$$

- Será apresentado a seguir alguns casos relacionados a suposição anteriormente assumida.
- Neste caso, sobre o conhecimento ou não de σ^2.

Caso 1: Intervalo para μ com σ^2 conhecido

- Neste caso, uma quantidade pivotal para θ, baseada na estatística suficiente $\sum_{i=1}^{n} X_i = n\bar{X}$, é dada por

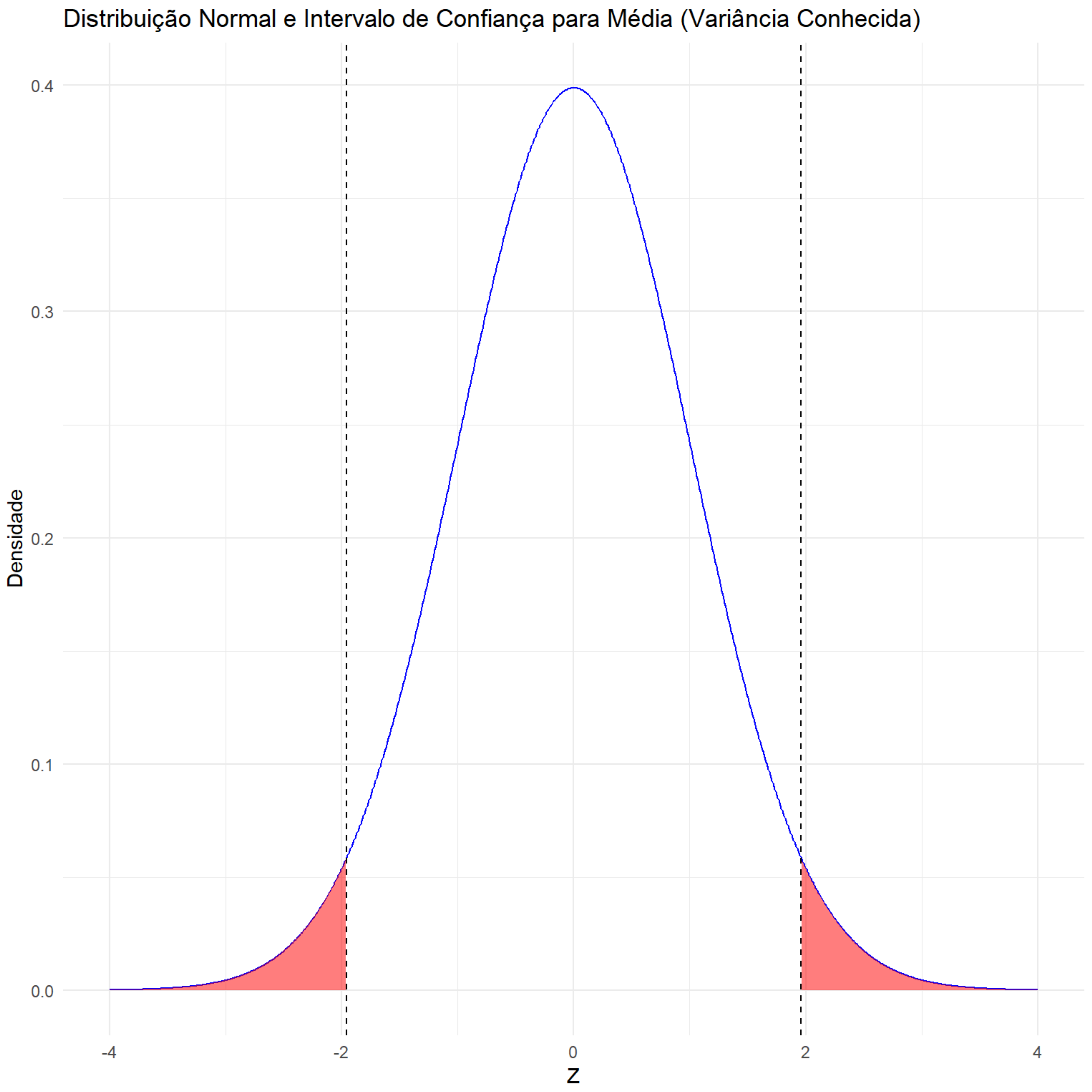
$$Q(\mathbf{X};\mu) = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

- Dessa forma, dado γ, determinamos λ_1 e λ_2 de modo que

$$P\left(\lambda_1 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \lambda_2\right) = \gamma.$$

- Como a distribuição $N(0,1)$ é simétrica, o intervalo de menor comprimento é o intervalo simétrico.
- Assim, sejam $\lambda_1 = -z_{\alpha/2}$ e $\lambda_2 = z_{\alpha/2}$ tais que $P(Z \geq z_{\alpha/2}) = \alpha/2$, com $Z \sim N(0,1)$ e $\alpha = 1-\gamma$.

Caso 1: Intervalo para μ com σ^2 conhecido



Caso 1: Intervalo para μ com σ^2 conhecido

Portanto, o intervalo de menor comprimento é dado por

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \Leftrightarrow -z_{\alpha/2}\sigma/\sqrt{n} \leq \bar{X} - \mu \leq z_{\alpha/2}\sigma/\sqrt{n}.$$

Em que,

$$-\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq -\mu \leq -\bar{X} + z_{\alpha/2}\sigma/\sqrt{n} \Leftrightarrow \bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}$$

Logo,

$$IC(\mu, \gamma\%) = \left[\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}; \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right].$$

Caso 2: Intervalo para μ com σ^2 desconhecido

- Vimos que

$$Q(\mathbf{X};\mu) = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

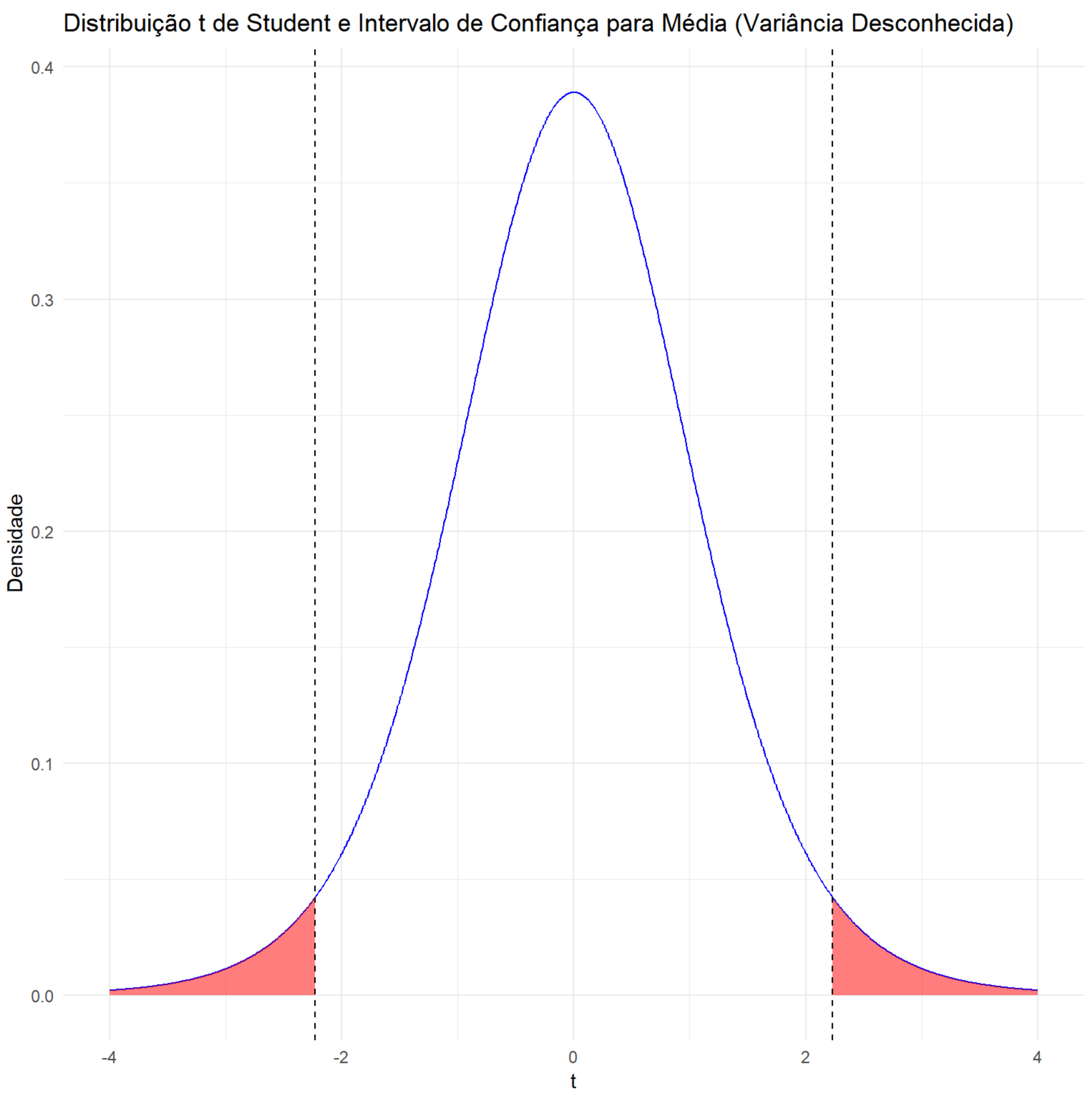
- Ou seja, é uma `quantidade pivotal` para θ.
- Assim, dado o coeficiente de confiança γ, podemos encontrar λ_1 e λ_2 na distribuição t_{n-1}, tais que

$$P\left(\lambda_1 \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq \lambda_2\right) = \gamma.$$

- Como a distribuição t_{n-1} é simétrica, tomamos $\lambda_1 = -t_{\alpha/2}$ e $\lambda_2 = t_{\alpha/2}$ tais que $P(T \geq t_{\alpha/2}) = \alpha/2$, com $T \sim t_{n-1}$ e $\alpha = 1-\gamma$. Assim,

$$IC(\mu, \gamma\%) = \left[\bar{X} - z_{\alpha/2}S/\sqrt{n}; \bar{X} + z_{\alpha/2}S/\sqrt{n}\right].$$

Caso 2: Intervalo para μ com σ^2 desconhecido



Caso 2: Intervalo de confiança para σ^2

- Vimos que

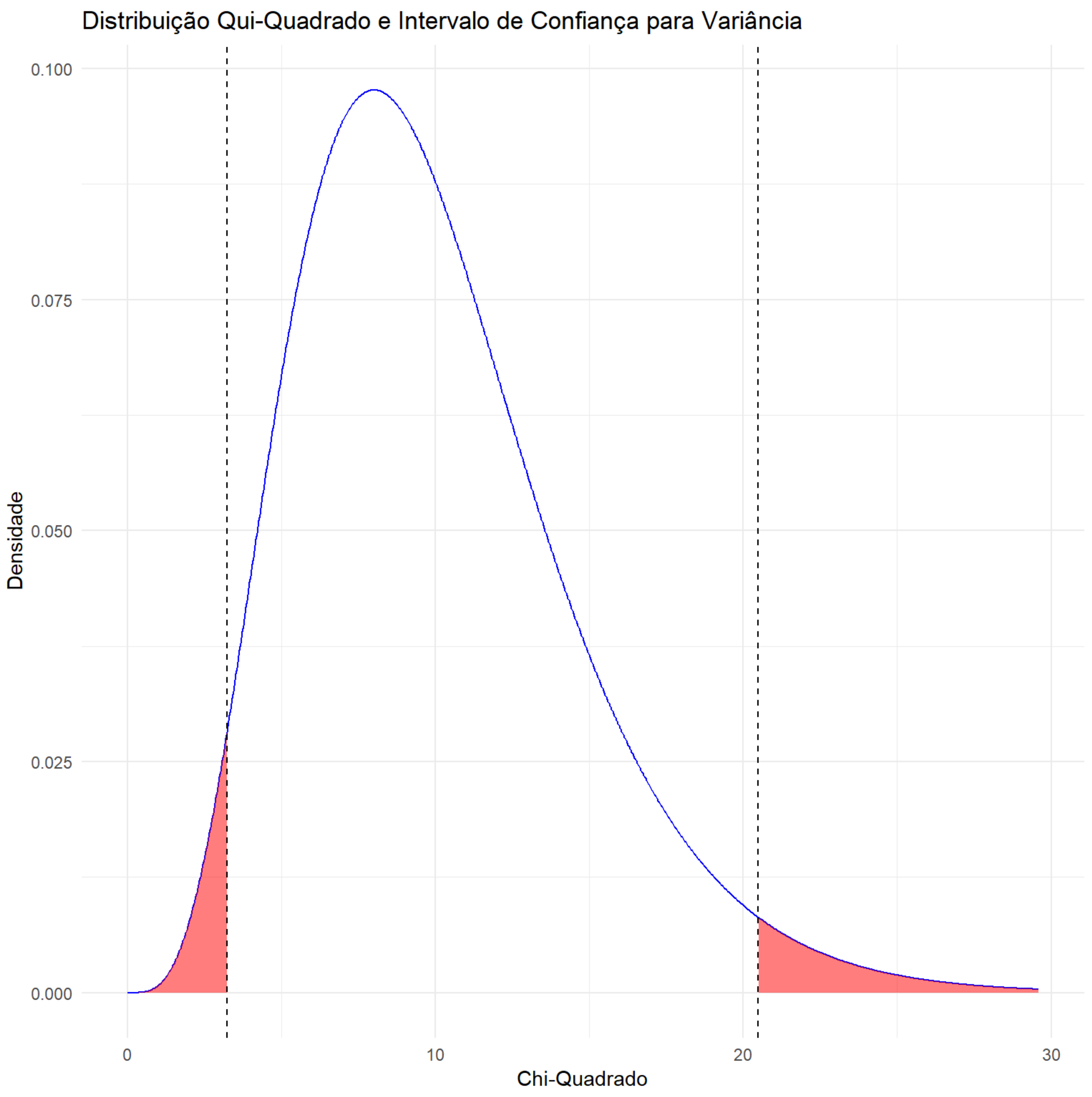
$$Q(\mathbf{X};\sigma^2) = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

- $Q(\mathbf{X};\sigma^2)$ é uma quantidade pivotal para σ^2.
- Assim, dado o coeficiente de confiança γ, podemos encontrar λ_1 e λ_2 na distribuição χ^2_{n-1}, tais que

$$P\left(\lambda_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq \lambda_2\right) = \gamma.$$

Considerando o intervalo simétrico, tomamos λ_1 e λ_2, tais que $P(\chi^2_{n-1} \geq \lambda_2) = P(\chi^2_{n-1} \leq \lambda_1) = \alpha/2$, com $\alpha = 1-\gamma$.

Caso 2: Intervalo de confiança para σ^2



Caso 2: Intervalo de confiança para σ^2

Assim

$$\lambda_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq \lambda_2 \Leftrightarrow \frac{\lambda_1}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\lambda_2}{(n-1)S^2}.$$

$$\frac{(n-1)S^2}{\lambda_2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\lambda_1}.$$

Logo,

$$IC(\sigma^2, \gamma\%) = \left[\frac{(n-1)S^2}{\lambda_2}; \frac{(n-1)S^2}{\lambda_1}\right].$$

Observações:

- A amplitude do intervalo de confiança é dada pela diferença entre o limite superior e o limite inferior;
- É usual referir-se à semi-amplitude, como o erro envolvido na estimação.
- Por exemplo, considerando o intervalo para a média com σ^2 conhecido, temos

$$\begin{aligned} \text{amplitude} &= \bar{X} - z_{\alpha/2}\sigma/\sqrt{n} - (\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}) \\ &= 2 \times z_{\alpha/2}\sigma/\sqrt{n}. \end{aligned}$$

e

$$\text{erro} = \frac{\text{amplitude}}{2} = z_{\alpha/2}\sigma/\sqrt{n}.$$

Relação entre medidas

Assim, observamos que a amplitude do intervalo depende de γ, σ e n:

- se aumentamos o γ, o valor de $z_{\alpha/2}$ aumenta e, consequentemente, a amplitude do intervalo também aumenta.
- uma variância grande indica a possibilidade de um considerável distanciamento dos possíveis valores da amostra em relação à média populacional. A amostra pode fornecer um valor de $\bar{x}$ muito influenciado por valores extremos.
- para uma mesma variabilidade σ e confiança γ, valores maiores de n produzem intervalos menores e, portanto, mais informativos.

Exemplos

Exemplo 1 Um provedor de acesso à Internet está monitorando a duração do tempo das conexões de seus clientes, com o objetivo de dimensionar seus equipamentos. São desconhecidas a média e a distribuição de probabilidade desse tempo, mas o desvio padrão, por analogia a outros serviços, é considerado igual a $\sqrt{50}$ minutos. Uma amostra de 500 conexões resultou num valor médio observado de 25 minutos. O que dizer da verdadeira média, com confiança 92%?

Exemplos

Exemplo 2 Uma amostra de 61 cidades brasileiras, de até 20 mil habitantes, forneceu o valor médio da hora aula para os professores do ensino fundamental em escolas municipais de R$ 2,5 e um desvio padrão igual a R$ 1,1. Obtenha um I.C. para o valor médio nacional da hora aula em cidades do tipo mencionado. Use $\gamma = 0,95$.

Exemplos

Exemplo 3 A vida média de baterias automotivas de uma certa marca está sendo estudada. Baseado em estudos similares, com outras marcas, é possível admitir que a vida dessas baterias segue a distribuição Normal com desvio padrão de 4,5 meses. De qual tamanho deverá ser a amostra, para que a amplitude do intervalo de 90% de confiança para a vida média seja de 3 meses?

Exemplos

Exemplo 4 Um fabricante deseja estudar a duração de baterias que são utilizadas em *smartwacths*. Uma amostra de vários lotes fabricados por uma mesma companhia foi submetida a testes acelerados e produziram os seguintes tempos de duração (em anos): 1,2; 1,4; 1,7; 1,3; 1,2; 2,3; 2,0; 1,5; 1,8; 1,4; 1,6; 1,5; 1,7; 1,5 e 1,3. Determine intervalos com 90% de confiança para a média e a variância do tempo de duração dessas pilhas.

Intervalos de confiança para populações não normais

Intervalo para a grandes amostras

- E se a variável X avaliada em uma determinada população não é normal.
- Como podemos construir intervalos de confiança para o nosso parâmetro de interesse?
- Há alguma aproximação usando o TCL, como vimos anteriormente?

- Sim podemos!!
- E dessa forma podemos obter intervalos de confiança para o parâmetro de interesse.
- Nesse caso, o intervalo será construído com um coeficiente de confiança aproximadamente igual à γ.
- Como na definição do `TCL`, quanto maior o tamanho amostral (n), melhor. (`Muito melhor`).

$$IC(\mu;\gamma) \simeq \left[\hat{\theta} - z_{\gamma/2}\frac{\sigma}{\sqrt{n}}; \hat{\theta} + z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\right].$$

$z_{\gamma/2}$: representa o quantil da distribuição $N(0,1)$.

Intervalo de confiança para a proporção p

Intervalo para a proporção p

- Vimos que pelo TCL, a distribuição amostral da proporção,

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

- Dessa forma o IC é dado por,

$$IC(p;\gamma) \simeq \left[\hat{p} - z_{\gamma/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}; \hat{p} + z_{\gamma/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right].$$

Intervalo para a proporção p

- Note que na formula acima p está como desconhecido.
- Uma solução é usar $\hat{p}(1-\hat{p})$ ao invés de $p(1-p)$.
- Assim, para um intervalo otimista:

$$IC(p;\gamma) \simeq \left[\hat{p} - z_{\gamma/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}; \hat{p} + z_{\gamma/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right].$$

- Outra abordagem é baseada no fato que a expressão $p(1-p)$ ter valor máximo igual à $1/4$. Assim, mais conservador

$$IC(p;\gamma) \simeq \left[\hat{p} - z_{\gamma/2}\frac{\sqrt{1/4}}{\sqrt{n}}; \hat{p} + z_{\gamma/2}\frac{\sqrt{1/4}}{\sqrt{n}}\right].$$

Exercícios

Em um certo lago, uma amostra aleatória de 1000 peixes acusou 290 Tilápias. Construa o intervalo de confiança de 95% para a proporção verdadeira de Tilápias. Interprete-o.

Intervalos de confiança para duas populações normais

Introdução

- Podemos calcular intervalos de confiança para duas amostras.
- Sempre com o objetivo de fazer comparações:
  - Médias;
  - Variâncias;
- Assuma agora que se tenha duas amostras aleatórias com a seguinte suposição

$$X_1, \ldots, X_n \text{ uma a.a. de } X \sim N(\mu_1, \sigma^2)$$

$$Y_1, \ldots, Y_m \text{ uma a.a. de } X \sim N(\mu_2, \sigma^2)$$

$$X \text{ e } Y \text{ independentes.}$$

Introdução

- Sabemos que

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

de modo que, sendo $\theta = \mu_1 - \mu_2$, consideramos a quantidade pivotal

$$Q(\mathbf{X}, \mathbf{Y}; \theta) = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N\left(0, 1\right).$$

Caso 1: σ^2 conhecido

Caso 1: σ^2 conhecido

- Nesse caso temos, como no caso de uma amostra, o intervalo

$$IC(\theta, \gamma\%) = \bar{X} - \bar{Y} \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

onde $z_{\alpha/2}$ é tal que $P(Z \geq z_{\alpha/2}) = \alpha/2$, com $Z \sim N(0,1)$ e $\alpha = 1 - \gamma$.

Caso 2: σ^2 desconhecido

Caso 2: σ^2 desconhecido

- Considere a seguinte quantidade pivotal:

$$Q(\mathbf{X}, \mathbf{Y}; \theta) = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} \sim t_{n+m-2}.$$

onde

$$S_p = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2},$$

em que S_x^2 e S_y^2 são as variâncias amostrais.

Caso 2: σ^2 desconhecido

- Dessa forma, o intervalo de confiança é dado por

$$IC(\theta, \gamma\%) = \bar{X} - \bar{Y} \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

em que $t_{\alpha/2}$ é o quantil de uma distribuição $t - student$.

Caso 2: IC para σ^2

- Como

$$\frac{(n-1)S_x^2}{\sigma^2} \sim \chi^2_{n-1} \text{ e } \frac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{m-1}$$

como S_x^2 e S_y^2 são independentes,

$$\frac{(n+m-2)S_p^2}{\sigma^2} = \frac{(n-1)S_x^2 + (m-1)S_y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

Caso 2: IC para σ^2

- Para construírmos um I.C. para σ^2, podemos considerar a quantidade pivotal

$$Q(\mathbf{X}, \mathbf{Y}; \sigma^2) = \frac{(n+m-2)S_p^2}{\sigma^2} = \frac{(n-1)S_x^2 + (m-1)S_y^2}{\sigma^2} \sim \chi^2_{n+m-2}.$$

- Temos então, $P(\chi^2_{n+m-2} \geq \lambda_2) = P(\chi^2_{n+m-2} \leq \lambda_1) = \alpha/2$, com $\alpha = 1 - \gamma$.
- O intervalo de confiança para σ^2 é dado por

$$IC(\sigma^2, \gamma\%) = \left[\frac{(n+m-2)S_p^2}{\lambda_2}; \frac{(n+m-2)S_p^2}{\lambda_1}\right].$$

IC para comparação de variâncias

IC para comparação de variâncias

- No caso em que $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ e o interesse é a construção de um IC para σ_1^2/σ_2^2.
- A quatidade pivotal é dada por:

$$Q(\mathbf{X}, \mathbf{Y}; \theta) = \frac{(m-1)S_y^2/\sigma_2^2(m-1)}{(n-1)S_x^2/\sigma_1^2(n-1)} \sim F_{m-1;n-1}.$$

onde $F_{m-1;n-1}$ denoda a distribuição F, com $m-1$ e $n-1$ graus de liberdade, é uma quantidade pivotal para $\theta = \sigma_1^2/\sigma_2^2$.

IC para comparação de variâncias

- Assim, dado γ, podemos determinar λ_1 e λ_2, na distribuição $F_{m-1;n-1}$, em que

$$P\left(\lambda_1 \leq \frac{\sigma_1^2 S_y^2}{\sigma_2^2 S_x^2} \leq \lambda_2\right) = \gamma$$

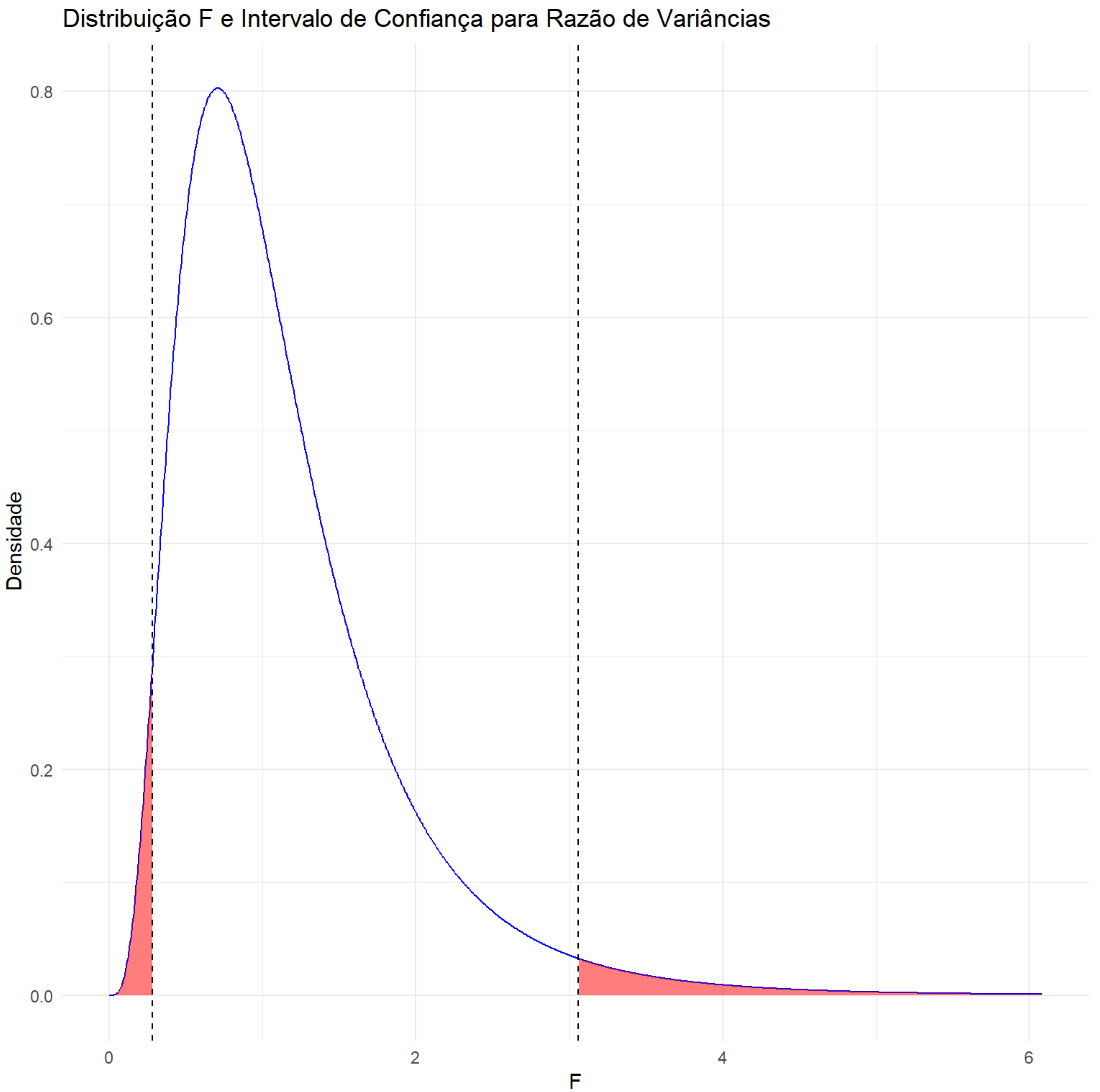
* Considerando um intervalo simétrico, ou seja, $\lambda_1 = F_1$ e $\lambda_2 = F_2$, de modo que, $P[F_{m-1;n-1} \geq F_2] = P[F_{m-1;n-1} \leq F_1] = \alpha/2$.

- O intervalo de confiança para a comparação de variâncias de $\gamma\%$

$$IC(\sigma^2, \gamma\%) = \left[F_1 \frac{S_x^2}{S_y^2}; F_2 \frac{S_x^2}{S_y^2}\right]$$

onde F_1 e F_2 são obtidos na tabela da distribuição F com $m-1$ e $n-1$ graus de liberdade, e $\alpha = 1-\gamma$.

IC para comparação de variâncias



Exemplos

Estão sendo estudados dois processos para conservar alimentos, cuja principal variável de interesse é o tempo de duração destes alimentos. No processo A, o tempo X de duração segue a distribuição $N(\mu_A; 100)$, e no processo B o tempo Y obedece à distribuição $N(\mu_B; 100)$. Sorteiam-se duas amostras independentes: a de A, com 16 latas, apresentou tempo médio de duração igual a 50, e a de B, com 25 latas, duração média igual a 60.

Para verificar se os dois processos podem ter o mesmo desempenho, decidiu-se construir um I.C. para a diferença $\mu_A - \mu_B$. Pode-se concluir que existe evidência de igualdade dos dois processos?

Exemplos

Admita que no exemplo anterior as variâncias dos dois processos sejam diferentes e desconhecidas. Suponha que a amostra das 16 latas do processo A produziu uma variância amostral $S_x^2 = 91$ e a amostra do processo B produziu uma variância amostral $S_y^2 = 100$. Construa um I.C. para a razão σ_1^2/σ_2^2, com coeficiente de confiança de 90%.