Exercício 1: *Bootstrap*

Fernando Mayer

2018-03-19

Introdução

O *bootstrap* é uma técnica de estimação de parâmetros desconhecidos de uma população, baseada em amostragem aleatória com reposição. A ideia é relativamente simples: com a amostra observada (de tamanho n), fazemos um grande número (r) de reamostragens (e.g. 1000) com reposição de tamanho $m \leq n$, e calculamos a estatística de interesse. Dessa forma, teremos r estimativas diferentes da estatística que temos interesse, e a partir disso, podemos obter a distribuição amostral dessa estatística e outras medidas pontuais (e.g. média, mediana), e medidas de variação (e.g. variância, intervalos de confiança). Para mais informações (gerais) veja: https://en.wikipedia.org/wiki/Bootstrapping_(statistics).

Objetivos

O objetivo deste exercício é programar um *bootstrap* para a estimativa da média de uma população simulada. O objetivo é programar o método sem ser rigoroso com os detalhes teóricos.

Procedimento

Copie o código abaixo para um script do R, substituindo a *string* `"<seu_numero_de_matricula>"` pelo número de sua matrícula na UFPR (numérico). Esse número será utilizado como semente para gerar a sua população de 1 milhão de números simulados a partir de uma distribuição Normal (objeto `pop`). Depois, são amostrados 1000 elementos dessa população para gerar a sua amostra (`amostra`).

```
## Insira aqui o número da sua matrícula para fixar uma semente
matricula <- "<seu_numero_de_matricula>"
## Gera 1 milhão de números aleatórios de uma distribuição normal
set.seed(matricula)
pop <- rnorm(n = 1e6, mean = 100, sd = sqrt(200))
## Retira uma amostra aleatória de tamanho n = 1000 da população
amostra <- pop[sample(1:length(pop), size = 1000)]
```

Note que, por ser um exercício de simulação, você conhece o verdadeiro valor da média populacional (a média de `pop`). A média da amostra já é uma estimativa da verdadeira média populacional. Na prática essa seria a única estimativa que teríamos, e toda inferência seria baseada nessa amostra. A ideia do *bootstrap* é então obter a distribuição amostral dessa estimativa e estudar seu comportamento.

A seguir serão apresentados os pseudo-códigos para dois exercícios de *bootstrap*. O primeiro faz as reamostragens para um tamanho de amostra m fixo, enquanto que o segundo permite variar esse número para avaliar o impacto desse número nas esimativas.

Para ambos exercícios, você deve fazer/criar:

- Um objeto com a classe e dimensão apropriados para armazenar os resultados de cada reamostragem.
- Uma função que calcule a diferença absoluta entre dois números, e usá-la para calcular as diferenças entre a média da população (verdadeira) e as médias obtidas via *bootstrap*.
- Um histograma das r estimativas (para cada valor de m no exercício 2), com uma linha vertical indicando a média populacional (verdadeira).

Exercício 1

Algoritmo geral para desenvolver o método, considerando o tamanho da amostra (m) fixo:

1. Com os dados da amostra, gere uma nova amostra aleatória (**com reposição**) de tamanho $m = 500$.
2. Calcule a média dessa nova amostra.
3. Repita esse procedimento $r = 100000$ vezes.
4. Faça um histograma das r estimativas, calcule a média e compare com a média verdadeira.

Exercício 2

Algoritmo geral para desenvolver o método, considerando o tamanho da amostra (m) variando entre quatro valores diferentes:

1. Com os dados da amostra, gere uma nova amostra aleatória (**com reposição**) com tamanhos: $m = 100, 300, 500, 700$.
2. Calcule a média dessa nova amostra, para cada valor de m.
3. Repita esse procedimento $r = 100000$ vezes, para cada valor de m.
4. Faça um histograma das r estimativas, calcule a média e compare com a média verdadeira (para cada valor de m).