Prof. Eduardo Gontijo Carrano - DEE/EE/UFMG

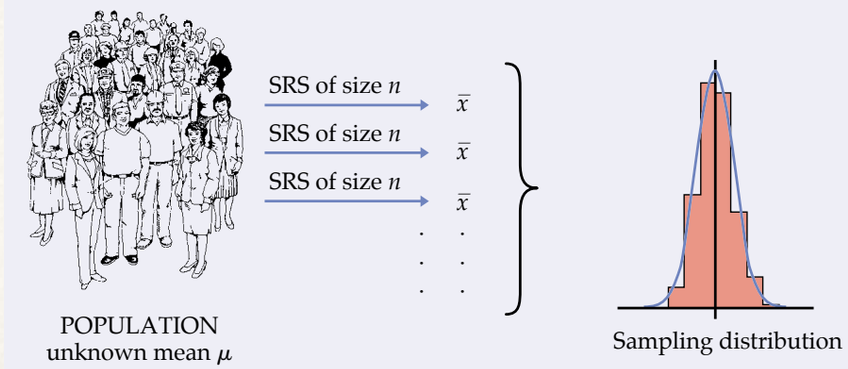
Confiabilidade de Sistemas

Bootstrapping

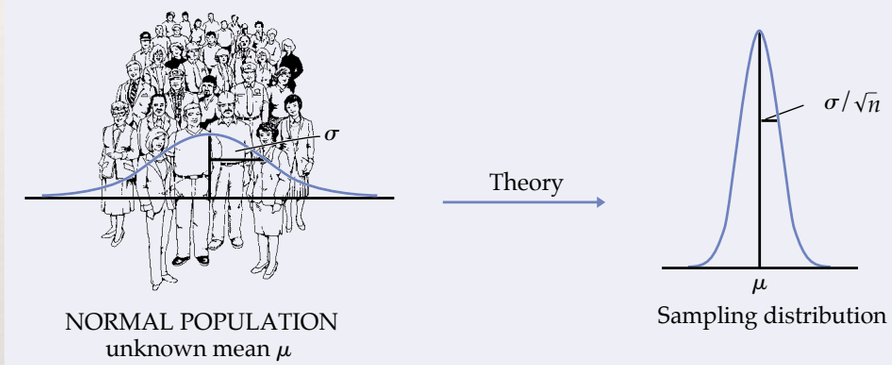
Introdução

- Bootstrapping: método de reamostragem comumente utilizado para estimar as propriedades de estimadores amostrais.
- Aplicações: estimativa de intervalos de confiança, polarização, variância, correlação e regressão.

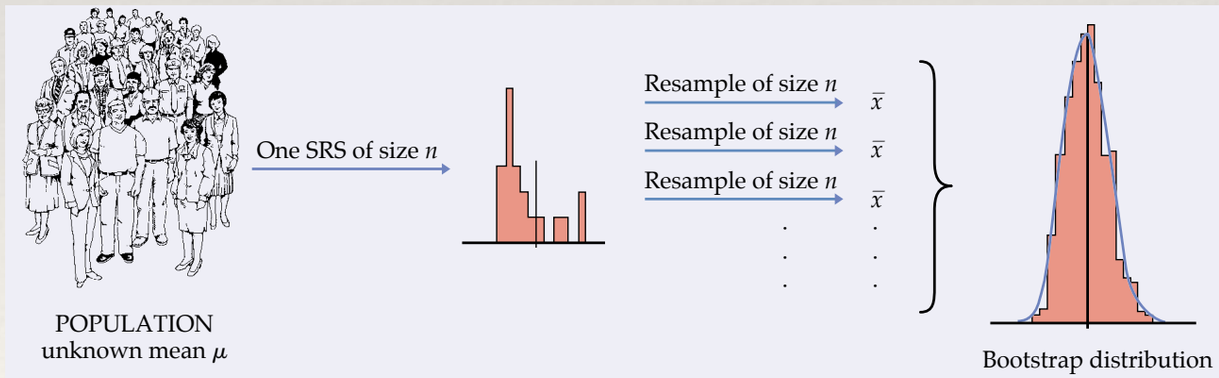
- Motivação:
  - média, variância (desvio padrão), mediana, percentil, etc.
- Amostra → População.
- O valor observado varia a cada amostra.
- Estimar a confiança da medida.


SRS of size n
SRS of size n
SRS of size n
$\bar{x}$
$\bar{x}$
$\bar{x}$
POPULATION
unknown mean μ
Sampling distribution

(a)


σ
NORMAL POPULATION
unknown mean μ
Theory
$\sigma/\sqrt{n}$
μ
Sampling distribution

(b)


One SRS of size n
Resample of size n
Resample of size n
Resample of size n
$\bar{x}$
$\bar{x}$
$\bar{x}$
POPULATION
unknown mean μ
Bootstrap distribution

(c)

Procedimento

Dados:

- população: x;
- n amostras iid: $X_i\ ,\ i = 1, \ldots, n$;
- o estimador de interesse: $\Theta = T(x)$;
- uma estatística definida para a amostra: $\hat{\Theta} = T(X)$.

- Para cada k de 1 até N:
  - extraia uma amostra de tamanho n de X, com repetição, para encontrar X_k^*;
  - encontre $\hat{\Theta}_k^* = T(X_k^*)$;
- Encontre $\hat{F}_b(x)$ a patir de $\hat{\Theta}^*$.

- Definição de N:
  - Método exato: todas as combinações possíveis das n amostras com repetição.

$$\begin{pmatrix} 2n-1 \\ n \end{pmatrix} \text{ combinações}$$

n	N
5	1,26E+02
10	9,24E+04
20	6,89E+10
30	5,913E+16
40	5,38E+22
50	5,04E+28

- Definição de N:
- Simulação de Monte Carlo:
  - combinações aleatórias.

- Forma: a forma da distribuição obtida no bootstrapping se aproxima da forma da distribuição da estatística considerada.
- Tendência central: a estimativa tende a ser polarizada caso a distribuição da amostra não esteja centrada no valor real do parâmetro.
- Dispersão: o erro padrão da estatística é o desvio padrão da distribuição do bootstrapping.

- O bootstrapping não tem a capacidade de gerar dados! O método estima como a grandeza amostral varia tendo em conta as n amostras disponíveis.
- Similar a abordagem média / erro padrão.
- Não depende de normalidade ou do TLC.

Premissas

- Independência das amostras;
- Representatividade da amostra utilizada.

Principais Vantagens

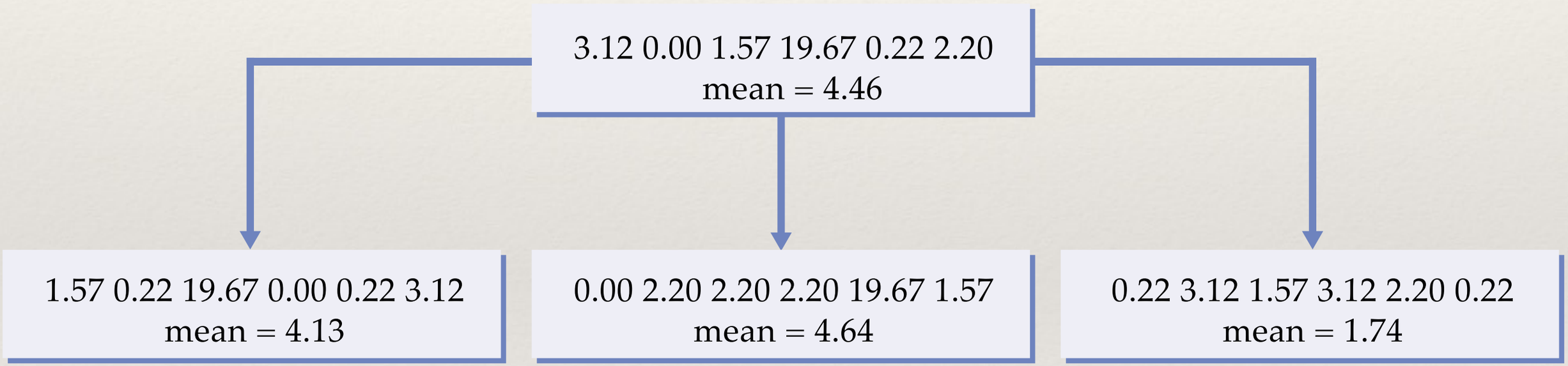
- Simplicidade;
- Aplicação à casos complexos;
- Independência de distribuição.

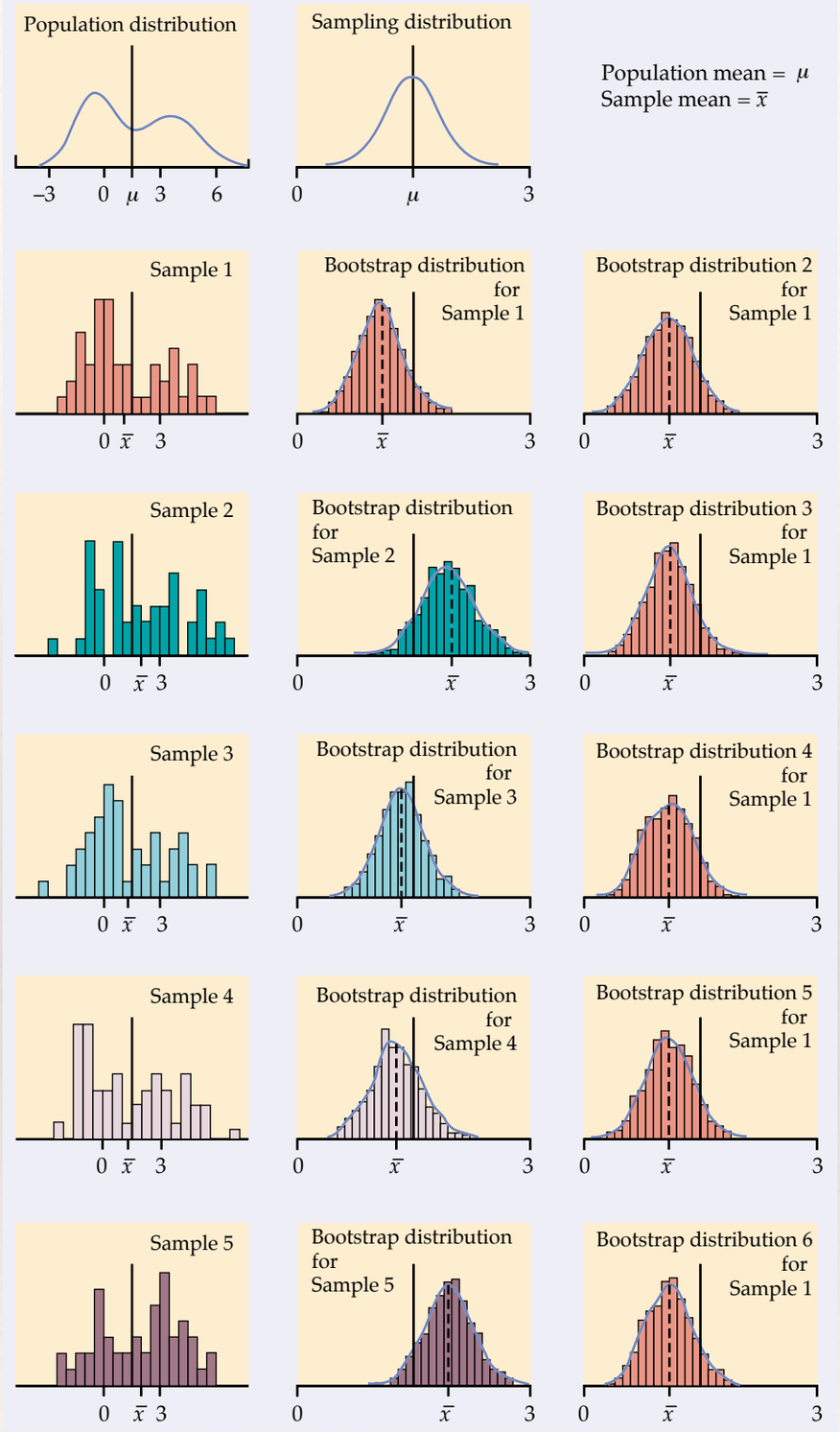
Principais Desvantagens

- Alta demanda computacional;
- Estimativa geralmente otimista;
- Duas fontes de imprecisão:
  - Amostra;
  - Conjunto de iterações do bootstrapping.

- Onde o Bootstrapping é particularmente útil?
  - Distribuição teórica da estatística de interesse é complexa ou desconhecida.
  - O tamanho da amostra é insuficiente para inferência direta.
  - Se faz necessário o cálculo de potência e uma pequena amostra está disponível.

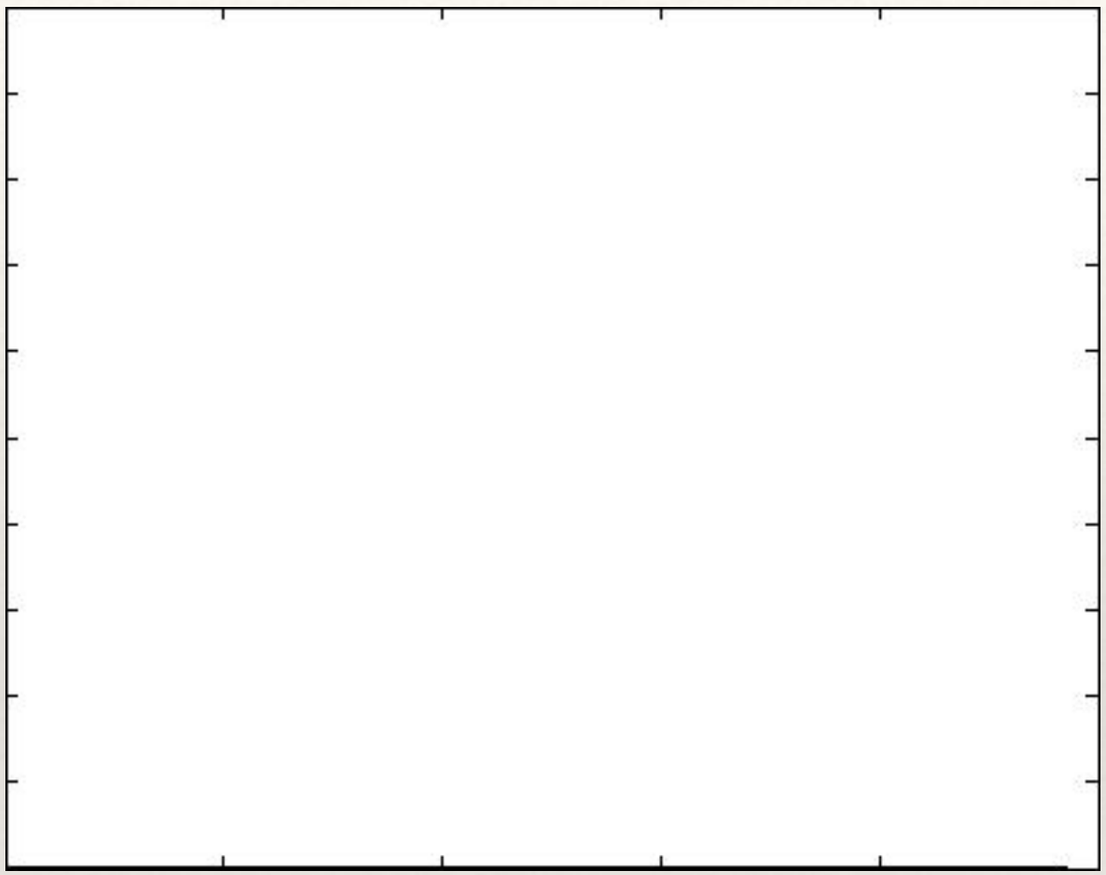
Exemplos


3.12 0.00 1.57 19.67 0.22 2.20
mean = 4.46
1.57 0.22 19.67 0.00 0.22 3.12
mean = 4.13
0.00 2.20 2.20 2.20 19.67 1.57
mean = 4.64
0.22 3.12 1.57 3.12 2.20 0.22
mean = 1.74

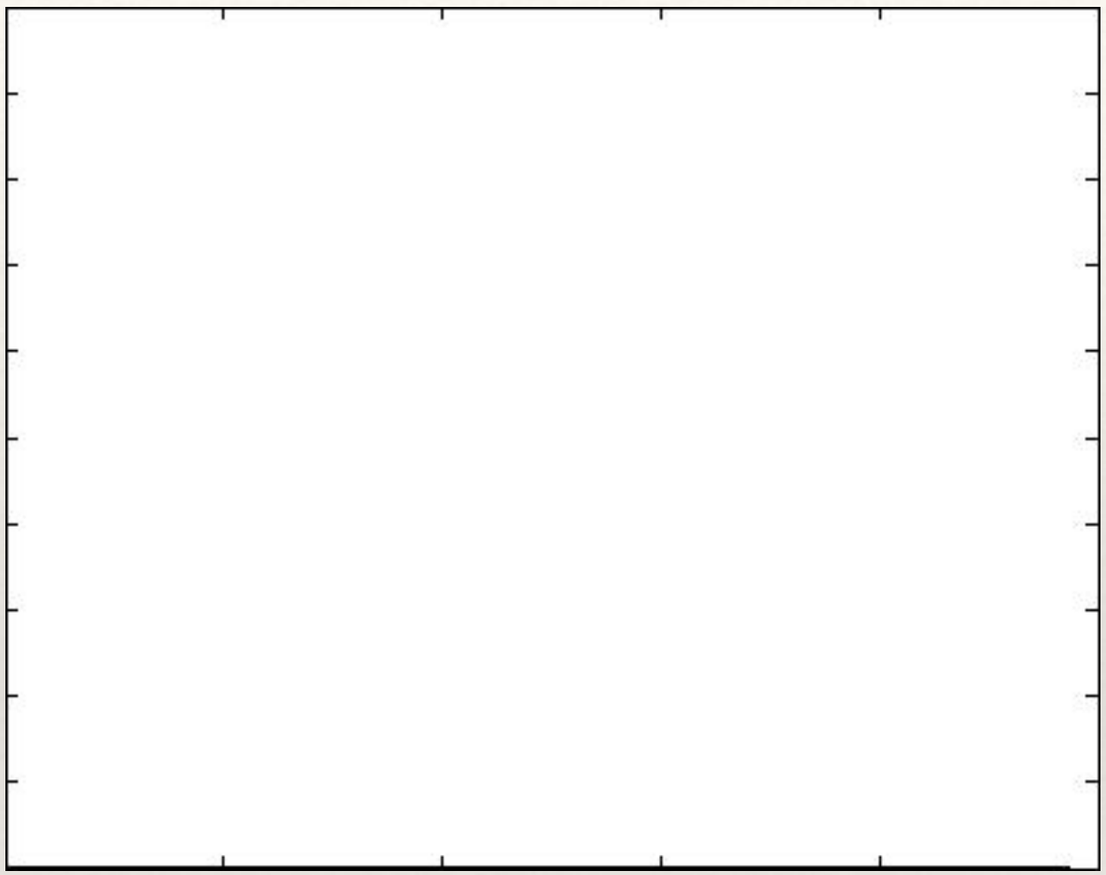

Population distribution
–3 0 μ 3 6
Sampling distribution
0 μ 3
Population mean = μ
Sample mean = x̄
Sample 1
0 x̄ 3
Bootstrap distribution for Sample 1
0 x̄ 3
Bootstrap distribution 2 for Sample 1
0 x̄ 3
Sample 2
0 x̄ 3
Bootstrap distribution for Sample 2
0 x̄ 3
Bootstrap distribution 3 for Sample 1
0 x̄ 3
Sample 3
0 x̄ 3
Bootstrap distribution for Sample 3
0 x̄ 3
Bootstrap distribution 4 for Sample 1
0 x̄ 3
Sample 4
0 x̄ 3
Bootstrap distribution for Sample 4
0 x̄ 3
Bootstrap distribution 5 for Sample 1
0 x̄ 3
Sample 5
0 x̄ 3
Bootstrap distribution for Sample 5
0 x̄ 3
Bootstrap distribution 6 for Sample 1
0 x̄ 3

- Distribuição uniforme [0.00;1.00];
- 1.000 iterações no bootstrapping.

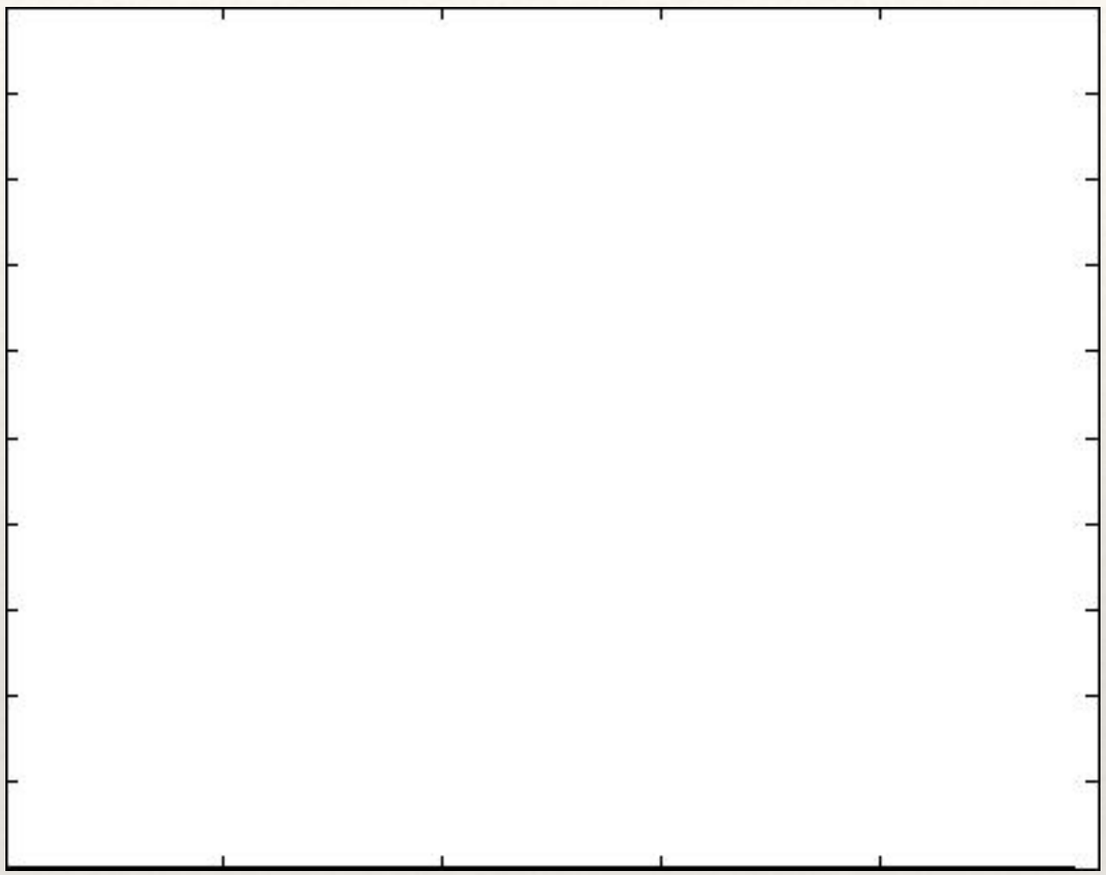
- 10 amostras:



- 30 amostras:



- 100 amostras:



Intervalos de Confiança

- Percentil:

Intervalos de confiança de $(1-\alpha)\cdot 100\%$:

- $\left[\hat{F}_b^{-1}(\alpha/2)\ ;\ \hat{F}_b^{-1}(1-\alpha/2)\right]$
- $\left[-\infty\ ;\ \hat{F}_b^{-1}(1-\alpha)\right]$
- $\left[\hat{F}_b^{-1}(\alpha)\ ;\ +\infty\right]$

$\hat{F}_b(\alpha)$: percentil α da distribuição obtida por bootstrapping.

- T-Student:

Intervalos de confiança de $(1-\alpha)\cdot 100\%$:

- $\left[\overline{x_b} - t_{\alpha/2,n-1}\cdot s_b \;;\; \overline{x_b} + t_{\alpha/2,n-1}\cdot s_b\right]$
- $\left[-\infty \;;\; \overline{x_b} + t_{\alpha,n-1}\cdot s_b\right]$
- $\left[\overline{x_b} - t_{\alpha,n-1}\cdot s_b \;;\; +\infty\right]$

$t_{\alpha,n-1}$: percentil α da distribuição T com $n-1$ graus de liberdade.
$\overline{x_b}$: média da distribuição obtida por bootstrapping.
$\overline{s_b}$: desvio padrão da distribuição obtida por bootstrapping.

- Bias-Corrected Accelerated (BCa):

$$Q(\alpha) = \hat{F}_b^{-1}\left\{\Phi\left[z_0 + \frac{z_0+z^\alpha}{1-a(z_0+z^\alpha)}\right]\right\}$$

Intervalos de confiança de $(1-\alpha)\cdot 100\%$:

- $[Q(\alpha/2)\ ;\ Q(1-\alpha/2)]$
- $[-\infty\ ;\ Q(1-\alpha)]$
- $[Q(\alpha)\ ;\ +\infty]$

Φ: CDF da normal padrão.
$z_0 = \Phi^{-1}\left[\hat{F}_b\left(\hat{\theta}\right)\right]$.
a: constante de aceleração.

Testes de Hipóteses

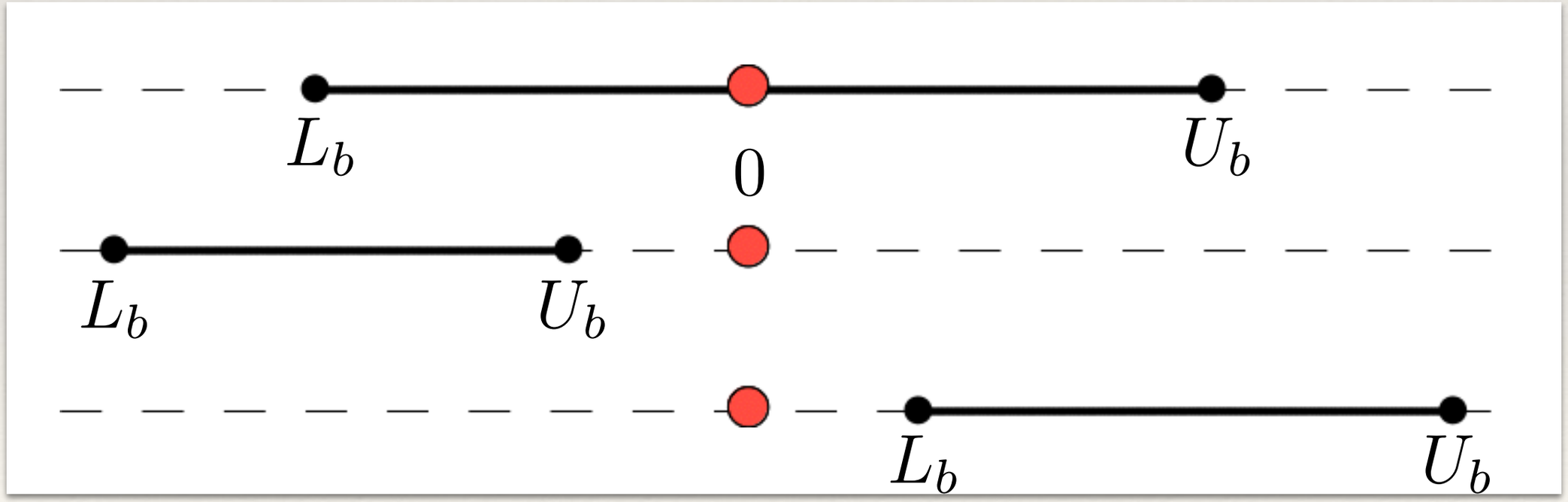
- Uma amostra:

$$\begin{cases} H_0 & : & \mu = \mu_0 \\ H_1 & : & \mu \neq \mu_0 \end{cases}$$

- Duas amostras:

$$\begin{cases} H_0 & : & \mu_1 = \mu_2 \\ H_1 & : & \mu_1 \neq \mu_2 \end{cases}$$

$$\begin{cases} H_0 & : & \mu_1 - \mu_2 = 0 \\ H_1 & : & \mu_1 - \mu_2 \neq 0 \end{cases}$$



- Várias amostras:

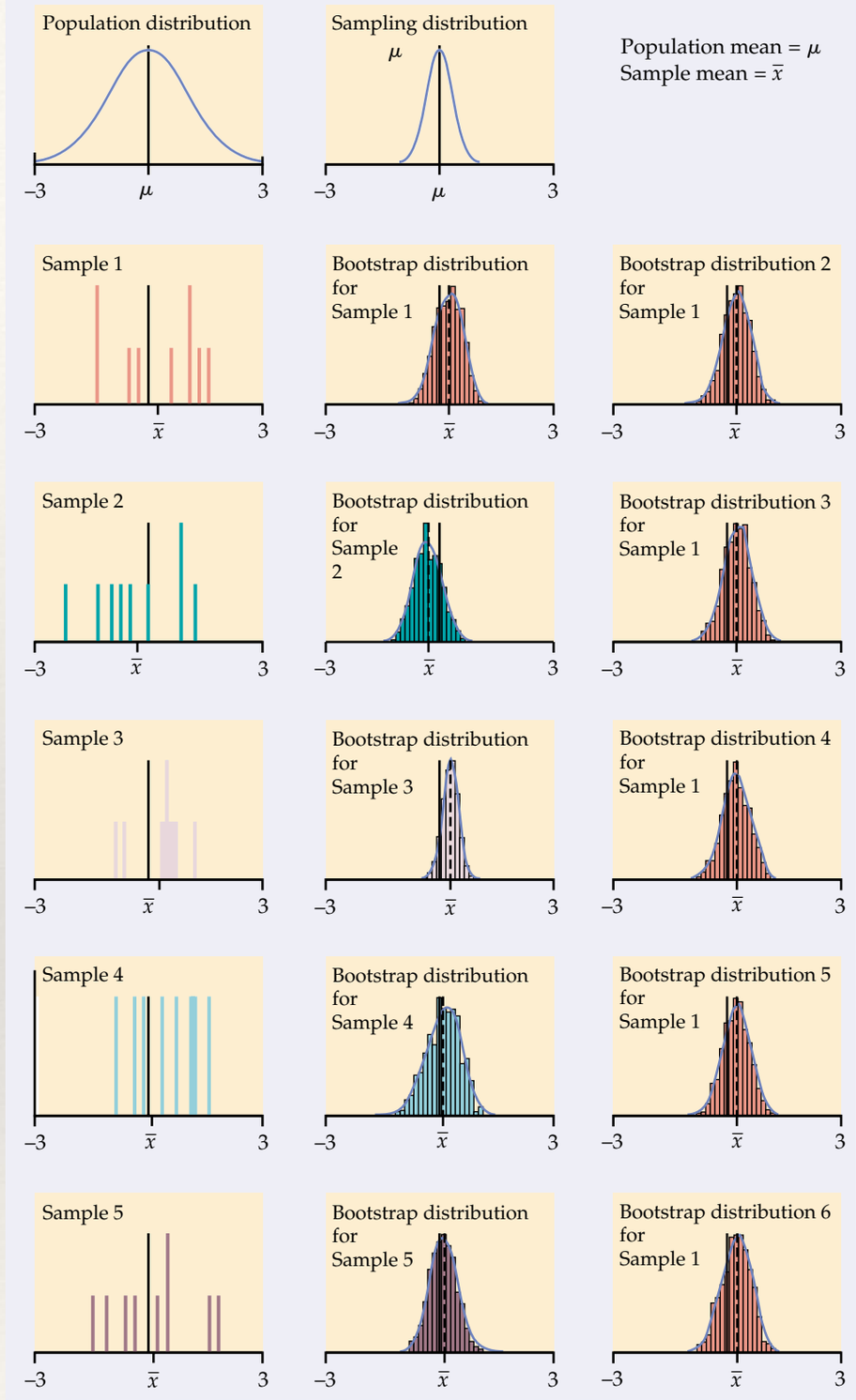
$$\begin{cases} H_0 & : & \mu_i = \mu_j & \forall i, j \in 1, \ldots, n \\ H_1 & : & \exists i, j \in 1, \ldots, n & \mid & \mu_i \neq \mu_j \end{cases}$$

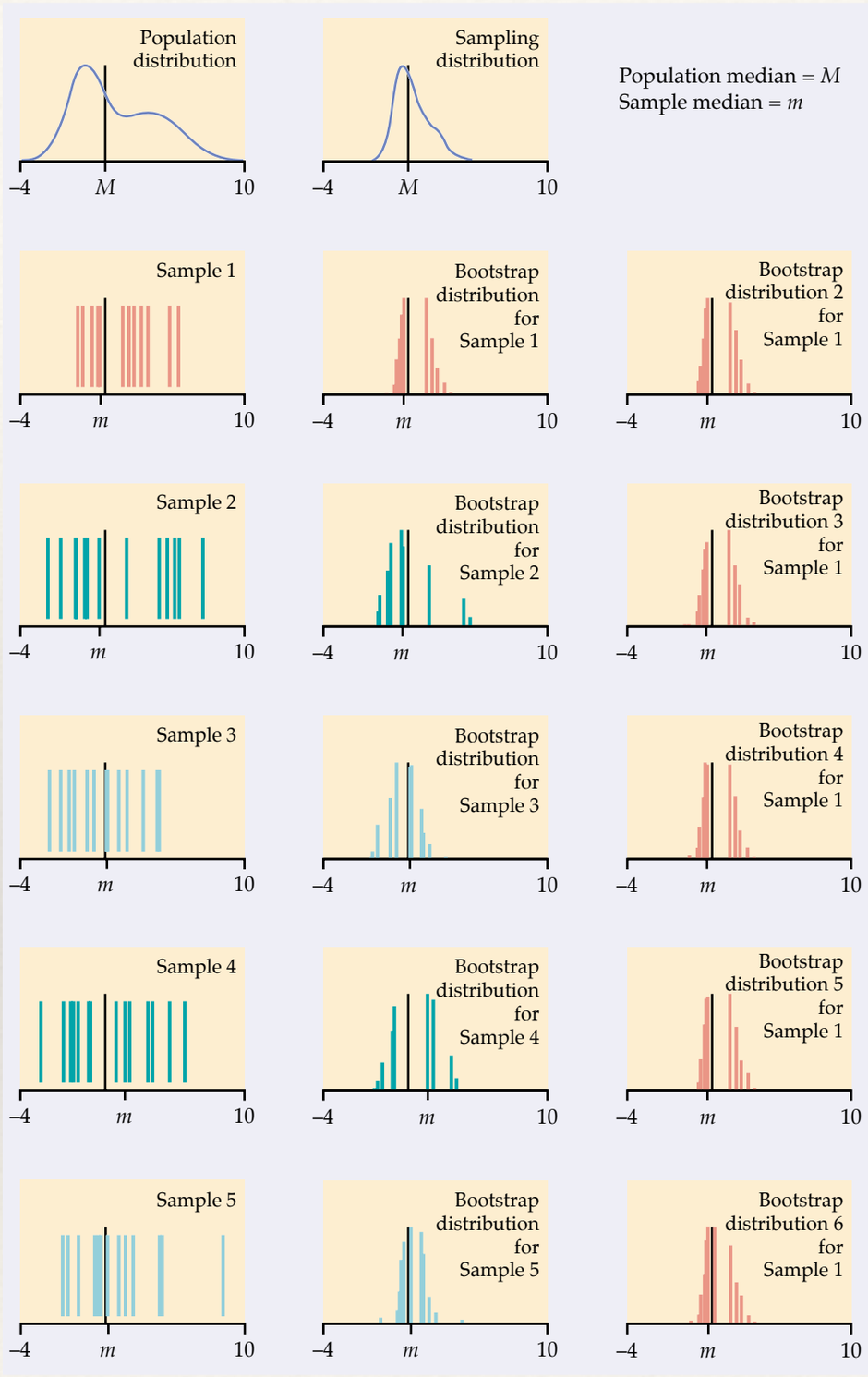
$$\begin{cases} H_0^1 & : & \mu_1 = \mu_2 \\ H_1^1 & : & \mu_1 \neq \mu_2 \end{cases} \quad \begin{cases} H_0^2 & : & \mu_1 = \mu_3 \\ H_1^2 & : & \mu_1 \neq \mu_3 \end{cases} \quad \cdots \quad \begin{cases} H_0^m & : & \mu_{n-1} = \mu_n \\ H_1^m & : & \mu_{n-1} \neq \mu_n \end{cases}$$

- Várias amostras:
  - Comparações múltiplas:
    - Ajuste da significância para construção dos intervalos.
    - Métodos de correção: Bonferroni, Holm-Bonferroni, Sidák, Dunnett, Tukey-Kramer, Nemenyi, Bonferroni-Dunn, Scheffe, etc.

Casos Críticos

- Amostras muito pequenas;
- Estimativa das propriedades de estimadores discretos: mediana, quantil, percentil, etc.

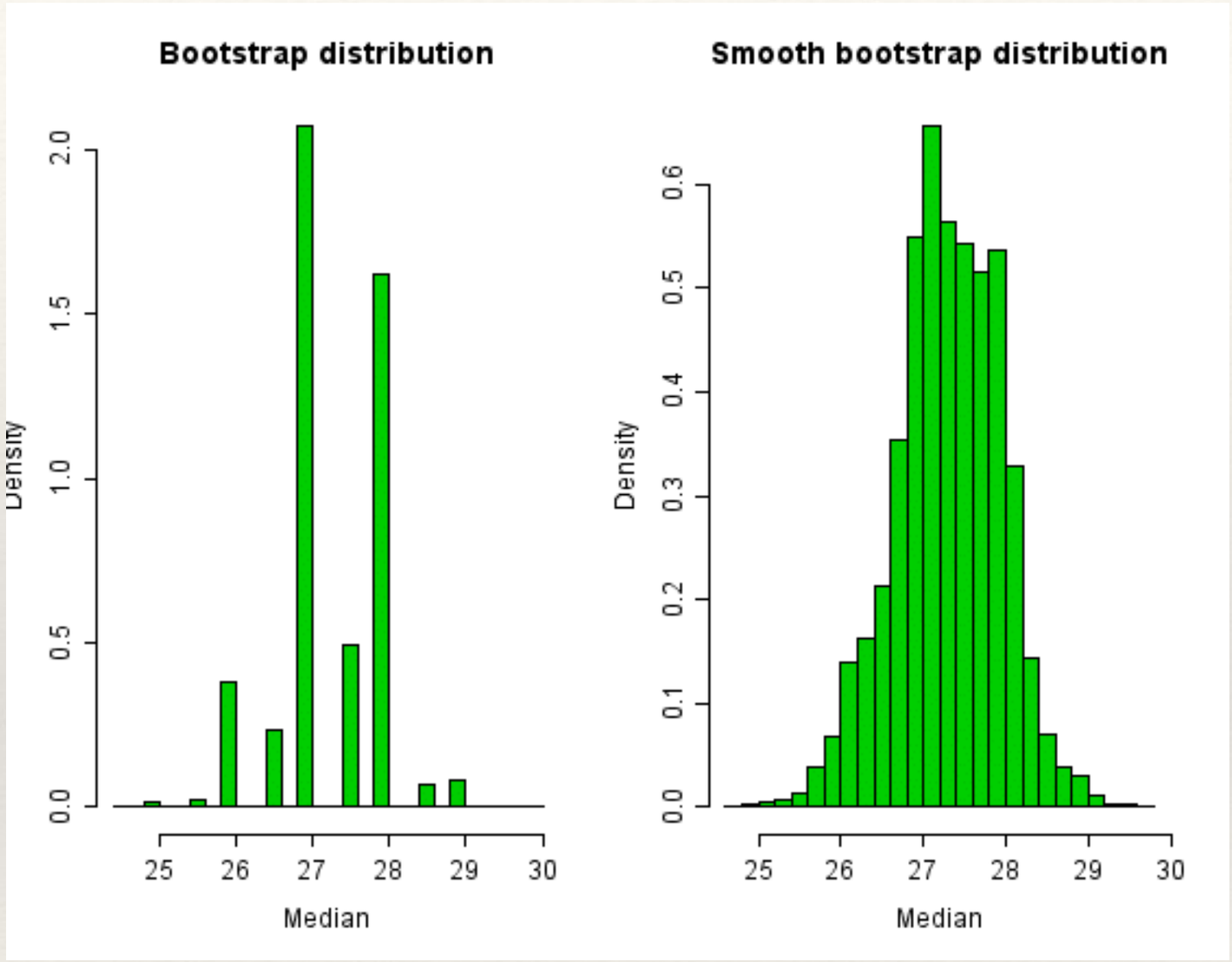

Population distribution
–3
μ
3
Sampling distribution
μ
–3
μ
3
Population mean = μ
Sample mean = x̄
Sample 1
–3
x̄
3
Bootstrap distribution for Sample 1
–3
x̄
3
Bootstrap distribution 2 for Sample 1
–3
x̄
3
Sample 2
–3
x̄
3
Bootstrap distribution for Sample 2
–3
x̄
3
Bootstrap distribution 3 for Sample 1
–3
x̄
3
Sample 3
–3
x̄
3
Bootstrap distribution for Sample 3
–3
x̄
3
Bootstrap distribution 4 for Sample 1
–3
x̄
3
Sample 4
–3
x̄
3
Bootstrap distribution for Sample 4
–3
x̄
3
Bootstrap distribution 5 for Sample 1
–3
x̄
3
Sample 5
–3
x̄
3
Bootstrap distribution for Sample 5
–3
x̄
3
Bootstrap distribution 6 for Sample 1
–3
x̄
3


Population distribution
–4
M
10
Sampling distribution
–4
M
10
Population median = M
Sample median = m
Sample 1
–4
m
10
Bootstrap distribution for Sample 1
–4
m
10
Bootstrap distribution 2 for Sample 1
–4
m
10
Sample 2
–4
m
10
Bootstrap distribution for Sample 2
–4
m
10
Bootstrap distribution 3 for Sample 1
–4
m
10
Sample 3
–4
m
10
Bootstrap distribution for Sample 3
–4
m
10
Bootstrap distribution 4 for Sample 1
–4
m
10
Sample 4
–4
m
10
Bootstrap distribution for Sample 4
–4
m
10
Bootstrap distribution 5 for Sample 1
–4
m
10
Sample 5
–4
m
10
Bootstrap distribution for Sample 5
–4
m
10
Bootstrap distribution 6 for Sample 1
–4
m
10

- Smoothed Bootstrapping
  - Adiciona-se um ruído gaussiano de baixa magnitude a cada observação reamostrada.

ruído: $N(0, \sigma^2)$

$$\sigma = \frac{1}{\sqrt{n}}$$


Bootstrap distribution
Smooth bootstrap distribution
Density
Median
0.0
0.5
1.0
1.5
2.0
0.1
0.2
0.3
0.4
0.6
25
26
27
28
29
30

Estudo de Caso - Comparação de Algoritmos

- Primeiras comparações: comparações mono-critério (convergência ou custo computacional) baseadas em valores médios.
- (Craenen et al, 2003), (Takahashi et al, 2003): comparações bi-critério (convergência e custo computacional) baseadas em valores médios.

- (Czarn et al, 2004), (Yuan et al, 2004): comparações mono-critério baseadas em testes paramétricos.
- (Shilane et al, 2006), (Garcia et al, 2008): comparações mono-critério baseadas em testes construídos por bootstrapping ou não paramétricos.
- (Carrano et al, 2007), (Carrano et al, 2008): comparações bi-critério baseadas em testes não paramétricos.

A Multicriteria Statistical Based Comparison Methodology for Evaluating Evolutionary Algorithms

Eduardo G. Carrano, Elizabeth F. Wanner, and Ricardo H. C. Takahashi

***Abstract*—This paper presents a statistical based comparison methodology for performing evolutionary algorithm comparison under multiple merit criteria. The analysis of each criterion is based on the progressive construction of a ranking of the algorithms under analysis, with the determination of significance levels for each ranking step. The multicriteria analysis is based on the aggregation of the different criteria rankings via a non-dominance analysis which indicates the algorithms which constitute the efficient set. In order to avoid correlation effects, a principal component analysis pre-processing is performed. Bootstrapping techniques allow the evaluation of merit criteria data with arbitrary probability distribution functions. The algorithm ranking in each criterion is built progressively, using either ANOVA or first order stochastic dominance. The resulting ranking is checked using a permutation test which detects possible inconsistencies in the ranking—leading to the execution of more algorithm runs which refine the ranking confidence. As a by-product, the permutation test also delivers *p*-values for the ordering between each two algorithms which have adjacent rank positions. A comparison of the proposed method with other methodologies has been performed using reference probability distribution functions (PDFs). The proposed methodology has always reached the correct ranking with less samples and, in the case of non-Gaussian PDFs, the proposed methodology has worked well, while the other methods have not been able even to detect some PDF differences. The application of the proposed method is illustrated in benchmark problems.**

***Index Terms*—Algorithm evaluation, evolutionary algorithms, multicriteria statistical comparison.**

ing with specific problems or classes of problems usually involves some kind of tradeoff between the computational effort associated with algorithm execution and the solution quality. In the case of deterministic algorithms, such a comparison is performed on the basis of algorithm results which are deterministic for each given problem instance. It is guaranteed under some assumptions that, starting from a given initial point, these algorithms always perform the same sequence of deterministic steps, and the algorithm converges (i.e., reaches a stop criterion) in a fixed number of algorithm iterations [1]. As a consequence, such algorithms are often evaluated on the basis of single-run results performed on sets of different problem instances.

The performance evaluation of non-deterministic algorithms, such as evolutionary algorithms, cannot be performed using such a kind of procedure. The stochastic nature of these methods introduces some random variability in the answer provided by the algorithm: the solution obtained by the algorithm can vary considerably from one run to another, and even when the same solution is reached, the computational time required for achieving such a solution is usually different for different runs of the same algorithm [1].

The flexible structure of the evolutionary algorithms makes it possible to build them in several different ways. Each operator variation inside an algorithm leads to a different algorithm version with its own associated performance. This

IEEE Transactions on Evolutionary Computation
Vol. 15
No. 6
pp. 848-870
December, 2011

- Compara K algoritmos evolucionários em um problema considerando C critérios de qualidade (fatores).
- Oferece como saída um ranking dos métodos e os p-valores associados a este ranking.
- Permite comparações de algoritmos “a posteriori” ou iterativamente.

- Repita:
  - Realize n execuções de cada um dos algoritmos.
  - Aplique PCA para descorrelacionar os critérios (*).
  - Aplique bootstrapping para construir as PDF.
  - Compare as PDF utilizando algum tipo de teste (T1).
  - Caso haja repetição do ranking obtido:
    - Compare as PDF utilizando permutações cíclicas (T2).
- Enquanto não houver repetição do ranking ou T1 e T2 levarem a conclusões distintas.

- Análise do método:
  - tamanho de amostra;
  - premissas;
  - robustez;
  - generalidade.

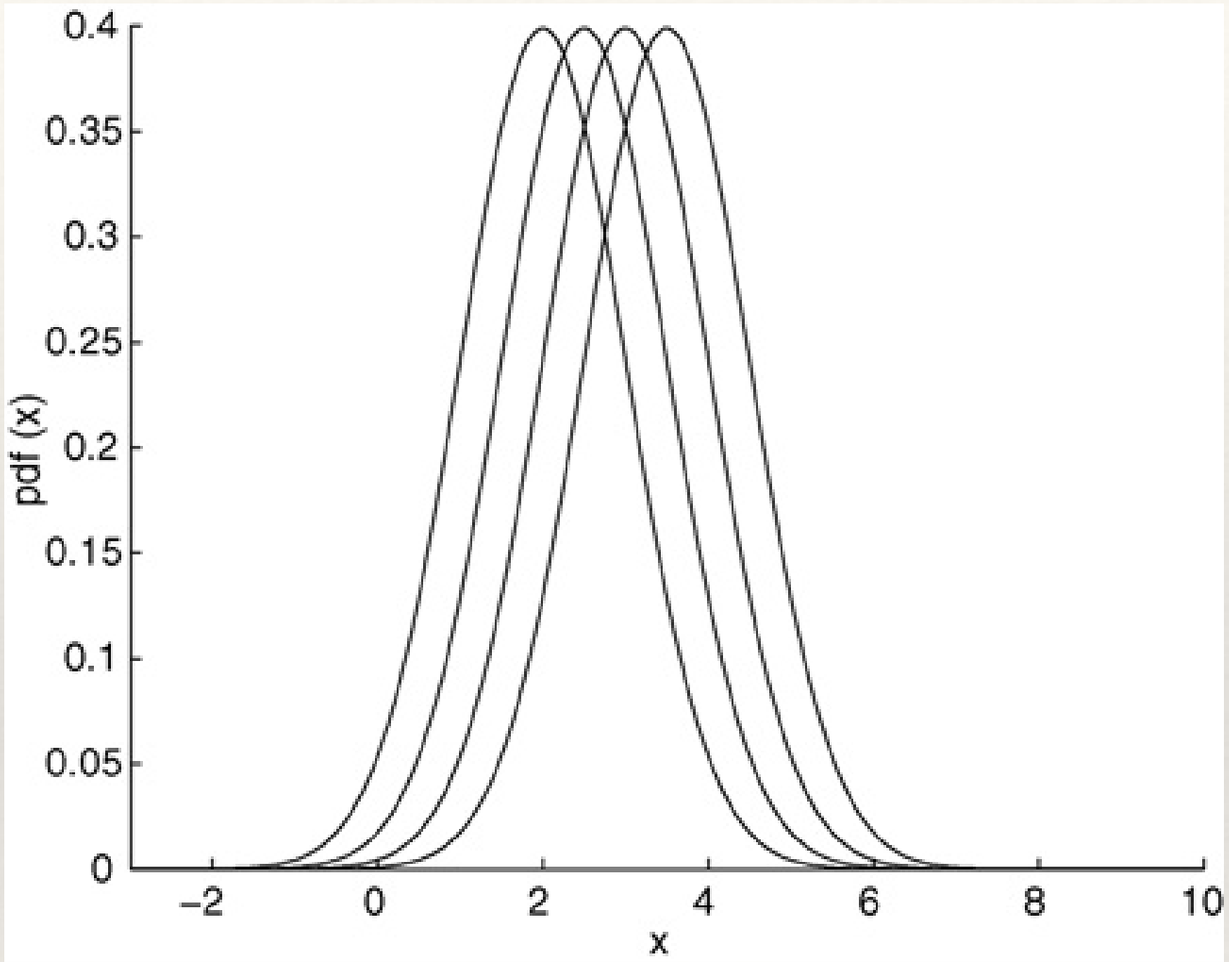
- Duas versões do método proposto:
  - PS-SD : T1 = Dominância Estocástica;
  - PS-AN : T1 = One-Way ANOVA.
- Comparações: “a posteriori” e iterativa.
- Métodos de referência:
  - One-Way ANOVA;
  - Kruskal-Wallis.

- Toy Problems:

reference problem		parameters				
		$\mathbf{A}_1$	$\mathbf{A}_2$	$\mathbf{A}_3$	$\mathbf{A}_4$	$\mathbf{A}_5$
Gaussian1	μ	3.00	2.50	3.50	2.00	2.50
	σ	1.00	1.00	1.00	1.00	1.00
Gaussian2	μ	3.00	2.50	3.50	2.00	2.50
	σ	2.00	1.50	3.00	1.00	1.50
Beta	α	0.60	0.40	0.75	0.50	0.40
	β	0.48	0.50	0.40	0.95	0.50
Binomial	p	0.60	0.50	0.70	0.40	0.50

$$\mathbf{A}_4 \quad \overline{\mathbf{A}_2 \quad \mathbf{A}_5} \quad \mathbf{A}_1 \quad \mathbf{A}_3$$

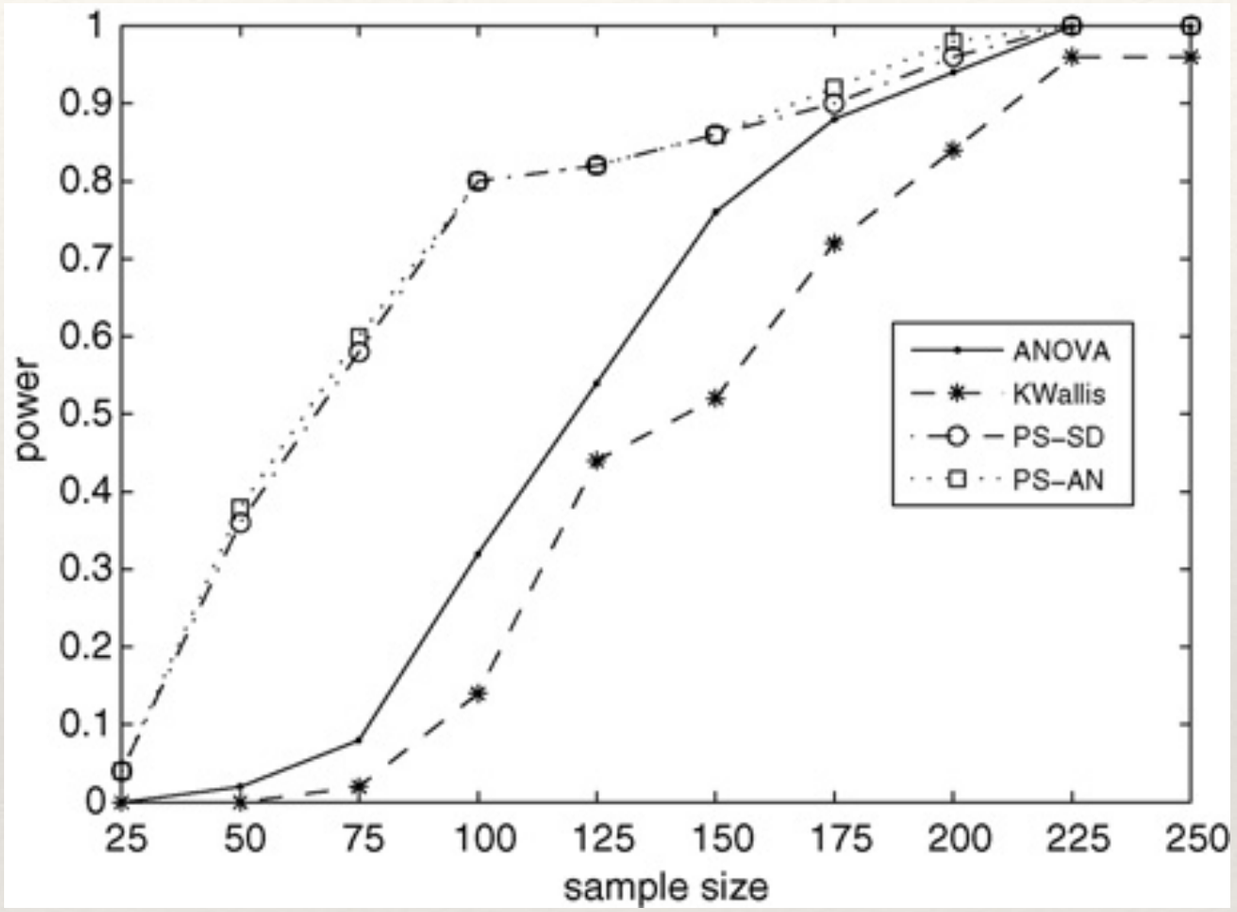
Gaussian1



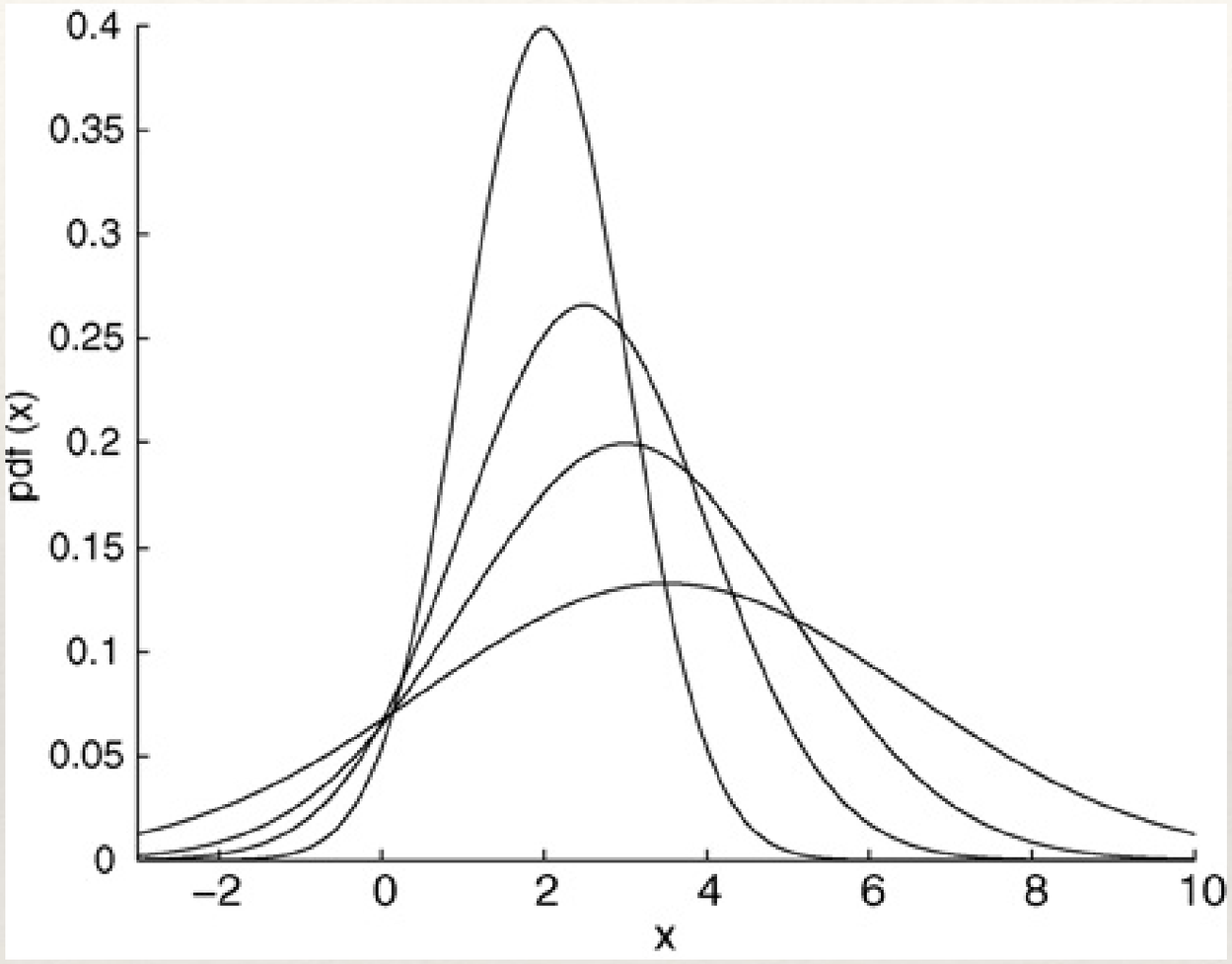
Gaussian1

comparison scheme	minimum sample	ranking repeatability
ANOVA	133	200, 300, 400 and 500
KWallis	175	200, 300, 400 and 500
PS-SD	68	100, 200, 300, 400 and 500
PS-AN	68	100, 200, 300, 400 and 500

Gaussian1


power
sample size
ANOVA
KWallis
PS–SD
PS–AN
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
25
50
75
100
125
150
175
200
225
250

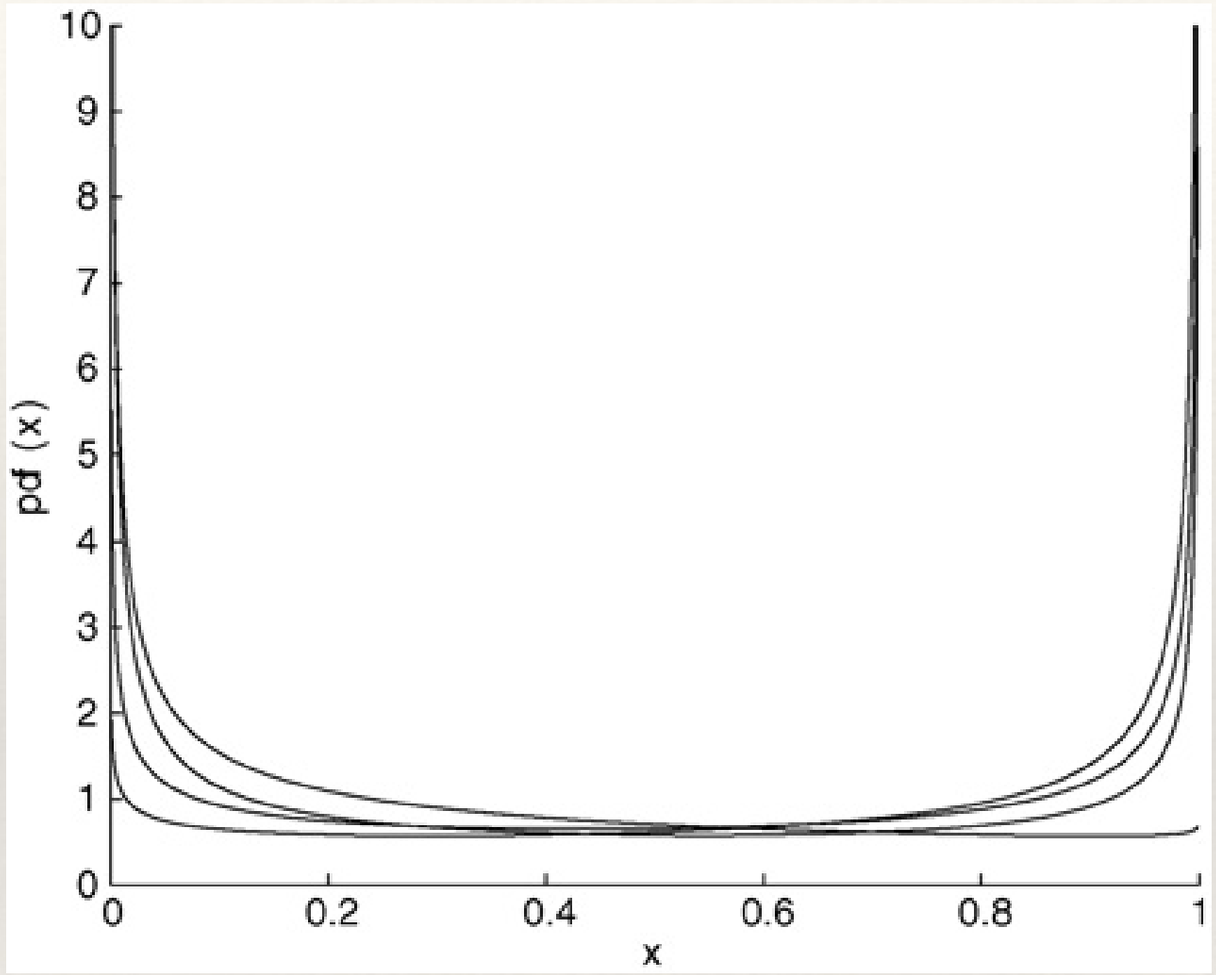
Gaussian2



Gaussian2

comparison scheme	minimum sample	ranking repeatability
ANOVA	342	400 and 500
KWallis	—	—
PS-SD	93	100, 200, 300, 400 and 500
PS-AN	88	100, 200, 300, 400 and 500

Beta



Beta

comparison scheme	minimum sample	ranking repeatability
ANOVA	—	—
KWallis	—	—
PS-SD	250	300, 400 and 500
PS-AN	250	300, 400 and 500

Binomial

comparison scheme	minimum sample	ranking repeatability
ANOVA	—	—
KWallis	—	—
PS-SD	220	300, 400 and 500
PS-AN	217	300, 400 and 500

- Problemas reais de otimização:
  - quatro problemas restritos;
  - três algoritmos evolucionários (GA, DE e ES);
  - três estratégias de tratamento de restrições;
  - três critérios de desempenho:
    - valor de função objetivo, número de avaliações necessárias para proporcionar 95% de melhoria e número de avaliações necessárias para alcançar a primeira solução factível.

prb.	alg.	c1	c2	c3
$g01$	1*	1	1	1
	2*	1	1	1
	3	1	2	2
	4	1	4	3
	5	1	3:4	3
	6	1	5	3
	7	3	4:5	4
	8	3	3:4	4
	9	2	3	4

prb.	alg.	c1	c2	c3
$g02$	1*	2	1	-
	2*	1	1:2	-
	3	1:2	2	-
	4	3	4	-
	5	3	4	-
	6	2	3	-
	7	5	6	-
	8	5	6	-
	9	4	5	-

prb.	alg.	c1	c2	c3
$g10$	1*	1	1	1
	2*	1	1	1
	3	2	2	2
	4	3	3	3
	5	3	3	4
	6	3	4	4
	7	4	5	4
	8	4	5	4
	9	4	4:5	4

prb.	alg.	c1	c2	c3
$g13$	1*	1	1	3
	2*	1	1	3
	3	1	2	4
	4	5	4	2
	5	3	4:5	2
	6	4	5	2
	7	2:3	3	1
	8*	2	3	1
	9*	2	3	1

Software


R


The MathWorks
MATLAB
The Language of Technical Computing
7

Referências

- DiCiccio, T. D. and Efron, B. (1996). Bootstrap confidence intervals. Statistical Science, v. 11, pp. 189-228.
- Davison, A. C. and Hinkley, D. (2006). Bootstrap Methods and their Application. Cambridge Series in Statistical and Probabilistic Mathematics, 8th ed..
- Moore, D. S., G. P. MacCabe and B. Craig (2012). Introduction to the Practice of Statistics. W. H. Freeman, 7th ed..

- B. G. W. Craenen, A. E. Eiben, and J. I. van Hemert, “Comparing evolutionary algorithms on binary constraint satisfaction problems,” IEEE Trans. Evol. Comput., vol. 7, no. 5, pp. 424–444, Oct. 2003.
- R. H. C. Takahashi, J. A. Vasconcelos, J. A. Ramirez, and L. Krahenbuhl, “A multiobjective methodology for evaluating genetic operators,” IEEE Trans. Magn., vol. 39, no. 3, pp. 1321–1324, May 2003.
- A. Czarn, C. MacNish, K. Vijayan, B. Turlach, and R. Gupta, “Statistical exploratory analysis of genetic algorithms,” IEEE Trans. Evol. Comput., vol. 8, no. 4, pp. 405–421, Aug. 2004.

- B. Yuan and M. Gallagher, “Statistical racing techniques for improved empirical evaluation of evolutionary algorithms,” in Proc. Parallel Problem Solving Nature, 2004, pp. 172–181.
- D. Shilane, J. Martikainen, S. Dudoit, and S. Ovaska, “A general frame- work for statistical performance comparison of evolutionary computation algorithms,” in Proc. Artif. Intell. Applicat. Conf., 2006, pp. 7–12.
- S. Garcia, D. Molina, M. Lozano, and F. Herrera, “A study on the use of non-parametric tests for analyzing the evolutionary algorithms’ behavior: A case study on the CEC’2005 special session on real parameter optimization,” J. Heuristics, vol. 15, no. 6, pp. 617–644, 2008.

- E. G. Carrano, C. M. Fonseca, R. H. C. Takahashi, L. C. A. Pimenta, and O. M. Neto, “A preliminary comparison of tree encoding schemes for evolutionary algorithms,” in Proc. IEEE Int. Conf. Syst. Man Cybern., Oct. 2007, pp. 1969–1974.
- E. G. Carrano, R. H. C. Takahashi, and E. F. Wanner, “An enhanced statistical approach for evolutionary algorithm comparison,” in Proc. Genet. Evol. Comput. Conf., 2008, pp. 897–904.