CÁLCULO DO DESVIO PADRÃO DE ESTIMADORES DE PARÂMETROS HIDROLÓGICOS

por

Fernanda Serra Costa [1], Jerson Kelman [1,2], Jorge M. Damázio [1,3]

RESUMO -- Neste artigo reporta-se experiências controladas para avaliar a eficiência de métodos de estimação de $S(\hat{\theta})$, desvio padrão do estimador de um parâmetro θ, quando aplicados a problemas de Hidrologia. A análise dos resultados obtidos permite concluir que as Técnicas de Reamostragem experimentadas (Bootstrap e Jackknife) são razoavelmente precisas, podendo ser usadas na construção de intervalos de confiança para parâmetros hidrológicos.

INTRODUÇÃO

A utilização de modelos probabilísticos em Hidrologia envolve dois tipos de incertezas: a incerteza na escolha do modelo que descreve o fenômeno em estudo (ex: distribuição normal ou lognormal) e a incerteza nos parâmetros, ou seja, incerteza na estimação dos parâmetros do modelo selecionado, causando o chamado "erro amostral de estimação".

Diversos pesquisadores já atentaram para o fato que os "erros amostrais de estimação" podem ser muito importantes. Uma forma tradicional de se considerar a incerteza nos parâmetros de modelos probabilísticos é o uso dos chamados "Intervalos de Confiança", USWRC (1974) e Ashkar (1986).

O intervalo de confiança de um parâmetro é dado em função do desvio padrão de seu estimador da seguinte forma:

$$\hat{\theta} - z\,(1-\alpha/2)\; S(\hat{\theta}) \leq \theta \leq \hat{\theta} + z(1-\alpha/2)\; S(\hat{\theta}) \qquad (1)$$

onde $\hat{\theta}$ é a estimativa do parâmetro θ, $S(\hat{\theta})$ é o desvio padrão de $\hat{\theta}$ e $z(1-\alpha/2)$ é o valor obtido de uma distribuição padronizada (usualmente assume-se a distribuição normal). Um método para o cálculo de $S(\hat{\theta})$ é sem dúvida fundamental para a construção de intervalos de confiança. No caso de parâmetros usuais, tais como média e variância de processos aleatórios, o hidrólogo encontrará

---

[1] Pesquisador do Centro de Pesquisas de Energia Elétrica - CEPEL Rio de Janeiro, RJ.

[2] Professor da COPPE/UFRJ, Rio de Janeiro, RJ.

[3] Professor Colaborador da COPPE/UFRJ, Rio de Janeiro, RJ.

na Estatística Clássica fórmulas já consagradas para o cálculo de $S(\hat{\theta})$. Já para parâmetros mais específicos tais como "valor esperado do tamanho do reservatório" estas fórmulas precisam ainda ser derivadas. Nestes casos o hidrólogo tem como alternativa o uso das chamadas Técnicas de Reamostragem (Efron, 1982). Como se verá adiante o cálculo de $S(\hat{\theta})$ por estas técnicas é extremamente intuitivo e simples.

O objetivo deste artigo é apresentar o uso de duas destas técnicas em problemas da Hidrologia e comparar os seus resultados com os fornecidos por fórmulas da Estatística Clássica.

JACKKNIFE

A idéia original do Jackknife foi proposta por Quenouille (1949), sendo seu uso generalizado por Tuckey (1958).

O Jackknife consiste em, dado uma amostra $\underset{\sim}{x} = \{x_1, x_2, .., x_n\}$ contendo n sorteios independentes da variável aleatória X, dividi-la em w grupos de tamanho h, onde n = wh, e tomar $\hat{\theta}_{-i}$ como a estimativa do parâmetro θ tendo omitido o i-ésimo grupo de observações. Exceto quando o volume de dados é muito grande, deve-se usar w = n e h = 1.

A determinação do desvio padrão $\hat{S}_{jackk}$ é feita obtendo-se da amostra x n "pseudo-amostras" x^i_{jackk}, i = 1,...,n., por omissão da i-ésima observação da amostra:

$$x^i_{jackk} = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\} \qquad (2)$$

A estimativa do parâmetro θ, calculada sobre x^i_{jackk} é então

$$\hat{\theta}_{-i} = g(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \qquad (3)$$

O estimador Jackknife de $S(\hat{\theta})$ é dado por:

$$\hat{S}_{jackk} = \left\{ \left(\frac{n-1}{n}\right) \Sigma_{i=1}^{n} (\hat{\theta}_{-i} - \hat{\theta}_{(.)})^2 \right\}^{1/2} \qquad (4)$$

onde $\hat{\theta}_{(.)}$ é a média dos $\hat{\theta}_{-i}$:

$$\hat{\theta}_{(.)} = \frac{1}{n} \Sigma_{i=1}^{n} \hat{\theta}_{-i} \qquad (5)$$

BOOTSTRAP

O Bootstrap foi desenvolvido por Efron (Efron, 1982) e exige um esforço computacional maior que o Jackknife. A metodologia do Bootstrap para determinar a distribuição empírica de probabilidades de $\hat{\theta}$, assim como para a estimação de seu desvio

padrão, pode ser resumida no seguinte algoritmo:

1 - Faz-se uma reamostragem com reposição das observações da amostra $\underset{\sim}{x}$ formando uma "pseudo-amostra" (ou amostra Bootstrap), x^b_{boot}

$$x^b_{boot} = \{ x^*_1, x^*_2, \ldots, x^*_n \} \qquad (6)$$

2 - A partir da "pseudo-amostra" x^b_{boot}, pode-se calcular a estimativa do parâmetro θ de interesse,

$$\hat{\theta}^b_{boot} = g \ (x^*_1, x^*_2, \ldots, x^*_n) \qquad (7)$$

Repetições independentes dos passos 1 e 2, fornecem $\hat{\theta}^1_{boot}$, $\hat{\theta}^2_{boot}$, ..., $\hat{\theta}^B_{boot}$, estimativas do parâmetro θ com as quais é possível determinar a distribuição empírica de probabilidades de $\hat{\theta}^b_{boot}$.

O estimador Bootstrap do desvio padrão de $\hat{\theta}$ é:

$$\hat{S}_{boot} = \{ \frac{1}{B-1} \Sigma^B_{b=1} (\hat{\theta}^b_{boot} - \hat{\theta}_{boot})^2 \}^{1/2} \qquad (8)$$

onde:

$$\hat{\theta}_{boot} = \frac{1}{B} \Sigma^B_{b=1} \hat{\theta}^b_{boot} \qquad (9)$$

ESTUDOS COMPARATIVOS

A aplicação do Jackknife e do Bootstrap foi avaliada na determinação do desvio padrão da vazão média e na determinação do desvio padrão do tamanho do reservatório para um nível de regularização de 90%.

Estes dois estudos foram realizados considerando dois conjuntos de vazões cada qual constituido de 100 séries de 40 vazões independentes, com distribuição Lognormal, média unitária, e desvio padrão 0.1 e 0.8 respectivamente. Estas vazões não são valores observados em um dado rio, mas sim séries sintéticas, geradas com a finalidade de realizar um estudo totalmente controlado. Estes conjuntos de vazões denominaremos cenário A (desvio padrão 0.1) e cenário B (desvio padrão 0.8).

A avaliação das técnicas, Jackknife e Bootstrap, baseou-se na comparação das variabilidades em torno do "valor verdadeiro", das estimativas obtidas em cada série por estas técnicas com a mesma variabilidade da estimativa obtida com fórmulas da Estatística Clássica. No primeiro estudo estas fórmulas são bastante conhecidas. No segundo foi necessário derivá-las. Aproveitou-se ainda estes estudos para avaliar o uso dos Métodos

Bayesianos para o cálculo de S ($\hat{\theta}$), embora estes não tenham sido desenvolvidos com este objetivo. Maiores detalhes da aplicação dos Métodos Bayesianos nestes casos podem ser encontrados em Costa, 1987.

Vazão Média

Pela Estatística Clássica temos que se X_i i = 1,2,.., são variáveis aleatórias independentes e identicamente distribuidas com desvio padrão σ_x, então o desvio padrão $S(\overline{X})$, é $\sigma_x / \sqrt{n}$.

Se σ_x é conhecido pode-se facilmente determinar $S(\overline{X})$. Para o caso dos cenários A e B temos, $S(\overline{X})$ = 0.01581 e $S(\overline{X})$ = 0.1265, respectivamente.

Em geral, porém, σ_x é quase sempre desconhecido e é estimado pelo correspondente valor amostral obtendo-se então o Estimador Clássico do desvio padrão de $\overline{X}$, $\hat{S}_{class}(\overline{X}) = s_x / \sqrt{n}$, onde s_x é o desvio padrão amostral de X e n o tamanho da série.

O cálculo da estimativa do desvio padrão pelo Método Jackknife, para o caso em que $\hat{\theta} = \overline{X}$ é dado por (Efron, 1982).

$$\hat{S}_{jackk}(\overline{X}) = \left[\left(\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \right]^{1/2} \qquad (10)$$

Pois, $\hat{\theta} = (1/n) \sum_{i=1}^{n} x_i$ e $\hat{\theta}_{-i} = \bar{x}_{-i}$

Então:

$$\hat{\theta}_{-i} = (n\hat{\theta} - x_i) / (n-1) \qquad \text{e}$$

$$\hat{\theta}_{(.)} = (1/n) \sum_{i=1}^{n} \hat{\theta}_{-i} = (1/n) \sum_{i=1}^{n} ((n\hat{\theta} - x_i) /(n-1)) = \hat{\theta} \text{ e}$$

$$\hat{\theta}_{-i} - \hat{\theta}_{(.)} = (n\hat{\theta} - x_i)/(n-1) - \hat{\theta} = (\bar{x} - x_i)/(n-1) \qquad (11)$$

Substituindo a Equação (11) na Equação (4) obtemos a Equação (10). Ou seja, neste caso o Estimador Jackknife do desvio padrão é igual ao Estimador Clássico.

O cálculo do estimador do desvio padrão por Bootstrap, para o caso em que $\hat{\theta} = \overline{X}$ consiste em fazer-se na Equação (7).

$$g(x_1^*, x_2^*, \ldots, x_n^*) = \frac{1}{n} \sum_{i=1}^{n} x_i^* \qquad (12)$$

Considerando os x_i^* como variáveis aleatórias independentes e igualmente distribuidas com distribuição igual a distribuição

empírica de x_i; $g(x_1^*, x_2^*, .., x_n^*)$ é uma variável aleatória cuja variância é dada por:

$$\text{Var}\,[\, g(x_1^*, x_2^*, ..)\,] = n\,\left(\frac{1}{n^2}\,\text{Var}(X_i^*)\right) \qquad (13)$$

Se interpretarmos o Estimador Bootstrap do desvio padrão como o desvio padrão de $g(x_1^*, x_2^*, \ldots, x_n^*)$, tem-se que:

$$\lim_{B\to\infty} \hat{S}_{boot}(\bar{X}) = \left(\frac{1}{n}\,\text{Var}\,[\, X_i^*\,]\right)^{1/2} \qquad (14)$$

Substituindo-se na Equação (14) a fórmula de $\text{Var}\,[\, X_i^*\,]$:

$$\lim_{B\to\infty} \hat{S}_{boot}(\bar{X}) = \frac{s_x}{\sqrt{n}} \cdot \sqrt{\frac{n-1}{n}} \qquad (15)$$

Ou seja, o Estimulador Bootstrap neste caso é proporcional ao Estimador Clássico.

Tamanho do Reservatório

Seja o problema da determinação do tamanho do reservatório que garanta uma vazão defluente igual a 90% da vazão média do histórico durante toda a vida útil do empreendimento.

Uma forma de abordar este problema é analisá-lo do ponto de vista probabilístico, ou seja, dada uma série de vazões anuais, $\underline{x} = (x_1, x_2, \ldots, x_n)$, afluentes a um reservatório inicialmente cheio, a capacidade requerida (K_n) é:

$$K_n = \max_{i=1,\ldots n} \{\, q_i \,\} \qquad (16)$$

$$q_i = \max \{\, 0,\; q_{i-1} + \frac{\alpha}{100}\,\bar{x} - x_i \,\} \quad i = ,..,n \qquad (17)$$

onde q_i é o volume necessário até o ano i, q_o é nulo e $\alpha = 90$.

Séries distintas de afluências resultam segundo as Equações (16) e (17) em diferentes capacidades. Desta forma K_n pode ser encarada como uma variável aleatória, de modo que o projetista pode optar pelo valor esperado de K_n ou pelo quantil a 95% desta distribuição de probabilidades. Gomide (1975) derivou, para afluências independentes e regularização qualquer a distribuição de K_n e seus momentos média e variância.

O método de Gomide exige (em geral) solução numérica ou o uso de distribuições discretas para aproximar distribuições contínuas. Costa (1987), resolveu este mesmo problema pelo

Método de Monte Carlo, obtendo o valor esperado, o desvio padrão e o quantil a 95% ($k_{40}(95\%)$) da distribuição de K_{40} para afluências lognormais e independentes em função do coeficiente de variação destas, para níveis de regularização de 70%, 90%, e 95%. As Figuras 1 e 2 apresentam as respectivas curvas em função do coeficiente de variação. Assim, por exemplo, para o cenário A é possível determinar m_{k40} = 0.11227 e $k_{40}(95\%)$ = 0.2099 através da figura 1. Entretanto, na prática os valores populacionais de média e desvio padrão das afluências não são conhecido e são substituídos pelos valores amostrais e neste caso interessa ao projetista considerar a variabilidade amostral de $\hat{m}_{k40}$ e $\hat{k}_{40}(95\%)$. Costa (1987) determinou ainda por Monte Carlo o desvio padrão de $\hat{m}_{k40}$ e $\hat{k}_{40}(95\%)$. A Figura 3 apresenta as respectivas curvas. Se a média e o desvio padrão das afluências são conhecidos temos para o cenário A $S(\hat{m}_{k40})$ = 0.0286 e $S(\hat{k}_{40}(95\%))$ = 0.0468 e para o cenário B $S(\hat{m}_{k40})$ = 0.910 e $S(\hat{k}_{40}(95\%))$ = 1.680.

Como anteriormente, o projetista não dispõe dos parâmetros populacionais e os estima pelos correspondentes valores amostrais $\bar{x}$ e s_x com os quais através da Figura 3 obtêm-se $\hat{S}_{class}(\hat{m}_{k40})$ e $\hat{S}_{class}(\hat{k}_{40}(95\%))$.

Dada uma série de afluências $x_{i,j}$, a determinação de $\hat{S}_{jackk}(\hat{m}_{k40})$ e $\hat{S}_{jackk}(\hat{k}_{40}(95\%))$ pelo Jackknife é feita conforme o seguinte algoritmo:

1- obter uma pseudo-amostra x^i_{jackk}, omitindo-se o i-ésimo elemento da série de $x_{i,j}$

2- com a pseudo-amostra obtida calcular $\bar{x}^i$ e s^i_x. Usar a Figura 1 para obter $\hat{m}_{k40}$ e $\hat{k}_{40}(95\%)$

3- repetir os itens (1) (2) n vezes

4- determinar $\hat{S}^j_{jackk}(\hat{m}_{k40})$ e $\hat{S}^j_{jackk}(\hat{k}_{40}(95\%))$ Equação (4) a partir dos n $\hat{m}_{k40}$ e $\hat{k}_{40}(95\%)$ obtidos.

A determinação de $\hat{S}_{boot}(\hat{m}_{k40})$ e $\hat{S}_{boot}(\hat{k}_{40}(95\%))$, é feita pelo algoritmo abaixo:

1- obter uma pseudo-amostra x^b_{boot}

2- com x^b_{boot} determinar $\bar{x}^b$ e s^b_x, usar a Figura 1 para obter $\hat{m}_{k40}$ e $\hat{k}_{40}(95\%)$

3- repetir os itens (1) e (2) B vezes (B = 1000)

4- determinar $\hat{S}^j_{boot}(\hat{m}_{k40})$ e $\hat{S}^j_{boot}(\hat{k}_{40}(95\%))$ equação, a partir dos B valores de $\hat{m}_{k40}$ e $\hat{k}_{40}(95\%)$ obtidos.

RESULTADOS

As tabelas 1 a 6 apresentam os resultados obtidos. No primeiro estudo, Vazão Média, observa-se que o Método Clássico subestima o valor verdadeiro. A tendência a subestimação é acentuada no Bootstrap, uma vez que o estimador do desvio padrão por Bootstrap é obtido através da multiplicação do valor obtido pelo Método Clássico por um fator de correção, $\sqrt{(n-1)/n}$, Equação (15). Tendência inversa ocorre na metodologia Bayesiana, onde existe uma superestimação em relação ao valor verdadeiro.

Quanto a variabilidade, medida pelo coeficiente de variação, os métodos Clássico e Bootstrap são equivalentes. O Método Bayesiano no caso do cenário B apresenta variabilidade superior.

Podemos observar ainda nas Tabelas 1 e 2, nas linhas correspondentes ao Bootstrap que os valores obtidos pela Equação (8), para $B = 1000$, e os obtidos para o caso limite $B = \infty$, Equação (15), são equivalentes.

As tabelas 3 a 6 mostram que os métodos apresentam comportamentos similares nos dois casos do segundo estudo. O Método Clássico apresentou uma tendência a subestimação em relação ao valor verdadeiro. Esta tendência é mais acentuada no Bootstrap e no Bayesiano. O Jackknife apresentou comportamento distinto para os cenários A e B. No cenário A o Jackknife apresentou uma tendência a superestimação com relação ao valor verdadeiro, enquanto que para o cenário B a tendência a subestimação apresentada pelos demais métodos foi mantida.

Quanto a variabilidade, o método que apresentou os menores valores foi o Bayesiano, seguido pelo Clássico, Bootstrap e Jackknife.

CONCLUSÃO

A análise dos resultados apresentados no item anterior permite concluir que as Técnicas de Reamostragem experimentadas (Bootstrap e Jackknife) são razoavelmente precisas, podendo ser usadas na construção de intervalos de confiança para parâmetros hidrológicos. A tendência destas técnicas à subestimação do tamanho do intervalo é apenas levemente superior à mesma tendência das fórmulas clássicas. Dentre as duas, a técnica de Bootstrap se revelou levemente superior. Os Métodos Bayesianos não se mostraram competitivos o que pode ser justificado pelo fato de não terem sido desenvolvidos para esta aplicação.

REFERÊNCIAS

ASHKAR, F., BOBÉE, B., e FORTIER, L. (1986), "Confidence Intervals for Design Flood Events under Different Stastical Flood Models", International Symposium on Flood Frequency and Risk Analyses, Baton Rouge, Louisiana.

COSTA, F.S. (1987), "Aplicações de Técnicas Estatísticas de

Reamostragem em Hidrologia", Tese de mestrado a ser submetida a COPPE/UFRJ.

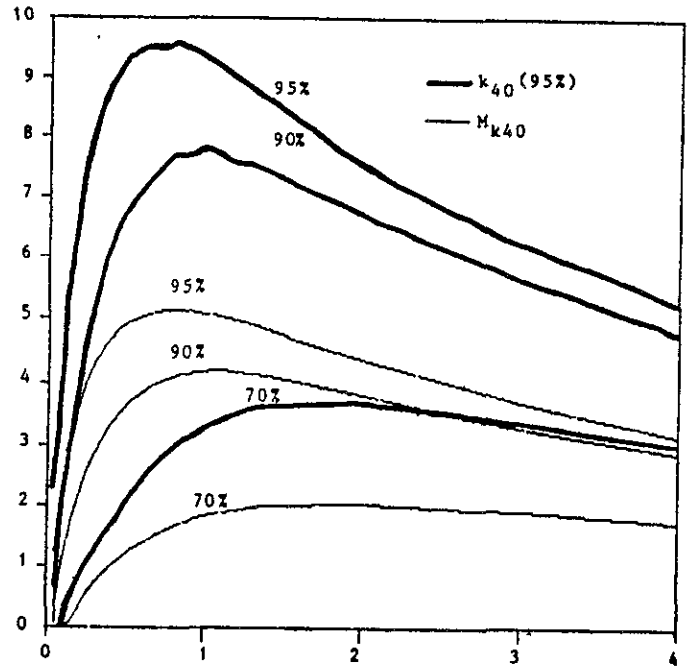
EFRON, B. (1982), The Jackknife, the Bootstrap and Other Resampling Plans, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

GOMIDE, F.L.S. (1975), "Range and Deficit Analysis Using Markov Chains", Hydrology Papers, Colorado State University, Fort Collins, Colorado.

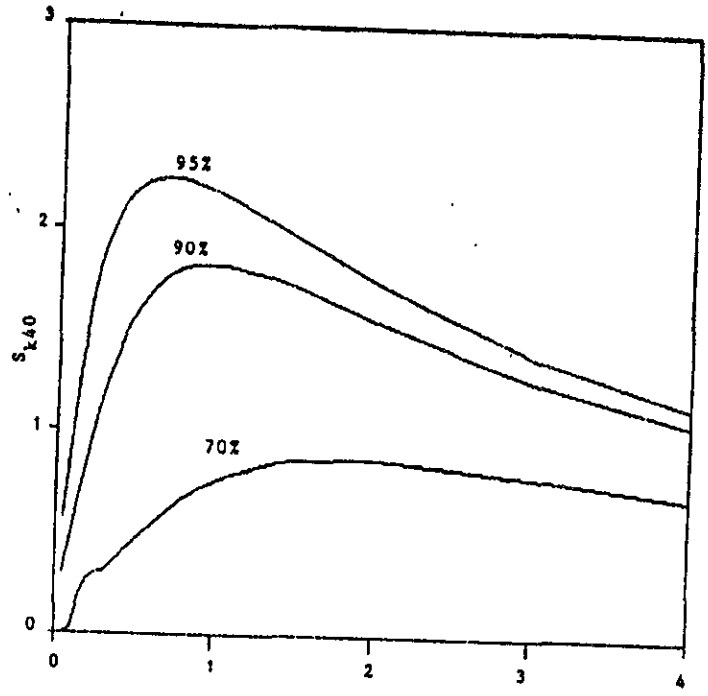
QUENOUILLE, M.H. (1949), "Approximate Tests of Correlation in Times Series", J. Roy. Statist. Soc. Ser. B, 11, pag.13-84.

TUKEY, J. W. (1958), "Bias and Confidence en Not Quite Large Sample", Ann. Math Statist., 29, pag 614.

USWRC (1977) Guidelines for Determining Flood Frequency, Bulletin 17A of the Hydrology Committe.

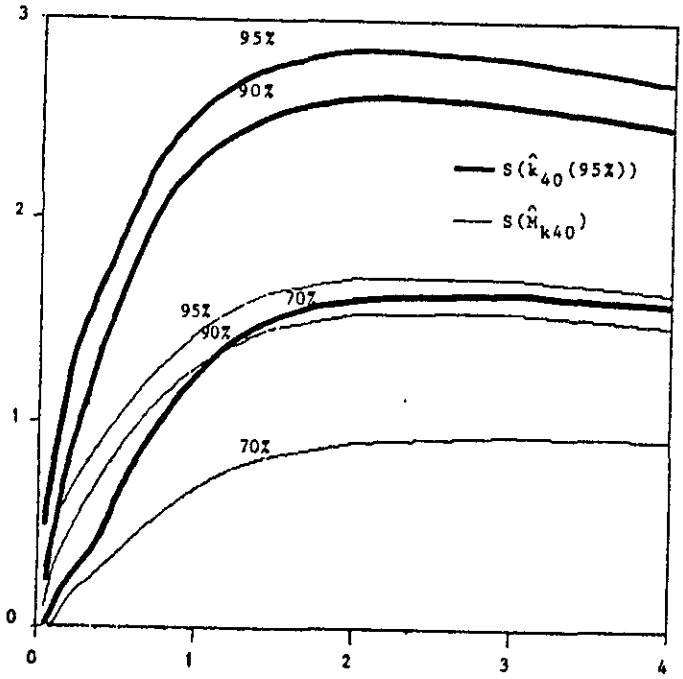


"Figura 1 - Curvas k_{40}(95%) x CV e M_{k40} x CV. Afluências Log-normais, níveis de regularização 95%, 90% e 70%. Valores em unidade do desvio padrão das afluências".



CV - (Coeficiente de Variação das Afluências)

"Figura 2 - Curvas S_{k40} x CV. Afluência Lognormais, níveis de regularização 95%, 90% e 70%. Valores em unidade de desvio padrão das afluências".



CV - (Coeficiente de Variação das Afluências)

"Figura 3 - Curvas $S(\hat{k}_{40}(95\%))$ x CV e $S(\hat{M}_{k40})$ x CV. Afluências Lognormais, níveis de regularização 95%, 90% e 70%. Valores em unidade de desvio padrão das afluências".

"Tabela 1 - Comparação dos Métodos. Parâmetro μ. Cenário A. (EMQ = erro médio quadrático)".

Método	Média	Desvio Padrão	Coef. Variação	$EMQ^{1/2}$
Clássico e Jackknife	15.76×10^{-3}	1.75×10^{-3}	0.111	1.75×10^{-3}
Bayesiano	16.34×10^{-3}	1.85×10^{-3}	0.113	1.92×10^{-3}
Bootstrap $B \to \infty$	15.37×10^{-3}	1.71×10^{-3}	0.111	1.77×10^{-3}
Bootstrap $B = 1000$	15.64×10^{-3}	1.81×10^{-3}	0.115	1.82×10^{-3}
Valor Verdadeiro	15.81×10^{-3}			

"Tabela 2 - Comparação dos Métodos. Parâmetro μ. Cenário B. (EMQ = erro médio quadrático)".

Método	Média	Desvio Padrão	Coef. Variação	$EMQ^{1/2}$
Clássico e Jackknife	0.1205	0.0325	0.269	0.0330
Bayesiano	0.1428	0.0329	0.231	0.0367
Bootstrap $B \to \infty$	0.1175	0.0317	0.269	0.0330
Bootstrap $B = 1000$	0.1198	0.0326	0.272	0.0333
Valor Verdadeiro	0.1265			

"Tabela 3 - Comparação dos Métodos. Parâmetro M_{k40}. Cenário A. (EMQ = erro médio quadrático)".

Método	Média	Desvio Padrão	Coef. Variação	$EMQ^{1/2}$
Clássico	0.0279	0.0060	0.2151	0.0060
Bayesiano	0.0117	0.0023	0.1966	0.0171
Jackknife	0.0295	0.0068	0.2305	0.0069
Bootstrap	0.0276	0.0063	0.2228	0.0064
Valor Verdadeiro	0.0286			

"Tabela 4 - Comparação dos Métodos. Parâmetro M_{k40}. Cenário B. (EMQ = erro médio quadrático)".

Método	Média	Desvio Padrão	Coef. Variação	$EMQ^{1/2}$
Clássico	0.0463	0.0095	0.0206	0.0095
Bayesiano	0.0328	0.0054	0.1638	0.0150
Jackknife	0.0482	0.0113	0.2344	0.0114
Bootstrap	0.0453	0.0106	0.2340	0.0107
Valor Verdadeiro	0.0468			

"Tabela 5 - Comparação dos Métodos. Parâmetro $K_{40}(95\%)$. Cenário A (EMQ = erro médio quadrático)".

Método	Média	Desvio Padrão	Coef. Variação	$EMQ^{1/2}$
Clássico	0.854	0.317	0.371	0.322
Bayesiano	0.433	0.131	0.303	0.495
Jackknife	0.745	0.488	0.655	0.515
Bootstrap	0.639	0.364	0.570	0.454
Valor Verdadeiro	0.910			

"Tabela 6 - Comparação dos Métodos. Parâmetro $K_{40}(95\%)$. Cenário B (EMQ = erro médio quadrático)".

Método	Média	Desvio Padrão	Coef. Variação	$EMQ^{1/2}$
Clássico	1.566	0.557	0.356	0.569
Bayesiano	1.023	0.326	0.319	0.733
Jackknife	1.387	0.894	0.645	0.941
Bootstrap	1.183	0.660	0.558	0.826
Valor Verdadeiro	1.680			

STANDARD DEVIATION OF HYDROLOGICAL PARAMETER'S ESTIMATORS CALCULATION

by

Fernanda da Serra Costa, Jerson Kelman, Jorge M. Damázio

ABSTRACT -- Controlled experiences to access the efficiencies of calculation methods for standard deviations of hydrological parameters estimators are made. The Resampling Technics tested are shown to be precise and useful in the development of confidence intervals.