Bootstrap

Bootstrap

Ferreira, 2013:

- Um dos principais temas envolvendo os métodos computacionalmente intensivos e, talvez, o maior responsável pela popularidade desses métodos é a técnica bootstrap;
- o método bootstrap envolve reamostragens com reposição da amostra original (bootstrap não-paramétrico);
- inspirado na teoria da amostragem, o bootstrap é utilizado para a realização de testes de hipóteses, estimação de parâmetros por intervalos e estimação de erros padrões;
- o bootstrap é útil para realizar inferências quando não conhecemos a distribuição de probabilidade da população e o modelo normal não é adequado para os dados ou resíduos;

Bootstrap não-paramétrico

Ferreira, 2013:

- No bootstrap não-paramétrico a ideia é realizar reamostragem da amostra original;
- supondo que tenhamos uma amostra aleatória de tamanho n, $Y_1, Y_2, \cdots, Y_n$, de uma população com distribuição desconhecida e cujo interesse é um parâmetro θ, podemos obter uma série de reamostragem com reposição de tamanho n e estimar o parâmetro de interesse por uma função $\hat{\theta}$ das amostras obtidas;
- essa série de estimativas obtidas é finita e representa uma amostra da distribuição do estimador, permitindo que possamos fazer inferência sobre θ;

Bootstrap não-paramétrico

Ferreira, 2013:

- se pudéssemos obter a distribuição de amostragem de $\hat{\theta}$ diretamente da teoria de amostragem, poderíamos fazer inferências sobre θ sem a necessidade de utilizar bootstrap. Para isso, devemos conhecer a distribuição de probabilidade de Y_j, $j = 1, 2, \cdots, n$ e, a partir dela, deduzirmos a distribuição de amostragem da função de interesse $\hat{\theta}$;

- nem sempre isso é possível, principalmente se não conhecemos a distribuição de probabilidade da variável aleatória que estamos amostrando;

Bootstrap não-paramétrico

Ferreira, 2013:

- para a maior parte dos modelos probabilísticos, teorias exatas não são conhecidas ou são intratáveis analiticamente. Nesse caso, usamos a distribuição de probabilidade empírica, pela qual atribuímos a cada uma das observações amostrais a probabilidade $1/n$;
- a ideia do bootstrap é substituir a distribuição desconhecida da população pela distribuição empírica. Por essa razão, que denominamos esse método de bootstrap não-paramétrico;
- o bootstrap não-paramétrico baseia-se no fato de que a amostra obtida da população contém toda a informação dessa população subjacente. Então, ela passa a representar a "população" para o processo de reamostragem.

Bootstrap não-paramétrico

Ferreira, 2013:

- o algoritmo geral das reamostragens bootstrap, para estimar um parâmetro θ desconhecido, cuja distribuição de probabilidade é $f(x;\theta)$, considerando, ainda, que o estimador $\hat{\theta}$ é uma função dos valores amostrais, ou seja, $\hat{\theta} = \hat{\theta}(Y_1, Y_2, \cdots, Y_n)$, é dado por:
    1. atribuir massa de probabilidade $1/n$ para cada observação amostral Y_j, $j = 1, 2, \cdots, n$, criando a distribuição de probabilidade empírica $\hat{f}(x;\theta)$;
    2. gerar amostras aleatórias com reposição da distribuição de probabilidade empírica $\hat{f}(x;\theta)$, denominada de amostra de bootstrap e dada por $\tilde{Y}_1, \tilde{Y}_2, \cdots, \tilde{Y}_n$;
    3. calcular uma estimativa de $\hat{\theta}$ por $\tilde{\theta}$, usando a amostra de bootstrap no lugar da amostra original, da seguinte forma $\tilde{\theta} = \tilde{\theta}\left(\tilde{Y}_1, \tilde{Y}_2, \cdots, \tilde{Y}_n\right)$;
    4. repetir os passos 2 e 3 B vezes.

Bootstrap não-paramétrico

Estimação (Ferreira, 2013):

- Devemos entender que os métodos bootstrap não melhoram as estimativas pontuais, mas fornecem mecanismos apropriados para estimar os erros padrões, intervalos de confiança e a distribuição de amostragem do estimador, por mais complexo que seja esse estimador;

- para a estimação do erro padrão de um estimador $\hat{\theta}$ de um parâmetro θ, devemos obter a distribuição bootstrap desse estimador, conforme descrito no algoritmo anteriormente apresentado;

Bootstrap não-paramétrico

Estimação (Ferreira, 2013):

- Se considerarmos uma amostra aleatória $Y_1, Y_2, \cdots, Y_n$, de tamanho n da população de interesse, que depende de um parâmetro θ, que pretendemos estimar, então aplicamos o algoritmo anterior e obtemos uma amostra bootstrap de tamanho B da distribuição desse estimador;
- se a amostra for $\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3, \cdots, \tilde{\theta}_B$, podemos estimar o erro padrão de $\hat{\theta}$ por

$$S_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{j=1}^{B}\left(\tilde{\theta}_j - \bar{\tilde{\theta}}\right)^2} = \sqrt{\frac{1}{B-1}\left[\sum_{j=1}^{B}\tilde{\theta}_j^2 - \frac{\left(\sum_{j=1}^{B}\tilde{\theta}_j\right)^2}{B}\right]}$$

Bootstrap não-paramétrico

Exercício:

Considerando o método bootstrap **não-paramétrico** e os dados:

4	5	8	0	5	0	64	13	10	18
4	13	4	1	17	54	19	4	12	11
15	6	12	7	10	28	3	2	17	5
9	7	2	1	4	12	10	33	2	4
0	37	15	9	7	0	2	4	24	16

Obtenha:
(i) histograma dos dados,
(ii) estimativa da média,
(iii) distribuição de bootstrap (histograma),
(iv) erro padrão de bootstrap,

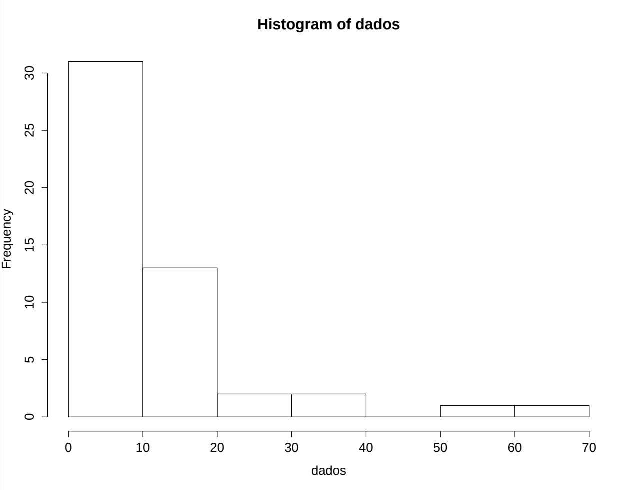
Bootstrap não-paramétrico

Exercício:

```
> dados <- c(4, 5, 8, 0, 5, 0, 64, 13, 10, 18,
+            4, 13, 4, 1, 17, 54, 19, 4, 12, 11,
+            15, 6, 12, 7, 10, 28, 3, 2, 17, 5,
+            9, 7, 2, 1, 4, 12, 10, 33, 2, 4,
+            0, 37, 15, 9, 7, 0, 2, 4, 24, 16)
> # (i)
> hist(dados)
```

Bootstrap não-paramétrico

Exercício:



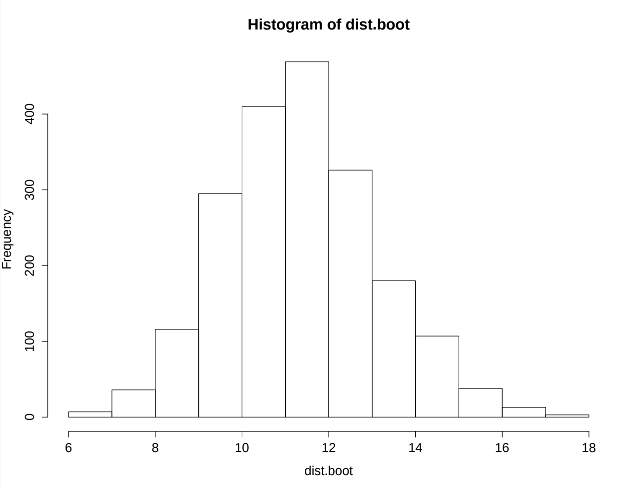
Bootstrap não-paramétrico

Exercício:

```
> # (ii)
> mean(dados)
[1] 11.38
> # (iii)
> # bootstrap não-paramétrico
> B <- 2000
> dist.boot <- NULL
> for(i in 1:B){
+   amostra <- sample(dados, replace = TRUE)
+   dist.boot <- c(dist.boot, mean(amostra))
+ }
> hist(dist.boot)
> # (iv)
> sd(dist.boot)
[1] 1.810682
```

Bootstrap não-paramétrico

Exercício:



Bootstrap

Correção de viés (Ferreira, 2013):

- devemos explorar no bootstrap a possibilidade de correção de viés para estimadores viesados;
- sabemos da teoria clássica que o viés δ de um estimador $\hat{\theta}$ é definido por
$$\delta_{\hat{\theta}} = E(\hat{\theta}) - \theta$$
- como na distribuição de bootstrap de um estimador $\hat{\theta}$, a média amostral representa a esperança de $\hat{\theta}$, e a estimativa de $\hat{\theta}$ na amostra original, faz o mesmo papel de θ na população, podemos estimar o viés por
$$\tilde{\delta}_{\hat{\theta}} = \bar{\hat{\theta}} - \hat{\theta}.$$
- É comum pensarmos que $\bar{\hat{\theta}}$ seria o estimador corrigido para viés, o que não é verdade.

Bootstrap

Correção de viés (Ferreira, 2013):

- Podemos obter um estimador ajustado por

$$\hat{\theta}_{aj.} = \hat{\theta} - \tilde{\delta}_{\hat{\theta}} = \hat{\theta} - \bar{\tilde{\theta}} + \hat{\theta} = 2\hat{\theta} - \bar{\tilde{\theta}}$$

$$\therefore \quad \hat{\theta}_{aj.} = 2\hat{\theta} - \bar{\tilde{\theta}}$$

Bootstrap

Intervalo de Confiança Padrão de Bootstrap (Ferreira, 2013):

- Em algumas ocasiões, podemos assumir que a distribuição de bootstrap do estimador $\tilde{\theta}$ é normal;
- A ideia é admitir o efeito do teorema do limite central:

$$\tilde{\theta} \overset{a}{\sim} N\left(\hat{\theta}, S^2_{\hat{\theta}}\right)$$

- podemos construir o intervalo bootstrap padrão com confiança de aproximadamente $1-\alpha$ por

$$IC_{1-\alpha}(\theta) : \left[\hat{\theta} - z_{1-\alpha/2} S_{\hat{\theta}}; \hat{\theta} + z_{1-\alpha/2} S_{\hat{\theta}}\right]$$

  em que $z_{1-\alpha/2}$ é o $1-\alpha/2$ quantil da distribuição normal padrão.
- infelizmente, a experiência tem mostrado que esse intervalo não é bom.

IC Baseado em Percentis Bootstrap (Ferreira, 2013):

- Assumir normalidade da distribuição bootstrap é subutilizar o potencial do método;
- Para aplica o intervalo de confiança baseado em percentis bootstrap, devemos obter $\tilde{\theta}_1, \tilde{\theta}_2, \cdots, \tilde{\theta}_B$;
- ordenar esses valores, obtendo as estatísticas de ordem por: $\tilde{\theta}_{(1)}, \tilde{\theta}_{(2)}, \cdots, \tilde{\theta}_{(B)}$;
- obter os percentis $100(\alpha/2)\%$ e $100(1-\alpha/2)\%$;
- Assim, construímos o intervalo de confiança por

$$IC_{1-\alpha}(\theta) : \left[\tilde{\theta}_{(k_1)}; \tilde{\theta}_{(k_2)}\right]$$

em que $k_1 = \lfloor (B+1)(\alpha/2) \rfloor$ e $k_2 = \lfloor (B+1)(1-\alpha/2) \rfloor$
- Esse intervalo possui propriedade melhores que o anterior e funciona bem na maioria dos casos mais simples.

Bootstrap

IC Básico de Bootstrap (Ferreira, 2013):

- Variação do intervalo baseado em percentis bootstrap.
- Construímos o IC básico de bootstrap com confiança de aproximadamente $1-\alpha$ por

$$IC_{1-\alpha}\left(\theta\right):\left[2\hat{\theta}-\tilde{\theta}_{(k_2)};2\hat{\theta}-\tilde{\theta}_{(k_1)}\right]$$

em que $k_1 = \lfloor (B+1)\left(\alpha/2\right) \rfloor$ e $k_2 = \lfloor (B+1)\left(1-\alpha/2\right) \rfloor$

Outros IC de Bootstrap:

- Intervalo de Confiança t de Bootstrap;
- Intervalo de Confiança Bootstrap com Correção de Viés Acelerado (BC_a)
- Intervalo de Confiança Bootstrap com Correção de Viés
- Detalhes: Ferreira, (2013).

Exercício:

Considerando o método bootstrap **não-paramétrico** e os dados:

4	5	8	0	5	0	64	13	10	18
4	13	4	1	17	54	19	4	12	11
15	6	12	7	10	28	3	2	17	5
9	7	2	1	4	12	10	33	2	4
0	37	15	9	7	0	2	4	24	16

Obtenha:

(v) estimativa da média com viés corrigido,

(vi) IC básico, percentílico e padrão de bootstrap,

(vii) as amplitudes dos IC's. Qual é o intervalo de menor amplitude?

Exercício:

```
> # Não-Paramétrico
> est.boot <- mean(dados) # a estimativa pontual de bootstrap é baseada na amostra
> est.boot
[1] 11.38
> # Estimativa ajustada, correção de viés
> est.boot.aj <- 2*est.boot - mean(dist.boot)
> est.boot.aj
[1] 11.39757
> # Intervalo de Confiança Padrão de Bootstrap:
> s <- sd(dist.boot) # Erro padrao de Bootstrap
> alpha <- 0.05
> z <- qnorm(1 - alpha/2)
> IC.padrao <- c(est.boot - z*s, est.boot + z*s)
> IC.padrao
[1]  7.807069 14.952931
```

Exercício:

```
> # IC  percentílico de Bootstrap:
> alpha <- 0.05
> z <- qnorm(1 - alpha/2)
> dist.boot.or <- sort(dist.boot)
> k1 <- trunc((B+1)*(alpha/2))
> k2 <- trunc((B+1)*(1 - alpha/2))
> IC.perc <- c(dist.boot.or[k1], dist.boot.or[k2])
> IC.perc
[1]  8.02 15.12
> # IC  Básico de Bootstrap:
> alpha <- 0.05
> z <- qnorm(1 - alpha/2)
> dist.boot.or <- sort(dist.boot)
> k1 <- trunc((B+1)*(alpha/2))
> k2 <- trunc((B+1)*(1 - alpha/2))
> IC.basico <- c(2 * est.boot - dist.boot.or[k2], 2 * est.boot - dist.boot.or[k1])
> IC.basico
[1]  7.64 14.74
> # Amplitude dos IC's
> IC.padrao[2] - IC.padrao[1]
[1] 7.145863
> IC.perc[2] - IC.perc[1]
[1] 7.1
> IC.basico[2] - IC.basico[1]
[1] 7.1
```

Bootstrap Paramétrico

Ferreira, 2013:

- Se conhecemos a distribuição da variável aleatória que está sendo amostrada, por que utilizar um procedimento (computacional) bootstrap paramétrico?

Bootstrap Paramétrico

Ferreira, 2013:

- Se conhecemos a distribuição da variável aleatória que está sendo amostrada, por que utilizar um procedimento (computacional) bootstrap paramétrico?
  - Conhecer a distribuição de probabilidade da variável aleatória não implica em conhecer a distribuição de amostragem do estimador.
  - Muitas vezes, a dedução teórica da distribuição de um estimador $\hat{\theta}$ não é uma tarefa simples ou exequível analiticamente.
  - Nessas circunstâncias, a utilização dos métodos bootstrap paramétricos se justifica.

Bootstrap Paramétrico

Ferreira, 2013:

- No bootstrap paramétrico é assumido como conhecida a distribuição da variável aleatória, mas não seus parâmetros.

- Devemos estimar os parâmetros da distribuição a partir da amostra aleatória disponível e utilizarmos as estimativas como parâmetros da função densidade correspondente e, assim, gerarmos dados da distribuição de interesse.

Bootstrap paramétrico

Exercício:

Considerando o método bootstrap **paramétrico** e os dados:

4	5	8	0	5	0	64	13	10	18
4	13	4	1	17	54	19	4	12	11
15	6	12	7	10	28	3	2	17	5
9	7	2	1	4	12	10	33	2	4
0	37	15	9	7	0	2	4	24	16

Obtenha:
(i) histograma dos dados,
(ii) estimativa da média,
(iii) distribuição de bootstrap (histograma), considerando a distribuição exponencial,
(iv) erro padrão de bootstrap,

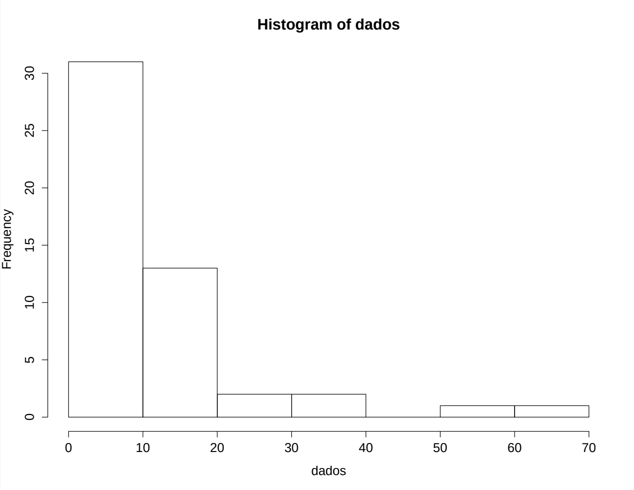
Bootstrap paramétrico

Exercício:

```
> dados <- c(4, 5, 8, 0, 5, 0, 64, 13, 10, 18,
+            4, 13, 4, 1, 17, 54, 19, 4, 12, 11,
+            15, 6, 12, 7, 10, 28, 3, 2, 17, 5,
+            9, 7, 2, 1, 4, 12, 10, 33, 2, 4,
+            0, 37, 15, 9, 7, 0, 2, 4, 24, 16)
> # (i)
> hist(dados)
```

Bootstrap paramétrico

Exercício:



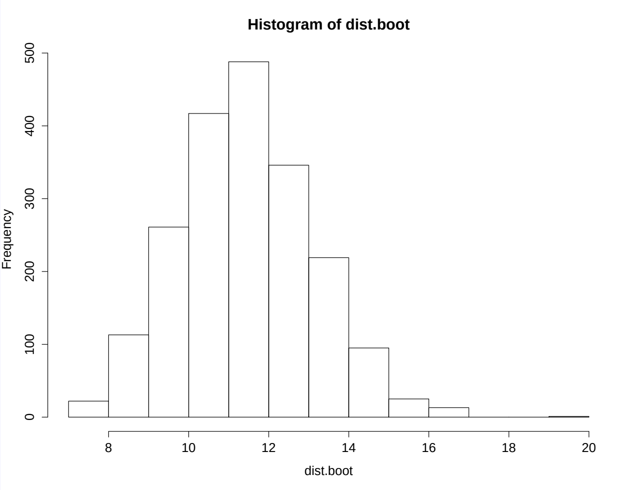
Bootstrap paramétrico

Exercício:

```
> # (ii)
> mean(dados)
[1] 11.38
> # (iii)
> # bootstrap paramétrico (supondo ~ Exp(lambda))
> n <- length(dados)
> lambda <- 1/mean(dados)
> B <- 2000
> dist.boot <- NULL
> for(i in 1:B){
+   amostra <- rexp(n, lambda)
+   dist.boot <- c(dist.boot, mean(amostra))
+ }
> hist(dist.boot)
> # (iv)
> sd(dist.boot)
[1] 1.590881
```

Bootstrap paramétrico

Exercício:



Exercício:

Considerando o método bootstrap **paramétrico** (distribuição exponencial) e os dados:

4	5	8	0	5	0	64	13	10	18
4	13	4	1	17	54	19	4	12	11
15	6	12	7	10	28	3	2	17	5
9	7	2	1	4	12	10	33	2	4
0	37	15	9	7	0	2	4	24	16

Obtenha:

(v) estimativa da média com viés corrigido,

(vi) IC básico, percentílico e padrão de bootstrap,

(vii) as amplitudes dos IC's. Qual é o intervalo de menor amplitude?

Exercício:

```
> # Paramétrico
> est.boot <- mean(dados) # a estimativa pontual de bootstrap é baseada na amostra
> est.boot
[1] 11.38
> n <- length(dados)
> lambda <- 1/mean(dados)
> B <- 20000
> dist.boot.par <- NULL
> for(i in 1:B){
+   amostra <- rexp(n, lambda)
+   dist.boot.par <- c(dist.boot.par, mean(amostra))
+ }
> s <- sd(dist.boot.par) # Erro padrao de Bootstrap
> # Estimativa ajustada, correção de viés
> est.boot.aj <- 2*est.boot - mean(dist.boot.par)
> est.boot.aj
[1] 11.36489
> # Intervalo de Confiança Padrão de Bootstrap:
> alpha <- 0.05
> z <- qnorm(1 - alpha/2)
> IC.padrao.par <- c(est.boot - z*s, est.boot + z*s)
> IC.padrao.par
[1]  8.233265 14.526735
```

Exercício:

```
> # IC  percentílico de Bootstrap:
> alpha <- 0.05
> z <- qnorm(1 - alpha/2)
> dist.boot.or <- sort(dist.boot.par)
> k1 <- trunc((B+1)*(alpha/2))
> k2 <- trunc((B+1)*(1 - alpha/2))
> IC.perc.par <- c(dist.boot.or[k1], dist.boot.or[k2])
> IC.perc.par
[1]  8.474533 14.759127
> # IC  Básico de Bootstrap:
> alpha <- 0.05
> z <- qnorm(1 - alpha/2)
> dist.boot.or <- sort(dist.boot.par)
> k1 <- trunc((B+1)*(alpha/2))
> k2 <- trunc((B+1)*(1 - alpha/2))
> IC.basico.par <- c(2*est.boot - dist.boot.or[k2], 2*est.boot - dist.boot.or[k1])
> IC.basico.par
[1]  8.000873 14.285467
> # Amplitude dos IC's
> IC.padrao.par[2] - IC.padrao.par[1]
[1] 6.293471
> IC.perc.par[2] - IC.perc.par[1]
[1] 6.284594
> IC.basico.par[2] - IC.basico.par[1]
[1] 6.284594
```

Referências

FERREIRA, D. F. Estatística computacional em Java. Editora UFLA, 2013.