Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Cientifica
Departamento de Estatística

Relatório

Hugo Calegari RA:155738
Leonardo Uchoa Pedreira RA:156231

Professor: Verônica

Campinas-SP, 29 de Junho de 2017

Introdução Bootstrap

O método de bootstrap faz parte de uma classe de métodos não-paramétricos de Monte Carlo que estimam a distribuição de uma população ou uma característica (parâmetro de interesse) por meio de reamostragem.

Métodos de reamostragem consideram as amostras como uma população finita, a partir da qual reamostras são tomadas para estimar características e realizar inferências a respeito desta população.

Inferências baseadas em percentis pelo método de bootstrap

Ao se comparar dois grupos independentes, o método é aplicado como segue. Gera-se amostra por bootstrap para cada grupo:

- Para o j-ésimo grupo, obter amostras de bootstrap via amostragem aleatória com reposição (n_j) da seguinte amostra: $X_{1j}, \dots, X_{nj}$, para obter a seguinte: $X^*_{1j}, \dots, X^*_{nj}$;

Seja $\hat{\theta}^*_j$ a estimativa por bootstrap de θ_j, tal que este parâmetro está associado com alguma característica de interesse. Seja, ainda, $D^* = \hat{\theta}^*_1 - \hat{\theta}^*_2$. Ao se repetir este processo B vezes (quantidade de réplicas) gera-se $D^*_1, \dots, D^*_B$. Defina $l = \frac{\alpha}{2}B$ (determinação do limite inferior do intervalo de confiança), arredonde para o inteiro mais próximo, e $u = B - l$ (limite superior). Com isso, um intervalo de confiança aproximado de $1-\alpha$ para a diferença entre os verdadeiros parâmetros ($\theta_1 - \theta_2$) é: $[D^*_{(l+1)}, D^*_{(u)}]$, em que $D^*_{(1)} \leq \dots \leq D^*_{(B)}$.

Uma vez que se quer testar a hipótese: $H_0 : \theta_1 = \theta_2$, pode-se utilizar as seguintes estruturas de acordo com o que segue. Para as estimativas de bootstrap de $\hat{\theta}^*_1$ e $\hat{\theta}^*_2$, seja $p^* = P(\hat{\theta}^*_1 > \hat{\theta}^*_2)$ (pode-se estimar esta probabilidade com o uso da proporção de $\hat{\theta}^*_1 > \hat{\theta}^*_2$).

Sob a hipótese nula (igualdade dos verdadeiros parâmetros), assintoticamente (para n e B suficientemente grandes), p^* tem distribuição uniforme. Assim, rejeita-se H_0 se $p^* \leq \alpha/2$ ou se $p^* \geq 1-\alpha/2$. Neste caso, a forma como foi estimado o valor de p^* é:

- Seja A número de valores que são maiores que zero para todos os valores das diferenças obtidos via bootstrap, isto é, entre os valores $D^*_1, \dots, D^*_B$. Consequentemente, pode-se estabelecer: $p^* = A/B$.

Por conveniência é adotado a o seguinte valor de p estimado: $p^*_m = \min(p^*, 1-p^*)$ (chamado de p-valor generalizado). Com isso, rejeita-se H_0 se $p^*_m \leq \alpha/2$.

Comparação de M-estimadores

Os M-estimadores que serão avaliados são os de locação. Quando se compara estes estimadores com dois grupos independentes, ainda se percebe que a inferência baseada nos percentis por meio do método de bootstrap é o melhor método. Um intervalo de confiança baseado na estimativa do erro padrão fornecerá boa probabilidade de cobertura quando o tamanho amostral é suficientemente grande, ou seja, para se ter razoável aproximação do erro padrão necessita-se de uma população para reamostragem (amostra) relativamente grande, para que características da variabilidade populacional seja captada. A boa cobertura também depende da suposição de que as diferenças etimadas são normalmente distribuídas, porém, é desconhecido o quão grande é o tamanho amostral deveria ser antes de que a aproximação seja considerada, particularmente quando a distribuição é assimétrica.

Quando os tamanhos amostrais são pequenos, todas as indicações são de que o método de percentil por bootstrap é o melhor, então este é recomendado, até existir boa evidência de que algum outro método possa ser utilizado em seu lugar.

Nota-se que com o objetivo de comparar dois M-estimadores, de dois grupos independentes, precisa-se a cada replicação obter uma estimatica do parâmetro de interesse. Com isso, será utilizado algum algoritmo, como M.P.I. (médias ponderadas iteradas), M.P.V.I. (média de pseudovalores iterados) ou N.R. (Newton Raphson) .

Comparação de média aparadas e medianas

Quando se compara médias aparadas e se tem pelo menos 20% dos dados desconsiderados para o seu cáculo, inferências baseadas no percentil pelo método de bootstrap é preferível quando comparado com o método de bootstrap-t (quando se utiliza a distribuição t de Student para determinar o valor crítico apropriado). A acurácia para o método de bootstrap-t é maior quando a quantidade de dados desconsiderados é pequena, mas há incertezas a respeito desse valor.

Para o caso no qual o objetivo é comparar as medianas, uma pequena mudança deve ser feita para quando se tem dados repetidos. Seja M_1^* e M_2^* medianas amostrais por bootstrap e $p^* = P(M_1^* > M_2^*) + 0,5P(M_1^* = M_2^*)$. De maneira semelhante ao que foi determinado anteriormente para p^*, entre B amostras de bootstrap se A é o número de vezes em que $M_1^* > M_2^*$, e C é o número de vezes em que $M_1^* = M_2^*$, uma estimativa para p^* é: $p^* = \frac{A}{B} + 0,5\frac{C}{B}$. Assim, o p-valor é definido como $2\min(p^*, 1-p^*)$.

Em termos de controle da probabilidade do erro do tipo 1, as indicações até o momento são de que o método de inferência baseada em percentil por bootstrap tem um bom desempenho independente de existir dados repetidos.

Com a dúvida levantada a respeito da precisão do método de bootstrap-t, de acordo com Keselman et al. (2004), este tem uma performance razoável quando se desconsidera um quantidade de 10% e 15% dos dados.

Por exemplo, considere a situação na qual se tem duas amostra definidas como segue: $n_1 = 40$ observações de uma amostra de distribuição normal padrão e $n_2 = 20$ observações de uma amostra de distribuição lognormal deslocada, tal que a média aparada tenha seja zero. Quando se testa a diferença entre os valores das médias, com nível de significância de $0,05$, e 10% da informação amostral retirada, observa-se que o nível verdadeiro para o método bootstrap-t é $0,066$ comparado com 0.050 para o método de percentil por bootstrap (ao usar 1000 réplicas). Ao se reduzir o tamanho amostral $n_1 = 20$ e $n_2 = 10$ as estimativas do nível de significância verdadeiros são: 0.082 e 0.074 para os métodos de bootstrap-t e o de percentil por bootstrap. Agora, para o último tamanho amostral fixo, e uma quantidade de informação retirada de 20% (isto é, nível de apara de 20%), as estimativas dos níves são 0.081 e 0.063, para o método de bootstrap-t e o percentil via bootstrap. Isto nos indica que o controle que se tem da probabilidade do erro do tipo 1 ao utilizar o método de percentil por bootstrap é maior quando comparado com o método de bootstrap-t. Este controle é maior ainda ao se considerar uma porcentagem de informação desconsiderda (nível de apara).