

Métodos de Reamostragem e Subamostragem no R-Project

Sara Morgado Nunes e João Renato Sebastião

Departamento de Matemática e Informática
Escola Superior de Gestão do Instituto Politécnico de Castelo Branco - Portugal
e-mail: sara@esg.ipcb.pt; jrenato@esg.ipcb.pt

RESUMO

Neste trabalho consideram-se o bootstrap e a subamostragem enquanto metodologias computacionalmente intensivas que podem constituir um meio de resposta eficaz na avaliação da precisão de uma determinada estatística. Num estudo de simulação aplicam-se estas metodologias a observações geradas por uma distribuição de probabilidade exponencial.

Palavras chave: bootstrap, subamostragem, intervalo de confiança e R-Project.

1. INTRODUÇÃO

Muitos dos problemas que surgem habitualmente em Estatística envolvem a estimação de um parâmetro desconhecido θ . Assim, tendo-se optado por uma estatística $\hat{\theta}$ para estimar um determinado parâmetro θ , ocorre a necessidade de avaliar a precisão de $\hat{\theta}$ enquanto estimador de θ . Nas situações em que a distribuição de $(\hat{\theta} - \theta)$ é conhecida, é possível determinar um intervalo de confiança exacto, porém, essa situação nem sempre se verifica. Numa tentativa de dar resposta a este problema, surgiram o bootstrap e a subamostragem que são metodologias computacionalmente intensivas, hoje amplamente utilizadas na Inferência Estatística, as quais facilitam a construção de intervalos ou regiões de confiança para um determinado parâmetro de interesse quando não se dispõem de métodos analíticos, os quais geralmente se baseiam em resultados assintóticos. O princípio básico destas metodologias é a utilização repetida da amostra original de modo a obter várias estimativas de θ que serão depois usadas para estimar ou aproximar a distribuição de $\hat{\theta}$ e as suas características.

Apesar de a ideia subjacente a estas técnicas ser bastante simples, a sua divulgação em larga escala só é exequível quando se procede à sua implementação em programas estatísticos muito utilizados, como o R-Project. O R-Project foi escrito em 1996 por Robert Gentleman e Ross Ihaka (Departamento de Estatística da Universidade de Auckland), tendo então ficado conhecido por R&R (Gentleman e Ihaka, 1996). Depois disso, inúmeros contributos têm sido dados para o aperfeiçoamento e desenvolvimento do R, tratando-se hoje de uma linguagem largamente utilizada no cálculo estatístico por pessoas de todo o mundo. Para esta rápida difusão do R tem contribuído, não só a qualidade dos *packages* aí disponibilizados, mas também o facto deste *software* estar acessível na web (no site www.r-project.org), o que, por sua vez, tem motivado a comunidade científica a implementar grande parte das metodologias estatísticas que desenvolve em *packages* utilizados no R.

A metodologia bootstrap foi inicialmente implementada por Efron e Tibshirani (1985) em Fortran e, mais tarde, adaptada para S-Plus e descrita por Efron e Tibshirani (1993), resultando o *bootstrap package*. Posteriormente, com base nos trabalhos de Davison e Hinkley (1997), Angelo Canty escreveu o *package boot* para S-Plus (Canty, 1998), o qual

foi depois adaptado para R por Brian Ripley. Hoje, o *package* `boot` é largamente utilizado no R e permite implementar muitas variantes da metodologia bootstrap, salientando-se neste trabalho a sua importância.

Em 1994 Politis e Romano propuseram a metodologia da subamostragem que, à semelhança do bootstrap, permite fazer inferência sobre um parâmetro θ pertencente a uma distribuição de probabilidade desconhecida. No entanto, até ao momento, o R não dispõe de um *package* específico para a aplicação desta metodologia, tendo portanto sido necessário proceder à sua implementação em R para a prossecução deste trabalho.

Neste artigo procede-se a uma revisão muito breve das metodologias bootstrap e subamostragem enquanto ferramentas úteis na construção de intervalos de confiança para um parâmetro associado a uma determinada distribuição de probabilidade. Num estudo de simulação, aplicam-se as referidas metodologias a conjuntos de dados gerados por uma distribuição exponencial, com o objectivo de calcular intervalos de confiança bootstrap e subamostragem para o parâmetro de escala.

2. BOOTSTRAP

O bootstrap faz parte da grande família dos métodos de reamostragem e foi originalmente proposto por Efron em 1979, tendo revolucionado a Estatística. Embora as ideias subjacentes ao bootstrap sejam antigas, o interesse da comunidade científica pelo método surgiu essencialmente após os trabalhos de Efron (1979) e Efron e Tibshirani (1986, 1993) onde o bootstrap é apresentado como um método de base computacional adequado a vários problemas de estimação. Desde então, o bootstrap tem vindo a ser aperfeiçoado, constituindo actualmente um método computacionalmente intensivo eficaz e amplamente usado na Inferência Estatística.

Seja $Y = \{X_1, X_2, \dots, X_n\}$ uma amostra aleatória pertencente a uma determinada distribuição de probabilidade desconhecida F . Com base em Y , estima-se um determinado parâmetro de interesse $\theta = \theta(F)$ obtendo-se $\hat{\theta}$. Pretende-se avaliar a precisão de $\hat{\theta}$ enquanto estimador de θ , porém a distribuição de $\hat{\theta}$ pode ser desconhecida. O bootstrap permite construir, de modo simples, estimadores para a distribuição de $\hat{\theta}$ com base na amostra Y . Assim, começa-se por construir a distribuição empírica $\hat{F}$, estimador de F , a qual se obtém atribuindo peso $\frac{1}{n}$ a cada um dos valores observados X_i , $i = 1, \dots, n$. Com $\hat{F}$ fixo, defina-se uma amostra bootstrap $Y^* = \{X_1^*, X_2^*, \dots, X_n^*\}$ como sendo uma amostra aleatória de tamanho n obtida de $\hat{F}$ com reposição a partir de Y . A notação $*$ indica que Y^* não é o conjunto de dados original Y mas sim uma versão reamostrada de Y . Assim, seleccionam-se B amostras bootstrap independentes de dimensão n , $Y^{*1}, Y^{*2}, \dots, Y^{*B}$, obtidas com reposição a partir de Y . Associada a cada conjunto de dados bootstrap Y^* está uma réplica bootstrap de $\hat{\theta}$ notada por $\hat{\theta}_b^*$, $b = 1, \dots, B$, a qual resulta do cálculo da estatística de interesse em Y^* .

A função `boot` permite, de forma simples, obter as B réplicas bootstrap, bastando para isso definir o conjunto de dados em causa, a estatística de interesse e o número de réplicas bootstrap pretendido. A partir das B réplicas bootstrap é possível obter ainda os estimadores bootstrap do viés e da variância de θ que são dados respectivamente por $b_B = B^{-1} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta})$ e $v_B = (B - 1)^{-1} \sum_{b=1}^B (\hat{\theta}_b^* - \bar{\theta}^*)^2$. O bootstrap pode também ser usado para calcular intervalos de confiança para uma determinada estatística de interesse sendo a ideia mais simples recorrer a uma aproximação normal com viés b_B e variância v_B . Neste caso é comum recorrer-se ao gráfico de probabilidade normal QQ-plot das B réplicas bootstrap com o objectivo de confirmar se tal aproximação é exequível. Assim, assumindo que a distribuição de $\hat{\theta}$ é aproximadamente normal, um **intervalo de confiança bootstrap**

normal a $(1 - 2\alpha)$ para θ é dado por

$$I_{norm} = [\hat{\theta} - b_B - \sqrt{v_B}z(1 - \alpha); \hat{\theta} - b_B + \sqrt{v_B}z(1 - \alpha)] \quad (1)$$

onde $z(\alpha)$ denota o quantil α da distribuição normal standard. Porém, se a distribuição de $(\hat{\theta} - \theta)$ não é suficientemente próxima da Normal, é possível construir outro intervalo de confiança assumindo que a distribuição de $(\hat{\theta}^* - \hat{\theta})$ aproxima a distribuição de $(\hat{\theta} - \theta)$. Neste caso, um **intervalo de confiança bootstrap basic** a $(1 - 2\alpha)$ é dado por

$$I_{basic} = [2\hat{\theta} - \hat{\theta}_{(B+1)(1-\alpha)}^*; 2\hat{\theta} - \hat{\theta}_{(B+1)\alpha}^*] \quad (2)$$

onde $\hat{\theta}_1^* < \hat{\theta}_2^* < \dots < \hat{\theta}_B^*$ denota os θ_b^* , $b = 1, \dots, B$ ordenados. Nos casos em que $(B + 1)\alpha$ não corresponde a um número inteiro, é possível recorrer a uma interpolação na escala de quantis normal (Davison e Hinkley, 1997). Se for possível obter um estimador consistente V para a variância de $\hat{\theta}$, pode escrever-se a versão studentizada de $\hat{\theta}$, $Z = \frac{\hat{\theta} - \theta}{\sqrt{V}}$, cujos quantis podem ser estimados a partir dos B valores simulados $z_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\sqrt{v_b^*}}$, obtendo-se um **intervalo de confiança bootstrap studentizado** a $(1 - 2\alpha)$

$$I_{stud} = [\hat{\theta} - \sqrt{v}z_{(B+1)(1-\alpha)}^*; \hat{\theta} - \sqrt{v}z_{(B+1)\alpha}^*]. \quad (3)$$

Ao contrário dos intervalos de confiança atrás referidos, existem outros que se mantêm invariantes sob transformações dos parâmetros, uma vez que se baseiam directamente nos quantis de $\hat{\theta}^*$. Assim, um **intervalo de confiança bootstrap percentil** a $(1 - 2\alpha)$ vem

$$I_{perc} = [\hat{\theta}_{(B+1)\alpha}^*; \hat{\theta}_{(B+1)(1-\alpha)}^*] \quad (4)$$

enquanto um **intervalo bootstrap percentil ajustado (BCa)** a $(1 - 2\alpha)$ é dado por

$$I_{BCa} = [\hat{\theta}_{(B+1)\alpha'}^*; \hat{\theta}_{(B+1)(1-\alpha'')}^*] \quad (5)$$

onde α' e α'' são escolhidos de forma a melhorar as propriedades do intervalo.

A partir da função `boot.ci` é possível calcular todos estes intervalos de confiança no R.

3. SUBAMOSTRAGEM

A ideia subjacente ao método de subamostragem introduzido por Politis e Romano (1994) é aproximar a distribuição amostral de uma estatística com base no cálculo da mesma em subamostras do conjunto de dados original e não em conjuntos de dados reamostrados como no bootstrap.

Sejam $\{X_1, X_2, \dots, X_n\}$ n observações de uma amostra aleatória que toma valores num espaço amostral $\mathcal{S}$ e seja $\mathcal{P}$ uma medida de probabilidade. Tendo-se optado por uma estatística $\hat{\theta}$ para estimar um determinado parâmetro $\theta = \theta(\mathcal{P})$, o objectivo é aproximar a distribuição amostral de $(\hat{\theta}_n - \theta(\mathcal{P}))$ pela distribuição empírica dos valores da estatística calculada nos $\binom{n}{b}$ subconjuntos de dimensão $b < n$ de observações amostradas sem reposição. Os valores resultantes do cálculo da estatística em cada uma destas subamostras são convenientemente normalizados de forma a aproximarem a distribuição amostral em causa. Deste modo, a subamostragem constitui um método muito geral para a construção de intervalos de confiança de primeira ordem assintoticamente válidos pois, se b for tal que $b/n \rightarrow 0$ e $b \rightarrow \infty$ quando $n \rightarrow \infty$, o método é válido sempre que a estatística original, convenientemente normalizada, converge em distribuição para uma determinada distribuição.

Note-se por $Y_1, Y_2, \dots, Y_{N_n}$ os $N_n = \binom{n}{b}$ subconjuntos de tamanho b de $\{X_1, X_2, \dots, X_n\}$ e por $\hat{\theta}_{n,b,i}$ o valor da estatística $\hat{\theta}$ calculada no conjunto de dados Y_i . Seja $J_n(\mathcal{P})$ a distribuição amostral de $\tau_n(\hat{\theta}_n - \theta(\mathcal{P}))$, onde τ_n é uma constante de normalização apropriada. Defina-se também a correspondente função distribuição acumulada como

$$J_n(x, \mathcal{P}) = \text{Prob}_{\mathcal{P}}\{\tau_n(\hat{\theta}_n - \theta(\mathcal{P})) \leq x\}.$$

A aproximação subamostragem a $J_n(x, \mathcal{P})$ é então definida

$$L_{n,b}(x) = \frac{1}{N_n} \sum_{t=1}^{N_n} I\{\tau_b(\hat{\theta}_{n,b,i} - \hat{\theta}_n) \leq x\}$$

onde $I\{E\}$ é a função indicatriz do evento $E = \{\tau_b(\hat{\theta}_{n,b,i} - \hat{\theta}_n) \leq x\}$.

Para cada i , Y_i é uma subamostra de tamanho b do verdadeiro modelo $\mathcal{P}$. Logo, a distribuição exacta de $\tau_b(\hat{\theta}_{n,b,i} - \theta(\mathcal{P}))$ é $J_b(\mathcal{P})$. A distribuição empírica dos N_n valores de $\tau_b(\hat{\theta}_{n,b,i} - \theta(\mathcal{P}))$ constitui uma boa aproximação a $J_n(\mathcal{P})$. Assim, sendo $\theta(\mathcal{P})$ desconhecido, podemos substituir $\theta(\mathcal{P})$ por $\hat{\theta}_n$ uma vez que $\tau_b(\hat{\theta}_n - \theta(\mathcal{P}))$ é de ordem τ_b/τ_n em probabilidade, assumindo-se que $\tau_b/\tau_n \rightarrow 0$.

Para $\alpha \in [0, 1]$, seja $c_{n,b}(1 - \alpha) = \inf\{x : L_{n,b}(x) \geq 1 - \alpha\}$. Então, a probabilidade de cobertura assíntótica sob $\mathcal{P}$ do intervalo $I_1 = [\hat{\theta}_n - \tau_n^{-1}c_{n,b}(1 - \alpha), \infty[$ é $1 - \alpha$.

Um intervalo de confiança bilateral pode ser obtido através da intersecção de dois intervalos unilaterais. Assim, o **intervalo de confiança subamostragem de caudas iguais** análogo a I_1 é

$$I_{ET} = \left[\hat{\theta}_n - \tau_n^{-1}c_{n,b}(1 - \frac{\alpha}{2}), \hat{\theta}_n + \tau_n^{-1}c_{n,b}(\frac{\alpha}{2}) \right]. \quad (6)$$

Como aproximação alternativa, também podem ser construídos intervalos de confiança simétricos, sendo, neste caso, a aproximação por subamostragem a $J_{n,|\cdot|}(x, \mathcal{P})$ definida por

$$L_{n,b,|\cdot|}(x) = \frac{1}{N_n} \sum_{t=1}^{N_n} I\{\tau_b|\hat{\theta}_{n,b,i} - \hat{\theta}_n| \leq x\}.$$

Assim, denotando um quantil $(1 - \alpha)$ de $L_{n,b,|\cdot|}$ por $c_{n,b,|\cdot|}(1 - \alpha)$, o **intervalo de confiança subamostragem simétrico** é dado por

$$I_{sym} = [\hat{\theta}_n - \tau_n^{-1}c_{n,b,|\cdot|}(1 - \alpha), \hat{\theta}_n + \tau_n^{-1}c_{n,b,|\cdot|}(1 - \alpha)]. \quad (7)$$

4. ESTUDO DE SIMULAÇÃO

Nesta secção apresentam-se os resultados de um estudo de simulação levado a efeito em R-Project, no qual se aplicam as metodologias bootstrap e subamostragem a observações geradas por uma distribuição exponencial, com vista à determinação de intervalos de confiança para o parâmetro de escala.

Seja X uma variável aleatória com distribuição exponencial uniparamétrica e com função densidade de probabilidade dada por

$$f(x; \theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right)$$

onde θ designa o parâmetro de escala. O estimador de máxima verosimilhança de θ é dado por $\frac{1}{n} \sum_{i=1}^n X_i$. Tendo em conta que $\frac{2n\hat{\theta}}{\theta}$ segue uma distribuição Qui-quadrado com $2n$ graus de liberdade (Lawless (1982) e Bain e Engelhardt (1991)), é possível determinar intervalos de confiança exactos para θ , os quais se calculam neste estudo.

Começaram por gerar-se amostras de 20 observações de uma distribuição exponencial nas quais o parâmetro de escala tomou sucessivamente os valores 0.5, 1 e 2. Às amostras geradas aplicaram-se as metodologias bootstrap e subamostragem. Na aplicação do bootstrap recorreu-se à função `boot` para o cálculo de 999 réplicas bootstrap e à função `boot.ci` para a determinação dos intervalos de confiança a 95% bootstrap normal I_{norm} (1), basic I_{basic} (2), studentizado I_{stud} (3), percentil I_{perc} (4) e percentil ajustado I_{BCa} (5). A subamostragem aplicou-se usando diferentes tamanhos de bloco ($b = 5$, $b = 10$ e $b = 15$), e construíram-se, para cada situação, intervalos de confiança a 95% de caudas iguais I_{ET} (6), e intervalos de confiança simétricos I_{sym} (7), tendo sido para tal necessário proceder à implementação desta metodologia no R. Estes procedimentos foram repetidos 2000 vezes registando-se o número de intervalos de confiança que continham o verdadeiro parâmetro com o objectivo de estimar a probabilidade de cobertura.

É ainda de referir que na estimação da variância de θ , necessária para a determinação dos intervalos I_{stud} , foi utilizado o estimador sugerido por Meeker e Escobar (1998)

$$Var(\hat{\theta}) = \frac{\hat{\theta}^2}{n}. \quad (8)$$

A fim de estabilizar a variância, é possível recorrer a uma transformação logarítmica do parâmetro θ . Assim, foram ainda calculados os intervalos de confiança (1), (2) e (3) utilizando a distribuição de $\log(\hat{\theta})$ em vez da distribuição de $\hat{\theta}$, resultando os intervalos $I_{log(norm)}$, $I_{log(basic)}$, $I_{log(stud)}$, para os quais se estimaram, de forma semelhante, as respectivas probabilidades de cobertura. Não se apresentam aqui os intervalos $I_{log(basic)}$ e $I_{log(stud)}$ por não diferirem dos intervalos I_{stud} , o que se deve ao facto de a distribuição em causa ser exponencial e o estimador para a variância de θ ser dado por (8). Além dos intervalos bootstrap e subamostragem, calcularam-se também intervalos de confiança exactos com base nos resultados apresentados por Lawless (1982) e Bain e Engelhardt (1991).

As probabilidades de cobertura estimadas para intervalos de confiança bootstrap, subamostragem e exactos apresentam-se na Tabela 1.

bootstrap				subamostragem				exactos		
θ	0.5	1	2	θ	0.5	1	2	θ	0.5	0.95
I_{norm}	0.90	0.91	0.89	b=5	I_{ET}	0.83	0.84		1	0.95
$I_{log(norm)}$	0.92	0.92	0.92		I_{sym}	0.85	0.86		2	0.95
I_{basic}	0.88	0.89	0.87	b=10	I_{ET}	0.87	0.87			
I_{stud}	0.91	0.92	0.92		I_{sym}	0.90	0.91			
I_{perc}	0.90	0.91	0.90	b=15	I_{ET}	0.61	0.63			
I_{BCa}	0.90	0.92	0.91		I_{sym}	0.63	0.64			

Table 1: Probabilidades de cobertura estimadas para os intervalos de confiança a 95% bootstrap, subamostragem e exactos para $\theta = 0.5$, $\theta = 1$ e $\theta = 2$.

Note-se que intervalos de confiança bootstrap, subamostragem e exactos apresentam probabilidades de cobertura comparáveis, destacando-se, contudo, os intervalos de confiança exactos. Nos intervalos bootstrap são os intervalos $I_{log(norm)}$ e I_{stud} que apresentam probabilidades de cobertura mais próximas de 0.95, enquanto nos intervalos subamostragem, intervalos de confiança simétricos (I_{sym}) apresentam probabilidades de cobertura superiores à dos intervalos de confiança de caudas iguais (I_{ET}), sendo o tamanho de bloco que conduz a melhores resultados $b = 10$. Tanto nas metodologias bootstrap como subamostragem se obtêm, de um modo geral, intervalos de confiança com melhor probabilidade de cobertura para $\theta = 1$ do que para os restantes valores atribuídos ao parâmetro de escala.

Com base nas amplitudes dos 2000 intervalos de confiança bootstrap, subamostragem e exactos determinados, construíram-se as respectivas boxplots. Na Figura 1 apresentam-

se apenas as boxplots relativas aos intervalos determinados para $\theta = 2$ por não se terem registado diferenças significativas nas amplitudes dos intervalos obtidos para os restantes valores de θ . Nos intervalos subamostragem o tamanho de bloco usado foi $b = 10$. Na apresentação das boxplots utilizou-se uma escala logarítmica.

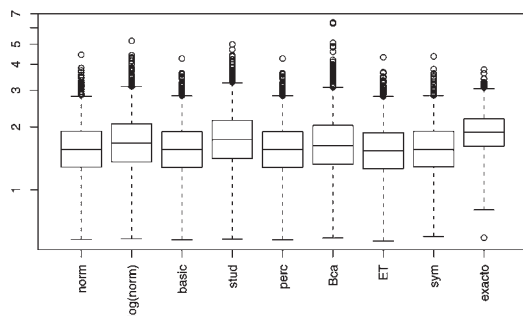


Figure 1: Boxplots das amplitudes dos intervalos de confiança a 95% bootstrap (I_{norm} , $I_{log(norm)}$, I_{basic} , I_{stud} , I_{perc} e I_{Bca}), subamostragem (I_{ET} e I_{sym}) e exactos para $\theta = 2$.

condições gerais, constituindo portanto instrumentos eficazes na construção de intervalos de confiança para parâmetros associados a uma determinada distribuição de probabilidade. Neste trabalho aplicam-se estas metodologias a observações geradas por uma distribuição exponencial, ficando para um trabalho futuro a aplicação das mesmas a outras distribuições e a diferentes conjuntos de dados entre os quais se destacam os dados censurados.

6. REFERÊNCIAS

- Bain, L. J. e Engelhardt, M. (1991). *Statistical analysis of reliability and life-testing models - theory and methods* (2nd ed.). Marcel Dekker. New York.
- Canty, A. J. (1998). An S-Plus Library for Resampling Methods. *Computing Science and Statistics: Proceedings of the 30th Symposium on the Interface*, 236-241.
- Davison, A. C. e Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press. Cambridge.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7, 1-26.
- Efron, B. e Tibshirani, R. J. (1985). The bootstrap method for assessing statistical accuracy. *Behaviormetrika* 17, 1-35.
- Efron, B. e Tibshirani, R. J. (1986). Bootstrap methods for standart errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1, 54-77.
- Efron, B. e Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall. New York.
- Ihaka, R. e Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5, 299-314.
- Lawless, J. F. (1982). *Statistical models and methods for lifetime data*. John Wiley and Sons. New York.
- Meeker, W. Q. e Escobar, L. A. (1998). *Statistical methods for reliability data..* John Wiley and Sons. New York.
- Politis, D. N. e Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* 22, 2031-2050.
- Politis, D. N., Romano, J. P. e Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.

Note-se que são os intervalos de confiança exactos aqueles que apresentam, por um lado, maiores amplitudes e, por outro, amplitudes com menor intervalo de variação. Dos restantes intervalos são os $I_{log(norm)}$ e I_{stud} aqueles que apresentam maiores amplitudes bem como maior intervalo de variação das mesmas (note-se que $I_{log(norm)}$ e I_{stud} apresentavam probabilidades de cobertura superiores). I_{norm} , I_{basic} , I_{perc} , I_{ET} e I_{sym} apresentam, não só amplitudes inferiores, como menor intervalo de variação das mesmas.

5. CONSIDERAÇÕES FINAIS

Constata-se que as metodologias bootstrap e subamostragem se aplicam sob