

Estatística I

Prof. Fernando de Souza Bastos
fernando.bastos@ufv.br

Departamento de Estatística
Universidade Federal de Viçosa
Campus UFV - Viçosa



Sumário

- 1 Introdução
- 2 Intervalos de Confiança para a Média: σ conhecido
- 3 Intervalos de confiança para a média: σ desconhecido
- 4 Intervalo de Confiança para Proporção
- 5 Determinação do tamanho amostral (σ conhecido)
- 6 Determinação do tamanho amostral (σ desconhecido)

Suponha que um clube de xadrez deseja estimar o QI médio de seus membros. A média de uma amostra aleatória dos membros é 115. Como essa estimativa consiste em um único número representado por um ponto em uma linha numerada, ele é chamado de estimativa pontual. O problema com a utilização de uma estimativa pontual é que ela raramente é igual ao parâmetro exato (média, desvio padrão ou proporção) da população. Aqui aprenderemos como fazer uma estimativa mais apropriada, especificando um intervalo de valores no conjunto dos números reais, juntamente com uma afirmação de quão confiante se está de que o intervalo contém o parâmetro populacional.

Suponha que o clube queira estar 90% confiante de sua estimativa para o QI médio de seus membros. Na Figura seguinte você pode ter uma visão geral de como construir uma estimativa por intervalo.

Construindo um intervalo de confiança.

Encontre a média
da amostra aleatória
 $\bar{x} = 115$



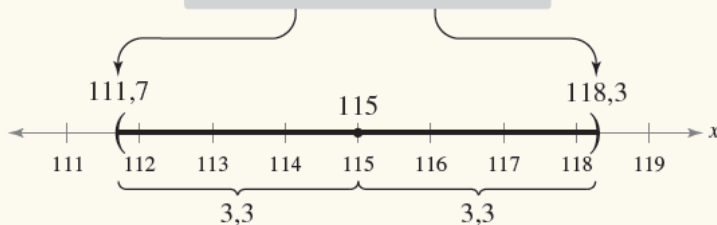
Encontre a
margem de erro
 $E = 3,3$



Encontre os limites do intervalo
Inferior: $115 - 3,3 = 111,7$
Superior: $115 + 3,3 = 118,3$



Expresse graficamente a
estimativa por intervalo
 $111,7 < \mu < 118,3$



Então o clube pode estar 90% confiante de que o QI médio de seus membros está entre 111,7 e 118,3.

Logo, existem dois tipos de estimativas que podemos 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). Com isso, podemos 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.

Suposições necessárias

- A amostra é uma **amostra aleatória simples**. (Todas as amostras de mesmo tamanho tem a mesma probabilidade de serem selecionadas);

Suposições necessárias

- A amostra é uma **amostra aleatória simples**. (Todas as amostras de mesmo tamanho tem a mesma probabilidade de serem selecionadas);
- O valor do desvio padrão populacional σ , é conhecido;

Suposições necessárias

- A amostra é uma **amostra aleatória simples**. (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 Amostral

Quando coletamos uma **amostra aleatória** e calculamos uma média, sabemos que o valor da média possui um desvio natural, em relação ao verdadeiro valor da média populacional (**erro amostral**), ou seja

$$e = \bar{X} - \mu \quad \Rightarrow \quad \bar{X} = \mu + e$$

Sabemos 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

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

podemos 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:

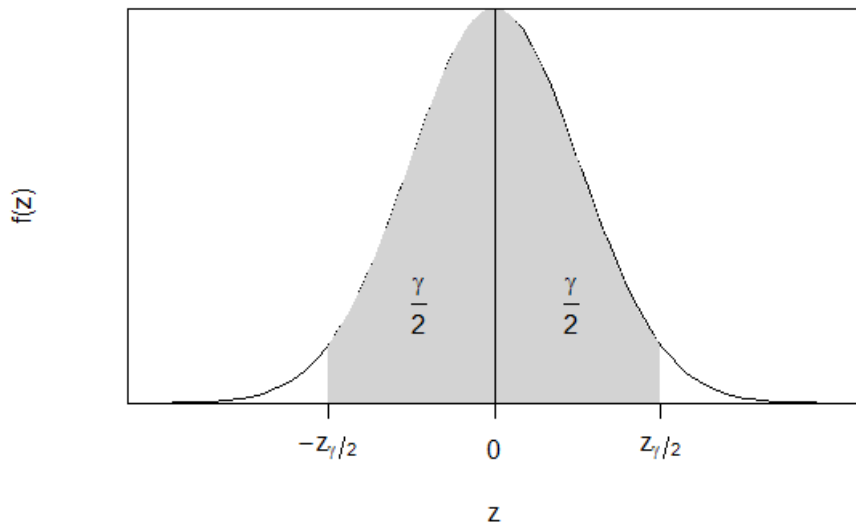
$$P(-z_{\gamma/2} < Z < z_{\gamma/2}) = \gamma$$

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

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

$$P(\bar{x} - e < \mu < \bar{x} + e) = \gamma$$

O valor crítico $z_{\gamma/2}$ é o valor de γ dividido por 2, uma vez que a “massa” γ deve ser distribuída igualmente em torno de 0.



A área γ determina o **coeficiente de confiança** associado ao intervalo de confiança que estamos construindo. O valor $z_{\gamma/2}$ pode ser obtido da tabela da Normal padrão, localizando o valor de $\gamma/2$ no corpo da tabela e obtendo o valor $z_{\gamma/2}$ nas margens correspondentes.

- Exemplo: $\gamma = 0,95$:

- Temos que $\gamma/2 = 0,475$ é a área que devemos procurar no corpo da tabela
- O valor de $z_{\gamma/2}$ será determinado pelos valores correspondentes nas margens da tabela. Nesse caso, $z_{\gamma/2} = 1,96$ é o valor crítico procurado.

Com estas definições, podemos construir um **intervalo de confiança** para μ , com coeficiente de confiança γ :

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

Procedimentos para a construção de intervalos de confiança

- 1 Verifique se as suposições necessárias estão satisfeitas
 - Temos uma AAS
 - σ é conhecido
 - A população tem distribuição normal ou $n > 30$
- 2 Determine o nível de confiança γ , e encontre o valor crítico $z_{\gamma/2}$
- 3 Calcule a margem de erro $e = z_{\gamma/2} \cdot (\sigma/\sqrt{n})$
- 4 Calcule $IC(\mu, \gamma)$

Interpretação de um intervalo de confiança

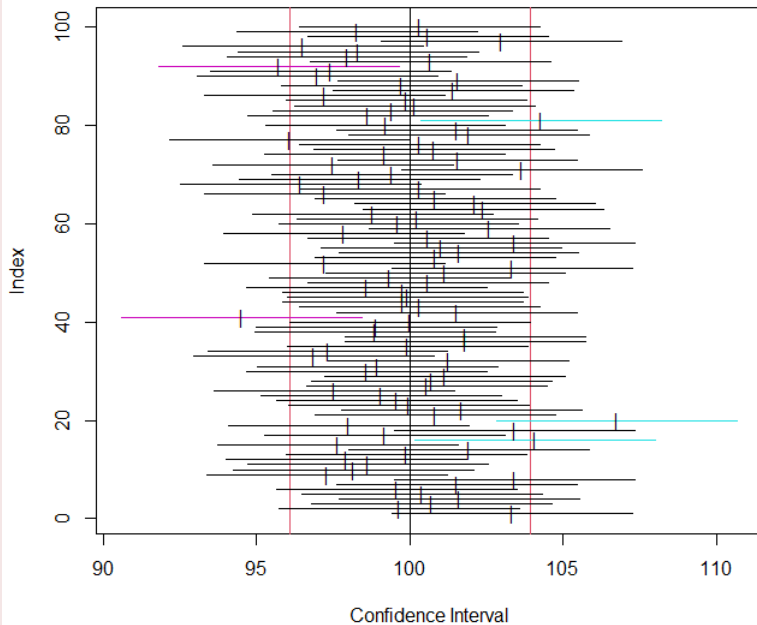
Como o intervalo de confiança é calculado a partir de uma amostra aleatória, este intervalo **também é aleatório!**

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 pudéssemos obter 100 amostras diferentes, e calcularmos um intervalo de confiança de 95% para cada uma das 100 amostras, esperaríamos que 5 destes intervalos **não** contenham o verdadeiro valor da média populacional μ .

Confidence intervals based on z distribution



Exemplo

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

Exemplo

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

$$(20.36164 \leq \mu \leq 24.43836)$$

Intervalos de confiança para a média: σ desconhecido

Estimativa da variância amostral

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. Sendo (X_1, \dots, X_n) VAs onde $X \sim N(\mu, \sigma^2)$, vimos que o “melhor” estimador para σ^2 é a variância amostral

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

que é não viciada e consistente para σ^2 .

A distribuição t de Student

Definindo a variável padronizada

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

o denominador S^2 fará com que a função densidade de T seja diferente da Normal. Essa nova densidade é denominada **t de Student**, e seu parâmetro é denominado **graus de liberdade**, que nesse caso é $n - 1$.

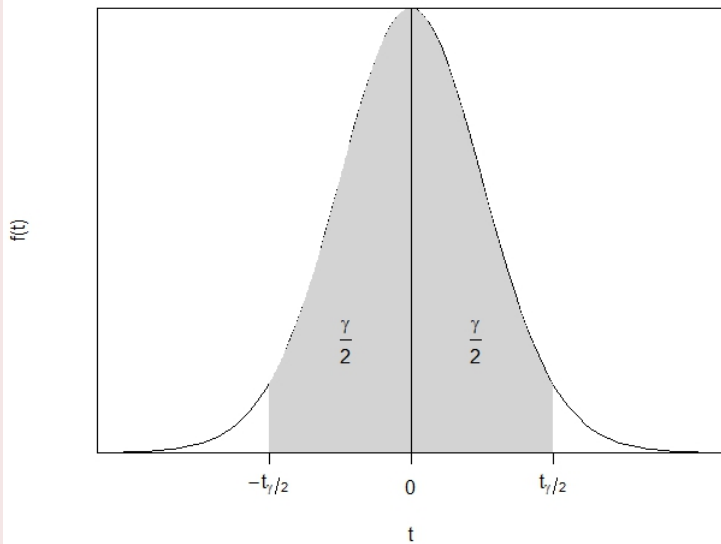
Assim:

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

Valores Críticos de t

Com a definição do **nível de confiança** e sabendo o tamanho da amostra n , sabemos então o valor de γ e dos gl, e devemos encontrar o **valor crítico** de $t_{\gamma/2}$. Usando como exemplo $\gamma = 0,95$ e uma amostra de $n = 7$

- Temos que $n = 7 \Rightarrow gl = n - 1 = 6$
- Na tabela da distribuição t de Student procure a linha correspondente aos gl, e coluna correspondente ao valor de $1 - \gamma = 1 - 0,95 = 0,05 = 5\%$
- O 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.



Intervalo de confiança

Com estas definições, podemos construir um **intervalo de confiança** para μ , com **coeficiente de confiança** γ , e σ desconhecido:

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

Procedimentos para a construção de intervalos de confiança

- 1 Verifique se as suposições necessárias estão satisfeitas
 - Temos uma AAS
 - Temos uma estimativa de s
 - A população tem distribuição normal ou $n > 30$
- 2 Determine o nível de confiança γ , e encontre o valor crítico $t_{\gamma/2}$
- 3 Calcule a margem de erro $e = t_{\gamma/2} \cdot (s/\sqrt{n})$
- 4 Calcule $IC(\mu, \gamma)$

Exemplo

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.

- a) Para um nível de confiança de 95%, calcule o intervalo para a verdadeira média das mudanças no nível de colesterol;
- b) O que o intervalo de confiança sugere sobre a eficácia do uso do alho na dieta para a redução do colesterol?
- c) Resolva o mesmo exemplo supondo que o $\sigma = s$ é conhecido (ou seja, usando a distribuição Z). Compare os dois métodos.

Intervalo de Confiança para Proporção

A proporção amostral

$$\hat{p} = \frac{x}{n} = \frac{\text{número de sucessos}}{\text{total de tentativas}}$$

é a “melhor estimativa” para a proporção populacional p . Através do estudo da distribuição amostral da proporção, chegamos aos seguintes resultados:

- A proporção amostral \hat{p} **tende** para o valor da proporção populacional p
- A distribuição das proporções amostrais tende a ser uma **distribuição normal**
- $E(\hat{p}) = \mu_{\hat{p}} = p$
- $\text{Var}(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$

Distribuição amostral da proporção \hat{p}

Assim, sabemos que

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

É possível mostrar que a quantidade

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

Intervalo de Confiança

Logo, podemos construir um **intervalo de confiança** para p , com **coeficiente de confiança** γ

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

Procedimentos para a construção de intervalos de confiança

- 1 Verifique se as suposições necessárias estão satisfeitas
 - Temos uma AAS
 - As condições para a distribuição binomial são satisfeitas:
 - as tentativas são independentes;
 - há duas categorias de resultado (“sucesso”, “fracasso”);
 - 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 γ , e encontre o valor crítico $z_{\gamma/2}$
- 3 Calcule a margem de erro $e = z_{\gamma/2} \cdot \sqrt{\frac{p(1 - p)}{n}}$
- 4 Calcule $IC(\mu, \gamma)$

Exemplo

Em uma pesquisa realizada por um instituto de pesquisa Norte-Americano, 1500 adultos foram selecionados aleatoriamente para responder à pergunta se acreditam ou não no aquecimento global. 1050 entrevistados responderam que sim. Com isso:

- a) 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.
- b) Com base nesses resultados, podemos concluir que a maioria dos adultos acredita no aquecimento global?

Determinação do tamanho amostral

Nosso objetivo é coletar dados para estimar a **média populacional** μ . A questão é:

Quantos elementos (itens, objetos, pessoas, ...) devemos amostrar?

Já vimos que, de maneira (bem) geral, $n > 30$ é um tamanho de amostra mínimo para a maioria dos casos. Será que podemos ter uma estimativa melhor de quantos elementos devem ser amostrados para estimarmos a média populacional com uma precisão conhecida?

A partir da equação do **erro máximo provável**

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

podemos isolar n e chegar na seguinte equação para a determinação do tamanho amostral

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

Note que, em

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

- O tamanho amostral n **não** depende do tamanho populacional N ;
- O tamanho amostral depende:
 - do nível de confiança desejado (expresso pelo valor crítico $z_{\gamma/2}$);
 - do erro máximo *desejado*
 - do desvio-padrão σ (embora veremos que não é estritamente necessário)
- Como o tamanho amostral precisa ser um número inteiro, arredondamos sempre o valor para o **maior** número inteiro mais próximo.

Exemplo

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

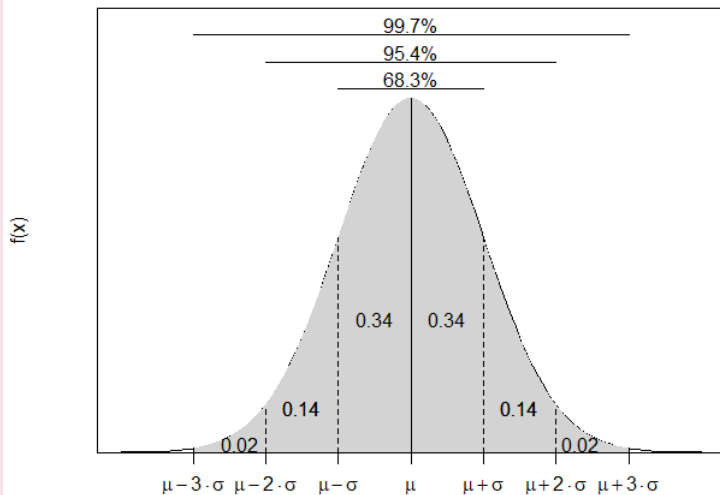
- a) Calcule o tamanho da amostra, para que com 95% de probabilidade, a média amostral não difira da média populacional por mais de
 - (i) 0,5 unidades (ii) 2 unidades
- b) Qual o impacto do erro máximo assumido para o tamanho da amostra?
- c) Calcule o tamanho da amostra, para que a diferença da média amostral para a média populacional (em valor absoluto) seja menor ou igual a 2 unidades, com níveis de confiança de
 - (i) 90% (ii) 95%
- d) Compare as estimativas do item anterior e analise o impacto do nível de confiança para a determinação do tamanho amostral.

Determinação do tamanho amostral (σ desconhecido)

Se σ for desconhecido?

- Estime o valor de σ com base em algum estudo feito anteriormente
- 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

Regra empírica para uma distribuição normal



Regra empírica para uma distribuição normal

Define-se **valores usuais** aqueles que são típicos e não muito extremos. 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

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

pode ser utilizado como uma estimativa para σ .

Exemplo

Um professor deseja estimar o salário médio de professores do Ensino Médio de uma cidade. Quantos professores devem ser selecionados para termos 90% de confiança que a média amostral esteja a menos de R\$30,00 da média populacional? Sabe-se apenas que os salários variam entre R\$800,00 e R\$1.200,00. Use

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

PETERNELLI, Luiz Alexandre. **Roteiro de Aulas da Disciplina Estatística 1**. [S.l.: s.n.], 2022.