Estimação por Intervalo

Luan Fiorentin

UFPR - DEST

2019-05-08

Sumário

Introdução

- Existem dois tipos de **estimativas** que pode-se obter a partir de uma amostra aleatória:
    - Estimativa **pontual**: fornecem como estimativa um único valor numérico para o parâmetro de interesse.
    - Estimativa **intervalar**: fornece um intervalo de valores "plausíveis" para o parâmetro de interesse.

- Por serem **variáveis aleatórias**, os estimadores pontuais possuem uma **distribuição de probabilidade** (distribuições amostrais).

- Pode-se apresentar uma estimativa mais informativa para o parâmetro de interesse, que inclua uma medida de precisão do valor obtido: **estimativa intervalar** ou **intervalo de confiança**.

- **Os intervalos de confiança são obtidos a partir da distribuição amostral de seus estimadores.**
  - Por exemplo: é possível obter um intervalo de confiança para a média amostral.

Intervalos de Confiança para a Média: σ Conhecido

Pressuposição inicial:

- A amostra é uma **amostra aleatória simples** (AAS). Portanto, todas as amostras de mesmo tamanho tem a mesma probabilidade de serem selecionadas.
- O valor do desvio padrão populacional σ é **conhecido**.
- Uma ou ambas das seguintes condições são satisfeitas:
  - A população é normalmente distribuída.
  - A amostra possui $n > 30$.

Erro a amostral:

- Quando coleta-se uma amostra aleatória e calcula-se uma média, o valor da média possui um **desvio natural** em relação ao verdadeiro valor da média populacional (**erro amostral**), ou seja:

$$e = \bar{X} - \mu$$

$$\bar{X} = \mu + e.$$

- Sabe-se que a distribuição amostral da média é uma **distribuição normal**, com média μ e variância σ^2/n:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Margem de erro:

- Usando a transformação

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{e}{\sigma/\sqrt{n}} \sim N(0, 1),$$

pode-se determinar o erro máximo provável que assumimos para a média amostral que estamos calculando.

- O **erro máximo provável** ou **margem de erro da média** é definido por

$$e = z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}},$$

onde $z_{\gamma/2}$ é chamado de **valor crítico**.

Intervalo de confiança:

- Fixando um valor γ tal que $0 < \gamma < 1$, podemos encontrar um valor $z_{\gamma/2}$ tal que:

$$P[|Z| < z_{\gamma/2}] = \gamma$$

$$P[-z_{\gamma/2} < Z < z_{\gamma/2}] = \gamma$$

$$P\left[-z_{\gamma/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\gamma/2}\right] = \gamma$$

$$P\left[\bar{x} - z_{\gamma/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + z_{\gamma/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = \gamma$$

$$P[\bar{x} - e < \mu < \bar{x} + e] = \gamma$$

- O valor crítico $z_{\gamma/2}$ é que determina um intervalo possível para o parâmetro de interesse.

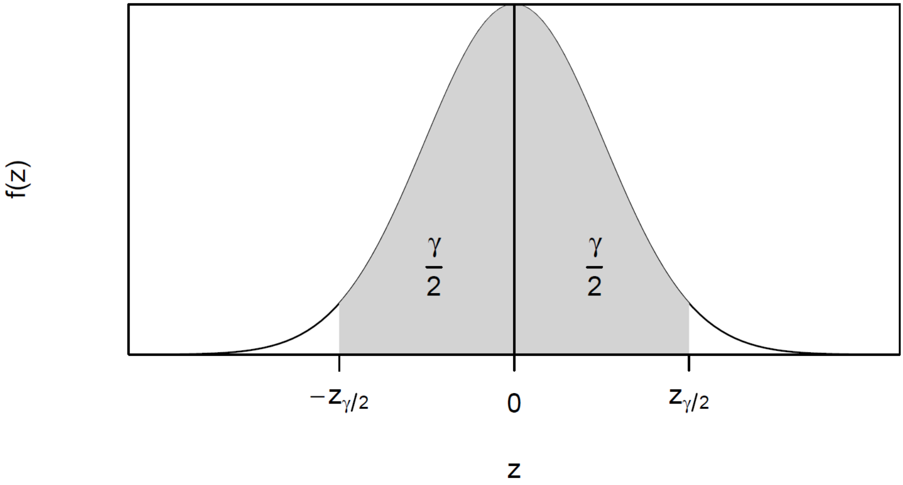

Figura 1: Distribuição normal padronizada.

- A área γ representa o **coeficiente de confiança** associado ao **intervalo de confiança** que está sendo construído.

- O valor $z_{\gamma/2}$ pode ser obtido da tabela da **Normal Padrão**, localizando o valor de $\gamma/2$ obtido na tabela e determinando o valor $z_{\gamma/2}$ nas margens.

Exemplo: considere o coeficiente de confiança $\gamma = 0.95$:

- Tem-se que $\gamma/2 = 0.475$ é a área a ser procurada na tabela.
- $\gamma/2 = 0.475 = 1.96$ é o valor crítico.

Construção do intervalo de confiança:

- Podemos construir um intervalo de confiança com um coeficiente de confiança:

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

$$IC[\mu, \gamma] = [\bar{X} - e; \bar{X} + e] = \bar{x} \pm e$$

$$\bar{x} - e < \mu < \bar{x} + e.$$

Como construir um intervalo de confiança???

- Verifique se as suposições são satisfeitas:
  - É uma amostra aleatória.
  - Desvio padrão é conhecido.
  - População tem distribuição normal ou n > 30.

- Determinar o nível de confiança γ, e o valor crítico $z_{\gamma/2}$.
- Determinar a margem de erro e.
- Calcular o $IC[\mu, \gamma]$.

Interpretação do intervalo de confiança:

Suponha que obtivemos um intervalo de 95% de confiança:
$IC[\mu, 0.95] = [95; 105]$

INTERPRETAÇÃO 1

Nós temos 95% de confiança que a verdadeira média μ se encontra entre 95 a 105.

INTERPRETAÇÃO 2

Nós temos 95% de confiança de que o intervalo entre 95 a 105 contém a verdadeira média populacional μ.

Interpretação do intervalo de confiança:

Suponha que obtivemos um intervalo de 95% de confiança:
$IC[\mu, 0.95] = [95; 105]$

INTERPRETAÇÃO 1 - ERRADA

Nós temos 95% de confiança que a verdadeira média μ se encontra entre 95 a 105.

INTERPRETAÇÃO 2 - CORRETA

Nós temos 95% de confiança de que o intervalo entre 95 a 105 contém a verdadeira média populacional μ.

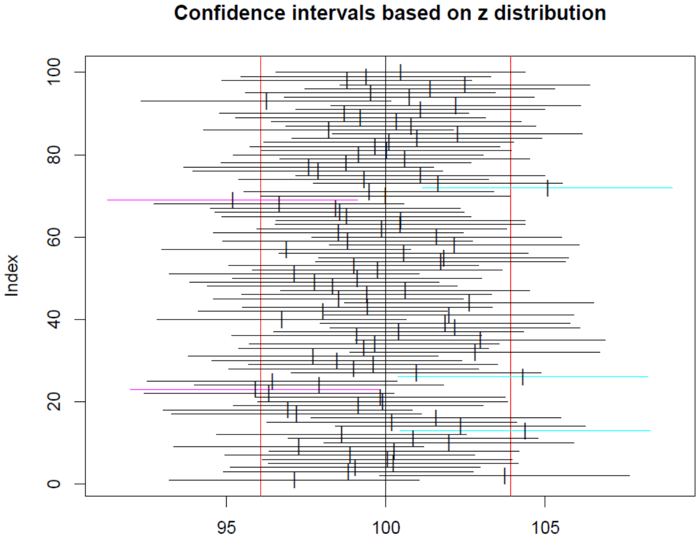

Figura 2: Intervalos de confiança.

Exercício 1

Uma empresa de computadores deseja estimar o tempo médio de horas semanais que as pessoas utilizam o computador. Uma amostra aleatória de 25 pessoas apresentou um tempo médio de uso de 22,4 horas. Com base em estudos anteriores, a empresa assume que $\sigma = 5,2$ horas, e que os tempos são normalmente distribuídos.
Construa um intervalo de confiança para a média μ com coeficiente de confiança de 95%.

Exercício 1

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

$$IC[\mu,\gamma] = \left[\bar{2}2,4 - 1,96 \cdot \left(\frac{5,2}{\sqrt{25}}\right); \bar{2}2,4 + 1,96 \cdot \left(\frac{5,2}{\sqrt{25}}\right)\right]$$

$$IC[\mu,\gamma] = \left[22,4 - 2,04; 22,4 + 2,04\right]$$

$$IC[\mu,\gamma] = 20,36 < \mu < 24,44$$

$$IC[\mu,\gamma] = [20,36; 24,44]$$

Amplitude do intervalo:

- A **amplitude** de um intervalo de confiança é dada pela diferença entre o limite superior e inferior, ou seja,

$$AMP_{IC} = \left[\bar{x} - z_{\gamma/2} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)\right] - \left[\bar{x} + z_{\gamma/2} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

$$AMP_{IC} = 2 \cdot z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

$$AMP_{IC} = 2 \cdot z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- $z_{\gamma/2}$: cada vez que aumenta-se a confiança γ, o valor de $z_{\gamma/2}$ fica maior e, consequentemente, a amplitude do intervalo aumenta.

- σ: um grande desvio padrão indica a possibilidade de um considerável distanciamento dos valores amostrais em relação à média populacional.

- n: quanto maior for o tamanho da amostra, maior será a quantidade de informação disponível. Com isso, valores maiores de n produzem intervalos mais informativos.

Exercício 2

Seja $X \sim N(\mu, 36)$.

1. Para uma amostra de tamanho 50, obtivemos média amostral 18,5. Construa intervalos de confiança de: i) 90%; ii) 95%; e iii) 99%.
2. Calcule as amplitudes dos intervalos acima e explique a diferença.
3. Para um nível de confiança de 95%, construa intervalos de confiança (admita a mesma média amostral 18,5) supondo tamanhos de amostra de: i) 15; e ii) 100.
4. Calcule as amplitudes dos intervalos acima e explique a diferença.

Dimensionamento de Amostra

- O objetivo do estudo é **estimar a média populacional μ desconhecida**. Logo, a pergunta é:

Quantos elementos (itens, objetos, pessoas, ...) eu devo amostrar?

- **Em geral, recomenda-se inicialmente $n > 30$ como um tamanho de amostra mínimo para grande parte dos problemas**.

- No entanto, pode-se obter uma **estimativa mais adequada** de quantos elementos amostrar.

- A partir do **erro máximo provável**

$$e = z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}},$$

pode-se isolar n e determinar uma **expressão** para estimar o tamanho amostral

$$n = \left[\frac{z_{\gamma/2} \cdot \sigma}{e} \right]^2.$$

- Tamanho da amostra n não depende do tamanho da **população** N.

- No entanto, o **tamanho amostral depende** do:
  - Nível de confiança desejado (γ).
  - Erro máximo desejado (e).
  - Desvio padrão (σ).

$$n = \left[\frac{z_{\gamma/2} \cdot \sigma}{e} \right]^2$$

- O tamanho amostral precisa ser um **número inteiro**, arredonda-se sempre o valor para o maior número inteiro mais próximo.

Exercício 3

Desejamos coletar uma amostra de uma variável aleatória X, com distribuição Normal, de média desconhecida e variância 30. Qual deve ser o tamanho da amostra para que, com 95% de probabilidade, a média amostral não difira da média populacional por mais de 1 unidade? E por mais de 4 unidades?

Intervalos de Confiança para a Média: σ Desconhecido

- Na maioria das situações práticas, não sabemos o **verdadeiro** valor do desvio padrão populacional σ.

- Se o desvio padrão é desconhecido, ele precisa ser estimado.

- Considere variáveis aleatórias $(X_1, X_2, ..., X_n)$, onde $X \sim N(\mu, \sigma^2)$, sabe-se que o "melhor" estimador para σ^2 é a variância amostral

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right),$$

a qual não é viciada e é consistente para σ^2.

- A **variável padronizada** é

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} = \frac{\bar{X} - \mu}{S\sqrt{n}}.$$

Logo, o denominador S^2 fará com que T seja diferente da Normal.

- Essa densidade é denominada **t de Student**, e seu parâmetro é denominado **graus de liberdade**, que nesse caso é $n - 1$. Portanto,

$$T = \frac{\bar{X} - \mu}{S\sqrt{n}} \sim t_{(n-1)}.$$

Características da distribuição t:

- É **simétrica** com média $t = 0$ (assim como $z = 0$).
- É **diferente** para tamanhos de amostra diferentes.
- Maior área nas caudas e menor área no centro, quando comparada com a distribuição normal, serve para incorporar as **incertezas**.
- **Desvio padrão** da distribuição t varia com o tamanho da amostra (ao contrário da distribuição z onde $\sigma = 1$):
    - Se n diminui, σ aumenta.
    - Se n aumenta, σ diminui.
- A medida que o n amostral aumenta, a distribuição t se aproxima cada vez mais de uma **distribuição normal padrão** Z.

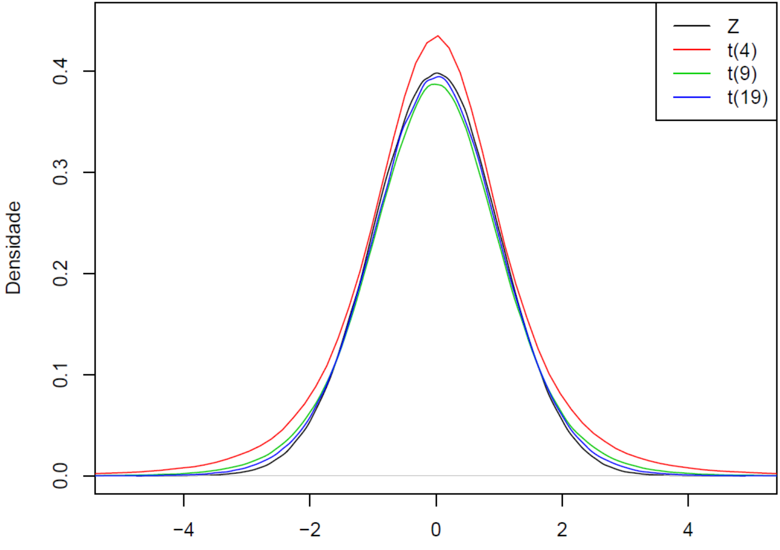

Figura 3: Distribuição t de Student.

Como encontrar os valores críticos de t?

- Com a definição do **nível de confiança** e sabendo o **tamanho da amostra** n, sabe-se então o valor de γ e dos G.L., podendo encontrar o **valor crítico** de $t_{\gamma/2}$.

- Exemplo: $\gamma = 0,95$ e uma amostra de tamanho $n = 7$:
    - se $n = 7$, o G.L. associado é $n - 1 = 6$.
    - Na tabela da distribuição t de Student procura-se a linha correspondente aos G.L., e coluna correspondente ao valor de $1 - \gamma = 1 - 0,95 = 0,05 = 5\%$.
    - Valor de $t_{\gamma/2}$ será determinado pelos valores correspondentes no corpo da tabela. Nesse caso, $t_{\gamma/2} = 2,447$ é o valor crítico procurado.

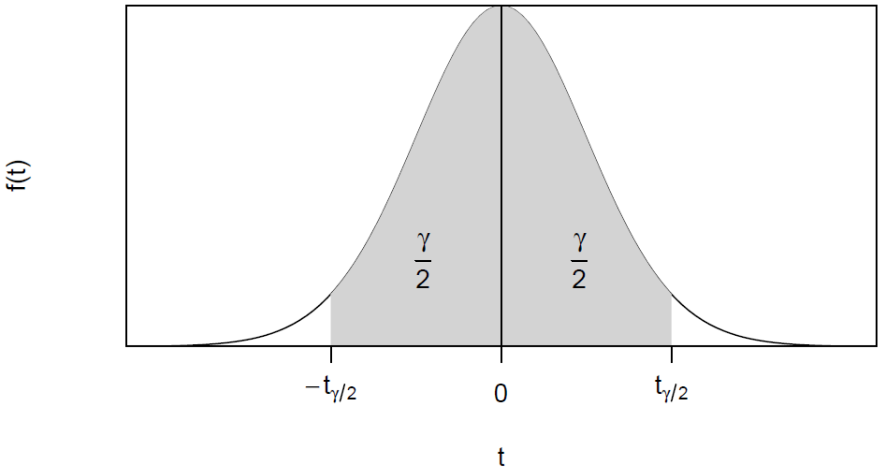

Figura 4: Distribuição t de Student.

Construção do intervalo de confiança:

- Podemos construir um intervalo de confiança com um coeficiente de confiança:

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

$$IC[\mu, \gamma] = [\bar{x} - e; \bar{x} + e] = \bar{x} \pm e$$

$$\bar{x} - e < \mu < \bar{x} + e.$$

Como construir um intervalo de confiança???

- Verifique se as suposições são satisfeitas:
  - É uma amostra aleatória.
  - Desvio padrão é **estimado**.
  - População tem distribuição normal ou n > 30.

- Determinar o nível de confiança γ, e o valor crítico $t_{\gamma/2}$.
- Determinar a margem de erro $e = t_{\gamma/2} \cdot (S/\sqrt{n})$.
- Calcular o $IC[\mu, \gamma]$.

Exercício 4

Em um teste da eficácia do alho na dieta para a redução do colesterol, 51 pessoas foram avaliadas e seus níveis de colesterol foram medidos antes e depois do tratamento. As mudanças nos níveis de colesterol apresentaram média de $0,4$ e desvio-padrão de 21.

1. Para um nível de confiança de 95%, calcule o intevalo para a verdadeira média das mudanças no nível de colesterol
2. O que o intervalo de confiança sugere sobre a eficácia do uso do alho na dieta para a redução do colesterol?
3. Resolva o mesmo exemplo supondo que o $\sigma = s$ é conhecido (usando a distribuição Z). Compare os resultados.

Dimensionamento de Amostra

Se σ é **desconhecido**:

- Estimar o valor de σ com base em estudo preliminar.
- Faça uma amostra piloto e estime o desvio padrão amostral s, e use-o como uma aproximação para o desvio-padrão populacional σ.
- Use a **regra empírica** da amplitude para dados com distribuição (aproximadamente) Normal:

Como sabemos que em uma distribuição (aproximadamente) normal, aproximadamente 95% dos dados encontram-se a 2 desvios-padrões acima e abaixo da média, temos que

$$4\sigma = (max - min)$$

$$\hat{\sigma} = s = \frac{(max - min)}{4}$$

.