

# Pacotes Estatísticos

---

Professor:

Frederico R. B. Cruz

*E-mail: [fcruz@est.ufmg.br](mailto:fcruz@est.ufmg.br)*

*Departamento de Estatística*

*Universidade Federal de Minas Gerais*

# Objetivos da disciplina

---

- Capacitar o estudante a usar **recursos avançados** do principal software estatístico disponível no Departamento:



# Ementa

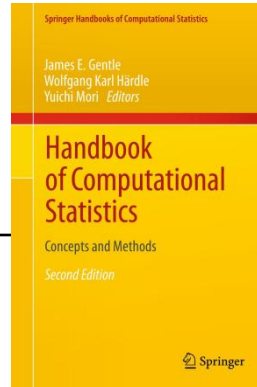
---

- Preparação de dados para análise;
- Manipulação de dados e arquivos;
- Geração de números aleatórios;
- Construção de tabelas estatísticas;
- Gráficos;
- Interface com editores de texto;
- Criação de *scripts*.

# Motivação

## □ Definições:

- Do *Handbook of Computational Statistics: Concepts and Methods*, por Gentle, J. E.; Härdle, W. K. e Mori, Y. (Eds.):
  - “*Computação estatística*” refere-se aos métodos computacionais auxiliares aos métodos estatísticos, tais como análise numérica, banco de dados, computação gráfica, engenharia de software e interface homem/máquina;
  - “*Estatística computacional*” tem significado mais amplo, por incluir não só os métodos de *computação estatística*, como também métodos estatísticos que são *computacionalmente intensivos*.



# Dois exemplos

---

- Em *Pacotes Estatísticos*, será aprofundado o estudo das ferramentas da **estatística computacional**, que podem ser úteis em situações de interesse prático em que:
  - os pressupostos para aplicação dos métodos clássicos **não** se aplicam ou
  - tais pressupostos são de **difícil** verificação.
- Exemplos:
  - 1 - testes de permutação;
  - 2 - *bootstrap*.

# Teste de Hipótese para Diferença de Médias

---

- As situações reais **diferenciam** a aplicação da teoria:
  - variâncias conhecidas *vs.* variâncias desconhecidas;
  - grandes amostras *vs.* pequenas amostras;
  - dados pareados *vs.* dados não pareados.
- As abordagens são **específicas** para cada situação.

# Teste de Hipótese para Diferença de Médias com Variância Conhecida

---

- Hipótese nula:

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

- Fazendo  $\Delta_0 = 0$ , estaremos testando a igualdade das duas médias  $\mu_1$  e  $\mu_2$ .
- A estatística de teste para este caso terá uma distribuição **normal padrão**,  $N(0,1)$ .
- Podemos ainda afirmar que encontrar uma diferença considerável entre os valores das médias  $\mu_1$  e  $\mu_2$  é uma evidência de que a hipótese alternativa,  $H_1$ , é verdadeira.

# Teste de Hipótese para Diferença de Médias com Variância Conhecida

- Estatística de teste:

$$Z_0 = \frac{X_1 - X_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- | □ Hipóteses Alternativas             | Critério de Rejeição                          |
|--------------------------------------|---|
| ■ $H_1: \mu_1 - \mu_2 \neq \Delta_0$ | $z_0 > z_{\alpha/2}$ ou $z_0 < -z_{\alpha/2}$ |
| ■ $H_1: \mu_1 - \mu_2 > \Delta_0$    | $z_0 > z_{\alpha}$                            |
| ■ $H_1: \mu_1 - \mu_2 < \Delta_0$    | $z_0 < -z_{\alpha}$                           |



# Teste de Hipótese para Diferença de Médias **via Testes de Permutação**

---

- Teste de permutação:
  - consiste em simples **permutações** dos dados originais;
  - comparamos então o valor da estatística observada nos dados **originais**, com as estatísticas obtidas destas amostras **permutadas**;
  - se estas saídas produzem estatísticas com valores **semelhantes** à estatística observada, esta terá sido um mero **acaso** e os dados são provenientes da **mesma** população.

# Algoritmo do Teste de Permutação

---

## algoritmo

**leia** amostra1, amostra2

misture as amostras em amostra3 (simulação da hipótese nula)

## repetir

distribuir amostra3 em amostra4 e amostra5

calcular a diferença das médias entre amostra4 e amostra5

acumular diferença

atualizar o valor- $p$  por comparação com a diferença observada

**até** número de replicações ser alcançado

escrever resultados

## fim algoritmo

# Resultados Experimentais: Dados Reais

---

- Utilizamos os dados reais apresentados em Santos (1998).
- Através de 18 amostras retiradas de camadas superficiais dos solos de dois locais diferentes (*Central Soil Salinity Research Institute*), foram coletados dados sobre seus devidos valores de pH.
- O processo utilizado foi a retirada de 9 amostras de cada uma das duas regiões pesquisadas.
- O objetivo é tentar provar se realmente as amostras podem ser classificadas em dois grupos distintos ou não.

# Resultados Experimentais: Dados Reais (continuação)

---

□ Localidade A:

8,53 8,52 8,01 7,99 7,93

7,89 7,85 7,82 7,80

□ Localidade B:

7,85 7,73 7,58 7,40 7,35

7,30 7,27 7,27 7,23

# Resultados Experimentais: Dados Reais (continuação)

---

- Chamada à macro e saída do teste de permutação:

```
MTB > %permuta c1 c2;  
SUBC> hist.  
Executing from file: permuta.MAC
```

## **Data Display**

Estatística original

k6 -0,595556

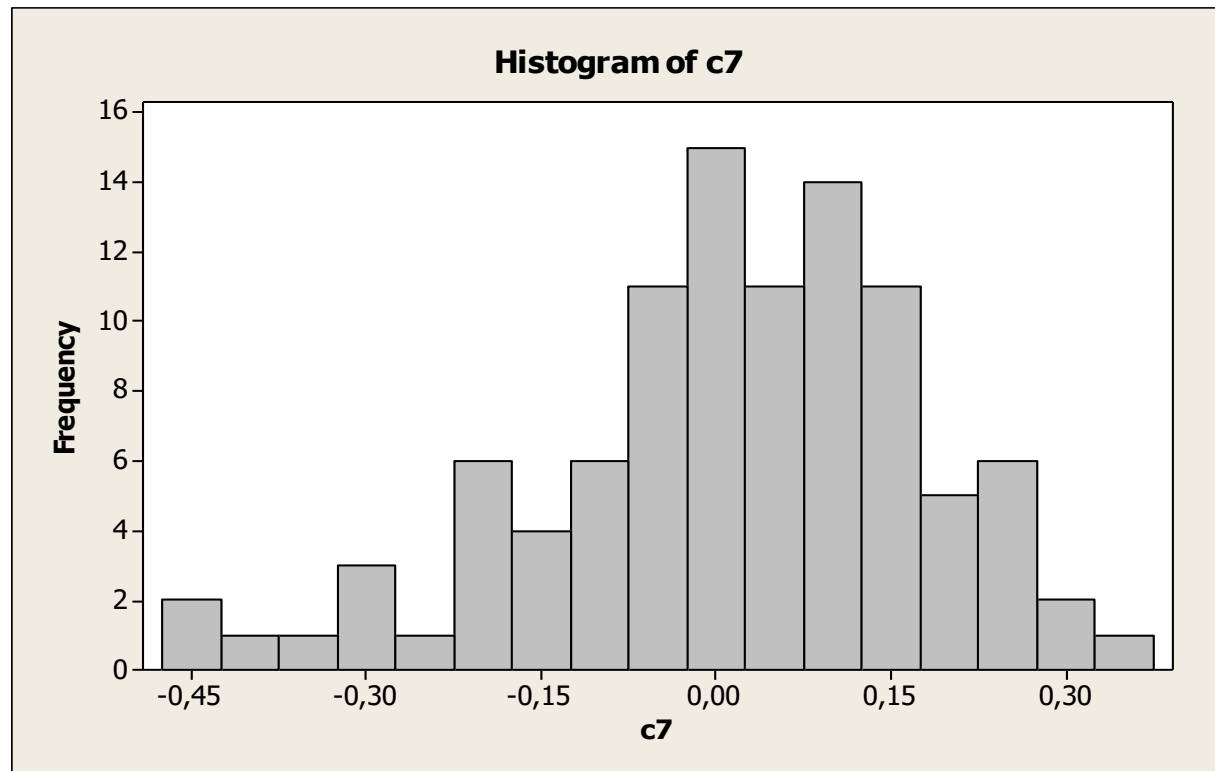
## **Data Display**

Valor-p

k7 0

# Resultados Experimentais: Dados Reais (continuação)

- Histograma gerado pela macro:



# Método **Clássico** para Construção de ICs

---

## □ ICs para a média amostral:

- Precisamos encontrar  $C_L$  e  $C_U$  tal que:

$$P[C_L \leq g(X_1, X_2, \dots, X_n; \theta) \leq C_U] = 1 - \alpha$$

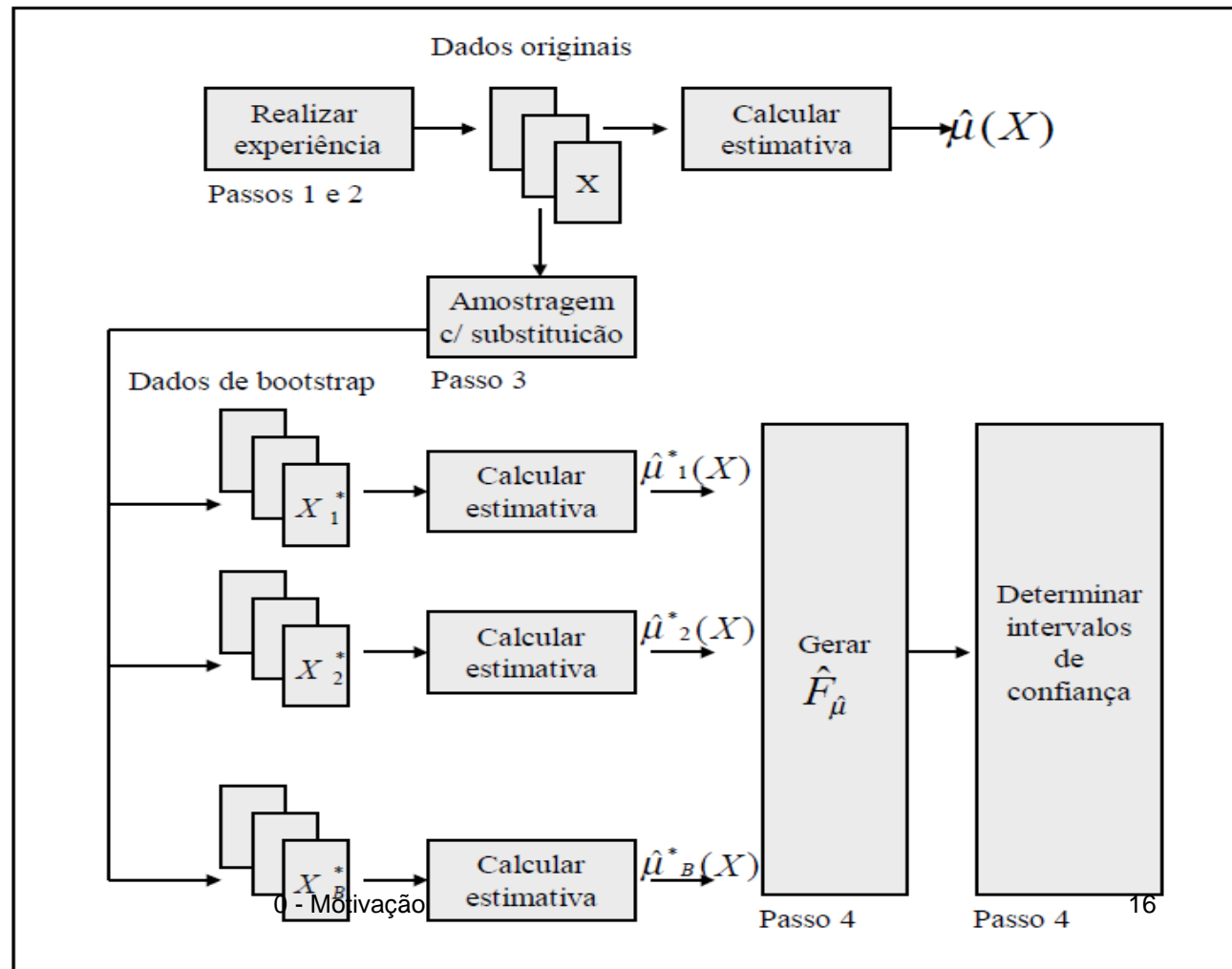
- Encontramos então:

$$L(X_1, X_2, \dots, X_n) = \bar{X} - Z_{\alpha/2} \sigma / \sqrt{n}$$

$$U(X_1, X_2, \dots, X_n) = \bar{X} + Z_{\alpha/2} \sigma / \sqrt{n}$$

# Método *Bootstrap* para Construção de ICs

## □ O método:





# Método *Bootstrap* para Construção de Ics (cont.)

---

- As seguinte estimativas são obtidas:

$$\hat{\mu}_1^*, \hat{\mu}_2^*, \dots, \hat{\mu}_B^*$$

- Elas são colocadas em ordem:

$$\hat{\mu}_{(1)}^* \leq \hat{\mu}_{(2)}^* \leq \dots \leq \hat{\mu}_{(B)}^*$$

- O IC de  $(1-\alpha)100\%$  é dado por:

$$\left( \hat{\mu}_{(q_1)}^*, \hat{\mu}_{(q_2)}^* \right)$$

em que  $q_1 = \text{parte inteira} \left( B \times \frac{\alpha}{2} \right)$  e  $q_2 = B - q_1 + 1$  .

# Resultados Experimentais

---

## □ Chamadas à macro:

#chamada para o valor *default*

MTB> %icboot c1 #primeira opção de chamada, para 95% de confiança

...

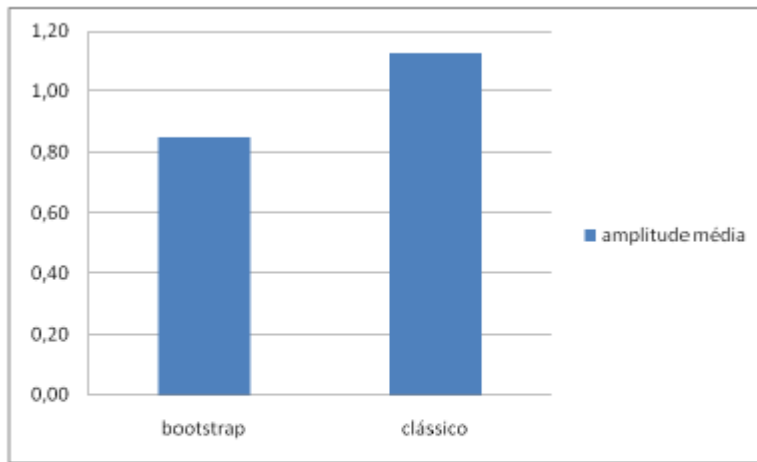
MTB> %icboot c2;

SUBC> conf 90. #segunda opção de chamada, para 90% de confiança

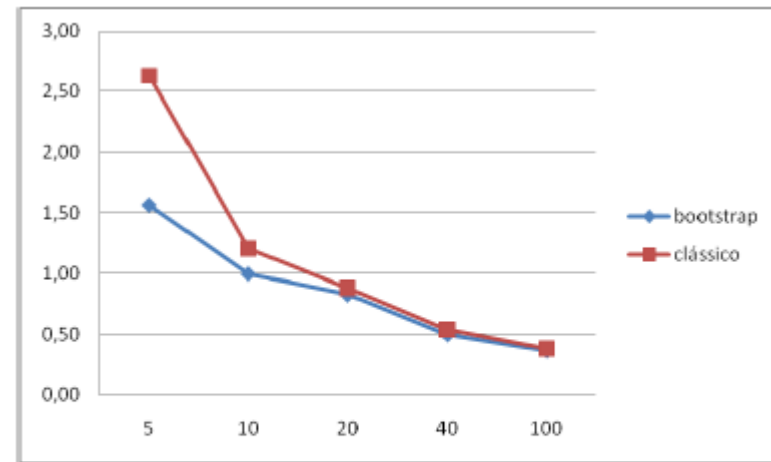
...

# Resultados Experimentais (cont.)

## □ Resultados simulados:



a) comparação em função do método



b) comparação em função do tamanho da amostra

**Figura 5.3:** Amplitudes médias dos intervalos de confiança

# Resultados Experimentais: Dados Reais

---

- Medidas de porcentagem de não enriquecimento de 12 bastões:

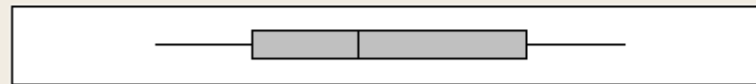
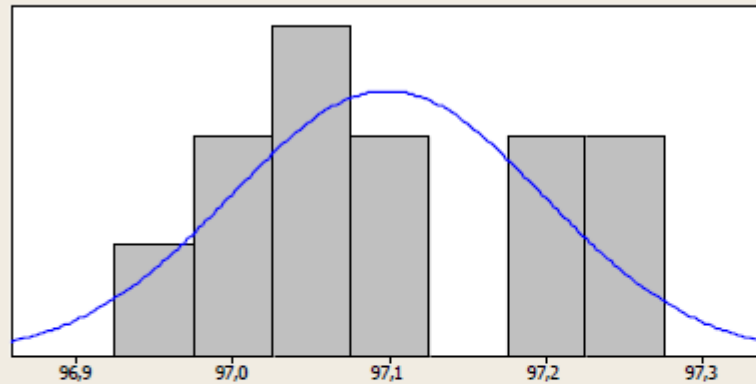
97,06; 97,25; 97,25; 97,19; 97,10; 97,10;  
97,18; 97,05; 97,00; 97,05; 97,00; 96,95

- **Observação:**

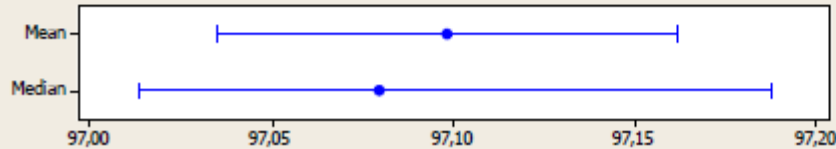
Caso de pequenas amostras e distribuição provavelmente não-normal.

# Resultados Experimentais: Dados Reais (cont.)

Sumário para Porcentagem de não enriquecimento



95% Confidence Intervals



Anderson-Darling Normality Test

A-Squared	0,34
P-Value	0,428

Mean	97,098
StDev	0,099
Variance	0,010
Skewness	0,31652
Kurtosis	-1,02851
N	12

Minimum	96,950
1st Quartile	97,013
Median	97,080
3rd Quartile	97,188
Maximum	97,250

95% Confidence Interval for Mean

97,035	97,161
--------	--------

95% Confidence Interval for Median

97,013	97,187
--------	--------

95% Confidence Interval for StDev

0,070	0,169
-------	-------

$$IC_1(\mu; 95\%) = (97,1; 97,2)$$

$$IC_2(\mu; 95\%) = (97,0; 97,2).$$

# Bibliografia

---

- Duczmal, L. H.; Bessegato, L. F.; Santos, M. A. C. e Ferreira, S. J. Santos, M. A. C. (2003). *Introducao as Pilhas e Filas e Teste de Permutacao*. RTE-03/2003, EST-ICEx-UFMG, Belo Horizonte, MG.
- Edgington, E. S. (1995). *Randomization tests*. Marcel Dekker, Inc., 3a Ed., New York, NY.
- Montgomery, D. C. & Runger, G. C. (2003). *Estatística Aplicada e Probabilidade para Engenheiros*. 2ª ed., LTC Livros Técnicos e Científicos Editora S.A., Rio de Janeiro, RJ.
- Santos, M. A. C. (1998). *Noções de Estatística Computacional*. RTE-03/1998, EST-ICEx-UFMG, Belo Horizonte, MG.
- Triola, M. F. (2005). *Introdução à Estatística*. 9ª ed., LTC Livros Técnicos e Científicos Editora S.A., Rio de Janeiro, RJ.