AMOSTRAGEM E ANÁLISE AMBIENTAL (2020/2021)

Manuela Neves

O *Jackknife* e o *Bootstrap* - Introdução

Métodos de reamostragem

Abordagem Tradicional em Estatística - - Procedimento habitual

$$X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} F \text{ (desconhecida)}$$

Análises baseadas em **estatísticas** (funções da amostra aleatória)

Distribuição de amostragem da estatística $\longleftarrow$ Inferência.

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

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

A Abordagem Tradicional em Estatística

Característica de interesse, por exemplo, **variância desta estatística**

$$Var[T_n] = \int \left[T_n(\underline{x}) - \int T_n(\underline{y}) d \prod_{i=1}^{n} F(y_i) \right]^2 d \prod_{i=1}^{n} F(x_i).$$

Se T_n é uma média (ou uma função simples da média)

$$T_n = \overline{X}_n \Rightarrow Var[T_n] = \frac{\sigma^2}{n} \text{ (função de quantidades desconhecidas)}$$

o procedimento clássico

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

A Abordagem Tradicional em Estatística

Regra geral, a expressão 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
- dificuldade na obtenção da fórmula teórica.
- ...

Métodos de reamostragem

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

$\downarrow$

Anos 40 /50 $\longrightarrow$ o computador começou a ser usado para efectuar simulações o que levou a:

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

Métodos de reamostragem

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 dos métodos de reamostragem mais populares usados na análise estatística

Métodos de reamostragem

Trabalho pioneiro $\rightarrow$ **Quenouille (1949)**

$\downarrow$

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.

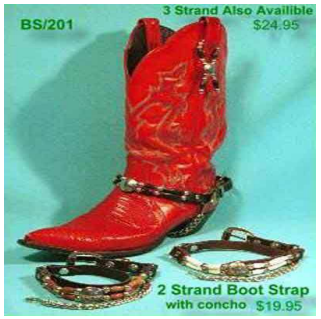
Finais dos anos 70, Efron unificou as ideias existentes e introduziu a metodologia *bootstrap não paramétrico simples*.

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.´´



A metodologia Jackknife – Justificação

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

Então $\overline{X}_{n-1,(-j)} = \frac{\sum_{i=1}^{n} X_i - X_j}{n-1}$ e tem-se

$$X_j = n\overline{X}_n - (n-1)\overline{X}_{n-1,(-j)}$$

i.e. os valores X_j podem ser obtidos à custa de $\overline{X}_n$ e de $\overline{X}_{n-1,(-j)}$

Tukey considerou então para um estimador qualquer $T_n = T_n(X_1, X_2, \cdots, X_n)$ de um parâmetro θ os **"pseudo-valores´´**

$$\widetilde{T}_j = nT_n - (n-1)T_{n-1,(-j)} \qquad j = 1, 2, \cdots, n$$

Os "pseudo-valores´´ desempenham o mesmo papel que os valores X_j no cálculo de $\overline{X}$.

A metodologia Jackknife

A média dos pseudo-valores

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{T}_i = nT_n - (n-1)\overline{T}_{(.)} \qquad \text{com} \qquad \overline{T}_{(.)} = n^{-1}\sum_{i=1}^{n} T_{n-1,(-i)}$$

é o **estimador jackknife de** θ e representa-se por $\boldsymbol{T_n^J}$.

Tukey conjecturou que os "**pseudo-valores**" $\widetilde{T}_i$ 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

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

Estimador Jackknife da variância

E quanto a estimação *jackknife* de $Var[T_n]$?

De novo (por analogia com a média)

$$\widehat{Var}^J[T_n] = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\widetilde{t}_i - t_n^J\right)^2 = \frac{n-1}{n} \sum_{i=1}^{n} \left(t_{n-1,(-i)} - \bar{t}_{(.)}\right)^2$$

Esta igualdade motivou Tukey (1958) a considerar como intervalo de confiança a $(1-\alpha)100\%$ para $\theta = \theta(F)$

$$t_n^J \pm t_{1-\alpha/2,n-1} \sqrt{\widehat{Var}^J[T_n]}$$

Porém... este I.C. só se tem revelado satisfatório assintoticamente.

A metodologia Bootstrap

Em *"Bootstrap methods. Another look at the jackknife´´*, Efron (1979)

$\longrightarrow$ técnica não paramétrica para a estimação do desvio padrão

$\longrightarrow$ recorrendo a métodos de computação intensiva.

Genericamente $\longrightarrow$ estimar viés, variância, quantis, distribuição de amostragem do estimador (ou melhorar estimadores existentes.)

A metodologia Bootstrap

Ideia básica da metodologia *bootstrap não paramétrico*

$\underline{X}_n = (X_1, X_2, \cdots X_n) \overset{i.i.d.}{\sim} F$ T_n um estimador de $\theta(F)$
aproximar este estimador pelo mesmo funcional da *amostra bootstrap*

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

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

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

A metodologia Bootstrap

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

$$\downarrow$$

$$\underline{X}^*_n = (X^*_1, X^*_2, \cdots, X^*_n).$$

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

$$P(X^*_i = X_j | \underline{x}_n) = \frac{1}{n} \qquad i, j = 1, \cdots n$$

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

A metodologia Bootstrap

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

$\longrightarrow$ 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 $\underline{x}_n = (x_1, x_2, \cdots x_n)$, constrói-se $\widehat{F}_n$ atribuíndo a cada x_i peso $1/n$;
- gera-se uma amostra *bootstrap* $\underline{x}^*_n = (x^*_1, x^*_2, \cdots x^*_n)$, de variáveis $X^*_i \overset{i.i.d.}{\sim} \widehat{F}_n$ e calcula-se $t^*_n = t_n(x^*_1, x^*_2, \cdots, x^*_n)$;
- repete-se, independentemente, B vezes o passo anterior, obtendo assim B réplicas $(t^{*,1}_n, t^{*,2}_n, \cdots t^{*,B}_n)$
- calcula-se as estimativas do viés, erro padrão e distribuição de amostragem *bootstrap*

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

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

A metodologia Bootstrap

Nota:

$\lim_{B\to\infty} \widehat{Vies}^*_B = Vies^*$ $\lim_{B\to\infty} \widehat{\sigma^*_B} = \sigma^*$ $\lim_{B\to\infty} \widehat{F^*_B} = F^*$

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.

A metodologia Bootstrap

Intervalos de confiança *bootstrap* — há vários procedimentos:

- intervalo *t-bootstrap* — $\left(t_n - t^*_{\alpha/2}\sigma^*_T, t_n + t^*_{\alpha/2}\sigma^*_T\right)$, com $t^*_{\alpha/2}$ quantis *bootstrap* da variável T^*_n estandardizada.
- o métodos dos percentis — $\left(t^*_{\alpha/2}, t^*_{1-\alpha/2}\right)$
- outros intervalos com redução de viés ...

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.

Bibliografia

- 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.