

O *Jackknife* e o *Bootstrap* - Breve introdução

Na Abordagem Tradicional em Estatística o procedimento habitual consiste em dispormos de:

$$X_1, X_2, \dots, X_n \sim F \text{ (desconhecida)}$$

As análises baseadas em **estatísticas** (funções da amostra aleatória)

Distribuição de amostragem da estatística $\xrightarrow{\quad}$ Inferência.

Mas . . . a distribuição de amostragem de uma estatística depende, geralmente, da distribuição da população subjacente à amostra, que é desconhecida.

Seja $T_n := T_n(X_1, X_2, \dots, X_n)$ uma estatística.

Característica de interesse, por exemplo, **variância desta estatística**, $Var[T_n]$

A Abordagem Tradicional em Estatística

Se T_n é uma média, $T_n = \bar{X}_n$, (ou uma função simples da média)
e para facilitar suponhamos (X_1, X_2, \dots, X_n) i.i.d.

$$T_n = \bar{X}_n \Rightarrow Var[T_n] = \frac{\sigma^2}{n} \text{ (função de uma quantidade desconhecida)}$$

$$Var[\bar{X}_n] \longrightarrow \text{estimada por } \frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Para outros parâmetros, a expressão regra geral é muito complicada

Abordagem tradicional: \longrightarrow simplificar o problema considerando aproximações ou desenvolvimentos de $Var[T_n]$.

Desvantagens: {
necessidade de valores muito elevados de n
dependência da fórmula teórica do modelo postulado \rightarrow validade(??)
dificuldade na obtenção da fórmula teórica.

Ideias iniciais de reamostragem e a orientação para uma linha de trabalho de análise estatística baseada em simulações.



Anos 40 /50 → o computador começou a ser usado para efectuar simulações:

- substituir complicadas e por vezes “grosseiras” aproximações por simulações
- despertar fortemente a atenção, quer de investigadores teóricos, quer de utilizadores de métodos estatísticos.

Métodos baseados em procedimentos repetidos sobre muitos conjuntos de réplicas dos próprios dados

são os chamados **métodos de reamostragem**

O jackknife e o bootstrap

são os métodos de reamostragem mais populares usados na análise estatística

Trabalho pioneiro → Quenouille (1949) → *jackknife*



Metodologia desenvolvida para **estimar** e portanto controlar **o viés de estimadores** e para construir **intervalos de confiança robustos**.

O termo *jackknife* foi introduzido por Tukey (1958), considerando que esta metodologia permitia testar hipóteses e calcular intervalos de confiança em situações em que não há melhores métodos que possam ser utilizados.



Finais dos anos 70, Efron unificou as ideias existentes e introduziu a metodologia *bootstrap não paramétrico simples*.

Métodos de reamostragem

A origem do termo **bootstrap** deriva da obra de Rudolph Raspe, autor do século XVIII, a quem se deve as *Aventuras do Barão Munchausen*.

Numa das suas obras ele diz:

“ The baron had fallen to the bottom of a deep lake. Just when he looked like all was lost, he thought to pick himself up by his own bootstraps. ”



Justificação baseada no **estimador média**, \bar{X}_n .

Seja $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} F$; $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$

Vamos construir amostras a partir daquela, retirando de cada vez uma das variáveis, i.e., retirando por exemplo a variável j , temos $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_n$

Define-se $\bar{X}_{n-1,(-j)} = \frac{\sum_{i=1}^n X_i - X_j}{n-1}$ tendo-se $X_j = n\bar{X}_n - (n-1)\bar{X}_{n-1,(-j)}$
i.e. os valores X_j podem ser obtidos à custa de \bar{X}_n e de $\bar{X}_{n-1,(-j)}$

Tukey considerou então para um estimador qualquer $T_n = T_n(X_1, X_2, \dots, X_n)$ de um parâmetro θ e chamou “**pseudo-valores**” a

$$\tilde{T}_j = nT_n - (n-1)T_{n-1,(-j)} \quad j = 1, 2, \dots, n$$

Os “**pseudo-valores**” desempenham o mesmo papel que X_j no cálculo de \bar{X}_n .

A média dos pseudo-valores

$$\frac{1}{n} \sum_{i=1}^n \tilde{T}_i = nT_n - (n-1)\bar{T}_{(\cdot)} \quad \text{com} \quad \bar{T}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n T_{n-1,(-i)}$$

é o **estimador jackknife de θ** que vamos representar por $T_n^J = \frac{1}{n} \sum_{i=1}^n \tilde{T}_i$.

Tukey conjecturou que os “**pseudo-valores**” \tilde{T}_i podiam ser considerados aproximadamente i.i.d. numa grande variedade de situações.

Para construir intervalos de confiança ou realizar testes de hipóteses para o parâmetro θ , sugeriu a estatística

$$\sqrt{n} \frac{(T_n^J - \theta)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (\tilde{T}_i - T_n^J)^2}} \sim t_{n-1}$$

Antes de Tukey ter sugerido o *jackknife* como instrumento para a inferência, em 1956, Quenouille sugeriu que usando a versão *Jackknife* de um estimador conseguíamos uma **redução do viés** de uma ordem de $1/n$.

Como sabemos o **viés de T_n** define-se como

$$E[T_n] - \theta$$

Quenouille propôs **o estimador jackknife do viés**

$$\widehat{Vies}^J [T_n] = T_n - T_n^J = (n - 1)(\bar{T}_{(.)} - T_n).$$

Nota: o estimador *jackknife* é o estimador corrigido do viés. De facto

$$T_n - (n - 1)(\bar{T}_{(.)} - T_n) = T_n^J$$

Em “*Bootstrap methods. Another look at the jackknife*”, Efron (1979)

—→ técnica não paramétrica para a estimação do desvio padrão

—→ recorrendo a métodos de computação intensiva.

Genericamente —→ estimar viés, variância, quantis, distribuição de amostragem do estimador (ou melhorar estimadores existentes.)

Ideia básica da metodologia **bootstrap não paramétrico**

$\underline{X}_n = (X_1, X_2, \dots, X_n) \stackrel{i.i.d.}{\sim} F$ T_n um estimador de θ

A ideia é construir a **amostra bootstrap**

$$\underline{X}_n^* = (X_1^*, X_2^*, \dots, X_n^*)$$

↑

são variáveis reamostradas de \underline{x}_n de acordo com a função de distribuição empírica,

$$\hat{F}_n(x) = \frac{\#\{i : x_i \leq x, 1 \leq i \leq n\}}{n}.$$

Considera-se $\underline{x}_n = (x_1, x_2, \dots, x_n) \sim \hat{F}_n$

$$\begin{array}{c} \downarrow \\ \underline{X}_n^* = (X_1^*, X_2^*, \dots, X_n^*). \end{array}$$

Nota: esta extracção é feita com reposição enquanto o *jackknife* faz extracções de amostras de dimensão $n - 1$, sem reposição.

Dada \underline{x}_n a distribuição da amostra *bootstrap* associada, \underline{X}_n^* , é

$$P(X_i^* = X_j | \underline{x}_n) = \frac{1}{n} \quad i, j = 1, \dots, n$$

$T_n^* := T_n(\underline{X}_n^*)$ **versão bootstrap do estimador T_n**

O comportamento da *versão bootstrap* deverá simular o comportamento de T_n .

→ A distribuição de T_n^* , é usada para aproximar a distribuição de amostragem (desconhecida), de T_n .

A metodologia Bootstrap não paramétrico

Recurso à simulação de Monte Carlo.

- dada uma amostra observada $\underline{x}_n = (x_1, x_2, \dots, x_n)$, constrói-se \hat{F}_n atribuindo a cada x_i peso $1/n$;
- gera-se uma amostra *bootstrap* $\underline{x}_n^* = (x_1^*, x_2^*, \dots, x_n^*)$, de variáveis $X_i^* \stackrel{i.i.d.}{\sim} \hat{F}_n$ e calcula-se $t_n^* = t_n(x_1^*, x_2^*, \dots, x_n^*)$;
- repete-se, independentemente, B vezes o passo anterior, obtendo assim B réplicas $(t_n^{*,1}, t_n^{*,2}, \dots, t_n^{*,B})$
- calcula-se as estimativas do viés, erro padrão e distribuição de amostragem *bootstrap*

$$\widehat{Vies}_B^*[T_n] = \sum_{i=1}^B t_n^{*,i} / B - \theta(\hat{F}_n) \quad \widehat{\sigma}_B^*[T_n] = \sqrt{\frac{1}{B-1} \sum_{i=1}^B (t_n^{*,i} - \overline{t_n^{*,i}})^2}$$

$$\widehat{F}_B^*(t) = \frac{\#\{i : t_n^{*,i} \leq t, 1 \leq i \leq B\}}{B}, \quad -\infty < t < +\infty$$

Nota: $\lim_{B \rightarrow \infty} \widehat{Vies}_B^* = Vies^*$ $\lim_{B \rightarrow \infty} \widehat{\sigma}_B^* = \sigma^*$ $\lim_{B \rightarrow \infty} \widehat{F}_B^* = F^*$

Desvio padrão — estimativa *bootstrap* é regra geral bastante boa, com um número pequeno de réplicas, Efron (1993) considera suficiente $B = 200$.

Viés — convergência é mais difícil de atingir.

Intervalos de confiança *bootstrap* — há vários procedimentos, mas aqui vamos apenas referir um deles:

- o métodos dos percentis — $\left(t_{\alpha/2}^*, t_{1-\alpha/2}^*\right)$, que são os percentis empíricos dos valores bootstrap $t_n^{*,i}$.

Problemas no uso do bootstrap

- Dimensão da amostra muito pequena (\hat{F}_n não é uma boa aproximação de F)
- Quando estamos com estruturas dependentes (séries temporais, dados espaciais, por ex.). O procedimento bootstrap aqui explicado assenta no pressuposto da **independência das variáveis**. Mas há formas de resolver a questão.
- Estimação de valores extremos (ex. $\max(X_i)$, o percentil de probabilidade 99.9%, etc.)
- Se existem outliers na amostra, então \hat{F} não é um bom estimador de F .
- Se pretendemos estimar quantidades não “suaves”
- Em dados multivariados - quando há um número elevado de dimensões há problemas com o estimador de F .

- Chernick, M. R. (1999) Bootstrap Methods: A Practitioner's Guide. New York: Wiley.
- Davison, A., Hinkley, D., (1997). Bootstrap Methods and Their Application. Cambridge University Press.
- Efron, B., Tibshirani, R., (1998). An Introduction to the Bootstrap. Chapman and Hall, Boca Raton.
- Good, P. (1998). Resampling Methods: A practical Guide to Data Anal
- Lahiri, S., (2003). Resampling Methods for Dependent Data. Springer-Verlag, New York.
- Manly, B. F. J. (1997) Randomization, Bootstrap and Monte Carlo Methods in Biology. Second edition. London: Chapman & Hall
- Singh, K., (1981). On the asymptotic accuracy of Efron's bootstrap. Annals of Statistics 9 (6), 1187-1195.