Aula 03. Intervalos de Confiança

Estatística Inferencial

MBA CDIA
ENAP - Escola Nacional de Administração Pública
2025

Agenda

1. **Revisão: Estimação Pontual e EMV**
2. **Intervalos de Confiança**
   - Derivação do IC para média
   - IC para média com σ conhecido
   - IC para média com σ desconhecido
   - IC para proporção populacional
3. **Interpretação e Propriedades dos ICs**
4. **Determinação do Tamanho Amostral**
5. **Exercícios práticos em Python**

Revisão: Estimação Pontual e EMV

Estimação Pontual:

- θ = parâmetro populacional
- $\hat{\theta}$ = estimador
- Propriedades desejáveis:
  - Não-viesado: $E[\hat{\theta}] = \theta$
  - Eficiente: menor variância
  - Consistente: $\hat{\theta} \to \theta$ quando $n \to \infty$

Máxima Verossimilhança (EMV):

- $L(\theta | dados) = \prod_{i=1}^{n} f(x_i | \theta)$
- $\hat{\theta}_{EMV} = \arg\max_{\theta} L(\theta)$
- Exemplos:
  - Normal: $\hat{\mu} = \bar{X}$
  - Exponencial: $\hat{\lambda} = 1/\bar{X}$
  - Bernoulli: $\hat{p} = \bar{X}$

Limitação: Uma estimativa pontual não fornece informação sobre a incerteza!

Intervalos de Confiança - Motivação

- Uma estimativa pontual $\hat{\theta}$ raramente será igual ao parâmetro populacional θ
- É mais informativo fornecer um **intervalo de valores plausíveis**
- O intervalo quantifica a incerteza da estimação

Exemplo prático:

- Estimativa pontual: "A satisfação média é 7.5"
- Intervalo de confiança: "Com 95% de confiança, a satisfação média está entre 7.3 e 7.7"

Definição

Um intervalo de confiança de nível $(1-\alpha)$ para um parâmetro populacional é uma regra de construção de intervalos baseada em amostras aleatórias, tal que, se repetirmos o processo de amostragem muitas vezes, a proporção de intervalos que conterá o verdadeiro valor do parâmetro será $(1-\alpha)$.

Objetivo: Construir um IC para μ quando σ^2 é conhecido.
Sabemos que: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
Padronizando: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
Para 95% de confiança:

$$P(-1.96 < Z < 1.96) = 0.95$$

Substituindo Z**:**

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Derivação do IC para Média (continuação)

Multiplicando por $\sigma/\sqrt{n}$:

$$P\left(-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Subtraindo $\bar{X}$ e multiplicando por -1:

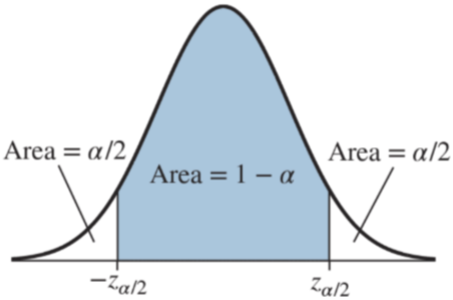
$$P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Forma geral do IC de $(1-\alpha)100\%$:

$$IC(\mu) = \bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

onde $z_{\alpha/2}$ é o quantil da $N(0,1)$ que deixa $\alpha/2$ na cauda superior.

Valores Críticos da Distribuição Normal



Confiança	α	$z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

Interpretação:

- 95% dos valores de uma $N(0,1)$ estão entre -1.96 e 1.96
- Áreas nas caudas: 2.5% cada

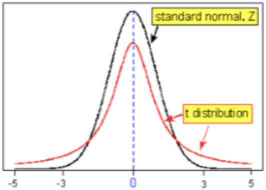
IC para Média - σ desconhecido

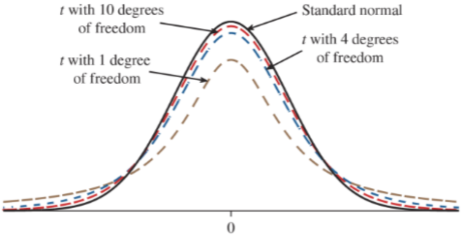
Problema: Na prática, raramente conhecemos σ
Solução: Usar o desvio padrão amostral s e a distribuição t de Student
Estatística: $T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$
IC de $(1-\alpha)100\%$**:**

$$IC(\mu) = \bar{X} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

Propriedades da distribuição t**:**

- Simétrica, centrada em 0
- Caudas mais pesadas que a normal
- Converge para $N(0,1)$ quando $n \to \infty$
- Graus de liberdade: $gl = n - 1$





Quando usar cada distribuição:

- **Distribuição Normal:** quando σ é conhecido
- **Distribuição t:** quando σ é desconhecido (caso mais comum)
- Para $n > 30$, as distribuições são praticamente idênticas

Interpretação:

- 95% dos valores de uma $N(0,1)$ estão entre -1.96 e 1.96
- Áreas nas caudas: 2.5% cada

Exemplo: IC para Média

Problema: Uma amostra de 25 tempos de atendimento em um órgão público teve média de 18.5 minutos e desvio padrão de 4.2 minutos. Construa um IC de 95% para o tempo médio de atendimento.

Solução:

- Dados: $n = 25$, $\bar{x} = 18.5$, $s = 4.2$
- Como σ é desconhecido, usar distribuição t
- Graus de liberdade: $gl = 25 - 1 = 24$
- Para 95% de confiança: $t_{0.025,24} = 2.064$
- Erro padrão: $\frac{s}{\sqrt{n}} = \frac{4.2}{\sqrt{25}} = 0.84$
- Margem de erro: $2.064 \times 0.84 = 1.73$
- IC: 18.5 ± 1.73 ou $(16.77, 20.23)$

Interpretação: Com 95% de confiança, o tempo médio de atendimento está entre 16.77 e 20.23 minutos.

Implementação em Python - IC para Média

```
import scipy.stats as stats
import numpy as np

# Dados do problema
n = 25
media_amostral = 18.5
desvio_padrao = 4.2
confianca = 0.95

# Método 1: Usando stats.t.interval() - mais simples!
erro_padrao = desvio_padrao / np.sqrt(n)
ic = stats.t.interval(confidence=confianca,
                      df=n-1,
                      loc=media_amostral,
                      scale=erro_padrao)

print(f"IC de 95%: ({ic[0]:.2f}, {ic[1]:.2f})")
# Saída: IC de 95%: (16.77, 20.23)

# Método 2: Usando stats.sem() para erro padrão
# erro_padrao = stats.sem(dados, ddof = 1)  # se tivéssemos os dados brutos
```

Objetivo: Estimar uma proporção populacional p

Estimador: $\hat{p} = \frac{X}{n}$ onde X = número de sucessos

Distribuição aproximada: Para n grande, $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

IC de $(1-\alpha)100\%$:

$$IC(p) = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Condições de validade:

- $n\hat{p} \geq 10$ e $n(1-\hat{p}) \geq 10$
- Garante boa aproximação normal

Exemplo: IC para Proporção

Problema: Em uma pesquisa com 500 cidadãos, 320 aprovaram um serviço público. Construa um IC de 99% para a proporção de aprovação na população.

Solução:

- Dados: $n = 500$, $X = 320$
- Proporção amostral: $\hat{p} = \frac{320}{500} = 0.64$
- Verificar condições:
  - $500 \times 0.64 = 320 \geq 10$
  - $500 \times 0.36 = 180 \geq 10$
- Para 99% de confiança: $z_{0.005} = 2.576$
- Erro padrão: $\sqrt{\frac{0.64 \times 0.36}{500}} = 0.0215$
- Margem de erro: $2.576 \times 0.0215 = 0.055$
- IC: 0.64 ± 0.055 ou $(0.585, 0.695)$

Interpretação: Com 99% de confiança, entre 58.5% e 69.5% da população aprova o serviço.

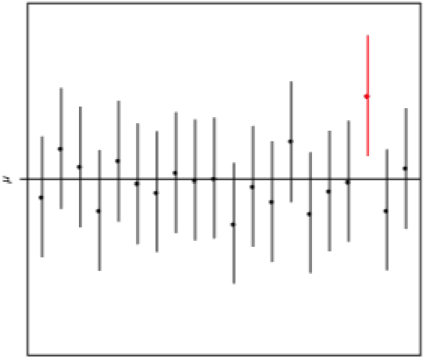
Dica Python: Use `statsmodels.stats.proportion.proportion_confint()` para cálculo direto!

Interpretação Correta do IC

O que o IC significa:

- Se repetíssemos o processo de amostragem muitas vezes
- E construíssemos um IC de 95% em cada vez
- Aproximadamente 95% desses intervalos conteriam θ

Interpretação prática: "Temos 95% de confiança que o verdadeiro valor do parâmetro está no intervalo"



Cuidado:

- NÃO diga: "Há 95% de probabilidade de θ estar no intervalo"
- O parâmetro é fixo, o intervalo é aleatório

Fatores que Afetam a Largura do IC

A largura do IC é determinada pela margem de erro: $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Fator	**Mudança**	**Efeito na Largura**
Tamanho da amostra (n)	Aumenta	Diminui
Nível de confiança ($1 - \alpha$)	Aumenta	Aumenta
Variabilidade (σ)	Aumenta	Aumenta

Trade-off fundamental:

- Maior confiança $\Rightarrow$ Intervalo mais largo (menos preciso)
- Maior precisão $\Rightarrow$ Menor confiança ou maior amostra necessária

Determinação do Tamanho da Amostra

Pergunta: Qual tamanho de amostra necessário para atingir uma margem de erro desejada?
Para estimar uma média: Dado margem de erro E e confiança $(1-\alpha)$:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Para estimar uma proporção:

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \cdot p(1-p)$$

Se p desconhecido, usar $p = 0.5$ (caso mais conservador)
Exemplo: Para estimar uma proporção com erro de 3% e 95% de confiança:

$$n = \left(\frac{1.96}{0.03}\right)^2 \cdot 0.5 \cdot 0.5 = 1067$$

Exercício 1: IC para Média com Dados Reais

Enunciado: Você tem os tempos de atendimento (em minutos) de 20 cidadãos em um órgão público. Calcule o IC de 95% para o tempo médio de atendimento.

```
import numpy as np
from scipy import stats

# Dados
tempos = [15.2, 18.5, 22.1, 16.8, 19.3,
          21.5, 17.2, 20.8, 18.9, 16.5,
          19.7, 23.2, 18.1, 17.9, 20.3,
          19.5, 18.8, 21.1, 17.6, 19.2]

# 1. Estatísticas descritivas
media = np.mean(tempos)
desvio = np.std(tempos, ddof=1)  # ddof=1 para amostra

# 2. IC de 95% usando stats.t.interval()
ic = stats.t.interval(confidence=0.95,
                      df=len(tempos)-1,
                      loc=media,
                      scale=stats.sem(tempos, ddof = 1))

print(f"Média: {media:.2f} minutos")
print(f"Desvio padrão: {desvio:.2f} minutos")
print(f"IC de 95%: ({ic[0]:.2f}, {ic[1]:.2f})")
# Resultado: IC de 95%: (18.36, 20.22)
# Interpretação: Com 95% de confiança, o tempo médio
# de atendimento está entre 18.36 e 20.22 minutos.
```

Exercício 2: IC para Proporção

Enunciado: Em uma pesquisa de satisfação com 200 servidores públicos, 156 se declararam satisfeitos com o ambiente de trabalho. Calcule o IC de 95% para a proporção de servidores satisfeitos.

```
from statsmodels.stats.proportion import proportion_confint
from scipy import stats
import numpy as np

# Dados
n = 200  # tamanho da amostra
x = 156  # número de satisfeitos

# 1. Proporção amostral
p_hat = x / n
print(f"Proporção amostral: {p_hat:.3f}")  # 0.780

# 2. Erro padrão
erro_padrao = np.sqrt(p_hat * (1 - p_hat) / n)
print(f"Erro padrão: {erro_padrao:.4f}")  # 0.0293

# 3. IC de 95% usando aproximação normal
z_critico = stats.norm.ppf(0.975)  # 1.96 para 95%
margem_erro = z_critico * erro_padrao

ic_inferior = p_hat - margem_erro
ic_superior = p_hat + margem_erro

print(f"IC de 95%: ({ic_inferior:.3f}, {ic_superior:.3f})")
# Resultado: IC de 95%: (0.723, 0.837)
```

Método Mais Simples - IC para Proporção

Usando statsmodels (mais fácil e robusto):

```
from statsmodels.stats.proportion import proportion_confint

# Mesmos dados do exercício anterior
n = 200  # tamanho da amostra
x = 156  # número de satisfeitos

# IC de 95% com apenas uma linha!
ic = proportion_confint(x, n, alpha=0.05, method='normal')

print(f"IC de 95%: ({ic[0]:.3f}, {ic[1]:.3f})")
# Resultado: IC de 95%: (0.723, 0.837)

# Outros métodos disponíveis:
# method='wilson' - melhor para amostras pequenas
# method='beta' - método exato
# method='agresti_coull' - ajuste para amostras pequenas
```

Enunciado: Compare os ICs quando você conhece ($\sigma = 1$) vs. não conhece o desvio padrão populacional.

```
from scipy import stats
import numpy as np

# Dados de uma amostra
dados = [23.5, 25.1, 24.8, 26.2, 25.5,
         24.3, 25.8, 24.9, 25.3, 26.0]
media = np.mean(dados)
n = len(dados)

# 1. IC com sigma conhecido (usar distribuição normal)
sigma_conhecido = 1.0
ic_normal = stats.norm.interval(0.95, loc=media, scale=sigma_conhecido/np.sqrt(n))

# 2. IC com  desconhecido (usar distribuição t)
ic_t = stats.t.interval(0.95, df=n-1, loc=media, scale=stats.sem(dados))

print(f"Média amostral: {media:.2f}")
print(f"IC ( conhecido):   ({ic_normal[0]:.2f}, {ic_normal[1]:.2f})")
print(f"IC ( desconhecido): ({ic_t[0]:.2f}, {ic_t[1]:.2f})")

print(f"\nLargura IC normal: {ic_normal[1] - ic_normal[0]:.3f}")
print(f"Largura IC t: {ic_t[1] - ic_t[0]:.3f}")
# O IC usando t é mais largo (mais conservador)
```

Resumo da Aula

Conceitos principais:

- **Limitação da estimação pontual:** não quantifica incerteza
- **Intervalo de Confiança:** fornece faixa de valores plausíveis
- **Componentes do IC:** estimativa $\pm$ margem de erro

Casos importantes:

- **IC para média (σ conhecido):** usar distribuição normal
- **IC para média (σ desconhecido):** usar distribuição t
- **IC para proporção:** usar aproximação normal quando n grande

Interpretação:

- IC quantifica a incerteza amostral
- Trade-off entre confiança e precisão
- Tamanho da amostra afeta diretamente a precisão

Próxima aula: Testes de Hipóteses