# Bootstrap

## Tópicos de Estatística

José Passos

ISEG-UTL

30 de Outubro de 2012

# Introdução

Um problema fundamental em estatística é conhecer a variabilidade de uma estatística:

- A inferência estatística assenta na distribuição por amostragem de uma estatística;
- A distribuição por amostragem de uma estatística é a distribuição da estatística em amostras repetidas da população;
- O bootstrap é particularmente interessante em situações onde não se conhece a população ou não é possível obter amostras repetidas dessa população;
- O bootstrap é uma técnica que nos permite conhecer a distribuição por amostragem, de forma aproximada, apenas a partir da amostra observada;
- O bootstrap permite-nos calcular o enviesamento, variância, intervalos de confiança, etc., de um estimador com base apenas na amostra observada.

# Bootstrap não-paramétrico

- O bootstrap não-paramétrico trata a amostra como se fosse a população, considerenado amostras repetidas (com reposição) a partir da amostra observada.
- A estimativa bootstrap de uma função da estatística pode ser vista como uma estimativa *plug-in* que usa a função de distribuição empírica, $\hat{F}_n$, no lugar de $F$.
- Suponha $T_n = \bar{X}$. Pretende-se calcular $Var_F(T_n) = \sigma^2/n$ que é um funcional estatístico, função de $F$.
- Aproximar $Var_{\hat{F}_n}(T_n)$ por simulação. Como se pode simular a partir da distribuição de $T_n$ quando os dados têm distribuição $\hat{F}_n$? A resposta consiste no seguinte:
  - simular $X_1^*, X_2^*, \ldots, X_n^*$ a partir de $\hat{F}_n$;
  - calcular $T_n^* = T(X_1^*, X_2^*, \ldots, X_n^*)$
- Portanto, simular $X_1^*, X_2^*, \ldots, X_n^* \sim \hat{F}_n$ é equivalente a gerar $n$ observações com reposição, a partir de $X_1, X_2, \ldots, X_n$

# Bootstrap não-paramétrico (cont.)

Algoritmo para a estimação bootstrap da variância e do enviesamento de uma estatística:

1. gerar $X_1^*, X_2^*, \ldots, X_n^* \sim \hat{F}_n$
2. calcular $T_n^* = T(X_1^*, X_2^*, \ldots, X_n^*)$
3. Repetir os passos 1 e 2, $B$ vezes, obtendo $T_{n,1}^*, T_{n,2}^*, \ldots, T_{n,B}^*$
4. Fazer,

$$v_{boot} = \frac{1}{B} \sum_{b=1}^{B} \left( T_{n,b}^* - \frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* \right)^2$$

$$b_{boot} = \frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* - T_n$$

# Bootstrap paramétrico

Principal diferença:

- no bootstrap não-paramétrico gera-se $X_1^*, X_2^*, \ldots, X_n^*$ a partir da distribuição empírica, $\hat{F}_n$;
- no bootstrap paramétrico, $X_1^*, X_2^*, \ldots, X_n^*$ é gerado a partir de $F(x \mid \hat{\theta})$ onde $\hat{\theta}$ é uma estimativa de $\theta$ (por exemplo, máxima verosimilhança).

# N° de amostras Bootstrap

- Numa amostra de dimensão $n$ existem $n^n$ sub-amostras possíveis de dimensão $n$.
- Quantas amostras bootstrap são suficientes?
- Em geral são suficientes $B = n(\log n)^2$ [Babu and Singh (1983)] sub-amastras.

# Situações em que o Bootstrap não funciona

- Distribuições com abas pesadas, por exemplo, $\hat{\theta} = \bar{X}$ e $E(X^2) = \infty$.
- Estatísticas não-lisas (non-smooth), por exemplo, $\hat{\theta} = \max\{X_i\}$.
- Amostras dependentes.

# Bibliografia


- Babu & Singh (1983), Inference on Means Using the Bootstrap, *Annals od Statistics*, Volume 11, Number 3, 999-1003.
- Boos & Stefanski (2010), Efron's bootstrap, *Significance*, Volume 7, Issue 4, pp. 186-188
- Efron, B & Tibshirami, R. J. (1993), *An introduction to the bootstrap*, Chapman & Hall, New-York.
- Moore, D & McCabe, G (2005), *Introduction to the Practice of Statistics*, 5th Edition, W H Freeman & Co. [Chap. 14].