Estatística aplicada à Biologia

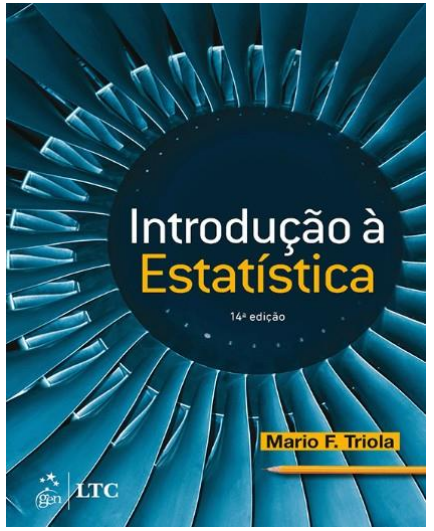
Aula 5 – Estimativa pontual e intervalar

Decifrando o Desconhecido nas Populações Biológicas

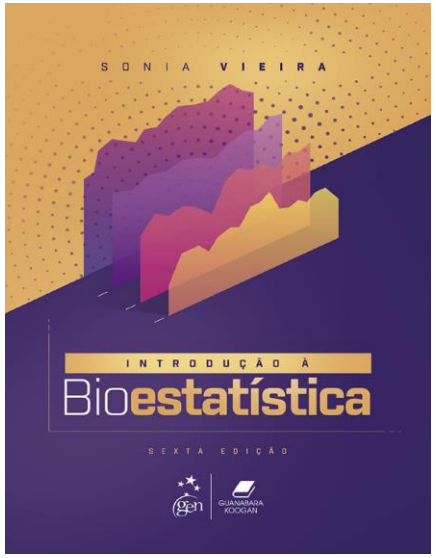
Objetivos da aula

Ao final, você será capaz de:

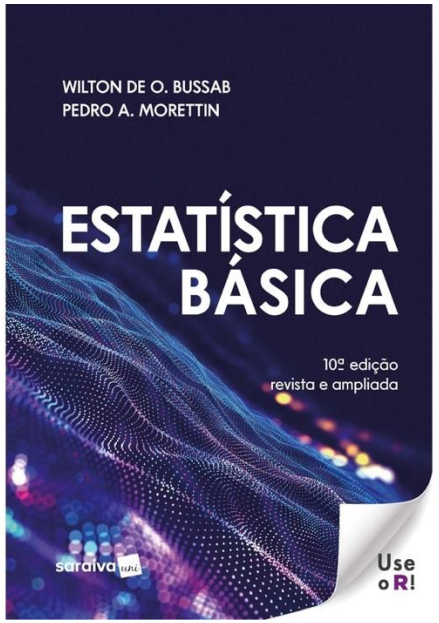
- Diferenciar estimativa pontual e intervalar
- Interpretar intervalos de confiança em estudos biológicos
- Aplicar conceitos em exemplos reais de biologia
- Praticar com exercícios interativos



Cap 7



Cap 9



Cap 10 e 11

Recapitulando a Aula Anterior: **Medidas de Amostra**

- Na aula passada, aprendemos a calcular:
  - **Média Amostral ($\bar{x}$):** Nosso melhor palpite para o centro dos dados da nossa amostra.
  - **Desvio Padrão Amostral (s):** Quão dispersos os dados estão na nossa amostra.
- **Lembre-se:** Nosso objetivo final com a estatística é entender o que se passa com a **população** inteira, não apenas com a amostra!

Introdução à inferência estatística

- Definições de estatística, estimador e estimativa
- Estimativa pontual e intervalar
  - Lei dos grandes números
- Teorema do limite central
  - simulações
- Erro e desvio padrão da média
- Intervalo de confiança

Relembrando conceitos chave

- **Amostra**: qualquer subconjunto da população.

- **População**: conjunto de todos os elementos ou resultados sob investigação.

Objetivo do procedimento de amostragem é estimar
um parâmetro da população a partir da amostra

Definições

- **Estimador pontual**: valor único que estima o parâmetro da população
  - Exemplo: média amostral, proporção amostral.

- **Estimativa**: valor observado de um estimador, ou seja, um número que é obtido quando uma amostra é obtida

Definições

- Uma *estatística* é uma característica da *amostra*, ou seja, um número que mede um determinado aspecto da amostra
  - Exemplos: média da amostra, variância etc

- Um *parâmetro* é uma medida usada para medir uma característica da *população*
  - Exemplos:
    - Média: $E(X)=\mu$;
    - Variância: $Var(X) = \sigma^2$

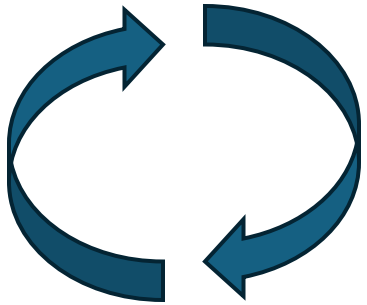
Imaginem um grande lago onde vivem milhares de tilápias. Como biólogos, queremos saber o peso médio *de toda a população* de tilápias. É impossível pesar todas, então vamos pescar amostras.

Atividade (15 min)

Eu sou a 'Mãe Natureza' e sei o peso médio real de todas as tilápias: **250 gramas**. Mas vocês, pesquisadores, não sabem disso.

O que fazemos para estimar o peso médio?

Eu sou a 'Mãe Natureza' e sei o peso médio real de todas as tilápias: **250 gramas**. Mas vocês, pesquisadores, não sabem disso.



Amostragens!

Cada grupo terá uma rede

- *Rede 1:* [245, 255, 250, 248, 252, 260, 240, 238, 251, 255]
- *Rede 2:* [230, 240, 245, 250, 235, 228, 241, 233, 229, 239]
- *Rede 3:* [260, 265, 258, 270, 255, 262, 268, 259, 261, 263]
- Rede 4: [231, 235, 252, 282, 241, 223, 254, 246, 239, 236]
- Rede 5: [262, 272, 268, 251, 270, 217, 242, 226, 231, 259]

Máximo de 10 pessoas por grupo

Atividade (10 min)

Cada grupo:

- Calculem a média do peso das Tilápias pescadas na sua rede (amostra)

- Cada grupo diz o resultado que encontrou

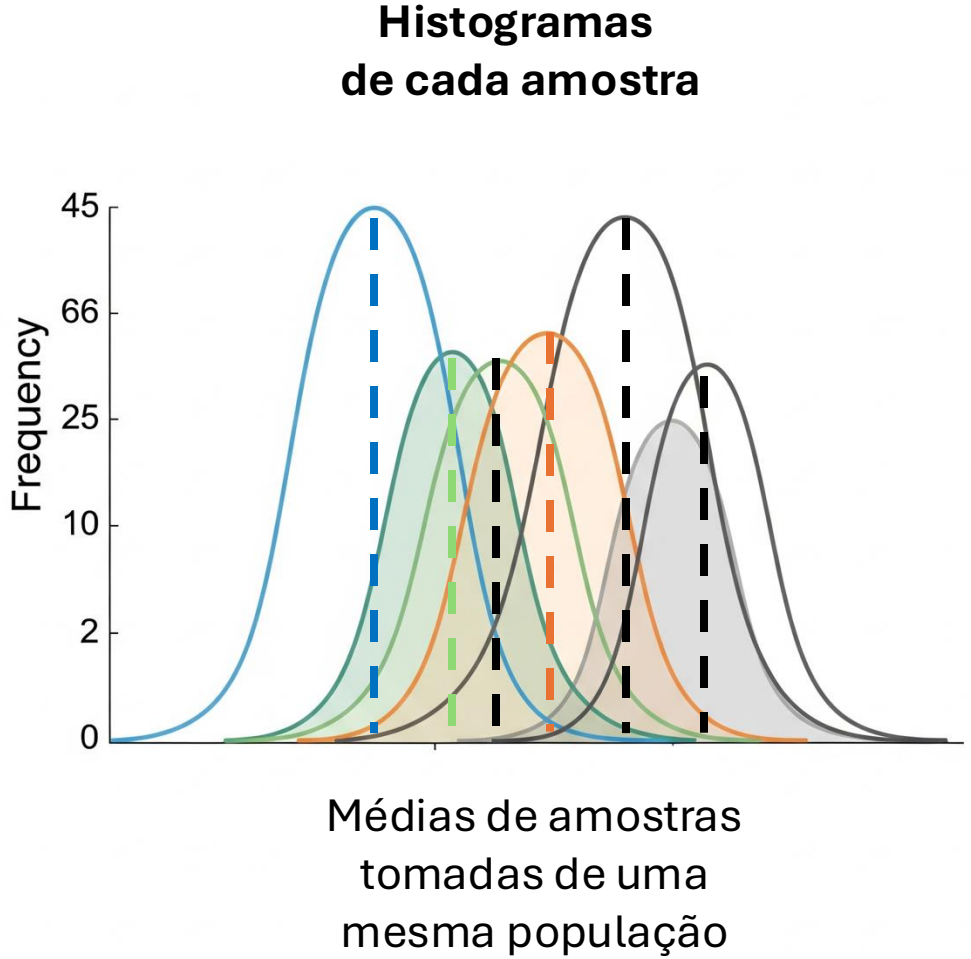
Todos pescaram no *mesmo lago*, mas cada um obteve uma média diferente.

Qual delas é a ‘correta’?

Podemos confiar em apenas um desses números para descrever a população inteira?

O Problema da Amostra Única

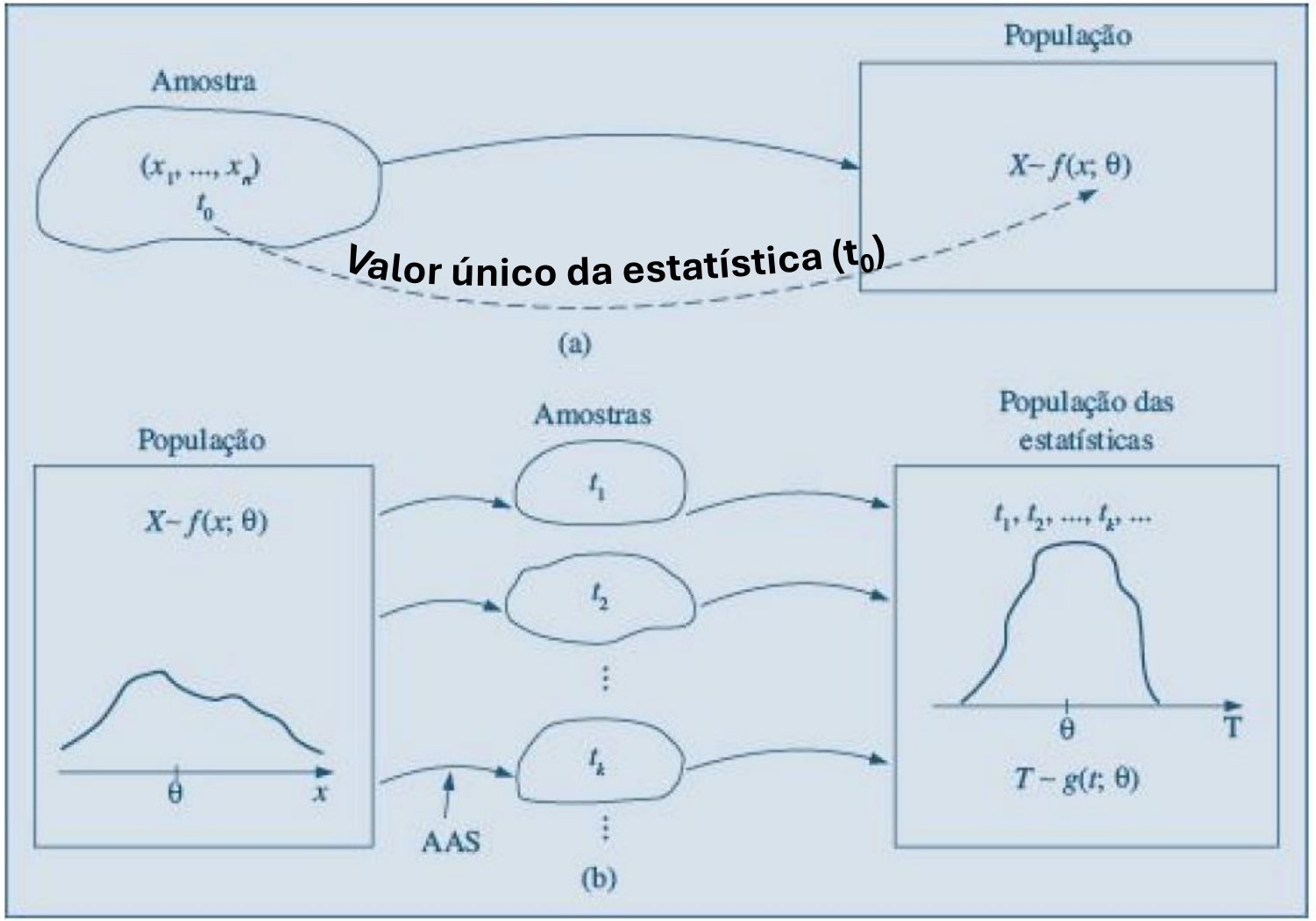
- Diferentes amostras da mesma população quase sempre terão médias e desvios padrões ligeiramente diferentes
- **Pense no experimento da "Pescaria":** Cada grupo obteve uma média de peso diferente para as tilápias, mesmo pescando no mesmo lago
- **Pergunta crucial:** Se temos apenas UMA amostra, como podemos usar suas estatísticas para inferir sobre a POPULAÇÃO inteira?



Estimativa Pontual: Nosso Melhor Palpite

- Uma **estimativa pontual** é um único valor calculado *a partir da amostra* que usamos como nossa melhor palpite (“chute”) para o parâmetro correspondente da população
- Determina o valor específico de um parâmetro, mede aonde a maioria dos dados estão
- Exemplos**:**
  - A **média amostral ($\bar{x}$)** é o melhor estimador não enviesado (BUE) para a **média populacional (μ)**
  - A **proporção amostral ($\hat{p}$)** é a melhor estimativa pontual para a **proporção populacional (p)**
  - O **desvio padrão amostral (s)** é o melhor estimador pontual para o desvio padrão **populacional (σ)**

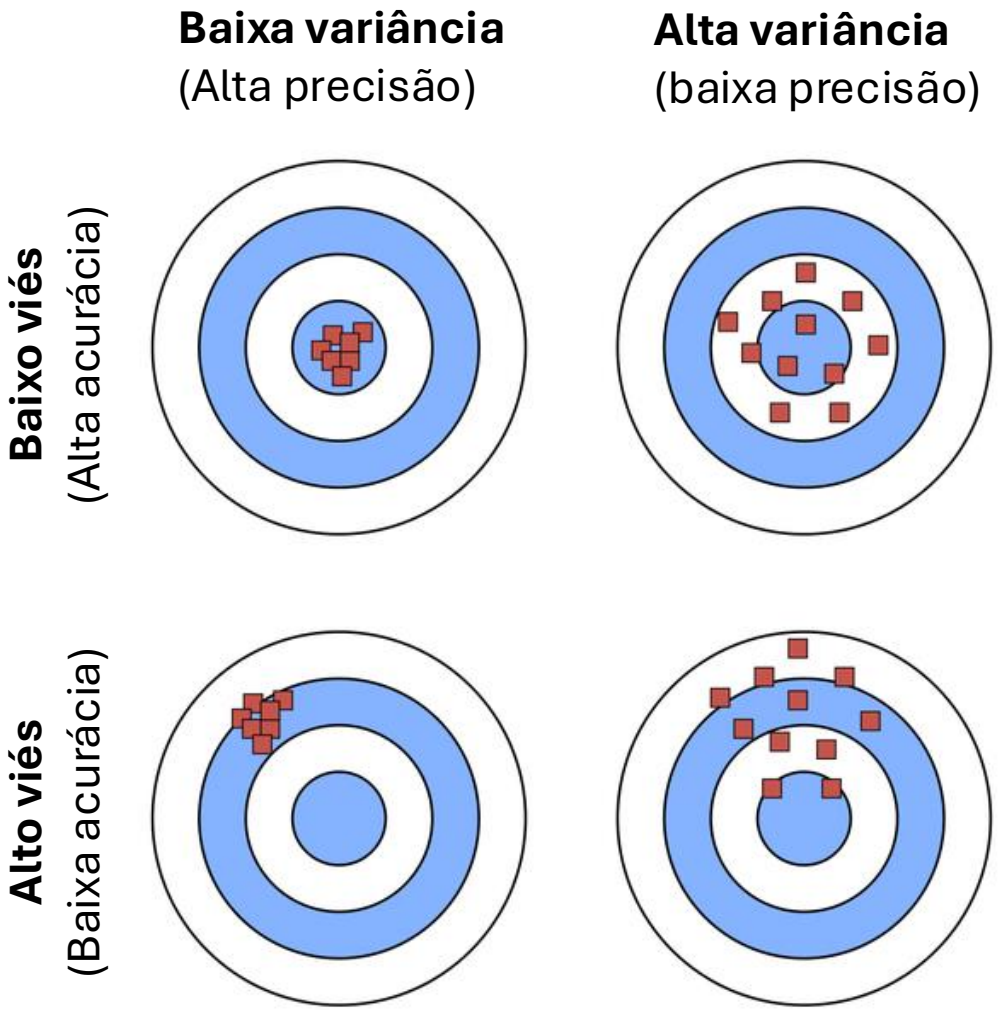
Figura 10.1 (a) Esquema de inferência sobre θ. (b) Distribuição amostral da estatística *T*.



Construção da distribuição amostral de uma estatística

Bussab & Morettin 2023

A Limitação de Estimadores Pontuais: Ausência da Margem de Erro



- Embora seja nosso melhor palpite, uma estimativa pontual isolada não nos diz o quão *precisa* ela provavelmente é
- **Imagine dizer:** "A altura média das árvores na floresta é de 15 metros."
- **A pergunta que fica:** Essa estimativa provavelmente está errada por alguns centímetros? Ou por vários metros?

Propriedades dos Estimadores Pontuais

- **Acurácia**: mede a proximidade entre cada observação e ***o valor alvo*** que se procura atingir

- **Precisão**: mede a proximidade entre cada observação e ***a média de todas as observações***

- Um estimador ideal para o parâmetro de interesse deve 1) ser *não enviesado*; 2) ter *pouca variância;* e 3) ser *consistente*

Bussab & Morettin 2023

“Um estimador não enviesado é uma estatística que tem como alvo o valor do parâmetro populacional correspondente, no sentido de que a distribuição amostral da estatística tenha uma média igual a esse parâmetro populacional correspondente”

Triola 2023

Estimativa pontual

- Se 3 condições forem satisfeitas, a média da amostra ($\bar{x}$) é um estimador não enviesado da média da população (μ):
  1. Observações são feitas em indivíduos escolhidos aleatoriamente
  2. Observações na amostra são independentes
  3. Observações são parte de uma população maior que pode ser descrita por uma *variável aleatória com distribuição normal*

Mais sobre isso nas próximas 2 aulas

Bussab & Morettin 2023

Lei dos Grandes Números

- Logo, à medida que o número de amostras **cresce**, a média da amostra ($\bar{x}$) **se aproxima** da média da população (μ)
- Esta é a **Lei dos Grandes Números (LGN)**
- Faz sentido então pensar que devemos (sempre que possível) ter o maior número de amostras que pudermos
- Isso serve para qualquer estimador pontual
  - Ou seja, se estivermos interessados em estimar a média da população, devemos tomar várias amostras e calcular a média delas porque a LGN garante que essas médias amostrais vão se aproximar da média da população à medida que aumentamos o número de amostras

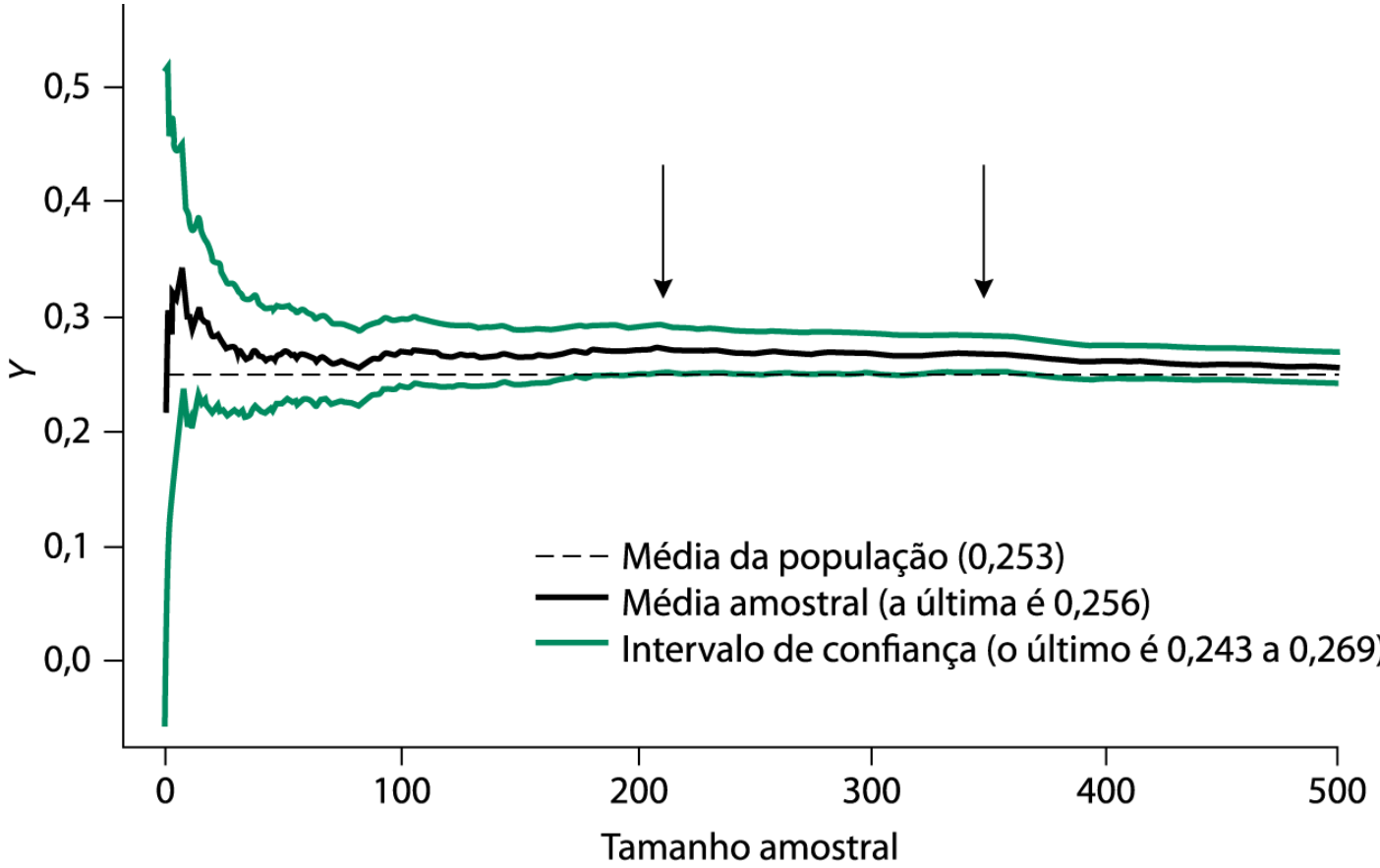


Ilustração do processo no R

Figura 3.1 Ilustração da Lei dos Grandes Números e a construção de intervalos de confiança usando os dados dos espinhos tibiais de aranhas da Tabela 3.1. A média da população (0,253) é indicada pela linha pontilhada. A média amostral para amostras de tamanhos crescentes (*n*) é indicada pela linha sólida central e ilustra a Lei dos Grandes Números: conforme o tamanho amostral aumenta, a média amostral se aproxima da verdadeira média da população. As linhas sólidas superior e a inferior ilustram o intervalo de confiança de 95% ao redor da média. A largura do intervalo de confiança decresce conforme o tamanho amostral aumenta. Intervalos de confiança de 95% construídos dessa forma devem conter a verdadeira média da população. Note, contudo, que existem amostras (entre as setas) para as quais o intervalo de confiança não inclui a verdadeira média da população. As curvas foram construídas usando algoritmos e códigos do S-Plus publicados por Blume e Royal (2003).

Gotelli & Ellison 2004

Distribuições amostrais

- O problema da inferência estatística é fazer uma afirmação sobre os parâmetros da população através da amostra
- Os valores encontrados para a amostra nem sempre serão iguais aos valores da população
- Para contemplar o erro possível, a estimativa do parâmetro pode ser feita de forma intervalar
- Para poder calcular o intervalo da estimação é preciso modelar e trabalhar com a **distribuição amostral do parâmetro estudado**

Estimativa intervalar

- **Definição**: intervalo de valores plausíveis para o parâmetro
- Mesmo sabendo que a **média da amostra** pode ser utilizada para estimar a **média da população**, precisamos ter uma medida de **incerteza da média da amostra**
- A média é uma **estimativa**. Logo, precisamos saber as **margens de erro** dessa estimativa
  - É construída adicionando e subtraindo uma **margem de erro** da estimativa pontual.
  - A largura do intervalo reflete a nossa **incerteza** sobre o valor verdadeiro
  - Precisa vir acompanhada de uma declaração de probabilidade

Margem de erro

*Ao usar uma estatística amostral para estimar um parâmetro populacional, a margem de erro (**E**), é a quantidade máxima provável de erro (a quantidade pela qual a estatística amostral erra o parâmetro populacional)*

Margem de erro da média (para variâncias desconhecidas)

$$E = t_{\alpha} \frac{s}{\sqrt{n}}$$

Valor crítico para o nível de confiança desejado

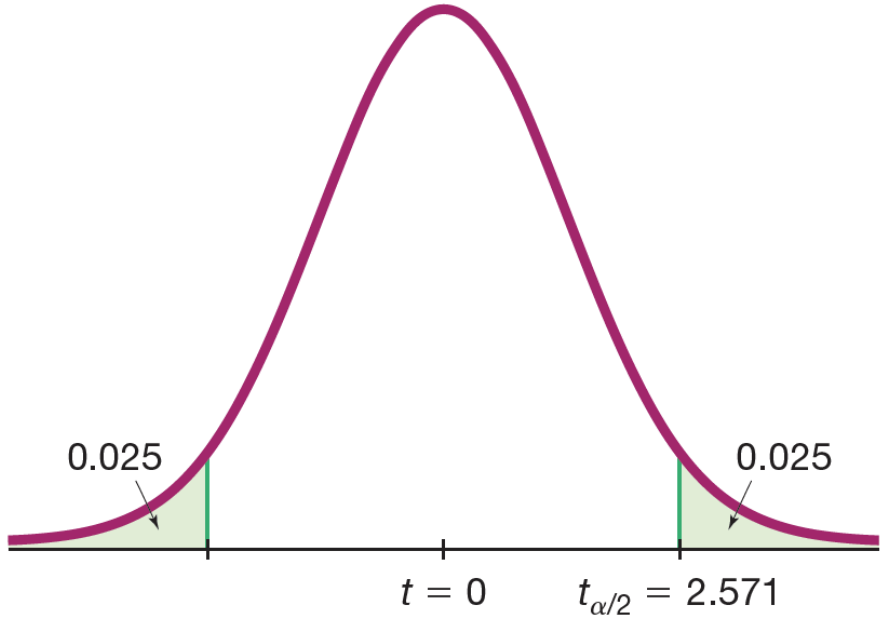


FIGURE 7-5 **Critical Value $t_{\alpha/2}$**

- S=desvio padrão
- n=tamanho da amostra
- t = valor crítico da distribuição *t* para dado grau de liberdade (gl) e nível de significância (α)

CAPÍTULO 7

Estimação de Parâmetros e Determinação de Tamanhos Amostrais

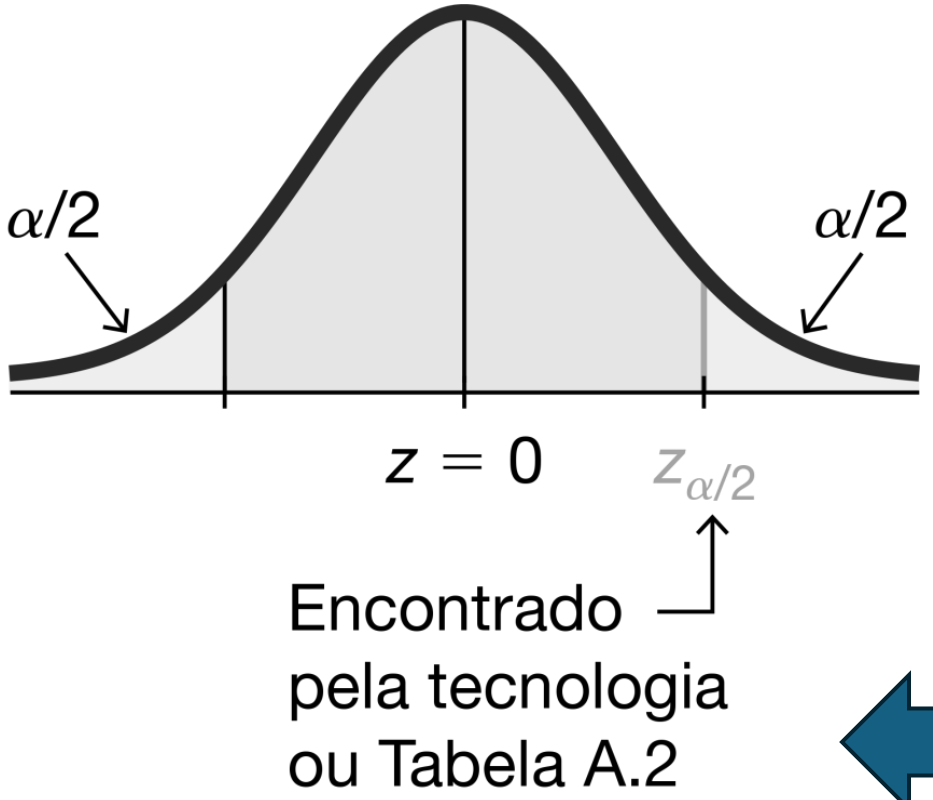


FIGURA 7.1 Valor Crítico $z_{\alpha/2}$ na Distribuição Normal Padrão.

Definição de valor crítico: Para a distribuição normal padrão, um valor crítico é o escore Z no limite que separa os escores Z significantemente baixos ou significantemente altos.

A figura mostra exemplo com nível de significância $\alpha = 0.05$ para um teste bicaudal (valor é dividido por 2). Isso corresponde a 2 desvios-padrões da média, para mais e para menos

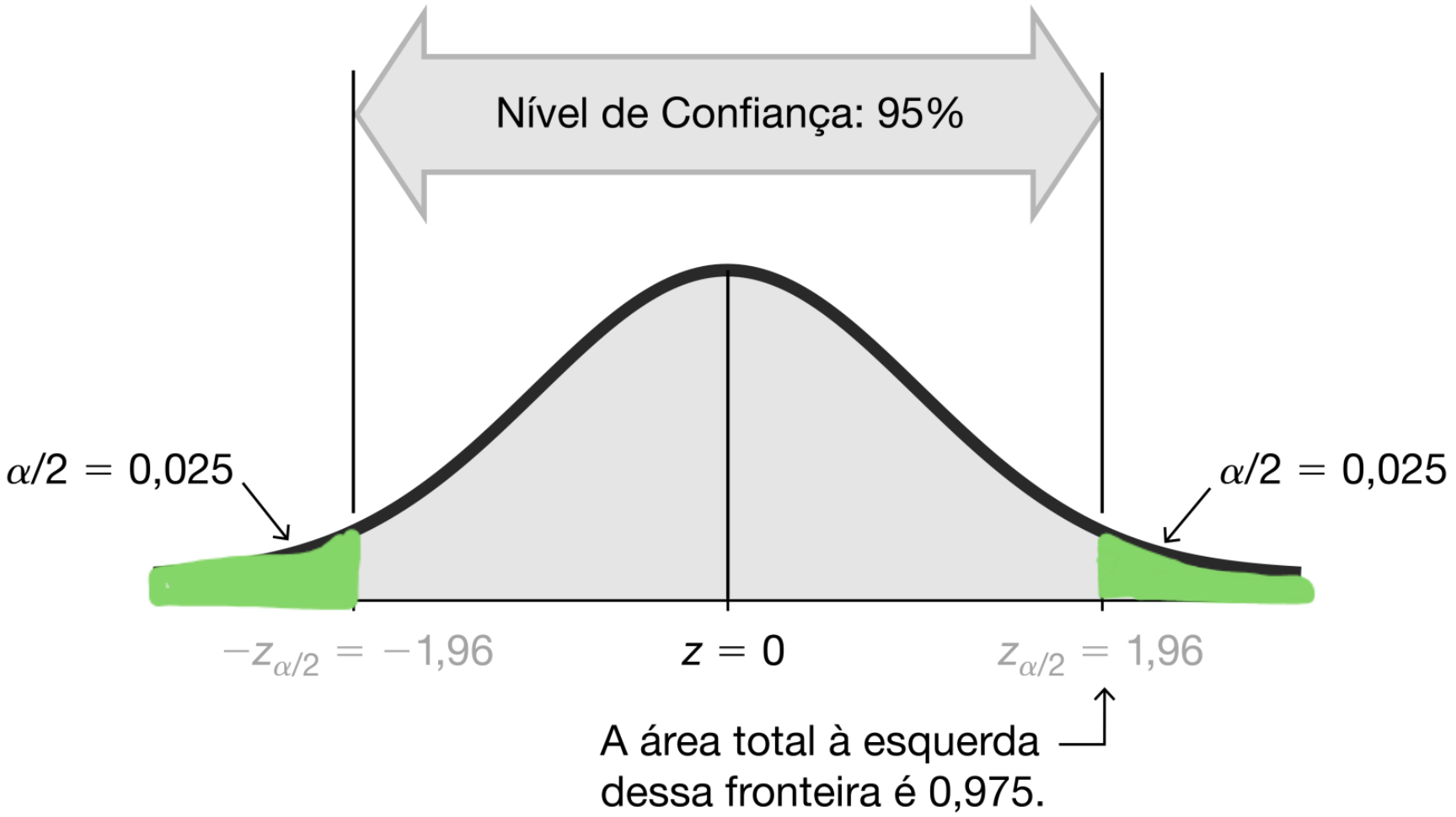


FIGURA 7.2 Determinação do Valor Crítico $z_{\alpha/2}$ para um Nível de Confiança de 95%.

Triola 2023

Variância e Erro padrão da média

- **Variância** da média: **estima a variabilidade das médias** que seriam obtidas, caso o pesquisador tivesse tomado, nas mesmas condições, todas as amostras possíveis

$$s_{\bar{x}}^2 = \frac{s^2}{n}$$

- **Erro padrão** da média:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- O **erro padrão (SE)** quantifica a **variabilidade das estatísticas amostrais** (como a média ou a proporção) se repetirmos a amostragem várias vezes.

Vieira 2022

Variância e Erro padrão da média

- Média das amostras têm variabilidade menor do que os dados

- Média das amostras de *n* dados têm dispersão menor do que os dados que as compõem

Vieira 2022

Estimativa intervalar

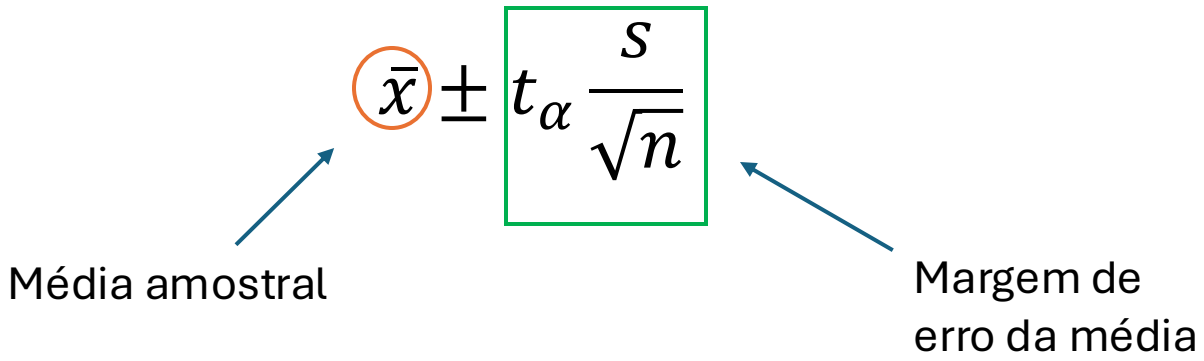
- Intervalo de confiança
  - Quando a variância da população é **conhecida**
  - Quando a variância da população é **desconhecida**

Intervalo de confiança

- Um **Intervalo de Confiança (IC)** é um tipo específico de estimativa intervalar usado para estimar o valor real do parâmetro populacional. Sempre vem acompanhado de um **nível de confiança** (geralmente 90%, 95% ou 99%)
  - Dá idéia de incerteza relacionada à estimativa pontual

- **Interpretação (Exemplo de 95% IC):** Se repetirmos o processo de amostragem muitas vezes e construirmos um IC de 95% para cada amostra, aproximadamente 95% desses intervalos conterão o verdadeiro parâmetro populacional

Construindo um Intervalo de Confiança para a Média (Amostras pequenas)

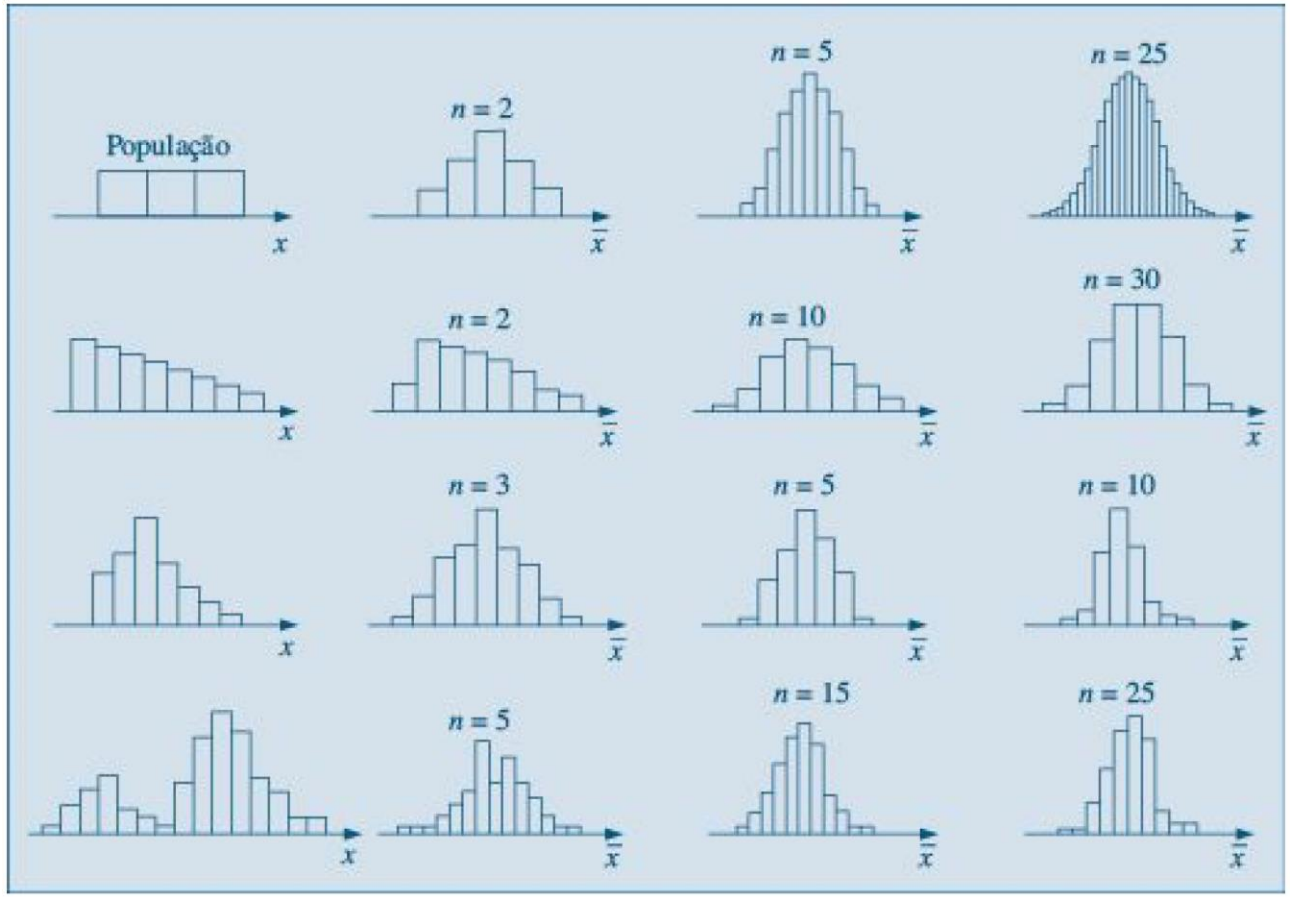


Teorema do Limite Central

- Se o parâmetro estudado for a média, é preciso analisar a distribuição das médias amostrais: ou seja, a forma como as frequências de médias amostrais calculadas costumam se distribuir

- Definição: Para Amostras Aleatórias Simples ($X_1, \ldots X_n$), retiradas de uma população com média μ, e variância σ^2 finita, a distribuição amostral da média x, **aproxima-se, para *n* grande, de uma distribuição normal**, com média μ e variância σ^2 /n.

4 distribuições diferentes... Convergem para a Normal (Gaussiana) ao aumentarmos o tamanho amostral

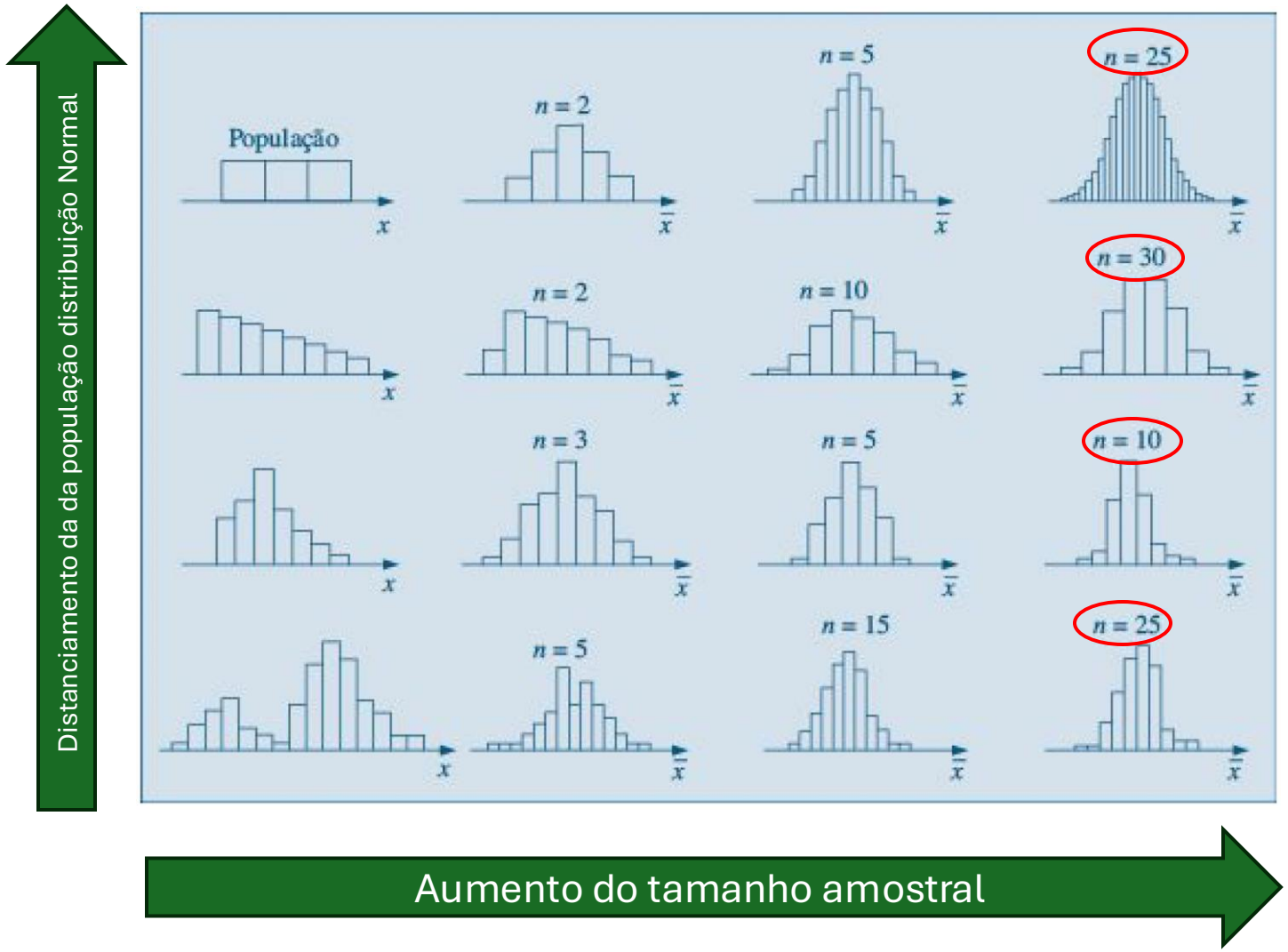


Aumento do tamanho amostral

Bussab & Morettin 2023

Vamos ver mais sobre distribuições estatísticas nas próximas aulas

4 distribuições diferentes... Convergem para a Normal (Gaussiana) ao aumentarmos o tamanho amostral



Quanto maior o distanciamento da população da distribuição Normal, maior tem de ser o tamanho amostral para que a distribuição da amostra convirja para a Normal

Bussab & Morettin 2023

Vamos ver mais sobre distribuições estatísticas nas próximas aulas

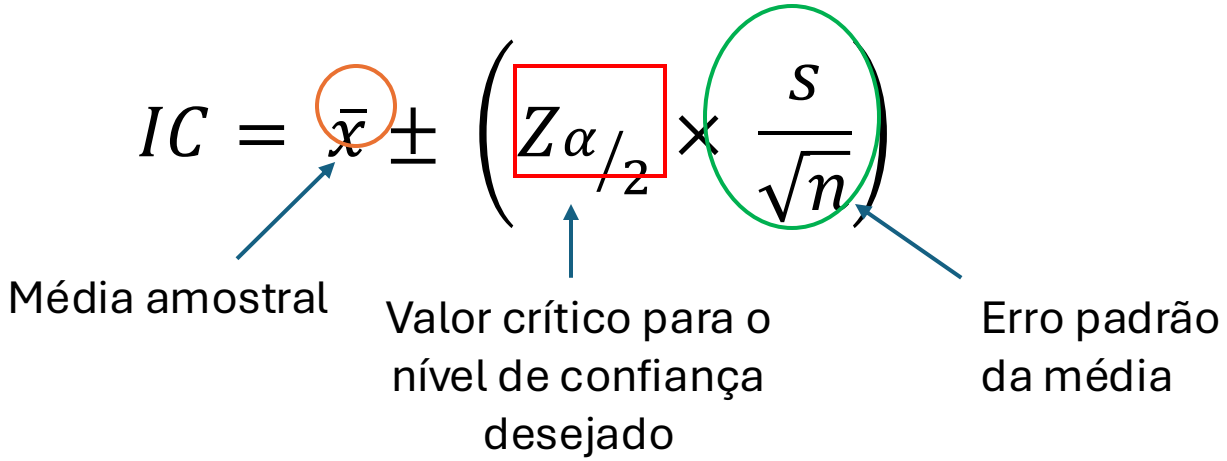
Usando simulações para entender o Teorema do Limite Central

- Vídeo introdutório
  - https://vimeo.com/75089338

- http://mfviz.com/central-limit/

Construindo um Intervalo de Confiança para a Média (Grandes Amostras)

- Usando a Distribuição Normal (Z-score)
- Para uma amostra grande (geralmente n ≥ 30), **podemos usar a distribuição normal para aproximar a distribuição das médias amostrais (Garantido pelo Teorema Limite Central)**



Escolha da Distribuição Correta

Na construção de uma estimativa de intervalo de confiança para a média populacional μ, é importante usar a distribuição correta. A Tabela 7.1 resume os pontos-chave a serem considerados.

TABELA 7.1 Escolha da Distribuição Correta

Condições	Método
σ desconhecido e população normalmente distribuída ou σ desconhecido e $n > 30$	Use a distribuição t de Student com $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$
σ conhecido e população normalmente distribuída ou σ conhecido e $n > 30$ (Na realidade, σ raramente é conhecido.)	Use a distribuição normal (z) com $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Exemplo Biológico: Nível de Glicose em Leões

- Um pesquisador coleta uma amostra aleatória de 40 leões adultos em uma reserva e mede seus níveis de glicose no sangue. Os resultados foram:
  - Média da amostra ($\bar{x}$): 95 mg/dL
  - Desvio padrão da amostra (s): 15 mg/dL
- Calcule o Intervalo de Confiança de 95% para o nível médio de glicose na população de leões
  - Erro padrão: $15/\sqrt{40} = 2.37$ mg/dL
  - Valor crítico: $Z_{0.025} = 1{,}96$
  - Margem de erro: 1,96 x 2,37 = 4,65 mg/dL
  - Intervalo de Confiança $95 \pm 4{,}65$ = [90,35; 99,65] mg/dL

Interpretação

Estamos 95% confiantes de que o verdadeiro nível médio de glicose na população de leões adultos da reserva está entre 90.35 mg/dL e 99.65 mg/dL.

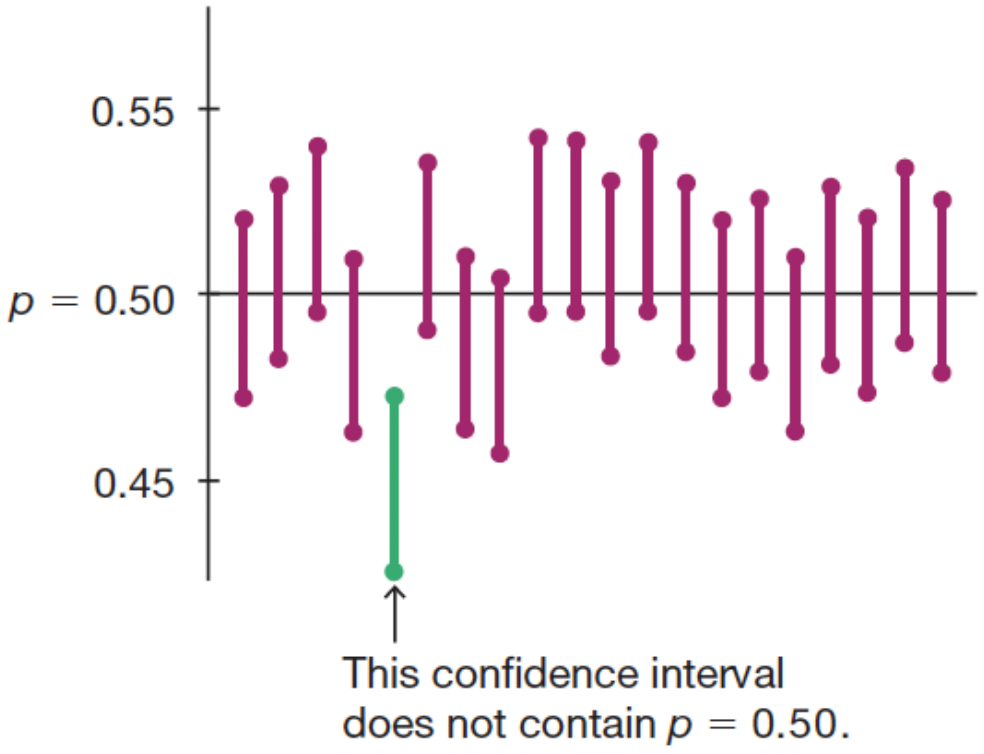
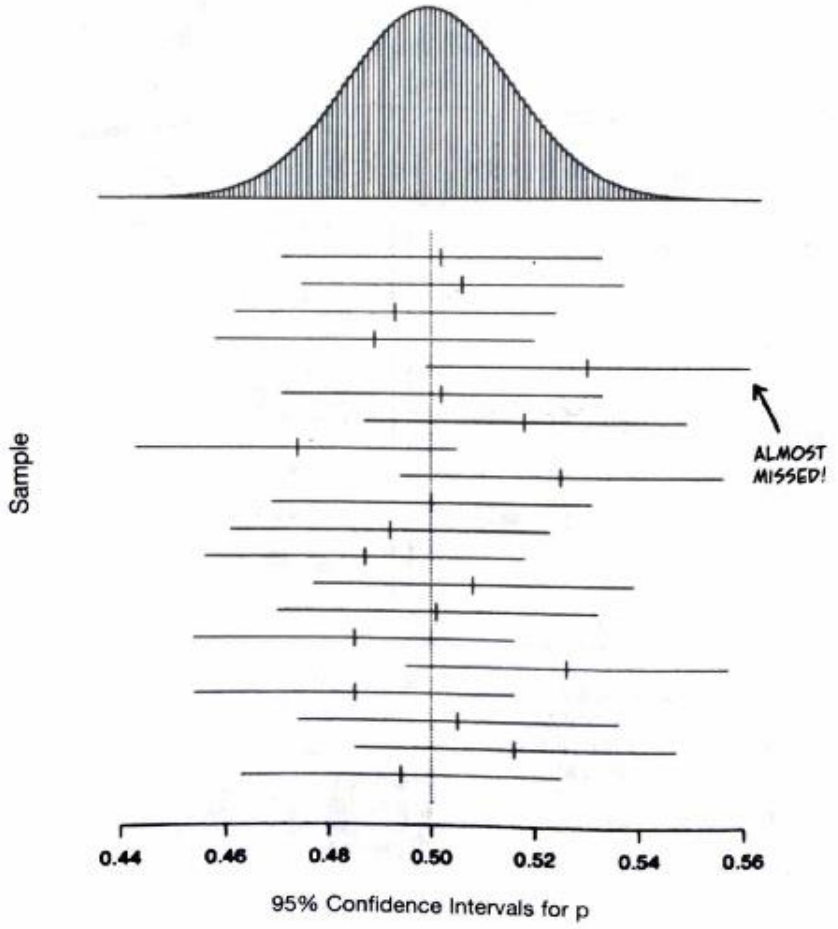


FIGURE 7-1 Confidence Intervals from 20 Different Samples

19 de 20 são 95%!

Triola 2023

THIS PAGE SHOWS THE RESULTS OF A COMPUTER SIMULATION OF TWENTY SAMPLES OF SIZE n = 1,000. THE SIMULATION SET THE TRUE VALUE OF p = 0.5. AT THE TOP YOU SEE THE SAMPLING DISTRIBUTION OF $\hat{P}$ (NORMAL, WITH MEAN p AND $\sigma = \sqrt{p(1-p)/n}$). BELOW ARE THE 95% CONFIDENCE INTERVALS FROM EACH SAMPLE. ON AVERAGE, ONE OUT OF TWENTY (OR 5%) OF THESE INTERVALS **WILL NOT COVER THE POINT** p = 0.5.



Gonick & Smith 2005

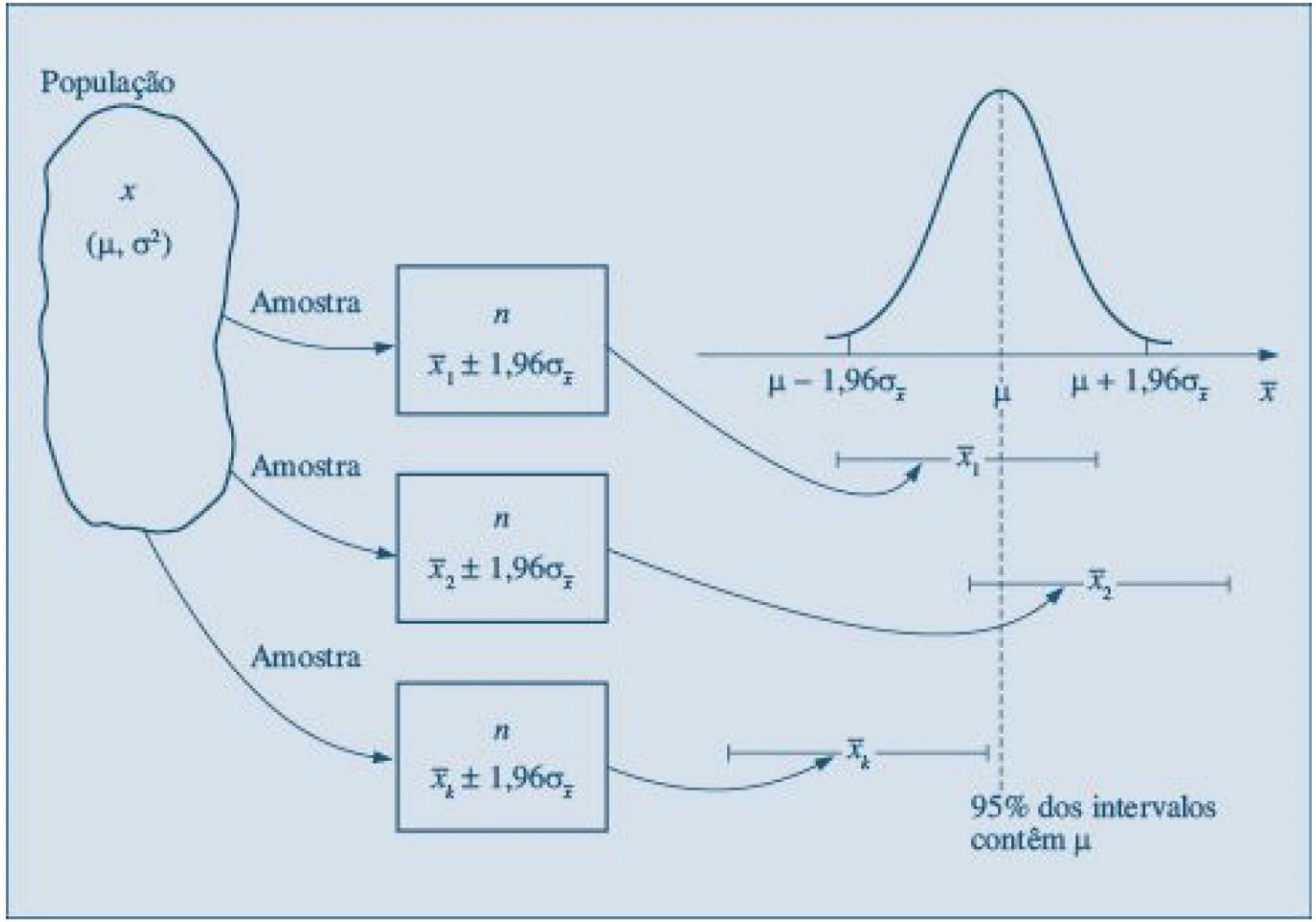


Figura 11.3 Significado de um IC para μ, com $\gamma = 0{,}95$ e σ^2 conhecido.

Bussab & Morettin 2023

Atividade: entendendo o intervalo de confiança para dados contínuos

- https://www.rossmanchance.com/applets/2021/confsim/ConfSim.html?

Atividade: entendendo o intervalo de confiança

- Cada grupo vai "coletar uma amostra" clicando no botão do simulador. A cada clique, o programa desenha uma linha que representa o IC de 95% para aquela amostra.
- Depois de ter uns 20-25 intervalos desenhados na tela, revele o valor verdadeiro da média da população (μ=250 g), que aparecerá como uma linha vertical.
- Conte quantos dos intervalos desenhados (as "redes de pesca") realmente "capturaram" o parâmetro verdadeiro (a linha vertical).

Construindo um Intervalo de Confiança para Proporções (Amostra Grande)

- Usando a Distribuição Normal (Z-score)
- Para uma amostra grande (geralmente $n\hat{p} \geq 10$ e $n(1-\hat{p}) \geq 10$), podemos usar a distribuição normal para aproximar a distribuição das proporções amostrais

$$IC = \hat{p} \pm \left(Z\alpha_{/2} \times SE_{\hat{p}} \right)$$

- Erro padrão para proporções:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Exemplo Biológico: Prevalência de um Alelo

- Em uma amostra aleatória de **200 borboletas** de uma população, **30** possuem **um alelo** específico para resistência a um pesticida
- Calcule o **Intervalo de Confiança de 99%** para a proporção populacional de borboletas com esse alelo
  - Proporção amostral: 30/200 = 0.15
  - Erro padrão: $\sqrt{\frac{0.15(1-0.15)}{200}}$ = 0.0252
  - Valor crítico (99%): $Z_{0.005} = 2.576$
  - Margem de erro: 2.576 x 0.0252 = 0.0649
  - Intervalo de confiança: 0.15 ± 0.0649 = [0.0851,0.2149]

Interpretação

Estamos 99% confiantes de que a verdadeira proporção de borboletas na população com o alelo de resistência está entre 8,51% e 21,49%.

Fatores que Afetam a Largura do IC

- **Nível de Confiança:** Quanto maior o nível de confiança, mais largo o intervalo (precisamos de uma "rede" maior para ter mais certeza de capturar o alvo)
- **Tamanho da Amostra (*n*):** Quanto maior o tamanho da amostra, menor o erro padrão e mais estreito o intervalo (mais informação leva a maior precisão)
  - O fato de o IC ficar menor não significa que contenha o parâmetro. Conter o parâmetro é uma probabilidade
- **Variabilidade da Amostra (*s* ou $\hat{p}(1-\hat{p})$):** Maior variabilidade na amostra leva a um erro padrão maior e um intervalo mais amplo

Atividade: entendendo o intervalo de confiança para proporções

- https://www.rossmanchance.com/applets/2021/confsim/ConfSim.html?

O Cenário Mendeliano (5 minutos)

- Gregor Mendel cruzou plantas de ervilha e previu que, no cruzamento de heterozigotos, a proporção de descendentes com a semente rugosa (traço recessivo) deveria ser de **25%** (p=0.25).
- Assim como no nosso lago de peixes, vamos assumir que nós somos a 'Natureza' e sabemos esse valor verdadeiro.
- Mas um pesquisador que repete o experimento hoje não sabe disso. Ele só tem acesso à sua amostra.

O Cenário Mendeliano (5 minutos)

- Um grupo vai “plantar 100 sementes” clicando no botão “Draw Sample(s)”. O simulador mostrará o resultado daquela amostra (ex: 28 sementes rugosas em 100, ou seja, p=0.28)
- Clique no botão várias vezes (ou digite “25” ou “100” e clique de uma vez) para simular vários pesquisadores repetindo o experimento. A tela se encherá com os intervalos de confiança
- Quantos dos nossos intervalos de confiança conseguiram 'capturar' a proporção verdadeira prevista por Mendel?

A lógica é idêntica à da média. Não importa se estamos estimando um peso médio ou uma proporção de sucesso. O intervalo de confiança é a nossa 'rede de pesca' para o parâmetro populacional verdadeiro. E ***o nível de confiança nos diz quão confiável é o nosso método de construção da rede***

Recapitulando

- **Estimativa Pontual:** Nossa melhor estimativa/suposição para um parâmetro populacional.
- **Erro Padrão:** Mede a variabilidade das nossas estimativas amostrais.
- **Estimativa Intervalar (IC):** Uma faixa de valores que provavelmente contém o parâmetro populacional verdadeiro, com um nível de confiança associado.
- O IC nos dá uma ideia da **precisão** da nossa estimativa pontual.

Kahoot!

https://play.kahoot.it/v2/lobby?quizId=5a154caf-a064-4482-9305-3970ebf3bdc6

Atividade de hoje (Parte da Lista de Exercício)

- Ler Item 11.6 do Cap 11 de Morettin & Bussab (pdf no Moodle)
- Fazer exercícios 14 a 16 e 20